Optical IP Switching

Marco Ruffini

A thesis submitted to the University of Dublin, Trinity College

for the degree of Doctor of Philosophy

June 2008

Computer Science Department
University of Dublin, Trinity College
IRELAND

Declaration

I, the undersigned, declare that this work has not previously been submitted to this or any other University, and that unless otherwise stated, it is entirely my own work.

Dated: June 14, 2008

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Dated: June 14, 2008

Summary

Improvements in the optical transmission technology, over the past fifteen years, have substantially facilitated the development and worldwide deployment of the Internet, by reducing the cost of data transport. The exponential growth in Internet traffic however, demands new solutions at the routing layer, as current IP routing technology struggles to deliver the necessary bandwidth at competitive costs.

Hybrid electro-optical architectures, where dynamical optical circuit switching is combined with legacy packet routing, have been introduced in the past few years as promising solutions to reduce costs at the IP layer, and to deliver new revenue-generating services and applications. Most of the architectures currently proposed however have focused on end-to-end lightpath provisioning, coordinated through a centralized management plane.

In this dissertation we propose Optical IP Switching (OIS), a hybrid electro-optical network architecture that combines IP routing and wavelength switching, using a distributed decision-making process. Each node constantly analyzes the IP traffic, and uses the information collected to adapt the optical network topology to the variable traffic demand. The main advantage of such distributed approach, besides the improved resiliency and scalability, is its close affinity with the interdomain structure of the Internet.

The simulation analysis we have performed, using real traffic traces and routing tables, emphasizes the efficacy of Optical IP Switching in redirecting IP traffic to dynamically created optical cut-through paths, considerably decreasing the network load on electronic routers. Such approach substantially reduces the capital expenditures for network operators.

We have also implemented the OIS model in prototype nodes, using inexpensive, commercially available components. We have proved, through the testbed we have set up, the feasibility of the OIS concept and the efficacy of the auto-configuration methods we have developed. Our testbed trials, in addition, showed that the impact of dynamic optical switching on the Internet transport protocols can be effectively minimized.

Acknowledgements

I am sincerely thankful to my supervisor, Professor Donal O'Mahony, for the guidance he provided me throughout my postgraduate work. His insight and wide knowledge were invaluable in keeping my aim focused in the right direction, while his very genuine interest in research was all along a great stimulus for my work.

I thank Linda Doyle for her overall support, for helping me set my initial ideas on the distributed and Ad-hoc network environment and above all for the great enthusiasm she showed while directing the "Emerging Network" group, thus creating a very pleasant working environment. I also thank all my colleagues, in particular Patroklos and Rob, for their assistance with the protocol stack and Tim for proof-reading this dissertation.

Many thanks to Dan Kilper from Bell Labs, for his hospitality during my visit to their laboratories and the valuable suggestions and discussions he provided during our joint work on network cost modeling.

I thank HEAnet for providing us with the optical links we have used for our demonstrations, and Eoin Kenny and Frank Smith for the help provided in setting up our testbed.

I thank Angel Sanchez and Sergi Figuerola from i2Cat, and Jonathan Mahady, for their cooperation in the development of the OIS platform integration with UCLP.

I thank Bruno Quoitin and Steve Hulig from University of Louvain-la-Neuve for providing the GÉANT dataset, and assisting me in its use.

Hearty thanks go to all my friends here in Dublin: their encouragement has been a great relief during the hard times and have made my stay in Ireland a very pleasant experience. I acknowledge in particular Giuseppe and Laura for our philosophical lunch-time discussions.

Very special thanks go to Natasha, for her immense support throughout this period and her deep understanding and patience, especially when I placed a greater emphasis on my work than my life.

Finally, I dedicate this thesis to my parents Bruno and Maria. Their full understanding and continuous, unconditional support have totally compensated for the long distance that separates us.

University of Dublin, Trinity College
June 2008

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the past fifteen years we have seen the Internet evolving from a niche network connecting academics and research institutes to the main global information and entertainment network. Such exponential evolution was made possible by the development of Wavelength Division Multiplexing (WDM) and Erbium-Doped Fiber Amplifier (EDFA) technologies, which, over ten years ago, started a deep revolution in optical networking. Combined together, in fact, they increased the overall bandwidth availability and drastically reduced the cost of data transfer, creating an ideal environment for the mass proliferation of the Internet.

As a result of such dramatic progress, existing applications, that have conventionally been delivered over different networks, are being re-developed to be implemented over the Internet (this is the case for VoIP and IPTV), while new applications continue to emerge (peer-to-peer networking and interactive gaming, to mention only a few). The increasing exploitation of the Internet has provoked an exponential increase in the bandwidth requirement on the underlying network architecture, a trend which is further emphasized by near-future applications (e.g., high definition IPTV and e-science applications).

Although the WDM/EDFA evolution has made the optical transport layer capable of supporting the amount of data generated by these applications, such dramatic progress has not yet occurred at the routing layer. As a result the network bottleneck has moved from the optical transport to the routing layer, as conventional electronic routers do not seem capable of offering a cost-effective solution to the increasing bandwidth demand.

Over the past decade, the increase in bandwidth requirement has been tackled by router vendors by increasing the speed of the electronic processing units, following the evolution predicted by Moore's Law. However, since the traffic exchanged in the Internet has grown at a much higher pace compared to that of transistor integration, we see today that routing equipment has reached a very high cost, footprint and power consumption per gigabit of data.

The idea of packet routing was originally developed to achieve high utilization of the transmission channels, through the use of effective statistical multiplexing. Because of the reduction in the transmission costs however, the influence of the transport channel efficiency on the overall network cost has considerably decreased. It is common belief that, under these circumstances, higher cost savings can be achieved by reducing the amount of electronic routing equipment in the network, redirecting part of the routing load towards an optical switching layer.

In the past few years the research community has focused on this direction, by developing hybrid electro-optical architectures aiming at effectively bridging the gap between optical transport and electronic routing. On a parallel track, the development of all-optical routers has also been considered, but the premature state of development of basic processing and buffering optical units makes its practical implementation several years away.

Hybrid electro-optical solutions proposed over the years generally fall into two major categories: in the Optical Circuit Switching (OCS) approach, optical lightpaths are dynamically setup when needed, to transport data transparently between any two nodes in the network. In the Optical Packet Switching (OPS) approach, data is routed packet-by-packet; the router processes the packet header electronically, determining the next hop node and activating the optical switch to route the packet payload in the optical domain. Other solutions (like for example optical burst switching) fall in between these two categories.

The work we present in this dissertation relates to optical circuit switching. We propose Optical IP Switching (OIS), a hybrid electro-optical routing architecture, which combines an enhanced electronic router and a transparent optical cross-connect, to dynamically create optical paths, redirecting part of the IP traffic load to the optical layer.

Although, as previously stated, maximum exploitation of transport channels has ceased to be the only relevant factor for a cost-effective architecture, a well balanced solution needs to optimize the tradeoff between the amount of data that can be switched optically and the channel usage. This tradeoff originates from the difference in granularity between electronic routing, where data is switched packet-by-packet (each in the order of the Kbyte in size), and wavelength switching, where data is switched at the channel rate (of the order of a few Gbps). We challenge this large gap (about six orders of magnitude) by first reorganizing the packets in IP flows, decreasing the granularity to values between hundreds of Kbps and few Mbps. We then aggregate the flows sharing a common route into the same dedicated optical cut-through paths, using a method we have developed, that groups flows depending on their destination network. These two operations, combined together, are capable of effectively bridging the packet and wavelength switched domains, reducing the granularity gap down to about one order of magnitude. The path creation algorithm finally selects the flow aggregates eligible for dedicated cut-through paths, taking into consideration the aggregate data rate, the resources available, the network policies of its own domain and, in the case of interdomain operations, the network policies of its neighboring domains.

One of the main novelties of the OIS approach is that path creation and cancellation are operated in a completely distributed fashion, following the analysis of local traffic at each node. This distributed approach, besides having advantages in terms of scalability and fault-tolerant design, is highly beneficial in the interdomain, where each autonomous system makes decisions following its own network policies. We believe that the current interdomain structure, which has provided the Internet with the heterogeneity and ubiquity that have contributed most to its exponential deployment, is the cornerstone that will persist all along its near and far-future evolution.

1.1 Key contributions

The key contribution of this work is the introduction of the Optical IP Switching architecture as a solution to the bottleneck at the IP layer in existing point-to-point routing systems. The most significant contributions on the design of our dynamic optical router architecture are:

- The development of methods and algorithms for distributed creation, extension and cancellation of lightpaths, that allow each node to provision paths automatically, based both on local traffic observation and individual domain policies.

- Our method for aggregating individual IP flows into larger categories, through a selection mechanism that avails of the network prefixes stored in the IP routing table.

- A signaling framework for the network nodes that supports and enables the dynamic creation, extension and cancellation of lightpaths.

- A port and link discovery mechanism that enables full self-configuration of each OIS node, providing the system with plug-and-play capability.

Other contributions are relevant to the techno-economical analysis of the overall OIS implementation:

- We have characterized the architecture performance simulating a real network scenario, using traces and routing tables collected from a production pan-European research network.
- We have studied the network economics of dynamic optical networks and implemented a cost model that shows the cost saving potentials of the approach we propose.
- We have developed functional OIS prototype nodes using off-the-shelf hardware, that allowed us to attest the practical feasibility of the OIS architecture and characterize the dynamics of the lightpath provisioning mechanisms.

The overall contribution of this dissertation is to propose, analyze and demonstrate the idea of dynamic and distributed provisioning of lightpaths based on local observation of IP flows; optimization of the individual algorithms and protocols is instead out of the scope of this work.

The Optical IP Switching architecture we present is therefore fully open to future developments.

1.2 Dissertation outline

This dissertation is structured as follows. In this introductory chapter we have defined the major issues in the development of future Internet architectures, defining the scenario around which we have developed our work. We also have shown how this work contributes to the current research activity in optical networking and design of future router architectures.

Chapter 2 offers an overall background on optical networks, describing both legacy and state of the art technologies. In particular it examines the latest developments in the optical control plane and its application in traffic engineering, offering a solid background for the development of our work. It also attempts to extrapolate, by analyzing the outcome of past research projects, the evolutionary line that has driven the research activity up to date, and to foresee, by looking at current projects, its future development.

Chapter 3 describes in full the Optical IP Switching network architecture we have developed. We describe in detail the algorithms and functions we have implemented for traffic analysis, path provisioning and node self-configuration. Finally we describe the impairments that dynamic provisioning of optical paths generates both at the transport protocol level and at the physical layer. Throughout this chapter we formulate research questions, derived from different aspects and functions of our OIS architecture. Chapters 4 and 5 will provide the answers to such questions both through simulations and testbed results.

Chapter 4 reports on the simulations we have carried on to evaluate the performance of the OIS algorithms and architecture, using real traffic traces and routing tables from the pan-European research network GÉANT. The chapter is divided into two parts. In the first part we analyze the technical performance of OIS, focusing on its ability to bypass the IP layer with dynamic optical cut-through paths. We analyze the impact that different network topologies have on the overall performance, and provide a comparison between OIS and a GMPLS-based end-to-end provisioning model, which operates on a centralized decision mechanism. In the second part we provide the results of the cost study we have performed, which emphasize the economical advantages of both optical bypass of the IP layer and dynamic reconfiguration of lightpaths.

Chapter 5 describes the testbed implementation of the OIS architecture. It illustrates the tests we have performed to demonstrate the practical feasibility of OIS, it analyzes the performance of the self-configuration mechanism we have developed, and studies the effects of dynamic switching of lightpaths on the Internet transport protocols. We also report on an integration project we have carried out with the UCLP bandwidth on demand platform.

Chapter 6 draws the conclusion of our work on Optical IP switching, summarizing the contributions of this dissertation. In the last section of this chapter we define some aspects of OIS that could be further developed in future research activities.

1.3 Publications arising from this work

- Marco Ruffini, Dan Kilpler, Donal O'Mahony, Linda Doyle. "Cost Study of Dynamically Transparent Networks". *OSA Optical Fiber Communication Conference*, OMG2, 2008.

- Marco Ruffini, Donal O'Mahony, Linda Doyle. "Automatic Configuration in Future Semi-Transparent Optical Network Nodes". *Springer Photonic Network Communications Journal*, vol. 14, no. 3, pp 241-251, 2007.

- Marco Ruffini, Donal O'Mahony, Linda Doyle. "Dynamic Optical Path Allocation in Multi-Layer Traffic Engineering". *In proceedings of the Workshop on Traffic Engineering in Next Generation IP Networks, IEEE International Conference on Communications*, pp 11-12, 2007.

- Marco Ruffini, Donal O'Mahony, Linda Doyle. "Optical IP Switching for dynamic traffic engineering in next-generation optical networks". *In proceedings of the conference on Optical Networks Design and Modeling*, Springer press, pp 309-318, 2007.

- Angel Sánchez, Sergi Figuerola, Gabriel Junyent, Eoin Kenny, Victor Reijs, Marco Ruffini. "A user provisioning tool for EoMPLS services based on UCLPv1.5". *In proceedings of the TERENA Networking Conference*, 2007.

- Marco Ruffini, Donal O'Mahony, Linda Doyle. "Feasibility of Flow-Based Optical Provisioning in GÉANT". *In proceeding of the OSA Optical Fiber Communication conference*, OWK3, 2007.

- Donal O'Mahony, Marco Ruffini. "Optical IP Switching - A Responsive Solution for Grid Interconnect". *In proceedings of the International Workshop on Autonomic Grid Networking and Management*, 2006.

- Marco Ruffini, Donal O'Mahony, Linda Doyle. "A cost analysis of Optical IP Switching in new generation optical networks". *In proceedings of the IEEE Photonics in Switching conference*, 2006.

- Gavin Mulvihill, Marco Ruffini, Frank Smith, Liam Barry, Linda Doyle, Donal O'Mahony. "Optical IP Switching a Solution to Dynamic Lightpath Establishment in Disaggregated Network Architectures". *In proceedings of the IEEE International Conference on Transparent Optical Networks*, pp 78-81, 2006.

- Marco Ruffini, Donal O'Mahony, Linda Doyle. "A Testbed Demonstrating Optical IP Switching (OIS) in Disaggregated Network Architectures". *In proceedings of the IEEE conference on Testbeds and Research Infrastructure for the Development of Network and Communities*, 2006.

Chapter 2

Optical Networks

2.1 Introduction

The first electrical telecommunication network was built in 1839 to support the telegraph system, while initial telephone services started to appear in 1878, with calls manually switched by operators and different channels transported on separate wires. A major evolutionary step was the introduction in 1938 of the electromechanical crossbar switches, which automated the switching process, and the use of frequency division multiplexing at the transport layer. As the silicon industry developed, the use of digital processors allowed voice signals to be converted from analogue to digital and also enabled the use of electronics in switching. At the same time, channels started being multiplexed in the time domain, originating the first commercial T-carrier services in 1962. The first digital system was the Plesiochronous Digital Hierarchy (PDH), characterized by non-perfect synchronization among the network elements. As electronic circuitry became more precise and atomic clocks could be used for tighter synchronization, the Synchronous Digital Hierarchy (SDH) was developed. The precise synchronization across an entire national network allowed more flexibility in the add/drop signal operation, and also greatly improved the interoperation among different countries. In the meantime optical communication became wide-spread, substantially increasing the bandwidth available and the quality of transported signals.

Although the technology and services of telecommunication networks had greatly improved, the operation mode followed the circuit-oriented approach, where a dedicated path was setup before the communication could be established. This modus operandi was suitable for voice services (a telephone call maps easily into a dedicated point-to-point electrical circuit).

Problems began when the telephone infrastructure started to be used for data communication between computer networks. Because of the bursty nature of computer generated traffic, data exchange time became comparable with circuit setup time, making the circuit-switching approach quite inefficient. Although, initially, research efforts in the area aimed at increasing the speed of circuit provisioning, Baran [Bar64] and Davies [DBSW67] (independently of each other) proposed a novel approach that completely revolutionized computer networking. The two scientists invented packet-switching, a network approach where information was transmitted in packets which carried the destination address in their header. No pre-established circuit was required as packets could be routed towards their destination with a distributed hop-by-hop forwarding mechanism.

Today's data networks are built upon protocols specifically designed for packet-switching and bursty traffic (IP, Ethernet and Frame Relay, to mention only a few). The optical transport layer however (especially in the MAN and WAN) is still mostly circuit-oriented. As data traffic has exponentially increased and has already exceeded traffic generated by voice services, the mismatch between packet and circuit-oriented architectures has become an important issue.

These issues and their implications in current optical networks are further elaborated in the next sections.

In section 2.2 we discuss legacy and state of the art layer 1 and 2 optical technologies. We first describe how legacy Synchronous Optical NETwork (SONET) and Synchronous Digital Hierarchy (SDH) networks have tackled the mismatch between packet-oriented and circuit-oriented data transport. We then present state of the art transport networks aiming at replacing the existing optical infrastructure.

In section 2.3 we address the issue of dynamic circuit provisioning, discussing current standardization efforts in the optical control plane for the development of next-generation reconfigurable optical networks. The expected economical impact of dynamic network reconfiguration is analyzed in section 2.4, where we focus our attention on its effects on revenue generation, operational and capital expenditures.

In section 2.5 we introduce the concept of traffic analysis and network engineering, and illustrate the multi-layer approach to traffic engineering allowed by next-generation reconfigurable networks.

Section 2.6 finally describes some relevant projects developed during past and current research activities. We first show how the research trend has evolved over the past few years.

We then attempt to extrapolate, from some of the major upcoming projects, a raw guideline for the near future research activity in the area of highly dynamic optical networks.

2.2 Optical Networks

2.2.1 Legacy SONET/SDH optical networks

Currently, Synchronous Optical NETwork (SONET) and Synchronous Digital Hierarchy (SDH) are the most widespread optical layer architectures. The SONET standard is defined by ANSI and used in North America, while SDH is an international standard defined by CCITT/ITU-T. Even though the two are similar and based on the same principles, they are not fully compatible; interworking is allowed for traffic transport but generally not for performance management. Many resources already exist that explain in detail the framing and structures of these optical signals [BRS03]. For our purpose, we will briefly remind the reader that these protocols are based on electrical multiplexing of signals in the time domain; smaller circuits can be aggregated into progressively larger ones and transported over a backbone network.

While this type of architecture is fully compatible with voice traffic, the transport of packet-based data is not very efficient and needs to go through link and packet-to-circuit adaptation mechanisms.

Packet-over-circuit adaptation mechanisms

The adaptation of the packet switched IP traffic to the circuit-oriented SONET/SDH optical transport network involves multiple intermediate layers. Figure 2.1 shows how an IP layer can be matched to a SONET/SDH optical transport layer.

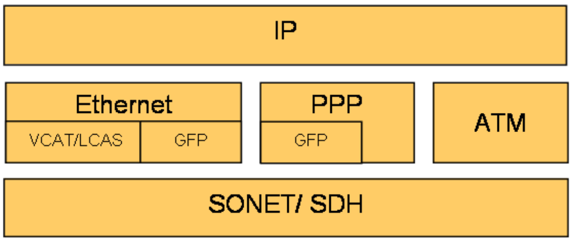


Figure 2.1: Matching between IP packets and SONET/SDH circuits

This can be achieved, for example, using a Point-to-Point Protocol (PPP) [Sim94], which simply creates a static end-to-end channel between two routers. The main problem of this solution is that it does not scale very well; the minimum allowable bandwidth is an STS-1 channel (∼52Mbps) and further allocations are either STS-3, STS-12, STS-48 or STS-192 (with bandwidth respectively of about 155Mbps, 622Mbps, 2.5Gbps and 10 Gbps).

Another possibility is to use an intermediate Asynchronous Transfer Mode (ATM) layer; the organization of dedicated circuits into Virtual Paths and Virtual Circuits allows a much finer bandwidth granularity. Implementing IP over ATM does, however, have some drawbacks, such as: bandwidth inefficiency of the ATM encapsulation process, broadcast not supported by ATM and the use of an additional addressing system ([Dan02, BRS03]).

An alternative solution uses an Ethernet layer directly on top of SONET/SDH. Initially this mapping was very inefficient ([Dan02]) due to the difference in the data rates supported by the two protocols (see Table 2.2).

Ethernet	*SONET*		*Efficiency*
Rate	Channels	Rate	
10 MBPS	STS-1	∼50 MBPS	∼20%
100 MBPS	STS-3c	∼150 MBPS	∼67%
1 GBPS	STS-48c	∼2.5 MBPS	∼40%

Table 2.2: Efficiency of Ethernet over SONET implementation without virtual concatenation

The introduction of the Virtual Concatenation (VCAT) technique allowed a more efficient channel mapping of Ethernet over SONET (Table 2.3, [Dan02]). This technique, implemented at the terminal nodes, enables aggregation of more physical channels in a virtual channel, allowing addition of bandwidth in steps of 50 (SONET) and 150 (SDH) Mbps. Additionally, only the source and destination nodes need to be provided with equipment that supports this feature. The Link Capacity Adjustment Scheme (LCAS), where implemented, is associated with the VCAT scheme and is used to dynamically adjust the bandwidth provided by virtual concatenation without losing any data.

Ethernet	*SONET*		*Efficiency*
Rate	Channels	Rate	
10 MBPS	VT-2.0-5v	∼10.88 MBPS	∼92%
100 MBPS	STS-1-2v	∼96.77 MBPS	∼100%
1 GBPS	STS-1-21v	∼1.02G GBPS	∼98%

Table 2.3: Efficiency of Ethernet over SONET implementation with virtual concatenation

Another solution is the use of the Generic Framing Procedure (GFP) [IT03a] protocol to map Ethernet packets into SONET circuits. The main characteristics of GFP are: its capability to encapsulate both block-coded and packet oriented data stream, its multiplexing capabilities of data from different clients and its compatibility with frames and packets (PDUs) of variable sizes. The high flexibility and low level of complexity of this protocol allows cost reduction in the hardware implementation and increased data transmission speeds.

2.2.2 Optical Transport Network (OTN)

The recent general migration to IP/Ethernet applications has led to the standardization of the "Optical Transport Network" (ITU-T G.709 [IT03b]). The main driver of this technology was the need for a more general optical transport architecture capable of efficiently carrying different types of network services. Besides the transport of SONET/SDH signals, the G.709 payload can efficiently map ATM and GFP frames, and allows efficient transport of IP/Ethernet packets. The capability of carrying different protocols helps to reduce capital and operational expenditures for network operators, as different types of traffic and services can all converge into the same optical transport network ([Bar03]).

One of the most important characteristics of OTN is the implementation of "Reed-Solomon" block coding for forward error correction. As the capacity of optical channels grows towards 40 Gbps bit rates, the optical signals suffer from degradation that depends on physical parameters of the optical fiber. Under these circumstances error correction decreases the BER of the transmitted channel, increasing the reach on long-haul links.

The G.709 standard also provides connection monitoring services. Being oriented towards the multiplexing of different types of protocols in fact, OTN offers monitoring activities for channels that, like Gigabit Ethernet or 10G Ethernet, do not yet provide efficient Operational Administration and Maintenance (OAM) capabilities.

2.2.3 Optical Ethernet

The IEEE 802.3 protocol standard (universally known as " Ethernet") was initially deployed in 1983 as a medium access control protocol for statistical multiplexing of different users' traffic onto a shared transmission medium. For several years its use has been restricted to the Local Area Network (LAN) environment as a user access technique and typically associated with IP traffic. As its transmission rate has reached 1 Gbps with the Gigabit Ethernet (IEEE 802.3z - 802.3ab) and in 2003 10 Gbps (802.3ae), its domain has expanded well beyond the LAN, first entering the Metropolitan Area Network (MAN) and then undergoing the necessary standardization activities to make it operational also in the Wide Area Network (WAN) environment.

Although Gigabit Ethernet can still be deployed over copper cable, the standardization of the fiber interfaces has allowed it to reach distances of tens of kilometers without amplification at bandwidths up to 10 Gbps, extending its operational domain from the LAN to the MAN.

The main reasons that motivated providers to extend the use of Ethernet beyond the LAN are: its packet-oriented operation mode, suitable for IP applications; its simplicity, with no need for global synchronization; and its low cost, achieved through an extremely wide market share ($ 5.9 billion sales in 2005 and estimated $ 22.5 billion in 2009 [Net07]). It is estimated that currently 97% of Internet traffic makes use of Ethernet interfaces.

Ongoing standardization of the Ethernet protocol focuses on the switching layer, with the Provider Backbone Bridging (PBB, IEEE 802.1ah standard), and on the Operations, Administration and Maintenance (OAM), which are covered by different standards. Moreover, Nortel is current developing and bringing into the standardization process the Provider Backbone Transport (PBT).

The aim of the OAM standardization for Ethernet is to match the monitoring and resilience capabilities that are currently deployed in SONET/SDH networks: IEEE 802.1ag is addressing Connectivity Fault Management, 802.1AB the topology discovery, ITU G.8031-SG 15 the Ethernet protection and the ITU-Y.1731-SG13 performance monitoring.

The Provider Backbone Bridging is an extension to the Virtual LAN standard (VLAN, IEEE 802.1Q) that increases the system scalability, creating a dual-hierarchy network model. VLANs are logical partitions in a local area network, which separate traffic belonging to different areas that share a common physical LAN infrastructure. Although the VLAN standard is appropriate for the LAN, the single level of hierarchy and the 12-bit label space of the LAN ID identifier (which would not allow more than 4096 distinctly addressable customers), make it unsuitable for the MAN environment. The PBB standard addresses this issue by creating a hierarchical model for the virtual LAN. On top of the MAC address and VLAN ID identifying the customer, PBB inserts an additional MAC address and VLAN ID to the protocol header, which identify the service provider. By stacking the provider header on top of the customer one, the provider can create tunnels across its network, completely masking the customer addresses. Such a mechanism simplifies the switching core equipment, which only processes the higher-hierarchy MAC header of the provider, while the information concerning the customers' MAC address and VLAN ID is only processed at the edge nodes.

The Provider Backbone Transport modifies the connectionless Ethernet into a connection-oriented protocol, providing it with deterministic Quality of Service (QoS) and traffic engineering capabilities. PBT on the one hand introduces virtual paths, similarly to ATM networks, identified by labels obtained using the bits of the VLAN identifiers. On the other hand it disables the automatic MAC learning mechanism, which would otherwise create scalability issues beyond the LAN. MAC destination addresses are in fact learned by the switches automatically through a flooding mechanism; when an Ethernet switch receives a packet destined to an unknown destination, it broadcasts the packet over every network segment. With the MAC learning disabled and the capability to create end-to-end virtual circuits, Ethernet nodes become capable of switching data using forwarding information externally provided by the management plane.

These enhancements to the protocol on the switching, transport and resilience domains constitute the foundations for the development of Ethernet in the MAN and WAN environments.

2.3 Circuit Provisioning

Traditionally, the provisioning of new optical paths over SONET/SDH networks is a lengthy process. In fact the time needed to setup a 155 Mbps (OC-3) or 622 Mbps (OC-12) optical connection can be up to six months in many locations [AM01].

In locations where there is insufficient capacity installed, the laying of a new optical fiber would require a set up time much longer than six months. Fortunately this is not the usual case, as the advances in Wavelength Division Multiplexing mean that the fiber already in place can accommodate many times more bandwidth than it previously did. However, the difficulty that network providers have is to coordinate and organize their own network elements (not to mention the problems in organizing a connection through different operators).

Current SONET/SDH networks do not have the intelligence necessary to speed up the creation of new links or optical paths. The lack of automated neighbor discovery and signaling support for creation of on-demand physical paths make updating a connection a slow and expensive process.

This problem is very well-known both to the network customers and providers, and is experienced by the former as a delay in fulfilling their bandwidth ambitions and by the latter as a loss of new revenue opportunities. The need for a more dynamic network environment became particularly evident as the world passed from an old economy model, based on production and consumption of physical objects, to the new economy, founded on the information and service sectors. The Internet in particular has played a primary role in the development of the new economy, passing from a network connecting academics and research institutes to a powerful interactive mass media generating a turnover of several billions of dollars every year [Bur06]. Under these circumstances, the idea of a dynamic network environment capable of supporting novel services, like deterministic quality of service requirements and bandwidth-on-demand, becomes a major milestone towards the evolution of our service-oriented economy.

2.3.1 Dynamic circuit provisioning

Next generation optical networks need to support functions like quality of service delivery and dynamic bandwidth allocation in order to provide the desired level of dynamicity.

Circuit provisioning can be implemented at different layers of a protocol stack. Depending on the type of service, location and bandwidth required, provisioning can be operated at the following levels: at layer 3 through MPLS paths; at layer 2 through VLAN or more recent layer-2 switched circuits; and at the physical layer with wavelength, waveband or fiber switching. In this section we will address the provisioning at the optical layer, while the higher layers will be addressed in section 2.5 on Traffic Engineering.

Dynamic provisioning of optical paths gives the network the capability to allocate new lightpaths connecting two arbitrary nodes belonging to the same operator (intradomain) or to different operators (interdomain path allocation). Its primary use is currently to satisfy large bandwidth requests from selected customers or to implement topology updates following a traffic engineering analysis.

One of the main reasons for establishing new optical paths in a network is to increase the capacity whilst minimizing the need for additional electronic routing and switching equipment; a lightpath can be switched transparently through different nodes, allowing the bypass of intermediate electronic layers. A more detailed description of how this is achieved will be provided in chapter 3, section 3.3.

The economical advantage is that transparent switching can be orders of magnitude cheaper than routing at the device level (e.g., comparing a 10Gbps IP port to an optical port on a Micro Electromechanical Mirror based switch). However, several other issues must be taken into account to evaluate the technical and economical feasibility of all-optical switching.

In the following sections we discuss some of the most relevant issues: the physical impairments, the routing and wavelength assignment, the path protection and restoration, and the link discovery problems.

Physical impairments

A primary issue to consider in dynamic lightpath provisioning is that of physical impairments. Optical links are usually engineered to operate on distances and conditions established during the design phase. Transparently switching two (or more) optical links in cascade can force the channel BER below acceptable values.

Chromatic dispersion can be compensated for, without major impact, by dispersion compensating fibers.

Path loss compensation is instead achieved through Erbium Doped Fiber Amplifiers (EDFA) that, adding noise to the channel, increase the noise figure, limiting the total length of the lightpath. The use of Raman amplifiers can help to mitigate this effect, increasing the optical reach by a factor of four (but at a cost that is currently double).

Polarization Mode Dispersion (PMD), consisting of the two orthogonal polarizations of the signal propagating at different speeds, is another physical impairment that can create signal degradation, and is currently the main obstacle to 40 Gbps channel rates. PMD originates from asymmetries of the fiber core and is dependent on external phenomena, such as changes in temperature or mechanical stress on the fiber. These non deterministic phenomena are difficult to analyze and compensate for.

Other physical impairments arise from fiber nonlinearities, the most relevant being Stimulated Brillouin Scattering (SBS), Stimulated Raman Scattering (SRS), Four Wave Mixing (FWM), Self-Phase Modulation (SPM), Cross-Phase Modulation (XPM), and intermodulation (mixing). For a thorough description of such phenomena the reader can refer to [Agr95].

In order to assure adequate transmission performance, optical channels need to be regenerated before their BER goes below acceptable values (with a threshold of generally 10^{-9}). Signal regeneration is currently performed electronically, which means each wavelength needs to be demultiplexed from the fiber, converted to the electrical domain, processed electronically and converted back to optical. All-optical regeneration of multiple wavelengths without electronic conversion is subject to much research at present [TKT07, LLB^{+}03].

Another open issue of general concern for dynamic transparent networks is the *gain mod-*

ulation transient effect on EDFAs that generates channel cross-talk. The sudden addition or cancellation of a wavelength in a fiber generates a temporary gain imbalance, that increases (or decreases) the signal level of other channels, creating additional noise at the receiver end [Kam03, KKHR04]. This issue is of major concern for highly dynamic reconfigurable optical networks. Early solutions to the problem were proposed over the past few years, [JPL$^+$04, TP05], while this research field remains open for further improvements.

In section 3.10 we will continue this discussion on physical impairments, briefly considering possible solutions for the dynamic reconfigurable optical architecture we propose in this dissertation.

Routing Wavelength Assignment

The Routing and Wavelength Assignment (RWA) problem consists of finding the optimal route on an available wavelength to provision an end-to-end lightpath between two network nodes, and is generally recognized to be an NP-complete problem. Finding the optimal route depends on which parameters the network operator wants to optimize. An optimal path, for example, can be the shortest path in terms of physical distance or number of hops, or more generally as a combination of these, and should avoid using congested links. The wavelength assignment problem becomes challenging when we assume the wavelength continuity constraint, i.e. using the same wavelength along the whole path, which is usually considered to avoid the expense of wavelength converters.

The aim of RWA algorithms is to optimize the use of available network equipment to satisfy as many end-to-end optical connections as possible. The performance of these algorithms is evaluated in terms of call blocking probability, a measure of the inefficiency in exploiting all the wavelengths of a WDM system. In some cases the wavelength continuity constraint is partially relaxed; it has been demonstrated that blocking probability can be highly improved by placing a small number of wavelength converters in the network [IM99]. An exhaustive overview of existing solutions to the RWA problem can be found in [ZJM00].

RWA algorithms assign routes and wavelengths to incoming lightpath requests, assuming an ideal network with no constraint imposed by the physical layer.

Impairment-aware routing addresses a more realistic problem. Besides finding suitable routes and wavelengths, the algorithms make sure that the selected paths are compatible with the physical constraints of the selected links, delivering a BER-free channel. The presence of the additional physical constraints make impairment-aware routing even more challenging than RWA and is currently the subject of investigation of new optimal solutions [YHM05, PAMS06, MPC$^+$06, Muk06].

Path protection and restoration

Path protection is the capability of a network to switch to a pre-provisioned, dedicated backup path when the associated working path reports a failure. Protection requires redundant capacity on the network, which is only exploited when a failure occurs. Restoration is the capability of establishing a new path after detecting a link failure, to re-route the traffic affected by the failure. This techniques does not rely on pre-established redundant links, as traffic is re-routed over available resources.

Restoration is generally more cost-effective, as it does not require redundant dedicated bandwidth, but path restoration times are at least 200 ms (and usually higher than that). Protection on the other hand can rely on an extremely fast switching time ($<$ 50 ms), which can get around link failures without disturbing network flows (a feature that many customers request from their providers).

The issue becomes more complicated in dynamic networks, as a protection path needs to be dynamically calculated for each new lightpath provisioned. Moreover provisioning a backup path is more complicated than provisioning the original path. A protection path in fact is subject to additional "diversity" constraints in order to assure an appropriate level of resilience against different types of failures. For example, if the faulty channel is due to a defective device working on a particular wavelength, the channel can be protected using a different wavelength on the same fiber. If the failure is due to a fiber cut, the protected channel needs to be provisioned on a different fiber. If the failure is due to a natural disaster like an earthquake, which can affect a large area, fiber diversity does not guarantee the protection, and geographical diversity is required to provide adequate resilience. The type of diversity required can be agreed in advance with the operator in the Service Level Agreement (SLA), and is implemented through the Shared Risk Link Group (SRLG). The idea behind this concept is to assign the same SRLG membership to all the links that are subject to a common risk [SYR05]. Although within the same network domain, link diversity information can be exchanged freely, interdomain path protection is a greater problem [YMBS$^+$06], as most of the topology information is usually confined within each domain.

Another issue that strongly influences the protection cost is the type of protection. SONET/SDH networks often use "1+1" protection, where the signal is duplicated over the original and backup paths; when the receiver loses the signal from the interface on the original path it simply uses the interface linked to the backup path. The "1:1" mechanism is very similar, with the difference that information is transmitted over the protection path only after the fault is detected on the original link. Although more efficient in terms of hardware required, "1:1" requires some additional time to switch the signal to the protection path. With the "1:N" scheme, the same backup path is shared among N working channels, allowing a cost reduction proportional to N. Finally the "M:N" scheme is a generalization of "1:N", where M backup paths are used to protect N working trails.

Current SONET/SDH systems generally adopt the "1+1" mechanism. Although this guarantees very fast protection, it doubles the link cost. The implementation of more efficient "M:N" protection mechanism, together with automatic path restoration, will be allowed by dynamic reconfiguration capability in next generation networks (see section 2.3.2).

Link discovery

Another issue raised by highly dynamic and transparent optical networks is the link discovery, which consists of determining how WDM wavelengths and neighbors are connected to the transparent optical ports. In current opaque networks (e.g., SONET/SDH) link discovery could be easily implemented because the switches are able to terminate all the optical signals locally and so can be aware of which port is being used to send or receive messages. When the network nodes are provided with transparent optical ports however, things are complicated by the fact that the optical switch is completely transparent to the data passing through its ports; for economical and practical reasons only a limited number of ports can be terminated at the same time.

Although a trivial solution would see the manual entry of the details of the transparent ports in a configuration file, this solution is not scalable. A real scenario could involve nodes with hundreds of ports, tens of wavelengths and multiple neighbors, operating in a network subject to continuous reconfiguration. Manually configuring even a medium size network would require several hours of work, coordination of more people at different locations, and would be highly prone to human errors.

The Common Control And Measurement Plane (CCAMP) of the Internet Engineering Task Force (IETF) has addressed the problem in the Link Management Protocol (LMP) [Lan04], a part of the Generalized Multi-Protocol Label Switching (GMPLS) protocol suite (described in the next section). LMP describes the steps required for link discovery both for opaque and transparent optical switches. One of the main assumptions in LMP is that a bidirectional channel is already established between the two nodes initiating the configuration procedure. The procedure for configuring transparent switches consists of one node sending test messages to another node, which terminates each input port in turn until the test message is received. When a message is received a link adjacency is noted and the transmitter starts the same procedure on a different port. Considering that a node is generally connected to more than one neighbor, the procedure has to be repeated for every other neighbor.

In [ELW^{+}04] the authors recognize the importance of this problem in semi-transparent optical networks and suggest that the discovery procedure could be initiated in parallel with all the neighbors in order to run the discovery process faster. The procedure however still requires that a control channel is pre-established and that the nodes are manually informed whenever the discovery procedure needs to be initiated. Moreover the authors do not explicitly suggest an algorithm implementing the idea.

In section 3.8 we will propose a solution to the link discovery problem in semi-transparent optical networks that does not require a pre-established control channel. This procedure represents a step forward towards the complete automatic configuration of future optical networks. In our scheme, the control channels between two peers is in fact selected during the link discovery process and embedded in one of the active wavelengths linking the two nodes.

2.3.2 Optical Control Plane (Next Generation Networks)

In order to achieve automatic and dynamic provisioning of optical paths, solving the technical problems associated with physical impairments and wavelength-route selection, a dynamic optical network needs a control suite capable of organizing the exchange of information among the network elements.

The control plane is the set of entities that controls and organizes the links between the network elements, provisioning optical paths among the network nodes.

Currently in the Internet, the control plane of the IP layer is completely independent from that of the optical layer. This condition doubles the amount of information exchanged between the network nodes and prevents the network and link layers from using a common routing mechanism. Moreover, while the IP layer implements algorithms to discover neighbors and find paths automatically, the optical layer needs a large amount of human intervention for any type of topology modification. Even the most basic functions, such as link and neighbor discovery, are either manually implemented or depend on proprietary non-standardized protocols, which do not represent either a complete or interoperable solution. This approach proves very inefficient each time a circuit has to be implemented or modified in the network, cost and time of provisioning being the primary issues.

Network service providers and vendors have been working hard over the past few years to overcome this deficiency by developing a standard for the optical control plane, under the guidance of ITU-T, IETF and OIF. In [SRB03] the authors present a complete review of the standardization activity of these bodies, emphasizing the interoperability aspects of their work. All three entities have already produced consistent documentation, in the form of recommendations, request for comments and implementation agreements, while more work is currently in progress. We give here a brief analysis of their work and mode of operation.

ITU-T

The Automatically Switched Optical Network (ASON) [IT01] is the control plane architecture being developed by the study group 15 of the International Telecommunication Union - Telecom's standardization sector (ITU-T). Their aim is to develop a baseline for an optical architecture, starting with the requirements and defining the abstract components of the model and their interactions.

The ITU-T group, having a telecommunication-oriented background, has shaped its optical architecture more from concepts derived from standards such as SDH, ISDN and ATM than from the Internet protocols [Lar02].

The ASON architecture makes use of three federation models to classify switching and routing between nodes and subnetworks, where a federation is defined as a "community of domains that co-operate for the purposes of connection management". The joint federation model (Figure 2.2) is based on a hierarchical scheme, where a parent controller exists that coordinates the other controllers within its own domain. This scheme can be iterated inside each domain to achieve a higher number of hierarchies.

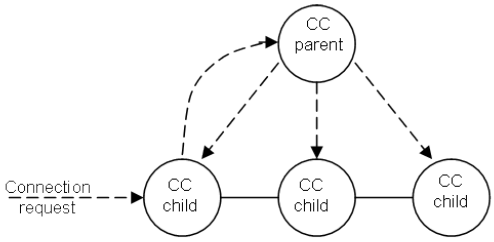


Figure 2.2: ASON - Joint federation model

In the cooperative model of Figure 2.3 no parent controller exists, and each controller can contact its direct neighbors to implement and organize the connection requested.

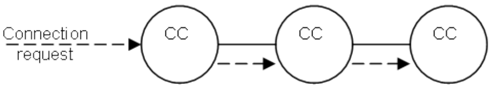


Figure 2.3: ASON - Cooperative federation model

There is also a combined model, illustrated in Figure 2.4, necessary to implement the joint federation model when there are multiple domains involved. In this case no general controller may exist that can coordinate the information of nodes spread over different domains. The highest levels of each domain use a cooperative model to exchange information between each other, while within each domain the hierarchical model applies.

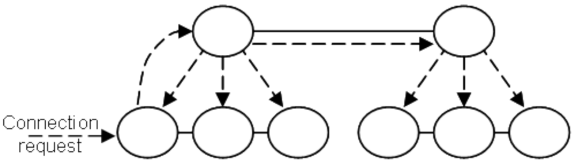


Figure 2.4: ASON - Combined federation model

The ASON architecture also defines the network components and their interaction. Setting up a connection through the network requires the interaction of three types of components: the connection controller, with the role of supervising the connection set-up, release and eventual adjustments of the connection parameters; the routing controller, which provides path information for connection and network management purposes; and the link resource manager, with the task of allocating and deallocating the links at the node and subnetwork level. The interaction between these components depends on the routing protocol implemented: hierarchical routing for the joint federation model, source routing for the cooperative model and step-by-step routing for the combined model.

Other components included in the ASON architecture have the task of supporting the network services. Among these, the traffic policing component assures the conformity between the agreed and actual traffic parameters. The call controller component manages the necessary associations between the end points to allow the utilization of the services provided by the network. Finally, there is a protocol controller component with the task of mapping the abstract instructions coming from a general controller into messages used by the communication protocol to exchange information.

The ASON recommendation aims at giving an architectural specification of the reconfigurable optical network, without going into the details of the protocol implementation. Such a task is coordinated by the Internet Engineering Task Force.

IETF

The Internet Engineering Task Force presents a more interactive approach in their GMPLS standard. In contrast to the ITU-T, where there is a designated group proposing and evaluating ideas for the architecture development, the IETF work group exchanges opinions with any Internet user who may want to actively participate in the discussions, through an Internet mailing list.

GMPLS [Man04] is derived from Multi Protocol Label Switching (MPLS) [Ros01], a technique that allows traffic engineering by creating switched virtual circuits through an IP network. In MPLS, each time a packet enters a label-switched path it is marked with a label (a packet header) that will indicate to the downstream neighbor to which direction the packet should be switched. After reading the label attached to the packet received, the downstream neighbor replaces it with a new label that indicates to its downstream neighbor how to switch the packet. The process goes on until the packet reaches the end of the switched circuit and is passed up to the IP layer. The practical advantages are that the limited label space allows faster switching and cost reduction at the hardware level, and that traffic engineering can be implemented without affecting the stability of the IP layer.

GMPLS generalizes this approach to support different types of switching, by allowing the label to take many different forms. Besides Packet Switch Capable (PSC) interfaces, GMPLS also introduces Layer-2 Switch Capable (L2SC), Time-Division Multiplex Capable (TDMC), Lambda Switch Capable (LSC) and Fiber-Switch Capable (FSC) interfaces. PSC concerns the switching of packets by reading information located in the packet header (e.g. IP/MPLS routers); L2SC refers to interfaces able to switch frames or cells (e.g., Ethernet, ATM switches); TDMC switches the data depending on an allocated time slot (e.g., SONET/SDH cross-connects); an LSC interface is able to switch an optical signal with wavelength granularity (e.g., wavelength selective switches); finally, FSC switches the entire data from a fiber using space division multiplexing (e.g., fiber switches).

The GMPLS signaling protocols RSVP-TE [Ber03b] and CR-LDP [Ber03a] handle the label distribution among the nodes to establish the switched paths.

LDP [And01], the Label Distribution Protocol, takes care of the label exchange between

Label Switched Routers (LSR), and was implemented to support the MPLS concept. Its extension, the Constraint-based Routed LDP (CR-LDP), allows the creation of traffic engineered paths; in this way the route between any two points in the network can take a different path than that obtained through legacy IP forwarding. Moreover this protocol can be used for resource reservation. The label switched path, for example, can be assigned a reserved bandwidth to accommodate traffic with specific bandwidth or time constraints.

RSVP [Bra97] (Resource Reservation Protocol), was developed initially for the standard IP network. The aim of this protocol was to allow IP routers to reserve bandwidth along the path for a particular IP flow (or class of flows). The Traffic Engineering extension (RSVP-TE) allows the protocol to exchange labels along the reserved path, supporting the GMPLS implementation.

The two protocols (CR-LDP and RSVP-TE) are different implementations of a similar idea. The main difference is that in the RSVP-TE the reservation request is made by the destination node. This approach may be better adapted to the Internet, where bandwidth constraints are usually on the network edge. This reason, together with the fact that RSVP is an already well-proved protocol, is probably the basis of the greater success of this protocol with respect to CR-LDP.

Among the other protocols of the GMPLS protocol suite, we wish to draw the reader's attention to OSPF-xTE [Kom05], the GMPLS Traffic Engineering extension of Open Shortest Path First (OSPF). By including in its header the necessary parameters for link and label-switched traffic engineering, this extension allows the selection of alternative routes that can be based on parameters different to the simple distance between routers.

Other relevant protocols include the Link Management Protocol (LMP) [Lan04] and its WDM extension [Fre05], which automate the link discovery process of the optical layer and allow the handling of a WDM link as a bundle. Since a WDM link can be made up of hundreds of channels connecting two points, the characterization of all or part of the channels as a unique bundle can reduce the signaling overhead of link and adjacency management.

Even though the GMPLS protocol suite is still a work in progress, a number of Request For Comments (RFC) documents have already been finalized and almost every vendor has implemented some of the specifications in their products. This positive attitude towards GMPLS is highly indicative of the beliefs that the network community has in the revenue generation potential of an automated control plane for service providers.

The signaling protocols developed within the GMPLS suite have generally the capability to operate both centrally and distributedly, but network management is still centrally operated. In this work we address the idea of distributed network management, where decisions are made distributedly at each node. The main advantages associated to distributed operations are: reduction of network traffic overhead, as traffic information is processed locally, higher scalability and a decision mechanism that is more compliant with the distributed nature of the Internet (especially for what it concerns interdomain operations). A centralized network management approach on the other hand can converge to solutions closer to the optimum (for example in traffic/network engineering), as the central management entity has a global view of the network resources. For such a reason a centralized approach might be favored in the intradomain, where node and link information is more likely to be shared among the nodes. It is possible to extend the view that a distributed management approach has about the global network by allowing nodes to forward update link and traffic information to each other. However, although this would allow better optimization, it would increase the network overhead, reducing the scalability advantages with respect to the centralized approach.

OIF

The Optical Internetworking Forum is a contribution-driven industry forum that encourages the cooperation between manufacturers, users and service providers. OIF is currently working on the specifications for the User-Network Interface (UNI) and Network-Network Interface (NNI).

The OIF is generally recognized as a merging point between the ITU-T and IETF works in the optical control plane. The main reason is that it implements concepts and protocols defined by both standardization bodies. Its user-network interface concept for example is drawn from the ASON architecture and from the ITU-T optical transport standards (like SONET/SDH); on the other hand, however, it uses GMPLS signaling protocols and implements an LMP-based auto-discovery mechanism.

The User Network Interface UNI [Jon04], defines the signaling specification between a user that requests a service and the network that provides it. The primary service it offers is the on-demand capability of creating, deleting and controlling optical connections. The control data between a UNI-C (on the client side) and UNI-N (on the server side) flows through a control link that can be an in-fiber or out-of-fiber channel, implemented either within a SONET/SDH link or through an external IP network. An important part of the UNI is the neighbor discovery procedure, which manages the links and control channels connecting to neighboring nodes. The service discovery procedure then helps the user understand which types of service the underlying transport network can provide and which attributes exist for the signaling and data channels. Support for SONET/SDH transparency and route/link diversity are examples of services supported, while reservation protocol used and type of link or concatenation implemented are examples of attributes.

The Network-Network Interface (NNI) defines the signaling specification between network elements and can be differentiated in E-NNI, used between nodes belonging to different domains, and I-NNI, used within a single domain. The main difference between the two is the amount of information shared between the nodes, which is more restrictive for the E-NNI. The E-NNI implementation agreement for intra-carrier signaling [Ong04] was officially published by the OIF in 2004. The specification of the I-NNI instead, being an internal interface, is left to vendors and service providers.

Working on the interface implementation, the OIF group is particularly active in practical demonstrations of the optical control plane. First initiated in 2004 their global interoperability demonstrations show practical implementations of the optical control plane, demonstrating the feasibility of their approach on a worldwide scale.

The 2004 and 2005 events brought together fifteen vendors and seven international carriers, connecting different laboratories located all over the world through trans-oceanic optical links. The 2004 event demonstrated the possibility of creating switched connections spanning multiple domains and triggered by users, employing the E-NNI and UNI implementations. The 2005 event focused on the adaptation of Ethernet services to SONET/SDH networks, utilizing the adaptation mechanisms described in section 2.2.1.

Those events represented the first real tests of the automatic control plane standardization and their success was essential to generate trust in an area that may have still been viewed with skepticism by many service providers around the world.

UCLP

The User Controlled Light Paths (UCLP) is a distributed network control and management system that allows users to self-provision lightpaths within a federation. UCLP is not part of the ASON - GMPLS standardization efforts, but is developed by independent organizations [CAN07], counting national research networks, communication research centers and universities.

UCLP can manage different types of routers and switches (SONET/SDH, Ethernet, MPLS or transparent optical cross-connects). Each physical port is identified as a Resource Object (RO), and the internal topology, linking the different ROs, can be set up by the user using the abstraction of Light Path Objects (LPO).

A running UCLP system is divided into federations. A federation is a logical partition composed by heterogeneous resources of different types of networks that are under the control of a single administrator. Once the Federation has been set up, the user can log into the UCLP server using a web interface and modify the network topology using a simple point-and-click approach.

The range of applications enabled by this platform is manifold, ranging from simple IPv4 routed traffic to e-science and grid applications.

2.4 Network economy

Feasibility and costs of dynamic circuit provisioning architectures are generally a function of the dynamicity we assume. For the applications and services we want to consider (for example lightpath provisioning for automated traffic engineering), path duration time could span from a few minutes to several hours. The exact value will depend on how variable is the traffic observed and how fast should the network react to such changes, originating a trade-off between routing efficiency and network stability.

The main reasons behind dynamic circuit provisioning in optical networks are the following. On the one hand it increases the network capacity, and supports novel revenue-generating applications (IPTV, bandwidth on demand, e-science). On the other hand it minimizes capital and operational expenditures, exploiting the economical advantage brought by transparent switching.

The economics of next generation optical networks have been extensively studied by researchers, which have focused their attention more on the Capital Expenditure (CapEx) analysis than on the Operational Expenditure (OpEx) and revenue generation. In the next sections we describe the various approaches that have been used, showing the advantage of transparent switching and analyzing the elements leading to such achievements.

2.4.1 Revenue generation

Aiming to increase their revenues by attracting new customers in a competitive environment, network operators are constantly looking for new services to offer to their clients. Leased lines and layer 2 virtual private networks are examples of the most successful services in terms of revenue generation that operators have provided to date. From this perspective, next generation optical networks are expected to provide faster and user-controlled access to those services, with increased levels of security and quality of service. In particular (transparent) wavelength services and optical grids are considered among the most relevant application in next generation networks.

From a consumer market point of view the convergence of many applications to the IP layer has also created new interesting opportunities for the operators. Although IPTV, targeting the mass market, might become a high revenue application, it is also envisioned as a killer application. Sending thousands (or millions) of high rate streams, generated by HDTV on-demand applications, implies provisioning of many dedicated peer-to-peer connections with guaranteed QoS. This situation creates new unsolved issues, posing colossal challenges to network designers.

Although user-on-demand end-to-end provisioning is generally implemented at the electronic layer, which guarantees finer granularity, optical provisioning is the key to support these layer 2 and 3 services at a large scale, providing fast upgrade of network capacity when existing links become congested.

2.4.2 Operational Expenditure (OpEx) savings

Economical analysis of OpEx for next generation optical networks is a field that lacks quantitative analysis in the literature. We believe the main reason is that calculating operational expenditures is a hard task as it includes many variables and parameters, like human resources, services and indirect costs [Mac06], which are difficult to collect and organize into a quantitative analysis. The general rule is that increasing the automation of the network decreases its operational expenses. Automatic fault location, for example, decreases the fault repairing time. User-controlled bandwidth-on-demand services speed up the provisioning process and reduce the need for manual intervention.

The improvement of operations facilitated by next generation networks is a topic that has recently been related [KIA$^+$05] to the value chain concept by Micheal Porter [Por85]. Applied to a network operator, the value creation can be identified as a two-step sequence: first, the transport network operator invests in the fiber infrastructure and sells coarse bandwidth to the carriers (as leased lines); then the carriers re-sell the bandwidth with finer granularities to end-customers or smaller carriers.

The primary activities of the value chain for network operators (following the observation in [KIA+05]) can be simplified into “Network Extension Process” and “Service Delivery Process”. The network extension concerns the planning and upgrade of the network capacity to support new requested services; it can be identified as the “Inbound Logistics” in figure 2.5. The service delivery instead covers the “Operations” and “Outbound logistics” primary activities. Service delivery operations are triggered by the sales department receiving the customer orders and sending them to the order management, which processes the orders and, if required, coordinates the activities with other providers (interdomain case). The delivery coordinator then proceeds with the actual planning of the physical path, which is finally revised and implemented by the network manager.

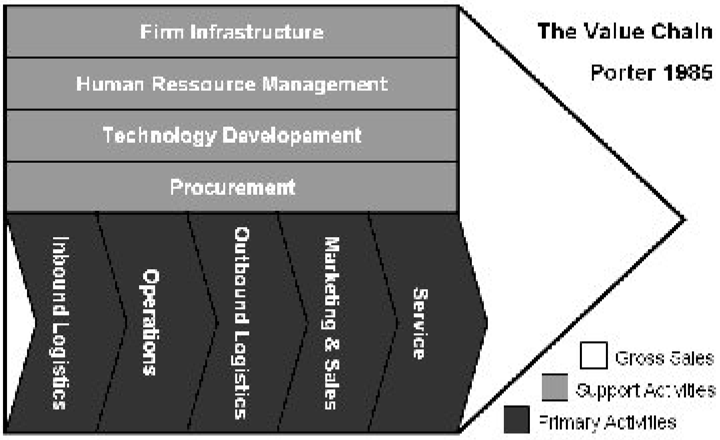


Figure 2.5: Micheal Porter’s value chain model

This value chain model can greatly benefit from an automated and standardized control plane. Considering the “Inbound Logistic” activity for example, the control plane, by providing advanced neighbor discovery, simplifies the network upgrade process, which becomes faster and requires less manual intervention. The standardized interfaces to the user (UNI described in section 2.3.2) allow a direct interaction between bandwidth supply and demand, increasing the user accessibility to the service and diminishing the expense for advanced over-provisioning of network resources (in [GSM05] we find a practical and quantitative analysis). The “Operations” activities are automatically handled by the GMPLS protocol that, making use of standardized Network-Network Interfaces (NNI), coordinates the interaction within a domain (I-NNI) and among other domains (E-NNI). The “Outbound Logistics” avails of similar benefits, as services are automatically delivered.

The sales process in general also benefits from automated operations, while marketing becomes more complex, as easier access to bandwidth increases competition among operators. Finally, the “Service” activity also benefits from the increased levels of resilience expected in next generation networks.

2.4.3 Capital Expenditure (CapEx) savings

In contrast to OpEx, savings on capital expenditures can be quantified more precisely as they directly relate to the cost of the network equipment needed to setup the network.

Many contributions modeling network capital expenditures for optical networks are available in the literature. They differ by topology considered, optical layer technology and type of resilience. In reconfigurable optical networks, the transparency advantage of optical switching over packet routing comes at the cost of a much coarser granularity (at the wavelength level), so that real economical advantage can only be achieved by efficient aggregation of packets into the dynamically created optical paths. In this section we introduce some of the most significant contributions in optical network cost modeling.

Although most recent work focuses on the advantages associated with optical transparency, some initial contributions [SKS03, VCPD04] showed how dynamic SONET/SDH or OTN switching could bring economical advantages to an IP-over-WDM network. The analysis shows that the benefits of optical switching are directly related to the amount of transit traffic in the network. Opaque optical cross-connects can switch transit traffic at a lower cost with respect to IP routing. However, because of the additional expenses of the switching equipment, the CapEx intersection point between the two models is only achieved when the transit traffic is higher than a certain threshold [VCPD04]. In [SKS03] the authors also consider the cost difference achieved by restoration schemes, considering MPLS restoration for the IP over WDM model and mesh restoration for OTN. Due to the lower transit traffic that hits the IP layer when restoration is operated at the optical transport layer, the cost advantages become even more noticeable for the OTN case.

The real economical advantage of optical switching however begins when we consider transparent switching, where the signal transits intermediate nodes without O-E-O conversion. Cost savings that transparent networks introduce are multifold and can be achieved at different layers. At the physical layer the number of O-E-O converters is highly reduced. At the IP layer the capacity required in the routers also decreases, as traffic is redirected towards transparent optical ports. In particular, if SONET/SDH switching allowed cost savings compared to IP routing, with a per-port cost difference of about three times, transparent switching can increase the per-port savings by tens of times (e.g., using a Wavelength Selective Switch, WSS). Optical transparency however, introduces issues such as physical impairments of the signal and an absence of traffic grooming, which radically change these values. In order to increase the reach of optical signals for example, allowing transparent bypass of a higher number of nodes, more expensive devices (like for example Raman amplifiers) need to be used. In addition, since traffic grooming is not allowed at the optical layer, channel occupation efficiency tends to decrease, and a higher number of optical channels might be required to satisfy the same traffic demand. For these reasons, for a better understanding of the actual benefits of transparent networking, cost analysis were carried out, which consider both benefits and limitations of transparent optical networking.

Network cost-optimization with physical impairments has been addressed in [BC06] and [Sim05], where the authors show the impact of the optical reach on the network cost depending on the average nodes distance.

In [JFN04] the authors compare the cost of an opaque network architecture against a transparent one for a European national network. The analysis, carried on for different levels of traffic, shows that the transparent architecture is more advantageous than opaque IP over WDM, and that the advantage increases with traffic. A similar analysis conducted over a US network [CY03], had previously shown that transparent architectures would only provide a modest economical advantage. The difference between the two models lies in the average link length. Networks characterized by smaller link distances (as is the case of European national networks, compared to a US network) can better exploit the advantages of transparent switching because the optical signal can be transported from source to destination without need for regeneration. Current regeneration systems in fact operate channel by channel, therefore their cost increases linearly with the number of wavelengths in a fiber.

In [FLM02] the authors make a more complete cost analysis of optical transparency, showing the higher revenue possibility allowed by transparent networks. Their analysis, comparing the network costs per Kbps of throughput, shows the net advantage of transparency also for low level of traffic. For yearly growth rates over 15% in user demand, the optical switched architecture is economically more advantageous because its high reconfigurability allows to accommodate a higher number of users.

This brief overview shows us where the economical advantages of transparent network architectures originate. The results discussed agree that the advantages are proportional to the quantity of transit traffic in the network. While the ports on optical switches are in fact bit-rate independent and their cost does not increase with the channel rate, the cost of the IP ports is proportional to the port rate. The average link distance also has a great impact on cost because transparency loses its advantages when regenerators, based on O-E-O devices, need to be used.

We envisage that in the future, the deployment of transparency in optical networks will be mostly determined by the cost evolution of optical reach and optical switching.

2.5 Traffic engineering

Traffic engineering assists the routing infrastructure to achieve more efficient utilization of the network resources. Communication networks need to be continuously monitored and updated in order to satisfy the increasing traffic demand and to avoid service denial and network congestion, which can arise because of hardware failure or traffic pattern variability.

Network engineering functions can be generally classified into three different categories, differentiated by their time-scale of operation. *Traffic management* operates on a day-by-day basis and involves tasks like QoS management, routing table management and dynamic routing. Its objective is to assure that short-term variations in the traffic patterns do not degrade network performances below acceptable levels. *Capacity management* involves routing design, bandwidth allocation and capacity design. It generally implies hardware upgrades and its operation timescale is in the order of several months. *Network planning,* finally, involves the re-arrangement of network nodes and transport links, requiring accurate planning for the updates in the physical topology, and is usually operated over some years.

Efficient management of these engineering functions has always been one of the main goals of any network architecture. Next Generation Networks (NGN), introduced in section 2.3.2, promise to deliver advanced mechanisms to facilitate traffic engineering operations.

In the remainder of the section we will address only those traffic management issues that are directly related to the Optical IP Switching architecture developed in this dissertation.

2.5.1 Traffic analysis

Traffic analysis is the first step in the whole traffic engineering process and is used to obtain a snapshot of the traffic condition at different nodes and links. Information can be collected and averaged over different time intervals, depending on the time granularity of interest. While, for example, traffic management would need information averaged over a few minutes, network planning would benefit more from information averaged over months, and predictions based over years of observation.

Traffic information is usually collected by the routers through traffic analysis tools, of which Cisco's Netflow [Sys07] is an example. Since analyzing packet-by-packet could become computationally very expensive, packet sampling is generally used. The IETF draft [DDC$^+$07] describes eight different types of sampling methods: systematic count-based sampling, systematic time-based sampling, random n-out-of-N sampling, uniform probabilistic sampling, property match filtering, hash based filtering using BOB, hash based filtering using IPSX and hash based filtering using CRC.

The difference between filtering and sampling is the following: filtering only considers a specific subset of packets with some defined property (e.g. packets with specific TCP port number), while sampling selects packets without considering any specific correlation [ZMD$^+$07].

Sampling in turn can be divided into systematic and random. Systematic sampling implies the use of a deterministic function to select the sample packet's population. Random sampling instead selects the sample population following a random-generation function. This eliminates the possibility of biased results, which can occur when using systematic methods. Both methods can be operated either in the space domain (e.g. select one packet in every n) or in the time domain (e.g. select a packet every x milliseconds).

Sampling methods can be fine-tuned by modifying their parameters. The most relevant is the packet sampling rate, which defines the average size of the sampled population with respect to the total number of packets. The sampling rate expresses very clearly the tradeoff between accuracy of traffic estimation versus processing power and memory consumed. A higher sampling rate for example will guarantee a more precise characterization of the traffic behavior, but at the same time will consume more computational and memory resources, which are generally very expensive in routers. Choosing an optimal sampling rate depends on available computational resources, on the traffic properties under evaluation and in general on the traces under analysis.

The information obtained from packet sampling can be used to monitor individual links or be collected together to build traffic demand matrices. In the first case, for example, an alarm can be generated in real-time when the link usage approaches the maximum link capacity, while a more thorough investigation can be carried out afterwards to reveal the causes of the problem. Traffic matrices instead give a more global view. Such information can be used to optimize the routing process, for example re-routing excess traffic through different paths to use links with more bandwidth available. Over longer time periods, traffic matrices are used for network planning, allowing off-line computation and simulations to determine how to efficiently distribute bandwidth along the network.

2.5.2 Flow characterization

Flow characterization can be considered a particular case of traffic analysis. Since it constitutes a fundamental block of the Optical IP Switching architecture we have included it in a separate section. Flow analysis categorizes the sampled packets into flows, grouping together packets with similar characteristics. They are generally categorized depending on the following five parameters: IP source, IP destination, transport protocol (e.g., TCP or UDP), source port and destination port. Flows can also be interpreted on a coarser granularity, aggregating all packets from the same IP address directed toward the same destination. At an even courser scale we can identify IP-prefix flows, where either the source or destination address (or both) is an IP prefix.

Flow characterization is particularly relevant because it represents a natural categorization of the route correlation property of IP traffic. For example, a TCP flow characterizes the traffic related to a single application running between two network terminals. An IP flow aggregates all the traffic generated by multiple applications between two network terminals. Finally, IP-prefix flows characterize all the traffic going from a network towards a certain destination node or network.

The Internet flow distribution has been, in the past decade, a topic of extensive analysis, which has revealed its "heavy tail" property [CAI03, SRS99, BC02, PTB+01].

A distribution function is heavy-tailed if it satisfies the following condition:

$$Pr[X > x] \sim x^{-\alpha} as\, x \to \infty,\ 0 < \alpha < 2 \tag{2.1}$$

i.e. if the asymptotic shape of the distribution is hyperbolic. In terms of traffic distribution, in order to better visualize the phenomenon, we can refer to figure 2.6 [MRS+06], which reports the results of a recent work we have conducted in our group. In this study we have examined recent Internet traffic traces taken from a trans-Pacific link on the WIDE backbone, available from the MAWI working group traffic archive [wgta05]. By comparing the red curve, indicating the flow distribution, the blue one, representing the data distribution, and the black line, representing the packet distribution, we can see that although only 1% of the flows contain more than ten thousand packets, they account for about 70% of the total traffic.

In other words, a very small number of large flows (elephants) carry most of the link traffic, while a large amount of small flows (mice) only account for a relatively small percentage of traffic.

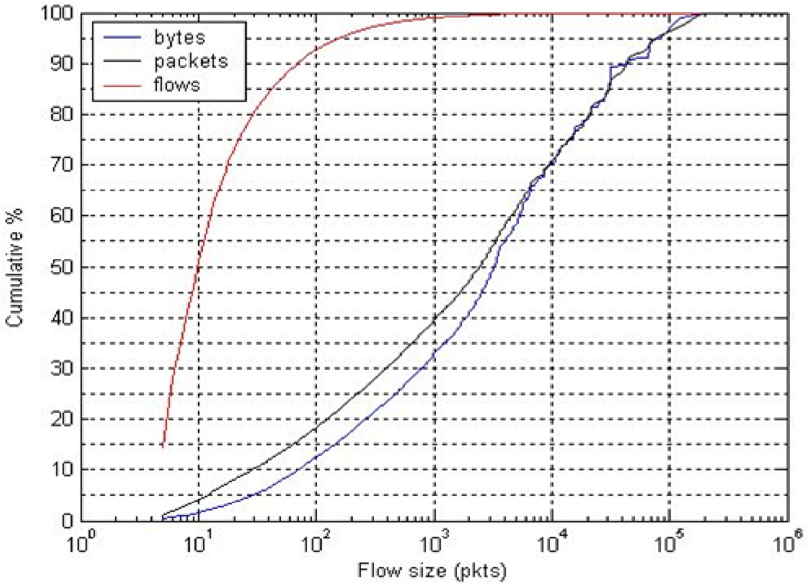


Figure 2.6: Cumulative data distribution for a trans-Pacific link

There have been some attempts in the literature to define a quantitative threshold to distinguish between elephants and mice. In [EV01] for example the authors fix the elephant threshold to 1% of the link utilization. In [SRB01] instead, an elephant is any flow whose peak rate exceeds the mean plus three standard deviations of the total link flow. We will not spend more time trying to define an absolute threshold between mice and elephants, as we believe that this boundary remains arbitrary, and an optimum definition depends on the type of traffic engineering being operated.

A more interesting approach was presented in [PTB$^+$01], where the authors focus on the methods to identify the threshold, rather than on its absolute value. They propose two separate methods based on a single metric, the flow bandwidth. In the first, they consider the heavy-tail property defined by equation 2.1, so that a flow is considered as an elephant if it is located in the tail of the flow distribution. The second method instead requires setting a parameter "α" indicating an arbitrary percentage of traffic that will be placed in the elephant class. Flows are sorted by size and added into the elephant class starting from the largest, until they account for α% of the total traffic. Their results show that the two methods give different threshold characterization behavior, with the former presenting a lower threshold value with respect to the latter. Their time dependency analysis shows that both the methods detect elephants with holding time between twenty and forty minutes. This implies that elephant detection needs to be performed very often (e.g. at least every twenty minutes). In order to target this high variability, the authors propose a “latent heat” algorithm that takes into consideration, besides the flow bandwidth, the traffic burstiness, averaging the flow rate over variable time intervals. This approach eliminates the flow reclassification due to transient traffic bursts, shifting the holding times towards values between one and two hours.

In [MUK$^+$04] the authors present a flow detection approach that considers the estimation error due to packet sampling. The algorithm they propose is based on Bayes theorem and assumes the knowledge a-priori of the traffic distribution. Their analysis shows the capability of the algorithm they introduce to effectively determine the elephant flows from the sampled population. Their main contribution is the observation that their threshold value is almost independent of the a-priori distribution considered, which makes the threshold invariant with respect to the network considered. They also identify 10^{-3} as an adequate sampling rate to identify elephant flows.

2.5.3 Flow-based traffic engineering

Conventional Interior Gateway Protocols (IGP), like OSPF and IS-IS, base their routing mechanism on the minimization of a single metric (e.g., number of hops or delay), whose value is usually established by the network administrator. A few years ago [FT00, AWK$^+$99] the idea of using variable metrics to make the routing protocol react to changing traffic conditions was considered, but it soon proved to be quite inappropriate. Fast changing metrics in fact created excessive control traffic overhead and route flapping issues, making the network unstable [EJLW01].

In the meantime a new idea was proposed, in [SRS99], to challenge this instability by applying the concept of elephant flows to dynamic routing. The relative small number of elephant flows in fact seemed particularly suitable to diminish the overhead needed for a flow-based traffic engineering approach. The large size of those flows would assure efficiency of such operations, while their long lifetime would decrease the update rates of link state information, diminishing routing table inconsistencies. According to the authors, the engineering process would only affect elephant flows, which could be re-routed towards paths with larger bandwidth available, while short-lived flows would follow default shortest-path routes. The authors compare their hybrid approach to ordinary static routing, where routes are not updated to reflect the changing network state, and to dynamic routing, where all the traffic is re-routed over periodically updated paths. The simulation results, focusing on the average network congestion, show that the flow-based approach has a clear advantage over the static approach and outperforms the dynamic approach for link-state update times over twenty seconds (i.e., for normal operating conditions).

What is clear from their study is that a dynamic routing algorithm needs very frequent updates of the link-state because it targets all traffic flows, including the highly variable mice. Short-lived flows are many and bring a relatively small percentage of traffic, which makes their engineering inefficient and difficult to implement.

Targeting the elephant flows instead increases the network stability, while keeping the number of dedicated paths relatively low. From a technical point of view, targeting elephant flows reduces the overhead by exploiting the route correlation of the IP packets. This is similar to how data compression algorithms exploit bit correlation to diminish storage space. It makes sense to consider a large stream of packets going from one source to the same destination as a unique flow rather than as a sequence of packets. This is the approach that we will follow in the design of our Optical IP Switching architecture.

Every approach that selectively adapts routes to traffic characteristics needs a mechanism to build and select alternative paths. This is the aim of the MPLS protocol, which allows creating dedicated paths without interfering with the original routing protocols.

2.5.4 Multi Protocol Label Switching

The MPLS protocol [Ros01] was originally conceived to allow faster switching of IP flows (compared to legacy IP routing) and to ease traffic engineering operations. Faster switching was introduced by aggregating packets into large flows; hardware speed could be achieved using a switching table much smaller than ordinary routing tables. However, in the meantime, advances in integrated circuit design for lookup tables allowed routers to operate at interfaces line rate, completely eliminating the issue.

Facilitating traffic engineering operations still remained a primary issue however and the standardization activities went on to produce a very functional protocol that is now being deployed by service providers all over the world. The advantage MPLS brings to operators is the ability to create arbitrary paths in the network, re-routing traffic flows without interfering with the traffic routed through the default paths selected by legacy shortest-path algorithms.

In this section we give a general overview of MPLS, focusing on its main characteristics, without going into the protocol details. For a more detailed description the reader can refer to [Tho02, Fau98].

MPLS carries traffic over Label Switched Paths (LSP). After receiving a packet from the IP layer belonging to a pre-established LSP, the ingress router adds a pre-defined label to it, sending it downstream towards the next hop. The next node will check the label on the incoming packet, look up its switching table, substitute the existing label with a new one and forward the packet downstream. The process continues until the packet reaches the egress router, where the MPLS label is finally stripped off and the packet routed following the default IP routing table.

The MPLS layer is closely integrated with the IP layer and located just below it. This allows the label switched paths to bypass the IP layer, while using information from the IP routing table at the ingress node to select the packets to be routed into the dedicated paths.

The labels assigned are only meaningful between a node and its downstream (or upstream) neighbor; this local assignment facilitates the use of distributed protocols for label distribution, like CR-LDP [Ber03a] and RSVP-TE [Ber03b]. The tasks of the distribution protocols include agreeing on the labels used for switching determined paths, and reserving the necessary bandwidth.

Label exchange however is operated after the routing protocol has established a suitable route for the switched path. Legacy routing protocols cannot be used for traffic engineering purposes as they only bring a very limited amount of information about the links. Traffic engineering instead needs up-to-date knowledge of the state of network links and more information like delay, capacity and protection-related parameters. For this reason, existing routing protocols have been updated (e.g. OSPF-xTE [SJ07], ISIS-TE [Smi04]) to allow the exchange of this type of information, and to implement routing algorithms capable of operating with multiple constraints.

In summary, the main advantage of MPLS is that it gives the capability to route selected traffic into dedicated switched paths that satisfy pre-established requirements, like reserved bandwidth, deterministic jitter and delay, and dedicated path protection. It is important however to emphasize that MPLS does not provide mechanisms to guarantee such requirements in the selected links. These have to be guaranteed either at layer 2 (for example through the ATM protocol) or by additional traffic conditioning mechanisms at layer 3 (e.g. shapers and droppers implemented in the DiffServ architecture).

The advantage that MPLS brings to traffic engineering is twofold: on one hand the operator has now full control over the routing mechanisms and can easily optimize the network resources while satisfying QoS traffic requirements. On the other hand it can easily set up and sell bandwidth-on-demand services to customers who might need, for example, point-to-point links with guaranteed bandwidth between two or more premises.

2.5.5 Interdomain vs. intradomain traffic engineering

The MPLS control capability we have described above functions well in the intradomain environment, where nodes are willing to share the necessary topology and link information, and to cooperate in order to optimize the usage of network resources.

The interdomain environment however is totally different, as each domain is an individual commercial entity, pursuing its own interests. This leads to a competitive model completely different from the cooperative intradomain.

Routing in the interdomain is achieved through the Border Gateway Protocol (BGP) [RLH06], which selects routing paths following a set of policies adopted by the network administrator. In selecting the routes, BGP gives higher priority to a local preference value manually selected by the operator, than to the number of hops (expressed in terms of autonomous systems). This is indicative of BGP's aim to satisfy specific operator's requirements rather than optimizing the overall network resources.

Hence, traffic engineering assumes a different meaning in the interdomain. Rather than optimizing the overall route, each operator can only directly influence the traffic distribution between its immediate upstream and downstream neighbors. Quite often, moreover, local decisions can cause unpredictable changes in traffic patterns in other domains.

Interdomain traffic engineering consists of balancing outgoing and incoming traffic to other Autonomous Systems (AS), of which [Quo06] gives an excellent overview. Since BGP was not originally developed to support traffic engineering operations however, traffic balancing is achieved indirectly through complex mechanisms. Incoming traffic, for example, is balanced using a mechanism called "path prepending", which consists of artificially increasing the hop count advertised for selected networks prefixes, in order to discourage their transit through the AS. The effects of path prepending are difficult to evaluate and its efficiency often depends on conditions that are beyond the influence of a single domain [LC07].

One of the goals of ASON/GMPLS networks (see section 2.3.2) is to provision dynamic paths across different domains (either at the electrical or optical level), following a standard circuit-oriented approach similar to that used in the intradomain. However we believe that the conceptual difference between intra and interdomain cannot be disregarded. The distributed nature of the Internet, its foundation upon independent autonomous systems and its distributed routing protocols are the basis of the network independence and ubiquity that have made it the primary worldwide communication medium. Although the GMPLS architecture is based on distributed protocols, its intelligence is still centralized. The decision to create a GMPLS optical path, for example, is made at the source node, even if the resulting path might traverse different domains [YMBS$^+$06]. We believe instead that interdomain optical paths should be created respecting the distributed nature of the Internet. In the Optical IP Switching architecture we present in this dissertation, transparent optical paths are created and extended in a distributed fashion, based upon distributed decisions made independently by the different domains.

2.5.6 Multi-layer traffic engineering

Because of the long time needed to provision additional bandwidth on network links, traffic engineering operations have so far focused on "moving traffic where the bandwidth is available". Packet routes are usually modified by acting on the link weight parameter of the IP routing protocols, or more recently through the creation of dedicated MPLS paths.

The GMPLS optical control plane standardization activity has focused on this issue, developing a protocol suite to allow dynamic link reconfiguration at progressively lower layers. Some initial implementations already provide reconfiguration capabilities in optical Ethernet, SONET/SDH and OTN networks. The practical implementation of dynamic lightpath reconfiguration is instead delayed by impairment issues at the physical layer, although control plane functions for transparent operations are currently under development.

This ability to dynamically reconfigure a network link will completely remodel the approach to traffic engineering, allowing network operators to "create bandwidth where required by the IP traffic".

In future NGN networks, Traffic engineering and Quality of service Optimization (TQO) will be performed at different levels of the protocol architecture [Ash06], which we have summarized in the diagram in figure 2.7.

The application layer takes care of the session initialization, providing information about destination address and class of service required. The transport layer optimizes the data transfer within a session, adapting the transfer rate to the condition of the transport channel (through the TCP collision avoidance mechanism). The task of the IP/MPLS layer is to find

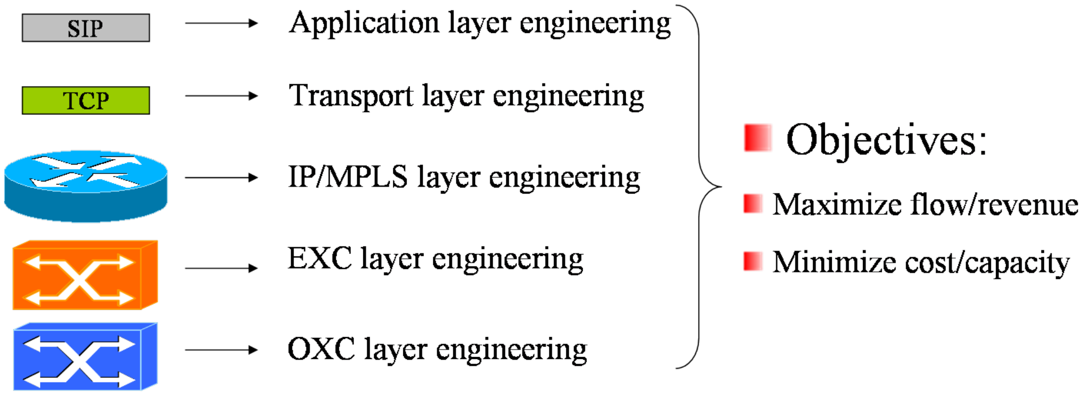


Figure 2.7: TQO at different layers of the protocol architecture

optimal paths to route packets towards their destination, given the link capacity imposed by the lower layers and the QoS constraint imposed by the application. Below the IP layer, the link capacity can be engineered through the use of Electrical Cross-Connects (EXC); they offer the capability of reconfiguring the link and path bandwidth with multiple granularities, allowing the bandwidth to be fine-tuned for the service being provided. For EXCs, the maximum bandwidth that can be provisioned is constrained by the lightpaths' configuration at the physical layer. Transparent Optical Cross-Connects (OXC), finally, allow the reconfiguration of physical paths, at the level of individual wavelengths, group of wavelengths or optical fibers. Since the cost of OXC ports is much lower compared to EXC and IP/MPLS ports, lightpath engineering plays a primary role in pursuing optimal network cost-effectiveness.

The benefits of multiple degrees of freedom for tuning network operations are: lower need for network over-provisioning, faster deployment of new services and applications, and the ability to customize the bandwidth for different services, leading altogether towards the optimal exploitation of the network resources.

The challenge on the other hand is the simultaneous optimization of the different layers, necessary to achieve a global optimum that maximizes the resource exploitation while maintaining the required QoS within the SLAs. Multi-layer traffic engineering is the research area that addresses this challenge.

We report some examples of multi-layer traffic engineering from the literature, where researchers have attempted the joint optimization of two or more network layers.

In [YKS$^+$02] the authors propose an architecture that merges traffic engineering at the IP/MPLS and OXC layers. They envisage the GMPLS-based Hikari router (described in more details on page 43), which allocates new lightpaths when the end-to-end demand between two nodes saturates the available bandwidth. These concepts are further developed in [OSS$^+$05], where the traffic engineering operations are extended also to the EXC layer.

In [DG07] the authors propose algorithms for optimization of routing strategies at the EXC and OXC layers. Taking as a reference model the route optimization for scheduled traffic (where demand arrival time and duration is known in advance), they show that the algorithms they introduce achieve similar optimal performance assuming only knowledge of the demand duration.

The work reported in [HBCR07] examines the interaction between the TCP and IP layers. The former operates distributedly to adapt the transfer rate to the network conditions, while the latter is centrally operated to re-route traffic over less congested links. The authors show that the current practice of engineering the two layers independently of each other, although it does not compromise network stability, does affect its robustness (with respect to traffic changes on a small time scale). The algorithm they propose, dubbed Distributed Adaptive Traffic Engineering (DATE), provides joint optimization of TCP and IP, offering the optimal trade-off in terms of stability, cost and robustness.

The approach we propose in this dissertation through our Optical IP Switching architecture, like the Hikari router, combines traffic engineering at the IP and OXC layers. Our solution however does not use a reactive approach, which creates novel paths when IP traffic saturates the available bandwidth. Rather, we use a proactive approach, which continuously aggregates IP flows into optical cut-through paths, reducing the packet forwarding cost through transparent bypass of the IP layer.

2.6 Current research projects

In this section we analyze how the research on optical networking has evolved over the past few years, describing its current trends by introducing some of the most relevant research projects recently developed or currently under development.

We focus primarily on projects relating to Optical Circuit Switching (OCS) because this is the main topic of this thesis, and represents a technology that can generally be implemented using off-the-shelf optical components. By presenting the projects in chronological order we will show the evolution of the research activity from the initial implementations of protocols and interfaces discussed within the ASON and GMPLS actions, towards the concept of highly dynamic transparent networks.

We also describe, in broad terms, the Optical Packet Switching (OPS), and Optical Burst Switching (OBS) architectures. OPS and OBS have been thoroughly investigated during the past decade, with the promise of delivering high networking performance by switching traffic optically at very fine granularities. However, the lack of a profitable ultra-fast switching technology (in the order of the microseconds for OBS and lower than a nanosecond for OPS) has prevented them (so far) from gaining commercial advantage over OCS networks.

Finally, in section 2.6.5, we briefly introduce the GENI initiative, a National Science Foundation project aimed at building a global network testbed facility. Once operational, GENI will allow researchers to test architectures, services and applications over a large-scale network, realistically reproducing their operation in the Internet.

2.6.1 Optical Circuit Switching (OCS)

The technology needed to implement OCS is generally mature, although challenges relating to transparent switching still remain at the physical layer (as described in section 2.3.1). Interdomain operations also remain an open challenge, both for the limited amount of link information exchanged among different domains [YMBS$^+$06] and for the lack of a business model specifically dedicated to NGN networks [VXHB05].

The strength of OCS is that it exploits the best of the electrical and optical domains, within the technology commercially available. The optical domain is used for optical transport and transparent switching of traffic at course granularity, while electronics are used for network administration and control, and for switching/routing at fine granularities.

This section gives an overview on how the concept of dynamic optical networks has evolved over the years, exploring the projects that have contributed to this evolutionary process.

LION project

The "Layers Interworking in Optical Networks" (LION) [IST00], was a European "Information Society Technologies" (IST) project funded under the Framework Program (FP) 5. LION was a three-year project that started in January 2000 with the aim of developing and testing a resilient and managed network infrastructure based on an the ASON and GMPLS architectures.

The main objectives of the project, which targeted multi-layer networks, were the following: testing novel technologies, like the Optical Transport Network described in section 2.2.2; implementing UNI and NNI signaling interfaces; designing an integrated optical control plane and a management architecture, to provide end-to-end view of the infrastructure over domains using different management technologies. The project terminated with the implementation of a testbed, described in [CBD+03], showing the feasibility of the concepts developed, and analyzing the performance of the resilience strategies adopted at the IP-MPLS (with fast restoration) and optical layer (using mesh protection).

This project, like the OIF interoperability demonstrations described in section 2.3.2, emphasized the potentials and feasibility of reconfigurable optical networks to operators.

The NTT Hikari router

The Hikari router is a GMPLS-based optical router developed at NTT research laboratories [SYT+02, OSS+05] in 2000, capable both of packet switching and wavelength path switching. One of its main characteristics is its ability to allocate bandwidth by creating new end-to-end optical paths when incoming traffic saturates the existing channels.

The lowest level of switched circuit is the packet label switched path, which can re-engineer the traffic transported through the network, creating MPLS circuits that bypass the IP layer. If the existing topology can no longer accommodate MPLS circuits, because some of the links are already working at their maximum capacity, the router can operate wavelength switching to generate a new virtual topology, capable of handling the requested MPLS paths. At a coarser level the router can also execute fiber switching, which allows more substantial alteration of the topology.

The Hikari router is probably one of the first GMPLS devices to reach the production environment, and is currently used by NTT in the Japanese national network.

DRAGON

The "Dynamic Resource Allocation via GMPLS Optical Networks" (DRAGON) testbed [Fun03b, LSJ06], funded in 2003 by the National Science Foundation (NSF), is an example of GMPLS-based grid network for e-science applications. The main focus of the network architecture is the capability of guaranteeing a deterministic level of service, achieved through resource reservation, and the provisioning of lightpaths connecting heterogeneous networks across multiple domains. The central element in the DRAGON architecture is the Network-Aware Resource Broker (NARB), which keeps updated resource information of the entire domain. Network nodes send end-to-end path requests to the NARB, which computes and returns the best route available, together with updated information on the selected links. If the end-to-end path is within the same domain, the signaling for path creation is achieved through a peer-to-peer approach, where nodes send information directly to each other in a distributed fashion. If the path crosses multiple domains the signaling is regulated by the NARB, following an overlay approach.

Another important function the NARB implements is the topology abstraction, which summarizes the internal domain topology, arbitrarily hiding detailed network information. This is used in interdomain provisioning, where the network operators do not want to disclose full topology information to competing adjacent domains.

The interdomain approach and the experience gained through the DRAGON testbed was of primary importance for the development of the Hybrid Optical and Packet Infrastructure network, described on page 49.

CHEETAH

The "Circuit-switched High-speed End-to-End Transport ArcHitecture" (CHEETAH) [Fun04, HSLR06], is an NSF project funded in 2004 aimed at developing the technology and infrastructure to support e-Science projects, and specifically the Terascale Supernova Initiative (TSI). The main aim of this project was to enable the dynamic creation of direct end-to-end paths, consisting of Ethernet circuits at the edges (i.e., in the LAN), carried by a SONET network in the core (i.e., in the WAN). This was achieved using a Multi-Service Provisioning Platform (MSPP) to map Ethernet frames into Ethernet-over-SONET (EoS) signals. The CHEETAH project developed an RSVP-TE client at the end-hosts that enables user applications to request the provisioning of dedicated Gigabit Ethernet circuits between end hosts or clusters connected to the GMPLS-based CHEETAH network. Additionally, a high-performance transport protocol called Circuit-TCP (CTCP) and application middleware was developed to provide guaranteed and stable throughput for file transfers, visualization, and control applications.

The CHEETAH testbed helped the understanding of relevant implementation issues in GMPLS networks, like for example the importance of IPv6 addressing, and the deployment of secure control channels over the Internet, while demonstrating the possibility of user-requested end-to-end dedicated circuits for high-rate data transfer.

MUPBED

The "MUlti-Partner european testBED for research networking" (MUPBED) [IST04a, SLS06] is an integrated EU IST project funded in 2004 that provides a test facility for ultra-broadband research networks. Its general objective is to investigate GMPLS technology and network solutions for the development of future European research infrastructures (grid networking being one of the primary objectives). The MUPBED testbed could be considered the European equivalent of the CHEETAH and DRAGON networks, as its main applications are the provisioning of end-to-end circuits across multiple domains, using the ASON/GMPLS control plane. The architectural design of the network is based on two different models, which provide service access following either an overlay or a peer-to-peer model. In the first case the grid access network is separated from the core, and the link between the two is provided through a Grid-User Network Interface (GUNI). In the peer-to-peer model instead there is a peering relation between access and grid core, and the user application can access the grid services directly through an Application Programming Interface (API). The fact that the MUPBED testbed makes use of GEANT2 as a core network, and different NRENs as local access points, has allowed the researchers involved in the consortium to carry out testbed demonstrations at different locations (usually during major networking conferences and events). The interesting aspects of these demonstrations were, on one hand, that end-to-end links could be created in real time, and on the other hand that the dedicated paths were provisioned through general purpose networks (GEANT2 and NRENs are networks carrying production traffic).

OptIPuter

The OptIPuter [TUC$^+$06] is a research project funded by NSF in 2002, which couples computational resources over parallel optical networks, in support of data intensive scientific research and collaboration [BSD$^+$06]. The scope of the OptIPuter project is to guarantee ultra-high bandwidth for e-Science applications like EarthScope and Biomedical Informatics Research Network, which require access to massive collections of distributed data objects that must be transferred with reliability and timeliness. The objective is to build a *Distributed Virtual Computer (DVC),* a virtual parallel machine where the processors are replaced by distributed clusters, the memory by large distributed data repositories and the peripherals by scientific instruments, visualization displays and sensor arrays. The whole system is interconnected by a system bus constituted by a dedicated lambda infrastructure, where data is transferred using the IP protocol. The lambda infrastructure, by offering transparent end-to-end optical paths to the applications, is the key for delivering ultra-high quality of service, including very large bandwidth (tens of gigabits per second) and controlled jitter/delay. Within the US the OptIPuter endpoints are connected through the National LambdaRail infrastructure (NLR), while the Global Lambda Integrated Facility (GLIF) provides links to the international community. Following the DVC approach, the OptIPuter creates dedicated end-to-end lightpaths in real-time (using a GMPLS-like approach) where lightpaths are requested directly by the applications, and bandwidth is provisioned for each application with guaranteed QoS.

NOBEL

The "Next generation Optical networks for Broadband European Leadership" (NOBEL) [IST04b] was an FP6 integrated IST project funded in 2004. Constituting a consortium of over thirty partners, the objectives of the Nobel project were very broad, targeting architectures for metro and core optical networks at different levels of the protocol stack. The reference architecture was based on a distributed ASON/GMPLS model, while its developments included multi-service management and integration, multi-layer traffic engineering, multi-layer resiliency, end-to-end Quality of Service, together with strategies for the end-to-end management and control of intra/interdomain connections.

The scope of the project however went beyond protocol implementation, and considered multiple aspects of networking. It targeted the social and techno-economic aspects related to the deployment of solution for intelligent reconfigurable optical networks, it identified the drivers behind the evolution of broadband services and it associated possible solutions with existing technologies and devices to optimize the network cost-effectiveness.

It has also traced a guideline for future network evolution. The short and medium term scenario will see initial coexistence of IP/MPLS, Ethernet, Next Generation SDH and OTN. This will progressively evolve in the long term towards a unified GMPLS control plane, which will completely integrate these network technologies. They also envisioned burst/packet switched optical networks for an extended long term scenario.

NOBEL-2 is the second (currently ongoing) phase of the NOBEL project, which makes use of the developments achieved during the first phase to target the network evolution in the long-term scenario. The project aims at implementing the concepts developed during the first phase into integrated testbeds, to demonstrate the feasibility of a unified control plane capable of supporting end-to-end services; this also involves the development of dynamic and transparent transport network architectures "for a pervasive introduction of broadband services in Europe". A second objective of NOBEL-2 is the exploration of future network concepts, addressing the tradeoff between optical vs. electrical, circuit vs packet, and routing vs. switching, to achieve an "optimum techno-economic balance".

TRIUMPH

The "Transparent Ring Interconnection Using Multi-wavelength PHotonic switches" (TRIUMPH) [IST06] is an FP6 IST "Specific Targeted Research Projects" (STReP) funded in 2006. The project's main objective is to build an architecture capable of transparently interconnecting core-metro rings (at rates up to 160 Gbps) and metro-access rings (at rates up to 40 Gbps). This objective is achieved through the use of a switching node located at the ring interconnection points, which provides novel transparent capabilities.

The project targets two of the most challenging aspects of transparent networking: signal regeneration and traffic grooming, directly in the optical domain.

TRIUMPH proposes multi-wavelength 2R (reamplification and reshaping) regeneration as a solution to the former issue. This technique optically regenerates multiple wavelengths at the same time, allowing sensible cost-reduction compared to current system (where each wavelength is regenerated separately).

All optical grooming is achieved using Optical Time Division Multiplexing (OTDM), a technique that allows grouping multiple lower rate WDM channels into a single higher rate stream and vice versa.

The importance of TRIUMPH among the various optical networking projects, is that it targets two physical issues (the optical regeneration and grooming), which constitute main obstacles towards the implementation of a cost-effective all-optical network architecture.

Clean slate Internet design

In 2003 the NSF launched a clean-slate program to invite researchers to propose new and revolutionary networking concepts not constrained by the current Internet architecture. It is widely believed in fact that the backward compatibility with existing Internet infrastructure is one of the main obstacles to the development of a robust, secure and economically profitable global network. The first initiative, jointly developed by CMU, Fraser Research, Stanford, Berkeley and Rice universities was the "100x100 Clean Slate program" [Fun03a], operational since 2003. The main objective focuses on re-developing the access and core architectures to create a global network capable of delivering 100 Mbps to 100 millions of houses.

In 2005 a similar initiative was developed at Stanford University, through the "Clean-Slate Design for the Internet" project [Uni05], aimed at funding project proposals fitting into one of the following, broadly-defined, areas: Heterogeneous Applications, Security, Heterogeneous Physical Layers, Economics & Policy, and Network Architecture (which links the others together).

Following the interest raised among the research community by the clean slate network idea, NSF has launched in 2006 a long-term initiative, dubbed "Future Internet Network Design" (FIND) [Fun06b]. This large-scale program invited the whole US research community to submit, in large scale, projects (for funding) aiming at exploring novel network architectures that could constitute valuable alternatives to the current Internet in fifteen years time.

The "Dynamic Optical Circuit Switched (DOCS) Networks for Future Large Scale Dynamic Networking Environments" [Fun06a] is a project funded under the NSF FIND program, led by the University of California (Santa Barbara and Davis) and Stanford University. Its aim is to develop novel technologies to enable dynamic optical circuit switching, such as hardware interfaces, network protocols and bandwidth allocation algorithms. One of the main objectives of DOCS, which constitutes its novelty, is to build the network using Photonic Integrated Circuit (PIC) technology to provide high connectivity, scalability and low cost.

Another relevant project funded under the FIND program is the "Future Optical Network Architectures" [Fun06c], led by MIT. The project emphasizes the fundamental physical differences between the electrical and optical technologies, asserting that it is unlikely that optical networking will follow a similar evolutionary path as electronic networking, from circuit to packet switching. Rather, the authors envisage the future Internet architecture as a circuit-switched network, based on the Optical Flow Switching concept described in the next section, integrated with impairment aware routing.

Optical Flow Switching (OFS)

Optical Flow Switching is an optical transport technique introduced at MIT ([CWM06]) that creates highly dynamic end-to-end lightpaths to transport traffic flows with a lifetime higher than 100 ms. The novelty of this approach is that optical paths are requested by the users to transfer data from source to destination LANs. Such paths transparently cross different MANs and WANs, and are relaxed when the transaction is completed.

Wavelength channels are statistically multiplexed in the backbone to achieve high utilization. The channel allocation is scheduled by dedicated processors located within each MAN.

Their scope is to accept lightpath requests from the users, select them through a scheduling mechanism, and coordinate the transmission of data through the WAN. As optical paths are created end-to-end, no buffering facilities are required in the network; all the traffic is queued at the end-user while it waits for the requested path to become available.

In [WCM06] the authors report a cost comparison between OFS and other network architectures: Electronic Packet Switching (EPS), where packets are routed electronically at each node; the GMPLS model, where packets are aggregated within the LAN and MAN, groomed electronically at the WAN edge, and switched transparently within the WAN backbone; and finally the Optical Burst Switching (OBS) model, where, like in OFS, packets are transmitted all-optically from source to destination, but unlike OFS the transmission is multiplexed randomly, causing occasional burst collisions at the switching nodes.

The cost modeling results show that OFS becomes economically convenient for high user's data rate (in the order of the Gbps), when the statistical multiplexing of wavelength channels becomes effective enough to justify the use of dedicated end-to-end wavelengths.

If we compare the OFS with the OIS architecture we will introduce in chapter 3, we see that the main difference is that the latter is an hybrid between optical circuit switching and electronic packet routing, while the former is only based on OCS. As a results OFS is only practical when the traffic rate at the user is very high (i.e., in the order of the Gbps), which allows exploitation of the high optical bandwidth. The benefits brought by OIS instead start at much lower traffic loads, as packet routing is used both to aggregate traffic into optical paths and to route small traffic flows, whose transmission rate does not justify the use of a dedicated optical path.

Research trends in circuit provisioning

In this section we have seen how the trend of research projects in optical (circuit-switched) networking was strongly correlated with the activities of the ASON/GMPLS standardization groups, and centered on the development of a control plane for the dynamic provisioning of dedicated end-to-end circuits over heterogeneous network architectures. These projects have focused on the implementation of the control plane standards on real networks, to evaluate the feasibility of dynamic networking, emphasize their efficacy in supporting future high-end applications, and discover the implementation problems and issues. Following the developments of the projects presented we now illustrate how the research activities have evolved over the past few years.

The first GMPLS-related projects, like LION and Hikari, had the task of producing a first implementation of the signaling protocols in switches and routers. These early testbeds were carried out in the laboratory, emulating a real network environment, as the main focus of the projects was testing the functionality of the GMPLS nodes.

Although these initial experiences proved the feasibility of the concept, network operators and equipment vendors needed to gain practical experience of the implementation of novel GMPLS mechanisms over real networks. This was achieved through the DRAGON and CHEETAH projects, which implemented the testbeds over existing networks, across multiple domains, and covering progressively larger areas. The experience gained through the DRAGON and CHEETAH testbeds was profitable for the development of the "Hybrid Optical and Packet Infrastructure" (HOPI) network [Int05], a project funded by the Internet2 consortium in 2005. The aim of HOPI is to build a hybrid network where default IP routing is combined with dynamic circuit allocation in a coherent and scalable architecture, which utilizes the existing Internet2 network infrastructure. In Europe a similar task is being carried out by the MUPBED project, which is bringing the dynamic provisioning idea into the "real world" by providing dedicated switched connections using main European network infrastructures.

On a parallel track, the idea of grid networks has emerged as a very appealing application of dynamic optical circuit switching. The idea is to use dedicated high-bandwidth connections to satisfy the large bandwidth requirements of high-end applications distributed around the globe, which need to exchange information at ultra-high data rates. Examples of such applications are: distributed computing, Very Long Baseline Interferometry (e-VLBI), High-energy physics and e-Health applications, to mention only a few.

Grid networking was already considered a main application for the DRAGON and CHEETAH testbeds. It is only with the OptIPuter however that the grid concept is fully embedded into the architecture, with the idea of the Distributed Virtual Computer, where applications automatically negotiate the end-to-end bandwidth with the network, through operations which are completely transparent to the end user. OptIPuter was also one of the first projects to promote end-to-end transparent wavelength switched circuits, a trend that is further being addressed by current research projects.

NOBEL(2) and TRIUMPH delineate important aspects in the near-future trend of research in optical networking, which we summarize with the terms "Transparency" and "Integration". Transparency, explicitly addressed in TRIUMPH, indicates the evolution of end-to-end dedicated circuits from the electronic to the optical domain, in a network where transparent paths are created with the same agility of MPLS or Ethernet switched circuits. This idea is also at the basis of the Optical Flow Switching model, which proposes an implementation of the transparency concept at the network and application layers.

Integration, on a parallel track, tries to merge the different switching and transport technologies to create an architecture that integrates optical, electronic, fixed and mobile domains into a unique service-oriented network. The user is only aware of the service he requests, while the network automatically provides the necessary bandwidth, dynamically choosing the underlying technology.

In this section we have given an overview of the major research topics that have characterized the evolution of OCS, analyzing some of the most meaningful networking projects. Other technologies, which might be applicable in the extended long term, are briefly described in the next section.

2.6.2 Optical Packet Switching (OPS)

Optical Packet Switching (OPS) is the forwarding of individual data packets, one by one, through a network directly in the optical layer. OPS, switching traffic at the packet granularity, aims at achieving the same levels of channel efficiency typical of electrically switched networks. Operating in the optical domain however gives OPS the advantage of higher bandwidth while avoiding O-E-O signal conversion, thus reducing power consumption and footprint [BBR+03].

Many of the current efforts at optical packet switching attempt to take all of the functions of electrical packet switching and transfer them to the optical domain. In order for this approach to be viable, many major technological obstacles must be overcome, in order to develop functional optical buffers and logic gates that would allow optical header recognition and clock recovery. Optical signal processing however is still in its infancy.

An approach that is more applicable to near-future optical technology instead tries to combine optical and electrical technology, where packet switching and transport are performed optically, while control and route processing are performed electronically.

In order to enable OPS, with currently available technology, Optical Label Swapping (OLS) has been proposed ([XY04, Yoo03]). OLS separates the data from the routing information, thereby giving transparency to payload data rate, protocol and modulation format. When the packet arrives at the optical router, the label, which contains the routing information and is transmitted together with the payload, but on a separate channel, is processed electronically. The optical switch is then activated by an electronic controller to switch the payload towards the correct output port (identified by the routing process), while the optical label is regenerated and transmitted beside the payload.

Although this approach is more feasible than all-optical packet switching, it is still limited by the speed needed at the optical switch (in the sub-nanosecond scale) and by the need for optical buffers (or Fiber Delay Lines, FDL), which store the optical payload while the label is being processed.

2.6.3 Hybrid OCS-OPS architectures

Due to the technical barriers to the achievement of pure optical packet switching, some researchers, aiming at developing near-future optical networks, have started to address hybrid optical packet/circuit switching architectures. This trend was first introduced by Hill and Neri in 2001 [HN01]: "Furthermore, from an overall networking perspective, a hybrid solution combining the merits of fast (optical) circuit switching with those of optical packet switching may offer better cost and performance. Indeed, such a solution may reduce the throughput requirements of packet switches". Since the concept of hybrid architecture is very generic, we clarify this matter by reporting on two network models implementing the idea.

ORION

The "Overspill Routing In Optical Networks" (ORION) architecture ([CBC+04, BCW+06]), developed by researchers at the University of Ghent in 2003, is based upon a wavelength switched optical network where lightpaths are used to transparently connect nodes that are not direct neighbors. The architecture design does not specify the rules for allocating the optical lightpaths, but focuses on a technique that enables the sending of packets to a certain destination, even when no network resources are available to provision a dedicated wavelength.

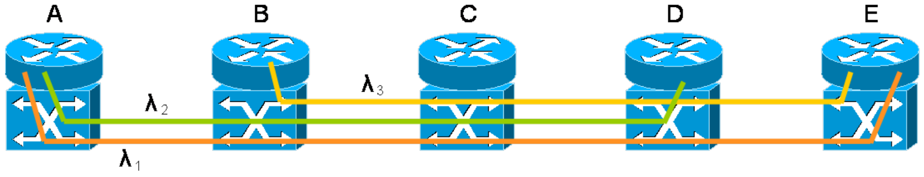


Figure 2.8: Overspill routing in the ORION architecture

Figure 2.8, illustrates the concept behind the ORION architecture. Node A needs to send data to D, but the wavelength λ_2 is fully used. Although there is capacity available in λ_1 between A and E, node D is transparently bypassed and cannot retrieve packets sent over this channel. In this case the ORION architecture allows node A to put packets directed to D on λ_1, marking them as "overspill packets". The special marking allows node B to recognize, extract and process the packets electronically. After node B has retrieved the routing information it puts the packet again in overspill mode, using one of the wavelengths available (λ_1 or λ_3). The process continues similarly for node C, which finally sends the packet to its destination D.

The main challenges of the ORION architecture are marking the overspill packets and finding an optical switching technology capable of recognizing and inserting/extracting them. The authors propose the use of an orthogonal labeling approach to mark the overspill packets, while the overspill mode is recognized electronically by a controller that examines all the packets on the transparent paths, and activates a fast optical switch to extract the packets detected in overspill mode.

Since ORION can be considered a hybrid between a point-to-point routed and a wavelength-switched architecture, the authors use these models as reference architectures for their simulations. The results show that the ORION architecture presents the best tradeoff in terms of wavelength resources and usage of routing processing power, among the architectures considered. This indicates that their hybrid solution is capable of effectively combining the individual advantages of point-to-point routed and wavelength switched networks.

OpMiGua

The "Optical Migration capable networks with service Guarantees" (OpMiGua) architecture ([BHS03, QY99]), launched in 2004 by Telenor R&D, is another hybrid optical circuit/packet switched architecture combining the advantages of wavelength routed networks (no buffer delay, jitter or contention, while requiring low processing power), and packet switched networks (efficient statistical multiplexing that guarantees high resource utilization).

OpMiGua differentiates the traffic in two distinct service classes: one circuit-switched Guaranteed Service Traffic (GST) class and one packet-switched Statistically Multiplexed (SM) service class. The packets in the GST class are switched through the dedicated lightpaths of the wavelength routed network, which can guarantee deterministic quality of service. Those in the SM class instead are serviced as best effort traffic and transmitted exploiting the empty gaps that GST packets or bursts leave in their channels. This mechanism uses the switched packets to increase the channel utilization, which tends otherwise to be low in wavelength-switched networks.

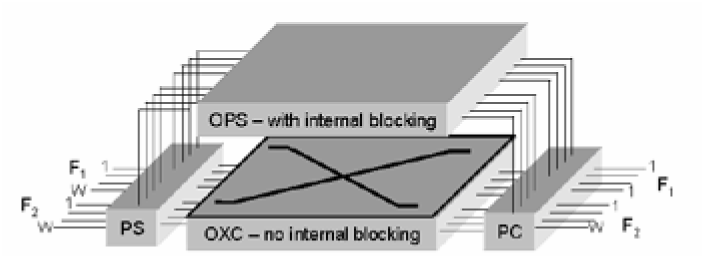


Figure 2.9: Illustration of an OpMiGua node

GST and SM packets are differentiated within the same wavelength channel by using orthogonal optical polarizations. At each node a polarization splitter (the PS block depicted in figure 2.9) separates the two traffic types, sending the former through the transparent optical cross-connect (the OXC block in the figure) and the latter to the optical packet switch (the OPS block). The main difference with the ORION architecture, described in the previous section, is that in OpMiGua, packets are differentiated by an optical device, which is faster and might in future prove to be cheaper compared to the electronic circuitry used in ORION.

Exiting the node, the packets coming from the OXC are mixed together with those coming from the OPS layer by a polarization controller (the PC block in the figure). Although GST and SM packets use different polarizations, they need to be multiplexed in the time domain, or Polarization Dependent Loss (PDL) and Polarization Mode Dispersion (PMD) would cause detrimental cross-talk when demultiplexing. In order to assure differentiated QoS classes, in case of collision the GST packets have the right of way with respect to the SM packets, which are either dropped or else buffered until a suitable time slot is available at the output.

2.6.4 Optical Burst Switching (OBS)

Optical Burst switching [QY99] is a technique that raised a lot of interest worldwide in the past few years and consists of aggregating packets with similar destination at an edge node. After collecting a sufficient amount of packets the node triggers the creation of an optical point-to-point connection, where the packets are sent as a burst to the network egress point. The network resources are released straight after the burst has been transferred.

Optical burst switching could be considered an intermediate step between the imminent dynamic optical circuit switching and the future optical packet switching. The time requirements for setting up and deleting optical paths are in this case of the order of milliseconds, while optical switching components need to operate at speeds tens of times faster.

Currently, for an all-optical network to work efficiently without ultra-fast hardware, the basic data unit has to be significantly larger than a single IP packet. In OBS networks this difficulty is overcome by assembling incoming packets into bigger entities, called bursts, at the edge node. Like in optical label switching, the control information is separated from the payload. The control packet is sent ahead of the actual data using a different channel. When received by a core node it is converted into electrical form, analyzed and sent to the next node. At each stage a routing decision is made and a connection is set up. The data burst is then sent by the edge node without waiting for any acknowledgment, using a one-way reservation protocol like the Tell-And-Go (TAG) described in [VS97, Wid95].

A guard time between the control packet and the data burst, called the offset time, is necessary to allow the intermediate nodes to prepare a lightpath for the incoming burst. It can be eliminated if, at each node, data bursts are delayed in a fiber delay line while the control packet is being analyzed.

One of the main issues of burst switched networks is the tradeoff between the burst size and network latency. Large bursts increase the network efficiency by increasing the network utilization (it minimizes the path setup time compared to the time needed to transport data); larger amount of data however are gathered by buffering packets for a longer time at the source, which increases the network latency.

The data burst traverses the entire network in the optical domain. Only at the destination node it is converted to electrical. It is then disassembled and all the IP packets are sent to their respective destinations. The control packet instead has to be converted to the electrical domain at each core node. Since each control packet however accounts for a multitude of data packets (those contained in the burst), the electrical routing hardware only processes a fraction of the packets traversing the node, highly reducing the bottleneck at the IP layer.

2.6.5 GENI

The "Global Environment for Network Innovation" (GENI) [Fun06d] project's objective is to build a shared, global testbed facility designed to facilitate research on network architectures, services and applications. The GENI project targets the well-known problem that the network research community lacks suitable technology to test new ideas rigorously, in a global environment. The Internet infrastructure is in fact too complex to be thoroughly simulated, while testbeds have so far only been implemented on small scale. This is among the main reasons why improvements to current protocols, like IP or TCP, for example, which are at the heart of the Internet, are generally accepted with difficulty by standard bodies and usually require a very long time to prove their robustness and efficiency.

The GENI facility will be constituted by an experimental platform that fills the gap between small-scale experiments in the lab and mature technology ready for commercial deployment. The GENI physical substrate will consist of a collection of networking components like optical links, switches and routers, processor clusters, storage areas and wireless subnets. It will constitute a large shared testbed that researchers will use to evaluate network systems on large-scale. The platform will be shared among different research institutions through a management framework that assigns different slices of the physical substrate to different experiments.

Four key ideas make this possible. First, the physical components will be programmable, to ensure that revolutionary network designs, radically different from today's Internet, can be tested. Second, the substrate will be virtualizable, so that different experiments can be run simultaneously and continuously, without the need for pre-reserved time slots. Third, GENI will allow external users to interact with the experiments being run, which will provide the tested application or architecture with traffic generated by real users. Finally, GENI will be modular, so that novel devices and technologies can be further integrated to keep the physical infrastructure up-to-date with the newest technologies available.

The relevance of the GENI platform is that it represents a unique facility, the first tentative attempt to build a shared, global-scale testbed. We believe that this effort will substantially speed up the evolution of the Internet towards new services, supported by more efficient protocols, transport and forwarding architectures.

2.6.6 Future research trends in optical networks

Section 2.6 has given a general overview of the latest achievements in optical networking, describing the main projects that in the past few years have demonstrated the feasibility of dynamic circuit provisioning, and proposed novel network concepts and architectures. We have followed the evolution of OCS from on-demand provisioning of end-to-end dedicated circuits to application-generated requests of transparent lightpaths over heterogeneous technologies and network domains.

We also have envisioned the network evolution in the extended long term, with OPS, hybrid OPS-OCS and OBS technologies. Although the OPS idea was developed more than a decade ago, the design of a commercial OPS network has been progressively postponed, due to major technological barriers. The OBS and hybrid OPS-OCS concepts were proposed as architectures that partially relaxed the fast switching requirement at the optical level.

Currently it seems, observing the FIND and GENI initiatives, that the trend in the long term scenario is towards the development of novel and revolutionary optical architectures, which go beyond the simple translation of electrical operations into the optical domain. Rather than trying to adapt the optical technology to the existing networking concept (all derived from the electronic domain), the new research trend is towards totally new network designs built around the available optical technologies. The best example is probably the concept of networks with little or no buffer requirements [EGG$^+$06, Øve07] to avoid building expensive, cumbersome and complex optical memories.

2.7 Summary

In the past few years the research trend in optical networks has been that of increasing the reconfiguration capability, to decrease the bandwidth provisioning time, reduce its cost, and enable novel network services. Dynamic reconfiguration, which in legacy networks was limited to the IP layer, has progressively involved lower layers, with the introduction first of MPLS, and more recently with the GMPLS protocol suite.

This evolution was driven by the common belief that dynamic reconfiguration would bring substantial economical benefits in optical networking. Faster provisioning, on the one hand, will increase the revenue of network operators, providing them with the ability to sell new bandwidth services with guaranteed QoS. Transparent switching, on the other hand, will introduce cost savings in capital expenditures by replacing part of the expensive layer-3 equipment currently deployed with transparent optical cross-connects.

Some issues however still exist, which need to be solved before dynamic lightpath reconfiguration can be implemented in production networks. One of the main problems is that dynamic lightpath provisioning causes signal impairments at the physical layer, because the optical transport layer is currently non-reconfigurable. Another relevant issue is that dynamic reconfigurations need to be performed simultaneously at different layers in order to optimize the traffic engineering performance. This issue, dubbed multi-layer traffic engineering, generates a complex multi-degree optimization problem.

Many research projects in the past have targeted these issues, gradually improving the reconfiguration capability in optical networks. Most of them however have only considered end-to-end lightpath provisioning. This approach implies that lightpaths crossing multiple nodes and domains are requested by a source node. Although end-to-end provisioning allows the delivery of high bandwidth services with guaranteed QoS to the user, its use becomes problematic in the interdomain. The low amount of information shared among different domains in fact makes it difficult to calculate optimal routes. Traffic engineering operations are also heavily affected by this lack of information.

The Optical IP Switching (OIS) architecture we propose in this dissertation tackles this problem by creating optical paths distributedly, following the observation of local traffic. The main advantage of this approach is that, being distributed, it is more consistent with the independent and distributed nature of the Internet, where each domain takes decisions according to its own network policies. In addition, the network becomes more scalable and easier to manage, as traffic observation and provisioning decisions are operated at each node rather than being centrally administered through a management layer.

Chapter 3

Optical IP Switching

Optical IP Switching is a technique we have developed that creates and deletes optical cut-through paths in a distributed fashion in response to a local analysis of IP traffic.

The idea of flows bypassing the IP layer is inherited from IP switching, which was first introduced in the electrical domain as a method of combining Asynchronous Transfer Mode (ATM) and IP technologies [NML98, LM97]. This idea, originally conceived by Ipsilon Networks Inc., was then further developed by Cisco's Tag Switching, and finally standardized by the IETF as Multi-Protocol Label Switching. This evolution has seen a dramatic change in the aim of such protocols. While IP switching detected IP flows automatically, operating a flow-by-flow switching (e.g., between two IP terminals or TCP sessions), MPLS creates end-to-end switched flows on demand for traffic engineering purposes. Optical IP Switching remains closer to the original IP switching idea, allocating dedicated paths to IP flows of suitable characteristics. The novelty of the OIS approach however is that the dedicated electronic circuits are substituted by wavelength switched paths. Switching data directly in the optical domain has important consequences. On one hand it can allow cost saving, as optical switch ports are data rate independent, and cost tens of times less than IP ports. On the other hand however, optical switching is operated at the wavelength granularity, which can become quite inefficient compared to the packet granularity offered by electronic routers and switches.

This chapter explains in detail the optical cut-through path allocation process.

Section 3.1 gives an initial overview of the OIS architecture, indicating the main functions we have developed in the protocol stack and the hardware required for a practical implementation.

In section 3.2 we go into the details of the traffic analysis process, introducing our prefix-based method for aggregating traffic depending on its destination network.

The path allocation process makes use of the information collected during traffic analysis, and is divided into three main functions. Path creation, described in section 3.3, creates an initial optical path between a node and its upstream and downstream neighbors. Path extension, described in section 3.4, allows a node to extend an existing path towards an upstream or downstream neighbor. Path cancellation, described in section 3.5, deletes existing paths that have become under-exploited due to traffic changes. The path allocation functions are presented in this chapter in a different order to how they appear in the functional diagram of figure 3.3, in order to give the reader a clearer understanding of the whole process.

We have previously anticipated that the distributed provisioning mechanism of OIS facilitates operations in the interdomain, as it allows every node to apply individual network policies. In Section 3.6 we illustrate how the original signaling methods for path extension and cancellation, described in sections 3.4 and 3.5, can be extended for interdomain provisioning.

Section 3.7 describes the flow re-classification function, which can be applied periodically to refresh the list of prefixes switched by the existing dynamic paths. The aim of this function is to increase the lifetime of each path, increasing the overall switching efficacy.

In section 3.8 we describe the Port and Link Discovery functions we have implemented. We believe in fact that a highly dynamic architecture like OIS should avail of highly automated discovery functions, approaching a plug-and-play model.

The final two sections deal with the main technical issues arising from dynamic reconfiguration of optical paths. Section 3.9 describes the impairments that OIS generates at the transport layer. In Section 3.10, we give a brief insight to the main open challenges at the physical layer; although we do not pretend to present a solution to the problem, we initiate a qualitative discussion on how some of the main issues could be tackled.

All throughout this chapter we point out the challenges we have identified during our work. After discussing these issues, we summarize them into well-defined research questions that will be answered in chapters 4 and 5 through our simulations and testbed trials.

3.1 OIS architecture

The idea behind Optical IP Switching is that of an IP network that adapts the underlying physical topology to the traffic flows encountered at the IP layer. The decision-making process is completely distributed and is only based on local traffic observation.

An OIS node monitors the traffic by sampling IP packets at a certain rate, using mechanisms similar to those adopted by Netflow. The goal is to identify elephant flows, aggregate them on the basis of their upstream and downstream directions, and evaluate the feasibility of switching those flows into a dedicated optical cut-through path. This path is then established between the selected upstream and downstream neighbors. As clearly illustrated in figure 3.1, the advantage of the cut-through path is that the middle node can switch the flows at the optical layer, without consuming expensive router resources. A path extension process then allows upstream and downstream neighbors to extend the optical path to their own neighbors, availing of the advantages of transparent switching.

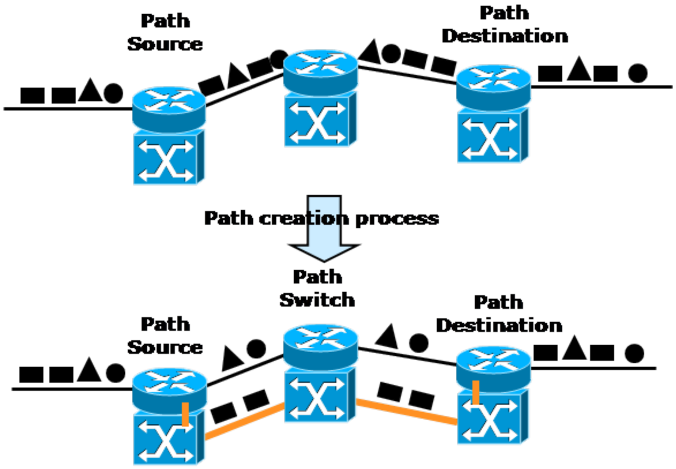


Figure 3.1: OIS path creation mechanism

The main advantage of this distributed decision mechanism is that every node can autonomously evaluate the convenience of switching or routing a flow aggregate, depending on its available electrical and optical resources. This makes OIS especially suitable for interdomain networking, and more generally for highly heterogeneous networks, because it allows every node to make its own traffic analysis and optimization.

Figure 3.2 gives a logical overview of the Optical IP Switching architecture. The upper part of the diagram specifies the IP/OIS functional elements together with their logical interconnections. The OIS functions are closely integrated with the IP layer entities, with read/write access to the routing table and ability to avail of the routing engine to forward signaling messages. Both the OIS and IP protocols can influence the routing mechanism by modifying the entries in the routing table independently from each other. The advantage brought about by the close integration of IP routing and optical switching is twofold. Firstly, it associates the routing at the IP and optical layers, allowing multi-layer traffic engineering. Secondly, it guarantees full backward compatibility with default IP networks, a crucial characteristic for the practical implementation of any Internet architecture.

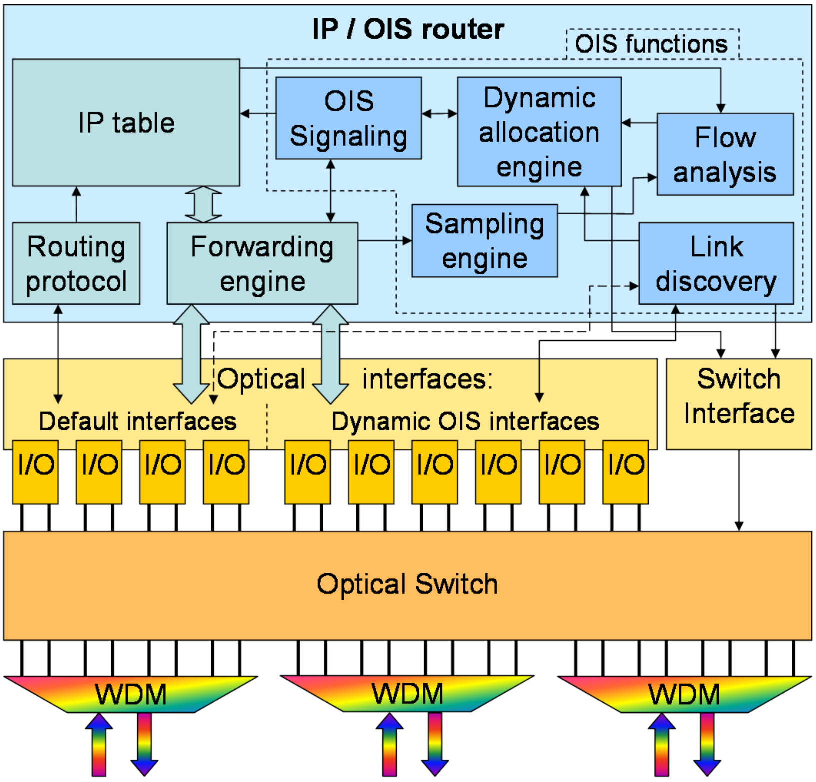


Figure 3.2: Overall Optical IP Switching architecture

The optical interfaces allow the router to create and terminate optical links, over which packets are transmitted and received. Some of these are used to create static default links to neighbors, through which normal IP traffic and control packets are exchanged. The remaining interfaces are used to accommodate the dynamic optical paths.

The optical switch is the element that physically links the node to the external world. On one hand it allows each interface to connect to any incoming or outgoing fiber; on the other hand it allows the transparent switch of an incoming to an outgoing fiber, creating optical bypasses of the IP layer. The switch is directly controlled by the dynamic allocation engine through a dedicated interface. In the figure, we have illustrated a fiber switch to clarify the interconnections of the optical elements. For a real implementation however, wavelength-selective switches (incorporating the WDM multiplexers) could be employed.

In the following sections we discuss the details of the different elements which make up the OIS architecture.

In sections 3.2 to 3.5 we describe the flow analysis and optical path creation functions. The functional diagram depicted in figure 3.3 gives an overview of the decision mechanism behind the Optical IP Switching operations. The first step is the traffic analysis and characterization, which examines traffic for a given period of time (in the order of minutes). At decision time the traffic information is used to determine which path can be canceled, which one can be extended, and which one can be newly created. Flow re-classification is an optional feature and can be used to update the flows switched on the optical path with finer time granularity with respect to the decision time.

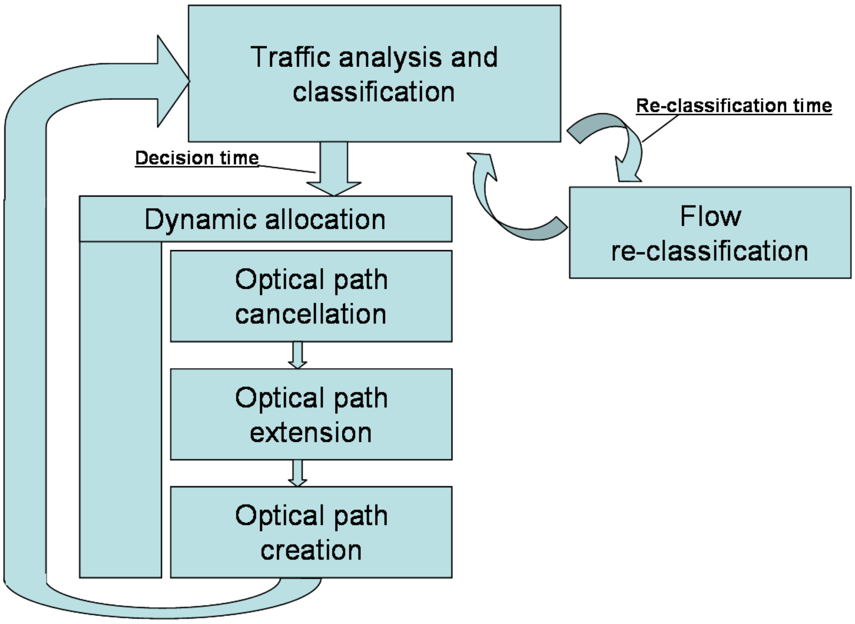


Figure 3.3: Functional diagram of Optical IP Switching

3.2 Traffic analysis

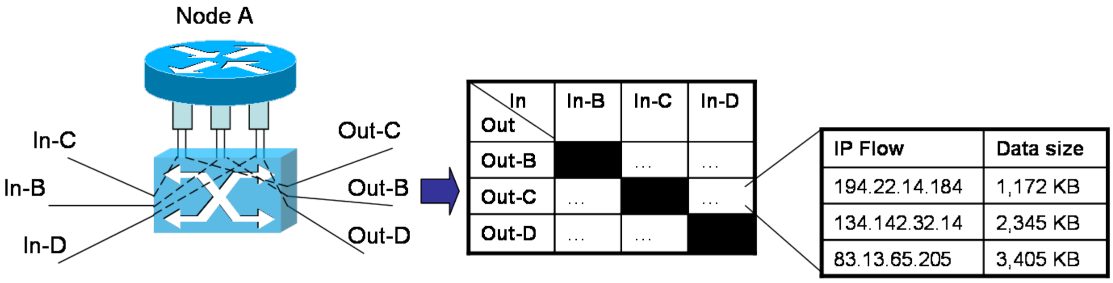
Traffic analysis is the first functional step of the Optical IP Switching mechanism. Every OIS node performs constant analysis of the IP traffic transiting the router, using one of the packet sampling mechanisms described in section 2.5.1. The information that needs to be collected includes: the interface from which the packet arrived, the output interface, which is selected by the router through the longest prefix matching algorithm, the payload size and the arrival time (time information, for example, can be processed to identify and estimate long lived IP flows [PTB$^+$01]). In sections 2.5.1 and 2.5.2 we have introduced some advanced methods, proposed in the literature, that analyze traffic flows and try to predict their behaviour. In our studies, for simplicity, we have used a simple mechanism that selects traffic based on a pre-established threshold, but more complex mechanisms could be adopted in future to improve the channel usage.

The basic idea, following the IP Switching approach, would be to create an optical path as soon as an IP flow of suitable size is observed at the IP layer. However, the average elephant flow rate and optical link bandwidth differ by three to four orders of magnitude, which makes this method inefficient. Although this approach would be suitable for grid networks, where high end applications might easily occupy entire wavelengths, we aim to design an optical architecture suitable for more general network models.

Wavelength utilization can be increased by aggregating multiple flows into the same lightpath. This process appears complex if operated in a centralized fashion because it would require full knowledge of the flows at one location. Besides the flow path, also information about flow data rate would need to be constantly updated, requiring large amounts of signaling overhead and centralized processing power. In contrast, our distributed approach spreads the processing power requirements among different nodes and reduces the signaling overhead to local communication (without transmitting all the information to a central entity).

3.2.1 Flow-based aggregation

In a flow-based aggregation approach the traffic analysis is operated by categorizing the sampled packets based on their arrival and departure interfaces. This classification is necessary because optical cut-through paths are created by wavelength or fiber switches, which transparently connect incoming links to outgoing links (from upstream to downstream neighbors, respectively); therefore they do not allow to groom traffic in the optical domain. The classification function works by building an “aggregation matrix”, illustrated in figure 3.4, with number of columns and rows equal, respectively, to the number of upstream and downstream neighbors (we assume for simplicity that each neighbor is connected through one default link). The generic matrix cell (i,j) stores information on traffic incoming from interface “i”, relayed through interface “j”. Each cell of the matrix collects information about all the flows that can be potentially aggregated into a single cut-through path, while, within each cell, sampled packets are sub-categorized depending on the IP flow they belong to. For flow-aggregation purposes only the information about the destination IP address of each flow is necessary, because this is the information used by the IP routers to forward packets. The 5-tuple (transport protocol, source port, destination port, source IP address and destination IP address) is instead useful for QoS-based flow differentiation, which will be addressed in the future work section (6.2).



In / Out	In-B	In-C	In-D
Out-B	■	…	…
Out-C	…	■	…
Out-D	…	…	■

IP Flow	Data size
194.22.14.184	1,172 KB
134.142.32.14	2,345 KB
83.13.65.205	3,405 KB

Figure 3.4: Flow-based aggregation mechanism

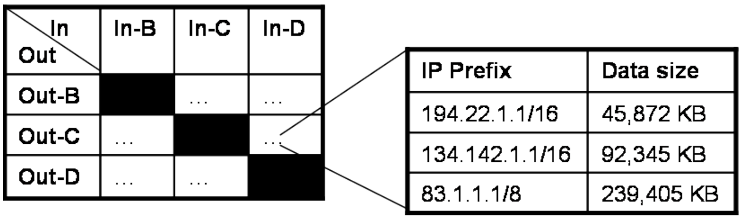
Although flow-based aggregation might seem adequate at a first glance, a deeper investigation reveals that it does not scale well with the number of flows. We base our deduction on the traces considered in section 2.5.2, which were taken from a trans-Pacific OC-3 signal (operating at the line rate of about 155 Mbps). Due to the heavy-tail distribution of the traffic, 70% of the data in the traces considered is carried by about 1% of the flows; this totals about 800 flows. If we scale these statistics to a 10 Gbps channel, a rate that is currently being deployed in wavelength links, we could infer that 1% corresponds to about 50,000 flows. Considering that a node's degree (i.e., the number of its direct neighbors) usually varies between two and six (or more), in the average such a node would see between 100,000 and 300,000 elephant flows simultaneously.

Such large numbers generate scalability issues at the source node, which has the task to inject packets into the optical cut-through paths (as it will be explained in section 3.3). Firstly, they create excessive overhead during the signaling phase, when the middle node signals the entire list of switched prefixes to the path source. Secondly, they substantially increase the size of the routing table at the path source, which adds an entry for each switched flow. Any increase in the size of routing tables is an issue of major concern for router administrators. Considering that current BGP routing tables have about 200,000 entries, switching elephant flows would imply increasing the routing table of source nodes by a factor that is at least of 50% and can increase to more than 250%.

In order to eliminate these two issues, we have proposed a different aggregation scheme that makes use of the IP prefixes stored in the routing tables.

3.2.2 Prefix-based aggregation

The flow aggregation approach we have developed for the OIS architecture is based on IP prefixes. Within each of the cells in the aggregation matrix, instead of classifying the packets considering the IP destination address (as in the flow-based aggregation approach), packets are categorized depending on the destination prefix they are directed to (as in figure 3.5). For each sampled packet, the router, similarly to the flow-based approach, checks its destination address and determines the output interface “j”, using the longest prefix matching algorithm. The size of the payload contributes to the total amount of data carried by the matching prefix within the cell (i,j) (“i” being the interface from which the packet arrived). More advanced algorithms might be developed here that also use packet timing information to predict the traffic behavior among the prefixes, in order to improve traffic characterization. The design of more performing algorithms will be addressed in the future work, as the primary objective of this dissertation is the overall design of the OIS architecture.



In / Out	In-B	In-C	In-D
Out-B	■	...	...
Out-C	...	■	...
Out-D	...	...	■

IP Prefix	Data size
194.22.1.1/16	45,872 KB
134.142.1.1/16	92,345 KB
83.1.1.1/8	239,405 KB

Figure 3.5: Prefix-based aggregation mechanism

The advantages introduced by aggregating flows through destination prefixes are many. The most important is that the size of the routing table is not unduly increased, since each prefix summarizes a large amount of class-D IP flows. Moreover, since most of the prefix entries in the BGP routing tables of peering BGP nodes are similar (what changes is instead the next hop value), the upstream node rarely needs to add the switched prefixes as new entries, and most of the time it will only modify the next-hop value (this will be clarified in the next section). Prefix summarization also diminishes the signaling overhead, as each prefix counts now for many IP flows. The traffic analysis phase is also simplified, as traffic information is processed at the granularity of the routing prefixes.

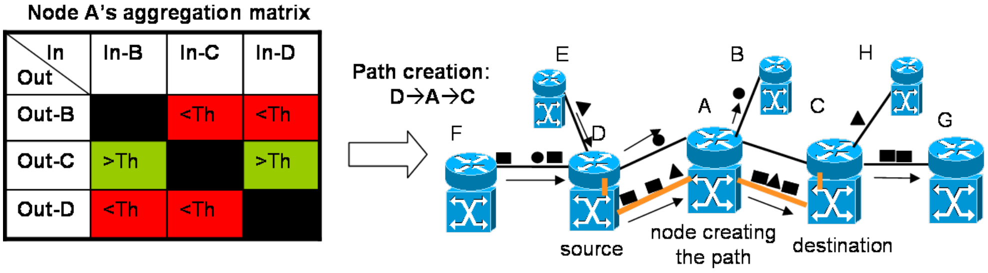
The disadvantage of prefix summarization is that information about each flow is disregarded, therefore it is not possible to guarantee quality of service to individual flows. Per-flow QoS could however be achieved operating layer 3 or layer 2 end-to-end virtual circuits, a concept developed in the “Flow Routing” architecture by Roberts [Rob03]. A hybrid solution would then see deterministic quality of service guaranteed electronically, while Optical IP Switching would operate traffic engineering at the optical layer on a coarser granularity. More details on this can be found in the “Future work” (section 6.2).

Another interesting aspect of prefix-based aggregation is that the heavy-tail distribution observed at the IP flow level occurs also at the prefix granularity; a small number of elephant prefixes carries most of the data. Its implications are that signaling and routing table updates can be further reduced if only elephant prefixes are taken into account. Elephant prefix identification can be implemented through a threshold mechanism that, for example, excludes from the aggregation all the prefixes that route traffic at a rate lower than a pre-established “prefix threshold”. We can summarize the issue with the following research question: how does the Internet heavy-tail flow distribution affect the prefix distribution in the IP routing tables? The answer to this question will be provided in section 4.2.2, where we analyze the prefix distribution of a major European core network.

3.3 Optical path creation

After collecting traffic information in the "Observation" state, the node passes to the "Decision" state. At decision time the router analyzes the statistics collected, summing up the amount of data carried by the different prefixes within each cell. Only cells whose aggregate data is above a pre-established “path creation threshold” value are eligible for Optical IP Switching.

The path creation process (figure 3.6) only considers cells whose collected data is higher than the path creation threshold, starting from the generic cell (i,j) showing the highest value.



Node A’s aggregation matrix

In / Out	In-B	In-C	In-D
Out-B	■	<Th	<Th
Out-C	>Th	■	>Th
Out-D	<Th	<Th	■

Figure 3.6: Path creation process

The signaling process sees the router requesting from its upstream and downstream neighbors (using interfaces “i” and “j”) the creation of a new optical cut-through path on a suitable wavelength. Eventual negative replies from the neighbors would carry the reason for the path denial. If the problem can be solved by selecting a different wavelength, the middle node will continue proposing alternative wavelengths until both neighbors agree. If no available wavelength can be found, the current path creation is aborted and the process moves to the next matrix cell (more details of the signaling implementation can be found in section 5.2.1).

If both neighbors acknowledge the request, the router sends to the path source the list of prefixes to be switched through the new optical path (figure 3.7).

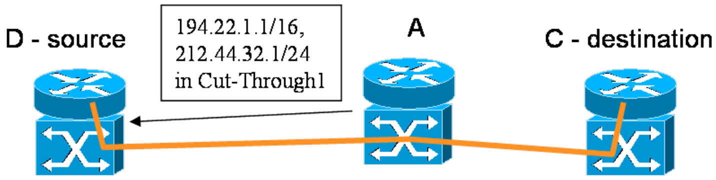
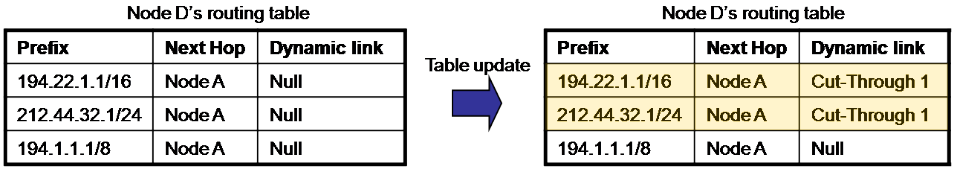


Figure 3.7: Signaling of switched prefixes

Once the path is created the upstream node updates its routing table (figure 3.8) and starts injecting the matching packets into the new transparent path. In order to have the legacy IP and the OIS protocols coexist without interfering with each other, we have slightly modified the structure of the IP table. We have added a field beside the "Next hop", named "Dynamic link". If the "Dynamic link" is not null, the packet will be forwarded through the interface indicated by this field, otherwise the "Next hop" value will be used. If the "Next hop" field of a table entry is modified by the IP routing protocol while a related cut-through path is active, this change will not affect the OIS path immediately; only after the dynamic link value is set back to null (for example, after the cut-through path is deleted) the change will become effective.


Node D's routing table

Prefix	Next Hop	Dynamic link
194.22.1.1/16	Node A	Null
212.44.32.1/24	Node A	Null
194.1.1.1/8	Node A	Null

Table update

Node D's routing table

Prefix	Next Hop	Dynamic link
194.22.1.1/16	Node A	Cut-Through 1
212.44.32.1/24	Node A	Cut-Through 1
194.1.1.1/8	Node A	Null

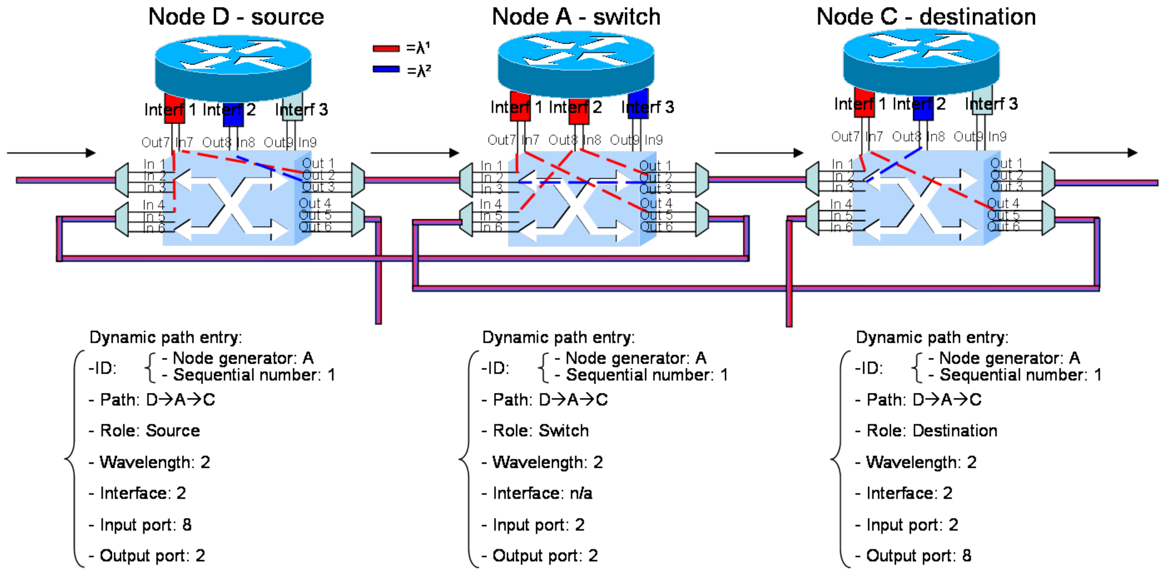
Figure 3.8: Routing table update of the path source

Since, as observed in the previous section, most routers have similar prefix entries in their routing tables, the update process will mostly consist of re-writing the "Dynamic link" value, without requiring a noticeable increase in the number of table entries. The entries that are instead added to the routing table by the OIS protocol should be deleted when the corresponding cut-through path is canceled. Such entries are easily recognizable because their "Next hop" value is null.

A similar path creation process is repeated for the remaining matrix cells with traffic above the path threshold.

While the path is being used, the source node periodically sends path refresh messages downstream. These are first processed and then relayed, hop-by-hop, down to the destination node. If the downstream nodes do not receive a refresh message within a pre-established period, they free the resources committed to that path.

All the nodes store information about the optical cut-through paths they are involved in, adding an entry for each path into a "dynamic path" list. Each entry uniquely defines the path through a path ID that identifies the path creator plus a path sequential number. The entries carry information about the path route, the role of the node in the path (i.e., source, switch or destination), the wavelength used, the optical ports involved and the interface used (only for source and destination nodes). Figure 3.9 shows an example of dynamic path entries for each of the nodes involved in the path.



Dynamic path entry:
-ID: - Node generator: A
- Sequential number: 1
- Path: D→A→C
- Role: Source
- Wavelength: 2
- Interface: 2
- Input port: 8
- Output port: 2

Dynamic path entry:
-ID: - Node generator: A
- Sequential number: 1
- Path: D→A→C
- Role: Switch
- Wavelength: 2
- Interface: n/a
- Input port: 2
- Output port: 2

Dynamic path entry:
-ID: - Node generator: A
- Sequential number: 1
- Path: D→A→C
- Role: Destination
- Wavelength: 2
- Interface: 2
- Input port: 2
- Output port: 8

Figure 3.9: Entries in the dynamic path list

It is very important to notice that the dynamically created optical paths should not be seen by the IP layer as new peering links. In our implementation the IP protocol does not exchange link discovery information over the optical cut-through paths. The reason is straightforward; every newly discovered link causes the generation of routing discovery messages (like the link state advertisements generated by the OSPF protocol). First, this makes the routing protocol unstable as the link update time approaches the protocol convergence time; too frequent route re-calculations moreover consume excessive processing power. Second, it increases the signaling overhead, leading to excessive bandwidth consumption. Although the issue is mitigated in the intradomain, where convergence times are usually of the order of a few seconds, it becomes a major problem in the interdomain. Due to the worldwide extension of the Internet, BGP convergence times can in fact vary from minutes to hours. In our approach, by hiding the dynamically created optical paths to the link advertisement process of the routing protocols, we eliminate the instability problem induced by highly dynamic path provisioning.

In our analysis there are three parameters that influence the dynamics and performance of the path creation process: the *prefix threshold*, the *path creation threshold* and the *length of the observation time*.

We have already clarified the meaning of the *prefix threshold* in the previous section and will report the relative simulation results in section 4.2.

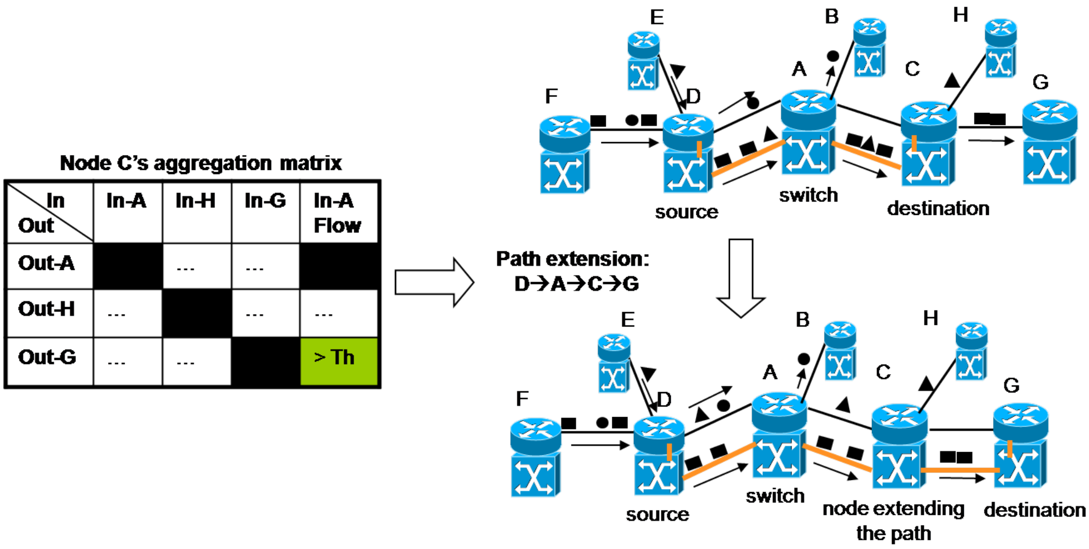
The second parameter, the *path creation threshold*, identifies the minimum amount of traffic that should trigger the creation of a novel optical cut-through path, and directly influences the channel efficiency. Its optimum value depends on the cost associated with each wavelength channel and optical switch port, versus the cost of IP ports. Intuitively, the lower the cost of optical channels and ports with respect to the IP ports, the lower we could set the threshold value. However, there are also other parameters, like number of WDM channels in the system, average link length and call-blocking probability (due to the wavelength continuity constraint) that come into play, making it difficult to infer proper results without a thorough analysis. We summarize the efficiency issue associated with the path threshold through the following question: how does the path creation threshold influence the efficiency of Optical IP Switching? The answer to this question will be provided by the simulation results in section 4.2, where we have analyzed the effect of different thresholds on the OIS network performance. In a practical implementation, we could see the threshold value dynamically adapted to the node's hardware availability. If, for example, there are many unused optical ports, while little routing processing power is available, the threshold could be lowered to increase the number of cut-through paths. Having different thresholds through the network however raises some issues that we will discuss in section 3.6.

The third parameter, the *length of the observation time*, regulates the overall dynamics of the OIS architecture. Its value should be short enough to follow relevant changes in the IP traffic pattern but long enough to be statistically meaningful and reduce the negative effects generated in the UDP and TCP transport protocols to acceptable levels (this issue is further discussed in section 3.9).

3.4 Optical Path extension

Since Optical IP Switching bases the optical provisioning only on local decisions, all the newly generated paths only involve three nodes: a source, a transparent switching node and a destination. The optical path extension process allows the paths to be extended to more nodes, increasing the number of transparent hops on each path.

Path extension, similar to the path creation, is also based on local analysis, therefore each node can only extend the path by one hop, either upstream or downstream. It also uses the same observation process of path creation. At decision time however, the node recognizes that one of the interfaces, represented by one of the (i,j) matrix cell indexes (the “In-A Flow” index in figure 3.6), is already involved in a cut-through path. If the "i" interface is the source of an existing path, the node will perform an upstream extension. If the "j" interface is the destination of an existing path, the extension will be towards the downstream node. This last case is represented in figure 3.10. In the example, the interface represented by the index “In-A Flow” is the destination interface for the existing flow $D \rightarrow A \rightarrow C$ and also the incoming interface (i.e. the source from node’s C perspective) of the flows that should be extended to G.



Node C’s aggregation matrix

In / Out	In-A	In-H	In-G	In-A Flow
Out-A	■	...	...	■
Out-H	...	■	...	...
Out-G	...	...	■	> Th

Figure 3.10: Path extension process

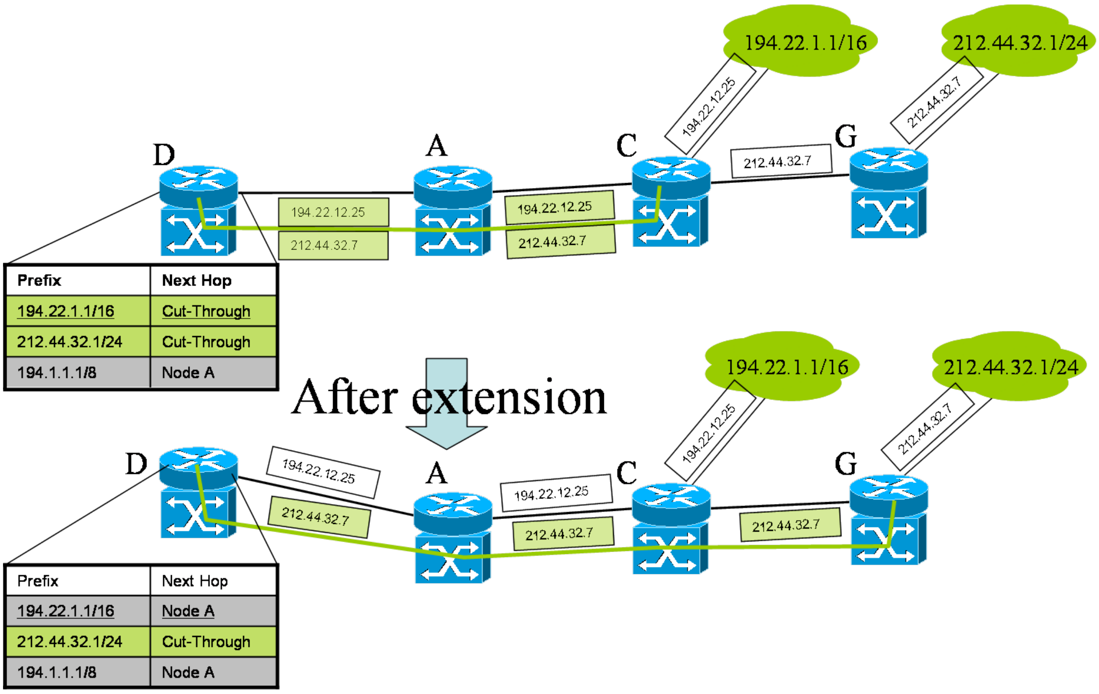
The extension procedure consists on node C reconfiguring the optical switch to transparently connect the incoming “In-A Flow” port to an outgoing port directed towards node G, creating a transparent $D \rightarrow A \rightarrow C \rightarrow G$ path from the existing $D \rightarrow A \rightarrow C$.

Although the extension follows the same traffic observation phase used for path creation, it accomplishes an additional prefix filtering purpose, by selecting a subset of the prefixes switched by the original paths. Only this subset is carried by the new extended path, while the remaining data is excluded from it and will be routed through the default IP links.

When node C decides to extend the existing $D \rightarrow A \rightarrow C$ path, following one of the extension algorithms described in the next section, it sends a signaling message to the downstream node requesting its availability to extend the path. Since the wavelength in the existing path is already defined, the downstream node cannot propose an alternative wavelength. This constraint can be released if node C has wavelength conversion capability.

Unlike in path creation, only the downstream node (or the one upstream, in the case of upstream extension) needs to cooperate to physically extend the path. After the extension is created, node C sends the new list of switched prefixes to source node D, which will update its routing table accordingly. The rest of the nodes only receive the necessary information to update the entry in their dynamic path's list (generally only the "path" field).

Figure 3.11 shows the prefix filtering associated with the path extension and the update of the IP table at the source. While in the original path the dynamic link carries the traffic destined to both networks 194.22.1.1/16 and 212.44.32.1/24, after the extension the first prefix needs to be filtered out. Node C in fact is transparently bypassed by the link and would not be able to extract the packets directed to 194.22.1.1/16.



Prefix	Next Hop
194.22.1.1/16	Cut-Through
212.44.32.1/24	Cut-Through
194.1.1.1/8	Node A

Prefix	Next Hop
194.22.1.1/16	Node A
212.44.32.1/24	Cut-Through
194.1.1.1/8	Node A

Figure 3.11: IP table update for the path extension

This example shows the trade-off between the length of the optical path and the amount of data carried in it. An optical cut-through path can in fact aggregate together only packets sharing a common path. When the existing path is extended, statistically, only a subset of the original packets will share the new longer path, diminishing the amount of data transported by the optical channel and consequently the channel efficiency. On the other hand longer cut-through paths increase the number of transparent hops, enhancing the cost-saving potentials of optical switching. From this perspective, the extension algorithm has the task of optimizing the cost-efficiency problem delineated by this trade-off.

The extension mechanism raises the following question: how does the extension trade-off affect the performance of the Optical IP Switching architecture? The answer to this question will be provided in section 4.2.1, where we analyze through simulations the behavior of the two extension algorithms we propose in the next sections.

3.4.1 Extension algorithms

The extension algorithm is used to determine whether an existing optical path should be further extended to an adjacent node. We describe here two algorithms derived during this work: the first is based on an absolute threshold, the second on a relative threshold.

Absolute threshold algorithm

The absolute threshold algorithm is quite simple and is derived from the path creation algorithm; if the traffic associated with the cell of the aggregation matrix is above a pre-established "path extension threshold", the path is extended. Although this algorithm works fine for the path creation, it raises some issues in the case of path extension. Figure 3.12 illustrates the problem.

In figure 3.12.A, the traffic in the original path $D \rightarrow A \rightarrow C$ fills 80% of the total channel rate (set, for example, at 10Gbps). Node C can extend the path either towards F or G; traffic to F amounts to 6 Gbps, while to G is 2 Gbps. C tries to extend the path first towards F, as the correspondent aggregation matrix cell reports higher traffic. If F cannot accept the extension however, for example because it does not have OIS capabilities, C tries to extend it towards G, since the traffic is above the extension threshold (set to 1 Gbps). In this case, depicted in figure 3.12.B, after the path is extended to G it can only carry 2 Gbps of traffic, while the remaining 6 Gbps have to be removed from the cut-through path to be routed hop-by-hop on the default IP links. In this case the extension might not be economically efficient,

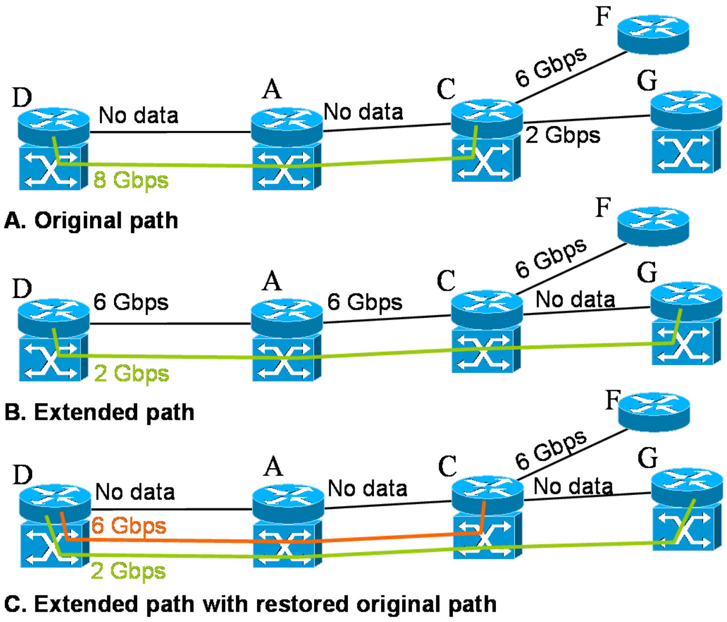


Figure 3.12: Path extension with absolute threshold

as the amount of traffic switched by the optical cut-through path after the extension is much lower than the traffic switched before.

This issue cannot be solved by simply using a path extension threshold higher than the path creation, because this would reduce the switching capability of OIS, by favoring path creation over path extension.

A better approach instead, after the path has been extended, is to restore the original path on a different wavelength, if the data removed from the extension path is above the path creation threshold. This is shown in figure 3.12.C, where a new path is allocated to carry the data that could not be included in the extended path.

The practical implementation of the this function, which we called "path restoration", is not straightforward, as it partly conflicts with our principle of local decision making. The local decision mechanism in fact would not allow node C to create a new end-to-end path $D \rightarrow A \rightarrow C$, since a node can only create a path between its upstream and downstream neighbors. However C could send a trigger message to node A, suggesting that an early provisioning decision be made regarding the path $D \rightarrow A \rightarrow C$.

The case is more complicated when the original path is already an extended path, involving four or more nodes. Figure 3.13 shows the case where the path $D \rightarrow A \rightarrow C \rightarrow G$ is extended to $D \rightarrow A \rightarrow C \rightarrow G \rightarrow H$. In this case node G needs to send the trigger message to node A, which originally created the path. After A has re-established the $D \rightarrow A \rightarrow C$ path, it can trigger C to consider the extension towards G.

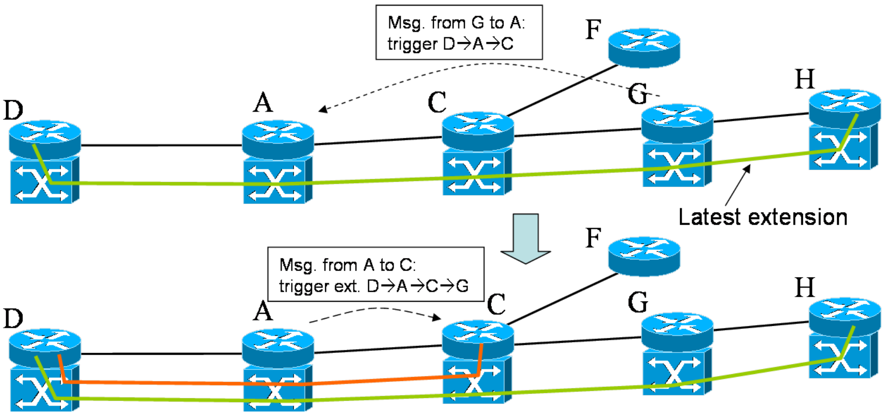


Figure 3.13: Example of path restoration

The process illustrated helps the node to quickly restore the original path when needed, increasing the switching performance of the absolute threshold algorithm.

Relative threshold algorithm

The second path extension algorithm we propose is based on a relative threshold. This algorithm directly targets the trade-off introduced in section 3.4 between the path length and the amount of data it carries.

The threshold value is derived by directly applying the constraint that the path extension should not increase the amount of data routed through the default links. The formula for the absolute threshold, expressed as a percentage is:

$$Threshold = \frac{N-2}{N-1} \cdot 100 \tag{3.1}$$

where "N" is the number of nodes committed in the cut-through path before the extension. The demonstration of equation 3.1 is straightforward. Figure 3.14 shows how data is switched and routed before and after an extension involving N nodes. Before the extension, the total amount of data "A" in the path is split at the output of node N: "x" is the part that can be extended (going towards an OIS-capable node), and "A-x" the remaining part.

Before the extension, node 1 routes all data "A" into the dynamic optical path, which is switched transparently up to node N. Node N will then route "A", splitting it into two flows: "x" and "A-x". We can see from figure 3.14.A that the total amount of data routed by the nodes in the networks before the extension is:

$$RoutData_{bef} = A + (A - x + x) + (A - x) + x = 3A \tag{3.2}$$

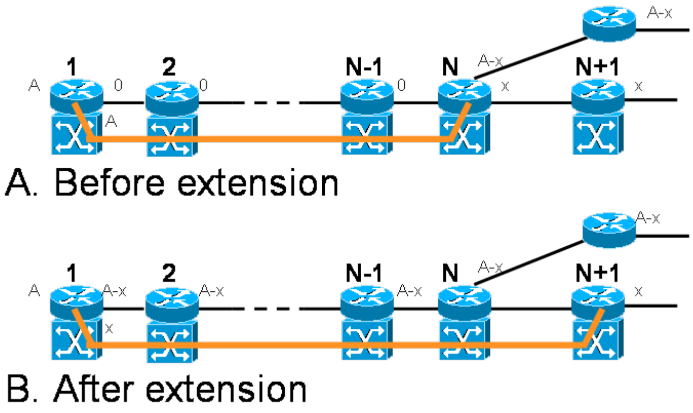


Figure 3.14: Calculation of the relative threshold algorithm

After the extension, only the quantity "x" of data addressed towards node N+1 can be injected into the extended path, while the remaining "A-x" needs to be routed hop-by-hop through all N nodes. The total amount of data routed after the extension, following figure 3.14.B, is:

$$RoutData_{aft} = (A-x+x)+(N-2)(A-x)+(A-x)+(A-x)+x = A+x+N(A-x) \quad (3.3)$$

The constraint we want to apply is that the data routed after the extension is not higher than the data routed before it:

$$RoutData_{Aft} \leq RoutData_{Bef} \Longrightarrow A+x+N(A-x) \leq 3A \quad (3.4)$$

which simplifies into:

$$\frac{x}{A} \geq \frac{N-2}{N-1} \quad (3.5)$$

Formula 3.1 can be directly derived from the inequality 3.5.

With the relative threshold algorithm the path extension does not depend on the absolute amount of data that can be extended, but on its value with respect to the data already in the dynamic path. In contrast to the absolute threshold algorithm, this algorithm tends to optimize the channel usage rather than the total amount of switched data. In section 4.2.1 we report the comparison results for the two extension algorithms we have presented here.

3.5 Optical path cancellation

Existing cut-through paths are deleted in order to free resources like interfaces, optical ports and wavelength channels that have become under-exploited due to traffic changes. Variations in the traffic pattern in fact change the amount of data destined towards the prefixes switched by dynamic paths. Such variations could reduce the traffic switched optically, decreasing the channel efficiency below acceptable levels.

The path deletion process makes use of the information stored in the aggregation matrix during the observation time. At decision time the node checks which cells, among those corresponding to interfaces that are sources or destinations of dynamic paths, present data values below a pre-established “path cancellation threshold”. This threshold needs to be lower than the path creation and extension thresholds, and like those, can be dynamically updated to optimize the use of optical versus IP resources available.

The path cancellation message is always sent by the path source or destination and is processed by each node before being relayed downstream (or upstream in case the cancellation message originated from a destination node). Any other node other than the source and destination in fact, being transparent switches, cannot collect any information about the traffic in the cut-through path.

Every time a dynamic path is deleted, the traffic that was being switched returns to be routed hop-by-hop through the default links. Although such traffic increases the occupancy of routing resources, the freed optical resources can be reused to allocate other traffic. If the path cancellation threshold is set lower than the path creation, the new optical path will carry more traffic than the path deleted (as far as there is traffic available to create a new optical path).

Looking back at the functional diagram in figure 3.3, we can now better understand the order of the path allocation sequence. Path cancellation is operated before the creation or extension of existing paths so that eventual unexploited resources can be relaxed and reused immediately for more productive cut-through paths. The extension is processed before the creation of new paths to favor the creation of longer trails. For each path in fact, two of the nodes (the source and destination) always consume routing resources equal to the data switched. Therefore, the higher the number of nodes involved in the path, the higher the amount of switched data with respect to the routed data (although this advantage is mitigated by the tradeoff described in section 3.4). Moreover, if wavelength conversion is not available, the extension operation is more restrictive than the creation, because a path can only be extended using the same wavelength as the original path. A new path instead can be created over any available wavelength. If the path creation was operated before the extension, the creation of new paths might occupy critical optical resources, precluding the extension of other paths. This is shown in the example illustrated in figure 3.15, where an existing path $A \rightarrow B \rightarrow D$ is active on wavelength λ^1. At decision time, node D creates the new path $C \rightarrow D \rightarrow E$ also on wavelength λ^1, precluding the extension of $A \rightarrow B \rightarrow D$ into $A \rightarrow B \rightarrow D \rightarrow E$ because λ^1 is already in use on the link $D \rightarrow E$. Had the precedence been given to the path extension, both operations could have been completed successfully.

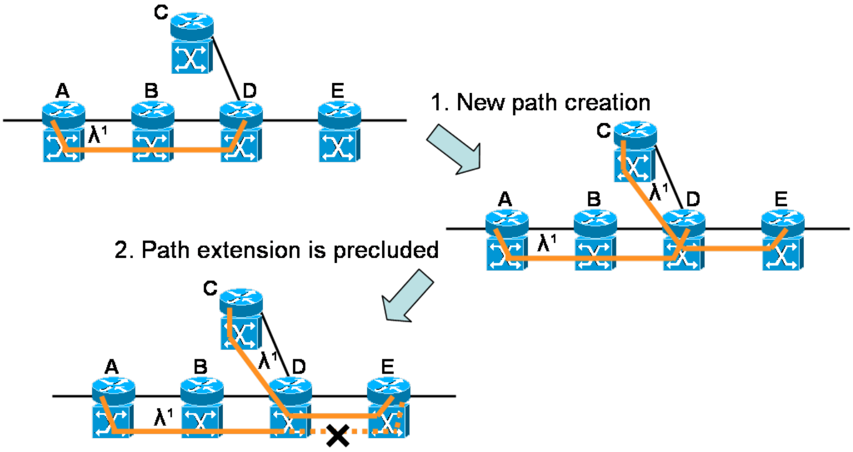


Figure 3.15: Path extension precluded by previous path creation

3.6 Local decision with shared acknowledgment

In the description of the path creation, extension and cancellation mechanisms we have emphasized the fact that the decision process is local and distributed. Each node can only decide to create paths involving its immediate downstream or upstream neighbors (or both), prior their acknowledgment. Since different nodes can apply different policies (considering the current interdomain routing mechanism), they could use different values for path creation, extension and cancellation threshold. Since however, due to the transparency of cut-through paths, the nodes cannot examine the traffic crossing their optical switches, the policies of some nodes might not be respected.

The problem is clarified by the following example. Considering figure 3.16, we set node A's path creation threshold to 1 Gbps and its cancellation threshold to 800 Mbps, while its extension follows the absolute algorithm with threshold equal to 1 Gbps. C uses the same threshold as A for path creation, but bases its extensions on the relative threshold algorithm described in section 3.4.1. If the aggregate rate of the current switched flows amounts to 1.5

Gbps, for example, and 750 Mbps of traffic are directed towards G, node C (according to equation 3.1) will extend the path towards G. This decision is however in contrast with the policy of node A, which, after the extension, would be using its optical ports to switch an amount of data lower than its cancellation threshold value of 800 Mbps.

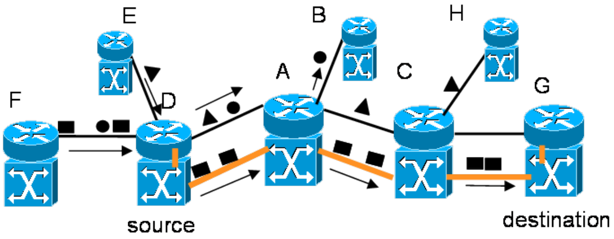


Figure 3.16: Example of OIS extended path

A similar case can occur during path cancellation; different cancellation thresholds set different limits to the minimum amount of data allowed in each optical path. If the cancellation decision is only made by the source or destination nodes, the differences in the nodes' policies would not be respected.

The problem can be solved by allowing all the nodes involved in the path to take part in the extension or cancellation decisions. Considering the extension process (figure 3.16), node C could inform D and A about the intention to extend the existing path towards node G. The information sent by C summarizes the parameters necessary for the extension decision, like data switched before and after the flow. Each node can process the information to evaluate if the extension is consistent with its own policies. If C does not receive any negative acknowledgment it proceeds with the path extension as described in section 3.4. However, a negative acknowledgment received from any node involved in the path, would be treated as a veto, aborting the extension process.

Path cancellation can be handled similarly. In this case, the source (or destination) node periodically sends downstream (or upstream) information about the state of the flows in the path. With this mechanism any node can evaluate if the path should still be in place and, if not, can send a request to the source node, asking to cancel the path. The source will then send the official path cancellation message to all the nodes involved in the path.

In both path extension and cancellation phases, one negative reply is enough to abort the extension process or trigger the removal of an entire cut-through path. A less dramatic approach, depicted in figure 3.17, would see an existing path being altered instead of being completely removed. In the first example, if node D decides to cancel the path, its request could be fulfilled by splitting the path in two. In the second case, if F takes the decision, the path could be reduced back to node E. Any modification in the existing path however, requires an update of the list of prefixes to be routed into the path, further increasing the complexity of the mechanism. For this reason, in our implementation we have followed the original OIS approach of deleting the path completely, leaving the original distributed OIS mechanism to rebuild it as appropriate.

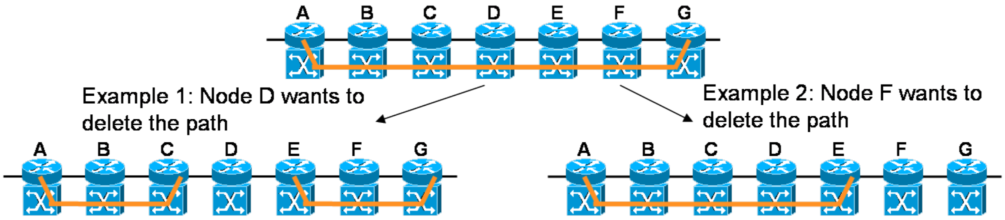

A
B
C
D
E
F
G
Example 1: Node D wants to delete the path
Example 2: Node F wants to delete the path
A
B
C
D
E
F
G
A
B
C
D
E
F
G

Figure 3.17: Path modification approach

In a practical implementation, a hybrid approach combining the advantages of both the shared acknowledgment and the original OIS method could be used.

In the intradomain, where nodes are willing to cooperate to increase the overall network performance, the original OIS method could be used, with threshold values properly coordinated among the nodes.

In the interdomain instead, the shared acknowledgment mechanism could be implemented to guarantee full compliance with the OIS routing policies by every node.

3.7 Flow re-classification

We have first introduced the flow re-classification in the functional diagram of figure 3.3. The idea is to periodically update the list of prefixes switched in a dynamic path, in order to increase its lifetime, giving more stability to the OIS network.

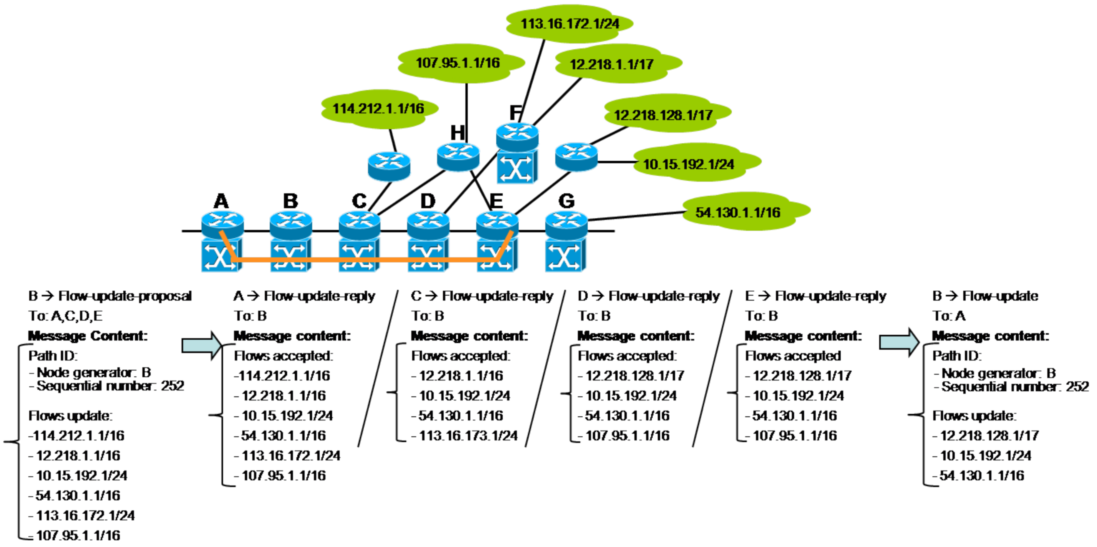
When creating a new path, a node sends a set of selected prefixes upstream (those carrying an amount of data higher than the prefix threshold described in section 3.2.2) to indicate which packets to inject into the transparent path. At each path extension the prefix list is modified by a filtering process that keeps only the subset of prefixes routed towards the destination node.

However, as traffic patterns change over time, some prefixes that were originally excluded from the initial path can rise above the prefix threshold. Such prefixes are not switched into the dynamic path because the source node is not aware of the traffic change. A similar situation occurs for prefixes that, due to route changes determined by the default IP routing protocol, do not get switched by the path although they might share the same route. The fact that packets that could avail of the transparent path are instead routed hop-by-hop constitutes an inefficiency.

In a situation where switched traffic decreases because of changes in the traffic distribution among prefixes, the OIS techniques of path creation, extension and cancellation, previously described, only react when the traffic drops below the cancellation threshold. In that case the existing path is completely deleted and a new one can be created with an updated list of prefixes. The whole process however might be quite slow, especially if the path involved more than three nodes. The time needed for the traffic to drop below the cancellation threshold, and for the node to recreate the path, step-by-step, following the OIS distributed approach, could considerably lower the average channel efficiency.

The flow re-classification mechanism intervenes in these situations to ensure higher exploitation of the dynamic paths.

The mechanism can be triggered by any node in the path at the expiration of the update timer, which can be shorter or equal to the decision timer.



B → Flow-update-proposal
To: A,C,D,E
Message Content:
Path ID:
- Node generator: B
- Sequential number: 252

Flows update:
-114.212.1.1/16
- 12.218.1.1/16
- 10.15.192.1/24
- 54.130.1.1/16
- 113.16.172.1/24
- 107.95.1.1/16

A → Flow-update-reply
To: B
Message content:
Flows accepted:
-114.212.1.1/16
- 12.218.1.1/16
- 10.15.192.1/24
- 54.130.1.1/16
- 113.16.172.1/24
- 107.95.1.1/16

C → Flow-update-reply
To: B
Message content:
Flows accepted:
- 12.218.1.1/16
- 10.15.192.1/24
- 54.130.1.1/16
- 113.16.173.1/24

D → Flow-update-reply
To: B
Message content:
Flows accepted:
- 12.218.128.1/17
- 10.15.192.1/24
- 54.130.1.1/16
- 107.95.1.1/16

E → Flow-update-reply
To: B
Message content:
Flows accepted
- 12.218.128.1/17
- 10.15.192.1/24
- 54.130.1.1/16
- 107.95.1.1/16

B → Flow-update
To: A
Message content:
Path ID:
- Node generator: B
- Sequential number: 252

Flows update:
- 12.218.128.1/17
- 10.15.192.1/24
- 54.130.1.1/16

Figure 3.18: Example of flow re-classification

In figure 3.18 the flow re-classification is triggered by node B, which has noticed in its aggregation matrix the presence of traffic from A to C on the default link. Since its dynamic path list shows that there is a transparent path joining A to C ($A \rightarrow B \rightarrow C$ is included in $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$), B considers the prefixes in the cell above the prefix threshold as candidates for the flow update. Since the path goes transparently down to E, the prefixes corresponding to the packets that do not need to pass through node E (e.g., 114.212.1.1/16, 12.218.1.1/17, 113.16.172.1/24, 107.95.1.1/16 in the figure) have to be filtered out. Node B however does not possess any information on how packets are routed by its neighbors and needs their cooperation to get the proper list of prefix updates, which will be sent to the source node. Node B sends a "flow-update-proposal" message to every node involved in the path, including in the message the "path ID" and the list of potential prefixes for the update. Every node receiving the message will filter out those prefixes that are not directed towards the path, and those not complying with its own policies, sending a "flow-update-reply" message back to B. At this point, B will select the subset of prefixes obtained by the intersection of the prefix sets included in all the "flow-update-reply" messages received, and sends it to the source node A with a "flow-update" message.

Figure 3.18 illustrates an interesting case. We see that node B has summarized the prefixes 12.218.1.1/17 and 12.218.128.1/17 into 12.218.1.1/16, a legitimate technique used by network administrators to reduce the size of routing tables. When node D receives the "flow-update-proposal" message, it realizes that only part of the 12.218.1.1/16 can be considered for update; in that case it splits the prefix, sending back only the subnet 12.218.128.1/17.

Some additional mechanisms could also be implemented to reduce the protocol overhead. For example, if the nodes keep track of the prefixes that were previously filtered out because of path extension, it can avoid including those in the flow-update-proposal. In addition, if most of the prefixes suggested by node B are suitable to C, C could include in its reply only the prefixes it wants to filter out, in order to minimize the size of the reply message.

3.8 Ports and link discovery

The port and link discovery functions aim at determining, in a practical OIS network implementation, which switch ports are connected to which router interfaces, and which neighbors (on which wavelengths) can be reached through each transparent port.

The general problem is illustrated in figure 3.19, where each node is connected to different neighbors through WDM links. Unlike opaque architectures (like SDH for example), where each port is always electrically terminated, the number of I/O interfaces is much lower than the number of transparent ports, and only a small number of ports can be electrically terminated at each time.

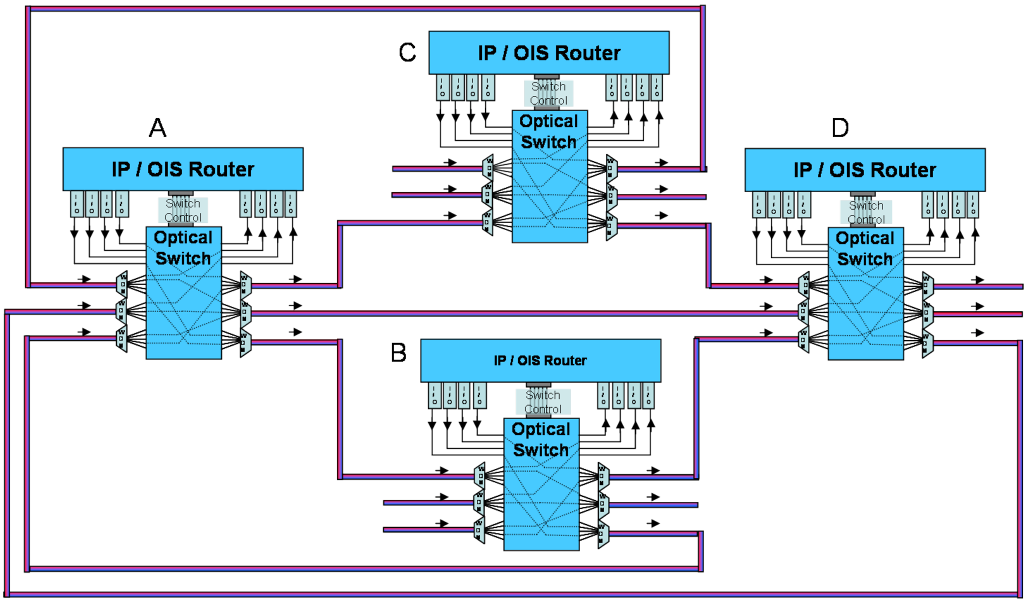


Figure 3.19: Port and link discovery issue in transparent networks

The core of our approach to port and link discovery can be summarized by the following research question: is it possible for transparent networks to be self-configuring? The answer to this question will be discussed in the following sections, where we propose two protocols giving separate solutions to the port and link discovery problems. In section 5.3 we will demonstrate the feasibility of our auto-configuration approach through simulations and testbed results.

3.8.1 Port discovery algorithm

Determining the connectivity between switch and router interfaces is a problem that only arises in semi-transparent networks, where a transparent optical switch is connected to an electrical router. Although a trivial solution would see a manual entry of the interface connectivity to the switch ports in a configuration file, this solution does not seem suitable and scalable for a scenario where a transparent node may have hundreds of ports and tens of interfaces. The manual process would in fact require hours of work and coordination of more people (in case router and switch are located in different places), and would be prone to human errors.

Automatic configuration is already widely deployed in electronic communication equipment. In optical systems however, since optical signal processing is still many years away, automatic configuration requires that optical signals are terminated and converted into electrical to be analyzed by the router. The limiting factor is that, both for economical and practical reasons, there is only a limited number of router interfaces, and only a subset of ports can be terminated at the same time.

We have developed a smart algorithm that executes the discovery process, taking advantage of the capability of photonic switches to execute multiple input-to-output port connections at the same time. The algorithm can be divided in two parts: the first collects all the information regarding the relative distance, in terms of port number, between each of the interfaces; the second determines the absolute position of a single TX-RX interface pair, and uses the information collected during the first phase to determine the position of the remaining interfaces. At the end of this section, in figures 3.25 and 3.26, we report a practical example of port discovery, in order to clarify the mechanisms of the algorithm we present.

Figure 3.20 shows the state of the connections during the first iteration of the discovery process.

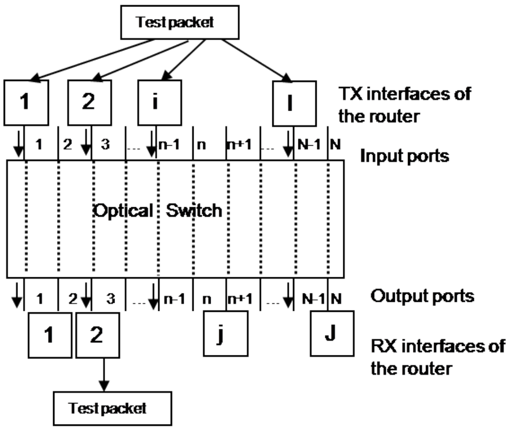


Figure 3.20: First iteration of the parallel algorithm

The node connects each input port "n" to the output port "n+k", where "k" is a number that is incremented after each iteration. For the first iteration, k is equal to "0": input port "1" is connected to output port "1", input port "2" to output port "2", and so forth. After configuring the switch, the node sends a "TEST" message on each TX interface, while the RX interfaces wait for incoming messages. In each "TEST" message is encoded the name (i.e. the interface number) of the TX interface transmitting the message. When RX interface "j" receives a message that was transmitted by TX interface "i", the value "k" is stored in the (i,j) position of the K matrix of dimensions I x J (i.e., the total number of TX and RX interfaces).

At each new iteration all the connections are re-established with "k" incremented by 1 unit. During the second iteration, for example, input port "1" is connected to output port "2", and in general, input port "n" to output port "n+1". The last input port "N" is connected to output port “1”, as illustrated in figure 3.21, i.e. barrel rotated. We can generalize this rule expressing the output ports as a function of the input port “n”:

$$\begin{cases} OutputPort(n) = (n+k)\,mod(N) & if(n+k \neq N) \\ OutputPort(n) = N & if(n+k = N), \\ where\,n\,is\,the\,input\,port \end{cases} \tag{3.6}$$

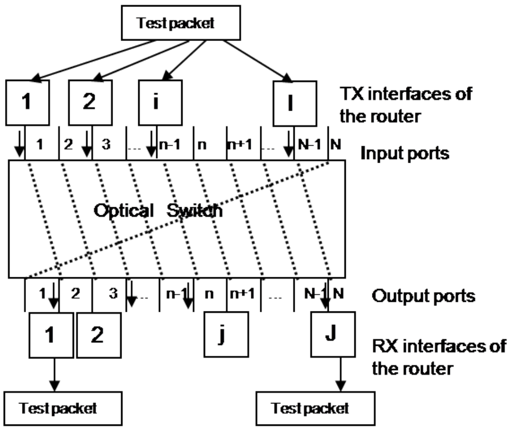


Figure 3.21: Second iteration of the parallel algorithm

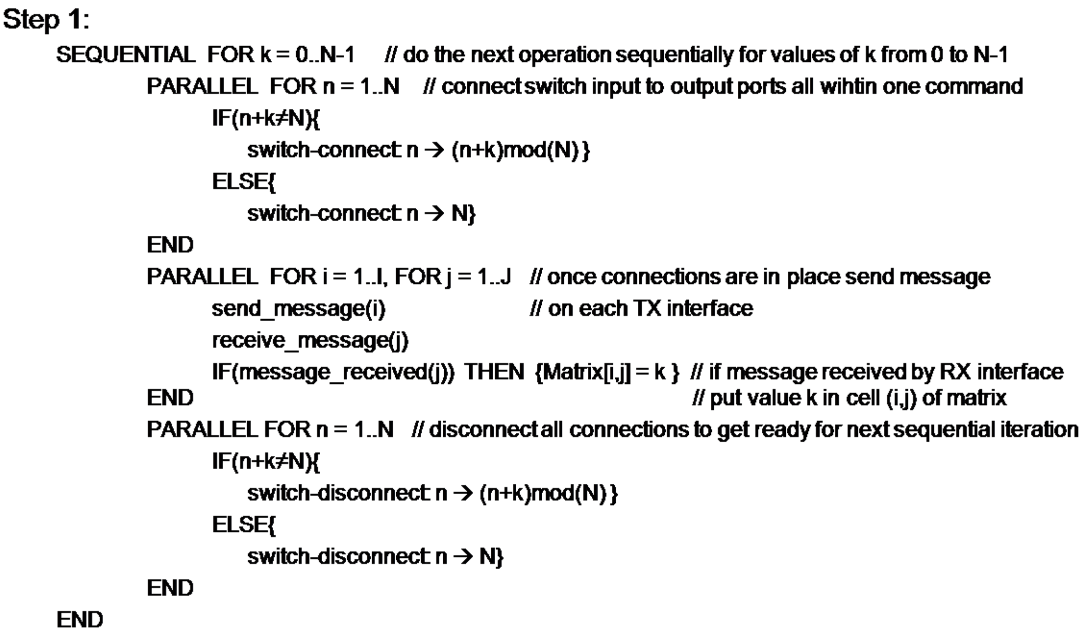
The process is iterated until “k” is equal to “N-1”, that is when all the possible connections have been tested. The first phase of the algorithm always terminates in N iterations. At this point the I x J matrix is completely filled in with values representing the relative distance between the ports connected to the interfaces.

The pseudocode for this first phase of the algorithm is reported in figure 3.22. In the code the terms “SEQUENTIAL” and “PARALLEL” emphasize how part of the instructions are operated sequentially and part in parallel.

The aim of the second phase of the algorithm is to determine the absolute location of a (TX,RX) interface pair, and to use it together with the values stored in the matrix to calculate the position of the remaining interfaces. This phase begins by picking a value “k^1” from an arbitrary cell of the matrix (even though choosing the value that most often appears would speed up the process by increasing the chances of receiving a message).

Each new iteration tests the connections one at a time, starting from input port “1”. Each input port “n” is connected in turn to output port “$n+k^1$”, following rule 3.6, until a message is delivered. Once one message is received, the positions of the i and j interfaces is uniquely identified. These operations are illustrated in part A of the pseudocode in figure 3.23.

The positions of the newly discovered TX and RX interfaces are used in the final part of the algorithm as a reference point to calculate the location of the remaining interfaces



```
Step 1:
    SEQUENTIAL  FOR k = 0..N-1     // do the next operation sequentially for values of k from 0 to N-1
            PARALLEL  FOR n = 1..N    // connect switch input to output ports all wihtin one command
                    IF(n+k≠N){
                        switch-connect n → (n+k)mod(N) }
                    ELSE{
                        switch-connect n → N}
            END
            PARALLEL  FOR i = 1..I, FOR j = 1..J   // once connections are in place send message
                    send_message(i)                  // on each TX interface
                    receive_message(j)
                    IF(message_received(j)) THEN {Matrix[i,j] = k }  // if message received by RX interface
            END                                                       // put value k in cell (i,j) of matrix
            PARALLEL FOR n = 1..N   // disconnect all connections to get ready for next sequential iteration
                    IF(n+k≠N){
                        switch-disconnect n → (n+k)mod(N) }
                    ELSE{
                        switch-disconnect n → N}
            END
    END
```

Figure 3.22: Pseudocode for the first part of the port discovery algorithm

without further switch activations. Taking as a reference point the matrix cell (i,j) (figure 3.24), corresponding to the newly discovered (TX,RX) interface pair, the algorithm can use the "k" value stored in the generic cell (i_a, j_b) to determine the position (i.e., the switch port where they are connected) of TX interface "i_a" or RX interface "j_b". The cells (i,j) and (i_a, j_b) however must always lie either on the same row or on the same column, in order to keep a TX or RX interface as common reference. If (i,j) and (i_a, j_b) lie in the same row ($i = i_a$), the (i_a, j_b) value can be used to discover the switch output port (i.e., the position) where RX interface "j_b" is connected:

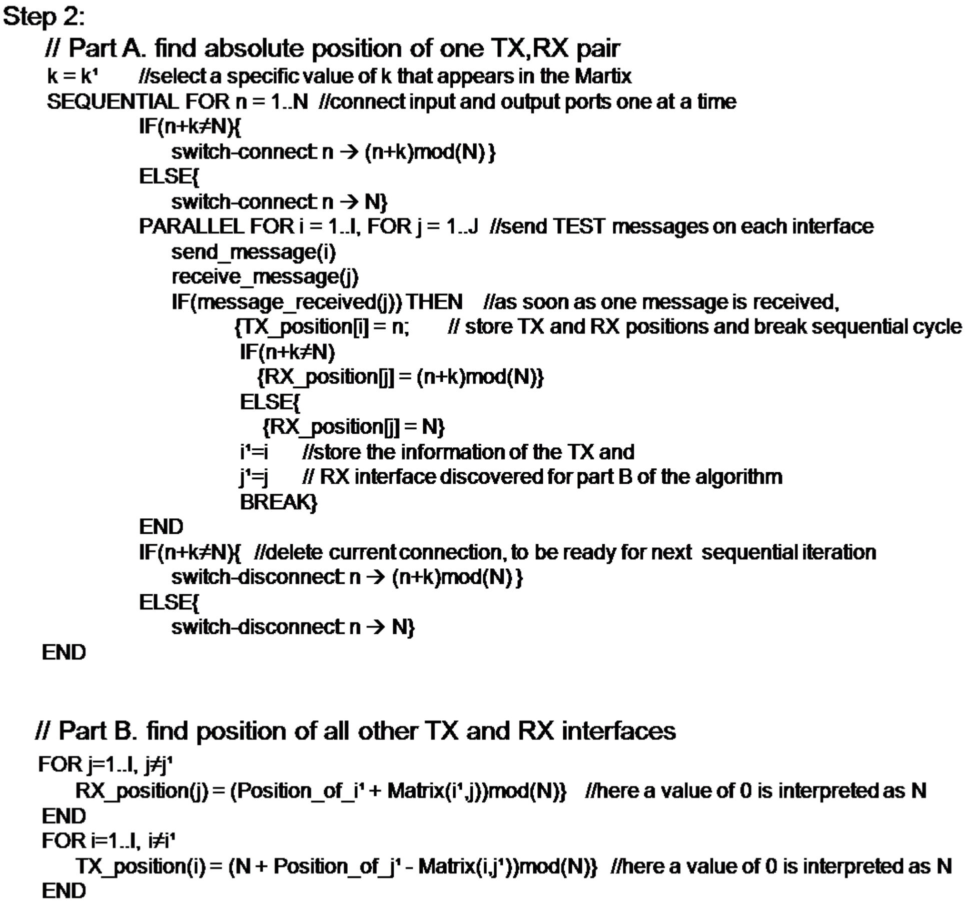
$$Position\_of\_j_b = (Position\_of\_i + k(i, j_b))\ mod(N) \tag{3.7}$$

where position "0" is identified with port "N" in the switch.

If instead (i,j) and (i_a, j_b) are in the same column ($i = j_b$), the value stored in (i_a, j_b) can be used to determine the switch input port (i.e., the position) where TX interface i_a is connected:

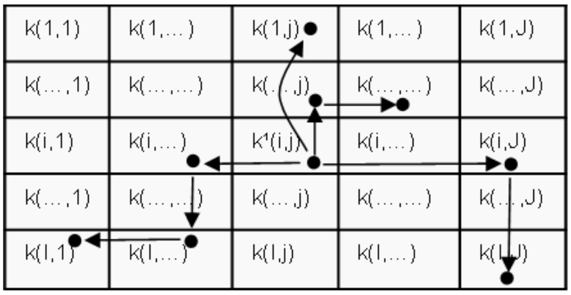
$$Position\_of\_i_a = (N + Position\_of\_j - k(i_a, j))\ mod(N) \tag{3.8}$$

At each step the newly discovered position of (i_a, j_b) can be used as a reference point for the next step, as shown in figure 3.24. These operations are illustrated in part B of the pseudocode in figure 3.23 (for simplicity in this case the walk through the matrix is operated



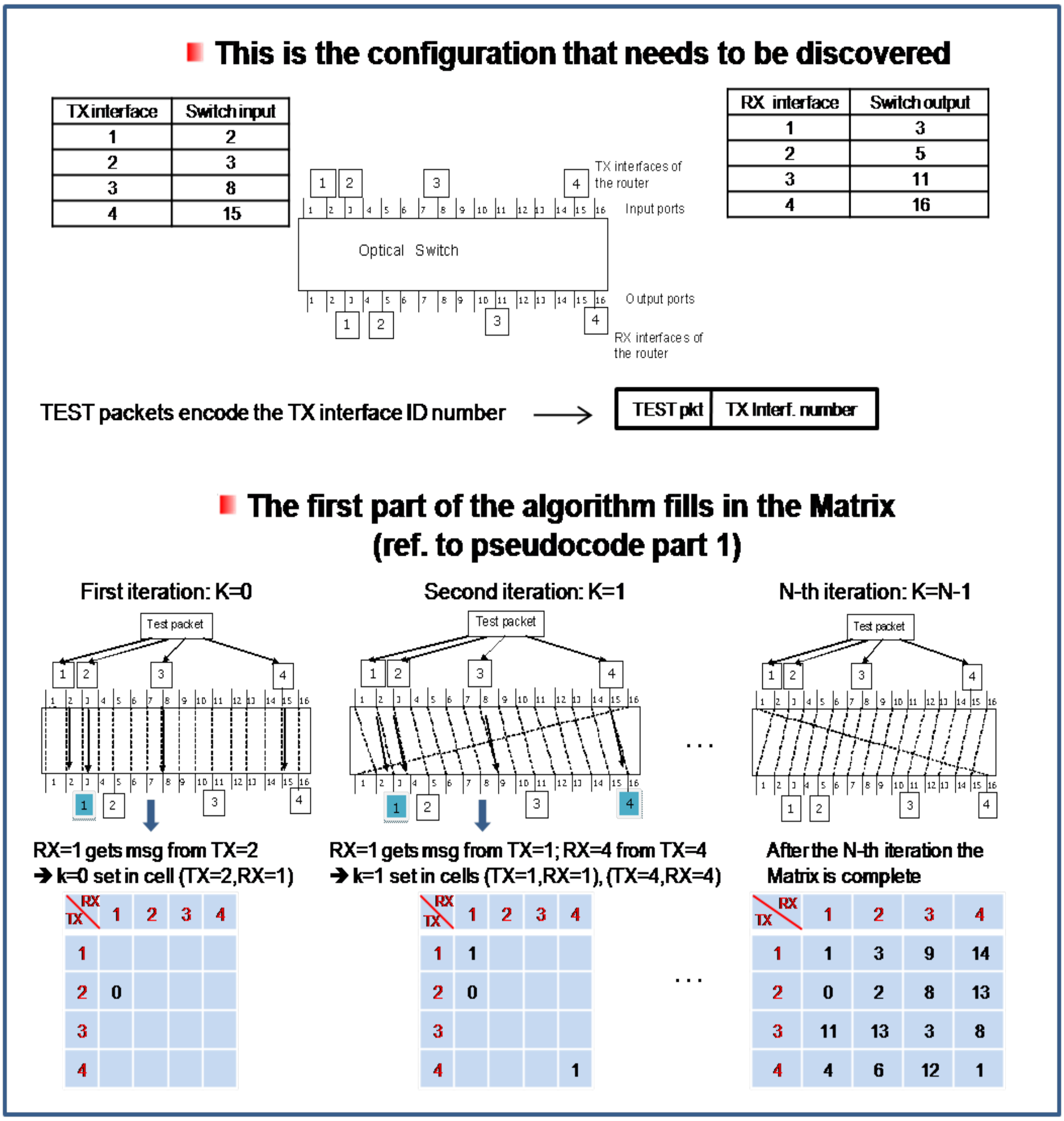
```
Step 2:
    // Part A. find absolute position of one TX,RX pair
    k = k'      //select a specific value of k that appears in the Martix
    SEQUENTIAL FOR n = 1..N  //connect input and output ports one at a time
            IF(n+k≠N){
                switch-connect n → (n+k)mod(N) }
            ELSE{
                switch-connect n → N}
            PARALLEL FOR i = 1..I, FOR j = 1..J  //send TEST messages on each interface
                send_message(i)
                receive_message(j)
                IF(message_received(j)) THEN    //as soon as one message is received,
                        {TX_position[i] = n;     // store TX and RX positions and break sequential cycle
                         IF(n+k≠N)
                           {RX_position[j] = (n+k)mod(N)}
                         ELSE{
                           {RX_position[j] = N}
                         i'=i     //store the information of the TX and
                         j'=j     // RX interface discovered for part B of the algorithm
                         BREAK}
            END
            IF(n+k≠N){  //delete current connection, to be ready for next  sequential iteration
                switch-disconnect n → (n+k)mod(N) }
            ELSE{
                switch-disconnect n → N}
    END

    // Part B. find position of all other TX and RX interfaces
    FOR j=1..I, j≠j'
        RX_position(j) = (Position_of_i' + Matrix(i',j))mod(N)}   //here a value of 0 is interpreted as N
    END
    FOR i=1..I, i≠i'
        TX_position(i) = (N + Position_of_j' - Matrix(i,j'))mod(N)}  //here a value of 0 is interpreted as N
    END
```

Figure 3.23: Pseudocode for the second part of the port discovery algorithm



k(1,1)	k(1,...)	k(1,j)	k(1,...)	k(1,J)
k(...,1)	k(...,...)	k(...,j)	k(...,...)	k(...,J)
k(i,1)	k(i,...)	k'(i,j)	k(i,...)	k(i,J)
k(...,1)	k(...,...)	k(...,j)	k(...,...)	k(...,J)
k(I,1)	k(I,...)	k(I,j)	k(I,...)	k(I,J)

Figure 3.24: K Matrix that stores the values of mutual distances between the TX and RX interfaces



This is the configuration that needs to be discovered

TX interface	Switch input
1	2
2	3
3	8
4	15

RX interface	Switch output
1	3
2	5
3	11
4	16

TEST packets encode the TX interface ID number →

TEST pkt	TX Interf. number

The first part of the algorithm fills in the Matrix (ref. to pseudocode part 1)

First iteration: K=0

RX=1 gets msg from TX=2
➔ k=0 set in cell (TX=2,RX=1)

TX \ RX	1	2	3	4
1				
2	0			
3				
4				

Second iteration: K=1

RX=1 gets msg from TX=1; RX=4 from TX=4
➔ k=1 set in cells (TX=1,RX=1), (TX=4,RX=4)

TX \ RX	1	2	3	4
1	1			
2	0			
3				
4				1

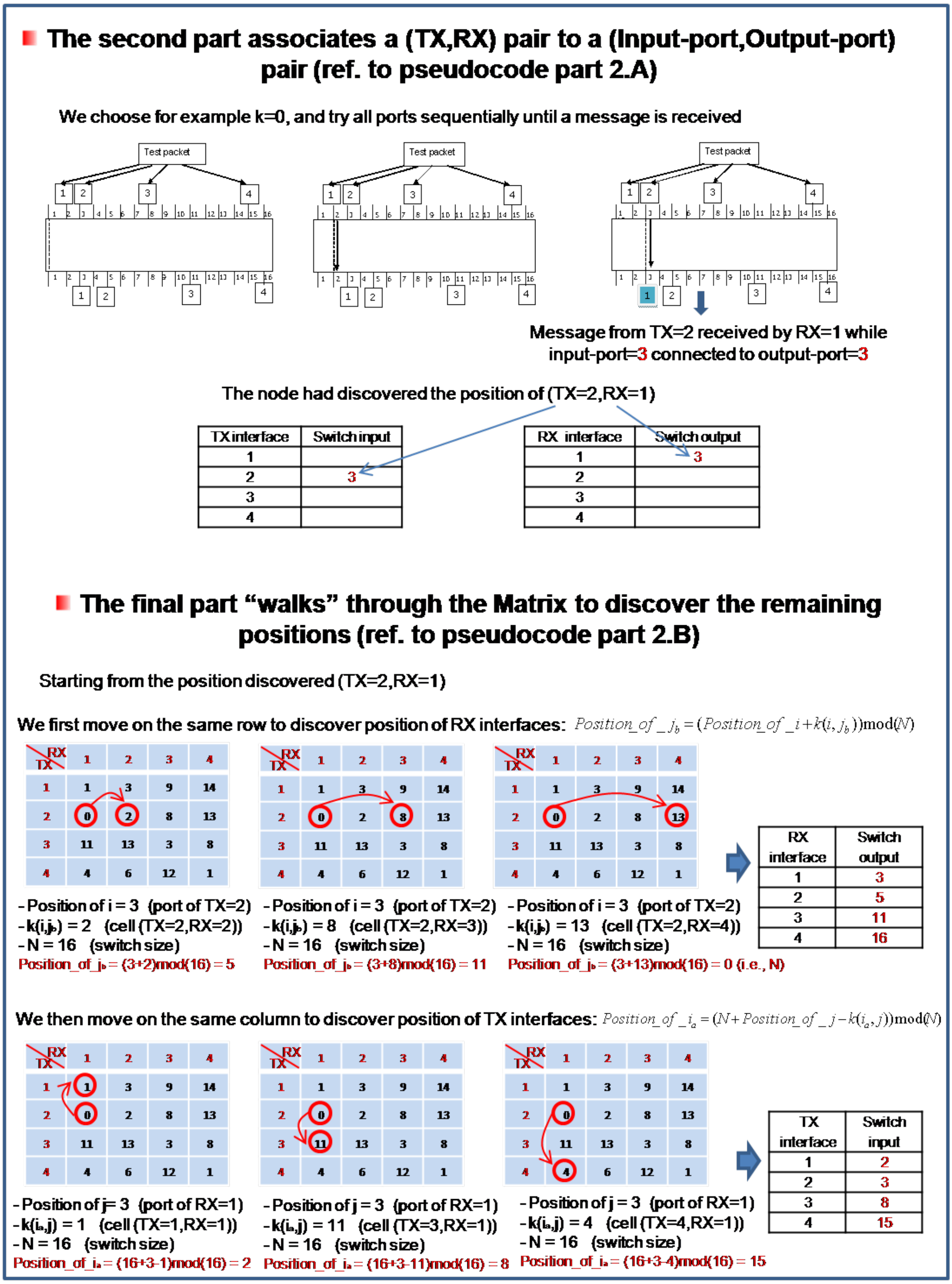
. . .

N-th iteration: K=N-1

After the N-th iteration the Matrix is complete

TX \ RX	1	2	3	4
1	1	3	9	14
2	0	2	8	13
3	11	13	3	8
4	4	6	12	1

Figure 3.25: Practical example of port discovery (part one)



The second part associates a (TX,RX) pair to a (Input-port,Output-port) pair (ref. to pseudocode part 2.A)

We choose for example k=0, and try all ports sequentially until a message is received

Message from TX=2 received by RX=1 while input-port=3 connected to output-port=3

The node had discovered the position of (TX=2,RX=1)

TX interface	Switch input
1	
2	3
3	
4	

RX interface	Switch output
1	3
2	
3	
4	

The final part "walks" through the Matrix to discover the remaining positions (ref. to pseudocode part 2.B)

Starting from the position discovered (TX=2,RX=1)

We first move on the same row to discover position of RX interfaces: $Position\_of\_j_b = (Position\_of\_i + k(i, j_b))\mathrm{mod}(N)$

TX \ RX	1	2	3	4
1	1	3	9	14
2	0	2	8	13
3	11	13	3	8
4	4	6	12	1

- Position of i = 3 (port of TX=2)
- k(i,j_b) = 2 (cell (TX=2,RX=2))
- N = 16 (switch size)

Position_of_j_b = (3+2)mod(16) = 5

TX \ RX	1	2	3	4
1	1	3	9	14
2	0	2	8	13
3	11	13	3	8
4	4	6	12	1

- Position of i = 3 (port of TX=2)
- k(i,j_b) = 8 (cell (TX=2,RX=3))
- N = 16 (switch size)

Position_of_j_b = (3+8)mod(16) = 11

TX \ RX	1	2	3	4
1	1	3	9	14
2	0	2	8	13
3	11	13	3	8
4	4	6	12	1

- Position of i = 3 (port of TX=2)
- k(i,j_b) = 13 (cell (TX=2,RX=4))
- N = 16 (switch size)

Position_of_j_b = (3+13)mod(16) = 0 (i.e., N)

RX interface	Switch output
1	3
2	5
3	11
4	16

We then move on the same column to discover position of TX interfaces: $Position\_of\_i_a = (N + Position\_of\_j - k(i_a, j))\mathrm{mod}(N)$

TX \ RX	1	2	3	4
1	1	3	9	14
2	0	2	8	13
3	11	13	3	8
4	4	6	12	1

- Position of j= 3 (port of RX=1)
- k(i_a,j) = 1 (cell (TX=1,RX=1))
- N = 16 (switch size)

Position_of_i_a = (16+3-1)mod(16) = 2

TX \ RX	1	2	3	4
1	1	3	9	14
2	0	2	8	13
3	11	13	3	8
4	4	6	12	1

- Position of j = 3 (port of RX=1)
- k(i_a,j) = 11 (cell (TX=3,RX=1))
- N = 16 (switch size)

Position_of_i_a = (16+3-11)mod(16) = 8

TX \ RX	1	2	3	4
1	1	3	9	14
2	0	2	8	13
3	11	13	3	8
4	4	6	12	1

- Position of j = 3 (port of RX=1)
- k(i_a,j) = 4 (cell (TX=4,RX=1))
- N = 16 (switch size)

Position_of_i_a = (16+3-4)mod(16) = 15

TX interface	Switch input
1	2
2	3
3	8
4	15

Figure 3.26: Practical example of port discovery (part two)

moving through adjacent cells, and the reference point is always the initial (i,j) cell).

All the interfaces are discovered when at least one cell per row and one cell per column of the matrix have been exploited. The total number of switch activations required by the parallel algorithm is between N+1 and 2N. The parallel algorithm illustrated in this section should be operated during the node startup phase, when no connectivity information between interfaces and switch ports is available. The discovery of additional interfaces that might be added while the node is operational (i.e., after the initial configuration procedure is accomplished), can instead be easily accomplished by serially testing each undiscovered port using an already discovered interface.

Redundancy of information

We can easily notice that part of the information stored in the matrix was not used during the discovery process. It is possible to eliminate this redundancy and save some iterations by re-arranging the discovery process. We could execute the first part of the algorithm until a message is received, and use the “k” value found to determine the interface-to-port relation of a pair (i,j) of TX-RX interfaces in the second part of the algorithm. Then, returning to the first phase, we could stop the iterations when just enough information is collected to complete the discovery process. Such operation however is only safe if none of the ports are connected to a neighboring switch.

An example is shown in figure 3.27, where the left node sends a "TEST" message through TX interface “2” and receives it through RX interface “j”, even if the two interfaces are not directly connected together. Because of the loop accidentally created by switch-2 in fact, the first node erroneously inserts a “k” value of “1” in the matrix position (2,j).

If the node had instead followed the procedure previously described, it would have received two different values for the same matrix cell. Although it would still not be able to recognize which of the two is the correct value, it would realize that an anomaly occurred with cell (2,j) and avoid using it for the calculations.

Cases with N x M switches and I $\neq$ J

A more general implementation, valid for N x M switches with I TX interfaces and J RX interfaces can be easily deduced. Having $I \neq J$ simply implies that the matrix we use to collect data in the first phase is rectangular with dimensions I x J instead of being square. This has no further implications in the use of the algorithm.

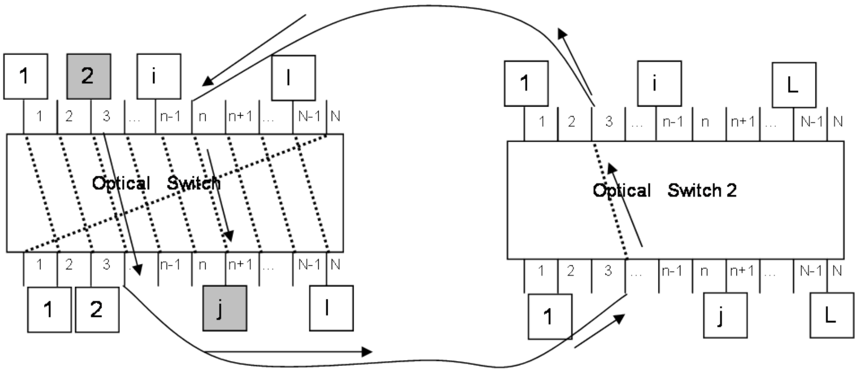


Figure 3.27: Example of configuration that could confuse the discovery process

If our switch has N input ports and M output ports instead, with $N \neq M$, we have to distinguish the case $N \leq M$ from the case $N > M$.

If $N \leq M$ we should arrange the connections in the first phase of the algorithm following the rules:

$$\begin{cases} OutputPort(n) = (n+k)\,mod(M) & if(n+k \neq M) \\ OutputPort(n) = M & if(n+k = M), \\ where\, n\, is\, the\, input\, port \end{cases} \tag{3.9}$$

At each iteration we shift the output ports with respect to the input ports, exactly as in the case of a N x N switch. The difference in the formulas is that we calculate the output ports modulo M (i.e., the total number of output ports).

In the case N > M instead, at each iteration we need to shift the input ports with respect to the output ports:

$$\begin{cases} InputPort(n) = n + k\,mod(N) & if(n+k \neq N) \\ InputPort(n) = N & if(n+k = N), \\ where\, n\, is\, the\, output\, port \end{cases} \tag{3.10}$$

Similar modifications apply to formulas 3.7) and 3.8) used in the second phase of the algorithm.

3.8.2 Link discovery algorithm

After completing the internal discovery, the node starts the discovery of its external links. Here we propose a solution to the link discovery problem in transparent switches that does not need the pre-provision of a control channel. The control channel is in fact provisioned in-band during the link discovery process and allocated in a wavelength that can also be used to transport data traffic (a prioritized virtual link could be established within the wavelength to favor control traffic over data traffic).

In the GMPLS architecture, where each node may be connected to a central management entity, an out of band control channel may be preferred. We believe however that the trend of semi-transparent optical networks, with the IP layer directly on top of the WDM layer, will follow a more distributed approach similar to current IP networks.

In distributed architectures, where control channels are only allocated between peering neighbors, the in-band approach may be preferred because it eliminates the need to set up a separate control channel and allows easier automation of discovery procedures.

The protocol we propose is made up of two distinct processes: one scans the receiving ports for incoming messages while the other sends "TEST" messages through the transmitting ports.

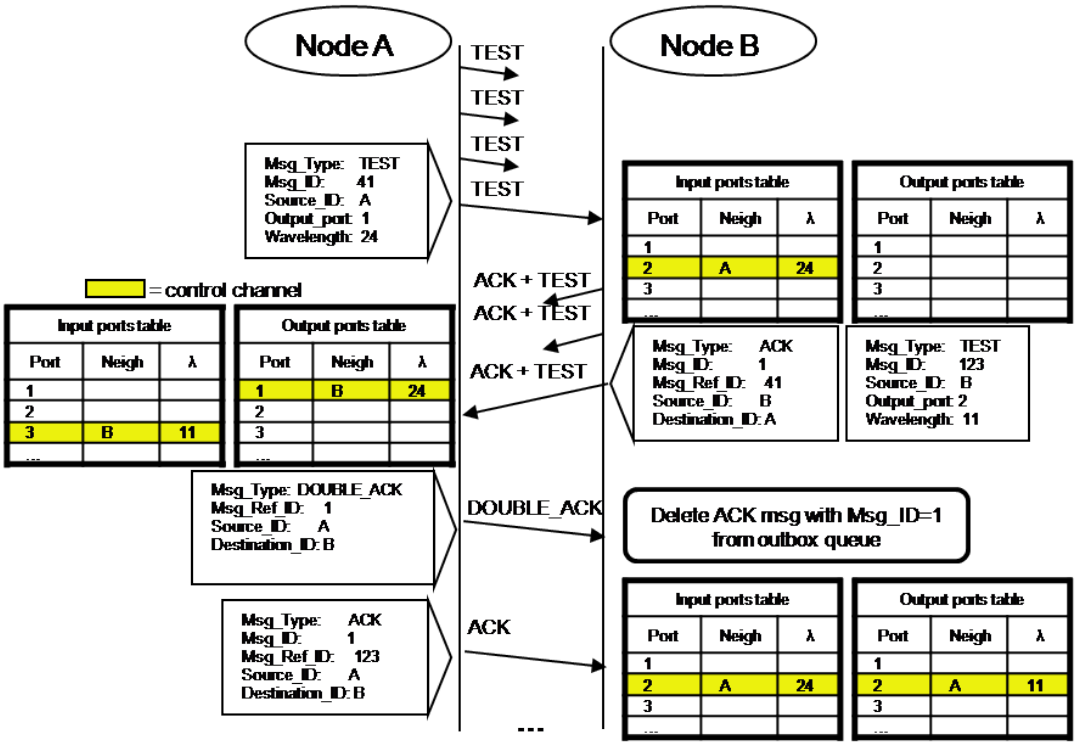
The process scanning the input ports runs permanently, in the background, on every active node and terminates, one after the other, all the undiscovered optical ports, using receiver router interfaces. All the receiving interfaces in idle mode (i.e. those not being used for data traffic) can be used simultaneously. After a pre-configured listening time, the receiving interfaces are connected to the next input ports; the process continuously scans the undiscovered input ports.

The process beaconing the output ports of the switch initiates when a node is started up and terminates once the discovery is completed. This procedure sees the router's optical transmitters beaconing "TEST" messages through the switch output ports. The parameters needed in the message are described in table 3.1.

TEST	**ACK**	**DOUBLE_ACK**
Message_type	Message_type	Message_type
Message_ID	Message_ID	Message_REF_ID
Source_ID	Message_REF_ID	Source_ID
Output_port	Source_ID	Destination_ID
Wavelength	Destination_ID	

Table 3.1: Content of the link configuration messages

As soon as a "TEST" message is received, the node will link the input port to the sender node and to the wavelength value encoded in the message, entering this information in its link table (as illustrated in figure 3.28). If the link is the first discovered from that node, it is designated as the control channel in the downstream direction and the router permanently connects a receiving interface to that input port. The message also serves as a trigger for the receiving node (B, in the figure), which starts beaconing on all the undiscovered outgoing ports. Node B should reply with an "ACK" message, making node A aware that the port over which that particular "TEST" message was sent is connected to node B. Since however no control channel has yet been discovered from B to A, node B will put the "ACK" on an outbox queue, whose content is relayed every time an output port is beaconed, before each "TEST" message. When it finally receives the "ACK" message, node A links it to the previously sent "TEST" message, discovering a new output port. Being the first output port that links to node B, A uses this port as an outgoing control channel to B.



Msg_Type: TEST
Msg_ID: 41
Source_ID: A
Output_port: 1
Wavelength: 24

Node A:

Input ports table		
Port	Neigh	λ
1		
2		
3	B	11
...		

Output ports table		
Port	Neigh	λ
1	B	24
2		
3		
...		

Msg_Type: DOUBLE_ACK
Msg_Ref_ID: 1
Source_ID: A
Destination_ID: B

Msg_Type: ACK
Msg_ID: 1
Msg_Ref_ID: 123
Source_ID: A
Destination_ID: B

Node B:

Input ports table		
Port	Neigh	λ
1		
2	A	24
3		
...		

Output ports table		
Port	Neigh	λ
1		
2		
3		
...		

Msg_Type: ACK
Msg_ID: 1
Msg_Ref_ID: 41
Source_ID: B
Destination_ID: A

Msg_Type: TEST
Msg_ID: 123
Source_ID: B
Output_port: 2
Wavelength: 11

Delete ACK msg with Msg_ID=1 from outbox queue

Input ports table		
Port	Neigh	λ
1		
2	A	24
3		
...		

Output ports table		
Port	Neigh	λ
1		
2	A	11
3		
...		

Figure 3.28: Example of messages exchanged by the link discovery protocol

The aim of the algorithm is to establish a control channel with each peering neighbor and to discover the wavelength associated with each output port (essential information when the switch ports are connected to WDM mux/demux). Wavelength tunability is an important feature of optical transmitters as it allows testing each port on all the wavelengths using the same transmitter (i.e., within a single switching time). If instead each transmitter can only beacon at one fixed wavelength, one switch activation is required every time a port is tested on a different wavelength.

The next step consists of node A sending a "DOUBLE_ACK" to B, using the newly established control channel, following a three-way handshake approach. The "DOUBLE_ACK" informs B to delete the related "ACK" message from its outbox list. Once the control channel is established in both directions, the two peers can exchange all the pending "ACK" messages. The procedure continues in the same way, in parallel with all the neighbors, operating asynchronously at all the nodes. A proper timer will terminate the beaconing process when no new link is discovered over a pre-defined time interval.

The parameter that most influences the algorithm performances is the time interval between the test of adjacent ports. Output ports should ideally be beaconed one after the other as fast as possible. For the input port scanning the best interval time (RX_SCAN_TIME) varies depending on the number of ports and interfaces. Intuitively, each node should test the same input ports until all the neighbors have beaconed all their output ports over the complete range of wavelengths. Such time increases with the number of switch ports and wavelengths, while it decreases with the number of interfaces available for beaconing the output ports. In operational environments however, nodes use switches of different sizes and different number of interfaces, making the optimization of the "RX_SCAN_TIME" impossible. In section 5.3.2 we will analyze the impact of this parameter on the overall link configuration time.

The link discovery procedure can also be used when new links are added to already operating nodes. In this case the output port scanning process needs to be triggered manually on at least one side of the link. Other implementations could see the output port discovery process triggered automatically after a link interruption, a link reconfiguration or at regular time intervals to keep the link table up-to-date. Since the process only uses idle interfaces, the discovery can be executed without interfering with regular router operations.

3.9 Effect of dynamic allocation on transport protocols

After having described the OIS architecture in the previous sections, we now focus on the effects that dynamic provisioning might have on the network transport protocols. Dynamic allocation of optical paths might create packet loss or out of order arrival. Out of order arrival, for example, might occur during path creation. In fact, after the optical path is created, the source node suddenly begins to inject packets over the new path, which, bypassing some of the nodes, presents lower delay than the original path. It might happen that the last packets sent through the default route arrive after the first packets sent through the dynamic path, even if those were sent before.

Packet loss could occur during path extension. The node extending the path activates the switch to modify a path where data is already being transmitted. All the packets that traverse the switch during the switching time (that is around 20 ms for a MEMs-based device) are lost.

Packet loss and out of order arrival cause problems to the transport protocols, affecting the connectionless UDP and connection-oriented TCP differently. In the following section we elaborate on the issue, differentiating the effects on the UDP and TCP protocols. Before going deeper into the details, we raise the following question: is it possible to eliminate packet loss and out of order arrival in highly dynamic optical networks? The answer to this question will be provided in section 5.4 where, through our optical testbed, we show that the OIS architecture is not particularly affected by out of order arrival during path creation, and we propose a mechanism to avoid packet loss during path extension.

3.9.1 Effects on the UDP protocol

UDP ([Pos80]) is a connectionless transport protocol where packets are sent as they arrive from the application layer without any guarantee of delivery. UDP does not receive any feedback from the destination node, so it is not able to react to network congestion and packet loss. Its main use is for real-time applications like voice calls or live video streaming, where there is no interest in recovering lost packets.

Packet loss deteriorates the quality of the voice or video received, and the impairments are somehow subjective, as they are associated with the quality perceived by the person at the receiver end. The ITU-T has defined the Mean Opinion Score (MOS) test [IT96] to evaluate the quality of an audio or video application, based on the perception of a group of testers. It defines a scale for the perceived impairment that goes from 1 to 5: 1 is considered “Very annoying”, 2 “Annoying”, 3 ”Slightly annoying”, 4 “Perceptible but not annoying”, 5 “Imperceptible”.

Several tests have been carried out in the literature ([HYM+99, DG03, CT99, BG98]), which examine the effects of packet loss on the perceived quality of audio and video signals, concluding that the results are dependent on the type of source, the data rate and the coder used. In general however, a packet loss of 1% is sufficient to deteriorate a video streaming application.

3.9.2 Effects on the TCP protocol

Unlike UDP, the TCP protocol is connection-oriented, as nodes establish a TCP session through a three-way handshake mechanism before exchanging data. TCP is considered a reliable protocol because it assures that the packets are all received in the correct order at the other end. This is achieved through a mechanism that resends the missing packets when the sender does not receive the expected acknowledgments. TCP uses a congestion window that determines the number of packets that should be sent before waiting for an acknowledgment. The size of the window, together with the Round Trip Time (RTT), determines the speed of the data transfer. TCP usually starts with a short window, which is gradually increased as the acknowledgments are correctly received. However, when the sender detects anomalies in the acknowledgments received, it reduces the size of the TCP window by a half. For a more detailed description of the TCP protocol the reader can refer to [Pos81].

The TCP protocol is generally used to transport non real-time applications or pre-buffered video streaming, where packets can be retransmitted when errors occur. For this reason, unlike the UDP protocol, the performance of TCP can be investigated by observing deterministic parameters like the average transfer rate.

The main issue with using TCP in dynamic optical networks is that the protocol always attributes packet loss or out of order arrival (in case out of order packets are received over multiple TCP windows) to network congestion, and reacts by reducing progressively the TCP window size. In OIS, where packet losses are caused by the switching time of the optical devices and the out of order arrival by the sudden re-routing of packets into a faster path, the TCP protocol erroneously attributes such anomalies to network congestion, reducing the TCP window.

In [Meh03] the authors discuss the out of order arrival issue for dynamic optical networks, showing quantitatively that dynamic path change noticeably decreases the data rate of the transport protocol. They show however that if the "Eifel" congestion algorithm, introduced in [LK00], is implemented in the TCP protocol, the packet reordering issue is completely eliminated and the TCP data rate is not affected by the path change.

In section 5.4, we show that out-of-order arrival of packets is a minor issue in our OIS architecture during path creation. We also illustrate the solution we have developed that targets the packet loss problem generated during path extension, and that does not require changes in the TCP protocol.

3.10 Impairments at the physical layer

One of the most challenging aspects of dynamic switching of optical paths is the compatibility of the paths created with the impairments at the physical layer, which we have briefly described in section 2.3.1. Although in this dissertation we do not address solutions to physical impairments, a field currently being investigated by many researchers around the world ([IST07, YHM05, PAMS06, MPC$^+$06, Muk06]), we describe briefly its implications for the OIS architecture.

We can divide the physical impairments between those caused by transient effects in the optical devices and those caused by joining two separate links into a single transparent path. The first problem is related to the gain transient effect present in the fiber amplifiers, where the addition or deletion of a wavelength creates transmission errors in the other WDM channels. Examples of this phenomenon can be found in [KKHR04]. Although this issue does have implications for the OIS architecture, we do not discuss it further, as possible solutions lie within the realm of the physical layer.

We address instead the impairments derived from the transparency of the optical cut-through paths, which can be counteracted with the aid of the signaling protocol. The problem here is that the effects of physical impairments, present on each link, add up when the links are transparently joined, bringing the BER to intolerable values at the receiver end.

The advantage of the Optical IP Switching architecture in building transparent paths is that, unlike end-to-end wavelength provisioning protocols, the physical constraints can be evaluated hop-by-hop as the path is created. The trade-off of this approach however is that if we only use local information, the decision will be sub-optimal compared to a centralized mechanism that makes use of global information. The efficiency of the distributed process can be improved if every node forwards information to all other nodes about impairments on their adjacent links. This would allow the nodes to take more optimal decisions, based on global link information. However this would also lead to an increase of the bandwidth occupied by the link messages and increase the path computation time. Therefore such trade-off between bandwidth and computation overhead versus global optimization of the solutions should be carefully evaluated during design phase.

In order to counteract the effects of physical impairments the nodes need first to estimate the physical characteristics of the links (e.g., path loss, dispersion). Each node can be provided for example with static information that describes such parameters for each outgoing link, or can learn them automatically if provided with adequate signal monitoring devices. The physical information about the links can then be exchanged between neighbors during link discovery (and periodically updated if monitoring devices are used).

When the decision process initiates the creation of a new cut-through path (or the extension of an existing one), the node estimates the physical impairments using an algorithm that considers the characteristics of each link constituting the path, together with the signal properties (e.g., transmission power, wavelength, modulation used, etc.). If the physical constraints allow for acceptable (estimated) BER at the receiver end, the node proceeds with the path creation mechanism. If instead the BER estimation is too high, the node could either abort the creation process or else try to solve the problem locally and proceed with the creation of the transparent path. In the second case, in order to counteract the signal impairments, the node needs to be equipped with devices like dispersion compensators, optical amplifiers and signal regenerators. If the problem is due to dispersion, the node could switch the signal into a dispersion-compensating device before sending it downstream. Similarly, if the problem is due to the high power loss, the signal can be amplified at the node. If the signal-to-noise ratio instead drops below acceptable values, it needs to be regenerated. Signal regeneration can be used in general to counteract any signal impairments. Its cost however is relatively high and increases linearly with the number of channels because current commercial solutions operate through O-E-O conversion. Nonetheless, initial prototypes of all-optical devices capable of regenerating entire WDM bundles are beginning to appear in research laboratories [LLB+03, TKT07].

The physical impairments described in this section currently constitute a major issue in dynamically transparent networks. Developments in this research area will therefore have high impact on the practical implementation of network architectures that, like OIS, are based on dynamic reconfiguration of lightpaths.

3.11 Summary

In this chapter we have introduced the Optical IP Switching architecture, describing algorithms, functions and protocols that enable dynamic and automatic provisioning of lightpaths. The distributed operating mode is the distinctive characteristic of our approach. Each node autonomously carries out local analysis of IP traffic, and based on the information collected and on its local networking policies, decides which optical paths should be created, extended or deleted.

Most of the optical circuit switching architectures proposed to date are based on centralized decision-making processes, and often use source routing to deliver end-to-end lightpath provisioning. We believe that this approach is not consistent with the distributed nature of the Internet and generates scalability issues, especially in the interdomain. Instead, the OIS architecture we propose adopts (at the optical layer) a distributed provisioning mechanisms that resembles IP routing operations. IP routing and optical provisioning are combined following a multi-layer engineering approach, through the prefix-based flow aggregation mechanism that we have developed. An effective method for node self-configuration was also developed, which gives OIS full plug-and-play ability.

Before OIS can be effectively implemented on a real network some open issues need to be solved. We have discussed the impairments at the transport layer caused by dynamic link reconfiguration. This issue can be effectively minimized through appropriate signaling operations, as it will be demonstrated in section 5.4.3. It is more challenging instead to counteract the signal impairments at the physical layer. Although a signaling mechanism can be introduced to estimate the signal disruption, the deployment of devices to improve the BER could considerably increase the overall network cost.

Chapter 4

OIS characterization and cost analysis

This chapter describes the simulation analysis of the Optical IP Switching architecture, answering some of the research questions raised in chapter 3.

Sections 4.1 to 4.4 focus on the technical analysis of the network performance, reporting results in terms of switched traffic, channel utilization and wavelength allocation efficiency.

In section 4.1 we illustrate the simulation tools we have used, the network topology and the traffic models.

In section 4.2 we report the simulation results, illustrating the performance of the OIS extension algorithms for the GÉANT network and the impact of the Internet heavy-tail prefix distribution.

Section 4.3 reports the performance analysis of OIS for network topologies different than GÉANT, covering both the intradomain and interdomain cases.

In section 4.4 we compare the performance of our distributed OIS model to a centralized approach that could be implemented through GMPLS, which provisions end-to-end optical circuits assuming global knowledge of the traffic demand matrices.

The last two sections deal with the cost analysis of the OIS architecture, with two distinct approaches.

The first, described in section 4.5, has already been considered in literature (see section 2.4.3) and is relative to the transparent bypass of the IP layer, which allows saving costs on expensive IP router equipment. The study we report has the novelty that it refers to a real network model (with real routing tables and traffic traces), and is calculated for our Optical IP Switching architecture.

The second approach, described in section 4.6, addresses the advantage of reconfigurable optical networks to reduce the (cost for) over-provisioning by dynamically adapting the optical paths to the variable traffic demand. Although this approach has already been considered for opaque networks ([SW06]), it has not yet been widely addressed in transparent optical networks.

In this chapter we limit the cost calculations to the savings on capital expenditures. Although it is widely recognized that dynamic optical networks will increase operators revenue by providing novel services and applications to their customers (as anticipated in section 2.4.1), we do not have precise information to conduct a quantitative analysis.

4.1 Network model for simulation analysis

4.1.1 GÉANT topology and data traffic

The reference topology we have used in our simulations is that of GÉANT, the pan-European data communication network, interconnecting different National Research and Education Networks in Europe, serving over 3500 research and education institutions. The network logical topology (illustrated in figure 4.1, available from [Top06]) includes 23 nodes, connected by links with capacities ranging from 155 Mbps to 10 Gbps.

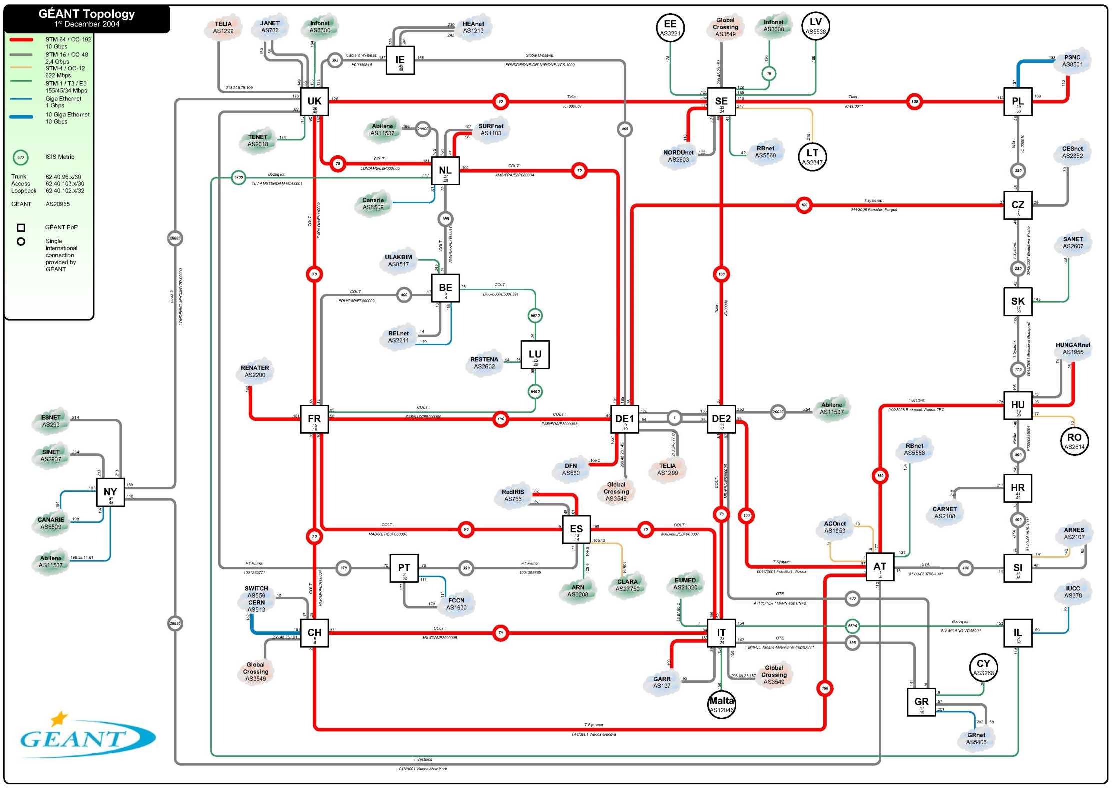


Figure 4.1: GÉANT logical layout topology

We have taken the GÉANT network as a model for our simulation study because it provides real traffic traces together with BGP routing tables. Other datasets are available, mainly collected from the CAIDA organization [CAI03], which provide either routing information or traffic traces from different networks around the globe. Since the goal of our study is to see how traffic flows are routed in a real network and the possibility of aggregating them into dedicated optical paths, we could not base our analysis only on traffic traces.

The dataset was made available by researchers from the Computer Science and Engineering Department at the University of Louvain-la-Neuve. They also provide C-BGP [QU05], a network simulator capable of reconstructing the BGP topology from the routes included in the dataset. C-BGP proved extremely valuable as it reconstructed the routing path of the traffic traces from the source node, where they were collected, towards their destination. The tool embeds a clustering algorithm that reduces the number of BGP entries by more than two orders of magnitude, making it possible to simulate a real network scenario. The traffic traces, collected using Netflow with sampling rate of 1/1000, are summarized depending on their source/destination prefix and only the total number of bytes over a fifteen minute period is provided. This has the two-fold effect of saving storage space for the data files while keeping the traces anonymous. The disadvantage is that information about the precise timing of the flows is lost. This condition however does not influence our study, since the mechanism that creates optical cut-though paths averages the observed traffic over a period of some minutes.

Traffic measurements on the GÉANT dataset showed that the transit traffic on the network nodes is 36% of the total (calculated on trace 1 showed in figure 4.2), while nine out of the 23 nodes only carry add/drop traffic. The average link distance is 797 km [SW06].

Although the model is based on the GÉANT network, we report in section 4.3.2 simulation results using different topologies, in order to examine the impact of different node degrees on the OIS performance. In section 4.3.3 we analyze the behavior of the OIS architecture in the interdomain, where transparent optical paths are created across the domain boundaries. Although this option tends to diminish the control that operators have on traffic, it greatly increases the amount of switched traffic within each domain, decreasing the use of costly router resources.

4.1.2 Optical IP Switching simulator

We have used the C-BGP simulator to recreate the original routes of the flows contained in the analyzed traces. The routes were then fed to our PERL-based simulator, to evaluate the performance of the OIS architecture.

The simulator we have built does not operate an a per-packet basis, as the dataset traces are averaged over a fifteen minute interval. We assume instead that for the fifteen minute duration, the traffic pattern is static (i.e., we have considered uniform, non-bursty traffic). Although this might not represent a real traffic situation, the OIS mechanisms use an observation time (described in section 3.2) that averages the traffic behavior over an arbitrary interval. By considering an OIS traffic observation time of fifteen minutes (i.e., equal to the trace duration), we can substantially reduce the dependence of the results on the traffic variation.

As simulation begins, the nodes initiate path creations and extensions one after the other, each considering the paths previously created by its upstream and downstream neighbors. The simulator also considers the wavelength constraint and keeps track of the wavelengths in use on each link. The wavelength assignment for the newly created paths follows a distributed first fit algorithm (see section 2.3.1 on RWA problem).

The results obtained show the performance of the OIS mechanism on the network, averaged over a fifteen minute interval, and focus on the overall amount of switched and routed traffic in the network. These simulations do not analyze the real-time behavior of the protocols, for which we provide dedicated testbed results in chapter 5.

In the simulations, the output values we consider to describe the OIS performances are: the routed data, the switched data allowed by OIS, the switched-to-routed traffic ratio, the channel occupancy and the wavelength efficiency. The routed data corresponds to the total amount of data routed by each node when no OIS switching is in operation (i.e., it characterizes the original GÉANT network). The switched data refers to the total amount of traffic that the OIS mechanism moves from the IP to the optical layer at each node. The switched-to-routed traffic ratio is the main parameter we consider in our simulations. This is obtained by dividing the previous two values, and shows the percentage of the total traffic that can be switched by the OIS architecture. The channel occupancy shows the amount of traffic that the OIS algorithms can aggregate into each wavelength channel. Considering that each additional path occupies costly resources, this is an important parameters to evaluate the cost-effectiveness of the architecture. Finally, the wavelength efficiency measures how well the system manages to exploit all the wavelengths available in each WDM system. Because of the wavelength continuity constraint, the links might need more wavelength channels than they actually use, increasing the link cost. In our simulations we show the wavelength efficiency obtained using the first fit wavelength assignment algorithm. We do not consider higher performance algorithms as the investigation of optimal distributed RWA algorithms falls outside the scope of this dissertation.

OIS parameter	*Value*
Observation time	15 minutes
Channel rate	1 Gbps
Path creation threshold	10% of channel rate
Path extension algorithm	Absolute threshold
Path extension threshold	100 Mbps
Trace date & time	4/5/05 @ 15:45

Table 4.1: Default simulation parameters

Table 4.1 reports the default setup parameters we have selected to run the simulations. The observation time determines the time interval during which each node examines the traces, and is set, as previously explained, equal to the duration of the traffic traces. The channel rate indicates the maximum rate that each wavelength channel can support. The path creation threshold is the minimum aggregate rate needed to trigger the creation of a new cut-through path. The path extension algorithm value indicates that we generally use the absolute threshold algorithm introduced on page 73 rather than the relative threshold described on page 75. The path extension threshold indicates the minimum amount of data needed to trigger the extension mechanism (in the absolute extension algorithm). Finally, the last field in table 4.1 indicates the date and time of the trace used as traffic pattern generator.

Since the simulation data is averaged over the entire trace duration and the traffic is considered uniform, we did not use the flow re-classification mechanism described in section 3.7. For the same reason, we have disregarded the path cancellation mechanism.

The values in table 4.1 are generally applied to all the simulations unless otherwise stated.

4.2 Switching performance of OIS

This section illustrates the switching performance of the OIS architecture. The first plot we analyze, in figure 4.2, shows both the amount of optically switched and electronically routed traffic in the GÉANT topology, considering ten traffic traces spaced differently in time. In the x-axis we report the different GÉANT nodes (identified by the initials of the country where the node is located). In the y-axis, the blue lines represent the routed traffic, while the red lines represent the optically switched traffic allowed by OIS (traffic is averaged over each entire trace and is expressed in Mbps). The different vertical lines within each node represent the results obtained for different traffic traces.

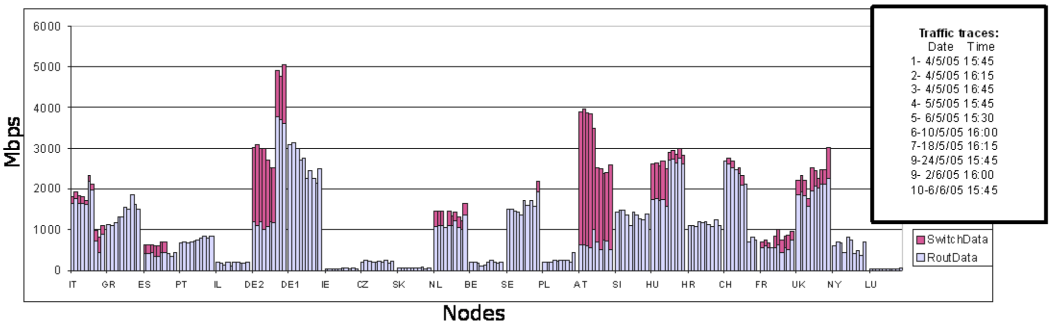


Figure 4.2: Switched vs. routed traffic for different traffic traces

It is interesting to observe that the switched-to-routed traffic ratio, at each node, has a mild variation over time. Analysis of the number of optical ports, transmitters and receivers used at each node support this observation, showing similar quasi-static behavior. This implies that the amount of optical resources required at each node is reasonably constant over time, a primary requirement in a practical implementation where hardware needs to be properly provisioned across the network. On more detailed examination of our traces, we noticed that nodes with higher ratio of switched traffic (e.g., AT, DE2, HU) mainly route data between GÉANT nodes (transit traffic), while others with little or no switched traffic (e.g., DE1, SI, CH) deal with data coming from or directed to external networks (add/drop traffic).

In some cases (most notably the last three traces for DE2, and the last five for AT) there is a sudden change in the amount of switched versus routed traffic. These changes appear to be step-like, as the values remain stable after the change. By observing the traces leading to this behavior, we could directly correlate the traffic increase at DE2 with the decrease at CH and IT, and conclude that probably the change was a consequence of an internal re-engineering of the GÉANT network links. The variation observed at the AT node instead is due to a general drop of transit traffic towards AT (mainly coming from CH).

Figure 4.3 plots the switched-to-routed ratio depending on the traffic level, averaged over all nodes, considering trace 1 of the GÉANT dataset. The different levels of traffic were obtained by multiplying the original traces (labeled "x1") by progressively increasing factors, reflecting the situation where traffic varies in volume and the distribution of the demand is unchanged. Although this simple method might not correctly represent the future evolution of traffic in the network, it allows us to analyze the impact of higher traffic volumes on the network performance.

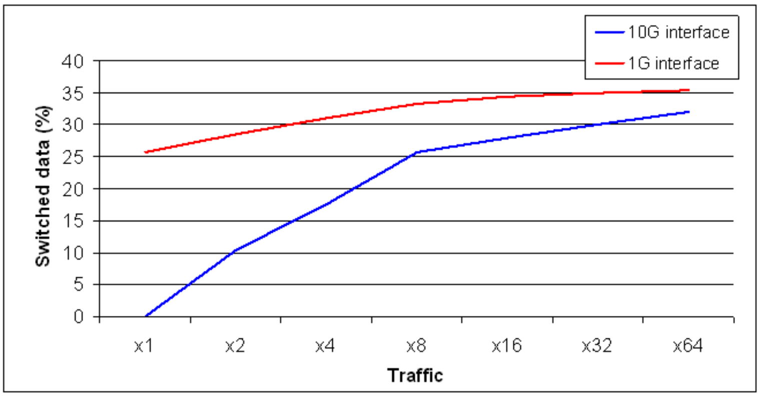


Figure 4.3: Switched-to-routed traffic ratio for different path creation thresholds

The red curve refers to the use of wavelength channels of 1 Gbps. As the traffic increases, the number of channels above threshold also grows, increasing the percentage of switched traffic. This behavior occurs until all the possible paths have been created, after which the ratio between routed and switched traffic tends to remain constant. The blue curve refers to wavelength channels of 10Gbps. The difference between the two curves is due to the different path creation (and extension) thresholds considered, which for the 10 Gbps channel is set to 1 Gbps (i.e., 10% of the channel rate).

The effect of using lower path creation thresholds can be inferred from the same graph. In terms of switched-to-routed traffic ratio, doubling the traffic level is equivalent to reducing the threshold by a half. For example, the switched-to-routed ratio obtained using a threshold of 50Mbps on the original traffic rate is equal to the value obtained from the 1 Gbps interface curve (with 100Mbps threshold) considering a "x2" traffic volume.

The results we have illustrated allow us to answer the research question formulated in section 3.3.

Decreasing the path creation threshold favors the provisioning of more cut-through paths because it increases the number of flow aggregates above the threshold. If however all the possible paths have already been created, lowering the threshold does not increase the switched-to-routed ratio any further, which explains why the curves saturate. Increasing the threshold has the opposite effect, decreasing the number of cut-through paths.

Figure 4.4 shows the wavelength efficiency relative to the distributed first fit algorithm we have used as the wavelength selection mechanism. The total height of each bar represents the number of wavelengths needed in the system averaged over the different links, while in red we display the number of wavelengths effectively used for cut-through paths. The difference between the two values reflects the inefficiency of the first fit algorithm in solving the wavelength constraint issue. The efficiency, calculated as the ratio of the two values, is displayed by the dark line.

The plot clearly shows how the efficiency decreases when the traffic demand increases, reaching values below 50%. This indicates that the first fit algorithm is quite inefficient and a better solution would be required for a practical implementation.

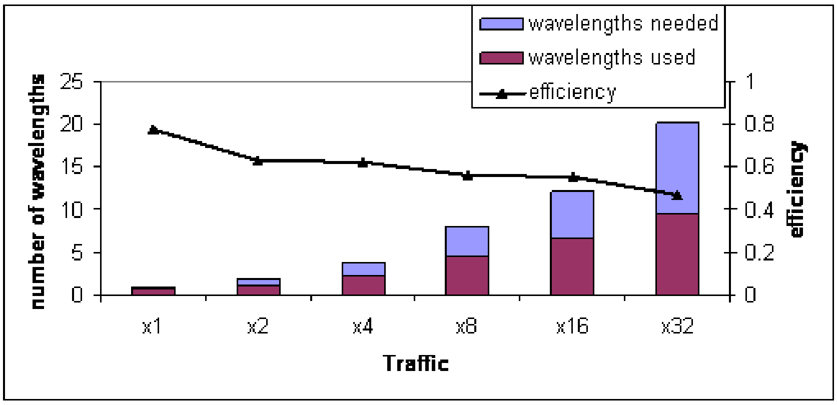


Figure 4.4: Wavelength efficiency using the first fit algorithm

4.2.1 Effect of the path extension algorithm on the OIS performance

In this section we analyze the difference between the two path extension algorithms described in section 3.4.

The plot in figure 4.5 shows that the absolute threshold algorithm has higher switching capability than the relative threshold algorithm. This is an expected behavior because the absolute threshold algorithm, as anticipated in section 3.4.1, aims at maximizing the amount of switched traffic by creating a higher amount of dedicated optical paths. It is interesting to notice that the absolute algorithm reaches a percentage of switched traffic of about 36%, which matches the percentage of transit traffic measured in the GÉANT network for this trace. This means that this algorithm reaches the maximum switching allowed in the network for traffic level 64 times the original.

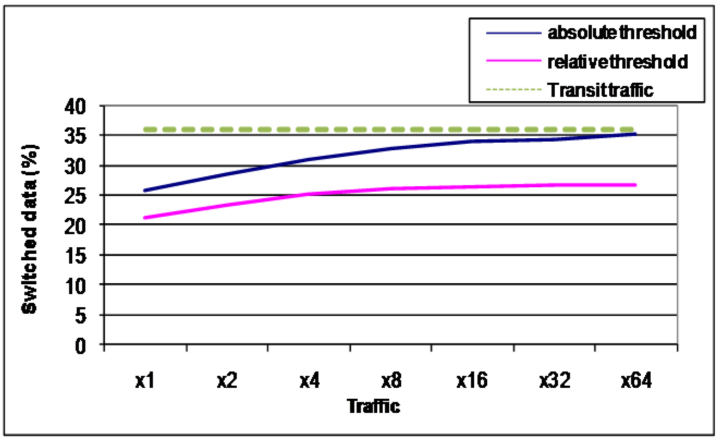


Figure 4.5: Switched-to-routed traffic ratio of the absolute vs. relative threshold extension algorithms

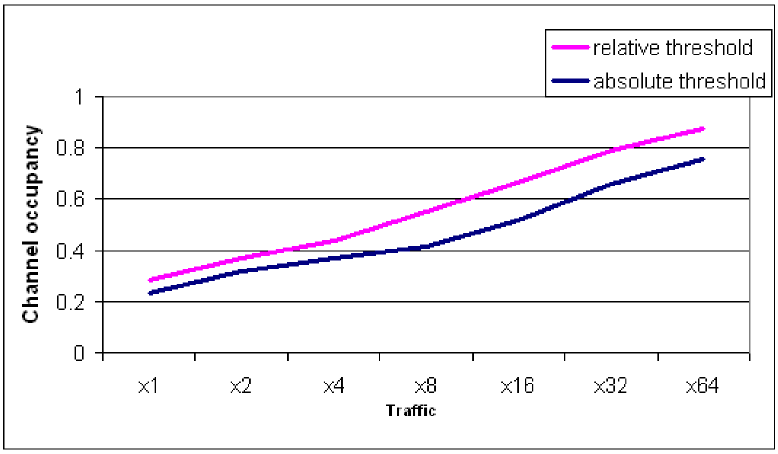


Figure 4.6: Channel occupation of the absolute vs. relative threshold extension algorithms

The situation is inverted in figure 4.6, which reports the average optical channel usage. The relative threshold algorithm shows a better channel occupancy with respect to the absolute threshold, allowing better exploitation of the optical bandwidth. This behavior was also anticipated in section 3.4.1, as the relative threshold algorithm was developed to optimize the amount of data in the paths at each extension process.

These results provide the answer to the research question formulated in section 3.4, about the implications of the path extension trade-off. The relative threshold algorithm, which directly addresses the trade-off between path length and switched data, optimizes the channel occupancy, making better use of the channels it creates. The absolute threshold algorithm instead maximizes the number of cut-through paths, increasing the switched data (thus saving routing resources) but decreasing the channel occupancy. Which of the two would represent a better approach in a practical implementation depends on how resources are distributed in the network. The extension algorithm could be dynamically adapted to the resources currently available. The absolute algorithm could be used, for example, when the availability of wavelengths and transmitters exceeds the availability of routing resources; the relative threshold in the opposite case.

4.2.2 Effect of prefix filtering

This section answers the research question formulated in section 3.2.2 regarding the effects of the heavy-tailed Internet traffic on how data is distributed among the IP routing prefixes. First we show that, similarly to the Internet flow distribution, the routing prefix distribution is heavy-tailed. Then we discuss the effects of filtering out the prefixes with low traffic volumes on the OIS performance.

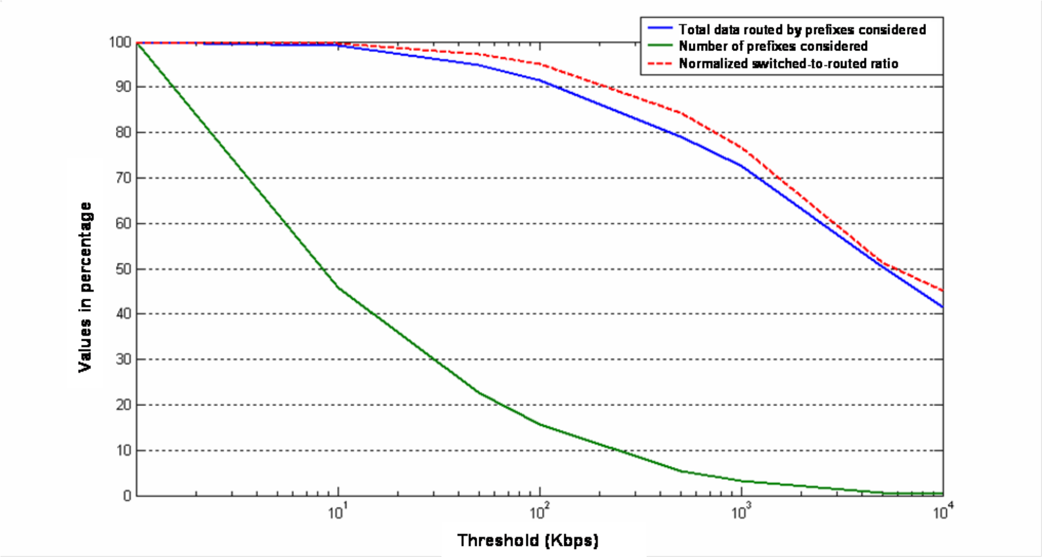


Figure 4.7: Heavy-tail distribution of routing prefixes in GÉANT

The green and blue curves in figure 4.7 illustrate respectively the percentage of the prefixes that carry data above the threshold value indicated in the x-axis, and the percentage of data carried by those prefixes (calculated on trace 1 of the GÉANT dataset). Comparing the two curves, we notice that if we apply a threshold to filter out the prefixes that route a small amount of traffic, the total number of prefixes diminishes substantially, while the reduction in the amount of data transported is minimal. If, for example, we set the prefix filtering threshold to 100 Kbps, the OIS path creation mechanism would only deal with 16% of the prefixes, with a substantial decrease in the signaling overhead and routing processing power. The blue curve however shows that those prefixes filtered out (84% of the total) only carry 8% of the overall traffic. Therefore OIS only disregards 8% of the traffic, with a minimal impact on the OIS switching capability.

The dashed red curve in particular shows how the OIS performance, expressed in terms of the switched-to-routed traffic ratio, is affected by the prefix-threshold mechanism. We see that a 100 Kbps threshold only causes a loss of 3% of switched traffic, which, compared to the 84% reduction on signaling overhead and processing power, represents an optimum result.

4.3 Effects of different topologies on the network performance

In this section we analyze how different topologies affect the performance of an Optical IP Switched network. Considering that the capability of creating dynamic paths highly depends on how traffic aggregates across the links, we expect different topologies, and in particular different node degrees, to have considerable impact on the system behavior.

4.3.1 Modified GÉANT topology

In experimenting with node degrees, the first topology we discuss was obtained by altering the original GÉANT topology. We have deleted five main bidirectional links, decreasing the average node degree from 3.2 to 2.8. Figure 4.8 shows the original topology together with that obtained by removing the following links: DE2-IT, CH-AT, UK-SE, FR-DE1 and DE1-DE2.

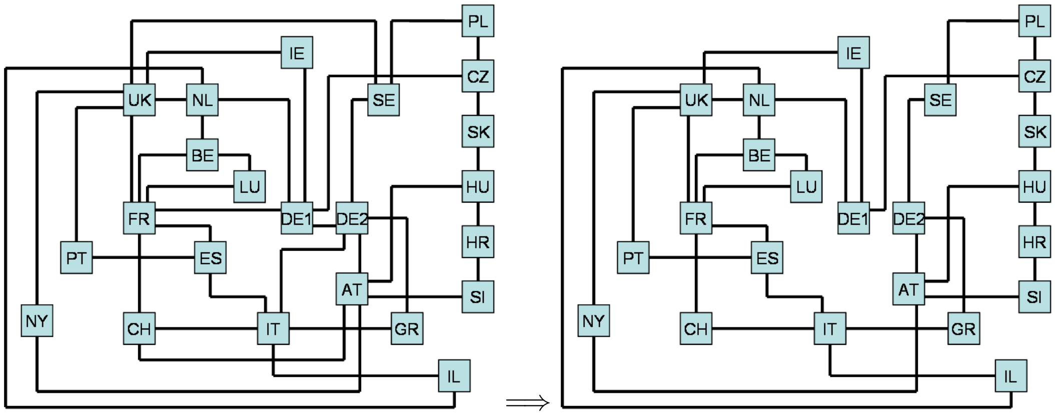


Figure 4.8: Original (left) and modified (right) GÉANT topology

Figure 4.9 compares the results obtained from the modified GÉANT topology with the original one. The plot shows that a less connected topology allows an increase in the switched-to-routed traffic ratio of about 15% (a relative increase of about 60%), while the channel efficiency increases between 4% and 13%.

The maximum switching OIS achieved in the modified topology was about 48%, which is close to the percentage of transit traffic measured for this topology (53%).

The fact that both the amount of traffic switched and the channel occupancy increase, shows that a decrease in the network connectivity allows OIS to achieve better efficiency. The reason is that with lower node degree, the average number of hops increases, as well as the aggregation of IP flows. The former increases the average length of the optical cut-through paths, while the latter makes sure that the channel occupancy remains high. Considering the tradeoff between path length and channel occupancy, introduced in section 3.4, we notice that a less connected architecture leads to an increase in both values.

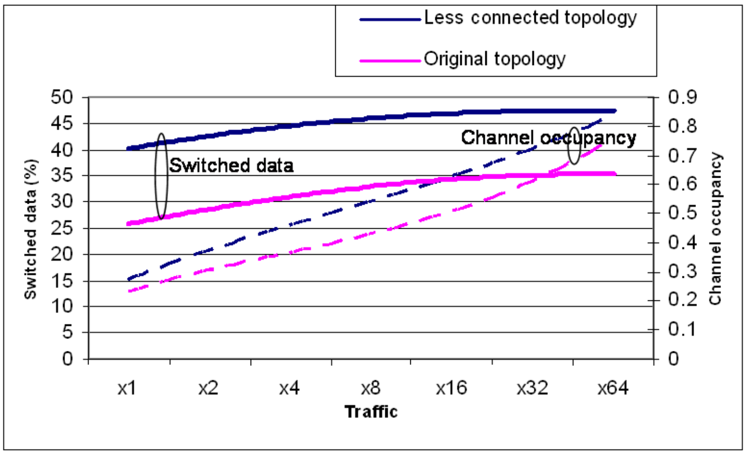
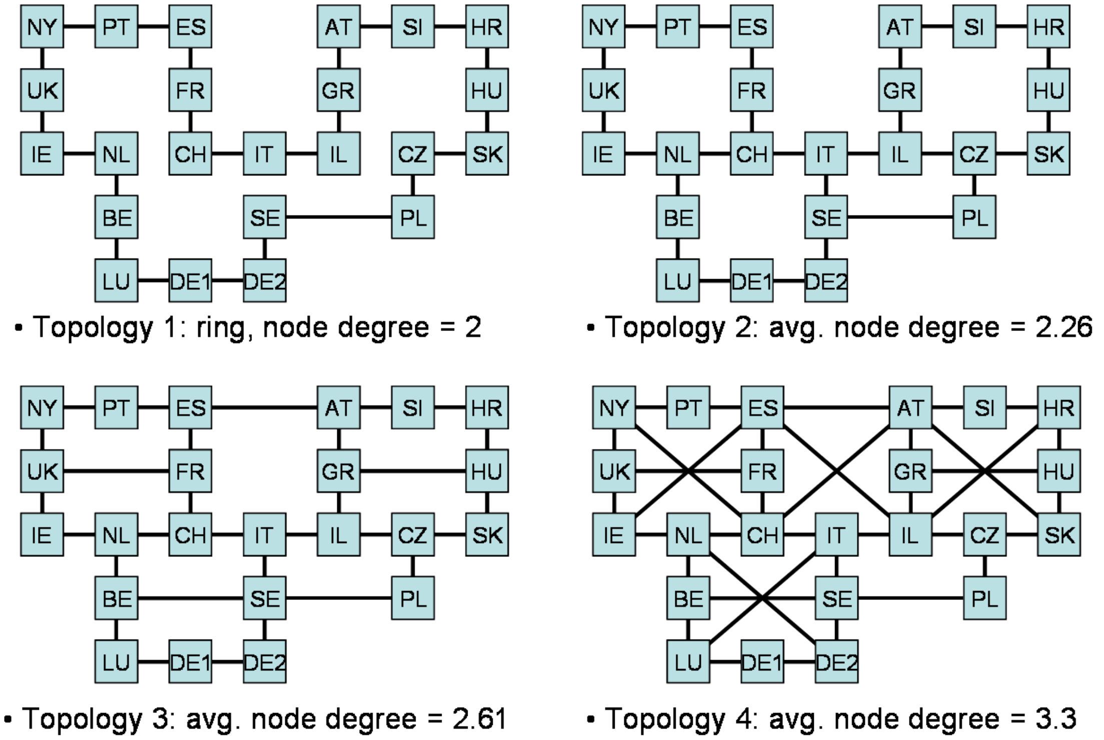


Figure 4.9: Network performance with reduced average node degree

The results we have shown could also be interpreted as the reaction of the OIS protocol to multiple links failure. Although the simultaneous failure of five uncorrelated and diverse links is highly improbable, for a given total traffic figure 4.9 shows how OIS would redirect excess traffic from the IP to the optical layer, to reduce the congestion at the routers. The reaction time, without the implementation of dedicated protection mechanisms, is proportional to the convergence time of the IP routing protocol and to the length of the OIS decision time.

4.3.2 Topological analysis

Based upon the results obtained in the previous section we have investigated in more depth the impact of the topology on the performance of OIS. We have considered four different topologies, obtained adding progressively more links to a ring topology. The nodes and traces are still those from the GÉANT dataset but we have modified the link directions and weight (giving same weight to all the links), obtaining networks where the nodes' degrees vary from 2 to 3.3. The topologies obtained are illustrated in figure 4.10.



• Topology 1: ring, node degree = 2

• Topology 2: avg. node degree = 2.26

• Topology 3: avg. node degree = 2.61

• Topology 4: avg. node degree = 3.3

Figure 4.10: Topologies under analysis

Figure 4.11 shows the results of the simulations obtained, reporting also the original GÉANT for comparison. The percentage of transit traffic measured for topologies 1 to 4 were respectively of 65%, 59%, 55% and 45%, which are generally close to the switched-to-routed traffic ratios achieved by OIS for traffic loads 64 times the original.

Lower degree networks show progressively increasing values of the switched-to-routed traffic ratio, confirming the trend we had already observed in section 4.3.1. The reason, as previously explained, is related to the increase in the average number of hops and to the decrease in the number of links, which allows better aggregation of the traffic flows. From the graph we also notice that, although topology 4 has higher degree than the original GÉANT network (3.3 against 3.2), the switched-to-routed ratio is higher in the former. This shows that, besides the average node degree, the particular link configuration also affects the switching capability.

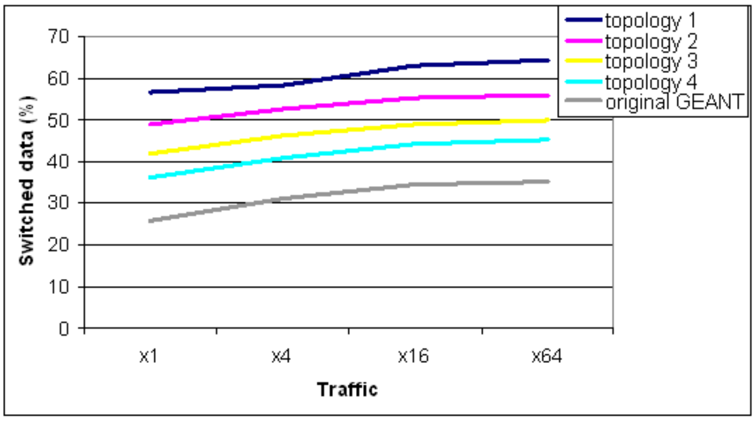


Figure 4.11: Simulation results using different network topologies

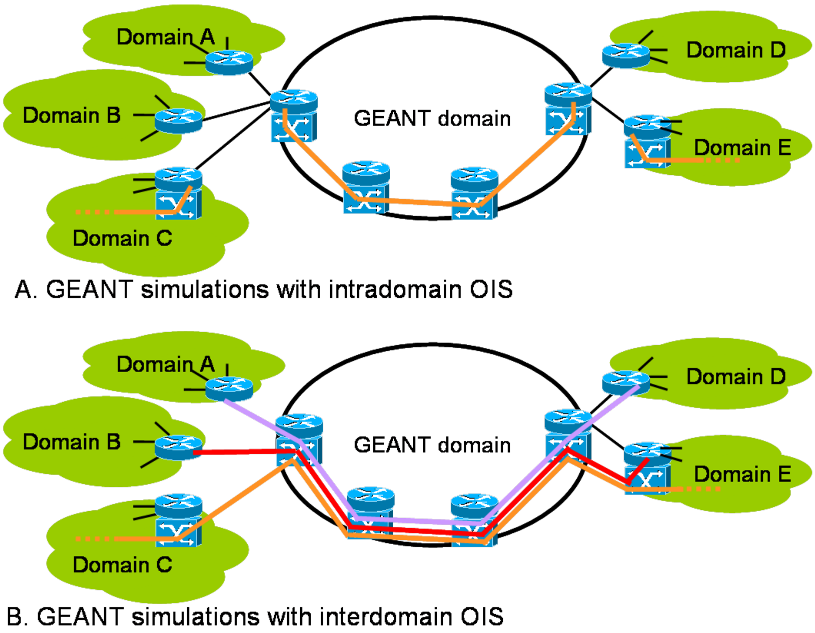
Although our results show that the ring topology is ideal to optimize the ratio between switched and routed traffic, in a practical implementation there are other issues, concerning the network resiliency, which must be taken into account. If on one hand the ring topology increases the nodes switching capabilities, on the other hand mesh topologies offer much better resiliency. A proper OIS design will therefore consider the tradeoff between the two aspects.

4.3.3 Interdomain OIS: extension to the user

From the simulation results we have illustrated in the previous sections considering the original GÉANT topology, we have seen that the switched-to-routed traffic ratio varied between 25 and 35 percent. For a practical implementation these values appear to be too low to be cost-effective, as will be proved in section 4.5.2 on cost analysis. One idea to improve the cost-efficiency could be to deploy the OIS switching capabilities only at the nodes with high estimated switched traffic; in fact, in figure 4.2 we have observed that most of the optical switching is concentrated in few nodes. Although we can avoid deploying optical switches on non-switching nodes, they are still involved in the optical cut-through paths as sources or destinations, and need to be provided with additional transceivers and multi-wavelength links. Therefore the cost reduction would only be partial.

We have observed that the low switching rate originates in part from traffic that cannot be switched in dedicated cut-through paths because the number of hops traversed is too short (i.e., less than three). Rather it is due to the fact that in each cut-through path, the source and destination nodes cannot avail of transparent switching, as they use electronic routing to add and drop packets in and out of the GÉANT domain.

In order to increase the switched traffic, we have considered the possibility, already discussed in section 3.6, of extending cut-through paths across different domains. The advantage of this approach is that the GÉANT access nodes acting as source or destinations of the optical paths, can extend the paths transparently to the external networks, gaining the benefit of optical bypass of the routing layer. The feasibility of this approach in a real implementation might be compromised by the fact that transparent switching across the domain boundaries does not allow network operators to control and filter the traffic entering and leaving the domain. The possibility of interdomain switching will thus depend on the development of new business models for next generation networks (as anticipated in section 2.6.1).



A. GEANT simulations with intradomain OIS

B. GEANT simulations with interdomain OIS

Figure 4.12: Intradomain and interdomain application of OIS in GÉANT

This situation is illustrated in figure 4.12. In the simulations reported in the previous sections, all the add/drop traffic was processed by the edge GÉANT nodes at the IP layer (figure 4.12.A). Interdomain cut-through paths allow the edge nodes to avoid routing that traffic, which is instead injected by the edge routers of the other domains. The drawback is that incoming traffic from different domains cannot be aggregated on the same wavelength at the GÉANT access nodes (figure 4.12.B).

Figure 4.13 compares the interdomain and intradomain performances of the OIS architecture, considering the nodes within the GÉANT domain. It is evident how the switched-to-routed ratio increases in the interdomain case, reaching values over 93%. Such a high increase is the effect of relaxing the constraint that path sources and destinations must be located in the GÉANT domain. In this case, GÉANT becomes a highly transparent core network, where traffic is mostly switched at the optical level while routing is mainly operated by the external domains. We have in fact calculated that the overall transit traffic, computed among the core GÉANT nodes, is over 98%. The analysis of the channel occupancy instead favors the intradomain approach (by a value between 8 and 11%). This is a consequence of the fact that, as visible from figure 4.12.B, edge nodes cannot aggregate traffic from different domains and the traffic is spread among a higher number of wavelengths.

Overall we can conclude that interdomain switching can be highly beneficial for OIS efficiency, as it will be demonstrated by the cost analysis reported in section 4.5.3.

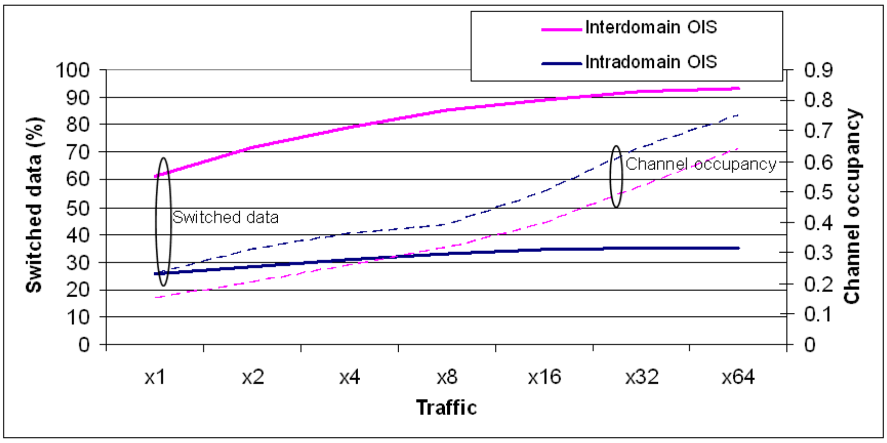


Figure 4.13: Interdomain vs. intradomain OIS performance in GÉANT

4.4 Comparison of OIS versus end-to-end provisioning architectures

In the preceding sections we have examined the performance of the OIS architecture, considering different algorithms and topologies. In this section we compare our distributed OIS model to a reference architecture that creates end-to-end dynamic optical paths with a centralized approach. A practical example of such a model could be a transparent network controlled by the GMPLS protocol, where a network administrator centrally provisions optical paths, following the analysis of traffic demand matrices.

For the end-to-end architecture we have assumed that the traffic demand generated by all nodes is available in a central database. A dedicated optical path is then provisioned for each end-to-end demand with average data rate above the threshold of 100 Mbps (for comparison with the OIS threshold).

The comparisons between the OIS and GMPLS architectures are illustrated in figures 4.14 and 4.15 respectively, for the intradomain and interdomain cases.

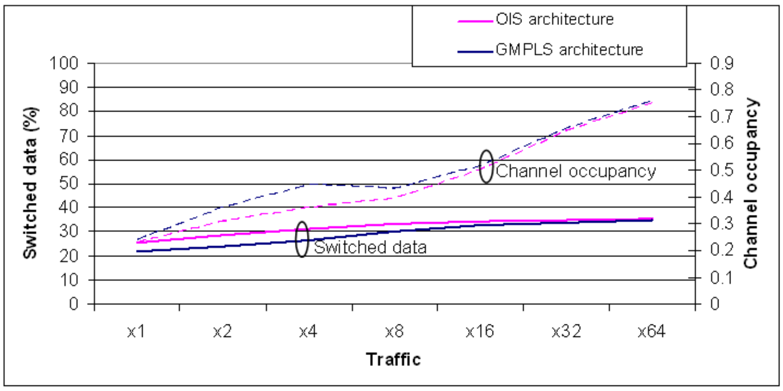


Figure 4.14: OIS vs. GMPLS architecture for intradomain switching

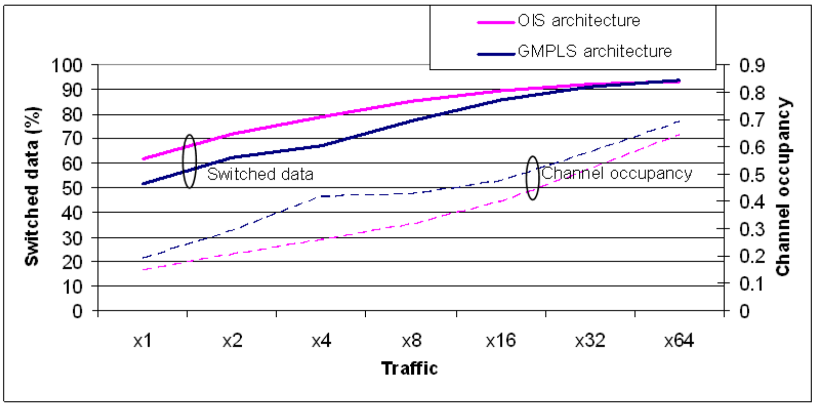


Figure 4.15: OIS vs. GMPLS architecture for interdomain switching

The capability of switching traffic optically is in general quite similar for the two architectures. OIS has an advantage over GMPLS for lower levels of traffic. This occurs because OIS, by building the optical paths in a distributed fashion, on a hop-by-hop basis, and analyzing the traffic locally, allows better traffic aggregation compared to an architecture that only considers end-to-end demands. As traffic increases however, more and more end-to-end demands grow above the provisioning threshold and the amount of switched traffic becomes similar for the two models. The channel occupancy is instead slightly in favor of the end-to-end provisioning architecture because of the lower number of paths created. In particular in the interdomain case, we note that for level of traffic 64 times the original, although the switched-to-routed ratios have similar values, the GMPLS architecture maintains an advantage in the channel occupancy of about 4%. We attribute this modest inefficiency to the fact that when a large amount of traffic is involved, complete visibility of the traffic demands allows a better path optimization, compared to our distributed approach where traffic is only known locally.

The simulations we have carried out show in general that the distributed approach performs equally or better than the centralized approach for what it concerns bypass of IP traffic. We have not operated however any optimization either in the distributed or centralized approach, as protocol optimization is outside the scope of this work. If optimization was considered, the centralized approach might present an advantage with respect to the distributed approach, as it can allow to take more optimal decisions, based on a global view of the network traffic and resources. As network size increases however, the centralized approach becomes less scalable, as the protocol overhead required to maintain updated global network information increases together with the processing power needed for optimal path computation (especially if impairment-aware routing is taken into consideration).

4.5 Cost savings of OIS: bypass of the IP layer

In the previous sections we have shown the results of the performance analysis on the OIS architecture under different situations, and compared it to a centralized end-to-end provisioning approach. We have analyzed the network performance in terms of switching capabilities, mainly focusing on the switched-to-routed traffic ratio and on the average channel occupancy.

In this section we analyze the performance of the OIS architecture from an economic point of view, focusing on the possible savings in capital expenditures allowed by the transparent bypass of the IP layer.

4.5.1 Cost model

The cost savings introduced by the optical bypass of the IP layer is the lower cost associated with optical switching compared to electronic IP routing, with a cost difference per port that can reach two orders of magnitude (e.g., considering 10 Gbps core router ports). However, as already anticipated in section 2.4.3, optical circuit switching operates on a much coarser granularity, which makes the economical advantage in a real environment unclear. An analysis is therefore necessary to evaluate the cost-effectiveness of the OIS architecture.

The cost model we have developed considers the reduction in routing equipment allowed by optical bypass, together with the increase in optical devices needed to implement the OIS concept. Besides the cost of optical switches, the model considers the increase of transport costs in terms of additional optical regenerators, longer reach transmitters and links, and higher capacity WDM systems. Longer reach transmitters are needed because transparent switching increases the distance covered by the optical signals and must be combined with the use of more expensive optical amplifiers offering better signal-to-noise ratio. When the distance covered by the optical path is above the optical reach of the transmitter (which we consider to be 2,500 km), the signal needs to be regenerated, adding a cost that increases proportionally with the number of channels. The increase in the number of WDM channels reflects the increase in the number of lightpaths used for optical bypass of the IP layer.

Figure 4.16 shows the basic equipment that we have considered in our cost model. Although not explicitly illustrated, all the devices are bidirectional.

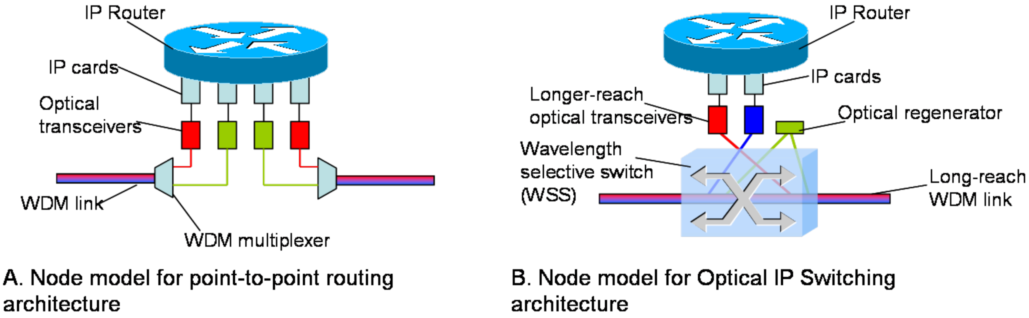

IP Router
IP cards
Optical transceivers
WDM link
WDM multiplexer
A. Node model for point-to-point routing architecture
IP Router
IP cards
Longer-reach optical transceivers
Optical regenerator
Wavelength selective switch (WSS)
Long-reach WDM link
B. Node model for Optical IP Switching architecture

Figure 4.16: Node models for the cost analysis

The costs we have considered for the network equipment, illustrated in table 4.2, were taken from [GLW$^+$06] and [WCM06]. When we considered high levels of traffic, for which systems with more than 80 wavelengths were needed, we extrapolated the cost, keeping the same relative cost as with the existing systems (for example, assuming that the cost ratio between a 160 and 80 wavelength system is equal to that between an 80 and 40 wavelength system). These calculations are based on the assumption that the cost of WDM and WSS technologies will increase linearly in the foreseeable future. This assumption is relaxed at the end of section 4.5.3, where we report the results of the sensitivity analysis of our model to cost changes, considering the effects of an additional increase in the cost of WSS devices. The link costs were calculated considering an average length of 800km, and that in the OIS model every link should be engineered to support an optical reach of 2,500km (in order to support

Device	*Cost (units)*
IP card (10G, core router)	120
IP router (per 160G traffic)	50
800km-reach transceiver	1.3
2,500km-reach transceiver	2.2
Signal regenerator	2.7
WDM multiplexer, 40λ	4.5
WDM multiplexer, 80λ	6.7
Link (short reach)	50
Link (long reach)	70
Transparent OXC port	2

Device	*Cost (units)*
WSS 9-ports 80λ	10
WSS 20-ports 80 λ	12*
WSS 40-ports 80λ	15*

*future estimation - effects of higher costs are evaluated through the sensitivity analysis reported at the end of section 4.5.3.

Table 4.2: Costs considered for the network devices

transparent switching). The link costs are calculated in terms of signal amplification and dispersion compensation, while fiber placement costs are not included as they are common to the models considered. The cost of the Wavelength Selective Switches (WSS) reported in the table was directly obtained by equipment vendors. Systems with higher degrees and number of ports are obtained by interconnecting more WSS, while a MEMs-based transparent OXC is introduced to ensure full steering ability of the optical transceivers (figure 4.17, [Ovi07]).

4.5.2 Cost comparison: point-to-point IP vs. end-to-end provisioning vs. Optical IP Switching

In this section we show the simulation results of the cost analysis we have performed on the following three models: the opaque point-to-point IP, which follows the original GÉANT architecture where packets are routed electronically at each hop; the OIS model; and the centralized end-to-end provisioning architecture (GMPLS) analyzed in section 4.4. The models use the same parameters described in table 4.1, apart from the channel rate, which is equal to 10 Gbps (creation and extension thresholds are both set to 100Mbps).

The results are illustrated in figure 4.18, showing the cost associated with the different architectures. The first observation we make is that the difference between the GMPLS and

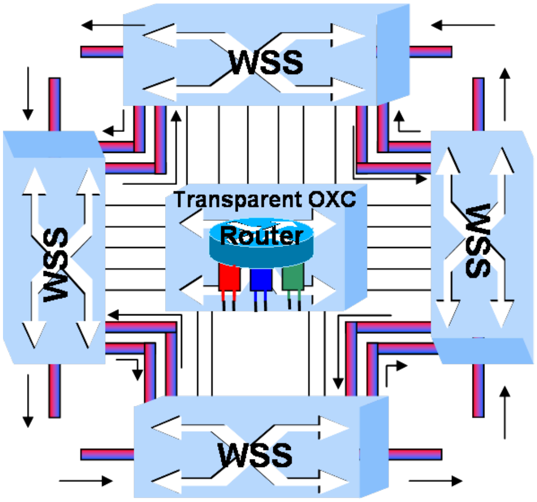


Figure 4.17: Four-degree node implemented combining WSS and transparent OXC

OIS models is negligible, indicating that both are capable of exploiting the advantages offered by transparent switching in a similar way. The value of OIS for traffic level equal to "x1024" was not available due to the high memory usage of the OIS simulator, but we can assume a value similar to that obtained with GMPLS.

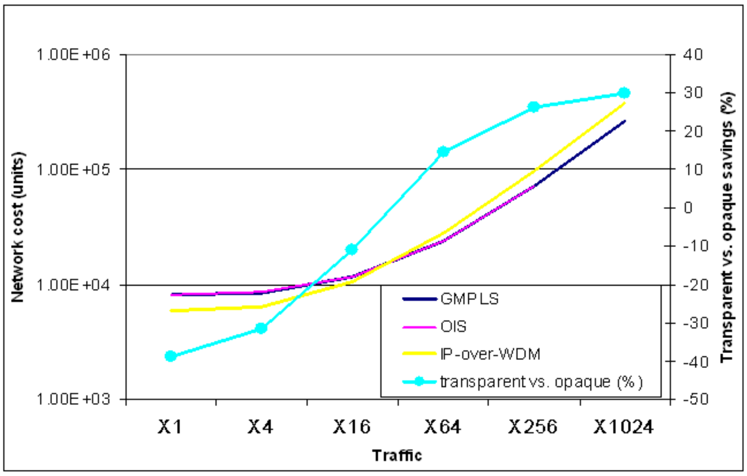


Figure 4.18: Cost comparison between the considered architectures

The second observation is the difference between the point-to-point and the transparent switched models. The light-blue curve represents the saving (expressed in percentage) allowed by the transparent architectures with respect to the opaque model. For low levels of traffic, the point-to-point IP model shows an advantage over the transparent models. If in fact traffic is too low, the dynamic optical paths cannot exploit the optical bandwidth enough to make the system cost-effective. As traffic increases however, the optical paths are capable of switching a higher amount of traffic, reducing the expenses for costly IP router cards.

It would be interesting to give a time dimension to the traffic increase we consider in our simulations. It is generally difficult to foresee how traffic will change in the future, and past observations have showed that the increase is quite variable over time. If however we assume that the current growth trend of 50% per year will continue steadily in the foreseeable future, the x1000 traffic level we have considered in our simulations should be reached in about 15 years time.

4.5.3 Cost comparison for the interdomain model

We have repeated the cost simulation for the interdomain case, with transparent paths crossing the domain boundaries. The model is similar to that used in the previous section, but we have also included the cost for additional WDM links and transceivers in the external domains. The results are illustrated in figure 4.19. The difference between the GMPLS and OIS models, similarly to the intradomain case, is negligible. The cost difference between the opaque and transparent architectures however is more remarkable, as the cost of the OIS and GMPLS models, unlike the point-to-point IP model, decreases with respect to the intradomain case. The reason is the increase in switched traffic that, as we have already observed in section 4.3.3, decreases the number of costly IP router cards in the system.

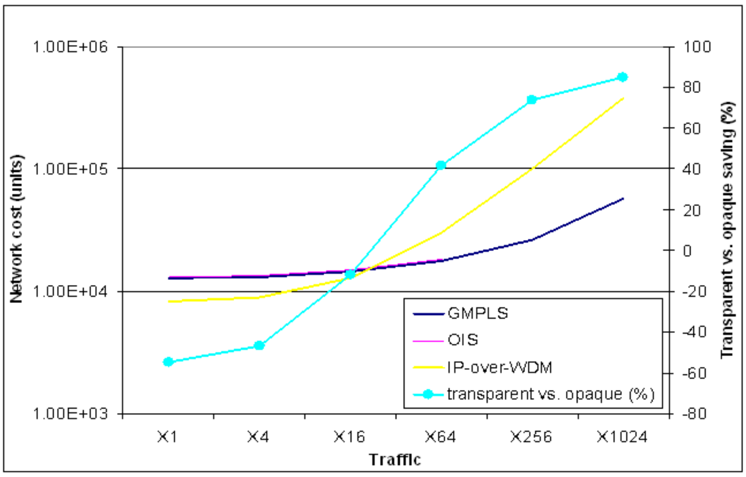


Figure 4.19: Cost comparison in the interdomain

Although the overall network cost decreases, if we separate the costs in the GÉANT and external domains, we realize that cost savings only occur in GÉANT, where the bypass of the IP layer is operated. The cost instead increases in the external domains due to the deployment of additional optical equipment. In the model examined the external domains are always considered to be source or destination of the transparent paths, routing packets in and out of the cut-through paths, and cannot benefit from optical bypass. Since domains are different economic entities, each pursuing their own profit, an architecture that decreases the cost in some domains and increases it in others is not realistic, even if the overall network cost decreases.

This issue however originates from the fact that our model is centered on the GÉANT domain. In a global interdomain model the transparent paths would be spread among the different domains following the Internet distribution of the traffic demand. Therefore every domain would take advantage of the cost savings of transparent switching.

We have carried out a sensitivity analysis, studying the effects of different equipment costs on the overall network cost. Decreasing the cost of IP cards and increasing that of optical switches diminishes the cost savings of the transparent architectures. Considering the interdomain model, for example, we have varied the costs by a factor of 4, halving the cost of the IP cards and doubling that of the optical switches. This led to a decrease of the maximum cost saving between transparent and opaque architectures from 83% to 67%. This modest decrease, compared to the high variation in cost considered, indicates that the results we have obtained would be moderately affected by possible cost changes in network equipment.

4.6 Cost savings of the OIS architecture: reconfigurability versus over-provisioning

In this section we consider the economical advantage brought about by the dynamic reconfiguration capability of Optical IP Switching. The scope of this analysis is to show how the cost of non-reconfigurable networks increases when we consider highly variable traffic demand, because of the over-provisioning necessary to satisfy the multiple traffic conditions that can occur. The reference models we use for this analysis are: point-to-point IP routing, static wavelength provisioning and OIS. Point-to-point routing, already described in section 4.5.2, presents the advantage of very fine routing granularity. However, as we have seen in the previous sections, when traffic increases, the cost of IP routing makes this approach not convenient economically. The static wavelength model is based on the pre-provisioning of dedicated optical links for end-to-end traffic demands above a pre-established threshold, similarly to the GMPLS model explored in section 4.5.2. However, since the lightpaths are non reconfigurable, the model requires higher over-provisioning to be able to cope with variations in the traffic demand. Finally, the OIS model represents the optical reconfigurable network, where lightpaths are dynamically adapted to the changing traffic demand.

4.6.1 Cost model

For this analysis we have not used the GÉANT dataset as a reference model. The reason is that we wanted to examine the cost variation of the competing architectures for variable traffic patterns, and we needed to arbitrarily control the variation of the traffic demand. We chose a ring topology interconnecting eleven nodes as a reference model, both for its inherent simplicity and to avoid topology and wavelength routing algorithm dependent results (lightpaths connect nodes either clockwise or counter-clockwise). Wavelength blocking is considered in this model, although the capacity of WDM links is automatically upgraded when more lightpaths are needed so that links are never exhausted.

Since the focus of this analysis is the effect of traffic variability on the different architectures, the simulation runs are based on the variation of the traffic demand matrix. The analysis starts by considering an arbitrary traffic demand, and calculating the solution (in terms of network equipment used) for the three network models described above. We then modify the demand matrix by randomly scrambling its elements, with the constraint that the amount of traffic added and dropped at each node remains constant (an approach known as "hose model" [DGG+02], [SW06]). In our case, for simplicity, we also consider that all the nodes add and drop the same amount of traffic. After modifying the demand matrix, a new solution is calculated for each of the three models. The amount of over-provisioning required (in terms of router cards, optical transceivers, link capacity and optical regenerators) in order to satisfy the different traffic demands is calculated as follows. For the point-to-point and OIS models we consider the union of the sets of equipment needed at each node, for each traffic pattern considered. The flexibility at the routing and optical layers in fact allows us to consider over-provisioning at the level of the routing or switching port (because the same ports can be re-used over different routing and optical paths). For the static wavelength provisioning, optical ports cannot be reconfigured to re-provision different paths, and the over-provisioning is operated at the level of the end-to-end path. In this model we have therefore considered the union of the end-to-end paths rather than the union of the ports used at each node.

In the static and dynamic wavelength models, dedicated optical paths are only considered for demands equal to the channel rate (i.e., 10Gbps). The demand matrix is constructed such that most of the traffic is exchanged between three nodes. The cost values are those illustrated in table 4.2. The assumptions we have considered throughout our analysis are generally consistent with the literature on cost modeling.

4.6.2 Results

The results of the reconfigurability cost model are shown in figures 4.20 and 4.21, expressing the network cost as a function of the variability of the traffic pattern, and considering, respectively, 100 Gbps and 400 Gbps of added and dropped traffic at each node. The percentage of transit traffic measured in the network configuration we have considered is the same at every node and is equal to 52%. We have considered four different levels of variability in the traffic demand, reported in the x-axis of the figures: static demand, low variation, medium variation and high variation. Static demand indicates that the network is only provisioned considering a single traffic demand. Low variation indicates that the network is required to operate on two (randomly generated) different traffic demands. Medium variation operates over four such demands, and high variation over eight. The higher the variation in the traffic demand, the higher the amount of over-provisioning needed and the more expensive the overall network cost.

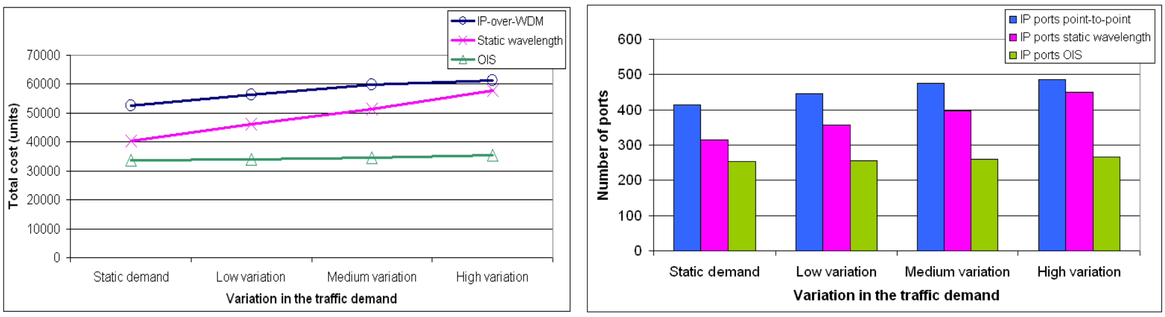


Figure 4.20: Cost analysis (left side) and ports count (right side) for 100 Gbps of add-drop traffic

By comparing the curves in the plots on the left side, we notice that the point-to-point model presents the highest cost compared to the transparent optical architectures. This result is the well-known benefit of optical bypass, analyzed in section 4.5.2. It is also in agreement with the results obtained in section 4.3.2, where the ring topology showed the highest values of switched-to-routed traffic, enhancing the cost saving capabilities of transparent architectures.

The cost for the static wavelength architecture is the most dependent on the traffic pattern because of its inability to reconfigure optical paths. The OIS architecture instead, being highly reconfigurable, does not show noticeable cost increase for traffic variations.

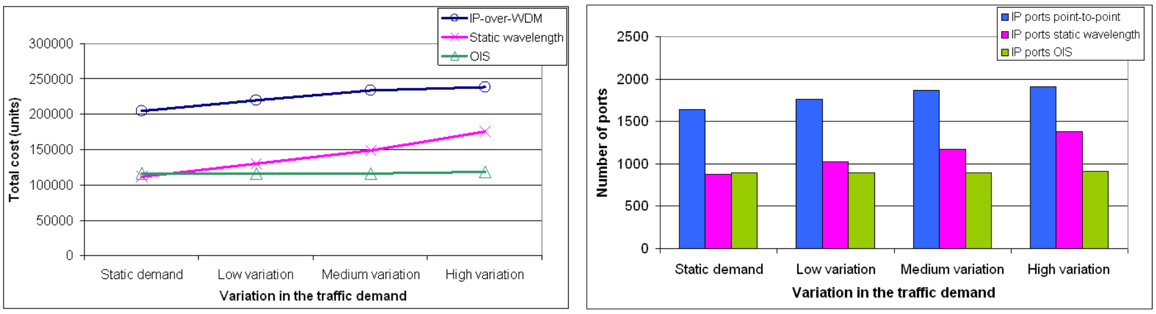


Figure 4.21: Cost analysis (left side) and ports count (right side) for 400 Gbps of add-drop traffic

The analysis of the number of IP ports used, reported on the plots on the right side of the figures, shows that the cost curves reflect very closely the trend of the IP port usage. This occurs because the cost of routing equipment is the highest in the network; therefore the overall cost is mostly influenced by the number of IP ports used.

Another aspect we notice, comparing the left plots of the two figures, is that the difference between the static wavelength provisioning and OIS models is more remarkable in the 100 Gbps add-drop case compared to the 400 Gbps. The effect, similarly to that observed in figure 4.15, is due to the fact that OIS can aggregate traffic better than optical end-to-end provisioning for lower amounts of traffic. When traffic increases however (figure 4.21) this advantage disappears, and for static traffic demand we see that the static-wavelength model is slightly more cost-effective than OIS. The reasons are that, on one hand, as observed in section 4.4, end-to-end provisioning allows a slightly more efficient path allocation for higher levels of traffic. On the other hand, when no traffic variation is considered, the static model exploits the advantage of deploying less expensive non-reconfigurable equipment. As we assume traffic variation however, the cost of over-provisioning sharply increases in the static wavelength model, leaving OIS the most cost-effective of the three architectures.

We have run simulations also considering higher levels of traffic and different traffic patterns. The results obtained showed that higher traffic leads to the same behavior illustrated in figure 4.21, and different traffic patterns (generated by uniformly distributed random variables) do not show considerable difference.

The cost sensitivity analysis carried out for this model has shown that even assuming considerable variation in the cost values (we have decreased the cost of IP cards by a factor of four and increased that of optical switches by the same amount), the dynamic provisioning architecture remains the most cost-effective under variable traffic demand.

4.7 Conclusions

In this chapter we have investigated, through simulations, the performance of the OIS architecture, analyzing its ability to bypass the IP layer with dynamic optical cut-through paths, and comparing its performance to other dynamic optical architectures. Here we summarize the most interesting results we have achieved.

Initially, we have shown that different path creation thresholds and extension algorithms generate a trade-off between the amount of data switched optically and the efficiency of channel occupancy. The values selected should depend, in a practical implementation, on the resources available in the network, and can be dynamically adapted to maximize the use of resources.

We have shown, through performance and cost analysis, that the effectiveness of the OIS architecture is highly dependent on the network topology considered. In general, topologies with lower node degrees enhance the switching ability of OIS.

The analysis carried out on the pan-European GÉANT network showed that the cost-effectiveness of OIS, although noticeable, is limited by the fact that source and destination nodes need to route the packets at the IP layer. If we allow however path source and destination to be located outside the GÉANT domain, extending the network transparency across the domain boundaries, OIS becomes highly convenient compared to legacy point-to-point routing.

The comparison of OIS to a dynamic, GMPLS-based model showed that the network performance and cost expenditures are not penalized by the distributed operations. The distributed OIS model in fact presents similar characteristics compared to a centralized network model.

The detailed analysis of network capital expenditures have demonstrated the ability of transparent reconfigurable optical networks to produce relevant cost saving in capital expenditure, when the traffic in the network increases and the demand matrix becomes more variable in time. First, by analyzing the pan-European GÉANT network, we have shown that optical transparency can allow savings of about 30% with respect to legacy point-to-point routed architectures, as traffic increases. It is only when transparency is allowed to cross the domain boundaries however, that the cost saving becomes very relevant, reaching values over 80%.

Finally, the cost analysis for highly variable traffic demand has shown that, although optical bypass of the IP layer can also be achieved with static wavelength provisioning, this approach is only convenient if the traffic demand is also static. If we consider variable demand instead, which more realistically characterizes the Internet behavior, the capability of provisioning paths dynamically becomes extremely important to maintain the cost-advantages of optical bypass.

Chapter 5

OIS optical testbed

In this chapter we describe the optical testbed we have assembled to implement the Optical IP Switching concept and the tests we have carried out to answer the remaining research questions raised in chapter 3. The testbed was used both to demonstrate the feasibility of the concept using off-the-shelf hardware, and to test the performance of the switching layer.

In section 5.1, we give a general overview of the testbed we have implemented between two university campuses in the city of Dublin.

We then illustrate in section 5.2 the node and network architecture, discussing the details of the hardware equipment and protocol stack used for the implementation. In this section we also describe the OIS layer, illustrating the implementation of the protocol mechanisms described in chapter 3.

Section 5.3 reports the results of the port and link discovery functions described in section 3.8.

In section 5.4 we analyze the path creation and extension times of our OIS implementation. These results are useful for estimating the effects of OIS path provisioning on Internet applications.

Finally, section 5.5 reports the implementation results of a platform integration between OIS and UCLP (described in section 2.3.2), giving a practical example of integration with an end-to-end Bandwidth-on-Demand (BoD) network.

5.1 OIS testbed network architecture

Figure 5.1 illustrates the network setup we have used to demonstrate the feasibility of the OIS architecture. The testbed contains equipment representing three core nodes of an OIS network, each connected to an access node. Two of the switching nodes are located in our laboratory in Trinity College, and the third is in the "Optical Communication Laboratory" in Dublin City University. The link between the two locations is 16 Km long and runs over a dark fiber supplied by HEAnet (Ireland's National Education and Research Network), on which we have implemented a three-wavelength dedicated bi-directional optical link. One of the channels is used as a default connection, carrying signaling and routed traffic, while the remaining two dynamically accommodate the optical cut-through paths.

Due to the short distance covered by the links, the physical impairments described at page 15 do not affect our testbed.

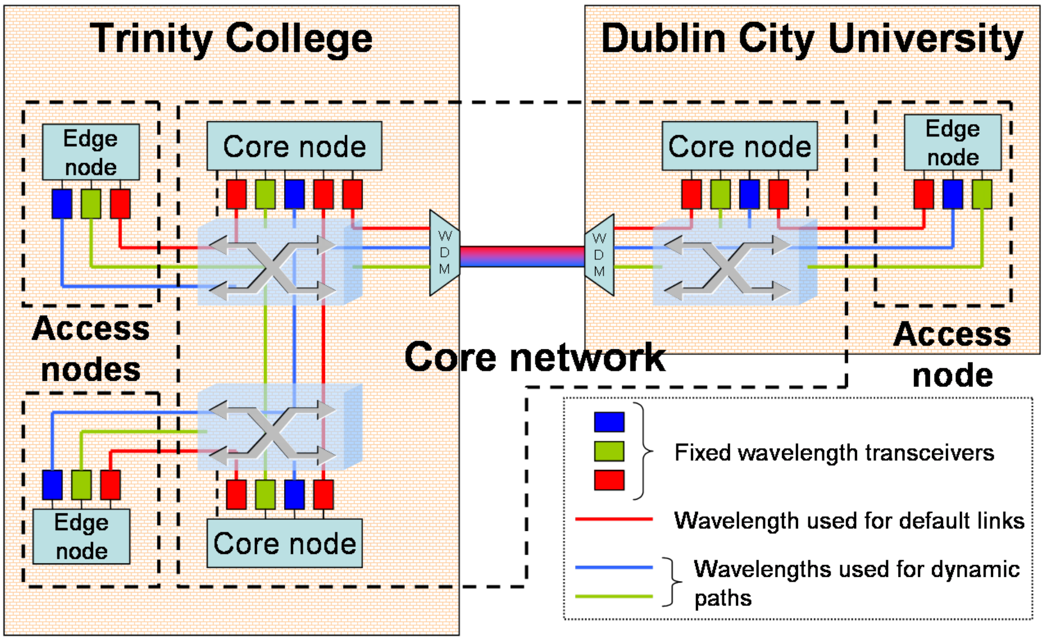


Figure 5.1: Network architecture of the OIS testbed

This set up aims at reproducing the interdomain model described in section 4.3.3, although in our experiments we have not considered differences in user-defined policies, assuming that all the nodes use similar threshold values. Only the core nodes are provided with switching capabilities, while the access nodes use transceivers operating on different wavelengths to create or terminate optical cut-through paths. Multiple fiber links between the nodes emulate WDM links.

For core and edge nodes we have used INTEL machines with 3-GHz Pentium 4 processors running the Windows operating system. Each machine was provided with multiple Gigabit Ethernet ports, each connected to an optical transceiver operating on one of the following CWDM wavelengths: 1510, 1530 and 1570 nm. The auto-negotiation function was disabled in the transceivers so that optical Ethernet packets could be sent without first assessing the presence of the receiver, a necessary condition to implement the port and link configuration mechanisms described in section 3.8. The optical switch is a Glimmerglass 16x16 port MEMs-based device with nominal switching time of 25 ms. Each server controls an optical switch through a dedicated TCP connection, using commands expressed in the Transaction Language 1 (TL1).

5.2 OIS testbed node architecture

Figure 5.2 shows the details of the OIS architecture we have implemented in the nodes. The OIS software was built on top of an existing protocol stack, used by researchers in our group to implement the Dublin Ad-hoc Wireless Network (DAWN). This stack runs as a normal user application and is made up of layers connected by a very simple interface.

Starting from the top, the "Application interface" constitutes the link between the protocol stack and the external applications, and operates through a Windows UDP socket opened on pre-established ports. The "Downmux layer" multiplexes the packets coming from different applications into the underlying protocol stack.

The "IP layer" performs the default routing operations, following the basics IPv4 protocol functions. Our implementation makes use of static IP tables for the routing operations. Since in fact our scope is the evaluation of the switching layer, we have only implemented the forwarding engine of the IP router.

Unlike the theoretical stack design illustrated in figure 3.2, we have implemented the optical layer as an independent entity below the IP layer rather than integrating the two. The main reason is that we wanted to have the option of reusing the Optilayer code with different routing algorithms and engines.

The "Upmux layer", similarly to the Downmux, multiplexes different I/O ports into the protocol stack. The Service Access Point (SAP) identifies the physical ports over which packets arrive and are relayed. Each packet is labeled with a "source_SAP" value, indicating the interface from which it was received, and a "dest_SAP" value obtained from the routing layer, which indicates the outgoing interface.

Finally, the "Datagram" and "Optical Datagram" are the elements that connect the protocol stack to the outside world through the Ethernet ports.

The modularity of the stack, although it reduces the throughput performance, has allowed us to integrate optical and wireless nodes together in a single highly dynamic network.

The Ethernet ports of the INTEL machines are connected to external Gigabit Ethernet optical transceivers, linked to the optical switch through single-mode optical fibers. The control connection between the router and the switch is achieved through a dedicated Ethernet connection, where a permanent TCP link allows the "Switch control" module to send the "TL1" commands.

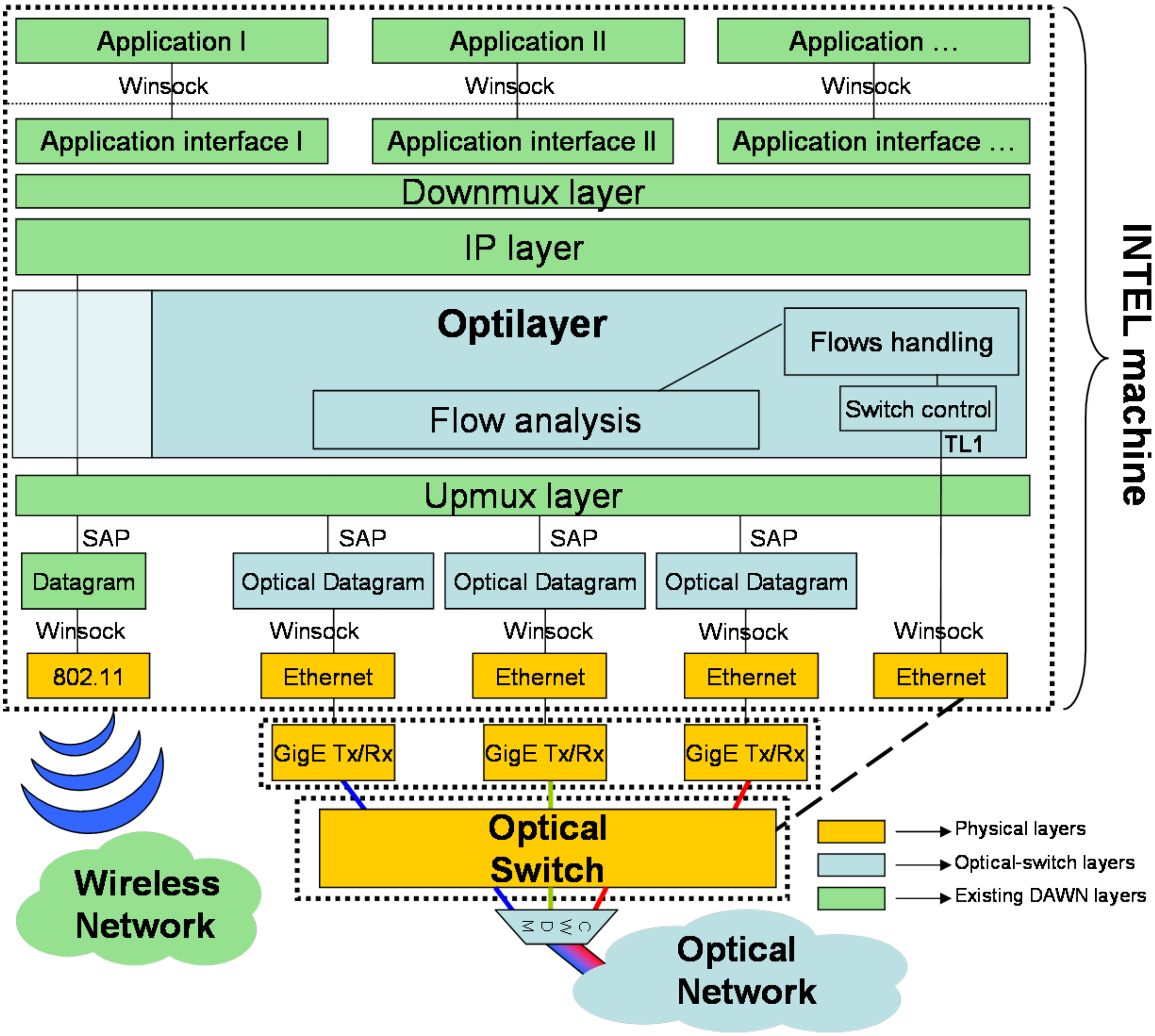


Figure 5.2: OIS testbed node architecture

5.2.1 Optilayer

The Optilayer is the part of the software stack that implements the Optical IP Switching concept. The first functions to be activated once the node starts up are the port and link discovery, described in section 3.8. These determine how the Ethernet ports are connected to the switch ports and how neighbors and wavelengths are associated with the external links, making the node completely self-configuring. After the configuration is complete the node is fully operational.

The "Flow analysis" block, pictured in figure 5.2, operates as an independent thread with the function of sampling and analyzing the IP packets. The packet sampling is operated after the routing destination has been determined by the IP layer, using the systematic count-based sampling method described in section 2.5.1. For each sampled packet, the “Flow Analysis” block determines its incoming interface (indicated by the source_SAP value) and its outgoing interface (dest_SAP), to associate the packet to the appropriate cell of the aggregation matrix (described in section 3.2.2). Within each cell, the packet contributes to determine the average data rate of the network prefix to which it is destined. In our implementation, because of the small number of nodes, we have used class-D IP addresses, but the implementation with network masks of different lengths is straightforward.

The information stored in the aggregation matrix is analyzed every time the decision timer, operated through the Windows timer function, expires. The first function to be operated when the decision process is triggered is the re-arrangement of the aggregation matrix, where the elements are sorted by decreasing order of average data rate. The prefix filtering operation described in section 3.2.2 can be applied before calculating the average rate.

The cell with the highest rate is the first to be considered for path extension or creation. After checking the availability of its own resources, the node selects a suitable wavelength, and it signals the upstream and downstream neighbors, assessing their capability of allocating an optical path as source and destination respectively. After both the nodes have acknowledged the request, the node sends them back a second acknowledgment, where it indicates the switched prefixes to the upstream node and the success of the path creation process to the downstream node. Thereafter, it activates the optical switch to create the dynamic path, and includes the path information in its “active path” list. Similar information is stored at the source and destination nodes. This signaling operation is illustrated in the message sequence chart in figure 5.3. The fields of the signaling messages used for the dynamic provisioning operations are illustrated in tables 5.1 and 5.2.

The source node is in charge of analyzing the traffic injected into the cut-through path (see section 3.3), and has the task of sending path refresh messages downstream (relayed hop-by-hop to the path destination). This maintains the reservation on the resources dedicated to the path, following a soft-state reservation mechanism that makes sure that resources are released in case of node failure. If however the source node decides that the path should be deleted, following the cancellation policy described in section 3.5, it sends a cancellation message downstream, following a hard-state mechanism, which assures fast release and reusability of the committed resources.

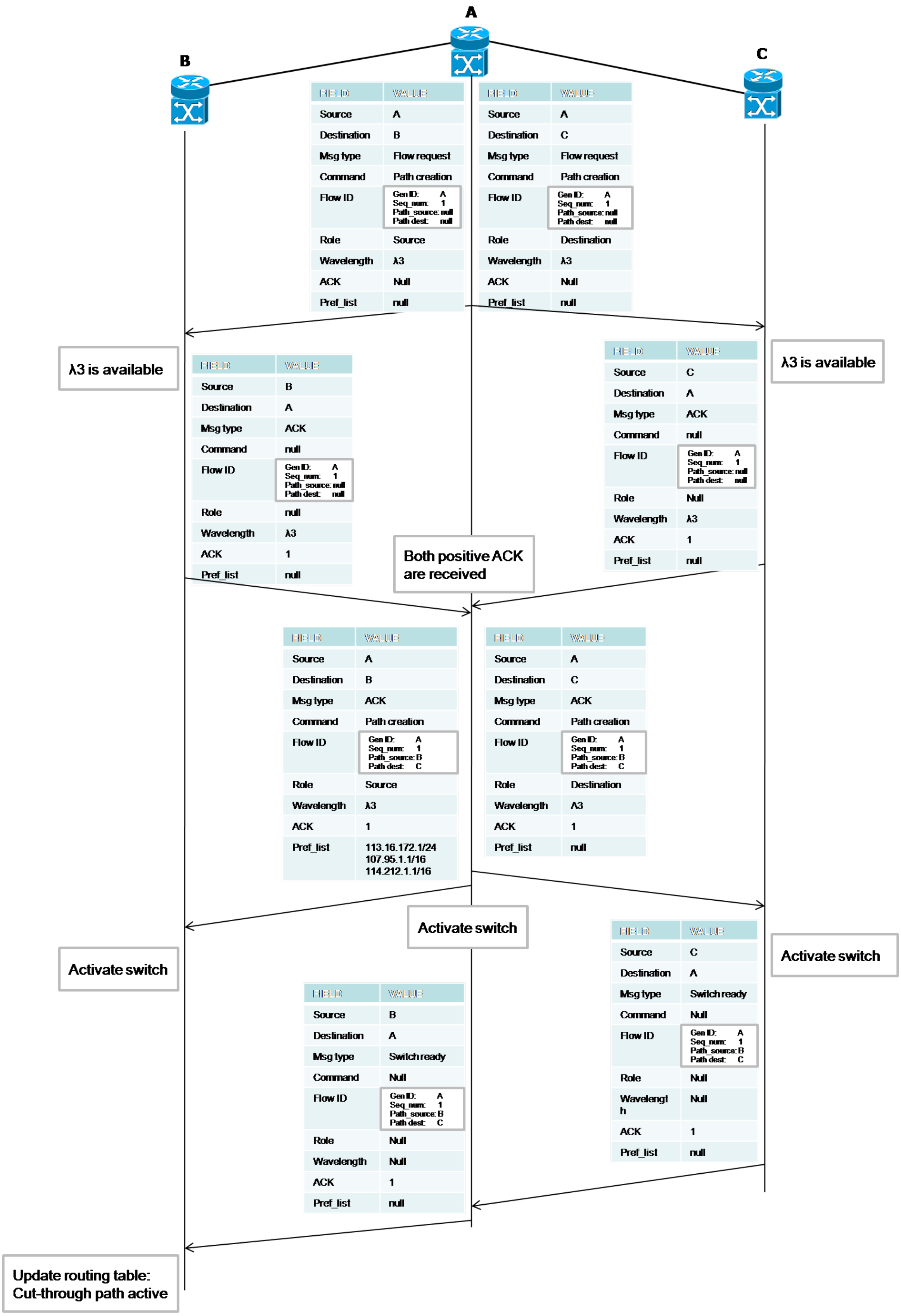
The "Switch control" block represents the interface with the optical switch, with the tasks of opening and maintaining an active TCP connection, and translating the requests received from the "Flows handling" block into TL1 commands.

Value	*Meaning*
Source	Network address of the message source
Dest	Network address of the message destination
Msg_Type	Defines the type of message: flow_request, ACK, prefix list, etc.
Command	Type of command: path creation, extension, cancellation, flow refresh, etc.
Flow_ID	see table 5.2
Role	Role of the receiving node: path source or destination
Wavelength	Wavelength proposed for the path creation or extension
ACK	1 for positive or 0 for negative acknowledgment
Prefix_list	List of prefixes switched by the current path

Table 5.1: Details of the Flow_signaling message

Value	*Meaning*
Generator_ID	Network address of the node that created the path
Seq_num	Unique flow sequence number (relative to the Generator_ID)
Path_source	Network address of the current path source
Path_dest	Network address of the current path destination

Table 5.2: Details of the Flow_ID field of the Flow_signaling message



FIELD	VALUE
Source	A
Destination	B
Msg type	Flow request
Command	Path creation
Flow ID	Gen ID: A Seq_num: 1 Path_source: null Path dest: null
Role	Source
Wavelength	λ3
ACK	Null
Pref_list	null

FIELD	VALUE
Source	A
Destination	C
Msg type	Flow request
Command	Path creation
Flow ID	Gen ID: A Seq_num: 1 Path_source: null Path dest: null
Role	Destination
Wavelength	λ3
ACK	Null
Pref_list	null

FIELD	VALUE
Source	B
Destination	A
Msg type	ACK
Command	null
Flow ID	Gen ID: A Seq_num: 1 Path_source: null Path dest: null
Role	null
Wavelength	λ3
ACK	1
Pref_list	null

FIELD	VALUE
Source	C
Destination	A
Msg type	ACK
Command	null
Flow ID	Gen ID: A Seq_num: 1 Path_source: null Path dest: null
Role	Null
Wavelength	λ3
ACK	1
Pref_list	null

FIELD	VALUE
Source	A
Destination	B
Msg type	ACK
Command	Path creation
Flow ID	Gen ID: A Seq_num: 1 Path_source: B Path dest: C
Role	Source
Wavelength	λ3
ACK	1
Pref_list	113.16.172.1/24 107.95.1.1/16 114.212.1.1/16

FIELD	VALUE
Source	A
Destination	C
Msg type	ACK
Command	Path creation
Flow ID	Gen ID: A Seq_num: 1 Path_source: B Path dest: C
Role	Destination
Wavelength	Λ3
ACK	1
Pref_list	null

FIELD	VALUE
Source	C
Destination	A
Msg type	Switch ready
Command	Null
Flow ID	Gen ID: A Seq_num: 1 Path_source: B Path dest: C
Role	Null
Wavelength	Null
ACK	1
Pref_list	null

FIELD	VALUE
Source	B
Destination	A
Msg type	Switch ready
Command	Null
Flow ID	Gen ID: A Seq_num: 1 Path_source: B Path dest: C
Role	Null
Wavelength	Null
ACK	1
Pref_list	null

Figure 5.3: Message sequence chart for the path creation mechanism

5.3 Automatic configuration

In this section we report the test performance of the automatic configuration mechanisms described in section 3.8, through simulation and testbed analysis. We have used simulations to examine the algorithm performance on systems with an arbitrary number of ports and neighbors and to compare them to reference algorithms. The testbed was used both to verify the feasibility of our approaches in a real network environment and to determine the exact timing of the operations.

5.3.1 Port discovery

The simulation results reported in this section compare the performance of the parallel port discovery algorithm introduced in section 3.8.1 with those of a default serial algorithm, which tests the switch ports one after the other until all have been discovered. The setup considers switches with N x N ports connected to routers with same number of TX and RX interfaces. We use the number of switch activations as a performance metric. Since in fact the switching time is the main source of delay in practical implementations, the configuration time can be considered proportional to the number of activations. We have expressed the number of TX-RX interfaces relatively to the number of ports, using values of N/8 and N/4. We expect typical values to be closer to N/8 so that most of the ports can be allocated to transparently switch traffic between adjacent nodes. For each configuration we have simulated 1000 trials using the Monte Carlo method to randomly connect the router interfaces to the switch ports.

The graphs in figure 5.4 report the average number of switch activations required to complete the discovery process, as a function of the switch dimensions (N).

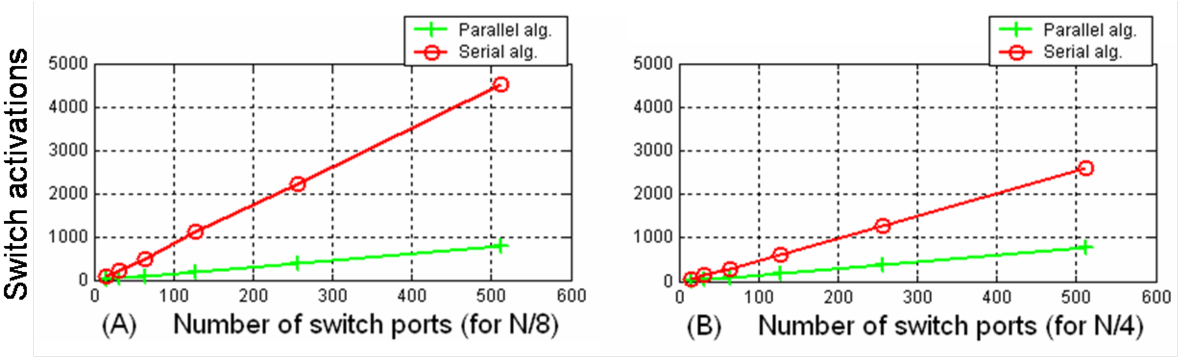


Figure 5.4: Average number of switch activations required to complete the discovery process

Figure 5.4.A reports the results when the number of TX and RX interfaces is N/8. We can see that the parallel algorithm we introduced performs much better than the serial scan, and the difference increases with the number of ports; we have a six-fold improvement when considering a 512-port switch. The improvement tends to decrease in figure 5.4.B, where the number of interfaces is N/4. The reason is that, while the performance of the parallel algorithm does not depend on the number of interfaces used, the serial algorithm improves when a higher number of interfaces is available.

In addition to their average values, it is interesting to examine the distribution of the simulation results. For this reason we have collected the results of the trials into histogram graphs.

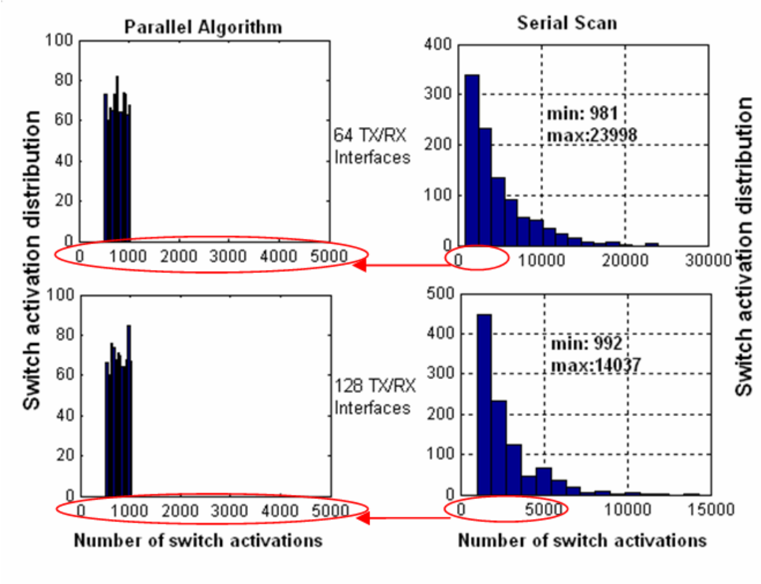


Figure 5.5: Distribution of the number of switch activations required

Figure 5.5 shows the distribution obtained with a 512-port switch (other cases present similar behavior). We can see that for the parallel algorithm the distribution is generally uniform and confined to a narrow interval of values (513 to 1024 switch activation, precisely as envisaged in section 3.8.1). The serial scan instead presents a decreasing exponential envelope and the results span over a much wider interval, which decreases when the number of interfaces increases.

From the results obtained we can conclude that the parallel algorithm outperforms the serial scan; the number of switch activations required is lower in the average and is distributed over a smaller interval. From a practical point of view, a system implementing the parallel algorithm would have shorter and less variable configuration times.

In the testbed analysis of our parallel discovery algorithm, we have achieved an average value of 47 milliseconds per iteration. Although the MEMs switching time is in the order of 20 ms, we had to introduce further delay to allow the switch to process two consecutive TL1 commands, and consider the time needed by the protocol stack to send, receive and process the test messages.

If we apply this value to the previous simulation results we can infer typical configuration times of larger systems. Considering a 512 x 512 ports switch with N/8 number of interfaces, for example, we can see that our algorithm would terminate in 36 seconds on average, where the serial algorithm would need 3 minutes 32 seconds. Moreover, while the parallel algorithm has also a very low maximum configuration time (48 seconds), the serial algorithm may require up to 18 minutes 48 seconds to configure the system, due to the long tail of the iteration distribution.

5.3.2 Link discovery

This section reports the simulation and testbed results of the link discovery protocol proposed in section 3.8.2, considering switches of different sizes, different number of peering neighbors, different number of router interfaces and different values for the RX_SCAN_TIME parameter.

The test topology considers M nodes connected in a full mesh using W wavelengths per link. Each node is constituted by a router with T TX and R RX interfaces and a transparent optical switch of N x N ports. We have chosen values of 16, 64 and 256 for N, and values of 3, 10 and 30 for W. We have considered the case of non-tunable transmitters (the worst case scenario). The number T of transmitters should be high enough to cover the M-1 control channels (one for each neighbour) and the complete range of wavelengths W. For convenience reasons, all the nodes use transmitters operating at the same wavelength to establish the control channels. The number of TX interfaces at each node is T = W+M-1, where M-1 represents the number of peering neighbors. We also assume T = R.

For comparison, we have considered the idea proposed in [ELW+04]. Even though the authors do not go into the details of the discovery algorithm, we can infer an approximate performance calculation. The idea they present is that on startup a node signals the start of a discovery session to all the peering neighbors. This node synchronizes its neighbors so that the scanning process can proceed in an orderly manner. We have approximated the number of switch reconfigurations needed by this algorithm with the following formula:

$$Reconf = \frac{(N-W)\cdot(N-R)}{2R} \tag{5.1}$$

Here follows a brief demonstration. The node has N-T undiscovered ports, which are tested using W of the T interfaces at the same time; (N-T)/W switch reconfigurations are required to cover all the output ports. Since we consider that all the control channels are out-of-band, we can assume T = W; therefore (N-T)/W = (N-W)/W. Considering that we use non-tunable transmitters, each port needs to be tested on each wavelength using a different TX interface; therefore the number of reconfigurations increases by a factor of W. For each of these reconfigurations the neighboring nodes need to scan all their incoming ports; assuming the use of broadband receivers, each operation can be completed in $\frac{(N-R-Discovered\_ports)}{R}$ steps. At the beginning, when many ports are undiscovered, this value will be close to (N-R)/R, while towards the end it will be close to "0" (assuming that most of the optical ports are used for external links). We average this behavior by using a value of (N-R)/2R. Since the process needs to be repeated similarly for the discovery of the input ports, we need to add a further multiplication factor of 2:

$$Reconf = 2\cdot\frac{(N-W)\cdot(N-R)}{2\cdot W\cdot R}\cdot W \tag{5.2}$$

which simplifies into:

$$Reconf = \frac{(N-W)\cdot(N-R)}{R} \tag{5.3}$$

This calculation however considers a worst-case scenario where a complete scan of all the ports is required each time. On average only half the ports need to be scanned, so we can again divide by 2, which brings us back to formula 5.1.

We have used our testbed to prove the feasibility of our approach in a real system, obtaining a practical value of 60 ms per switch activation. The simulations were used to consider more complex topologies with multiple nodes and several links.

Figure 5.6 reports the results of the link discovery simulations that we have carried out using the simulation parameters shown in Table 5.3.

On the x-axis we report the sequential number of the input and output ports, and on the y-axis the time of the discovery expressed in milliseconds. The graphs on the left side are

	Sim A	*Sim B*	*Sim C*
Number of neighbors	3	5	7
Switch size (N)	16 x 16	64 x 64	256 x 256
Wavelengths per link (W) (excluding the control channel)	2	9	29
Number of interfaces (T=R)	5	14	36

Table 5.3: Parameters used for the simulations

averaged over different simulation runs, and each line is obtained using a different value for the RX_SCAN_TIME parameter. On the right side we report the best and worst results (representing the lower and upper limits), obtained using the most performing RX_SCAN_TIME value (e.g., 500 ms for figure 5.6.A). Figures B and C report the results obtained using the other simulation parameters illustrated in table 5.3.

As we can see, the choice of the RX_SCAN_TIME parameter is moderately relevant as long as the value selected is not excessively small (red line in the left graphs). By comparing the best values for the RX_SCAN_TIME with the simulation parameters in table 5.3, we can confirm that, as already anticipated in section 3.8.2, the best value is directly proportional to the number of ports of neighbor switches and to the link wavelengths; it is inversely proportional instead to the number of TX interfaces used by the neighbors:

$$RX\_SCAN\_TIME \propto \frac{N_{neigh} \cdot W_{towards-neigh}}{T_{neigh}} \tag{5.4}$$

For a practical implementation, considering that a different optimum value can be estimated for each neighbor, and that choosing a value too low has much worse effects than choosing a value too high, the best tactic is to choose the largest of all the estimated optimum RX_SCAN_TIME values.

Table 5.4 reports the results obtained by integrating testbed experimentation and simulation analysis of the link configuration algorithm. Our discovery algorithm is compared to the synchronous algorithm described in [ELW+04], using the formula 5.1. Notice that, being the control channel out-of-band in the synchronous algorithm, we have chosen to set the number of transmitters T equal to the number of wavelengths W; the same consideration applies to the number of receivers R.

The synchronous algorithm completes the task faster than our algorithm. However we have to consider that the algorithm we propose also establishes the control channel within the

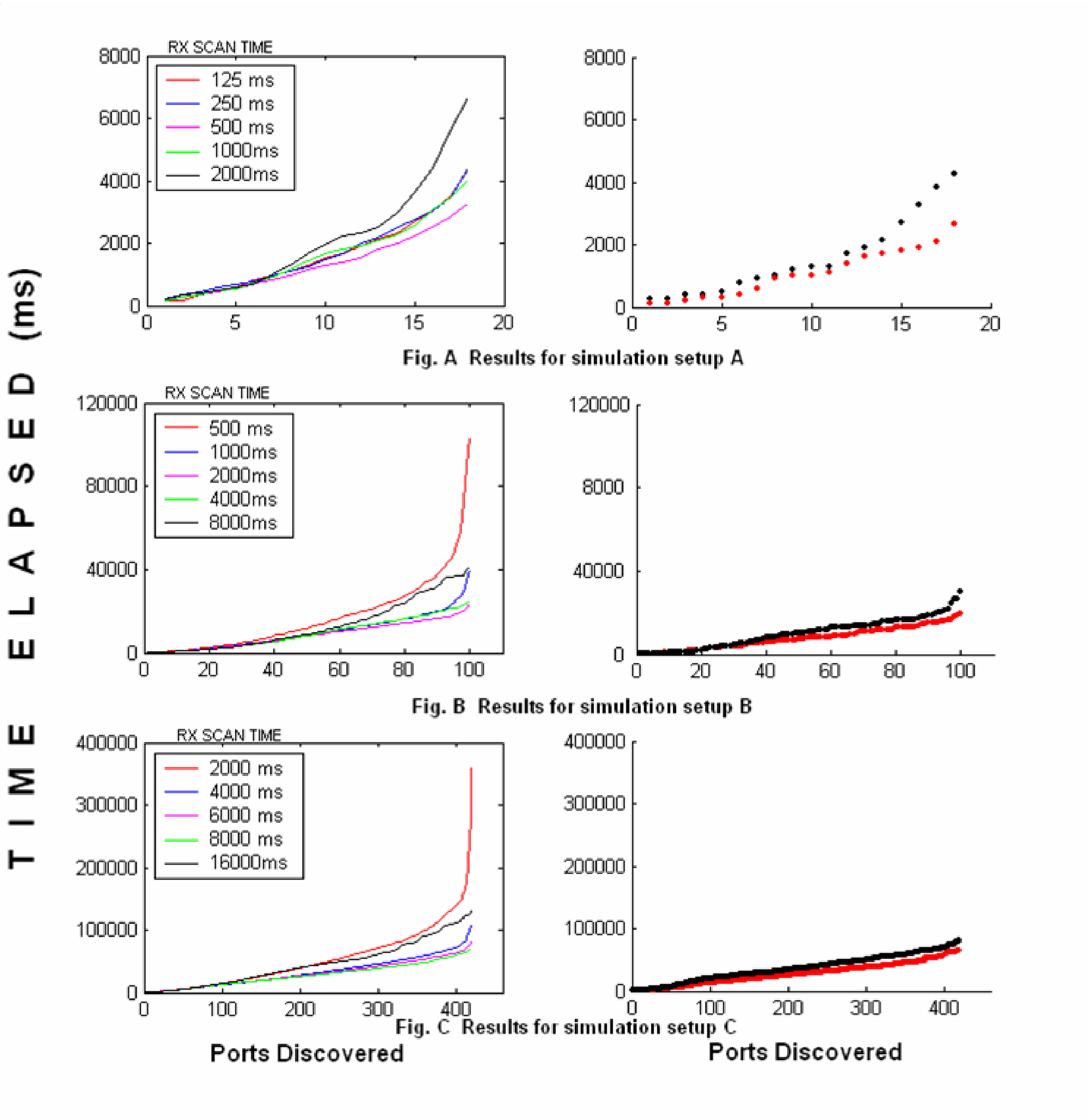


Figure 5.6: Simulation results of our synchronous algorithm

Simulations	*Synchronous Algorithm*	*Asynchronous Algorithm*
Sim A	2.9	3.3
Sim B	10.0	23.7
Sim C	53.3	72.5

Table 5.4: Comparison between the synchronous and our asynchronous algorithm (time values are displayed in seconds)

same time, while the reference algorithm assumes that control channels are manually setup. If we add the time needed to manually setup a control channel for each peering neighbor (which we have estimated, based on our experience, to be approximately five minutes), we can see that the algorithm we propose could reduce the link discovery time between one and two orders of magnitude (the values for Sim A-C would be approximately 903, 1510 and 2153 respectively).

The results obtained with our asynchronous algorithm considered in-band control channels. If the network administrator decides to use out-of-band signaling, then the synchronous algorithm appears to be a better choice.

In conclusion, the results obtained allows us to answer the research question raised in section 3.8: efficient self-configuration is possible in transparent networks.

5.4 Switching times of dynamic paths

In this section we report the testbed results on dynamic path allocation, showing typical path creation and extension times obtained with our OIS platform. We use these results to analyze the effects of dynamic path provisioning on the end-to-end packet transfer for different transmission rates. In fact, as we have anticipated in section 3.9, dynamic creation, extension and cancellation of optical paths might cause packet loss and out of order arrival, compromising the performance of Internet transport protocols.

Following an initial performance investigation, we also propose and analyze mechanisms that help reduce the packet loss observed during path extension.

The packet transfer tests are performed through the UDP protocol, which allows us to analyze delay, loss and arrival order of individual datagrams. The effects on the performance of UDP-based applications and of the TCP protocol are discussed in section 5.4.4.

5.4.1 Path creation: effects of switching from default to dedicated link

Disruption to the transport protocols can originate during path creation when packets are switched from the default IP-routed to the dedicated optical switched path. Although there is no strict constraint on the path creation time, because packets can be routed over the default links until the optical path is setup, passing from a routed to a switched path could cause out of order arrival, as packets on the dedicated link travel faster than packets on the routed paths.

Packet loss could occur if the source node starts sending packets over the dedicated cut-through path before the path is effectively created. However, our use of double acknowledgments makes sure that packets are transmitted only after the cut-through path is in place.

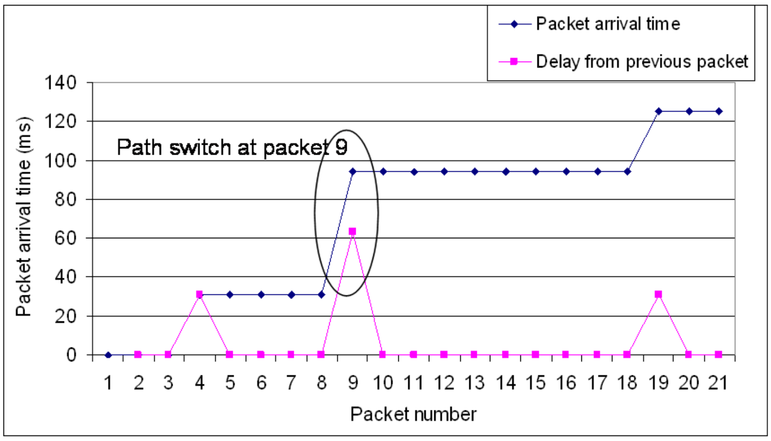


Figure 5.7: Path creation test results for burst interval times of 31 ms

Figure 5.7 illustrates the results obtained with packets sent every 31 ms in bursts of 5. The size of each packet is 2 Kbytes, originating a flow rate of 2.58 Mbps. The blue line in the plot shows the arrival time of the packets, while the purple line illustrates the relative delay with respect to the previous packet. In the x-axis we report the packet sequential number. As we can see from the picture, there is a delay on the received packets during the provisioning operation. Such a delay is related to our stack implementation, and originates from the signaling and packet processing required by the path creation mechanism, which delays the stack from delivering data packets.

In figure 5.8 we have further decreased the packet interval to 15 ms, emulating a flow rate of 5.16 Mbps; the results are similar to those in figure 5.7.

As we can see from the plots typical path creation times are in the order of 30ms.

From the tests we have performed, we have not observed any out of order packet arrival when packets are switched from the routed path to the optical cut-through path. The main reason is that OIS creates paths between three nodes, where only one node is optically bypassed, making the time difference between the two routes sufficiently small.

In general, packet reordering should not occur as long as the flows are a few Mbps in size and the packet forwarding time in the middle node remains low. If however such time increases, for example due to link congestion, the delay in forwarding the packet could cause out of order arrival at the receiver. In this case the source node, after sending the last packet

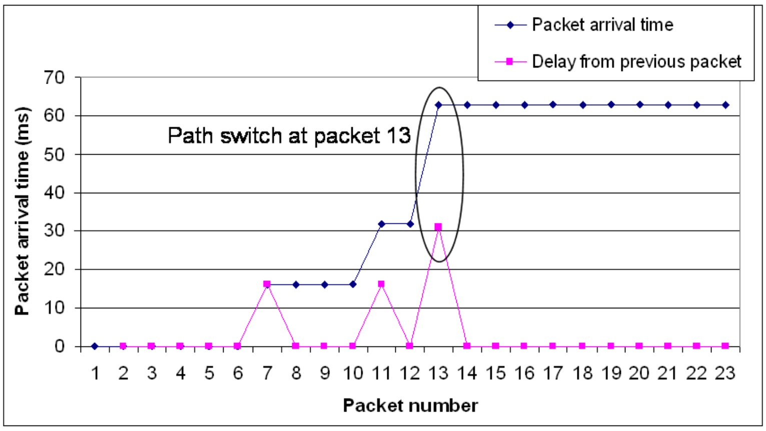


Figure 5.8: Path creation test results for burst interval times of 15 ms

over the default IP link, should wait for an appropriate time interval before sending packets over the cut-through path, storing such packets in a local buffer. A similar solution will be introduced in section 5.4.3 for the path extension mechanism.

5.4.2 Path extension: modification of an active path

This section examines the impairments at the transport layer that can originate from the extension process. The issue arises because the extension modifies a path that is already transporting data traffic; all the packets traversing the switch during the switching time are lost.

Figure 5.9 illustrates the packet arrival time during a path extension operation, with packet inter-arrival time of 31 ms. Similarly to the path creation observed in figure 5.7, we notice an increase of the delay during the provisioning operation.

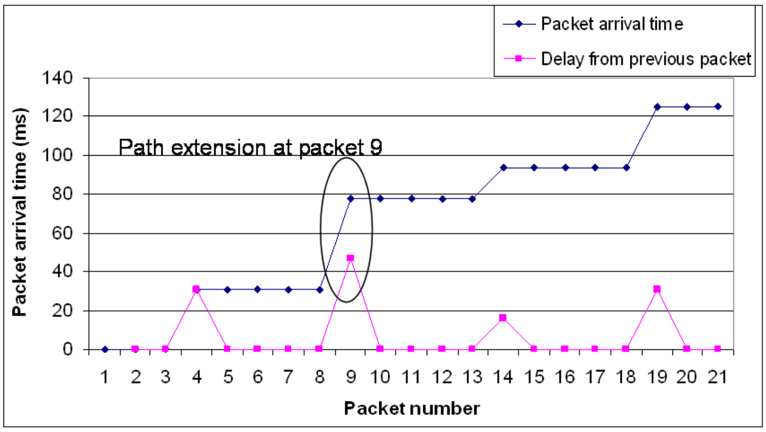


Figure 5.9: Path extension test results for burst interval times of 31 ms

When the packet interval time gets close to the switching time of the optical switch (for most cases we have measured actual switching times between 16 and 18 ms), the probability that packets are transmitted during the switching operation becomes high, causing packet loss. This case is illustrated in figure 5.10, where the packet inter-arrival time is 15 ms. In this case the extension operation causes the loss of the burst transmitted during the switching time.

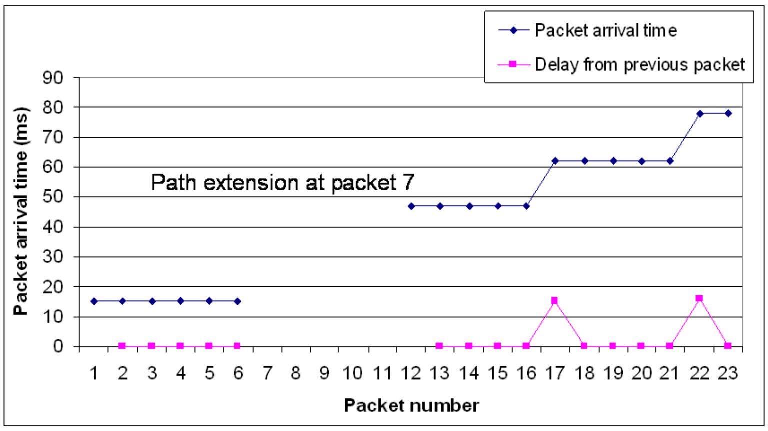


Figure 5.10: Path extension test results for burst interval times of 15 ms

5.4.3 Path extension without packet loss

Since packet loss has a high impact on the perceived quality of UDP applications and on the performance of the TCP protocol, we have implemented a mechanism to avoid packet loss during path extension. The idea is to avoid packets crossing the optical switch while switching is in operation.

In the case of downstream extension, just before triggering the optical switch the node sends a "STOP_flow" message to the path source, which, on receipt, stops sending traffic towards the selected path, buffering all the packets for an appropriate period of time. This stop time, which includes the switching time plus a guard interval, was set in our testbed to 30 ms. After sending the "STOP_flow" message, the node extending the path needs to wait, before activating the switch, long enough for the path source to receive and process the message, and for the packets already in the optical path to reach and traverse the switch. This waiting time is proportional to the distance (both in terms of number of hops and physical distance) between extending node and path source, and needs to be estimated by the node.

In some cases, especially when the number of hops is large or a node is congested, the routing time might be highly variable and it might be difficult to make a correct estimation. A more complex signaling mechanism should then be implemented to reduce the chances of packet loss. In this case, after receiving the "STOP_flow" message, the path source sends back an "Acknowledgment" message and starts buffering the packets. Only after receiving the "Acknowledgment" the node extending the path actuates the switch to physically extend the path. After the operation is complete, the node sends back to the path source a "RESTART_flow" message, which triggers the path source to start using again the cut-through path.

In the case of upstream extension the node extending the path is also the path source. In this case, after having agreed on the path extension with the upstream node, the node sends a "STOP_flow" message upstream, following the procedure described for the downstream extension. Since in this case the path source and its upstream node are direct neighbors, the signaling procedure is faster and does not need to be operated using a complex signaling mechanism.

Figure 5.11 shows the results of implementing the "STOP_flow" signaling on the path extension. The mechanism eliminates the packet loss we observed in figure 5.10, introducing instead a longer delay in the packets' arrival. Typical extension times are in the order of 50 ms.

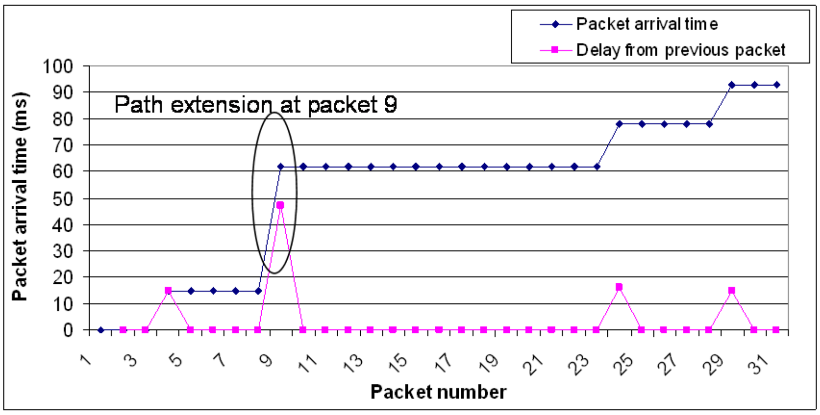


Figure 5.11: Path extension test results with “STOP_flow” mechanism, for burst interval times of 15 ms

The drawback of this method is the buffer required at the source node to store all the packets during the "STOP_flow" time interval. The buffer size depends both on the aggregate flow rate in the cut-through path and on the time needed to complete the switching operation. A worst case scenario, considering an aggregate flow rate of 10Gbps and an extension time of 100 ms, would require a buffer size of 125 Mbytes. The estimated cost for the additional router memory is less than $ 100, which is negligible compared to the overall router cost.

Buffering packets at the source node could also be used, as already anticipated in section 5.4.2, to avoid out of order arrival on path creation when the middle router is congested. The challenge in this case is to estimate correctly the packet latency time at the router, which is generally highly variable when congestion occurs.

In the next session we will discuss the effects of the observed delay and packet loss on UDP and TCP applications, and also the effects of the "STOP_flow" mechanism we have developed.

5.4.4 Effects of the observed packet loss and delays on UDP and TCP applications

Packet loss, variable delay and out of order arrival can degrade the quality of applications transported both over UDP and TCP protocols. In our tests we have not observed out of order arrival but we have experienced packet loss and variable delay. In this section we try to understand the impact of the measured values on the transport protocols and evaluate the benefits of the "STOP_flow" mechanism introduced in section 5.4.3.

We have seen in section 3.9.1 that for live stream applications transported through UDP, even low packet loss or jitter can deteriorate the perceived quality of the signal. In our case however, the jitter or packet loss is not constant over time but only concentrated around the switching time, so that the effect on a streaming application might be noticeable but very limited in time. For UDP-based applications, since jitter and packet loss have similar effects [CT99], there would be no advantage in implementing the "STOP_flow" mechanism.

The issue is different for the TCP protocol, where packet loss causes both the retransmission of packets (all those included in the same TCP window where the loss occurred) and the shrinking of the acknowledgment window, significantly slowing down the data transfer process. The effect of packet jitter on protocol performance instead is usually negligible, as long as it is not above the TCP timeout for the acknowledgment (usually of the order of a few seconds).

These considerations suggest that, unlike for UDP, it is advantageous for TCP to implement the "STOP_flow" mechanism, as it can compensate for the transport protocol impairments originating from the path extension.

With the results we have obtained we can answer the research question raised in section 3.9: it is possible to eliminate the packet loss and out of order arrival by buffering packets at the source for an appropriate time interval. Since however the algorithm we have proposed introduces delay in the packet delivery, it might not be effective in improving the signal quality in the case of live streaming applications without pre-buffering.

5.5 OIS interface with UCLP

Since the interdomain environment is highly heterogeneous, one of the main requirements for novel Internet architectures is the capability of efficiently interfacing with different network models. An Optical IP Switched network would therefore coexist with other optical architectures in a real implementation.

Most of next-generation network architectures implement the Bandwidth-on-Demand (BoD) concept, where end-to-end channels with arbitrary bandwidth can be requested by a user. In this section we consider the case where different domains, some of them implementing OIS, are connected through BoD networks. The situation is illustrated in figure 5.12, where the OIS domains are pictured in green, the BoD domains in blue and other generic Internet domains in yellow .

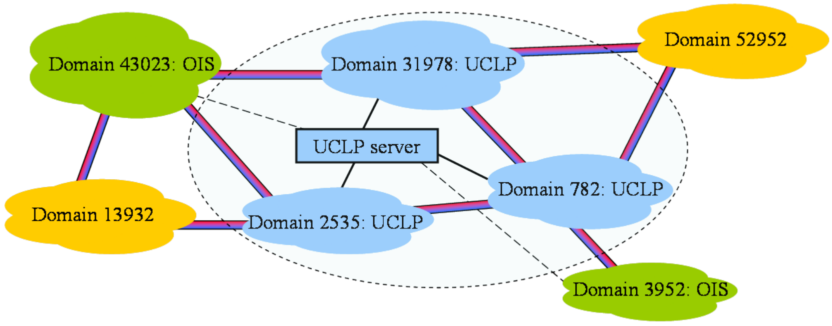


Figure 5.12: Example of multi-domain heterogeneous network environment

Although most of current BoD implementations follow the IETF GMPLS standard described on page 22, we have considered the UCLP platform (see page 25). The reason is that an opportunity arose for a collaborative project with HEAnet (Ireland's National Education & Research Network [INERN07]), the i2Cat Foundation (a platform for developing Internet technologies in Catalonia [iF07]) and Glimmerglass (vendor of MEMs-based optical switches [Gli07]), to work on the extension of the UCLP software to transparent optical switches.

The first part of the project involved testing the UCLP implementation over a large-scale testbed, connecting the i2Cat laboratory in Barcelona (Spain) with our CTVR laboratory in Dublin, using a dedicated link provided by the European GÉANT network. The results, reported in [SFJ+07], showed the ability of the UCLP platform to provide user-requested BoD services through a graphical user interface, creating dedicated path over heterogeneous links (combining MPLS, Ethernet, and all-optical switching technologies).

The second part of the project involved the integration of UCLP and OIS, so that BoD services could be requested automatically by the OIS nodes, following traffic flow analysis. In this model OIS acts as a client of the external UCLP networks. When an OIS node detects traffic destined for a specific external domain, it can automatically request the bandwidth needed to reach the desired destination. The UCLP server is the central unit in charge of signaling and scheduling operations, receiving the bandwidth requests, calculating the best routes and signaling the network elements to provision the optical paths.

The interface we have developed for this experiment allows the OIS node to log into the UCLP server, download the network topology and request/release dedicated end-to-end paths. The novelty of our implementation is that we have provided the UCLP server with a mechanism that stores an updated list of network prefixes reachable through each node. In a practical implementation the prefix list could be engineered by the UCLP network administrator to optimize the resource usage.

The connection and login into the UCLP server is initiated at startup time by the OIS nodes during the auto-configuration process. The OIS operations of flow analysis and path creation, extension and cancellation proceed as described in chapter 3. If there is a suitable traffic aggregate directed towards one of the networks advertised by the UCLP server, the OIS protocol will request the UCLP server to provision a direct optical link from the OIS ingress point to the appropriate UCLP node. If the destination point is a gateway to another OIS domain the path created can be further extended within that domain, through distributed OIS operations.

5.5.1 OIS-UCLP testbed result

Figure 5.13 shows the testbed setup for the experimentation of the OIS-UCLP interface. The UCLP network is manually configured with the information about the nodes connected to the network, together with the destination prefixes advertised. The connections between the OIS gateway, the UCLP server and the UCLP nodes are operated through a generic transport network (emulating the Internet). In order to perform the test over a real environment we set up the testbed between two university campuses, emulating two separate domains connected through the optical links described in section 5.1.

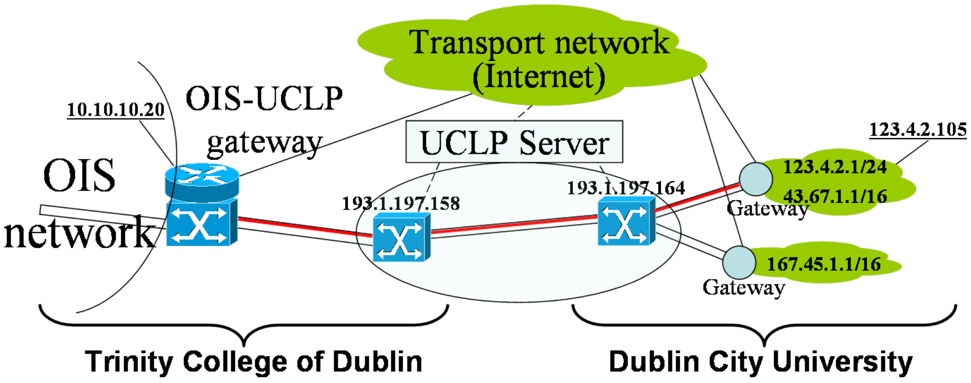


Figure 5.13: OIS-UCLP testbed

At startup, the OIS gateway logs into the UCLP server, requesting the list of prefixes reachable through all the switches registered to the UCLP network.

During operations, low rate flows are transported from source to destination through the default transport network. When we start a multimedia application, which generates higher rate traffic from node 10.10.10.20 towards the address 123.4.2.105, the OIS node detects the IP flow, triggering the path creation mechanism. Since the 123.4.2.1/24 prefix appears between the list of networks reachable through UCLP, the OIS gateway requests an optical path from its UCLP access point (an optical input port of the switch 193.1.197.158) towards the gateway associated with the prefix 123.4.2.1/24 (an optical output port of the switch 193.1.197.164).

After receiving the request, the UCLP server computes the path route and creates the optical path by sending TL1 commands to the optical switches. The optical path remains active as long as the traffic towards the UCLP destination remains above the path cancellation threshold. The OIS gateway, the source of the path requested, is responsible for deleting the UCLP path when the associated OIS cut-through path is deleted.

This experiment demonstrated the feasibility of integrating the OIS architecture with a realistic bandwidth on demand network, where bandwidth requests are automatically forwarded by the nodes without user intervention.

One of the main achievements was testing our idea of abstracting the UCLP topology by considering the IP network prefixes reachable through the nodes connected to the UCLP federation. Although the list of prefixes was manually configured in our case, in future implementations this information could be automatically collected and processed by the UCLP server.

Finally, during our tests we measured UCLP path creation times between 1 and 1.4 seconds. This delay is generated by the internal operation of the UCLP platform, for which we noticed that the path creation time is proportional to the number of hops in the UCLP path.

5.6 Conclusions

In this section we have described the testbed we implemented to demonstrate the practical feasibility of the OIS concept and to investigate the performance of the OIS architecture in transporting real data.

Initially, we demonstrated that automatic configuration of optical ports and links is possible in transparent networks. We have proposed an efficient and fast algorithm that uses parallel switch operations to determine the port configuration of an OIS node. We have also introduced an effective method of link discovery, capable of determining both data and (in-band) control channels. The combination of both methods provides OIS with full plug-and-play capability.

As a consequence of the availability of the testbed we have been able to report on typical path creation and extension times. We have also analyzed the effects of the switching operations on packet loss, delay and out of order arrival, in order to investigate the impact of OIS on the UDP and TCP transport protocols. Our investigation showed that, for flows measuring a few Mbps (i.e., a typical size for an elephant flow), optical path creation does not generate transport impairments unless router congestion occurs at the switching node. The situation is different for the path extension process, where packet loss generally occurs as switching operations are performed on optical paths carrying data traffic. We have proposed a method that interrupts the traffic flow in the path while the switching is in operation and that can be implemented by introducing small additional memory at the IP router. This method eliminates any packet loss originated by the extension operation. However, it introduces a small jitter on the transport channel that could create minor disruptions to live stream applications transported through the UDP protocol.

Finally, we have analyzed a network scenario where the distributed OIS network interacts with centralized bandwidth-on-demand networks to automatically create dedicated lightpaths across different domains. We have developed the interfaces for the integration of the OIS and UCLP platforms, proposing a method for abstracting the UCLP network topology in terms of network prefixes reachable through the UCLP domain.

Chapter 6

Conclusions and future work

Recent advances in optical technology brought to the market new devices like Reconfigurable Optical Add Drop Multiplexers (ROADM) and MEMs-based optical cross-connects, increasing the reconfiguration capabilities of optical networks.

The benefits of implementing transparent switching in data networks is twofold: it increases the network flexibility, allowing faster deployment of bandwidth services, and it generates large cost savings, both by decreasing the need for over-provisioning and reducing considerably the switching cost (compared to electrical cross-connects and IP routers).

Although the practical implementation of network transparency is still limited by the low flexibility of the optical transport layer, many transparent network architectures have already been proposed and standardization bodies are defining a protocol suite to manage and coordinate optical link reconfiguration.

Current proposed optical architectures generally focus on the ability of provisioning end-to-end lightpaths, either under request of the network administrator (e.g., to implement traffic engineering decisions) or of an external user (for bandwidth on demand services). Although this concept is in line with intradomain operations, where full topology information is shared among the nodes (which can therefore cooperate for the overall network benefit), its application in the interdomain poses several issues.

The interdomain structure on which the Internet is built is a highly competitive environment where domains share a very limited amount of information. Under such circumstances, provisioning end-to-end paths across several domains through a legacy source routed approach that follows a centrally operated decision seems very ambitious.

In this dissertation we have introduced Optical IP Switching, an optical network architecture we have developed at CTVR that provisions lightpaths with a distributed approach that is both fully compatible with existing IP architectures and compliant with the interdomain networking concept. Distributed operations in fact, besides adding scalability and fault-tolerance to the network design, allow the nodes to make individual decisions, consistent with their domain policies and resource availability.

A distinctive characteristic of our OIS approach is that the traffic engineering mechanism is integrated in each node; the nodes are capable of provisioning lightpaths autonomously without the need for a centralized management plane that computes the lightpath topology. Autonomous operations are based on the distributed observation (at every node) of data packets, which are first sampled and analyzed as IP flows, and then aggregated taking into account their route and destination network. Flow aggregates that are compliant with the node policies are re-routed through dynamically provisioned optical cut-through paths, bypassing the IP routing layer.

The simulations we have performed on the OIS model have shown its ability to redirect traffic from the electronic IP layer down to the optical switching layer, displaying how the architecture performance varies for different algorithms and network topologies. For the main reference topology that we have used (the pan-European GÉANT network), the ratio of switched-to-routed data increases with traffic volume, saturating to a value of about 36% (i.e., it reaches asymptotically the percentage of transit traffic in the network). However, if we allow optical paths to cross the domain boundaries, simulating interdomain transparent operations, the saturation value increases to 93%. The importance of such results becomes clearer in the cost analysis we have performed. The results of the cost analysis show that while in the intradomain model we have obtained maximum savings of about 30% (compared to the legacy point-to-point routing model), in the interdomain model the saving possibility was well over 80%. The performance of our OIS architecture was also compared to a centralized optical provisioning (GMPLS-based) model. Although OIS utilizes only local traffic information, it allows similar cost saving compared to the centralized GMPLS model, which instead makes use of centralized global knowledge of the traffic demand distribution to provision the optical paths.

The Optical IP Switching architecture was also implemented in an optical testbed, made up of three core nodes and three edge nodes. Besides demonstrating the practical feasibility of the OIS concept using off-the-shelf hardware components, our tests demonstrated the efficacy of the auto-configuration mechanisms we have designed, and allowed us to analyze the impact of transparent switching on the UDP and TCP protocols. These results have suggested to us some modifications to the path provisioning protocols that minimize the impact of transparent switching on these Internet transport protocols.

6.1 Summary of contributions

The main contributions of this dissertation are the following:

- We have introduced the Optical IP Switching concept of distributed lightpath provisioning, where the nodes dynamically engineer traffic routes and network links through local observation of IP flows.
- We have developed distributed algorithms for optical path creation, extension and cancellation, which can operate either on a cooperative intradomain model or can be configured for operation in the competitive interdomain environment.
- We have addressed the large gap in granularity between IP packet routing and wavelength switching by developing a method that uses routing prefixes to accumulate multiple flows into larger aggregates. We have shown how the heavy-tail distribution typically observed on Internet traffic flows also occurs at the IP-prefix granularity. This allows us to extend some of the results that apply to flow-based traffic engineering (for example, rate-based flow filtering) also to prefix aggregates.
- We have implemented a signaling mechanism enabling the distributed creation, extension and cancellation of lightpaths. We have also indicated how the local decision mechanism can be integrated with shared acknowledgments to facilitate interdomain operations.
- We have implemented a port and link discovery mechanism that makes use of parallel switch operations to achieve fast self-configuration for the OIS nodes.
- We have shown through simulations that the capability of switching IP traffic optically is for OIS highly dependent on the network topology considered. For the reference pan-European GÉANT network (but similar results are applicable to core networks of similar size) we have shown that the switching ability increases considerably if lightpaths are allowed to cross the domain boundaries.
- The network cost analysis we have performed shows that transparent and reconfigurable switching operations increase the cost savings, compared to legacy point-to-point rout-

ing and static wavelength provisioning network models. Our cost model confirms the results obtained in the technical analysis, showing that cost savings of the OIS architecture increase with its switching capability and are considerably more evident in the interdomain.

- We have developed an OIS prototype network using inexpensive, commercially available components. Besides demonstrating the practical feasibility of the overall OIS concept and the methods we have developed, we have analyzed the impact of highly dynamic optical switching on the transport protocol performance. Our results show that the effects of OIS switching operations can be reduced to a minimum by fine-tuning the signaling parameters of the path creation and extension processes.

6.2 Future work

The primary scope of this dissertation has been the overall design of the Optical IP Switching architecture. During the development and test phases we identified some research topics worth further investigation. They are summarized below:

- While we have developed distributed signaling protocols and algorithms for dynamic lightpath provisioning, and implemented them on a testbed prototype, we have not focused our efforts on the algorithm optimization. We have noticed in section 4.2, for example, that the first fit wavelength assignment algorithm shows poor performance in terms of wavelength efficiency, which in some case drops below 50%. Higher performance wavelength selection algorithms could be developed, drawing on the distributed networking research field. Similarly, more optimal solutions could be identified for the path creation, extension and cancellation mechanisms.

- The lightpath provisioning mechanisms we have described in chapter 3, by following the routing information available from the IP table, always use the routes selected by the default IP routing algorithms. For this reason the OIS routes optimize the same parameters (e.g., number of hops or delay) as the underlying routing protocol. If the IP layer is capable of handling multiple routes for quality of service differentiation however (for example implementing the MPLS protocol), OIS paths could be built over routes that are different from those selected by the default routing protocol. The OIS provisioning mechanisms could therefore be modified to follow the routes established by traffic engineering decisions made by the network operator at the IP/MPLS layer.

- A research area that should be further explored is the policy specification for interdomain lightpath provisioning. We have already indicated in section 4.3.3 how the distributed mechanisms of Optical IP Switching facilitates interdomain operations, but we have not defined explicitly how current domain policies should be modified to include optical provisioning. Since interdomain operations are carried out in a highly competitive environment, optimal solutions to this issue could be derived from the game theory field of research.

- Although the prototype we have developed fully implements the concept of transparent optical switching, the routing layer has very low throughput because the protocol stack we used was not optimized for performance. It would be interesting to re-develop the concept on a faster router software (e.g., Quagga/Zebra, XORP or the Click Modular Router Project) and implement time-critical functions in FPGA hardware. The Click Modular Router Project ([Koh00]) seems to be particularly suitable for this task, both for its modularity and because the functions implemented through the "Click" language can be more easily mapped into FPGA hardware.

- As we have indicated in section 3.2.2, the OIS architecture does not allow quality of service differentiation as it does not control individual IP flows. We have recently developed, together with the "High Speed Networks and Optical Communications Group" of the Athens Information Technology center, the idea of a hybrid optical architecture that merges the concepts of flow routing ([Rob03]) and Optical IP Switching. In the model developed, while OIS performs optimization of the optical resources by re-routing lightpaths depending on the traffic encountered, the flow routing layer operates flow-by-flow QoS differentiation. The main issue associated with this approach is that the constraints imposed by the QoS requirements of the IP flows increase the complexity of the lightpath provisioning algorithms of the OIS architecture.

In section 2.6.1, we have indicated how the near-future trend of research in optical networking seems oriented, on one hand towards the developments of dynamic transparent provisioning, and on the other hand on the close integration of multiple network technologies into a unique, service-oriented architecture. Following this evolutionary line, we envisage that in future optical networks the distributed operations of OIS would be closely integrated with other provisioning technologies. While end-to-end oriented platforms like GMPLS and UCLP could satisfy specific point-to-point bandwidth requests (generated, for example, by high-end grid applications), the OIS distributed mechanism could operate multi-layer traffic engineering, automatically provisioning the most cost-effective lightpath configuration, while taking into account the requirements of IP flows and switched paths at the higher layers.

Bibliography

[Agr95] G. P. Agrawal. *Nonlinear Fiber Optics.* Academic Press, 1995.

[AM01] D. J. Arent and A. Martin. Third-generation DWDM networks near reality. *PenWell Lightwave Magazine*, 18(3), 2001.

[And01] L. Andersson. LDP specification. IETF RFC 3036, 2001.

[Ash06] G. R. Ash. *Traffic Engineering and QoS Optimization of Integrated Voice and Data Networks.* Elsevier, Morgan Kaufmann publishers, 2006.

[AWK$^+$99] G. Apostolopoulos, D. Williams, S. Kamat, R. Guerin, A. Orda, and T. Przygienda. QoS routing mechanisms and OSPF extensions. IETF RFC 2676, 1999.

[Bar64] P. Baran. On distributed communications networks. *IEEE Transactions on Communications*, 12(1):1–9, 1964.

[Bar03] G. Barlow. A G.709 Optical Transport Network tutorial. Technical report, Innocor Ltd., 2003.

[BBR$^+$03] D. J. Blumenthal, J. E. Bowers, L. Rau, C. Hsu-Feng, S. Rangarajan, W. Wang, and K. N. Poulsen. Optical signal processing for optical packet switching networks. *IEEE Communications Magazine*, 41(2):S23–S29, 2003.

[BC02] N. Brownlee and K. C. Claffy. Understanding Internet traffic streams: Dragonflies and tortoises. *IEEE Communications Magazine*, 40(10):110–117, 2002.

[BC06] M. P. Belanger and M. Cavallari. Network cost impact of solution for mitigating optical impairments: Comparison of methods, techniques, and practical deployments constraints. *Journal of Lightwave Technology*, 24(1):150–157, 2006.

[BCW$^+$06] E. Van Breusegern, J. Cheyns, D. De Winter, D. Colle, M. Pickavet, F. De Turck, and P. Demeester. Overspill routing in optical networks: a true hybrid optical network design. *IEEE Journal on Selected Areas in Communications*, 24(4):13–25, 2006.

[Ber03a] L. Berger. Generalized Multi-Protocol Label Switching (GMPLS) signaling Constraint-based Routed Label Distribution Protocol (CR-LDP) extensions. IETF RFC 3472, 2003.

[Ber03b] L. Berger. Generalized Multi-Protocol Label Switching (GMPLS) signaling resource ReSerVation Protocol-Traffic Engineering (RSVP-TE) extensions. IETF RFC 3473, 2003.

[BG98] J. M. Boyce and R. D. Gaglianello. Packet loss effects on MPEG video sent over the public Internet. In *Proceedings of ACM International Conference on Multimedia*, pages 181–190, 1998.

[BHS03] S. Bjørnstad, D. R. Hjelme, and N. Stol. A highly efficient optical packet switching node design supporting guaranteed service. In *Proceedings of European Conference on Optical Communication*, pages 110–111, 2003.

[Bra97] R. Braden. Resource ReSerVation Protocol (RSVP)–version 1 functional specification. IETF RFC 2205, 1997.

[BRS03] G. Bernstein, B. Rajagopalan, and D. Saha. *Optical Network Control. Architecture, Protocols, and Standards*. Addison-Wesley, 2003.

[BSD$^+$06] M. Brown, L. Smarr, T. DeFanti, J. Leigh, M. Ellisman, and P. Papadopoulos. The OptIPuter: A national and global-scale cyberinfrastructure for enabling lambdagrid computing. In *Proceedings of TeraGrid Conference*, 2006.

[Bur06] Interactive Advertising Bureau. Internet advertising revenues close to 4 billion dollars for Q1 2006. http://www.iab.net/news/, 2006.

[CAI03] Cooperative association for Internet data analysis CAIDA. Anonymized OC-48 traces. http://www.caida.org, 2003.

[CAN07] Canada's advanced network CANARIE. User Controlled LightPath (UCLP). http://www.uclp.ca, 2007.

[CBC$^+$04] J. Cheyns, E. Van Breusegem, D. Colle, M. Pickavet, P. Demeester, and D. De Winter. Controlling LSPs in an ORION network. In *Proceedings of Broadnets conference*, pages 74–81, 2004.

[CBD$^+$03] C. Cavazzoni, V. Barosco, A. D'Alessandro, A. Manzalini, S. Milani, G. Ricucci, R. Morro, R. Geerdsen, U. Hartmer, G. Lehr, U. Pauluhn, S. Wevering, D. Pendarakis, N. Wauters, R. Gigantino, J. P. Vasseur, K. Shimano, G. Monari, and A. Salvioniet. The IP/MPLS over ASON/GMPLS test bed of the IST project LION. *IEEE Journal of Lightwave Technology*, 21(11):2791–2803, 2003.

[CT99] M. Claypool and J. Tanner. The effects of jitter on the peceptual quality of video. In *Proceedings of ACM international conference on Multimedia*, pages 115–118, 1999.

[CWM06] V. W. S. Chan, G. Weichenberg, and M. Medard. Optical Flow Switching. In *Proceedings of International Workshop on Optical Burst/Packet Switching*, 2006.

[CY03] A. Chiu and C. Yu. Economic benefits of trasparent OXC network as compared to long systems with OADMs. In *Proceedings of Optical Fiber Communication conference*, 2003.

[Dan02] M. Dannhardt. Ethernet over Sonet. Technical report, PMC-Sierra, 2002.

[DBSW67] D. W. Davies, K. A. Barlett, R. A. Scantlebury, and P. T. Wilkinson. A digital communication network for computers giving rapid response at remote terminals. In *Proceedings of ACM Symposium on Operating Systems Principles*, pages 1–17, 1967.

[DDC$^+$07] T. Dietz, F. Dressler, G. Carle, B. Claise, and P. Aitken. Information model for packet sampling exports. IETF Draft 'draft-ietf-psamp-info-06', 2007.

[DG03] L. Ding and R.A. Goubran. Assessment of effects of packet loss on speech quality in VoIP. In *Proceedings of IEEE Workshop on Haptic Audio and Visual Environments and Their Applications*, pages 49 – 54, 2003.

[DG07] E. A. Doumith and M. Gagnaire. Impact of traffic predictability on WDM EXC/OXC network performance. *IEEE Journal on Selected Areas in Communications*, 25(5):895–904, 2007.

[DGG+02] N. G. Duffield, P. Goyal, A. Greenberg, P. Mishra, K. K. Ramakrishnan, and J. E. Van der Merwe. Resource management with hoses: point-to-cloud services for virtual private networks. *IEEE/ACM Transactions on Networking*, 10(5):679–692, 2002.

[EGG+06] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, and T. Roughgarden. Routers with very small buffers. In *Proceedings of IEEE Infocom conference*, pages 1–11, 2006.

[EJLW01] A. Eldid, C. Jin, S. Low, and I. Widjaja. MATE: MPLS Adaptive Traffic Engineering. In *Proceedings of IEEE Infocom conference*, pages 1300–1309, 2001.

[ELW+04] G. Ellinas, J.F. Labourdette, J.A. Walker, S. Chaudhuri, L. Lin, E. Goldstein, and K. Bala. Network control and management challenges in opaque networks utilizing transparent optical switches. *IEEE Communications Magazine*, 42(2):S16–S24, 2004.

[EV01] C. Estan and G. Varghese. New directions in traffic measurement and accounting. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, pages 75–80, 2001.

[Fau98] F. Le Faucheur. IETF multiprotocol label switching (MPLS) architecture. In *Proceedings of IEEE International Conference on ATM*, pages 6–15, 1998.

[FLM02] P. Ferreira, W. Lehr, and L. McKnight. Optical networks and the future of broadband services. *Elsvier, Technological Forecasting and Social Change Journal*, 69(7):741–758, 2002.

[Fre05] A. Fredette. Link Management Protocol (LMP) for Dense Wavelength Division Multiplexing (DWDM) optical line systems. IETF RFC 4209, 2005.

[FT00] B. Fortz and M. Thorup. Internet traffic engineering by optimizing OSPF weights. In *Proceedings of IEEE Infocom conference*, pages 519–528, 2000.

[Fun03a] National Science Fundation. 100x100 clean slate program. http://www.100x100network.org/, 2003.

[Fun03b] National Science Fundation. Dynamic Resource Allocation via GMPLS Optical Networks (dragon). http://dragon.east.isi.edu, 2003.

[Fun04] National Science Fundation. Circuit-switched High-speed End-to-End Transport ArcHitecture (CHEETAH). http://cheetah.cs.virginia.edu, 2004.

[Fun06a] National Science Fundation. Dynamic Optical Circuit Switched (DOCS) networks for future large scale dynamic networking environments. http://www.nets-find.net/DynamicOptical.php, 2006.

[Fun06b] National Science Fundation. Future Internet Network Design (FIND). http://www.nets-find.net/, 2006.

[Fun06c] National Science Fundation. Future optical network architectures. http://www.nets-find.net/Future.php, 2006.

[Fun06d] National Science Fundation. Global Environment for Network Innovation (GENI). http://www.geni.net, 2006.

[Gli07] Glimmerglass. http://www.glimmerglass.com/, 2007.

[GLW$^+$06] M. Gunkel, R. Leppla, M. Wade, A. Lord, D. Schupke, G. Lehmann, C. Furst, S. Bodamer, B. Bollenz, H. Haustein, H. Nakajima, and J. Martensson. A cost model for the WDM layer. In *Proceedings of Photonics in Switching conference*, 2006.

[GSM05] S. F. Gieselman, N. K. Singhal, and B. Mukherjee. Minimum-cost virtual-topology adaptation for optical WDM mesh networks. In *Proceedings of IEEE International conference on communications*, pages 1787–1791, 2005.

[HBCR07] J. He, M. Bresler, M. Chiang, and J. Rexford. Towards robust multi-layer traffic engineering: Optimization of congestion control and routing. *IEEE Journal on Selected Areas in Communications*, 25(5):868–880, 2007.

[HN01] A. Hill and F. Neri. Optical switching networks: from circuits to packets. *IEEE Communications Magazine*, 39(3):107–108, 2001.

[HSLR06] I. W. Habib, Q. Song, Z. Li, and N. S. V. Rao. Deployment of the GMPLS control plane for grid applications in experimental high-performance networks. *IEEE Communications Magazine*, 44(3):65–73, 2006.

[HYM$^+$99] T. Hayashi, S. Yamasaki, N. Morita, H. Aida, M. Takeichi, and N. Doi. Effects of IP packet loss and picture frame reduction on MPEG1 subjective quality. In

Proceedings of IEEE Workshop on Multimedia Signal Processing, pages 515 – 520, 1999.

[iF07] i2Cat Fundation. http://i2cat.net, 2007.

[IM99] J. Iness and B. Mukherjee. Sparse wavelength conversion in wavelength-routed WDM optical networks. *Springer Photonic Network Communications Journal*, 1(3):183–205, 1999.

[INERN07] Ireland's National Education and (HEANET) Research Network. http://www.heanet.ie/, 2007.

[Int05] Internet2. Hybrid Optical and Packet Infrastructure (HOPI). http://networks.internet2.edu/hopi/, 2005.

[IST00] European Framework Program 5 IST. Layers interworking in optical networks. http://www.telecom.ntua.gr/lion, 2000.

[IST04a] European Framework Program 6 IST. MUlti-Partner european testBED for research networking (MUPBED). http://www.ist-mupbed.org, 2004.

[IST04b] European Framework Program 6 IST. Next generation Optical networks for Broadband European Leadership (NOBEL). http://www.ist-nobel.org, 2004.

[IST06] European Framework Program 6 IST. Transparent Ring Interconnection Using Multi-wavelength PHotonic switches (TRIUMPH). http://www.ihq.uni-karlsruhe.de/research/projects/TRIUMPH, 2006.

[IST07] European Framework Program 7 IST. Dynamic Impairment COnstraint NETwork for transparent mesh optical networks (DICONET). http://www.diconet.eu/homepage.asp, 2007.

[IT96] ITU-T. Methods for subjective determination of transmission quality. Recommendation P.800, 1996.

[IT01] ITU-T. Architecture for the Automatically Switched Optical Network (ASON). Recommendation G.8080, 2001.

[IT03a] ITU-T. Generic Framing Procedure (GFP). Recommendation G.7041, 2003.

[IT03b] ITU-T. Interfaces for the Optical Transport Network (OTN). Recommendation G.709, 2003.

[JFN04] G. L. Jones, W. Forysiak, and J. H. B. Nijhof. Economic benefits of all-optical cross connects and multi-haul DWDM systems for european national networks. In *Proceedings of Opical Fiber Communication Conference*, 2004.

[Jon04] J. D. Jones. User Network Interface (UNI) 1.0 signaling specification, release 2. OIF implementation agreement, 2004.

[JPL+04] H. C. Ji, K. J. Park, J. Lee, H. Chung, E. Son, K. Han, S. Jun, and Y. Chung. Optical performance monitoring techniques based on pilot tones for WDM network applications. *OSA Journal of Optical Networking*, 3(7):510–533, 2004.

[Kam03] Kamelian. Semiconductor Optical Amplifiers (SOAs) in multi-channel environments. Technical report, Kamelian, 2003.

[KIA+05] A. Kirstädter, A. Iselt, A. Autenrieth, D.A. Schupke, R. Prinz, and B. Edmaier. Business models for next generation transport networks. *Springer Photonic Network Communications Journal*, 10(3):283–296, 2005.

[KKHR04] M. Karasek, J. Kanka, P. Honzatko, and J. Radil. Protection of surviving channels in all-optical gain-clamped lumped raman fibre amplifier: modelling and experimentation. *Elsevier Optics Communication journal*, 231(39234):309–317, 2004.

[Koh00] E. Kohler. *The Click modular router.* PhD thesis, Massachussets Institute of Technology, 2000.

[Kom05] K. Kompella. OSPF extensions in support of generalized multi-protocol label switching. IETF RFC 4203, 2005.

[Lan04] J. Lang. Link Management Protocol (LMP). IETF RFC 4204, 2004.

[Lar02] N. Larkin. ASON and GMPLS - the battle of the optical control plane. Technical report, Data Connection Ltd., 2002.

[LC07] S. S. M. Lo and R. K. C. Chang. Measuring the effects of route prepending for stub autonomous systems. In *Proceedings of IEEE Traffic Engineering in*

Next Generation IP Networks workshop, International Conference on Communications, pages 3–4, 2007.

[LK00] R. Ludwig and R.H. Katz. The Eifel algorithm: Making TCP robust against spurious retransmissions. *ACM SIGCOMM Computer Communication Review*, 30(1):30–36, 2000.

[LLB+03] O. Leclerc, B. Lavigne, E. Balmefrezol, P. Brindel, L. Pierre, D. Rouvillain, and F. Seguineau. Optical regeneration at 40 Gb/s and beyond. *Journal of Lightwave Technology*, 21(11):2779–2790, 2003.

[LM97] S. Lin and N. McKeown. A simulation study of IP switching. *ACM SIGCOMM Computer Communication Review*, 27(4):15–24, 1997.

[LSJ06] T. Lehman, J. Sobieski, and B. Jabbari. DRAGON: A framework for service provisioning in heterogeneous grid networks. *IEEE Communications Magazine*, 44(3):84–90, 2006.

[Mac06] C. M. Machuca. Expenditures study for network operators. In *Proceedings of International Conference on Transparent Optical Networks*, pages 18–22, 2006.

[Man04] E. Mannie. Generalized Multi-Protocol Label Switching (GMPLS) architecture. IETF RFC 3945, 2004.

[Meh03] V. Mehta. TCP transmission over a reconfigurable optical acces network. In *Proceedings of Annual Meeeting of IEEE Lasers and Electro-Optics Society, LEOS.*, pages 965 – 966, 2003.

[MPC+06] R. Martinez, C. Pinart, F. Cugini, N. Andriolli, L. Valcarenghi, P. Castoldi, L. Wosinska, J. Cornelias, and G. Junyent. Challenges and requirements for introducing impairment-awareness into the management and control planes of ASON/GMPLS WDM networks. *IEEE Communications Magazine*, 44(12):76–85, 2006.

[MRS+06] G. Mulvihill, M. Ruffini, F. Smith, L. Barry, L. Doyle, and D. O'Mahony. Optical ip switching a solution to dynamic lightpath establishment in disaggregated network architectures. In *Proceedings of International Conference on Transparent Optical Networks*, pages 78–81, 2006.

[MUK$^+$04] T. Mori, M. Uchida, R. Kawahara, J. Pan, and S. Goto. Identifying elephant flows through periodically sampled packets. In *Proceedings of Internet Measurement Conference*, pages 115–120, 2004.

[Muk06] B. Mukherjee. *All-Optical Impairment-Aware Routing.* Springer US, 2006.

[Net07] Nortel Networks. Provider backbone transport. Technical report, 2007.

[NML98] P. Newman, G. Minshall, and T. L. Lyon. IP switching-ATM under IP. *IEEE/ACM Transactions on Networking*, 6(2):117–129, 1998.

[Ong04] L. Y. Ong. Intra-carrier E-NNI signaling specification. OIF implementation agreement, 2004.

[OSS$^+$05] E. Oki, K. Shiomoto, D. Shimazaki, N. Yamanaka, W. Imajuku, and Y. Takigawa. Dynamic multilayer routing schemes in GMPLS-based IP+optical networks. *IEEE Communications Magazine*, 43(1):108–114, 2005.

[Ovi07] Ovisor. http://www.ovisor.com/Products-WSS-Arch.html, 2007.

[PAMS06] C. Politi, V. Anagnostopoulos, C. Matrakidis, and A. Stavdas. Physical layer impairment aware routing algorithms based on analytically calculated Q-factor. In *Proceedings of OFC/NFOEC conference*, 2006.

[Por85] M. E. Porter. *Competitive Advantage.* Collier Macmillan, 1985.

[Pos80] J. Postel. User Datagram Protocol (UDP). IETF RFC 768, 1980.

[Pos81] J. Postel. Transmission Control Protocol (TCP). IETF RFC 793, 1981.

[PTB$^+$01] K. Papagiannaki, N. Taft, S. Bhattacharya, P. Thiran, K. Salamatian, and C. Diot. On the feasibility of identifying elephants in Internet backbone traffic. Technical report, Sprint ATL, Sprint Labs, 2001.

[QU05] B. Quoitin and S. Uhlig. Modeling the routing of an autonomous system with C-BGP. *IEEE Networks*, 19(6):12–19, 2005.

[Quo06] B. Quoitin. *BGP-based Interdomain Traffic Engineering.* PhD thesis, Université catholique de Louvain-la-Neuve, 2006.

[QY99] C. Qiao and M. Yoo. Optical Burst Switching (OBS) - a new paradigm for an optical Internet. *Journal of High Speed Networks*, 8(1):69–84, 1999.

[RLH06] Y. Rekhter, T. Li, and S. Hares. A Border Gateway Protocol 4 (bgp-4). IETF RFC 4271, 2006.

[Rob03] L. G. Roberts. The next generation of IP - flow routing. In *Proceedings of SSGRR International Conference*, 2003.

[Ros01] E. Rosen. Multiprotocol label switching architecture. IETF RFC 3031, 2001.

[SFJ$^+$07] A. Sánchez, S. Figuerola, G. Junyent, E. Kenny, V. Reijs, and M. Ruffini. A user provisioning tool for EoMPLS services based on UCLPv1.5. In *Proceedings of Terena networking conference*, 2007.

[Sim94] W. Simpson. The Point-to-Point Protocol (PPP). IETF RFC 1661, 1994.

[Sim05] J. M. Simmons. On determining the optimal optical reach for a long-haul network. *Journal of Lightwave Technology*, 23(3):1039–1048, 2005.

[SJ07] P. Srisuresh and P. Joseph. OSPF-xTE: Experimental extension to OSPF for traffic engineering. IETF RFC 4973, 2007.

[SKS03] S. Sengupta, V. Kumar, and D. Saha. Swithed optical backbone for cost-effective scalable core IP networks. *IEEE Communications Magazine*, 41(6):60–70, 2003.

[SLS06] P. Szegedi, Z. Lakatos, and J. Spath. Signaling architectures and recovery time scaling for grid applications in IST project MUPBED. *IEEE Communications Magazine*, 44(3):74–82, 2006.

[Smi04] H. Smit. Intermediate system to intermediate system (IS-IS) extensions for Traffic Engineering (TE). IETF RFC 3784, 2004.

[SRB01] S. Sarvotham, R. Riedi, and R. Baraniuk. Connection-level analysis and modeling of network traffic. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, pages 99–103, 2001.

[SRB03] D. Saha, B. Rajagopalan, and G. Bernstein. The optical network control plane: State of the standards and deployment. *IEEE Communications Magazine*, 41(8), 2003.

[SRS99] A. Shaikh, J. Rexford, and K.G. Shin. Load sensitive routing of long-lived IP flows. *ACM SIGCOMM Computer Communication Review*, 29(4):215–226, 1999.

[SW06] F. B. Shepherd and P. J. Winzer. Selective randomized load balancing and mesh networks with changing demands. *OSA Journal of Optical Networking*, 5(5):320–339, 2006.

[SYR05] L. Shen, X. Yang, and B. Ramamurthy. Shared risk link group (SRLG)-diverse path provisioning under hybrid service level agreements in wavelength-routed optical mesh networks. *IEEE/ACM Transactions on Networking*, 13(4):918–931, 2005.

[Sys07] Cisco Systems. Introduction to Cisco IOS NetFlow - a technical overview. Technical report, 2007.

[SYT$^+$02] K. Sato, N. Yamanaka, Y. Takigawa, M. Koga, S. Okamoto, K. Shiomoto, E. Oki, and W. Imajuku. GMPLS-based photonic multilayer router (Hikari router) architecture, an overview of traffic engineering and signaling technology. *IEEE Communications Magazine*, 40(3):96–101, 2002.

[Tho02] S.A. Thomas. *IP Switching and Routing Essentials.* John Wiley and Sons, Inc., 2002.

[TKT07] I. Tomkos, C. Kouloumentas, and S. Tsolakidis. Performance studies of multi-wavelength all-optical 2R regeneration subsystems based on highly non-linear fibers. In *Proceedings of International Conference on Transparent Optical Networks*, pages 132–135, 2007.

[Top06] The GÉANT Network Topology. http://www.geant.net/server/show/nav.159, 2006.

[TP05] Y. Taing and L. Pavel. Application of H control on pilot tones in erbium-doped fiber amplifiers. In *Proceedings of Optical Fiber Communication Conferece*, 2005.

[TUC$^+$06] N. Taesombut, F. Uyeda, A. A. Chien, L. Smarr, T. A. DeFanti, P. Ppadopoulos, J. Leigh, M. Ellisman, and J. Orcutt. The OptIPuter: High-performance, QoS-guaranteed network service for emerging e-science applications. *IEEE Communications Magazine*, 44(5):38–45, 2006.

[Uni05] Stanford University. Clean-slate design for the Internet. http://cleanslate.stanford.edu/index.php, 2005.

[VCPD04] S. Verbrugge, D. Colle, M. Pickavet, and P. Demeester. Capex comparison between link-by-link and end-to-end grooming in a european backbone network. In *Proceedings of International Symposium on Telecommunications Network Strategy and Planning*, pages 309–322, 2004.

[VS97] E.A. Varvarigos and V. Sharma. The ready-to-go virtual circuit protocol: A loss-free protocol for multigigabit networks using FIFO buffers. *IEEE/ACM Transactions on Networking*, 5(5):705–718, 1997.

[VXHB05] V.K. Gurbani V.K., S. Xian-He, and A. Brusilovsky. Inhibitors for ubiquitous deployment of services in the next-generation network. *IEEE Communications Magazine*, 43(9):116–121, 2005.

[WCM06] G. Weichenberg, V.W.S. Chan, and M. Mèdard. On the throughput-cost tradeoff of multi-tiered optical network architectures. In *Proceedings of IEEE Globecomm conference*, pages 1–6, 2006.

[wgta05] MAWI working group traffic archive. Working group traffic archive. http://tracer.csl.sony.co.jp/mawi/, 2005.

[Wid95] I. Widjaja. Performance analysis of burst admission control protocols. *IEEE Proceedings on Communications*, 142(1):7–14, 1995.

[XY04] F. Xue and S. J. B. Yoo. High capacity multiservice optical label switching for the next generation Internet. *IEEE Optical Communications*, 42(5):S16–S22, 2004.

[YHM05] H. Yurong, J.P. Heritage, and B. Mukherjee. Connection provisioning with transmission impairment consideration in optical wdm networks with high-speed channels. *Journal of Lightwave Technology*, 23(3):982–993, 2005.

[YKS$^+$02] N. Yamanaka, M. Katayama, K. Shiomoto, E. Oki, and N. Matsuura. Multilayer traffic engineering in photonic-GMPLS-router networks. In *Proceedings of Global Telecommunications Conference*, pages 2731–2735, 2002.

[YMBS$^+$06] M. Yannuzzi, X. Masip-Bruin, S. Sanchez, J. Domingo-Pascual, A. Orda, and A. Sprintson. On the challenges of establishing disjoint QoS IP/MPLS paths across multiple domains. *IEEE Communications Magazine*, 44(12):60–66, 2006.

[Yoo03] S. J. B Yoo. Optical label switching, MPLS, MPLambdaS and GMPLS. *SPIE Optical Networks Magazine*, 4(3):17–31, 2003.

[ZJM00] H. Zang, J. P. Jue, and B. Mukherjee. A review of routing wavelength assignment approaches for wavelength-routed optical WDM networks. *SPIE Optical Network Magazine*, 1(1), 2000.

[ZMD$^+$07] T. Zseby, M. Molina, N. Duffield, S. Niccolini, and F. Raspall. Sampling and filtering techniques for IP packet selection. IETF Draft 'draft-ietf-psamp-sample-tech-10', 2007.

[Øve07] H. Øverby. Traffic modelling of asynchronous bufferless optical packet switched networks. *Elsevier Computer Communications journal*, 30(6):1229–1243, 2007.