ADONIS project
Algorithms for Dynamic Optical Networks based on Internet Solutions

Technical Report related to 2nd year activities

Version: January 2005

Telecommunications Networks Group at Politecnico di Milano,
Telecommunications Networks Group at Politecnico di Torino,
Computer Networks and Mobility Group at Università di Trento

Contents

Chapter 1

Introduction

Objectives of the research project

ADONIS (Algorithms for Dynamic Optical Networks based on Internet Solutions) is a FIRB research project, sponsored by the Italian Ministry for University and Scientific Research, aiming at proposing novel methodologies for the design and management of optical networks in order to support the migration from the current static assignment and routing techniques to adaptive, dynamic techniques based on IP-centric control.

According to the original project proposal, the research activities are organized into two workpackages:

WP1 *Intelligent static network design*, whose main objective is to suggest a unified evolutionary design approach, which supports the current static network structure, but is aware of the dynamic future requirements. By a proper design of the static network, these techniques allow a seamless migration to dynamic management as soon as equipment and protocols become available.

WP2 *Evolutionary design of dynamic networks*, which aims at developing efficient algorithms and protocols for establishing lightpaths in wavelength-routed networks with dynamic traffic demands. The proposed protocols will consider traffic engineering capabilities and guarantee effective protection/restoration mechanisms. Furthermore, new switching architectures are proposed and evaluated to support the dynamic traffic patterns.

In the following section, an overview of the activities developed during the second year of the project is given. The activities pursued by each Research Unit participating in the project are described in greater detail in the internal chapters of the Technical Report.

Research activities developed during the second year

The activities developed during the second year of the project are divided in the following according to the specific workpackage to which they refer.

Activities in WP1 - Intelligent static network design

In the framework of the first workpackage the theme of network planning under strict constraints on connection availability has been studied [1]. The guarantees of a predetermined maximum outage time and a maximum connection-failure probability for a connection become fundamental components of the service level agreement between an operator and its customers. Being an Optical Transport Network (OTN) system composed of reparable subsystems, quality is usually measured in terms of connection availability, rather than reliability, though this second parameter could have its relevance. Network planning consists of allocating resources (in terms of WDM channels) to both working and protection lightpaths. Two optimization objectives are pursued: a) maximization of the availability level guaranteed to the users and b) minimization of the total number of fibers that must be installed to deploy the network (assuming that every fiber carries the same prefixed number of WDM channels). Different strategies to achieve these objectives are shown: initially, no availability target is specified and all the connections are individually routed so to maximize the availability of each one, with no constraints on physical network resources; then the total number of fibers is minimized, by assuming the availability values obtained by the first strategy as constraints; finally, physical-resource minimization is repeated in the same way as before, but relaxing the availability constraints by a prefixed ratio.

Another activity in the context of WP1 focused on the performance comparison of of guided-wave architectures for space-division photonic switching [2]. The core of an OOO cross-connect is an optical switch that is independent of data rate and protocol. Various technologies have been proposed and studied, which can be subdivided into two large categories: free-space and guided-wave systems. In this contribution several architectures have been evaluated and compared considering a possible physical implementation based on guided-wave structures realized with integrated optics technology. Some properties including number of switching elements required, blocking performance, number of waveguide crossovers, system attenuation, and signal-to-noise ratio are evaluated and analyzed. The main purpose of this study is to review the state-of-the-art of optical guided-wave space-switching architectures and to provide a relevant set of technical elements useful in the selection of architectures to be used in all-optical cross-connect implementation.

Activities in WP2 - Evolutionary design of dynamic networks

In this workpackage, dynamically reconfigurable wavelength routed networks are considered, in which lightpaths carrying IP traffic are on demand established.

A first contribution in this framework refers to the Routing and Wavelength Assignment problem considering as constraints the physical impairments that arise in all-optical wavelength routed networks [8]. In particular, the impact of the physical layer

when establishing a lightpath in transparent optical network is studied. Because no signal transformation and regeneration at intermediate nodes occurs, noise and signal distortions due to non-ideal transmission devices are accumulated along the physical path, and they degrade the quality of the received signal. A simple yet accurate model for the physical layer is proposed which considers both static and dynamic impairments, i.e., nonlinear effects depending on the actual wavelength/lightpath allocation. A novel algorithm to solve the RWA problem that explicitly considers the physical impairments is proposed and its effectiveness evaluated by simulation. Indeed, when the transmission impairments come into play, an accurate selection of paths and wavelengths which is driven by physical consideration is mandatory.

Part of the activities related to this workpackage are related to Traffic Grooming (TG). TG is the multiplexing capability aimed at optimizing the capacity utilization in transport systems by means of the combination of low-speed traffic streams onto high-speed (optical) channels. Dynamic grooming is basically a routing problem in a multi-layer network architecture (e.g. an IP over Optical network), since the objective is to find the "best" path to route traffic requests arriving dynamically to grooming nodes.

RFC 3717 defines three interconnection models for IPO networks: overlay, augmented and peer. The peer and the augmented models are appealing because sharing the knowledge base between the two layers allows running an integrated routing function, by using, for instance, an auxiliary graph. The integrated management enables a better usage of the network resources. However, both models seem not feasible in the near term due to the tight integration between the two levels and scalability issues regarding the amount of exchanged information. The overlay model is instead technically feasible, since it only requires the definition of an interface between the IP and optical level and dynamic lightpath capabilities in the optical level.

Only a few dynamic grooming algorithms based on the overlay model have been proposed so far, and the interaction between different routing strategies adopted in the IP and Optical levels was never assessed. However, none of the works on grooming in IPO networks adopted a realistic traffic model. The traffic loading the IPO network is always modelled like a traditional circuit switched traffic, i.e. CBR (Constant Bit Rate) connections characterized by the bit rate and duration. Any realistic evaluation of algorithms to be deployed within the Internet, should instead capture at least the basic characteristics of Internet traffic.

In [11] a formal definition of dynamic grooming based on graph theory is defined in a general interconnection model and specialized to the case of the overlay model. A family of grooming policies is proposed in the overlay architecture and the existing dynamic grooming proposals are assessed as a special case of it. The performance analysis of these policies by considering realistic Internet (elastic) traffic is evaluated in [9]. In particular, a simple analytical model highlighting the interaction effect between the IP routing and the optical layer is analyzed to assess the impact of traffic elasticity on dynamic grooming. Finally, performance and tradeoffs of different policies are discussed and explained in both regular and irregular topologies, also discussing the impact of adopting TE techniques in the IP or optical level and unfairness issues inherent to overlay dynamic grooming. A separate contribution [10] describes in detail the simulation tool (GANCLES) used to perform these comparative studies on dynamic

grooming algorithms.

Other contributions in this framework concern the long-term view of a full packet switching network performing IP packet transport, in which optical operations are performed as much as possible exploiting the currently available optical device technology. Apparently most of the operations related to the packet header processing need to be done in the electronic domain.

A node architecture for optical packet-switched transport networks based on an AWG device is analyzed in [4]. Packet buffering is made possible by fiber delay lines accomplishing either input queueing only, or combined input/shared queueing. In this contribution it is shown that, unless the input buffer length exceeds the maximum packet size, optical scheduling and optical FIFO buffering give almost the same performance. On the other hand, when the input queue can hold at least one packet of maximum size, optical scheduling yields a better performance than optical FIFO buffering, because the output links can be more efficiently exploited.

In [5] different node architectures for the switching of IP packet flows are considered that are based on current optical routing devices. The traffic performance of a mesh network is evaluated with the various node structures, assuming that nodes employ either shortest path routing or deflection routing to forward packets to the addressed destinations. This contribution shows how the different node structures behave in terms of packet loss probability with different network configurations when the node parameters are varied.

Joint activities in both WP1 and WP2

The activities described in this section are related to both workpackages WP1 and WP2.

Among them, one is related to Metropolitan network architectures based on single-hop WDM optical ring networks operating in packet mode. They are one of the most promising architectures for MAN since they permit a cost-effective design, with a good combination of optical and electronic technologies, while supporting features like restoration and reconfiguration that are essential in any Metro scenario. In [7] the tunability requirements that lead to an effective resource usage and permit reconfiguration in optical WDM Metros are addressed. Some reconfiguration algorithms are introduced that adapt the network configuration to traffic demands to optimize performance on the basis of traffic measurements. Using a specific network architecture as a reference case, this contribution aims at the broader goal of showing which are the advantages fostered by innovative network designs exploiting the features of optical technologies.

In the transition towards full dynamic traffic, WDM networks optimized for a specific set of static connections will most likely also be used to support on-demand lightpath provisioning. The paper [3] investigates the issue of routing of dynamic connections in WDM networks which are also loaded with high-priority protected static connections. By discrete-event simulation various routing strategies are compared in terms of blocking probability and a new heuristic algorithm is propose based on an occupancy cost function which takes several possible causes of blocking into account. The behavior of this algorithm is tested in well-known mesh networks, with and without wavelength conversion and under different types of network traffic.

Another activity is related to optical switch architectures which are able to handle variable-length packets such as IP datagrams, based on an AWG device to route packets and equipped with a fiber delay-line stage as optical input buffer. Unfortunately, extensive simulations of optical networks built with switches of this type showed that considerable buffering capability is required in order to achieve acceptable performance. The paper [6] builds upon these previous two, by studying the effectiveness of packet deflection as a mean for solving packet contentions on outputs of optical switches. Optical transport networks are simulated, by evaluating the performance of packet deflection routing, based on a traffic model adherent to real IP traffic measurements. In this contribution full-mesh and wheel network topologies are considered, comparing results to assess deflection effectiveness. The simulation results show that deflection routing leads to satisfying performance even using buffers with limited size. Furthermore, the average delivery delay does not suffer heavy penalty from packet deflection, even under heavy traffic conditions.

Bibliography

[1] M. Tornatore, G. Maier, and A. Pattavina, "Availability design of optical networks," *IEEE Journal on Selected Areas on Communication - Optical Communication Networks Series*, to appear.

[2] L. Savastano, G. Maier, A. Pattavina, and M. Martinelli, "Performance comparison of guided-wave architectures for space-division photonic switching," in *Proc. of Broadnets 2004*, Oct. 2004, pp. 202–211.

[3] G. Maier, A. Pattavina, L. Barbato, F. Cecini, and M. Martinelli, "Routing algorithms in wdm networks under mixed static and dynamic lambda traffic," *"Photonic network communications"*, vol. 8, no. 1, pp. 69–87, July 2004.

[4] A. Pattavina, "Architecture and performance of optical packet switching nodes for ip networks," *IEEE Journal of Lightwave Technology*, to appear.

[5] A. Pattavina, and M. Aste, "Performance of optical packet switching nodes in ip transport networks," in *Proc. of Broadnets 2004*, Oct. 2004, pp. 84–91.

[6] S. Bregni and A. Pattavina, "Performance evaluation of deflection routing in optical ip packet-switched networks," *Cluster Computing*, vol. 7, pp. 239–244, July 2004.

[7] A.Bianco, J.M.Finochietto, G.Giarratana, F.Neri, C.Piglione, "Measurement Based Reconfiguration in Optical Ring Metro Networks", *submitted to JLT*, december 2004.

[8] R.Cardillo, V.Curri, M.Mellia, "Considering Transmission Impairments in Wavelength Routed Networks", *Optical Network Design and Models - ONDM*, Milan, February 7-9, 2005.

[9] R. Lo Cigno, E. Salvadori, Z. Zsóka, "Elastic Traffic Effects on WDM Dynamic Grooming Algorithms," *In Proc. of GLOBECOM 2004*, Dallas, TX, USA, Dec. 2004.

[10] E. Salvadori, Z. Zsóka, and R. Lo Cigno. Dynamic Grooming in IP over WDM Networks: A Study with Realistic Traffic based on GANCLES Simulation Package. In *Proceedings of the 9th IFIP/IEEE ONDM*, Milan, Italy, February 2005.

[11] E. Salvadori, Z. Zsóka, R. Lo Cigno, and R. Battiti. A Framework for Dynamic Grooming in IPO Overlay Architectures. Submitted for publication to Networking, 2005.

Chapter 2

Politecnico di Milano Research Unit

A. Pattavina, M. Tornatore, S. Bregni
Dipartimento di Elettronica – Politecnico di Milano
Via Ponzio, 34/5- 20133 Milano, Italy
{pattavin,tornator,bregni}@elet.polimi.it

Abstract

The tasks of the unit of Politecnico di Milano in this project are focused mainly on the following topics:

- development of methodologies for designing optical networks under static traffic demand and with reliability requirements;
- development of methodologies for the evolution of optical networks able to provide transmission capacity under dynamic traffic demand;
- development of methodologies for designing highly dynamic optical networks and study of switching architectures for IP traffic.

According to the previous objectives our works have been carried on in order to obtain significant results in these three areas.

In recent years the importance of WDM networks has rapidly grown, leading them to become the core of the telecommunication infrastructure, able to face the increase of data and video traffic. If WDM technology can offer a solution to the large bandwidth demand, WDM protocols (management and control systems) have been developed in order to guarantee that the WDM layer will act in the future as a common transport platform able to operate in an integrated multi-protocol environment, providing a high quality of service.

The theme of network planning under strict constraints on connection availability has been studied in [1]. The guarantees of a prefixed maximum outage time and a maximum connection-failure probability [2] for a connection become fundamental components of the service level agreement between an operator and its customers. Being an OTN system composed of reparable subsystems, quality is usually measured in terms of connection availability (a measure of outage time), rather than reliability (a measure of disconnection probability), though this second parameter could have its relevance [3]. Many authors have proposed various ad-hoc reliability parameters to be employed to networking problems [2, 4, 5, 6].

In this work the planning of an OTN starting from a known set of optical connection requests and from a green-field physical installation is considered. The sets of nodes and of WDM links connecting them are given, together with the physical lengths of the links. Each connection is *end-to-end protected*, i.e. it is implemented by setting up a pair of lightpaths from the source to the destination node, one of which is used in working conditions and the other is for protection. Network planning consists in allocating resources (in terms of WDM channels) to both working and protection lightpaths. Two optimization objectives are pursued: a) maximization of the availability level guaranteed to the users and b) minimization of the total number of fibers that must be installed to deploy the network (assuming that every fiber carries the same prefixed number of WDM channels). We show different strategies to achieve these objectives: initially, no availability target is specified and all the connections are individually routed so to maximize the availability of each one, with no constraints on physical network resources; then the total number of fibers is minimized, by assuming the availability values obtained by the first strategy as constraints; finally, physical-resource minimization is repeated in the same way as before, but relaxing the availability constraints by a prefixed ratio.

The second topic that has been analyzed focuses on the performance comparison of of guided-wave architectures for space-division photonic switching [7]. The core of an OOO cross-connect is an optical switch that is independent of data rate and protocol. Various technologies, e.g. microelectromechanical systems (MEMS) [8] , electrooptical [9] , thermooptical [10] , liquid-crystal [11] , bubble-jet [12] and acoustooptical [13] , have been proposed and studied for realization of optical switches. All these technologies can be subdivided into two large categories: free-space and guided-wave systems.

Our analysis has two main guidelines: first, the identification of some evaluation criteria based on specific performance parameters to compare different guided-wave switching architectures; second, the presentation of an overview of the main guided-wave architectures currently state-of-the-art in technical literature. Up to now, several guided-wave architectures have been proposed for optical space switching. We have collected them in this unified framework and we have calculated the parameters we have considered significative to allow a technical comparison among different possible architectural options for OOO implementation. This will set the basis for a switching-technology aware project of the optical networks under static offered traffic assumption, after the availability-aware design method developed in the first activity.

On the other hand, the past few years (up to 2001) witnessed the flourishing of telecommunication industry as one of the fastest-growing and most wide-spread phe-

nomena in economy ever recorded. When the general crisis hit the telecommunication market it found WDM networks at the dawn of a new evolution. Optical transport networks in the past have mainly been designed and operated as static systems: optical connections were used as long-distance trunks mostly to carry large aggregates of telephone traffic, usually serving only customers of the network operator itself. Traffic was thus highly predictable.

The scenario is much different today. Data traffic is going to overcome traditional telephone traffic in volume. The former is characterized by less regular flows than the latter, which are more and more independent of geographical distances. This change in traffic statistics is further amplified in the regional and metro area, where flows are less aggregated and more sensitive to traffic relations due to single, large bandwidth applications. Finally, many WDM network operators are beginning to offer the "lambda service" (i.e. optical connections for lease) and the carrier-of-carriers service to support the so-called "bandwidth-trading" business. This implies that their network infrastructure is no longer used solely from their own final customers, making a quota of connection demands no longer deterministically predictable. The recent development of the Generalized Multi-Protocol Label Switching (GMPLS) technique seems to convey the idea that lightpaths in the future will have to be set up and torn down on a very short time-scale, even few seconds, perhaps paving the way to a possible optical packet-switching (or optical burst-switching) era.

All the facts mentioned above are pushing research in security, management and network design to re-focus its attention from the simple static Optical Transport Network (OTN) to Automatic Switched Optical Network (ASON). While OTN is already well-defined by the main standard bodies [14, 15], the new ASON model, able to set-up and release lightpaths on-demand based on on-line requests, is still undergoing an intense research and standardization activity. This justifies the birth of dynamic traffic as a new subject of research in optical networks.

In our research we have addresses both static and dynamic traffic paradigm in a new very particular context. Given the evolution from OTN to ASON as an actual process, this will surely occur gradually, in any case always preserving the investments of the network operators. In the transition the two paradigms of static and dynamic traffic have to co-exist and to be supported by the same WDM network infrastructure. So this part of our work falls in the first two areas of our activity.

In our investigation [16], we propose and discuss a heuristic strategy for routing lightpaths for dynamic traffic that allows to increase the acceptance rate of dynamic connection requests (or, equivalently, to decrease the blocking probability) compared to other previously known routing algorithms. Such a new algorithm is based on a global network function which provides an estimation of the available network resources according to different criteria. This allows to assign resources to the new lightpath so that chances of having congestion in the most critical spots of the network are kept as low as possible. We are going also to study the performance of the heuristic algorithm proposed under different types of dynamic traffic.

The fourth and fifth point described in this report concern the long-term view of a full packet switching network performing IP packet transport, in which optical operations are performed as much as possible exploiting the currently available optical device technology. Apparently most of the operations related to the packet header pro-

cessing needs to be done in the electronic domain. The last two areas that have been analyzed concern our third area of activity. Our work on this topic is divided into two main parts: both of them deal with the architecture of an optical packet switching node first proposed in [17], which is equipped with a fiber delay line stage used as an input buffer for optical packets.

Two papers, namely [18, 19], focus optical packet switching in a full-IP transport network scenario. For the switching of IP packet flows different node architectures are considered that are based on current optical routing devices. The traffic performance of a mesh network is evaluated with the various node structures, assuming that nodes employ either shortest path routing or deflection routing to forward packets to the addressed destinations.

In previous papers [20, 21], an optical switch architecture was proposed to handle variable-length packets such as IP datagrams, based on an AWG device to route packets and equipped with a fiber delay-line stage as optical input buffer. Unfortunately, extensive simulations of optical networks built with switches of this type showed that considerable buffering capability would be required in order to achieve acceptable performance. In [22], therefore, we studied the effectiveness of packet deflection as a mean for solving packet contentions on outputs of optical switches. Optical transport networks were simulated, evaluating the performance of packet deflection routing, based on a traffic model adherent to real IP traffic measurements. Full-mesh and wheel network topologies have been considered, comparing results to assess deflection effectiveness. Our simulation results show that deflection routing leads to satisfying performance even using buffers with limited size. Furthermore, the average delivery delay does not suffer heavy penalty from packet deflection, even under heavy traffic conditions.

In order to carry out the simulative analysis in our three activity fields, two separate software tools have been developed.

The first one is a C++ optical network simulator, which runs under Linux operative system. This software allows (1) to emulate the dynamic behavior of a network loaded by an on-demand offered traffic, under the assumptions that the capacity distribution in the link of the network is given and (2) to design a network loaded by a static traffic matrix. This simulator is able to start either from a green-field or considering a fixed initial distribution of optical circuits in the network. The static version of this software tool has been developed in order to include an availability-based and availability-constrained design of the network. In the dynamic version, confidence level can be set to guarantee the accuracy of the results.

A second software allows to simulate a optical packet switching node, based on a structure tat is hypothesized in our studies. Two main properties, its modularity and the choice of an object-based programming language like C++, allow for a flexible adaptation of our software to several distinct network scenarios. This simulator tool is able to modelize both the single node performances and the behavior of a structure composed by many nodes, interconnected by several distinct network pattern. Furthermore, nodes with different internal structures has been set up by this tool, according to various patterns of connections between AWG's and wavelength converters. The offered traffic has been also shaped with respect to various typologies, as well as different structural parameters (e.g., the maximum delay in the ingress buffer, port number, wavelength

number per channel and the structure of the optical packet) can be specified by means of a text file.

In the next references section, we report all the publications that are related to this project. Items in **boldface** are publications generated at Politecnico di Milano in the second year of the Adonis project.

Bibliography

[1] **M. Tornatore, G. Maier, and A. Pattavina, "Availability design of optical networks,"** ***IEEE Journal on Selected Areas on Communication - Optical Communication Networks Series,*** **to appear.**

[2] A. Fumagalli, M. Tacca, F. Unghvary, and A. Farago, "Shared Path Protection with Differentiated Reliability," in *IEEE International Conference on Communications*, vol. 4, April 2002, pp. 2157–2161.

[3] E. E. Lewis, *Introduction to Reliability Engineering*, J. W. . Sons, Ed. John Wiley & Sons, 1987.

[4] A. Kiss, G. Vesztergombi, J. Levendovszky, and L. Jereb, "Adaptative Statistical Algorithms in Network Reliability Analysis," in *International Conference on Telecomunications Systems*, 2001.

[5] L. Jereb, T. Jakab, and F. Unghvary, "Availability Analysis of Multi-Layer Optical Networks," *Optical Networks Magazine*, March-April 2002.

[6] M. Clouqueur and W. D. Grover, "Availability Analysis of Span-Restorable Mesh Networks," *IEEE Journal on Selected Areas In Communications*, vol. 20, pp. 810–821, May 2002.

[7] **L. Savastano, G. Maier, A. Pattavina, and M. Martinelli, "Performance comparison of guided-wave architectures for space-division photonic switching," in** ***Proc. of Broadnets 2004*****, Oct. 2004, pp. 202–211.**

[8] L.-Y. Lin, E. L. Goldstein, and R. W. Tkach, "Free-Space Micromachined Optical Switches for Optical Networking," *Journal of Selected Topics in Quantum Electronics*, vol. 5, no. 1, pp. 4–9, January-February 1999.

[9] A. Dugan, L. Lightworks, and J. C. Chiao, "The optical switching spectrum: A primer on wavelength switching technologies," *Telecommun. Magazine*, May 2001.

[10] D. K. Cheng, Y. Liu, and G. J. Sonek, "Optical switch based on thermally-activated dye-doped biomolecular thin films," *IEEE Photon. Technol. Lett.*, vol. 7, pp. 366–369, April 1995.

[11] S. Hardy, “Liquid-crystal technology vies for switching applications,” *Lightwave*, pp. 44–46, December 1999.

[12] A. Ware, “New photonic-switching technology for all-optical networks,” *Lightwave*, pp. 92–98, March 2000.

[13] D. A. Smith, A. Alessandro, J. E. Baran, D. J. Fritz, J. L. Jackel, and R. S. Chakravarthy, “Multiwavelength performance of an adopized acoust-optic switch,” *Journal of Lightwave Technology*, vol. 14, pp. 2044–2051, September 1996.

[14] ”ITU-T Intern. Telecom. Union - Telecom. Standard. Sector”, *Architecture of Optical Transport Networks*, 1999, no. G.872.

[15] ——, *Network Node Interface for the Optical Transport Network (OTN)*, 2001, no. G.709.

[16] **G. Maier, A. Pattavina, L. Barbato, F. Cecini, and M. Martinelli, “Routing algorithms in wdm networks under mixed static and dynamic lambda traffic,” *”Photonic network communications”*, vol. 8, no. 1, pp. 69–87, July 2004.**

[17] S. Bregni, G. Guerra, and A. Pattavina, “Optical Packet Switching of IP Traffic,” in *Proceedings of* 6^{th} *Working Conference on Optical Network Design and Modeling (ONDM)*, 2002.

[18] **A. Pattavina, “Architecture and performance of optical packet switching nodes for ip networks,” *IEEE Journal of Lightwave Technology*, to appear.**

[19] **A. Pattavina, and M. Aste, “Performance of optical packet switching nodes in ip transport networks,” in *Proc. of Broadnets 2004*, Oct. 2004, pp. 84–91.**

[20] S. Bregni, G. Guerra, and A. Pattavina, “Optical Switching of IP Traffic Using Input Buffered Architectures,” *Optical Network Magazine*, vol. 3, no. 6, pp. 20–29, 2002.

[21] S. Bregni, A. Pattavina, and G. Vegetti, “Architectures and Performances of AWG-based Optical Switching Nodes for IP Networks,” *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 7, pp. 1113–1121, 2003.

[22] **S. Bregni and A. Pattavina, “Performance evaluation of deflection routing in optical ip packet-switched networks,” *Cluster Computing*, vol. 7, pp. 239–244, July 2004.**

Architectures and Performance of Optical Packet Switching Nodes for IP Networks

Achille Pattavina
Dept. of Electronics and Information, Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milan, Italy
pattavina@elet.polimi.it

***Abstract*— As new bandwidth-hungry IP services are demanding more and more capacity, transport networks are evolving to provide a reconfigurable optical layer in order to allow fast dynamic allocation of WDM channels. To achieve this goal, optical packet-switched systems seem to be strong candidates as they allow a high degree of statistical resource sharing, which leads to an efficient bandwidth utilization. In this work, we propose an architecture for optical packet-switched transport networks, together with an innovative switching node structure based on the concept of per-packet wavelength routing. The traffic performance of such node when loaded by a typical IP traffic is evaluated through computer simulation; packet loss probability and average delay performance are shown for various load conditions.**

***Index Terms*— WDM network, optical switching, arrayed waveguide grating (AWG), IP packets, traffic performance.**

I. Introduction

Telecommunication networks are currently experiencing a dramatic increase in demand for capacity, driven by new bandwidth-hungry IP services. This will lead to an explosion of the number of wavelengths per fiber, that can't be easily handled with conventional electronic switches. To face this challenge, networks are evolving to provide a reconfigurable optical layer, which can help to relieve potential capacity bottlenecks of electronic-switched networks, and to efficiently manage the huge bandwidth made available by the deployment of dense wavelength division multiplexing (DWDM) systems.

As current applications of WDM focus on a relatively static usage of single wavelength channels, many works have been carried out in order to study how to achieve switching of signals directly in the optical domain, in a way that allows fast dynamic allocation of WDM channels, so as to improve transport network performance.

Two main alternative strategies have been proposed to reach this purpose: optical packet switching [1]–[5], and optical burst switching [6]–[8]. In this article, we describe an 'almost' all-optical switching architecture that uses an arrayed waveguide grating (AWG) as the packet router device. As usual, packet buffering is accomplished by fiber delay line units. Preliminary papers have already reported some traffic performance results for this switch, when the node supports packet buffering only at switch inputs (input queueing) [9] or by sharing delay lines among all switch inputs through packet recirculation (shared queueing) [10]. Here we present the overall switch structure in a deeper detail and evaluate the switch performance also with combined input/shared queueing.

[0]Work partially supported by MIUR, Italy, under FIRB project ADONIS.

In Section 2 we first introduce the basic concepts supporting the two types of switching, that is optical packet switching and optical burst switching. Then in Section 3 we introduce the optical network environment envisioned for the long-haul transport scenario, for which Section 4 describes the architecture of the AWG-based switching node. Traffic performance is evaluated in Section 5 for a single node supporting either pure input queueing or combined input/shared queuing.

II. Optical Packet and Burst Switching

Optical packet switching makes it possible to exploit single wavelength channels as shared resources, with the use of statistical multiplexing of traffic flows, helping to efficiently manage the huge bandwidth of WDM systems. Two different basic approaches have been proposed to this aim, which differ in the switching matrix unit: *broadcast-and-select* architectures or *wavelength routing* architectures.

The first project based on broadcast-and-select principle is KEOPS [11]; Figure 1 shows its architecture that adopts input buffering. It is composed of two stages performing optical buffering and switching. In the first stage packets are delayed by a suitable amount of time in order to avoid collisions at the switch output ports; this function is accomplished by a set of tunable wavelength converters (TWC) whose task is to select the proper delay line to be accessed through the demultiplexers. Optical packets emerging from the multiplexers are given a new wavelength by the second set of TWCs so as to select the addressed switch outlet.

Since this solution does not allow packet recirculation, it cannot efficiently support different packet priorities, because, once a packet has been sent to a delay line, it cannot be stored longer than the fiber delay to eventually transmit a new packet with higher priority. This is a crucial shortcoming of this solution, since the need for some methods of providing differentiated classes of service for Internet traffic is growing, with the explosion of new possible applications. Actually the IPv4 TOS field or the IPv6 Traffic Class field are already used to give packets a particular forwarding treatment at each network node, and the availability in the network nodes of such feature is a fundamental requirement.

Recently another project has been proposed, further elaborating the broadcast-and-select solution: the DAVID [12]

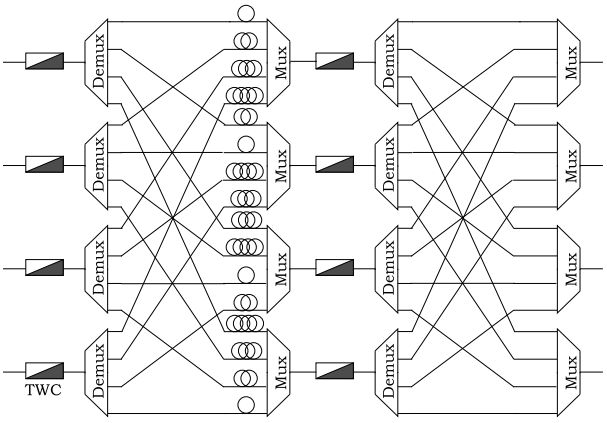


Fig. 1. KEOPS node architecture.

project proposes an optical networking solution viable for both metropolitan and wide-area networks.

Using a wavelength-routing device for optical packet switching has been proposed in the projects OPERA [13] and WASPNET switch [14]. In the former case only one arrayed waveguide grating (AWG) device is used for optical packet routing, whereas two AWGs are adopted in the latter project, which presents a more complete view from a networking standpoint. Figure 2 shows the WASPNET architecture that includes two switching stages. The first stage is used to route packets to the delay lines buffer, for contention resolution, or to the second stage. This latter stage is used to properly route packets to the desired output port. In both stages an AWG device is used to switch (route) packets to an outlet which is jointly identified by the AWG incoming port and the adopted transmission wavelength. TWCs at the AWG inputs perform this function, while the other TWCs feeding the demultiplexers are used to select the amount of recirculation delay. This architecture allowing packet recirculation accomplishes shared queueing, but the need for a second AWG to route packets to their addressed output link yields a considerable hardware overhead.

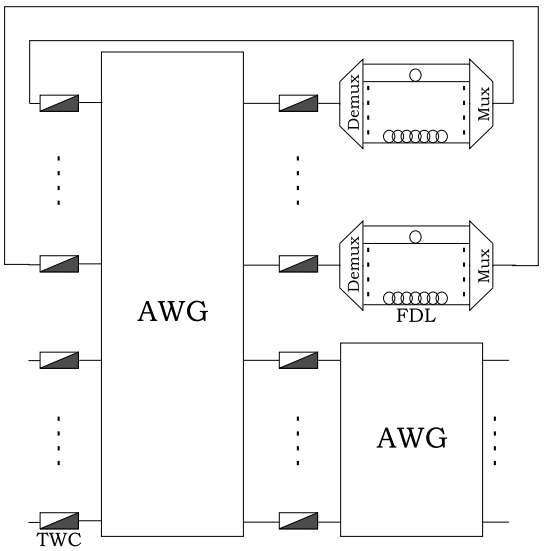


Fig. 2. WASPNET node architecture.

The systems presented insofar carry out header processing and routing functions electronically, while the switching of optical packet payloads takes place directly in the optical domain. This eliminates the need for many optical-electrical-optical conversions, which call for the deployment of expensive opto-electronic components, even though most of the optical components, needed to achieve optical packet switching, still remain too crude for commercial availability.

Optical burst switching aims at overcoming these technological limitations. The basic units of data transmitted are bursts, made up of multiple packets, which are sent after control packets, carrying routing information, whose task is to reserve the necessary resources on the intermediate nodes of the transport network (see Figure 3). This results in a lower average processing and synchronization overhead than optical packet switching, since packet-by-packet operation is not required. However packet switching has a higher degree of statistical resource sharing, which leads to a more efficient bandwidth utilization in a bursty IP-like traffic environment.

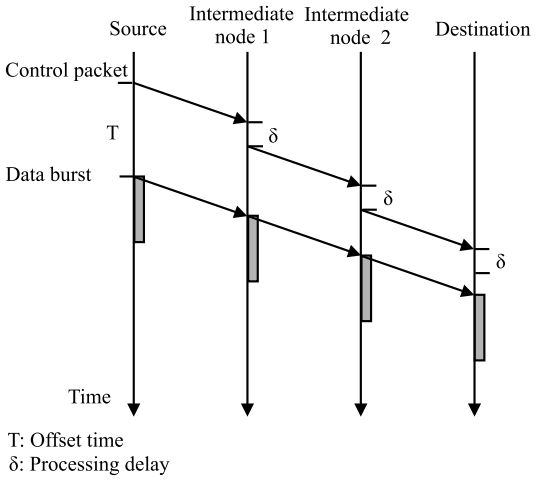


Fig. 3. The use of an offset time in optical burst switching.

Since optical packet-switching systems still face some technological hurdles, the existing transport networks will probably evolve through the intermediate step of burst-switching systems, which represent a balance between circuit and packet switching, making the latter alternative a longer term strategy for network evolution. In this work, we have focused our attention on optical packet switching, since it offers greater flexibility than the other relatively coarse-grained WDM techniques, aiming at efficient system bandwidth management.

All the mentioned solutions accomplishing optical packet switching (KEOPS, WASPNET, DAVID, OPERA) define an environment suitable to switch fixed-length packets, whose transmission time is called *slot*, whose duration is in the order of 1 μs. This implies the development of complex segmentation and reassembly protocols at the optical network edges, if the offered traffic is composed of variable-length information units whose transmission time exceeds the slot time. On the other hand if the slot time is selected in such a way to fit the largest information unit, it is very likely that most of the slots will be partially used when small-size information

units are sent, thus wasting network resources.

The solution we propose here is to an optical packet switching node capable of switching variable-length optical packets; so the client layer (we assume IP in our scenario) can be interfaced more easily with the optical layer, thus avoiding a heavy packet processing overhead at the optical transport network edges. A *slot* concept is introduced also in our case, but here refers to the minimum size of optical packet that can be switched in a TCP/IP network environment.

III. Optical Transport Network Architecture

The architecture of the optical transport network we propose consists of M *optical packet switching nodes*, each denoted by an optical address made of $m = log_2\lceil M \rceil$ bits, which are linked together in a mesh-like topology. Edge systems (ES) interface the optical transport network with IP legacy (electronic) networks (see Figure 4).

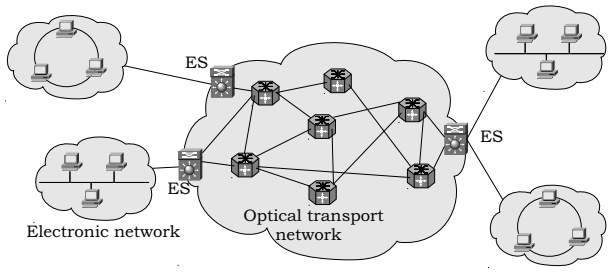


Fig. 4. The optical transport network architecture.

The optical packet is composed of a simple optical header, which comprises the m-bit destination address, and an optical payload containing an IP packet. In principle multiple IP packets could be packed in the same optical packet payload, if they are all addressed to the same ES. The optical packets are buffered and routed through the optical transport network to reach their destination ES, which delivers the traffic it receives to its destination electronic networks. At each intermediate node in the transport network the optical packet headers are received and electronically processed, in order to provide routing information to the control electronics, which will properly configure the node resources to switch packet payloads directly in the optical domain.

The transport network operation is *asynchronous*; that is, packets can be received by nodes at any instant, with no time alignment. The internal operation of the optical nodes, on the other hand, is *synchronous* (that is *slotted*), meaning that the optical packet switching must start at the beginning of a time slot. In the model we propose the time slot duration, T, to be equal to the amount of time needed to transmit an optical packet with a 40-byte payload from an input WDM channel to an output WDM channel. Such payload has been chosen as it is the minimum-size packet that can be transmitted in an IP-based network; actually it is an IP datagram transporting a TCP acknowledgement. Supposing a bit rate of 10 Gbps per wavelength channel, a 40 ns slot duration seems appropriate, since the 40-byte payload is transmitted in 32 ns (payload time, T_p) and the additional time can be used for the optical packet header transmission and to provide guard times.

Optically transporting variable-length packets in such a slotted environment is made possible by allowing an optical packet to engage several consecutive slots. We assume that the optical packet header needed for packet routing is present only in the first slot of the optical packet and that the payload per slot is always the same, that is $T_p = 32$ ns. For example an IP packet of 1500 bytes is transported by an optical packet engaging 38 slots. Therefore the bandwidth usage is kept under control by selecting carefully T and T_p.

Slotted operation has been assumed by all projects mentioned in the previous section, with slot duration equal to the time needed to transmit fixed-size packets. Only the project DAVID foresees the possibility of switching variable-length packets with a slot aggregation similar to that assumed here. Assuming slotted operation for variable-size packets compared to unslotted switching, makes simpler the switch control, for example handling the transition between different switch permutations. Other project proposals of optical packet switching with unslotted operations are not available in the technical literature, as far as the author's knowledge is concerned. In this paper we do not address the issue of comparing complexity or traffic performance between slotted and unslotted switching.

In our switch model a contention occurs every time two or more packets are trying to leave a switch from the same output port. How contentions are resolved has a great influence on network performance. Three main schemes are generally used to resolve contention: wavelength conversion, optical buffering and deflection routing.

In a switch node applying *wavelength conversion*, two packets trying to leave the switch from the same output port are both transmitted at the same time but on different wavelengths. Thus, if necessary, one of them is wavelength converted to avoid collision. In the *optical buffering* approach, one or more contending packets are sent to fixed-length fiber delay lines, in order to reach the desired output port only after a fixed amount of time, when no contention will occur. Finally, in the *deflection routing* approach, contention is resolved by routing only one of the contending packets along the desired link, while the other ones are forwarded on paths which may result in paths longer than the minimum-distances.

Implementing optical buffering gives good network performance, but involves a great amount of hardware and electronic control. On the other hand, deflection routing is easier to implement than optical buffering, but network performance is reduced since a portion of network capacity is taken up by deflected packets.

In the all-optical network proposed, in order to reduce complexity while aiming at attaining good network performance, the problem of contention is resolved combining a small amount of optical buffering with wavelength conversion and, eventually, deflection routing. Our policy can be summarized as follows:

1) When a contention occurs, the system first tries to transmit the conflicting packets on different wavelengths.
2) If all of the wavelengths of the correct output link are busy at the time the contention occurs, some packets are scheduled for transmission in a second time, and are forwarded to the fiber delay lines.

3) Finally, if no suitable delay line is available at the time the contention occurs for transmission on the correct output port, a conflicting packet is lost or, if a suitable deflection algorithm is implemented, it can be deflected to a different output port than the correct one.

IV. Node Architecture

The general architecture of a network node is shown in Figure 5. It consists of N incoming fibers with W wavelengths per fiber. The incoming fiber signals are demultiplexed and G wavelengths from each input fiber are then fed into one of the W/G switching planes, which constitute the switching fabric core. Once signals have been switched in one of the second-stage parallel planes, packets can reach every output port on one of the G wavelengths that are directed to each output fiber. This allows the use of wavelength conversion for contention resolution, since G packets can be transmitted at the same time by each second-stage plane on the same output link. Apparently hardware simplicity requirements suggest to feed each plane with the same wavelengths from any input fiber. Nevertheless in principle there is no restriction in selecting the value of G, even if it will be shown that it has a significant impact on the switch traffic performance.

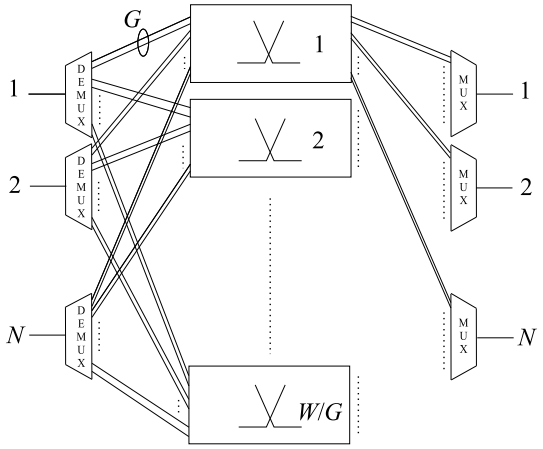


Fig. 5. Optical packet-switching node architecture.

The structure of one of the W/G parallel switching planes is presented in Figure 6. It interfaces single-wavelength input and output links and consists of three main blocks: an input *synchronization unit*, as the node is slotted and incoming packets need to be aligned, a *fiber delay lines unit*, used to store packets for contention resolution, and a *switching matrix unit*, to achieve the switching of signals.

These three blocks are all managed by an *electronic control unit* which carries out the following tasks:

- optical packet header recovery and processing;
- managing the synchronization unit in order to properly set the correct path through the synchronizer for each incoming packet;
- managing the tunable wavelength converters (TWCs) in order to properly delay and route incoming packets in the second and third unit of the system, respectively.

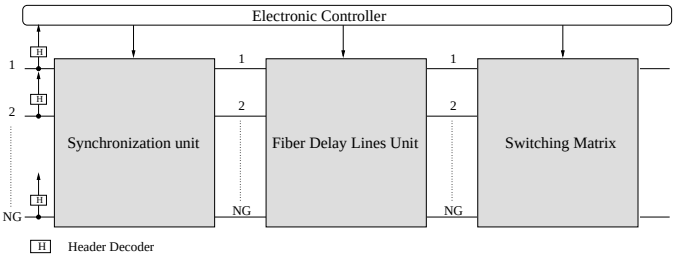


Fig. 6. Structure of one of the W/G parallel switching planes.

When packet recirculation is allowed, the AWG is used to switch packets to the output ports or, if necessary, to a recirculation port, in order to store them for an additional amount of time, to avoid collisions. Moreover recirculation ports allow the switch to support different priority classes, with service preemption. In fact, an optical packet, traveling through a recirculation port delay line, can always be preempted by a higher priority packet and be redirected to a recirculation port, instead of being transmitted.

We will now describe the second-stage switching units mentioned above, detailing their implementation.

A. Synchronization Unit

The synchronization unit is shown in Figure 7 and consists of a series of 2×2 optical switches interconnected by fiber delay lines of different lengths. These are arranged in a way that, depending on the particular path set through the switches, the packet can be delayed by a variable amount of time, ranging between $\Delta t_{min} = 0$ and $\Delta t_{max} = 2(1 - (1/2)^{n+1}) \times T$, with a resolution of $T/2^n$, where T is the time slot duration and n the number of delay line stages.

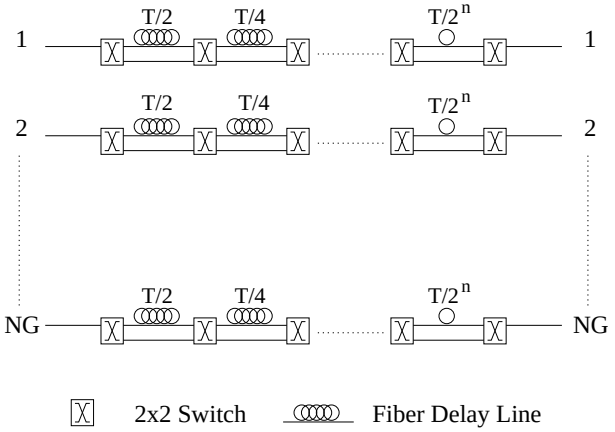


Fig. 7. Structure of the synchronization unit

The synchronization is achieved as follows: once the packet header has been recognized and packet delineation has been carried out, the packet start time is identified and the control electronics can calculate the necessary delay and configure the correct path of the packet through the synchronizer.

Due to the fast reconfiguration speed needed, fast 2×2 switching devices, such as 2×2 semiconductor optical amplifier (SOA) switches [15], which have a switching time in

the nanosecond range, must be used. SOAs are all-optical amplification devices that are already used in a wide range of applications; they can be arranged in a particular structure (as shown in Figure 8), in order to achieve switching of optical signals. In this configuration SOAs are used as gates that let the signals go through or stop, depending on the permutation required. An interesting characteristic of SOA switches is that these devices allow the amplification of the traveling signals making it possible, besides routing functionalities, to restore a required given signal level.

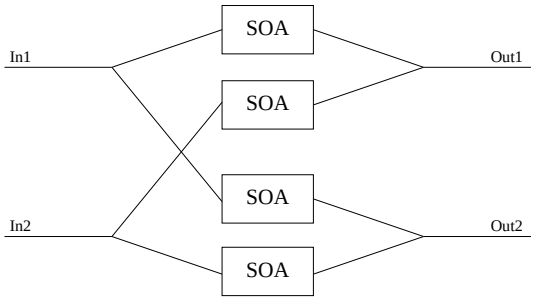


Fig. 8. 2×2 SOA switch.

B. Fiber Delay Lines Unit

After packet alignment has been carried out, the routing information carried by the packet header allows the control electronics to properly configure a set of tunable wavelength converters (TWCs), in order to deliver each packet to the correct delay line to resolve contentions (see Figure 9). On each of the NG inputs of the plane a delay can be applied that is multiple of the basic slot duration T and ranges up to D_{max} slots. An optical packet can be stored for a time slot, with a 40 ns duration, in about 8 meters of fiber at 10 Gbps. To achieve wavelength conversion several devices are available [16]–[19].

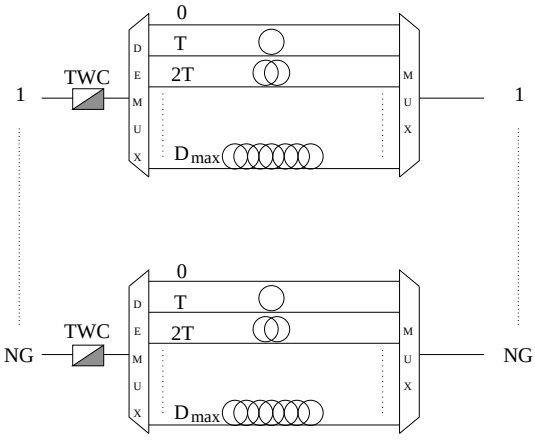


Fig. 9. Structure of the fiber delay unit

Depending on the managing algorithm used by control electronics, the fiber delay lines stage can be used as an*optical scheduler* or as an *optical first-in-first-out* (FIFO) *buffer*.

- *Optical scheduling*: this policy uses the delay lines in order to schedule the transmission of the maximum number of packets onto the correct output link. This implies that an optical packet P_1, entering the node at time αT from the i-th WDM input channel, can be transmitted after an optical packet P_2, entering the node on the same input channel at time βT, being $\beta > \alpha$. For example, suppose that packet P_1, of duration $l_1 T$, must be delayed d_1 time slots, in order to be transmitted onto the correct output port. This packet will then leave the optical scheduler at time $(\alpha + d_1)T$. So, if packet P_2, of duration $l_2 T$, has to be delayed for d_2 slots, it can be transmitted before P_1 if $\beta + d_2 + l_2 < \alpha + d_1$ since no collision will occur at the scheduler output.
- *Optical FIFO buffering*: in the optical FIFO buffer the order of the packets entering the fiber delay lines stage must be maintained. This leads to a simpler managing algorithm than the one used for the optical scheduling policy, yielding, however, a sub-optimal output channel utilization. In fact, suppose that optical packet P_1, entering the FIFO buffer at time αT, must be delayed for d_1 time slots. This implies that packet P_2, behind packet P_1, must be delayed of at least d_1 time slots, in order to maintain the order of incoming packets. Due to this rule, if packet P_2 could be delayed for $d_2 < d_1$ slots to avoid conflict, its destination output port will be idle for $d_1 - d_2$ time slots, while there would be a packet to transmit.

C. Switching Matrix Unit

Once packets have crossed the fiber delay lines unit, they enter the switching matrix stage in order to be routed to the desired output port. This is achieved using a set of tunable wavelength converters combined with an arrayed waveguide grating (AWG) wavelength router [20], as is shown in Figure 10a.

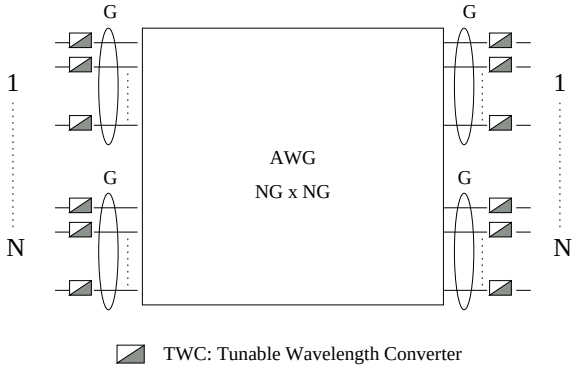
This device consists of two slab star couplers, interconnected by an array of waveguides. Each grating waveguide has a precise path difference with respect to its neighbors, ΔX, and is characterized by a refractive index of value n_w.

Once a signal enters the AWG from an incoming fiber, the input star coupler divides the power among all waveguides in the grating array. As a consequence of the difference of the guides lengths, light traveling through each couple of adjacent waveguides emerges with a phase delay difference given by:

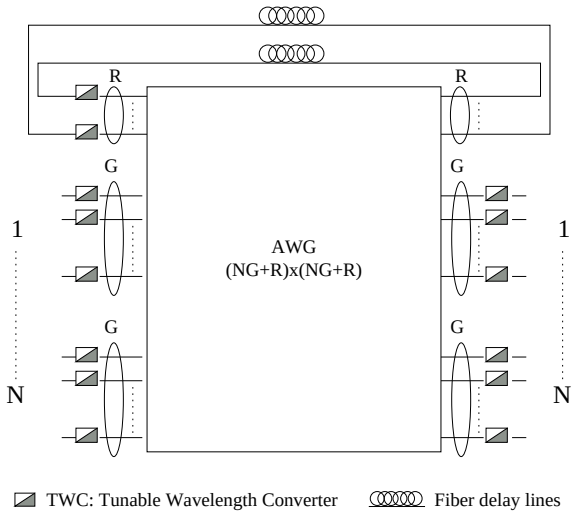
$$\Delta\phi = 2\pi n_w \times \frac{\Delta X}{\lambda}$$

where λ is the incoming signal central wavelength. As all the beams emerge from the grating array they interfere constructively onto the focal point in the output star coupler, in a way that allows to couple an interference maximum with a particular output fiber, depending only on the input signal central wavelength.

Figure 11 shows the mechanism described above. Two signals of wavelength λ_0 and λ_3 entering an 8×8 AWG, from input fibers number 6 and number 1 respectively, are correctly switched onto the output fibers number 0 and number 3, the



(a)



(b)

Fig. 10. Structure of the switching unit: input queueing (a); combined input-shared queueing (b)

wavelength and the input port of the signals being the only parameters determining the switch permutation.

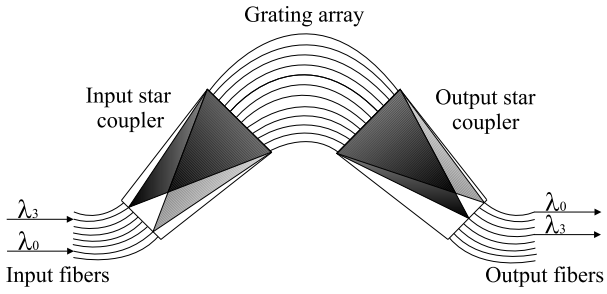


Fig. 11. Arrayed waveguide grating.

The AWG is used as it gives better performance than a normal space switch interconnection network, as far as insertion losses are concerned. This is due to the high insertion losses of all the high-speed all-optical switching fabrics available at the moment, that could be used to build a space switch interconnection network. Moreover AWG routers are strictly non-blocking and offer high wavelength selectivity. Commercially available 40 channel devices have a channel spacing of 100 GHz and show a typical insertion loss of 7.5 dB.

As we said before, to improve the system performance and to eventually support different priority classes with service preemption, some of the AWG ports can be reserved to allow packet recirculation (see Figure 10(b)). To this purpose R AWG output ports are connected, via fiber delay lines, to R input ports. Packet recirculation is then managed using tunable wavelength converters.

After crossing the three stages previously described, packets undergo a final wavelength conversion, to avoid collisions at the output multiplexers, where W WDM channels are multiplexed on each output link.

V. Simulation Results

We present now some simulation results of the operation of an optical node with a single switching plane and hence $W = G$ is always assumed. These results have been obtained assuming that the node receives its input traffic directly from N edge systems. The edge system buffer capacity is supposed to be large enough to make packet loss negligible and each WDM channel is supposed to have a dedicated buffer in the edge system.

The packet arrival process has been modelled as a Poisson process, with packet interarrival times having a negative exponential distribution. As the node operation is slotted, the optical packets duration is always assumed to be multiple of the time slot duration T, which is equal to the amount of time needed to transmit an optical packet, with a 40-byte payload, from an input WDM channel to an output WDM channel.

As far as packet length is concerned, the following probability distributions were considered:

1) *Empirical distribution*. Based on real measurements on IP traffic [21], [22], we have assumed the following probability distribution for the packet length, L:

$$\begin{cases} p_0 = Pr\{L = 40 \text{ bytes}\} = 0.6 \\ p_1 = Pr\{L = 576 \text{ bytes}\} = 0.25 \\ p_2 = Pr\{L = 1500 \text{ bytes}\} = 0.15 \end{cases}$$

In this model, packets have average length equal to 393 bytes. Since a 40-byte packet is transmitted in one time slot of duration T, the average duration of an optical packet is approximatively $10T$. Moreover, p_0, p_1 and p_2 represent the probability that the packet duration is T, $15T$ and $38T$ respectively.

2) *Uniform distribution*. To show a comparison with the empirical model described above, we have modeled the optical packet length as a stochastic variable, uniformly distributed between 40 bytes (duration T) and 760 bytes (duration $19T$). Also in this model, packets have average duration of $10T$.

No deflection routing algorithm has been implemented. Under this assumption, a packet is supposed to be lost if it cannot be delayed by a suitable amount of time, in order to transmit it onto the correct output port, on any of the G available wavelengths. We will now present the performance results of both architectures, with and without packet recirculation ports in the AWG, remarking that all the plotted values have a 95% confidence interval not larger than 40% of the plotted values.

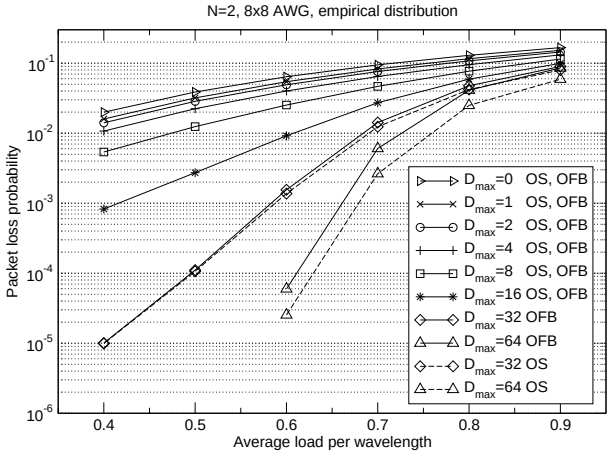


Fig. 12. Packet loss probability for the empirical distribution: OS vs. OFB.

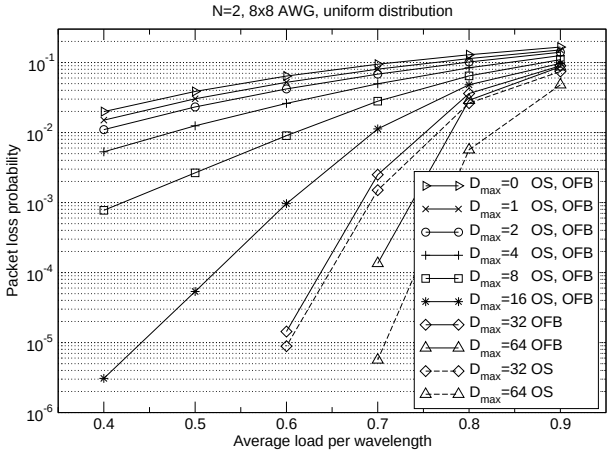


Fig. 13. Packet loss probability for the uniform distribution: OS vs. OFB.

A. Optical node without packet recirculation

A node without recirculation line is considered first in which we evaluate the effect on traffic performance of the two managing policies optical scheduling (OS) and the optical FIFO buffering (OFB). Figures 12 and 13 show the packet loss probability at different traffic loads per wavelength for various values of the input line delay D_{max} of an 8×8 AWG, with $N = 2$ and $W = G = 4$ wavelengths per fiber, for the empirical and uniform distributions. It can be seen that, regardless of packet length distribution, the OS policy yields a better performance than the OFB policy only when the maximum delay, D_{max}, becomes large enough to allow efficient packet scheduling. It must be remarked that an optical packet P_2, of duration l_2T entering the node at time βT from the i-th WDM input channel, can be transmitted before a packet P_1, entering the node on the same input channel at time $\alpha T < \beta T$, only if D_{max} is large enough to avoid collision at the fiber delay lines output, that is $D_{max} \geq (\beta + l_2 + d_2 - \alpha)T$. Further performance improvements are expected by equipping longer delay lines on the input side.

Figure 14 shows the packet loss probability for the empirical and uniform distributions, which have an L_{max} value of $38T$ and $19T$ respectively, for the OFB policy. It can be pointed out

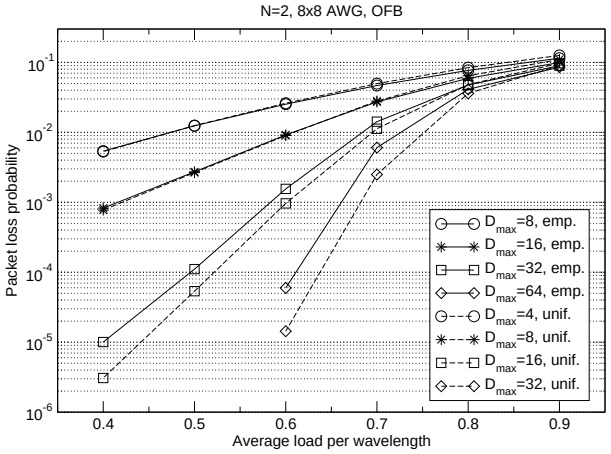


Fig. 14. Packet loss probability for the empirical and uniform distributions.

that, regardless of packet length distribution, the node almost shows the same loss probability for the same value of the D_{max}/L_{max} ratio. Furthermore, performance improves as this ratio increases. In a variable packet length environment, then, it is convenient to use the OFB policy for the fiber delay lines unit management, as it is simpler to implement than the OS policy and gives almost the same performance.

We evaluate how the parameter G, which represents the number of channels per input/output fibers handled by a single plane, affects the overall packet loss performance of the node under optical scheduling operation. To this aim we have selected a node architecture with a single switching plane ($W = G$). We have compared four switch configurations with the same external lightpath number ($N \cdot W$), and no recirculation lines ($R = 0$). By assuming the availability of a 32×32 AWG and an FDL stage with maximum delay $D_{max} = 8T$, the switch size varies in the set $N = \{2, 4, 8, 16\}$ and the channel group size in the set $G = \{16, 8, 4, 2\}$, in such a way that $N \cdot G = 32$. Figure 15 shows that for a given offered load the packet loss performance improves as G increases. In particular for low levels of the offered traffic the improvement can be of several orders of magnitude. This improvement is nothing else than that attained in any multiple-server system, in which all users fully share the set of servers. Traffic engineers well know this phenomenon under various names, among which perhaps the most common is "channel grouping" (or "trunk grouping").

B. Optical node with packet recirculation

Here we present the simulation results for an optical switching plane with R recirculation ports with OS policy. Two different structures for the recirculation delay lines have been tested: the constant delay recirculation (CDR) and the variable delay recirculation (VDR). In the CDR structure all the recirculation ports delay each packet by the same amount of time, which, unless stated otherwise, is given by $D_{rec} = T$. In the VDR structure D_{rec} doubles every two ports, that is the first couple of ports will then have a recirculation delay of T, the second couple of $2T$, and so on. Such structure of delay lines for the VDR case has been selected to enable

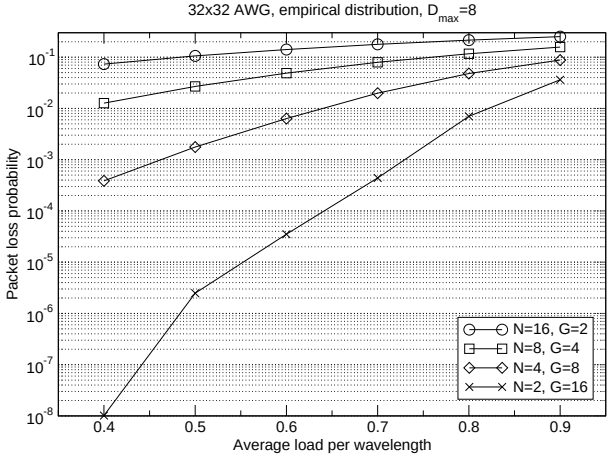


Fig. 15. Packet loss performance for different grouping factor and node size values.

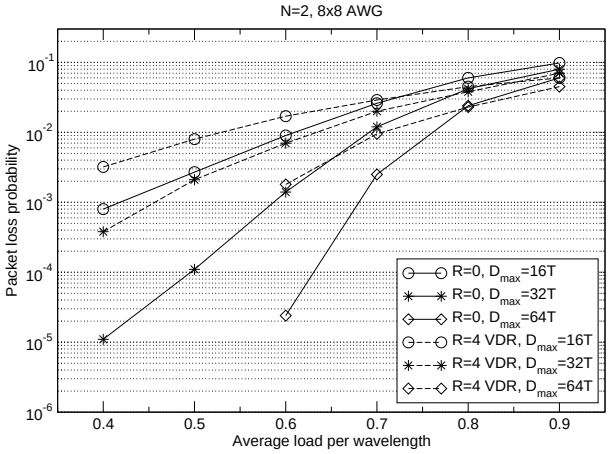


Fig. 17. Packet loss probability: 8×8 AWG, $R = 0$ vs. 8×8 AWG, $R = 4$ VDR.

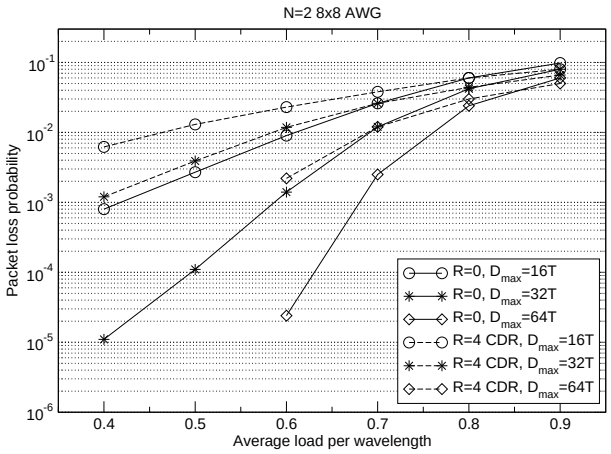


Fig. 16. Packet loss probability: 8×8 AWG, $R = 0$ vs. 8×8 AWG, $R = 4$, CDR.

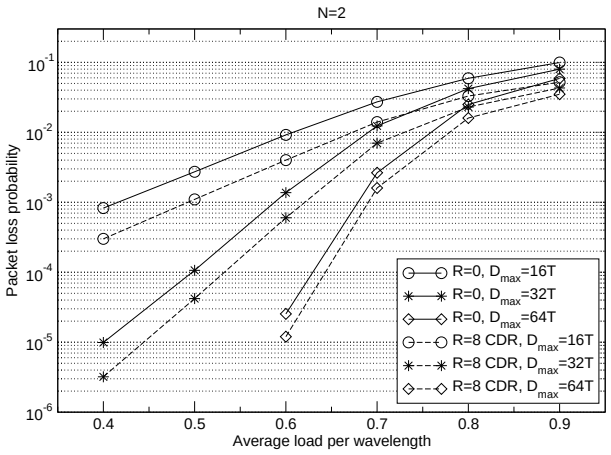


Fig. 18. Packet loss probability: 8×8 AWG, $R = 0$ vs. 16×16 AWG, $R = 8$, CDR.

the switch performance evaluation with recirculation lines of reasonable length for the AWG sizes considered here, e.g. up to $D_{rec} = 16T$. Moreover multiple recirculations are allowed only if the packet duration L is lower than the recirculation delay to prevent long packets to occupy simultaneously more than one recirculation port.

Figures 16 through 19 show the packet loss probability of an optical switching node with recirculation ports ($R > 0$) and compare them with a structure without recirculation ports ($R = 0$), with different values of D_{max} for the empirical distribution.

Figures 16 and 17 plot the loss probability of an 8×8 AWG, with $R = 0$ and $R = 4$, for constant delay recirculation (CDR) and variable delay recirculation (VDR) structures, respectively. As the AWG dimension does not change, the system with the recirculation ports always gives higher loss probabilities than the other one since, in order to make packet recirculation possible, the grouping factor G has to be reduced, reducing the number of packets that can be transmitted at the same time on one output link.

Figures 18 and 19 show the comparison between an 8×8 AWG, with $R = 0$, and a 16×16 AWG, with $R = 8$. Now, as the grouping factor, G, does not change, the packet recirculation effect on the system performance is apparent: larger buffers in the recirculation lines (adopted in the VDR case) improve the packet loss performance. As the introduction of the recirculation ports allows a longer packet storage, the average delay grows (see Figure 20), yielding a lower loss probability for both the CDR and VDR structures.

We have then compared the two configurations, with and without recirculation lines, both equipped with an 8×8 AWG. The number of input/output fibers is kept constant ($N = 2$), while the grouping factor varies in the set $G = \{2, 4\}$.

Figure 21 shows the packet loss probability and average delay for a node with and without fiber recirculation lines, for different values of the offered load; the FDL stage maximum delay is $D_{max} = 16T$ and the recirculation lines configuration is the CDR configuration. It can be pointed out that the reduction of the grouping factor G, from $G = 4$ ($R = 0$) to $G = 2$ ($R = 4$ and $D_{rec} = \{T, 4T, 16T\}$), yields higher loss probability and average delay. This performance worsens more as the average traffic load decreases, since the effect of the grouping factor variation is more evident for low levels of the offered load, as we pointed out before.

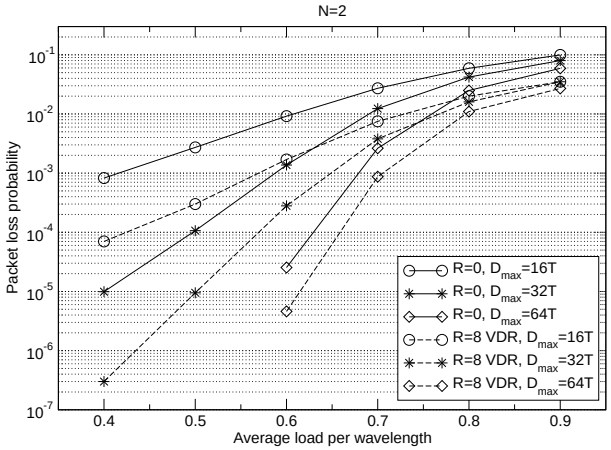


Fig. 19. Packet loss probability: 8×8 AWG, $R = 0$ vs. 16×16 AWG, $R = 8$, VDR.

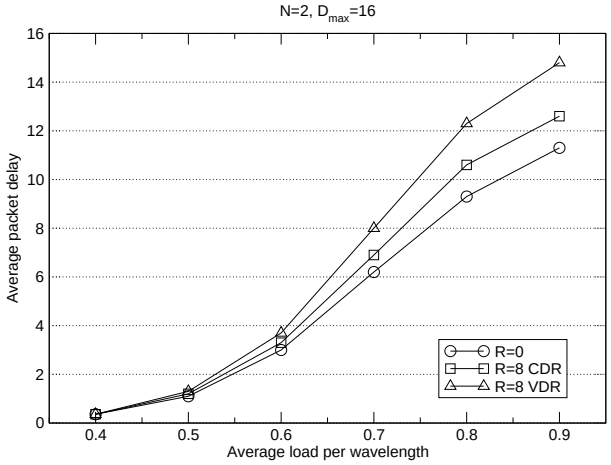


Fig. 20. Average packet delay at different loads per wavelength for an 8×8 AWG, $R = 0$ and for a 16×16 AWG, $R = 8$, with CDR and VDR structures.

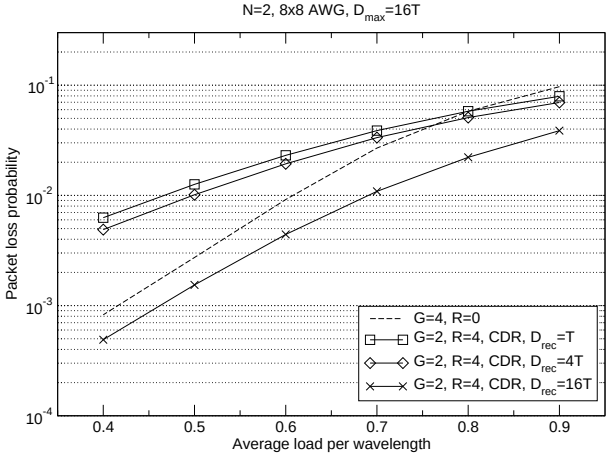


Fig. 21. Packet loss performance for different grouping factor values, with and without recirculation lines ($R = 0$ vs. $R = 4$ CDR).

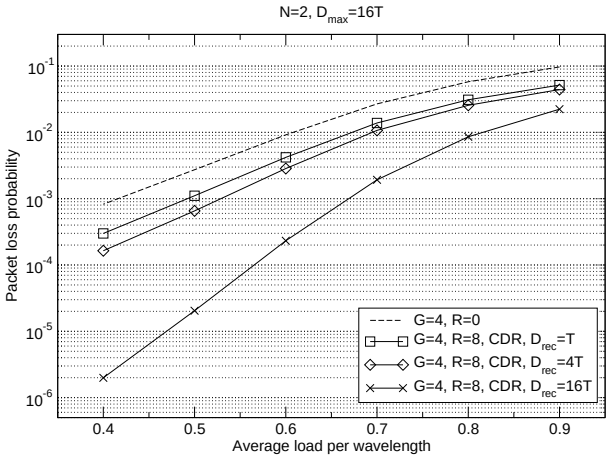


Fig. 22. Packet loss performance for the same grouping factor value, with and without recirculation lines ($R = 0$ vs. $R = 8$ CDR).

Finally, we have compared the performance of two architectures, with and without recirculation fibers, with the same number of input/output fibers $N = 2$, the same value of the grouping factor $G = 4$, varying the AWG dimension. An 8×8 AWG, without recirculation lines and a 16×16 AWG, with $R = 8$ recirculation lines have been selected to this aim. Figure 22 shows that, as the grouping factor does not change, the nodes with recirculation lines always give a better performance, since they have a higher buffering capability than the nodes without recirculation lines, while the same number of contentions can be resolved in the wavelength domain.

VI. Conclusions and Topics for Further Research

A node architecture for optical packet-switched transport networks has been proposed that is based on an AWG device. Packet buffering is made possible by fiber delay lines accomplishing either input queueing only, or combined input/shared queueing. It has been shown that, unless the input buffer length exceeds the maximum packet size, optical scheduling and optical FIFO buffering give almost the same performance. On the other hand, when the input queue can hold at least one packet of maximum size, optical scheduling yields a better performance than optical FIFO buffering, because the output links can be more efficiently exploited. Adding shared buffering through recirculation lines can improve the system performance only if the grouping factor G is not reduced.

Many issues will have to be addressed in the future, such as the improvement attainable with the introduction of different priority classes adopting service preemption. Moreover, the behavior of a meshed optical transport network will have to be investigated, in which deflection routing policies can be adopted to enhance the overall system performance.

References

[1] S. Yao, B. Mukherjee, and S. Dixit. Advances in Photonic Packet Switching: An Overview. *IEEE Communications Magazine*, 38(2):84–94, Feb 2000.

[2] D. K. Hunter and I. Andonovic. Approaches to Optical Internet Packet Switching. *IEEE Communications Magazine*, 28(9):116–122, Sep. 2000.

[3] L. Xu, H. G. Perros, and G. Rouskas. Techniques for Optical Packet Switching and Optical Burst Switching. *IEEE Communications Magazine*, 39(1):136–142, Jan. 2001.

[4] M. J. O'Mahony, D. Simeonidou, D. K. Hunter, and A. Tzanakaki. The Application of Optical Packet Switching in Future Communication Network. *IEEE Communications Magazine*, 39(3):128–135, Mar. 2001.

[5] M. A. Bourouha, M. Bataineh, and M. Guizani. Advances in Optical Switching and Networking: Past, Present, and Future. In *Proceedings of IEEE SoutheastCon 2002*, pages 405–413, 2002.

[6] C. QIao and M.Yoo. Optical Burst Switching OBS - A New Paradigm for an Optical Internet. *Journal of High Speed Networks*, 8:69–84, 1999.

[7] C. Qiao. Labeled Optical Burst Switching for IP-over-WDM Integration. *IEE Commun. Mag.*, pages 104–114, Sep. 2000.

[8] M. Yoo and C. Qiao. Optical Burst Switching for Service Differentiation in the Next Generation Optical Internet. *IEEE Communication Magazine*, 32(2):98–104, Feb. 2001.

[9] S. Bregni, G. Guerra, and A. Pattavina. Optical Switching of IP Traffic Using Input Buffered Architectures. *Optical Network Magazine*, 3(6):20–29, Nov.-Dec. 2002.

[10] S. Bregni, G. Guerra, and A. Pattavina. Optical Packet Switching of IP Traffic. In *Proceedings of* 6^{th} *Working Conference on Optical Network Design and Modeling (ONDM)*, 2002.

[11] C. Guillemot et al. Transparent Optical Packet Switching: the European ACTS KEOPS Project Approach. *IEEE Journal of Lightwave Technology*, 16(12):2117–2134, Dec. 1998.

[12] L. Dittman et al. The European IST Project DAVID: A Viable Approach Toward Optical Packet Switching. Architecture and Performance of AWG-based Optical Switching Nodes for IP Networks. *J. on Selected Areas in Commun.*, 21(7): 1026–1040, Sept. 2003.

[13] A. Carena, M.D. Vaughn, R. Gaudino, M. Shell, D.J. Blumenthal. An Optical Packet Experimental Routing Architecture with Label Swapping Capability. *IEEE Journal of Lightwave Technology*, 16(12):2135–2145, Dec. 1998.

[14] D. K. Hunter, K. M. Guild, and J. D. Bainbridge. WASPNET: a Wavelength Switched Packet Network. *IEEE Communication Magazine*, 37(3):120–129, Mar. 1999.

[15] F. Dorgeuille, B. Mersali, M. Feuillade, S. Sainson, S. Slempkès, and M. Fouche. Novel Approach for Simple Fabrication of High-Performance InP-Switch Matrix Based on Laser-Amplifier Gates. *IEEE Photoncs Technology Letters*, 8:1178–1180, 1996.

[16] M.W.K. Mak and H.K. Tsang. Polarization-insensitive Widely Tunable Wavelength Converter Using a Single Semiconductor Optical Amplifier. *IEE Electronics Letters*, 36:152–153, 2000.

[17] A. Tzanakaki and M.J. O'Mahony. Analysis of Tunable Wavelength Converters Based on Cross-Gain Modulation in Semiconductor Optical Amplifiers Operating in the Counter Propagating Mode. In *IEE Proceedings-Optoelectronics*, volume 147, pages 49–55, Feb. 2000.

[18] I. White, R. Penty, M. Webster, Y. J. Chai, A. Wonfor, and S. Shahkooh. Wavelength Switching Components for Future Photonic Networks. *IEEE Communications Magazine*, 40(9):74–81, Sep. 2002.

[19] O. A. Lavrova, L. Rau, and D. J. Blumenthal. 10-Gb/s Agile Wavelength Conversion With Nanosecond Tunig Times Using a Multisection Widely Tunable Laser. *IEEE Journal of Lightwave Technology*, 20(4):712–717, Apr. 2002.

[20] C. Parker and S.D. Walker. Design of Arrayed-Waveguide Gratings Using Hybrid Fourier-Fresnel Transform Techniques. *IEE Journal on Selected Topics in Quantum Electronics*, 5:1379–1384, 1999.

[21] K. Thompson, G. J. Miller, and R. Wilder. Wide-Area Internet Traffic Patterns and Characteristics. *IEEE Network Magazine*, pages 10–23, Nov. 1997.

[22] Generating the Internet Traffic Mix Using a Multi-Modal Length Generator. Spirent Communications white paper at http://www.netcomsystems.com.

Performance Comparison of Guided-Wave Architectures for Space-Division Photonic Switching

Luigi Savastano, Guido Maier, Mario Martinelli
CoreCom
Via Colombo 81
20133 Milan, Italy
{*t_luisav, maier, martinelli*}*@corecom.it*

Achille Pattavina
Politecnico di Milano,
P.za Leonardo da Vinci, 32
20133 Milan, Italy
pattavina@elet.polimi.it

Abstract

The paper[1] *presents and compares various unicast non-blocking architectures to be used into space-domain photonic switching networks. All the analyzed architectures have been evaluated and compared considering a possible physical implementation based on guided-wave structures realized with integrated optics technology. Some properties including number of switching elements required, blocking performance, number of waveguide crossovers, system attenuation, and signal-to-noise ratio are evaluated and analyzed. The main purpose of this work is to review the state-of-the-art of optical guided-wave space-switching architectures and to provide a relevant set of technical elements useful in the selection of architectures to be used in all-optical cross-connect implementation.*

1 Introduction

Recently, the growth of network traffic has stimulated the deployment of long-haul optical network systems which employ wavelength-division multiplexing (WDM) to achieve enormous transport capacity. Such systems, having tens of wavelengths per fiber with each wavelength modulated at 2.5 Gb/s, 10 Gb/s or more [25], rely mainly upon electronics to implement the switching functions. In every switching node, optical signals are converted to electrical form (O/E conversion), switched electronically and converted back to optical form (E/O conversion). Switching systems that execute these operations are called OEO cross-connects. Although electronic switching is highly reliable, it has many disadvantages as the dependence of switching hardware upon data bit-rate and transmission protocol, and high costs due to E/O and O/E conversion devices. The transition of the switching functions from electronics to optics with the deployment of all-optical (OOO) cross-connects will potentially reduce the network-equipment complexity and increase the flexibility, provided that the cost of OOO switches will be competitive with the cost of their OEO counterparts. The main cost advantages of the OOO solution can be envisioned in the absence of E/O (O/E) converters and in their transparency to the signal format. As a matter of facts, optoelectronic conversion represents an important cost component in today networks [19]. Moreover, several subsystems and components of OEO switches are subject to be substituted at any protocol or bit-rate variation.

The core of an OOO cross-connect is an optical switch that is independent of data rate and protocol. Various technologies, e.g. microelectromechanical systems (MEMS) [15], electrooptical [5], thermooptical [3], liquid-crystal [6], bubble-jet [28] and acoustooptical [21], have been proposed and studied for realization of optical switches. All these technologies can be subdivided into two large categories: free-space and guided-wave systems. For example, MEMS systems belong to the free-space category while electrooptical and thermooptical switches are guided-wave systems.

This paper has two main purposes: first, the identification of some evaluation criteria based on specific performance parameters to compare different guided-wave switching architectures; second, the presentation of an overview of the main guided-wave architectures currently state-of-the-art in technical literature. The first aspect is analyzed in section 2, while architectures are described in section 3. In this last section, after a brief description of the technology, the selection of switching architectures is characterized by providing for each case formulas to evaluate various parameters, focusing the aspects that are more peculiar for the given technology. Up to now, several guided-

[1] Work partially supported by MIUR, Italy, under FIRB project ADONIS.

wave architectures have been proposed for optical space switching. We have collected them in this unified framework and we have calculated the parameters we have considered significative to allow a technical comparison among different possible architectural options for OOO implementation.

Although some described architectures have multicasting or broadcasting capability, in this paper we have focused our attention only on space-domain switching networks with point-to-point connection capacity.

2 Performance parameters

Optical switching architectures can be compared using as benchmark four different classes of characteristics: blocking properties, physical structure, signal-transfer impairments and cost. Let us introduce for each class the parameters we will use in Sec. 3, briefly discussing them one-by-one in this section to form a qualitative point of view.

2.1 Blocking properties

According to their switching capability, switching networks can be subdivided into two categories: blocking and non-blocking [20]. A network is said to be non-blocking if an unused input port can always be connected to any unused output port. Thus, a non-blocking network is capable to realize every permutation of input ports on output ports. If at least one of these permutations can never be realized, the network is said to be blocking. Most applications require non-blocking architectures.

Non-blocking networks can be further distinguished in three subclasses according to their dynamic behavior in the transitions from a switching state to another. In Rearrangeable Non-Blocking (RNB) architectures the setup of a new connection between an unused input and an unused output may require the reconfiguration of the entire switching network, with rearrangement of other already active connections. Since existing connections must be interrupted, though only for the switching time, rearrangeability is often considered unacceptable by operators. Strict-sense Non-Blocking (SNB) and Wide-sense Non-Blocking (WNB) architectures can always setup a new connection between an unused i/o pair, regardless of the current switching state, thus avoiding rearrangements. WNB can achieve this feature only provided that any new connection is routed according to a specific setup rule. In SNB switching networks, any free route can be assigned to a new connection, indifferently. There is generally a trade-off between switching capability and complexity: the avoidance of connection-rearrangement disruptions is paid in terms of a higher number of switches in the WNB and an even higher number in the SNB networks. Despite its scarce popularity, the RNB solution is nevertheless interesting, as it can remain the sole possibility to achieve large switch-dimensions in those cases in which technology sets hard bounds to scalability.

We shall point out that this study has on purpose overlooked the wavelength switching domain. Since our investigation is dedicated to space-switching optical fabrics, we make no special assumption on the colors of the optical signals crossing the switches. They can be either all at the same wavelength or they can be WDM connections composed by several multiplexed channels. The general basic condition valid for all the cases we are going to analyze is that the switching operation is always wavelength insensitive, i.e. routing never depends on wavelength. Moreover, multiplexed input signals remain so at the output of the switch and no wavelength conversion operation is carried out. Under this hypotheses, the blocking classification reported above always refers to the sole space-switching and space-multiplexing domain.

2.2 Physical structure

A switching architecture is usually composed of a pattern of several optically-interconnected basic Switching Elements (SEs) (as directional couplers), arranged according to a specific topology. The switching fabric can be realized entirely on a unique substrate (single-substrate implementation) or by distributing the SEs among many modules (multiple-substrate implementation).

Advantages of the single-substrate implementation are network compactness and construction simplification. On the other hand, separation on many substrate often allows essential physical-performance improvements. For example, for guided-wave systems the multiple-substrate implementation allows a great reduction in the number of waveguide crossovers (see later), resulting in lower insertion loss and crosstalk. An important drawback of the multiple-substrate structure is the need for external systems interconnecting the different modules. They are very often implemented by arrays of optical fibers (usually organized in ribbons). If propagation loss on short fiber-spans (as expected to be inside an OOO) is normally negligible, coupling loss can be an issue in guided systems due to optical-mode mismatching between fiber and waveguide.

It should be noted that, while any multistage architecture can in principle be implemented on multiple substrate, there are some architectures which are topologically more suitable of being decomposed than others. The multi-substrate implementation is thus more frequent for modular topologies. Incidentally, modular fabrics are often also among the cheapest possible (in terms of number of SEs required to achieve a given switch size).

2.3 Signal-transfer impairments

The signal-transfer performance class includes some parameters related to the degradation suffered by optical signals that cross the switching network.

The first impairment on signal usually considered is attenuation. Contributions to the global attenuation come from several sources: the SEs, the waveguide-bends, the substrate/fiber interfaces, etc., that a given signal crosses inside a switching fabric, all dissipate fractions of its optical power. Thus, attenuation depends on the path the signal has been routed on through the switch: different routings may result in different attenuation values. We define *insertion loss* L the worst-case value among all those attenuations a signal can possibly face by following a path through the switch. Insertion-loss evaluation is performed from the input to the output port connected by the worst-case path.

Switch loss-features are also characterized by another important parameter: the *insertion loss difference* Δ. This is defined as the differential attenuation between the most and the least lossy paths through the fabric. Frequently, insertion loss difference is a more critical impairment than insertion loss itself. This is because a high common attenuation can be compensated for with the addition of optical amplifiers, while compensation of insertion loss differences requires a more complex distributed equalization system. A high differential attenuation adversely affects also the optical receiver, since it must be designed for a wider amplitude-dynamic and it must be adapted to react to high fluctuations of the signal-to-noise ratio.

Loss possible contributions are so many that accurate attenuation values can be obtained only by experimental measurements. Since the purpose of our paper is to provide analytical expression for many different switch architectures, we have necessarily to simplify the physical problems resorting to approximated analysis. In the following of the paper we will estimate insertion loss and insertion loss difference by considering only a limited number of phenomena as contributions to attenuation. We are going to complete this discussion in the next section providing specific expressions for L and Δ for each architecture.

Optical signals are degraded through a space switch also for the accumulation of noise, which lower the Signal-to-Noise Ratio (SNR), increasing error probability. Guided-wave switching architectures are seriously affected by this signal impairment, as the majority of integrated-optics systems. Being a switching fabric a multiport device, where many signals converges together in a limited physical space, the dominant noise source is the interference of a given optical channel with the other channels that are simultaneously present in the switch. The *crosstalk* measures the total optical power transferred to an output of a space-switching matrix from all the inputs different from the one from which comes the expected signal. In order to measure crosstalk, a common approach is to consider the worst case in which all connections are active at the same time. Moreover, crosstalk originating in the interstage interconnections due to optical beam intersection is usually negligible [26]: only power leakage in non-ideal SEs is thus considered. When two connections are active at the same time in the two channels switched by a 2×2 SE, each of them receives a so called *first-order* crosstalk contribution. If a connection crosses alone a 2×2 SE, it does not receive first-order crosstalk. However, the free input of the SE may be connected to a free output of another 2×2 SE which is crossed by another active connection. In this case the first connection receives a *second-order* crosstalk contribution. In our analysis we will always consider crosstalk only up to the second order: this is generally considered sufficiently accurate.

Each SE of the fabric (generally, a directional coupler) is characterized by the so called *extinction ratio* m, which is the ratio of power that is leaked from each channel to the other channel. In the SNR calculations we will use m and the equivalent parameter X, which is the inverse of the extinction ratio in dB

$$X = 10 \log_{10} \frac{1}{m}$$

The majority of papers dealing with optical guided-wave switching fabrics assumes an extinction ratio $m = 0.01$ ($X = 20$ dB). SNR is then evaluated assuming crosstalk as the only source of noise, which is reasonable if we are interested to characterize the fabric itself, considering any other possible noise source (e.g. ASE of optical amplifiers which may be present at inputs or outputs of the fabric) as external. Moreover, ideal optical signals with equal power are assumed to be applied to all the inputs. SNR, as attenuation, generally depends on the path followed by a signal inside the switch. The worst-case SNR among all the possible connections of the switching-network is regarded as the interesting figure, while differential SNR is usually not relevant.

To easily evaluate the entity of SNR a common approximated procedure is followed: second-order crosstalk is only calculated when it is known that no first-order contribution reaches the switch output of the worst-case connection. Otherwise, second-order crosstalk evaluation is useless, since its contribution would be negligible compared to the first-order components. The choice between first and second order calculation is readily taken by inspecting the architecture under exam: first-order crosstalk is evaluated when at least one SE along the worst-case path accommodates two active connections at the same time; second-order crosstalk is computed instead when all the SEs along the worst-case path (and thus in the whole network) are crossed by no more than one connection. The second situation occurs in the so called "dilated" architectures. Under the

above approximations, in order to evaluate the total first-(second-) order crosstalk noise at the output of the worst-case connection, it is sufficient to count the number of SEs along its path that give first- (second-) order crosstalk contributions. If P_i is the common input power for all the connections, each of these contributions will be equal to P_i/m (P_i/m^2). It should be noted that in the approximated conditions the attenuation experienced by a connection will equally apply to both signal and crosstalk-noise contributions and will cancel out of the final SNR. Thus, if A is the number of crosstalk-active (contributing) SEs along the worst-case path, SNR is given by

$$\mathrm{SNR} = \frac{P_o}{P_{\mathrm{xtalk}}} = \frac{1}{A \cdot m^{\gamma}}$$

where $\gamma = 1$ or $\gamma = 2$ if the evaluated crosstalk is of the first or the second order, respectively. Usually, SNR is expressed in dB as follows

$$\mathrm{SNR[dB]} = 10 \log_{10} \left[\left(\frac{1}{m} \right)^{\gamma} A^{-1} \right] = \gamma X - 10 \log_{10} A$$

In our crosstalk approximated evaluation we have considered only SEs. Actually, in guided-wave architectures two signals also interacts at waveguide crossovers, where a small crosstalk noise is generated. Crossover crosstalk is often difficult to measure and strictly depends not only on waveguide and material properties, but also on the geometry of the intersections (and in particular on the intersection angle). It has been traditionally not considered, being regarded as negligible compared to the SE crosstalk. It can become however a major cause of impairment in those architectures in which the number of crossovers is extremely high. We are currently studying this problem in order to define a less approximate model which takes also crossovers into account in crosstalk evaluations.

We shall finally point out that for the reasons mentioned at the end of Sec. 2.1, in the framework of this paper it is impossible to distinguish between crosstalk at the same (homodyne) or at a different wavelength (heterodyne) of the signal. It is however well known that homodyne and heterodyne crosstalks have very different impacts on the quality of signal.

2.4 Cost

The cost of an optical space-switching matrix can be evaluated by counting its elementary components and attributing to each one of them a unitary cost. This evaluation is not always accurate. Especially with integrated optics, costs of the fabrication process (and its production yield) is related to parameters such as substrate area and waveguide width, which are not necessarily directly proportional to the number of SEs or their density. However our rough cost evaluation may be useful to have an idea of the scalability of each particular architecture.

Characteristics correlated to the cost of the architectures comprise: number of 2×2 switching elements, total number of driver devices.

As already mentioned above, the switching architectures we are considering are obtained by replicating many times a single elementary SE (usually a directional coupler). The *number of SEs* S necessary to achieve a given switch size is the first total-cost parameter we will consider to compare the switching architectures.

The state of each SE (cross or bar) is electrically controlled by some kind of driving actuator, according to the specific switching technology adopted (electromechanical, thermal, electromagnetic, electrostatic, etc.). Any driver is per-se a source of cost. The number of drivers is not always equal to the number of controlled SEs. In fact, there is a number of architectures (e.g. the tree-type networks) in which many SEs can be controlled by the same driver, since the switching network is devised in such a way that there are groups of SEs which always change their state coherently and simultaneously. Therefore, we have considered the *number of drivers* D as the second relevant cost-parameter (of course, $D \leq S$ in any architecture).

3 Guided-wave switching architectures

Guided-wave space switches are often fabricated on lithium niobate substrates on which light-guiding structures are created by titanium indiffusion (Ti:$LiNbO_3$ technology)[23]. Various types of electrically-controlled optical SEs can be integrated on the substrate, including electrooptical and thermooptical switches[8]. Complex and very powerful architectures can be realized exploiting 1×2, 2×1 and 2×2 SEs. These devices are relatively reliable and capable of changing their state extremely rapidly, guaranteeing low switching time values. Unfortunately, they suffer high insertion loss and possible polarization dependence.

In order to evaluate attenuation, three types of contributions have been taken into account: power dissipation of each SE, waveguide-to-fiber and fiber-to-waveguide coupling loss[2] and crossover loss. The last contribution is a peculiarity of waveguided architectures: in these fabrics, SEs and interconnection-stage waveguides are integrated on a common substrate. Unlike integrated electronic circuits, in which connections between the various components can be made at multiple levels and separated by dielectric material, in integrated optics waveguides physically cross each other on the same substrate. Waveguide intersection is the cause of many undesirable effects, one of which is power loss.

[2] It should be noted that in all the architectures considered, coupling loss is path-independent (equal for all the paths).

Let us consider the worst-case path, that crosses K SEs, I (fiber-waveguide + waveguide-fiber) interfaces and W waveguide crossovers. The total *insertion loss* (in dB) is given by

$$L[\text{dB}] = s \cdot K + c \cdot I + w \cdot W$$

where s is the loss of a SE, c is the loss of a fiber-waveguide (waveguide-fiber) interface and w is the loss experienced by a signal propagating along a waveguide each time this crosses another waveguide. The *insertion loss difference* is calculated by subtracting from L the loss (in dB) obtained applying the above equation to the best-case path.

3.1 Switching architectures

We have considered 22 different guided-wave network architectures and we have reported their characteristics into three different tables, grouping them according to their blocking features. Fig. 1, 2 and 3 concern SNB, WNB and RNB architectures, respectively. For each architecture, we have analyzed the seven features introduced in Sec. 2: physical structure, insertion loss, insertion loss difference, SNR, number of SEs, number of drivers and number of waveguide crossovers. The choice of the physical structure type has been made according to the intrinsic properties of each architecture. For modular networks we have preferred multi-substrate realizations while we have chosen single-substrate structures where it has been difficult or useless to identify simple modules inside the switch. In the tables, letters S and M indicate single-substrate and multiple-substrate structures, respectively. When the multiple-substrate configuration has been chosen for a certain architecture, the last column of the three tables indicates the additional number of waveguide crossovers between the input and output ports that the worst-case connection would cross if the same architecture had been realized on a single substrate.

In the tables, values of insertion loss, insertion loss difference and SNR are expressed in dB while the other values are dimensionless numbers. All the parameters are given as functions of the number of inputs and outputs of the architectures (the size of the network) N.

Let us list the architectures we have considered, providing for each the reference to the paper in which it has been presented:

- SNB: classical Clos [4]; Double Layer [16]; Dilated Double Layer [16]; Extended Baseline [29]; Strict-sense non-blocking NWN (named here NWN-S) [27]; the numerous class of tree-type networks comprising a number of architectures, as indicated in Fig. 1 [22, 17, 11, 10, 9].
- WNB: classical Crossbar [7]; Double Crossbar [14]; Modified Double Crossbar [12]; Wide-sense non-blocking NWN (named here NWN-W) [26].
- RNB: classical Benes [1]; Dilated Benes [18]; Modified Dilated Benes [13]; classical Slepian-Duguid [20]; N-Stage Planar [24].

The association of the Crossbar architecture to the wide-sense non-blocking network class needs an additional comment. The optical guided-wave implementation of the Crossbar network is usually realized by using N^2 2×2 SEs, where every SE is equivalent to a crosspoint of the electronic network. Since every SE has two possible active inlets and two possible switching states, such construction offers multiple routes for various i/o pairs. Only one of these possible routes guarantees the non-blocking condition for the network, being the others possible causes of blocking states. So, a routing algorithm must exist (although very simple) that guarantees network non-blockingness. In this way, the network is wide-sense non-blocking. However, the border between SNB and WNB networks is so thin that the same architecture becomes SNB considering a different type of SEs. For example, 2-D MEMS Crossbar systems with single-face-reflective micromirrors fall in the SNB category. In this case, in fact, there is a unique possible path between every input and output, making blocking states impossible without the need of any routing algorithm.

In Clos and Slepian-Duguid networks, given a value of N, another independent parameter, n, must be chosen in order to fully specify the topology. n is the number of input ports of the switching matrices of the first stage (the number of output ports of the third stage matrices is chosen equal to n). We have chosen $n = \sqrt{N/2}$, a value which minimizes the total number of SEs.

For the multiple-substrate implementations of the Clos and Slepian-Duguid architectures, we have assumed that guided-wave Crossbar networks are used to realize each module of the three stages: modules are connected by fibers arranged in EGS patterns.

3.2 Architecture comparison

We are now going to numerically compare the various architectures on the basis of the described characteristics. The values of the physical parameters s, c, w and X have been chosen according to the following considerations. s depends on the type of SEs employed. A typical value for directional couplers is $s = 0.5$ dB. It is convenient, however, to consider the increased value $s = 1$ dB to keep waveguide-propagation loss into account. This choice is based on the assumption that the path length of a signal across the switching architecture is roughly proportional to the number of crossed SEs. Realistically, c could be around 1.5 dB, but we have chosen the more conservative value $c = 2$ dB. Extinction ratio is strongly dependent on the switching element type. In literature, a value of $X = 20$ dB is suggested for directional-coupler-based switches [16].

	Architecture	Physical structure	Insertion loss [dB]	Insertion loss difference [dB]	SNR [dB]	Number of SEs	Number of drivers	Additional waveguide crossovers (single substrate)
	Clos	M	$5[(2N)^{0.5}-1]s+6c$	$[5(2N)^{0.5}-8]s$	$X-10\log_{10}[(5/2)*(2N)^{0.5}-4]$	$4N[(2N)^{0.5}-1]$	$4N[(2N)^{0.5}-1]$	$4N-6(2N)^{0.5}+4$
	Double Layer	S	$(2\log_2N-1)s+2c+w(3N-2\log_2N-4)$	0	$X-10\log_{10}[1+m(\log_2N-1)]$	$(5/4)N^2-2N$	$N^2/4+2N\log_2N-2N$	–
	Dilated Double Layer	S	$(2\log_2N)s+2c+w(3N-2\log_2N-3)$	0	$2X-10\log_{10}(\log_2N)$	$2N(N-1)$	$2N\log_2N$	–
	Extended Baseline	S	$(3\log_2N-2)s+2c+w(4N-3\log_2N-5)$	0	$X-10\log_{10}(2\log_2N-1)$	$(3N^2/2)-(5N/2)$	$6N\log_2N-4N$	–
	NWN-S	M	$(4\log_2N-3)s+c(4\log_2N-2)+2w\log_2(N/2)$	0	$X-10\log_{10}(2\log_2N-1)$	$3^{\log_2N-1}+2^{2\log_2N+1}x[(3/2)^{\log_2N-1}-1]$	$3^{\log_2N-1}+2^{2\log_2N+1}x[(3/2)^{\log_2N-1}-1]$	$4N-4\log_2N-4$
Tree-type networks	Conventional AS/AC "Type1"	M	$(2\log_2N)s+4c$	0	$2X-10\log_{10}(\log_2N)$	$2N(N-1)$	$2N\log_2N$	$(N-1)^2$
	Conventional PS/AC, AS/PC "Type2"	M	$(3+s)\log_2N+4c$	0	$X-10\log_{10}(\log_2N)$	$N(N-1)$	$N\log_2N$	$(N-1)^2$
	Enhanced PS/AC	M	$(3+s)\log_2N+s+4c$	0	$2X-10\log_{10}(\log_2N)$	$N(2N-1)$	$N\log_2N+N^2$	$(N-1)^2$
	Enhanced PS/PC	M	$6\log_2N+s+4c$	0	$X-10\log_{10}(N-1)$	N^2	N^2	$(N-1)^2$
	Simplified AS/AC	M	$(2\log_2N-1)s+6c$	0	X	$N(5N/4-2)$	$N(5N/4-2)$	$N^2/2-N$
	Simplified PS/AC	M	$(3+s)\log_2N-3+6c$	0	$X-10\log_{10}(\log_2N)$	$N(3N/4-1)$	$N(3N/4-1)$	$N^2/2-N$
	Simplified AS/PC	M	$(3+s)\log_2N-3+6c$	0	$X-10\log_{10}(N/2)$	$N(3N/4-1)$	$N(3N/4-1)$	$N^2/2-N$
	Two-Active-Stage "Jajszczyk s"	M	$6(\log_2N-1)+2s+6c$	0	$X-10\log_{10}(N/2)$	$3N^2/4$	$3N^2/4$	$3N-2\log_2N-4$

Figure 1. Complete characteristics of strict-sense non-blocking guided-wave architectures.

Architecture	Physical structure	Insertion loss [dB]	Insertion loss difference [dB]	SNR [dB]	Number of SEs	Number of drivers	Additional waveguide crossovers (single substrate)
Crossbar	S	$(2N-1)s+2c$	$(2N-2)s$	$X-10\log_{10}(N-1)$	N^2	N^2	–
Double Crossbar	S	$(N+1)s+2c+wN(N-1)$	0	$2X-10\log_{10}(N-1)$	$2N^2$	$2N^2$	–
Modified Double Crossbar	S	$(3N/4+2)s+2c+w(N\log_2N+N/4-2)$	0	$2X-10\log_{10}(\log_2N-1)$	$2N^2$	$3N^2/2$	–
NWN-W	M	$(4\log_2N-3)s+c(2+4\log_2N/2)$	$(2\log_2N-2)s$	$X-10\log_{10}(2\log_2N-1)$	$13(3^{\log_2N-1})-6N$	$4N(3/2)^{\log_2N-1}-4N+3^{\log_2N-1}$	$4N-4\log_2N-4$

Figure 2. Complete characteristics of wide-sense non-blocking guided-wave architectures.

Architecture	Physical structure	Insertion loss [dB]	Insertion loss difference [dB]	SNR [dB]	Number of SEs	Number of drivers	Additional waveguide crossovers (single substrate)
Benes	S	$(2\log_2 N-1)s$ $+2c$ $+w(2N-2\log_2 N-2)$	0	X $-10\log_{10}(2\log_2 N-1)$	$(N/2)(2\log_2 N-1)$	$(N/2)(2\log_2 N-1)$	–
Dilated Benes	S	$(2\log_2 N)s$ $+2c$ $+w(4N-2\log_2 N-5)$	0	$2X$ $-10\log_{10}[(2\log_2 N-1)$ $\log_2 N]$	$2N\log_2 N$	$2N\log_2 N$	–
Modified Dilated Benes	S	$(2\log_2 N+1)s$ $+2c$ $+w(4N-2\log_2 N-5)$	0	$2X$ $-10\log_{10}[(\log_2 N-1)$ $\log_2 N]$	$2N(\log_2 N+1)$	$2N(\log_2 N+1)$	–
Slepian-Duguid	M	$[4(2N)^{0.5}-3]s$ $+6c$	$2[2(2N)^{0.5}-3]s$	X $-10\log_{10}[2*$ $(2N)^{0.5}-3]$	$(2N)^{1.5}$	$(2N)^{1.5}$	$2N-3(2N)^{0.5}+2$
N-Stage Planar	S	$Ns+2c$	$(N/2)s$	X $-10\log_{10}N$	$N(N-1)/2$	$N(N-1)/2$	–

Figure 3. Complete characteristics of rearrangeable non-blocking guided-wave architectures.

Despite the fact that modern fabrication techniques achieve a much lower level of crosstalk, we prefer keeping $X = 20$ dB as reference, to be conservative. Finally, considering 2μm-wide waveguides that intersect each other at an average angle of $50°$, a realistic value for signal loss due to every single crossover is $w = 45$ mdB [2]. It is important to say that such value refers to a specific case (InP-based photonic integrated circuits) and that different materials could give different results. However, the contribution of a single crossover to the total loss is so low that different values of w would give rise to significative differences only with very large architectures.

In the following, we show graphs for some network characteristics in which many architectures of the three classes (SNB, WNB and RNB) are plotted: this allows an easy observation of the differences among architectures belonging to the various classes. Curves corresponding to SNB architectures have been drawn with a continuous line, while a long-dashed line is used for WNB and a short-dashed line for RNB.

Fig. 4 shows the insertion loss. We can see that, in general, guided-wave architectures have large attenuation values also for small network dimensions. The loss increases very rapidly with the size. Large dimension networks can be used only if coupled with optical amplifiers capable to compensate the loss introduced by the switch. In the graph, some architectures display a linear trend while others present a parabolic trend, the latter being a relevant obstacle to scalability. The network architecture with the smallest value of insertion loss is the Benes network for $N \leq 64$, while for $N > 64$ the conventional AS/AC becomes the best one. The explanation of this behavior is that every signal in a Benes network must pass through $2\log_2 N - 1$ SEs while the same signal must pass through $2\log_2 N$ SEs in a conventional AS/AC architecture. Since we have chosen the single-substrate Benes implementation, the effect of the crossovers is rather negligible for small network sizes. On the opposite, when the network dimension increases, the signal loss due to the large number of waveguide crossovers cannot be neglected anymore and so the total attenuation becomes high. In the conventional AS/AC waveguide crossover is not relevant, because of the presence of optical fibers between diverse substrates. Fig. 4 does not show relevant differences in the insertion loss values for the architectures belonging to the three non-blocking classes. In other words, we can find architectures with a relative low or high insertion loss independently of the blocking property of the network (i.e. if it is SNB, WNB or RNB). It is interesting to note that several switching networks are roughly equivalent for small values of N under the loss point of view. For example, Benes, Double Layer, Dilated Double Layer, Dilated Benes and Modified Dilated Benes have similar attenuation values for $N \leq 32$. On the opposite, only the Conventional AS/AC architecture is capable to keep insertion loss on acceptable values when N becomes high. All the other architectures, in fact, reach very high attenuation levels because of the large number of SEs crossed by the connections and mainly because of the huge number of waveguide crossovers along the paths.

Only five architectures present a differential attenuation among input-output connections. For these architectures the differential attenuation rapidly increases to very large amounts, causing high insertion loss differences. The Crossbar architecture has the worst behavior under this aspect because the insertion loss difference is almost equal to the insertion loss value. This is due to the fact that the shortest path inside the matrix always crosses just one SE and hence it is independent of the network dimension. In a

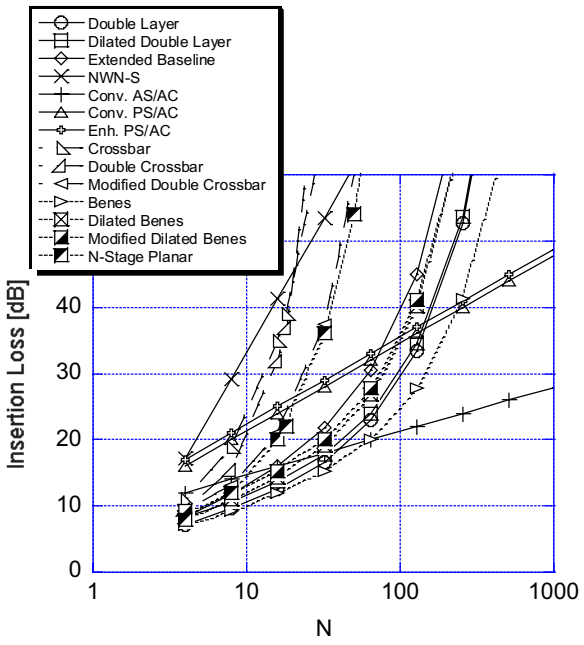


Figure 4. Insertion loss for various guided-wave switching architectures.

large dimension Crossbar, the longest path suffers losses of the order of many tens of dB while the attenuation on the shortest path, crossing only one SE, becomes almost negligible. This large insertion loss difference would impose so much stress on the optical receiver amplitude-dynamic that the use of large Crossbar architectures for switching becomes practically unfeasible. Clos and Slepian-Duguid architectures decrease the insertion loss difference with respect to the Crossbar network, but improvements are limited by the use of Crossbar networks in every block they are composed of. N-Stage Planar networks have a particular behavior: it can happen that $N-2$ i/o connections (out of N) cross exactly N SEs, while the other two connections cross $N/2$ SEs. For this reason in this type of networks the insertion loss difference is strongly dependent on the network size N and increases linearly with it. NWN-W networks have acceptable values of differential loss, especially for small architectures. Fortunately, all the other described switching networks have an insertion loss difference exactly equal to zero.

In Fig. 5 SNR is plotted as a function of the network size. Best results are obtained by the architectures which guarantee zero first-order crosstalk. These are: Modified Double Crossbar (having the best behavior in absolute), Enhanced PS/AC, Modified Dilated Benes, Dilated Benes and Double Crossbar. The case of the Double Layer Networks is very interesting: they have an almost constant high value of SNR, despite they belong to the class of non-zero first-order crosstalk networks. As the previously analyzed characteristics, also the signal-to-noise ratio is independent of the connection property of the network. For example, the three architectures that present the best SNR performance are wide-sense, strict-sense and rearrangeable non-blocking respectively. It could be proved that the minimum acceptable SNR to achieve a bit error rate of 10^{-9} is about 11 dB [10]. For this reason, all the architectures with a lower SNR value are hardly of practical use. Unfortunately, many guided-wave switching architectures have SNR below 11 dB also with a very small network size. Some architectures have SNR also below 0 dB, which means that the noise power is larger than the signal power in the optical signals outgoing by the switch. So, if the goal is to construct a large architecture, we have a very limited number of wave-guided switching architectures to rely upon.

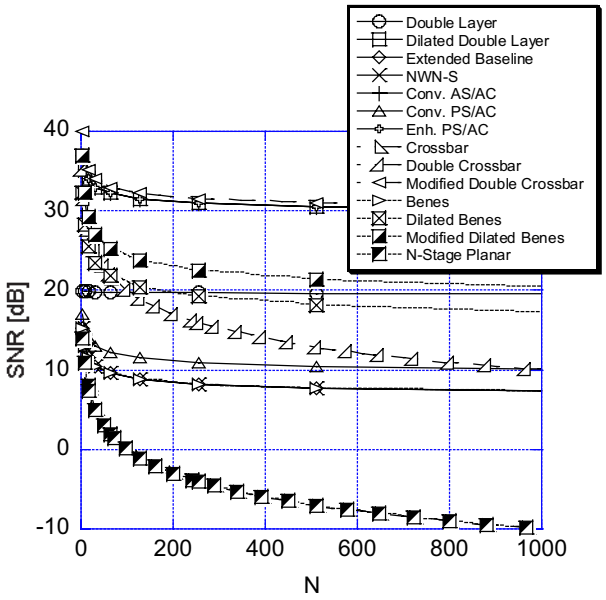


Figure 5. Signal-to-Noise characteristics for various guided-wave switching architectures.

Let us now consider the cost parameters. The analyzed architectures behave very differently under the point of view of the total number of SEs (Fig. 6). As it is well known, the very large number of SEs, increasing with the square of the network size, is one of the main problems of the Crossbar. Fig. 6 shows that many architectures require a number of SEs even greater than that of the Crossbar. On the other hand, there are some architectures whose complexity increases slowly with N, resulting nicely scalable. Two important aspects can be noted in the graph. First, an enormous difference in the number of SEs exists between the

category of RNB architectures and the other two categories. Second, the difference in the number of SEs between SNB and WNB networks is not very relevant. So, if the disruption of optical connections inside the space switch is not a problem, RNB networks are strongly to be preferred. On the opposite, if connection rearrangement has to be avoided, SNB architectures are the best choice since their control is very easy and their cost is similar to that of WNB architectures.

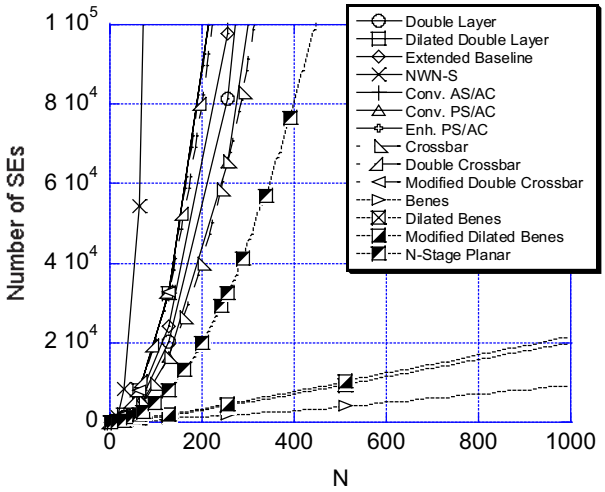


Figure 6. Number of SEs required for various guided-wave switching architectures.

Let us now consider the number of drivers with the size of the networks. Dilated Double Layer, Extended Baseline, Conventional AS/AC and PS/AC architectures, interestingly, require a very small number of drivers even with a large number of SEs. This is because in tree-type networks all the SEs belonging to the same column in a splitter or in a combiner change their state at the same time and then they can be driven by a common control signal. In the other two types of networks, even if there are neither splitters nor combiners, some SEs change their state at the same time and can be driven by the same system. RNB architectures ensure the best performance also for this characteristic. The sole exception is represented by the N-Stage Planar networks that require a large number of SEs and drivers, even greater than that of many SNB architectures.

Optical signals passing through single-substrate architectures usually encounter a very high number of waveguide crossovers. Under this aspect, a multiple-substrate implementation is the best solution. Fig. 7 represents the maximum number of waveguide crossovers between input and output ports in various networks with different values of N. We have considered only architectures for which the physical structure reported in Fig. 1-3 is the single-substrate type, except the NWN-S. For networks having a number of crossovers in the order of some thousands, the attenuation term due to waveguide crossovers cannot be neglected in the expression of the insertion loss.

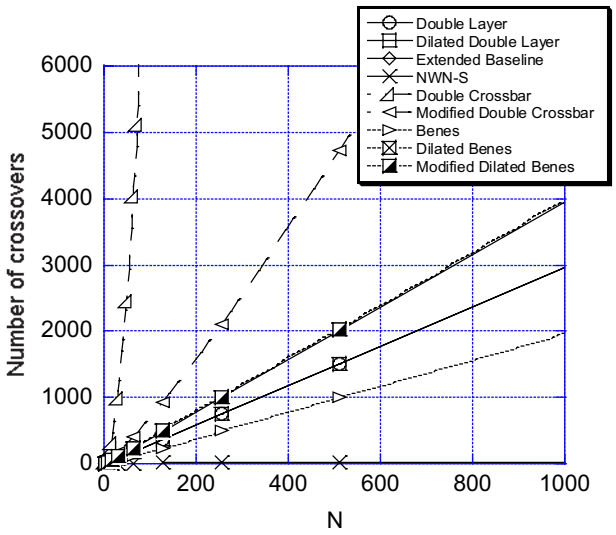


Figure 7. Number of waveguide crossover for various guided-wave switching architectures.

4 Conclusion

Optical switching node design is a complex task, perhaps even more complex than the design of electronic switches. In the electronic domain, the unique target is to minimize the total number of switching elements required by the network. On the opposite, in the optical domain, there are other additional parameters to be optimized as system attenuation, number of crossovers, and signal-to-noise ratio. Therefore, the optimal architectures for photonic switching are not necessarily the optimal architectures for electronic switching, since the two applications undergo different sets of constraints. This study addresses the considerations involved in selecting a guided-wave architecture to be used as an optical space switch fabric. The relevant characteristics of optical switching architectures have been defined and a great number of previously proposed structures have been reviewed. To successfully compare different architectures, many issues have still to be investigated. However, we can say in general that there is no best switching architecture in absolute, but only architectures most suitable for each specific application, depending on the relative constraints the specific application requirements impose to the various characteristics. If all-optical switching will actually

prove to be cost effective compared to electronic switching in the future, some of the switching architectures reviewed in this paper could be probably adopted in the next generation OXCs.

References

[1] V. E. Benes. *Mathematical Theory of Connecting Networks and Telephone Traffic*. New York: Academic Press, 1965.

[2] H. G. Bukkems, C. G. P. Herben, M. K. Smit, F. H. Groen, and I. Moerman. Minimization of the Loss of Intersecting Waveguides in InP-Based Photonic Integrated Circuits. *IEEE Photonics Technology Letters*, 11(11):1420–1422, November 1999.

[3] D. K. Cheng, Y. Liu, and G. J. Sonek. Optical switch based on thermally-activated dye-doped biomolecular thin films. *IEEE Photon. Technol. Lett.*, 7:366–369, April 1995.

[4] C. Clos. A study of non-blocking switching networks. *Bell System Tech. Journal*, 32:406–424, Mar. 1953.

[5] A. Dugan, L. Lightworks, and J. C. Chiao. The optical switching spectrum: A primer on wavelength switching technologies. *Telecommun. Magazine*, May 2001.

[6] S. Hardy. Liquid-crystal technology vies for switching applications. *Lightwave*, pages 44–46, December 1999.

[7] H. S. Hinton. A nonblocking optical interconnection network using directional couplers. In *GLOBECOM 1984*, volume 26, pages 26.5.1–26.5.5, 1984.

[8] H. S. Hinton. Photonic switching using directional couplers. *IEEE Communication Magazine*, 25(5):16–26, May 1987.

[9] A. Jajszczyk. A Class of Directional-Coupler-Based Photonic Switching Networks. *IEEE Transactions on Communications*, 41:599–603, April 1993.

[10] A. Jajszczyk and H. T. Mouftah. Tree architectures for photonic switching. In *Global Telecommunications Conference*, volume 3, pages 1605–1609, December 1992.

[11] A. Jajszczyk and H. T. Mouftah. Tree-Type Photonic Switching Networks. *Network*, 9:10–16, January 1995.

[12] W. Kabacinski. Modified Double Crossbar Photonic Switching Networks. In *Electrotechnical Conference MELECON 98, 9th Mediterranean*, volume 2, pages 754–758, 1998.

[13] W. Kabacinski. Modified Dilated Benes Networks for Photonic Switching. *IEEE Transactions on Communications*, 47(8):1253–1259, August 1999.

[14] M. Kondo, N. Takado, K. Komatsu, and Y. Ohta. 32 switch elements integrated low-crosstalk $LiNbO_3$ 4×4 optical matrix switch. In *IOOC-ECOC '85*, pages 361–364, Venice, 1985.

[15] L.-Y. Lin, E. L. Goldstein, and R. W. Tkach. Free-Space Micromachined Optical Switches for Optical Networking. *Journal of Selected Topics in Quantum Electronics*, 5(1):4–9, January-February 1999.

[16] C.-C. Lu and R. A. Thompson. The Double-Layer Network Architecture for Photonic Switching. *Journal of Lightwave Technology*, 12(8):1482–1489, August 1994.

[17] H. Okayama, A. Matoba, R. Shibuya, and T. Ishida. Optical Switch Matrix With Simflified N×N Tree Structure. *Journal of Lightwave Technology*, 7(7):1023–1028, July 1989.

[18] K. Padmanabhan and A. Netravali. Dilated Networks for Photonic Switching. *IEEE Transactions on Communications*, 35:1357, December 1987.

[19] G. I. Papadimitriou, C. Papazoglou, and A. S. Pomportsis. Optical Switching: Switch Fabrics, Techniques, and Architectures. *Journal of Lightwave Technology*, 21(2):384–404, February 2003.

[20] A. Pattavina. *Switching Theory: Architectures and Performance in Broadband ATM Networks*. John Wiley & Sons, Baffin Lane, Chichester, West Sussex, England, first edition, 1998.

[21] D. A. Smith, A. Alessandro, J. E. Baran, D. J. Fritz, J. L. Jackel, and R. S. Chakravarthy. Multiwavelength performance of an adopized acoust-optic switch. *Journal of Lightwave Technology*, 14:2044–2051, September 1996.

[22] R. A. Spanke. Architecture for Large Nonblocking Optical Space Switches. *IEEE Journal of Quantum Electronics*, 22(6):964–967, June 1986.

[23] R. A. Spanke. Architecture for Guided-Wave Optical Space Switching Systems. *IEEE Communication Magazine*, 25(5):42–48, May 1987.

[24] R. A. Spanke and V. Benes. An N-stage planar optical permutation network. *Applied Optics*, 26, April 1987.

[25] T. E. Stern and K. Bala. *Multiwavelength Optical Networks: A Layered Approach*. Addison-Wesley, 1999.

[26] F. M. Suliman, A. B. Mohammad, and K. Seman. A New Nonblocking Photonic Switching Network. In *GLOBECOM '01*, volume 4, pages 2071–2076, 2001.

[27] F. M. Suliman, A. B. Mohammad, and K. Seman. A Space Dilated Lightwave Network - A New Approach. In *International Conference on Telecommunications*, volume 2, pages 1675–1679, February-March 2003.

[28] A. Ware. New photonic-switching technology for all-optical networks. *Lightwave*, pages 92–98, March 2000.

[29] C.-S. Wu, G.-K. Ma, and B.-S. P. Lin. Extended Baseline Architecture for Nonblocking Photonic Switching. *Journal of Lightwave Technology*, 15(5):771–778, May 1997.

Performance of Optical Packet Switching Nodes in IP Transport Networks

Maurizio Aste, Achille Pattavina
Dept. of Electronics and Information,
Politecnico di Milano
e-mail:pattavina@elet.polimi.it

Abstract

This paper[1] deals with optical packet switching in a full-IP transport network scenario. For the switching of IP packet flows different node architectures are considered that are based on current optical routing devices. The traffic performance of a mesh network is evaluated with the various node structures, assuming that nodes employ either shortest path routing or deflection routing to forward packets to the addressed destinations. The paper shows how the different node structures behave in terms of packet loss probability with different network configurations when the node parameters are varied.

1 Introduction

In the latest years, the outbreak of Internet and broadband services for the development of electronic commerce, entertainment and education has involved the increase of demand for transmission bandwidth. Nowadays, routing of traffic flows in transport networks occurs by processing electronically data and transmitting them in optical fibers; optics is therefore used exclusively at the physical level. Advent of the wavelength multiplexing technologies (WDM) has concurred an increase of the transmission capacity per fiber but at the same time has shown the limits of actual network infrastructures based on electronic devices. Today networks use only a small fraction of the large-capacity made available by each fiber (in the order of terabits per second), since electronic switching, processing and storage technologies do not allow to manage fully the huge size of data transported by fibers. At the same time the current need for transporting very large amounts of traffic based on the IP protocol has shown all the limits of SONET/SDH technologies, born as solutions for circuit switching and thus unable to guarantee efficient management of the IP traffic flows.

The development of new switching systems is therefore important in order to face such new networking scenario. The advent of optical switching devices is going to define a new generation of network elements in which routing is operated without optical/electronic conversions. In this context optical circuit and packet switching technologies play a different role. With *optical circuit switching* every connection needs the reservation of an entire WDM channel in order to realize *end-to-end* circuits, with the inefficient bandwidth usage typical of circuit switching. *Optical packet switching* on the other hand enables a high and efficient exploitation of the available capacity thanks to the bandwidth sharing typical of statistical multiplexing. This latter technology would moreover be consistent with the new *paradigma* of an IP protocol that supports any kinds of telecommunication service.

Unfortunately, today optical devices used in market equipment are still too crude to allow packet-by-packet operation. An interesting solution which tries to represent a balance between circuit switching low hardware complexity and packet switching efficient bandwidth utilization is *optical burst switching* [10], [8]. In an optical burst switching system, the basic units of data are bursts, made up of multiple packets, which are sent after control packets, carrying routing information, whose task is to reserve electronically the necessary resources on the intermediate nodes of the transport network.

This paper addresses the future-looking scenario of optical packet switching by exploiting the optical switching technologies available today. In particular it considers the architecture of optical packet switching nodes already proposed in [1][2][3], which exploit arrayed waveguide grating devices for packet routing and are equipped with fiber delay lines used either for input buffering or for shared buffering of optical packets. The paper is organized as follows. Sections 2 describes the envisioned optical network architecture, while section 3 details the proposed structures of the

[1]Work partially supported by MIUR, Italy, under FIRB project ADONIS and by EU IST Network of Excellence e-Photon/One.

optical packet switching nodes. Finally section 4 provides a traffic performance comparison of the different node architectures in a mesh network with different packet routing strategies.

2 Network Architecture

The architecture of the optical transport network we propose consists of *optical packet-switching nodes*, which are mutually connected in a mesh-like topology. A number of *edge systems* (ES) interfaces the optical transport network with IP legacy (electronic) networks (see figure 1). An ES receives packets from different electronic networks (or local hosts) and performs *optical packet* generation. The optical packet is composed of a simple optical header, which comprises the destination address, and of an optical payload made of a single IP packet, or, alternatively, of an aggregate of IP packets. The optical packets are then buffered and routed through the optical transport network to reach their destination ES, which delivers the traffic it receives to its intended destination in the electronic domain. At each intermediate node in the transport network, packet headers are received and electronically processed, in order to provide routing information to the control electronics, which will properly configure the node resources to switch packet payloads directly in the optical domain.

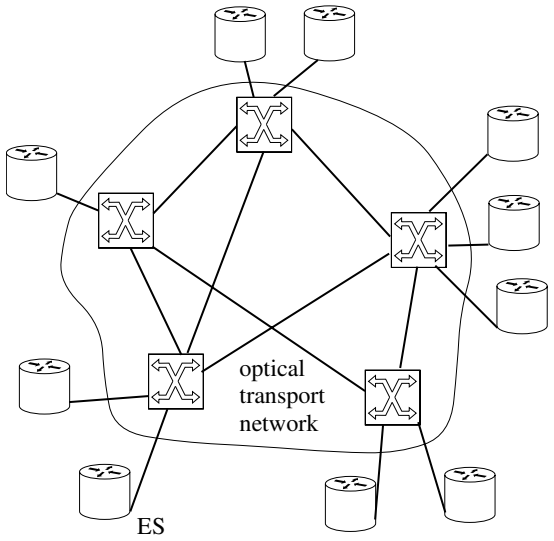


Figure 1. The optical transport network architecture

The transport network operation is *asynchronous*; that is, packets can be received by nodes at any instant, with no time alignment. The internal operation of the optical nodes, on the other hand, is assumed to be *synchronous*, or *slotted*, since the switching of packets in an unslotted node is less regulated and more unpredictable, resulting in a larger contention probability. In our model we propose, the time slot duration, T, to be equal to the amount of time needed to transmit from an input WDM channel to an output WDM channel an optical packet with a 40-bytes payload which corresponds to the smallest payload enabled by an IP packet. Supposing a bit rate of 10 Gbit/s per wavelength channel, a 40 ns slot duration seems appropriate, since the 40-bytes payload is transmitted in 32 ns, and the additional time can be used for the optical packet header transmission and to provide guard times. For a deeper discussion about this issue the reader is referred to [3].

In order to relieve the complexity of the problem related to the arbitrary mesh topology of the transport network, we assume here that all nodes of the transport network have the same *nodal degree* N_h, that is the same number of adjacent nodes. Since we have selected eight optical nodes in the transport network, we are going to examine here the four network topologies represented in figure 2, where the number N_h of adjacent nodes is N_h = 2, 3, 5, 7. Hence these regular topologies span from the ring network (N_h = 2) up to the full-mesh network (N_h = 7). The other two networks identify topologies with intermediate meshing degree. If we define the connectivity factor α as the ratio between the number of links in the network and that of a full-mesh network, it follows that the eight-node networks in figure 2 are characterized by a connectivity factor α_8 = 0.29, 0.43, 0.71, 1.

3 Node Architecture

The general architecture of a network node is shown in figure 3. It consists of three stages: a first stage of channel demultiplexing, a second stage of switching and a third stage of channel multiplexing. The node is fed by N incoming fibers each having W wavelengths. In the first stage the incoming fiber signals are demultiplexed and G wavelengths from each input fiber are fed into each one of the $N_p = W/G$ second-stage switching planes, which constitute the switching fabric core. Once signals have been switched in one of the parallel planes, packets can reach every output port through multiplexing carried out in the third stage using any of the G wavelengths that are directed to each output fiber. We note that the number of inlets of each third-stage multiplexer varies, depending on the specific structure of the switching planes. Wavelength conversion must be used for contention resolution, since at most G packets can be concurrently transmitted by each second-stage plane on the same output link.

The detailed structure of one of the W/G parallel switching planes is presented in figure 4. It consists of three main blocks: an input *synchronization unit*, as the packet switch-

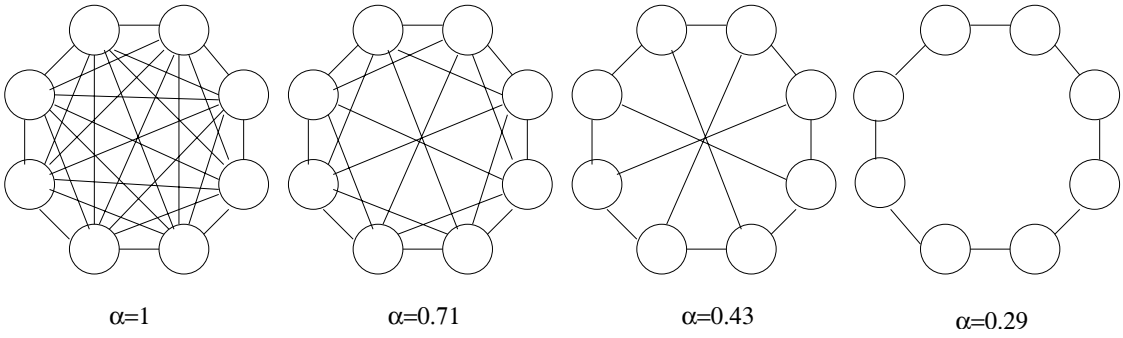


Figure 2. Network topologies

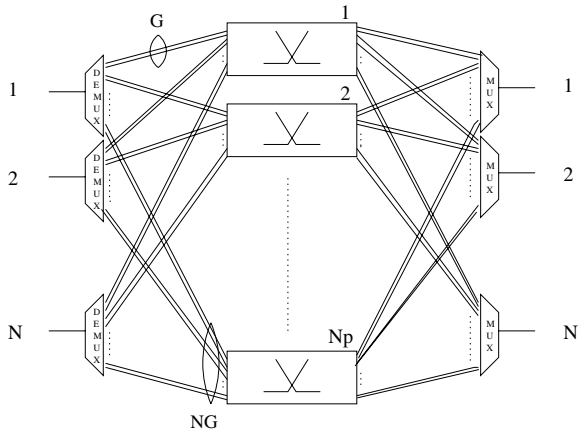


Figure 3. Optical packet-switching node general architecture

ing is slotted and incoming packets need to be slot-aligned, a *fiber delay lines unit*, used to store packets for contention resolution, and a *switching matrix unit*, adopted to achieve the switching of signals.

These three blocks are all managed by an *electronic control unit* which carries out the following tasks:

- optical packet header recovery and processing;
- managing the synchronization unit in order to properly set the correct path through the synchronizer for each incoming packet;
- managing the tunable wavelength converters inside the switching matrix, in order to properly delay and route incoming packets.

The details of the synchronization and of the fiber delay unit can be found in [3]. We simply recall here that the delay lines are used as an *optical scheduler* that, thanks to the electronic control, maximizes the number of packets transmitted on the requested output links. Given the maximum achievable delay D_{max} slot, for each switch input $D_{max}+1$ delay lines are needed in each plane, with delays growing from 0 to D_{max}. Once packets have crossed the fiber delay lines unit, they enter the switching matrix stage in order to be routed to the desired output port. This is achieved using a set of tunable wavelength converters combined with an arrayed waveguide grating (AWG) wavelength router [7].

The AWG is used here as it gives better performance than a normal space switch interconnection network, as far as insertion losses are concerned. This is due to the high insertion losses of all the high-speed all-optical switching fabrics available at the moment, that could be used to build a space switch interconnection network. Commercially available 40 channel devices have a channel spacing of 100 GHz and show an insertion loss of less than 7.5 dB [5]. Other proposals of switching nodes based on AWGs can be found in [4][6].

Three different structures are proposed here for the implementation of the switching matrix unit, referred to as structures Basic (B), Enhanced (E) and Optimized (O). In all these structures a *shared buffer* is implemented through recirculation lines in order to enable a much more effective contention resolution. R denotes the number of AWG ports destined to recirculation lines, each one delaying the packets by a fixed amount D_{ric} slot and R_{max} denotes that maximum number of recirculations allowed to a packet in the node.

3.1 Basic structure (B)

The simplest switching matrix (Basic structure), first proposed in [1], is shown in figure 5, referred to a single plane. It consists of $2NG + R$ tunable wavelength converters and an AWG with size $(NG + R) \times (NG + R)$. Only one packet is routed to each AWG outlet and this packet must finally be converted to one of the wavelengths used in the WDM channel, paying attention to avoid contention

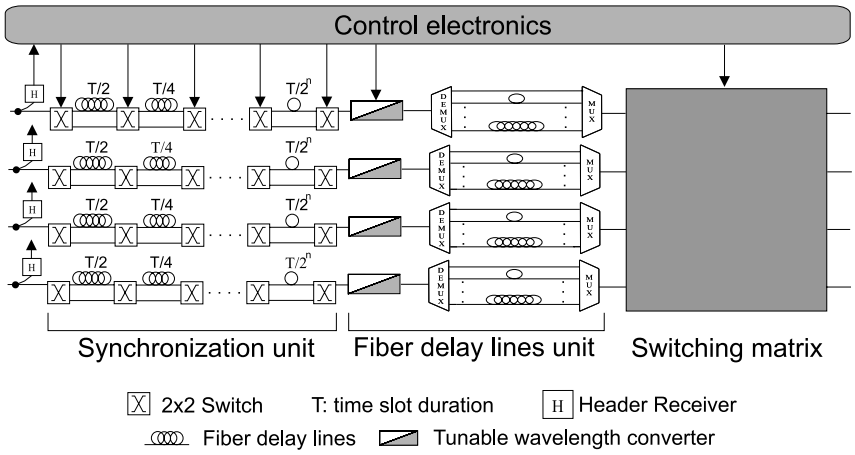


Figure 4. Detailed structure of one of the W/G parallel switching planes

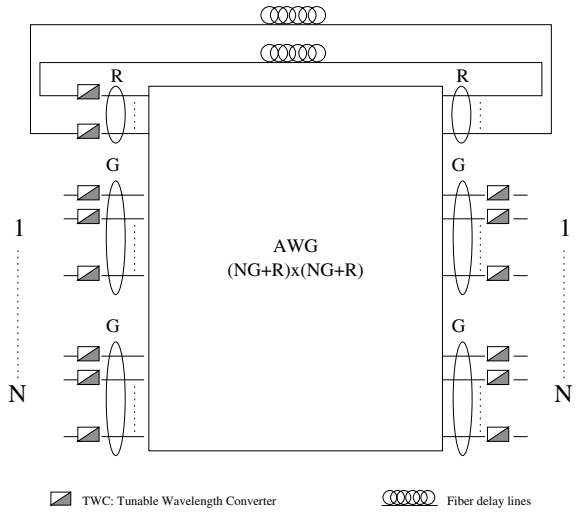


Figure 5. Basic switching matrix

with other packets of the same channel.

3.2 Enhanced structure (E)

In order to reduce the number of planes of the node and thus to better exploit the "channel grouping" effect (i.e. the sharing of different channels for transmitting a large number of packets, the load per channel being constant), more than one packet can be routed through each AWG inlet; apparently the packets sharing the same input must be transmitted on different wavelengths. The structure of the AWG is such that different wavelengths entering the same input port will emerge on different output ports.

In the Enhanced switching matrix structure illustrated in figure 6, up to k different packets are sent to the same AWG inlet using different wavelengths. A simple node design requires k to be an integer that divides G. From AWG input port i, the output channel j can be reached by G/k different packets, since there are exactly G/k AWG outlets connected to that channel. During each time slot, up to G packets can be routed to the same AWG outlet using different wavelengths. Hence, demultiplexers are needed to split the different signals and to route them to the last stage of wavelength converters. If $k \leq G/k$, no contention can happen in the multiplexing stage, so this structure behaves exactly as a structure Basic with size $NG \times NG$. On the other hand, when $k > G/k$, events of packet blocking occur, considering the fact that G/k paths are available to reach a tagged output for up to k packets per inlet. So, when more than G/k packets in the same AWG inlet are destined to the same output channel, a contention happens, even if the total number of packets addressed to that output is smaller than G.

3.3 Optimized structure (O)

The Enhanced node structure be simplified by selecting $k = N$, so that each AWG input can receive up to N packets using different wavelengths, thus implying that the node structure includes just one plane ($N_p = 1$). Therefore, the number of AWG inlets is now exactly W. In this structure, shown in figure 7 and referred to as Optimized structure, the last TWC stage isn't needed anymore, provided the employed AWG works on the same wavelengths used in the outgoing fibers. In fact, if the electronic controller takes care of avoiding wavelength contention between AWG outlets connected to the same output channel, packets are ready to be transmitted as soon as they exit the AWG. Therefore, a packet entering the AWG inlet i and destined to the output

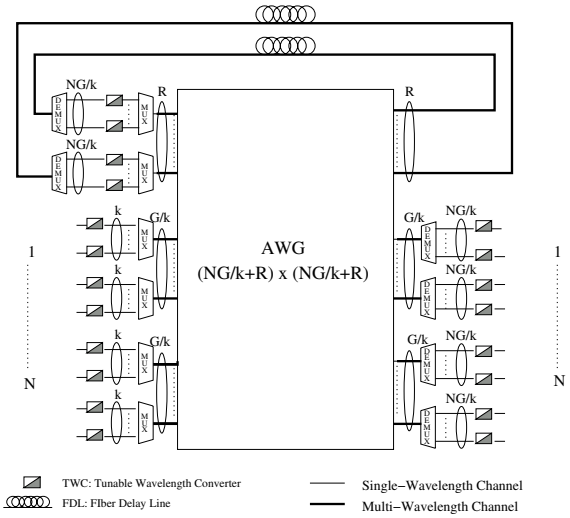


Figure 6. Enhanced switching matrix

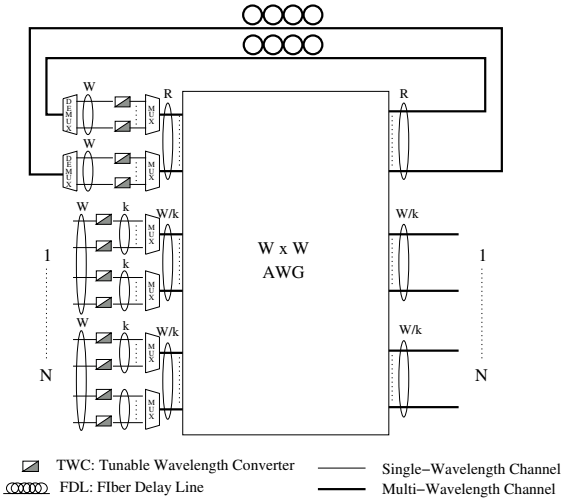


Figure 7. Optimized switching matrix

WDM channel j can not be transmitted using every color in the WDM channel, but only using a subset which consists of the W/N wavelengths through which the packet can reach the desired output channel, thus reducing the benefits of *channel grouping* (when N and W are kept constant).

4 Performance Evaluation

We show now some traffic performance results given by the different node architecture configurations obtained through computer simulation. In order to evaluate the performance provided by the three different switching structures, we will first examine the packet loss probability in 8-node networks with different connectivity factors. In particular we first consider the case of *shortest path routing*, in which each optical packet must exit the node on a specific output link (either to an adjacent node or to a local edge system). Then we will examine the effect of adopting *deflection routing*, in which an optical packet is forwarded to a link different from that identifying the shortest path if the intended link is already busy in spite of the input buffering possibilities enabled by the node. Such deflection routing is operated choosing random the output link. If the node is equipped also with shared buffering, this last storage possibility is exploited if all the outgoing links are busy. Finally we will examine the effect of varying one single network parameter, that is the number of nodes, the network connectivity factor α and the AWG size.

In order to obtain performance results mutually comparable in spite of the different connectivity factors, we assume that each network node interfaces a number N_h of edge systems (ES) generating the same amount of traffic ρ. In such a way the total number of ESs interfacing a node is equal to the number of network links outgoing from the node and the total (normalized) load offered to the network is ρ.

Traffic levels and shared buffer capacities are assumed equal for all nodes, so that Basic and Enhanced nodes require more planes ($N_p > 1$). The AWG size is set equal in all nodes and this value depends on the network connectivity. The number of wavelengths per plane has been set to $G = 2$ for the Basic node with shared buffer and the number of recirculation lines R equals half of the AWG size (if $R = 0$ only input buffering is exploited by the node). Given the previous assumption about the ESs (the internode links equal the number of local ESs), it follows that we are assuming an AWG size $8N_h \times 8N_h$. Then it follows that the AWG sizes are 16×16, 24×24, 40×40 and 56×56 for the nodal degrees $N_h = 2, 3, 5, 7$, respectively. Furthermore the number of wavelengths W equals the AWG size, since the Optimized node includes just one plane.

The buffering capacity of each node has been set according to the following parameters: $D_{max} = 8$ slot in input buffers, while $D_{ric} = 2$ slot and $R_{max} = 4$ when shared buffering is equipped ($R > 0$).

As far as the offered traffic distribution is concerned, packet interarrivals for each ES wavelength has been modelled as a Poisson process with negative exponential distribution. Based on measurement of real IP traffic [9], the following distribution of packet length L has been assumed

$$\begin{cases} p_0 = Pr(L = 40 \text{ bytes}) = 0.6 \\ p_1 = Pr(L = 576 \text{ bytes}) = 0.25 \\ p_2 = Pr(L = 1500 \text{ bytes}) = 0.15 \end{cases}$$

so that the resulting average packet length is 393 bytes.

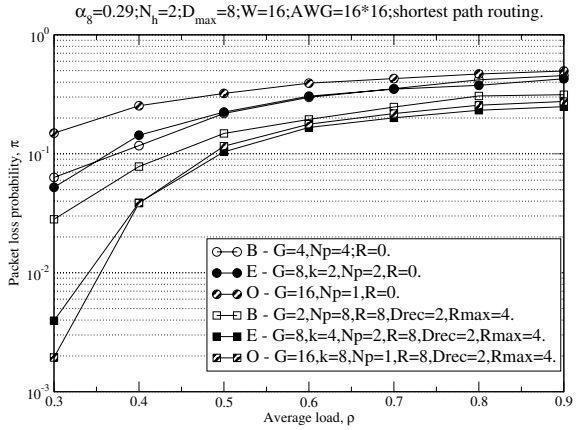


Figure 8. Packet loss performance with shortest path routing and $\alpha = 0.29$

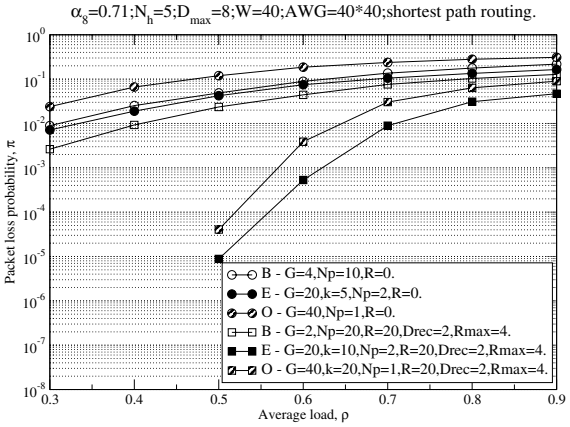


Figure 10. Packet loss performance with shortest path routing and $\alpha = 0.71$

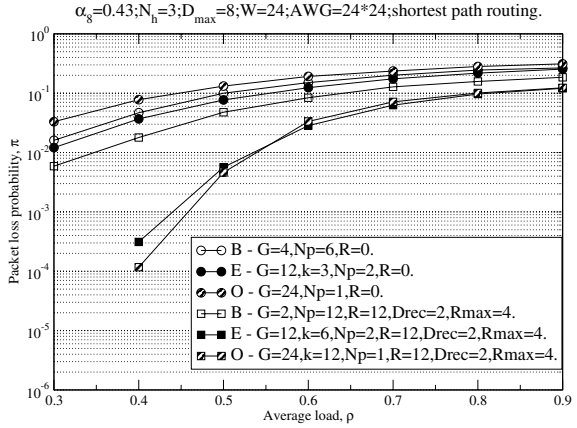


Figure 9. Packet loss performance with shortest path routing and $\alpha = 0.43$

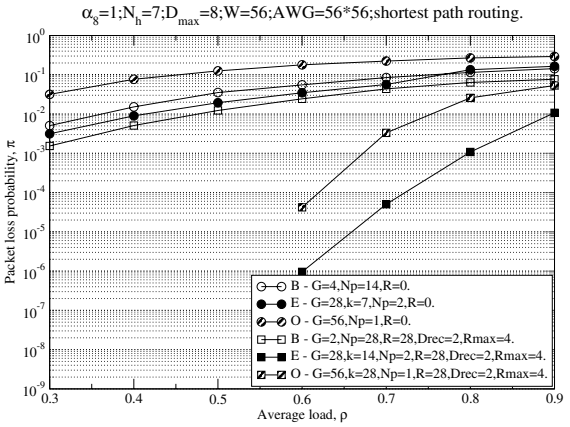


Figure 11. Packet loss performance with shortest path routing and $\alpha = 1$

Packets are assumed to be equally likely to be addressed to any destination ES.

4.1 Shortest path routing

The packet loss probability with shortest path routing is shown in figures 8, 9, 10, 11, for $\alpha_8 = 0.29, 0.43, 0.71, 1$, respectively. We notice that, as we might expect, solutions with shared buffering give better performance than without it. The Optimized node gives the worst performance without shared buffers, the other two solutions providing similar behaviour. An explanation of this is that packets coming from a local ES or a node cannot be routed to the same node outlet (ES or node). So this implies that with the Enhanced and Optimized solutions more contentions arise for accessing each single node outlet compared to a case in which all node outlets are equally addressable, as assumed in [3].

On the other hand when shared buffering is employed, the Enhanced and Optimized structures outperform the Basic one especially for large network connectivity factors. This is due to the fact that the shared buffer is distributed among a much larger number of planes in the former node structures and then becomes less efficient than a larger buffer in a single plane (Optimized solution).

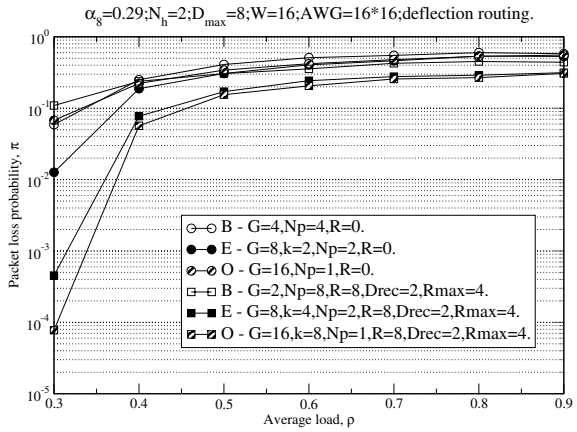


Figure 12. Packet loss performance with deflection routing and $\alpha = 0.29$

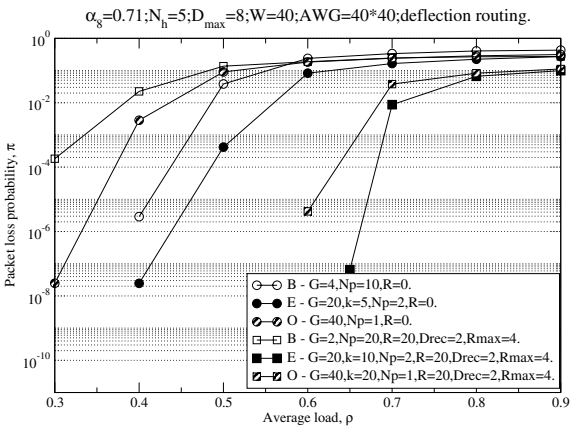


Figure 14. Packet loss performance with deflection routing and $\alpha = 0.71$

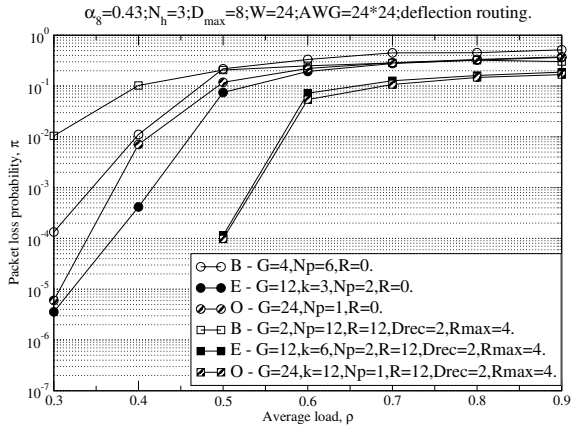


Figure 13. Packet loss performance with deflection routing and $\alpha = 0.43$

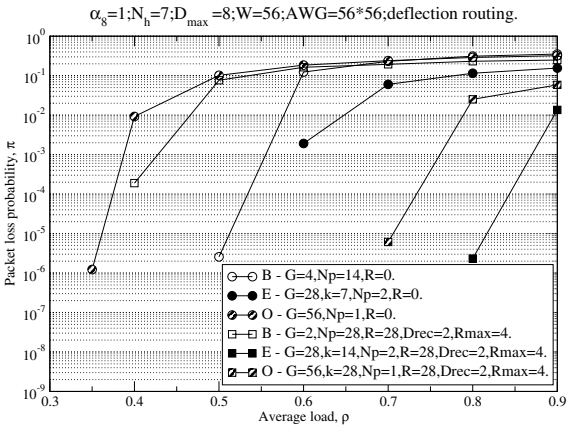


Figure 15. Packet loss performance with deflection routing and $\alpha = 1$

4.2 Deflection routing

The three switching node structures are now compared with deflection routing for the same values of the network connectivity and the corresponding results are shown in figures 12, 13, 14, 15.

It is quite interesting to note that now the network performance improves significantly as the network connectivity grows due to the fact that deflection routing can be exploited better with more output links from the node. When shared buffering is employed Enhanced and Optimized node structures give again the best performance due to the larger buffers they can exploit, as with shortest path routing. Without shared buffering, the Enhanced node gives the best performance, whereas the Optimized node behaves the worst. The reason is analogous to that given for shortest path routing. In fact the load offered to the node outputs is not evenly distributed, since now packets entering the node from local outlets cannot exit the node on the same outlet. Note that now, due to deflection routing, packets entering the node from an upstream node can be routed on any output. Hence it follows that the load offered to inter-node outlets is larger than that received from local outlets of the node. Notice that such behaviour applies to all node structures.

Unlike networks based on Enhanced and Optimized nodes, adding shared buffering with the Basic solution provides worse loss performance. As already observed in [3], shared buffering with Basic node is beneficial only if the (fixed) delay of recirculation lines exceeds a given threshold, for example 16 slots (recall that we have assumed here a delay $D_{ric} = 2$ slots).

5 Conclusions

In this work, we have analyzed and compared the packet loss performance of three types of optical nodes that differ substantially for the switching matrix. We have found that the Enhanced node performs always better than the others without or with shared buffering. In the former case the Basic node provides better results, whereas in the latter case the situation is reversed. Adding shared buffering to a node by keeping the same AWG size (and hence increasing the number of planes with Basic solution or increasing the multiplexing factor in Enhanced and Optimized solutions) provides large benefits with all nodes except in the case of deflection routing and Basic structure. Adopting deflection routing improves significantly the loss performance especially for low traffic loads and high connectivity factors.

References

[1] S. Bregni, G. Guerra, and A. Pattavina. Optical Packet Switching of IP Traffic. In *Proceedings of* 6^{th} *Working Conference on Optical Network Design and Modeling (ONDM)*, 2002.

[2] S. Bregni, G. Guerra, and A. Pattavina. Optical Switching of IP Traffic Using Input Buffered Architectures. *Optical Network Magazine*, 3(6):20–29, 2002.

[3] S. Bregni, A. Pattavina, and G. Vegetti. Architectures and Performances of AWG-based Optical Switching Nodes for IP Networks. *IEEE Journal on Selected Areas in Communications*, 21(7):1113–1121, 2003.

[4] J. Cheyens, J. Jennen, E. V. Breusegem, M. Pickavet, and P. Demeester. Optical Packet Switches Based on a Single Arrayed Waveguide Grating. *Conference on High Performance Switching and Routing (HPSR)*, pages 5–9, Jun. 2003.

[5] S. Electronics. Fiberoptics products: Arrayed Waveguide Grating. http://www.samsungelectronics.com/fiberoptics/.

[6] L. Li, S. Scott, and J. Deogun. Performances Analysis of WDM Optical Packet Switches with a Hybrid Buffering Architecture. *Conference on Optical Networking and Communications (OptiComm)*, 5285:346–356, Oct. 2003.

[7] C. Parker and S. Walker. Design of Arrayed-Waveguide Gratings Using Hybrid Fourier-Fresnel Transform Techniques. *IEEE Journal on Selected Topics in Quantum Electronics*, 5:1379–1384, 1999.

[8] C. Qiao. Labeled Optical Burst Switching for IP-over-WDM Integration. *IEEE Communications Mag.*, pages 104–114, Sep. 2000.

[9] K. Thompson, G. J. Miller, and R. Wilder. Wide-Area Internet Traffic Patterns and Characteristics. *IEEE Network Magazine*, pages 10–23, Nov. 1997.

[10] M. Yoo and C. Qiao. Just-Enough-Time(JET): A High Speed Protocol for Bursty Traffic in Optical Networks. In *Proc. IEEE/LEOS Tech. for a Global Info Infrastructure*, pages 26–27, Aug. 1997.

Cluster Computing 7, 239–244, 2004

Performance Evaluation of Deflection Routing in Optical IP Packet-Switched Networks *

STEFANO BREGNI and ACHILLE PATTAVINA **
Politecnico di Milano, Department of Electronics and Information, Piazza L. Da Vinci 32, 20133 Milano, Italy

Abstract. In previous papers [5,6], an optical switch architecture was proposed to handle variable-length packets such as IP datagrams, based on an AWG device to route packets and equipped with a fiber delay-line stage as optical input buffer. Unfortunately, extensive simulations of optical networks built with switches of this type showed that considerable buffering capability would be required in order to achieve acceptable performance. In this work, therefore, we studied the effectiveness of packet deflection as a mean for solving packet contentions on outputs of optical switches. Optical transport networks were simulated, evaluating the performance of packet deflection routing, based on a traffic model adherent to real IP traffic measurements. Full-mesh and wheel network topologies have been considered, comparing results to assess deflection effectiveness. Our simulation results show that deflection routing leads to satisfying performance even using buffers with limited size. Furthermore, the average delivery delay does not suffer heavy penalty from packet deflection, even under heavy traffic conditions.

Keywords: Arrayed Waveguide Grating (AWG), Internet Protocol (IP), optical networking, optical switching

1. Introduction

The exponential growth of Internet users and the introduction of new broadband services have been fostering an unprecedented increase of network capacity. On the other hand, the IP architecture is being seen as the unifying paradigm for a variety of services and for the Broadband Integrated Services Network (B-ISDN), which has been foreshadowed since the 1980's. To face this challenge, considerable research is currently devoted to design IP full-optical backbone networks, based on Wavelength Division Multiplexing (WDM) technology, in order to relieve the capacity bottleneck of classical electronic-switched networks.

Photonic packet switching represents a potential solution [1–3]. Today, unfortunately, optical devices available on the market are still not mature enough to allow packet-by-packet operation in the optical domain. Optical burst switching has been proposed as intermediate solution between pure packet and circuit switching [4]. However, packet switching features an higher degree of statistical resource sharing, which should lead to a better bandwidth utilization when the network carries bursty traffic such as IP traffic.

This work is based on the optical switch architecture proposed in [5,6], based on an Array Wavelength Guide (AWG) to route packets to outlets and equipped with a fiber delay-line stage as optical input buffer. This optical switch was designed to handle variable-length packets, such as IP datagrams. Its performance was also evaluated for some typical statistical distribution empirically verified in the Internet.

The network architecture proposed in [5,6] simplifies the encapsulation of IP datagrams in optical packets by eliminating fragmentation issues. Moreover, it allows ultra-fine statistical resource allocation, being able to switch independently 40-bytes packets. Unfortunately, this switch would require considerable buffering to achieve acceptable performance, thus relying on expensive optical hardware and control electronics.

A possible solution, studied in this paper, is to implement efficient packet deflection inside the optical network, as a mean for solving packet contentions on outputs of optical switches. Thus, optical networks have been simulated to assess deflection effectiveness, based on a traffic model adherent to real IP traffic measurements. Full-mesh and wheel network topologies have been considered.

This paper is organized as follows. In section 2, the architecture of optical network studied in this work is introduced, summarizing the optical packet format and the switching architecture. In section 3, the system and traffic simulation models are described. In section 4, several simulations results are presented. Finally, section 6 draws some conclusions.

2. Architecture of the optical transport network

The general architecture of the optical network, as proposed in [5,6], is shown in figure 1 and consists of *M optical packet switching nodes*, each denoted by an unique optical address made of $m = \lceil \log_2 M \rceil$ bits, linked together according to a suitable topology. A number of *Edge Systems* (ES) interfaces the optical transport network with IP legacy electronic net-

* This paper is mainly based on the paper "Deflection Routing Effectiveness in Full-Optical IP Packet Switching Networks", by M. Baresi, S. Bregni, A. Pattavina and G. Vegetti, included in the *Proceedings of the IEEE Conference ICC 2003*, Anchorage, AK, USA (May 2003).

** Corresponding authors.
E-mail: {bregni,pattavina}@elet.polimi.it

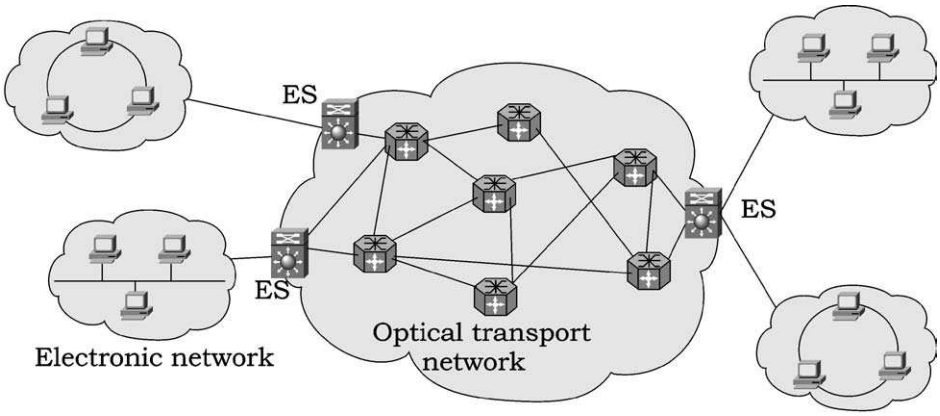


Figure 1. Architecture of the optical transport network.

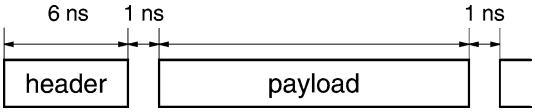


Figure 2. Optical packet format.

works. In our model, N ESs are connected to each optical node. Therefore, the total number of ESs is $N \cdot M$.

Edge systems multiplex IP datagrams from electronic networks and encapsulate them into optical packets with no fragmentation. Optical packets are then routed through the optical network to reach their destination ES, which delivers them to the destination electronic networks. The network operation is asynchronous: packets are transmitted between nodes without enforcing any time alignment. Conversely, internal operation within optical nodes is synchronous (slotted), to achieve lower contention probability [5,6].

2.1. Optical packet format

An optical packet is composed of a simple header, carrying the m-bits destination address, and a payload made of a single IP packet, as shown in figure 2. The optical header has fixed length while the payload size is not constrained.

The minimum time slot T of operation in optical nodes is the time needed for the smallest optical packet, carrying a 40-bytes IP datagram, to go from input to output ports. A 40 ns time slot seems appropriate, since 40 bytes are transmitted in 32 ns at the base speed 10 Gbit/s and 8 ns can be used for optical header transmission and to provide guard times. These are required between payload and header and between contiguous packets, to allow header processing and to account for some packet temporal skew inside switching nodes. Duration of guard intervals has been set to 1 ns. Therefore, header has duration 6 ns. It is transmitted at fixed rate 10 Gbit/s, while payload transmission can be set to higher rates, since the network is totally transparent to payload format and bit rates except for optoelectronic stages integrated into Edge Systems.

2.2. Optical switch architecture

The internal operation of optical nodes is synchronous, to achieve lower contention probability. Therefore, all packets entering input ports have to be aligned first to time slots, of duration $T = 40$ ns to accommodate smallest IP packets (40 bytes), before being routed by the switching fabric.

The structure of the optical switch is shown in figure 3. For a detailed description of its architecture and operation, the reader is referred to [5,6]. In this section, only its main features are highlighted.

Input WDM channels are demultiplexed, so that each wavelength enters the switch from a different inlet. At the switch output, W adjacent outlets, being W the number of wavelength per channel, are then multiplexed on the output WDM channel.

At the switch input, headers are first read and sent to the control electronics (H blocks). An n-stages *synchronization unit*, consisting of a series of 2×2 Semiconductor-Optical-Amplifier (SOA) switches interconnected by fiber delay lines of different lengths, aligns incoming packets to time slots.

The second stage is the *fiber delay lines (FDL) unit*, which stores packets to accomplish optical buffering and scheduling for coping with contention resolution on output ports. Tunable Wavelength Converters (TWCs) are used to route packets to the chosen delay line. The optical scheduling algorithm sets variable delays for packets entering the switching matrix. This algorithm even allows two packets entering the switching matrix in inverted order compared to that in which they entered the FDL unit, supposed that a sufficient maximum delay is available (buffer depth $D_{\max}$).

Finally, the third stage is the *switching matrix unit*, based on an Arrayed Waveguide Grating (AWG) device and two stages of TWCs, where the first stage is needed to route packets to the desired output and the second is responsible to convert the signal to a suitable wavelength, in order to avoid two packets to be transmitted using the same color.

2.3. Packet deflection

Packet deflection extends internal switch buffering, using network links as longer optical delay lines. Nevertheless, de-

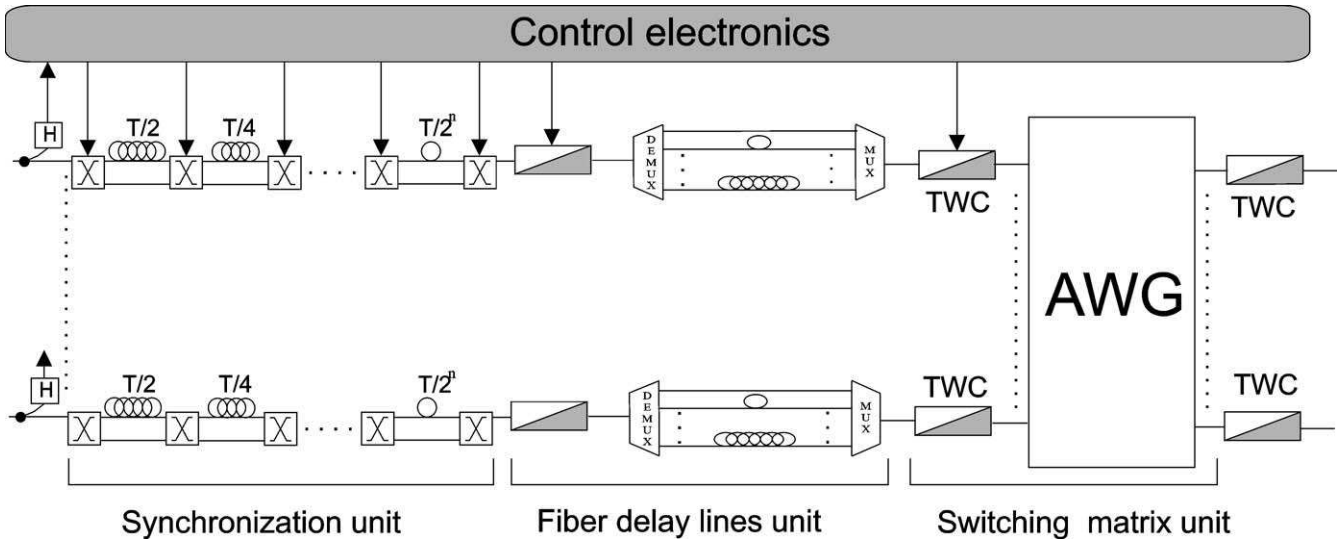


Figure 3. Structure of optical switch based on Arrayed Waveguide Grating (AWG).

flection generally leads to increasing network load. Thus, optimal deflection algorithms should direct packets to links scarcely loaded first, aiming at uniforming load among network links.

In this work, uniform packets deflection has been implemented: when a packet is deflected, it is routed with equal probability to one of the output links that are able to propagate it without further contention. At every switching node, deflected packets are handled as normal packets and are routed toward destination without any special processing. A *hop limit* H (i.e., *time to live*) is also enforced, to discard packets pinging too long inside the network.

3. Simulation model

According to the general network architecture shown in figure 1, we simulated the operation of optical transport networks for different network topologies and by varying the number of ESs, which generate and receive the IP traffic. Moreover, since the purpose of this work was to assess the performance of the transport network, we chose the simplest *star topology* to connect Edge Systems to optical switches.

In this work, we aimed at assessing the effectiveness of packet deflection in our optical transport network architecture. Therefore, we chose to simplify switch hardware complexity. In the FDL unit, we set the maximum buffering depth to $8T$. On the other hand, it has been shown in [5,6] that W should be set large enough in order to obtain satisfying performance, due to the *channel grouping* phenomenon. For this reason, the number of WDM channels used for the single input-output fiber has been set to $W = 20$.

For the aggregated traffic generated on each wavelength by the ES, we adopted a Poisson model, with interarrival times exponentially distributed. The length L of IP datagrams generated is a random variable, with empirical distribution according to real IP traffic measurements [7]:

$$\begin{cases} p_1 = P(L = 40\,\text{byte}) = 0.60 \\ p_2 = P(L = 552\,\text{byte}) = 0.25 \\ p_3 = P(L = 1500\,\text{byte}) = 0.15. \end{cases} \quad (1)$$

Hence, average packet length is 387 bytes. Moreover, the traffic pattern has been assumed addressed uniformly to all possible destinations of the network: therefore, the destination address of each packet is a random variable uniformly distributed between all possible ES addresses.

4. Simulation results

In this section, we present a selection of results obtained by the extensive simulations carried out. Full-mesh and wheel network topologies have been considered. Finally, results obtained in the two cases are compared.

All simulation results reported in this section are the central values of confidence interval estimates, with confidence level set to 95% and interval width on the order of 5%.

4.1. Full-mesh networks

We considered mesh networks with size $M = 3, 6, 9$. If not otherwise indicated, the number of Edge Systems connected to each transport switching node was set to $N = M - 1$. Hence, every switching node is connected to $M - 1$ ESs and $M - 1$ other switches. In this way, the traffic A [Erlang] offered by each ES equals the traffic offered on the average to each network link (network load). The packet hop limit has been arbitrarily set to a multiple of the network size M ($H = 0$, $H = 9$ or $H = 18$).

Figures 4 and 5 plot the packet loss probability, evaluated for networks with $M = 3$ and 6 nodes and hop limit $H = 0, 9$ and 18, versus the offered load A. The network exhibits better performance for higher levels of deflection. Conversely,

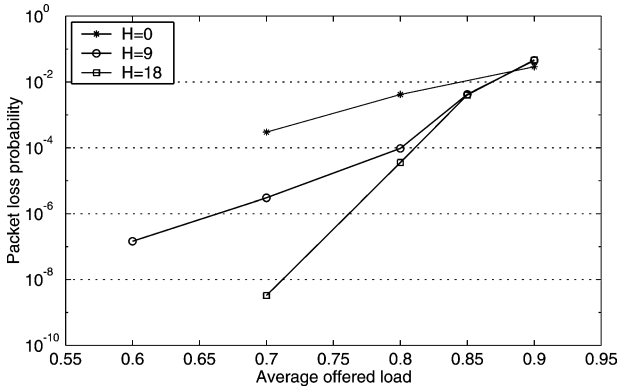


Figure 4. Packet loss probability in a full-mesh network with $M = 3$ nodes.

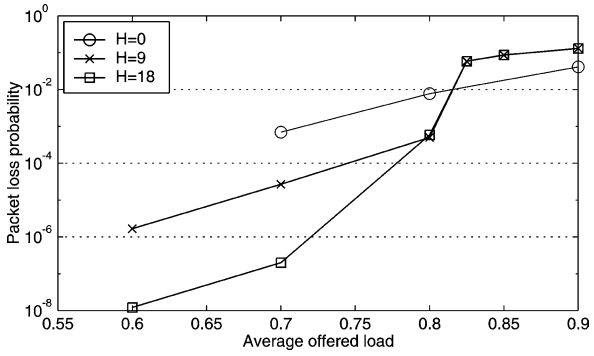


Figure 5. Packet loss probability in a full-mesh network with $M = 6$ nodes.

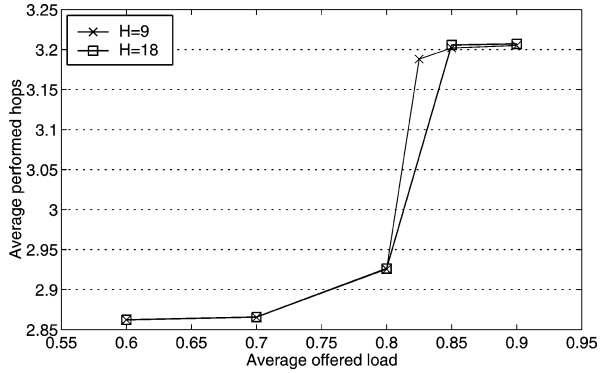


Figure 6. Average number of hops counted by delivered packets in a full-mesh network with $M = 6$ nodes.

under heavy traffic conditions, packet deflection worsens network performance. Deflected packets, in fact, represent a further load for single nodes, which leads to higher packet loss especially when heavy traffic is offered to the network.

The average number of hops required to deliver a packet is plotted versus the offered load A in figure 6, for $M = 6$ and $H = 9$, 18. Deflection routing does not increase drastically the average hop count. Even under congestion, a limited number of hops is sufficient to deliver packets in most cases.

In figures 7 and 8, networks with $M = 3$ and 6 nodes are directly compared, by setting the H parameter respectively to $H = 9$ and 18. We can see that loss probability increases as the number of nodes grows. This behavior is not determined by deflection routing, but is common to all switching systems

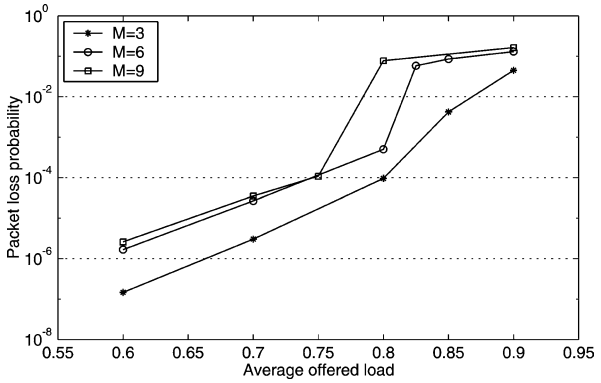


Figure 7. Packet loss probability in full-mesh networks with $M = 3, 6, 9$ nodes and $H = 9$.

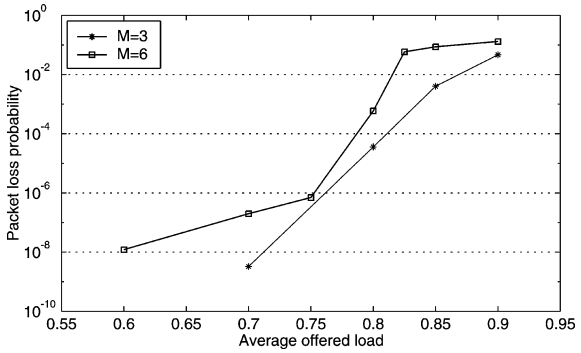


Figure 8. Packet loss probability in full-mesh networks with $M = 3, 6$ nodes and $H = 18$.

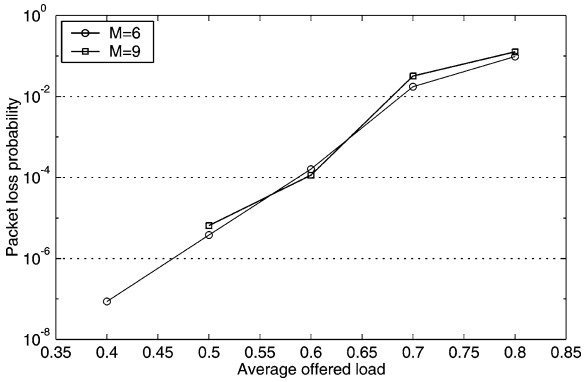


Figure 9. Packet loss probability in the network with nodes with $D_{\max} = 0$.

featuring input queuing (head-of-line blocking).

This consideration is supported also by figure 9, which shows the performance of networks with $M = 3$ and 6, where the maximum buffer depth has been set to $D_{\max} = 0$ (no input queuing) and the hop limit to $H = 18$. In these cases, the loss probability does not depend on the network size M.

To better understand network behavior under heavy load, we can examine the results shown in figure 10, where the loss probability versus the number of Edge Systems is plotted keeping constant the network load $A = 0.8$ Erlang.

Increasing the number of hosts per single transport node yields better performance. Since we are keeping network load constant, in fact, we are decreasing the traffic offered by sin-

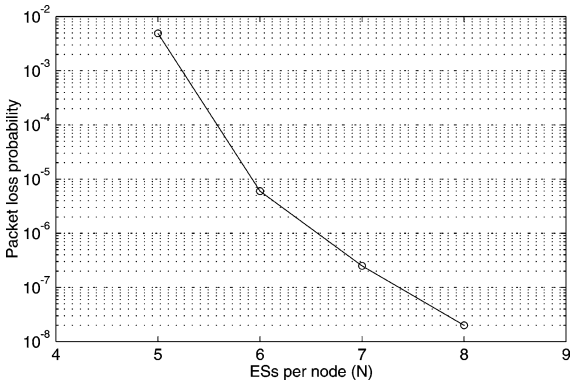


Figure 10. Packet loss probability in a full-mesh network with $M = 6$, $H = 18$, $N = 5, 6, 7, 8$ and constant offered load $A = 0.8$ Erlang.

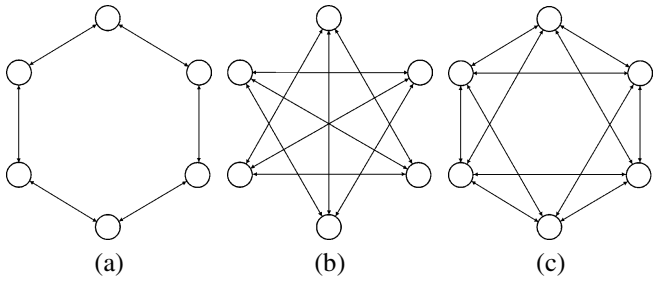


Figure 11. Wheel network topologies considered ($M = 6$).

gle ESs and therefore also the amount of traffic addressed to each ES. Thus, when a packet reaches the switching node directly linked to its destination, it has an higher probability to immediately delivered.

4.2. Wheel networks

In this section we present some results obtained for partially-meshed networks. Wheel networks are a particular class of regular network topologies that are easily represented placing nodes around a wheel. In particular, we considered 6-nodes wheel networks, with the three connection topologies depicted in figure 11. Table 1 summarizes the values of some characteristic parameters of these three network topologies.

The *connectivity factor* α is defined as

$$\alpha = \frac{2l}{M(M-1)}, \tag{2}$$

where l is the number of bidirectional links and M is the number of nodes. Therefore, α represents the ratio between the number of links in a wheel network and the number of links in a full-mesh network having same number of nodes.

The *network diameter* D is the maximum distance between two nodes. The *network order* Δ is the maximum number of links connected to a node. Finally, the *network number of hops* N_{H} is defined as the average distance seen from a node divided by the number of nodes of the network.

Figure 12 shows simulation results for these three kinds of network with a deflection limit $H = 18$ hops. Network performance worsens rapidly as α decreases. In fact, removing links from the network reduces deflection possibilities inside

Table 1
Characteristic parameters of the wheel network topologies shown in figure 11.

	l	α	D	Δ	N_{H}
Figure 11(a)	6	0.4	3	2	1.8
Figure 11(b)	9	0.6	2	3	1.4
Figure 11(c)	12	0.8	2	4	1.2

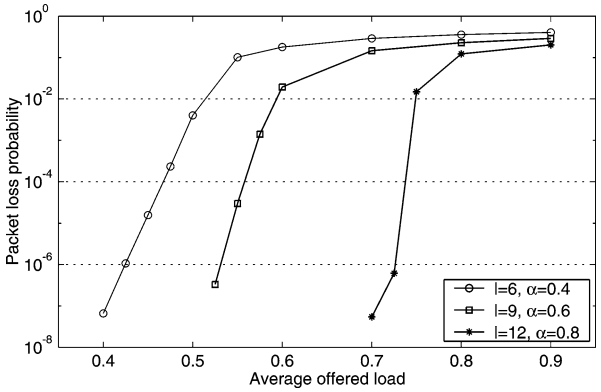


Figure 12. Packet loss probability in wheel networks with $M = 6$, $H = 18$.

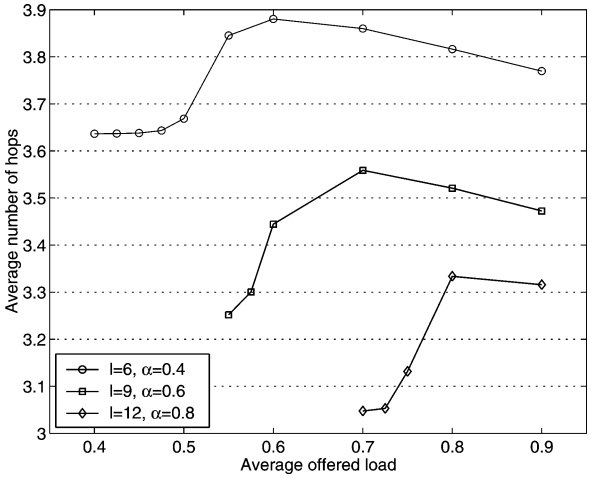


Figure 13. Average number of hops counted by delivered packets in a wheel network with $M = 6$, $H = 18$.

switches and this, combined with higher N_{H}, limits deflection effectiveness. The higher is the distance between two nodes, the higher is the chance for a packet to get deflected in the wrong direction around network topology, and thus also the probability to get lost.

Figure 13 displays the average number of hops performed by packets delivered to destination. All curves show a maximum: this is due to the fact that at very high loads many packets are discarded even before reaching the hop limit H, due to lack of available output links to any direction.

4.3. Comparison of full-mesh and wheel topologies

In figure 14, the performance of the full-mesh network is compared to that of the wheel network with $\alpha = 0.8$ ($M = 6$ and $H = 18$ in both cases).

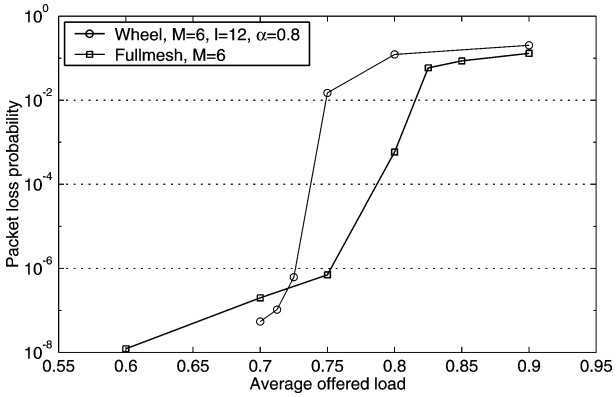


Figure 14. Comparison of full-mesh and wheel network ($\alpha = 0.6$) with $M = 6$ and $H = 18$.

Surprisingly, for medium-light loads the wheel topology outperforms the full mesh network. This behavior is explained observing that switching nodes have fewer input/output links in the wheel topology than in the full-mesh topology. Thus, input queuing may introduce a significant penalty.

5. Conclusions

In this work, we have studied the impact of packet deflection on the performance of optical IP packet switching networks with reduced buffer capacity. Based on the switch architecture proposed in [5,6], optical transport networks were simulated to assess deflection effectiveness, based on a traffic model adherent to real IP traffic measurements. Full-mesh and wheel network topologies were considered, comparing results to assess deflection effectiveness.

We have shown that, in full-mesh networks, deflection routing leads to satisfying performance even using buffers with limited size.

Furthermore, we pointed out that average delivery delay does not suffer heavy penalty from packet deflection, even in heavy traffic conditions.

Simulation results also confirmed that reducing the connectivity factor impairs substantially the performance of the optical transport network and the effectiveness of deflection routing. If the connectivity factor is low, cleverer deflection policies should be investigated, to avoid deflecting packets to nodes far from destination.

Acknowledgements

Work partially supported by the Italian Ministry of Education, University and Research (MIUR) under the FIRB project ADONIS.

References

[1] M.M. Renaud, F. Masetti, C. Guillemot and B. Bostica, Network and system concepts for optical packet switching, IEEE Communications Magazine 35(4) (1997) 96–102.

[2] D.K. Hunter and I. Andonovic, Approaches to optical Internet packet switching, IEEE Communications Magazine 38(9) (2000) 116–122.

[3] S. Yao, B. Mukherjee and S. Dixit, Advances in photonic packet switching: An overview, IEEE Communications Magazine 38(2) (2000) 84–94.

[4] C. Quiao, Labeled optical burst switching for IP-over-WDM integration, IEEE Communication Magazine 38(9) (2000) 104–114.

[5] S. Bregni, G. Guerra and A. Pattavina, Optical switching of IP traffic using input buffered architectures, Optical Networks Magazine 3(6) (November/December 2002) 20–29.

[6] S. Bregni, A. Pattavina and G. Vegetti, Architectures and performance of AWG-based optical switching nodes for IP networks, IEEE Journal on Selected Areas in Communications 21(7) (September 2003) 1113–1121.

[7] K. Thompson, G.J. Miller and R. Wilder, Wide-area Internet traffic patterns and characteristics, IEEE Network 11(6) (1997) 10–23.

Stefano Bregni was born in Milano, Italy, in 1965. He received his *Dott.Ing.* degree in telecommunications engineering from Politecnico di Milano. Since 1991, he has been involved in SDH transmission systems testing and in network synchronization issues, with special regard to clock stability measurement. Since 1999, he has been an Assistant Professor at Politecnico di Milano, where he teaches telecommunications networks.

He has been Senior Member of IEEE since 1999. He served on ETSI and ITU-T committees on digital network synchronization. He is author of the book *Synchronization of Digital Telecommunications Networks*, published by Wiley. He is Distinguished Lecturer on this subject of the IEEE Communications Society. He was Vice-Chair of the Transmission, Access and Optical Systems Committee of the IEEE Communications Society. He is Co-Chair of the *Access and Home Networks Symposium of the IEEE Conference ICC 2004* (Paris, France). He served in the Technical Program Committees of several ICC and GLOBECOM Conferences.
E-mail: bregni@elet.polimi.it

Achille Pattavina received the degree in Electronic Engineering (Dr.Eng. degree) from University "La Sapienza" of Rome (Italy) in 1977. He was with the same University until 1991 when he moved to "Politecnico di Milano", Milan (Italy), where he is now Full Professor. He has been author of more than 100 papers in the area of Communications Networks published in international journals and conference proceedings. He has been author of the book *Switching Theory, Architectures and Performance in Broadband ATM Networks* (Wiley). He has been Editor for Switching Architecture Performance of the IEEE Transactions on Communications since 1994 and Editor-in-Chief of the *European Transactions on Telecommunications* since 2001. He is a Senior Member of the IEEE Communications Society. His current research interests are in the area of optical networks and wireless networks.
E-mail: pattavina@elet.polimi.it

Routing Algorithms in WDM Networks under Mixed Static and Dynamic Lambda-Traffic

Guido Maier

CoreCom

Via Colombo 81 - 20131 Milan, Italy (maier@corecom.it)

Achille Pattavina

Department of Electronics and Information (DEI), Politecnico di Milano

Piazza Leonardo da Vinci 32 - 20133 Milan, Italy (pattavina@elet.polimi.it)

Luigi Barbato

CoreCom (t_luibar@corecom.it)

Francesca Cecini

CoreCom (Francesca.Cecini@rsisistemi.it)

Mario Martinelli

CoreCom; DEI, Politecnico di Milano (martinelli@corecom.it)

Abstract

Dynamic traffic is becoming important in WDM networks. In the transition towards full dynamic traffic, WDM networks optimized for a specific set of static connections will most likely also be used to support on-demand lightpath provisioning. Our paper investigates the issue of routing of dynamic connections in WDM networks which are also loaded with high-priority protected static connections. By discrete-event simulation we compare various routing strategies in terms of blocking probability and we propose a new heuristic algorithm based on an occupancy cost function which takes several possible causes of blocking into account. The behavior of this algorithm is tested in well-known case-study mesh networks, with and without wavelength conversion. Moreover, Poissonian and non-Poissonian dynamic traffics are considered.

F. Cecini is now with RSI Sistemi, Altran Group, Corso Stati Uniti, 29 - 10129 Torino, Italy.

Keywords

Wavelength Division Multiplexing, dynamic traffic, routing algorithms

I. INTRODUCTION

THE past few years (up to 2001) witnessed the flourishing of telecommunication industry as one of the fastest-growing and most wide-spread phenomena in economy ever recorded. The core of that technological revolution has been Wavelength Division Multiplexing (WDM) optical networks, with their unique capability to offer the best solution to the large bandwidth demand on one side and to guarantee high quality of service and high reliability on the other side. Thanks to these features the area of applications of WDM networks drastically expanded from the original voice traffic transport to become a common high speed transport platform carrying data, video and voice. WDM protocol has been developed in order to operate in an integrated multi-protocol environment, serving Sonet, SDH, GMPLS, IP, etc.

When the general crisis hit the telecommunication market it found WDM networks at the dawn of a new evolution. Optical transport networks in the past have mainly been designed and operated as static systems: optical connections were used as long-distance trunks mostly to carry large aggregates of telephone traffic, usually serving only customers of the network operator itself. Traffic was thus highly predictable.

The scenario is much different today. Data traffic is going to overcome traditional telephone traffic in volume. The former is characterized by less regular flows than the latter, which are more and more independent of geographical distances. This change in traffic statistics is further amplified in the regional and metro area, where flows are less aggregated and more sensitive to traffic relations due to single, large bandwidth applications. Finally, many WDM network operators are beginning to offer the "lambda service" (i.e. optical connections for lease) and the carrier-of-carriers service to support the so-called "bandwidth-trading" business. This implies that their network infrastructure is no longer used solely from their own final customers, making a quota of connection demands no longer deterministically predictable. Lambda-service customers such as Internet Service Providers (ISP) may requests limited duration connections with a wide range of possible time-scales. Duration can be for example: several months to provide temporary connectivity to an ISP waiting for the installation of his new network; few days to cover some special event (e.g. the Soccer World Cup or the Olympic Games); few hours to

respond to daily traffic variations; tens of minutes to perform the backup of a large data-storage facility. The recent development of the Generalized Multi-Protocol Label Switching (GMPLS) technique seems to convey the idea that lightpaths in the future will have to be set up and torn down on a very short time-scale, even few seconds, perhaps paving the way to a possible optical packet-switching (or optical burst-switching) era.

All the facts mentioned above are pushing research in security, management and network design to re-focus its attention from the simple static Optical Transport Network (OTN) to Automatic Switched Optical Network (ASON). While OTN is already well-defined by the main standard bodies [1], [2], the new ASON model, able to set-up and release lightpaths on-demand based on on-line requests, is still undergoing an intense research and standardization activity. Despite economy may impose a slower evolution pace today, this change of paradigm from static to dynamic systems does not seem to be reversible for WDM networks. This justifies the birth of dynamic traffic as a new subject of research in optical networks. This area accounts today probably for less works than those published on static planning and it offers several open issues, such as efficient routing and traffic modelling.

In this paper we are going to address both the above issues, but in a new very particular context. Given the evolution from OTN to ASON as an actual process, this will surely occur gradually, in any case always preserving the investments of the network operators. In the transition the two paradigms of static and dynamic traffic have to co-exist and to be supported by the same WDM network infrastructure. Mostly likely, these infrastructure has been designed and configured in order to support a given original static traffic and it is then employed also to provide lightpaths on-demand. We have more precisely imagined the situation of a WDM network operator who has optimized his network according to a given set of protected static connections, adopting WDM path protection (in the two alternative versions, dedicated and shared) as survivability technique. He still wants to keep its original customers of the static connections, regarding these as high-priority traffic which should not be disrupted. In addition, the operator wants to do the best to accommodate as many lightpaths as possible to satisfy dynamic traffic requests by exploiting the same WDM network.

In this paper we propose and discuss a heuristic strategy for routing lightpaths for dynamic traffic that allows to increase the acceptance rate of dynamic connection requests (or, equivalently, to decrease the blocking probability) compared to other previously known routing al-

gorithms. Such a new algorithm is based on a global network function which provides an estimation of the available network resources according to different criteria. This allows to assign resources to the new lightpath so that chances of having congestion in the most critical spots of the network are kept as low as possible. We are going also to study the performance of the heuristic algorithm proposed under different types of dynamic traffic. The paper can be summarized in the following way. In Sec. II we present our network model and introduce our simulation method. In Sec. III we describe the models employed to simulate different types of traffic conditions of dynamic lightpath requests. Then, in Sec. IV, the new heuristic routing algorithm is proposed, showing its differences from other classical algorithms. In Sec. V we will show the results obtained by simulating case-study networks under dynamic traffic.

II. NETWORK MODEL

Let us describe in detail the network model we are referring to in this paper. The physical topology (either ring or mesh) is composed of WDM transmission links and WDM switching nodes connected according to a given graph. In the ring case the nodes of the physical topology correspond to Optical Add Drop Multiplexers (OADMs), while in the mesh network they represent Optical Cross Connects (OXCs). A WDM link represents a multifiber bidirectional cable: some fibers are used in one propagation direction, and some others (not necessarily the same number) in the opposite direction. All the fibers of the network carry the same number W of WDM channels, each transmitted on a different wavelength.

Traffic is carried by means of circuit-switched transport entities, optically routed on the basis of their wavelength. These entities are the lightpaths, each composed of a sequence of WDM channels connecting a source node to a destination node. In the present work we will consider two types of WDM networks, according to their wavelength conversion capability [3]:

- Virtual Wavelength Path (VWP) network: all the OXC are able to perform full wavelength conversion, i.e. an incoming optical signal having any wavelength can be converted to an outgoing optical signal having any possible transmission wavelength;
- Wavelength Path (WP) network: no wavelength conversion is allowed in the whole network. All the WDM channels composing a lightpath must be at the same wavelength (wavelength continuity constraint).

According to the scenario described in Sec. I, our network model must be able to represent

both static and dynamic optical connections. In this paper we will concentrate on the dynamic aspects of the problem (low-priority traffic), while all the static aspects have been solved by applying a tool that we developed and which is described in a previously published work [4]. We need only to briefly summarize the main features of this tool.

The set of requests for static connections (virtual topology) is known "a priori" and is fed to the design tool, together with a description of the physical characteristics of the network (topology, wavelength conversion, etc). Then a protection strategy is selected for the static optical connections. For simplicity all the static connections are protected with the same WDM-layer path-protection strategy. For each optical connection between a source and a destination node a working and a protection lightpath are set up. In order to guarantee the recovery of the connection in case of a single link failure, the two lightpaths must be in physical route diversity, i.e. they can not share any common link [1]. For brevity, we will say that a static request is satisfied by allocating resources in order to set-up a working/protection pair (w/p pair).

In the following of the paper we will consider two different implementations of path protection: 1:1 dedicated path-protection and mesh shared path-protection. In the dedicated case a spare WDM channel is rigidly allocated to a specific protection lightpath, so that two different w/p pairs never intersect. In shared path-protection two or more protection lightpaths belonging to different w/p pairs can share one or more WDM channels, provided that their respective working lightpaths are link-disjoint. It is well known that shared protection has the advantage of reducing the amount of resources used for protection and thus the global cost of the network.

Once the protection technique is selected the design tool proceeds to evaluate Routing, Fiber and Wavelength Assignment (RFWA) for all the requested w/p pairs, jointly solving the dimensioning problem of the physical topology. This is done by exploiting a heuristic optimization cycle which assumes the total number of fibers installed in the network as cost function. The cost optimization technique we have defined leads to very good suboptimal results, i.e. returning a network design (lightpath configuration and link capacity) very close to the one necessary and sufficient to support the given set of static connection requests (further details on the tool are given in [4]).

We recall that our aim is to model a realistic situation in which an operator wishes to employ

[1] A stronger condition of node-sharing prevention can be imposed if protection to node failures has to be enforced. This can be easily implemented with the same tool, but any further investigation of this aspect is left for future papers.

its WDM network initially designed for static traffic to offer lightpath provisioning service. As we are going to explain below, in such a scenario dynamic traffic has to use resources that are sparse in the network. This is much different from the typical situation considered so far in most of the papers on dynamic traffic in which dynamic connections occupy an initially empty network with the same amount of WDM channels in all the links. In our case available capacity varies from link to link because of the presence of static traffic and because of the optimization. Therefore the simulations of dynamic traffic we are going to describe in this paper are carried out on network systems obtained as results of the design tool. We believe this is the most faithful and simplest way to model the scenario we have assumed.

The network resources available to support dynamic traffic comprise: (a) WDM channels assigned to protection lightpaths (spare WDM channels); (b) unassigned WDM channels. The two path-protection implementations we have considered leave protection lightpaths idle in absence of failure. They can thus be used to host low-priority traffic (dynamic connections in our case) [2]. The low-priority qualification is due to the fact that when a failure occurs some low-priority connections have to be sacrificed to activate the protection lightpaths for all the static working paths hit by the failure. Since dynamic connections are low-priority traffic, they are satisfied by the setup of a single lightpath, without requiring any protection technique.

The second type of resources, i.e. unassigned channels, are present in the optimized WDM network as a result of cost minimization. Due to the finite granularity given by the fixed number of wavelengths per fiber W, some fibers are not fully occupied by static traffic. Unused capacity, which tends to increase with W, is usually small and widely scattered over the network links, making its exploitation very inefficient to route dynamic lightpaths. This is the main reason why we considered protection techniques for static traffic (1:1 and shared) that offered extra capacity to host dynamic connections.

To conclude the description of our model, let us explain how dynamic traffic is managed. Dynamic requests for optical connections arrive to the network control system from the upper transport protocol layers at random time. Each request is characterized by a source node, a destination node and a finite random duration of the connection. At each arrival of a new request the

[2]The exploitation of spare WDM channels would not be possible with other protection techniques such as, for instance, 1+1 dedicated path protection, in which the signal always propagates on both the working and spare lightpaths also in absence of failures.

control system applies a heuristic RFWA algorithm trying to setup the corresponding lightpath using the available WDM channels. No disruption of high-priority static working lightpaths is admitted to accommodate the new circuit, as well as no reconfiguration of already active dynamic lightpaths. If available resources are not sufficient the request is blocked and lost forever [5], [6]. Otherwise, the lightpath is setup by allocating the suitable sequence of WDM channels that will be released after the given connection duration. The chance of being able to accept a new connection is the blocking probability parameter. The main purpose of the simulation experiments we have performed is to compare the effectiveness of different RWFA algorithms (that will be presented in Sec. IV) in terms of average blocking probability. This parameter is estimated by measuring the ratio between the number of refused connection requests and the total number of requests received by the network during the simulation time. The duration of a simulation is chosen in order to be able to observe the evolution of the system long enough to reach statistical equilibrium under a constant dynamic traffic load. In such a condition the average value of the blocking probability changes very little in time.

Beside routing algorithm comparison, a second objective of this work is the evaluation by simulation of the sensitivity of the network to the type of dynamic traffic. In Sec. III we are giving details on our traffic modelling system and better explain what we mean by type of traffic.

III. Traffic modelling

As we stated in Sec. I, ASON *dynamic lambda traffic* service is a new emerging feature for WDM networks. To our knowledge no traffic measures of real cases have been yet reported in literature. Therefore a statistical description suitable for the future scenario is very difficult to predict [7]. Because of this lack of actual traffic characterizations, traffic models developed in the traditional circuit-switching theory for telephone networks are usually employed [8].

In our work we assume that each node of the WDM network (i.e. each OXC) is a generator of dynamic optical connection requests, having the node itself as source and a destination which is randomly chosen among the other nodes of the network with equal probability. This last assumption is appropriate in the case all the nodes of the network have the same importance. The analysis of particular situations of polarized destination choice (e.g. hub-like traffic) is left for a future upgrade.

In order to be able to represent a wide range of traffic situations, we adopted the “moment-

matching" technique, commonly used in tele-traffic theory to study overflow streams in telephone networks [9], [10]. The traffic offered by a generator to the network is defined by assigning two parameters which corresponds to the first two moments of x, the random number of active connections in the system having the node as source in statistical equilibrium. These two parameters, per each network node, are:

- *offered load* $A_0 = \mathrm{E}[x]$, is the average number of active connections;
- *peakedness factor*, also called Variance to Mean Ratio, VMR $= \sigma_x^2/\mathrm{E}[x]$, where σ_x^2 is the variance of x.

The "moment-matching" technique then allows to model a generator choosing any *equivalent process* that yields the first two moments of the offered load defined above, independently of the real nature of the generator itself.

The earliest and most widely accepted application of moment-matching [11], [12] exploits the Bernoulli-Poisson-Pascal processes as equivalent processes. Such birth-death processes are in general characterized by the mean inter-arrival time $1/\lambda_x$ between two consecutive connection requests and the mean duration $1/\mu_x$ of each single connection. The holding time of each single connection is an exponentially distributed random variable whose mean value $\mu_x = \mu$ is independent of the state of the generator and of the arrival process. The inter-arrival generation process is instead more complicated and it varies according to the specific traffic type.

Most of the dynamic traffic models in the literature have assumed a Poisson traffic type (*regular* traffic), that in the "moment-matching" technique is obtained by choosing VMR $= 1$. In this case inter-arrival time is exponentially distributed with a mean value $\lambda_x = \lambda$ that is constant and independent of the state of the generator, like the holding time. For this traffic type a simple relation holds between the average load and the two generator parameters: $A_0 = \lambda/\mu$.

Poisson process, however, may not be representative of the input traffic in a wide area optical network [7], [13], [14]. The Bernoulli and Pascal processes can be used to model *smooth* and *peaked* traffic types, respectively. While in the Poisson process births are quite regularly distributed in time, Pascal traffic is characterized by burst periods with high density of births and Bernoulli traffic, on the opposite, is characterized by burst periods of "silence". In the "moment-matching" technique the peakedness factor determines which type of process has to be used to model the generators: VMR < 1 and VMR > 1 correspond to the Bernoulli and Pascal processes, respectively.

The exact relations between λ_x, A_0 and VMR in the Bernoulli and Pascal cases [7] are quite complicated and will be omitted here for brevity. Qualitatively, λ_x increases with x in the *smooth* case and decreases in the *peaked* case.

In this work we have developed most of simulation experiments by assuming Poisson traffic-type. In section V-C however we are going to show some results obtained with Pascal and Bernoulli traffic types.

IV. Heuristic Routing algorithms

At the arrival of a new dynamic connection request at a given time, the control system of the network must solve a RFWA problem. This consists in the identification of the route (sequence of links) from the source to the destination and in the selection of a WDM channel in a particular fiber of each link to be allocated to the new connection.

Several techniques have been proposed to perform RFWA. In principle, techniques based on mathematical programming could be applied, but these methods require very high computational efforts. Moreover, the formal definition of an objective function to be minimized which is directly related to blocking probability is not an easy task, as we will explain later on. Therefore most of the works in literature regarding dynamic WDM networks propose heuristic methods to solve RFWA [15]. Some of these methods are based on disjoint solution first of routing and then of fiber and wavelength assignment. In these cases routing can be constrained: only a limited set of pre-computed routes for each pair of nodes in the network is considered for possible routing. This simplifies the RWFA operation but limits the routing options, potentially affecting blocking probability.

In our work we adopted a heuristic method which jointly solves routing, fiber and wavelength assignment without imposing constraints on the viable routes (unconstrained routing [16]). It is based on the *multifiber layered graph* (MLG) representation of the network, also used in our static-network design tool [4] and derived from the layered graph method, well-known from many published papers (e.g. Ref. [17] for dynamic RWA). Each WDM channel of the network is represented by an arc of the MLG, and each node by a set of image nodes. In order to jointly solve RFWA, all the arcs of the MLG are assigned proper weights prior to setup the new lightpath. Then the Dijkstra algorithm is run on the MLG. This allows to find the least-total-weight route between the image node corresponding to the source OXC and the image node

corresponding to the destination OXC: the MLG arcs belonging to this route correspond to the WDM channels that must be allocated to the new lightpath.

The key point of our heuristic approach is weight assignment to the MLG arcs. First of all an infinite (actually, very large) weight is given to arcs corresponding to unavailable WDM channels. We recall that WDM channels available for dynamic traffic at a given instant are those left idle by lightpaths that are active at that instant (either static working lightpaths and dynamic lightpaths) plus those allocated to static protection lightpaths. Secondly the MLG weight system is used to support a wide range of RFWA heuristic criteria. In our approach specific criteria for routing, for fiber assignment and for wavelength assignment can be combined together with a given priority order. To do this each MLG arc is actually assigned an array of weights instead of a single scalar weight. Each weight of the array is determined by a specific criterion. We have modified the Dijkstra algorithm so that the criteria can be applied in a prioritized sequence. Each time several alternative MLG routes have an equal total weight according to the primary criterion, they are compared according to the secondary criterion, and so on.

The focus of this paper is routing algorithm evaluation under dynamic traffic. Therefore, in all the simulations we have performed we have always assigned the highest RFWA priority to routing; fiber and wavelength assignment follow in order of decreasing priority. Moreover, with all the tested routing criteria, as we will explain further below, we always used the same criterion both for fiber and for wavelength assignment. This is the FIRST FIT criterion [5], according to which weights are assigned to MLG arcs so that, during RFWA, the first idle wavelength (fiber) is selected after having sorted all the wavelengths (fibers) in each link according to an *a priori* fixed order. FIRST FIT is simple, since it does not require information about the instantaneous network state; however it proved to be a good criterion as for blocking performance.

A. Simple routing algorithms

As mentioned above, the main objective of this work is to compare different routing algorithms in terms of their effectiveness in achieving low blocking probability of the dynamic connections. The high flexibility of our network model allowed us to test several algorithms, starting from "classical" solutions and then proceeding to newly defined advanced algorithms. The two best-known and simplest routing algorithms for dynamic traffic in a WDM network are the Shortest Path Routing (SPR) and the Least Loaded Routing (LLR).

The first one routes the lightpath on the minimum distance available path between source and destination: distance is evaluated as the number of hops (WDM links) crossed by the lightpath. It is very easily implemented by setting to 1 the weights of all the available WDM channels. This routing algorithm is *static* since the corresponding weights do not depend on the state of the network.

LLR tries instead to route the new lightpath on a path which carries the lowest possible amount of traffic generated by already active connections at the time of connection establishment. It obviously requires a knowledge of the network state: it can therefore be classified as an *adaptive* routing algorithm. To perform LLR WDM channels are weighted by the so-called *link congestion parameter*: if a given channel belongs to link j, then it is assigned a weight b_j equal to the number of busy WDM channels on that link. The algorithm allocates the new lightpath on the route having the least possible *route congestion parameter*. This latter variable is equal to the maximum link congestion parameter among all the links crossed by the route itself .

It should be noted that each the above algorithms is effective in reducing blocking probability of dynamic connections on a single different front. SPR tends to minimize the amount of resources that a new connection is going to subtract from the pool of available WDM channels of the network. LLR tends to uniformly distribute the load over the links of network. A very interesting option offered by our network model is that more criteria concerning the same aspect of RFWA can be applied in sequence, taking advantage of the best heuristic quality of each one. We applied this to routing, creating a new algorithm from the combination of LLR and SPR in a prioritized sequence, named LLR/SPR. The highest priority is given to LLR; when two routes are equal according to the least-loaded criterion, the shortest one is selected according to SPR. We can expect that cascading LLR and SPR can improve blocking probability compared to both the single algorithms.

Several other algorithms have been proposed in literature [15]. We consider only the ones presented above for brevity, knowing however that they are the most frequently used.

B. An advanced heuristic routing algorithm

Most of the studies on WDM dynamic traffic presented so far in literature propose routing algorithms which take present or past network states into account [18], [5], [19], [20]. Recent works however, suggested a different approach which focus on the prediction of the future

network state based on the network history. In [21] this task is accomplished by a method requiring quite a complex mathematical formulation. In this paper we try to accomplish the same task following an alternative heuristic approach, by seeking a parameter which allows to evaluate the impact of a present routing choice of an optical path on the following connection requests.

The phenomena which can potentially combine to contribute to increase blocking probability may be identified as: (a) general shortage of free WDM channels; (b) exhaustion of available resources on some particular link cut-set, leaving parts of the network disconnected; (c) saturation of all input and/or output links of a particular node. Finding an exact mathematical formulation to represent these causes is not so easy. The identification of critical cut-sets in the network, for instance, is a complex problem that cannot be exactly solved with polynomial algorithms [22]. So we have defined a global network function which can be used as a heuristic measurement of the incidence of all three combined causes. This function $\varphi(t)$, which we name Occupancy Cost Function (OCF), is defined as follows:

$$\varphi(t) = \sum_{j=1}^{L} \frac{1}{f(d_j)} \cdot \frac{b_j}{WF_j} \cdot \frac{\hat{n}_j(t)}{\hat{R(t)}\bar{S}} \tag{1}$$

where symbols have the following meanings:

- L: total number of links;
- $f(d_j)$: value of the probability density function of the network link lengths corresponding to the length d_j of link j;
- b_j: link congestion parameter of link j;
- W: number of wavelengths per fiber;
- F_j: number of fibers installed on link j;
- t: time at which the network is observed, assuming $t = 0$ as the instant from which the network, loaded only with the static lightpaths, is used to carry dynamic traffic; $t = 0$ is also the beginning of all the network simulations;
- $\hat{n}_j(t)$: number of lightpaths that were routed on link j from $t = 0$;
- $\hat{R(t)}$: total number of dynamic connections setup in the network from $t = 0$;
- $\bar{S}$: average shortest path (in number of hops) taking into account all the couples of nodes of the network

The normalized link congestion parameter $\frac{b_j}{WF_j}$ measures an instantaneous utilization state.

The term $\frac{\hat{n}_j(t)}{\hat{R(t)}\bar{S}}$ estimates which share of all the WDM channels of the network allocated to dynamic traffic until time t has been supported by link j. This term is a measure of the link utilization rate derived from the past history. The product of these two factors is a sort of "historical" link congestion parameter. Since the network model we have assumed is a pure loss system, it will reach statistical equilibrium when kept under constant mean dynamic-traffic load for a sufficiently long period. In such a condition the third factor of Eq. 1 tends to become constant in time for each link.

The factor $1/f(d_j)$ may appear unexpected at a first glance, since link lengths seems irrelevant when blocking is the concern. This factor introduces a cost of the utilization of a link of a given length which is inversely proportional to the chances of finding other links of the same length in the network. As it will be shown by examples in Sec. V, in many realistic topologies network-cut-sets are often composed of few links having similar length: in this cases a dramatic increase of OCF due a $f(d_j)$ contribution denotes high chances of cut-set exhaustion.

The most important aspect of the OCF function is that it has been defined in order to be used in an operative way for dynamic lightpath routing. For this purpose, the variation of the function due to a new lightpath setup, more than $\varphi(t)$ itself, is of interest. Based on $\varphi(t)$ variation, a new interesting routing algorithm can be defined, as explained in the following. Let t be the arrival time of a new connection request and t^+ the time after the new lightpath has been set up, assuming that the request is not blocked. Furthermore be γ the set of links crossed by new connection (i.e. its route). The value $\varphi(t^+)$ of the function after the new lightpath setup is:

$$\begin{aligned}
\varphi(t^+) &= \sum_{\forall j \notin \gamma} \frac{1}{f(d_j)} \cdot \frac{b_j}{WF_j} \cdot \frac{\hat{n}_j(t)}{\hat{R(t)}\bar{S}} + \\
&\quad + \sum_{\forall j \in \gamma} \frac{1}{f(d_j)} \cdot \frac{(b_j+1)[\hat{n}_j(t)+1]}{WF_j[\hat{R(t)}+1]\bar{S}} \\
&= \sum_{\forall j \notin \gamma} \frac{1}{f(d_j)} \cdot \frac{b_j}{WF_j} \cdot \frac{\hat{n}_j(t)}{\hat{R(t)}\bar{S}} + \\
&\quad + \sum_{\forall j \in \gamma} \frac{1}{f(d_j)} \cdot \frac{b_j \hat{n}_j(t)}{WF_j[\hat{R(t)}+1]\bar{S}} + \\
&\quad + \sum_{\forall j \in \gamma} \frac{1}{f(d_j)} \cdot \frac{[b_j + \hat{n}_j(t)+1]}{WF_j[\hat{R(t)}+1]\bar{S}}
\end{aligned}$$

Let us consider the equation above after an initial transient time τ. In realistic traffic con-

ditions we can assume that, for $t > \tau$, $\hat{R(t)} \gg 1$ and therefore $\hat{R(t)} + 1 \approx \hat{R(t)}$. The first two terms can be grouped together, thus obtaining $\varphi(t)$ again. In conclusion the increment of function $\varphi(t)$ due to a new lightpath setup can be written as

$$\Delta\varphi(t) = \varphi(t^+) - \varphi(t) = \sum_{\forall j \in \gamma} \frac{b_j + \hat{n}_j(t) + 1}{f(d_j) W F_j \hat{R(t)} \bar{S}} \tag{2}$$

Equation 2 gives us the hint and the opportunity to define a new routing algorithm whose optimality criterion is the minimization of the increment of the OCF itself. This algorithm can easily be implemented on our network model by assigning the following weight to each link j:

$$w_r(j,t) = \frac{1}{f(d_j)F_j} \cdot [b_j + \hat{n}_j(t) + 1]$$

The other normalization factors $W\hat{R(t)}\bar{S}$ can be omitted since they are common to all the links. It should be noted that towards the beginning of the simulations the algorithm behaves similarly to LLR, due to the presence of b_j, while as the system approaches statistical equilibrium $\hat{n}_j(t)$ becomes more and more relevant.

The function $\varphi(t)$ is obviously far from being a blocking probability. However, when $\varphi(t)$ is evaluated for a network in statistical equilibrium conditions and for different values of average load, its behavior as a function of the load is extremely similar to the corresponding behavior of the blocking probability. The exact comprehension of the relation between OCF and Π_p would require the definition of a suitable analytical model. Several analytical models for evaluating blocking probability of a WDM network under dynamic traffic have been published. Several of these models (e.g. Ref. [23]) assume fixed routing (one single possible path between each source and destination). An early model considering adaptive routing has been proposed for classical circuit-switched networks (thus also VWP WDM networks) in Ref. [24]: however in this case routing is constrained (also, the topology must be fully meshed). The most advanced analytical models for WDM networks seems today those based on link correlation. A first model was proposed in Ref. [19] for fixed routing, then it was extended in Ref. [25] to multifiber WDM networks; in Refs. [16], [26] the correlation model was applied to adaptive routing, but again considering a limited predetermined set of routes for each node pair. All the models mentioned so far require to (iteratively) solve a set of Erlang fixed-point equations: constraining the routing choices is needed in order to limit the complexity of this solution. In

conclusion, none of the theoretical approaches mentioned above is directly applicable to evaluate blocking probability when routing is based on OCF. In fact this new routing criterion is unconstrained, adaptive, multifiber and takes the link length distribution into account; besides, the presence of the high-priority static traffic further increases the complexity. Thus, the present paper is based on heuristic considerations supported by simulation results: an analytical model, currently under study, will be proposed in a future work in which we wish to provide theoretical justifications to the present results.

In the next section we will compare all the routing algorithms described in this section by simulating their performance on various case-study networks.

V. Case-study results analysis

The results we are going to present in this paper were obtained by a C++ discrete-event network simulator implemented according to the network model described in the previous sections and which has been integrated with the WDM network design tool.

Three case-study networks have been considered. The first two, namely the USA National-Science-Foundation Network (NSFNET) and the European Optical Network (EON), shown in Fig. 1 and Fig. 2 respectively, have been designed, optimized (minimizing the total number of fibers) and preloaded with static traffic using the tool described in [4], assuming $W = 32$ wavelengths per fiber. Data regarding physical and virtual topology for these two realistic examples were taken from [27] and [28].

The third case-study network (Fig. 3) has been created *ad hoc* as an example of network having cut-sets composed of few links of similar length. Though it does not physically exists, its topology is realistic, since it corresponds to a global infrastructure composed of three metro-area networks interconnected by a wide-area network. For this reason we named it Wide Plus Metro Area Network (WPMNET). WPMNET has only three types of links: short (MAN links), medium and long (WAN links). The number of short links is much higher than that of medium and long links, since MAN networks usually have a larger connectivity. This particular length distribution allows us to highlight the importance of introducing the cost factor $1/f(d_j)$ in the advanced routing algorithm we have proposed in this paper. Thanks to this factor the routing algorithm will try to avoid routes such as the one shown in Fig. 3 from S to D, which unnecessarily overload long links and can bring to a quick congestion of critical cut-sets.

The particular structure of WPMNET is well shown in Fig. 4, in which the probability distribution function of the length of the links is plotted, normalizing the length to the longest link in the network. By comparison, also the probability distribution functions of the other two networks are reported.

We have performed three different sets of simulation experiments on the three chosen case-study networks, in order to analyze three different aspects of the ASON scenario: (a) effects of the protection mechanism adopted for the static connections; (b) effectiveness of the routing algorithms; (c) effects of the type of dynamic traffic. In all the cases network blocking probability Π_p has been considered as the basic performance parameter. Curves are plotted using the average offered traffic (in Erlang) *per generator* A_0 as x-axis variable. In simulations concerning WPMNET no static traffic has been pre-loaded on the network; all the links have 5 fibers with $W = 5$ wavelengths per fiber. Poisson traffic has been always used, except for the experiments presented in the last subsection, and all the OXCs have been assumed to be active independent dynamic traffic generators. Finally, all the results displayed are obtained when the network system has reached the statistical equilibrium under constant average load.

A. *Dedicated and shared protection of high-priority traffic*

Fig. 5 shows the blocking probability of dynamic connections for NSFNET and EON (VWP scenario) in the two cases of dedicated and shared path-protection of the high-priority static connections. LLR/SPR has been used as routing algorithm.

Curves clearly show in both cases that blocking probability is higher with shared path protection. This is quite expected. The same number of static w/p pairs are supported by the optimized networks in the dedicated and shared cases. If the number of WDM channels occupied by the working lightpaths is roughly the same; protected lightpaths require far less channels in the latter case, due to sharing of several channels by more w/p pairs. As a consequence, in a shared-protected network far less WDM channels are available to host dynamic traffic. We can conclude that for a network operator the shared protection strategy is beneficial to save on initial installation costs of the network, since the total number of fibers to support a given static demand is less than in the dedicated protection case. This saving however has the drawback of limiting the operator ability of satisfying future lambda-service customers.

As stated above, blocking events can be caused either by shortage of WDM available channels

or by cut-set congestion. A lower bound of blocking probability which takes only the first cause into account is given by the following simple analytical method. The total number C of WDM channels that are initially available (at $t = 0$) for dynamic traffic is a known quantity that results from static network optimization. The total number of channels allocated to dynamic connections at any time is obviously bounded by C. All the N network nodes are identical traffic generators that are demanding connections randomly selecting destinations between all the other nodes. In the ideal case we can assume that all these connections are routed on the shortest path between source and destination in order to occupy the minimum possible amount of resources. The average accepted traffic per node is A_s, which is related to A_0 by: $A_s = (1 - \Pi_p)A_0$. Given this scenario, the minimum possible number of WDM channels allocated to dynamic lightpaths having a certain node as source is $A_s \cdot \bar{S}$, where we recall that $\bar{S}$ is the average shortest path (in number of hops) of the network. After simple math we can finally write

$$\Pi_p \geq 1 - \frac{C}{N \cdot A_0 \cdot \bar{S}} \tag{3}$$

In Fig. 5 the curves derived form Eq. 3 have been plotted for the two networks in the two dedicated and shared cases. By comparing these curves with those obtained by the simulation we can conclude that cut-set congestion is a strong cause of blocking in networks pre-loaded with static traffic.

After the comparison presented in this subsection, only dedicated path-protection will be considered for static traffic for the rest of the paper.

B. Routing algorithm comparison

Fig. 6 allows to compare the performance of the various routing algorithms considered in the paper and for the three case-study networks, in both WP and VWP wavelength-conversion scenario. As expected from the lower-bound analysis reported above, the worse routing algorithm is SPR, since it is not effective in avoiding congestion. LLR alone is a little better performing, while fair improvements are obtained when the two algorithms are combined together, giving LLR a higher priority. The better behavior of LLR compared to SPR with dynamic traffic is well known [5], [6], [16], [20] [3]. Some papers [16], [26] proposed the combined LLR/SPR

[3] Ref. [26] is an exception, showing that SPR is better than LLR. This is probably due to the fact that routes are constrained to the shortest-path length + one hop.

algorithm (or similar versions) showing their advantages.

The new algorithm we proposed, based on OCF increment minimization is the best performing algorithm, especially when the cost associated to the link length probability density function $f(d_j)$ is taken into account.

We have further analyzed this latter aspect by comparing the blocking probabilities obtained by applying the OCF based routing algorithm without and with taking the cost factor associated to $f(d_j)$ into account. The ratios displayed in Fig. 7 measures the penalty that must be paid in terms of blocking probability when the distribution of the link lengths is not considered in routing. As expected, this penalty gets worse for networks which contain cut-sets composed of few links of similar length, while is less severe for networks that have a rather uniform link length distribution (e.g. the NSFNET).

Another interesting aspect we wished to better understand is the relation between the occupancy cost function $\varphi(t)$ and the blocking probability when the algorithm based on OFC is adopted for routing. Fig. 8 compares Π_p and $\varphi(t_e)$ evaluated at a time t_e in which the network is in equilibrium. The graph clearly shows that, even if OFC assumes a completely different set of values, its dependence on A_0 is very similar to that of the blocking probability. Π_p and $\varphi(t_e)$ result to be almost proportional, especially for small loads.

As a final comment we shall add that all the graphs displayed in this subsection show that wavelength conversion is quite important for WDM dynamic traffic (while it is not so in the static case), especially for low traffic loads, for which the presence of the converters can reduce blocking probability of one order of magnitude.

C. Traffic type comparison

We have simulated the behavior of the NFSNET under the three types of dynamic traffic (Bernoulli, Poisson and Pascal) described in Sec. III. All these simulations are carried out adopting the OCF based algorithm for routing. In the case of Bernoulli and Pascal the chosen values of the peakedness ratio VMR are 0.5 and 5 respectively.

In Fig. 9 the results of the simulations are displayed in three cases: (a) with a single active generator; (b) with 5 active generators; (c) when all the 14 generators are active. Results with one single active generator (which could model the real situation of a large hub node connected to small customers) indicates that blocking probability is strongly dependent on the type of

traffic. Most of the blocking events are probably due to the saturation of the links connected to the active generator. With a higher number of independent generators the blocking probability becomes more and more insensitive to the type of traffic. This behavior is probably due to the fact that the network links aggregate uncorrelated traffic from many different sources which are characterized by "classical" random processes such as those we employed for our simulations.

It should be observed that though the global blocking probability tends to be the same, the dynamic behavior of a network loaded with the three types of traffic is different, even with a large number of active generators. In fact with Bernoulli process blocking events are regularly distributed in time, with Poisson process losses occur less regularly, while with Pascal process there are periods in which the system accepts all the connection requests and some others in which loss is very high (further results regarding this aspect are not reported here for brevity).

VI. Conclusions

We have considered the future scenario of WDM networks designed and optimized for static traffic and then employed to provide lambda-connection service on demand. This leads dynamic low-priority optical connections to co-exist together with static high-priority connections on the same optical network. This paper simulates and compares the performance of several well-known algorithms for dynamic lightpath routing in such a network environment. We have proposed and tested an advanced routing algorithm able to reduce blocking of future connections in different ways. By means of dynamic traffic simulations we have shown the performance improvement of the advanced routing algorithm compared to other well-known classical solutions. Our study also includes an evaluation of the behavior of WDM networks under non-Poissonian traffic type, as well as a comparison between cases in which high-priority static connections are protected by shared and dedicated WDM path-protection, respectively.

References

[1] ITU-T Intern. Telecom. Union Telecom. Standard. Sector, *Architecture of Optical Transport Networks*, Number G.872. 1999.

[2] ITU-T Intern. Telecom. Union Telecom. Standard. Sector, *Network Node Interface for the Optical Transport Network (OTN)*, Number G.709. 2001.

[3] N. Wauters and P. M. Deemester, "Design of the optical path layer in multiwavelength cross-connected networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, pp. 881–891, june 1996.

[4] A. Dacomo, S. De Patre, G. Maier, A. Pattavina, and M. Martinelli, "Design of static resilient WDM mesh networks with multiple heuristic criteria," in *Proceedings, IEEE INFOCOM*, June 2002.

[5] Ahmed Mokhtar and Murat Azizoğlu, "Adaptive Wavelength Routing in All-Optical Networks," *IEEE/ACM Transaction on Networking*, vol. 6, pp. 197–206, apr 1998.

[6] Ezhan Karasan and Ender Ayanoglu, "Effects on wavelength routing and selection algorithms on wavelength conversion gain in WDM optical networks," *Networking, IEEE/ACM Transactions on*, vol. 6, no. 2, pp. 186 –196, Apr. 1998.

[7] S. Subramaniam, A. K. Somani, and M. Azizoğlu, "A Performance Model for Wavelength Conversion with Non-Poisson Traffic," in *Proceedings, IEEE INFOCOM'97*, 1997, vol. 2, pp. 499–506.

[8] J. Späth and S. Bodamer, "Performance Evaluation of Photonic Networks under Dynamic Traffic Conditions," in *Proceedings of the 2nd IFIP TC6 Conference on optical network design and modelling*, Feb. 1998, vol. Post Conference Paper, pp. 15–19.

[9] H. Heffes and D Lucantoni, "A Markov- modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *jsac*, vol. 4, no. 6, pp. 856–868, Sep 1986.

[10] A. Kuczura and D Bajaj, "A method of moments for analysis of switched communication network's performance," *toc*, vol. 25, no. 2, pp. 185–193, Feb 1977.

[11] R. I. Wilkinson, "Theories for toll traffic engineering in the U.S.A," *Bell Systems Technical Journal*, vol. 35, pp. 421–514, Mar 1956.

[12] L. E. N Delbrouck, "A unified approximate evaluation of congestion functions for smooth and peky traffics," *toc*, vol. 29, no. 2, pp. 85–91, Feb 1981.

[13] G. Maier, M. Martinelli, A. Pattavina, and M. Scappini, "Performance of WDM rings with partial and sparse wavelength conversion under general dynamic traffic," *European Transaction on Telecommunications*, vol. 11, no. 1, pp. 91–98, Jan.-Feb. 2000.

[14] G. Maier, M. Martinelli, A. Pattavina, and M. Scappini, "Performance of WDM rings with wavelength conversion under non-Poisson traffic," in *Proceedings, IEEE International Conference on Communications*, June 1999, p. s51.4.

[15] H. Zang, J. P.Jue, and B. Mukherjee, "A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks," *SPIE Optical Network Magazine*, pp. 47–60, jan 2000.

[16] Ling Li and A.K. Somani, "Dynamic wavelength routing using congestion and neighborhood information ," in *Networking, IEEE/ACM Transactions on*, Oct. 1999, vol. 7(5), pp. 779 –786.

[17] Chien Chen and Subrata Banerjee, "A New Model for Optimal Rouiting in All-Optical Networks with Scalable Number of Wavelength Converters," in *Proceedings,IEEE INFOCOM '96*, 1996, pp. 164–171.

[18] A. Birman, "Computing Approximate Blocking Probabilities for a Class of All-Optical Networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 5, pp. 852–857, June 1996.

[19] S. Subramaniam, M. Azizoğlu, and A. K. Somani, "Connectivity and Sparse Wavelength Conversion in Wavelength-Routing Networks," in *Proceedings,IEEE INFOCOM '96*, 1996.

[20] Hiroaki Harai, Masayuki Murata, and Hideo Miyahara, "Performance Analysis of Wavelength Assignment Policies in All-Optical Metworks with Limited-Range Wavelength Conversion," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 1051–1060, sept 1998.

[21] E. Hyytiä and J. Virtamo, "Dynamic Routing and Wavelenght Assignment Using First Policy Iteration," in *IEEE Symposium on Computers and Communication,2000*, 2000, pp. 146–151.

[22] Stefano Baroni, *Routing and wavelenght allocation in WDM optical networks*, Ph.D. thesis, University College of London, Departement of Electronic and Electrical Engineering, May 1998.

[23] T Tripathi and K.N Sivarajan, "Computing approximate blocking probabilities in wavelength routed all-optical networks with limited-range wavelength conversion," in *Proceedings, IEEE INFOCOM '99*, 1999, vol. 1, pp. 329–336.

[24] S.P Chung, A Kashper, and K.W Ross, "Computing approximate blocking probabilities for large loss networks with state-dependent routing," *IEEE/ACM Transactions on Networking*, vol. 1, no. 1, pp. 105–115, Feb 1993.

[25] L Li and A.K Somani, "A new analytical model for multifiber WDM networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, pp. 2138–2145, Oct 2000.

[26] Ching-Fang Hsu, Te-Lung Liu, and Nen-Fu Huang, "Performance of adaptive routing strategies in wavelength-routed networks," in *Performance, Computing, and Communications, 2001 IEEE International Conference on*, April 2001, pp. 163 –170.

[27] Y. Miyao and H. Saito, "Optimal design and evaluation of survivable WDM transport networks," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 1190–1198, sept 1999.

[28] A. Fumagalli, I. Cerutti, M. Tacca, F. Masetti, R. Jagannathan, and S. Alagar, "Survivable networks based on optimal routing and WDM self-heling rings," *Proceedings, IEEE INFOCOM '99*, vol. 2, pp. 726–733, 1999.

LIST OF FIGURES

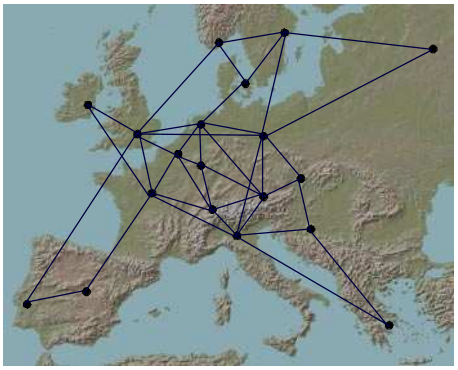


Figure 1. European Optical Network (EON).

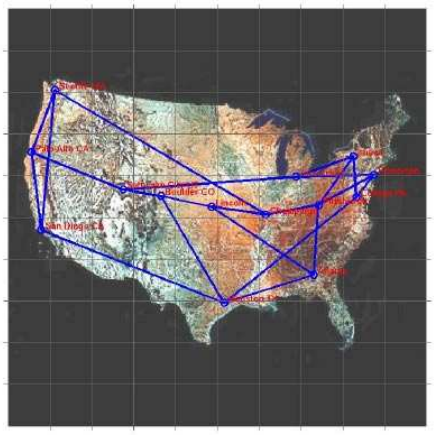


Figure 2. National Science Foundation Network (NSFNET).

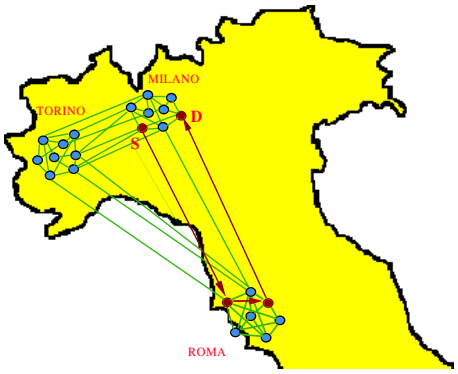


Figure 3. Wide Plus Metro Area Network (WPMNET). S-D is an example of an undesirable route.

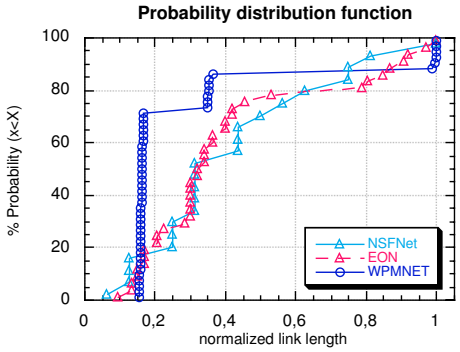


Figure 4. Statistical distance distribution for NSFNET, EON and WPMNET.

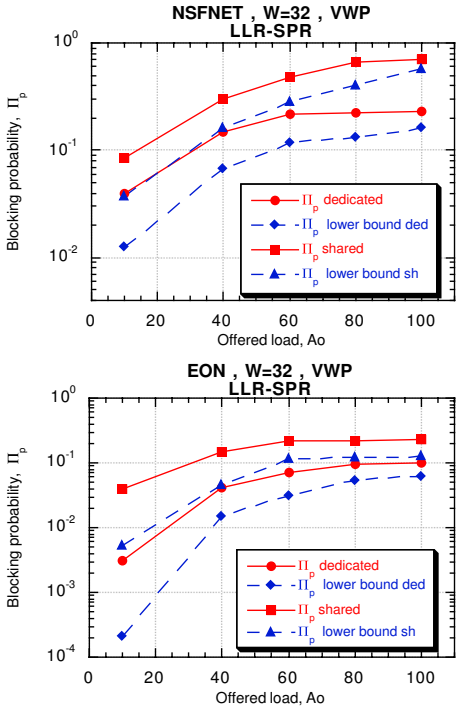
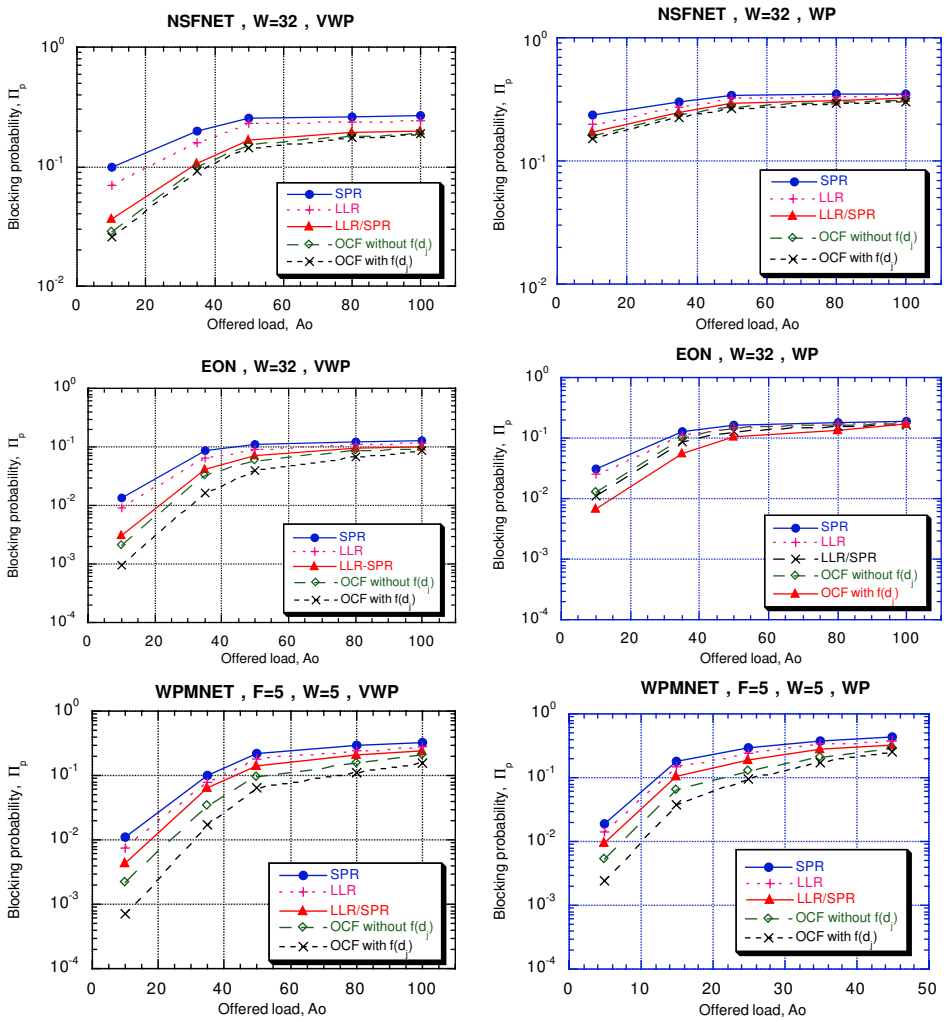


Figure 5. Blocking probability of dynamic connections in NSFNET and EON in the two cases of dedicated and shared path-protection of the static traffic. Lower-bound curves are also plotted



(a) (b)

Figure 6. Blocking probability of NSFNET, EON and WPMNET for different routing algorithms, with (a) and without (b) wavelength conversion capability.

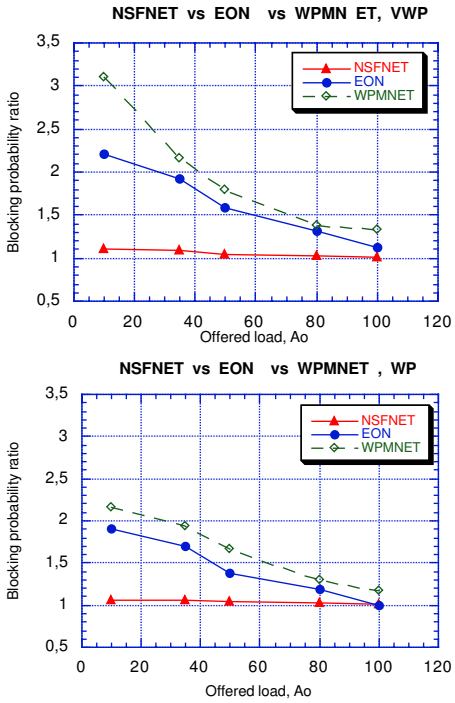


Figure 7. Ratio between blocking probability evaluated not including and including the cost factor associated to $f(d_j)$ in the OCF-based routing algorithm.

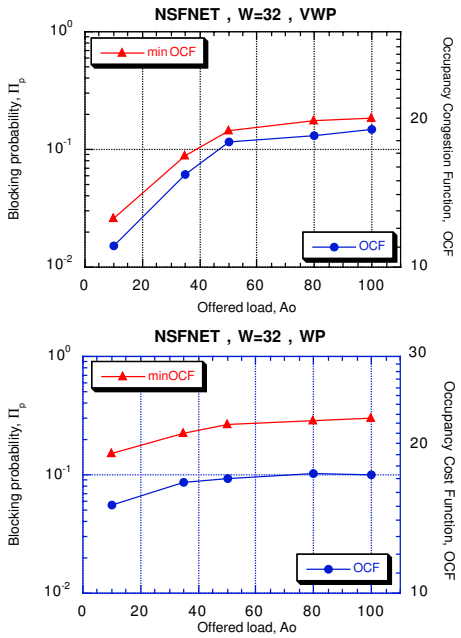


Figure 8. Blocking probability obtained by the OCF-based routing algorithm and values of the OCF itself as functions of the average offered traffic.

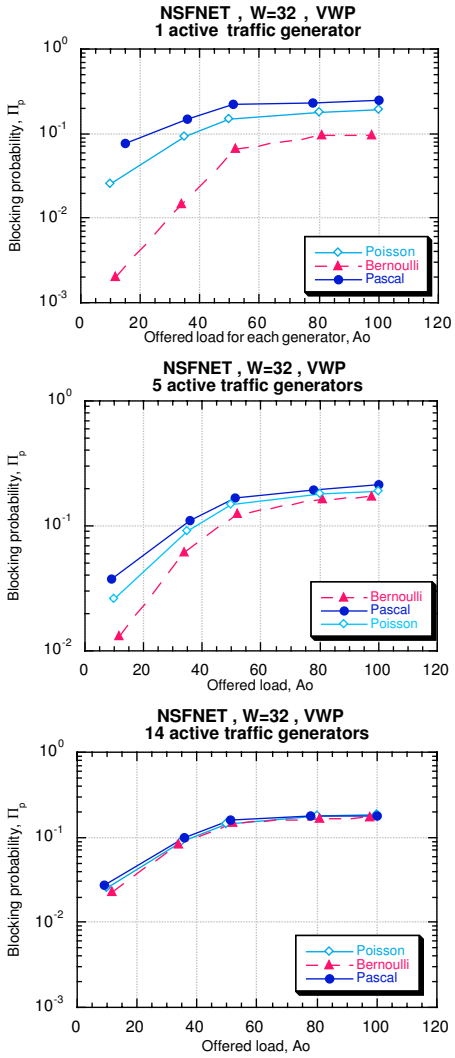


Figure 9. Blocking probability of VWP NSFNET under Bernoulli, Poisson and Pascal dynamic traffic type, with 1, 5 and 14 active generators

Chapter 3

Politecnico di Torino Research Unit

A.Bianco,V.Curri, J.Finochietto, E.Leonardi, M.Mellia, F.Neri, C.Piglione,
Dipartimento di Elettronica – Politecnico di Torino
Corso Duca degli Abruzzi, 24 – I-10129 Torino, Italy
{lastname}@mail.tlc.polito.it

Abstract

During the second year of the project, the research unit of Politecnico di Torino focused its research efforts in two directions:

- **Impact of physical layer constraints on the Routing and Wavelength assignment problem**: we consider dynamically reconfigurable wavelength routed networks in which lightpaths carrying IP traffic are on demand established. We face the Routing and Wavelength Assignment problem considering as constraints the physical impairments that arise in all-optical wavelength routed networks.
- **Resource Allocation in MANs**: we discuss capability of optical Metros, and introduce alternative architectural designs based on rings, which can be seen as reconfigurable networks. Reconfiguration algorithms are devised as well, to adapt network configuration to traffic demands to optimize performance.

As regards the first item, belonging to the second workpackage of this project, we face the Routing and Wavelength Assignment problem considering as constraints the physical impairments that arise in all-optical wavelength routed networks. In particular, we study the impact of the physical layer when establishing a lightpath in transparent optical network. Because no signal transformation and regeneration at intermediate nodes occurs, noise and signal distortions due to non-ideal transmission devices are accumulated along the physical path, and they degrade the quality of the received signal. We propose a simple yet accurate model for the physical layer which consider

both static and dynamic impairments, i.e., nonlinear effects depending on the actual wavelength/lightpath allocation. We then propose a novel algorithm to solve the RWA problem that explicitly considers the physical impairments.

Simulation results show the effectiveness of our approach. Indeed, when the transmission impairments come into play, an accurate selection of paths and wavelengths which is driven by physical consideration is mandatory. See [5] for more details.

For what regards instead the second item, belonging to the first and second project workpackages, we investigated two aspects on MAN. The first one presents RingO, a WDM, ring-based, optical packet network suitable for a high-capacity Metro environment. We present three alternative architectural designs, and elaborate on the effectiveness of optical with respect to electronic technologies, trying to identify an optimal mix of the two technologies. We present the design and prototyping of a simple but efficient access control protocol, based upon the equivalence of the proposed network architecture with input-buffering packet switches. We discuss the problem of node allocation to WDM channels, which can be viewed as a particular optical network design problem. We finally briefly illustrate the fault protection properties of the RingO architecture. See [3] for more details.

The second work build upon the previous one, and it is motivated by the idea that slow tunable receivers in addition to fast tunable transmitters are suitable for packet WDM metro networks and are a fundamental asset that permit to follow long term traffic variations while preserving low network complexity. Indeed, receivers do not have to react to single user needs but to their aggregate demands. As a consequence, since groomed demands change slowly, slow tunability is enough to pursue this goal. The increased cost is marginal with respect to traditional fixed devices.

For a specific WDM packet ring Metro network, we discuss reconfiguration schema that permit to re-allocate receivers to wavelengths according to long term variations in the traffic pattern. The propose reconfiguration schema address the reconfiguration problem with two different measurements approaches. While the incoming traffic measurement one uses a complex but more accurate measurement mechanism (at the cost of requiring periodic signaling messages), the in-transit measurement approach implements a simpler scheme that, however, may become unreliable under highly loaded conditions. In this second scenario, we show that it is wise to run simpler algorithms that try to improve network performance by a trial and error procedure rather than using an optimization approach that implies a complete redefinition of receiver allocation.

The proposed algorithms show good ability in tracking traffic variations by receiver reconfigurations, accurately balancing performance and reconfiguration costs. As expected, distributed measurement schema perform better than in-transit one, even if the difference is significant only in overloaded conditions. While results were presented only in a specific dynamic traffic setup, they are well representing general trends. see [4] for more details.

Most of the results presented in this project were obtained by the means of computer simulations. To this purpose, a high-performance cluster of Personal Computer was built, and configured using LINUX as Operating System.

Several software tools have also been produced, some for internal usage, but some is made available to the research community. In particular, the tool developed to assess the performance of several RWA algorithms in case of physical impairments is

available on request. This tool has been called DOTDM - WDM/OTDM Dynamic Simulator. It allows the user to test the blocking probability on a given network topology, and for a given traffic pattern. OTDM SuperLightpath and traditional lightpath are supported. Confidence level and accuracy of results are tested by the means of statistical techniques to guarantee the user the accuracy of the results.

Several master and PhD students were involved in the project, and presentation of the results have been given to both public conferences and workshops and to people working in research centers that hosted people involved in this project. In particular, Ing. Chiara Piglione, who is following her 3rd year of PhD course, was visiting the Department of Electrical Engineering, Arizona State University (ASU), Tempe, AZ, under the supervision of Prof. Martin Reisslein and collaborated with them for eight months on topics related to this project. Similarly, Prof. Neri visited Cisco Systems, San Jose (CA), while Prof. Emilio Leonardi visited the Sprint Labs at Burlingame (CA) during summer 2003.

In the next references section, we report all the publications that are related to this project. Items in **boldface** are publications generated at Politecnico di Torino in the second year of the Adonis project.

Bibliography

[1] P. Petracca, M. Mellia, E. Leonardi, F. Neri, "Design of WDM Networks Exploiting OTDM and Light Splitters", Proceedings *2*$^{\text{nd}}$ *International Workshop on QoS in Multiservice IP Networks (QOS-IP 2003)*, Milan, February 2003, available in M. Ajmone Marsan, G. Corazza, M. Listanti, A. Roveri (eds.), *Quality of Service in Multiservice IP Networks*, Lecture Notes in Computer Science, vol. 2601, Springer, pp. 433-446, 2003

[2] L.Calafato, M.Mellia, E.Leonardi, F.Neri, "Exploiting OTDM Traffic Grooming in Dynamic Wavelength Routed Networks", *ONDM 2004 - Eighth Working Conference on Optical Network Design and Modelling*, Ghent - Belgium, 2-4 February 2004.

[3] **C.Andrea, V.De Feo, J.M.Finochietto, R.Gaudino, F.Neri, C.Piglione, P.Poggiolini, "RINGO: An Experimental WDM Optical Packet Network for Metro Applications",** *IEEE Journal on Selected Areas in Communications, Advances in Metropolitan Optical Networks (Architectures and Control)***, Vol. 22, No. 8, pp. 1561–1571, October 2004.**

[4] **A.Bianco, J.M.Finochietto, G.Giarratana, F.Neri, C.Piglione, "Measurement Based Reconfiguration in Optical Ring Metro Networks",** *submitted to JLT***, december 2004.**

[5] **R.Cardillo, V.Curri, M.Mellia, "Considering Transmission Impairments in Wavelength Routed Networks",** *Optical Network Design and Models - ONDM***, Milan, February 7-9, 2005.**

Measurement Based Reconfiguration in Optical Ring Metro Networks

A. Bianco, J. M. Finochietto, G. Giarratana, F. Neri, and C. Piglione
Dipartimento di Elettronica, Politecnico di Torino
Cso. Duca degli Abruzzi, 24 10138
Email: *lastname*@mail.tlc.polito.it

***Abstract*—Single-hop WDM optical ring networks operating in packet mode are one of the most promising architectures for the design of innovative Metro (Metropolitan network) architectures. They permit a cost-effective design, with a good combination of optical and electronic technologies, while supporting features like *restoration* and *reconfiguration* that are essential in any Metro scenario. In this article, we address the tunability requirements that lead to an effective resource usage and permit reconfiguration in optical WDM Metros. We introduce reconfiguration algorithms that, on the basis of traffic measurements, adapt the network configuration to traffic demands to optimize performance. Using a specific network architecture as a reference case, the paper aims at the broader goal of showing which are the advantages fostered by innovative network designs exploiting the features of optical technologies.**

Keywords: all-optical networks, WDM packet rings, logical topology design, dynamic network configuration

I. Introduction

As Internet usage continues its growth, carriers continue to see a steady increase of packet data traffic in their Metros. Today's network solutions are mostly based on circuit-switched SONET/SDH rings that are not efficient in carrying data traffic, due to their inherent asymmetry, and bursty and self-similar behavior. Several evolutions of legacy SONET/SDH to packet-switched technologies are currently being proposed. For example, the IEEE 802.17 RPR (Resilient Packet Ring) standard aims at solving problems from which SONET/SDH networks suffer in supporting packet data by optimizing bandwidth sharing. However, as higher rates need to be supported, both SONET/SDH and RPR node costs increase, since all incoming/outgoing and in-transit traffic needs always to be processed electronically. Similar scaling problems arise in Metro infrastructures based upon switched Gigabit Ethernet, with additional concerns related to fair resource allocation and QoS control. Basically, in current solutions network scalability is limited because nodes must switch/process the full network bandwidth.

Due to advances in optical technology [1], new packet-switched networks can be devised that can sustain cost-effectively larger bandwidths. Metros seem to be one of the best arenas for an early penetration of these technologies. On the one hand, high capacity requirements can be satisfied by exploiting fiber bandwidth by means of Wavelength Division Multiplexing (WDM), without requiring node interfaces to access and electronically process the full network bandwidth. On the other hand, packet traffic can be handled by temporally sharing WDM channels, either by dynamically setting up *lightpaths* between nodes willing to communicate, or by exploiting statistical packet multiplexing in static channels.

In this context, single-hop optical ring networks operating in packet mode are considered a promising architecture for future Metros [2]–[7]. The ring topology has been extensively proposed in the literature because of its simplicity and since it easily satisfies restoration requirements. Besides, the single-hop approach avoids complex switching in the optical domain and thus permits a cost-effective balance of optics and electronics. In these networks, nodes are equipped with few (typically one) transceivers, and each transceiver operates at the data rate of a single WDM channel. Paths between nodes are created by dynamically sharing on a packet-by-packet basis WDM channels, without requiring nodes to process the full network bandwidth. However, tunability at transceivers is required to exploit the fiber bandwidth by temporally allocating all-optical single-hop bandwidth between nodes in all available channels.

Due to the cost of tunability at transceivers, media access protocols that require packet-by-packet tunability only at one end of the all-optical path (i.e., either only at the transmitter, or only at the receiver) have been studied to save the cost of the still quite expensive tunable devices. Usually these protocols assume a *fastly* tunable transmitter and a fixed receiver, permanently

tuned to a WDM channel [8], [9]. When a node needs to send a packet, it simply tunes its transmitter to the receiver's destination wavelength. This implies that transmitter tuning times must be negligible with respect to the packet duration to obtain a good efficiency. Simple distributed access protocols can be designed for these tunable-transmitter/fixed-receiver architectures.

If the number of nodes is larger than the number of WDM channels, a decision problem arises concerning the allocation of the different receivers to WDM channels to equalize the traffic among the available channels. If fixed receivers are considered, any allocation is permanent and cannot be updated in response to long-term changes in the traffic pattern, which are typical in Metros. Therefore, it may be worthwhile to re-allocate, i.e. tune to different wavelengths, receivers, to dynamically keep the network in an optimal operation point. One elegant way of achieving this result is to introduce *slow* (hence cheap) tunability in receivers. This tunability does not need to be *fast*, since it must not track packet-by-packet variations, but longer-term variations of the traffic pattern. Low-cost devices available today (e.g., mechanical or thermo-optic filters) can be suitable to implement this *slow* receiver tunability feature.

If *slow* tunability is present at receivers, the impact of reconfiguring network receivers must be taken into proper account [10] when solving the problem of allocating receivers to match traffic conditions. In fact, retuning a receiver implies introducing a period of service disruption during which nodes cannot transmit to that receiver. As a consequence, the reconfiguration must not only bring the network to an optimal operation point, but also minimize service disruption.

The scope of this work is to introduce reconfiguration algorithms in a single-hop optical ring network with slow receiver reconfiguration capabilities. Although a similar problem has been studied in [10], our work looks at defining proper reconfiguration algorithms relying on traffic measurements to detect the traffic pattern, which is not assumed to be known. The main contribution is the introduction of reconfiguration schema which aim at keeping the network at an optimal operation point with minimum service disruption.

We would like to remark that the issue addressed in the paper has an interest beyond the application to the specific considered Metro architecture. Indeed, the rate at which network resources should be reallocated (which often translates into the speed at which transceivers and switches must tune in optical networks) cannot keep up with the continuous reduction of packet durations with increasing line rates. Packet-by-packet control and retuning introduces strong technological challenges, and consequently high costs, which may not be strictly necessary to provide acceptable levels of quality of service to users. Indeed, service requirements do not scale with transmission speeds and packet durations, and most network control dynamics need to be matched to service requirements instead of to optical transmission rates.

The paper is organized as follows. In Sec. II we describe the considered network architecture and the system model, providing motivations for the importance of network reconfigurability, i.e., adaptability to long term traffic fluctuations. The problem is formalized in terms of *Mixed Integer Linear Programming* (MILP) in Sec. III. In Sec. IV we introduce the basic reconfiguration mechanisms, assuming two different measurement schema, named incoming and in-transit traffic measurements. Next, in Sec. V we discuss simulation results to assess the properties of the proposed algorithms. Finally, we draw some conclusions in Sec. VI.

II. System Model

We consider a specific WDM optical packet network, physically made of two counter-rotating rings. This architecture was proposed and is currently being studied and prototyped in the framework of the Italian national project called WONDER [11]. Each ring comprises N nodes and conveys W wavelengths. The network is assumed to be synchronous and time-slotted. During a time slot, at most one packet can be transmitted in one of the W available slots, one for each wavelength channel. Rings are used in a peculiar way: one ring is used for transmission only while the other one is used for reception only. To provide connectivity between the two rings, a folding point is needed, where transmission wavelengths are switched to the reception path, as sketched in Fig. 1. Transmitted packets travel towards the folding point in a first ring traversal, are switched to the reception path, and then received during a second ring traversal. If each node can become the folding point (i.e., if each node has a switching capability) then the network preserves the interesting restoration property of rings, as described in [12]. Although this architecture does not exploit wavelength spatial reuse, it avoids transmission impairments (e.g., noise recirculation) typical of ring topologies, while guaranteeing that all the network traffic accepted prior to the fault can be supported also after restoration (note that this may not be the case in ring networks with spatial reuse).

Nodes are equipped with a *fastly* tunable transmitter and exploit WDM to partition the traffic directed to disjoint subsets of destination nodes, each subset comprising the destinations whose receivers are currently tuned

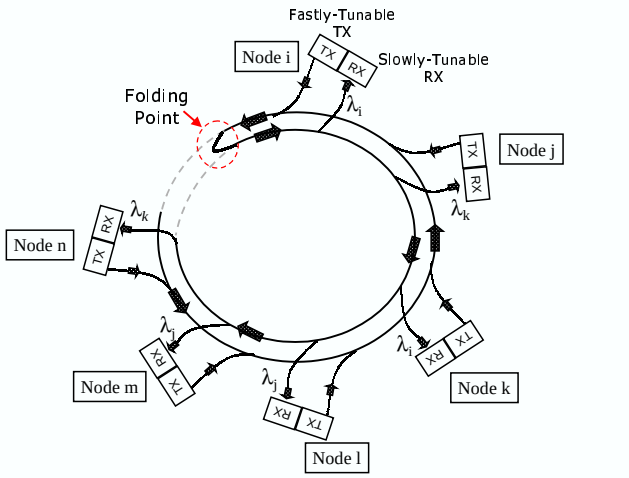


Fig. 1. Logical network model

to the same wavelength. Nodes tune their transmitters to the receiver's destination wavelength, and establish a temporary single-hop connection lasting one time slot. Sharing of wavelength channels is therefore achieved according to a statistical Time Division Multiple Access (TDMA) scheme. Access decisions are based on a channel inspection capability (similar to the carrier sense functionality in Ethernet – see [12]), by which nodes know which wavelengths were not used by upstream nodes in each time slot. Priority is given to in-transit traffic, i.e. a multi-channel empty-slot protocol is used.

While *fast* tunability at transmitters is required to provide full connectivity between nodes, *slow* tunability at receivers is needed to follow traffic changes in time, since nodes' receivers may have to be re-tuned to balance the offered load among the available WDM channels. As an example, a configuration where N/W nodes are tuned on each of the W available wavelength channels is optimal for uniform traffic distribution, since the load on all available channels is equalized, but it may be highly inefficient for a different traffic distribution, as for example in presence of heavy-traffic servers. Thus, if the traffic distribution is unknown or time-variable, the slow tunability becomes a must to allow a dynamic network configuration to be obtained.

We consider that a *fastly* tunable transmitter is capable of re-tuning itself in negligible time between two consecutive time slots, while a *slowly* tunable receiver is characterized by a tuning latency τ.

Fast tunability is not assumed at the receiver to limit both network costs and reconfiguration algorithm complexity. Due to the receiver tuning latency, all nodes willing to transmit to receivers that are currently involved in a re-tuning process must refrain from transmission for a period of time at least equal to the tuning latency τ. This is to avoid packet losses, since retransmission costs may be very high in networks with large delay$\times$bandwidth products.

We focus in this paper on the control of slowly tunable devices, to adapt network configuration to slowly-varying traffic distributions. The reconfiguration mechanism proposed is conceptually centralized, i.e., it runs in a given master node which is responsible for reconfiguration decisions as well as for collecting information used to drive the reconfiguration process. The master node collects information on the network status, as channel loads and node transmission needs, by proper traffic measurements described later in more details. If the new network status requires or suggests that receivers reconfiguration may be useful, a proper reconfiguration algorithm is run and a new network configuration, i.e., a new assignment of receivers to wavelengths, is computed. The master node broadcasts signaling messages containing the new configuration to the other, slave, nodes. To account in a simple way for the propagation time of signaling messages, we assume the worst case situation, in which a receiver becomes unavailable due to a re-tuning for twice the RTT (ring Round Trip Time, equal to one ring traversal latency) to allow for signaling messages propagation, plus the tuning latency to allow the slowly tunable device to tune to the proper wavelength. More precisely, the reconfiguration process goes through three steps: first the master disables all transmissions towards nodes that must be reconfigured; after two RTTs (to permit to in-flight packets to reach the receiver after traversing the transmission and the reception paths), nodes re-tune their receivers, and finally, after the tuning latency, transmissions towards destinations involved in the reconfiguration process are re-enabled. Note that, under light load conditions, these periods of receiver unavailability only affect transmission delays, while under high load conditions they may also cause buffer overflows and packet losses due to network capacity reductions.

In Sec. IV, we define both the measurement techniques adopted to monitor network status, and the algorithms used to determine first the need of reconfiguration and then the new network configuration.

III. Problem formulation

In this section we formally model the reconfiguration problem. We focus on a given, measured, traffic pattern, and formalize the problem of finding an optimal network configuration, i.e. an optimal allocation of receivers to wavelengths. The optimality is related to network

performance: thus, balancing the traffic load on the available wavelengths by properly assigning receivers to wavelengths is an obvious optimization goal. However, it is important to observe that reconfiguration decisions must take into account not only traffic balance among WDM channels but also service disruption due to temporary receiver unavailability. Whereas traffic balance depends on the aggregate traffic of each receiver and on the assignment of receivers to wavelengths, service disruption is related to network RTT and to receiver tuning latency. The algorithms used to determine whether a reconfiguration is needed or not should carefully consider these two aspects.

We will formalize the reconfiguration problem via a two-objectives problem, taking care first of wavelength's loads and then of re-configuration costs. Indeed, if we assume that the traffic pattern is fully known, i.e. we assume to known also the newly detected traffic pattern duration, we could compute exactly the benefit of each network configuration in terms of bandwidth utilization, taking into account both the bandwidth gain due to the new receiver configuration and the bandwidth loss due to RTTs and tuning latencies. This would lead to an optimization problem with a single objective function that considers overall network bandwidth only. However, this is clearly not possible in practice; moreover, sometimes it can be interesting to study the problem with cost functions induced by management issues, and not directly related to performance. For these reasons, we keep costs and loads as two separate objectives, rather than merging them in a linear combination and moving to a scalar problem.

The output of the measurement process is stored on the traffic matrix$\mathbf{T} = [t_{ij}]$, where t_{ij} is the traffic rate from node i to node j normalized to the channel bandwidth, and $t_j = \sum_{i=1}^{N} t_{ij}$ the aggregate receiver bandwidth of node j.

The cost of tuning receiver j to wavelength k is denoted by c_{jk} while the cost of a network reconfiguration is the sum of the costs of all re-tuned receivers. This cost can refer to the total number of receivers to be re-tuned, or to the amount of bandwidth lost, or to other different metrics. Although the reconfiguration algorithms we present can be adapted to different cost definitions, in the remainder of the paper we define the reconfiguration cost as:

$$c_{jk} = \begin{cases} 0 & \text{iff node } j \text{ is currently on wavelength } k \\ 1 & \text{otherwise} \end{cases}$$

The MILP problem formulation is obtained by introducing a set of control variables x_{jk} that consider the potential allocation of the receivers after the reconfiguration:

$$x_{jk} = \begin{cases} 1 & \text{iff node } j \text{ will receive on wavelength } k \\ 0 & \text{otherwise} \end{cases}$$

Thus, the mathematical model of the problem becomes:

$$\text{Minimize} \quad [\mathcal{L}_{\max};\ \textstyle\sum_{j=1}^{N} \sum_{k=1}^{W} c_{jk} x_{jk}] \tag{1}$$

subject to the following constraints:

$$\mathcal{L}_{\max} \geq \mathcal{L}(k) = \sum_{j=1}^{N} t_j x_{jk} \qquad \forall k,\ 1 \leq k \leq W \tag{2}$$

$$\sum_{k=1}^{W} x_{jk} = 1 \qquad \forall j,\ 1 \leq j \leq N \tag{3}$$

Eq. (2) guarantees that no wavelength has a load $\mathcal{L}(k)$ larger than $\mathcal{L}_{\max}$, while Eq. (3) ensures that each receiver is allocated only to one wavelength, since nodes are equipped with a single receiver. The problem is solved in a lexicographic fashion, i.e, finding out first the minimum $\mathcal{L}_{\max}$ for the current traffic pattern, and then looking for the least disruptive allocation of receivers that equals $\mathcal{L}_{\max}$. In particular, the problem of finding a well-balanced allocation of receivers to wavelength channels is equivalent to the well-known problem of scheduling jobs on identical parallel machines [13], where receiver's aggregate traffict_j represents job's duration, and wavelengths represent machines. This is an NP-hard problem; hence, our problem, being a generalization, is also NP-hard.

IV. The Reconfiguration Mechanism

In this section, we describe heuristic approaches to solve the reconfiguration problem. We classify reconfiguration mechanisms on the basis of the traffic measurement approach used to collect information on network status. Two different approaches are defined to collect measured data:

- *in-transit* traffic measurement scheme: the master node observes the amount of traffic transmitted on all WDM channels
- *incoming* traffic measurement scheme: all nodes measure the amount of data arriving for transmission to the Metro, and send this information to the master node

The first scheme is simpler to implement and requires less signaling. However, only the traffic that was able to access the network can be observed and measured; thus, no overload situation can be measured. In the

second one, the real traffic offered to the network can be detected, thereby leading to a more reliable estimate of the traffic pattern, at the price of signaling overhead to convey this information to the master node. In particular, the estimate is accurate also in overloaded conditions.

Since reconfiguration mechanism are triggered by measurements, a key point is the estimation of the current traffic pattern. We assume long-term variations of the traffic with respect to packet dynamics; thus, measurements can be done periodically over a measurement window long enough to estimate steady-state traffic conditions. If traffic measurements are done by each node considering incoming traffic addressed to the Metro, then a straightforward mean of the measured samples is accurate enough to characterize traffic conditions, provided that the measurements window size is properly set. However, if measurements are done looking at in-transit traffic, then measured samples can be affected by transient phenomena. In fact, nodes could buffer packets either if the network configuration is not matched to the traffic pattern (i.e., prior to reconfiguration) or if a node is re-tuning. As a consequence, after network reconfiguration, the traffic sent on the network represents not only current incoming traffic conditions but also transient ones. Therefore, an exponentially weighted mean is used to mitigate the effect of transient measured values. More details on the measurement process are provided later.

A. Incoming Traffic Measurement Scheme: the 3-step algorithm

In this scheme, nodes measure the traffic groomed from the local area, and willing to travel over the Metro. Each node estimates one row of the traffic matrix $\mathbf{T}$, which is communicated to the master node to keep an updated estimation of the whole traffic matrix. The estimation is done periodically over a measurement window of duration T_m.

The reconfiguration scheme we propose for this case is named *3-step* algorithm; it aims at a full network reconfiguration at the end of each measurement window. In the first step, named LB (Load Balancing), the problem of balancing receiver loads is solved without accounting for service disruption. Once a solution has been found, Step 2, named WA (Wavelength Assignment), and Step 3, named RS (Receiver Swapping), try to improve the LB solution by jointly keeping channels balanced and avoiding unnecessary reconfigurations.

1) LB: Load Balancing: As previously stated, the problem of finding a well-balanced allocation of receivers to wavelength channels can be mapped to the problem of scheduling jobs on identical parallel machines [13]. Although the problem falls in the class of NP-hard problems, approximation algorithms do exist which limit the distance from the optimal solution. The *Longest Processing Time* (LPT) algorithm is one of these, which guarantees that any solution is at most $4/3$ greater that the optimal one [14], thus providing a upper bound on the distance of our algorithm from the optimal solution.

LPT works on W bins, representing wavelengths to which receivers must be allocated. Receiver aggregate bandwidths, obtained by matrix $\mathbf{T}$, are loaded in initially empty bins following these steps:

1 Sort nodes by decreasing $t_j,\ \forall j = 1, \ldots, N$.
2 Allocate largest t_j to least loaded bin.
3 If unassigned receivers do exist, go to 2.

The LPT algorithm is run each time a new traffic matrix is detected at the master node, to find out if a new configuration is needed. Whether to schedule or not a reconfiguration depends on how much the new allocation improves network performance. A threshold is defined to decide whether it is worthwhile to reconfigure the network by comparing the old and the new allocation: the threshold is applied to $\sum_{k=1}^{W} \min(1, \mathcal{L}(k))$, the maximum overall traffic that can be handled by the network. Once LB has found a solution that improves over the previous one more than the threshold, the next step is to associate each of the loaded bins with wavelengths to minimize the number of nodes that must be reconfigured.

2) WA: Wavelength Assignment: The next step is to associate the bins filled by LB with a proper wavelength, so as to minimize the number of nodes that should retune their receiver. The wavelength assignment problem can be seen as a bipartite matching problem [15]. A bipartite graph has two sets of nodes: edges may not connect nodes in the same set. A matching is a subset of edges with the constraint that at most one edge in the matching can be connected to each node. If a weight is associated to each edge, the weight of a matching is the sum of the weights associated with the selected edges. A matching with weight w^* has maximum weight if no other matching exist with weight larger than w^*.

In Fig. 2, numbers close to left nodes represent the receivers that must be allocated to the corresponding bin while numbers close to right nodes represent the receivers currently allocated to the corresponding wavelength. Each edge from bin i to wavelength j has a weight w_{ij} equal to the number of allocated receivers on bin i currently tuned to wavelength j. By running a Maximum Weight Matching (MWM) algorithm, we obtain the wavelength assignment that minimizes the number of nodes involved in the reconfiguration, thus minimizing the cost function. As a result, we minimize the number of reconfiguration required to obtain the load

balancing on wavelength channels as determined by LB in the previous step.

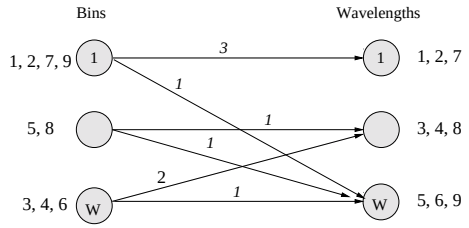


Fig. 2. Bipartite graph for Wavelength Assignment

3) RS: Receiver Swapping: Even if WA tries to minimize the number of receiver reconfiguration, there may exist some un-needed reconfigurations. To understand why, let's assume that two receivers, i and j, are similarly or equally loaded. Since no notion of previous wavelength allocation is used when determining bin assignments via LB, and since WA cannot modify receiver allocations to bins, it may happen that receivers i and j are assigned to wavelengths in such a way that both need a reconfiguration, whereas it would be possible, by exchanging their bin assignment, to avoid their reconfiguration. Thus, RS is a simple local search algorithm that swaps almost equally loaded receivers to avoid unnecessary reconfigurations.

The outcome of the 3-step algorithm is a new configuration with better load balancing properties than the previous one, and with low service disruption. Although the proposed heuristic may not lead to the optimal solution with minimum service disruption, it clearly eliminates unwanted reconfigurations that otherwise would cause larger blackout periods.

B. In-transit Traffic Measurement Scheme: the First-Fit algorithm

Instead of measuring incoming traffic at each node, the in-transit measurement approach for traffic estimation measures in-band traffic. One way of implementing this scheme is to equip the master node with a receiver on each wavelength, so that measurements can be centralized in this node, thereby eliminating the exchange of signaling messages at the cost of increased hardware complexity. However, under high load conditions, measurements may not estimate accurately the traffic matrix. Indeed, the master is not able to observe all arrived traffic, due to contentions among nodes willing to access an overloaded channel.

When this scheme is adopted, measurements are exponentially weighted to smooth transient values that do not represent steady-state conditions. As a consequence, the master node divides the measurement window of duration T_m in K sub-windows, each one weighted by w_k. If m_{jk} is the mean value of bandwidth directed to receiver j in the k-th sub-window, the resulting aggregate traffic bandwidth on this receiver is computed as:

$$t_j = \frac{\sum_{k=1}^{K} m_{jk} w_k}{\sum_{k=1}^{K} w_k} \tag{4}$$

where $w_k = e^{k\tau}$, τ being the slow receiver tuning latency, since larger tuning latencies imply longer transient periods.

If there is no congestion on channels, the 3-step algorithm described in Sec. IV-A could be used since the traffic matrix can be correctly estimated. However, if congestion is present, the algorithm frequently exhibits convergence problems due to uncertainties in the estimation of the measured bandwidth. Moreover, running a fairly complex algorithm when the measured data are potentially affected by errors may not be the best solution.

As a consequence, we propose a simpler algorithm that aims at considering both load balancing and service disruption simultaneously. Instead of taking full reconfiguration decisions at once, we try to improve the network configuration by successive, partial modifications, each modification being possibly triggered at the end of each measurement window. Although this approach may not lead to the optimal solutions, it reconfigures the network in a conservative way, so as to see how the new configuration accommodates traffic, later considering another partial reconfiguration that takes into account the effect of the previous modification.

1) First-Fit Algorithm: The key idea of the FF (First-Fit) algorithm, which is an extension of the algorithm presented in [16], is to schedule only partial reconfigurations that improve traffic balance among channels. In fact, each partial reconfiguration aims at re-tuning the least loaded receiver from the most loaded wavelength M to the least loaded one m. While it is straightforward to find out$m$, M cannot be easily determined if there is congestion in more than one channel. Indeed, it is difficult to distinguish a channel close to congestion from one strongly congested since both would look alike considering in-transit measurements (i.e., both measured loads are approximately equal to 1). Thus, there is risk that a node is re-tuned from a channel not congested (but close to congestion) to the least loaded channel without any overall improvement. In fact, if the channel was not suffering congestion, the performance in the next measurement window is not improved. The FF algorithm exploits this idea to keep track of the channels that are

actually not congested. For this purpose, a list $\mathcal{W}$ of potentially congested channels is used; while m is picked up among all channels, M is selected from the list $\mathcal{W}$. To evaluate the performance improvement of a previously scheduled reconfiguration, we look at the mean measured channel load defined as:

$$\bar{\mathcal{L}} = \frac{\sum_{k=1}^{W} \mathcal{L}(k)}{W}$$

Every time a reconfiguration has to be done, we save the current mean load as $\bar{\mathcal{L}}_{old}$ and the values of wavelengths M and m. After a measurement window, we evaluate the new mean channel load as $\bar{\mathcal{L}}_{new}$ from the in-transit measurements. If $\bar{\mathcal{L}}_{new}$ is greater than $\bar{\mathcal{L}}_{old}$, then the previous reconfiguration helped; otherwise, an un-needed re-tuning was done and channels M and m must be removed from the list $\mathcal{W}$, since we know that they are not congested channels.

More formally, the FF algorithm runs through the following steps:

1 Evaluate $\bar{\mathcal{L}}_{new}$
2 If $\bar{\mathcal{L}}_{new} > \bar{\mathcal{L}}_{old}$ goto 4
3 Update $\mathcal{W} = \mathcal{W} \backslash \{M, m\}$
4 Find
$$\mathcal{L}(M) : M = \arg \max_{k \in \mathcal{W}} \mathcal{L}(k)$$
$$\mathcal{L}(m) : m = \arg \min_{1 \le k \le W} \mathcal{L}(k)$$
5 Select smallest t_j on wavelength M such that :
$\mathcal{L}(M) + \epsilon > \mathcal{L}(m) + t_j$
6 If t_j does exist then:
re-tune node j to wavelength m and $\bar{\mathcal{L}}_{old} = \bar{\mathcal{L}}_{new}$;
else, $\bar{\mathcal{L}}_{old} = 0$ and $\mathcal{W} = \{1, \ldots, W\}$

Initially, the list $\mathcal{W}$ contains all channels; if, after step 3, the list becomes empty, then the list is replenished. Once M and m have been selected after step 4, we still need to find out the smallestt_j in M that does not load m beyond the current load on M; otherwise, we are increasing traffic unbalance. When comparing the current maximum load $\mathcal{L}(M)$ with the possible load on m, a small value ϵ is used to account for inaccuracies in the estimation of channel loads. Finally, a re-tuning is scheduled at step 6 if there exists a t_j that satisfies the previous described condition; otherwise, the algorithm parameters are re-set to the initial values.

V. Results and Analysis

In this section we present performance results obtained by simulation when considering a network with $W = 4$ wavelengths and a total of $N = 16$ nodes, where the distance between two adjacent nodes is about 27 km, i.e., 90μs; thus, the ring RTT is 1.45ms. Slots last 1μs, corresponding to a packet size of about 1250 bytes at 10 Gbit/s. The slowly tunable receiver has a tuning latency $\tau = 10$ms. Nodes adopt a VOQ (Virtual Output Queuing) architecture to avoid HoL (Head of the Line) blocking of packets waiting for access; thus, each node keeps a separate FIFO queue for each destination node, with a queue size of 32000, fixed size, packets.

The duration of the measurement window is set to $T_m = 50$ms. In the case of incoming traffic measurements, the threshold to determine whether the new allocation is worth the reconfiguration cost is set to 5% of the previous allocation. For the in-transit traffic case, each measurement window is divided in $K = 5$ intervals, exponentially smoothed as previously described.

We look at transient scenarios, i.e., algorithm behaviors when the traffic pattern changes, according to a predefined scheduling, from an initial traffic pattern to a final one. We assume a linear transition from the initial traffic pattern to the final one, i.e., the transition occurs in a given number S of subsequent steps of uniform duration T_S. At a given step, the intermediate traffic matrix (representing the current traffic pattern) is obtained from a linear combination of the initial traffic matrix and the final traffic matrix, weighted by$S - i/S$ and i/S respectively, where i is an integer ranging from 0 to $S-1$ which increases at each step. In our simulations, we set $S = 10$ and $T_S = 100$ms, i.e. 100000 slot times; thus, the whole transition process lasts $S \times T_S = 1$s, starting at simulation time 1s and ending at simulation time 2s. Although a one-second traffic variation cannot be considered a long-term one, we use it to describe the properties of our algorithms; moreover, it can represent a limit case of fastly changing pattern that our algorithms are able to cope with.

The initial traffic pattern is an uniform traffic pattern; since the same number of nodes is initially assigned to each wavelength, the whole capacity of the network is equally shared by all nodes. Thus, the element t_{ij}, $1 \le i, j \le N$, of the uniform traffic matrix is given by:

$$t_{ij} = \lambda_{in} \frac{W}{N} \frac{1}{N-1}$$

where λ_{in} represents the normalized input load. The final traffic pattern is named "2-server"; in this scenario, nodes are partitioned into two separated subsets: servers $\mathcal{S}$ and clients $\mathcal{C}$. The two nodes belonging to $\mathcal{S}$ transmit at a high rate, equal to the capacity of one wavelength per node, with equal probability to the other $N - 2$ nodes belonging to $\mathcal{C}$. The remaining network capacity is shared by client nodes, that transmit only to the servers with equal probability. In other words, t_{ij}, $1 \le i, j \le N$,

is given by:

$$t_{ij} = \lambda_{in} \begin{cases} 0 & \text{iff } i \in \mathcal{S} \wedge j \in \mathcal{S} \\ \frac{1}{N-2} & \text{iff } i \in \mathcal{S} \wedge j \in \mathcal{C} \\ \frac{W-2}{N-2}\frac{1}{2} & \text{iff } i \in \mathcal{C} \wedge j \in \mathcal{S} \\ 0 & \text{iff } i \in \mathcal{C} \wedge j \in \mathcal{C} \end{cases}$$

In the sequel, we analyze algorithm performance when varying the traffic pattern according to the previously described scheduling, starting from the initial uniform traffic and finally leading to the 2-server traffic pattern. In all subsequent figures, with the exception of Fig. 7, we plot the instantaneous throughput on the left and the normalized cumulative throughput, i.e., the total amount of data sent normalized to the input load, on the right. In left figures, each traffic pattern transition is highlighted at the top of the plot by a cross point when the transition occur.

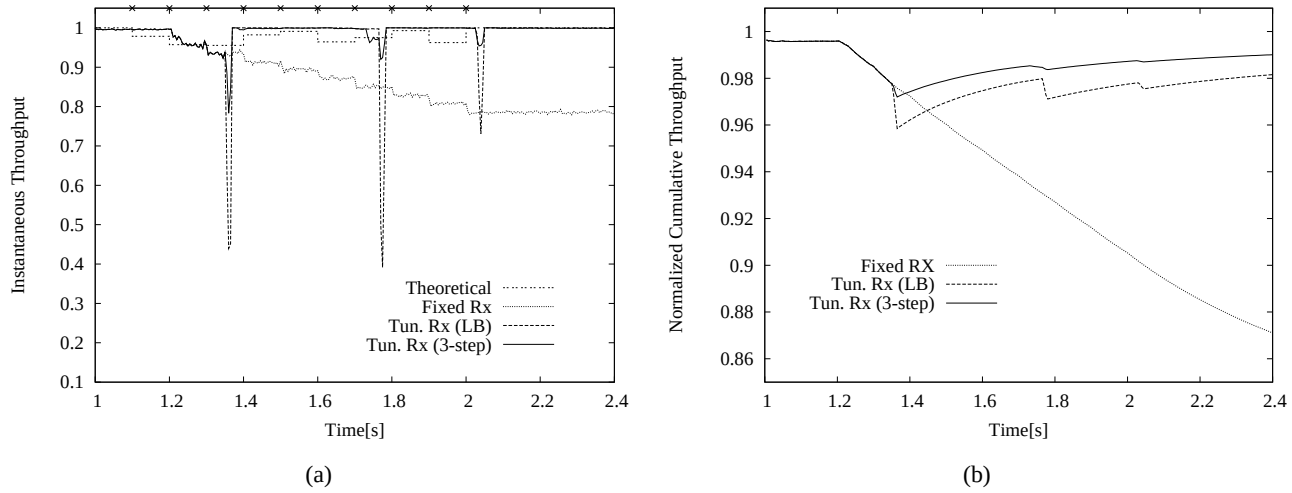
A. Incoming traffic measurement

We first analyze the behavior when the incoming traffic measurement scheme is considered. Several algorithms are considered and compared: fixed receiver ("Fixed RX"), theoretical, tunable receiver with LB only ("Tun. Rx (LB)") and tunable receiver with the full 3-step algorithm ("Tun. Rx (3-step)"). The fixed receiver refers to the case when no reconfiguration takes place, with the configuration being matched to the initial uniform traffic pattern. The theoretical case refers to an ideal case, when the new network configuration is, instantaneously, optimally matched to the current real traffic pattern. This curve is not obtained by simulation; it simply refers to the maximum achievable throughput given the traffic matrix under consideration. Therefore, this case does not consider buffering effects; as a consequence, the curve may provide worse performance than other algorithms in simulation situations where buffering may temporarily increase throughput. Note that this is not equivalent to a simulation under an idealized scheme, i.e., where the network configuration is immediately adapted to traffic changes without waiting for the measurement algorithm to detect the new traffic pattern, and both RTT and tuning latencies are neglected. The two other cases refer to the full 3-step algorithm and to the same algorithm when using the LB step only; in this last case, bin i, $1 \leq i \leq W$, is directly associated to wavelength i.

In Fig. 3(a), the instantaneous network throughput (averaged over all wavelengths) is plotted when the network load is 100%, i.e., $\lambda_{in} = 1$. The theoretical curve shows that some of the intermediate traffic matrices, obtained as a linear combination of the uniform and of the 2-server scenarios, cannot be completely scheduled, i.e., they are not admissible (the instantaneous throughput becomes less than 1). The case of fixed receivers highlights the performance degradation when traffic changes occur and the network configuration becomes increasingly unsatisfactory with respect to the current traffic pattern. When slow tunability is present at receivers, reconfigurations are scheduled to keep the configuration matched to traffic variations. Although these reconfigurations introduce blackout periods causing throughput falls, they put the network close to an optimal operational point with respect to the new traffic pattern. As shown on Fig. 3(a), reconfiguration algorithms produce fairly different results; in particular, the 3-step algorithm clearly outperforms the algorithm using LB only. The performance difference between the algorithms can be better appreciated in Fig. 3(b), where we report the normalized cumulative throughput. The 3-step algorithm performs best, thanks to the ability of scheduling reconfigurations that minimize the number of re-tuned receivers, whereas the fixed solution is clearly unacceptable.

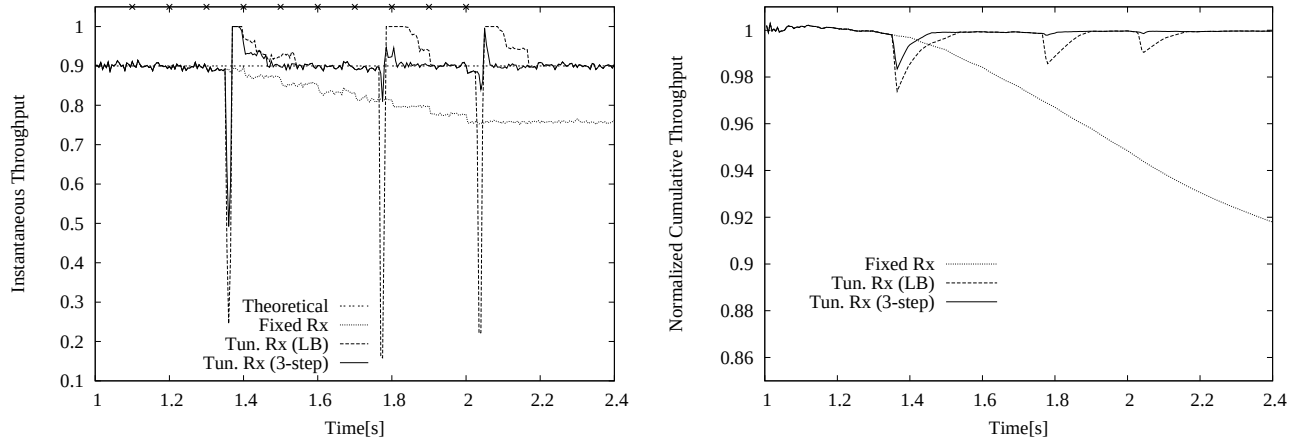
Coming back to Fig. 3(a), observe that only three network reconfigurations take place at around 1.35s, 1.75s, and 2.85s. This means that even if the algorithms detect all 10 traffic modifications, only three times the reconfiguration process provides a significantly better channel equalization than the previous configuration, thus suggesting that the reconfiguration may be worthwhile. Reconfigurations are clearly detected by the significant instantaneous throughput drop caused by the inability of transmitting toward receivers involved in the reconfiguration process, and are delayed, with respect to traffic matrix changes, roughly by one measurement window. It is worth noticing that sometimes the throughput of both 3-step and LB algorithm is higher than the theoretical case; this phenomenon is due to the buffering process at nodes. After a traffic matrix change, additional traffic with respect to the newly generated one is offered temporarily due to packets stored in buffers; thus, nodes may end up transmitting more than it would be possible with the current packet generation process alone (as it is for the theoretical case curve). Finally, after the third reconfiguration, traffic becomes steady according to the 2-server scenario and the network configuration becomes optimal for both algorithms.

In Fig. 4 the same scenario is simulated with an input load $\lambda_{in} = 0.9$. In this case all traffic matrices are admissible, and the buffered data during reconfiguration time can be transmitted once the new configuration has been set up, using the excess bandwidth available. This phenomenon is clearly visible when the instantaneous



(a) (b)

Fig. 3. Instantaneous and normalized cumulative throughput with incoming traffi c measurement scheme; λ_{in} = 1



(a) (b)

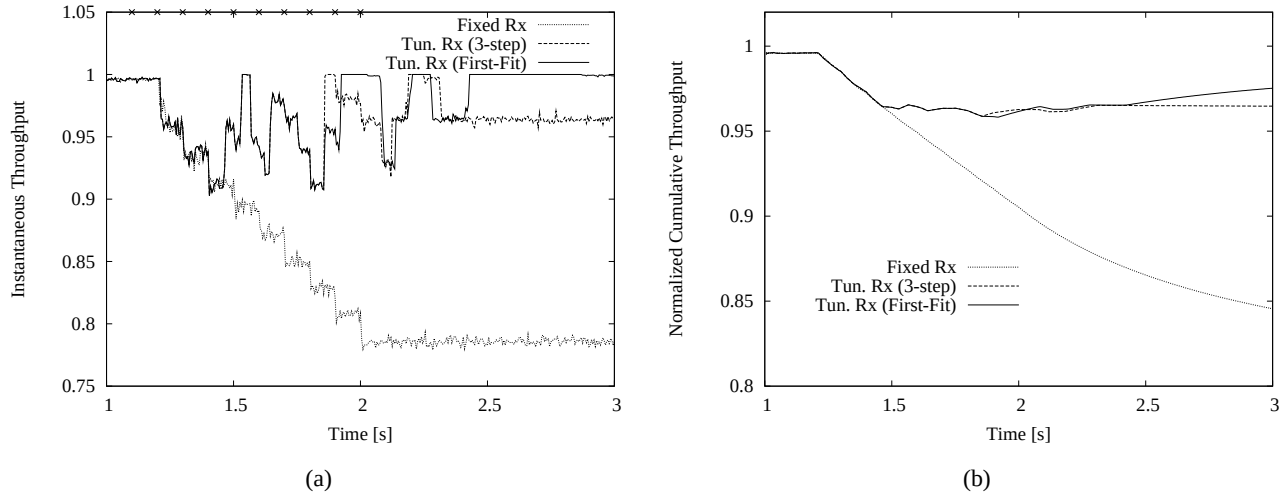
Fig. 4. Instantaneous and normalized cumulative throughput with incoming traffi c measurement scheme; λ_{in} = 0.9

throughput goes above the input load after each reconfi guration. In this scenario, the difference between both algorithms are highlighted not only during the blackout periods but also immediately afterwards. Both slow reconfi guration algorithms allow nodes to transmit all generated packets, as shown in Fig. 4(b): however, the 3-step algorithm helps the network to recover faster than LB.

B. In-transit traffic measurement

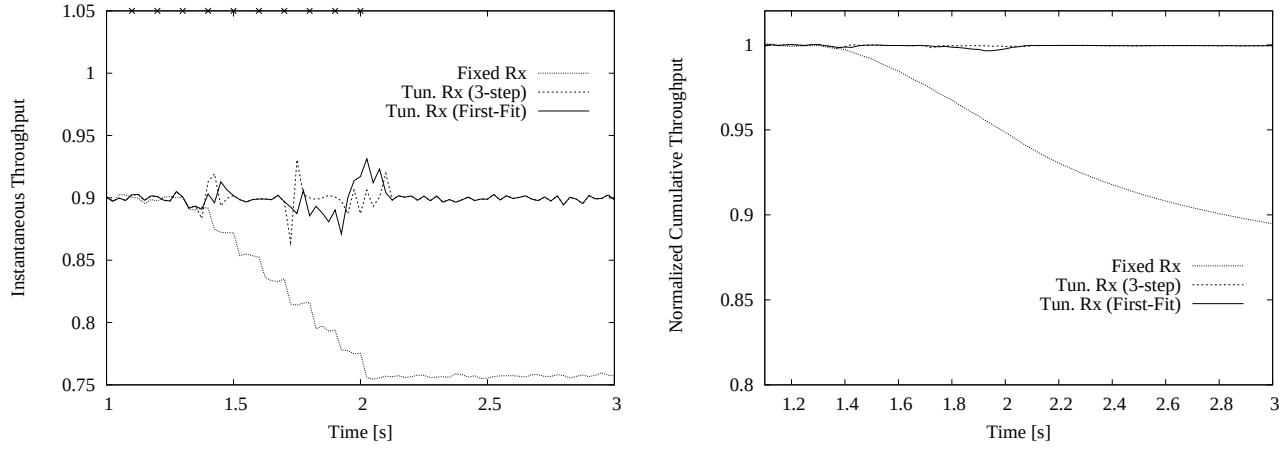
Let us focus on the in-transit measurement scheme, plotted in Fig. 5 for the same traffi c pattern described above, with $\lambda_{in} = 1$. Recall that in this scheme the measurement information is obtained by looking at wavelength channels only, thus being infl uenced by nodes' ability to successfully access wavelength channels. We report results for the fi xed receiver scheme as a " worst-case" reference and for two reconfi guration algorithms: the 3-step algorithm, defi ned for the incoming traffi c measurement scheme, and the simpler First-Fit algorithm described in Sec. IV-B.1. As expected, throughput performance decrease as compared to the previous scheme. Due to the large amount of buffered data, reconfi guration instants are not as evident as in the previous case, and can be identifi ed by a sharp throughput increase or decrease.

Uncertainties on traffi c estimates delay reconfi guration decisions and even force wrong ones, especially in the case of the 3-step algorithm. In fact, Fig. 5(a) shows that this approach does not even fi nally set the network to its



(a) (b)

Fig. 5. Instantaneous and normalized cumulative throughput with in-transit measurement scheme; λ_{in} = 1



(a) (b)

Fig. 6. Instantaneous and normalized cumulative throughput with in-transit traffi c measurement scheme;λ_{in} = 0.9

optimal operational point, while the First-Fit approach does it.

As mentioned in Sec. IV-B, full reconfi guration decisions, as those taken in the 3-step algorithm, when based on uncertain data, may be dangerous, since they can temporally improve network throughput, but, being not matched to real traffi c could sometimes produce more harms than benefi ts. In fact, as seen in Fig. 5(a), initially both algorithms perform similarly and schedule the same reconfi gurations (i.e., re-tune only one node at each reconfi guration). However, during the last three pattern changes, the 3-step algorithm cannot fi nd a well-matched confi guration since measurements on congested wavelengths become very inaccurate. Indeed, performance degrades to the point that it is not possible to fi nd the optimal operational point once the traffi c pattern is in steady state. However, it is worth to notice that the First-Fit may not always fi nd a better operational point in other traffi c scenarios, since it schedules partial reconfi gurations only. Finally, the situation becomes really diffi cult when the network is in overload only. Indeed, Fig. 6 shows that when the network load $\lambda_{in} = 0.9$, both schema perform well and no throughput loss is observed.

C. Effects of tuning latency

We wish to discuss the effect of tuning latency values on network throughput. The receiver tuning latency is a key variable, whose duration infl uences the hardware

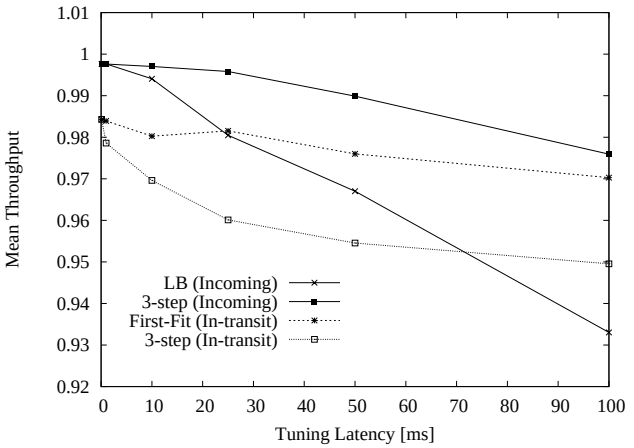


Fig. 7. Influence of tuning latency on the throughput at 100% load

architecture, thus, network cost, as well as network performance. In Fig. 7 we present the results of several simulations with different receiver tuning latencies and report the mean network throughput. The simulation scenario is the same as above, and the instantaneous throughput is averaged over a time window starting at 1s of simulation time and ending at 3s of simulation time.

At a glance, the incoming traffic measurement scheme performs better than the in-transit one. While both approaches based on incoming traffic measurements show a stronger throughput decrease with increasing tuning latencies, both algorithms based on in-transit measurements are less sensible to the tuning latency. Indeed, the tuning latency has a dominant effect in the incoming traffic measurement case given to the ability of computing a reconfiguration well matched to the traffic pattern. Conversely, in the in-transit measurement case, the dominant effect is the difficulty in obtaining a good network configuration, thus leading to larger throughput losses, in a manner less dependent from tuning latencies.

VI. Conclusions

Our work was motivated by the idea that slowly tunable receivers in addition to fastly-tunable transmitters are suitable for packet WDM Metros, and are a fundamental asset that permit to follow long-term traffic variations while preserving low network complexity. Indeed, receivers do not have to be reallocated to WDM channels on a packet-by-packet basis, but only to react to user aggregate demands. As a consequence, since groomed demands change slowly with respect to packet dynamic, slow tunability is enough to pursue this goal. The increased cost due to slowly tunable devices can be considered to be marginal with respect to traditional fixed devices.

Likewise, it is worth to remark that several switching and network configuration functions do not need to scale with the packet duration, but should instead be matched to constraints related to the quality of the offered service, hence can act on a less stringent time scale.

For a specific WDM packet ring Metro architecture, we discussed reconfiguration schema that permit to reallocate receivers to wavelengths according to long-term variations in the traffic pattern. The proposed reconfiguration schema address the reconfiguration problem with two different measurements based approaches. While the incoming traffic measurement uses a complex but more accurate measurement mechanism (at the cost of requiring periodic signaling messages), the in-transit measurement approach implements a simpler scheme that, however, may become unreliable under highly loaded conditions. In this second scenario, we showed that it is wiser to run simpler algorithms that try to improve network performance by a trial and error procedure, rather than using an optimization approach that implies a complete redefinition of receiver allocation.

The proposed algorithms show good ability in tracking traffic variations by receiver reconfigurations, accurately balancing performance and reconfiguration costs. As expected, incoming measurement schema perform better than in-transit one, even if the difference is significant only in overloaded conditions. While results were presented only in a specific dynamic traffic setup, they are well representing general trends.

Looking beyond the details of the proposed algorithms, the considered network architecture exhibits very interesting features that well exploits optical technologies.

Acknowledgment

This paper was partially funded through the Italian FIRB project ADONIS, the Italian PRIN project WONDER, and the FP6 European Network of Excellence e-Photon/ONe.

References

[1] S. Yao, B. Mukherjee, and S. Dixit, " Advances in Photonic Packet Switching: an Overview," *IEEE Commun. Mag.*, Feb. 2000, pp. 84-94

[2] J. Cai, A. Fumagalli, and I. Chlamtac, " The Multitoken Interarrival Time (MTIT) Access Protocol for Supporting Variable Size Packets Over WDM Ring Network," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, Oct. 2000, pp. 2094-2104

[3] Y. Cai, R. M. Fortenberry, and R. S. Tucker, " Demonstration of Photonic Packet–Switched Ring Network with Optically Transparent Nodes," *IEEE Photonics Technology Letters*, vol. 6, no. 9, 1994, pp. 1139-1141

[4] J. Fransson, M. Johansson, M. Roughan, L. Andrew, and M. A. Summerfi eld, “ Design of a Medium Access Control Protocol for a WDMA/TDMA Photonic Ring Network,” in *Proc., IEEE GLOBECOM*, vol. 1, Nov. 1998, pp. 307-312

[5] C. S. Jelger and J. M. H. Elmirghani, “ Photonic Packet WDM Ring Networks Architecture and Performance,” *IEEE Communications Magazine*, vol. 40, no. 11, Nov. 2002, pp. 110-115

[6] A. Smiljanić, M. Boroditsky, and N. J. Frigo, “ High–Capacity Packet–Switched Optical Ring Network,” *IEEE Communications Letters*, vol. 6, no. 3, Mar. 2002, pp. 111-113

[7] L. Dittman *et al.*, “ The European IST Project DAVID: a Viable Approach towards Optical Packet Switching,” *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 7, Sept. 2003, pp. 1026-1040

[8] M. Ajmone Marsan, A. Bianco, E. Leonardi, M. Meo, F. Neri, “ MAC Protocols and Fairness Control in WDM Multi-Rings with Tunable Transmitters and Fixed Receivers,” *IEEE J. Lightwave Tech.*, Jun. 1996, pp. 1230-1244

[9] K. V. Shrikhande, I. M. White, D.-R. Wonglumsom, S. M. Gemelos, M. S. Rogge, Y. Fukashiro, M. Avenarius, and L. G. Kazovsky, “ HORNET: A Packet–Over–WDM Multiple Access Metropolitan Area Ring Network,” *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, Oct. 2000, pp. 2004-2016

[10] I. Baldine, G. Rouskas, “ Traffi c Adaptive WDM Networks: A study of Reconfi gurations Issues,”*IEEE J. Lightwave Tech.*, vol. 19, no. 4, Apr. 2001, pp. 433-454

[11] http://www.tlc-networks.polito.it/wonder/

[12] A. Carena, V. De Feo, J. Finochietto, R. Gaudino, F. Neri, C. Piglione, P. Poggiolini, “ RingO: An Experimental WDM Optical Packet Network for Metro Applications,” *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 8, Oct. 2004, pp. 1561-1571

[13] M. Pinedo, “*Scheduling: Theory, Algorithms, and Systems*,” Prentice Hall, 2002

[14] R. L. Graham, “ Bounds on multiprocessing timing anomalies,” *SIAM J. on Applied Mathematics*, 1969, Vol. 17, pp. 416-429

[15] R.E. Tarjan, *Data Structures and Network algorithms*, Society for Industrial and Applied Mathematics, Pennsylvania, Nov. 1993

[16] I. Alfouzan, A. Jayasumana, “ Dynamic reconfi guration of wavelength-routed WDM networks,” in *Proc. IEEE LCN*, Nov. 2001, pp. 477-485

Considering Transmission Impairments in Wavelength Routed Networks

R. Cardillo, V. Curri, M. Mellia
Dipartimento di Elettronica - Politecnico di Torino
Torino, Italy
email: {cardillo,curri,mellia}@mail.tlc.polito.it

***Abstract*— We consider dynamically reconfigurable wavelength routed networks in which lightpaths carrying IP traffic are on demand established.**

We face the Routing and Wavelength Assignment problem considering as constraints the physical impairments that arise in all-optical wavelength routed networks. In particular, we study the impact of the physical layer when establishing a lightpath in transparent optical network. Because no signal transformation and regeneration at intermediate nodes occurs, noise and signal distortions due to non-ideal transmission devices are accumulated along the physical path, and they degrade the quality of the received signal. We propose a simple yet accurate model for the physical layer which consider both static and dynamic impairments, i.e., nonlinear effect that depends on the actual wavelength/lightpath allocation. We then propose a novel algorithm to solve the RWA problem that explicitly considers the physical impairments.

Simulation results show the effectiveness of our approach. Indeed, when the transmission impairments comes into play, an accurate selection of path and wavelength which is driven by physical consideration is mandatory. physical impairments.

I. Introduction

Wavelength Routed (WR) networks are considered the best candidate for the short-term implementation of a high-capacity IP infrastructure, since they permit the exploitation of the huge fiber bandwidth, but do not require complex processing functionalities in the optical domain.

In WR networks, remote high-capacity (electronic) routers are connected through IP-tunnels. IP tunnels are implemented by optical pipes called *lightpaths* that may extend over several physical links. Lightpaths are routed in the optical layer through the physical topology using a single wavelength (we do not assume to exploit wavelength conversion); at intermediate nodes, incoming wavelengths belonging to in-transit lightpaths are switched to outgoing fibers through an optical cross-connect that does not process in-transit information. At the IP layer, lightpaths are seen as data-link channels through which packets are moved from a router to another router toward their destinations following the classic IP forwarding procedure. Therefore, in a WR network, an *IP layer topology* (also called logical topology), whose vertexes are IP routers and whose edges are lightpaths, is overlayed to the *physical topology*, made of optical fibers and optical cross-connects (OXC). If the OXC node implementation requires opto/electronic conversions, the technology is usually called "opaque". Otherwise, if switching of lightpaths is fully performed in the optical domain, the term "transparent" is used. In this second case, the cost of switching a lightpath is almost independent on the transmission data-rate [1]. In this paper we consider the latter technology, which is also the most promising one.

This work was supported by the Italian Ministry for University and Scientifi c Research under the ADONIS project.

Lightpaths can either be semi-permanent [2], or be allocated in on-demand fashion [3]. In the first case a static topology is seen at the IP layer, while in the second case more adaptivity can be gained at the cost of additional complexity both at the optical layer and the IP layer. In this paper we consider dynamically reconfigurable WR networks in which lightpaths are on demand established.

In classic WR networks that support the dynamic allocation of lightpaths according to user requests, the Routing and Wavelength Assignment (RWA) problem must be faced. Indeed, for each connection request, a route across the physical topology must be found, and a wavelength must be selected with the constrains that i) two (or more) lightpaths sharing the same fiber must be identified by two (or more) different wavelengths (also called "wavelength integrity constraint") and ii) a lightpath must be identified by the same wavelength on all the physical fibers along the path (also called "wavelength continuity constraint"). If such a path/wavelength exists, a point-to-point lightpath is established for the duration of the connection. On the contrary, the connection may be blocked given the limited number of wavelengths supported by fibers and OXCs. The goal of the RWA is therefore to minimize the connection blocking probability, and several algorithms have been proposed to address this problem [4].

RWA problem is a classic problem in the context of wavelength routed networks. However, despite several solutions have been proposed, most of them fail to consider the impact of the physical layer on the data transmissions. Indeed, in the definition of the RWA problem, only the *availability* of a wavelength is considered as constraint in the formulation of the problem itself. Considering opaque networks, this is a realistic assumption, as the optical signal is regenerated ad each node, and transmission impairments are therefore compensated at each node. But this is not anymore the case when transparent optical networks are considered.

In a transparent all-optical network, because no signal transformation and regeneration at intermediate nodes occurs, noise

and signal distortions incurred due to non-ideal transmission devices are accumulated along the physical path, and they degrade the quality of the received signal. Noise accumulation actually decreases the Optical Signal to Noise Ratio ($OSNR$) increasing the corresponding Bit Error Rate (BER). Distortions due to fiber propagation modify the shape of the received pulse inducing performance impairments equivalent to a reduction of the $OSNR$. In this paper, besides considering the noise accumulation, we evaluate the impact of the linear and nonlinear fiber propagation with the purpose to obtain an equivalent $OSNR$ characterizing each lightpath of the considered transparent optical network. If for a certain lightpath the $OSNR$ is too low, the corresponding BER may exceed the maximum tolerable BER imposed by the transmission techniques employed. In that case the lightpath becomes not usable and such an information must be taken into account by the RWA algorithms. The $OSNR$ information can be also used as *soft* parameter giving a weight of the goodness of the a lightpath allowing to implement RWA algorithms based on the choice of the lightpath with the best $OSNR$ among all the usable ones.

In this paper, we consider a transparent optical network, in which lightpath requests are dynamically set-up. When solving the RWA problem, we explicitly take into account the physical impairments imposed by the optical layer. In particular, for the first time to the best of our knowledge, we consider the effect of *nonlinearities* which arise when considering dynamic wavelength allocation on optical fibers. In particular, nonlinearities strongly depend on the current allocation of wavelength on a given fiber (and path), and therefore on the current status of allocated lightpaths on the top of the physical topology. This intuitively affect the RWA problem solution of new lightpath request: the selection of a suitable path and suitable wavelength may fail to meet the minimum transmission requirement. But it may also affect already established lightpaths whose transmission properties are negatively affected by the new establishing lightpath. We therefore propose a novel routing and wavelength assignment algorithm (called Best-OSNR) which explicitly tries to minimize the impact of physical impairments.

In the remaining of the paper, Section II describes the physical layer model used to evaluate the transmission quality of a lightpath, including a brief comparison with related work. Section III focuses on the RWA algorithm adopted in this paper whose performance results are presented in Section IV. Finally, Section V summarize our findings.

II. Physical model

In order to analyze the evolution of the electromagnetic signals through a transparent optical network based on the Wavelength Division Multiplexing (WDM) technique, the wave equation for the fiber optic propagation should be solved for every optical link. Since the optical fiber is a nonlinear medium, the wave equation that regulates the propagation is the so called Nonlinear Shroedinger Equation (NLSE) [5] whose expression is:

$$\frac{\partial A(z,t)}{\partial z} = -\alpha A(z,t) + \jmath\frac{1}{2}\beta_2\frac{\partial^2 A(z,t)}{\partial t^2} - \jmath\gamma\left|A(z,t)\right|^2 A(z,t) \quad (1)$$

where $A(z,t)$ is the modal amplitude of the electromagnetic field propagating in the optical fiber, α is the fiber loss coefficient, β_2 is the dispersion coefficient, γ is the nonlinear coefficient, and z and t are the propagation direction and time, respectively. Note that $A(z,t)$ must include all the modulated signals associated to the wavelengths in use because the nonlinear nature of the problem does not allow to solve separately - wavelength by wavelength - the signal propagation in optical fibers. Besides the model for the propagation of optical signals through the fiber, the other component that must be accurately considered is the optical amplifier, e.g., the Erbium-Doped Fiber Amplifier (EDFA). EDFA's are used to recover the fiber loss introduced by the fiber spans but impair the system performance by introducing a certain amount of noise, that is called Amplified Spontaneous Emission (ASE) Noise. Given the amount of gain G and the spontaneous emission factor n_{sp}, the power spectral density of noise introduced by the amplifier is [6]:

$$G_{ASE}(f) = 2\, n_{sp}(G-1)hf \quad (2)$$

where h is the Planck constant and f is the operation frequency.

As well as the transmission components, i.e., fiber and amplifiers, the transmitters and receivers should be modeled in order to include in the performance analysis their effects and potential system impairments.

The other network blocks to be modeled are the passive components such as filters, and, in general, all the elements performing optical network operations. For instance, the add-drop multiplexers and the optical cross-connects.

Due to the nonlinear nature of Eq. (1), the evolution of the optical signals along a transparent optical network should be studied as a single complex problem. Eq. (1) should be solved simultaneously for all the fiber links considering the boundary conditions, i.e., transmitters and receivers, and, in general, network nodes. Furthermore, Eq. (1) does not admit analytical solutions, therefore it must be integrated numerically using simulators that typically are based on the *Split-Step Fourier Method* [7], [8]. It means that the performance evaluation of a single network configuration could require a relevant computational effort, e.g., hours of CPU time with the present state-of-the-art computers. Hence, it is not possible to setup a RWA analysis that requires to evaluate the network performance for possible millions different network configurations, i.e., millions extremely time consuming simulations of the physical layer.

In order to overcome the computational limits introduced by the complexity of the exact analysis of the physical level of transparent optical networks, many approximated solutions have been presented in the technical literature.

In [9], [10], the authors consider independently the impairments due to the effect of Polarization Mode Dispersion (PMD) and accumulated ASE noise. The authors considered

the use of Raman amplifiers besides EDFAs. The analysis is done for each lightpath and they consider that lightpath performs well if both the requirements in terms of noise accumulation (ASE) and PMD are satisfied. In these works the effect of fiber nonlinearities is not considered: it implies neglecting the fundamental tradeoff between increasing of transmitted power to overcome noise impairments and limiting the power to avoid the impact of nonlinearities. Similarly, in [11], [12] the authors considered only the impairments of optical ASE noise introduced by the in-line EDFAs and of electrical noise of the receivers.

We target our analysis to the inclusion in performance evaluation of lightpaths the effect of accumulated ASE noise, linear and nonlinear propagation. To the best of our knowledge this is the first time nonlinear effects are included in the performance evaluation of physical layer of optical networks in order to drive the RWA algorithms with the physical impairments on each lightpath. The simplified model we propose is based on the separation of the effects impairing the signal propagation in order to evaluate the Optical Signal-to-Noise Ratio ($OSNR$) penalty induced by each effect. We start from the assumption that the performance in terms of Bit Error Rate (BER) of an optical link based on the optical amplification is well approximated by :

$$BER \approx \frac{1}{2}\mathrm{e}^{-\eta\,\mathrm{OSNR}} \tag{3}$$

where η is a coefficient assuming values in $[0,1]$ that takes into account how close to the ideal one is the receiver used; $\eta = 1$ for the ideal receiver based on the optical filter matched to the transmitted pulse. Using Eq. 3 we neglected the influence of receiver electric noise. It is a reasonable assumption for optical networks based on the optical amplification, since the ASE noise is typically widely prevalent with respect to the electric noise. In case of studying networks without an extensive use of optical amplification, Eq. 3 can be replaced by a more complex one including the electric noise without varying the general structure of the presented analysis.

For the optimal receiver, the exact expression can be analytically derived and it is [13]:

$$BER = \frac{1}{2}\left\{\mathrm{e}^{-\phi}\left(1+\phi\right) + 1 - \mathrm{Q}_2\left(\sqrt{8\ \mathrm{OSNR}}, \sqrt{2\phi}\right)\right\} \tag{4}$$

where Q_2 is the Marcum Q-function of order 2 [13] and ϕ is the normalized decision threshold that must be optimized for each value of the $OSNR$. Eq. 3 derives from a fitting of Eq. 4. The $OSNR$ is given by:

$$OSNR = \frac{P_S}{P_N} \tag{5}$$

where P_S is the power of the modulated signal carrying the information and P_N is the overall power of the ASE noise introduced by the in-line optical amplifiers, i.e.,

$$P_N = \sum_{i=1}^{i=N_{span}} 2n_{sp,i}(G_i - 1)hfB_n \tag{6}$$

where M is the number of amplifiers for the lightpath under analysis, $n_{sp,i}$ is the spontaneous emission factor for the i-th amplifier, G_i is the gain for the i-th amplifier and B_n is the equivalent noise bandwidth of the receiver.

Using Eq. 3, BER of a lightpath is directly related to the $OSNR$. Therefore, if we define BER_{max} as the maximum error probability tolerable by the transmission technique used by the network under analysis, a lightpath can be considered as *in service* if presents a BER smaller than BER_{max}. Alternatively, the lightpath is *in service* if

$$OSNR\ >\ OSNR_{min}\ =\ \frac{1}{\eta}\ln\left(\frac{1}{2\ BER_{max}}\right), \tag{7}$$

therefore, to distinguish between different lightpaths within the application of a RWA, the $OSNR$ is a parameter to be maximized in order to minimize the error rate. Furthermore, the use of a certain lightpath must be discarded if the related $OSNR$ results to be smaller than $OSNR_{\mathrm{min}}$. This approach is the one we followed in order to implement the RWA algorithms described in details in Sec. III.

In case of propagation impairments besides the ASE noise accumulation, performance for each lightpath can be still evaluated using Eq. 3, substituting the Optical Signal-to-Noise Ratio with an *equivalent* coefficient $OSNR_{eq}$ that wants to include the effects of the considered impairments. Therefore, the expression of $OSNR_{eq}$ in dB units can be described as follows:

$$OSNR_{eq,dB} = OSNR_{dB} - OSNR_{pen,l} - OSNR_{pen,nl} \tag{8}$$

where $OSNR_{dB}$ is 10 times the logarithm of the $OSNR$ value due to the ASE noise accumulation expressed in dB units. $OSNR_{pen,l}$ and $OSNR_{pen,nl}$ are the penalties - expressed in dB units - introduced by the linear (dispersion, PMD) and nonlinear (Kerr effect) propagation effects, respectively. $OSNR$ penalties are caused by the pulse distortions induced by the propagation effects that impairs the decision signal - eye-diagram closure - inducing a performance impairments equivalent to a certain amount of extra noise. Either ASE noise accumulation, either the eye-diagram closure due to the propagative linear effects act separately on different wavelength independently on the number of wavelengths in use on the fiber span under analysis. Therefore, $OSNR_{dB}$ and $OSNR_{pen,l}$ depend only on the path π and on the wavelength λ, while $OSNR_{pen,l}$ depends also on the number of wavelengths N_λ actually turned on - for the considered network configuration - per each fiber span used by the lightpath λ. It means that the overall $OSNR_{eq,dB}$ function must be evaluated per each lightpath per each possible network configuration and not just for each lightpath independently of the network configuration. It is clearly understandable how the problem complexity dramatically grows with the inclusion of the propagation nonlinear effects.

A rigorous analysis of the physical effects on the performance of an optical network should require the simulation of the entire network for every possible configuration that the RWA algorithms may take into account. As previously

explained, such a task should require millions of hours of computation time. Hence, we decided to evaluate separately the ASE noise accumulation, the impairments of linear effects and the impairments of nonlinear effects. Here is the description of the approximations we used in order to derive the impairments due to the considered effects.

- *ASE Noise accumulation.*
  The graph describing the network is analyzed in order to individualize the amplifiers, fiber losses, and lumped losses. Then, for each physical path, the accumulated ASE noise is evaluated together with the signal level. As a result, each lightpath is targeted with the corresponding $OSNR_{ASE}$.
- *Impairments of linear propagation effects.*
  In order to evaluate the impairments of linear effects (PMD and dispersion) for each lightpath is evaluated the amount of accumulated dispersion and PMD. Then penalties are evaluated according to the results presented in [14], [15]. If the dispersion compensation is applied and the overall PMD is summed with respect to the bit duration, impairments of linear propagation effects can be neglected. In general from the analysis of linear propagation the penalty $OSNR_{pen,l}$ is derived. In case of linear effects negligible, $OSNR_{pen,l} = 0$ dB.
- *Impairments of nonlinear propagation effects.*
  Nonlinearities in optical fibers are caused by the physical effect called *Kerr Effect*. Its effect is a locale change of the refractive index as a function of the overall propagating optical power. *Kerr effect* induces well know impairments on the propagating signal that can be classified as [5]: Self Phase Modulation (SPM), i.e., the modulation of the phase of a signal induced by variation in time of the power of the signal itself; Parametric Gain (PG), i.e., the transfer of power from a signal to the adjacent spectral components; Cross-Phase Modulation (XPM), i.e., the modulation of the phase of a signal induced by variation in time of the the overall power of the comb of WDM channels propagating in the fiber; Four Wave Mixing (FWM), i.e., the generation of spurious tones at new frequencies. In commercial systems, the nonlinear limiting effect is typically the XPM [16], [17], [18], [19]. Therefore, we focus our attention in the evaluation of the $OSNR$ penalty due to the XPM. In order to pursue such a target, we assume that this penalty is a monotone increasing function with number of wavelength actually in use on the fiber and with power per channel. Whereas we assume it decreases with the increasing of dispersion and channel spacing. These are well known general behaviors, but the exact expression of the function is not known. Therefore, we performed a series of Monte-Carlo simulations on a defined test-link using the optical system simulator *OptSim*^{TM}. From the results of these simulations we deduced an empirical function giving $OSNR_{pen,nl}$ from the knowledge of the fiber characteristics, the number of wavelengths turned on, the length of the fiber span and the transmitted power. From this function, knowing the network characteristics from its graph description and the wavelength assignment $OSNR_{pen,nl}$ is evaluated. Of course this penalty depends on the dynamic reconfiguration of the network because it depends the number of wavelength in use per each fiber and on their spectral assignment.

Considering the separate evaluation of impairments due to the considered effects, for each possible lightpath of the network, the physical layer analysis was able to provide to the RWA algorithms a function $OSNR(\pi, \lambda) = OSNR_{ASE} - OSNR_{pen,l} - OSNR_{pen,nl}$. The value of such a function, given a path π and a wavelength λ, is a constant for a static network, while changes in case of dynamic re-configuration of the network because it depends also on the number of wavelengths actually in use on each fiber span.

III. RWA ALGORITHMS

To gauge the impact of physical impairments on the RWA solution, we compare the performance of traditional RWA algorithms to the one obtained by a novel algorithm which considers the physical impairments when solving the RWA problem. We first describe traditional algorithms while also introducing the notation, and then describe the novel algorithm.

A. Traditional Algorithms

To solve the RWA problem, we selected two algorithms that were shown to give good performance: the *First Fit-Minimum Hop* (FF-MH) and *First Fit-Least-Congested* (FF-LC) [4]. These are traditional algorithm, which split the RWA problem into two simpler sub-problems: first a suitable path is selected, and then a suitable wavelength is allocated if available on the selected path.

In more details, when searching for available wavelengths on a given path, a First-Fit strategy is used: a lower numbered wavelength is considered before higher numbered wavelengths, and the first available wavelength is then selected by both algorithms.

As regards the path selection, for each source/destination pair, the FF-MH algorithm considers only one possible path, which has been preselected to be the minimum hop path. In case more than one minimum hop path is present between the same source/destination pair, only one is considered (in particular, the first minimum hop path found is selected). Dijkstra algorithm can be used to obtain the minimum hop path.

The FF-LC algorithm, instead, considers a pre-ordered list of available paths for each source/destination pair. Paths are dynamically sorted, so that always the least congested path is tested first. The "congestion" metric counts the number of wavelengths already used on a fiber, so that the path with the largest number of unused wavelengths is chosen. In case more than least congested path exists, (one at ramdom among) the shortest path will be selected For the purpose of providing a formal description of the algorithms, we use a standard graph theory formalism. Thus, we refer to the generic physical

network as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertexes (nodes, in our case), and $\mathcal{E}$ is the set of edges (links)[1].

A path $\pi(s,d)$ of length $n(\pi(s,d)) = ||\pi(s,d)||$ is defined as a sequence of n distinct edges e_i joining s and d, where $s, d \in \mathcal{V}$, $e_i \in \mathcal{E}$, $\pi(s,d) = \{e_1, e_2, ..., e_n\}$.

Let $\Pi(s,d) = \{\pi_i(s,d)\}$ be the set of available loop-free paths from node s to node d. Let $W(e_i)$ be the number of wavelength already allocated on link e_i.

Given those definitions, the Minimum Hop routing will select the path $\pi^{MH}(s,d)$ such that

$$\pi^{MH}(s,d) = \min_{\pi \in \Pi(s,d)} n(\pi)$$

On the contrary, the Least Congested path $\pi^{LC}(s,d)$ will be selected such that:

$$\pi^{LC}(s,d) = \min_{\pi \in \Pi(s,d)} \left(\max_{e_i \in \pi} (W(e_i)) + \frac{1}{c} n(\pi) \right)$$

The constant c must be selected such that

$$c > \max_{\pi \in \Pi(s,d)} (n(\pi))$$

Notice that the MH path selection can be performed off-line, being $n(\pi)$ constant with respect to wavelength allocation. On the contrary, the implementation of the LC path selection criterion requires each route to be selected for each lightpath request, thus entailing a much larger complexity, both in term of computational power and signaling.

Once a path has been selected, the wavelength allocation is performed using the first-fit approach by both algorithms. Let $\Lambda(e_i) = \{\lambda_j, j = 1, \ldots, L\}$ be the ordered set of supported wavelength on link e_i. Let $F(\lambda_j(e_i))$ take the value 0 if the j-th wavelength is free on link e_i, 1 otherwise. Then, the set $\mathcal{F}$ of available wavelength on path $\pi(s,d)$ is defined as

$$\mathcal{F} = \{\lambda_j \text{ such that } F(\lambda_j(e_i)) = 0 \ \forall e_i \in \pi(s,d)\}$$

Then, lightpath request will be allocated using wavelength $\hat{\lambda}$ on path $\pi(s,d)$ such that:

$$\hat{\lambda} = \min_j (\lambda_j \in \mathcal{F})$$

B. B-OSNR algorithm

Traditional algorithms fails to consider the physical impairments that may affect the transmission on a given path/wavelength. We therefore propose a novel algorithm, called *Best-Optical Signal Noise Ratio* (B-OSNR), which will jointly assign to a given request a path and a corresponding wavelength. In particular, the path/wavelength solution which will present the maximum OSNR will be selected. Let $OSNR(\pi(s,d), \lambda_j)$ be the OSNR on wavelength λ_j on path $\pi(s,d)$. $OSNR(\pi(s,d), \lambda_j) = -\infty$ if λ_j is not usable on path $\pi(s,d)$. Then, the path $\pi^{OSNR}(s,d)$ and the wavelength λ^{OSNR} will be selected such that:

$$(\pi^{OSNR}(s,d), \lambda^{OSNR}) = \max_{\pi \in \Pi(s,d)} \left(\max_{\lambda \in \Lambda} OSNR(\pi, \lambda) \right)$$

As can be noticed, the B-OSNR algorithm *jointly* assigns a path and a wavelength to a given lightpath request. Its complexity grows linearly with the number of paths and the number of wavelengths that must be checked to find the best solution.

[1]In this paper we interchangeably use the terms 'edges' and 'links' and the terms 'vertexes' and 'nodes'.

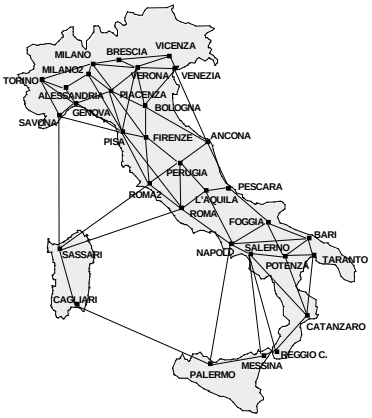


Fig. 1. Physical topology.

IV. Performance Analysis

To gauge the impact of the physical constraints on the routing and wavelength assignment, we developed a simulator which implements all the RWA algorithms described in the previous section, and performs the evaluation of the OSNR as described in Section II. To this purpose, the description of the physical topology by means of a graph $\mathcal{G}$, which includes the definition of fibers, amplifiers, optical cross connects, etc., is given as input. In particular we assumed that the network is cabled using Non-Zero Dispersion Shifted fibers. In order to recover fiber losses we considered to use EDFAs spaced L_{span} km that perfectly recover the loss introduced by the fiber span. We supposed the employed EDFAs are perfectly spectrally equalized and have flat transfer functions, providing the same amount of gain for all the wavelengths. We explored different scenarios analyzing the network behaviors for L_{span} = 40, 60, 80 km. We assumed to use dispersion compensation techniques and that the PMD effect is negligible at the supposed bit-rate of 10 Gbit/s. Therefore, we supposed to be negligible the propagation linear effects focusing our analysis on considering the limiting effects of noise accumulation and impairments of fiber nonlinearities. Regarding the effects of passive components performing network operations within the nodes (filters, add-drop multiplexers, optical cross-connects, etc...) we considered the extra losses that they introduce. We did not include they filtering effect.

A description of the traffic pattern completes the scenario whose performance indexes will be analyzed during the

simulation. The traffic description includes a traffic matrix $T = \{t_{s,d}\}$ whose elements $t_{s,d}$ represent the fraction of lightpath requests from node s to node d. Lightpath requests are generated according to a Poisson process of rate $\rho t_{s,d}$, in which ρ represent the average arrival rate in connection per seconds. Connection holding time is exponentially distributed, with average set to 1 which therefore fixes the time reference in the simulation.

Once a connection request is generated, the corresponding RWA problem is solved according to the selected algorithm. If a path π and a free wavelength λ are available, the corresponding OSNR is evaluated, and if it is above to a given $OSNR_{\min}$ threshold, then the lightpath is accepted, and the corresponding λ is allocated on all links of path π. Otherwise, the lightpath request is blocked and no reservation occurs. Allocated resources will then be released at the end of the connection lifetime.

As performance indexes, the average blocking probability P_b is evaluated. In particular, to asses the impact of the OSNR limitation, the simulator evaluates the blocking probability due to physical impairments (P_b^{OSNR}) and the blocking probability due to lack of available wavelength (P_b^{λ}). The first one is defined as the ratio between the number of lightpath requests which were blocked because the OSNR level on the selected (free) wavelength was below the minimum threshold with respect to the total number of lightpath requests. P_b^{λ} accounts for blocked lightpath requests due to lack of available free wavelength. Clearly $P_b = P_b^{OSNR} + P_b^{\lambda}$.

In the simulation result reported in this paper, we considered as physical topology the Italian Optical Network sketched in Fig. 1 which was derived from a possible evolution of the Telecom Italia network topology. Nodes reflect the real position of cities and link lengths reflect the real distances among cities. All fiber and nodes are assumed to be physically equal. Maximum number supported wavelength L is set to 16.

We consider three different physical configurations, which differ by the maximum *span* of fibers that is admissible without requiring regeneration, i.e, the maximum length of optical fiber between two adjacent amplifiers. In particular, spans of 40 km, 60 km, 80 km will be considered. The longer is the fiber span, the larger is the amount of gain required to recover fiber losses. Hence, the larger is the amount of noise introduced by the amplifiers. To restore the target $OSNR$ a larger amount of transmitted power can be employed, but with the increasing of transmitted power the effect of nonlinearities progressively grows inducing a stronger impairment on performance.

Regarding the traffic pattern, we consider in this paper a simple uniform traffic, in which all $t_{s,d} = 1$. We set $OSNR_{\min} = 20dB$, corresponding to $BER = 10^{-12}$ with an $OSNR$ margin of about 4 dB. During the path search phase, the sets $\Pi(s,d)$ are build by considering only those paths whose minimum OSNR is larger than $OSNR_{\min}$. The minimum OSNR of a given path is evaluated by not considering the nonlinearities, i.e., by considering $OSRN(\pi,\lambda)$ when no other lightpaths is established on any other paths. A limited

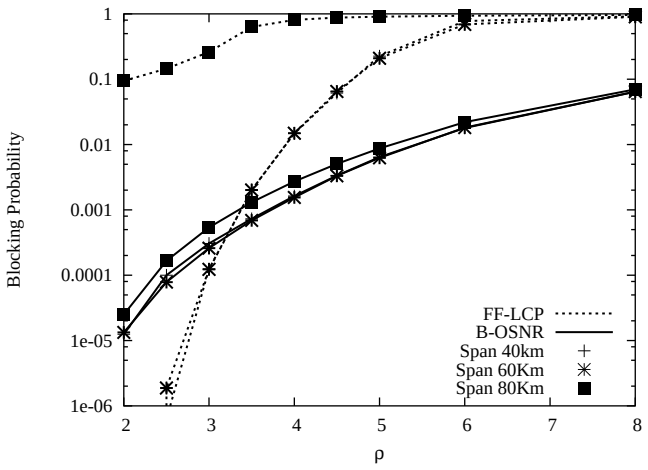


Fig. 2. Total average blocking probability versus offered load for different algorithms. Fiber span (L_{span} of 40 km, 60 km, 80 km are presented.

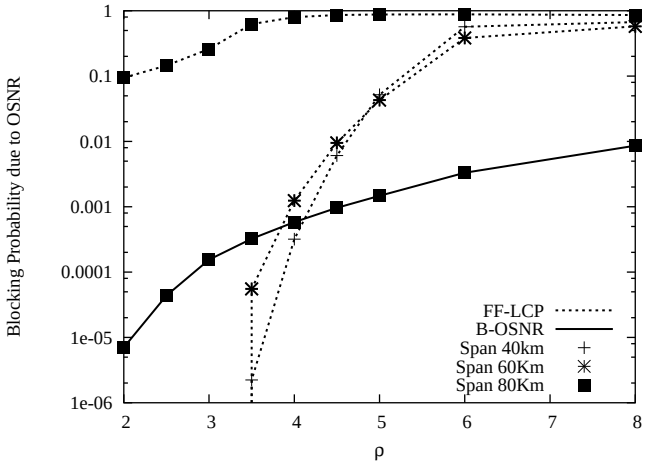


Fig. 3. Average blocking probability due to OSNR impairment versus offered load for different algorithms. Physical span of 40km, 60km, 80km are presented.

number of path is considered for each source destination pair, so that the complexity of finding π^{LC} and π^{OSNR} is limited: paths in $\Pi(s,d)$ are sorted in decreasing number of hops, and then only the first 30 paths are considered[2].

Finally, to get accurate results, each simulation was ended when the performance indices were such that the 95% confidence interval was within 5% of the point estimate.

A. Blocking probability

In Figures reported in this section, dashed lines refers to the blocking probability obtained when the FF-LC algorithm is considered, while solid lines report results considering the B-OSNR. Different points are used to highlight different span values.

Figure 2 plots the average blocking probability versus offered load. Comparing the results obtained by the FF-LC or

[2]We considered larger sets of paths, but without observing major differences on the results.

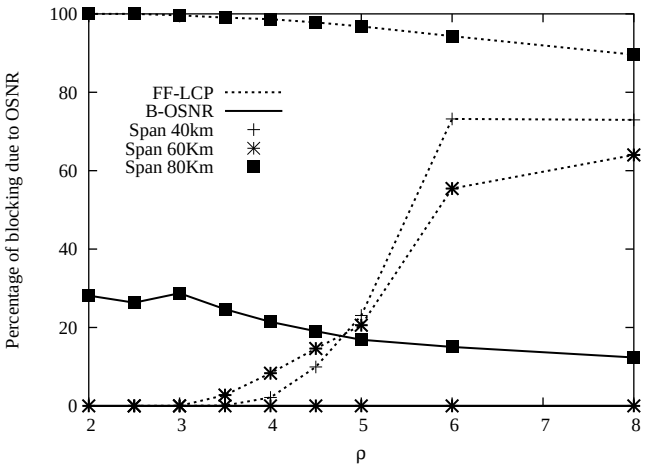


Fig. 4. Percentage of blocking probability due to OSNR degradation versus offered load for different algorithms. Physical span of 40km, 60km, 80km are presented.

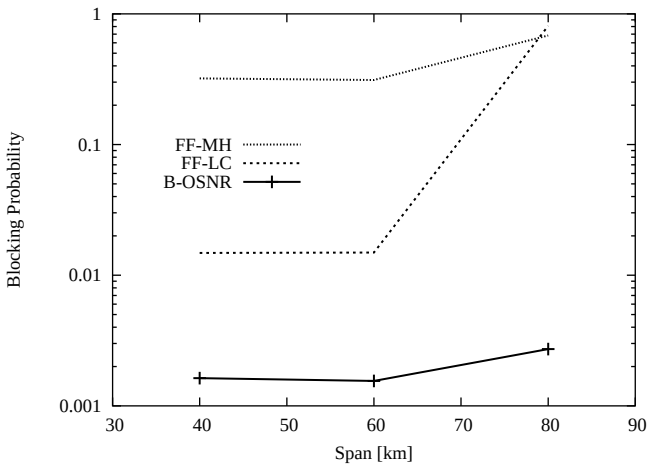


Fig. 5. Total average blocking probability versus physical span for different algorithms. Offered load set to 4.

the B-OSNR algorithm, it can be noticed that when the impact of the OSNR introduced by the physical layer is negligible, the FF-LC algorithm performs better than the B-OSNR approach. Indeed, for small values of the offered load and for small span values the FF-LC takes the lead, while for both larger values of ρ and for span value set to 80km, the B-OSNR algorithm clearly outperforms the FF-LC approach.

The intuition behind this is that the better allocation of wavelength used by the FF approach tends to better pack wavelength usage so that the change of obtaining a free wavelength is larger. On the contrary, the wavelength allocation performed by the B-OSNR algorithm tends to spread out the wavelength as much as possible, so to minimize the noise introduced by adjacent channels. This leads to a larger blocking probability when the cause of blocking is due to lack of wavelength.

On the contrary, for larger values of the offered load, the effects due to nonlinearities clearly affect the blocking probability faced by a FF-LC algorithm. Indeed, its more compact wavelength allocation criterion maximizes the noise due to interfering wavelengths. Therefore, when the blocking probability is largely due to physical impairments, the FF-LC algorithm cannot find any good solution.

Similarly, considering different network span configuration, the B-OSNR approach shows little differences, showing that it is able to overcome physical configuration which offers worse OSNR. On the contrary, the FF-LC algorithm present almost identical results when 40km and 60km span long networks are considered, while the 80km span network performance are much worse. This is due to the path selection choice, which allows the FF-LC algorithm to select longer paths which will cause larger transmission noise that will be accumulated along the path itself, finally resulting in a blocked lightpath due to lack of OSNR.

To better highlight this effects, Figure 3 plots the blocking probability due to physical impairments. Considering the 40km and 60km span, the B-OSNR presents no blocking due to lack of OSNR, while the FF-LC algorithm shows a steep increase of the blocking probability due to transmission impairments. This confirms the intuition the the nonlinearities faced by the FF-LC wavelength allocation (and path selection) are the largest cause of blocking.

Similarly, considering the 80km span long network, the FF-LC algorithm is not able to find any suitable path and wavelength solution to the RWA problem even when the nonlinearities are small, i.e., when then offered load is small so that few lightpath are present at the same time.

Finally, to gauge the ratio between the blocking due to wavelength lack or to OSNR lack, Figure 4 plot the percentage of blocking probability due to OSNR degradation versus the offered load. It confirms the previous observation, by showing that the B-OSNR algorithm is only marginally affected by the lack of OSNR. On the contrary, the FF-LC approach faces the majority of blocking probability because the selected wavelength and path cannot offer an adequate OSNR level.

To better observe the effect of nonlinearities on the blocking probability, Figure 5 plots the total average blocking probability versus the span for offered load equal to 0.4. The plot also reports results considering the FF-MH algorithm. Its performance are in general limited when compared to algorithms that allow to test more than a single path, as already well-known [4]. The B-OSNR algorithm presents the best results, about one or two order of magnitude better than results presented by classic algorithms which fail to consider physical impairments.

In particular, considering span smaller than 80km, the static impairments due to the physical layer are negligible, as no major differences are observed moving from 40km long span to 60km long span physical configuration. Increasing the span length to 80km, on the contrary the blocking probability of the FF-LC algorithm increases. This performance downgrade is largely due to the selection of possibly longer and more noisy paths. The FF-MH algorithm is little affected by this, as it always select the minimum hop path which in general is also the shortest one and therefore the one which presents the smaller noise due to linear effects. Still, a little increase

in the blocking probability is due to the smaller static OSNR ratio which, combined with the nonlinearity noise, increases the chance of observing a OSNR larger then $OSNR_{\min}$.

V. Conclusions

In this paper we considered a transparent optical network. By using wavelength routed technology, we considered the routing and wavelength assignment problem under transmission impairments. We considered a dynamic scenario, in which lightpath requests arrive and leave the network. Because in transparent optical network no signal transformation and regeneration at intermediate nodes occurs, noise and signal distortions due to non-ideal transmission devices are accumulated along the physical path, and they degrade the quality of the received signal. This affects the availability of the optical channel, and therefore must be considered during the RWA solution. We presented a novel simple physical model to evaluate the OSNR ratio which considers both static noise due to optical components and nonlinearity effects due to the current wavelength allocation and usage.

We then presented a novel algorithm which tries to minimize the effect of transmission impairments when solving the RWA problem for each lightpath requests. Simulation results showed that, when the transmission impairments comes into play, an accurate selection of path and wavelength which is driven by OSNR is mandatory.

In particular, both static effects and nonlinearities can largely affect the blocking probability: the first one depend on the physical configuration and must be considered for any offered load to the network; the latter one rapidly degrades the quality of the transmission layer when the number of lightpath already established is large, i.e., when the offered load is higher. In such scenarios, the proposed B-OSNR algorithm outperforms traditional algorithms which fails to consider the physical impairments.

VI. Acknowledgment

The authors would like to thank RSoft Design Group, Inc for supplying the simulation tool `OptSim`$^{\mathrm{TM}}$.

References

[1] R. Ramaswami, K.N. Sivarajan, *"Optical Networks: A Practical Perspective,"* The Morgan Kaufmann Series in Networking, February 1998.

[2] B.Mukherjee, D.Banerjee, S.Ramamurthy, A.Mukherjee, "Some Principles for Designing a Wide-Area WDM Optical Network," *ACM/IEEE Transactions on Networking*, Vol.4, n.5, pp. 684-695, Oct. 1996.

[3] H.Zang, J.P.Jue, L.Sahasrabuddhe, S.Ramamurthy, B.Mukherjee, "Dynamic Lightpath Establishment in Wavelength-Routed WDM Networks," *IEEE Communications Magazine*, Sept. 2001.

[4] H.Zang, J.P.Jue, B.Mukherjee, "A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks," *SPIE Optical Networks Magazine*, vol. 1, no. 1, Jan. 2000.

[5] Govind P. Agrawal, "*Nonlinear fiber optics*," Academic Press, San Diego, 2^{nd} edition, 1989.

[6] E. Desurvire, "*Erbium-Doped Fiber Amplifiers*," John Wiley & Sons, New York, 1994.

[7] A. Carena, V. Curri, R. Gaudino, P. Poggiolini and S. Benedetto, "A time-domain optical transmission system simulation package accounting for nonlinear and polarization related effects in fiber", *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 4, pp. 751-765, May 1997.

[8] G. Boggio, M. Burzio, N. Portinaro - ARTIS Software J. Cai, I. Cerutti, A. Fumagalli, M. Tacca, L. Valcarenghi, A. Carena, R. Gaudino, "NetworkDesigner - Artifex - OptSim: a suite of integrated software tools for synthesis and analysis of high speed networks," *Optical Networks Magazine*, September-October 2001, pp 27-41.

[9] Y. Huang, A. Gencata, J. P. Heritage, B. Mukherjee, "Routing and Wavelength Assignment with Quality-of-Signal Constraints in WDM Networks," *European Conference on Optical Communications (ECOC '02)*, Copenhagen, Sept. 2002.

[10] Y. Huang , W. Wen , J. P. Heritage, B. Mukherjee "Signal-Quality Consideration for Dynamic Connection Provisioning in All-Optical Wavelength-Routed Networks" *Optical Networking and Communications conference (OptiComm)*, Dallas, TX, Oct. 2003

[11] B. Ramamurthy, D. Datta, H. Feng, J. P. Heritage, and B. Mukherjee, "Impact of Transmission Impairments on the Teletraffic Performance of Wavelength-routed Optical Networks", *IEEE/OSA Journal of Lightwave Technology*, vol. 17, no. 10, pp. 1713-1723, October 1999.

[12] R. Sabella, E. Iannone, M. Listanti, M. Berdusco, and S. Binetti, "Impact of Transmission Performance on Path Routing in All-optical Transport Networks", *IEEE/OSA Journal of Lightwave Technology*, vol. 16, pp. 1965-1971, Nov. 1998

[13] J. G. Proakis, *Digital communications*, 2^{nd} edition, New York, McGraw-Hill, 1989.

[14] I. P. Kaminow, T. L. Koch, *Optical Fiber Telecommunications IIIA*, San Diego, Academic Press, 1997.

[15] L. Kazovsky, S. Benedetto, A. Willner, *Optical Fiber Communication Systems*, Boston, Artech House, 1996.

[16] S. Ten, K.M. Ennser, J.M. Grochocinski, S.P. Burstev, and V.L. da Silva, "Comparison of Four-Wave Mixing and Cross Phase Modulation penalties in dense WDM systems," *Proceedings of OFC 1999* ThC4, Vol. 3, pp. 43-45, 1999.

[17] G. Bellotti, M. Varani, C. Francia, and A. Bononi, "Intensity distortion induced by cross-phase modulation and chromatic dispersion in optical-fiber transmissions with dispersion compensation," *IEEE Photonics Technology Letters*, Vol. 10, pp. 1745-1747, Dec. 1998.

[18] A. V. T. Cartaxo, "Cross-phase modulation in intensity modulation-direct detection WDM systems with multiple optical amplifiers and dispersion compensators," *IEEE/OSA Journal of Lightwave Technology*, Vol. 17, pp. 178 -190, Feb. 1999.

[19] A. Carena, P. Cobetto Ghiggia, V. Curri, "A Novel Model of Cross Phase Modulation in Long-Haul WDM Optical Systems," Proceedings of SubOptic 2004, Monaco (Montecarlo), March 29th - April 1st 2004.

RingO: An Experimental WDM Optical Packet Network for Metro Applications

Andrea Carena, *Member, IEEE*, Vito De Feo, *Student Member, IEEE*, Jorge M. Finochietto, *Student Member, IEEE*, Roberto Gaudino, *Member, IEEE*, Fabio Neri, *Member, IEEE*, Chiara Piglione, *Student Member, IEEE*, and Pierluigi Poggiolini, *Member, IEEE*

***Abstract*—This paper presents Ring Optical Network (RingO), a wavelength-division-multiplexing (WDM), ring-based, optical packet network suitable for a high-capacity metro environment. We present three alternative architectural designs and elaborate on the effectiveness of optic with respect to electronic technologies, trying to identify an optimal mix. We present the design and prototyping of a simple but efficient access control protocol, based upon the equivalence of the proposed network architecture with input-buffering packet switches. We discuss the problem of node allocation to WDM channels, which can be viewed as a particular optical network design problem. We, finally, briefly illustrate the fault protection properties of the RingO architecture.**

The main contribution of this paper is the identification and experimental validation of an innovative optical network architecture, which is feasible and cost effective with technologies available today, and can be a valid alternative to more consolidated solutions in metro applications.

***Index Terms*—Metropolitan area networks, optical packet networks, optical testbeds, wavelength-division-multiplexing (WDM) rings.**

I. Introduction

THE MARKET segment of metropolitan high-speed networks is alive despite the current telecom crisis. According to several studies, the provision of low-cost broadband access in metropolitan areas has the potential for fast returns on investments, and can foster the development of new bandwidth-hungry applications, which in turns should lead to the long-sought return to the fast increase of user demands that can revitalize the telecom market.

Metro networks are characterized by high dynamism of traffic patterns, relatively high aggregate bandwidths, and relatively short covered distances. Technical solutions for metro architectures are far from being consolidated, and range from classical circuit-switched synchronous optical network/synchronous digital hierarchy (SONET/SDH) rings, to extensions of traditional high-speed local area networks (LANs), such as the resilient packet ring IEEE 802.17, to switched (multi) Gigabit Ethernets, to Broadband Passive Optical Networks ITU-T G.983, to more innovative optical packet switching proposals. The latter are considered by many researchers the only approach capable of withstanding in the long term the continuous growth of aggregate capacities.

Wavelength-division multiplexing (WDM) is today a well-established technique to exploit the fiber bandwidth in both core and metro networks, and all major vendors in this field offer a wide range of products and commercial solutions. The development of optical technologies for applications beyond point-to-point transmission has instead suddenly slowed down due to the telecom market downfall. Nevertheless, at research and standardization levels, a large effort is being devoted to exploit optical technologies also for the implementation of network functions such as switching, protection, and restoration [1].

Nowadays, the most advanced products essentially provide optical *circuit* switching at the wavelength level (see, for example, [2]), in the sense that end-to-end optical lightpaths are dynamically set up and torn down upon network, or even user requests. On the contrary, the implementation of optical *packet* switching functions [3] (i.e., of an optical layer that can handle and switch data traffic on time scales in the order of microseconds or less) is still at an earlier development stage, although several prototypes and testbeds have already been demonstrated [4]–[6]. This is certainly due to the high technological challenges inherent in dealing with packets directly at the optical level. Indeed, although optical devices allow huge potential in terms of available bandwidth, they do not easily offer substantial features in terms of very fast switching, processing speed, and storage of digital signals, which are instead necessary for packet switching and are very natural and easy to implement in the electronic domain.

Metropolitan area networks are one of the best arenas for an early penetration of advanced optical technologies. Indeed, their large traffic dynamism requires packet switching to efficiently use the available resources; their high-capacity requirements justifies WDM use; and their limited geographical distances lowers the impact of fiber transmission impairments. From a research view point, designing innovative architectures for metro networks often means finding cost-effective combinations of optic and electronic technologies and new networking paradigms that better suit the constraints dictated by available photonic components and subsystems.

Our research group has designed and prototyped network architectures for metro applications, taking an approach based

Manuscript received July 30, 2003; revised March 10, 2004. This work was supported in part by the Italian Ministry for Education, University and Research (MIUR) under the PRIN Projects "RingO" and "Wonder," in part by the FIRB Project "Adonis," and in part by the European Commission under the FP5 IST Project "David."

A. Carena, R. Gaudino, and P. Poggiolini are with the PhotonLab, Dipartimento di Elettronica, Politecnico di Torino, Torino 10129, Italy (e-mail: andrea.carena@polito.it; roberto.gaudiono@polito.it; pierluigi.poggiolini@polito.it).

V. De Feo, J. M. Finochietto, F. Neri, and C. Piglione are with the Dipartimento di Elettronica, Politecnico di Torino, Torino 10129, Italy (e-mail: vito.defeo@polito.it; jorge.finochietto@polito.it; fabio.neri@polito.it; chiara.piglione@polito.it).

Digital Object Identifier 10.1109/JSAC.2004.830479

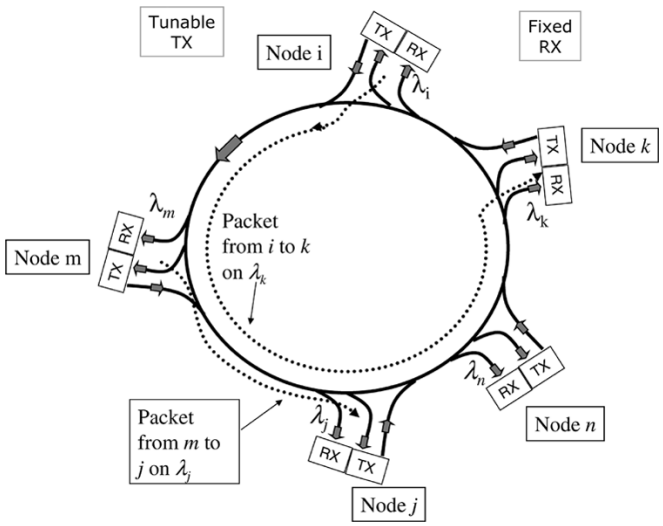


Fig. 1. Architecture of the RingO network.

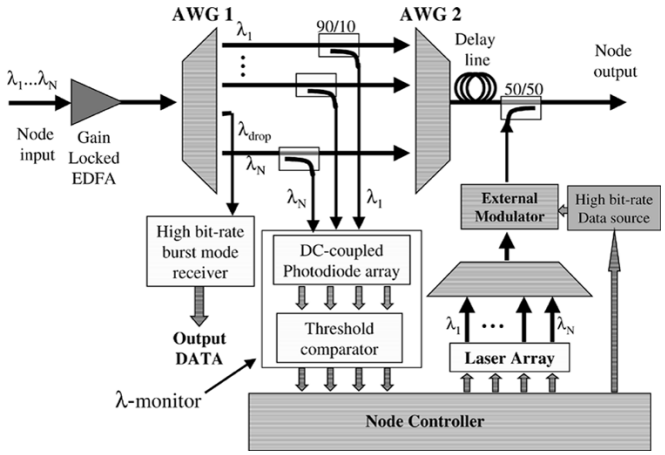


Fig. 2. First structure of RingO nodes.

upon optical packets, but limiting optical complexity to a minimum and trying to use only commercially available components. To best exploit the advantages of available technologies, the bulk of raw data is kept in the optical domain, while more complex network control functions are mostly implemented in the electronic domain. Likewise, neither distributed resource allocation nor contention resolution is performed in the optical domain, thereby taking a radically different perspective with respect to traditional electronic packet-switched architectures.

In this paper, we introduce the rationale, the network architecture and design of the ring optical network (RingO) project, carried out by a consortium of Italian Universities coordinated by the Optical Communications Group (OPTCOM) and the Telecomunication Network Group (TNG) of Politecnico di Torino. The RingO project is focused on experimentally studying the feasibility of a WDM optical packet network based on a ring topology. The presentation will evolve through three different network designs, both to follow the project history, and to ease the description for the reader.

The paper is organized as follows. The RingO general architecture, medium access control (MAC) protocol and node structure are explained in Section II. Section III briefly overviews physical-layer issues related to transmission impairments and network scalability. Then, in Section IV, we describe the current RingO experimental setup, presenting the demonstrator and some details of the node controller hardware implementation. In Section V, we present an interesting evolution of the node design, and discuss problems related to allocating nodes to the available WDM channels. Finally, in Section VI, we briefly discuss fault recovery mechanisms.

II. RingO Architecture

The general architecture of the RingO network is illustrated in Fig. 1, while the structure of a node is depicted in Fig. 2 (and described in more detail in Section II-A). As mentioned above, we will step through three different versions of the network architecture in this paper; they all preserve the same rationale and basic subsystems design.

The first version of the RingO network is based on a unidirectional WDM fiber ring with N network nodes equipped with an interface between the electronic domain and the optical domain. The main features of this first RingO architecture are the following:

- packets transmission is time-slotted and synchronized on all wavelengths; as reference values in RingO, the slot duration is 1 μs and the transmission bit rate is 2.5 Gb/s;
- packets have fixed length corresponding to one time slot: the packet format adaptation, possibly including segmentation/reassembly, or concatenation, is left to higher (electronic) layers of the node protocol, it is outside the scope of this paper;
- the number N of nodes in the network in this first design is equal to the number W of wavelengths (which will be often indicated in the following as "channels"): a given node i is, thus, identified by a wavelength λ_i, it is the only node able to receive this wavelength, and it is also responsible for physically removing it from the ring, using a fixed-wavelength optical drop filter;
- each node is equipped with a tunable transmitter since, in order to communicate to node k, a node must tune its transmitter to send a packet on λ_k, as shown in Fig. 1; tuning times are assumed to be short with respect to the slot duration;
- each node is able to check the state (busy/free) of all wavelengths (a feature called λ-monitoring) on a slot-by-slot basis, and avoids collisions and contentions by electronically queueing input packets, and by accessing channels using a suitable access protocol, as discussed in Section II-B.

In the architecture described above, the fixed relation between a destination node and a wavelength allows a significant simplification on the optical hardware with respect to most of other packet network proposals. First, packet headers are not required, at least for addressing functions, since the destination address is "coded" into the used wavelength. Second, packets do not need to be actively routed along the network, but are simply passively dropped at the destination by the node optical drop filter. As a result, our proposal is able to take advantage of packet statistical multiplexing without requiring optical switches. Third,

λ-monitoring can be obtained by simply measuring the power level in each slot and wavelength, without again requiring the presence of an optical header.

For what regards wavelength tunability requirements, it is easy to understand that full tunability either at the transmitter, or at the receiver, is required to provide full node-to-node connectivity avoiding a multihop operation, that would increase the amount of electronic processing in the network. Tunability is a characteristic feature of optical networks, leading to interesting and well-understood logical topology design and fault recovery approaches, but it is still very costly, specially if very high switching rates are necessary. All RingO designs chose to have fast-tunable transmitters, which are considered to be easier to implement than tunable receivers. The tunable transmitters are made by an array of ON–OFF switchable fixed lasers in the lab testbed, as discussed later.

The proposed architecture combines, in an efficient way, optic and electronic technologies: the aggregate bandwidth is handled in the photonic domain by working on a wavelength granularity, while packet queueing, MAC protocol, and statistical time multiplexing are handled in the electronic domain at the speed of a single data channel. Due to the *optical* simplicity of our solution, the resulting architecture does not offer all the networking features of other more complex optical network proposals [7], like a large Internet protocol (IP)-like addressing base, label swapping, arbitrary mesh topology, etc. However, we carefully selected a solution that is only based on commercially available optical components, and that at the same time offers a set of interesting features for metropolitan area networks connecting a limited number of very high-capacity nodes over a ring.

Our architecture does not require any advanced optical component, such as fast optical switches or wavelength converters. Moreover, it does not require at all optical buffering. In fact, packet buffering is implemented in the electronic domain at the boundary of the optical cloud. In our opinion, this is an important aspect, since it allows to both reduce optical complexity *and* to implement electronically efficient access algorithms.

The RingO structure is an evolution of certain WDM ring packet network proposals presented in the mid 1990s. In 1993, the first such proposal appeared in [8]. It featured a WDM ring architecture with time slotting and as many wavelengths as the number of nodes. It relied on fixed transmitters and tunable receivers, as opposed to later proposals that did the opposite. Shortly afterwards, in [9], a new network structure, using instead tunable transmitters and fixed receivers, was proposed. It already featured a λ-monitor-based protocol to avoid collisions, based on subcarriers: each wavelength carried a different and unique subcarrier frequency, which could be probed in a given time slot to assess the presence or absence of a packet on that wavelength. Later, the same basic ideas of [9] were independently brought to actual implementations at Stanford University, in the hybrid opto-electronic ring network (HORNET) project [5], and at Politecnico di Torino, in the RingO project. The two groups comprise some of the original authors of [9].

A. Node Structure

The structure of the first RingO node design is shown in Fig. 2. Some basic subsystems are common to all node architectures presented in this paper. Scanning the node structure from input to output, the main functions supported by the node are the following.

1) Amplification of the optical signals in order to compensate for the losses of the node passive elements and of the downstream fiber link.
2) Demultiplexing of the WDM comb after the amplifier. Devices which have been used for this purpose for the first RingO testbed are arrayed waveguide grating (AWG) filters.
3) Monitoring the state of channels on each slot. This is done by tapping a fraction of the power on each fiber at the output of the demultiplexer and by sending it to a DC-coupled photodiode array. This λ-monitoring electronics requires a much smaller bandwidth than the data bit-rate, since it should only detect the received average power on a slot-by-slot basis. Our solution for packet detection is easier to implement than other approaches, such as the often proposed subcarrier-tone detection [10].
4) Burst-mode detection of the incoming data-stream on the wavelength λ_i associated with node i. Note that the shift from continuous-wave operations of traditional optical network to our burst-mode operation is a major increase in complexity, but it is a price that we chose to pay to allow high efficiencies in resource utilization via statistical multiplexing.
5) Local packet traffic generation. We used a laser array driven by the node controller. The lasers are turned on for each time slot by direct current injection when a packet has to be generated. Data bits are then "written" inside the packet by an external modulator. This transmitter architecture has several motivations:
   - the use of an array of lasers, rather than a single fast tunable laser, allows using commercial and reliable devices on the ITU wavelength grid [2]; this choice was due to the difficulty in finding commercial fast-tunable lasers, and to the interesting opportunity to implement multicasting (see next item);
   - to allow efficient multicast, i.e., to send the same packets to multiple destinations. Multicasting is currently seen as an important requirement, since it is crucial to video conferencing and groupware, and indeed it is implemented in most of today commercial top-level routers [11]. In our situation, multicasting means to replicate the same packet on different wavelengths, possibly in the same time slot. With our structure, bits can be written *simultaneously* by the external modulator on all wavelengths that are generated by the laser array in a given time slot. In this way, multicasting in a single time slot can be implemented without increasing electrical bandwidth requirements at the transmitter, since the "replication" of packets is obtained in the optical domain (in which bandwidth efficiency is less critical). For what regards the electronic part of the transmitter, the cost of sending a packet to multiple destinations is the same as for sending a packet to a single destination.

As it can be seen from the description above, our architecture requires an electrical data path bandwidth, on both the transmitter and receiver side, that is equal to a single channel data rate. In fact, even when multicasting is implemented, the high-speed electrical interface of the transmitter and receiver need only to handle data traffic carried by a single wavelength, and not the aggregate bit rate of all wavelengths passing through the node. This is true for all RingO designs, and part of the RingO rationale: a metro architecture capable of scaling at large aggregate capacities must avoid to process at each node the whole network bandwidth. This was one of the problems that prevented a straightforward extension of the original LAN protocols and architectures (which assume to process the entire network bandwidth at each node interface) to metros. This is also one of the advantages of our architectures with respect to current SONET/SDH circuit-switched solutions.

B. MAC Protocol

Our architecture requires a suitable MAC protocol to allocate time slots to transmitters. From the MAC protocol design perspective, RingO is a multichannel network, in which packet collisions must be avoided and some level of fairness in resource sharing must be guaranteed together with acceptable levels of network throughput.

A collision may arise when a node inserts a packet on a time slot and wavelength which have already been used. This is avoided by giving priority to upstream nodes, i.e., to in-transit traffic, via the λ-monitoring capability.

Fairness is obtained by implementing an efficient *a posteriori* [12] packet selection strategy exploiting a virtual output queueing (VOQ) structure. While standard single-channel protocols use a single first-in–first-out (FIFO) electrical queue, in multichannel scenarios, where channels are associated with destination nodes, FIFO queueing performs poorly due to the head-of-line (HOL) problem [13]: a packet at the head of the queue may block following packets which could be transmitted on other channels. The HOL problem has been carefully studied and can be solved using one of the VOQ [13] structures. The basic VOQ idea, applicable to the RingO architecture, consists in storing packets waiting for ring access in separated FIFO queues, each corresponding to a different destination (or to a different set of destinations), and to appropriately select the queue that gains access to the channels for each time slot. It is worth noting that VOQ was demonstrated to be able to achieve 100% throughput for uniform and unicast traffic when suitable packet selection algorithms are implemented [13].

Another problem common to ring and bus topologies is the fact that an upstream node can "flood" a given wavelength, as shown in [14], reducing (or even blocking) the transmission opportunities of downstream nodes, thus generating a significant fairness problem. The fairness problem has also been investigated in detail in several previous papers, where it was shown that, again, it can be solved by using separate input queues, by selecting them with some form "round-robin" strategy (called antiresonant ring (ARR) or split ring resonator (SRR) in [14]), and by using a fairness control algorithm suited for this multichannel setup (called multimetaring).

It is not difficult to observe that our multichannel ring is equivalent to a distributed input-queued packet switch, in which node interfaces correspond to input/output line cards, and the fiber ring behaves as a distributed switching fabric. When one wavelength channel is associated with each receiver (as in Figs. 2 and 3), this switching fabric is functionally equivalent to a crossbar, capable in each time slot of delivering at most one packet to each destination, and of allowing at most the transmission of one packet from each source. In other words, in each time slot at most an input/output permutation can be served. Building upon this equivalence, the optimal packet selection criteria would be the outcome of a centralized maximal weight matching (MWM) algorithm, with weights equal to queue sizes [13]. Since this would have led to excessive complexities, our packet selection criteria is a distributed heuristic maximal approximation of MWM: each node transmits in a given slot the packet at the head of the longest of its several queues, neglecting queues whose HOL packets could not be transmitted because of the λ-monitor information. The implementation of the MAC protocol in RingO is further described in Section IV.

The complexity of the proposed MAC algorithm is mainly confined to the electronic domain, without stringent requirements on optical devices.

III. Transmission Issues in RingO

We studied RingO physical-layer design and performance in a previous paper [15], by detailed simulative analysis. Although the full set of results cannot be shown in this paper due to space limitations, the major results are briefly outlined in the following. We assume that the system is limited by the accumulation of ASE noise along the ring. We require a 2-dB margin over a reference bit-error rate equal to 10^{-9} for any receiver. Under reasonable assumptions, the cascadability of the node structure (shown in Fig. 2) was verified to reach 16 nodes, using 16 wavelength-division-multiplexing (WDM) channels each working at 10 Gbit/s, with a distance of 25 km between nodes. This transmission limit comes from ASE noise accumulation, and it is determined by the combination of the high insertion loss at each optical node, mainly due to the presence of two AWGs, and the limited signal power level at the output of the Erbium-doped fiber amplifiers (EDFAs). This cascadability is not large, but should be sufficient in a metro environment, which is the target of the RingO project.

Anyway, our paper showed that any optical impairment on the node input–output path is critical, since signals propagate all-optically along the ring without 3R regeneration. For example, in order to reach 16 nodes, polarization dependent loss (PDL) must be less than 0.6 dB per node. In addition, self-filtering effects can be critical: with the commercial AWGs used in our experiment, the required wavelength alignment accuracy for a 16 node ring should be better than 30 GHz for a 200-GHz WDM spacing.

The node structure based on AWGs, see Fig. 2, was first proposed because of the major flexibility given by fully demultiplexing all channels on separate fibers. Although some more advanced network functionalities could be envisioned with

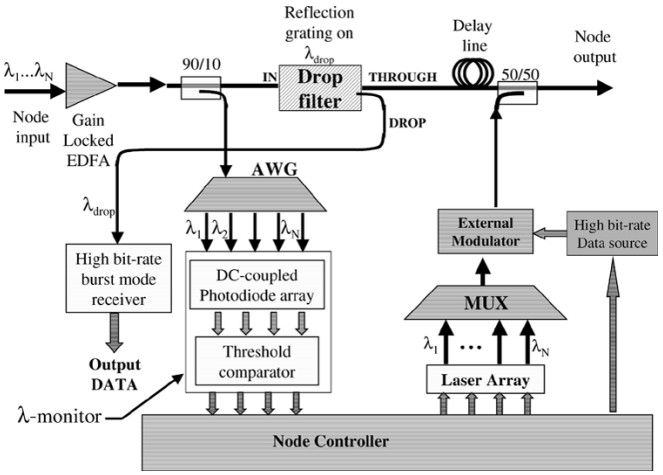


Fig. 3. Second structure of RingO nodes based on fiber-grating add-drops filters.

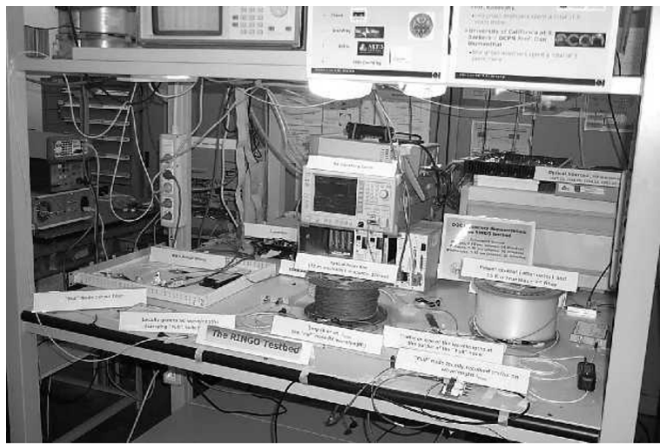


Fig. 4. RingO testbed in the PhotonLab at Politecnico di Torino.

such a structure, our simulative analysis and experimental measurements showed significant physical-layer performance limitations. In order to increase the scalability of the proposed network in terms of maximum number of nodes, we need to reduce the node insertion loss, PDL, and self-filtering effects.

These results motivate the introduction of our second node design, which is based on an add–drop filter, see Fig. 3, allowing for better cascadability and less stringent physical constraints.

While this structure is similar to the previous one for network functionalities, and most subsystems are directly derived from Fig. 2, it is significantly different from the physical layer point of view. The input–output optical path is greatly simplified, and now consists only of a passive optical splitter and a fixed add–drop filter tuned on the wavelength λ_{drop} that must be received locally. This setup greatly reduces node attenuation, self-filtering, and PDL effects, and allows a higher node cascadability.

IV. Experimental Testbed

The RingO network experimental testbed, shown in Fig. 4, was implemented in the PhotonLab at Politecnico di Torino, and was based upon nodes having the structure shown in Fig. 3. The RingO testbed goals are:

- the demonstration of the proposed architecture and MAC protocol;
- the availability of an experimental setup, where RingO physical transmission properties can be easily studied.

The testbed is currently based on two nodes, as depicted in Fig. 5, exchanging information on four different wavelengths, spaced at 200 GHz. The first one is used to generate random packet data traffic, while the second one implements all RingO protocol functions. We are, thus, able to generate an arbitrary stream of packets on any wavelength using the first node, and to demonstrate the MAC protocol operations in the second one. Since the *optical* details of the demonstrator were already shown in [15], we focus in this paper on the implementation of the node controller, which is based on a high-performance FPGA mounted on a custom-designed electronic board. The FPGA is an Altera APEX20KE600–3, with 600 000 gates, 24 320 flip-flops, four internal PLLs, 588 I/O. The working frequency can be set in the range of 30–133 MHz, thanks to the aid of the internal broadband PLL.

The node controller logical structure is shown in Fig. 6. In the following, we present the node functionality, focusing on multicast transmission. Unicast transmission can be seen as a particular case of multicast transmission in our architecture, requiring only a subset of the described logic functions.

When a packet arrives from the PCI bus (we assume that segmentation/reassembly, or packet concatenation, if necessary, occurs in higher layers, typically in the operating system of the attached PC), it is stored into an input FIFO buffer. This buffer is needed to separate the activity of the PCI bus from the activity of the on-board logic, which are not synchronous. Every packet is formatted in a fixed size protocol data unit (PDU), which contains the payload bits [service data unit (SDU)], and a fan-out set, which contains packet destination information. Since four wavelengths are used in this first prototype, the fan-out set is composed of 4 bits; a bit set to 1 means that the packet must be transmitted on the corresponding wavelength. Eight FIFO queues store packets waiting for ring access, four unicast queues, and four multicast queues. The chosen number of queues stems from our previous studies in [16]. A reference fan-out set is associated with each queue, and a simple criterion based on the minimum Hamming distance is used to build a lookup table which associates all possible multicast fan-out sets with one of the eight destination queues. The reference fan-out sets for the eight queues are shown at queue-heads in Fig. 6. For example, queue Q6 stores packets whose fan-out sets are at minimum distance from the fan-out set comprising destinations 2 and 4. For each packet entries in the queues comprise a pointer to the SDU and the corresponding fan-out set. The fan-out set of the HOL packet can be a residual, since a fan-out-splitting service policy [17] is implemented according to which destinations in the fan-out set may be reached with more than one transmission.

The length of each queue is stored in a special register file $(L0, \ldots, L7)$. On the rising edge of the slot synchronization signal, the channel state is acquired by the λ-monitor. In Fig. 6, an available wavelength is coded by a logic "1", in this example,

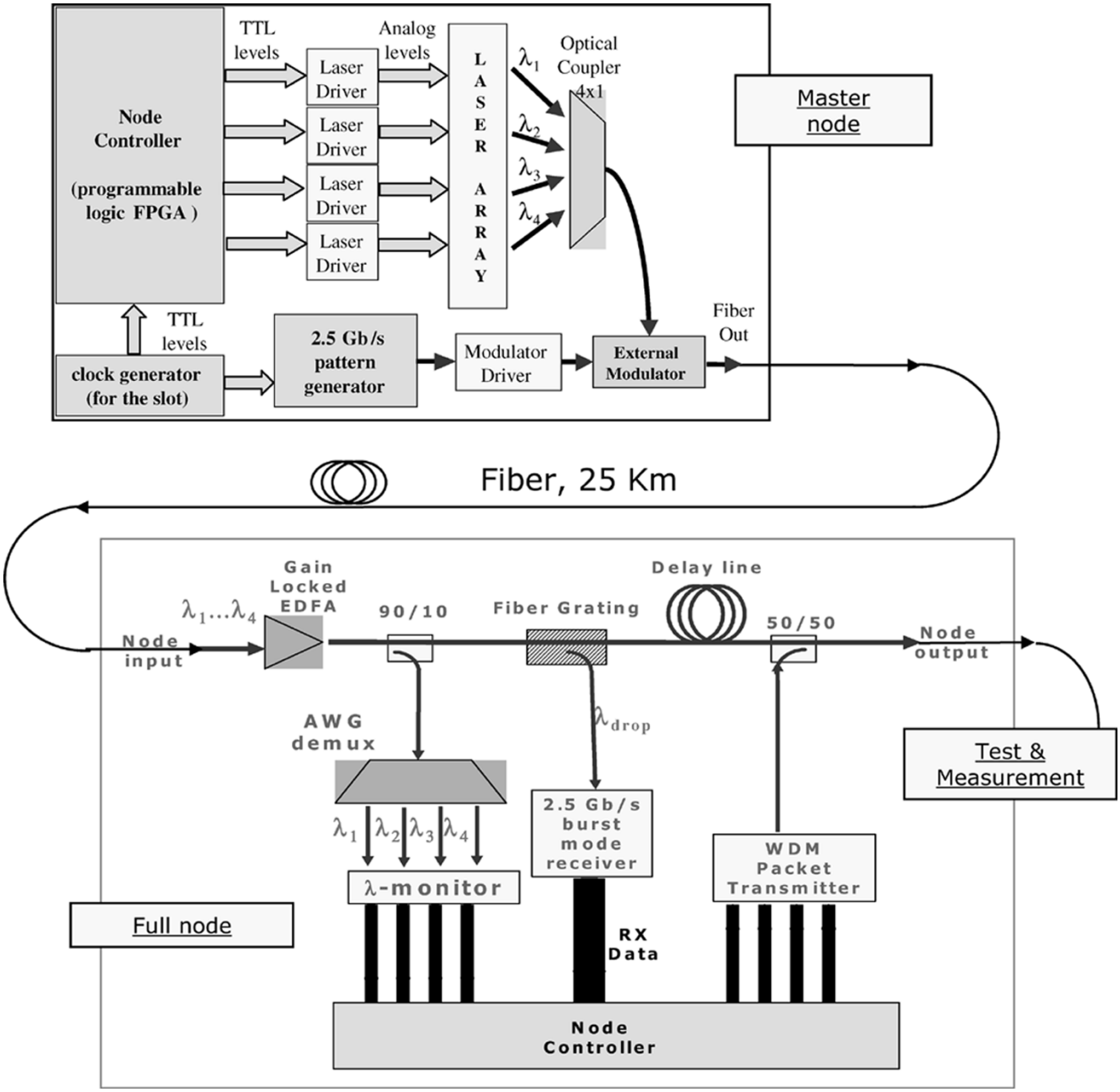


Fig. 5. Topology of the RingO testbed.

λ_2 and λ_4 are not available. A bitwise AND operation is computed between the channel state vector and the residual fan-out set vector of the HOL packet of each queue. The result is loaded into a support vector. Some queues cannot be chosen for transmission (in the example Q1, Q3, Q6) because all the wavelengths of their fan-out set are not available. The next step is to find the queue with maximum length among queues having transmission possibility. The maximum length is found by a tournament algorithm. Transmission requests of the "winner" queue, in the example Q4, have to be served in accordance with the channel wavelength availability. Our *a posteriori* packet selection policy, thus selects, among HOL packets that can be transmitted to at least one destinations in the fan-out set, the packet belonging to the longest queue. It is demonstrated that this choice guarantees the maximum throughput when the inputs and outputs are not overloaded [13], i.e., when no node is transmitting nor receiving more than the capacity of one channel. In the example, the node controller enables laser 1 and laser 3 to transmit, and sends the PDU of the first packet in queue 4 to the external laser modulator. The last step is to refresh the queue content. In the example, the fan-out set of the first packet in queue 4 has to be changed from 1110 to 0100 because λ_2 is the only transmission request not served. When all transmission requests are served, the packet is removed from the queue-head.

The control logic takes about 370 ns to do all these operations. Hence, an optical delay line of about 75 m was placed in the node demonstrator between the point where λ-monitoring is performed and the point where the locally generated packets are inserted (see Fig. 3).

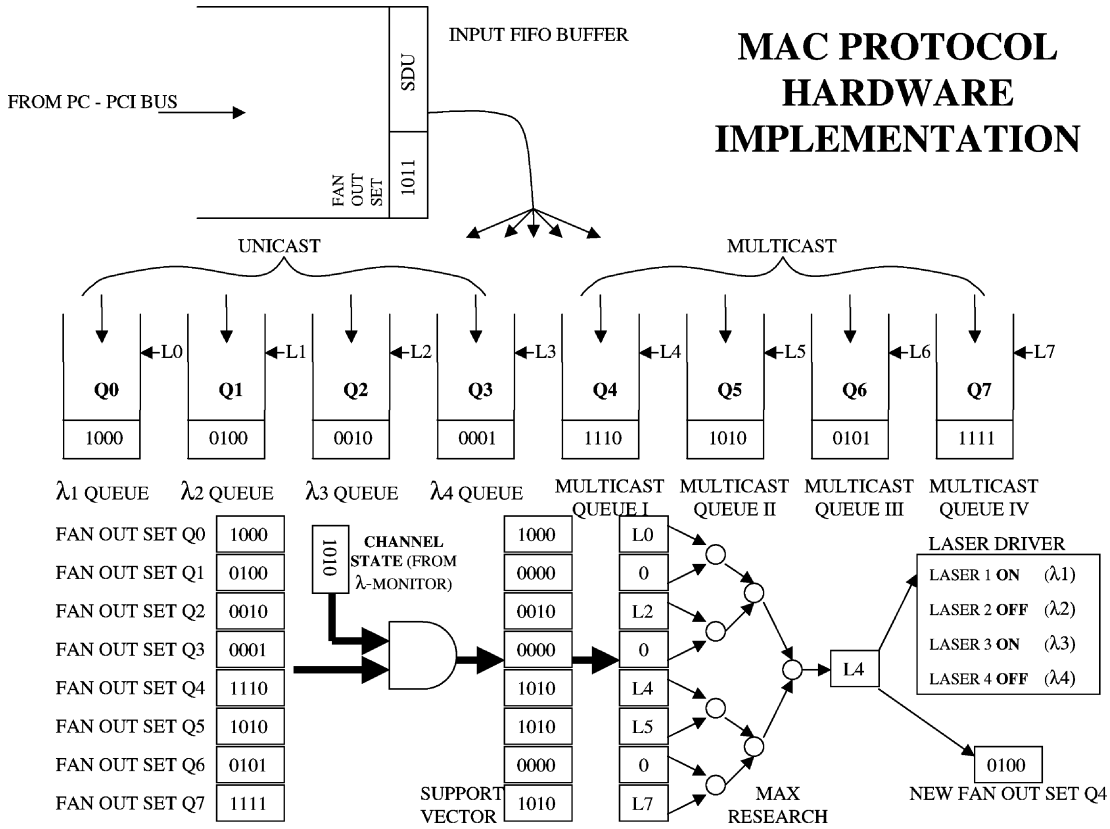


Fig. 6. Node controller logical structure.

V. RingO Scalable Architecture

An important limitation of the two previously presented RingO architectures is the fact that the number of nodes must not be greater than the number of wavelengths available on the ring, i.e., $N = W$. This largely impairs the scalability and the flexibility of our proposal. This observation leads us to the introduction of the third design for RingO nodes, which overcomes the above limitation by means of statistically time multiplexing packets to several destinations on the same wavelength channel (that is, the same wavelength can be used to transmit to different nodes). This can be achieved without changing the node's hardware in a significant way: the same basic node architecture, with fast tunable transmitter and a fixed receiver can be used, as discuss below.

A possible physical node architecture, with N nodes and W wavelengths, when $W < N$ are shown in Fig. 8. The major difference of this new design is the separation between resources devoted to transmission and resources devoted to reception. Transmitted packets traverse the ring a first time, are switched to the reception path, and then received during a second ring traversal. The transmission/reception separation can be obtained in wavelength (using different wavelength bands), in time (using different time frames), or in space (using two different fibers). We pick here the third option because it is easier to implement. This means that two physical fiber rings are used (see Fig. 7): packet transmissions occur on the first ring and receptions occur on the second ring. At some point, the two rings are interrupted and a connection between the transmission ring and the reception ring is done. This means that the ring is indeed transformed into two busses or into a folded bus, with

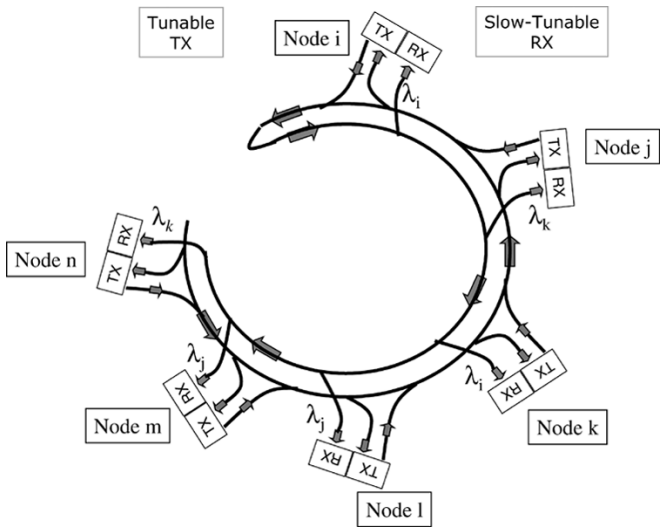


Fig. 7. Scalable architecture of the RingO network: two fiber rings topology.

significant advantages from the optical transmission viewpoint. Note that also in previous node designs the ring was broken into a set of staggered busses, one per wavelength, terminating at different receivers (each receiver terminates one wavelength). A given node must not drop from the ring the packets carried on its own receiver wavelength, and should select them (possibly in the electronic domain) according to a destination address.

The architecture of Fig. 8 requires extra optical capacity in the network, but no increase in the node complexity, nor in the capacity of transmitter and receivers, and of the data path toward applications. A negative effect of this transmission/reception separation is the loss of the space reuse capability typical of

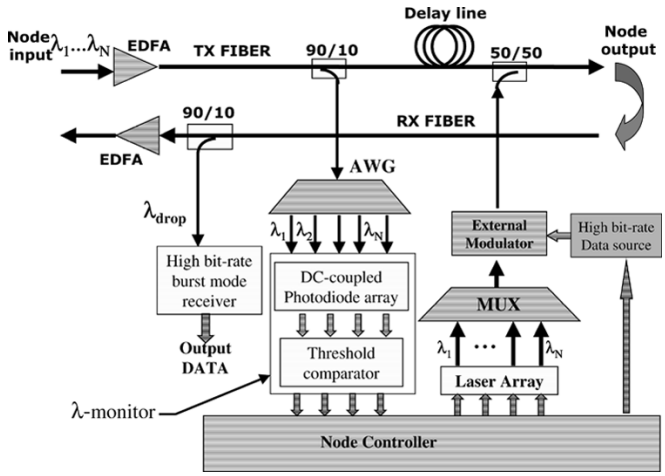


Fig. 8. Third structure of RingO nodes based on two fiber rings.

ring topologies. The two previous RingO architectures did not exploit space reuse on WDM channels due to the presence of a single receiver per channel. Space reuse becomes instead possible with multiple receivers per WDM channel. Space reuse can bring a significant throughput gain, which depends on the traffic distribution, it is around 100% in uniform traffic with a large number of nodes, less for hot-spot client/server traffic, but more for highly localized traffic.

The loss of the space reuse opportunity is the price that has to be paid with multiple receivers per channel if no optical switching in the data path is introduced. Indeed, another possible RingO architecture is currently under investigation, which keeps the single-ring topology of Fig. 1, and selectively drops packets at a receiver depending on destination addresses, allowing space reuse. To this end, at least two significant extra features should be added to the architecture of Fig. 8: an optical packet header, to carry the information on the packet destination, and a fast optical switching functionality to select packets to be dropped. Such an approach needs much more careful design of the physical layer. Moreover, fast optical switches are far from being mature components, so that this solution has been not further analyzed in this paper.

Another important feature of the architecture considered in this section is the fact that single-fault recovery comes at no extra cost, as discussed later in Section VI.

From a physical layer perspective, the architecture shown in Fig. 8 simplifies even more the node input–output optical path, which now consists only in (possibly sparse) EDFAs and optical splitter/combiners, while optical filters or add-drops are avoided. As discussed in Section III, the reduction of the complexity of optical components on the data path greatly reduces physical impairments such as PDL and self-filtering. Moreover, the two-fiber topology shown in Fig. 7 does not generate optical loops, thus avoiding potentially detrimental effects such as ASE noise recirculation, ring lasing, etc.

A. Allocation of Receivers to WDM Channels

From a network dimensioning perspective, since more than one node can receive on the same wavelength, a decision problem arises concerning the allocation of the different receivers to WDM channels. Good solutions to this problem should aim at equalizing the load on the different channels, that is the maximum load among all channels must be minimized.

It is straightforward to notice that the solution of the node allocation problem depends on the traffic on the network. Although this traffic matrix could be dynamically estimated, we suppose for simplicity that the traffic matrix is known.

The problem can be formalized in terms of integer linear programming (ILP), and it can be shown to be equivalent to the well-known problem of scheduling jobs on identical parallel machines, which falls in the class of NP-hard problems [18]. The problem states that given W wavelengths and N nodes, the receiver bandwidth load can be expressed as

$$l_i = \sum_{j=1}^{N} p_{ji} r_j \qquad \forall i,\ 1 \le i \le N$$

where r_j represents the transmission rate of node j and, p_{ji} its transmission probability to node i. A set of control variables x_{ik} can be defined, where

$$x_{ik} = \begin{cases} 1, & \text{iff node } i \text{ receives on wavelength } k \\ 0, & \text{otherwise} \end{cases}.$$

Receivers allocation is to be done trying to minimize $L_{\max}$, i.e., the load on the most loaded wavelength $L_{\max} = \max_k \sum_{i=1}^{N} l_i x_{ik}$. Thus, our problem formulation becomes

$$\text{Minimize } L_{\max}$$

subject to the following constraints:

$$L_{\max} \ge \sum_{i=1}^{N} l_i x_{ik} \qquad \forall k,\ 1 \le k \le W \tag{1}$$

$$\sum_{k=1}^{W} x_{ik} = 1 \qquad \forall i,\ 1 \le i \le N \tag{2}$$

$$x_{ik} \in \{0,1\} \qquad \forall i,\ 1 \le i \le N \quad \forall k,\ 1 \le k \le W. \tag{3}$$

Equation (1) ensures that no wavelength has a load larger than $L_{\max}$. Equation (2) ensures that each receiver must be allocated to only one wavelength.

Performance results are plotted in Fig. 9, where a scenario with 16 nodes and 4 wavelengths (four for each fiber ring, since we obtain transmission/reception separation using separated fiber rings) was simulated. In this simple scenario, two nodes named servers transmit at high load, equal to the capacity of one wavelength per server, with equal probability to the remaining 14 nodes, called clients. Client nodes transmit only to servers at a lower rate, equal to 1/14 of the channel capacity. Hence, the input and output load for all servers and for all clients are the same.

In Fig. 9, we show the throughput versus input load (both normalized to the available network capacity) for three different modes of allocating nodes to wavelengths. In particular, we compare the optimal receiver allocation obtained with the ILP

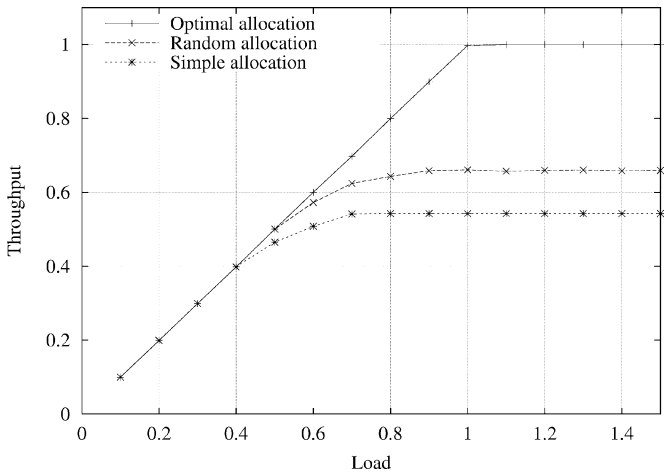


Fig. 9. Normalized network throughput versus input load for three different allocations of node receivers to WDM channels.

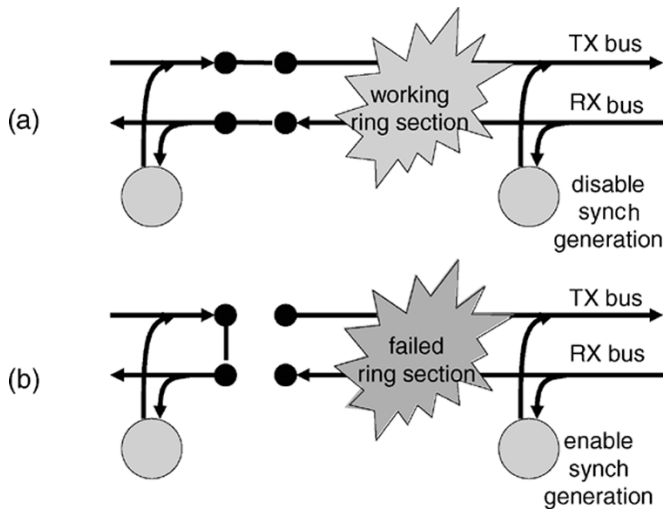


Fig. 10. Logical functionalities (a) for normal nodes and (b) for the nodes around the bus folding point (or the faulty section).

model described above with two other allocations. In the first one, called *random allocation*, each node is randomly allocated to one of the available wavelengths. In the second one, called *simple allocation*, we force that the number of allocated nodes on each wavelength is the same. We can observe that a nonoptimal solution to the allocation problem may lead to significant reductions of the total network throughput.

The complexity of the optimal solution may be too large. We sketch a simple but effective heuristic (for the scenario of Fig. 9 it provides the same solution of the ILP model) that solves the problem of receiver allocation with a low complexity [it can be shown to be $O(NW\log(W))$]. The algorithm follows the next three steps.

Step 1) Order all receiver loads as a nonincreasing sequence.

Step 2) Allocate the first receiver of the sequence on the least loaded wavelength, and delete it from the sequence.

Step 3) If the sequence is empty, then EXIT; else GOTO Step 2).

This algorithm is known, in operational research, as longest processing time (LPT) [18].

Despite the fact that the traffic matrix upon which the receiver allocation is chosen must be known *a priori*, it can show variations over time, i.e., it can behave as a dynamic matrix. In this case, it may be worthwhile to reallocate receivers dynamically in order to keep the network in an optimal operation point. One elegant way of achieving this result is to introduce (slow) tunability in node receivers. This tunability does not need to be fast, and does not need to track packet-by-packet variations. Low-cost devices available today (e.g., mechanical or thermo-optic filters) can be suitable to implement this slow receiver tunability feature. It is out of the scope of this paper to deal with problems concerning reconfiguration issues, but it should be clear that a tradeoff arises between keeping the receiver allocation well matched to the dynamic traffic matrix to optimize performance, and throughput losses due to blackouts when receivers are tuning.

VI. Fault Recovery

The node architecture described in the previous section has interesting fault recovery properties. The capability of recovering faults is considered an essential feature in all high-speed networks. Recovering from single faults requires biconnected topologies and bidirectional rings are the simplest such topologies. Ring-based networks typically require two counterrotating fiber rings to be able to protect single faults.

In the design of Fig. 7, we already have two counterrotating fiber rings, and single-fault protection can be provided by logically moving the folding point between the transmission bus and the reception bus to just before the fault (which can be either a fiber cut or a node failure) on the transmission bus. Although we did not discuss synchronization issues in this paper, around the folding point between two busses, and more precisely at the beginning of the transmission bus, the slot synchronization information must be injected in all wavelength channels. Fault protection implies equipping all nodes with this capability, and enabling it only at the first node of the transmission bus. Fig. 10 logically depicts the switching and synch signal generation capabilities that should be available at all nodes to recover from single faults. Note that switching must not necessarily be very fast: a reasonable target for error recovery is the SONET/SDH 50 ms figure. We do not further discuss on this issue due to space limitations, but proper fault detection procedures, fault signaling protocols, and fault recovery algorithms must be identified and implemented. Although this architecture does not allow space reuse, we observe that, in presence of a fault, all current traffic can be rerouted in the restored topology (i.e., overloading is avoided); the same is in general not true when rings exploiting space reuse are to be protected.

In absence of faults, the position of the folding point can be selected according to straightforward algorithms, and the configuration of the nodes around the folding point are exactly the same as for the nodes around a faulty network section.

As previously noted, however, the nice property of the architecture depicted in Fig. 8 is the sharing of network resources between the multireceiver per wavelength feature, and the fault

recovery mechanisms. We also remark that no additional transceivers are required for fault protection, and that the amount of fiber for protection is minimal.

VII. Conclusion

Metropolitan area networks are an arena where researchers and network architects have the opportunity to speculate on the best utilization of optical technologies in the implementation of switching and control functions.

Our work was motivated by the trust that optical packet transmission, though not yet standardized and commercially available, may become in the medium term a promising alternative to the current approach of building WDM networks with a high degree of fast circuit-switching reconfigurability, but where packet switching is still completely handled at the electronic level. At the same time, we do not believe that all packet switching functions can be *completely* moved from the electrical to the photonic domain in a reliable way without fundamental improvements in optical components technology. A good compromise between the two domains (optical and electrical) is the major goal of the RingO project presented in this paper.

We have presented three alternative node designs, which exhibit several common features, but trace an evolutionary path toward a final design that can be engineered in a successful and cost-effective manner. Several important issues (e.g., signaling for fault recovery, synchronization) were not discussed in this paper due to space limitations.

An interesting contribution of the RingO architecture was the definition of the access protocol, which builds upon previous experiences in scheduling packets in input-queued switching architectures, and offers good performance at complexities that are compatible with available technologies, as proved in our lab experiments.

References

[1] R. Ramaswami and K. N. Sivarajan, *Optical Networks–A Practical Perspective*. San Mateo, CA: Morgan Kaufman, 1988.

[2] *ITU-T Recommendation G.872*.

[3] S. Yao, B. Mukherjee, and S. Dixit, "Advances in photonic packet switching: An overview," *IEEE Commun. Mag.*, vol. 38, pp. 84–94, Feb. 2000.

[4] A. Carena, M. Vaughn, R. Gaudino, M. Shell, and D. J. Blumenthal, "OPERA: An optical packet experimental routing architecture with label swapping capabilities," *IEEE/OSA J. Lightwave Technol.*, vol. 16, pp. 2135–2145, 1998.

[5] K. V. Shrikhande, I. M. White, D. Wonglumsom, S. M. Gemelos, M. S. Rogge, Y. Fukashiro, M. Avenarius, and L. G. Kazovsky, "HORNET: A packet-over-WDM multiple access metropolitan area ring network," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 2004–2016, Oct. 2000.

[6] C. Guillemot, M. Renaud, P. Gambini, C. Janz, I. Andonovic, R. Bauknecht, B. Bostica, M. Burzio, F. Callegati, M. Casoni, D. Chiaroni, F. Clerot, S. L. Danielsen, F. Dorgeuille, A. Dupas, A. Franzen, P. B. Hansen, D. K. Hunter, and A. K. Kloch, "Transparent optical packet switching: The European ACTS KEOPS project approach," *IEEE/OSA J. Lightwave Technol.*, vol. 16, pp. 2117–2134, Dec. 1998.

[7] D. J. Blumenthal, B. E. Olsson, G. Rossi, T. E. Dimmick, L. Rau, M. Masanovic, O. Lavrova, R. Doshi, O. Jerphagnon, J. E. Bowers, V. Kaman, L. A. Coldren, and J. Barton, "All-optical label swapping networks and technologies," *IEEE/OSA J. Lightwave Technol.*, vol. 18, pp. 2058–2075, Dec. 2000.

[8] I. Chlamtac, A. Fumagalli, L. K. Kazovsky, and P. Poggiolini, "A multi-Gbit/s WDM optical packet network with physical ring topology and multi-subcarrier header encoding," in *Proc. Eur. Conf. Optical Communications*, Montreux, Switzerland, Sept. 1993, pp. 121–124.

[9] ——, "A contention/collision free WDM ring network for multi-Gbit/s packet switched communication," *J. High Speed Networks*, vol. 1, no. 4, pp. 1–19, Apr. 1995.

[10] M. Cerisola, T. K. Fong, R. T. Hofmeister, L. G. Kazovsky, C. L. Lu, P. Poggiolini, and D. J. M. Sabido IX, "CORD-A WDMA optical network: Subcarrier-based signaling and control scheme," *IEEE Photonics Technol. Lett.*, vol. 7, pp. 555–557, May 1995.

[11] W. Parkhurst, *Cisco Multicasting Routing & Switching*. New York: McGraw-Hill, 1999.

[12] A. Bianco, E. Di Stefano, A. Fumagalli, E. Leonardi, and F. Neri, "*A posteriori* access strategies in all-optical slotted rings," in *Proc. IEEE GLOBECOM*, Sydney, Australia, Nov. 1998, pp. 300–306.

[13] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," *IEEE Trans. Commun.*, vol. 47, pp. 1260–1267, Aug. 1999.

[14] M. A. Marsan, A. Bianco, E. Leonardi, M. Meo, and F. Neri, "MAC protocols and fairness control in WDM multi-rings with tunable transmitters and fixed receivers," *J. Lightwave Technol.*, vol. 14, pp. 1230–1244, June 1996.

[15] A. Carena, V. Ferrrero, R. Gaudino, V. De Feo, F. Neri, and P. Poggiolini, "RingO: A demonstrator of WDM optical packet network on a ring topology," in *Proc. IFIP Optical Network Design and Modeling Conference ONDM 2002*, Turin, Italy, Feb. 2002, pp. 183–197.

[16] A. Bianco, P. Giaccone, E. Leonardi, F. Neri, and C. Piglione, "On the number of input queues to efficiently support multicast traffic in input queued switches," in *Proc. IEEE Workshop High Performance Switching Routing*, Turin, Italy, June 2003, pp. 111–116.

[17] R. Ahuja, B. Prabhakar, and N. McKeown, "Multicast scheduling for input-queued switches," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 855–866, June 1997.

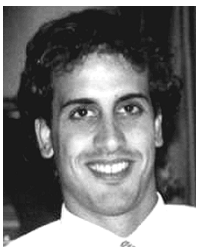
[18] M. Pinedo, *Scheduling: Theory, Algorithms, and Systems*. Englewood Cliffs, NJ: Prentice-Hall, 2002.

Andrea Carena (M'98) was born in Carmagnola, Torino, Italy, on October 7, 1970. He received the Laurea degree in electronic engineering (*summa cum laude*) and the Ph.D. degree in electrical engineering (optical communications) from Politecnico di Torino, Torino, Italy, in 1995 and 1999, respectively.

In 1998, he spent a year as a Visiting Researcher, first, in the Optical Communications and Photonic Network (OCPN) Group, Georgia Institute of Technology, Atlanta, and then at the University of California at Santa Barbara working in the realization of OPERA, an optical label swapping network testbed. He has collaborated in the development and implementation of OptSim, an optical transmission system simulator. He is currently an Assistant Professor in the Optical Communication Group, Politecnico di Torino. He has coauthored more than 40 papers in the field of optical fiber transmission. His research interests are in the field of new modulation formats, long-haul DWDM systems, fiber nonlinearity, performance analysis, and computer simulation of lightwave transmission systems.

Vito De Feo (S'03) was born in Salerno, Italy, in 1972. He received the Laurea degree in electronic engineering (*summa cum laude*) from Politecnico di Torino, Torino, Italy, in 2001 (thesis on the electronic control logic of an all-optical network). In 2001, he began working toward the Ph.D. degree in the Optical Communications Group and in the Telecommunication Network Group, Department of Electronics, Politecnico di Torino. He is currently a Visiting Ph.D. Student working toward the Ph.D. degree in the Optical Communication and Optical Network Group, Department of Electrical Engineering, Stanford University, Stanford, CA.

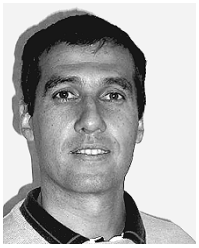
In 1997, he was a Visiting Student at Trinity College, Dublin, Ireland. His interests involve experimental demonstration of optical networks, subsystems for next-generation all-optical networks, and scheduling in optical network.



Jorge M. Finochietto (S'99) was born in Buenos Aires, Argentina, in 1978. He received the degree in electronics engineering from the Universidad Nacional de Mar del Plata, Mar del Plata, Argentina, in 2000. Since 2002, he has been working toward the Ph.D. degree in the Dipartimento di Elettronica, Politecnico di Torino, Torino, Italy.

From 2000 to 2001, he was with the Engineering Group of Techtel, Buenos Aries, Argentina, as a South American Network Operator, in the areas of routing performance, quality-of-service (QoS), and ATM. His research interests include the design of all-optical networks and switch architectures.

Roberto Gaudino (M'98) is currently an Assistant Professor in the Optical Communications Group, Politecnico di Torino, Torino, Italy, where he works on several research topics related to optical communications. His main research interest is in the metro and long-haul DWDM systems, fiber nonlinearity, modeling of optical communication systems, and on the experimental implementation of optical networks. Currently, he is investigating new optical modulation formats, such as polarization or phase modulation, and packet switched optical networks. In 1997, he spent one year as a Visiting Researcher in the Optical Communication and Photonic Network (OCPN) Group, Georgia Institute of Technology, Atlanta, where he worked in the realization of the MOSAIC optical network testbed. From 1998, for two years, he has been with the team that coordinates the development of the commercial optical system simulation software OptSim. He is author or coauthor of more than 60 papers in the field of optical fiber transmission and optical networks. He is a consultant for several companies of the optical sector, and he is also involved in professional continuing education.



Fabio Neri (M'98) received the Dr.Ing. and Ph.D. degrees in electrical engineering from Politecnico di Torino, Torino, Italy, in 1981 and 1987, respectively.

He is a Full Professor in the Electronics Department, Politecnico di Torino. His teaching includes graduate-level courses on computer communication networks and on the performance evaluation of telecommunication systems. He leads a research group on optical networks at Politecnico di Torino. He has coauthored over 100 papers published in international journals and presented in leading international conferences. His research interests are in the fields of performance evaluation of communication networks, high-speed and all-optical networks, packet switching architectures, discrete event simulation, and queueing theory.

Dr. Neri was General Co-Chair of the 2001 IEEE Local and Metropolitan Area Networks (IEEE LANMAN) Workshop and General Chair of the 2002 IFIP Working Conference on Optical Network Design and Modeling (ONDM).

Chiara Piglione (S'03) was born in Racconigi, Italy, in 1977. She received the degree in telecommunication engineering (*summa cum laude*) from Politecnico di Torino, Torino, Italy, in July 2002. In 2003, she began working toward the Ph.D. degree in the Electronics Department, Politecnico di Torino. She is currently a Visiting Ph.D. Student working toward the Ph.D. degree in the Department of Electrical Engineering, Arizona State University, Tempe, AZ.

From October 2002 to December 2002, she had a research contract with the Italian National Inter-University Consortium for Telecommunications (CNIT). Her research interests include the study of multicast traffic, input queued switches, and all-optical networks.

Pierluigi Poggiolini (S'90–M'93) was born in Torino, Italy, in 1963. He received the M.S. (*cum laude*) and the Ph.D. degrees from Politecnico di Torino, Torino, Italy, in 1988 and 1993, respectively.

From 1990 to 1992, he was a Visiting Scholar with the Optical Communications Research Laboratory, Stanford University, Stanford, CA, where he mainly worked on the STARNET optical network research project. In 1994 and 1995, he was again at Stanford University, as a Postdoctoral Fellow, working on the ARPA-funded CORD all-optical packet network project. Since 1998, he has been an Associate Professor at Politecnico di Torino. He is the Coordinator of the Optical Communications Group, Politecnico di Torino. He has coauthored over one hundred papers in leading journals and conferences. His research interests include new modulation formats for optical transmission, optical packet networks, nonlinear fiber effects and modeling, and simulation of optical communications systems.

Dr. Poggiolini was awarded the International Italgas Prize for Scientific Research and Technological Innovation in 1998.

Chapter 4

Università di Trento Research Unit

R. Battiti, R. Lo Cigno, E. Salvadori
Dipartimento di Informatica e Telecomunicazioni - Università degli Studi di Trento
Via Sommarive 14 - 38050 Povo (TN), Italy
{battiti,locigno,salvador}@dit.unitn.it

Abstract

During the second year of the ADONIS project, the University of Trento Research Unit concentrated its activities on these two main topics:

- Impact of traffic elasticity on dynamic grooming algorithms in IP over Optical networks
- The definition of a formal framework for dynamic grooming policies

The IP protocol and optical transmission techniques are going to play a fundamental role in the next generation Internet. It is a widespread prediction that all other intermediate network management layers (ATM, SDH, SONET, ...) will gradually disappear, leaving a scenario where IP packets are carried directly on high speed WDM-based optical connections, the so-called IP over Optical (IPO) network. In this scenario, the interaction between routing and control of the circuit-switched optical network and that of the packet-switched IP network is of the utmost importance for the end-to-end performance and the efficient use of network resources.

Traffic grooming is the multiplexing capability aimed at optimizing the capacity utilization in transport systems by means of the combination of low-speed traffic streams onto high-speed (optical) channels. This problem is a variant of the well-known virtual topology design problem and has received a lot of attention in recent years (see [1] for a review).

There are two main approaches to study traffic grooming: *static* and *dynamic*. Static grooming refers to some network usage optimization when the traffic matrix is known in advance and it was proved to be an NP-hard problem. Dynamic grooming is a routing problem in a multi-layer network architecture, since the objective is to find the "best" path to route traffic requests arriving dynamically to grooming nodes. In this case equivalent requests arriving at different times may be treated differently because of different network conditions.

In IPO networks, IP routers are attached to an optical core network and connected to other peers via dynamically established lightpaths [2]. Considering the *control plane*, different architectures can be envisioned according to the amount of information exchanged between the IP and optical layer. RFC 3717 defines three interconnection models: overlay, augmented and peer. In the *peer* model, the topology and other network information are completely shared in a unified control plane, while in the *overlay* model, each layer performs its own routing functions because no information is exchanged between them. An intermediate architecture is the *augmented* model, where some aggregated information from one routing instance is passed to the other.

The peer and the augmented models are appealing because they allow running an integrated routing function, by using, for instance, an auxiliary graph, as done in [3]. However, both models do not seem feasible in the near term due to the tight integration between the two levels and scalability issues regarding the amount of exchanged information. The overlay model is instead technically feasible, because it only requires the definition of an interface between the IP and optical level and dynamic lightpath capabilities in the optical level, which are being experimented in laboratories and research projects [4]. Surprisingly, most of the dynamic grooming algorithm proposed in the literature implicitly consider such models [3, 5, 6, 7], while only a few dynamic grooming algorithms based on the overlay model have been proposed so far [8, 9]. These papers explore the two extreme policies of privileging always the optical level exploitation or the other way around. In [7] the authors propose new algorithms that improve performance in the overlay model with respect to [8] and in the augmented model with respect to a grooming algorithm in peer models proposed in [5].

All these works, however, simply disregard the elastic nature of TCP/IP traffic: IP over WDM is indeed modelled like a traditional circuit switched traffic.

As shown in [10], considering the adaptivity of traffic has a deep impact on the network performance and on routing algorithms in particular. The reason lies in the feedback nature of the interaction of elastic traffic with the network: the network status (e.g., congestion) induces a reaction in the source behavior that, depending on the control signal[1] can be a positive or a negative feedback. It is obvious that a positive feedback has, to say the least, a noxious impact on performance, because congestion, or any other performance detrimental status, is exasperated by the positive feedback.

As usual in closed loop systems with delay, the nature of feedback (positive or negative) can change with changing conditions, so that, for instance, a negative feedback at low loads can change to a positive feedback at high loads, leading to instability phenomena.

[1] We use the term "*control signal*" though it is not necessary to have a notification protocol to have feedback. Implicit signals, network measures, or simply source-destination interaction can carry the feedback information.

Paper [11] investigates how traffic elasticity impacts on some basic grooming algorithms and assesses their performance in dynamic networking scenarios where the optical and IP level of the network interact with one another. The problem is rather complex, since it requires to take into account how competing groomed flows interact, e.g., sharing resources following a max-min criterion, as well as how the optical management plan behaves and assigns resources to traffic relations.

Elasticity in groomed traffic can arise due to a number of reasons and in very different scenarios. In emerging metro-area optical network, the foreseen trend is a very dynamic and aggressive use of optical paths, thus leading to traffic relations that are very close to a simple host-to-host IP flow[2].

In more traditional wide-area optical networks, where it is generally assumed that traffic relations are peering contracts between operators with highly aggregated flows, the elasticity still arises from the fact that all the flows within a traffic relation are elastic: if congestion arises, then all flows react by reducing their offered load and the result is the overall elasticity of the aggregation.

Two very simple grooming algorithms have been considered, one privileging the opening of new optical paths, named *OptFirst*, the other one privileging the use of the already available IP logical topology, named *VirtFirst*.

In both grooming algorithms the impact of elastic traffic, included with a sophisticated model in the simulations tool, is dramatic, showing clearly that approximating IP traffic with CBR-like traffic can lead to wrong conclusions when routing and grooming are considered. The different performance induced in the network by the elastic traffic is such that conclusions drawn with traditional traffic models can be completely misleading.

The focus of the paper was on the impact of the traffic elasticity, thus little attention was placed on the "suitability" of the grooming algorithms analysed. Both the *OptFirst* and *VirtFirst* algorithms, however, have clearly shown that they are not suited for the management of an IP over WDM network, since the lack of coordination between the IP and the optical level leads to waste resources. As shown on the NSFNET topology the *OptFirst* policy may even lead to block requests with very low network loads because a very aggressive use of optical resources may lead to IP-level virtual topologies that are not completely connected.

A second contribution [12] builds upon the previous one, proposing a comparative study of dynamic grooming algorithms in realistic scenarios with a data-based traffic model including elasticity, together with the simulation tool (GANCLES) used to perform it.

The first part of this contribution is devoted to GANCLES presentation [13], highlighting its innovative features and the management of different architectural models of IPO networks through the explicit simulation of the optical and IP network levels.

The second part discusses instead performance results of different dynamic grooming algorithms on two topologies: a ring and a modification of NSFNet. First of all it is shown how the traffic model impacts on results. Then several performance indices, including the throughput of elastic flows, the probability that the service they receive falls

[2]We are not interested here in discussing whether such flows are based on UDP, TCP or whatever other transport protocol, we just notice that any recent discussion and proposal on transport protocols includes elasticity and end-to-end congestion control

below an acceptable threshold causing the flow starvation, and fairness are compared for the chosen grooming policies assuming an overlay IPO model.

The results show that the presence of a double network layer (optical and IP) does not alleviate traditional fairness problems associated with best-effort, elastic traffic. Some form of compensations are possible through the use of smart grooming policies; however, in an overlay model, where no information is shared between the optical and the IP level, it is not easy to find the appropriate and definitive solution.

Further studies on dynamic grooming enabled by GANCLES include comparison between different IPO architectures, studying what is the amount of information that needs to be exchanged to allow "intelligent" resource use. In addition, grooming strategies, policies, and algorithms can be implemented and studied in the simulator as we did for the *HopCons* policy that, although very simple, allows overcoming some of the shortcomes of *OptFirst* and *VirtFirst*. Finally, one major question is related to the use of QoS routing either in the optical or IP network layer in order to understand how "intelligent" routing strategies do interact one another through grooming policies.

A third contribution [14] introduces a formal description of dynamic grooming policies based on graph theory, clearly defining the limits between grooming in overlay architectures and grooming in peer or augmented architectures, where there is total or partial integration of the optical and IP control planes.

Furthermore a family of grooming policies based on constraints on the number of hops and bandwidth available at the virtual topology level are defined and analyzed in different regular and irregular topologies, discussing parameters setting and the impact of the number of available wavelengths per fiber on the grooming policy. It has been shown that dynamic grooming policies previously presented in literature are particular cases of the family we defined, and that it is possible to define grooming parameters that lead to good performance regardless of the topology and that allow good scaling with the amount of optical resources.

The impact of traffic engineering techniques applied at the optical or IP level was discussed, highlighting that constrained based routing techniques in the IP level, which is characterized by quicker dynamics, ensures better performance with respect to the use of adaptive routing in the optical level. In any case the well known problem of performance degradation at high loads when constrained-based routing is used is present also in IPO networks.

The activities presented here have been developed thanks to the collaboration with Zoltan Zsóka, from the Department of Telecommunications of the Budapest University of Technology and Economics.

In the next references section, we report all the publications that are related to this project. Items in **boldface** are publications generated at Università di Trento in the second year of the Adonis project.

Bibliography

[1] R. Dutta, G.N. Rouskas, "Traffic grooming in WDM networks: past and future," *IEEE Network Magazine*, 16(6):46–56, Nov./Dec. 2002.

[2] B. Rajagopalan, J. Luciani, D. Awduche, "IP over Optical Networks: A Framework," RFC 3717, IETF, Mar. 2004.

[3] H. Zhu, H. Zang, K. Zhu, B. Mukherjee, "A novel generic graph model for traffic grooming in heterogeneous WDM mesh networks," *IEEE/ACM Trans. on Networking*, 11(2):285–299, Apr. 2003.

[4] C. Cavazzoni, et al., "The IP/MPLS over ASON/GMPLS test bed of the IST project LION," *IEEE/OSA Jou. of Lightwave Technology*, 21(11):2791–2803, Nov. 2003.

[5] M. Kodialam, T.V. Lakshman, "Integrated Dynamic IP and Wavelength Routing in IP over WDM Networks," *In Proc. of INFOCOM 2001*, pp. 358–366, Anchorage, AK, USA, Apr. 22–26 2001.

[6] R. Srinivasan, A.K. Somani, "Dynamic Routing in WDM Grooming Networks," *Photonic Network Communications*, 5:123–135, Mar. 2003.

[7] S. Koo, G. Sahin, S. Subramaniam, "Dynamic LSP Provisioning in Overlay, Augmented, and Peer Architectures for IP/MPLS over WDM Networks," *In Proc. of INFOCOM 2004*, Hong Kong, China, Mar. 7–11 2004.

[8] C. Assi, A. Shami, M.A. Ali, Y. Yue, S. Dixit, "Integrated Routing Algorithms for Provisioning "Sub-Wavelength" Connections in IP-Over-WDM Networks," *Photonic Network Communications*, 4(34):377–389, Jul./Dec. 2002.

[9] X. Niu, W.D. Zhong, G. Shen, T.H. Cheng, "Connection Establishment of Label Switched Paths in IP/MPLS over Optical Networks," *Photonic Network Communications*, 6(1):33–41, Jul. 2003.

[10] C. Casetti, R. Lo Cigno, M. Mellia, M. Munafò, and Z. Zsóka. A Realistic Model to Evaluate Routing Algorithms in the Internet. In *IEEE Proceedings of GLOBECOM*, volume 3, pages 1882–85, San Antonio, Texas, USA, 2001.

[11] **R. Lo Cigno, E. Salvadori, Z. Zsóka, "Elastic Traffic Effects on WDM Dynamic Grooming Algorithms," *In Proc. of GLOBECOM 2004*, Dallas, TX, USA, Dec. 2004.**

[12] **E. Salvadori, Z. Zsóka, and R. Lo Cigno. Dynamic Grooming in IP over WDM Networks: A Study with Realistic Traffic based on GANCLES Simulation Package. In *Proceedings of the 9th IFIP/IEEE ONDM*, Milan, Italy, February 2005.**

[13] GANCLES - Grooming cApable Network Call-Level Simulator. URL: `http://netmob.unitn.it/gancles.html`

[14] **E. Salvadori, Z. Zsóka, R. Lo Cigno, and R. Battiti. A Framework for Dynamic Grooming in IPO Overlay Architectures. Submitted for publication to Networking, 2005.**

Elastic Traffic Effects on WDM Dynamic Grooming Algorithms

R. Lo Cigno[1], E. Salvadori[1], Z. Zsóka[2],
[1]Dipartimento di Informatica e Telecomunicazioni – Università di Trento
[2]Department of Telecommunications – Budapest University of Technology and Economics

***Abstract*— Traffic grooming in IP over WDM networks introduces a coupling between the optical and the IP layer. Grooming algorithms are normally studied with a very simple traffic model that completely ignores this interaction. This paper compares the performance of two simple grooming algorithms with a traditional, Poisson based traffic model and a more complex one that takes into account the IP traffic elasticity and the inherent interaction between IP and the optical layer. Simulation results, supported by heuristic considerations highlighting the interaction effects, show that ignoring the two layer interaction may lead to wrong conclusions and waste of resources.**

I. Introduction

IP over WDM is one of the racehorses that pulls the train of large bandwidth networking. New services are continuously deployed over IP, and WDM evolution [1], [2] provides the transmission speed needed to pump the information through the network. One of the main issues in IP over WDM architectures, is the traffic aggregation or *grooming*. The traffic is generated as tiny trickles over IP, while the transmission pipe over a single λ within an optical fiber is enormous.

Many grooming algorithms were proposed in recent years (see [3], [4], [5] to cite just a few), and compared one another. Some works assume static grooming [5], [6], and generally tackle the problem with some optimization technique, while others assume that grooming is dynamic [4], [7], [8], [9]. All these works, however, disregard the elastic nature of TCP/IP traffic: IP over WDM is indeed modeled like a traditional circuit switched multiplexing problem!

As shown in [10], considering the adaptivity of traffic has a deep impact on the network performance and on routing algorithms in particular. The reason lies in the feedback nature of the interaction of elastic traffic with the network: the network status (e.g., congestion) induces a reaction (feedback) in the source behavior. If the feedback is positive it has, a noxious impact on performance, since congestion, is exasperated by the positive feedback.

As usual in closed loop systems with delay, the nature of feedback (positive or negative) can change with changing conditions, so that, for instance, a negative feedback at low loads can change to a positive feedback at high loads, leading to instability phenomena.

The aim of this paper is to investigate how traffic elasticity impacts on grooming. The problem in itself is rather complex, since it requires to take into account how competing groomed flows interact one another, e.g., sharing resources following a max-min criterion, as well as how the optical management plan behaves and assigns resources to traffic relations.

Elasticity in groomed traffic can arise due to a number of reasons and in very different scenarios. In emerging metro-area optical network, the foreseen trend is a very dynamic and aggressive use of optical paths, thus leading to traffic relations that are very close to a simple host-to-host IP flow. In more traditional wide-area optical networks, where traffic relations are peering contracts with highly aggregated flows, the elasticity still arises from the fact that all the flows within a traffic relation are elastic: if congestion arises, then all the flows will react reducing their offered load.

This work was supported by the Italian Ministry of Education and Research (MIUR) through the GRID.IT and ADONIS projects, and by the Hungarian Italian Intergovernmental S&T Cooperation Programme '04–'07 (I-17/03).

II. Grooming Algorithms

The network architecture considered in this work is IP over WDM with dynamic optical routing, i.e., optical paths are opened on demand. We assume an overlay model [11].

The optical level is based on Optical Crossconnects (OXC) interconnected by fiber links. Routing is shortest path with First-Fit wavelength assignment for the establishment of lightpaths. OXCs do not have wavelength conversion capabilities. The search for a lightpath is greedy, but it is terminated when a predefined maximum number of crossed links N_l is reached. This threshold is very important both to limit the complexity of routing and wavelength assignment and the waste of optical resources on very long lightpaths.

The IP level assumes traditional routers with shortest path routing based on the number of hops. An optical path is seen as a single hop regardless of the number of OXCs it crosses.

There are two node architectures: i) pure OXC, which allows to switch entire lightpaths from an ingress port to an egress port; ii) Grooming OXC (G-OXC), which supports sub-wavelength traffic flows and groom them onto wavelength channels. A G-OXC is also an IP router. Low-speed traffic can then be transmitted or received only in G-OXCs.

In this architecture, a path connecting two routers in the IP layer is called a *virtual* or logical path, because it is created over some established lightpath in the optical layer. IP traffic dynamically follows the virtual topology build by the optical level underneath. Using this information and the grooming strategy defined below G-OXCs decide whether a new traffic relation must be routed at the IP level or a new lightpath should be opened.

The decision to route the incoming requests over the existing virtual topology or to establish new lightpaths to create more room for them can lead to different network performances. A general analysis of different "grooming policies" is carried out in [4] under the hypothesis of bandwidth-guaranteed (circuit-based) traffic. When elastic traffic is considered, there is no obvious upper limit to the possible number of flows which is routed onto the existing logical layer. In this case, the need for the establishment of new lightpaths must be introduced based on some suitable parameter. We introduce a parameter, called *optical opening threshold* th_o, as a threshold on the instantaneous throughput obtained by connections, defined as a fraction of the peak rate B_M required by each flow.

In this work we consider the following two grooming policies.

Virtual-topology First (*VirtFirst*) **—** Each time a new IP request arrives in some router, the current virtual topology is considered first to route the request. If, once routed, the amount of bandwidth for some flow (not necessarily the one being routed) is less than th_o, a new lightpath is set-up between source and destination (if possible). If the setup is successful, the IP request is routed over it (it is a one hop route at the IP level) and a new virtual topology is computed at the IP level. The new topology does not affect already routed requests (i.e., no re-routing is considered), but will be used for routing all new requests. If the new lightpath cannot be set up, then the request is routed based on the current virtual topology. Whenever a closing flow leaves a lightpath empty, the lightpath is closed too (after a suitable timeout) and the virtual topology is re-computed.

Optical-level First (*OptFirst*) **—** Each time a new IP request arrives in some router, the G-OXC always attempts first to set up a new lightpath in the optical layer, in order to route the request over it. If no free wavelengths are available, the incoming request is routed on the current virtual topology. As in *VirtFirst* if a closing flow leaves a lightpath empty, the lightpath is closed.

These two opposite policies have been often considered by different authors to perform comparisons with new proposals or to study the impact of some specific network constraints, such as OXC node's architecture. In this paper we considered them to study the impact of elastic traffic and analyze whether it affects them differently.

III. The Simulation Tool

The simulator we developed for this study, named GANCLES is described extensively in [12] and a web page [13] is maintained where the software is available. GANCLES is an extension of the connection level simulator ANCLES [14].

Several improvements were made to ANCLES over the years, some regarding the introduction of best-effort, elastic traffic as described in [10] and some related to the introduction of optical routing capabilities [15].

The key point required to jointly study the IP and the optical level is the capability of handling both a physical topology (a directed graph of links and OXCs) and a virtual topology (a directed graph of virtual links and the routers embedded in G-OXCs). The links of the virtual topology match lightpaths provided dynamically by the lower level. Dynamic grooming solutions presented in Sect. II (and others being added) correlate the optical and IP level during simulations.

Several routing and management schemes are available at both levels. We use only very simple routing algorithms both at the optical and at the IP level to highlight clearly the interaction of grooming algorithms and elastic traffic.

When elastic traffic is considered, no admission control is enforced, and no backpressure on traffic sources is available, the network can become instable, as the number of flows within the networks grows to infinity and their individual throughput goes to zero. To avoid this risk, and to build a more realistic scenario, we introduce a starvation threshold th_s expressed as a fraction of the peak bandwidth B_M required by the flow[1]. If at some time instant one or more flows receive a throughput smaller than th_s (due to the arrival of a new flow), the elastic flow with the highest backlog is immediately closed. An important performance meter is the rate of flows interrupted this way. We call this meter *starvation probability*. Notice that if admission control is enforced, this simply means refusing the arriving flow instead of closing a flow as just described. The two actions are however not equivalent, because: i) the arriving flow may not be a starved one (e.g., has a smaller required B_M); ii) blocking is not influenced by the flow dimension, while the starvation is higher for larger flows (in bytes); iii) starved flows waste network resources and may influence overall throughput, which is computed only on completed flows.

A. Traffic Models

We introduce two different models of elastic traffic. Both share the characteristic that a flow i arrives to the network with a backlog of data D_i to transmit and both include some form of elasticity, though very different one another.

The first model, that we name *time-based* (TB), assumes that the elasticity is taken into account only reducing the transfer rate when congestion arises. The flow duration τ_i is determined when the flow arrives to the network, based on its backlog D_i and its "requested bandwidth" B_{Mi} (e.g., the peak negotiated rate, or the access link speed) $\tau_i = (D_i)/(B_{Mi})$. During the connection lifetime bandwidth is then shared according to the max-min criterion. Congestion reduces the throughput, but the closing time is not affected. A consequence is that the data actually transferred by a flow i is generally less than the "requested" amount D_i. This model is very simple and does not grab all the complexity of the closed-loop interaction between the sources and the network. It simply models the fact that the more congested is the network, the smaller is the throughput the flows get.

The second model, that we name *data-based* (DB), assumes instead an ideal max-min sharing of the resources within the network at any given instant. Flows still arrive to the network with a backlog D_i, but the acceptance of a new flow will affect not only all the other flows on the same path, but indeed

[1]Notice that th_s is structurally identical to th_o introduced in Sect.II, but its meaning is very different and its numerical value can be different too.

all the flows in the network, since the max-min fair share is completely recomputed updating the estimated closing time of all the flows in the network. The same applies when flows close, freeing network resources. This model includes the most important feature of elastic traffic, which is the feedback on the flows duration. The more congested is the network, the longer flows remain in the network. Congestion spreads over time enhancing the possibility that still further flows arrive in the network worsening congestion.

The DB model is more accurate, mimicking the behavior of an ideal congestion control scheme; however, its complexity and computational burden are larger, specially for high loads. Investigating whether (or under which conditions) the simpler TB model is accurate enough in the context of IP over WDM with dynamic grooming, or if it leads to gross approximations can be very important.

IV. Numerical Examples

As we discuss in Sect. IV-B, the phenomena involved in routing/grooming elastic traffic are complex, and often far from intuitive. As performance parameters we consider the following five: three at the IP level and two at the optical level.

- T: The average throughput per flow $T = \frac{1}{N_c}\sum_{i=1}^{N_c} T_i$ where N_c is the number of observed flows. Notice that in a resource sharing environment this is not the average resource occupation divided by the number of flows, since flows have all the same weight, regardless of their dimension.
- p_s: The starvation probability as defined before.
- R_o: The ratio between the opening rate of optical paths and the arrival rate of flows at the IP level. It is a measure of the optical level routing effort. For an optical routed network without grooming $R_o = 1$, while for a purely IP routed network $R_o = 0$.
- N_{lo}: Average number of links per optical path.

The goal of a grooming algorithm is maximizing T while minimizing p_s, R_o and N_{lo}.

Before discussing results on a mesh topology, we highlight some peculiar behavior of the *VirtFirst* grooming in a very simple scenario, that will help in interpreting results in more complex scenarios.

A. A Trivial Example

Consider the simple 3-node topology of Fig. 1 a), where only a single wavelength per link is present and the active traffic relations are only A–B, B–C, and A–C. Assume that *VirtFirst* grooming is used and, starting with the network empty, the following sequence of flows arrives: AB, BC, AC, AC, AC, AC, ...(AB identifies a flow originating in A with destination B and so on). We set $th_o = 0.2$ and $th_s = 0$ and all flows are able to fully exploit the optical path capacity. The average throughput obtained by flows is represented by the solid line with cross marks in Fig. 2 (this curve refers to the top x-axis (numer of flows) and left y-axis (normalized throughput)), as

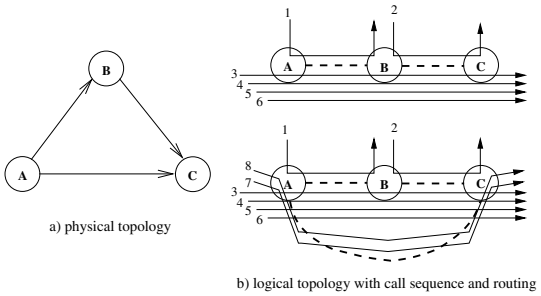


Fig. 1. Simple 3-node topology used for the theoretic verification of results

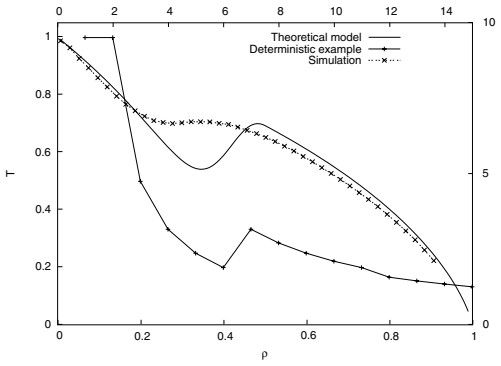


Fig. 2. Average throughput computed deterministically, via simulation and with a simple stochastic model for the scenario depicted in Fig. 1 a)

can be easily seen following the logical topology evolution reported in Fig. 1 b).

This example show that with *VirtFirst* grooming, it is possible that T increases while the load increases due to the interaction between the IP and optical layer. However, a deterministic example is not enough to draw conclusions. In order to investigate further in the behavior, we have set up a simple (and approximate) queuing model of the same scenario based on processor sharing queues that mimic the max-min resource division. Details about the model can be found in [16].

Fig. 2 reports, beside the simple deterministic example, results obtained with the simple stochastic model and with simulations (DB traffic model) for $th_o = 0.2$. These two curves are plotted versus the bottom (ρ) and right (absolute throughput) axes. The simulation curve does not show the same increase in throughput around the load $\rho = 0.5$ displayed by the model (however, we have observed it for smaller thresholds th_o). The reason is that the dynamic routing of flows makes the transition from routing the AC traffic mainly through A–B and B–C to routing it mainly over A–C smoother than in the approximate model. In this case we set $th_s = 0$, so that $p_s = 0$. Given the simple scenario $N_{lo} = 1$, while R_o and N_h are not of much interest.

This simple example give some insight on the complex behavior of grooming associated with elastic traffic, which, to the best of our knowledge was never observed in other works, that, using constant-bit-rate like traffic models, cannot observe throughput performance. In the following we study a more realistic scenario, to gain more insight on grooming and

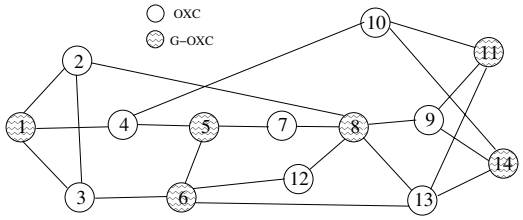


Fig. 3. NSFNET network

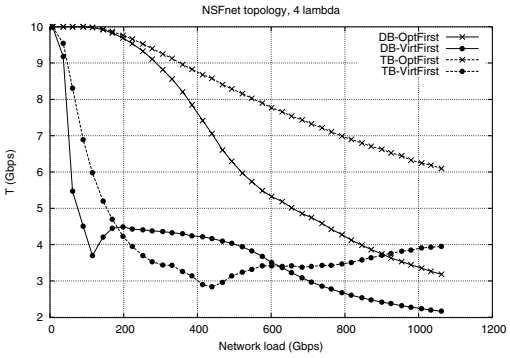


Fig. 4. Average throughput T for the DB and TB traffic models for the two grooming policies

elastic traffic interaction.

B. Results for a Mesh Topology

We present results obtained on the NSFNET network shown in Fig. 3, which has 14 nodes and 21 fiber links. Each fiber carry up to 4 wavelengths, and only 6 nodes out of 14 are G-OXCs. Each wavelength has a capacity of 20 Gbit/s. A traffic source is connected to each G-OXC, opening flows with $B_M = 10$ Gbit/s; each flow transfer data whose size is randomly chosen from an exponential distribution with average 12.5 Gbytes. A uniform traffic pattern is simulated, i.e., when a new traffic relation is generated, the source and destination are randomly chosen with the same probability; $th_s = 0.1$ in all simulations and $th_o = th_s$ for the sake of simplicity. All simulations are run until performance indices reach a 95% confidence level over a $\pm 5\%$ confidence interval around the point estimate. We have run simulations on other topologies obtaining similar results, not reported here for lack of space [16].

Fig. 4 presents a comparison of the average throughput T obtained modeling best-effort traffic relations using the TB approach (dotted lines) and the DB approach (solid lines) when the two grooming algorithms *VirtFirst* (round marks) and *OptFirst* (cross marks) are used. With the same graphic rules, Fig. 5 reports the starvation probability. The difference in performance results of the two approaches is striking.

Let's consider first the *OptFirst* grooming policy. Both approaches show T starting from 10 Gbit/s when the offered load is low; however, they immediately diverge as the offered load increases. Indeed, the DB traffic model shows much faster decrease in T as soon as the offered load increases and this is due to the spreading of congestion over time with a sort of

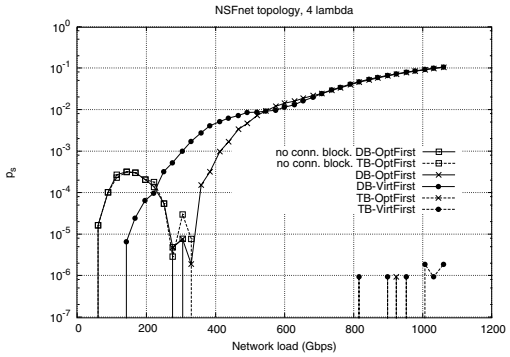


Fig. 5. Starvation probability p_s for the DB and TB traffic models for the two grooming algorithms

snow-ball effect. On the contrary, the TB traffic model shows a smoother decrease of the average bandwidth.

Analysing the starvation probability in Fig. 5 adds more insight. When the traffic is very low (below 350 Gbit/s) both traffic models show the same, very strange behavior: the starvation increases and then decreases sharply. This form of blocking is independent of the traffic model and it is due to a very aggressive and dynamic use of optical resources that sometimes leads to have no connectivity at the IP level, i.e., a flow request arrive and there is no possible path, neither optical, nor through multiple IP hops, between the source and the destination. When the load increases, however, lightpaths become more stable (because there is always traffic keeping lightpaths open) and the probability that the virtual topology is not completely connected becomes negligible. To highlight the difference of this phenomenon from the real starvation, in Fig. 5 the curves relative to it are plotted with square marks. When the load increases further, the two traffic models behavior diverges: the TB model show no starvation at all, apart from points at very high loads, which show a blocking probability around 10^{-6}, while the DB model show a starvation probability increasing steadily.

When considering the *VirtFirst* grooming policy instead, the behavior of both traffic model is different from the previous one. Both DB and TB T decrease sharply even when the offered load is low, due to the conservative policy of *VirtFirst*. In fact, *VirtFirst* sets up the minimum number of lightpaths in order to guarantee the minimum network connectivity, and keeps this configuration unchanged until some flow crosses the starvation threshold th_s. Only in this case *VirtFirst* increases the resources at IP level by setting up new lightpaths. In particular, the T for DB traffic relations decreases very rapidly, causing an earlier set-up of new lightpaths compared to TB traffic. This lead the DB traffic throughput T to "bounce" taking advantage of the higher number of lightpaths in the network, at a load much smaller than for the TB model, that starts increasing again at higher loads. Obviously, both models would show another (and definitive) decrease in T for higher loads, not shown here. Notice that the starvation rate (see Fig. 5) for the TB model is in this case always zero (apart from a single point around load 900 Gbit/s), while the starvation rate of the DB model increases steadily and shows a behavior

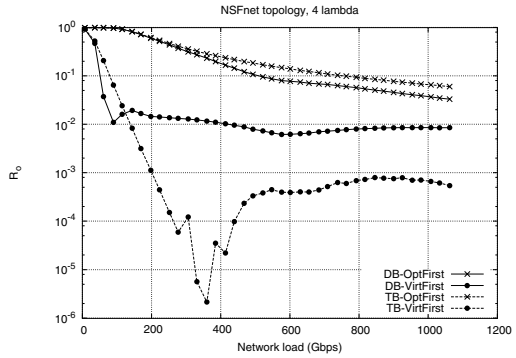


Fig. 6. Ratio R_o between the opening rate of optical paths and the arrival rate of flows at the IP level.

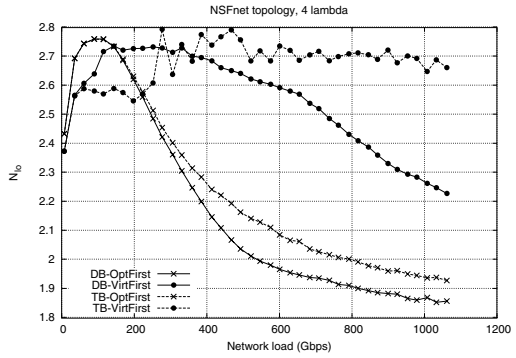


Fig. 7. Average number of links N_{lo} per optical path.

similar to the DB model in the *OptFirst* case.

Fig. 6 shows the ratio R_o. As expected, when *VirtFirst* grooming policy is used, R_o decreases quickly with the load, indicating a burden for the optical level that does not increase with the traffic (indeed, it might also decrease when the load is high). When the *OptFirst* grooming policy is adopted, R_o decreases slowly and smoothly, indicating a much higher burden for the optical level.

Fig. 7, finally, plots N_{lo}. Once again the behavior of the *OptFirst* policy is more predictable, with the number of links that decreases steadily with the load, and roughly converges to the weighted average distance in number of links between G-OXCs. The *VirtFirst* policy shows instead very long optical paths. This effect is due to the intrinsic behavior of this grooming policy: most of the lightpaths set up by *VirtFirst* are in fact never torn-down since they are carrying traffic almost all the time. Then, when new lightpaths must be established, it is more likely that they would be set up in the optical network through longer routes.

V. Discussion & Conclusion

This paper introduced the analysis of dynamic grooming algorithms in IP over WDM with elastic traffic. The elasticity of traffic interacts with the grooming algorithms as well as with the routing both at the IP and optical level, leading to unexpected results.

Two basic grooming policies were considered, one privileging the opening of new optical paths (*OptFirst*), the other privileging the use of the already available IP logical topology (*VirtFirst*).

In both grooming algorithms the impact of elastic traffic, included with a sophisticated model in the simulations tool, is dramatic, showing clearly that approximating IP traffic with CBR-like traffic can lead to wrong conclusions when routing and grooming are considered.

The focus of the paper was on the impact of the traffic elasticity, thus little attention was placed on the "suitability" of the grooming algorithms analyzed. Both the *OptFirst* and *VirtFirst* algorithms, however, have clearly shown that they are not suited for the management of an IP over WDM network, since the lack of coordination between the IP and the optical level leads to waste resources. As shown on the NSFNET topology the *OptFirst* policy may even lead to block requests with very low network loads because a very aggressive use of optical resources may lead to IP-level virtual topologies that are not completely connected.

This observation open new and interesting questions on the heuristics that dynamic grooming algorithms in IP over WDM networks should pursue in order to optimize the use of resources and, at the same time, maximize the satisfaction of the end users.

References

[1] N. Ghani, S. Dixit, T.S. Wang, "On IP-WDM Integration: A Retrospective," *IEEE Communications Magazine*, 41(9):42–45, Sept. 2003.

[2] P. Molinero-Fernandez, N. McKeown, H. Zhang, "Is IP going to take over the world (of communications)?" *ACM SIGCOMM Comp. Comm. Review*, 33(1):113–119, Jan. 2003.

[3] X. Zhang, C. Qiao, "An Effective and Comprehensive Approach to Traffic Grooming and Wavelength Assignment in SONET/WDM Rings," *IEEE/ACM Trans. on Networking*, 8(5):608–617, Oct. 2000.

[4] H. Zhu, H. Zang, K. Zhu, B. Mukherjee, "A novel generic graph model for traffic grooming in heterogeneous WDM mesh networks," *IEEE/ACM Trans. on Networking*, 11(2):285–299, Apr. 2003.

[5] K. Zhu, B. Mukherjee, "Traffic grooming in an optical WDM mesh network," *IEEE JSAC*, 20(1):122–133, Jan. 2003.

[6] M. Brunato, R. Battiti, "A multistart randomized greedy algorithm for traffic grooming on mesh logical topologies," *In Proc. of the 6th IFIP ONDM*, Torino, Italy, Feb. 2002.

[7] M. Kodialam, T.V. Lakshman, "Integrated Dynamic IP and Wavelength Routing in IP over WDM Networks," *In Proc. of INFOCOM 2001*, pp. 358–366, Anchorage, AK, USA, Apr. 22–26 2001.

[8] X. Niu, W.D. Zhong, G. Shen, T.H. Cheng, "Connection Establishment of Label Switched Paths in IP/MPLS over Optical Networks," *Photonic Network Communications*, 6:33–41, July 2003.

[9] R. Srinivasan, A.K. Somani, "Dynamic Routing in WDM Grooming Networks," *Photonic Network Communications*, 5:123–135, Mar. 2003.

[10] C. Casetti, R. Lo Cigno, M. Mellia, M. Munafò, Z. Zsóka "A Realistic Model to Evaluate Routing Algorithms in the Internet," *In Proc. IEEE Globecom 2001*, S. Antonio, TX, USA, pp. 1882–85, Nov. 25–29, 2001.

[11] B. Rajagopalan, J. Luciani, D. Awduche, "IP over Optical Networks: A Framework," RFC 3717, IETF, Mar. 2004.

[12] E. Salvadori, Z. Zsóka, D. Severina, R. Lo Cigno, "GANCLES: A Network Level Simulator to Study Optical Routing, Wavelength Assignment and Grooming Algorithms," DIT-04-017, 2004. `www.dit.unitn.it/locigno/preprints/SZL4-17.pdf`

[13] GANCLES - Grooming cApable Network Call-Level Simulator. `ardent.unitn.it/gancles`

[14] ANCLES - A Network Call-Level Simulator. `www.tlc-networks.polito.it/ancles`

[15] ASONNCLES - ASOn Network Call-Level Simulator. `www.hit.bme.hu/~zsoka/asoncles`

[16] E. Salvadori, R. Lo Cigno, Z. Zsóka, "Analysis of Elastic Traffic Effects on WDM Dynamic Grooming Algorithms (extended version)," DIT-04-091, 2004. `www.dit.unitn.it/locigno/preprints/SLZ4-91.pdf`

Dynamic Grooming in IP over WDM Networks: A Study with Realistic Traffic based on GANCLES Simulation Package

E. Salvadori, R. Lo Cigno
Dipartimento di Informatica e Telecomunicazioni
Università di Trento
Via Sommarive, 14, 38050 Povo, Trento, Italy
e-mail: {salvador,locigno}@dit.unitn.it

Z. Zsóka
Department of Telecommunications
Budapest University of Technology and Economics
Magyar tudósok körútja 2, 1117 Budapest, Hungary
e-mail: zsoka@hit.bme.hu

***Abstract*— Dynamic grooming capabilities lies at the hearth of many envisaged scenarios for IP over Optical networks, but studies on its performance are still in their infancy. This work addresses two fundamental aspects of the problem.**

First of all it presents a novel tool for the study of IP over Optical networks. The tool, freely available on-line, is a network level simulator named GANCLES that includes several innovative features allowing the study of realistic scenarios in IP over Optical networking, making it an ideal tool for Traffic Engineering purposes. GANCLES architecture enables the simulation of dynamic traffic grooming on top of a realistic network model that correctly describes the logical interaction between the optical and the IP layer, i.e., the mutual relationship between routing algorithms and lightpath assignment procedures at the optical layer and routing at the IP layer. Adding or removing lightpaths changes the logical IP topology, which affects IP routing and traffic patterns. The simulator allows for the description of Overlay, Augmented or Peer IP over Optical architectures, depending on the amount of information shared between the IP and optical domain.

Second it analyzes and discusses several performance indices and aspects of different grooming policies in the IPO Overlay model, using different traffic models, some of them including elasticity of best effort traffic. Both regular and mesh topologies are analyzed, and results clearly show that the correct evaluation of dynamic grooming policies in IPO networks requires a sophisticated level of modeling, since simplistic assumptions like Poisson traffic, or the incorrect representation of the interaction of IP and Optical control planes, may induce misleading results.

I. Introduction and Related Work

IP over WDM networks [1] are spreading very fast and even tier 1 providers have begun to encapsulate IP packets directly in the optical layer, avoiding the use of sophisticated middle layers such as ATM (Asynchronous Transfer Mode) or SONET/SDH (Synchronous Optical NETwork/Synchronous Digital Hierarchy).

The optical layer is managed through protocols like GMPLS (Generalized Multi-Protocol Label Switching) [2] or ASON (Automatically Switched Optical Network) [3], while the IP layer is either integrated in a MPLS framework or uses standard intra-AS IP routing protocols such as OSPF or IS-IS.

Dynamic grooming of IP traffic over a wavelength routed optical network means that the two routing layers (IP and optical) interact, with deep impacts on Traffic Engineering (TE) and QoS provisioning. The interaction nature depends on the grooming algorithm, as well as on the amount of information (if any) exchanged between the two layers. This situation is very complex and its study is normally done via simulations with a modeling effort to reduce the problem complexity, e.g., without simulating packet level traffic, but with fluid models. Sivalingam et al. have presented an *ns*-2 based simulation tool for performance studies of WDM networks [4]. This simulator does not consider the problem of grooming and the WDM management layer is seen as a logical layer on top of an IP network (the standard *ns*-2 network layer) building virtual circuits on the packed switched routing layer, thus its "philosophy" is completely different from GANCLES.

From the perspective of performance analysis, a few recent studies started considering the problem of dynamic grooming [5], [6], [7], [8], [9], but, to the best of our knowledge, only simplistic models of the network dynamics associated to CBR Poisson traffic sources were used in these studies. More traditional studies such as [5], [10], [11], [12] are based on the simplifying assumption that the traffic is static and "circuit-like." In a previous contribution [13] we discussed in detail the impact of realistic traffic models on dynamic grooming, showing the inherent interaction between the IP and the optical layer and its effect on the overall performance of the network. Traffic flows in this models

This work is supported by the Italian Ministry of Education and Research (MIUR) through the GRID.IT and ADONIS projects, and by the Hungarian Italian Intergovernmental S&T Cooperation Programme '04–'07 (I-17/03); GANCLES is developed under the E-NEXT NoE umbrella.

share the resources on a virtual topology path following the max-min fairness criterion [14], thus mimicking the ideal behavior of a bundle of TCP connections.

The contribution of this paper is twofold. In the first part we present GANCLES, making it available to the community for use in research and applications. A major achievement in GANCLES is that it allows the study of the interaction between four major drivers of the overall network performance:

- The optical layer Routing and Wavelength Assignment (RWA) algorithms;
- The IP level routing;
- The grooming policies used to mediate between the IP level and the optical level;
- The adaptability of current data applications, that, being based on TCP, modify the transmission rate following the availability of resources.

In particular the clear separation of the IP and Optical control planes, enables the correct definition of the IPO model (peer, augmented, overlay) [15] addressed with the simulation experiment, and the presence of elastic traffic makes the experiments representative of IP over Optical, relaxing the usual approximation based on CBR, bandwidth guaranteed traffic.

In the second part we present results for different, albeit not entirely new, dynamic grooming policies, comparing their behavior and highlighting the scenarios, conditions and performance indices that require proper, in-depth studies for the assessment of dynamic grooming policies and architectures. In this work we consider only the Overlay model, in part because it is impossible to include all different facets in a single work and in part because we think the Overlay model is the only one which is technologically foreseeable in the next future, and for this reason, the most interesting from the implementation point of view.

The remaining part of the paper is organized as follows. Sect. II describes thoroughly the simulation environment and the innovative features we introduced to study dynamic IP over WDM networks. In Sect. III we define some performance figures of IPO networks which can be studied by using GANCLES. Sect. IV is devoted to the discussion of results, both in a ring and on the NSF topology; Sect. V ends the paper and discusses future work enabled by GANCLES.

II. The Tool Features and Architecture

GANCLES [16] is an event-driven asynchronous simulator derived from ANCLES [17]. ANCLES has gradually evolved over the years to allow the simulation of elastic connections over IP networks, with a flow-level granularity, as described in [18]. A separate extension allows the study of different lightpath-level granularity in ASON-based wavelength-routed networks [19].

GANCLES integrates the two network layers (from now on: the *data-layer* and the *optical-layer*), thus allowing the in-depth study of the interaction between them when a multi-layer network environment such as IP over WDM is considered. The objective of the simulator is to give researchers a useful tool to study new algorithms and protocols, to analyze network performance, to implement traffic engineering criteria, and to design QoS provisioning means in this multi-layer environment.

The simulator includes advanced tools to perform statistical analysis based on the "batch-means" technique [20]. Simulation experiments are stopped when the desired accuracy is reached on a selected set of performance parameters.

The tool allows the computation of a large number of performance indexes, both for the entire network and for selected traffic relations. Some of them are illustrated in more detail in Sect. III.

The network models simulated by the tool are composed of instances of three basic entities.

- NODES, which perform the routing functions at the IP and at the WDM layer, and implement the CAC and the grooming algorithms; NODES can be either pure OXCs (Optical Crossconnect), switching entire lightpaths, or include an IP router on top of the OXC, allowing for data-layer traffic injections/extraction and performing grooming operations; these nodes are named G-OXC (Grooming-OXC).
- CHANNELS, that accommodate the information transfer between either adjacent NODES or USER-NODE pairs; a CHANNEL can accommodate up to W independent lightpaths.
- USERS, that acts as sources and sinks for the traffic flowing through the network; USERS can be both at the optical-layer, generating "circuit-like" requests of entire lightpaths, and connected directly to OXCs, or at the data-layer, connected to G-OXCs only and generating sub-wavelength requests that can follow one of three different "models:" i) traditional circuit-like requests, ii) Time Based (TB) best-effort requests, and iii) Data Based (DB) best-effort requests (See Sect II-C for further details).

Simulations are specified with a formal grammar inherited from [17] and named ND. The number of CHANNELS and their data rates are expressed in number of fiber per link, number of wavelengths per fiber and finally, transmission capacity in Gbit/s per wavelength. The NODE architecture is described in detail discussing the interaction between the data-layer and optical-layer in Sect. II-B

A. IP over Optical architectures

One of the most important feature of GANCLES is enabling the implementation of dynamic grooming algorithms which refer to different IP over Optical (IPO) architectures as defined by RFC 3717 [15]. According to this RFC, when considering the *control plane* of an IPO network,

three different interconnection models can be envisioned according to the amount of information exchanged between the IP and optical layer: *Overlay, Augmented and Peer*.

In the *peer* model, IP routers and OXC are considered peer network elements, thus the topology and other network information are completely shared by a unified control plane. In the *overlay* model, each layer performs its own routing functions since no information is exchanged between them. An intermediate architecture between these two is the *augmented* model, where some aggregated information from one routing instance is passed to the other, in general only from the optical to the IP layer, in order to allow this latter to take more informed decisions when submitting requests for additional resources.

The peer and the augmented model are appealing because sharing the knowledge base between the two layers allows running an integrated routing function, using, for instance, an auxiliary graph, as done in [5]. The integrated management enables a better usage of the overall network resources. However, both models seem not feasible in the near term due to the tight integration between the two levels and scalability issues regarding the amount of exchanged information. The overlay model is instead technically feasible, since it only requires the definition of a clear interface between the IP and optical level and dynamic lightpath capabilities in the optical level, which are being experimented in laboratories and research projects [21].

Due to its specific implementation characteristics, GANCLES enables the study of scenarios with full, partial or no exchange of information between the data-layer and the optical-layer, therefore allowing the description of any of the grooming algorithms proposed so far in literature.

B. Physical and Logical Topology Management and Interaction

As mentioned before, in GANCLES nodes can be pure OXCs, that switch entire lightpaths from an ingress port to an egress port, or they can be G-OXCs that support sub-wavelength traffic flows and multiplex them onto wavelength channels through a grooming fabric. OCXs and G-OXCs can be mixed freely into a simulation experiment. It is possible to have both *full opaque* or *full transparent* crossconnects. Opaque OXCs allows full wavelength conversion; transparent OXCs have no conversion capabilities. Partially opaque OXCs, with limited conversion capabilities, are being implemented. A G-OXC is *also* a router, hence the transit traffic (not terminated in the router), can be groomed with incoming traffic. Sub-wavelength traffic can be generated and received only in G-OXCs.

When considering a multi-layer environment, with connections flowing between IP level nodes through an optical network, we need to distinguish the *physical* topology and the *logical* topology. The latter one is made of all the lightpaths established between G-OXCs over the physical topology according to some optical-level routing algorithms. The logical topology is used for routing at the data-layer (IP) and it is modified each time the grooming management entity triggers the establishment or release of some lightpaths.

Simulation in GANCLES are driven by the USERS, which collect requests from their associated call generators and forward them to the network, while acting as destinations for the connections coming from remote users.

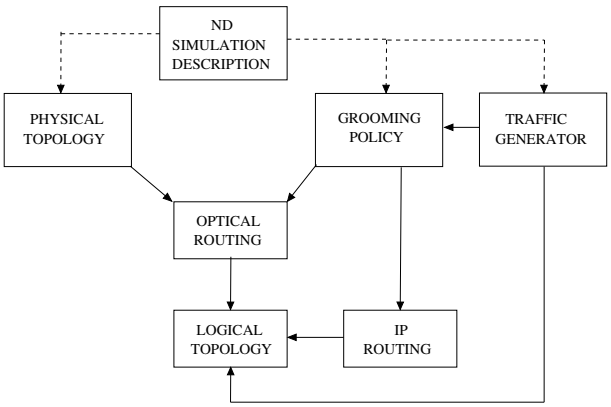


Fig. 1. Logical interaction between different high-level modules in GANCLES, the management of the optical-layer is mediated by the grooming strategies and algorithms

Fig. 1 represents the interaction between different GANCLES parts. A simulation starts after GANCLES has acquired the simulation experiment description in ND. The description includes: (i) the network topology in terms of a weighted graph connecting USERS and NODES through CHANNELS; (ii) the traffic relations between USERS and the statistical characterization of sources; (iii) the selection of the routing, CAC and grooming algorithms adopted for the simulation run; (iv) a number of options concerning both the network operation and the simulation session management; (v) the performance indices to be measured.

When the simulation starts an optical-layer function independent from the selected RWA algorithm defines the set of available physical paths for each pair of OXCs/G-OXCs ordering them according to some specific criterion. Only this set can be used by the routing algorithm, reducing the routing problem complexity.

Every time a new lightpath is added or removed from the network by the optical-layer, the data-layer (logical) topology is changed. As done at the optical-layer, also in this case the set of possible logical paths between each G-OXC (router) pair is computed following a user-specified criterion. This task is extremely critical, because path-computing is very time consuming, hence the path selection criterion must be carefully defined depending on the data-layer routing algorithm selected (e.g., only one path need to be computed if Fixed Shortest Path routing is used).

Notice that the overall setup (and performance) of IPO networks is heavily influenced by technology constraints. We already mentioned opaque or transparent OXCs, but other constraints, such as simplex- or duplex-only lightpath management also come into play at the optical-layer. Similar constraints may arise if TE techniques are used at the data-layer. GANCLES allows the simulation of an arbitrary mix of these constraints.

Each time a USER generates a connection request, the grooming entity decides whether: (i) route it over the current topology; or (ii) ask the optical-layer to open one or more lightpaths (thus modifying the logical topology) and then route the request at the data-layer. In the first case, a data-layer CAC algorithm (if any) is executed for each of the paths being considered by the data-layer routing algorithm; if no path is found to route the incoming request in agreement with its QoS requirements, the request is dropped. In the second case, after the logical topology modification, the same data-layer routing and CAC are applied. As a general rule, in this case a connection can be refused only if the optical-layer was not able to modify the logical topology meeting the grooming algorithm requirements. For each request the grooming procedure may require an arbitrary mix of actions of type (i) and (ii), depending on the complexity of the algorithm implemented and on the model (peer, augmented, overlay) assumed for the IPO.

The interaction between the optical-layer and the data-layer lies at the core of dynamic grooming problems. This interaction is described in GANCLES (as in real networks) by the optical/IP control interface within G-OXCs. This simple and clear, though realistic, interface implementation is one of the innovative features of GANCLES.

Modifying the logical topology, the question arises whether all the traffic is re-routed or if only new connections follow the new available routes. GANCLES presently assumes the second option, but its modification for re-routing is trivial.

When an active connection terminates, the corresponding resources in the logical topology are released. This operation can lead to the release of some lightpaths if they are not carrying traffic anymore. In this case, paths at the IP level need to be recalculated again over the new logical topology. The specific lightpath releasing criteria (e.g., a given time-lapse without traffic) can be specified by GANCLES user.

GANCLES implicitly assumes the utilization of a separated control plane (e.g., GMPLS) to keep each node informed of the network status. The control traffic is not considered in the performance measures.

C. Traffic Sources

The present release of GANCLES provides different types of call generators at both the data-layer and optical-layer. Each connection request is associated with the identifier of the destination USER, which is chosen accordingly to the traffic relations specified for the experiment.

The main traffic models implemented for the data-layer USERS are: CBR, ON-OFF CBR, Uniform VBR, Video VBR or Best-Effort with TB and DB model (see [17]), while for the optical-layer USERS it is possible to generate lightpath requests according to Poisson traffic generators, but it is also possible to generate permanent and semi-permanent lightpaths (see [19]), a feature that enables adding constraints or static portions to the logical topology.

Since GANCLES has been developed mainly to study the interaction between IP (as data-layer) and the optical-layer, a more detailed explanation of the *elastic* traffic features of Best-Effort USERS is needed. As a general rule IP does not implement CAC functions and the congestion control is done reactively by TCP. As we discussed in [13] the elastic nature of present-day data applications cannot be disregarded if dynamic grooming is considered, because the feedback introduced by the closed-loop nature of TCP (and traffic aggregation does not destroy the feedback) has an enormous impact on the overall performance. We introduce two different models of elastic traffic. Both share the characteristic that a flow i arrives to the network with a backlog of data D_i to transmit. Both include some form of elasticity, though very different one another.

The first model, named *time-based* (TB), assumes that the elasticity is taken into account only reducing the transfer rate when congestion arises. The flow duration is determined when the flow arrives to the network, based on its backlog D_i and its peak transmission rate B_{Mi} (e.g., the access link speed) $\tau_i = \dfrac{D_i}{B_{Mi}}$. The effect of congestion is just that the throughput of flows is reduced, but their closing time is not affected. A consequence of this behavior is that the data actually transferred by a flow i is generally less than the "requested" amount D_i, thus reducing the actual network load and relieving congestion. This model is very simple and does not grab all the complexity of the closed-loop interaction between the sources and the network.

A more accurate model, named *data-based* (DB), assumes instead an ideal max-min sharing of the resources within the network at any given instant. Flows still arrive to the network with a backlog D_i and with a peak transmission rate B_{Mi}. The acceptance of a new flow will affect not only all the other flows on the same path, but indeed all the flows in the network, since the max-min fair share is completely recomputed updating the estimated closing time of all the flows in the network. The same applies when flows close, freeing network resources. This model includes the most important feature of elastic traffic, which is the positive feedback on the flows duration. The more congested is the network, the longer the accepted flows remain in the network.

Without a CAC, at high loads the network can become

instable, as the number of flows within the network can grow to infinity and their individual throughput goes to zero. To build a more realistic scenario, a second attribute has been introduced to characterize any IP flow i: a minimum requested rate b_{mi}. If at any time the bandwidth assigned to flow i falls below it, then flow i closes and is counted as a "starved" flow, because the network was not able to guarantee its correct completion. The attributes b_{mi} and B_{Mi} are included in some SLA (Service Level Agreement) at the IP/Optical interface (see [22] for initial works on Optical-SLA).

A new performance measure can therefore be introduced, called the *starvation probability* p_s, which complements more traditional metrics such as throughput, blocking probability, optical-layer overhead in opening and closing lightpaths, etc.

D. Routing Algorithms

Inheriting the terminology in [17], whatever criterion is used to order the paths at both the optical-layer or data-layer, a *primary path* is always defined for each pair of NODES. All the other allowed paths are referred to as *secondary* paths.

Data-layer Routing Several alternatives are available to route calls at data-level so as to take into account the dynamic load of the network. An important property of GANCLES is the possibility to simulate both source-based (e.g., MPLS-like) or hop-by-hop routing.

The following list enumerates the main routing algorithms implemented in GANCLES. The reader is referred to [17] for more details on this part and the relative references to the literature.

- *Single Path Routing*: only the primary path between the nodes is considered to route the connection. This is also known as Fixed Shortest Path.
- *Controlled & Uncontrolled Alternate Routing*: if there is no space for the connection along the primary path, the secondary ones are investigated with different constraints (see [23] for details).
- *Minimum Distance Routing*: for each source-destination pair, the path π is chosen that minimizes the following quantity: $C_\pi = \sum_{l \in \pi} \frac{1}{b_l}$ where b_l is the max-min fair bandwidth that is available to a new connection over link l belonging to path π (see [24] for details).
- *Widest-Shortest Routing*: this algorithm first identifies the minimum-hop-count paths and breaks ties by choosing, among the paths with the minimum hop count, the one with the maximum available bandwidth (see [25] for details).

Optical-layer Routing There are different static and dynamic algorithms implemented for the routing of optical-layer requests. In the following only single-path routing algorithms are described, but the tool provides also for protective routing, with both *dedicated* or *shared* protection mechanisms (see [19] for more details).

If we use opaque OXCs, the wavelength continuity is necessary only in the transparent sections, i.e. the part of route connecting two opaque nodes. Allowing wavelength conversion in the opaque nodes implies that the wavelength assigned to the connection can be different in each transparent section. The main optical routing algorithms implemented in GANCLES are:

- *Fixed Shortest Path*: it routes the lightpath request always on the dedicated primary path between the endpoints.
- *Shortest-Widest Path*: it selects the paths with the largest number of available optical channels; if there are more possibilities, it routes the lightpath on the shortest among them.
- *Alternate Shortest Path*: it selects the shortest from those paths where there is at least one wavelength available.

Furthermore, Wavelength Assignment (WA) algorithms can be freely associated to any routing. WA algorithms include *Random* and *First-Fit*, and others can be added easily if required (details and references on optical-layer routing and RWA can be found in [19]).

E. Grooming Algorithms

Grooming policies are the ensemble of algorithms and protocols that takes the decisions regarding possible changes of the current logical topology each time a data-layer request arrives or leaves the network. When new logical links need to be installed, two factors must be determined: how many of them must be set up and between which OXCs.

The decision to route the incoming requests over the existing logical topology or to establish new lightpaths to create more room for them can lead to different network performances. A general analysis of different "grooming policies" is carried out in [5] under the hypothesis of bandwidth-guaranteed traffic within a peer IPO model. When *elastic traffic* is considered, there is no obvious upper limit to the possible number of flows which is routed onto the existing logical layer. In this case, the need for the establishment of new lightpaths must be introduced based on some suitable parameter. We introduce this parameter, called *optical opening threshold* τ_o, as a threshold on the throughput that the incoming request would achieve on the selected route, defined as a fraction of the peak rate B_{Mi} required by each flow i present on the path analyzed.

Three grooming solutions are implemented in the current version of the simulator:

- *Virtual-topology First* (*VirtFirst*). This grooming algorithm aims to open new optical lightpaths only if the

current virtual (logical) topology does not have enough resources to carry the incoming request. In case of new lightpath setup, if it is successful, the IP request is routed over it, otherwise the request is routed based on the current logical topology.

- *Optical-level First* (*OptFirst*). Each time a new data-level request arrives in some router, the G-OXC always attempts to set up a new lightpath in the optical layer, in order to route the request over it. If no free wavelengths are available, the IP router routes the incoming request over the current logical topology. Indeed, a logical topology is defined only when optical resources for the considered source-destination pair are exhausted.
- *HopConstrained* (*HopCons*). In this case, the decision on opening a new connection is taken based on path-based constraints. Given a number of IP-level hops K, a new lightpath between source s and destination d is set up if it is impossible to find a path on the current logical topology with less than $K+1$ hops and more than $B_{Mi} \cdot \tau_o$ bandwidth available. *HopCons* can be considered an intermediate policy between the previous two, in fact it behaves as *VirtFirst* for $K \to \infty$ and as *OptFirst* for $K = 0$. *HopCons* grooming is described in detail and its implications are discussed in [26].

Note that in all these cases if the source-destination pair is disconnected on the logical topology and no new lightpath can be installed between them, the incoming connection request must be refused even if it is elastic. This phenomenon is particularly evident at low/medium loads with grooming algorithms using aggressively the optical resources, as it was shown in [13]. Such unacceptable situations could be avoided by including a pre-defined spanning tree or any other basic logical topology to ensure the connectivity in the data-layer, a feature GANCLES is provided with, defining different pre-defined logical topologies (see [16]).

When a lightpath needs to be released because it is not carrying traffic, the simulator gives the possibility of delaying the closure using a timeout, called *optical closing timeout* τ_{cl}. When no traffic is carried over some lightpath, it is kept open for the timeout period and gets closed only if its state does not change. This parameter can be specified by the user and can be adapted to the traffic characteristics, e.g. it can be set according to the mean interarrival time of data-layer requests. This dependency on the load structure can be very useful to avoid excessive oscillations in the logical topology, which are notoriously harmful to IP routing. Using management information it can be easily implemented in real networks.

III. Performance indices

As discussed in Sect. IV, the phenomena involved in routing/grooming elastic traffic are rather complex, and often far from intuitive. The following performance indices blend both user-perceived performance and network operation costs, thus helping in understanding the global performance of the network.

p_b — Blocking probability. It is the probability that a flow is not accepted, either because of some CAC function decision or (in best-effort traffic) because no connectivity can be found between the source and the destination.

p_s — Starvation probability. It is the probability that a best-effort flow closes during its life because it is not receiving service with acceptable quality. A flow i closes and drops the network if its instantaneous throughput falls below b_{mi}.

T — The average throughput per flow.

$$T = \frac{1}{N_c} \sum_{i=1}^{N_c} T_i$$

where N_c is the number of observed flows (e.g., during a simulation). Notice that in a resource sharing environment this is not the average resource occupation divided by the number of flows, since flows have all the same weight in the average, regardless of their dimension.

R_c — Routing table change rate. Each time a new lightpath is established, some of the routing tables must be recomputed due to the new virtual topology. The rate of such changes is a good measure of the joint grooming and routing cost within the network.

N_h — Average number of IP hops per flow.

N_{lo} — Time weighted average number of links per optical path. Each lightpath is weighted by its holding time, so that lightpaths lasting longer are correctly accounted for.

U_d — Distance unfairness index.

$$U_d = \frac{\max_{0<r<N} T^r - \min_{0<r<N} T^r}{T}$$

measures if the resource assignment is fair with respect to physical distance. T^r is the average throughput calculated for node-pairs with hop distance r in the physical topology, while N is the number of nodes in the network. It ranges from zero to ∞; any value larger than one indicates unacceptable unfairness. This parameter is evaluated only for regular topologies and uniform traffic, since in other cases it is can be influenced by factors external to the grooming policy.

Many other unfairness indices can be defined, based on different flow characteristics, e.g., the flow granularity (see for instance [27]), or in irregular topologies based on the source-destination $s-d$ traffic relation. Distance based unfairness is however typical of elastic traffic and it can be interesting, as discussed in Sect. IV-A on a ring topology, to investigate whether grooming policies can relieve, at least in part, the skewed behavior induced by the max-min sharing criterion.

TABLE I
PARAMETERS USED THROUGHOUT THE SIMULATIONS

Simulation parameters		Value
Fibers per link		1
OXCs		transparent
Number of wavelengths per fiber	W	8 for 8-nodes Ring 4 for NSFNet
Data rate per wavelength	g	20 Gbit/s
Duplex lightpaths		
Request arrival process	λ	Variable with load
Request backlog distribution: exponential	η	12.5 Gbit/s
Requested peak rate	B_M	10 Gbit/s
Requested minimum rate	b_m	1 Gbit/s
optical opening threshold	τ_o	0.1
optical closing timeout	τ_{cl}	0
Traffic pattern		Uniform among G-OXCs
Data flows		mono-directional
IP routing		Fixed Shortest Path
optical-layer routing		*Shortest-Widest* for 8-Ring *alternate shortest path* for NSFNet
Maximum number of hops in *HopCons*	K	1
Confidence interval		1% of point estimate
Confidence level		99%

In general the goal of a grooming algorithm is maximizing T while minimizing p_s and U_d, while it is not always straightforward to define "a goal" for other performance parameters, since they can be subject to contrasting needs.

IV. A SAMPLE STUDY

In this work, we are interested in understanding the fundamental aspects of dynamic grooming in overlay IPO networks, and the inherent interaction with data-based, elastic, best-effort traffic. At the same time we highlight the features we have introduced in GANCLES.

As already mentioned we limit the study to overlay networks because of their inherent simplicity and because we deem they will be the first ones to be deployed. In overlay networks the control planes of the IP and optical network are completely separated and there is no information exchange between IP and optical routing entities: the two layers interact only through the lightpath setup or tear-down. Lightpaths setup is requested by the IP layer whenever it needs additional resources, and lightpaths are teared down when they are not needed anymore[1].

We investigate two different scenarios, that are complementary one another.

8-Ring — This is an 8-node (all G-OXCs) bidirectional ring topology, whose regularity features help in the interpretation of results. Besides, rings are among the topologies of choice for the realization of metropolitan area optical networks.

Modified NSFNet — or NSFNet for short (see Fig. 6). This is a modification of the NSF network topology, where we have limited the number of G-OXCs to study whether an irregular, wide area, mesh topology with additional optical resources (the pure OXCs), with respect to the traffic generation points (G-OXCs) does influence the performance of dynamic grooming in overlay IPO.

Table I summarizes the main simulation parameters that are kept constant throughout the simulations unless otherwise stated. Other parameters are defined and explained when needed.

A. 8-Ring topology

One of the wavelengths is reserved to realize a pre-established logical ring topology to ensure connectivity in the data-layer. Some results on this topology were included in [26]; here we focus on the impact of the traffic model and on fairness issues.

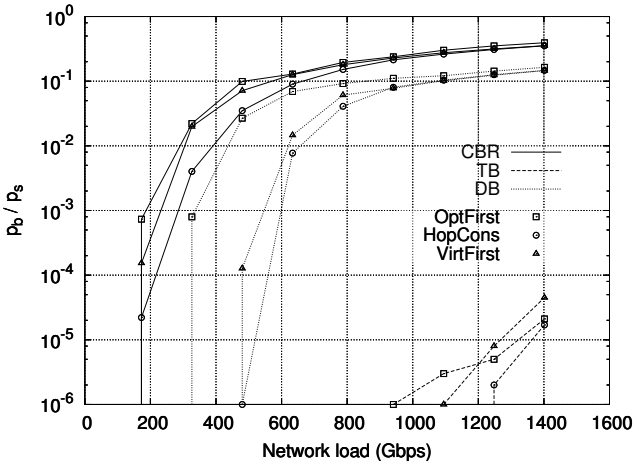


Fig. 2. Blocking probability p_b for the CBR and starvation probability p_s for both the elastic traffic models

Let's first of all analyze the importance of using a realistic traffic model. Fig. 2 presents a comparison between the blocking probability p_b obtained using a circuit-based CBR traffic model (solid lines) with the starvation probability p_s obtained modeling best-effort traffic relations using the TB approach (dashed lines) and the DB approach (dotted lines) when the three grooming algorithms *VirtFirst* (triangle marks), *OptFirst* (square marks) and *HopCons* (round marks) are used. For both elastic traffic models, the minimum requested rate b_m is fixed to 1 Gbit/s.

The difference in performance results yielded by the three approaches is dramatic, showing that the performance

[1]The lightpaths tear-down can be done independently by the optical layer or upon request from an IP entity, but this does not affect either performances or the overall architecture of the network.

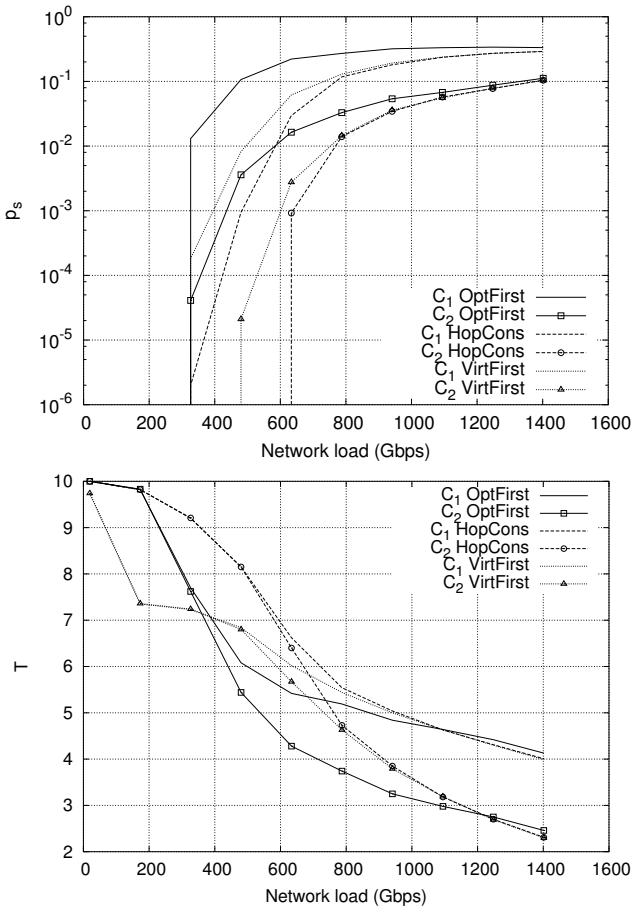


Fig. 3. Evaluating bandwidth requests fairness. Starvation probability p_s (upper plot) and average throughput T (lower plot) users requesting different minimum bandwidth

obtained by the considered grooming algorithms with CBR traffic is very conservative, leading to intolerable levels of blocking probability even for medium to low traffic loads. Instead, when considering a simplistic model of elasticity such as TB, the starvation probability is very low, almost negligible, and this is due to the approximation that a flow request is closed without considering if the data has been completely transmitted or not, thus reducing the actual network load. Therefore, we argue that using a more sophisticated, *data-based* traffic model exhibiting a realistic behavior is essential to evaluate the performance of any dynamic grooming algorithm. From now on we use only the DB traffic model.

We do not include throughput plots because it is useless to compare the throughput of CBR connections (constant!), with the one of elastic traffic. Moreover even the throughput of TB and DB model are not comparable, since the overall amount of transferred data is different.

We now concentrate on fairness issues, analyzing how different grooming algorithms in overlay IPO networks influence fairness. In particular we investigate two main aspects that influence fairness: the minimum bandwidth b_m required by incoming flow requests, and the physical distance (in terms of the number of crossed OXCs) between the end-nodes of a flow. This latter problem was addressed in [27] in the context of peer IPO networks with a bandwidth-guaranteed traffic model. Resource sharing in general exacerbate the unfairness, because longer flows share the resources with more flows with respect to short ones, thus they remain longer within the network and the more they remain the more are the flows they compete with.

A first set of results regarding the fairness of different grooming algorithms for clients with different SLA requirements is shown in Fig. 3. These results have been obtained differentiating the users in terms of the minimum bandwidth b_m they require to the network to avoid starvation. In particular, we consider two classes of users. Class C_1 with minimum bandwidth request $b_{m\,1} = 2\,\text{Gbit/s}$ and class C_2 with $b_{m\,2} = 1\,\text{Gbit/s}$.

The upper plot shows the starvation probability and the lower one the average throughput. It is clear that the relative merits of grooming policies are unchanged by the presence of classes, but the main result to highlight here is that when considering a realistic traffic scenario, none of the grooming algorithms allows to improve the fairness with respect to the minimum requested bandwidth. The explanation is trivial, but the solution seems to be far more complex, like introducing some form of proportional scheduling in nodes or some form of CAC. In fact, when the traffic shares resources with a max-min fairness criterion, the available network resources are fully shared among all the accepted flows, and it is not possible to distinguish between traffic flows with different minimum bandwidth requirements. In other words, a flow with higher minimum requested rate gets starved with higher probability.

As easily predictable and analyzed in [27], dynamic grooming is prone to an unfair behavior toward user pairs with a longer physical distance in terms of crossed OXCs. A ring is the ideal topology to analyze this behavior, since the regularity of the topology does not introduce any distortion effect. In general the dominating effect is that user pairs distant one another have less chances to setup a new lightpath. An intelligent grooming policy should compensate for, at least in part, this inherent unfairness by *sparing* optical resources to dedicate to longer flows. However, in an overlay model and without costly traffic measurements, this might not be easy to implement. A possible, partial solution, may reside in a more conservative closing procedure for long lightpaths. We have defined a *LEngth DEpendent optical closing (LEDE)* policy that defines the delay of closure for a lightpath of H physical hops as $(H-1)\tau_{cl}$.

We now study the fairness of the three grooming al-

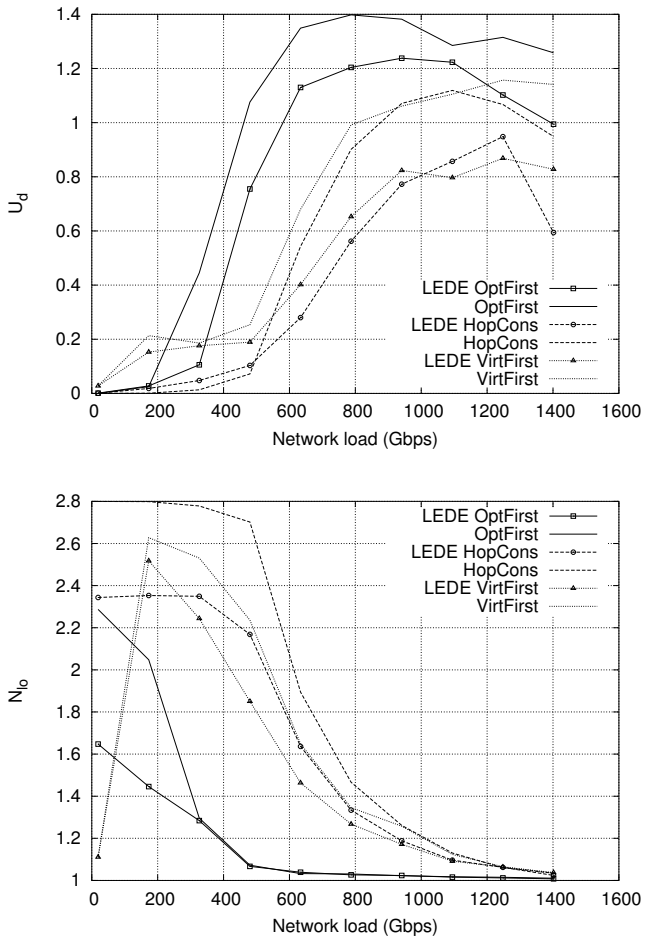


Fig. 4. Comparison for grooming algorithms with and without *LEDE* option: impact over the distance unfairness index U_d(top plot), and average number of links per optical path N_{lo}(bottom plot).

gorithms according to the distance between the end-nodes of a flow, with and without the *LEDE* option as defined above. The base closing timeout τ_{cl} is computed according to the average flow interarrival time as explained in Sect. II-E and it automatically depends on the network load. In each simulation we set $\tau_{cl} = t_{ia}/2$, where t_{ia} is the mean value of the time spent between the arrival of two flow-requests with the same USERS pair as source and destination.

The upper plot in Fig. 4 shows that the *LEDE* option effectively alleviates the unfairness, by uniformly reducing the fairness index. In particular, at very high load (for $p_s \simeq 20\%$), U_d is always below 1 for both *VirtFirst* and *HopCons*.

The impact of the *LEDE* option over the average number of links per optical path is well illustrated in the lower plot, which shows an average increase of N_{lo} for all the grooming algorithms. This behavior proves that increasing the closure timeout on longer optical routes increases the amount of resources dedicated to longer routes and reduces the unfairness toward longer flows in IPO networks.

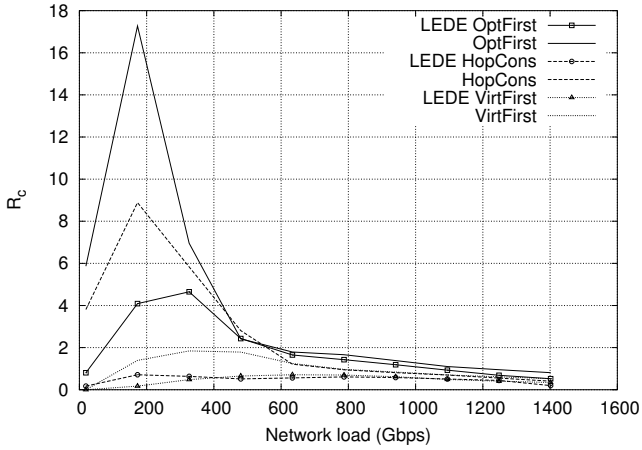


Fig. 5. Comparison for grooming algorithms with and without *LEDE* option: impact over the routing table change rate R_c.

Fig. 5 shows instead a useful 'side-effect' of *LEDE*. In fact, keeping lightpaths not carrying active traffic in the network open for longer periods, the frequency of routing table updates due to virtual topology changes is reduced, which is an important cost factor in an IPO network [15]. Starvation probability and throughput (not shown for the sake of brevity) show that this is obtained without average performance losses.

B. NSFNet topology

We consider the NSFNet topology with a mix of OXCs and G-OXCs as shown in Fig. 6. We concentrate on the impact of IP flow requests granularity and the threshold τ_o on the grooming algorithms performance, referring the reader to [13] for additional results on this topology.

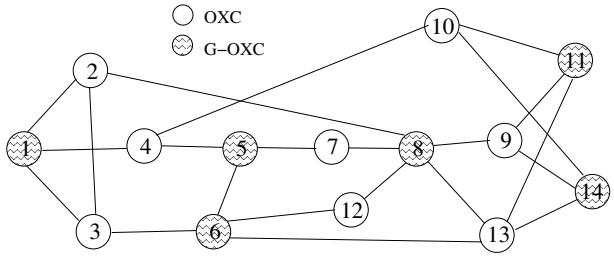


Fig. 6. Topology NSFNet

We first investigate the performance of *VirtFirst* and *OptFirst* as a function of the data-layer requests granularity. Fig. 7 presents T, p_s and p_b of data-layer flows for $B_M = 20, 10$, and 4 Gbit/s, i.e., when requests have a granularity of 1, 1/2, and 1/5, of the wavelength capacity.

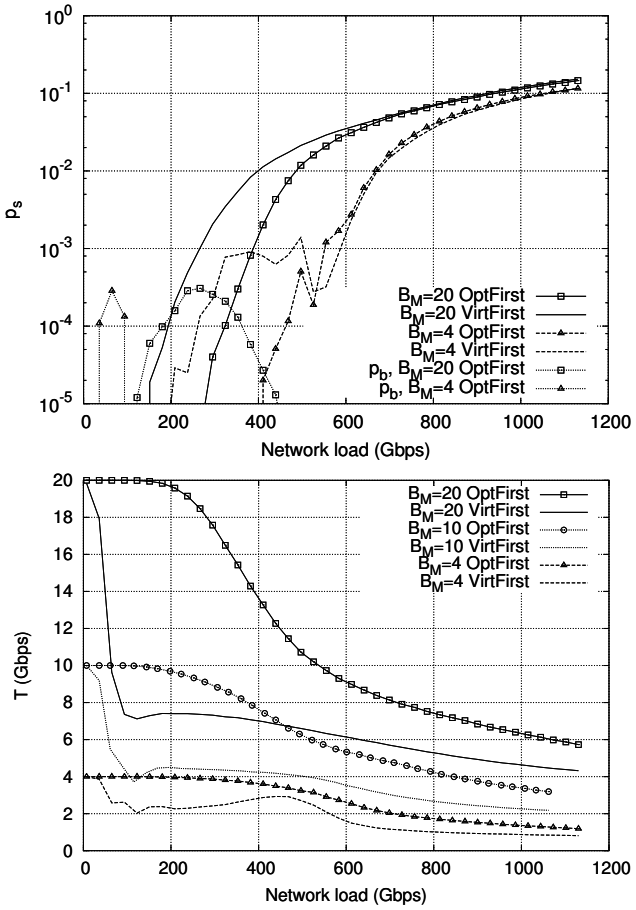


Fig. 7. Throughput T and starvation/blocking probability p_s/p_b as a function of the network load and for different flow granularity

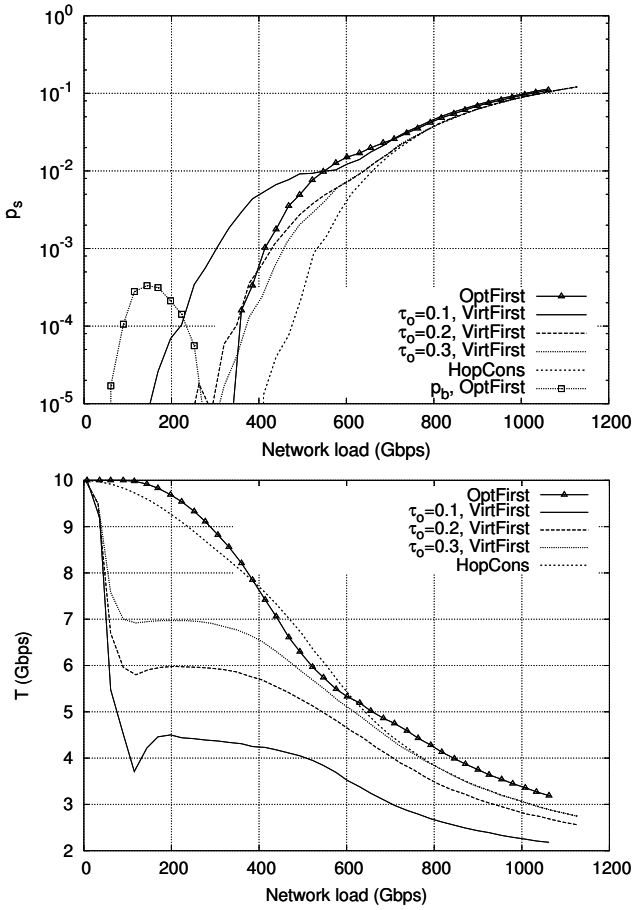


Fig. 8. Throughput T and starvation – blocking probability $p_s - p_b$ as a function of the network load and the threshold τ_o for *VirtFirst* grooming

The lower plot shows the throughput T and the upper plot shows the starvation p_s and the blocking p_b. The upper plot reports the curves only for $B_M = 20$ and 4 Gbit/s to avoid cluttering the graph. It is interesting that the absence of a pre-established logical topology does not allow *OptFirst* to ensure $p_b = 0$ at low loads. On the other hand *OptFirst* ensures a higher throughput, but its advantage tends to disappear as the flows granularity decreases. As we did in the 8-Ring topology the "network disconnection" phenomenon can be avoided by pre-establishing a fully connected topology of lightpaths. Notice, however, that in this general mesh topology the choice of the pre-established topology can be non-obvious. For instance a minimum spanning tree can create artificial bottlenecks, while the presence of pure OXCs makes other choices, like for instance rebuilding the physical topology at the logical level, far less obvious than in a regular topology like the ring.

The last set of results is related to the impact of the threshold τ_o. Fig. 8 presents the throughput and the starvation probability obtained with the *VirtFirst* grooming for $\tau_o = 0.1, 0.2, 0.3$. *VirtFirst* is most sensitive to this threshold, but we also report the curves relative to *OptFirst* and *HopCons* with $\tau_o = 0.3$. *OptFirst* is obviously insensitive to τ_o. Increasing τ_o improves *VirtFirst* throughput and starvation probability as expected, but the performance remains well below *OptFirst* and *HopCons*. Indeed, the effect of limiting the data-layer path with *HopCons* policy seems to be the dominant effect, confirming that this grooming policy has a user perceived quality similar to that provided by *OptFirst*, while limiting the cost for the network.

Fig. 9 reports the average number of hops N_h (upper plot) and the average number of links per lightpath N_{lo} (lower plot) in the same scenario. The impact of τ_o on these parameters is much smaller than on the average throughput. Finally note that *HopCons* performs much better than *VirtFirst* not only in term of throughput but also when considering the number of links per lightpath, at least for low/medium network loads, as shown in the upper plot of

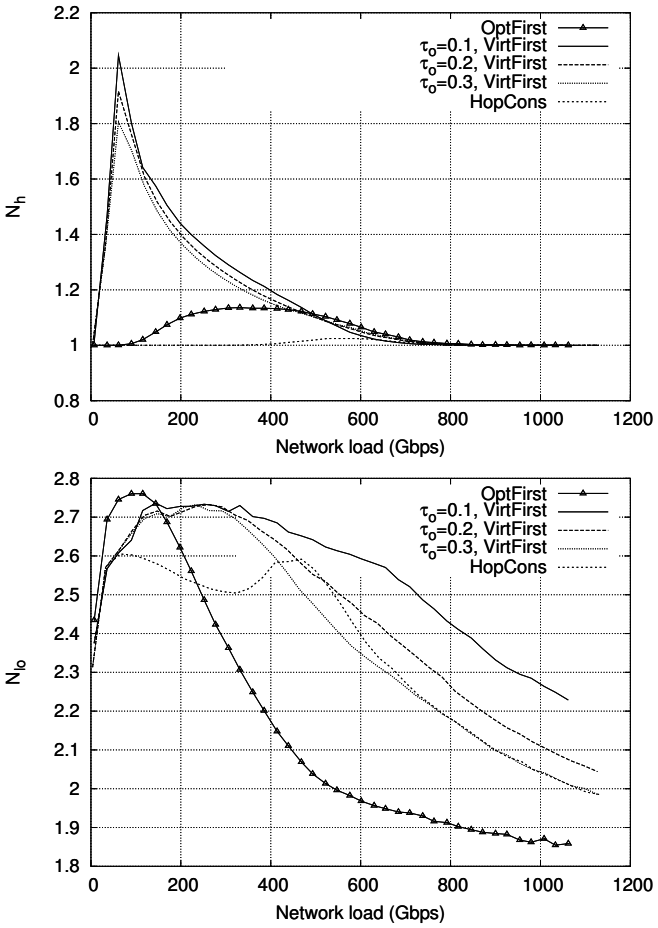


Fig. 9. Average number of hops N_h and average number of links per lightpath N_{lo} for different values of τ_o

Fig. 9. In some cases this grooming policy performs even better than *OptFirst*. The range of the average number of links per lightpath is very compressed as shown by the scale of the lower plot in Fig. 9. The detailed behavior is not easily explained, but the *OptFirst* shows a tendency to open shorter lightpaths at high loads.

V. Conclusion & Future Work

This paper has presented a comparative study of dynamic grooming algorithms in realistic scenarios with a data-based traffic model including elasticity, together with the simulation tool (GANCLES) used to perform it.

The first part of the paper is devoted to GANCLES presentation, highlighting its innovative features and the management of different architectural models of IPO networks through the explicit simulation of the optical and IP network levels.

The second part of the paper discusses performance results of different dynamic grooming algorithms on two topologies: a ring and a modification of NSFNet. First of all it is shown how the traffic model impacts on results. Then several performance indices, including the throughput of elastic flows, the probability that the service they receive falls below an acceptable threshold causing the flow starvation, and fairness are compared for the chosen grooming policies assuming an overlay IPO model.

The results show that the use of simplistic traffic model in the design of grooming algorithms may bring to misleading conclusions. Moreover, it is shown that the presence of a double network layer (optical and IP) does not alleviate traditional fairness problems associated with best-effort, elastic traffic. Some form of compensations are possible through the use of smart grooming policies; however, in an overlay model, where no information is shared between the optical and the IP level, it is not easy to find the appropriate and definitive solution.

Further studies on dynamic grooming enabled by GANCLES include comparison between different IPO architectures, studying what is the amount of information that needs to be exchanged to allow "intelligent" resource use. In addition, grooming strategies, policies, and algorithms can be implemented and studied in the simulator as we did for the *HopCons* policy that, although very simple, allows overcoming some of the shortcomes of *OptFirst* and *VirtFirst*. Finally, one major question is related to the use of QoS routing either in the optical or IP network layer in order to understand how "intelligent" routing strategies do interact one another through grooming policies.

References

[1] K.H. Liu, *IP Over WDM*, John Wiley, Nov. 2002

[2] A. Banerjee, J. Drake, J.P. Lang, B. Turner, K. Kompella, Y. Rekhter, *Generalized Multiprotocol Label Switching: an Overview of Routing and Management Enhancements*, IEEE Communications Magazine, 39(1):144–150, Jan. 2001.

[3] S. Tomic, B. Statovci-Halimi, A. Halimi, W. Muellner, J. Fruehwirth, *ASON and GMPLS: Overview and Comparison*, Photonic Network Communications, 7(2):111–130, March 2004.

[4] B. Wen, N.M. Bhide, R.K. Shenai, K.M. Sivalingam, *Optical Wavelength Division Multiplexing (WDM) Network Simulator (OWns): Architecture and Performance Studies*, SPIE Optical Networks Magazine Special Issue on "Simulation, CAD, and Measurement of Optical Networks",Sep/Oct. 2001.

[5] H. Zhu, H. Zang, K. Zhu, B. Mukherjee, *A novel generic graph model for traffic grooming in heterogeneous WDM mesh networks*, IEEE/ACM Transactions on Networking, 11(2):285–299, Apr. 2003.

[6] M. Kodialam, T.V. Lakshman, *Integrated Dynamic IP and Wavelength Routing in IP over WDM Networks*, In Proc. of INFOCOM 2001, pp. 358–366, Anchorage, AK, USA, Apr. 22–26 2001.

[7] X. Niu, W.D. Zhong, G. Shen, T.H. Cheng, *Connection Establishment of Label Switched Paths in IP/MPLS over Optical Networks*, Photonic Network Communications, 6:33–41, July 2003.

[8] R. Srinivasan, A.K. Somani, *Dynamic Routing in WDM Grooming Networks*, Photonic Network Communications, 5:123–135, Mar. 2003.

[9] S. Koo, G. Sahin, S. Subramaniam, *Dynamic LSP Provisioning in Overlay, Augmented, and Peer Architectures for IP/MPLS over WDM Networks*, In Proc. of INFOCOM 2004, Hong Kong, China, Mar. 7–11 2004.

[10] R. Dutta, G.N. Rouskas, *Traffic grooming in WDM networks: past and future*, IEEE Network Magazine, 16(6):46–56, Nov./Dec. 2002.

[11] X. Zhang, C. Qiao, *An Effective and Comprehensive Approach to Traffic Grooming and Wavelength Assignment in SONET/WDM Rings*, IEEE/ACM Transactions on Networking, 8(5):608–617, Oct. 2000.

[12] K. Zhu, B. Mukherjee, *Traffic grooming in an optical WDM mesh network*, IEEE Journal on Selected Areas in Communications, 20(1):122–133, Jan. 2003.

[13] R. Lo Cigno, E. Salvadori, Z. Zsóka, *Elastic Traffic Effects on WDM Dynamic Grooming Algorithms*, In Proc. IEEE Globecom 2004, Dallas, Texas, USA, Nov. 27 – Dec. 03, 2004.

[14] D. Bertsekas, R. Gallager, *Data Networks*, Prentice-Hall, 1987.

[15] B. Rajagopalan, J. Luciani, D. Awduche, *IP over Optical Networks: A Framework* IETF RFC 3717, Mar. 2004.

[16] GANCLES - Grooming cApable Network Call-Level Simulator. URL: http://netmob.unitn.it/gancles.html

[17] ANCLES - A Network Call-Level Simulator. URL: http://www.tlc-networks.polito.it/ancles

[18] C. Casetti, R. Lo Cigno, M. Mellia, M. Munafò, Z. Zsóka, *A Realistic Model to Evaluate Routing Algorithms in the Internet* In Proc. IEEE Globecom 2001, San Antonio, Texas, USA, pp. 1882–85, Nov. 25–29, 2001.

[19] ASONCLES - ASOn Network Call-Level Simulator. URL: http://www.hit.bme.hu/~zsoka/asoncles

[20] K. Pawlikowski, *Steady-state simulation of queueing processes: a survey of problems and solutions*, ACM Computer Surveys, 22(2):123-170, June 1990.

[21] C. Cavazzoni, V. Barosco, A. D'Alessandro, A. Manzalini et al. *The IP/MPLS over ASON/GMPLS test bed of the IST project LION*, IEEE/OSA Journal of Lightwave Technology, 21(11):2791–2803, Nov. 2003.

[22] M. Fawaz, B. Daheb, O. Audouin, M. Du-Pond, G. Pujolle, *Service level agreement and provisioning in optical networks*, IEEE Communications Magazine, 42(1):36–43, Jan. 2004.

[23] S.Sibal, A. DeSimone, *Controlling Alternate Routing in General-Mesh Packet Flow Networks*, in Proc. ACM SIGCOMM'94, London UK, August 1994.

[24] Q. Ma, P. Steenkiste, and H. Zhang, *Routing High-Bandwidth Traffic in Max-Min Fair Share Networks,* in Proc. of the ACM SIGCOMM'96, pp. 206–217, Stanford, CA, USA, Aug. 1996.

[25] Z. Wang, J. Crowcroft, *QoS Routing for Supporting multimedia applications,* IEEE JSAC, 14(7):1228–1234, Sept. 1996.

[26] E. Salvadori, Z. Zsóka, R. Lo Cigno, R. Battiti, *Dynamic Grooming in IP over Optical Networks based on the Overlay Architecture*, Submitted for publication, preprint available: http://dit.unitn.it/locigno/preprints/SaZsoLoBa04-3.pdf.

[27] T. Cinkler, C. Gáspár, *Fairness Issues of Routing with Grooming and Shared Protection*, In Proc. of IFIP ONDM 2004, Gent, Belgium, Feb. 2004.

A Framework for Dynamic Grooming in IPO Overlay Architectures*

E. Salvadori[1], Z. Zsóka[2], R. Lo Cigno[1], and R. Battiti[1]

[1] Dipartimento di Informatica e Telecomunicazioni – Università di Trento
[2] Department of Telecommunications, Budapest University of Technology and Economics

Corresponding Author: Renato Lo Cigno, DIT, Università di Trento Via Sommarive, 14 – 38050 Povo, Trento, Italy Tel: +39(0461)882026 – E-mail: locigno@dit.unitn.it

Abstract. This paper defines a formal framework for the definition of dynamic grooming policies in IP over Optical networks. The formal framework is then specialized for the Overlay Architecture, where the control plane of the IP and optical level are separated, and no information is shared between the two.
We define a family of grooming policies for the Overlay Architecture based on constraints on the number of hops and on the bandwidth sharing degree at the IP level, and we analyze the performance as a function of the grooming parameters in regular topologies.
Results are derived by using realistic traffic models that depart from the circuit-like traffic traditionally used in grooming studies, and fairness issues versus the flow physical length are also discussed.

Keywords: IP over Optical, Grooming, Overlay Architecture, Elastic traffic

1 Introduction

The IP protocol and optical transmission techniques are going to play a fundamental role in the next generation Internet. It is a widespread prediction that all other intermediate network management layers (ATM, SDH, SONET, ...) will gradually disappear, leaving a scenario where IP packets are carried directly on high speed WDM-based optical connections, the so-called IP over Optical (IPO) network. In this scenario, the interaction between routing and control of the circuit-switched optical network and that of the packet-switched IP network is of the utmost importance for the end-to-end performance and the efficient use of network resources.

Traffic grooming is the multiplexing capability aimed at optimizing the capacity utilization in transport systems by means of the combination of low-speed traffic streams onto high-speed (optical) channels. This problem is a variant of the well-known virtual topology design problem and has received a lot of attention in recent years (see [1] for a review).

There are two main approaches to study traffic grooming: *static* and *dynamic*. Static grooming refers to some network usage optimization when the traffic matrix is known

* This work is supported by the Italian Ministry of Education and Research (MIUR) through the GRID.IT and ADONIS projects, and by the Hungarian Italian Intergovernmental S&T Cooperation Programme I-17/03. The project is developed under the EU E-NEXT NoE.

in advance and it was proved to be an NP-hard problem. Dynamic grooming is a routing problem in a multi-layer network architecture, since the objective is to find the "best" path to route traffic requests arriving dynamically to grooming nodes. In this case equivalent requests arriving at different times may be treated differently because of different network conditions.

In IPO networks, IP routers are attached to an optical core network and connected to other peers via dynamically established lightpaths [2]. Considering the *control plane*, different architectures can be envisioned according to the amount of information exchanged between the IP and optical layer. RFC 3717 defines three interconnection models: overlay, augmented and peer. In the *peer* model, the topology and other network information are completely shared in a unified control plane, while in the *overlay* model, each layer performs its own routing functions because no information is exchanged between them. An intermediate architecture is the *augmented* model, where some aggregated information from one routing instance is passed to the other.

The peer and the augmented models are appealing because they allow running an integrated routing function, by using, for instance, an auxiliary graph, as done in [3]. However, both models seem not feasible in the near term due to the tight integration between the two levels and scalability issues regarding the amount of exchanged information. The overlay model is instead technically feasible, since it only requires the definition of an interface between the IP and optical level and dynamic lightpath capabilities in the optical level, which are being experimented in laboratories and research projects [4]. Surprisingly, most of the dynamic grooming algorithm proposed in the literature implicitly consider such models [3, 5–7], while only a few dynamic grooming algorithms based on the overlay model have been proposed so far [8, 9]. These papers explore the two extreme policies of privileging always the optical level exploitation or the other way around. In [7] the authors propose new algorithms that improve performance in the overlay model with respect to [8] and in the augmented model with respect to a grooming algorithm in peer models proposed in [5].

None of the works on grooming in IPO networks adopted a realistic traffic model. The traffic loading the network is always composed of CBR (Constant Bit Rate) connections characterized by the bit rate and duration. Any realistic evaluation of algorithms to be deployed within the Internet, should instead capture at least the basic characteristics of Internet traffic. From the routing point of view, the most important features of present Internet applications are the capacity to adapt the rate to changing network conditions (elasticity) and the need to transfer a given amount of data (compared to the duration of, for instance, a conversational application). The holding time of a flow becomes a *consequence* of the network conditions and not a property of the flow. In a previous contribution [10] we discussed in detail the impact of realistic traffic models on dynamic grooming, showing the inherent interaction between the IP and the optical layer and its effect on the overall performance. In this paper we use only the realistic model we called *data-based* in [10]. Traffic flows in this models share the resources on a virtual topology path following the max-min fairness criterion [11], thus mimicking the ideal behavior of a bundle of TCP connections.

This paper contributes to the field of dynamic grooming policies in at least four distinct ways. I) A formal definition of dynamic grooming based on graph theory is defined in

a general interconnection model and specialized to the case of the overlay model; II) A family of grooming policies is proposed in the overlay architecture, simple existing proposals are assessed as a special case of it; III) Performance and tradeoffs of different policies are discussed and explained; IV) Unfairness issues inherent to dynamic grooming and arising from different physical distance between flow end points are discussed and hints on the problem solution are given.

2 Problem Formulation

2.1 A formalism for dynamic grooming

IPO networks are based on two layers: the *optical*- and the *data-layer*. The optical-layer is based on OXCs interconnected by fiber links. G-OXCs, which support sub-wavelength traffic multiplexing onto wavelength channels, are the bridge between the optical-layer and data-layer. A G-OXC is *also* an IP router, hence transit traffic (not terminated in the router), can be groomed with incoming traffic. The data-layer consists of routers interconnected with a virtual topology made of all the lightpaths which have been set up in the optical-layer.

It is therefore necessary to define a topological graph for each of the layers:

- $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ is the physical topology, where $\mathcal{N}$ is the set of vertices v_j (OXCs) and $\mathcal{E}$ is the set of edges e_{jk} (fiber links) connecting vertices v_j and v_k, without loss of generality we compact multiple fibers into a single edge;
- $\mathcal{G}^\nu = (\mathcal{N}^\nu, \mathcal{E}^\nu)$ is the virtual topology, where $\mathcal{N}^\nu$ is the set of vertices v_j^ν (IP routers), $\mathcal{N}^\nu \subseteq \mathcal{N}$, and $\mathcal{E}^\nu$ is the set of edges $e_{jk,q}^\nu$ (virtual links) connecting vertices v_j^ν and v_k^ν. Each edge in $\mathcal{E}^\nu$ corresponds to a lightpath in the optical-layer. Because there can be more than one lightpath between any two nodes an additional identifier q is required for uniqueness[3].

In the rest of the paper, the superscript ν is used to specify vertices, edges and paths belonging to the virtual topology.

Each edge e_{jk} in $\mathcal{G}$ is assigned a vector of properties $\bar{w}_{jk}$ describing any static or dynamic (possibly vectorial) metrics pertaining to physical or traffic-related characteristics of the link. Similarly, a property vector $\bar{w}_{jk,q}^\nu$ is assigned to each edge of $\mathcal{G}^\nu$.

In the optical-layer, a path $\pi_p(v_j, v_k)$, or simply π_p, of length n is defined as a sequence of n distinct edges e_{ih} joining v_j and v_k where $v_j, v_k \in \mathcal{N}$, $e_{ih} \in \mathcal{E}$, $\pi_p(v_j, v_k) = \{e_{ji}, e_{ih}, ..., e_{zk}\}$. The value of p is unique in $\mathcal{G}$ and identifies explicitly the path. This identifier is required since several parallel paths may exist between the nodes v_j and v_k. The path π_p is a *lightpath*, and it corresponds to a specific wavelength if no wavelength conversion is considered.

Let $\models$ be the operator that maps a lightpath in the physical topology onto an edge of the virtual topology: $e_{jk,q}^\nu \models \pi_p$ if the path π_p joins the two vertices $v_j^\nu, v_k^\nu \in \mathcal{N}^\nu$. In the data-layer, a path $\pi_t^\nu = \pi_t^\nu(v_s^\nu, v_d^\nu)$ is a sequence of n distinct edges $e_{ih,q}^\nu \in \mathcal{E}^\nu$, t is a unique identifier to distinguish multiple parallel paths.

[3] The virtual topology varies in time as lightpaths are set up and torn down. Notice also that $\mathcal{G}^\nu$ has nothing to do with the *auxiliary graph* defined in [3], which is an abstract representation of both levels assuming complete sharing of the two control planes.

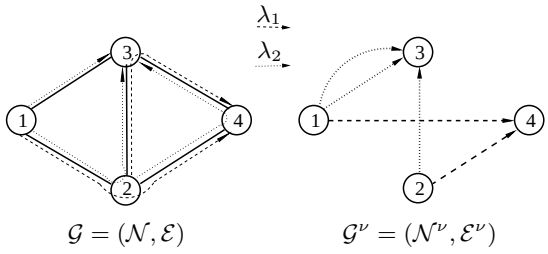


Fig. 1. Physical and virtual topology for a small 4 nodes network

Fig. 1 presents an example. The maximum number of wavelengths per link is $W = 2$. Let's assume that the following five paths have been set up in the optical-layer, where the superscript (i) is used here to specify the corresponding wavelength[4]:

$$\pi_1^{(1)} = \{e_{12}, e_{24}\},\ \pi_2^{(1)} = \{e_{23}, e_{34}\},$$
$$\pi_3^{(2)} = \{e_{12}, e_{23}\},\ \pi_4^{(2)} = \{e_{24}, e_{34}\},\ \pi_5^{(2)} = \{e_{13}\}.$$

The corresponding set of edges in the data-layer is: $\mathcal{E}^\nu = \{e^\nu_{14,1}, e^\nu_{24,2}, e^\nu_{13,3}, e^\nu_{23,4}, e^\nu_{13,5}\}$, where

$$e^\nu_{14,1} \models \pi_1^{(1)},\ e^\nu_{24,2} \models \pi_2^{(1)},\ e^\nu_{13,3} \models \pi_3^{(2)},\ e^\nu_{23,4} \models \pi_4^{(2)},\ e^\nu_{13,5} \models \pi_5^{(2)}.$$

The identifiers q in edges and p in paths are conceptually different as they are defined in different logical planes, but, as seen in the example, they can assume the same value for simplicity.

For each pair v_j, v_k, the set $\mathcal{P}_\mathcal{G}^{jk}$ is defined as the set of all paths existing between v_j and v_k: $\mathcal{P}_\mathcal{G}^{jk} = \{\pi_p(v_j, v_k) \mid v_j, v_k \in \mathcal{N}, j \neq k\}$. Similarly, a set $\mathcal{P}_{\mathcal{G}^\nu}^{sd}$ is defined as the set of all paths existing between v_s^ν and v_d^ν: $\mathcal{P}_{\mathcal{G}^\nu}^{sd} = \{\pi_t^\nu(v_s^\nu, v_d^\nu) \mid v_s^\nu, v_d^\nu \in \mathcal{N}^\nu, s \neq d\}$.

Given some Routing and Wavelength Assignment (RWA) algorithm $\boldsymbol{\Lambda}$ in use in the optical-layer, a cost $c^{\boldsymbol{\Lambda}}(\pi_p)$ is assigned to a path π_p by using a combination of the properties $\bar{w}_{ih}$ of its links $e_{ih} \in \mathcal{E}$. Given $\mathcal{P}_\mathcal{G}^{jk}$ the algorithm $\boldsymbol{\Lambda}$ selects the minimum cost path available in the set: $\dot{\pi}_{jk} = \boldsymbol{\Lambda}(\mathcal{P}_\mathcal{G}^{jk})$ that we assume to be unique, possibly by breaking ties with random choices. If no path is available $\boldsymbol{\Lambda}$ returns $\emptyset$ identifying the empty path.

Let us illustrate the path cost assignment and routing mechanism by using the adaptive routing FPLC (Fixed-Paths Least-Congestion) presented in [12]. A contiguous wavelength is one concurrently available on each link of the path. The cost $c^{\boldsymbol{\Lambda}}(\pi_p)$ is calculated as the number of contiguous wavelengths on the path, breaking ties with the number of hops in the path:

$$c^{\mathrm{FPLC}}(\pi_p) = u(\pi_p) + \sum_{e_{ih} \in \pi_p} \frac{1}{|\mathcal{N}| + 1} \tag{1}$$

[4] This notation is used here only to help the reader, and it won't be used anymore in the following, since this information can be embedded in the path characteristics.

where $u(\pi_p)$ is a function counting the contiguous wavelengths currently available on the whole path π_p and $|\mathcal{N}|$ is the number of vertices in $\mathcal{G}$ and it provides a smaller-than-unit weight to break ties based on the number of hops.

In the virtual topology $\mathcal{G}^\nu$ we can operate in the same way. Given an IP-based routing algorithm $\boldsymbol{\Omega}$, the minimum cost path between any two nodes is selected as $\dot{\pi}^\nu_{sd} = \boldsymbol{\Omega}(\mathcal{P}^{sd}_{\mathcal{G}^\nu})$.

As an example, the cost function used for the *Minimum Distance* (MD) routing presented in [13] is:

$$c^{\mathrm{MD}}(\pi^\nu_t) = \sum_{e^\nu_{ih,q} \in \pi^\nu_t} \frac{1}{B(e^\nu_{ih,q})} \tag{2}$$

where $B(e^\nu_{ih,q})$ is the bandwidth that is available on the lightpath $e^\nu_{ih,q}$ for the request.

$\mathcal{P}^{sd}_{\mathcal{G}^\nu}$ can include a set $\mathcal{P}_{pre}$ of pre-established lightpaths in $\mathcal{G}$. $\mathcal{P}_{pre}$ defines a set of permanent lightpaths which guarantee the connectivity of the virtual topology regardless of the traffic pattern and lightpath establishment dynamics. The corresponding set of edges in the virtual topology is called $\mathcal{E}^\nu_{pre}$. Although not necessary for defining grooming policies, the introduction of lightpaths that cannot be teared down can help to avoid the risk that an aggressive use of optical resources builds a virtual topology that is not completely connected. We observed this behavior in a previous work [10] and we deem that it was never reported before in the literature because studies that use circuit-like traffic models cannot easily distinguish between the blocking of a connection for the lack of resources and its blocking because of a badly defined virtual topology.

A generic *grooming policy* is a procedure

$$\boldsymbol{G}\left(\boldsymbol{\Lambda}(\mathcal{P}^{jk}_{\mathcal{G}}), \boldsymbol{\Omega}(\mathcal{P}^{sd}_{\mathcal{G}^\nu}), \Delta\right) \tag{3}$$

where Δ is a set of criteria defining the interaction between the optical and IP level, determining the collaboration between $\boldsymbol{\Lambda}$ and $\boldsymbol{\Omega}$. For instance Δ defines v_j and v_k for the lightpath setup, which can be different from the source-destination pair v^ν_s,v^ν_d of the incoming flow f_{sd}. If present, admission control criteria can be integrated in Δ, which is normally expanded as a set of **if-then-else** clauses.

The set of criteria Δ is influenced by the integration level of the IP and optical control planes (Overlay, Peer or Augmented models). The information base (e.g., status of the optical resources, or cost of the different IP level connections) used by Δ depends on the control planes integration. For example, in the overlay model, reasonable assumptions are $j = s$, $k = d$, and a single lightpath established for each request f_{sd}.

Releasing a lightpath between v_j and v_k means recomputing the set $\mathcal{P}^{jk}_{\mathcal{G}}$ in the optical-layer and consequently deleting an edge e^ν_{jk} from the virtual topology and recomputing all the sets $\mathcal{P}^{sd}_{\mathcal{G}^\nu}$ that included e^ν_{jk}.

2.2 Detailing G for overlay architectures

In IPO networks based on the overlay architecture the control planes of the optical and IP levels are separated. Each time an incoming request f_{sd} needs to be routed, there are only two possible options: (i) route it over the current virtual topology $\mathcal{G}_\nu$ invoking $\boldsymbol{\Omega}$

or (ii) set-up a new lightpath $e^{\nu}_{sd,q} \models \dot{\pi}_{sd}$ invoking $\boldsymbol{\Lambda}$ and route the request over the new virtual topology $\mathcal{G}_\nu \bigcup e^{\nu}_{sd,q}$, invoking $\boldsymbol{\Omega}$ in a second phase.

Fig. 2 (left part) specifies the procedure (3) for an overlay IPO network, without detailing the criteria set Δ. Notice that policies privileging the use of already established lightpaths can always resort to invoke $\boldsymbol{\Lambda}$ either because no $\dot{\pi}^{\nu}_{sd}$ was found or because the result of $\boldsymbol{\Omega}$ is refused for any reason.

3 Grooming Policies

In the scenario depicted above, independently from the definition of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Omega}$, the set of rules Δ must define how and when to invoke $\boldsymbol{\Lambda}$. Although many criteria can be envisaged, we propose a simple rule based on the number of IP hops between s and d. In other words, Δ defines as rule the invocation of $\boldsymbol{\Lambda}$ only if the path selected by $\boldsymbol{\Omega}$ has more than K hops. We call this policy *Hop Constrained Grooming* $\boldsymbol{HC}(\cdot)$. Additionally, $\boldsymbol{HC}$ can include rules for refusing a logical path π^{ν}_{t} based on congestion measures. With a realistic elastic model of Internet traffic, the definition of congestion is not trivial, since it cannot refer directly to the amount of resources requested by flows. Bandwidth overbooking is a normal practice and we assume that a new lightpath is opened when accepting the flow f_{sd} on the virtual topology would result in assigning it a bandwidth smaller than some given amount applying max-min sharing. The dynamic grooming policies studied in [9] and named "Optical-layer-first" and "IP/MPLS-layer-first" are simply the extreme cases for $K = 0$ and $K = \infty$, and have been studied in the simpler case of bandwidth guaranteed traffic.

In order to fix ideas, let's assume that flow requests arrive to the network with two attributes: a peak transmission rate B_M and a minimum requested rate expressed as a fraction th_s of B_M. If at any time the bandwidth assigned to flow f_{sd} falls below $th_s(f_{sd}) \cdot B_M(f_{sd})$ then the flow will close and counted as a "starved" flow, because the network was not able to guarantee its correct completion. B_M and th_s can be included in some SLA (Service Level Agreement) at the IP/Optical interface. We thus define a *starvation probability* and not a blocking probability, since the adaptive and

0. request f_{sd} arrives
1. **if** *exploit optical resources*
2. $\quad e^{\nu}_{sd,q} \models \dot{\pi}_{sd} = \boldsymbol{\Lambda}(\mathcal{P}^{sd}_{\mathcal{G}})$
3. $\quad \dot{\pi}^{\nu}_{sd} = \boldsymbol{\Omega}(\mathcal{P}^{sd}_{\mathcal{G}^{\nu} \bigcup e^{\nu}_{sd,q}})$
4. **else** *exploit virtual resources*
5. $\quad \dot{\pi}^{\nu}_{sd} = \boldsymbol{\Omega}(\mathcal{P}^{sd}_{\mathcal{G}^{\nu}})$
6. $\quad$ **if** $\dot{\pi}^{\nu}_{sd} = \emptyset$ **or** $\dot{\pi}^{\nu}_{sd}$ *is refused*
7. $\quad\quad e^{\nu}_{sd,q} \models \dot{\pi}_{sd} = \boldsymbol{\Lambda}(\mathcal{P}^{sd}_{\mathcal{G}})$
8. $\quad\quad \dot{\pi}^{\nu}_{sd} = \boldsymbol{\Omega}(\mathcal{P}^{sd}_{\mathcal{G}^{\nu} \bigcup e^{\nu}_{sd,q}})$

procedure $\boldsymbol{HC}(\boldsymbol{\Lambda}, \boldsymbol{\Omega}, \delta(K, \tau_o), f_{sd})$

1. $\dot{\pi}^{\nu}_{sd} = \boldsymbol{\Omega}(\mathcal{P}^{sd}_{\mathcal{G}^{\nu}})$
2. **if** $\{\dot{\pi}^{\nu}_{sd} = \emptyset\}$**or**$\{b(\dot{\pi}^{\nu}_{sd}) < \tau_o\}$
   $\quad$ **or**$\{H(\dot{\pi}^{\nu}_{sd}) > K\}$
3. $\quad e^{\nu}_{sd,q} \models \dot{\pi}_{sd} = \boldsymbol{\Lambda}(\mathcal{P}^{sd}_{\mathcal{G}})$
4. $\quad \dot{\pi}^{\nu}_{sd} = \boldsymbol{\Omega}(\mathcal{P}^{sd}_{\mathcal{G}^{\nu} \bigcup e^{\nu}_{sd,q}})$
5. **return** $\dot{\pi}^{\nu}_{sd}$

Fig. 2. General definition of dynamic grooming policies in overlay IPO networks (left); the grooming procedure $\boldsymbol{HC}$ (right).

elastic nature of Internet traffic does not allow the easy definition of strict admission procedures.

Besides choosing an appropriate K, a network operator may choose to open a new lightpath when routing f_{sd} in virtual topology means it will receive from the beginning a bandwidth smaller than τ_o, possibly a function of f_{sd}. We call τ_o *optical opening threshold*. Clearly, $\tau_o \geq th_s \cdot B_M$.

The above criteria specify a Δ with two parameters, say $\delta(K, \tau_o)$, defining the **if-then-else** rules of the generic grooming policy in the left hand part of Fig. 2. This leads to the implementable grooming procedure described in the right hand part of Fig. 2, and compactly written as $\boldsymbol{HC}(\boldsymbol{\Lambda}, \boldsymbol{\Omega}, \delta(K, \tau_o), f_{sd})$.

With reference to Fig. 2 (right), $H(\pi_t^\nu)$ returns the number of hops in virtual topology path π_t^ν and $b(\pi_t^\nu) = \min_{e_{ih,q}^\nu \in \pi_t^\nu} B(e_{ih,q}^\nu)$.

4 Results and discussion

A theoretic performance analysis is not feasible due to complexity of the system, thus we resort to simulations for the performance evaluation.

The implementation of grooming policies in a packet level simulator such as *ns-2* is not convenient for efficiency reasons. Starting from existing tools in our research groups, we have developed a simulator capable of handling the layered topological structure of IPO networks as well as several different $\boldsymbol{\Lambda}$ and $\boldsymbol{\Omega}$ functions. The description of the tool, named GANCLES, goes beyond the scope of this paper, and we refer the interested reader to GANCLES web site [14].

Among possible performance indices we have selected the following ones.

P_s — Starvation probability. It is the probability that a flow closes during its life because it is not receiving service with acceptable quality. A flow f_{sd} closes and drops the network if its instantaneous throughput falls below $th_s \cdot B_M$.

T — Average normalized throughput of completed flows.

$$T = \frac{\sum_{v_s^\nu \in \mathcal{G}^\nu} \sum_{v_d^\nu \in \mathcal{G}^\nu} \sum_{\forall f_{sd}} T(f_{sd})}{\sum_{v_s^\nu \in \mathcal{G}^\nu} \sum_{v_d^\nu \in \mathcal{G}^\nu} \sum_{\forall f_{sd}} 1}$$

$T(f_{sd})$ is the average normalized throughput of flow f_{sd} provided it was not starved. Notice that in a resource sharing environment T is not the average resource occupation divided by the number of flows.

U_d — Distance unfairness index.

$$U_d = \frac{\max_{0<r<|\mathcal{N}|} T^r - \min_{0<r<|\mathcal{N}|} T^r}{T}$$

measures if the resource assignment is fair with respect to physical distance. T^r is the average throughput calculated for node-pairs with hop distance r in the physical topology. It ranges from zero to ∞; any value larger than 1 indicates unacceptable unfairness.

The goal of a grooming algorithm is maximizing T while minimizing P_s and U_d.

4.1 Networking Scenarios

We use two different regular topologies, where all nodes have grooming capabilities:

- *R8* is an 8 nodes bidirectional ring;
- *MT16* is a 16 nodes mesh-torus network with connectivity four, i.e., each node is connected to the four adjacent nodes in a regular, closed lattice mapped on the surface of a bi-dimensional torus — or a doughnut in pop terms.

The number of wavelength W is one of the parameters of major interest to investigate the generality and scalability of solutions with respect to the amount of available resources. Each wavelength has a capacity $g = 20\,\text{Gbit/s}$. A data-layer traffic source is connected to each G-OXC, generating requests with $B_M = 10\,\text{Gbit/s}$ following a Poisson process. Each flow transfer data whose amount is randomly chosen from an exponential distribution with average 12.5 GBytes; $th_s = 0.1$ in all simulations and $\tau_o = 3\,\text{Gbit/s}$. Dynamically opened lightpaths are immediately torn down if they are not used. All simulations are run until performance indices reach a 99% confidence level over a $\pm 1\%$ confidence interval around the point estimate. Estimations are carried out with the *batch means* technique. Results are plotted versus the total load L offered to the network, and also versus the relative traffic ρ offered to each network node, normalized to the total capacity of its egress links. Given the total number of nodes $|\mathcal{N}|$ in the network, the connectivity degree D, the number of wavelengths per fiber W and their data-rate g, we have: $\rho = \dfrac{L}{|\mathcal{N}|DWg}$.

Unless otherwise stated, $\boldsymbol{\Lambda}$ is the FPLC algorithm described in Sec. 2.1 with first-fit wavelength assignment and $\boldsymbol{\Omega}$ is the standard fixed shortest path (FSP) algorithm. A uniform traffic pattern is simulated, i.e., when a new flow request is generated, the source and destination are randomly chosen with the same probability.

The virtual topology connectivity is guaranteed with pre-established lightpaths that are never closed. We populate $\mathcal{P}_{pre}$ with a 'Physical-Topology' of lightpaths, which sets up the lowest order wavelength among each pair of adjacent nodes.

4.2 Grooming policies behavior on different topologies

The first set of results shows how the policy $\boldsymbol{HC}$ behaves when varying K and W, considering the impact of W on *R8*. Fig. 3 and Fig. 4 show the impact of using different values of K on T and P_s for $W = 4$ and $W = 8$ respectively. On the bottom x-axis we use L and on the top x-axis we use ρ.

In both cases $K = 1$ ensures the best performance. The performance spread increases with W; results for $W = 12$ confirm this result. This behavior comes from the aggressive use of optical-layer resources with $K = 0$. Setting up lightpaths even when not needed, the optical-layer becomes overcrowded with lightpaths, which leads to blocking lightpath set-up requests when congestion is impending, resulting in a poorly connected virtual topology, reduced throughput T and increased starvation P_s. Increasing K above 1, the performance tends to be similar to $K = \infty$. This is due to the small average distance between nodes, but it also confirms that the best way to use optical resources is trying to build a fully connected mesh in the virtual topology. Although

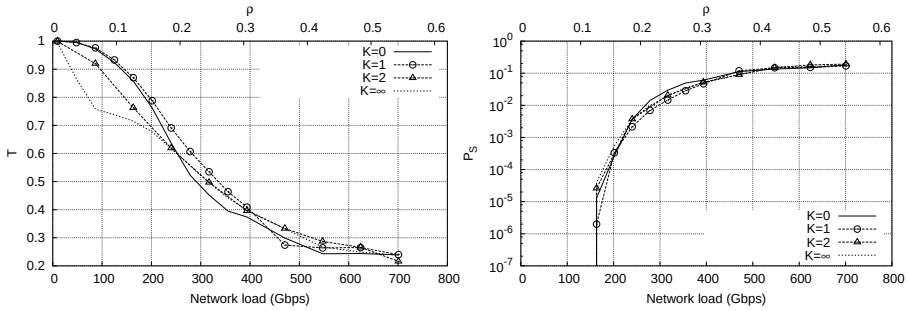


Fig. 3. Per-flow average normalized throughput T (left plot) and starvation probability P_s (right plot) for *R8* with $W = 4$

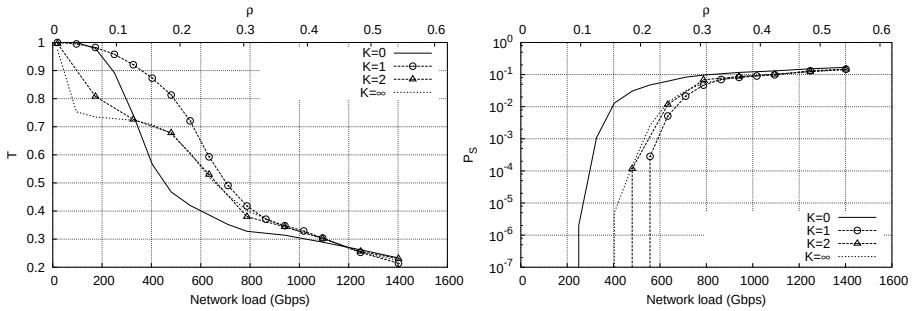


Fig. 4. Per-flow average normalized throughput T (left plot) and starvation probability P_s (right plot) for *R8* with $W = 8$

difficult to prove formally, we have observed that by using $K = 1$ the grooming algorithm tends to build a full-mesh virtual topology when W provides enough resources, and this independently from the physical topology. In our opinion, this is the main reason why setting $K = 1$ guarantees better performance.

Fig. 5 shows T and P_s for *MT16* with $W = 8$. The behavior is similar to the one observed in *R8* confirming that the relative merit of grooming policies is not related to the topology.

Restricting the analysis to $K = 0, 1$, we analyze the effect of W on performance as a function of the normalized load ρ. Fig. 6 compare the behavior on *R8* with $W = 4, 8, 12$. The figure clearly shows that increasing W the choice of $K = 1$ is indeed the best one, keeping the efficiency almost constant as the amount of resources and the traffic increases, while for $K = 0$ the performance decreases drastically.

Fig. 7 presents a similar set of results for *MT16* and $W = 2, 4, 8$. These results confirm the conclusions above. Most interesting is that for low W the performances are similar and in both topologies tend to diverge as W increase.

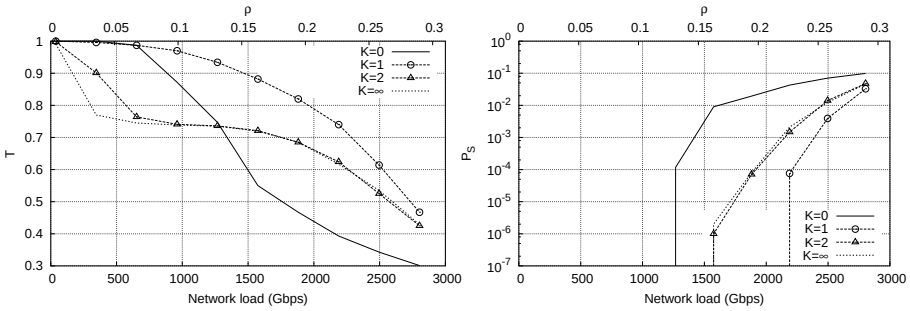


Fig. 5. Per-flow average normalized throughput T (left plot) and starvation probability P_s (right plot) for *MT16* with $W = 8$

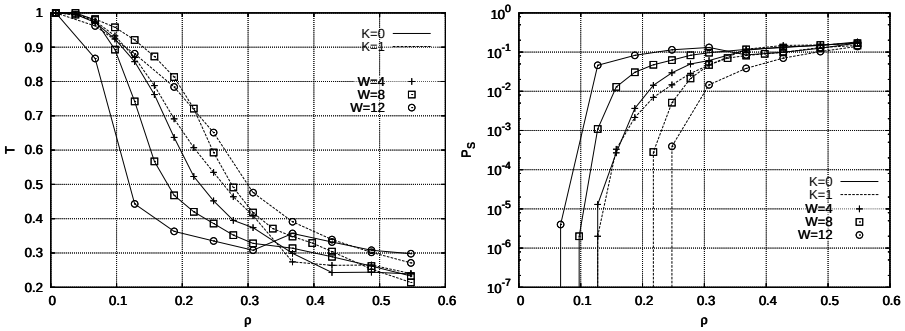


Fig. 6. Per-flow average normalized throughput T (left plot) and starvation probability P_s (right plot) for *R8* with varying W and $K = 0, 1$

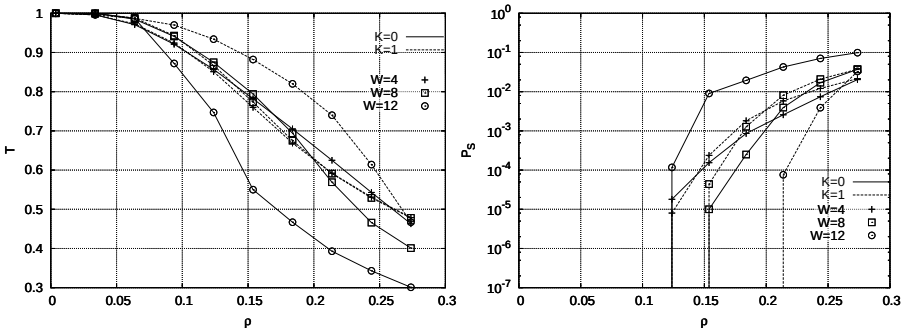


Fig. 7. Per-flow average normalized throughput T (left plot) and starvation probability P_s (right plot) for *MT16* with varying W and $K = 0, 1$

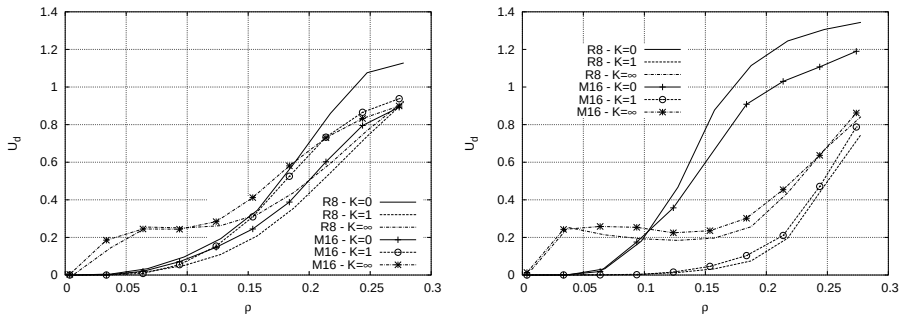


Fig. 8. Fairness comparison between *R8* and *MT16* with $W = 4$ (left plot) and $W = 8$ (right plot)

4.3 Fairness issues with elastic traffic

One of the major drawbacks of dynamic network management, as it is well known from Internet experience, is the unfair behavior of the network toward longer connections (both in terms of physical distance and in terms of logical hops). In [15] the problem is addressed in the context of peer IPO networks (for the first time to the best of our knowledge) with a circuit-like traffic model. Resource sharing in general exacerbate the unfairness, because longer connections compete with more connections with respect to short ones.

An interesting question is whether the physical topology and available number of wavelengths impacts on the problem. Fig. 8 reports results for $W = 4$ (left plot) and $W = 8$ (right plot) for *R8* and *MT16*. The hop constraint K is 0, 1, ∞.

On the one hand, it is clear that none of the parameter setting can guarantee perfect fairness, and that the physical topology does not have a major impact. On the other hand, increasing W does help in keeping a certain degree of fairness, specially if the grooming policy tends to build a logical topology that approaches a full mesh ($K = 1$). Instead, the very aggressive use of optical resources operated with $K = 0$ gobbles resources to build unnecessary parallel paths between adjacent node pairs, exacerbating the unfairness at high loads.

5 Conclusions

This paper introduces a formal description of dynamic grooming policies, defining the limits between grooming in overlay architectures and grooming in peer or augmented architectures, where there is total or partial integration of the optical and IP control planes.

A family of grooming policies for the overlay model, based on constraints on the number of hops and bandwidth available at the virtual topology level, is defined and analyzed, discussing parameters setting and the impact of the number of available wavelengths per fiber on the grooming policy. Several grooming policies previously presented in literature are particular cases of the family we defined.

Results analysis proves that it is possible to define grooming parameters that lead to good performance regardless of the topology and that allow good scaling with the amount of optical resources. The inspection of the virtual topology that are build by the grooming policies hints to the fact that a good policy should try to build a full mesh at the virtual topology level to keep the balance between the traffic pattern (uniform) and the virtual topology. This observation may pave the road for the definition of more performing strategies that the ones we analyzed, which can also adapt to asymmetric and time varying traffic.

Fairness issues in dynamic grooming were also discussed, proposing guidelines to solve them.

References

1. R. Dutta, G.N. Rouskas, "Traffic grooming in WDM networks: past and future," *IEEE Network Magazine*, 16(6):46–56, Nov./Dec. 2002.
2. B. Rajagopalan, J. Luciani, D. Awduche, "IP over Optical Networks: A Framework," RFC 3717, IETF, Mar. 2004.
3. H. Zhu, H. Zang, K. Zhu, B. Mukherjee, "A novel generic graph model for traffic grooming in heterogeneous WDM mesh networks," *IEEE/ACM TON*, 11(2):285–299, Apr. 2003.
4. C. Cavazzoni, et al., "The IP/MPLS over ASON/GMPLS test bed of the IST project LION," *IEEE/OSA Jou. of Lightwave Technology*, 21(11):2791–2803, Nov. 2003.
5. M. Kodialam, T.V. Lakshman, "Integrated Dynamic IP and Wavelength Routing in IP over WDM Networks," *In Proc. of INFOCOM 2001*, pp. 358–366, Anchorage, AK, Apr. 2001.
6. R. Srinivasan, A.K. Somani, "Dynamic Routing in WDM Grooming Networks," *Photonic Network Communications*, 5:123–135, Mar. 2003.
7. S. Koo, G. Sahin, S. Subramaniam, "Dynamic LSP Provisioning in Overlay, Augmented, and Peer Architectures for IP/MPLS over WDM Networks," *In Proc. of INFOCOM 2004*, Hong Kong, China, Mar. 2004.
8. C. Assi, A. Shami, M.A. Ali, Y. Yue, S. Dixit, "Integrated Routing Algorithms for Provisioning "Sub-Wavelength" Connections in IP-Over-WDM Networks," *Photonic Network Communications*, 4(34):377–389, Jul./Dec. 2002.
9. X. Niu, W.D. Zhong et al. "Connection Establishment of Label Switched Paths in IP/MPLS over Optical Networks," *Photonic Network Communications*, 6(1):33–41, Jul. 2003.
10. R. Lo Cigno, E. Salvadori, Z. Zsóka, "Elastic Traffic Effects on WDM Dynamic Grooming Algorithms," *In Proc. of GLOBECOM 2004*, Dallas, TX, USA, Dec. 2004
11. D. Bertsekas, R. Gallager, *Data Networks*, Prentice-Hall, 1987.
12. Ling Li, A. K. Somani, "Dynamic wavelength routing using congestion and neighborhood information," *IEEE/ACM TON*, 7(5):779-786, Oct. 1999
13. Q. Ma, P. Steenkiste, H. Zhang, "Routing high-bandwidth traffic in max-min fair share networks," *In Proc. of ACM SIGCOMM 1996*, pp. 206–217, Palo Alto, CA, USA, Aug. 1996.
14. GANCLES - Grooming cApable Network Call-Level Simulator http://netmob.unitn.it/tools/gancles.html.
15. T. Cinkler, C. Gáspár, "Fairness Issues of Routing with Grooming and Shared Protection," *In Proc. of IFIP ONDM 2004*, Gent, Belgium, Feb. 2004.