UNIVERSITY OF MODENA AND REGGIO EMILIA

Department of Engineering "Enzo Ferrari"

Doctorate School in Information and Communication Technologies (ICT)

Cycle XXVI

PhD Dissertation

Architectures for Energy Efficient Optical Networks

Candidate: **MATTEO FIORANI**

Advisor: **Prof. MAURIZIO CASONI**

The Coordinator of the Doctorate: **Prof. GIORGIO M. VITETTA**

The Director of the School: **Prof. GIORGIO M. VITETTA**

Architectures for Energy Efficient Optical Networks

Matteo Fiorani

Abstract

Although the Information and Communications Technologies (ICT) sector can play a fundamental role in enabling a low carbon economy, the energy and carbon impact of the sector itself is significant and is expected to grow rapidly with the proliferation of devices connected to the Internet and emergence of new services. The energy consumption of the ICT sector can be divided in: (*i*) energy consumed by the user devices, (*ii*) energy consumed by the telecommunication network infrastructure, and (*iii*) energy consumed by the data centers. Even if end-user devices (such as computers, printers, etc.) are the major contributors, the sum of the energy consumed by the telecommunication networks and data centers is today estimated to be more than a half of the total ICT energy consumption. With the expected growth in the Internet and data center traffic the energy consumption of telecommunication networks and data centers is destined to drastically increase if the energy efficiency would not be improved. In this work energy efficient network architectures for telecommunication networks and data centers are proposed and analyzed.

Telecommunication networks can be divided into three domains: (*i*) access, (*ii*) metro, and (*iii*) core. Core networks are the central part of the network hierarchy, providing nationwide or global coverage, and are based on optical transmission technologies. They contribute significantly to the energy consumption of the telecommunication infrastructure since they must support very high capacities. To increase the energy-efficiency in optical core networks, in the first part of the thesis a novel core network paradigm, referred to as Hybrid Optical Switching (HOS), is presented. HOS integrates optical packet, burst, and circuit switching on the same network and envisages the use of two parallel switches, a slow optical switch for the transmission of circuits and long bursts, and a fast switch for the transmission of packets and short bursts. The most appropriate

switching method is selected for the traffic generated by different applications and the less power consuming elements are utilized for transmission, ensuring flexibility, QoS differentiation, and low energy consumption. The HOS architecture is analyzed and compared with traditional solutions based on electronic switching through a combined simulation and analytical approach. Results show the effectiveness of the proposed solution.

The energy consumption of a data center is divided in (*i*) energy consumed by the ICT equipment, (*ii*) energy consumed by the cooling system and (*iii*) energy consumed by the power supply chain. According to the latest specifications data centers are designed in such a way that the ICT equipment consumes nearly all the energy within the data center. As a consequence, major energy savings in modern data centers can be achieved by reducing the energy consumption of the ICT equipment and in particular the energy consumption of the internal interconnection network. To this aim, in the second part of this thesis two novel optical switched data center interconnects are proposed. The first, referred to as HOS data center (HOSDC) interconnect, is based on the HOS switching paradigm while the second, referred to as Elastic Optical Data Center (EODC) interconnect, is based on the elastic optical networking concept. The energy consumption of the two architectures is evaluated in order to show the advantages with respect to existing solutions. Finally, an integrated all-optical network that provides both intra-data-center and inter-data-center connectivity together with interconnection toward legacy IP networks is proposed for achieving high overall energy improvements.

Contents

1 Introduction **9**

1.1 Telecommunication Core Networks 10

1.2 Data Centers 13

1.3 Outline of the Thesis 16

2 Energy-efficiency in Core Networks **19**

2.1 Hybrid Optical Switching 20

2.1.1 Forwarding Plane Architecture 21

2.1.2 Core Node Architectures 26

2.1.3 Power Consumption Model 29

2.1.4 Core Node Simulation Setup 34

2.1.5 Numerical Results 37

2.1.6 Conclusions 43

2.2 HOS Core Network Employing GMPLS 45

2.2.1 Network Overlay Model 46

2.2.2 Architectures and Power Consumption 49

2.2.3 Core Network Simulation Setup 53

2.2.4 Numerical Results 55

2.2.5 Conclusions 57
2.3 HOS Edge Network and QoS 59
2.3.1 Quality of Service 60
2.3.2 Edge Node Architectures 62
2.3.3 Edge Nodes Power Consumption 64
2.3.4 Edge and Core Network Simulation Setup 66
2.3.5 Numerical Results 69
2.3.6 Conclusions 78

3 Energy-efficiency in Data Centers 81
3.1 HOS for Data Center Networks 83
3.1.1 Data Center Interconnects 83
3.1.2 Power Consumption Model 89
3.1.3 HOS Data Center Simulation Setup 93
3.1.4 Numerical Results 97
3.1.5 Conclusions 106
3.2 Elastic Optical Data Center 109
3.2.1 Elastic Optical Interconnect 110
3.2.2 Power Consumption Model 113
3.2.3 Numerical Results 115
3.2.4 Conclusions 117
3.3 Carrier Cloud Network 118
3.3.1 Integrated HOS Network Architecture 119
3.3.2 Edge Caching 122
3.3.3 Power Consumption Model 123
3.3.4 Carrier Cloud Network Simulation Setup 127
3.3.5 Numerical Results 130
3.3.6 Conclusions 140

4 Conclusions **143**

Bibliography **147**

Publications List **157**

Journal Papers 157

Conference Papers 158

Acknowledgments **159**

Chapter 1

Introduction

Tackling climate change is a huge challenge for the world and an increasing issue in political agendas. The Information and Communication Technology (ICT) sector can make an immense contribution to low-carbon development through the deployment of innovative products and applications that provide the same or better service while significantly reducing carbon emissions. However, the energy and carbon impact of the ICT sector itself is already significant and is expected to expand rapidly over the coming decade due to the rapid proliferation of devices connected to the Internet and emergence of new services. According to the International Telecommunications Union (ITU), the ICT sector is today responsible for about 2.5% of the global Greenhouse Gas (GHG) emissions and its impact is expected to nearly double by the year 2020 [1]. Therefore, there are many efforts to improve energy efficiency in communication networks, ranging from the component technology to the architectural and service level approaches.

The energy consumption of the ICT sector can be divided in: (*i*) energy consumed by the user devices, (*ii*) energy consumed by the telecommunication network infrastructure, and (*iii*) energy consumed by the data centers. The user devices include Personal Computers (PC), portable devices (e.g. laptops, tablets and smartphones), peripherals, and printers and are the major contributor to the energy consumption of the ICT sector, accounting for 49% of the total [2]. On the other side, the telecommunication network infrastructure and the data centers are responsible together of 51% of the total energy consumption of the ICT sector, i.e. more than a half of the total. With the expected growth in the

Internet and data center traffic [3, 4] the energy consumed by telecommunication networks and data centers is destined to drastically increase if solution for increasing the energy efficiency would not be taken into account.

This thesis addresses the energy consumption of telecommunication networks and data centers and proposes new energy efficient architectures based on optical technologies that can cope with the increasing Internet traffic demand in a sustainable way. The rest of Chapter 1 is organized as follows. In Section 1.1 we give an overview of energy issues in current core network architectures and we point out the contribution of the thesis in this area. In Section 1.2 we address the energy limitations of current data center architectures and highlight the contribution of the thesis in this area. Finally, in Section 1.3 we describe the structure of the thesis.

1.1 Telecommunication Core Networks

Telecommunication networks can be divided into three domains: (*i*) access, (*ii*) metro, and (*iii*) core. An access network connects end users to the rest of the network infrastructure and spans a few kilometers. Metro networks span a metropolitan region, covering areas of tens of kilometers. Core networks are the central part of the network hierarchy, providing nationwide or global coverage. Recent research activity has concentrated on assessing energy consumption in telecommunications networks and evaluating the impact of the different network domains [5–8]. It was shown that although access networks are currently the major contributor, the energy consumption of core networks is expected to grow rapidly to be able to support future capacity requirements, which are in the range of several hundreds of Terabit per second (Tbps) or even Petabit per second (Pbps) per node [9, 10].

Core networks usually rely on optical transmission. In particular, the Internet Protocol (IP) over Wavelength Division Multiplexing (WDM) technology promises to meet the growing Internet bandwidth requirements [11]. In fact, the optical WDM layer can exploit the huge capacity that optical fibers offer by dividing the available bandwidth into multiple independent channels, each capable of carrying high data rate signals. Current implementations of IP over WDM networks employ electronic routers. Accordingly, data transmission occurs in the optical domain, whereas data switching and control information processing

are deployed in the electronic domain. Consequently, at each node along their path toward the destination, data undergo optical-to-electrical-to-optical (OEO) conversion. This solution offers several advantages, such as high network flexibility, high bandwidth usage, and the ability to deploy advanced quality-of-service (QoS) and traffic engineering policies. Furthermore, electronic buffers could enable us to implement highly efficient scheduling algorithms that introduce negligible data losses. That said, scaling current electronic router architectures to capacities of several tens or hundreds of Tbps leads to high energy consumption and consequently to high GHG emissions.

To reduce energy consumption in current IP over WDM networks, we can employ optical switching solutions. Optical switches consume potentially less energy than electronic routers, and their energy consumption does not increase significantly as the bit rate increases, leading to greater network scalability [12–14]. The lack of efficient optical buffering techniques, however, means that optical switches can hardly provide the flexibility and advanced data processing capability current electronic IP routers offer. Considerable research has demonstrated that a pure optical packet switch that can provide performance similar to its electronic counterpart would require a large number of complex optical components [15, 16]. Consequently, the energy consumption of a high-capacity, all-optical packet router would be the same or even higher than that of a comparable electronic IP router. Thus, we need a novel network concept that can exploit optical switching solutions more efficiently to significantly reduce energy consumption while fulfilling Internet application requirements.

Three optical switching solutions are available for IP over WDM networks. Optical Circuit Switching (OCS) offers low costs and energy consumption but limits flexibility and bandwidth utilization and is thus unsuitable for Internet applications, which generate highly variable traffic. To address this shortcoming, Optical Burst Switching (OBS) technology assembles data into bursts at the network ingress edge nodes [17, 18]. Before the edge node sends a burst, it forwards a control packet toward the destination to make the resource reservation. The actual burst transmission occurs after a fixed delay called the offset time. A pure OBS network suits Internet traffic bursty nature while keeping network costs and energy consumption bounded. However, it also introduces high data losses at core nodes, leading to low throughputs and a need for overprovisioning. Finally, Optical Packet Switching (OPS) achieves the highest flexibility

and resource exploitation and thus represents the best alternative to electronic switching; however,it requires complex optical components that introduce high costs and energy consumption. To combine the advantages of OCS, OBS, and OPS on the same network, hybrid optical switching (HOS) approaches have been recently proposed [19–22]. In HOS long traffic flows are carried over circuits or long bursts, whereas short traffic flows are carried over packets or short bursts. HOS could reach the flexibility and bandwidth use that future Internet applications require, while reducing power consumption and increasing the scalability of current core network solutions.

In general, we can divide HOS networks into two categories: optical/electronic and all-optical. The first category envisages using slow optical switches to forward circuits and long bursts, and fast electronic switches to forward packets and short bursts. Consequently, the network transmits large and long-lasting traffic flows transparently, while OEO converting short and more dynamic traffic flows at the core nodes. In contrast, in the all-optical solution, both the slow and fast switches are implemented via optical technologies, so all data are transmitted transparently with regard to OEO conversion. Most research efforts on HOS focus on evaluating network performance, with only a few works studying the network architecture and its energy consumption.

In this work we propose a novel data and control plane architecture for HOS networks. The data plane makes use of a slow optical switch, based on three-dimensional (3D) Micro Electro-Mechanical Systems (MEMS), and a fast switch, which can be either optical and based on Semiconductor Optical Amplifiers (SOA) or electronic and based on fast CMOS switching elements. The control plane is organized in an overlay model ad is composed of two layers, namely control layer and forwarding layer. The control layer performs routing, signaling, and link management as defined in the Generalized Multiprotocol Label Switching (GMPLS) standard [23]. The forwarding layer employs advanced data scheduling and resource reservation algorithms that guarantee high resource utilization and low energy consumption. We study performance and energy consumption of the proposed network using a combined simulation and analytical approach. The simulator is an event-driven *C++* software simulator developed specifically for the study of HOS networks. The performance and energy consumption of our HOS network are evaluated and compared with respect to traditional solutions based on electronic switching.

1.2 Data Centers

A data center refers to any large, dedicated cluster of computers that is owned and operated by a single organization. Mainly driven by emerging cloud computing applications data center traffic is showing an exponential increase. Cisco [4] reports that while the amount of traffic crossing the Internet is projected to reach 1.3 zettabytes per year in 2016, the amount of data center traffic has already reached 1.8 zettabytes per year, and by 2016 will nearly quadruple to about 6.6 zettabytes per year. This corresponds to a Compound Annual Growth Rate (CAGR) of 31% from 2011 to 2016. The main driver to this growth is cloud computing traffic that is expected to increase six-fold by 2016, becoming nearly two-thirds of total data center traffic. To keep up with these trends, data centers are improving their processing power by adding more servers. Already now large cloud computing data centers owned by online service providers such as Google, Microsoft, and Amazon host tens of thousands of servers in a single facility. With the expected growth in data center traffic, the number of servers per facility is destined to increase posing a significant challenge to the data center interconnection network.

Another issue rising with the increase in the data center traffic is energy consumption. The direct electricity used by data center has shown a rapid increase in the the last years. Koomey estimated that the aggregate electricity use for data centers worldwide doubled from 2000 to 2005 [24, 25]. The rates of growth slowed significantly from 2005 to 2010, when the electricity used by data centers worldwide showed an increase by about 56%. Still, data centre are a main contributor the energy consumption of the ICT sector.

The overall energy consumption of a data center can be divided in energy consumption of the IT equipment, energy consumption of the cooling system and energy consumption of the power supply chain. The ratio between the energy consumption of the IT equipment and the overall energy consumption represents the power efficiency usage (PUE). The PUE is an important metric that shows how efficiently companies exploit the energy consumed in their data centers. The average PUE among the major data centers worldwide is estimated to be around 1.80 [26] meaning that for each Watt of IT energy 0.8 W are consumed for cooling and power distribution. However, modern data centers show higher efficiency. Google declares that its most efficient data center shows

a PUE as low as 1.12. We can then conclude that the major energy savings in modern data centers can be achieved by reducing the power consumption of the IT equipment. The energy consumption of IT equipment can be further divided in energy consumption of the servers, energy consumption of the storage devices and that of the interconnection network. According to [27] current data centers networks consume around 23% of the total IT power. When increasing the size of data centers to meet the high requirements of future cloud services and applications, the internal interconnecting network will most likely become more complex and power consuming [28, 29]. As a consequence, the design of more energy efficient data center networks is of utmost importance for the scope of building greener data centers.

Current data centers networks rely on electronic switching elements and point-to-point (ptp) interconnects. The electronic switching is realized by commodity switches that are interconnected using either electronic or optical ptp interconnects. Due to the high cross talk and distance dependent attenuation very high data-rates over electrical interconnects can be hardly achieved. As a consequence, a large number of copper cables are required to interconnect a high-capacity data center, thereby leading to low scalability and high power consumption. Optical transmission technologies are generally able to provide higher data rates over longer transmission distances than electrical transmission systems, leading to increased scalability and reduced power consumption. Hence, recent high-capacity data centers are increasingly relying on optical ptp interconnection links. According to an IBM study [30] only the replacement of copper-based links with VCSEL-based ptp optical interconnects can reduce the power consumption of a data center network by almost a factor of 6.

However, the energy efficiency of ptp optical interconnects is limited by the power hungry electrical-to-optical (EO) and optical-to-electrical (OE) conversion required at each node along the network since the switching is performed using electronic packet switching. When considering future data center requirements, optical switched interconnects that make use of optical switches and WDM technology, can be employed to provide high communication bandwidth while reducing significantly the power consumption with respect to ptp solutions. It has been demonstrated in several research papers that solutions based on optical switching can improve both scalability and energy efficiency with respect to ptp interconnects [28, 29] As a result, several optical switched interconnect

architectures for data centers have been recently presented [31–40]. Some of the proposed architectures [31, 32] are based on hybrid switching with packet switching in the electronic domain and circuit switching in the optical domain. The others are based on all-optical switching elements, and rely either on optical circuit switching [33, 34] or on optical packet/burst switching [35–40]. Only a few of these studies evaluate the energy efficiency of the optical interconnection network and make comparison with existing solutions based on electronic switching. Furthermore, only a small fraction of these architectures are proven to be scalable enough to keep up with the expected increase in the size of the data centers. Finally, none of this study addresses the issue of flexibility, i.e. the capability of serving efficiently traffic generated by different data centers applications.

With the worldwide diffusion of cloud computing, new data center applications and services with different traffic requirements are continuously rising. As a consequence, future data center networks should be highly flexible in order to serve each application with the required service quality while achieving efficient resource utilization and low energy consumption. To achieve high flexibility, in telecommunication networks HOS approaches have been recently proposed. HOS combines optical circuit, burst and packet switching on the same network and maps each application to the optical transport mechanism that best suits to its traffic requirements, thus enabling service differentiation directly in the optical layer. In this thesis we propose an optical data center network architecture, referred to as HOS data center (HOSDC) interconnect, which provides high energy efficiency and QoS differentiation in order to satisfy the requirements of future data center applications. The HOSDC interconnect can be realized by adding minimal hardware modifications to current data center interconnect and represent a promising solution for the short and mid term. We study performance and energy consumption of the HOSDC interconnect and we compare to traditional solutions.

New applications and services are pushing for higher transmission capacities at the server side. As a consequence, the transmission rate of the Network Interface Cards (NIC) of the servers is increasing. Even if today the majority of commercially available servers are equipped with 1 Gbps NIC [41], it is expected that in some years servers will be mainly equipped with 10 Gbps or higher capacity NIC. To be able to support these capacities in an energy efficient and flexible

manner we propose a novel optical network interconnect for data centers which is realized by integrating two advanced optical networking technologies: broadcast-and-select approach and Elastic Optical Networking (EON). The proposed data center network architecture is referred to as Elastic Optical Data Center (EODC) interconnect. The EODC interconnect is able to support capacities of 100 Gbps per server or even higher, but is based on advanced optical components and require a complete change in the structure of current data center networks. As a consequence, we propose the EODC architecture as a long term solution for data centers. We study the energy consumption of the EODC interconnect and we compare both with traditional solutions and with the HOSDC interconnect.

Recently, communication service providers are looking for cloud solutions to reduce costs and create new level of efficiency. In this context, one of the most promising solutions is carrier cloud [42]. In carrier cloud a unique operator owns and runs both the core network and the data center internal interconnect. Through network virtualization techniques, the carrier cloud operator sells services to different Internet Service Providers (ISP) which act as tenants. In such a scenario, the definition of a unique network architecture able to provide both intra and inter data center connectivity could lead to increased resource utilization and increased energy efficiency. A main advantage of the integration of core and intra-data-center networks comes from the possibility to avoid the energy inefficient electronic interfaces between data centers and telecommunication network. To this aim we propose a unified network architecture that provides both intra-data-center and inter-data-center connectivity together with interconnection toward legacy IP networks. The proposed architecture is based on the HOS concept. We study the performance and energy consumption of the proposed solution and we compare to a similar network solution that still rely on electronic interfaces to interconnect the core network and data centers.

1.3 Outline of the Thesis

The rest of the thesis is organized as follows.

In **Chapter 2**, we analyze the energy consumption of core networks and we propose a new energy efficient network architecture based on the HOS paradigm.

- In **Section 2.1**: (*i*) we define the proposed HOS forwarding plane architecture, (*ii*) propose two possible HOS core node architectures, (*iii*) develop an analytical model for evaluating the power consumption of optical core nodes, (*iv*) evaluate the data losses and energy consumption of the proposed HOS core nodes, and (*v*) make extensive comparison with traditional core nodes based on electronic switching.

- In **Section 2.2**: (*i*) we propose an integration of the HOS forwarding plane and the GMPLS control plane in order to perform routing and network management in a HOS core network and (*ii*) we study the architecture and energy consumption of the integrated control plane.

- Finally, in **Section 2.3**: (*i*) a novel HOS edge node architecture is proposed in order to interconnect the HOS core network to legacy IP networks, (*ii*) a technique for mapping different Internet applications to the most suited HOS transport mechanism is studied, (*iii*) the data losses, end-to-end delay and energy consumption of an HOS network with high-capacity edge and core nodes are studied and compared against traditional network solutions.

In **Chapter 3**, we address the energy consumption of data centers and we propose two novel optical data center interconnect architectures and an integrated intra-data-center and core network for carrier cloud applications.

- In **Section 3.1**: (*i*) a novel optical interconnect for data centers based on the HOS concept (HOSDC) is proposed for short/mid term applications and (*ii*) its energy consumption and performance are evaluated and compared with respect to traditional data center networks.

- In **Section 3.2**: (*i*) a novel optical interconnect for data centers based on optical broadcast-and-select and EON concepts is proposed for long term applications and (*ii*) its energy consumption is evaluated and compared against the HOSDC interconnect and other traditional data center networks.

- Finally, in **Section 3.3**: (*i*) we propose a unified network architecture that provides both intra-data-center and inter-data-center connectivity together with interconnection toward legacy IP networks, (*ii*) we evaluate performance and energy consumption and we compare against standard non-integrated networks, and (*iii*) we estimate the impact of edge caching.

Finally, in **Chapter 4** the conclusions are drawn and the main achievements of the thesis are summarized.

Chapter 2

Energy-efficiency in Core Networks

In the future, the Internet may ultimately be constrained by energy consumption and the capability to provide QoS. As regards the Internet core, HOS is promising to provide service differentiation and reduced energy consumption in respect to current electronic switching solutions. HOS aims to combine optical circuit, burst and packet switching on the same network.

In this Chapter, we present a novel integrated control plane for a HOS core network. The HOS network is organized in an overlay model with the HOS control layer performing routing, signaling, and link management, and with the HOS forwarding layer managing the reservation of resources and data scheduling. The HOS forwarding layer makes use of a unified control packet able to carry the control information for all the different data formats and employs an appropriate scheduling algorithm for each incoming data type. The proposed HOS network envisages the use of two parallel switches, a slow optical switch for the transmission of circuits and long bursts, and a fast switch for the transmission of packets and short bursts. The most appropriate switching method is selected for the traffic generated by different applications and the less power consuming elements are utilized for transmission, ensuring flexibility, QoS differentiation, and low energy consumption. Performance and energy efficiency of the proposed HOS network are assessed by means of a combined analytical and simulation approach. The obtained results show that HOS has a very high energy efficiency and allows a sustainable growth of the Internet.

2.1 Hybrid Optical Switching

New applications and services have led to the need of high-speed and high-capacity switches and routers. Also, the availability of new generation high-speed access networks has put more pressure under backbone networks and backbone nodes in particular, which are required to support loads of several tens of Tbps and above.

Wavelength division multiplexing (WDM) represents the reference technology for current backbone networks. High-capacity optical backbones usually comprise edge nodes and core nodes. Edge nodes are located at the periphery of the network and are used to link the optical network to the legacy networks. Core nodes are the internal nodes of the optical network and are used to route data from ingress to egress edge nodes. Three different optical switching schemes have been proposed for WDM networks, namely OCS, OBS, and OPS.

In an OCS network, a dedicated path (called lightpath) connecting source and destination nodes is created before the actual data transmission. A lightpath reserves one specific wavelength in each link toward the destination and it is established using a two-way reservation mechanism. A pure OCS network perfectly fits long and stable traffic flows, but it leads to poor bandwidth utilization in presence of bursty sources.

In an OBS network [17, 18, 43] data are assembled into bursts at the ingress edge nodes. Before a burst is sent a control packet is forwarded toward the destination in order to make the resource reservation. The actual transmission of the burst takes place after a fixed delay called offset time. A pure OBS network is effective for both constant and bursty sources, but introduces a high loss rate at the core nodes, which can be only avoided by implementing expensive contention resolution techniques.

In an OPS network [44–46], optical packets are sent toward the destination without any resource reservation in advance. The control information is encapsulated into the packet header, which is separated from the packet payload by a time-guard. A pure OPS network provide high bandwidth utilization, but introduces high costs and power consumption.

Nowadays, one of the the main limiting factors in realizing nodes able to support capacities of several tens of *Tbps* is their high energy consumption. Scaling current routers to capacities of more than hundred *Tbps* would lead to

power consumptions above 1 MW [10], which is very difficult to provide. Different approaches and technologies have been addressed to be possible candidates for implementing next generation high performance nodes. For this reason, the concept of green (i.e. energy-efficient) optical networks has been introduced and several solutions for decreasing energy consumption have been proposed, e.g. [47–50]. The authors in [47] present an analysis of the current research activity on green optical networking and propose a model based on optical bypass technology and waveband switching. The authors in [48] devised a Mixed Integer Linear Programming (MILP) and two heuristics based on lightpath bypass technology for minimizing energy consumption in IP over WDM networks. The obtained results show that IP routers are the major contributors to the total power consumption, using more than 90% of the total power. The authors in [49] and [50] introduce the concept of power-awareness in traffic grooming. In [49] an Integer Linear Programming (ILP) formulation of the traffic grooming problem is given, while in [50] an heuristic approach is proposed. However, these solutions rely on OCS and static resource assigment and for this reason they are not able to serve future application requirements in a flexible manner.

In order to combine the advantages of OCS, OBS and OPS on the same network and reduce the energy consumption of the current electronic switching core nodes, HOS approaches can be employed. However, an efficient control plane and node architecture for HOS networks has not yet been defined. For this reason, in this Section we propose a novel HOS forwarding plane architecture and two possible HOS core node architectures. The models and results presented in this Section are part of the work presented in [19]. In Section 2.1.1 the proposed HOS forwarding plane is presented. In Section 2.1.2 two HOS core node architectures are proposed and the architecture of a traditional electronic switching node is illustrated. In Section 2.1.3 an analytical model for the power consumption evaluation of the considered nodes is introduced. In Section 2.1.4 the performance metrics used to analyze the and compare the different node architectures are introduced. In Section 2.1.5 numerical results obtained through simulation are shown and in Section 2.1.6 the main conclusions are drawn.

2.1.1 Forwarding Plane Architecture

In this Section we describe the HOS forwarding plane architecture [19]. Firstly, we present the data and control packet formats and, secondly, we describe the data scheduling algorithms.

2.1.1.1 Data and Control Packet Formats

The HOS node is supposed to handle all three switching paradigms (OCS, OBS and OPS) in an effective manner. In the case of OCS, we assumed the use of time division multiplex (TDM) circuits. In a TDM-circuit, time is divided in frames, each of which is divided in a fixed number of time-slots. Different traffic flows, sharing the same circuit, use different time-slots in a time-domain multiple access (TDMA) manner. If the number of flows sharing the same TDM-circuit is lower than the number of time-slots in a frame, some slots are not used. In this case, a core node along the circuit path can fill the unused slots in the frame with incoming optical packets of suitable length having the same destination as the circuit. This technique aims to increase the bandwidth utilization for two reasons: firstly, because the circuit utilization is raised and, secondly, because packets scheduled within a circuit do not consume new resources, letting more space available for other traffic flows.

As mentioned before OCS makes use of a two-way reservation mechanism. Therefore, the source node transmits a lightpath establishment request toward the destination node and waits for the acknowledgment (ACK) before starting the transmission of data. When a core node receives the lightpath establishment request, it reserves the required resources, if available, and waits for the incoming data. The time interval between the arrival of the lightpath establishment request and the arrival of data could be very long, usually in the order of milliseconds or hundreds of milliseconds, and during this time the reserved resources cannot be used by other connections. Therefore, in a HOS network, where circuits, bursts, and packets share the same resources, the use of the two-way reservation mechanism for circuits would lead to a low bandwidth utilization. Consequently, aiming to improve the network efficiency, an offset-time for circuits is defined. This offset-time informs the nodes along the path about the exact time in which data are going to arrive at the input ports of the switch, so that the nodes reserve the resources only for the actual duration of the circuit. The offset-time for circuits is defined to be greater than a round trip time (RTT), because the data transmission can start only after the ACK has reached the source node. Furthermore, the circuit-offset is defined to be greater than the sum of the maximum burst-offset and the maximum burst-length, so that circuits result to have the higher priority with respect to bursts and packets.

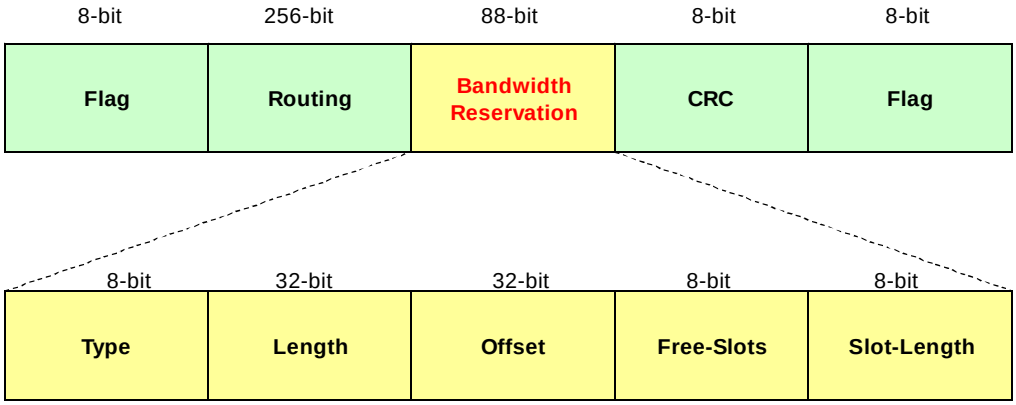

Figure 2.1: Control packet format.

Regarding switching of optical bursts, the just enough time (JET) reservation mechanism [18] has been chosen. Here, when a core node receives the control packet it schedules the burst by reserving the resources only for its actual length, thus providing efficient resource utilization. Because bursts are scheduled a priory due to the offset-time, this results in a kind of prioritized handling in comparison to packets. Basing on the discussion above, circuits result to have the highest priority and are thus designed to carry premiere traffic. On the contrary, packets result to have the lowest priority and are thus designed to carry best-effort traffic. Bursts have different priority basing on their offset-time and therefore can be used for implementing QoS differentiation.

The format of the unified control packet defined for the management of all data types is reported in Figure 2.1. The packet is 368 $bits$ long and comprises two main fields, namely routing information and bandwidth reservation. The routing information field contains the source and destination addresses in the IPv6 format (128 $bits$). It is also possible that a different addressing method is used, such as MPLS, but in order to be flexible we decided to reserve more space for the address section. The bandwidth reservation field is 88 $bits$ long and carries the information required by the core nodes for deploying the data scheduling. It is divided into the following five subfields.

The type sub-field (8 $bits$) is used by the core node to recognize the type of the incoming data. Depending on its value the node learns if the control packet is a lightpath establishment request, a burst control packet or an optical packet header and decides which scheduling algorithm to use. The length sub-field (32 $bits$) contains the length of the data corresponding to the control packet

expressed in $kBytes$. In this work, the data length is assumed to be a random value in a specific range. Different ranges have been defined for different data types. In case of a lightpath establishment request or a burst control packet, the offset sub-field ($32\,bits$) carries the offset-time expressed in μs. The burst-offset is defined as a random value in a specific range, whereas the circuit-offset is fixed. In case of a data packet header, the offset sub-field is ignored by the node. Finally, the free-slots ($8\,bits$) and the slot-length subfields ($8\,bits$) carry the information necessary to the core nodes to be able to schedule optical packets in unused slots of circuits. The free-slots subfield contains the number and the position of the unused TDM slots in each frame of the circuit, while the slot-length subfield contains the length of a TDM-slot expressed in $kBytes$.

Different coding techniques have been analyzed and compared with each other concerning achievable performance and practicability [51, 52]. One of the most promising is the optical sub-carrier modulation (SCM) technique, which we selected for multiplexing data and control packet in the proposed network solution. In SCM the control packet and the data payload are modulated on the same optical carrier, by treating the payload as a baseband signal while modulating the control information on a sub-carrier frequency. Several implementations of this technique for optical label switching and optical packet switching have been proposed in literature [53, 54].

2.1.1.2 Data Scheduler

The control plane employs an appropriate scheduling algorithm for each of the different incoming data types [19]. In the following the implemented algorithms are presented.

Circuit scheduling: since circuits have the highest priority, they do not compete with bursts and packets, so that they can be blocked only by other circuits. A new circuit is then scheduled on the first output channel in which no other circuit has been previously scheduled. However, if too many circuits are scheduled on the same output fiber, the burst and packet traffic on that fiber would be almost completely blocked. For this reason, the maximum number of circuits that can be simultaneously scheduled on the same fiber has to be limited.

Burst scheduling: several algorithms for burst scheduling have been proposed [17, 18, 55–57]. In this study, two particular algorithms have been selected

and implemented. The first one, proposed in [56], is the First Fit Unscheduled Channel with Void Filling (FFUC-VF). In this algorithm, the scheduler keeps track of all the void intervals and assigns a new burst to the first suitable void, which is found by checking all the available wavelength channels in a sequential manner. FFUC-VF has been chosen because it provides a good bandwidth utilization while introducing a relatively small complexity and processing time. Then performance with the Best Fit with Void Filling (BF-VF) algorithm, described in [18], have also been evaluated. BF-VF keeps track of all the voids created before and after the new inserted burst. When a suitable void is found, the scheduler computes the difference between the arriving time of the burst and the starting time of the void, and the difference between the ending time of the void and the ending time of the burst. The BF-VF algorithm selects the void in which the sum of these values is lower. BF-VF achieves higher bandwidth utilization with respect to FFUC-VF, but introduces also higher complexity and processing times.

Packet scheduling: when a new packet reaches the node, firstly the scheduler tries to insert it on a free TDM-slot of an already established circuit. In order to do that, the scheduler selects all the circuits with the same destination as the packet and checks if one of them has a suitable free slot. If there is no such a circuit, the packet is scheduled on the first idle channel in which no reservation has been made for the time required for transmission. Firstly, no optical buffers for packets contention resolution were used. In this case, the scheduler drops any incoming packet for which no available output resource is found. Afterwards, aiming to reduce the loss rate, fiber delay lines (FDLs) for packets contention resolution have been added. Two possible architectures for FDLs can be considered: the feed-forward and the feed-back. In the feed-forward one, packets are fed into FDLs of different lengths that are grouped in FDL banks. Each FDL bank is assigned to an output port. Once packets come out of the FDL, they must be scheduled. In the feed-back architecture, a packet can re-circulate in a loop created by one FDL bank and a pair of switch ports as long as an available output resource is found. The feed-back architecture is more efficient than the feed-forward, but requires an increased switch size. In this work, the feed-back configuration has been chosen. The maximum number of re-circulations for a packet has been fixed to three, in order to prevent the optical signal from becoming too poor to be transmitted.

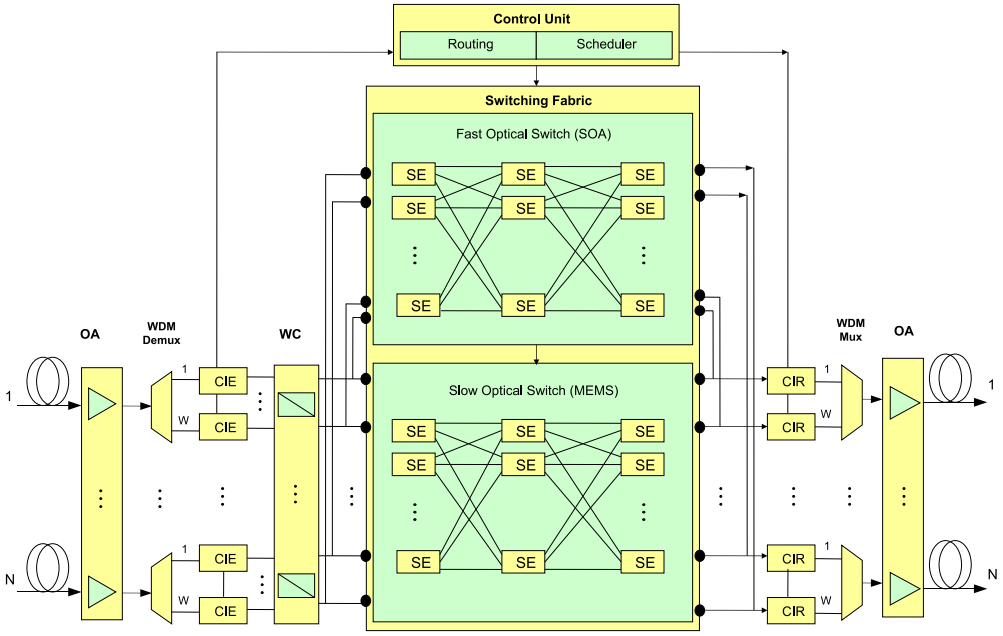

Figure 2.2: All-optical HOS core node architecture.

2.1.2 Core Node Architectures

In this Section, the three node architectures considered in the study are presented. We propose two possible realizations of a HOS core node, namely all-optical and optical/electronic, and a traditional architecture of a core node based on electronic switching [19]. The architectures are shown is Figure 2.2, 2.3 and 2.4, respectively.

All the considered architectures are composed of an electronic control unit and a switching fabric. Each of N incoming fibers, carrying W wavelength channels, is connected to a WDM-DEMUX in which the wavelength channels are separated. Each of $N \cdot W$ incoming wavelength channels corresponds to an input port of the node and it is supposed to carry simultaneously data and its corresponding control information. The node is supposed to be symmetric, i.e., the number of input ports is equal to the number of output ports. The control information associated to each wavelength channel is extracted and sent to the control unit by the control information extractor (CIE), while the corresponding data are sent towards the switching fabric. The control unit converts the control signal into the electronic domain and processes it in order to select the proper output port for the corresponding data. It is divided in two functional blocks:

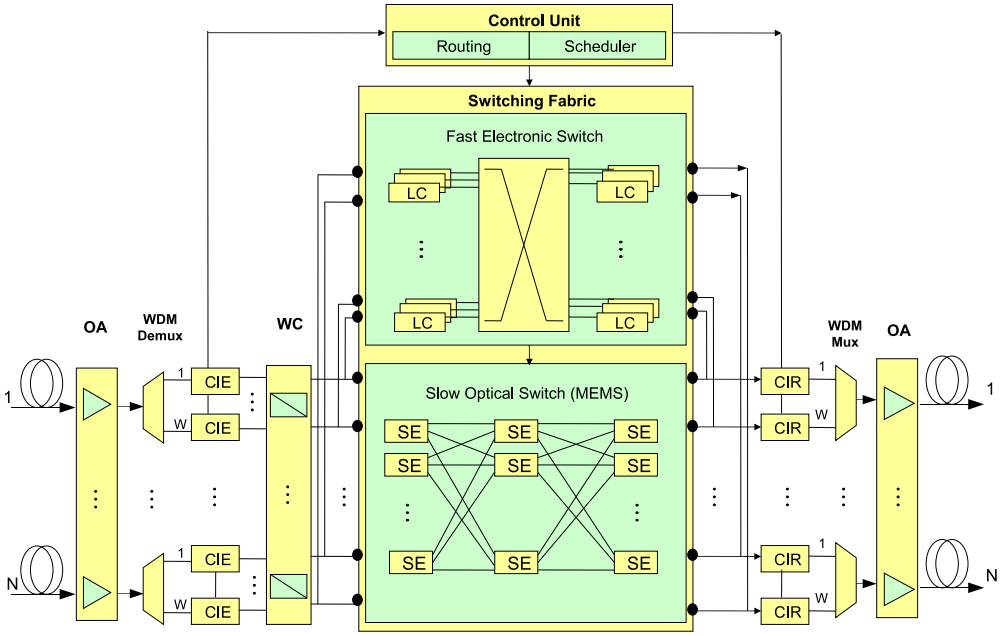

Figure 2.3: Optical/electronic HOS core node architecture.

the routing unit and the scheduler. Wavelength converters and optical buffers can be used for solving eventual contentions at output ports.

In the hybrid architectures, the switching fabric is divided in two parts: a slow and a fast switch. The slow switch is realized using switching elements whose switching time is in the order of milliseconds or hundreds of microseconds and it is used to transmit slow traffic (i.e., circuits and long bursts). The fast switch is realized using switching elements whose switching time is in the order of nanoseconds and it is used for carrying fast traffic (i.e., packets and short bursts).

In the all-optical hybrid architecture (Figure 2.2), both the fast and the slow switch are realized using optical elements. The fast optical switch is obtained by deploying switching elements (SE) organized in a three-stages Clos architecture. The Clos architecture used here (Figure 2.5) is a strictly non-blocking network, in which each switching element of one stage is connected to all the switching elements of the next stage. Each switching element in the Clos network is internally realized using SOA organized in a Spanke architecture. SOA have switching times in the order of nanoseconds, so that they can be used for switching both slow and fast traffic.

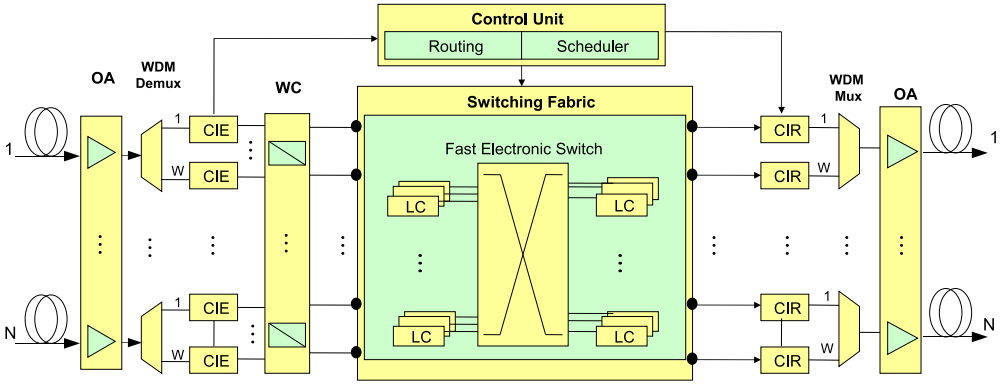

Figure 2.4: All-electronic packet switching core node architecture.

The slow switching fabric is realized using MEMS. MEMS are miniature movable mirrors made in silicon that transmit or deflect the optical signal depending on their position. MEMS have switching time in the order of milliseconds or hundreds of microseconds and for this reason can be used only for slow switching. Anyway MEMS have several advantages if compared to SOAs. Firstly, using MEMS it is possible to realize switching fabrics of large dimension. Secondly, MEMS consume much less power with respect to SOAs. Finally, MEMS are made in silicon, which is a very mature technology, and allow to obtain more reliable and stable devices. For all these reasons in the all-optical hybrid architecture, the fast traffic only is switched using SOAs while the slow traffic is switched using MEMS.

In the optical/electronic hybrid architecture (Figure 2.3) the switching fabric is made up by a slow optical switch and a fast electronic switch. The former is based on MEMS while the latter is composed of electronic line cards (LC) and electronic switching elements. The fast traffic is routed to the input LCs where it is converted and processed in the electronic domain. The structure of a fast electronic LC is reported in Figure 2.6 and comprises transceivers, PHY (physical layer) devices, framers/mappers, MAC chips, a traffic processor/forwarding engine (TP/FE), memory devices, and fabric interfaces.

The electronic switch permits to realize advanced data processing, including functions of the physical, data-link and network layer. As a consequence, the electronic switching fabric reaches high levels of traffic differentiation and QoS support that nowadays cannot be met by optical switching fabrics. Another

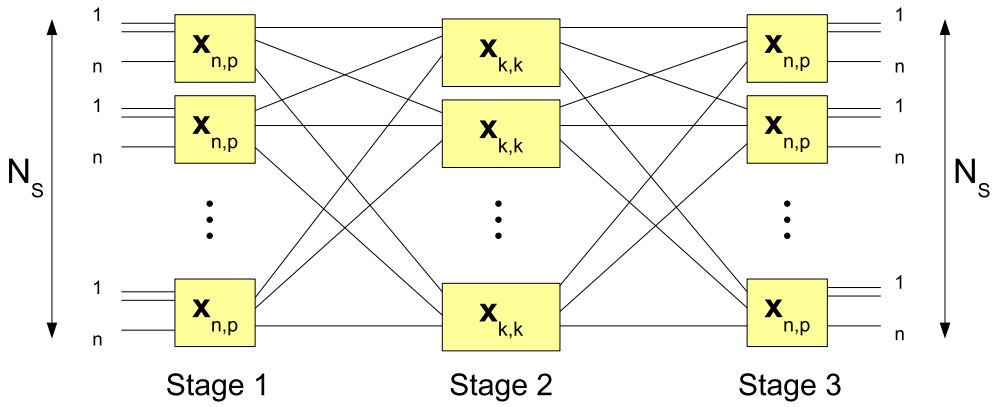

Figure 2.5: Example of a Clos network.

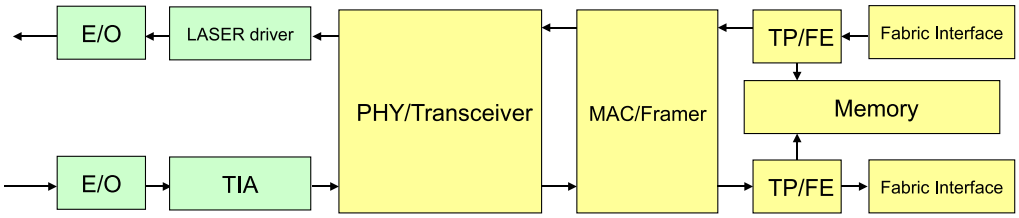

Figure 2.6: Block diagram of a fast electronic line card.

advantage of electronic switches over optical switches is the possibility of implementing electronic buffers that provide higher storage capacity in a reduced area when compared to FDLs. The drawback of using electronic switches is their high power consumption. To limit the overall power consumption, in the optical/electronic hybrid architecture only the fast traffic is switched within the electronic switch.

Finally, the all-electronic architecture is composed of a unique fast electronic switch (Figure 2.4) and represents the architecture of a current IP-router. Here, all the incoming traffic is routed to electronic line cards and forwarded by electronic switching elements. This provides high throughput, QoS differentiation, and enhanced data processing, but it also leads to higher overall power consumption.

2.1.3 Power Consumption Model

This Section describes the analytical model that we use to evaluate the power consumption of the HOS and the electronic core node architectures.

The total power consumption of the core node P_{TOT} is defined as the total power consumed by the active elements within the node and can be computed as the sum of the power consumed by the switching fabric, P_{SF}, and the power consumption of the other active components, P_{OAC}:

$$P_{TOT} = P_{SF} + P_{OAC} \tag{2.1}$$

The term P_{OAC} is common for all the considered architectures and is given by the following formula:

$$P_{OAC} = P_{CP} + P_{RP} + P_{SC} + NWP_{CIE/R} + 2NP_{OA} + N_{TWC}P_{TWC} \tag{2.2}$$

Here, P_{CP}, P_{RP}, P_{SC}, and $P_{CIE/R}$ represent the power consumption of the control plane, the route processor, the switch control unit, and the unit for control information extraction and reinsertion, respectively. P_{OA} is the power consumption of optical amplifiers (OA) and N is the number of input/output fibers. OAs are applied to compensate the losses suffered by the optical signal passing through the switch and we assumed the use of one OA at each input/output fiber.

As already mentioned before, we assumed that the control information is transmitted together with the data on the same optical carrier using the subcarrier (SCM) multiplexing technique. For both extraction and reinsertion of control information a subsystem based on Fiber Bragg Gratings (FBGs) such as those proposed and experimentally demonstrated in [53, 54] can be used. The power consumption of all active components such as VCO oscillators, RF up- and down-converters, modulator drivers, RF amplifiers, and optical receivers has to be taken into account. Thus, the power consumption of such a module can be estimated to be approximately $\approx 17W$. One CIE/R module must be placed for each wavelength channel.

Finally, P_{TWC} is the power consumption of tunable wavelength converters (TWC) and N_{TWC} is the number of active wavelength converters. TWCs are utilized to solve data contentions and we assumed one TWC per switch port in case of the full wavelength conversion option. Only active TWCs are supposed to consume power. The number of active TWCs (N_{TWC}) is an output

Table 2.1: Power consumption of optical active components.

Component	**Power**
Control Plane (P_{CP})	300 W
Route Processor (P_{RP})	200 W
Switch Control Unit P_{SC}	300 W
Optical Amplifier (P_{OA})	14 W
Tunable Wavelength Converter P_{TWC}	1.69 W
Control Information Extraction and Reinsertion ($P_{CIE/R}$)	17 W

of the developed simulator. The values for power consumption of the control plane module, the route processor, the switch control unit and the optical amplifier are estimated by collecting data of a number of commercially available components and modules of conventional large-scale switching and routing systems. The numbers given in Table 2.1 are obtained by averaging and rounding the power consumption values of modules that implement the same or a similar functionality.

The term P_{SF} depends on the considered architecture. In the hybrid architectures, P_{SF} is obtained by summing the power consumption of the slow and the fast switches, which are computed separately. The power consumption of the switching fabric is obtained by multiplying the power consumption of a single switch port, P_{Port}, and the number of active ports, N_{AP}:

$$P_{SF} = P_{Port} \cdot N_{AP} \tag{2.3}$$

N_{AP} is an output of the HOS network simulator, while P_{Port} has been computed utilizing the formulas reported in the following (based on the matehmatical model proposed in [10, 19].

2.1.3.1 SOA-Based Switch

Let N be the number of fibers and W the number of wavelengths per fiber. The total number of switch ports, N_S can be then calculated from:

$$N_S = N \cdot W \tag{2.4}$$

As reported before, the SOA-based switch is organized in a non-blocking three-stage Clos network. Through the total number of ports N_S it is possible

to compute the number of ports belonging to a single switching unit of the three-stages Clos network:

$$n = \sqrt{\frac{N_S}{2}} \tag{2.5}$$

Referring to the Clos network shown in Figure 2.5, in order to have a non-blocking three-stages Clos network the parameters p and k are defined through the following formulas:

$$k = \frac{N_S}{n} \tag{2.6}$$

$$p = 2n - 1 \tag{2.7}$$

With reference to [10], the number of active SOA within the active paths is defined by the following expression:

$$N_{SOA} = N_S(4\,log_2 p + 2\,log_2 k) \tag{2.8}$$

Finally, the number of active SOAs per switch port is obtained by dividing N_{SOA} by N_S:

$$N_{SOA\times Port} = \frac{N_{SOA}}{N_S} \tag{2.9}$$

In order to obtain the power consumption of a fast port based on SOAs, also temperature stabilization circuits (TEC) and 3R regenerators must be considered. The number of active TECs is defined through the following formula:

$$N_{TEC} = 2k + p \tag{2.10}$$

which is nothing but the number of switching elements within the Clos network. Consequently, the number of active TECs per switch port is given by:

$$N_{TEC\times Port} = \frac{N_{TEC}}{N_S} \tag{2.11}$$

By assuming a maximum allowed optical signal-to-noise ratio (OSNR) degradation of $-20\,dB$, 3R regenerators need to be placed after each ninth SOA within the switch. To obtain the number of active 3R regenerators per port the following expression is then used:

$$N_{3R\times Port} = \frac{N_{SOA\times Port}}{9} \tag{2.12}$$

With reference to [29], the following assumptions have been made for the power consumption of SOA, TEC and 3R regenerators. The power consumption of an active SOA is assumed to be $0.24\,W$, that of a TEC $4\,W$ and of a 3R regenerator $2.78\,W$. Consequently, the power consumption per port P_{Port} of a SOA-based switch expressed in W is then given by:

$$P_{Port} = 0.24 \cdot N_{SOA\times Port} + 4 \cdot N_{TEC\times Port} + 2.78 \cdot N_{3R\times Port} \tag{2.13}$$

It is worth noting that the number of active elements per switch port, and thus the power consumption per port, depends on the switch size. For instance, considering a unidirectional aggregate switch capacity of $76.8\,Tbps$ with 24 input/output fibers carrying 80 wavelength channels each, the power per switch port of the SOA-based switch, P_{Port}, is $19.9\,W$. Because the feed-back architecture has been chosen for implementing FDLs, if optical buffers are used, the number of ports of the switch is increased by the number of buffer ports. To evaluate the power consumption of a SOA-based architecture implementing FDLs, the previous procedure can be used, but the definition of N_S must be changed to the following:

$$N_S = N \cdot W + N_b \tag{2.14}$$

where the number of active buffer ports, N_b, is an output of the developed simulator.

2.1.3.2 MEMS-Based Switch

In this work, 3D-MEMS have been considered for the realization of the MEMS-based switches. The power per switch port of a MEMS-based switch corresponds then to the power consumption of two mirrors of a 3D-MEMS switch, which is assumed to be $0.1\,W$ [10].

Table 2.2: Power consumption of a unidirectional line card at 40 Gbps.

Component	**Function**	**Power**
Transceiver	LASER driver, TIA, postamp., equalization, clock and data recovery, mux/demux	5.9 W
PHY	Encoding/decoding, scrambling/descreambling, FEC	3.4 W
Framer/MAC	Mapping, framing, OH processing, payload processing, MAC	30.6 W
TP/FE	Packet processing, classifying, table lookup, packet forwarding	183.6 W
Fabric	Interface Switch fabric procontrol, port queuing, TP/FE interface	61.2 W
Memory	Packet memory, lookup table, configuration parameter	13.6 W
Total	**Line card at 40 Gbps**	**298.3 W**

2.1.3.3 Fast Electronic Switch

The power consumption per switch port of the fast electronic switch is calculated by summing the power of one LC and the power per port of an electronic switching unit. In this paper unidirectional line cards working at $40\,Gbps$ are considered. To obtain the power consumed by a line card the values for each functional block reported in [10] and summarized in Table 2.2 have been used. Assuming that the power consumption per $80\,Gbps$ bidirectional switching capacity ($40\,Gbps$ unidirectionally) of an electronic packet switch is $8\,W$ [10], the power consumption per switch port, P_{Port}, of the $76.8\,Tbps$ fast electronic switch is calculated to be $306.3\,(= 298.3 + 8)\,W$.

2.1.4 Core Node Simulation Setup

The simulation tool developed for the analysis of the proposed node is a time-discrete *C++* simulator. Time is divided in clock cycles, which has been fixed to the time needed for the transmission of $1\,kByte$ of data. The data rate has been set to $40\,Gbps$ and 80 wavelength channels per fiber have been considered, consequently, a simulator clock-time corresponds to $0.2\,\mu s$ and each input/output fiber of the switching node has a capacity of $3.2\,Tbps$. The simulator generates control packets and deploys the functionalities described in previous Sections to

forward the incoming traffic. The produced outputs are for both the performance and the power consumption analysis.

The simulation model defines the following parameters for keeping track of the generated and scheduled traffic. $A_{Counter}$ and $S_{Counter}$ store respectively the number of $Bytes$ arrived to the node and the actual number of $Bytes$ successfully scheduled. $P_{Arrival}$, $B_{Arrival}$ and $C_{Arrival}$ represent respectively the total number of packets, bursts and circuits arrived to the node during the entire simulation. Similarly, the parameters $P_{Scheduled}$, $B_{Scheduled}$, $C_{Scheduled}$, keep track of the numbers of successfully scheduled packets, bursts and circuits, respectively. To evaluate the node performance the following outputs are produced by the simulator: input load, normalized throughput, packet loss rate, burst loss rate and circuit establishment failure rate. The input load and the normalized throughput (NT) are respectively calculated using the following formulas:

$$Input\,Load = \frac{A_{Counter}}{t_{Sim} \cdot N_S} \tag{2.15}$$

$$NT = \frac{S_{Counter}}{t_{Sim} \cdot N_S} \tag{2.16}$$

where t_{Sim} represents the total simulation time in seconds. The input load and normalized throughput will be expressed in percentage. To evaluate the packet loss P_{Loss}, burst loss B_{Loss} and circuit establishment failure $C_{Failure}$ rates, the following formulas have been used:

$$P_{Loss} = \frac{P_{Arrival} - P_{Scheduled}}{P_{Arrival}} \tag{2.17}$$

$$B_{Loss} = \frac{B_{Arrival} - B_{Scheduled}}{B_{Arrival}} \tag{2.18}$$

$$C_{Failure} = \frac{C_{Arrival} - C_{Scheduled}}{C_{Arrival}} \tag{2.19}$$

The results of the performance study have been obtained by varying one specific network parameter per time. The parameters that have been varied are: input load, traffic pattern, aggregate node capacity, and wavelength conversion

capacity. The traffic pattern is defined as the percentage of traffic carried respectively by packets, bursts, and circuits. The aggregate node capacity is the maximum amount of traffic that could be theoretically handled by the node and it is obtained by multiplying the line data rate and the number of input ports. The node is supposed to be equipped with one TWC per switch port, which is able to convert the input signal to a fixed number of the output wavelength channels. The ratio between the number of channels to which an input signal can be converted and the total number of available output channels represents the wavelength conversion capacity. The lower the wavelength conversion capacity the lower the cost and the power consumption of the node.

To better compare the considered node architectures we introduce the concept of increase in power efficiency. For a certain node's aggregate capacity (AC) the achievable throughput (AT) can be calculated through the following formula:

$$AT = NT \cdot AC \tag{2.20}$$

It is then possible to define the power efficiency P_{EFF} of the node as the ratio between the achievable throughput and the total power consumption:

$$P_{EFF} = \frac{AT}{P_{TOT}} \tag{2.21}$$

The increase in power efficiency between the all-optical hybrid and the optical/electronic hybrid architectures is then defined through the following formula:

$$IE_{O,OE} = \frac{P_{EFF}\,|_{All-optical} - P_{EFF}\,|_{Optical/electronic}}{P_{EFF}\,|_{Optical/electronic}} \tag{2.22}$$

and the increase in power efficiency between the optical/electronic hybrid and the all-electronic architectures is accordingly defined by:

$$IE_{OE,E} = \frac{P_{EFF}\,|_{Optical/electronic} - P_{EFF}\,|_{All-electronic}}{P_{EFF}\,|_{All-electronic}} \tag{2.23}$$

The main parameters used in simulations are reported in the following. The packet and burst lengths have been defined as random uniform integers in the range of $[1; 4]\,kBytes$ and $[50; 5000]\,kBytes$, respectively. Regarding the circuits, the number of TDM-slots in a frame has been fixed to 10 and the slot-length

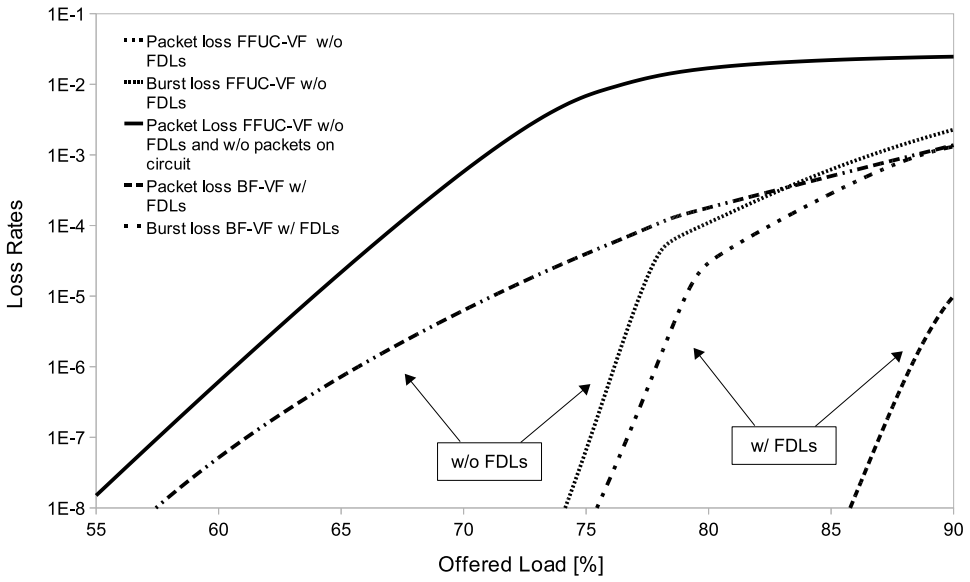

Figure 2.7: Packet and burst loss probabilities as a function of the offered load for two burst scheduling algorithms (FFUC-VF and BF-VF) as well as with and without FDL for contention resolution.

has been defined as a random uniform integer in the range of $[2;6]\,kBytes$. Due to the fact that the number of frames has been set to $5,000$, the total circuit duration results to be a random value among $20\,ms$, $30\,ms$, $40\,ms$, $50\,ms$ and $60\,ms$. The number of free TDM-slots in a frame has been defined as a random uniform integer in $[0;3]$. The burst-offset has been set as a random value in the range of $[0.2;2]\,ms$ and the circuit-offset has been set to $3\,ms$. The simulation time has been set to $1\,s$.

2.1.5 Numerical Results

This section reports some results of the performance and power consumption analysis.

Firstly the results as a function of the offered load will be shown. The aggregate node capacity has been set to $76.8\,Tbps$, corresponding to 24 input/output fibers with a capacity of $3.2\,Tbps$ each. The traffic generated at the sources include 40% of circuits, 30% of bursts, and 30% of packets. Additionally, full wavelength conversion has been assumed.

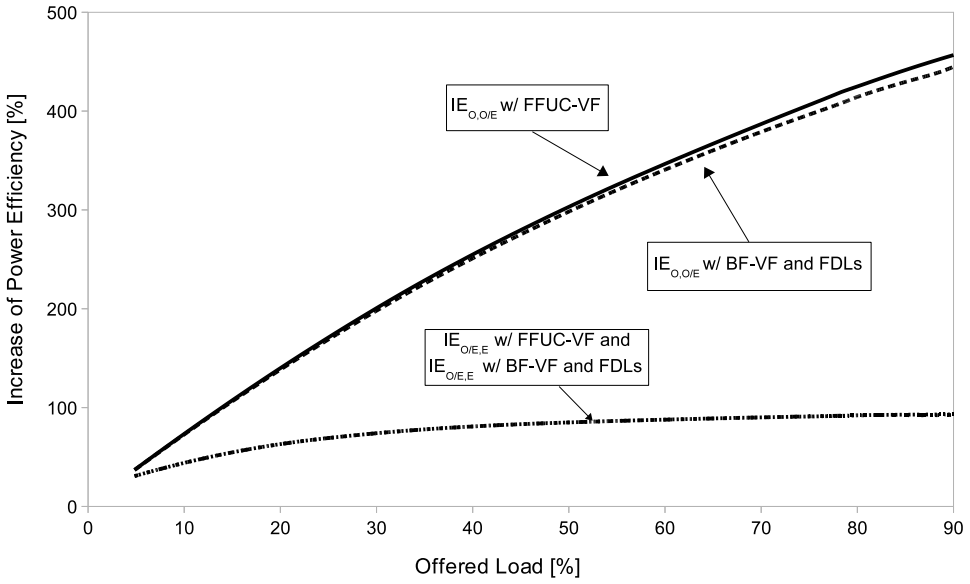

Figure 2.8: Increase of efficiency between all-optical hybrid and optical/electronic hybrid and increase of efficiency between optical/electronic hybrid and all-electronic architectures as a function of the input load.

Two different configurations have been analyzed. In the first one, the FFUC-VF algorithm is used for scheduling of bursts with no buffering space for packets contention resolution. In the second configuration, aiming to decrease the loss probabilities, we implemented a buffering space of $1\,kByte$ per switch port and we implemented the BF-VF algorithm instead of the FFUC-VF, for burst scheduling.

Figure 2.7 shows the packet and burst loss probabilities as a function of the input load. Consider first the packet and the burst loss probabilities with the FFUC-VF algorithm. The Figure shows that both curves remain below $1\,E^{-8}$ for offered loads under 55%. For loads between 55% and 70% the packet loss increases from $1\,E^{-8}$ to $1\,E^{-5}$, whereas the burst loss rate is still lower than $1\,E^{-8}$. From 70% to over 90% of offered load the packet loss increases further from $1\,E^{-5}$ to $1\,E^{-3}$. Simultaneously, the burst loss increases very fast from $1\,E^{-8}$ to $1\,E^{-3}$, becoming higher than the packet loss for loads higher than 90%.

This effect can be explained as follows. When the load is very high, a large percentage of the available output resources is occupied by circuits, which have the highest priority. Because packets can be scheduled within free-slots of already

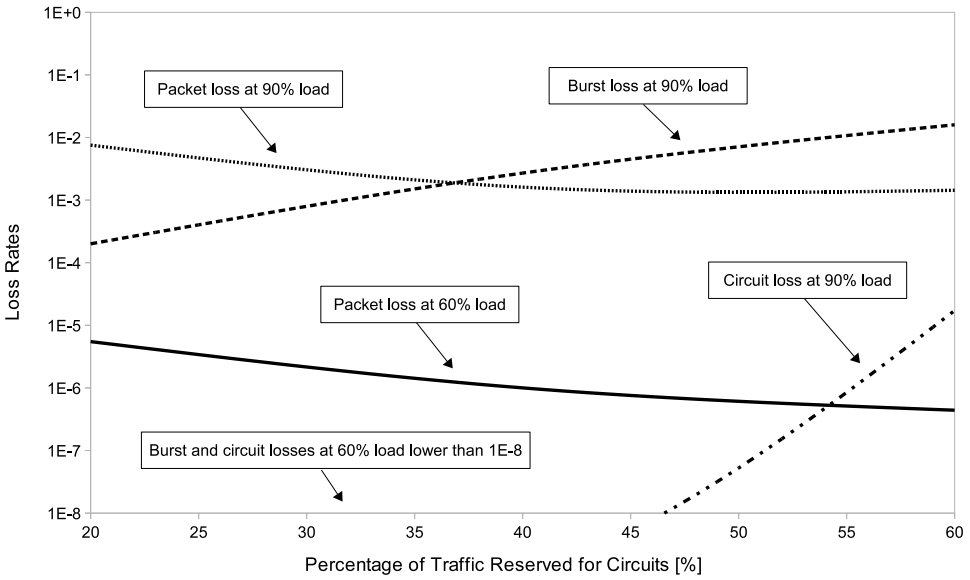

Figure 2.9: Loss probabilities as a function of the percentage of traffic reserved for circuits. The FFUC-VF algorithm is considered for burst scheduling with no FDLs for packet contention resolution.

established circuits, the packet loss does not increase too rapidly at high loads. On the contrary, bursts suffer the increased number of circuits showing a rapidly increasing loss for very high input loads.

As confirmation of this trend, Figure 2.7 reports also the packet loss rate in case that optical packets cannot be scheduled into unused slots of TDM-circuits. In this case, we have that the three switching schemes share the same resources but act independently. The Figure shows that the packet loss rate is significantly higher than in the previous case, highlighting that employing a unique integrated control plane leads to better overall performance with respect to three separated control planes.

Implementing the BF-VF algorithm and $1\,kByte$ of buffering space per switch port, the burst loss rate shows a slight decrease and remains lower than $1\,E^{-8}$ until almost 75% of offered load. On the other hand, the packet loss is drastically reduced, remaining lower than $1\,E^{-8}$ until 85% of offered load and not higher than $1\,E^{-5}$ for loads higher than 90%.

A final note regards the circuit establishment failure rate, which is lower than $1\,E^{-8}$ for every value of the offered load.

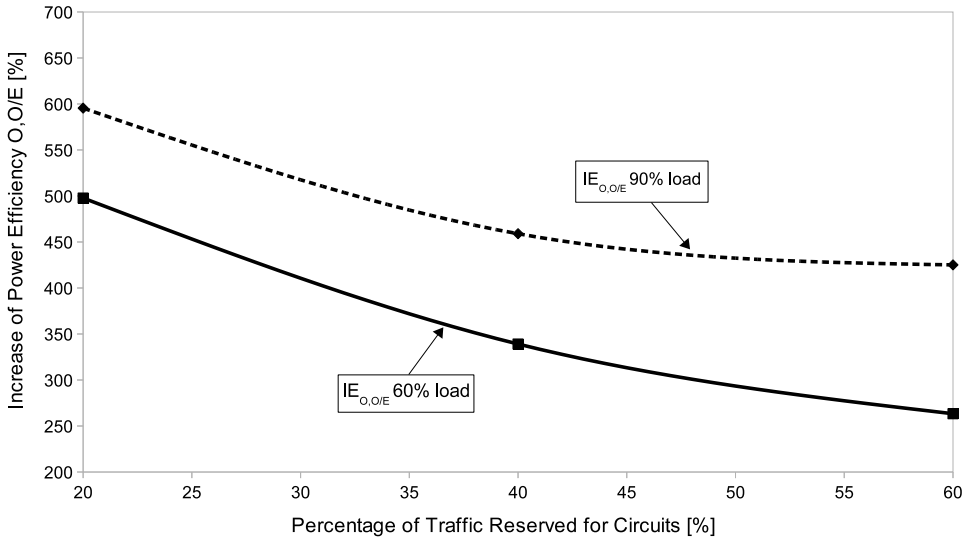

Figure 2.10: Increase in efficiency between all-optical hybrid and optical/electronic hybrid architectures as a function of the traffic pattern.

Figure 2.8 reports $IE_{O,O/E}$ and $IE_{O/E,E}$ as a function of the offered load. The curves show that $IE_{O,O/E}$ increases with the load reaching almost 460% for loads higher than 90%. The curve obtained with BF-VF and FDLs presents a lower slope with respect to the one obtained with FFUC-VF algorithm. This is due to the fact that in the FDL feed-back architecture utilized here the introduction of FDLs leads to an increased size of the SOA-based switch and, consequently, the power consumption per port becomes larger. As a consequence, FDLs reduce the advantage of using a SOA-based switch instead of the electronic one. Also $IE_{O/E,E}$ increases with the offered load, but presents a lower slope with respect to $IE_{O,O/E}$, remaining lower than 100% even for loads higher than 90%.

To evaluate the results as a function of the traffic pattern the following assumptions have been made. The node capacity has been set to $76.8\,Tbps$ and two values of offered load have been considered, namely at 60% and 90%. The FFUC-VF algorithm has been used for scheduling of bursts. Full wavelength conversion and no FDLs for packet contention resolution were assumed.

In order to show the effects of changing the traffic pattern, the percentage of traffic carried by circuits has been assumed as the reference parameter. Three different traffic patterns have been analyzed. In the first pattern, 20% of input

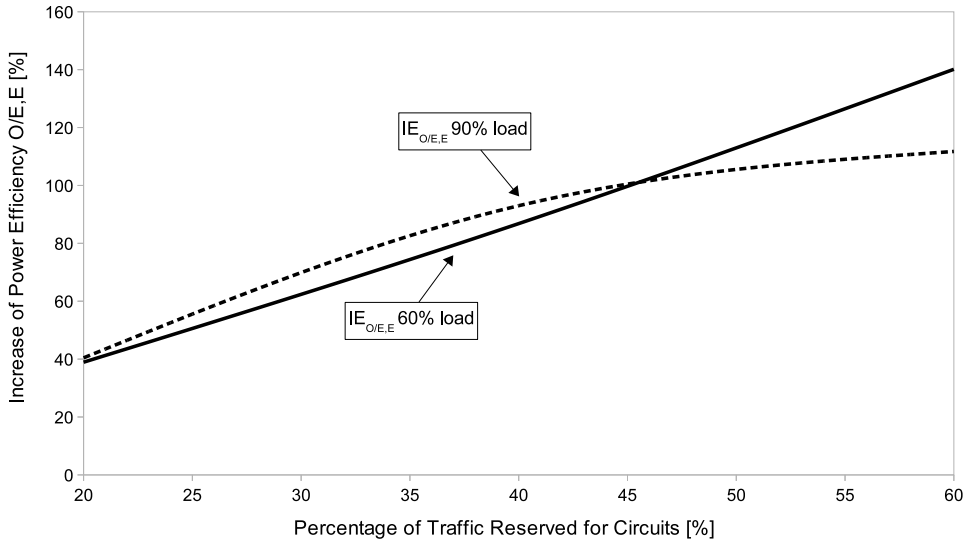

Figure 2.11: Increase in efficiency between optical/electronic hybrid and all-electronic architectures as a function of the traffic pattern.

traffic is carried by circuits, 40% by bursts and 40% by packets. In the second pattern, we assume that 40% of input traffic is carried by circuits, 30% by bursts and 30% by packets. Finally, a pattern composed by 60% circuits, 20% bursts and 20% packets was considered.

Figure 2.9 shows the loss probabilities as a function of the traffic pattern. The packet loss decreases while increasing the percentage of incoming traffic reserved for circuits, whereas the burst loss increases. The explanation is the following. While increasing the circuit traffic, the rate for optical packets to be scheduled in an established circuit increases leading to a lower packet loss rate. Instead, when many output wavelengths are occupied by circuits it becomes more difficult for bursts to find an available output resource.

Consequently, at high loads and high percentage of circuit traffic, the burst loss becomes appreciably higher than the packet loss. The burst loss rate at 60% of input load is always lower than $1\,E^{-8}$, showing that at lower loads bursts are advantaged with respect to packets independently from the percentage of circuit traffic. These results confirm the trends already shown in Figure 2.7.

Finally, the circuit establishment failure rate increases while increasing the circuit traffic. As shown in Figure, the circuit loss at 90% of offered load slightly exceeds $1\,E^{-5}$ when the 60% of the input traffic is reserved for circuits.

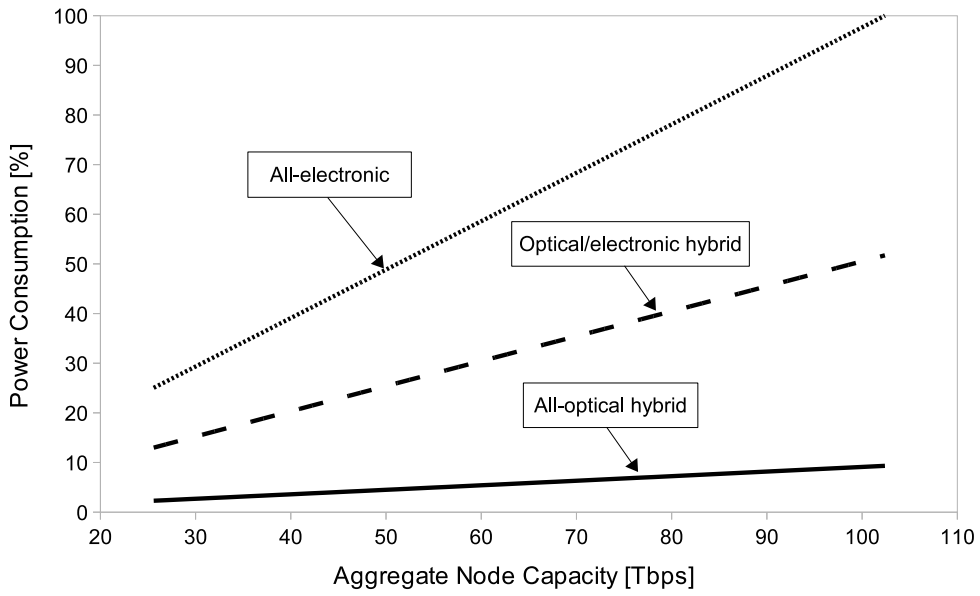

Figure 2.12: Power consumption of the considered node architectures as a function of the aggregate node capacity.

In Figure 2.10, $IE_{O,O/E}$ as a function of the analyzed traffic patterns is shown. The graph shows that $IE_{O,O/E}$ decreases while increasing the percentage of traffic carried by circuits. To explain this, consider that the higher the percentage of circuit traffic the higher the number of slow ports used with respect to fast ports, which leads to a lower $IE_{O,O/E}$. Furthermore, the Figure shows that the curve taken at 60% of the input load is lower and decreases faster with respect to that taken at 90%. The lower slope is due to the fact that at higher loads some circuit establishment requests are refused because the maximum number of circuits is reached. As a consequence, increasing the percentage of traffic carried by circuits does not increase strongly the number of slow ports used and thus $IE_{O,O/E}$ presents a lower slope.

Figure 2.11 reports $IE_{O/E,E}$ for the different considered traffic patterns. In this case, the gain in efficiency increases while increasing the circuit traffic because the higher the number of slow ports used the higher the advantage of implementing a slow switch based on MEMS. Also, the Figure shows that at 60% of offered load the curve is almost linear while at 90% the slope is lower because some circuit establishment requests are rejected.

In Figure 2.12, the relative power consumption of the considered architectures

is shown as a function of the aggregate node capacity. The power consumption values are normalized with respect to the power consumption of the all-electronic architecture at $102.4\,Tbps$, which has been computed to be $765\,kW$. The Figure shows the advantage of using optical technologies. Using the optical/electronic hybrid architecture instead of the all-electronic one, the overall power consumption at $102.4\,Tbps$ is almost halved, while using the all-optical hybrid architecture the power consumption is reduced to 10% of the total.

2.1.6 Conclusions

In this Section a novel forwarding plane for HOS core nodes has been proposed, which is able to provide an efficient handling of data packets, bursts, and circuits directly in the optical domain. The control plane makes use of a unified control packet and employs a suitable scheduling algorithm for each incoming data type. Two different node architectures have been defined to be managed by this control plane: an all-optical hybrid architecture and an optical/electronic hybrid architecture. Furthermore the architecture of a traditional electronic packet switching node, namely, all-electronic architecture, has been modeled and considered as a reference. A simulation model has been developed to evaluate both the performance of the nodes and their power consumption. The model has been used to assess the considered architectures in terms of their potential for increased power efficiency.

The performance analysis shows that the packet loss rate remains lower than $1\,E^{-8}$ up to almost 60% of offered load, while the burst loss rate is lower than $1\,E^{-8}$ for almost 75% of the load. Furthermore, the packet loss decreases while increasing the percentage of traffic reserved for circuits. This is due to the possibility of scheduling optical packets on suitable unused TDM-slots in already established circuits. On the contrary, the burst loss increases while increasing the circuit traffic. Finally, it is highlighted that, employing FDL for packets contention resolution implies a significant increase in the node performance, and using a more efficient algorithm for burst scheduling (like BF-VF) reduces the burst loss.

The power consumption analysis has shown that the increase in power efficiency when using an all-optical hybrid instead of the optical/electronic hybrid architecture can be as high as 600% (with 60% of traffic carried by circuits),

but it decreases to about 420% while increasing the percentage of traffic carried by circuits. On the other side, the increase in power efficiency between optical/electronic hybrid and all-electronic architectures increases while increasing the circuit traffic, reaching almost 140% when the 60% of traffic is carried by circuits.

In general the results of this Section show that HOS has the potential of achieving high performance while reducing the energy consumption of traditional core nodes. In the next Section the HOS paradigm will be integrated with the GMPLS control plane and tested on a network composed of a cascade of core nodes.

2.2 HOS Core Network Employing GMPLS

In Section 2.1 we demonstrated that HOS has the potential to optimize the overall network design and to allow considerable energy savings and improved node scalability. However, a single core node has been taken into consideration so far. To include network operation and management functions, in this Section we propose the integration of our HOS forwarding plane with the GMPLS control plane [23].

GMPLS is widely recognized as a suitable control plane solution for optical networks since it provides efficient end-to-end provisioning, fast forwarding and traffic engineering (TE) decisions. Furthermore, the GMPLS control plane ensures multi-domain and multi-technology interoperability.

To the best of our knowledge, this is the first attempt of integration between HOS and GMPLS control planes, even if several proposals for the integration of the GMPLS control plane with the OBS control plane have been made [58–61]. In [58] an integrated approach, namely labeled optical burst switching (LOBS), is proposed while in [59–61] overlay solutions are considered. The overlay model proposed in [59, 60] is composed of three layers: the GMPLS control layer, the OBS control layer and the OBS data layer. The GMPLS control layer is responsible for GMPLS TE tunnels establishment and maintenance, but it does not entail resource reservation, which is delegated to the OBS control plane. The authors in [61] propose a model composed of a GMPLS control plane, a path computation engine (PCE) layer and a OBS transport (OBST) layer.

The integration of the HOS forwarding plane and the GMPLS control plane is realized using an overlay network model. The models and results presented in this Section are part of the works presented in [62–65]. In Section 2.2.1, we describe the proposed network overlay model. Section 2.2.2 proposes a possible implementation of the proposed model using as a reference architecture the optical/electronic architecture presented in Section 2.1. In Section 2.2.3, the simulator and the metrics used for the analysis of a core network composed of a cascade of optical/electronic nodes are introduced. Finally, in Section 2.2.4 numerical results are shown and discussed, and in Section 2.2.5 the conclusions of this work are drawn.

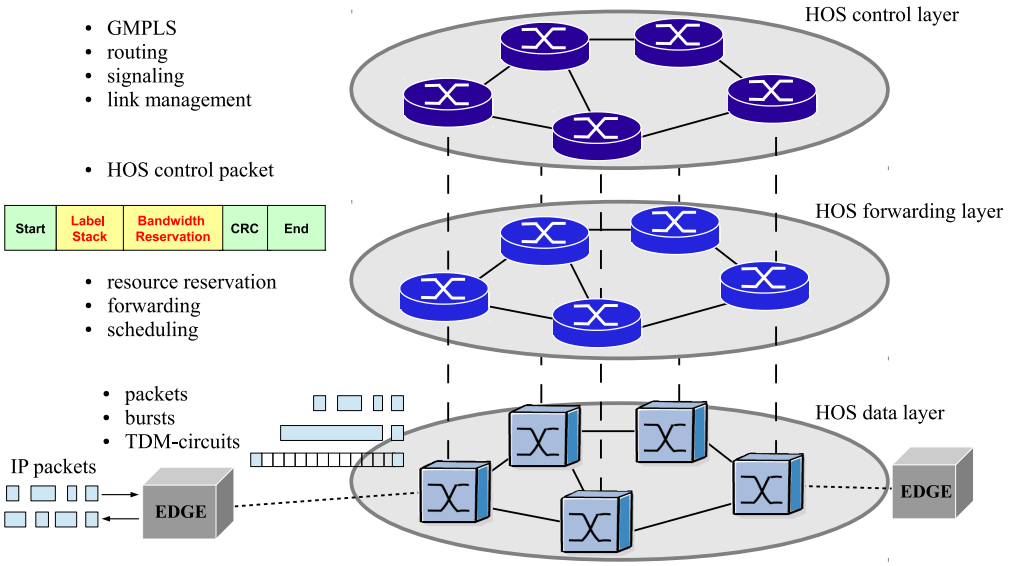

Figure 2.13: HOS network overlay model.

2.2.1 Network Overlay Model

The proposed architecture of a HOS network is divided into three separated layers and can be represented with an overlay model, as depicted in Figure 2.13. At the highest layer there is the GMPLS control plane that is in charge of configuring the virtual topology and setting up and tearing down the Label Switching Paths (LSP). The GMPLS control layer makes use of a dedicated network and is physically and logically separated from the layers below. The intermediate layer is given by the HOS control plane. The HOS control plane carries the information for properly scheduling and forwarding different data types. The HOS control plane shares the same network as the HOS data plane, but it is logically divided. The control information are carried over a hybrid optical/electronic network, i.e., it is O/E/O converted at each node along the path toward the destination. Finally, the HOS data plane is an optical network able to support different traffic granularities, i.e., circuits, bursts and packets.

2.2.1.1 GMPLS Control Layer

The GMPLS control plane consists of several building blocks. The building blocks include routing, signaling and link management.

The task of the routing block is to distribute and maintain the network topology and information about the resource usage. Standard IP routing protocols, such as Open Shortest Path First (OSPF) or Intermediate System-to-Intermediate System (IS-IS) with GMPLS-TE extensions, can be used to reliably exchange the information. The extensions to OSPF and IS-IS add additional information about links and nodes into the link-state database. The OSPF-TE for instance determines the link usage in an aggregated way through using bandwidth units (bit/s) and flood the information using TE link state advertisement (LSA).

The information collected by the routing block is used by the signaling block to establish LSPs. In the proposed HOS network we assume that a link between two adjacent nodes is composed of a group of fibers. A LSP is defined as a sequence of fiber links along the path connecting source and destination nodes where data are forwarded using label switching. As a consequence, an entry into the switching table of a node along the path is a correspondence between the input label and fiber, and the output label and fiber. In traditional GMPLS implementations the signaling block is in charge of exchange control information among nodes, to distribute labels, and to reserve resources along the path. To this aim, two signaling protocols have been extended, i.e., Resource Reservation Protocol TE (RSVP-TE) and Constraint based Label Distribution Protocol (CR-LDP). However, in the proposed HOS network we assume that the resource reservation is delegated to the HOS control plane. As a consequence, the GMPLS signaling block is responsible only for label distribution and management. In this context, the RSVP-TE and CR-LDP protocols can be used to set up, tear down and maintain LSPs, but without performing resource reservation.

Finally, the link management block is based on the link management protocol (LMP). LMP runs between adjacent systems for link provisioning and fault isolation and is used to automatically generating and maintaining associations between links and labels for use in label swapping.

2.2.1.2 Modified HOS Forwarding Layer

The HOS forwarding layer is responsible for managing the transmission of circuits, bursts and packets along the LSPs. Once LSPs have been established by the GMPLS control layer, it is the HOS control layer that schedules data

on specific wavelengths along the LSPs and makes the resource reservations. A specific scheduling algorithm and a specific reservation mechanism for each data type is used. The proposed HOS forwarding plane has been described in 2.1. In the following its most important features will be reported, together with some changes introduced to facilitate its integration with GMPLS.

The control plane makes use of TDM circuits. In a TDM-circuit, time is divided in frames, each of which is divided in a fixed number of time-slots. The nodes along the circuit path can fill unused TDM-slots of a circuit with optical packets of suitable length that have the same destination as the circuit. In GMPLS, a packet and a circuit have the same destination if they belong to the same Forwarding Equivalence Class (FEC) and thus they share the same LSP. This technique aims to increase the bandwidth utilization for two reasons: firstly, because the circuit utilization is raised and, secondly, because packets scheduled within a circuit do not consume new resources.

A unified control packet has been defined to manage the three data types. The format of this control packet is depicted in Figure 2.14 and comprises two main fields, namely label stack and bandwidth reservation. The label stack field is used to carry the GMPLS labels and substitutes the routing information field proposed in 2.1, which contained the source and destination IP addresses. This field is able to carry up to 4 labels of 32-bits each, providing the capability of nesting up to 4 LSPs. The bandwidth reservation field contains the information required by the core nodes for deploying the data scheduling. In particular, the type subfield is used to distinguish among circuits, bursts and packets. The length and offset subfields carry the data length and offset-time respectively. The free-slot and slot-position carry the information about the number and position of unused TDM-slots in a circuit. The control packet is assumed to be encoded together with the corresponding payload on the same optical carrier, using the optical SCM technique [53].

The control plane uses appropriate scheduling algorithm for each incoming data type. The highest priority is given to circuits, which are scheduled on the first output wavelength in which no other circuit has been previously scheduled. However, if too many circuits are scheduled on the same output fiber, the burst and packet traffic on that fiber would be almost completely blocked. For this reason, the maximum number of circuits that can be simultaneously scheduled on the same fiber has been limited to 60% of the number of wavelengths. As regards

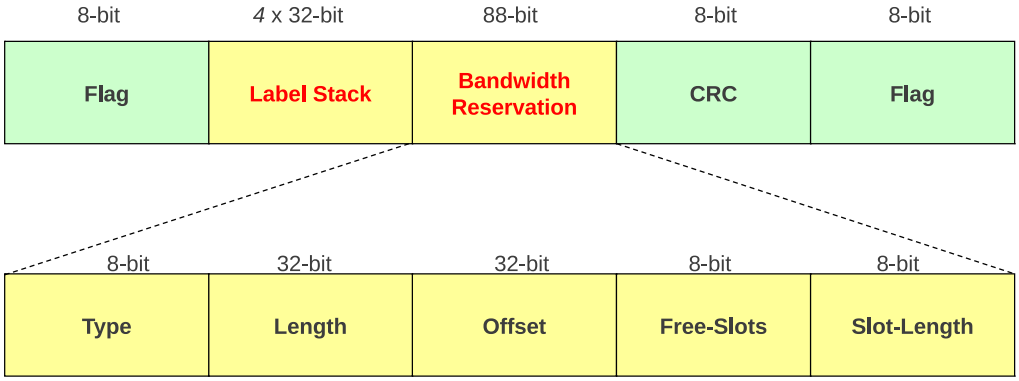

Figure 2.14: Modified control packet format for the integration of GMPLS and HOS control layers.

the scheduling of bursts, the FFUC-VF algorithm is used because it provides a good bandwidth utilization while introducing a relatively small processing time. Packets can be scheduled either on an unused TDM-slot of a circuit that belongs to the same FEC, or on the first output wavelength for which no reservation has been made for the time required for transmission.

2.2.2 Architectures and Power Consumption

In this Section, we extend the optical/electronic hybrid architecture, presented in 2.1, to include the GMPLS control plane. Furthermore, we derive a model for the evaluation of the power consumption introduced by the GMPLS control plane and compare the power consumption of the optical/electronic hybrid architecture with that of a current electronic IP router.

2.2.2.1 Node and Control Plane Architectures

The optical/electronic hybrid architecture is depicted Figure 2.15. It can be divided into three building blocks: the switching fabric, the control logic and the other active components.

The switching fabric is divided into a slow switch and a fast switch. The slow switch is realized using MEMS, whose switching time is in the order of milliseconds or hundreds of microseconds, and it is used to forward circuits and long bursts. The fast switch is made up by electronic LC and electronic switching

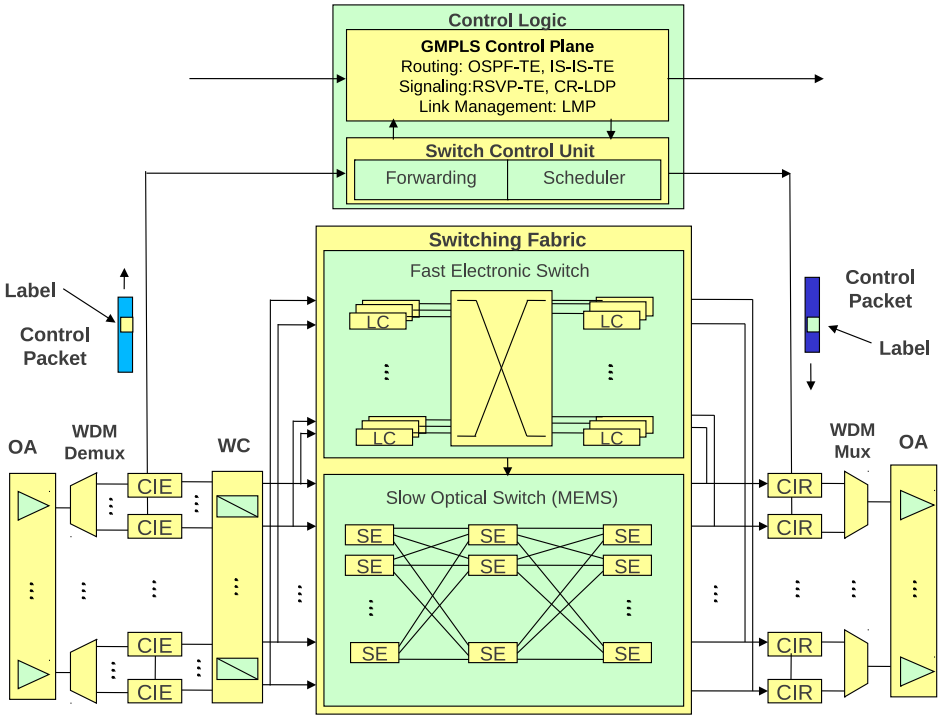

Figure 2.15: Extended optical/electronic HOS architecture for integration with the GMPLS control plane.

elements, whose switching time is in the order of nanoseconds, and it is used to forward packets and short bursts. As a consequence, circuits and long bursts are transmitted transparently, while packets and short bursts are O/E/O converted at each node along the path.

The electronic control logic implements the GMPLS and HOS control planes. The structure of the control logic is shown in Fig. 2.16, and it comprises several building blocks: the GMPLS routing, signaling and link management block, the label processing, table lookup and forwarding block, the data type check, for the distinction of the different data types, and the packet, burst and circuit schedulers.

Finally, in the other active components we include: OA, TWC, CIE/R, switch control and management cards.

The optical/electronic hybrid architecture is compared with a traditional electronic IP router. The architecture of a current IP router can be also divided into three building blocks: the switching fabric, the route processor and

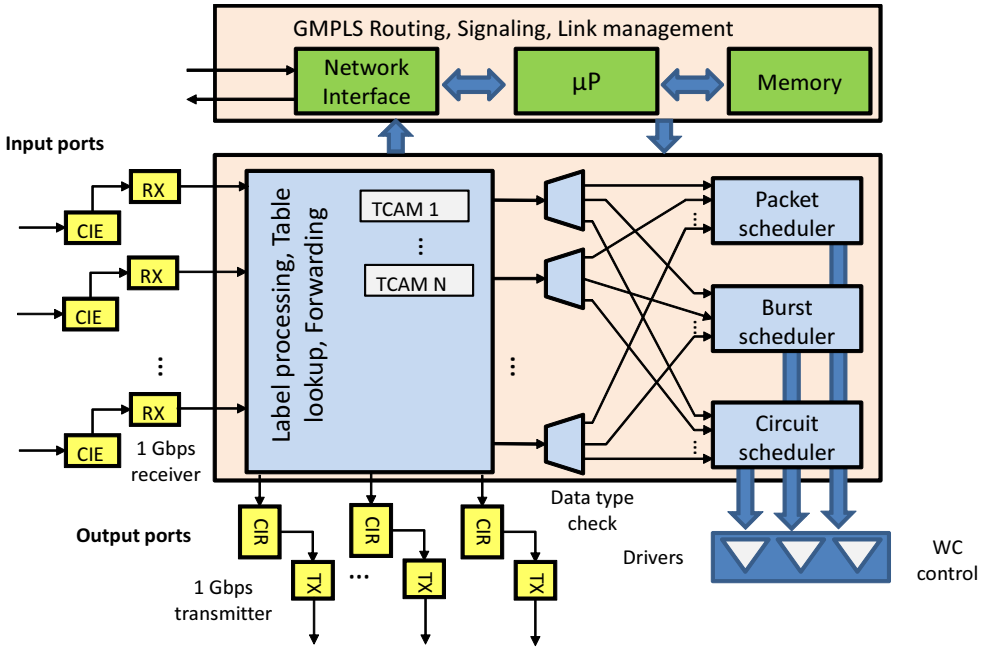

Figure 2.16: Block diagram of the control logic (RX: Receiver, TX. Transmitter, μP: Microprocessor, WC: Wavelength Controller, CIE: Control Information Extraction, CIR: Control Information Reinsertion).

the other active components. The switching fabric is composed of a large fast electronic switch, made up by electronic switching elements. The large electronic switching fabric interconnects a large number of line interface cards (LC). The route processor deploys the MPLS control plane and data forwarding functionalities. The other active components include the OAs, the switch control and the management cards. The TWC and CIE/R are not included here since all the data are O/E/O converted.

2.2.2.2 Power Consumption Model

The total power consumption of a core node P_{TOT} is defined as the total power consumed by the active elements within the node and it is derived through the following formula:

$$P_{TOT} = P_{SF} + P_{CL} + P_{OAC}, \tag{2.24}$$

where P_{SF} is the power consumption of the switching fabric, P_{CL} is the power consumption of the control logic and P_{OAC} is the power consumption

of the other active components. In this paper we make use of the analytical model introduced in [10,19] to evaluate P_{SF} and P_{OAC}. Note that we assume all components that are inactive for a period of time are switched off. Furthermore, we estimate P_{CL} as follows.

Let N be the number of input/output fibers and W be the number of wavelengths per fiber. The power consumption of the control logic is computed through the following formula:

$$P_{CL} = P_{GMPLS} + P_{Scheduler} + N \cdot W \cdot P_{Transceiver}, \tag{2.25}$$

where P_{GMPLS} is the power consumption of the GMPLS control plane, $P_{Scheduler}$ is the power consumption of the scheduler and $P_{Transceiver}$ is the power consumption of a DWDM long reach transceiver used for receiving/transmitting the control information. The power consumption of the GMPLS control plane is in turn obtained through the following formula:

$$P_{GMPLS} = P_{Offline} + N \cdot P_{Online} + N \cdot W \cdot P_{SearchEngine}, \tag{2.26}$$

where $P_{Offline}$ is the power consumption required for routing, signaling and link management. P_{Online} is the power consumption introduced by label processing, table lookup and forwarding. $P_{SearchEngine}$ is the power consumption of a search engine used for table lookup.

Finally, the power consumption of the scheduler $P_{Scheduler}$ is given by the sum of the power consumption of the packet, burst and circuit schedulers. The values for the power consumption of the above mentioned elements are estimated by collecting data of a number of commercially available components and modules for conventional large-scale switching and routing systems. The numbers given in Table 2.3 are obtained by averaging and rounding the power consumption values of modules that implement the same or a similar functionality.

Table 2.3: Power consumption of the control logic

Function	**Component**	**Power**
Transceiver	1 Gbit/s DWDM long reach transceivers	1.25 W
GMPLS off-line processing	Microprocessor, DRAM, network interface	150 W
GMPLS on-line processing	Large programmable logic device (FPGA)	40 W
Search engine	Ternary content addressable memory (TCAM) chip	4.5 W
Packet scheduler	Large programmable logic device (FPGA)	40 W
Burst scheduler	Large programmable logic device (FPGA)	40 W
Circuit scheduler	Large programmable logic device (FPGA)	40 W

As regards the electronic IP router, the total power consumption is obtained through the following formula:

$$P_{TOT} = P_{SF} + P_{RP} + P_{OAC}, \tag{2.27}$$

where P_{RP} is the power consumption of the route processor. It is worth noting that in this case the power consumption of the others active components, P_{OAC}, does not include the power consumption of TWCs and CIE/R, which are not used in the electronic router. The term P_{RP} is obtained by summing up the power consumption of all the route processor cards P_{RPC} implemented within the node. We assume the use of one route processor card each 16 wavelength ports, and consequently the power consumption of the route processor is given by the following formula:

$$P_{RP} = \frac{N \cdot W}{16} \cdot P_{RPC}. \tag{2.28}$$

The power consumption of a route processor card comprising the MPLS control plane functionalities has been estimated to be $P_{RPC} = 200\,W$.

2.2.3 Core Network Simulation Setup

The performance and power consumption of the proposed HOS network have been assessed by means of an event-driven simulator, developed specifically for this scope. The simulated network is depicted in Figure 2.17, and consists of a cascade of core nodes connected by a group of fibers. At each node along

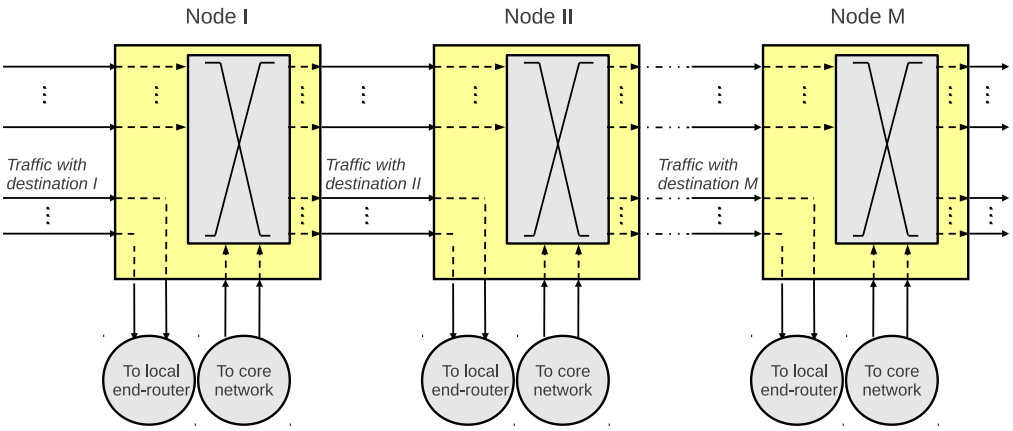

Figure 2.17: Core network simulation setup.

the cascade, traffic addressed to this particular node is dropped and sent to a local electronic end-router. At the same time, new traffic is generated by the node and injected into the core network. Data are forwarded by the nodes on a proper output fiber according to their destination addresses and employing the appropriate scheduling algorithm.

To analyze the proposed HOS core network we use the same metrics that were introduced in 2.1.4. Specifically, we evaluate the network performance in terms of data loss rates, which are defined as follows. The packet (burst) loss rate is defined as the ratio between the number of dropped packets (bursts) and the number generated packets (bursts). Finally, the circuit establishment failure probability is defined as the ratio between the number of negative-acknowledged and the number of generated circuit establishment requests. As regards the energy consumption, to better compare the considered architectures, we employ the definition of increase in power efficiency between the optical/electronic hybrid and an all-electronic architecture, i.e. typical architecture of a current IP router ($IE_{O/E,E}$).

The main parameters used in simulations are reported in the following. The packet and burst lengths have been defined as random uniform integers in the range of $[1;4]\,kBytes$ and $[50;5000]\,kBytes$, respectively. Regarding the circuits, the number of TDM-slots in a frame has been fixed to 10 and the slot-length has been set to $4\,kBytes$. The number of free TDM-slots in a frame has been defined as a random uniform integer in $[1;4]$. The total circuit duration is a random value among $4\,ms$, $6\,ms$, $8\,ms$, $10\,ms$ and $12\,ms$.

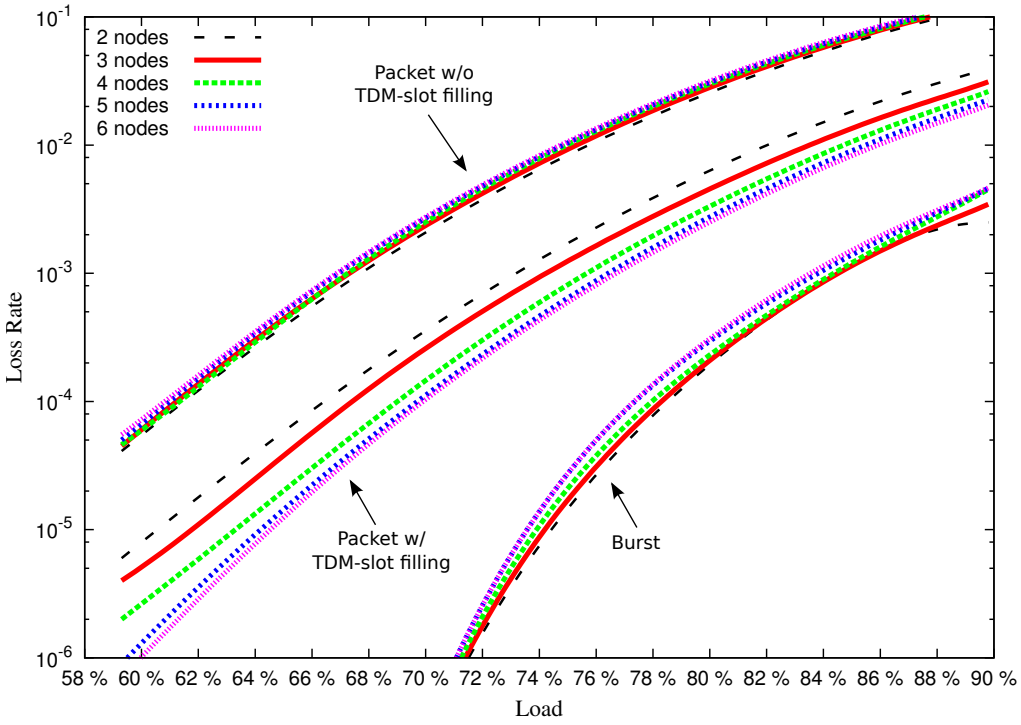

Figure 2.18: Packet and loss burst rates as a function of the input load and for different numbers of nodes composing the core network.

2.2.4 Numerical Results

This section reports some selected results of the performance and power consumption analysis. The number of input/output fibers of each node along the cascade has been set to 24, with 80 wavelengths per fiber. The line data rate has been set to $40\,Gbps$, leading to an aggregate capacity of $76.8\,Tbps$ per node. At each node along the cascade a half of the incoming traffic is dropped and sent to the local end-router. Concurrently, the same amount of traffic is generated by the node and inserted into the network. The traffic pattern includes 40% of circuits, 30% of bursts and 30% of packets.

In Figure 2.18, the packet and burst loss rates are shown in regard to the input load and for different numbers of nodes composing the cascade. The number of cascaded nodes has been varied from 2 to 6. The figure shows that the burst loss rate is lower than 10^{-6} for input loads below 70%. For higher loads the burst loss rates increase rapidly, reaching values in the order of 10^{-3} for input loads close to 90%. The packet loss rates are always higher than the burst loss rates,

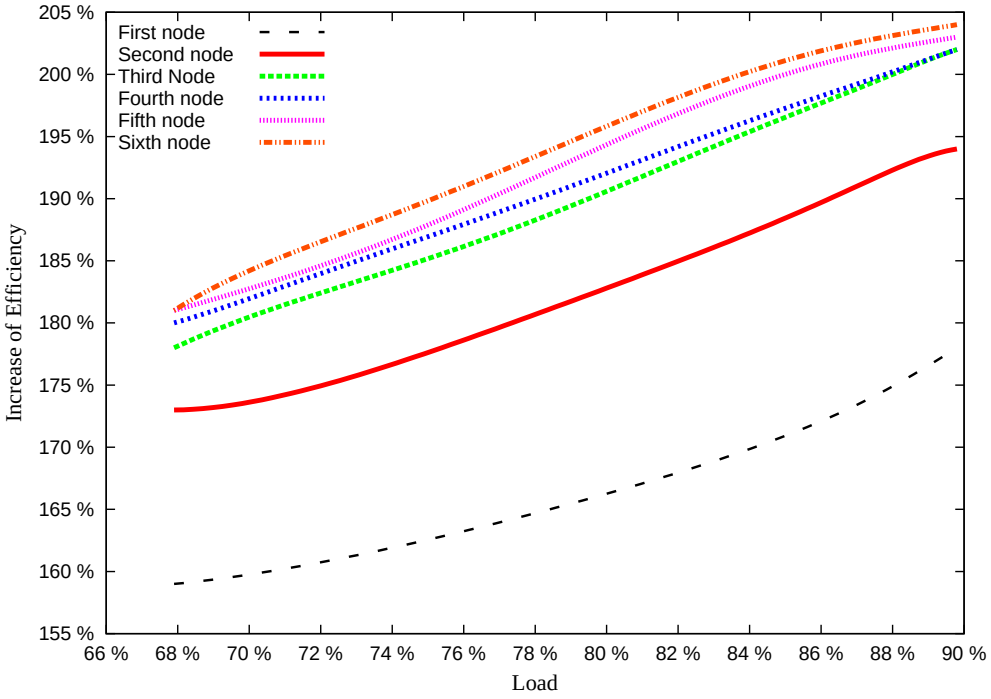

Figure 2.19: Improvement in power efficiency between the optical/electronic HOS core network and current network based on IP routers.

and they reach values in the order of 10^{-2} for loads close to 90%. Bursts are scheduled a priory due to the offset-time and this results in a kind of prioritized handling in comparison to packets. However, at high loads the burst loss rates increase very fast because a large portion of available resources is occupied by circuits which have the highest priority. On the contrary, the packet loss rates do not increase very fast at high loads because packets can be scheduled into unused TDM-slots of circuits. A final note regards the circuit establishment failure rates, which are always null in the considered configurations.

Figure 2.18 also shows that the number of nodes in the cascade has no significant influence on the packet loss rates. This again is due to the fact that packets can be inserted into unused TDM-slots of circuits. The simulations demonstrate that the higher is the number of nodes in the cascade and the higher is the probability for a packet to find a circuit with the same destination that has a free TDM-slot. To prove this fact, in Figure 2.18 also the packet loss rates without the possibility for packets to fill unused TDM-slots are shown. The Figure relates that the packet loss rates without the possibility of filling unused TDM-

slots are one order of magnitude higher reaching values in the order of 10^{-1} for loads close to 90%. Furthermore, in this case the packet loss rates increase significantly while increasing the number of nodes in the cascade. We can then conclude that the possibility for packets to be transmitted in unused TDM-slots of already established circuits leads to improved overall performance.

Figure 2.19 reports the $IE_{O/E,E}$, in percentage, as a function of the input load. The number of nodes in the cascade has been set to 6 and the improve of efficiency of each node is plotted separately. The $IE_{O/E,E}$ depends largely on the number of slow ports used. In fact, the optical/electronic hybrid architecture employs MEMS switches for forwarding circuits and short bursts and it is well known that optical MEMS switches have a much lower power consumption than conventional electronic switches. The higher is the input load and the higher is the number of slow ports used, and in turn the higher is $IE_{O/E,E}$. In addition, the figure shows that the last node in the cascade is the one that presents the highest gain in efficiency. This is due to the fact that at the end of the cascade a large portion of the initially free TDM slots are already occupied by packets sent by the preceding nodes, which leads to more active slow ports and a higher utilization of circuits in the last node.

2.2.5 Conclusions

In this Section we discussed the integration between the HOS forwarding plane introduced in Section 2.1 and the GMPLS control plane. The proposed HOS core network is organized in an overlay model that is composed of three layers: the GMPLS control layer, the HOS control layer and the HOS data layer. The GMPLS is in charge of configuring the virtual topology and setting up and tearing down the LSPs. The HOS control plane makes the resource reservation and performs the scheduling and forwarding functionalities for the different supported traffic granularities.

The optical/electronic hybrid HOS node architectures analyzed in 2.1 has been extended to include the GMPLS control plane and the power consumption of the GMPLS module has been estimated. The performance and power consumption analysis of the HOS network has been carried out through the use of an event driven simulator. Results show that our control plane, capable of inserting packets onto unused TDM-slots of circuits, provides improved performance. Furthermore, the results show that the deployment of optical switching

elements instead of electronic elements enables an increase of the network energy efficiency.

In the next Section we will include in our analysis also the HOS edge nodes, i.e., the nodes that provide the connectivity between the HOS core network and external networks. We will study their architecture and evaluate their impact on performance and energy consumption.

2.3 HOS Edge Network and QoS

Considerable research [7,66] has already highlighted that core networks may ultimately be constrained not by the capacity of optical and electronic technologies, but rather by their energy consumption and capability to provide QoS. In fact, an effective control plane that is able to exploit the huge capacity offered by the optical WDM layer in order to provide QoS in an energy-efficient manner, is still missing.

On the one hand, traditional electronic packet switching core networks provide the possibility to implement advanced QoS policies, that are usually based on traffic classification and Per-Hop Behaviors (PHB) like in the *DiffServ* model [67, 68]. However, as we already demonstrated in Section 2.1 and 2.2, electronic switching networks are characterized by low energy efficiency and are not able to support the future Internet growth in a sustainable manner. On the other hand, optical switching core networks have the potential to provide high energy efficiency, but mainly due to the lack of effective optical buffering techniques, are not able to support advanced QoS policies.

The HOS network proposed in Section 2.1 and extended in Section 2.2 has the potential of providing both high energy efficiency and effective QoS differentiation. In fact, the use of optical bursts in combination with packets and circuits enables for the dynamic implementation of different service classes, thus adapting the network to the current traffic characteristics and leading to an efficient QoS differentiation.

However, so far we only considered the HOS core network and thus we neglected the architecture of the HOS edge network and the implementation of specific QoS policies. The HOS edge network is responsible for providing connectivity toward external networks and applying QoS policies on incoming and outgoing data. In this Section, we extend our analysis by introducing a novel HOS edge node architecture and evaluating its impact on network performance and energy efficiency. Furthermore, we define four service classes for our HOS network and map a possible set of future services into these classes. We will prove that HOS has the potential of providing the QoS differentiation needed by the future Internet applications, while providing a higher energy efficiency with respect to electronic switching solutions.

The rest of the Section is organized as follows. Firstly, we describe the QoS classes implemented in the HOS network in Section 2.3.1. Secondly, in Section 2.3.2 we present the HOS edge node architecture. Then, in Section 2.3.3 the analytical model employed for the power consumption evaluation of the HOS edge network is introduced. In Section 2.3.4 we define the simulation model and the performance metrics, and in Section 2.3.5 we present results of performance and energy efficiency. Finally, Section 2.3.6 draws conclusions. The results shown in this Section are part of the work presented in [69].

2.3.1 Quality of Service

Future networks will be required to support a wide range of services and applications with different requirements in term of QoS. Any attempt to classify future services into a limited number of typical descriptions of traffic characteristics and QoS demands can never be exhaustive. However, a rough classification of future services can be provided taking into consideration their tolerance to loss, delay and jitter. Basing on these metrics, in [70], the user/subscriber traffic has been classified in ten classes of service and a standard mapping between each class and a Differentiated Code Point (DSCP) value has been proposed. The DSCP is a 6-bit field that is used to perform packet classification in *DiffServ* [67,68] and replaces the type of service (TOS) field in the IPv4 header or the traffic class field in the IPv6 header. In *DiffServ* a different Per-Hop Behavior (PHB), i.e. a different forwarding treatment, is associated to each DSCP.

In this paper, we propose a possible mapping of the ten classes of service defined in [70] onto the four optical transport mechanisms employed in the proposed HOS networks, i.e. circuit, long burst, short burst and packet. In other words, each of the ten classes defined in [70] is mapped onto the most appropriate optical transport mechanism. The traffic classification is performed basing on the DSCP field of IP packets arriving at the HOS network. When an IP packet arrives at the edge node, a traffic classifier processes its DSCP field, associates the packet to one of the service classes defined in [70] and then selects the most appropriate transport mechanism. Table 2.4 reports the ten user traffic classes with their characteristics and a possible mapping with the HOS transport mechanisms.

The TDM-circuits are used to transport the services that have the most stringent requirements in terms of QoS and that are characterized by large and

Table 2.4: Traffic classes and their mapping on the HOS transport mechanisms.

Service name	**Tolerance to loss**	**Tolerance to delay**	**Tolerance to jitter**	**Transport type**
Telephony	Very low	Very low	Very low	TDM-Circuit
Signaling	Low	Low	Yes	Short burst
Multimedia conferencing	Low-Medium	Very low	Low	TDM-Circuit
Real-time interactive	Low	Very low	Low	Short burst
Multimedia streaming	Low-Medium	Medium	Yes	Short burst
Broadcast video	Very low	Medium	Low	Long burst
Low latency data	Low	Low-Medium	Yes	Long burst
High throughput data	Low	Medium-High	Yes	Long burst
Low priority data	High	High	Yes	Packet
Standard	Not specified	Not specified	Not specified	Packet

stable flows. Data are queued at the ingress edge node until the circuit has been established through the network using the two-way reservation mechanism, and once the circuit has been established, data are carried without losses, jitter and delays, except from the propagation delay, thereby ensuring high QoS. Long bursts are associated to applications generating long flows that do not have stringent requirements in terms of delay. In fact, due to long assembly times and long offset-times, long bursts introduce a relatively high delay. On the other side, the long offset-times gives to long bursts a kind of prioritized handling in comparison to short bursts and packets that results in low loss rates. Short bursts are well suited for applications requiring a minimum guaranteed level of QoS in terms of both losses and delays and characterized by rapidly changing traffic. Finally, packets are used to transport services that have a relatively high tolerance to loss, delay and jitter.

In the all-optical HOS network packets and short bursts are forwarded by means of fast optical switching elements and without buffers for contentions resolution. Consequently, packets and short bursts are expected to have relatively high loss rates, but, since they are not buffered at the core nodes they will introduce relatively small delays and jitters. In the optical/electronic HOS network packets and short bursts are transmitted using fast electronic switching elements

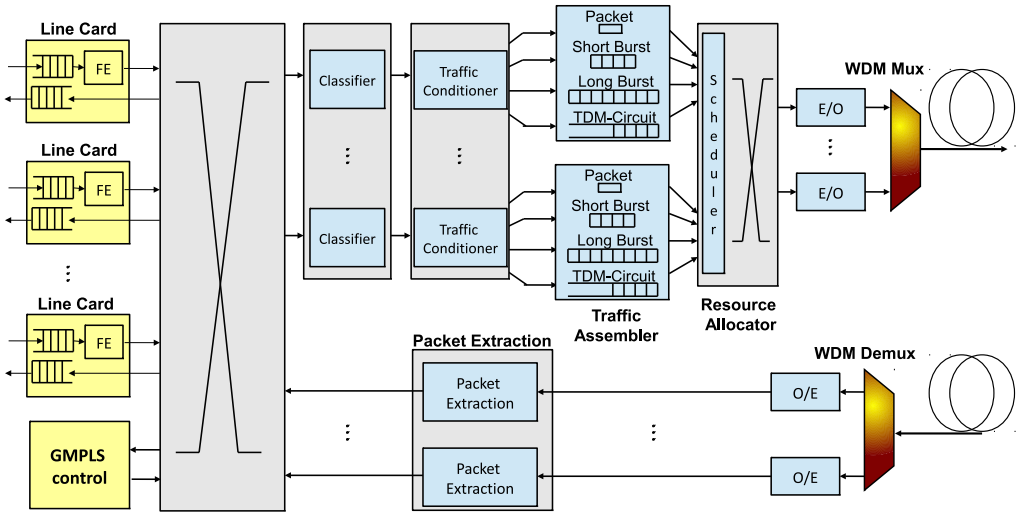

Figure 2.20: Architecture of a HOS edge node.

and electronic buffers are used for solving contentions at the core nodes. This will results in negligible losses, but will increase significantly delays and jitters.

2.3.2 Edge Node Architectures

Edge nodes perform the tasks required for the interoperability between the core network and the legacy networks, and in particular they are responsible for the traffic aggregation and classification. Edge nodes can have a strong impact on performance and energy efficiency, and thus we decided to include them in our analysis. For the all-electronic network we consider a traditional implementation of an electronic edge router with a switching fabric interconnecting a large number of electronic LC. The electronic LC deploy all the functions needed for the interoperability toward the legacy networks, and perform traffic classification and shaping according to the QoS policies adopted in the network core.

The HOS edge node architecture is depicted in Figure 2.20. When a packet arrives at a LC from a legacy network, its header information is extracted and passed through the switch to the GMPLS control, while the remainder of the packet remains in the inbound LC. The GMPLS module processes the information in the IP header and associates the packet to a proper FEC. If there exists an already established LSP associated to the FEC, the packet is assigned to this LSP. Otherwise, the GMPLS module sends a request to a PCE that creates a

new LSP and connects it to the FEC. Once the LSP has been identified, the information is sent back to the inbound LC and the packet is forwarded through the switch to the proper output LC. It is worth remembering that the GMPLS operates at the fiber level, and thus the GMPLS module selects only the output fiber where to forward the packet and it does not select a specific wavelength that will be selected later by the resource allocator.

The output line cards are composed by the following blocks: classifier, traffic conditioner, traffic assembler and resource allocator. The traffic classifier processes the DSCP of incoming packets and associates each packet to an optical transport mechanism as already described in Section 2.3.1. The second block is the traffic conditioner. Core networks are usually managed by carrier providers that offer services to many Internet Service Providers (ISP). The main task of the traffic conditioner is to ensure that ISPs comply with the Service Level Agreement (SLA) for employing the core network. As an example, a maximum number of circuits that can establish simultaneously through the HOS network could be assigned to each ISP. The traffic conditioner keeps track of all the circuits that are established by the ISPs. When an incoming packet requires for the establishment of a new circuit, the traffic conditioner checks if the ISP is allowed to establish the circuit according to its SLA. In case the ISP is not allowed, the packet will be remarked and transmitted using a different transport mechanism.

The third block is the traffic assembler, which in turn is divided into 4 parallel sub-blocks, one for each transmission scheme. The first sub-block is the optical packet generator, which simply converts the incoming IP packets into the optical domain. The second sub-block is the short bursts assembler that is composed by a queue for the burst formation and employs a mixed timer-length assembly algorithm [71]. When an IP packet arrives at the queue a timer is started and as soon as the timer expires or the length of the queue achieves a certain threshold, the burst is completed. The third sub-block is the long bursts assembler, which is similar to the previous one but employs a different assembly algorithm. In this case, we define two thresholds L_{min} and $L_{max} > L_{min}$. When the length of the queue overcomes L_{min} a timer is started. The burst is generated as soon as the timer expires or the queue length becomes equal or greater than L_{max}. This mechanism ensures that long bursts are always longer than L_{min}. This is important because long bursts are forwarded using slow optical switching elements that have switching times in the order of milliseconds. If the long bursts

are too short, i.e. their duration is lower than a few milliseconds, this would lead to high inefficiencies. On the other side, L_{min} must not be too high. Otherwise, if the arrival rate of IP packets is low the burst generation would require a lot of time leading to high delays. Finally, the fourth sub-block is the circuit generator that performs traffic grooming functionalities and is equipped with a buffer for storing packets until the lightpath has been established.

The last block is the resource allocator that schedules circuits, bursts and packets on specific output wavelengths. The resource allocator is composed by a switch and a scheduler and employs electronic buffers to solve contentions. Data are stored until an available output resource is found, and consequently the resource allocator does not introduce losses but it may introduces extra-delays.

In the direction toward the legacy networks, the packet extractor is used to extract IP packets from bursts and TDM-circuits.

2.3.3 Edge Nodes Power Consumption

Let us consider a core network with a mesh topology, like the Pan-European network [72]. We assume that each node in the network is composed of a core switch, which forwards the traffic through the core network, and an edge router, which provides traffic aggregation and interoperability between core and legacy networks. Let M be the number of ingress/egress fibers at each core switch and let N be the number of wavelengths per fiber. We assume that at each core switch a fraction R of the incoming fibers is connected to the edge node and the corresponding traffic is dropped and received by the electronic edge router. Correspondingly, the same number of fibers is used to add new traffic into the network by the edge router. The number of fibers that are added/dropped at each node is then given by $D = R \cdot M$, and each edge node is connected to the corresponding core node by D fibers with N wavelength channels each. In this scenario, we define the energy consumption of a network node as the sum of the energy consumed by the edge router and the energy consumed by the core switch:

$$P_C = P_{Edge} + P_{Core} \tag{2.29}$$

The energy consumptions of the edge router and core switch are obtained by summing the power consumption of all their active components. In the following

Table 2.5: Power consumption of the HOS edge node.

Components	**Power**
Line Card (LC) at 40 Gbps	300 W
Route Processor (RP)	200 W
Switch	8 W
Traffic Classifier and Conditioner (TCC)	62 W
Traffic Assembler (Ass)	62 W
Resource Allocator (RA)	680 W
Packet Extractor (PE)	25 W

we will introduce an analytical model for evaluating the power consumption of the edge nodes. As regards the core nodes, we refer the architectures and power consumption models introduced in Section 2.1 and 2.2.

2.3.3.1 HOS Edge Node

The HOS edge node architecture can be logically divided in two blocks. The first block includes all the functions for interfacing toward the legacy networks and can be referred to as the general edge functions block. The second block is responsible for traffic aggregation and shaping and it can be referred to as the traffic aggregation block. We can then define the power consumption of the edge nodes through the following formula:

$$P_{Edge} = P_{GEF} + P_{TA} \tag{2.30}$$

where P_{GEF} is the power consumption of the general edge function block and P_{TA} is the power consumption of the traffic aggregation block. The power consumption of the general edge function block is given by the following formula:

$$P_{GEF} = D \cdot N \cdot (P_{LC} + \frac{P_{RP}}{16} + P_{Switch}) \tag{2.31}$$

where P_{LC}, P_{RP} and P_{Switch} are the power consumption of the input LCs, the route processors, and the switch, respectively. The route processors are used to perform the functions of the GMPLS module in a distributed manner, with one route processor every 16 LCs. The switch is used to interconnect the LCs, and P_{Switch} represents the power consumption per switch port. The power

consumption of the traffic aggregation block is obtained through the following formula:

$$P_{TA} = \frac{D \cdot N}{2}(P_{TCC} + P_{Ass} + \frac{P_{RA}}{N} + P_{PE}) \tag{2.32}$$

where P_{TCC} is the power consumption of the traffic classifier and traffic conditioner that are realized using a unique large field programmable gate array. The P_{Ass}, P_{RA} and P_{PE} are the power consumptions of the traffic assembler, resource allocator and packet extractor, respectively. The dividing factor two is due to the fact that the traffic aggregation functions are implemented only at the output ports. The energy consumptions of all the considered network components are reported in Table 2.5 and have been obtained by collecting data of a number of commercially available components as well as from research papers.

2.3.3.2 All-electronic Edge Node

The energy consumption of the all-electronic edge router can be also obtained using formula 2.30. However, in this case the traffic aggregation block includes only the traffic classifier and the traffic conditioner. In fact, since the all-electronic network is a pure packet switched network the functions such as traffic assembly, resource allocation and packet extraction are not required. As a result, the power consumption of the all-electronic edge router will be lower than the power consumption of the HOS edge routers.

2.3.4 Edge and Core Network Simulation Setup

In this Section we describe the performed analysis by describing the simulation setup and presenting the performance metrics that are used to compare the considered network solutions.

The performance and power consumption of the proposed networks have been assessed by means of an event-driven simulator, developed specifically for this scope. The network setup is depicted in Figure 2.21, and consists of a cascade of core nodes each connected to an edge router.

The edge routers receive IP packets coming from legacy networks and perform traffic classification, traffic assembly and resource allocation. The IP packets are

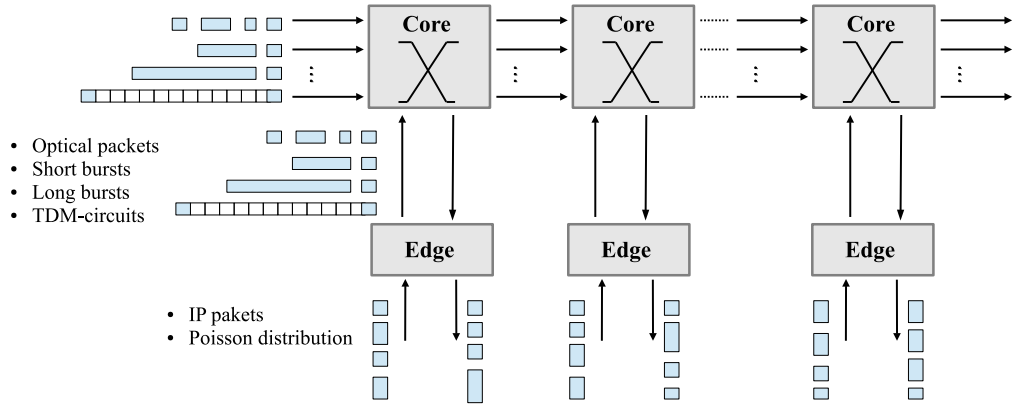

Figure 2.21: Simulation setup.

generated according to a Poisson distribution. No specific transport protocol is considered because it does not influence our analysis and theoretically any transmission control mechanism could be employed. The DSCPs of the incoming IP packets are randomly generated and are modeled in order to ensure that in the HOS core networks the resources are equally distributed among circuits, long bursts, short bursts and packets. At each core node, traffic addressed to this particular node is received and sent to the edge router. The add/drop ratio R is set to 0.25, i.e. at each core node one fourth of the fibers are terminated and the corresponding traffic is sent to the destination edge node. Core nodes forward data on the proper output fiber according to their destination and employing the appropriate scheduling algorithm.

The setup of Figure 2.21 can be considered as a link in a mesh network, such as the Pan-European network. We assume a possible future design of the Pan-European network with both nodes and links of high-capacity. Because the average node degree is 3 and the add/drop ratio is 0.25, each core node has in average 24 input and 24 output fibers, i.e. $M = 24$. The number of wavelengths per fiber N is set to 80 and each wavelength channel is operated at $40\,Gbps$, resulting in an average aggregate capacity per core node of $76.8\,Tbps$. Each edge node is connected to the corresponding core node by $D = 6$ fibers. The considered network architectures are compared by means of the following metrics: average loss rates, maximum delays, maximum jitters, and energy efficiency. Loss rates and delays are considered in order to prove the effectiveness of HOS of supporting the classes of service introduced in Section 2.3.1. We consider connections with 4 and 8 hopes, which are respectively the average and the

maximum number of hops in the Pan-European network.

The average loss rates are defined as in previous Sections. The delay is defined as the time between the arrival of an IP packet at the source edge node and the time in which the IP packet is received by the destination edge node. In the HOS networks the packet delay is the delay experienced by IP packets that are transmitted as optical packets through the core network. Similarly, the short burst delay, long burst delay, and circuit delay, are the delays experienced by IP packets that are transmitted through the core network within a short burst, a long burst or a circuit, respectively. In the all-optical HOS network the delays are obtained through the following formulas:

$$\begin{cases} D_{IPpacket}^{Packet,O} = D_{RA} + D_{Prop} \\ D_{IPpacket}^{ShortBurst,O} = D_{Ass} + D_{Off} + D_{RA} + D_{Prop} \\ D_{IPpacket}^{LongBurst,O} = D_{Ass} + D_{Off} + D_{RA}D_{Prop} \\ D_{IPpacket}^{Circuit,O} = D_{Prop} \end{cases} \tag{2.33}$$

Here, D_{RA} is the delay introduced by the resource allocator at the network edge. D_{Ass} and D_{Off} are the delays introduced by the bursts assembler and the bursts offset-time, respectively. Finally, D_{Prop} is the propagation delay, i.e. the time required to propagate the optical signal through the links. In the optical/electronic HOS network electronic buffers are employed at the core nodes for packet and short burst contention resolutions. Consequently, the delays are obtained using the following formulas:

$$\begin{cases} D_{IPpacket}^{Packet,OE} = D_{RA} + D_{Buff} + D_{Prop} \\ D_{IPpacket}^{ShortBurst,OE} = D_{Ass} + D_{Off} + D_{RA} + +D_{Buff} + D_{Prop} \\ D_{IPpacket}^{LongBurst,OE} = D_{Ass} + D_{Off} + D_{RA} + D_{Prop} \\ D_{IPpacket}^{Circuit,OE} = D_{Prop} \end{cases} \tag{2.34}$$

where D_{Buff} is the delay introduced by the electronic buffers. We assume that the TDM-circuit durations are sufficiently longer than the time required for their establishment using the two-way reservation mechanism. As a consequence,

in formulas 2.33 and 2.34 we set the circuit delay equal to the propagation delay. The all-electronic network is a pure packet switched network and the delays are given by the sum of the buffering delays at the core nodes and the propagation delay:

$$D_{IPpacket}^{Packet,E} = D_{Buff} + D_{Prop} \tag{2.35}$$

In the all-electronic network the edge nodes introduce no extra-delays because no resource allocation functions are needed. It can be noted that the propagation delay D_{Prop} is an additive constant that is equal for all the considered delays. Since we are interested in comparing the delays in the different network architectures, we decided to neglect D_{Prop} because it has no relevance in our analysis.

In accordance with the definition of delays given above, we define the maximum jitter as the maximum delay variation between IP packets. Consequently, we refer to maximum packet jitter as the difference between the maximum and the minimum packet delays. In the HOS network, we refer then to short burst, long burst and circuit maximum jitter, as the difference between the maximum and the minimum long burst, short burst, and circuit delays, respectively.

To compare the energy efficiency of the considered network architectures, we use the definition of improvements in energy efficiency between the HOS networks and the all-electronic network that we already introduced in Section 2.1.4. In addition, we define the improvement in energy efficiency between the all-optical HOS and the all-electronic network as:

$$IE_{O,E} = \frac{\frac{T_h}{P_C}\,|_O - \frac{T_h}{P_C}\,|_E}{\frac{T_h}{P_C}\,|_E} \cdot 100\,\% \tag{2.36}$$

where T_h is the achievable throughput and P_C the total power consumption of a network node.

2.3.5 Numerical Results

In this Section we present and compare the performance of the all-optical HOS, optical/electronic HOS and all-electronic core and edge networks. Firstly, we will take into consideration the average loss rates, secondly the maximum delays and jitters, and finally the energy efficiency.

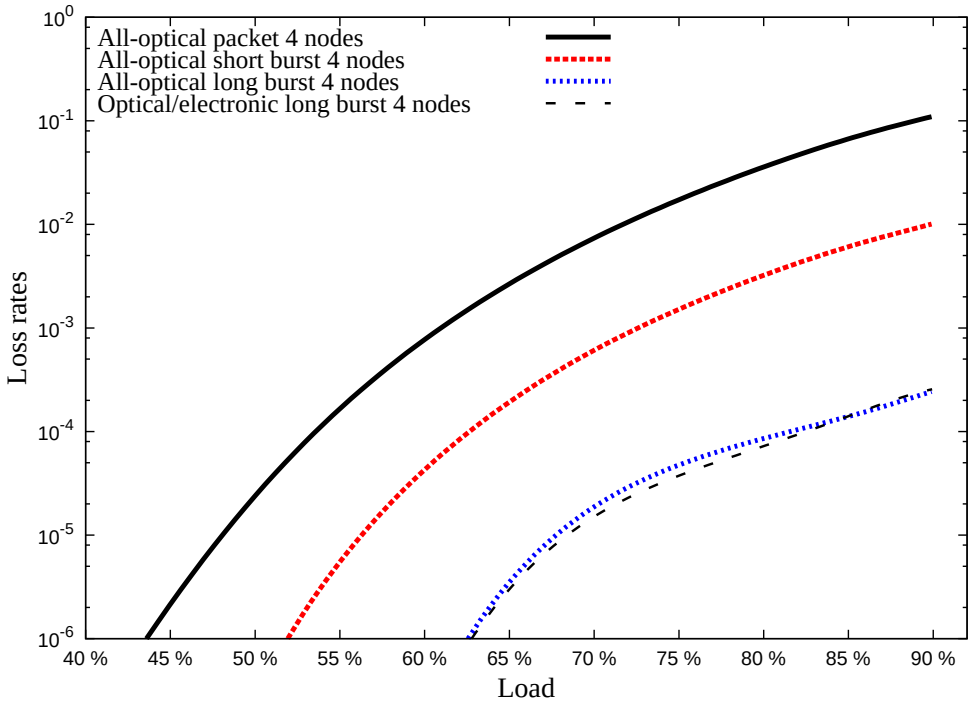

Figure 2.22: Data loss rates as a function of the input load in a link with 4 cascaded nodes.

2.3.5.1 Loss Rates

Figure 2.22 shows the average data loss rates as a function of the input load in a link with 4 cascaded nodes. The Figure shows that in the all-optical HOS network the packets have the highest loss probability, which is in the order of 10^{-1} for loads higher than 90%. Due to shorter offset-times, short bursts have loss rates that are almost two order of magnitude higher with respect to long bursts, but always at least one order of magnitude lower than the packet loss rates, providing a minimum level of QoS. Long bursts have the lowest loss rates that are always lower than 10^{-3}, even for very high input loads. As regards the optical/electronic HOS network, the packets and short bursts are transmitted using the fast electronic switch and electronic buffers are used to solve contentions at the core nodes, so we can assume that their loss rates are negligible. Long bursts, which are instead forwarded using slow optical switching elements, show almost the same loss rates as in the all-optical HOS network. Finally, the circuit establishment failure probability is always zero in both the all-optical and

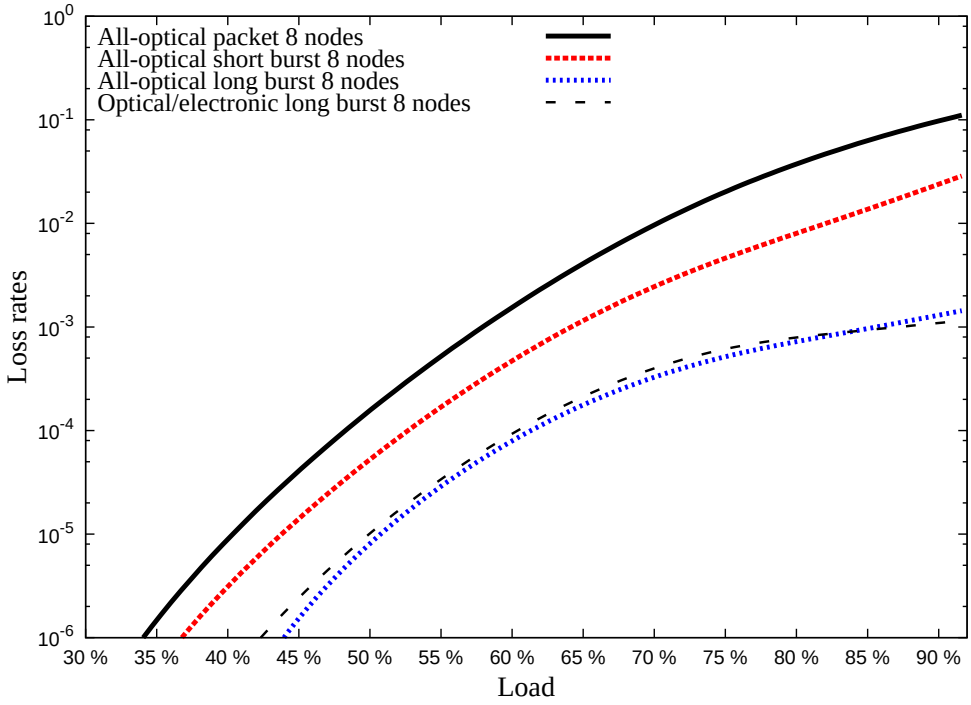

Figure 2.23: Data loss rates as a function of the input load in a link with 8 cascaded nodes.

optical/electronic HOS networks. The all-electronic network is a pure packet switched network with negligible data losses.

Figure 2.23 shows the data loss rates as a function of the input load in a link with 8 cascaded nodes. The same observations as in the previous case can be made, but here the loss probabilities are noticeably higher. In particular, the short and long bursts suffer from the increased number of cascaded nodes showing much higher losses especially for relatively low input loads. On the other hand, packets do not suffer too much from the increased number of cascaded nodes because they can be inserted into unused TDM-slots of circuits. In fact, the simulation results demonstrate that the higher is the number of nodes in the cascade and the higher is the probability for a packet to find a circuit with the same destination that has a free TDM-slot. The circuit establishment failure probability is again always null.

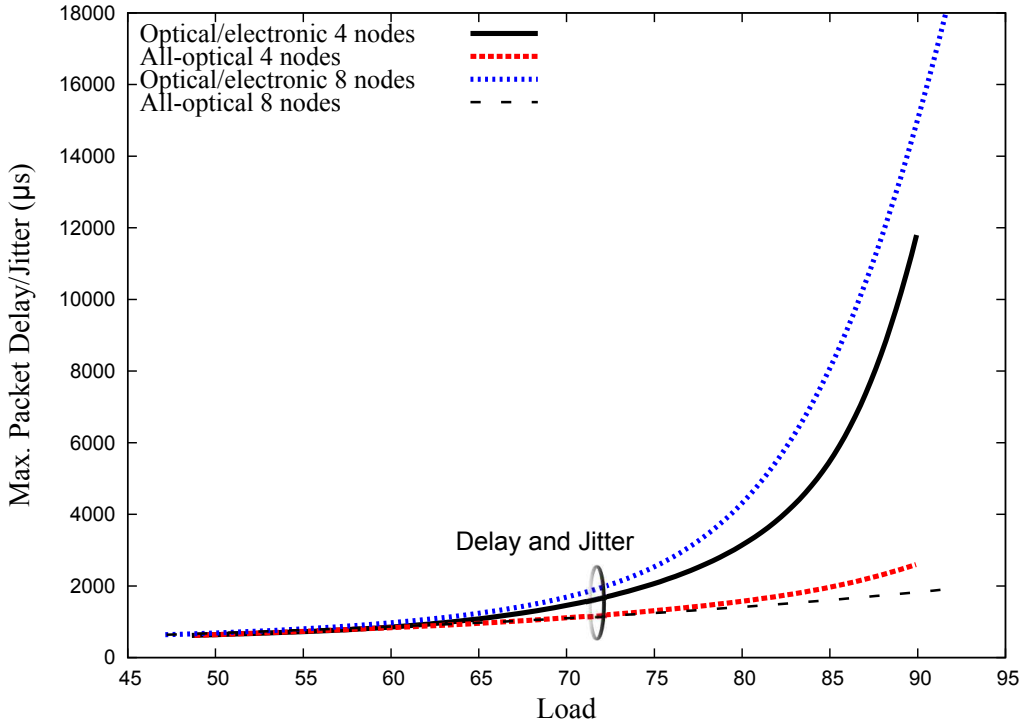

Figure 2.24: Maximum delay experienced by IP packets that are transmitted through the HOS networks as optical packets.

2.3.5.2 Delays and Jitters

In Figure 2.24 the maximum packet delays and jitters are shown for the three considered network architectures in links with 4 and 8 cascaded nodes. The Figure shows that the maximum packet delay and the maximum packet jitters always coincide. This is due to the fact that in our simulations, the minimum packet delay is always null and therefore we obtain that the maximum packet jitters is always equal to the maximum packet delay. To understand why, it is worth remembering that we neglect the propagation delay D_{Prop} and we consider only the delays introduced for data processing. Packets that do not contend for output resources neither in the edge node nor in the core nodes are then transmitted without delays and thus we obtain that the minimum packet delay is null.

The Figure also shows that packet delays increase exponentially while increasing the input load. When the load is lower than 60% the packet delays in the all-optical and optical/electronic HOS networks are almost the same. In the

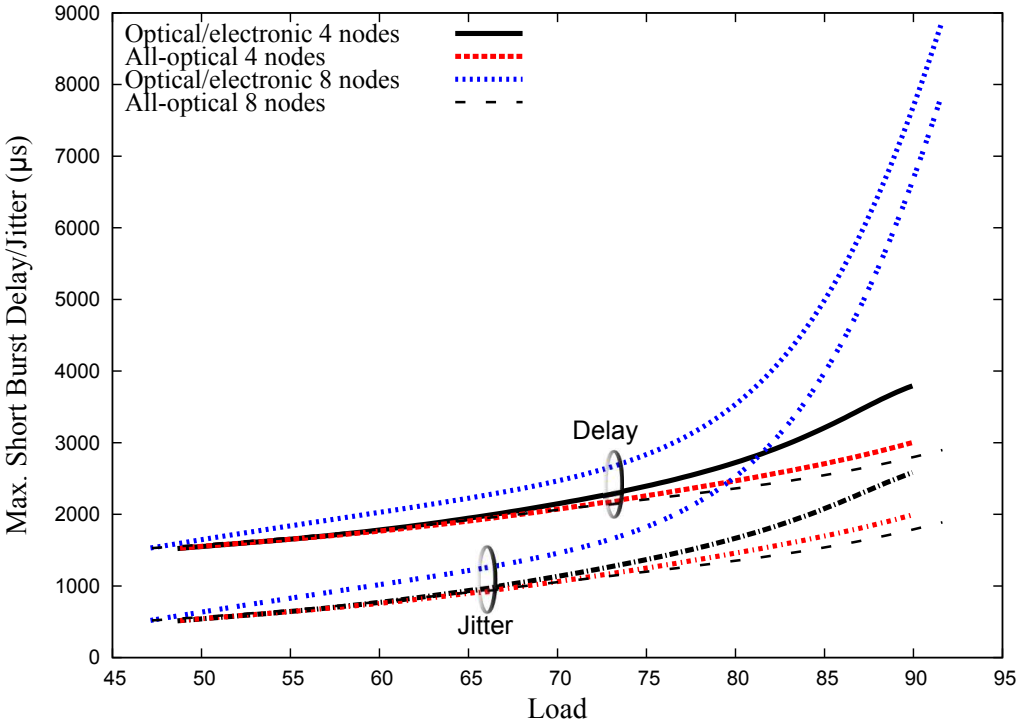

Figure 2.25: Maximum delay experienced by IP packets that are transmitted through the HOS networks within short bursts.

optical/electronic HOS network, while increasing the load from 60% to 90% the packet delays increase very fast, due to the delay introduced by the electronic buffers, D_{Buff}. This is due to the fact that the higher is the input load and the higher is the packets contention probability, resulting in longer waiting times in the electronic buffers. The Figure also shows that with 8 nodes the maximum delays in the optical/electronic network are much higher due to the fact that the longer is the cascade and the higher is the probability of contention. On the other side, in the all-optical HOS network the packet delays do not increase much with the increase in the input load and the maximum delay is not influenced by the length of the cascade. This is due to the fact that the core nodes do not employ buffers for contention resolution and thus they do not introduce delays. Here, the delay is only due to the resource allocation D_{RA} at the edge nodes.

Figure 2.25 shows the short bursts delays and jitters in the HOS networks for links with 4 and 8 cascaded nodes. In this case, delays and jitters show the same trend but different values. In fact, the minimum short burst delay is always greater than null due to the assembly delay D_{Ass} and the offset-time D_{Off}. D_{Ass}

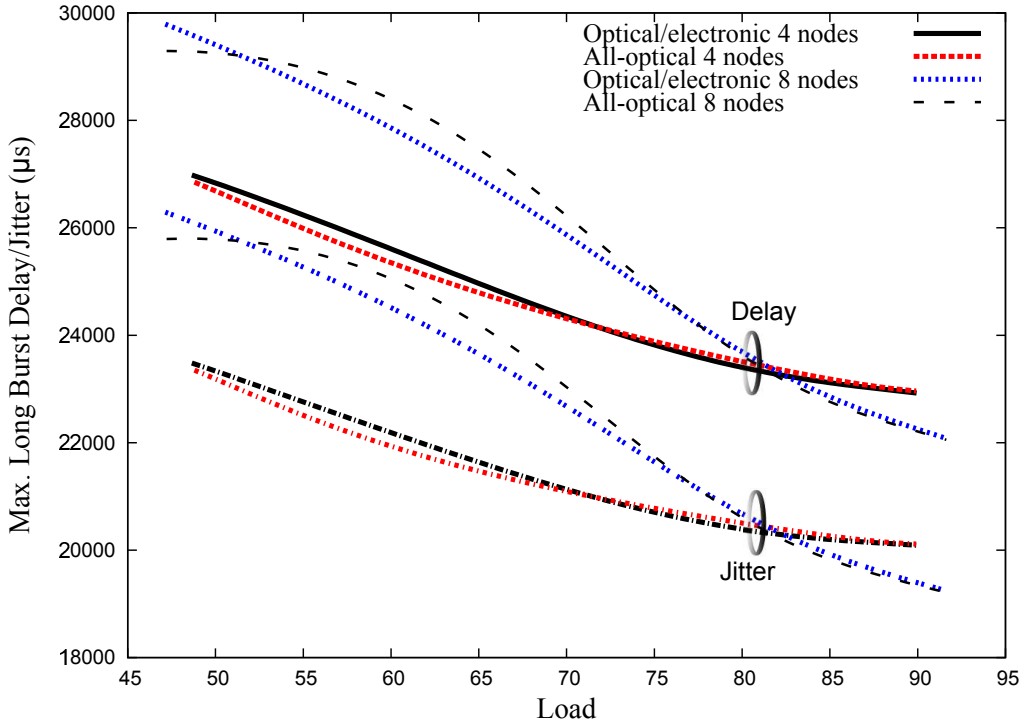

Figure 2.26: Maximum delay experienced by IP packets that are transmitted through the HOS networks within long bursts.

depends on the arrival rate of IP packets at the edge node, while D_{Off} depends on the priority given to the burst and in our simulations is a uniform random value within a specified interval. The Figure shows that while increasing the network load the difference between the maximum and minimum delays increases and the curves representing the jitters approach the curve representing the delays.

Again we observe that for relatively low loads, below 50%, the delays in the all-optical and optical/electronic HOS networks are similar. While increasing the load the probability of contention becomes higher, and this leads to increasing the short burst loss rate in the all-optical HOS network and the short burst delay in the optical/electronic HOS network. As a result, for loads higher than 50% $D_{IPpacket}^{ShortBurst,OE}$ becomes much higher than $D_{IPpacket}^{ShortBurst,O}$, especially in the link with 8 cascaded nodes. Comparing Figures 2.24 and 2.25 it can be observed that for loads below 70% short bursts show always higher delays with respect to packets. This is due to the time that is required for the short burst assembly D_{Ass} and to the offset-time D_{Off}. In particular, D_{Ass} is higher at low loads because when the arrival rate of IP packets is low the assembler takes more time for

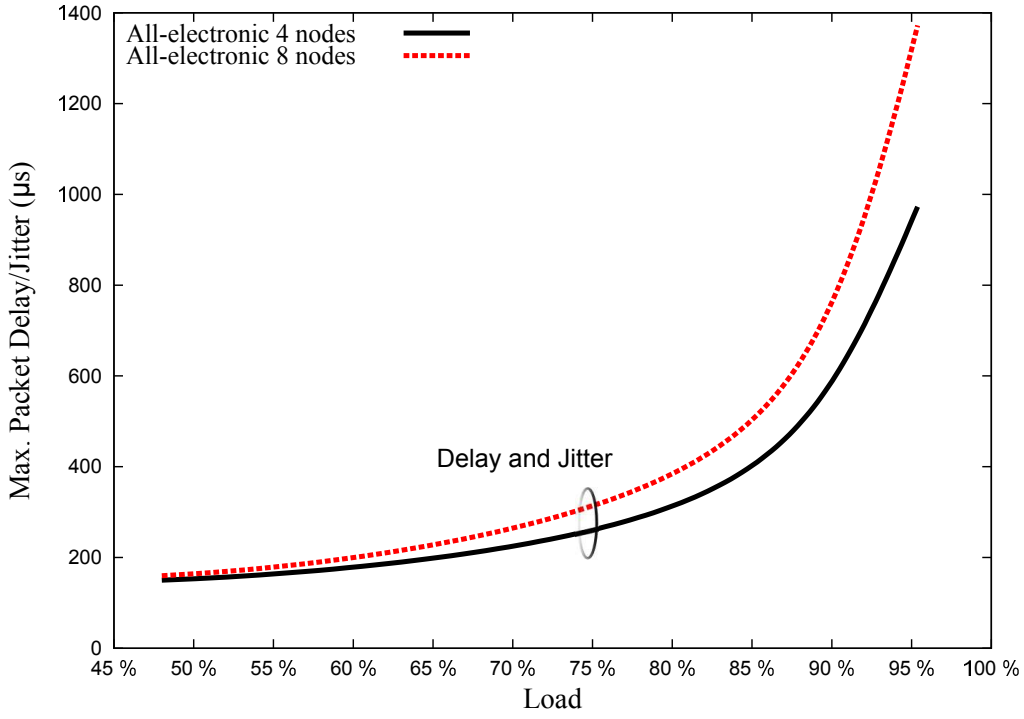

Figure 2.27: Maximum IP packets delays in the all-electronic network.

generating the bursts. However, when the load is higher than 70%, $D_{IPpacket}^{Packet,OE}$ becomes much higher than $D_{IPpacket}^{ShortBurst,OE}$, due to the fact that packets are scheduled with lower priority.

In Figure 2.26 the long bursts delays and jitters as a function of the input load are shown for the HOS networks. The long bursts delay show a completely different trend with respect to packets and short bursts delays. This is because the main contributor to the long bursts delay is the assembly delay D_{Ass}. The higher is the input load and the higher is the arrival rate of the IP packets to the edge nodes, which in turn reduces the long bursts assembly time. As a consequence while increasing the input load the long bursts delay decreases. The all-optical and optical/electronic long burst delays show almost the same trend and both are lower in the link with 4 cascaded nodes, except that for very high input loads. As for the short bursts, long burst jitters and long burst delays show the same trend as a function of the network load. Again, the minimum short burst delay cannot be null due to D_{Ass} and D_{Off}. Increasing the network leads to both lower maximum and minimum D_{Ass}, however the Figure shows that the difference between maximum and minimum D_{Ass} is always relatively

large leading also high long burst jitters.

Finally, in Figure 2.27 the maximum delays and jitters of the IP packets in the all-electronic network are shown. All the packets are supposed to have the same priority and the delays are only due to the electronic buffering at the core nodes. As in the HOS networks, the minimum packet delay is always null and thus the maximum jitters and the maximum delays coincide. The Figure shows that the all-electronic network is able to keep the delays bounded to less the $2\,ms$ even for very high input loads. This is due to the fact that no delays are introduced at the edge nodes for burst assembly or resource allocation. Furthermore, in the all-electronic network packets share the resources efficiently leading to a low probability of congestion at the network core and low buffering delays at low and moderate loads. At high loads of more than 80%, the IP packet delays increase exponentially as we have already observed in the optical/electronic hybrid network. However, in the all-electronic network the entire traffic is experiencing increased delays, while the hybrid optical networks are capable of guaranteeing low and deterministic delays for the IP packets transmitted over circuits. Thus, the HOS networks are able to provide an efficient service differentiation directly at the optical layer.

2.3.5.3 Energy Efficiency

In Figure 2.28 we show the improvements in energy efficiency between the HOS networks and the all-electronic network as a function of the input load. The $IE_{O,E}$ and the $IE_{O/E,E}$ are shown when considering both core and edge nodes, and additionally, for a network containing core nodes only. In the former case, the power consumption of a network node and the increase in energy efficiency are defined as in Section 2.3.3 and 2.3.4. In the latter case, the power consumption of the edge nodes is neglected and we assume $P_C = P_{Core}$.

The Figure shows that the lower is the load and the higher is the improvement in energy efficiency provided by the hybrid architectures. This is explained by the fact that in the HOS networks it is possible to schedule the switch off of the unused ports, which can lead to large power savings. The lower is the load and the higher is the number of ports that can be switched off, which leads to a higher energy efficiency with respect to the all-electronic network. The Figure also shows that when considering only the core nodes the improvement of the

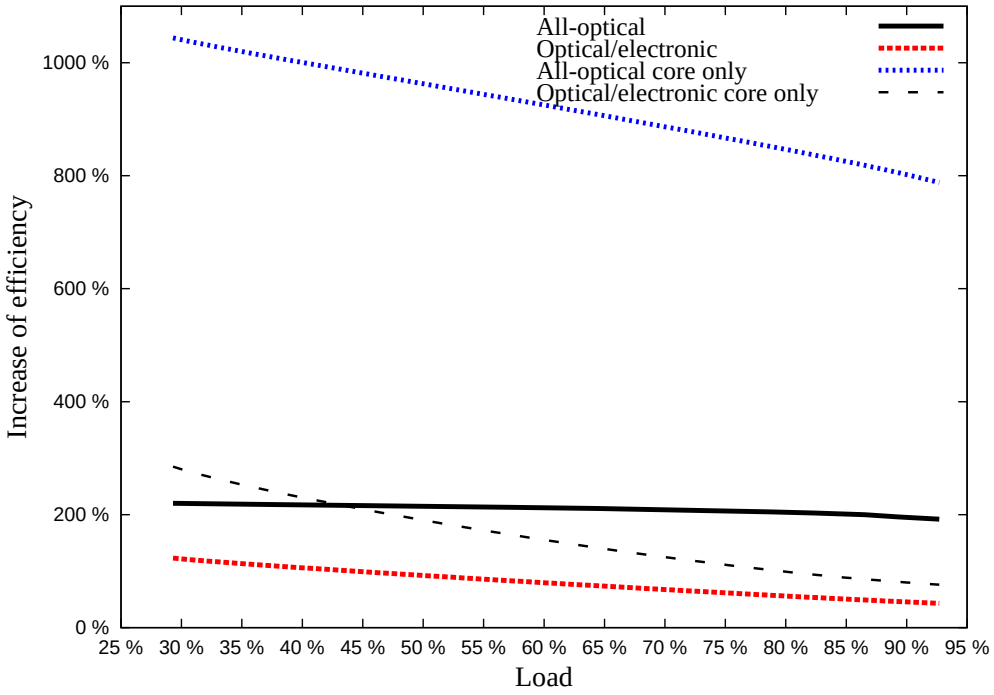

Figure 2.28: Increases of efficiency between the HOS networks and the all-electronic network as a function of the input load.

energy efficiency is much higher. The energy efficiency of an all-optical and an optical/electronic HOS core node is, respectively, around 9 and 2 times higher than the energy efficiency of an all-electronic core node even for very high input loads. However, when we take into consideration also the power consumption of the edge nodes the increases of efficiency are lower. This is due to the fact that the HOS edge nodes implement complex and power consuming traffic assembly functions that decrease the energy efficiency of the HOS networks. This effect is more evident for $IE_{O,E}$ that changes from 790% to 190% when the input load is as high as 90%. The trend is less evident for the optical/electronic network, where the curves with and without the edge nodes become very similar at high input loads. This is due to the fact that at high input loads the power consumptions of both the all-electronic and optical/electronic HOS networks are dominated by the power consumption of the core electronic switch.

In Figure 2.29 the improvements of energy efficiency are shown as a function of the add/drop ratio R. The Figure shows that the higher is the add/drop ratio and the lower are the improvements of efficiency provided by the HOS

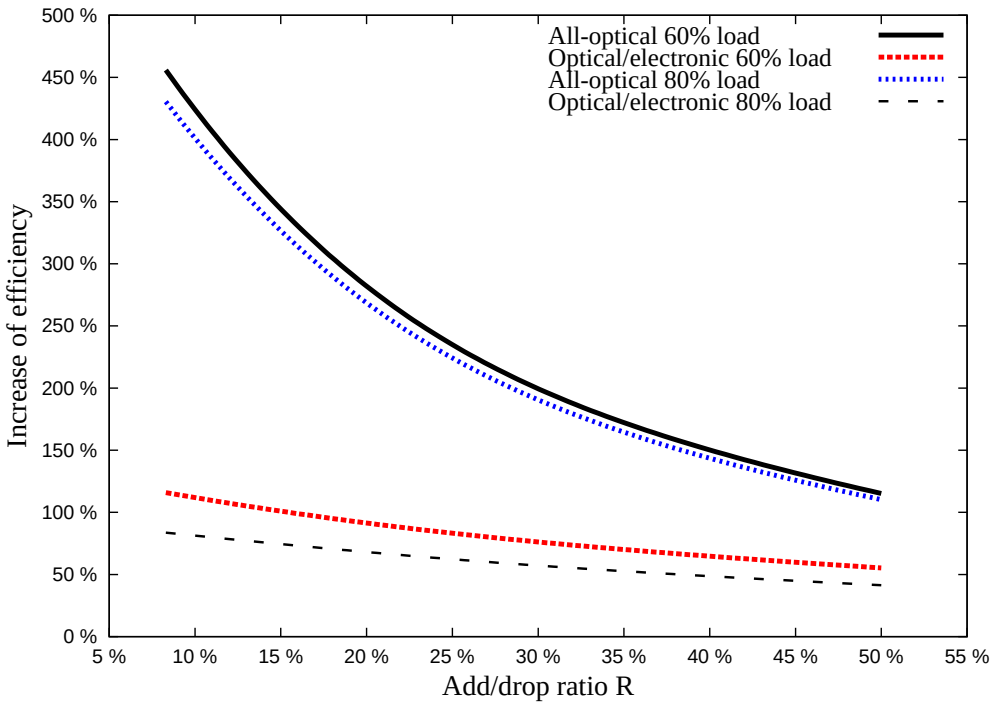

Figure 2.29: Increases of efficiency between the HOS networks and the all-electronic network as a function of the percentage of fibers that are terminated at each node.

networks. In fact, the higher is R and the higher is the number of fibers that are terminated at each core node, and in turn the higher is the size and the capacity of the edge nodes. The higher is the size of an edge node and the higher is its energy consumption. Since the energy consumption of the edge nodes has a strong impact on the HOS networks, an increase of the add/drop ratio results in a decrease of both $IE_{O,E}$ and $IE_{O/E,E}$. In the all-optical HOS network the energy consumption of the edge nodes is dominant even at very high input loads, and consequently $IE_{O,E}$ decreases more abruptly than $IE_{O/E,E}$.

2.3.6 Conclusions

In this paper we studied the performance and energy efficiency of two possible implementations of a HOS network, namely all-optical HOS and optical/electronic HOS, which include edge and core nodes. The HOS edge node architecture has been studied and an analytical model has been proposed for the evaluation of

its energy consumption. Four different HOS classes of service have been defined, namely TDM-circuits, long bursts, short bursts and packets, and a set of possible future Internet applications have been mapped into these classes. The HOS edge nodes map each application onto the most suited optical transport mechanism using the content of the DSCP field of the IPv4 or IPv6 packet header. Performance and energy efficiency of the analyzed HOS networks have been evaluated through a combined analytical and simulation approach, with particular attention to their capability of supporting different classes of service while reducing the energy consumption of current solutions based on electronic switching. The metrics that have been introduced to perform this analysis are average loss rate, maximum delay, and improvement in energy efficiency toward an electronic switching solution.

Results show that HOS has the potential for supporting future Internet applications with the required QoS. Premier applications generating long and stable traffic flows are associated to TDM-circuits that provide no losses, negligible delays and no jitter. Applications requiring low loss rates and that are not sensitive to delays are carried over long bursts. Short bursts are associated to applications that require minimum guaranteed requirements in terms of both loss and delay. Finally, packets are used to transmit applications that have no specific QoS requirement.

Furthermore, we demonstrated that both the all-optical HOS and optical/electronic HOS networks are able to improve considerably the energy efficiency of current networks based on electronic switching. In particular, the all-optical HOS network always improves the energy efficiency by at least a factor of 3. The edge nodes have a strong impact on the energy consumption of the HOS networks and reduce strongly the improvement of energy efficiency with respect to the all-electronic network. The improvements in energy efficiency decrease while increasing the input load and while increasing the percentage of fibers that are terminated at each node.

Chapter 3

Energy-efficiency in Data Centers

Current data centers networks rely on electronic switching and point-to-point interconnects. When considering future data center requirements, these solutions will raise issues in terms of flexibility, scalability, performance and energy consumption. For this reason several optical switched interconnects, which make use of optical switches and WDM, have been recently proposed. They can be categorized in three groups: hybrid solutions, which envisage the use of optical circuit switching in combination with electronic packet switching, optical circuit switching solutions, optical burst/packet switching solutions. However, hybrid and optical packet/burst switching solutions are limited by their relatively high energy consumption. On the other side, optical circuit switching solutions are characterized by low flexibility and low resource utilization.

In this Chapter we introduce two novel data center networks, which are respectively based on the HOS and EON concept. The former, referred to as HOSDC, envisages the use of a novel and highly scalable optical switch in the core tier, which enable to interconnect a high number of electronic aggregation switches. The electronic aggregation switches require minimal hardware modifications with respect to current commodity switches. As a consequence, the HOSDC interconnect is a applicable in the short/mid term and, as we will prove in this Section, achieves high performance and low energy consumption. However, in the long term HOSDC interconnect may be limited by the capability to scale with the capacity of the servers. As shown in [41], while in 2013 the

majority of the servers deployed worldwide are equipped with 1 Gbps NICs, the capacity of the servers will increase rapidly in the future. To keep up with this trend we propose the EODC interconnect, which is based on the EON paradigm and is able to support capacities up to 100 Gbps per server. The EON concept ensures flexibility and high bandwidth utilization, while the proposed EODC architecture is studied for achieving low energy consumption.

Finally, we propose for the first time an integrated network architecture that provides both intra-data-center and inter-data-center connectivity together with interconnection toward legacy IP networks. The main advantage of the integration of core and intra-data-center networks comes from the possibility to avoid the energy inefficient electronic interfaces between data centers and telecommunication network. The integrated network utilizes an optimized edge caching deployment for achieving maximum energy savings. The results will prove that the proposed network achieves higher energy efficiency with respect to current non-unified solutions and represent a promising solution for future carrier cloud operators.

3.1 HOS for Data Center Networks

In this Section we propose and evaluate the HOSDC interconnect for data centers. The HOS switching paradigm ensures a high network flexibility that we have not found in the data center solutions proposed so far in technical literature. We evaluate the proposed HOS architecture by analyzing its performance, energy consumption and scalability. We compare the energy consumption of the proposed HOS network with a traditional network based on optical ptp interconnects. We demonstrate that HOS has potential for satisfying the requirements of future data centers networks, while reducing significantly the energy consumption of current solutions. The rest of the paper is organized as follows.

The rest of this Section is organized as follows. In Section 3.1.1 we describe a reference optical ptp architecture and the proposed HOSDC interconnect. In Section 3.1.2 we present the model used for the evaluation of energy consumption. In Section 3.1.3 we describe the performed analysis and discuss data center traffic characteristics. In Section 3.1.4 we present and discuss the results and, finally, in Section 3.1.5 draw conclusions.

3.1.1 Data Center Interconnects

In the following we will, firstly, describe the architecture of a traditional data center interconnect based on optical point-to-point interconnects and electronic switching, secondly, introduce the HOSDC network architecture.

3.1.1.1 Optical ptp Architecture

Figure 3.1 shows the architecture of a current data center based on electronic switching and optical ptp interconnects. Here, multiple racks hosting the servers are interconnected using a fat-tree 3-Tier network architecture [73]. The 3 tiers of the data center network are the edge tier, the aggregation tier and the core tier. In the edge tier the Top-of-Rack (ToR) switches interconnect the servers in the same rack. We assume that each rack contains N_S servers and that each server is connected to a ToR switch through a 1 Gbps link. Although in future data centers, servers might be connected using higher capacity links, the majority of current data centers still use 1 Gbps links. In future works we plan

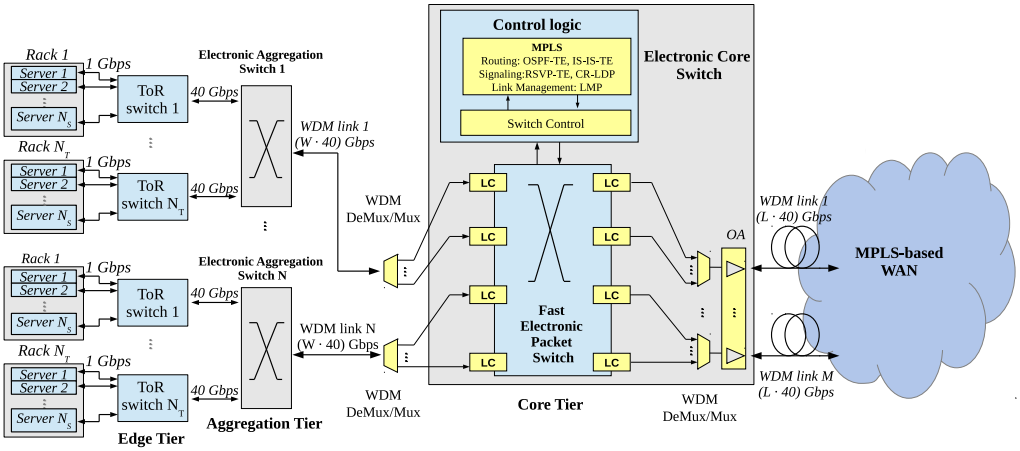

Figure 3.1: Architecture of a data center employing an optical ptp interconnection network. ToR: top of the rack, OA: optical amplifiers.

to consider higher capacity per server port and evaluate the effect of increased server capacity on the network performance.

As many as N_T ToR switches are connected to an aggregation switch using 40 Gbps links. The aggregation switches interconnect the ToR switches in the edge tier using a tree topology and are composed of a CMOS electronic switching fabric and electronic modules that implement traffic aggregation and classification functions. Each aggregation switch is connected to the electronic core switch through a WDM link composed of W wavelengths channels operated at 40 Gbps. The core switch is equipped with $N \cdot W \cdot 40$ Gbps ports for interconnecting as many as N aggregation switches. Furthermore, the core switch employs $M \cdot L \cdot 40$ Gbps ports for connecting the data center to a Wide Area Network (WAN). We assume that the data center is connected to a WAN employing the MPLS control plane. It is worth noting that the considered optical ptp architecture employs packet switching in all the data center tiers. The electronic core switch is a large electronic packet switch that comprises three building blocks, namely control logic, switching fabric and other optical components. The control logic comprises the MPLS module and the switch control unit. The MPLS module performs routing, signaling and link management as defined in the MPLS standard. The switch control unit performs scheduling and forwarding functionalities and drives the electronic switching elements. The switching fabric is a single electronic switch interconnecting a large number of electronic line cards (LCs). Finally, the other optical components include the WDM demultiplex-

ers/multiplexers (WDM DeMux/Mux) and the optical amplifiers (OA) used as boosters to transmit toward the WAN.

In data centers with many thousands of servers, failures in the interconnection network may lead to losses of a high amount of important data. Therefore, resilience is becoming an increasingly critical requirement for future large-scale data center networks. However, the resilience is out of scope of this study and we do not address it in this paper, leaving it as an open issue for a future work.

3.1.1.2 HOS Architecture

The architecture of the proposed HOS optical switched network for data centers is shown in Figure 3.2. The HOS network is organized in a traditional fat-tree 3-Tier topology, where the aggregation switches and the core switches are replaced by the HOS edge and core node respectively. The HOS edge nodes are electronic switches used for traffic classification and aggregation. The HOS core node is composed by two parallel large optical switches. The HOS edge node can be realized by adding some minimal hardware modifications to current electronic aggregation switches. Only the electronic core switches should be completely replaced with our HOS core node. As a consequence, our HOS data center network can be easily and rapidly implemented in current data centers representing a good mid-term solution toward the deployment of a fully optical data center network. When higher capacities per server, e.g. 40 Gbps, will be required, operators can just connect the servers directly to the HOS edge switches without the need of passing through the electronic ToR switches. In this way it will be possible to avoid the electronic edge tier, meeting the requirements of future data centers and decreasing the total energy consumption. In the long term, it is possible also to think about substituting the electronic HOS edge switches with some optical devices for further increase the network capacity. This operation will not require any change in the architecture of the HOS core node, which can be easily scaled to support very high capacities. Furthermore, for increased overall performance and energy efficiency we assume that the HOS core node is connected to a HOS WAN, even if in general the core node could be connected to the Internet using any kind of network technology.

The architecture of a HOS edge node is shown in Figure 3.3. In the direction toward the core switch the edge node comprises three modules, namely classifier,

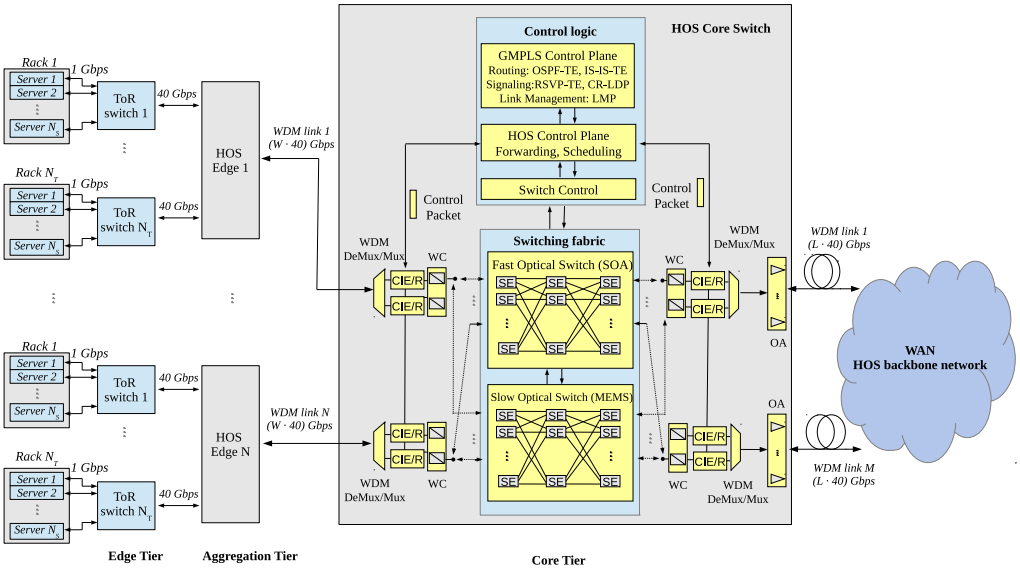

Figure 3.2: Architecture of a data center employing a HOS interconnection network.

traffic assembler and resource allocator. In the classifier, packets coming from the ToR switches are classified basing on their application layer requirements and are associated with the most suited optical transport mechanism. The traffic assembler is equipped with virtual queues for the formation of optical packets, short bursts, long bursts and circuits. Finally, the resource allocator schedules the optical data on the output wavelengths according to specific scheduling algorithms that aim at maximizing the bandwidth usage. In the direction toward the ToR switches a HOS edge node comprises packet extractors, for extracting packets from the optical data units, and an electronic switch for transmitting packets to the destination ToR switches.

As for the electronic core switch, we can divide the HOS core node in three building blocks, i.e. control logic, switching fabric and other optical components. The control logic comprises the GMPLS module, the HOS control plane and the switch control unit. The GMPLS module is used to ensure the interoperability with other core nodes connected to the WAN. The GMPLS module is needed only if the HOS core node is connected to a GMPLS-based WAN, such as the WAN proposed in Chapter 2. The HOS control plane manages the scheduling and transmission of optical circuits, bursts and packets. Three different scheduling algorithms are employed, one for each different data type, for optimizing the

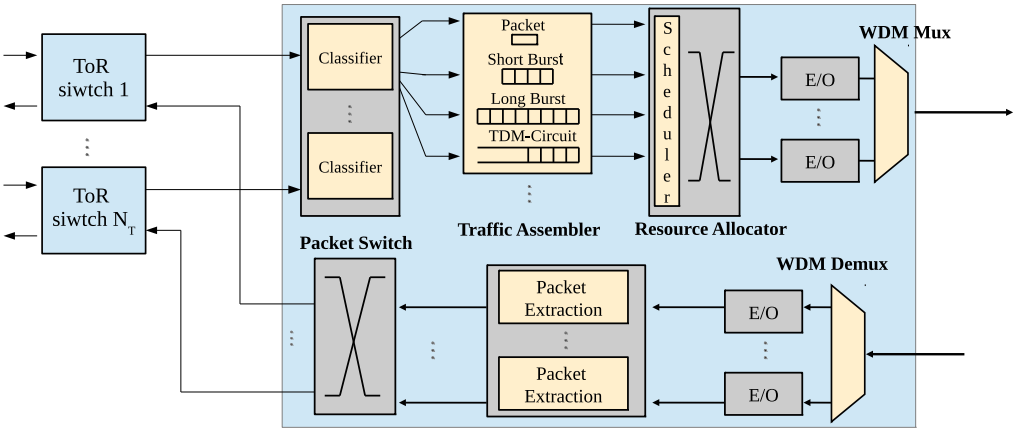

Figure 3.3: HOS edge node architecture.

resource utilization and minimizing the energy consumption. A unique feature of the proposed HOS control plane is that packets can be inserted into unused TDM-slots of circuits with the same destination. This technique introduces several advantages, such as higher resource utilization, lower energy consumption and lower packet loss probability. For a detailed description of the HOS scheduling algorithms the reader is referred to Section 2.1. Finally, the switch control unit creates the optical paths through the switching fabric. The switching fabric is composed of two optical switches, a slow switch for handling circuits and long bursts and a fast switch for the transmission of packets and short bursts. The fast optical switch is based on semiconductor optical amplifiers (SOA) and its switching elements are organized in a nonblocking three-stage Clos network. The slow optical switch is realized using 3D micro electromechanical systems (MEMS). Finally, the other optical components include WDM DeMux/Mux,OAs, tunable wavelength converters (TWCs), and control information extraction/reinsertion (CIE/R) blocks. TWCs can convert the signal over the entire range of wavelengths, and are used to solve data contentions.

3.1.1.3 HOS Transport Mechanisms

The proposed HOS network supports three different optical transport mechanisms, namely circuits, bursts and packets. The different transport mechanisms share dynamically the optical resources by making use of a common control

packet that is sub-carrier multiplexed with the optical data. The use of a common control packet is a unique feature of the proposed HOS network that ensures high flexibility and high resource utilization. Each transport mechanism employs then a particular reservation mechanism, assembly algorithm and scheduling algorithm according to the information carried in the control packet. For a detailed description of the control plane the reader is referred to Section 2.1.

Circuits are long lived optical connections established between the source and destination servers. Circuits are established using a two-way reservation mechanism, with incoming data being queued at the HOS edge node until the reservation has been made through the HOS network. Once the connection has been established data are transmitted transparently toward the destination without any losses or delays other than the propagation delay. In the HOS network circuits are scheduled with the highest priority ensuring a very low circuit establishment failure probability. As a consequence, circuits are well suited for data centers applications with high service requirements and generating long-term point-to-point bulk data transfer, such as virtual machine migration and reliable storage. However, due to relatively long reconfiguration times, optical circuits provide low flexibility and are not suited for applications generating bursty traffic.

Optical burst switching has been widely investigated in telecommunication networks for its potential in providing high flexibility while keeping costs and power consumption bounded. In optical burst switching, before a burst is sent a control packet is generated and sent toward the destination to make an one-way resource reservation. The burst itself is sent after a fixed delay called offset-time. The offset-time ensures reduced loss probability and enables for the implementation of different service classes. In this paper we distinguish between two types of bursts, namely short and long bursts, which generate two different service levels. Long bursts are characterized by long offset-times and are transmitted using slow optical switching elements. To generate a long burst incoming data are queued at the HOS edge node until a minimum queue length L_{min} is reached. After L_{min} is reached, the burst is assembled using a mixed timer/length approach, i.e. the burst is generated as soon as the queue reaches $L_{max} > L_{min}$ or a timer expires. The long offset-times ensure to long bursts a prioritized handling in comparison to packets and short bursts leading to lower loss probabilities. On the other side, the long offset-times and the long times required for burst assembly lead to large

end-to-end delays. Short bursts are characterized by shorter offset-times and are transmitted using fast optical switching elements. To generate a short burst we use a mixed/timer length approach. The short burst is assembled as soon as the queue length reaches a fixed threshold or a timer expires. No minimum burst length is required, as was the case for the long bursts. The shorter offset-times and faster assembly algorithm lead to a higher loss probability and lower delays with respect to long bursts. In Chapter 2 we observed that bursts are suited only for delay-insensitive data center applications because of their high latency. Here, we were able to reduce the bursts latency by acting on the thresholds used in the short and long burst assemblers. Still, the bursts present remarkably higher delays than packets and circuits and thus are suited for data-intensive applications that have no stringent requirement in terms of latency, such as MapReduce, Hadoop, and Dryad.

Optical packets are transmitted through the HOS network without any resource reservation in advance. Furthermore, packets are scheduled with the lowest priority. As a consequence they show a higher contention probability with respect to bursts, but on the other hand they also experience lower delays. However, the fact that packets are scheduled with the lowest priority leads to extra buffering delays in the HOS edge nodes, giving place to higher latency with respect to circuits. Optical packets are mapped to data centers applications requiring low latency and generating small and rapidly changing data flows. Examples of data center applications that can be mapped to packets are those based on parallel fast Fourier transform (MPI FFT) computation, such as weather perdition and earth simulation. MPI FFT requires data-intensive all-to-all communication and consequently requires frequent exchange of small data entities.

For a more detailed description of the HOS traffic characteristics we refer the reader to Chapter 2.

3.1.2 Power Consumption Model

We define the power consumption of a data center as the sum of the energy consumed by all of its active elements. In our analysis we consider only the power consumed by the network equipment and thus we exclude the power consumption of the cooling system, the power supply chain and the servers.

3.1.2.1 Optical ptp Architecture

The power consumption of the optical ptp architecture is defined through the following formula:

$$P_{Net} = N_T \cdot N \cdot P_{ToR} + N \cdot P_{Aggr} + P_{Core} \tag{3.1}$$

where P_{ToR} is the power consumption of a ToR switch, P_{Aggr} the power consumption of an aggregation switch and P_{Core} the power consumption of the core switch. The ToR switches are conventional electronic Ethernet switches. Several large companies, such as HP, Cisco, IBM and Juniper, offer specialized Ethernet switches for use as ToR switch in data center networks. We estimated the power consumption of a ToR switch by averaging the values found in the data sheets released by these companies. With reference to Figures 3.1 and 3.2, without loss of generality we assume $N_T = W$. As a consequence we can assume that the aggregation switches are symmetric, i.e. they have the same number of input and output ports. From now on we will then use N_T to indicate also the number of wavelengths in the WDM links connecting the aggregation and core tiers. The power consumption of an aggregation switch P_{Aggr} is then given by the following formula:

$$P_{Aggr} = N_T \cdot (P_{CMOS} + P_{LC}) \tag{3.2}$$

Here, N_T is the number of input/output ports, P_{CMOS} is the power consumption per port of an electronic CMOS-based electronic switch and P_{LC} is the power consumption an electronic LC at 40 Gbps.

The power consumption of the electronic core switch is given by the sum of the power consumed by all its building blocks:

$$P_{Core} = P_{CL} + P_{SF} + P_{OC} \tag{3.3}$$

where P_{CL} is the power consumption of the control logic, P_{SF} is the power consumption of the switching fabric and P_{OC} is the power consumption of the other optical components. P_{CL} includes the power consumption of the MPLS module and the switch control unit. When computing P_{SF} we assume that the electronic ports are always active. This is due to the fact that current electronic

switches do not yet support dynamic switching off or putting in low power mode of temporarily unused ports. The reason for that is because the time interval between two successive packets is usually too short to schedule the switching off of the electronic ports. As a consequence, we compute P_{SF}through the following formula:

$$P_{SF} = (N \cdot N_T + M \cdot L) \cdot (P_{CMOS} + P_{LC}) \tag{3.4}$$

where P_{LC} is the power consumption of an electronic LC and P_{CMOS} is again the power consumption per port of an electronic CMOS-based electronic switch. Finally, P_{OC} includes the power consumption of the OAs only, since the WDM DeMux/Mux are passive components. In Table 3.1 the power consumption of all the elements introduced so far is reported. The values were obtained by collecting and averaging data from a number of commercially available components and modules of conventional switching and routing systems as well as from research papers. A more detailed explanation on how to compute the power consumption of the electronic core switch is given in Chapter 2.

3.1.2.2 HOS Architecture

The power consumption of the HOS network architecture is obtained through the following formula:

$$P_{Net}^{HOS} = N_T \cdot N \cdot P_{ToR} + N \cdot P_{Edge}^{HOS} + P_{Core}^{HOS} \tag{3.5}$$

where P_{Edge}^{HOS} is the power consumption of the HOS edge node and P_{Core}^{HOS} is the power consumption of the HOS core node. The power consumption of the HOS edge node is obtained by summing the power consumption of all the blocks shown in Figure 3.3:

$$P_{Edge}^{HOS} = N_T \cdot (P_{Cs} + P_{As} + P_{PE} + P_{CMOS}) + P_{RA} \tag{3.6}$$

where P_{Cs} is the power consumption of the classifier, P_{As} is the power consumption of the traffic assembler, and P_{PE} is the power consumption of a packet extraction module. To compute the power consumption of the classifier and assembler we evaluated the average buffer size that is required for performing

Table 3.1: Values of power consumption of the components within the optical ptp and the HOS data center networks.

Components	**Power [W]**
Top of the Rack Switch (P_{ToR})	650
Aggregation Switch	
Electronic Switch (P_{CMOS})	8
Line card (P_{LC})	300
Electronic Core Switch	
Control logic (P_{CL})	27,096
Optical Amplifiers (1 × port)	14
HOS Edge Node	
Classifier (P_{Cs})	62
Assembler (P_{As})	62
Resource Allocator (P_{RA})	296
Packet Extractor (P_{PE})	25
HOS Core Switch	
Control logic (P_{CL}^{HOS})	49,638
SOA switch (P_{SOA})	20
MEMS switch (P_{MEMS})	0.1
Tunable Wavelength Converter (1 × port)	1.69
Control Info Extraction/Re-insertion (1 × port)	17

correct classification and assembly. We obtained an average required buffer size of 3.080 MByte. The assembler and classifier are realized with two large FPGAs equipped with external RAM blocks for providing the total required memory size of 3.080 MByte. P_{RA} represents the power consumption of the resource allocator. Again, P_{CMOS} is the power consumption per port of an electronic CMOS-based electronic switch. The power consumption of the HOS core node is obtained by summing the power consumption of the control logic, switching fabric and other optical components:

$$P_{Core}^{HOS} = P_{CL}^{HOS} + P_{SF}^{HOS} + P_{OC}^{HOS} \tag{3.7}$$

Here, P_{CL}^{HOS} is the sum of the power consumed by the GMPLS module, the HOS control plane and the switch control unit. When computing P_{SF}^{HOS}, we assume that the optical ports of the fast and slow switches are switched off when they are inactive. This is possible because when two parallel switches are in use, only one must be active to serve traffic from a particular port at a specified time. In addition, because circuits and bursts are scheduled a priori, the traffic

arriving at the HOS core node is more predictable than the traffic arriving at the electronic core switch. We then compute the power consumption of the HOS switching fabric through the following formula:

$$P_{SF}^{HOS} = N_{fast}^{AV} \cdot P_{SOA} + N_{Slow}^{AV} \cdot P_{MEMS} \tag{3.8}$$

Here, N_{fast}^{AV} and N_{Slow}^{AV} are respectively the average number of active ports of the slow and fast switches obtained through simulations. P_{SOA} and P_{MEMS} are respectively the power consumption per port of the SOA-based and MEMS-based switches. The average number of active ports for a specific configuration is obtained through simulations. Finally, P_{OC}^{HOS} includes the power consumption of OAs, TWCs and CIE/R blocks. The values used for the power consumption evaluation of the HOS data center network are included in Table 3.1. A more detailed explanation on how to compute the power consumption of the HOS core node is given in Chapter 2.

3.1.3 HOS Data Center Simulation Setup

To evaluate the proposed HOS data center network we developed an event-driven C++ simulator. The simulator takes as inputs the parameters of the network and the data center traffic characteristics. The output produced by the simulator includes the network performance and energy consumption.

3.1.3.1 Data Center Traffic

In general traffic flowing through data centers can be broadly categorized into three main areas: traffic that remains within the data center, traffic that flows from data center to data center and traffic that flows from the data center to end users. Cisco [4] claims that the majority of the traffic is the one that resides within the data center accounting for 76% of all data center traffic. This parameter is important when designing the size of the data center and in particular the number of ports of the core node that connects the data center to the WAN. Basing on the information provided by Cisco, we designed our data center networks so that the number of ports connecting the core node to the WAN is 24% of the total number of ports of the core node.

In this paper we analyze the data center interconnection network, thus we simulate only the traffic that remains within the data center. To the best of our knowledge a reliable theoretical model for the data center network traffic has not been defined yet. However, there are several research papers that analyze data collected from real data centers [74–76]. Basing on the information collected in these papers, the inter-arrival rate distribution of the packets arriving at the data center network can be modeled with a positive skewed and heavy-tailed distribution. This highlights the difference between the data center environment and the wide area network, where a long-tailed *Poisson* distribution typically offers the best fit with real traffic data. The best fit [76] is obtained with the *lognormal* and *weibull* distributions that usually represent a good model for data center network traffic. We run simulation using both the lognormal and weibull distributions. In order to analyze the performance at different network loads, we considered different values for the mean and standard deviation of the lognormal distribution as well as for the shape and scale parameters of the weibull distribution.

In the considered data center networks, the flows between servers in the same rack are handled by the ToR switches and thus they do not cross the aggregation and core tiers. We define the *intra-rack traffic ratio (IR)* as the ratio between the traffic directed to the same rack and the total generated traffic. According to [74–76], the IR fluctuates between 20% and 80% depending on the data center category and the applications running in the data center. The IR impacts both performance and energy consumption of the HOS network and thus we run simulations with different values for the IR. The IR ratio has instead a negligible impact on the energy consumption of the optical ptp network. This is due to the fact that in the optical ptp network we do not consider switching off of the core switch ports when inactive and thus the power consumption is constant with respect to the network traffic characteristics.

In our analysis we set the number of blade servers per rack to 48, i.e. $N_S = 48$, that is a typical value used in current high-performance data centers. Although a single rack can generate as much as 48 Gbps, the ToR switches are connected to the HOS edge nodes by 40 Gbps links leading to an over-subscription ratio of 1.2. Over-subscription relies on the fact that very rarely servers transmit at their maximum capacity because very few applications require continuous communication. It is often used in current data center networks to reduce the

overall cost of the equipment and simplify data center network design. As a consequence, the aggregation and core tiers of a data center are designed to have a lower capacity with respect to the edge tier.

When simulating the HOS network, we model the traffic generated by the servers so that about 25% of the flows arriving at the edge nodes require the establishment of a circuit, 25% are served using long bursts, 25% are served with short bursts, and the remaining 25% are transmitted using packet switching. We do not to consider in this paper the impact of different traffic patterns, i.e. the portions of traffic served by circuits, long bursts, short bursts and packets. In fact, we already evaluated this effect for core networks in 2, where we showed that an increase in traffic being served by circuits leads to slightly higher packet losses and a more evident increase of burst losses. Since in this paper we employ the same scheduling algorithms as in 2, we expect a similar dependence of the performance on the traffic pattern.

3.1.3.2 Performance Metrics

In our analysis we evaluate the performance, scalability and energy consumption of the proposed HOS data center network.

As regards the performance, we evaluate the average data loss rates and the average delays. When computing the average loss rates, we assume that the ToR switches and HOS edge nodes are equipped with electronic buffers with unlimited capacity and thus they do not introduce data losses. As a consequence, losses may happen only in the HOS core node. The HOS core node does not employ buffers to solve data contentions in the time domain, but is equipped with TWCs for solving data contentions in the wavelength domain. We consider one TWC per port with full conversion capacity, i.e. each TWC is able to convert the signal over the entire range of wavelengths. We define the packet (burst) loss rate as the ratio between the number of dropped packets (bursts) and the total number of packets (bursts) that arrive at the HOS core switch. Similarly, the circuit establishment failure probability is defined as the ratio between the number of negative-acknowledged and the total number of circuit establishment requests that arrive at the HOS core switch.

The delay is defined as the time between a data packet is generated by the source server and when it is received by the destination server. We assume that

the IR traffic is forwarded by the ToR switches with negligible delay, and thus we analyze only the delay of the traffic between different racks, i.e. the traffic that is handled by the HOS edge and core nodes. The delay is given by the sum of the propagation delay and the queuing delay, i.e. $D = D_p + D_q$. The propagation delay D_p depends only on the physical distance between the servers. The physical distance between servers in a data center is usually limited to a few hundreds of meters, leading to negligible values for D_p. We then decided to exclude D_p from our analysis and consider $D = D_q$. The queuing delay includes the queuing time at the ToR switch and the delays introduced by the traffic assembler and resource allocator in the HOS edge switch ($D_q = D_{ToR} + D_{as} + D_{ra}$). The HOS optical core switch does not employ buffers and thus does not introduce any queuing delay. We refer to the packet delay as to the average delay of data packets that are transmitted through the HOS core node using packet switching. Similarly, we define the short (long) burst delay as the average delay of data packets that are transmitted through the HOS core node using short (long) burst switching. Finally, the circuit delay is the average delay of data packets that are transmitted through the HOS core node using circuit switching.

As regards the scalability, we analyze our HOS network for different sizes of the data center. In general, data centers can be categorized in three classes: university campus data centers, private enterprise data centers, and cloud computing data centers. While university campus and private enterprise data centers have usually up to a few thousands of servers, cloud computing data centers, operated by large service providers, are equipped with up to tens or even hundreds thousand of servers. In this paper we concentrate on large cloud computing data centers. As a consequence, we vary the data center size from a minimum of 25K servers up to a maximum of 200K servers.

As regards the energy consumption, we compute the total power consumed by the HOS and the optical ptp networks using the analytical model described in previous Section. To highlight the improvements introduced by our HOS approach, we compare the two architectures in terms of energy efficiency and total GHG emissions. The energy efficiency is expressed in Joule of energy consumed per bit of successfully transmitted data. The GHG emissions are expressed in metric kilotons (kt) of carbon dioxide equivalent (CO_{2e}) generated by the data center networks per year. To compute the GHG emissions, we apply the conversion factor of 0.356 $KgCo_{2e}$ emitted per KWh, which was found in [77].

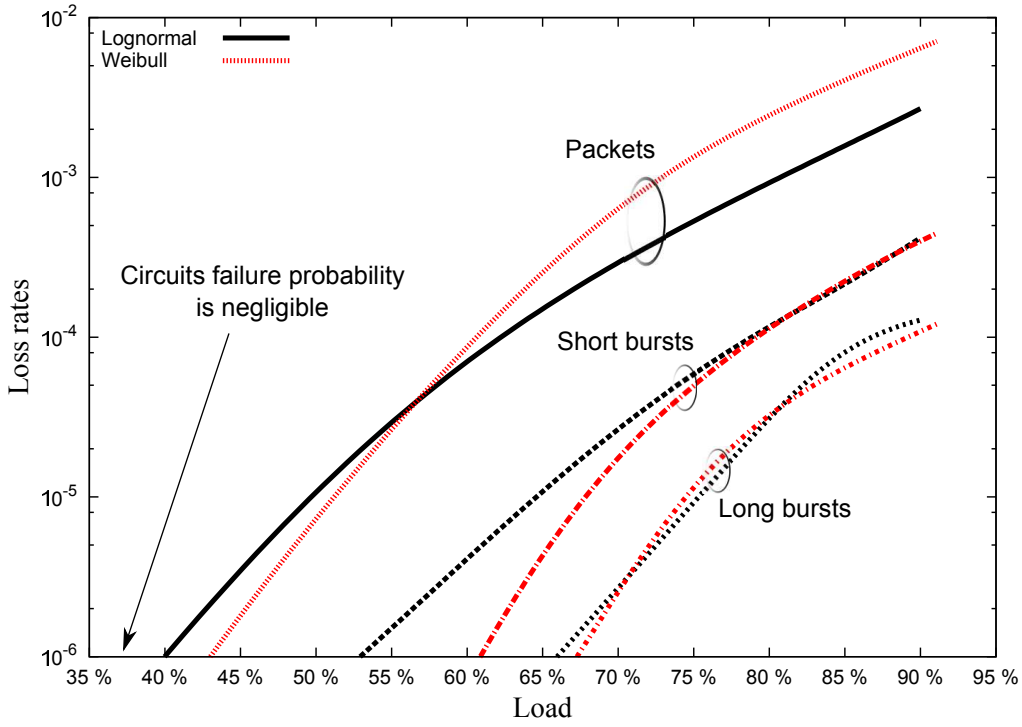

Figure 3.4: Average data loss rates in the HOS network as a function of the input load. Two different traffic distributions, i.e. lognormal and weibull, are considered.

3.1.4 Numerical Results

In this Section we show and discuss the results of the performed study. Firstly, we present the data loss rates, secondly, we report the network delays, and, finally, we analyze the energy consumption.

3.1.4.1 Loss Rates

In this Section we show and discuss the average data loss rates in the HOS network.

In Figure 3.4 we show the average data loss rates in the HOS network as a function of the input load. Two different distributions for the inter-arrival time of the traffic generated by the servers are considered, i.e. lognormal and weibull. We set $N_T = W = 64$ and $N = 32$ and considering that the number of severs per rack N_S is fixed to 48, the total number of servers in the data center is

98,304. The IR is set to 40%. Figure 3.4 shows that the data loss rates with the lognormal and weibull distributions present the same trend and very similar values. In the case of the weibull distribution the loss rates are slightly lower at low and medium loads, but they increase more rapidly with increasing the load. At high loads the loss rates obtained with the weibull distribution are similar or slightly higher than the loss rates obtained with the lognormal distribution. This effect is particularly evident for the packet loss probability, where the loss rates obtained with the two distributions are more different. Figure 3.4 also shows that the packet loss rates are always higher than the burst loss rates. This is due to the fact that for packets there is no resource reservation in advance. Due to shorter offset-times, the short bursts show higher loss rates with respect to long bursts, especially for low and moderate loads. Finally, we observe that the circuit establishment failure probability is always null. We conclude that data center applications having stringent requirements in terms of data losses can be mapped on TDM-circuits or long bursts, while applications that are less sensitive to losses can be mapped on optical packets or short bursts.

In Figure 3.5 the average data loss rates as a function of the IR are shown for two different values of the input load, namely 65% and 80%. The IR has been varied from 20% to 60%, while the number of servers in the data center has been set to 98,304. The input traffic distribution is lognormal. The Figure shows that the higher is the IR and the lower are the data loss rates. This is due to the fact that a higher IR leads to a lower amount of traffic passing through the core switch, thus leading to a lower probability of data contentions. While increasing IR from 20% to 60% the packet and short burst loss rates decrease respectively by two and three orders of magnitude. It can also be observed that the difference between the loss rates at 65% and 80% of input load becomes more evident at higher IRs. The circuit established failure probability is always null.

Finally, in Figure 3.6 we show the data loss rates as a function of the number of servers in the data center. Again, two values for the input load, namely 65% and 80%, are considered. The IR is set to 40% and the input traffic distribution is lognormal. When changing the size of the data center, we changed both the number of ToR switches per HOS edge node (N_T) and the number of HOS edge nodes (N). We always consider $N_T = W$, in order to have symmetric HOS edge nodes. As a consequence, the higher is N_T and the higher is the number of wavelengths in the WDM links. The number of servers per rack is fixed to

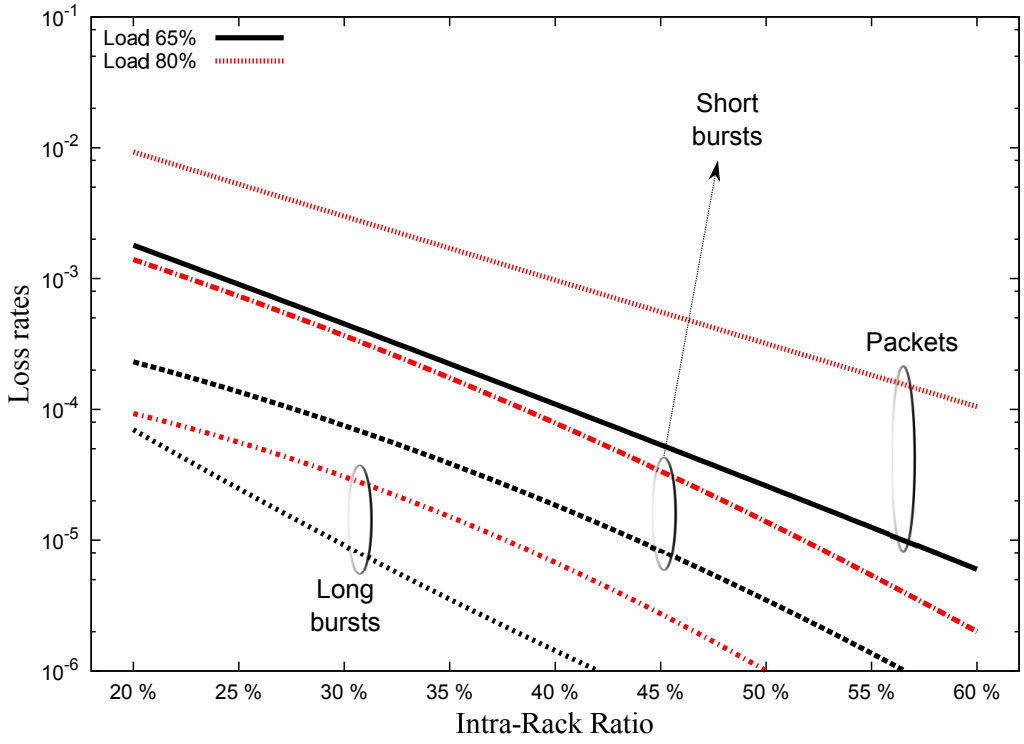

Figure 3.5: Average data loss rates in the HOS network as a function of the IR at 65% and 80% of offered load.

$N_S = 48$. The smallest configuration was obtained by setting $N = 22$ and $N_T = 24$, achieving a total number of 25,344 servers in the data center. The largest configuration was obtained by setting $N = 50$ and $N_T = 84$ achieving a total number of 201,600 servers. Figure 3.6 shows that the higher is the size of the data center network and the lower are the loss rates introduced by the HOS core node. This is due to the fact that in our analysis a higher data center size corresponds to a higher number of wavelengths per WDM link. Since the HOS core node relies on TWCs to solve data contentions, the higher is the number of wavelengths per fiber and the higher is the probability to find an available output resource for the incoming data. This is a unique and very important feature of our HOS data center network, that results in high scalability. In fact, increasing the number of wavelengths per fiber ($N_T = W$) we can scale the size of the data center while achieving an improvement in the network performance. Figure 3.6 shows that the loss rates, especially the loss rates of the long bursts, decrease by more than one order of magnitude while increasing the number of

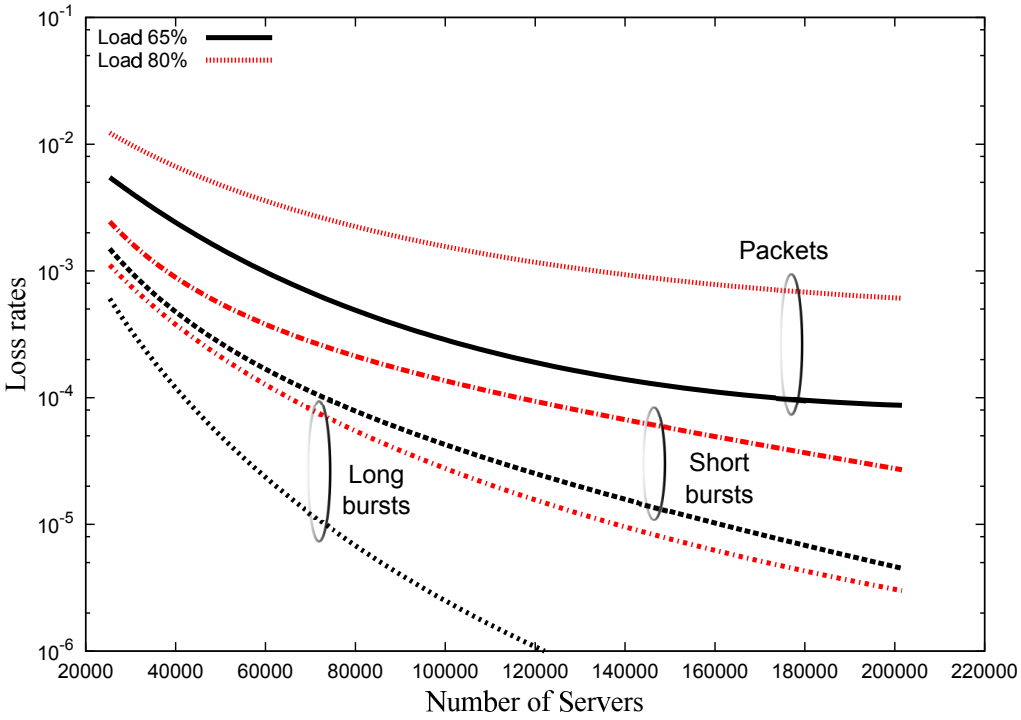

Figure 3.6: Average data loss rates in the HOS network as a function of the number of servers in the data center. Two different values for the input load, namely 65% and 80%, are considered.

servers from 25K to 200K.

3.1.4.2 Delays

In this Section we address the network latency. Since there are differences of several orders of magnitude between the delays of the various traffic types, we plotted the curves using a logarithmic scale.

In Figure 3.7 the average delays as a function of the input load are shown for two different distributions of the inter-arrival times of packets generated by the servers. The IR is set to 40% and the number of servers in the data center is 98,304. The Figure shows that the delays obtained with the lognormal and weibull distributions show the same trends. The largest difference is observed for the delays of packets at high input loads, with the delays obtained with the weibull distribution being slightly higher. Figure 3.7 also shows that circuits introduce the lowest delay. To explain this result let us recall the definition of

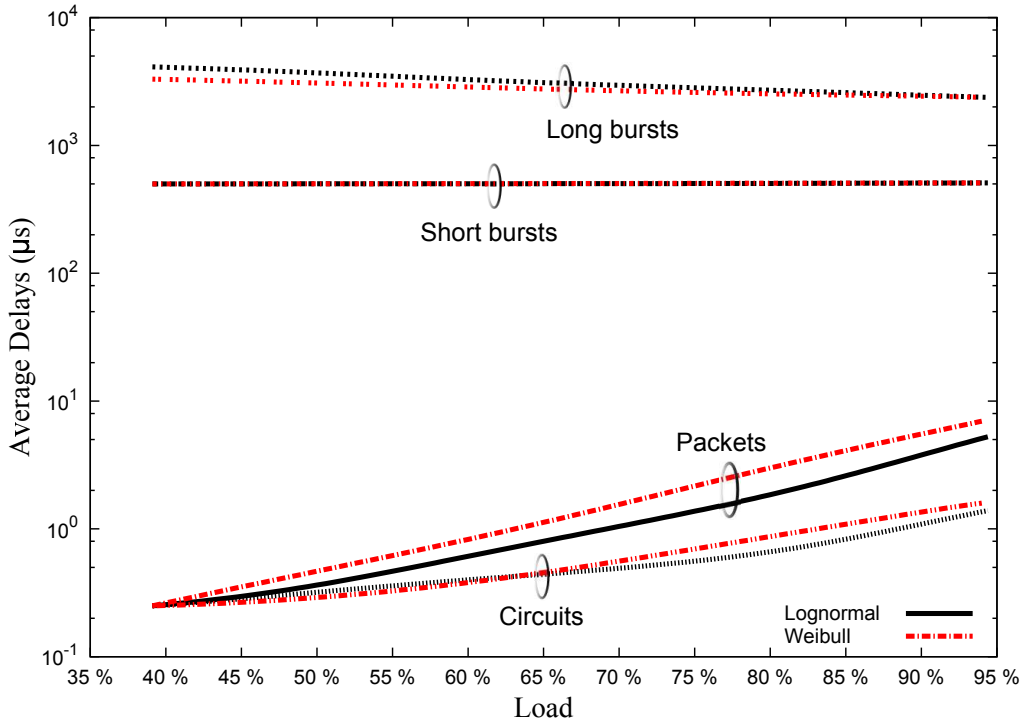

Figure 3.7: Average delays in the HOS network as a function of the input load and for two different input traffic distributions, namely lognormal and weibull.

end-to-end delay $D = D_{ToR} + D_{as} + D_{ra}$. For circuits the assembly delay D_{as} is related to the circuit setup delay. Since in our network the circuit setup delay is several orders of magnitude lower than the circuit duration, its effect on the end-to-end delay is negligible. Furthermore, circuits are scheduled with the highest priority by the resource allocator resulting in negligible D_{ra}. As a consequence, the circuit delay is determined mainly by the delay at the ToR switches D_{ToR}. As can be seen from Figure 3.7, circuits ensure an average delay below 1.5 μs even for network loads as high as 90% and thus are suitable for data center applications with strict delay requirements. Packets also do not suffer from any assembly delay, i.e. $D_{as} = 0$, but they are scheduled with low priority in the resource allocator resulting in non negligible values for D_{ra}. However, it can be observed that the packet delay remains below 1 μs up to 65% of input load. For loads higher that 65% the packet delays grow exponentially, but they remain bounded to a few tens of μs even for loads as high as 90%. We can then conclude that packets are suitable for the majority of todays delay-sensitive data center applications.

Short and long bursts are characterized by very high traffic assembler delays D_{as}, which are given by the sum of the time required for the burst assembly and the offset-time. The traffic assembler delay is orders of magnitude higher than D_{ToR} and D_{ra} and thus the end-to-end delay can be approximated with D_{as}. In order to reduce the bursts keep delays bounded we acted on the timers and the length thresholds of the burst assemblers. We optimized the short and long burst assemblers and strongly reduced the bursts delays. Still, short and long bursts delays are respectively one and two order of magnitude higher than packets delays, making bursts suitable only for delay-insensitive data center applications. Figure 3.7 shows that short bursts present an almost constant delay attested around 500 μs. Instead, the long burst delay decreases while increasing the input load. This is due to the fact that the higher is the rate of the traffic arriving at the HOS edge node and the shorter is the time required for reaching the long burst threshold L_{min} and starting the process for generating of the burst. The minimum long burst delay, which is obtained for very high input loads, is around 2 ms. This delay is quite high for the majority of current data center applications and raises the question if it is advisable or not to use long bursts in future data center interconnects. On the one hand long bursts have the advantage of introducing low loss rates, especially at low and moderate loads, and reducing the total power consumption, since they are forwarded using slow and low power consuming switching elements. On the other hand, it may happen that a data center provider does not have any suitable application to map on long bursts due to their high latency. If this is the case, the provider could simply switch off the long burst mode and run the data center using only packets, short bursts and circuits. This highlights the flexibility of our HOS approach, i.e. the capability of the HOS network to adapt to the actual traffic characteristics.

In Figure 3.8 we show the average delays in the HOS network as a function of the IR. We consider two values for the input load, namely 65% and 80%. The input traffic distribution is lognormal and the number of servers is set to 98,304. The Figure shows that the circuits and packets delay decrease while increasing the the IR traffic. This is due to the fact that the higher is IR and the lower is the traffic that crosses the ToR switches and the HOS edge nodes in the direction toward the HOS core node. This leads in turn to lower D_{ToR} and lower D_{ra}. In particular, when IR is as high as 60% the D_{ra} for packets becomes almost negligible and the packets delays become almost equal to the circuits delays. As for the long bursts, the higher is IR and the higher are the delays. In fact, a

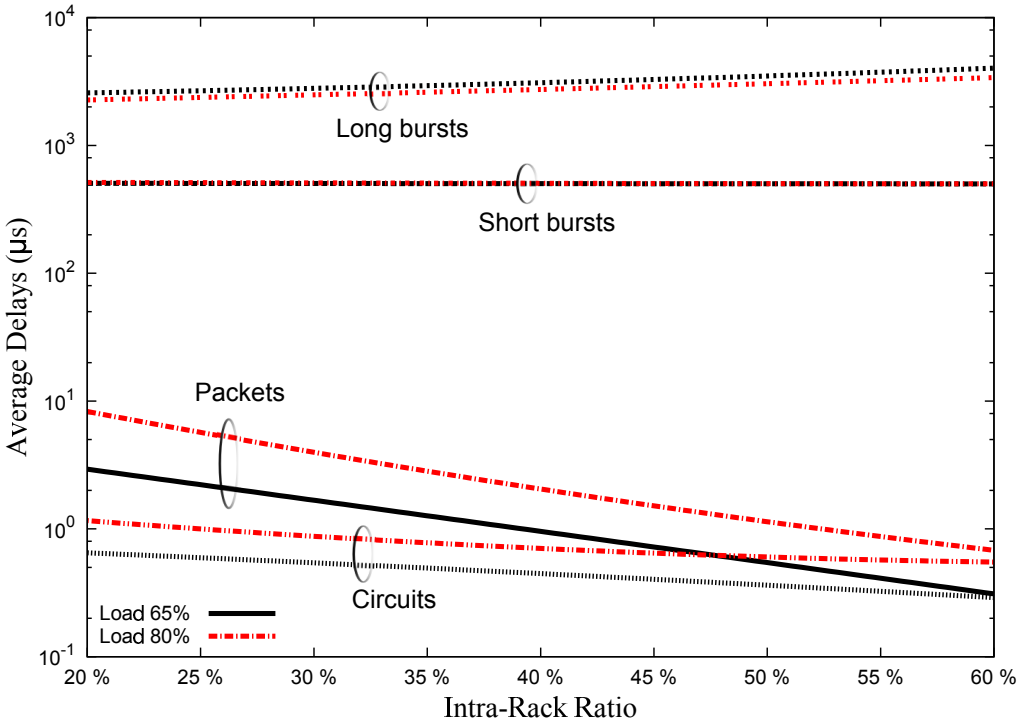

Figure 3.8: Average delays in the HOS network as a function of the IR at 65% and 80% of offered load.

higher IR leads to a lower arrival rate at the HOS edge nodes and, consequently, to a longer assembly delay D_{as}. Finally, the short burst delay is almost constant with respect to IR.

In Figure 3.9 we show the average delays as a function of the number of servers in the data center. We assume lognormal traffic distribution and IR equal to 40%. We consider input loads of 65% and 80%. The Figure shows that increasing the size of the HOS data center leads to a slight decrease of the end-to-end delays. To explain this fact it is worth remembering that when increasing the number of servers we also increase the number of wavelengths per fiber $W = N_T$ in the WDM links. The higher is N_T and the lower is the time required by the resource allocator to find an available output resource where to schedule the incoming data, i.e. the higher is N_T and the lower is D_{ra}. This fact again underlines the scalability of the proposed HOS solution.

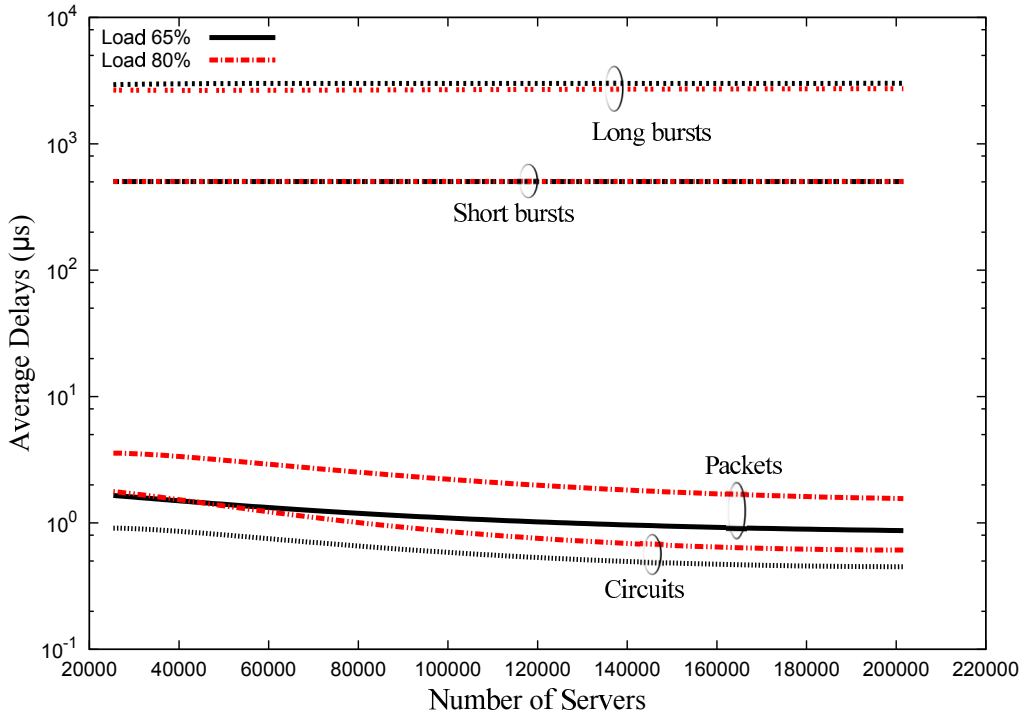

Figure 3.9: Average delays in the HOS network as a function of the number of servers in the data center. Two different values for the input load, namely 65% and 80%, are considered.

3.1.4.3 Energy Consumption

In this Section we present and compare the energy efficiency and GHG emissions of the HOS and the optical ptp data center networks.

In Figure 3.10 the energy consumption per bit of successfully delivered data is shown as a function of the input load. In the case of the HOS network we consider three different values for IR, namely 20%, 40% and 60%. The energy consumption of the optical ptp network is independent with respect to the IR. The number of servers in the data center is set to 98,304. Firstly, we consider the overall energy consumption of the data center network and thus we include in our analysis the power consumption of the ToR switches. The electronic ToR switches are the major contributor to energy consumption especially for the HOS network where they consume more than 80% of the total. In the optical ptp network ToR switches are responsible for around 50% of the total energy consumption. Figure 3.10 shows that the proposed HOS network provides energy

savings in the range between 31.5% and 32.5%. The energy savings are due to the optical switching fabric of the HOS core node that consumes considerably less energy with respect to the electronic switching fabric of the electronic core switch. Furthermore, the HOS optical core node is able to adapt its power consumption to the current network usage by switching off temporarily unused ports. This leads to additional energy savings especially at low and moderate loads when many ports of the switch are not used. However, the improvement in energy efficiency provided by HOS is limited by the high power consumption of the electronic ToR switches. In order to evaluate the relative improvement in energy efficiency provided by the use of HOS edge and core switches instead of traditional aggregate and core switches, we show in Figure 3.10 also the energy efficiency obtained without the energy consumption of the ToR switches. It can be seen that the relative gain offered by HOS is between 75% and 76%. The electronic ToR switches limit then by more than two times the potential of HOS in reducing the data center power consumption, raising the issue for a more energy efficient ToR switch design. Finally, Figure 3.10 shows that the energy efficiency of the HOS network depends only marginally on the IR traffic ratio. While increasing the IR ratio the energy consumption decreases because a higher IR ratio leads to a lower amount of traffic crossing the HOS core node. Due to the possibility of switching off unused ports, the lower is the amount of traffic crossing the HOS core node and the lower is its energy consumption.

Figure 3.11 shows the GHG emissions per year of the HOS and the optical ptp networks versus the number of servers in the data center. The IR is set to 40% and the input load is set to 65%. Again we show both the cases with and without the ToR switches. The Figure illustrates that the GHG emissions increase linearly with the number of servers in the data center. In both the cases with and without the ToR switches the GHG emissions of the HOS architecture are significantly lower than the GHG emissions of the optical ptp architecture. In addition, the slopes of the GHG emission curves of the HOS network are lower. In fact, while increasing the number of servers from 25K to 200K the reduction in GHG emissions offered by the HOS network increases from 30% to 32.5% when the power consumption of the ToR switches is included and from 71% to 77% when the power consumption of the ToR switches is not included. This is due to the fact that the power consumption of all the electronic equipment depends linearly on the size, while the power consumption of the optical slow switch does not increase significantly with the dimension. As a consequence,

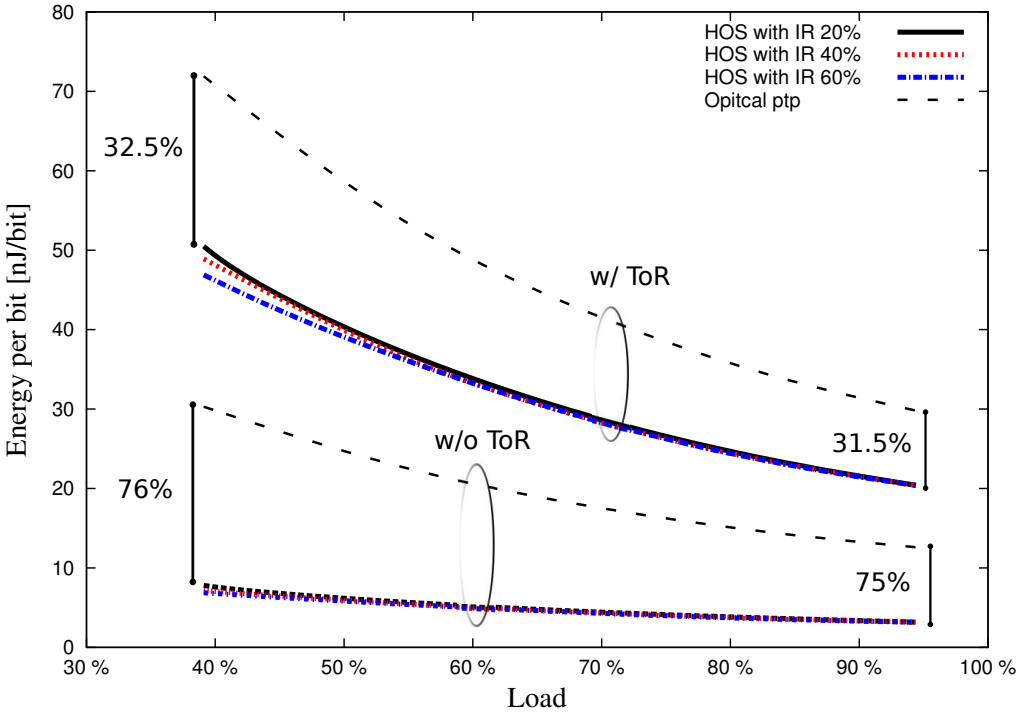

Figure 3.10: Energy consumption per bit of successfully transmitted data for the HOS and optical ptp networks. Both the cases with and without the ToR switches are shown.

the power consumption of the HOS core node increases slower than the power consumption of the electronic core switch. This leads to a higher scalability of the HOS network with respect to the optical ptp network. Figure 3.11 also shows that when including the energy consumption of the ToR switches the gain offered by the HOS architecture is strongly reduced, highlighting again the need for a more efficient ToR switch design.

3.1.5 Conclusions

With the expected increase in the data center traffic current data center networks based on electronic switching and ptp interconnects will raise issues in terms of flexibility, scalability, performance and energy consumption. As a consequence, several solutions based on optical switched interconnects, which make use of optical switches and WDM technology, have been recently proposed. However, the solutions proposed so far do not take into consideration the flexibility

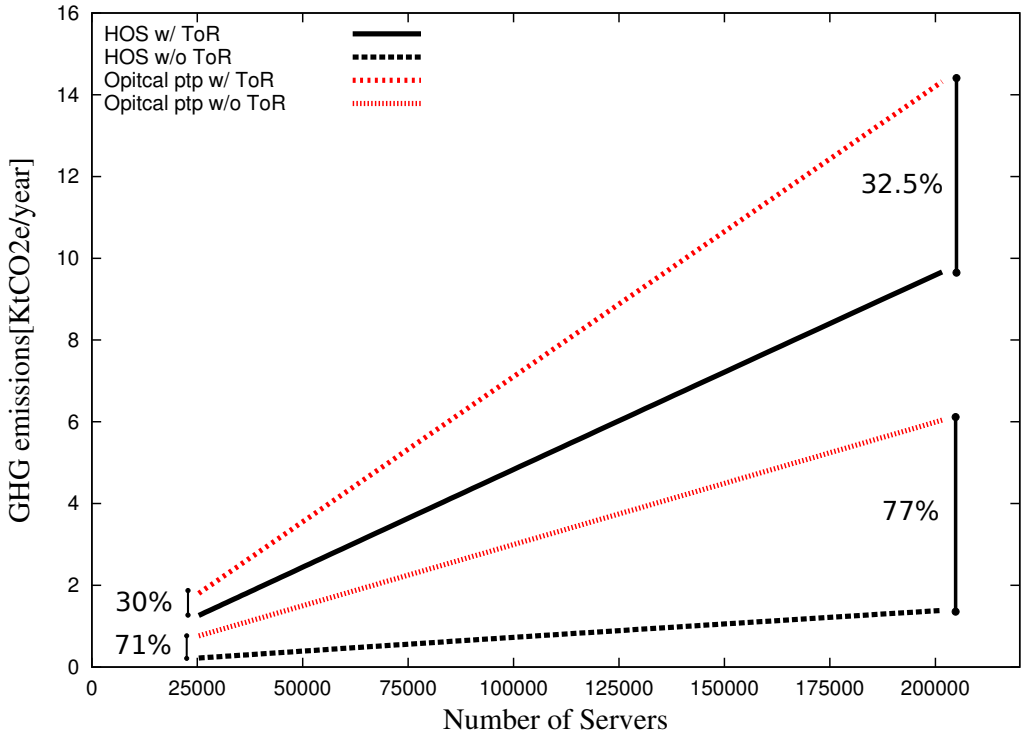

Figure 3.11: Greenhouse gas emissions per year of the HOS and the optical ptp networks as a function of the size of the data center.

needed to support future data centers applications in an efficient manner. Furthermore, only a few studies address scalability and energy consumption and make comparison with current ptp solutions.

In this Section, we proposed a novel optical switched interconnect network for data centers based on hybrid optical switching (HOS). HOS integrates optical circuit, burst and packet switching within the same network. Different data center applications are mapped to the optical transport mechanism that best suits to their traffic characteristics, ensuring high flexibility and efficient resource utilization. Furthermore, the proposed HOS network envisages the use of two parallel core optical switches, of which one is a slow and low power consuming switch for the transmission of circuits and long bursts while the other is a fast switch for the transmission of packets and short bursts. The use of two parallel optical switches enables switching off or putting in low power mode temporarily unused ports, thus providing high transmission efficiency and reduced power consumption. We presented both the architecture of a traditional data center network based on electronic switches and optical ptp interconnects

and the architecture of the proposed HOS network. We described an analytical model for the power consumption evaluation and we presented the simulation approach for evaluating the performance of the HOS network.

The proposed HOS network was evaluated in terms of average loss rates, average delays and energy consumption. The scalability of the HOS network was proved by changing the number of servers in the data center from 25K to 200K. Different values for the intra-rack traffic ratio (IR) were considered. The obtained results prove that the HOS network achieves relatively low loss rates that are suitable for today's data center applications. Circuits introduce negligible losses and are suited for premiere applications, while packets introduce the highest loss probability and are suited for best-effort traffic. As regards the delays, circuits and packets obtain the best performance and are suitable for delay-sensitive applications. On the contrary, bursts introduce relatively high delays and can be used for applications without stringent requirements in terms of latency. Finally, the results show that HOS is able to considerably increase the energy efficiency and reduce GHG emissions with respect to a conventional optical ptp architecture.

A main drawback is the capability of the proposed HOS interconnect to scale efficiently with the expected increase in server capacity. In fact, the proposed architecture has been studied for representing a short/mid term solution for operator, that want to save energy with limited investments and relatively small changes in their current data center infrastructure. In the long term, when the peak capacity per server will be much higher than today, more energy-efficient solutions will probably need to be defined. In the next Section we propose a novel architecture that will be able to satisfy the long term data centers requirements.

3.2 Elastic Optical Data Center

Driven by the rapid development of new cloud applications, network operators are continuously increasing the number and performance of their servers. As a consequence, data center traffic is expected to experience a tremendous increase in the next years. Furthermore, the server capacity is rapidly growing. While the majority of the severs operating today are based on 1 Gb/s network cards, the majority of last generation servers are equipped with 10 Gb/s ports and in the near future 40 Gb/s and 100 Gb/s ports are expected to become popular.

This poses a significant challenge to the networking of the data center and creates the need for highly efficient interconnection systems. In current data centers, servers are grouped in racks and are interconnected using electronic switches, typically organized in a fat-tree three-tiers topology. These architectures are not able to scale to meet the expected growth in data center traffic mainly because of their complexity, large number of required cables and high energy consumption. For this reason, in Section 3.1 we proposed a novel energy efficient optical switched interconnect architecture. However, the proposed architecture still rely on electronic top-of-rack (ToR) switches for interconnecting the servers within the same rack and for aggregating the traffic toward the aggregation/core tiers. Electronic ToR switches consume the largest amount of power in a current data center network and consequently limit the reduction in energy consumption that can be achieved by employing the HOS optical switched interconnect.

In this Section, we propose an all-optical data center network where the electronic ToR switches are replaced by an optical broadcast-and-select interconnect. This solution guarantees low energy consumption and, at the same time, provides high flexibility, facilitating the transmission of multicast traffic. The elastic optical network (EON) concept is employed for achieving high bandwidth utilization while meeting the demand for higher servers capacities. In Section 3.2.1 the design of the elastic optical data center will be illustrated. In Section 3.2.2 we introduce the power consumption model for its performance evaluation. Finally, in Section 3.2.3 we present some selected results and in Section we draw conclusions 3.2.4.

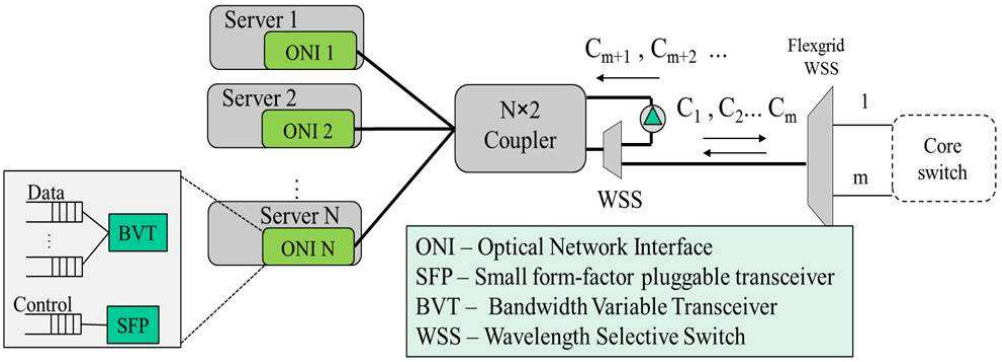

Figure 3.12: Optical interconnect at the ToR.

3.2.1 Elastic Optical Interconnect

In the following, we will refer to the proposed interconnect as elastic optical data center network (EODCN). In this Section, the data and control planes of the EODCN are presented. The data plane is composed of two tiers, where the first is given by the optical interconnects at the ToR and the second is given by the elastic optical core switch. The optical interconnects at ToR is shown in Figure 3.12. Here, optical network interfaces (ONIs) that are dedicated to different servers in the same rack, are connected to N input ports of an $N \times 2$ coupler, where N represents the number of servers in one rack. By passing through a flexgrid wavelength selective switch (WSS), which supports elastic channel spacing the signals of the channels assigned for intra-rack communications (e.g. C_{m+1}, C_{m+2}... shown in Figure 3.12) are sent back to the coupler and broadcasted to all the connected ONIs in the rack. In this way the multicast capabilities offered by the broadcast-and-select nature of the coupler. The channels assigned for inter-rack communications as well as the traffic directed to Internet (e.g. C_1, C_2,...,C_m) shown in Figure 3.12) is sent to (or received from) the core switch. The flexgrid WSS considered here could easily support flexible resource allocation for the traffic within the rack and the one from/to the outside of the rack. The ONIs are composed of three building blocks: the data and control queue management, the small form-factor pluggable transceiver (SFP), and the bandwidth variable transceiver (BVT). The SFP is used for transmitting/receiving the control information. The BVT transmits and receives data and provides wavelength tunability and flexibility of varying the number of the occupied spectral slots. Thus, the capacity of a link can be changed from 1 Gbps

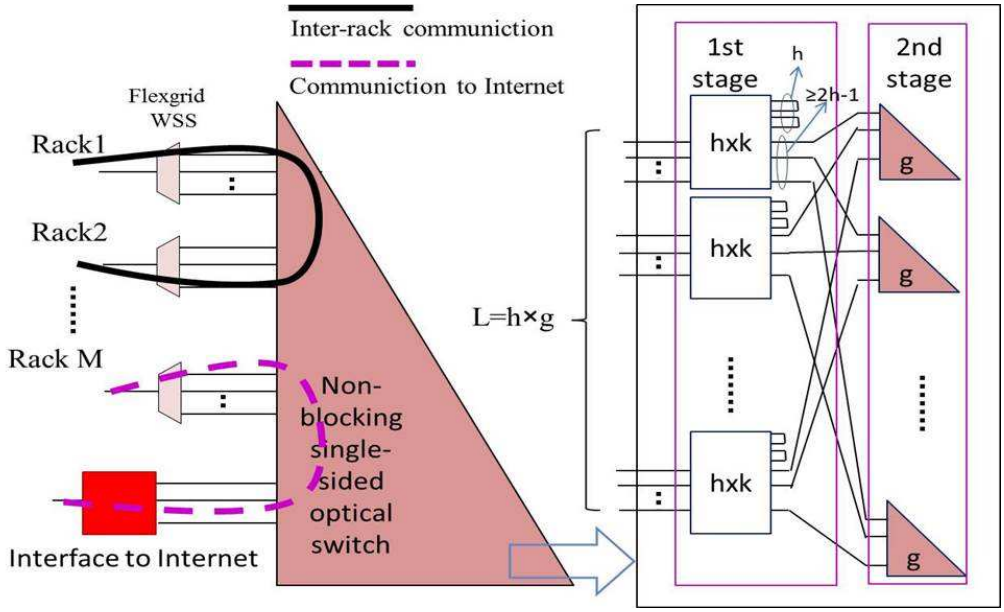

Figure 3.13: Elastic optical core switch.

to 100 Gbps upon request on a per-server level.

Figure 3.13 shows the proposed structure of the elastic optical core switch. The input ports of this core switch are directly connect to the output ports of the optical interconnects at ToR. In Figure 3.13, the elastic channels from each rack are first de-multiplexed by the flexgrid WSS and then pass a single-sided switch in order to be routed to the corresponding destinations (either another rack in the data center or an interface to Internet). Here, single-sided means each port of a switch can be connected to one of any other ports if available. Thus, it could significantly simplify the configuration (i.e. no need to distinguish the ports as input and output) compared to the conventional dual-sided optical switch, where each input port could only connect to the ports at the output side. Currently, single-sided switches are usually fabricated using the beam steering technology [78], where the maximum achieved number of switch ports so far is 192, while a 500-port matrix is under development. Larger single-sided switches can be realized by combining several stages of smaller switch matrices as shown in Figure 3.13, which depicts a large 2-stage switch. In this configuration, the switches in the first stage are hk dual-sided switches, while the second stage comprises single-sided switches with g ports. The total number of ports for this configuration is hg. In order to achieve strictly non-blocking feature, it should

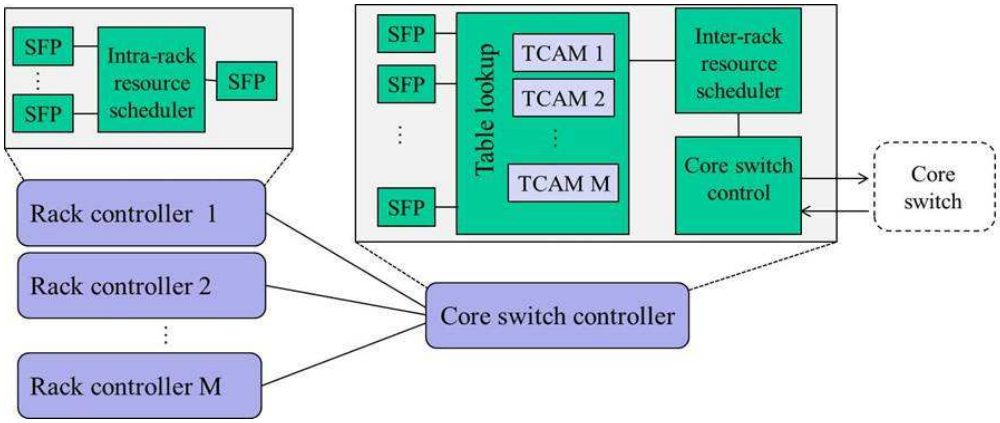

Figure 3.14: Control plane architecture of the elastic optical interconnect for data center.

be ensured that $k \geq 3h-1$. Here, h out of k output ports should be connected in pairs to provide internal interconnections within the switch, while the remaining h - k output ports are connected to the switches in the second stage, which are used to establish paths between the switches in the first stage. Note that all paths through the switch are bidirectional. It should also be noted that all switches used to implement the large optical core switch can be of the same size and type. The only limitation on the maximum size of the optical core switch is the degradation of signals passing several stages due to insertion loss and cross-talk.

The control plane architecture is shown in Figure 3.14. It consists of several rack controllers (RCs) and one core switch controller (CSC). Each RC is connected to N servers inside a rack and is responsible for assigning the time slots and the EON spectral slices to manage the intra-rack communications. SFPs are employed to transmit the signals between the RC and the servers, and between the RCs and the CSC. The CSC manages the spectrum resources inside the core switch and routes the traffic between different racks and to/from the Internet. The CSC is composed of: (*i*) search engines, using ternary content addressable memory (TCAM) for table lookup and path computation, (*ii*) inter-rack resource scheduler, using field programmable gate arrays (FPGAs) for managing EON spectral slices inside the core switch, and (*iii*) the switch control unit, for setting up the paths through the switch matrix.

3.2.2 Power Consumption Model

To evaluate the power consumption of the proposed EODCN we consider 12.5 GHz as the smallest spectrum slice. Using a single spectral slice a server can transmit data at either 1 or 10 Gb/s, according to the bandwidth requirement of the applications. When requested by the application and given that the network has enough available resources a server can be assigned 2 or 3 spectral slices dynamically for transmitting data at 40 Gbps or 100 Gbps, respectively. Therefore, in the proposed EODCN each server is always guaranteed to have at least one slice to transmit at 10 Gbps and can reach the peak capacity of 100 Gbps if available. To compute the total power consumption of the proposed EODCN (P_{EODCN}) we sum the power consumed at the ToR (P_{ToR}), the power consumed by the elastic optical core switch (P_{EOC}), and the power consumed by the control plane (P_{CP}), as shown in the following formula:

$$P_{EODCN} = P_{ToR} + P_{EOC} + P_{CP} \tag{3.9}$$

The power that is consumed by the broadcast-and-select matrices at the ToR is obtained through the following equation:

$$P_{ToR} = N_{server} \cdot (P_{QM} + P_{BVT} + P_{SFP}) + N_{rack} \cdot P_{WSS} \tag{3.10}$$

where N_{server} and N_{rack} are the total number of servers and the total number of racks in the data center, respectively. On the other side, P_{QM}, P_{SFP}, P_{BVT} and P_{WSS} are the power consumption of the queue management block, SFP, BVT and WSS, respectively.

The power consumption of the core switch is calculated according to the following equation:

$$P_{EOC} = N_{switch} \cdot P_{switch} \tag{3.11}$$

where $N_{switch}$and P_{switch} are respectively the number and power consumption of the switching matrices inside the elastic optical core switch. We assume that the switch matrices are equipped with 192 ports (where $3h - 1 \leq 192$) and are based on the beam steering technology. The number of switching matrices

Table 3.2: Power consumption of the network equipment.

Components	**Power [W]**
Data and control queue management (P_{QM})	3
Small form-factor pluggable transceiver (P_{SFP})	1
Bandwidth Variable Transceiver (P_{BVT})	8
Flexgrid Wavelength Selective Switch (P_{WSS})	15
Beam steering technology switch (192 ports) (P_{switch})	40
Intra-rack scheduler (FPGA) (P_{schI})	40
Search Engine(TCAM) (P_{SE})	4.5
Inter-rack scheduler (FPGA) (P_{schO})	40
Switch control unit (P_{SC})	300

depends on the number of required stages and is obtained through the following formula:

$$N_{switch} = \begin{cases} 1 & \text{if 1-stage} \\ \frac{L}{h} + 2h - 1 & \text{if 2-stage} \\ \frac{L}{h} + 2h - 1 \cdot (\frac{L}{h} + 2h - 1) & \text{if 3-stage} \end{cases} \quad (3.12)$$

The required number of stages is calculated according to the number of servers in the data center and the number of connections toward the Internet. In our calculations we vary the number of servers, while always assuming that 24% of the ports of the core switch are reserved to connect to the Internet, basing on the data reported in [4].

Finally, the power consumption of the control plane is given by the following formula:

$$P_{CP} = (N_{server} + 2 \cdot N_{rack}) \cdot P_{SFP} + N_{rack} \cdot (P_{schI} + P_{SE} + P_{schO}) + P_{SC} \quad (3.13)$$

where P_{schI} , P_{SE} , P_{schO} and P_{SC} are the power consumption of the intra-rack scheduler, search engine, inter-rack scheduler and switch control unit, respectively. To compute the total power consumption of the proposed EODCN we take into account the maximum values of commercially available network devices that are found in datasheets from different vendors (see Table 3.2).

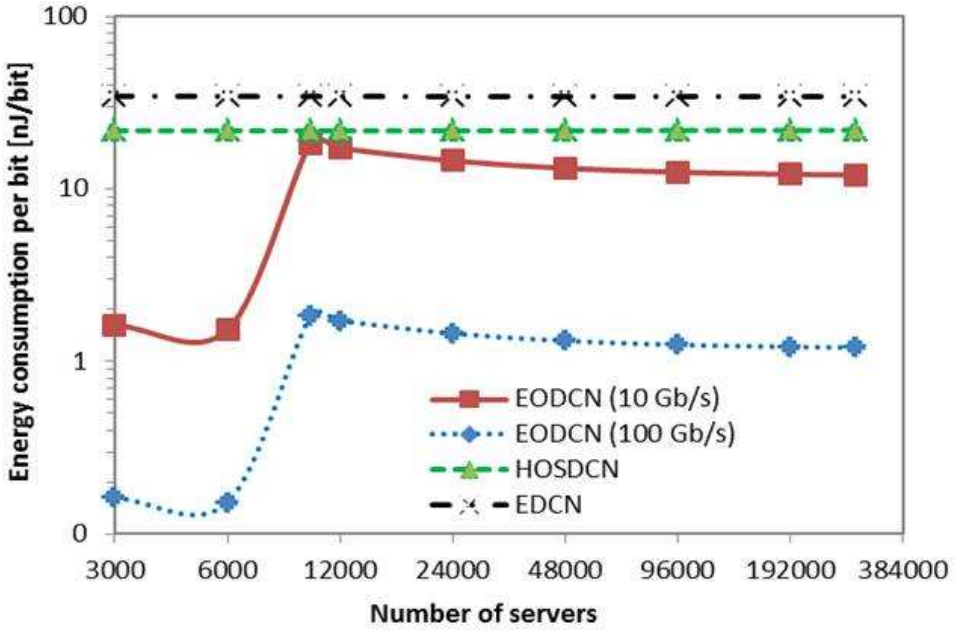

Figure 3.15: Energy consumption per bit of EODCN, HOSDCN and EDCN.

3.2.3 Numerical Results

In this section we evaluate the power consumption of the proposed EODCN and compare the results with a traditional electronic DC network (EDCN) and a hybrid optical DC network (HOSDCN) proposed in Section 3.1, which uses optical circuit and packet switching in core tier but still relies on electronic switches at the ToR. The EDCN and the HOSDCN guarantee a capacity of 1 Gb/s per server and their energy consumption model as well as the corresponding data for calculation are referred to Section 3.1.

Figure 3.15 shows the energy consumption per bit of the three considered DC interconnect architectures as a function of the number of severs within the DC. While the energy consumption per bit in EDCN and HOSDCN is almost independent on the number of servers, the EODCN shows a maximum when the number of servers is 10,000. This is due to the fact that taking into account the maximum size (i.e., 192 ports) of optical switch matrix based on the current beam steering technology, a 3-stage core switch structure is needed (replacing a 2-stage switch of Figure 3.12)) starting from 10,000 servers. In the beginning 3-stage

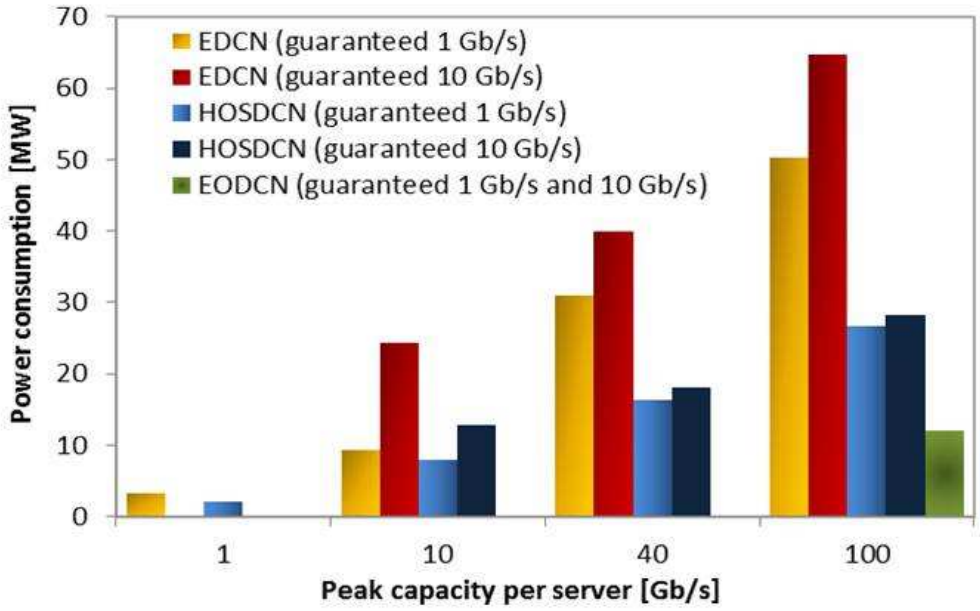

Figure 3.16: Total power consumption of EODCN, HOSDCN and EDCN.

structure is not efficiently utilized because of a low sharing factor. Increasing the number of servers leads to a higher amount of traffic crossing the core switch and hence the EODCN could achieve higher energy efficiency. We observe that if we consider the minimum guaranteed capacity per server, i.e., 10 Gb/s, the EODCN reduces the energy consumption by at least 4 nJ/bit compared with the HOSDCN and by at least 16 nJ/bit compared with the EDCN. On the other hand, if fully utilizing the maximum capacity, i.e. 100 Gb/s, the reduction of energy consumption per bit is obviously much higher.

Figure 3.16 shows the total power consumption for different values of the peak capacity per server with a fixed total number of servers (i.e., 96,000) in the data center. We scale the EDCN and HOSDCN by increasing the data rate at the servers to offer high peak capacity while dimensioning the switching capacity in the core tier according to the guaranteed bandwidth per server (where two cases, namely 1 Gb/s and 10 Gb/s, are considered). Figure 3.16 shows that to provide a peak capacity of 100 Gb/s per server the EODCN requires less than 1/2 the power of the HOSDCN and around 1/6 the power of the EDCN. It can be seen clearly that the EODCN is more beneficial when the guaranteed capacity required by the servers is higher.

3.2.4 Conclusions

In this Section we propose a novel data center network architecture that employs optical broadcast-and-select interconnects at the ToR and elastic optical core switch. The optical broadcast-and-select architecture at the ToR could significantly reduce the energy consumption compared to the conventional commodity switches, which currently takes the largest part of energy consumption in DC networks. Furthermore, the elastic optical core switch offers high flexibility and resource utilization using currently available optical components for telecommunication applications. The results have confirmed that the proposed overall architecture for data center networks consumes much less energy than the existing ones relying on electronic switches at the ToR, in particular for the case with high peak or guaranteed capacity per server, e.g., 10 Gb/s and beyond. We conclude that our solution is able to meet future data center requirements, with respect to the number of servers and bandwidth capacity, in a sustainable and effective manner. Furthermore, it should be noted that in this paper we present conservative results, which are obtained for the maximum energy consumption regardless of the traffic load. A proper strategy of dynamic power management could definitely further reduce the energy consumption. In our future work we plan to devise efficient and flexible resource allocation schemes tailored for the EODCN.

3.3 Carrier Cloud Network

Carrier cloud is a novel network model where data centers and core network are operated by the same entity, which then sell services to different operators through network virtualization techniques. Many companies are today considering to move to this network model [42]. Carrier cloud responds to the weaknesses of existing cloud solutions, including performance, availability, security and service level agreements [79]. A first attempt of an integrated control plane for intra-data-center and core networks for carrier cloud has been very recently proposed in [80]. This control plane is based on the Software Defined Network (SDN) concept. However, this work only shows preliminary results and much more needs to be done in order to define an efficient control and data plane for carrier cloud.

There are many efforts to improve energy efficiency in communication networks, ranging from the component technology to the architectural and service level approaches. In this Section, we address network energy efficiency at both the architectural and service levels and propose a unified network architecture that provides both intra-data-center and inter-data-center connectivity together with interconnection toward legacy IP networks. The architecture is mainly thought for future carrier cloud operators, which own both the intra-data-center and inter-data-center networks. It is based on the HOS concept for achieving high performance and energy efficiency and is referred to as integrated HOS network. The main advantage of the integration of core and intra-data-center networks comes from the possibility to avoid the energy inefficient electronic interfaces between data centers and telecom network.

At the service level, recent studies demonstrated that the use of distributed video cache servers can be beneficial in reducing energy consumption of intra-data-center and core networks. However, these studies only take into consideration conventional network solutions based on IP electronic switching, which are characterized by relatively high energy consumption. When a more energy efficient switching technology, such as HOS, is employed, the advantage of using distributed video cache servers becomes less obvious. In this Section we evaluate the impact of video servers employed at the edge nodes of the integrated HOS network to understand if a carrier cloud operator that rely on our network solution could be motivated to make use of edge caching.

To summarize, the contribution of this Section is twofold, i.e., (*i*) we propose and evaluate an integrated intra-data-center and core network architecture based on the HOS concept for application in carrier cloud along with a study of its benefits, and (*ii*) we assess the impact of distributed video cache servers on the proposed integrated HOS network architecture. The remainder of the Section is organized as follows. In Section 3.3.1 we describe the proposed integrated core and intra-data-center HOS network. Section 3.3.2 introduces the approach used to model the video cache servers. In Section 3.3.3 the energy consumption model is described. In Section 3.3.4 the the reference network used for the simulations is presented. Section 3.3.5 presents the simulation results, while Section 3.3.6 contains some concluding remarks.

3.3.1 Integrated HOS Network Architecture

Figure 3.17 shows a high level representation of the proposed integrated core and intra-data-center network based on HOS. The integrated network provides three different types of interconnections using a unified all-optical infrastructure and a common control plane. The first type of interconnection is between servers inside the same data center, referred to as an intra-data-center interconnection. In Figure 3.17 it is represented by a red dotted line to highlight the path over which data are sent using the HOS paradigm. The second type of interconnection is between servers located in different data centers. We refer to it as inter-data-center interconnection and we use a blue dashed line in Figure 3.17 to indicate the path performed in the HOS domain. The third type of interconnection is between servers inside a data center and HOS edge nodes, i.e., it provides the server-to-edge interconnections. An example is indicated in Figure 3.17 by a green solid line for the HOS path.

It should be noted that for all the considered types of interconnections the data remain in the optical domain, i.e., no O/E/O conversion takes place along all the paths established using the HOS paradigm. This is possible because in the proposed integrated network, core and data centers employ the same unified control plane. A first attempt of an integrated control plane for intra-data-center and core networks for carrier cloud has been recently proposed in [80]. Here, the authors create a proof of concept for an integrated control plane based on the SDN mechanism, but they don't evaluate the network performance nor they compare against a standard non-integrated solution. In this Section we propose

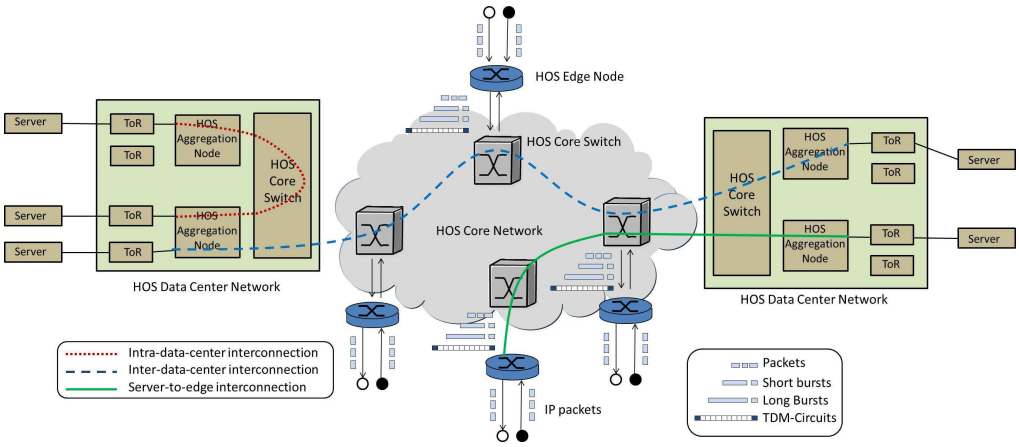

Figure 3.17: Optical interconnect at the ToR.

a novel integrated control plane for intra-data-center and core networks based on the HOS control model. It consists of two layers, the GMPLS control layer and the HOS forwarding layer. The GMPLS control layer is in charge of configuring and managing the network virtual topology. It consists of three building blocks: routing, signaling, and link management. The HOS forwarding layer performs data aggregation, data scheduling, and resource reservation. It supports three different optical transport mechanisms, i.e., circuits, bursts, and packets. The HOS forwarding layer has the unique feature of employing a common control packet for managing all three switching paradigms, enabling circuits, bursts, and packets to share dynamically the optical resources. The use of optical bursts in combination with packets and circuits allows the dynamic implementation of different service classes, leading to an efficient QoS differentiation.

3.3.1.1 HOS Core Network

The HOS core network provides connectivity among different data centers as well as between data centers and legacy IP networks. As shown in Figure 1, each node in the HOS core network includes a HOS core switch. If the node is located at the edge of the HOS core network, it is equipped with an electronic switch for inter-domain connectivity.

An electronic switch in the HOS edge node ensures interoperability between the core network and the legacy IP networks. In the direction toward the HOS

core network, the HOS edge node performs traffic classification and traffic aggregation. In other words, each incoming IP packet is classified based on the value of the differentiated service code point (DSCP) field in the IP header and mapped over the best suited optical transport mechanism. In the direction toward the legacy IP networks the HOS edge node extracts IP packets and performs IP routing. The HOS edge node is divided in two logical building blocks, one which consists of an electronic switch to perform IP routing and the second one which includes all the electronic components required to: (*i*) perform traffic aggregation and classification in the direction toward the HOS core network, and (*ii*) to perform IP packet extraction in the direction toward the legacy IP networks. For simplicity, we will refer to this block as the traffic aggregation block.

High capacity optical switches provide connectivity inside the core network. A HOS core switch can be logically divided in two building blocks, i.e., the electronic control logic and the optical switching fabric. The electronic control logic consists of three electronic blocks for implementing the GMPLS control layer, the HOS forwarding layer, and the switch control unit. The optical switching fabric is composed of two large optical switches. A fast optical switch, based on semiconductor optical amplifiers (SOA), takes care of the transmission of packets and short bursts. A slow optical switch, based on micro electro-mechanical systems (MEMS), handles the transmission of circuits and long bursts. In the optical switching fabric block we also include the following active optical components: optical amplifiers (OA), tunable wavelength converters (TWC), and control information extraction/re-insertion blocks (CIE/R), in order to compensate for signal losses in components, reduce blocking probability, and encode the control information together with the data payload on the same optical carrier, respectively.

For a detailed description of the HOS core network we refer to Chapter 2.

3.3.1.2 HOS Intra-Data-Center Network

The HOS intra-data-center network provides connectivity among the servers inside a data center and connects the data centers to the HOS core network. It is organized in a 3-tiers fat-tree topology. The first tier consists of electronic top-of-rack (ToR) switches. In a conventional high-end data center, servers are organized in racks, with each rack hosting typically 48 blade servers. The ToR

switches interconnect the servers inside a rack and connect the racks to the second tier of the intra-data-center network, which is composed of the HOS aggregation nodes. The HOS aggregation nodes perform the same functions inside a data center as HOS edge nodes in the HOS core network. In particular, in the direction toward the network core, the HOS aggregation nodes perform traffic classification and traffic aggregation, while in the direction toward the data center servers, the HOS aggregation nodes extract the IP packets and perform IP routing. The HOS aggregation nodes consist of the same logical building blocks as the HOS edge nodes. The main difference between HOS edge and the HOS aggregation nodes is that the HOS edge nodes could also include the video cache servers, which will be further elaborated in Section 3.3.2. The third tier of the intra-data-center network is represented by a single large HOS core node. This node has exactly the same architecture as the HOS core switch used in the core network. For more details we refer to Section 3.1.

3.3.2 Edge Caching

To evaluate the impact of distributed video cache servers on the proposed integrated intra-data-center and core HOS network, we extended the HOS edge node architecture described in Section II in order to include the video cache servers. The extended architecture of HOS edge nodes with cache servers is shown for the first time in Figure 3.18. It can be logically divided in three building blocks. Two of them have already been mentioned in Section 3.3.1, i.e., the electronic switch block and the traffic aggregation block. The former one includes the switch, the GMPLS module and the input electronic line cards while the latter block comprises the classifier, the conditioner, the assembler, the resource allocator, and the packet extractor. The last block is related to the caching operations and consists of the content tracker, the ToR switch, and the video cache servers. The content tracker is a novel network element which interacts with the HOS control plane in order to keep track of all the video content inside the cache servers, process the incoming video requests and update the cache servers.

The impact of distributed cache servers on the network energy efficiency has already been addressed in previous studies [81–86], mainly focused on electronic switched networks. The rationale behind these works is that distributed cache servers reduce the traffic load leading to the lower number of electronic

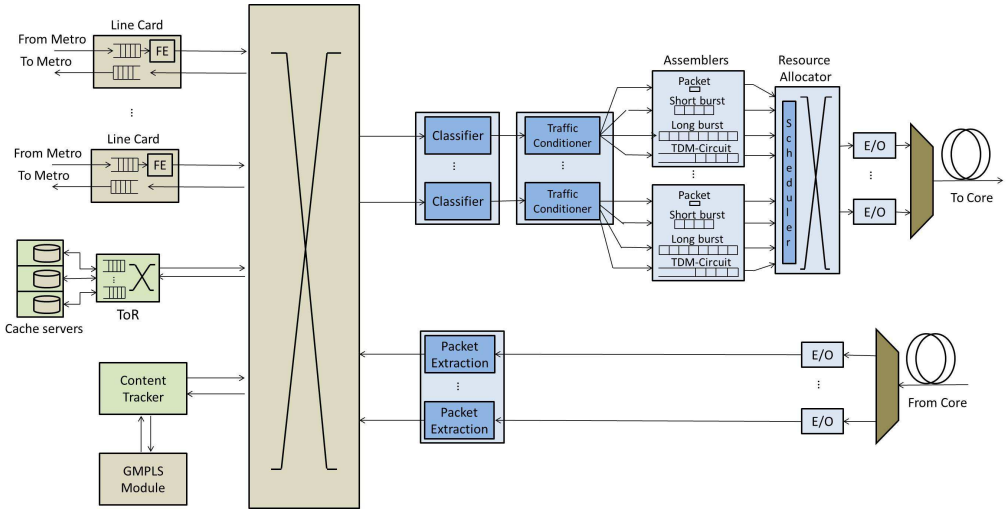

Figure 3.18: Optical interconnect at the ToR.

switch ports used in the core and intra-data-center networks. However, the electronic switching devices commercially available today do not implement dynamic switching-off of the ports and thus their energy consumption is almost independent on the traffic load. Techniques for dynamically switching-off the line-cards (LCs) have been proposed in [87], but their efficiency in real network scenarios has still to be proven. In fact, scheduling the switching-off the LCs in a packet switching network is a very challenging task because of the stochastic nature of the traffic and usually the inter-arrival time between two successive packets is very small. The novelty of our approach consists in applying the caching concept to HOS network, where we assume that all the optical components (in the optical switching fabric of the HOS core nodes) are turned off when they are inactive. This is not as challenging as turning off electronic switch ports. In fact, with two parallel optical switches, only one needs to be active to serve traffic from a particular port at a specified time. In addition, in a HOS network circuits and bursts are scheduled a priori, thus the incoming traffic is better predictable than in a traditional packet switched network, i.e., where the traffic is processed on a packet-by-packet basis.

3.3.3 Power Consumption Model

The total power consumption of the integrated core and intra-data-center HOS network is given by the sum of the power consumed by each node in the core

network (P^i_{Node}) and the power consumed by data centers (P^j_{DC}):

$$P_{Network} = \sum_{i=1}^{N_{Node}} P^i_{Node} + \sum_{j=1}^{N_{DC}} P^j_{DC} \tag{3.14}$$

where N_{Node} is the number of nodes in the core network and N_{DC} is the number of data centers. Each node in the HOS core network performs both edge and core functions. The power consumption of the *i-th* node in the network is determined by:

$$P^i_{Node} = P^i_{Edge} + P^i_{Core} \tag{3.15}$$

where P^i_{Edge} is the power consumption of the *i-th* HOS edge part if any and P^i_{Core} is the power consumption of the *i-th* HOS core switch. The power consumption of the *i-th* HOS edge part is given by the sum of the power consumption of its building blocks:

$$P^i_{Edge} = N_F^{Edge,i} \cdot N_W \cdot (P_{ES} + P_A) + P^i_{Cache} \tag{3.16}$$

where $N_F^{Edge,i}$ is the total number of fibers connected to the HOS edge node i and N_W is the number of wavelengths per fiber, which are assumed to be the same for all nodes. In the formula, P_{ES} is the power consumption of the electronic switch block per port and P_A is the power consumption of the traffic aggregation block per port. The number of ports of the switch is given by the product of the number of wavelength channels per fiber and the number of fibers ($N_F^{Edge,i} \cdot N_W$), i.e., it represents the total number of wavelength channels at a HOS edge node. Finally, P^i_{Cache} is the power consumption of the cache block. The power consumption of the cache block of the *i-th* HOS edge node is obtained through the following formula:

$$P^i_{Cache} = P_{CT} + P_{ToR} + N^i_{CS} \cdot P_{CS} \tag{3.17}$$

where P_{CT} is the power consumption of the content tracker, P_{ToR} is the power consumption of the ToR switch, and P_{CS} is the power consumption of a cache server. Finally, N^i_{CS} represents the number of cache servers hosted in the *i-th* HOS edge node. The cache servers are assumed to have a fixed storage

capacity of 1 TByte. Also the power consumption of the *i-th* HOS core switch is computed by summing up the power consumption of its building blocks, as defined by Equation 3.18:

$$P^i_{Core} = P^i_{ECL} + P^i_{OSF} \tag{3.18}$$

where P^i_{ECL} is the power consumption of the electronic control logic and P^i_{OSF} is the power consumption of the optical switching fabric of the *i-th* HOS core switch. The power consumption of the control logic of the *i-th* HOS core switch is given by Equation 3.19:

$$P^i_{ECL} = N_F^{Core,i} \cdot N_W \cdot P_{GMPLS} + P_{HOS} + P_{SC} \tag{3.19}$$

where $N_F^{Core,i}$ is the total number of fibers connected to the HOS core node *i*. In Equation 3.19, P_{GMPLS} is the power consumption of the GMPLS block per port, P_{HOS} is the power consumption of the HOS forwarding layer, and P_{SC} is the power consumption of the switch control unit. The power consumption of the optical switching fabric of HOS core nodes depends on the traffic because we assume that optical switch ports can be turned off when they are inactive. To compute the power consumption of the optical switching fabric of the *i-th* HOS core switch we use Equation 3.20:

$$\begin{aligned} P^i_{OSF} = N_{SOA}^{active,i} \cdot P_{SOA} + N_{MEMS}^{active,i} \cdot P_{MEMS} + \\ N_{TWC}^{active,i} \cdot P_{TWC} + N_F^{core,i} \cdot (N_W \cdot P_{CIE/R} + 2 \cdot P_{EDFA}) \end{aligned} \tag{3.20}$$

Here, $N_{SOA}^{active,i}$, $N_{MEMS}^{active,i}$ and $N_{TWC}^{active,i}$ represent the number of active SOA-switch ports, MEMS-switch ports, and TWCs of the *i-th* HOS core node, respectively. These values depend on the traffic load and are computed through simulations. In Equation 7, P_{SOA}, P_{MEMS}, and P_{TWC} are the power consumption of the SOA-switch per port, the MEMS-switch per port and the TWC, respectively. Finally, $P_{CIE/R}$ and P_{EDFA} are the power consumption of the CIE/R block and the OAs. When computing the power consumption of the HOS intra-data-center networks we exclude from our analysis the power consumed by servers and consider only the power consumed by the network equipment, i.e., by the intra-data-center network. The power consumption of the *j-th* intra-data-center network is computed using Equation 3.21:

Table 3.3: Power consumption of the network components in the integrated intra-data-centers and core network with edge caching.

Components	**Power [W]**
Electronic switching block per port (P_{ES})	320
Traffic aggregation block per port (P_A)	159
Content tracker (P_{CT})	330
Top-of-Rack switch (P_{ToR})	650
Cache server (P_{CS})	450
GMPLS control layer per port (P_{GMPLS})	6.75
HOS forwarding layer (P_{HOS})	570
Switch control unit (P_{SC})	300
SOA switch per port (P_{SOA})	20
MEMS switch per port (P_{MEMS})	0.1
Tunable WC (P_{TWC})	1.69
Control information E./R. ($P_{CIE/R}$)	17
Optical amplifiers (P_{EDFA})	14

$$P^j_{DC} = N^j_{ToR} \cdot P_{ToR} + N^j_{Aggr} \cdot P_{Agrr} + P^j_{Core} \tag{3.21}$$

where N^j_{ToR} and N^j_{Aggr} are the numbers of ToR switches and HOS aggregation switches in the *j-th* data center, respectively. Here, P_{Aggr} represents the power consumption of an HOS aggregation switch and P^j_{Core} represents the power consumption of the HOS core switch inside the *j-th* data center. We assume that each HOS aggregation switch is connected to the corresponding HOS core switch in the data center using one fiber and that the number of ToR switches connected to the corresponding aggregation node is equal to the number of wavelength channels per fiber (N_W). To calculate the power consumption of a HOS aggregation switch we use Equation 3.22:

$$P_{Aggr} = N_W \cdot (P_{ES} + P_A) \tag{3.22}$$

Finally, the power consumptions of the HOS core switch inside the *j-th* data center is computed using Equation 3.18 and replacing the index i with the index j. The energy consumptions of all the considered network components are reported in Table 3.3 and have been obtained by collecting data from data sheets as well as from research papers.

3.3.4 Carrier Cloud Network Simulation Setup

In this Section, we describe assumptions used to model and evaluate performance of the proposed integrated intra-data-center and core HOS network with edge caching. First, we present the reference network scenario, and secondly, we introduce the performance metrics.

3.3.4.1 Reference Network Scenario

To assess the performance of the proposed integrated intra-data-center and core HOS network with edge caching, we developed a custom event-driven C++ simulator. In the following we report the main parameters that we used in our simulations and present the model that is applied to generate the network traffic.

We denote N_{Node} as the number of nodes in the network and N_{DC} as the number of nodes connected to the data center. We consider the Pan-European network [72] composed of 28 nodes (i.e., $N_{Node} = 28$) and 41 links as the reference network topology. We assume that 25% of the network nodes are connected to a data center, i.e. $N_{DC} = 7$. In each simulation we randomly connect the data centers to different nodes of the network. We assume that all the data centers have the same size and are equipped with 76,800 servers organized in racks. In each rack, 48 servers are connected to a ToR switch using dedicated 1 Gbps links. The number of ToR switches per data center is given by the ratio between the number of servers and the number of racks, i.e., $N^j_{ToR} = N_{ToR} = 1600 \; \forall j \in N_{DC}$. As many as 64 ToR switches are connected to a HOS aggregation switch using 40 Gbps links. We obtain that each data center is equipped with $N^j_{Aggr} = N_{Aggr} = 25 \; \forall j \in N_{DC}$ HOS aggregation nodes. Each HOS aggregation node is connected to the HOS core node inside the data center using one fiber. The HOS core switch inside a data center is equipped with 25 fiber ports, for interconnecting all the HOS aggregation switches. In addition it employs 7 fiber ports, for the interconnection toward the Pan-European network. Thus, it has in total 32 fiber ports. The number of fiber ports for the interconnection between a data center and the Pan-European network has been chosen according to [4], where it is reported that currently 76% of the traffic generated inside a data center is directed to a server within the same data center (internal traffic). We assume that each core node in the Pan-European network also provide edge functionality.

As described before, each data center is connected to a network node of the Pan-European network using 7 fibers. To assure that the network nodes have enough capacity to support the connection toward a data center without becoming a bottleneck, we assume that each link in the network is composed of 4 fibers. We also assume that each HOS core node is connected to the corresponding HOS edge node using a number of fibers that is equal to the node degree. As a result, the number of fibers attached to the i-th HOS edge node ($N_F^{Edge,i}$) is equal to the node degree. The number of fibers connected to the i-th HOS core node ($N_F^{Core,i}$) is equal to five times the node degree (four times the node degree for the interconnection toward other HOS core nodes and one time the node degree for the interconnection toward the HOS edge node), plus 7 fibers in case that the HOS core node is directly connected to a data center. Each fiber carries 64 wavelength channels ($N_W = 64$), each of which is operated at 40 Gbps. As for the edge caching, we assume that the network nodes that are not directly connected to a data center are equipped with the same number of cache servers ($N_{CS}^i = N_{CS}\ \forall i \in N_{Node}$). The network nodes that connect data centers with the HOS core network do not comprise any cache servers.

The cache size of a HOS edge node is defined as the sum of the storage capacities, expressed in Bytes, of all the video servers hosted in the node. Furthermore, we define the video content hit rate as the probability that a video request arriving at a HOS edge node determines a cache hit and thus is served using the local cache servers. The cache hit rate depends mainly on the cache size. Several studies evaluate how the video content hit rate in real network settings is related to the cache size [88, 89]. The results of these studies show that high video hit rates can be achieved even with small cache sizes and that the cache hit rate exhibits a logarithmic grow as a function of the cache size. As a consequence, increasing the cache size over a certain value has a limited impact on the video content hit rate. In particular, the authors in [89] study the video content hit rate by varying the cache size at an edge node of a large core European network and obtain a curve of the cache hit rate as a function of the cache size. The cache hit rate estimation in our study is based on this curve, i.e. given a cache size the cache hit rate is defined according to the log-like curve presented in [89].

The IP traffic arriving at the HOS edge nodes from legacy IP networks is modeled using Poisson distribution. We assumed that 57% of this traffic consists

of requests for video content [3]. A request for video content can be either served locally by the video cache servers in the HOS edge node if the required content is available in the cache, or can be forwarded to the original server located in one of the data centers. In our simulations we also take into account the possibility that some of the traffic that arrives to a HOS edge node is destined to another network node, i.e., not to the data center. We refer to this traffic as cut-through traffic. Even if it is not directly related to our analysis, the cut-through traffic is important because it has an impact on the data losses and the delays as well as on the energy consumption. For the traffic generated by the servers, we implemented a more complex traffic model. According to [73], the inter-arrival rate distribution of the packets generated inside a data center can be modeled using a lognormal distribution. We then model the servers as finite-state machines with two states, namely lognormal state and video-transfer state. In the lognormal state, the servers generate IP packets with a lognormal distributed inter-arrival time. The IP packets generated by the servers in the lognormal state can be addressed either to a server in the same data center, or to a server in a different data center, or to a specific legacy IP network connected to a HOS edge node. When a server receives a request for a video content from an edge node, it switches to the video-transfer state. In the video-transfer state the server transmits IP packets at a constant bit-rate to the requesting HOS edge node. When all the video content has been transmitted, the server automatically switches back to the state with the longnormal inter-arrival rate distribution. The size of the video contents is uniformly distributed between 100 MByte and 500 MByte.

3.3.4.2 Performance Metrics

The performance of the proposed integrated intra-data-center and core network architecture based on HOS is assessed in terms of energy consumption, average delay, and average data loss. The energy consumption is measured in terms of Joule per bit (J/b) and it is computed as the ratio between the total network power consumption in Watts and the total network throughput in bit per second.

The delay is defined as the time difference between an IP packet is generated (i.e., either by a server in a data center or a cache server in a HOS edge node or a user of a network connected to edge HOS nodes) and the time the IP packet is received (i.e., either by the destination server or the destination HOS

edge node). The global average network delay is defined as the mean value of the delays over all IP packets measured during a simulation run. The IP packets that traverse the HOS network can be carried over different transport mechanisms. We refer then to the packet delay as the delay experienced by IP packets that are transmitted as optical packets through the HOS network. Similarly, the short burst delay, long burst delay, and circuit delay are the delays experienced by IP packets that are transmitted through the HOS network over a short burst, a long burst, or a circuit, respectively.

While computing the data loss rates, we assume that all the electronic switches introduce negligible losses. As a consequence, the losses in the core and intra-data-center networks may happen only in the HOS core switches. We define the packet loss rate as the ratio between the number of optical packets that are lost along a path through the HOS network and the total number of generated packets. Similarly, the short burst and the long burst loss rates are defined as the ratio between the number of lost and the number of generated short and long bursts, respectively. Circuits are established using a two-way reservation mechanism and consequently the data transmitted over circuits do not experience any losses. However, in heavily loaded networks a circuit establishment request could be refused (i.e., blocked) by a core node. As a consequence, we define the circuit establishment failure probability as the ratio between the number of blocked and the number of generated circuit.

We evaluate the above mentioned performance for different values of the network load. We define the load as the ratio between the total amount of traffic offered to the network by external sources (servers and legacy IP networks) and the maximum amount of traffic that can be handled by the network, i.e., the network capacity.

3.3.5 Numerical Results

This Section presents a performance analysis of the proposed integrated intra-data-center and core HOS network architecture with edge caching. First we comment on the benefits that a carrier cloud operator can achieve in employing the integrated HOS network when compared to a non-integrated HOS network and a conventional IP network, where core switches in a data center and core network are based on electronic switching. Then, we present and discuss the impact of using distributed cache servers on an integrated HOS network.

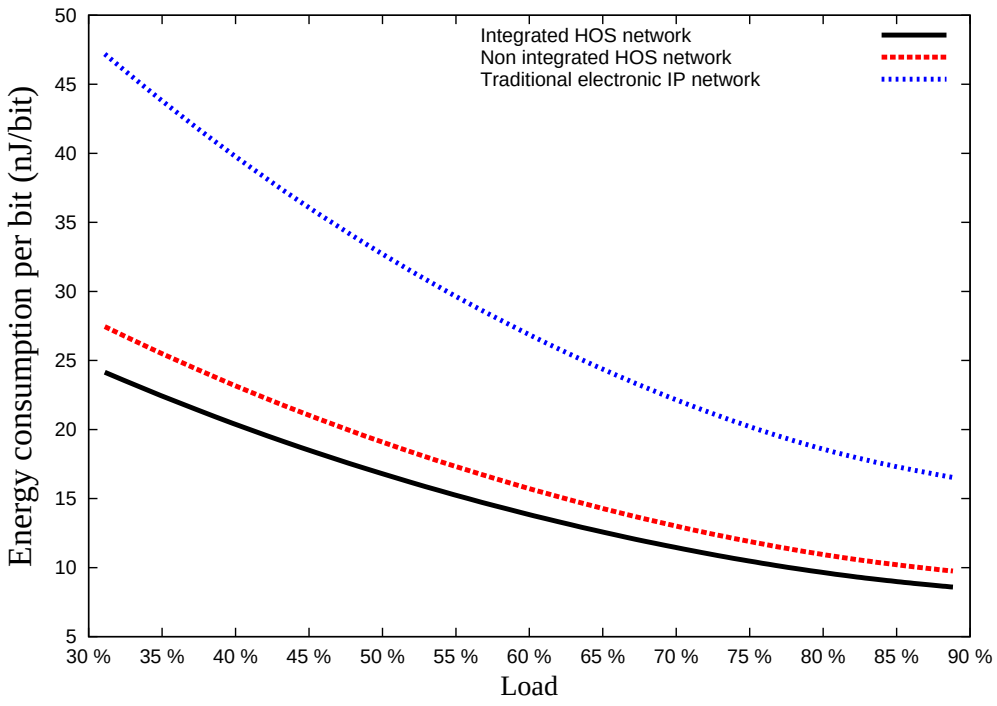

Figure 3.19: Energy consumption per bit as a function of the input load. Overall for core and intra-data-center networks.

3.3.5.1 Integrated HOS Network

To understand the results shown in this Section, in the following we explain in detail how the intra-data-center and core networks are interconnected in the non-integrated HOS architecture. In the non-integrated HOS network the data centers are equipped with additional HOS aggregation switches used for the interconnection toward the core network. Meanwhile, the core network is equipped with additional HOS edge nodes dedicated to the interconnection toward the data centers. Let us consider first the direction from the data center to the core network. Optical data exiting from the HOS core switch inside the data center are: firstly, de-assembled using the HOS aggregation switches, and then, re-assembled in the HOS edge nodes and according to the policies used in the core network. Similarly, in the direction from the core to the data center, optical data are: firstly, de-assembled using the HOS edge nodes, and then, re-assembled using HOS aggregation switches and according to the policies used in the data center.

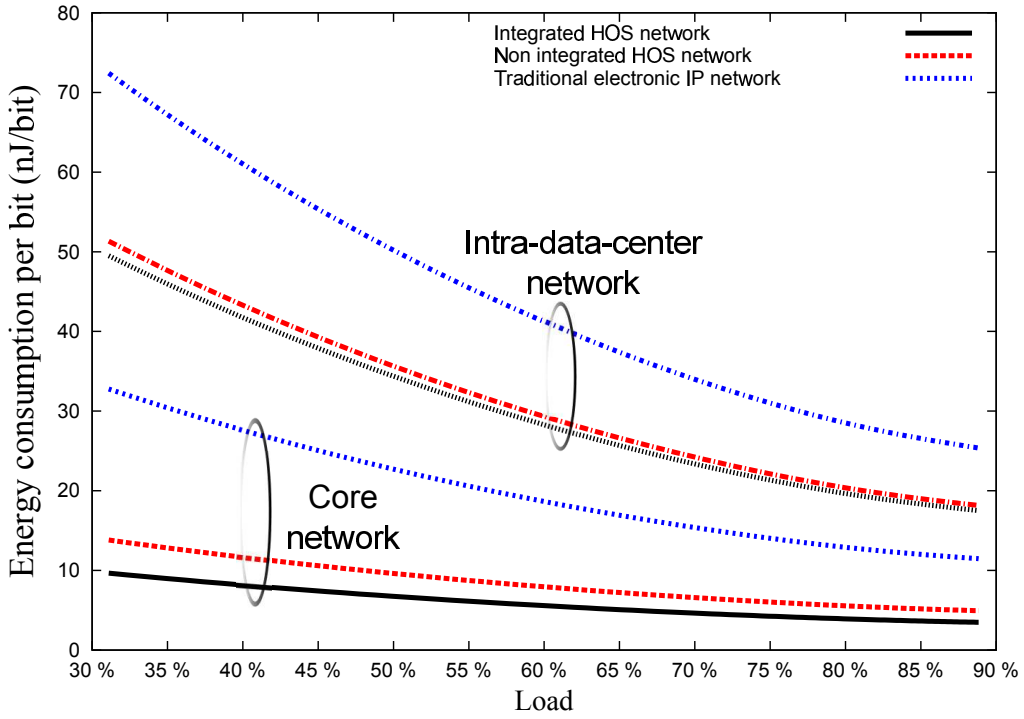

Figure 3.20: Energy consumption per bit as a function of the input load. Core and intra-data-center network shown separately.

In Figure 3.19 and 3.20, we compare the energy consumption per bit as a function of the network load for the integrated HOS network, the non-integrated HOS network, and a conventional IP network. The non-integrated HOS network has the same architecture as the integrated one, with the only difference in the use of electronic interfaces for the interconnection between the data centers and the core network, i.e., in our case the Pan-European network. On the other hand, the conventional IP network has a core and an intra-data-center network based on electronic switching. In our simulations, we assumed that all the network nodes which are not directly connected to a data center are equipped with N_{CS} = 10 cache servers, resulting in a total cache size of 10 TByte.

In Figure 3.19 we show the overall energy consumption per bit. The Figure shows that both integrated and non-integrated HOS networks drastically reduce the energy consumption of a conventional electronic IP network. The HOS networks consume roughly a half the energy per bit compared to a conventional IP network. This is due to the fact that the HOS networks employ an energy-efficient optical switching technology that benefits from transmitting circuits and

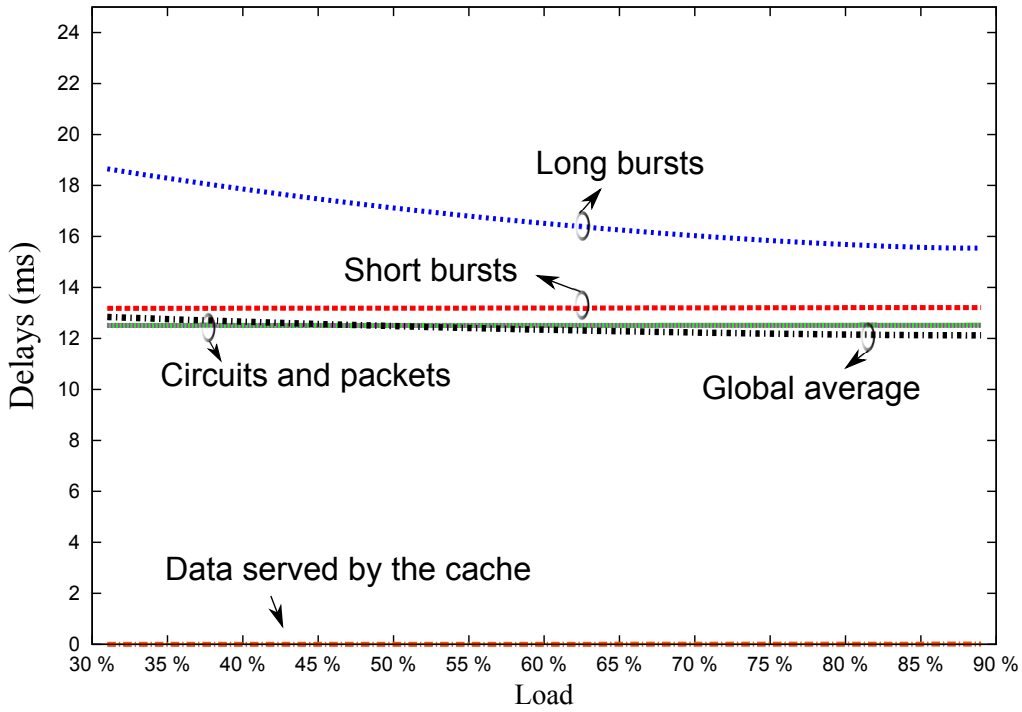

Figure 3.21: Average network delays as a function of the input load. Integrated HOS network.

long bursts using a slow and low power consuming optical switch, while using a less number of fast optical switches for the transmission of packets and short bursts. Figure 3(a) also shows the improvement in energy efficiency offered by the integrated HOS network with respect to the non-integrated HOS network. This increment in energy efficiency may seem small, but it is worth noting that at very high amount of network traffic, such as assumed here, even a reduction of few nJ/b can result in significant overall energy savings. For instance, at a network load of 35% the integrated HOS network consumes 4 nJ/b less than the non-integrated HOS network. This translates into a total of almost 2 MegaWatt (MW) saved.

In Figure 3.20 we show separately the energy consumption per bit in the core network and the intra-data-center networks. The energy consumption per bit of the core network is given as the ratio of the core network power consumption and the core network throughput. In the non-integrated HOS network the power consumption of the core includes the HOS edge nodes dedicated to the interconnection toward the data centers. Similarly, the energy consumption per bit of the

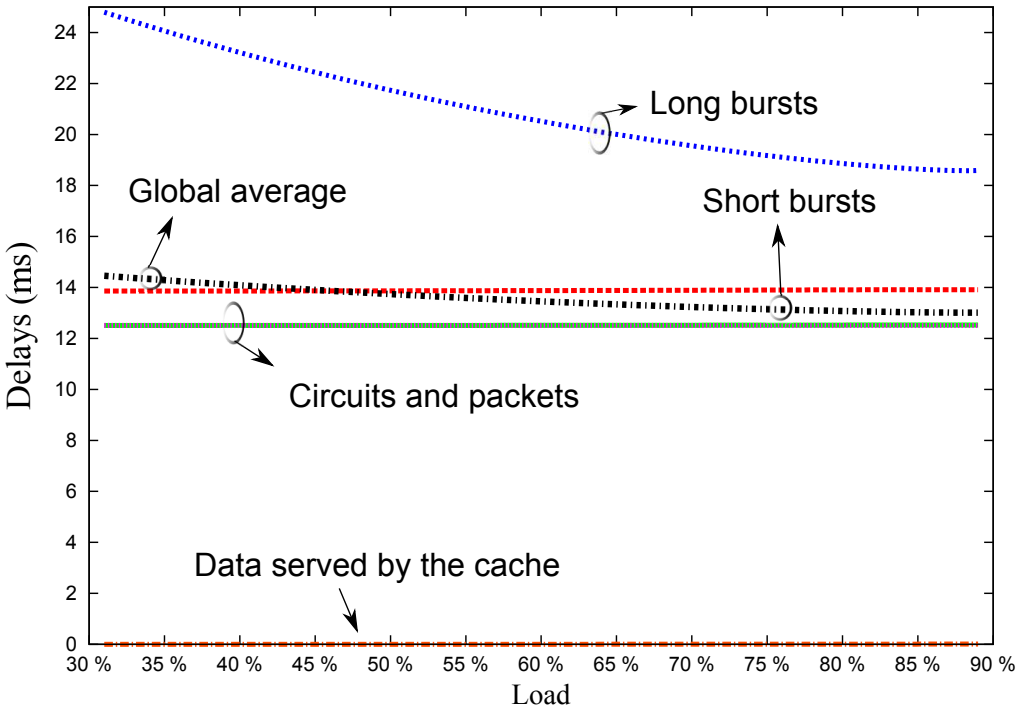

Figure 3.22: Average network delays as a function of the input load. Non-integrated HOS network.

intra-data-center networks is given as the ratio of the total power consumption of the intra-data-center networks and the total throughput of the intra-data-center networks. In the non-integrated HOS architecture the power consumption of the intra-data-center networks include the HOS aggregation switches dedicated to the interconnection toward the core.

From Figure 3.20 we can have two important observations. First, core networks are more energy efficient than intra-data-center networks. This fact is more evident for the integrated HOS network where we observe that at network load of 35% the energy consumption per bit of the core network is 5 times lower compared to the intra-data-center network. Second, the integrated approach has a higher beneficial impact on the energy consumption per bit of the core network than of the intra-data-center networks. In fact, when comparing the energy consumption of the integrated and the non-integrated HOS networks we observe that at a network load of 35%, the integrated approach reduces the energy consumption per bit by 30.5% of the core network and by 3.5% in case of the intra-data-center network. This is because the additional HOS edge nodes,

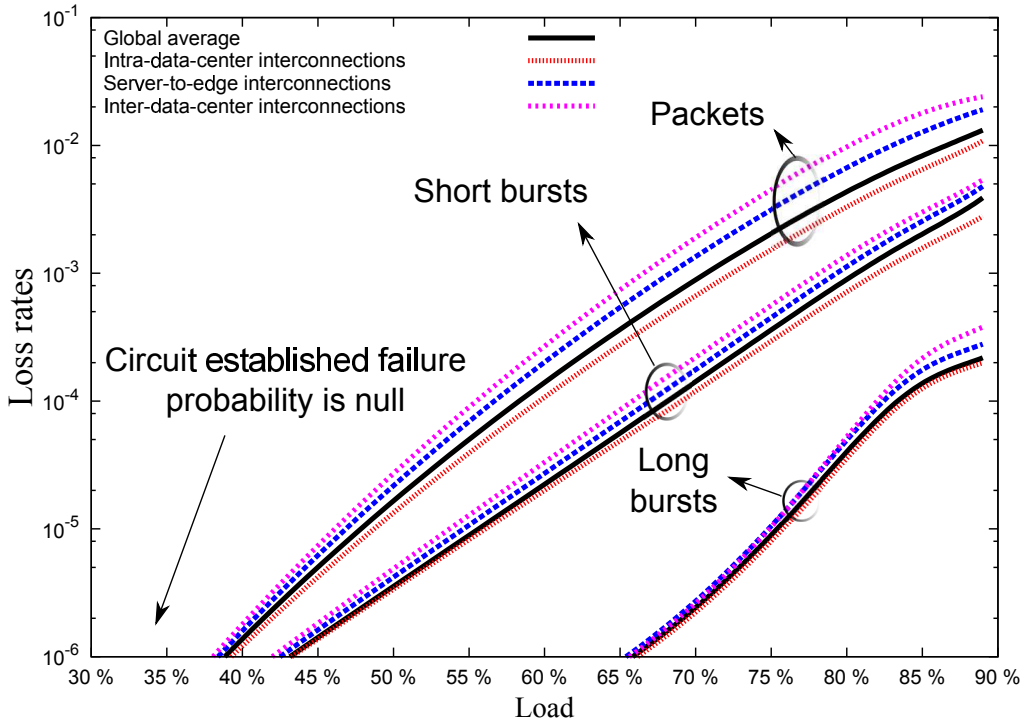

Figure 3.23: Average data loss rate as a function of the input load.

used in the non-integrated HOS network to connect toward the data centers, have a strong impact on the total power consumption of the core network. This impact is higher than the impact of the additional HOS aggregation switches used inside the intra-data-center networks.

In Figure 3.21 and 3.22, we compare the values of the average network delays as a function of the network load. Figure 3.21 shows the average delays in the integrated HOS network, while Figure 3.22 presents the average delays in the non-integrated HOS network. The Figures demonstrate clearly that the integrated approach leads to a better delay performance and reduces the global average delays of IP packets by always more than 1 millisecond (ms). In particular, the integrated approach significantly reduces the delays of IP packets transmitted over short and long bursts. This is due to the fact that bursts employ a mixed timer-length assembly algorithm that may take from several hundreds of microseconds up to a few milliseconds. In the non-integrated HOS network, the bursts must be de-assembled and assembled again in the electronic interfaces between a data center and the core network leading to a strong increase in the overall network delay.

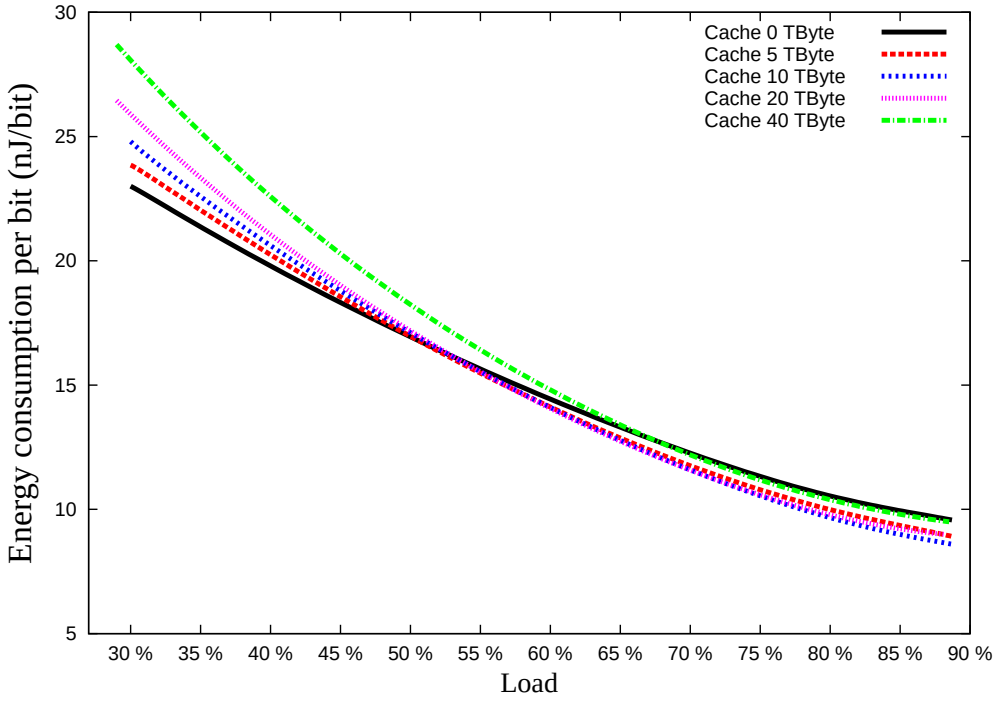

Figure 3.24: Energy consumption per bit against the input load for different cache sizes.

In this paper we assume that the electronic components introduce negligible losses. As a consequence, the data loss rates in the integrated HOS network and in the non-integrated HOS network are the same. In Figure 3.23 we show the average data loss rates as a function of the network load. The optical packets are scheduled with the lowest priority and thus they experience the highest losses. Optical bursts are scheduled a priori due to the offset-time so that they receive a sort of prioritized handling in comparison to packets. In particular, long bursts are characterized by long offset-times and show loss rates almost three orders of magnitude lower than packets and almost two orders of magnitude lower than short bursts. Finally, circuits are scheduled with the highest priority and achieve a lossless operation and negligible establishment failure probabilities in our simulations. To understand where in the network we observe the highest losses, we plot in Figure 3.23 the average loss rates in the intra-data-center, inter-data-center, and server-to-edge interconnections. We observe that the average loss rates in inter-data-center interconnections are always the highest. This is because in the inter-data-center interconnections the data needs to cross on

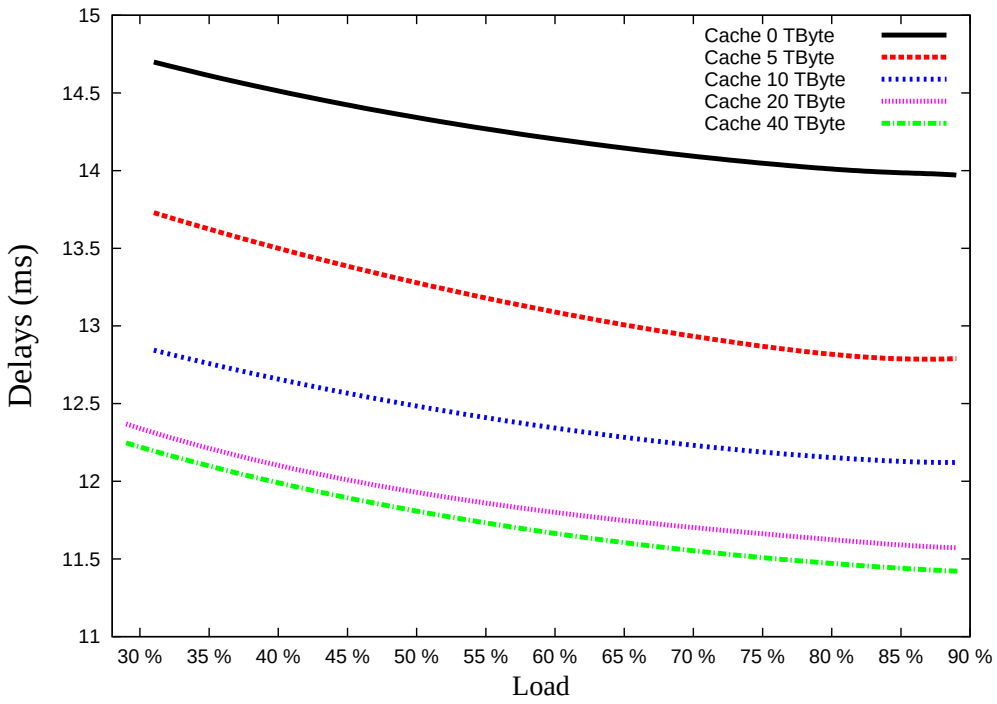

Figure 3.25: Average delays as a function of the input load for different values of the cache size.

average the highest number of HOS core switches (in both HOS core network and intra-data-center network). The lowest average loss rates are instead achieved by the intra-data-center interconnections where data always have to cross a single HOS core switch inside the data center.

3.3.5.2 Impact of Edge Caching

In Figure 3.24 we show the energy consumption per bit of the integrated HOS network against the network load and for different values of the cache size. To vary the cache size we change the number of cache servers per HOS edge node, i.e., the value of N_{CS}. We always assume that the network nodes connected to a data center are not equipped with local cache servers. To understand the results shown in Figure 3.24 it should be noted that we do not consider dynamic switching-off the electronic LCs and consequently the energy consumption of the electronic components is independent on the network load. Only the power consumption of the optical switching fabric of the HOS core nodes, i.e., P_{OSF},

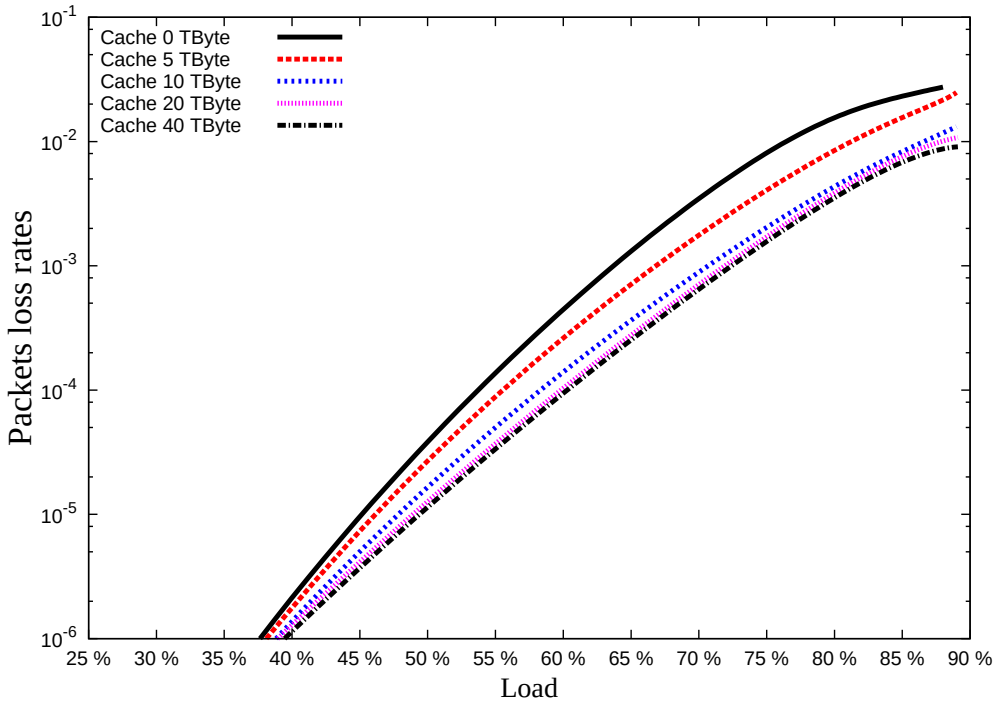

Figure 3.26: Average packet loss rates as a function of the input load for different values of the cache size.

changes with the network load. Furthermore, it should be noted that the energy consumption per bit is defined as the ratio between the network power consumption given in Watt [W] and the network throughput given in bit per second [b/s]. Figure 3.24 shows that at low and moderate loads, the higher is the cache size the higher is the energy consumption per bit. In fact, in our simulations, the increase in the storage energy consumption introduced by the distributed cache servers (P_{Cache}) is always higher than the reduction of the transport energy that is obtained by switching-off the unused optical switch ports of the HOS core nodes. When increasing the load, we observe that the larger the cache size the faster the decrement of the energy consumption per bit. This is due to the fact that increasing the number of the distributed cache servers reduces the average data loss rates in the network. The larger the cache size the higher the network throughput, especially at high loads. However, the network throughput does not increase linearly with the cache size. In fact, according to [88] [89] the network throughput increases in a log-like way with the increase in the cache size. It means that the network throughput becomes saturated when increasing

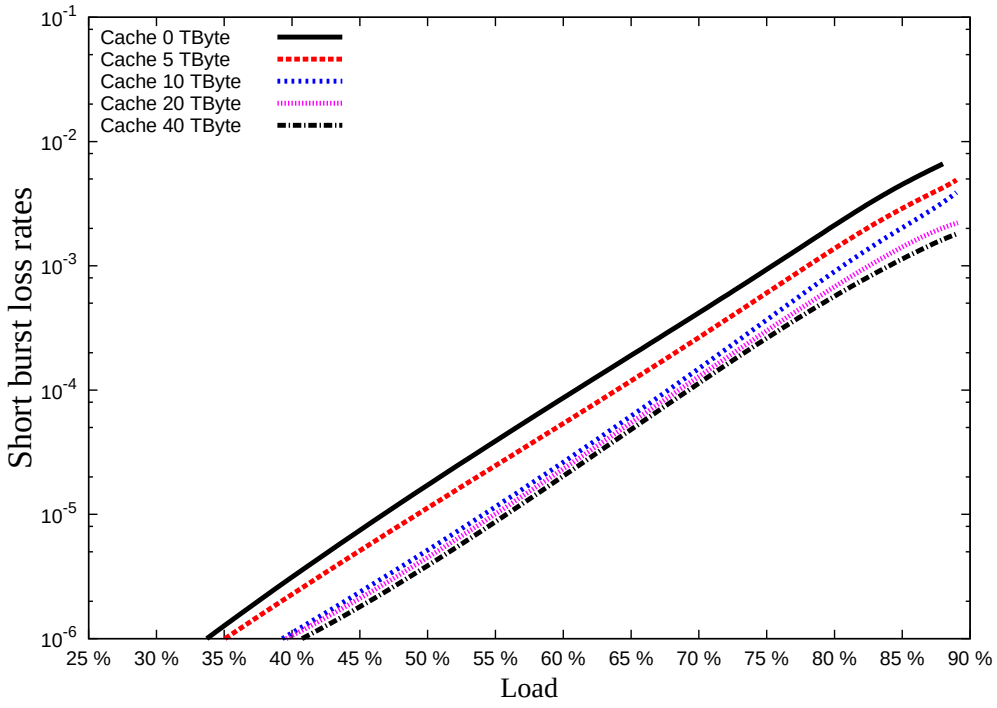

Figure 3.27: Average short burst loss rates as a function of the input load for different values of the cache size.

the cache size over a certain value. On the other and, increasing the cache size leads to an almost linear increase of the storage power consumption. As a consequence, at high loads we observe that there is a trade-off between cache size and energy consumption per bit. In our simulations, when the load is higher than 50%, the best results in terms of energy consumption are achieved using 10 TByte of cache size, i.e., setting $N_{CS} = 10$.

In Figure 3.25 we present the global average network delay as a function of the network load for different values of the cache size. The Figure highlights that the larger the cache size the lower is the global average network delay. In particular, increasing the cache size from 0 to 10 TByte leads to a reduction of the global average delay in the network by about 2 ms. A further increase of the cache size from 20 to 40 TByte has a very limited impact on the global average network delays.

Finally, in Figure 3.26, 3.27 and 3.28 we show the average loss rates of packets, short bursts and long bursts as a function of the network load for different values of the cache size. The circuit establishment failure probability is always null in

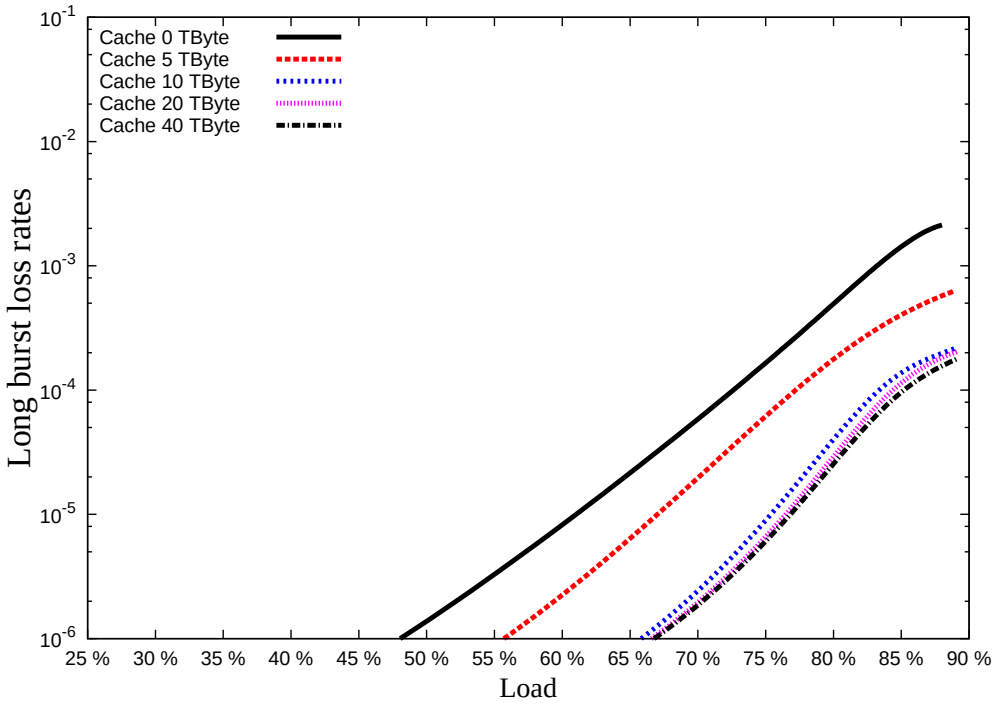

Figure 3.28: Average long burst loss rates as a function of the input load for different values of the cache size.

the considered configurations. The Figures show that the larger the cache size the lower the average loss rates. This is due to the fact that increasing the cache size keeps the traffic more locally, which corresponds to a higher amount of requests from the end users served by the cache servers. This leads to a reduction of the traffic in the core and in the intra-data-center networks and consequently to lower loss ratios. Figure 7 also shows that by increasing the cache size from 0 to 10 TByte we achieve a high reduction in the loss rates, while increasing the cache size over 10 TByte has a very limited impact on the loss rates.

3.3.6 Conclusions

In this paper we have proposed a unified network architecture that provides both intra-data-center and inter-data-center connectivity together with interconnection toward legacy IP networks. This architecture is studied for future carrier cloud operators, which run both the data centers and the core network. The architecture is referred to as integrated core and intra-data-center network and is

based on the Hybrid Optical Switching (HOS) technology. The main advantage in the integration of core and intra-data-center networks in a unique infrastructure is in avoiding electronic interfaces between the data centers and the core network. We evaluated the energy consumption along with the delay and loss performance of the integrated HOS network and made extensive comparisons with respect to a non-integrated HOS solution and a conventional IP network based on electronic switching. We conclude that the integrated HOS network achieves by far the highest energy efficiency. Furthermore, we demonstrated that the integrated HOS network reduces considerably the average network delays with respect to a non-integrated HOS solution. As a consequence, we conclude that our integrated HOS network suits very well for application in carrier cloud.

Furthermore, we studied the impact of distributed video cache servers on the energy consumption as well as the delay and loss performance of the integrated HOS network. The existing literature on this topic only takes into account conventional core and intra-data-center networks based on IP electronic switching, which are characterized by low energy efficiency. We aim to discover if a carrier cloud operator that relays on our integrated HOS network is motivated in using edge caching for optimizing its energy efficiency. We then propose an extended HOS edge node architecture that includes the cache servers and a novel network element, referred to as content tracker, which interact with the HOS control plane for updating the servers and processing incoming video requests. We also develop a novel analytical model for evaluating the power consumed by the cache. From the performed analysis we conclude that to achieve both low delay and low loss as well as high energy efficiency in an integrated HOS network, a careful dimensioning of the cache size is needed, e.g. at low and moderate loads we achieve the highest energy efficiency without any edge caching. Furthermore, our analysis can lead also to a more general conclusion: when deciding to move from traditional electronic switching to a more energy efficient network solution, operators will probably have to reconsider their edge caching strategy.

Chapter 4

Conclusions

This thesis addressed the energy consumption of telecommunication core networks and data centers and proposed new energy efficient architectures based on optical technologies. The proposed architectures are able to cope with the expected increase in Internet traffic demand in an effective and sustainable manner.

In the first part of the thesis, the energy efficiency in optical core networks has been considered. Firstly, an innovative forwarding plane for Hybrid Optical Switching (HOS) core nodes, namely HOS forwarding plane, has been proposed and investigated. The proposed HOS forwarding plane combines efficiently three different optical switching paradigms, namely circuit, burst, and packet switching, on the same node. It employs a unique control packet for carrying the control information of the different data types, ensuring efficient resource utilization and high flexibility. The most appropriate switching method is selected for the transmission of traffic generated by different applications enabling efficient QoS differentiation at the optical layer. Furthermore, a novel optical core node architecture based on two parallel switches, a slow and low power consuming optical switch and a fast switch, has been proposed. The architecture is studied for reducing the energy consumption and optimizing the transmission efficiency. A combined analytical and simulation model has been developed for the evaluation of the data losses and the energy consumption of the proposed HOS core node. The results have shown that the HOS core node is able to achieve high performance while drastically reducing the consumption of current core nodes based on electronic switching.

As a second step, in order to extend the HOS paradigm to a core network (e.g. the Pan-European network), an integration of the HOS forwarding plane and the GMPLS control plane has been proposed. GMPLS is widely recognized as an effective control plane for optical networks since it provides efficient end-to-end provisioning, fast forwarding and traffic engineering decisions. The proposed HOS network is based on an overlay model and is composed of three layers. At the highest layer there is the GMPLS control plane that is in charge of configuring the virtual topology. The GMPLS control plane makes use of a dedicated network and is physically and logically separated from the layers below. The intermediate layer is given by the HOS forwarding plane, which carries the information for properly scheduling and forwarding different data types. Finally, the HOS data plane is an optical network able to support different traffic granularities, i.e., circuits, bursts and packets. The performance and energy consumption of the HOS core network have been evaluated using a combined simulation and analytical approach and the results show that the HOS core network achieves high performance and drastically reduces the energy consumption of current solutions

Finally, the analysis of the HOS core network has been extended by introducing a novel HOS edge node architecture and evaluating its impact on the performance and the energy efficiency. The HOS edge nodes perform the tasks required for the interoperability between the core network and the legacy networks, and, in particular, they are responsible for traffic classification and traffic assembly. Four service classes for the HOS network have been defined and a possible mapping of future Internet services into these classes has been proposed. The performance and energy consumption of the HOS core and edge network have been evaluated using a combined simulation and analytical approach. The results showed that the HOS edge nodes reduce the improvement of energy efficiency against traditional electronic networks. This is due to a higher complexity of the HOS edge nodes with respect to traditional edge nodes. However, it has also been demonstrated that the HOS core and edge network is still able to reduce the energy consumption of current networks by a great amount and, furthermore, it is able to support an efficient QoS differentiation at the optical layer.

In the second part of the thesis, the energy efficiency in optical data center networks has been addressed. Firstly, a novel intra-data-center interconnect

architecture based on the HOS paradigm has been proposed for increasing the flexibility and reducing the energy consumption of current point-to-point solutions. The HOS interconnect is organized in a fat-tree 3-Tier topology, where at the edge tier the Top-of-the-Rack (ToR) switches manage the communications among the servers inside the same rack. At the aggregation tier, the HOS aggregation switches perform traffic aggregation, classification, and assembly and connect the ToR switches to the core tier. The latter, is composed by a large HOS core switch which manages the communications between servers in different racks and is responsible of the interconnection toward the Internet. The HOS interconnect requires minimal hardware modifications with respect to current solutions and can be implemented with limited investments from the operators. The HOS interconnect has been studied using a combined simulation and analytical approach and the results showed that it achieves high transmission efficiency while reducing the energy consumption of current point-to-point solutions. However, the ToR switches, which consume the largest amount of energy inside the data center network, limit the achievable energy gain. Furthermore, the HOS interconnect is not able to scale efficiently with the servers' capacity. It can be concluded that the HOS interconnect is an optimal solution for the short/mid term, but that in the long term a more scalable solution may be needed.

To address the long term need for high-capacity, scalable and energy-efficient data center solutions, in this thesis a novel data center network architecture, realized by combining broadcast-and-select approach with elastic channel spacing technology, has been proposed. The interconnection among the servers within the same rack is carried out by an optical broadcast-and-select switching matrix, which presents higher transmission efficiency and lower energy consumption than traditional ToR switches. The inter-rack communication are managed by an elastic switch realized by interconnecting a high number of optical switching element based on the beam steering technology. The proposed elastic optical interconnect requires a complete change of the hardware used in current data centers networks and thus it needs high investments from the operators. The energy consumption of the proposed elastic optical data center has been proposed using an analytical approach. It has been demonstrated that the proposed architecture is able to scale efficiently with the number of servers and the servers' capacity and offers higher flexibility and lower energy consumption compared with both the HOS interconnect and the traditional point-to-point interconnects.

Finally, in the last part of the thesis, a unified network architecture that

provides both intra-data-center and inter-data-center connectivity together with interconnection toward legacy IP networks, is proposed. The architecture is well suited for the carrier cloud model, where both data-center and telecom infrastructure are owned and operated by the same entity. It is based on the HOS concept for achieving high network performance and energy efficiency. Therefore, we refer to it as an integrated HOS network. The main advantage of the integration of core and intra-data-center networks comes from the possibility to avoid the energy inefficient electronic interfaces between data centers and telecom network. The results have verified that the integrated HOS network introduces high benefits in terms of energy efficiency and network delays compared to the conventional non-integrated solution. Furthermore, the impact of distributed video cache servers on the energy consumption as well as the delay and loss performance of the integrated HOS network has been studied. The aim has been to identify whether a carrier cloud operator that relies on the integrated HOS network concept could increase energy efficiency by employing edge caching. According to the results we conclude that to achieve both low delay and data loss as well as high energy efficiency in an integrated HOS network, a careful dimensioning of the cache size is needed. In particular, at low and moderate loads we observed the highest energy efficiency is achieved in the case without any edge caching.

Bibliography

[1] "Report on climate change," *International Telecommunications Union (ITU), available online at www.itu.int.*

[2] "Smart 2020: Enabling the low carbon economy in the information age," *technical report, the Climate Group, Global eSustainability Initiative, 2008; availabe online at www.smart2020.org.*

[3] "Cisco visual networking index: Forecast and methodology, 2012-2017," *Cisco white paper*, May 2013.

[4] "Cisco global cloud index: Forecast and methodology, 2011-2016," *Cisco white paper*, May 2012.

[5] C. Lange, D. Kosiankowski, R. Weidmann, and A. Gladisch, "Energy consumption of telecommunication networks and related improvement options," *IEEE J. Selected Topics in Quantum Electronics*, vol. 17, no. 2, pp. 285–295, May 2011.

[6] Y. Zhang, P. Chowdhury, M. Tornatore, and B. Mukherjee, "Energy efficiency in telecom optical networks," *IEEE Communications Surveys and Tutorials*, vol. 12, no. 4, pp. 441–458, 2010.

[7] J. Baliga, R. Ayre, K. Hinton, W. Sorin, and R. Tucker, "Energy consumption in optical ip networks," *IEEE J. Lightwave Technology*, vol. 27, no. 13, pp. 2391–2403, 2009.

[8] S. Aleksic, "Energy efficiency of electronic and optical network elements," *IEEE J. Selected Topics in Quantum Electronics*, vol. 17, no. 3, pp. 296–308, 2011.

[9] A. Tzanakaki, K. Katrinis, T. Politi, A. Stavdas, M. Pickavet, P. van Daele, D. Simeonidou, M. O'Mahony, S. Aleksic, L. Wosinska, and P. Monti, "Dimensioning the future pan-european optical network with energy efficiency considerations," *IEEE/OSA J. of Optical Communications and Networking*, vol. 3, no. 4, pp. 272–280, April 2011.

[10] S. Aleksic, "Analysis of power consumption in future high-capacity network nodes," *IEEE/OSA J. Optical Communications and Networking*, vol. 1, no. 3, pp. 245–258, 2009.

[11] R. Ramaswami, K. Sivarajan, and G. Sasaki, *Optical Networks: A Practical Perspective*. Morgan Kaufmann Publishers, 2009.

[12] D. Blumenthal, J. Barton, N. Beheshti, J. Bowers, E. Burmeister, L. Coldren, M. Dummer, G. Epps, A. Fang, Y. Ganjali, J. Garcia, B. Koch, V. Lal, E. Lively, J. Mack, M. Masanovic, N. McKeown, K. Nguyen, S. Nicholes, H. Park, B. Stamenic, A. Tauke-Pedretti, H. Poulsen, and M. Sysak, "Integrated photonics for low-power packet networking," *IEEE J. Selected Topics in Quantum Electronics*, vol. 17, no. 2, pp. 458–471, 2011.

[13] R. Tucker, "Green optical communications - part I: Energy limitations in transport," *IEEE J. of Selected Topics in Quantum Electronics*, vol. 17, no. 2, pp. 245–260, 2011.

[14] ——, "Green optical communications - part II: Energy limitations in networks," *IEEE J. of Selected Topics in Quantum Electronics*, vol. 17, no. 2, pp. 261–274, 2011.

[15] ——, "The role of optics and electronics in high-capacity routers," *IEEE J. Lightwave Technology*, vol. 24, no. 12, pp. 4655–4673, 2006.

[16] S. Aleksic, "Energy-efficient communication networks for improved global energy productivity," *Telecommunication Systems, Springer*, vol. 54, no. 2, pp. 183–199, 2013.

[17] J. Turner, "Terabit burst switching," *J. High Speed Networks*, vol. 8, no. 1, pp. 3–16, 1999.

[18] M. Y. Chunming Qiao, "Optical burst switching (OBS) - a new paradigm for an optical internet," *J. High Speed Networks*, vol. 8, no. 1, pp. 69–84, 1999.

[19] M. Fiorani, M. Casoni, and S. Aleksic, "Performance and power consumption analysis of a hybrid optical core node," *IEEE/OSA J. Optical Communications and Networking*, vol. 3, no. 6, pp. 502–513, 2011.

[20] M. Takagi, H. Li, K. Watabe, H. Imaizumi, T. Tanemura, Y. Nakano, and H. Morikawa, "400gb/s hybrid optical switching demonstration combining multi-wavelength ops and ocs with dynamic resource allocation," in *Conference on Optical Fiber Communication - incudes post deadline papers, 2009. OFC 2009.*, 2009, pp. 1–3.

[21] R. Veisllari, S. Bjornstad, and D. Hjelme, "Experimental demonstration of high throughput, ultra-low delay variation packet/circuit fusion network," *Electronics Letters*, vol. 49, no. 2, pp. 141–143, 2013.

[22] R. Cafini, W. Cerroni, C. Raffaelli, and M. Savi, "Standard-based approach to programmable hybrid networks," *IEEE Communications Magazine*, vol. 49, no. 5, pp. 148–155, 2011.

[23] IETF RFC 3945, "Generalized Multi-Protocol Label Switching (GMPLS) Architecture," October 2004.

[24] J. Koomey, "Worldwide electricity used in data centers," *Environmental Research Letters*, no. 034008, September, 2008.

[25] ——, "Growth in data center electricity use 2005 to 2010," *The New York Times*, vol. 49, no. 3, p. 24, 2011.

[26] *Google report, available online at: www.google.com/about/datacenters/efficiency/internal.*

[27] "Where does power go?" *GreenDataProject, available online at: www.greendataproject.org*, 2008.

[28] S. Aleksic, G. Schmid, and N. Fehratovic, "Limitations and perspectives of optically switched interconnects for large-scale data processing and storage systems," *MRS Proceedings, Cambridge University Press*, vol. 1438, Spring 2012.

[29] S. Aleksic and N. Fehratovic, "Requirements and limitations of optical interconnects for high-capacity network elements," in *International Conference on Transparent Optical Networks (ICTON)*, 2010, pp. 1–4.

[30] A. Benner, "Optical interconnect opportunities in supercomputers and high end computing," in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC)*, 2012, pp. 1–60.

[31] G. Wang, D. Andersen, M. Kaminsky, K. Papagiannaki, T. Ng, M. Kozuch, and M. Ryan, "c-through: Part-time optics in data centers," *Proceedings of ACM SIGCOMM*, pp. 327–338, 2010.

[32] N. Farrington, G. Porter, S. Radhakrishnan, H. Bazzaz, V. Subramanya, V. Fainman, G. Papen, and A. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers," *Proceedings of ACM SIGCOMM*, pp. 339–350, 2010.

[33] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: a topology malleable data center network," *Proceedings of ACM SIGCOMM*, pp. 8:1–8:6, 2010.

[34] O. Liboiron-Ladouceur, I. Cerutti, P. Raponi, N. Andriolli, and P. Castoldi, "Energy-efficient design of a scalable optical multiplane interconnection architecture," *IEEE J. Sel. Topics Quantum Electronics*, no. 99, pp. 1–7, 2010.

[35] X. Ye, Y. Yin, S. Yoo, P. Mejia, R. Proietti, and V. Akella, "Dos: A scalable optical switch for datacenters," *Proceedings of ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, pp. 24:1–24:12, 2010.

[36] K. Xia, Y. Kaob, M. Yangb, and H. Chao, "Petabit optical switch for data center networks," *Technical report, Polytechnic Institute of NYU*, 2010.

[37] R. Luijten, W. Denzel, R. Grzybowski, and R. Hemenway, "Optical interconnection networks: The osmosis project," *17th Annual Meeting of the IEEE Lasers and Electro-Optics Society*, 2004.

[38] A. Shacham and K. Bergman, "An experimental validation of a wavelength-striped, packet switched, optical interconnection network," *IEEE J. Lightwave Technology*, vol. 27, no. 7, pp. 841–850, April 2009.

[39] J. Luo, S. Di Lucente, J. Ramirez, H. Dorren, and N. Calabretta, "Low latency and large port count optical packet switch with highly distributed

control," in *Conference on Optical Fiber Communication - incudes post deadline papers, 2009. OFC 2009.*, 2012, pp. 1–3.

[40] C. Kachris and I. Tomkos, "Power consumption evaluation of hybrid wdm pon networks for data centers," in *European Conference on Networks and Optical Communications (NOC)*, 2011, pp. 118–121.

[41] "Connectivity solutions for the evolving data center," *white paper, Emulex*, May 2011.

[42] D. Cai and S. Natarajan, "The evolution of the carrier cloud networking," *IEEE International Symposium on Service-Oriented System Engineering*, pp. 286–291, 2013.

[43] M. Casoni and M. Merani, "On the performance of tcp over optical burst switched networks with different qos classes," *J. High Speed Networks*, vol. 3375, pp. 574–585, 2005.

[44] 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. Danielsen, F. Dorgeuille, A. Dupas, A. Franzen, P. Hansen, D. Hunter, A. Kloch, R. Krahenbuhl, B. Lavigne, A. Le Corre, C. Raffaelli, M. Schilling, J. C. Simon, and L. Zucchelli, "Transparent optical packet switching: the european acts keops project approach," *IEEE J. Lightwave Technology*, vol. 16, no. 12, pp. 2117–2134, 1998.

[45] Y. Yamada, K. Sasayama, K. Habara, A. Misawa, M. Tsukada, T. Matsunaga, and K. Yukimatsu, "Optical output buffered atm switch prototype based on frontiernet architecture," *IEEE J. Selected Areas in Communications*, vol. 16, no. 7, pp. 1298–1308, 1998.

[46] D. Cotter, J. Lucek, and D. Marcenac, "Ultra-high-bit-rate networking: from the transcontinental backbone to the desktop," *IEEE Communications Magazine*, vol. 35, no. 4, pp. 90–95, 1997.

[47] X. Wanga, W. Houa, L. Guoa, J. Caoc, and D. Jianga, "Energy saving and cost reduction in multi-granularity green optical networks," *Elsevier Computer Networks*, vol. 55, no. 3, pp. 676–688, 2011.

[48] G. Shen and R. Tucker, "Energy-minimized design for ip over wdm networks," *IEEE/OSA J. Optical Communications and Networking*, vol. 1, no. 1, pp. 176–186, 2009.

[49] E. Yetginer and G. Rouskas, "Power efficient traffic grooming in optical wdm networks," in *IEEE Global Telecommunications Conference, 2009. GLOBECOM 2009*, 2009, pp. 1–6.

[50] M. Hasan, F. Farahmand, and J. Jue, "Energy-awareness in dynamic traffic grooming," in *Conference on Optical Fiber Communication (OFC), collocated National Fiber Optic Engineers Conference (OFC/NFOEC)*, 2010, pp. 1–3.

[51] C. Raffaelli, S. Aleksic, F. Callegati, W. Cerroni, A. Pattavina, and M. Savi, "Optical packet switching," *Enabling Optical Internet with Advanced Network Technologies, Springer*, pp. 31–85, 2009.

[52] S. Aleksic and V. Krajinovic, "Comparison of optical code correlators for all-optical mpls networks," in *European Conference on Optical Communication, 2002. ECOC 2002*, vol. 1, 2002, pp. 1–2.

[53] A. Martinez, D. Pastor, J. Capmany, B. Ortega, P. Fonjallaz, M. Popov, T. Berceli, and T. Banky, "Experimental demonstration of subcarrier multiplexed optical label swapping featuring 20 gb/s payload speed and 622 mb/s header conveyed @18.3 ghz," in *European Conference on Optical Communication, 2005. ECOC 2005.*, vol. 4, 2005, pp. 959–960 vol.4.

[54] N. Deng, Y. Yang, C.-K. Chan, L.-K. Chen, and W. Hung, "All-optical ook label swapping on ofsk payload in optical packet networks," in *Optical Fiber Communication Conference, 2004.*, vol. 2, 2004, p. 3 pp. vol.2.

[55] Y. Xiong, M. Vandenhoute, and H. Cankaya, "Control architecture in optical burst-switched wdm networks," *IEEE J. Selected Areas in Communications*, vol. 18, no. 10, pp. 1838–1851, 2000.

[56] K. Dozer, C. Gauager, J. Spath, and S. Bodamer, "Evaluation of reservation mechanism for optical burst switching," *AEU International Journal of Electronic and Communication*, vol. 55, no. 1, January, 2001.

[57] J. Xu, C. Qiao, J. Li, and G. Xu, "Efficient channel scheduling algorithms in optical burst switched networks," in *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, vol. 3, 2003, pp. 2268–2278 vol.3.

[58] C. Qiao, "Labeled optical burst switching for ip-over-wdm integration," *Communications Magazine, IEEE*, vol. 38, no. 9, pp. 104–114, 2000.

[59] P. Pedroso, J. Sole-Pareta, D. Careglio, and M. Klinkowski, "Integrating gmpls in the obs networks control plane," in *International Conference on Transparent Optical Networks, 2007. ICTON '07.*, vol. 3, 2007, pp. 1–7.

[60] P. Pedroso, D. Careglio, R. Casellas, M. Klinkowski, and J. Sole-Pareta, "An interoperable gmpls/obs control plane: Rsvp and ospf extensions proposal," in *International Symposium on Communication Systems, Networks and Digital Signal Processing, 2008. CNSDSP 2008.*, 2008, pp. 418–422.

[61] J. Triay, G. Zervas, C. Cervello-Pastor, and D. Simeonidou, "Gmpls/pce/obst architectures for guaranteed sub-wavelength mesh metro network services," in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC)*, 2011, pp. 1–3.

[62] M. Fiorani, M. Casoni, and S. Aleksic, "Analysis of a GMPLS enabled hybrid optical switching network," in *International Conference on Optical Network Design and Modeling (ONDM), 2012*, 2012, pp. 1–6.

[63] S. Aleksic, M. Fiorani, and M. Casoni, "Energy efficiency of hybrid optical switching," in *International Conference on Transparent Optical Networks (ICTON), 2012*, 2012, pp. 1–4.

[64] M. Fiorani, M. Casoni, and S. Aleksic, "Hybrid optical switching for an energy-efficient internet core," *IEEE Internet Computing*, vol. 17, no. 1, pp. 14–22, 2013.

[65] S. Aleksic, M. Fiorani, and M. Casoni, "Adaptive hybrid optical switching: Performance and energy efficiency," *Journal of High Speed Networks*, vol. 19, no. 1, pp. 85–98, 2013.

[66] T. Miyazawa, H. Furukawa, K. Fujikawa, N. Wada, and H. Harai, "Development of an autonomous distributed control system for optical packet and

circuit integrated networks," *IEEE/OSA J. Optical Communications and Networking*, vol. 4, no. 1, pp. 25–37, 2012.

[67] IETF RFC 2474, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," December 1998.

[68] IETF RFC 2475, "An Architecture for Differentiated Services," December 1998.

[69] M. Fiorani, M. Casoni, and S. Aleksic, "Hybrid optical switching for energy-efficiency and qos differentiation in core networks," *IEEE/OSA J. Optical Communications and Networking*, vol. 5, no. 5, pp. 484–497, 2013.

[70] IETF RFC 4594, "Configuration Guidelines for DiffServ Service Classes," August 2006.

[71] M. Casoni, E. Luppi, and M. Merani, "Impact of assembly algorithms on end-to-end performance in optical burst switched networks with different qos classes," in *International Workshop on Optical Burst Switching*, 2004.

[72] A. Betker, C. Gerlach, R. Hlsermann, M. Jger, M. Barry, S. Bodamer, J. Spth, C. Gauger, and M. Khn, "Reference transport network scenarios," in *MultiTeraNet Report*, July, 2003.

[73] C. Kachris and I. Tomkos, "A survey on optical interconnects for data centers," *IEEE Communications Surveys Tutorials*, vol. 14, no. 4, pp. 1021–1036, 2012.

[74] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The nature of data center traffic: measurements and analysis," in *ACM SIGCOMM conference on Internet measurement conference, Proc. IMC 2009*, 2009, pp. 202–209.

[75] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," in *ACM workshop on Research on enterprise networking*, 2009, pp. 65–72.

[76] T. Benson, A. Akella, and D. Maltz, "Network traffic characteristics of data centers in the wild," in *IMC 2010*, 2010, pp. 267–280.

[77] American Economics Association (AEA), "Guidelines to Defra/DECCs GHG Conversion Factors for Company Reporting," 2009.

[78] EU project, "Discus (The DIStributed Core for unlimited bandwidth supply for all Users and Services) ," http://www.discus-fp7.eu/ 2013.

[79] CRN, "Cloud Services: Carriers Want Cloud," http://www.crn.com July, 2011.

[80] A. Autenrieth, J.-P. Elbers, P. Kaczmarek, and P. Kostecki, "Cloud orchestration with sdn/openflow in carrier transport networks," in *International Conference on Transparent Optical Networks (ICTON)*, 2013, pp. 1–4.

[81] J. Baliga, R. Ayre, K. Hinton, and R. Tucker, "Architectures for energy-efficient iptv networks," in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC)*, 2009, pp. 1–3.

[82] C. Jayasundara, A. Nirmalathas, E. Wong, and C. A., "Energy efficient content distribution for vod services," in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC)*, 2011, pp. 1–3.

[83] C. Chan, E. Wong, A. Nirmalathas, A. Gygax, and C. Leckie, "Energy efficiency of on-demand video caching systems and user behavior," *Optics Express*, vol. 19, no. 26, pp. B260–B269, 2011.

[84] N. Osman, T. El-Gorashi, and J. Elmirghani, "Reduction of energy consumption of video-on-demand services using cache size optimization," in *Conference on Wireless and Optical Communications Networks (WOCN)*, 2011, pp. 1–5.

[85] ——, "The impact of content popularity distribution on energy efficient caching," in *International Conference on Transparent Optical Networks (ICTON)*, 2013, pp. 1–6.

[86] ——, "Caching in green ip over wdm networks," *Journal of High Speed Networks*, vol. 19, pp. 33–53, 2013.

[87] F. Idzikowski, S. Orlowski, C. Raack, H. Woesner, and A. Wolisz, "Saving energy in ip-over-wdm networks by switching off line cards in low-demand scenarios," in *International Conference on Optical Network Design and Modeling (ONDM)*, 2010, pp. 1–6.

[88] M. Zink, k. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network - measurements, models, and implications," *Elsevier Computer Networks Journal*, vol. 53, no. 4, pp. 501–514, 2009.

[89] L. Braun, A. Klein, G. Carle, H. Reiser, and J. Eisl, "Analyzing caching benefits for youtube traffic in edge networks - a measurement-based evaluation," in *IEEE Network Operations and Management Symposium (NOMS)*, 2012, pp. 311–318.

Publications List

Journal Papers

1. **M. Fiorani**, M. Casoni, S. Aleksic, "Performance and Power Consumption Analysis of a Hybrid Optical Core Node", IEEE/OSA Journal of Optical Communications and Networking, Vol. 3 , Issue 6, pp. 502-513, June 2011.

2. **M. Fiorani**, M. Casoni, S. Aleksic, "Energy-Efficient Internet Core Employing Hybrid Optical Switching", IEEE Internet Computing Magazine, Special Issue on "Sustainable Internet", Vol. 17, Issue 1, pp. 14-22, January/February 2013

3. S. Aleksic, **M. Fiorani**, M. Casoni, "Adaptive Hybrid Optical Switching: Performance and Energy Efficiency", (invited) IOS Journal of High Speed Networks, Special Issue on "Green Networking and Computing", Vol.19, No.1, pp. 85-98, 2013.

4. **M. Fiorani**, M. Casoni, S. Aleksic, "Hybrid Optical Switching for Energy-Efficiency and QoS Differentiation in Core Networks", IEEE/OSA Journal of Optical Communications and Networking, Vol. 5, Issue 5, pp. 484-497, 2013.

5. **M. Fiorani**, M. Casoni, S. Aleksic, "Hybrid Optical Switching for Data Center Networks", (available online) Hindawi Journal of Electrical and Computer Engineering, Special Issue on "Innovative Techniques for Power Consumption Saving in Telecommunication Networks", 2014.

6. W. Cerroni, **M. Fiorani**, M. Casoni, "TCP Performance in Multi-EPON Access Networks under Different Optical Core Switching Paradigms", accepted to Elsevier Optical Switching and Networking Journal, 2014.

Conference Papers

1. **M. Fiorani**, M. Casoni, W. Cerroni, "Transport Layer Performance of Hybrid Networks Combining Multiple EPONs and OBS", Proc. of the 18th IEEE LANMAN Workshop, October 13-14, 2011, Chapel Hill, North Carolina, U.S.A.

2. **M. Fiorani**, M. Casoni, S. Aleksic, "Analysis of a GMPLS enabled hybrid optical switching network", Proc. of the 16th IEEE International Conference on Optical Networks Design and Modeling (ONDM), April 17-20, 2012, Colchester, England.

3. S. Aleksic, **M. Fiorani**, M. Casoni, "Energy efficiency of hybrid optical switching", Proc. of the 14th IEEE International Conference on Transparent Optical Networks (ICTON), July 2-5, 2012, Warwick, England.

4. **M. Fiorani**, M. Casoni, S. Aleksic, "Large Data Center Interconnects Employing Hybrid Optical Switching", Proc. of the 18th IEEE European Conference on Network and Optical Communications (NOC), July 10-12, 2013, Graz, Austria.

5. S. Tombaz, P. Monti, F. Farias, **M. Fiorani**, L. Wosinska, J. Zender, "Is Backhaul Becoming a Bottleneck for Green Wireless Access Networks?", Proc. of the IEEE International Conference on Communications (ICC), June 10-14, 2014, Sydney, Australia.

Acknowledgments

It is not easy to write about all the people that contributed to your personal and professional growth over a period of almost 4 years. For this reason I have always been determined in not writing any acknowledgment for my PhD thesis, convinced that this would be a more fair choice than writing an inevitably incomplete list of thanks. However, one day I stumbled upon a sentence of Albert Einstein that says "*Learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning*". Given my unlimited esteem for his figure, whose amazing discoveries have made possible the research presented in this thesis, I started wondering on the meaning of these words. They reflect the real essence of research; but not only. They represent an approach to life. In particular, putting my attention to the opening words, I have considered that in a learning process there is always a time in which one should stop and look back to what has been done and to the people that contributed to it. What could be the best time to do so, if not after the achievement of an important milestone, which the PhD surely represents? Convinced by Albert's words, I will then report in the following an incomplete list of thanks to all the people that somehow contributed to this elaborate.

First, I would like to express my deepest gratitude to my supervisor, Prof. Maurizio Casoni, for accepting me as his PhD student and always being very positive from the start. Without his constant support this work would not have been possible. I am very grateful to Dr. Slavisa Aleksic, from Vienna University of Technology, for guiding me through the years of my PhD and being an endless source of knowledge and information. I am also very grateful to Dr. Walter Cerroni, from University of Bologna, for all his kind help especially at the beginning of my PhD studies. In addition, I would like to express my gratitude to the ONLab people, KTH Royal Institute of Technology, Stockholm,

and in particular to Prof. Lena Wosinska, Dr. Paolo Monti and Dr. Jiajia Chen, for our intense and prolific collaborations.

I would then like to thank all my friends "Amici miei" in Modena. It is impossible to tell in a few words how their constant presence and support have helped to live happily these last years. I would like to write about each of them and tell about our stories, but this would require another entire PhD thesis long at least twice as the present one. I will then limit myself to say thank to all of them for our long-time and everlasting friendships. I will just mention a few names apologizing for those who are not here: Cave, Lance, Alle, Pit, Frux, Ansa, Andre, Aba, Giuly, Vale, Ippo, Checco, Davo, Lombo, Vercio, Bea, Simo. I would also like to thank the friends and colleagues of the Elecom Lab in Modena, especially my "companion in misfortune" Fabio. Also, I would like to acknowledge my dear friends from Vienna for all the life lessons that they taught me. In addition, I am very grateful to Dr. Lelio Baldeschi. Last, but first in my heart, I would like to say thank you to Sibel for making me feel the luckiest man in the world.

I would like to reserve the last slot for thanking all my family. In particular, I will be endlessly grateful to my parents for their constant love and encouragement. Nothing would have been possible without them. Grazie per essere i migliori genitori del mondo.