Scalable electro-optical solutions for data center networks

Scalable Electro-Optical Solutions for Data Center Networks

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een
commissie aangewezen door het College voor
Promoties, in het openbaar te verdedigen
op donderdag 22 maart 2018 om 11.00 uur

door

Gonzalo Guelbenzu de Villota

geboren te Madrid, Spanje

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. A.B. Smolders
1e promotor:	prof.ir. A.M.J. Koonen
co-promotoren:	dr. O. Raz
	dr. N. Calabretta
leden:	prof.dr. L. Dittmann (Technical University of Denmark, Denmark)
	dr. S. Spadaro (Universitat Politecnica de Catalunya, Spain)
	dr. B.M. Sadowski
adviseur:	dr. Y. Ben-Itzhak (IBM Research and Development Labs, Israel)

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

A catalogue record is available from the Eindhoven University of Technology Library.

Title: Scalable Electro-Optical Solutions for Data Center Networks
Author: Gonzalo Guelbenzu de Villota
Eindhoven University of Technology

ISBN: 978-90-386-4464-6

Dedicated to my parents.

Summary

Data centers are the underlying infrastructure supporting the exponential growth of cloud applications. Large-scale data centers interconnect more than one hundred thousand servers with an optical network and they are reaching the economically feasible limits in terms of power consumption and cost. On top of that, it is foreseen that future bandwidth demands will require the interconnection of even more devices. Data center networks play a critical role in the performance of these complex systems, especially because most of the data center traffic happens within the data center. This thesis investigates the scalability of optical data center networks and suggests a number of solutions to improve the scaling of such networks and interconnect additional devices without incurring extra power consumption, space, and cost.

The first part of this thesis focuses on the scaling of electronic switches, the main building block of data center networks deployed at present. Our investigation follows a research-through-design approach, i.e., we embed the design knowledge in an artifact that attempts to transform the world from the current state to a preferred one. Our analysis points out the relevance of shrinking the size and the power consumption of electronic switches in order to keep increasing the number of servers in data center networks limited in space and power. In order to achieve these goals, we suggest the integration of On-Board Optics transceivers (instead of the typical front-panel transceivers) and the packaging of multiple compact electronic switches per rack unit (instead of only one at most like in the front-panel transceiver approach). Our prototype demonstrates the feasibility of the approach, and the result is one of the most compact 1.28 Tbps electronic data center switches ever implemented. The single board prototype has only 20 cm by 20 cm size and integrates 12-port On-Board Optics transceivers responsible for the electronic-to-optical and optical-to-electronic conversion. The 20-layer Printed Circuit Board generates all voltages required from a single 12V input power supply, making the power supply from the rack unit redundant. The extra space

gained in the rack unit allows integrating four of these prototypes in our packaged demonstrator, achieving a front plate density of 4 x 128-ports and 5.12 Tbps. The whole platform is enabled for software-defined networking (SDN) thanks to the control plane processor included in each switch. Our measurements validate this approach both in operational temperature range and in power consumption. Regarding temperature, the rack unit operates below 40°C due to the transceivers all spread through the rack area and the free space in the front-plate for additional ventilation. This contrasts with the traditional front-panel transceiver approach, limited in scaling by the front panel area and in cooling by a front-panel fully blocked with the transceivers. Regarding power, a single prototype including transceivers consumes just around 100W, well below other approaches based on front panel transceivers needing 300W. This difference is obtained because On-Board Optics transceivers can be placed very close to the electronic switching ASIC, allowing shorter traces and enabling more energy efficient electrical interfaces with reduced power consumption. We conclude that devices similar to our demonstrator enable further upscaling of data centers, interconnecting additional devices without requiring extra space, cost, or power.

The second part of this thesis focuses on the scaling of hybrid networks, integrating a combination of electronic and all-optical switches. We present a novel analytic model describing three extensions of the Fat-Tree (FT) topology, namely Extended Fat-Tree (EFT), Hybrid Fat-Tree (HFT), and Extended Hybrid Fat-Tree (EHFT) topologies. These architectures explore the introduction of optical switching and wavelength-division multiplexing technologies in Fat-Tree like topologies in order to reduce the power consumption and cost while maintaining important features such as the scalability and full bisection bandwidth. The flexibility of our model, given by its configuration parameters, allows generating multiple flavours of the architectures mentioned which include other hybrid topologies found in the literature lacking a common mathematical framework. On top of that, the equations included in the model accurately compute the number of switches, transceivers, and fibers of each topology which enables comparing the architectures in terms of devices, power consumption, and cost.

EFT explores the introduction of WDM technologies in FT networks in order to reduce the number of fibers. Our studies on the scaling in a 25G real case scenario with technologies available at present show that 25% of fibers can be saved with 4-port transceivers. HFT and EHFT topologies integrate also optical switching technologies and reduce the number of switches, transceivers, and fibers compared to FT. Our 25G real case scenario investigation concludes that the minimum hybrid topologies achieve savings of 45% in switches, 60% in transceivers, 50% in fibers,

55% in power consumption, and 48% in cost. We conclude that the introduction of optical switching and wavelength-division multiplexing technologies into data center networks enable further scaling by reducing power consumption and cost.

We experimentally validate the feasibility of the integration of these technologies by implementing the IPI-TU/e (Institute for Photonic Integration - Technical University of Eindhoven) hybrid data center demonstrator. It integrates FOX (our optical core unit), with electronic switches, servers, and an SDN data center controller. FOX is a fast optical switch based on semiconductor optical amplifiers and integrates the same control plane processor included in our electronic switch in order to leverage the same control network. The central controller orchestrates the behaviour of the system and implements E-WDM, a software control technique meant for hybrid networks with optical switches and wavelength division multiplexing which may suffer traffic pattern restrictions when the optical switches provide pure spatial switching. E-WDM leverages the electronic switches present in hybrid networks in order to recover traffic patterns with granularity at the server level without requiring wavelength selective switches.

Overall, the results achieved in this thesis demonstrate promising solutions for the scaling of next-generation data center optical networks.

Contents

Chapter 1

Introduction

1.1 The rise of cloud computing

Traditionally, most applications have resided in the client, including email, photo and video storage, and office applications. The emergence of popular Internet services such as web-based email, search engines, e-commerce, and social networks plus the increased worldwide availability of high-speed connectivity has accelerated a trend toward server-side or cloud computing [1]. Cloud computing is defined by [2] as *a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.*

The five essential characteristics of cloud computing bring advantages for users and vendors. From the user perspective, cloud computing provides *on-demand self-service* (i.e. a consumer can unilaterally provision computing capabilities, such as server time and network storage, without requiring human interaction with each service provider), *broad network access* (i.e. capabilities are available over the network and accessed through standard platforms such as mobile phones, tablets, laptops, and workstations), and *rapid elasticity* (i.e. capabilities can be elastically provisioned and released to scale rapidly outward and inward according to demand). From the vendor's perspective, it enables efficient use of equipment with *resource pooling* (i.e. the provider's computing resources such as storage, processing, memory, and network bandwidth are pooled to serve multiple consumers, dynamically assigned and reassigned according to consumer demand), and *measured service* (i.e. metering capabilities enable optimal monitoring, control, and resource allocation for provider and/or consumer).

Cloud computing has three services models: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). They differ in the degree of control that the consumer has over the applications running or the resources used in the cloud infrastructure. With SaaS, the consumer can access provider's applications running on a cloud infrastructure without having control over the cloud infrastructure or application capabilities; with PaaS, the consumer can deploy onto the cloud infrastructure consumer-created applications implemented using programming languages, libraries, services, and tools supported by the provider; with IaaS, the consumer is able to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, including operating systems and applications.

1.2 Problem definition: data centers scaling

Data Centers are the underlying and essential infrastructure supporting the exponential growth of cloud services. They provide storage, processing time, and networking capabilities to the growing number of networked devices, users, and business processes in general. They are basically a collection of servers (providing the computational and storage capabilities), switches and fibers (providing the inter-connectivity between the servers and/or the Wide Area Network (WAN)), and the power and cooling systems required to operate the equipment.

According to Cisco's Global Cloud Index [3] forecast, reproduced in Fig. 1.1, the amount of data center traffic will triple from 2015 to 2020, being the traffic within the data center the main contributor. According to this prediction, the amount of annual global data center traffic in 2020 will reach 15.3 ZB per year (1 ZB = 10^{21} bytes), compared to the 2.3 ZB projected for the total Internet and WAN networks.

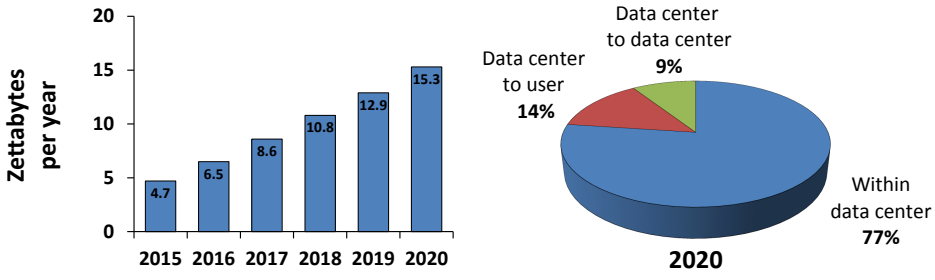

Fig. 1.1 Data center traffic growth and traffic distribution forecast.

Regarding the traffic destination, most of the traffic remains within the data center. 11.8 ZB (77%) are within the data center (e.g. moving data from a development environment to a production environment within a data center, or writing data to a storage array), 2.1 ZB (14%) from data center to user (e.g. streaming video to a mobile device or Personal Computer (PC)), and 1.4 ZB (9%) from data center to data center (e.g. moving data between clouds or copying content to multiple data centers as part of a content distribution network).

The efficient and effective use of data center technologies such as Network Function Virtualization (NFV) (which decouples the logical requirement of the computation from the actual physical infrastructure with faithful abstraction) [4], Software Defined Network (SDN) (which separate the control and forwarding of data center traffic) [5], and the growing importance of data analytics and Internet of Things (IoT) are pushing further the growth of data centers. SDN/NFV will be responsible for 44% of the traffic within the data center by 2020, and big data will be responsible for 17%. Regarding data center storage, the installed capacity will grow from 382 EB (1 EB = 10^{18} bytes) in 2015 to 1.8 ZB in 2020.

With such predictions of increasing demands of storage and computing cloud resources, data centers are forced to continuously expand in size and complexity, leading to the development of large-scale cloud data centers called hyperscale data centers. They are deployed by organizations such as IBM, Microsoft, Amazon, Facebook and Google, and contain hundreds of thousands of servers stored in warehouse-scale buildings consuming tens of megawatts of power. These hyperscale data centers are foreseen to grow from 259 in number at the end of 2015 to 485 by 2020, as shown in Fig. 1.2. By 2020, they will represent 47% of all installed data center servers, 68% of all data center processing power, 57% of all data stored in data centers, and 53% of all data center traffic.

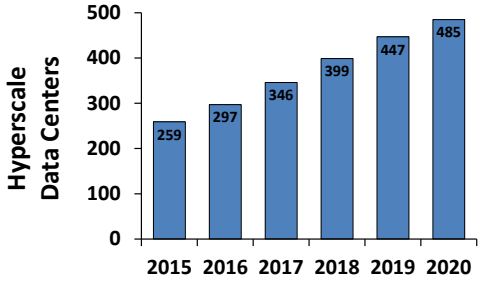

Fig. 1.2 Growth in number of hyperscale data centers forecast.

These hyperscale data centers integrate an increasing number of servers with high-performance and energy-efficient multicore processors, providing higher pro-

cessing capabilities. As the workloads become more distributed in nature, there is an increasing dependence on the networking interconnects between the servers. Consequently, the data center interconnection network should scale accordingly because otherwise the overall system performance may be degraded as indicated by the Amhdal's balanced system law [6] (i.e. Amdahl's Law states that in an efficient computing system there must be a balance between the platform clock speed, the capacity of main memory, and the bit rate of the input/output bandwidth; if any one of these three resources becomes constrained, a computation will be forced to wait). In effect, the interconnects and switching elements should guarantee a balanced bandwidth performance among the underlying compute nodes [7–9].

However, data center networks usually introduce some degree of oversubscription, which limits by construction the performance of these networks [10]. Ideally, the network fabric including the switches connecting servers in a data center should provide full bisection bandwidth, i.e., the available bandwidth between two bisected sections equals to the aggregated bandwidth of the servers in one of the sections. Besides, data center networks are based mainly on electronic switches with front-panel optical transceivers, which result in power hungry and expensive devices dominating power consumption and cost of these networks.

From all the above, it is clear the relevance of investigating the scalability of large-scale, full bisection bandwidth data center networks, especially because most of the traffic happens within the data center. This thesis presents a number of solutions improving the scalability of such networks based exclusively on electronic switches (first part of this dissertation), and hybrid data center networks combining electronic and optical switches (second part of this dissertation).

1.3 Contributions and organization of the thesis

1.3.1 COSIGN project

The project *Combining Optics and SDN In next Generation data center Networks (COSIGN)* was supported by the European Union's Seventh Framework Program for Research (EU FP7). COSIGN investigated solutions to address the emerging demands on data center infrastructures, stressed by data volumes, service provisioning, and consumption trends. It proposed the use of advanced optical technologies to demonstrate novel solutions capable of sustaining the growing resource and operational demands required by next-generation data center networks.

The COSIGN consortium was composed of a unique combination of expertise: Technical University of Denmark, Interoute Communications Ltd., Nextworks, I2CAT, Polatis, University of Bristol, Venture Photonics, Universitat Politecnica de Catalunya, University of Southampton, Technical University of Eindhoven, PhotonX Networks B. V., IBM Israel - Science and Technology Ltd., and OFS.

The work of this thesis was developed as part of the work package *WP2 - High-performance optical subsystems and devices*, led by Technical University of Eindhoven, and responsible for developing the enabling technology for future data center networks by exploiting the latest research and innovation in optical component technology.

1.3.2 Main contributions

Design and implementation of compact low-power electronic switches with On-Board Optics transceivers.
At present, electronic switches based on front-panel optical transceivers suffer from two important bottlenecks: the switching Application-Specific Integrated Circuit (ASIC) and the front-panel bottlenecks [11]. Both bottlenecks end up producing devices of at least one rack unit size integrating a single switching ASIC with 32 front-panel optical transceivers providing a total of 128 ports.

In this work, we analyze how this approach is limiting further scaling of data center networks, and present the importance of shrinking the size and power consumption of the switches. In order to further scale data centers without extending the power consumption and space constraints, we suggest it is required the adoption of compact electronic switches, i.e. switches smaller than one rack unit, with reduced power consumption, and place multiple of them per rack unit.

We demonstrate the feasibility of this approach with a proof-of-concept prototype requiring only one-fourth of the rack space and less than 150 W power consumption. We package four of these devices in a single rack unit, multiplying by four the number of ports and bandwidth density per rack unit (4 x 128 ports, and 4 x 1.28 Tbps). Assuming similar results with the servers, a data center scales to the double number of devices requiring half the space without needing extra power consumption.

Analytic model describing electronic and hybrid topologies.
Recently, hybrid topologies integrating a mixture of electronic and optical switches have been proposed to solve the limitations of data center networks deployed at present, based on electronic switches. Unfortunately, most of the hybrid topologies

lack an analytic model describing them accurately in terms of devices, i.e. in terms of switches, transceivers, and fibers. This fact makes more difficult the investigation of the scaling of such topologies, and also, the comparison with other electronic and hybrid topologies in terms of devices, power consumption, and cost.

In this work, we present an analytic model describing a number of electronic and hybrid topologies with the same model parameters. The topologies follow a Fat-Tree like architecture, and explore the introduction of Wavelength Division Multiplexing (WDM) and optical switching technologies into data center networks. They are named Fat-Tree (FT), Extended Fat-Tree (EFT), Hybrid Fat-Tree (HFT), and Extended Hybrid Fat-Tree (EHFT).

By making use of the presented analytic model, we analyze and compare the scaling in terms of devices, power consumption, and cost of these architectures, concluding that optical switches and wavelength division multiplexing are promising technologies improving the scaling of data center networks.

Dynamic wavelength assignment to WDM links in hybrid data center networks.

At present, Optical Circuit Switches are the only optical switching technology commercially available providing the large port count required for large data center deployments. They provide pure spatial switching, and when used in conjunction with wavelength division multiplexing techniques, they are able to switch together groups of wavelengths. This brings benefits in terms of scaling since fewer switches, transceivers, and fibers are required. Unfortunately, switching together multiple wavelengths restricts the communication patterns and may lead to underutilized links. The solution to this problem is the dynamic assignment of wavelengths to the links, which is usually addressed by the integration of some kind of wavelength selective switches in the network.

In this work, we propose a different approach, a control technique that we name E-WDM. It leverages the benefits of software-defined networking and exploits the electronic switches present in hybrid networks to dynamically assign the traffic to different wavelengths in the high-capacity links of optical circuit switches. We demonstrate the feasibility of this approach in our ECO-IPI Hybrid Data Center demonstrator, which integrates electronic switches, servers, a central controller, and FOX, our fast optical circuit switch.

1.3.3 Outline of the thesis

The remaining of this thesis is structured as follows. Chapter 2 introduces basic concepts and related work regarding data center network technologies (i.e. fiber optics, optical transceivers, electronic switches, and optical switches), and data center network topologies (i.e. topologies based on electronic switches and hybrid topologies based on a mixture of electronic and optical switches). Chapter 3 presents our analysis and solution for electronic switches, the main building block of data center networks deployed at present. Taking Fat-Tree topology as the driving thread, we analyze the impact on the number of ports, different packaging approaches, and the relevance of shrinking size and power consumption of the devices. Furthermore, we present our prototype of compact low-power electronic switch integrating On-Board Optics transceivers and a packaging demonstrator including four of these devices in a single rack unit. Chapter 4 introduces the analytic model governing electronic and hybrid data center networks following a Fat-Tree like architecture, accurately describing the impact of introducing optical switching technologies and wavelength division multiplexing techniques to such topologies. After the definition of the model configuration parameters, and the analysis of the impact of each one of them, a set of equations defining the topologies in terms of switches, transceivers, and fibers is presented. Finally, a number of examples for different configuration parameters is included to demonstrate the flexibility of the model. Chapter 5 explores the scaling in terms of devices, power consumption, and cost of the topologies described by the analytic model. Chapter 6 presents our experimental work regarding hybrid data center networks. It includes FOX (our fast optical circuit switch), the ECO-IPI Hybrid Data Center demonstrator (integrating a central controller, servers, electronic and optical switches), and E-WDM (our software control technique to dynamically assign the traffic to different wavelengths in the high-capacity WDM links of hybrid networks with pure spatial optical switching). Finally, Chapter 7 presents a summary of the thesis and suggests a number of directions for further research.

Chapter 2

Background

Data centers are complex systems comprising the power system, the cooling system, the cabinets with servers, and the network interconnecting them. They are often classified depending on the reliability of the underlying infrastructure into TierI to TierIV groups [1]. For instance, data centers may vary from not including any redundancy in the single power and cooling paths (TierI), to including redundant components in the two active power and cooling paths, tolerating any single equipment failure (TierIV). Typical availability ranges from 99.7 % to 99.995% and above.

The power system transforms the received input *medium-voltage*, typically 10-20 kV, down to *low-voltage*, typically 200-600 V. In parallel, emergency diesel generators may generate similar voltages in case of failure of the main supply system. Both inputs are fed into the Uninterruptible Power Supply (UPS) units, which usually integrate batteries to help with the transition from one source to the other, ensuring the output is always active in case of emergency. The output of the UPS is connected to a number of Power Distribution Unit (PDU), providing the required voltage to servers and switches.

The cooling system usually divides the data center floor into hot and cold aisles because it is more efficient than cooling down the whole data center space. The racks filled with equipment are placed in the cold aisles, which typically receive the cold air through openings in the raised floor. The operation of the equipment situated in the cold aisles generates heat and removes the hot air into the hot aisles. This hot air is collected by the cooling units, sometimes called Computer Room Air Conditioning (CRAC), responsible for cooling it down and pumping it back into the raised floor [12].

Servers provide the computational capabilities and services of the data center, and are responsible for answering the requests from clients in a cloud computing environment. Servers integrate multicore processors, with a number of Dynamic Random Access Memory (DRAM) modules and hard disks. Large-scale data centers integrate more than one hundred thousand servers [13–15] which are packaged in cabinets or racks with forty to fifty rack unit spaces. Typically, they are one rack unit size (1 U), but there is an ongoing effort to shrink its size to half rack unit and beyond [16, 17].

Regarding data storage, there are mainly two approaches. The hard-disks providing the data storage capabilities may be directly attached to the servers, or directly attached to the network switches. In the first case, data is accessed by a globally distributed file system, such as Google File System (GFS) or Microsoft's Distributed File System (DFS). In the second case, the storage is decoupled from servers, and is part of Network-Attached Storage (NAS) devices directly attached to the network using protocols such as Network File System (NFS) on Unix systems or Common Internet File System (CIFS) on Microsoft's systems. The first approach trades off higher write overheads for lower cost, higher availability, and higher read bandwidth [1]. In this work, we assume the first approach with storage included in the servers.

Although the power system, the cooling system, and the servers are a crucial part of data centers, this work focuses in the data center network, the substrate over which the servers inter-operate and are connected to the WAN. As data centers continue to scale in size, the critical performance bottleneck has shifted from the server to the network [8]. Indeed, the data center network should keep pace with multicore processors advancements because otherwise it will become the system performance bottleneck, as stated by Amdahl's Law [6, 9].

The basic building block of the network is the switch (or router) that interconnects the servers or processing nodes according to some prescribed topology. The past 20 years have seen orders of magnitude increase in off-chip bandwidth spanning from several gigabits per second up to several terabits per second today. The switches are interconnected with optical fiber and deployed according to certain topology, which must be carefully selected given its important implications on scalability, cost, and latency and throughput performance.

The remainder of this chapter is organized as follows. First, an overview of data center network technologies is presented, namely fiber optics, optical transceivers, electronic switches, and optical switches. Then, a survey of data center network topologies is carried out, organized in electronic networks (with only electronic switches) and hybrid networks (with a combination of electronic and optical switches).

2.1 Data center network technologies

2.1.1 Fiber optics

Fiber optics plays a critical role as the transmission medium in data centers. With data rates at 10 Gbps and beyond, passive and active copper cables are impractical above a few meters of reach due to their bulky size, frequency-dependent loss, and high power consumption transceivers [18]. With the ever-increasing demands for speed, the optical interconnects probably will replace the traditional copper-based solutions even for the links between servers and rack switches [19].

There are mainly two types of fibers in data centers at present: Multi-Mode Fiber (MMF) and Single-Mode Fiber (SMF). MMF is more expensive to manufacture than SMF due to its complex refractive index profile [20]. However, MMF has substantially larger core diameters (e.g. 50 to 60 microns) than SMF (e.g. 8 to 9 microns). Consequently, MMF offers relaxed alignment tolerances, enabling low-cost assembly and packaging of VCSEL-based transceivers dominating the short reach interconnects within data centers [21].

One disadvantage of MMF when increasing the data rates is the decrease in link reach length due to effects of modal and chromatic fiber dispersion that significantly distort the signal [22]. For instance, the maximum distance reduces from around 500 m at 1G to around 100 m at 25G [23]. The introduction of lower attenuation and higher bandwidth optical fibers (from OM3 with maximum attenuation of 3.5 dB/km and 2700 MHz·km bandwidth to OM4 with 3.0 dB/km and 4700 MHz·km) enables further reach: e.g. at 25G, 100 m with OM3 and 150 m with OM4. Longer distances typically require SMF together with singlemode lasers [18].

2.1.2 Optical transceivers

Optical transceivers are responsible for the Electronic-to-Optical (E/O) and Optical-to-Electronic (O/E) conversion. They are required between the electrical interfaces of the (electronic) switching ASIC and the optical fibers used as the transmission medium. The basic optical transmitter includes a Laser Driver (LD) and a Vertical Surface Emitting Laser (VCSEL) or Distributed Feedback (DFB) laser. The basic optical receiver includes a Photodiode (PD) and a Transimpedance Amplifier / Limiting Amplifier (TIA/LA) [24]. On top of that, transmitter and/or receiver may include additional circuits to better overcome the impairments of the transmission medium, such as Clock Data Recovery (CDR), Feed Forward Equalization (FFE), and Decision Forward Equalization (DFE) circuits [25].

Optical transceivers may be classified according to different considerations. Perhaps the most fundamental classification relies on the type of fiber intended to be used: Multi-Mode (MM) transceivers are normally based on VCSEL arrays and dominate Short Reach (SR) interconnections within the data center; Single-Mode (SM) transceivers are typically based on more expensive, higher-power DFB lasers, and dominate the Long Reach (LR) connections within the data center. A number of approaches have been suggested to overcome the inherent reach limitation of MM transceivers. For instance, the integration of a mode filter to reduce the spectral width of the VCSEL and thereby mitigate the effects of fiber dispersion demonstrating transmission distances over 500 m at 25 Gbps [26]. Another study utilizes digital signal processing techniques such as FFE and DFE to recover the signal integrity loss [27]. Looking forward, the operation of VCSELs at higher data rates may need more advanced modulation formats such as Pulse-Amplitude Modulation - 4 levels (PAM4) [28], multicore fibers [29], and/or WDM techniques.

Regarding the use of WDM, optical transceivers typically transmit data at (or around) one of the three primary wavelengths: 850 nm, 1310 nm, and 1550 nm (corresponding to the first, second and third transmission windows in optical fiber, respectively). MM transceivers typically use the first window, and SM transceivers normally use the second and third windows. In general, the narrower spectrum of DFB lasers used in conjunction with SMF makes them more suitable for WDM technologies [19]. Such devices are often classified depending on the channel spacing: 20 nm channel spacing with Coarse Wavelength Division Multiplexing (CWDM) and 0.8-0.4 nm channel spacing with Dense Wavelength Division Multiplexing (DWDM). There are also efforts to introduce WDM in MMF [30–35], called Shortwave Wavelength Division Multiplexing (SWDM). However, SWDM is limited at present to four wavelengths, distances in the order of hundreds of meters, and faces challenges for further scaling in the number of wavelengths.

Regarding format factors and placement considerations, data center switches are dominated at present by front-panel pluggable transceivers. Networks deployed with 10G technologies integrate SFP+ (1 x 10G) and QSFP+ (4 x 10G) transceivers; networks deployed with 25G technologies select SFP28 (1 x 25G) and QSFP28 (4 x 25G) transceivers. The QSFP - Double Density (QSFP-DD) standard has been recently released [36], and it provides eight interfaces at 25G with Non-Return-to-Zero (NRZ) modulation or 50G with PAM4 modulation. Although there are other format factors, such as the 100G Form-factor Pluggable (CFP) family [37], they are not well suited for the high-density interconnects required in data centers due to its density. For instance, only four CFP2 modules can be integrated into the front-panel of a rack unit, which is not enough to provide access to the number of ports and bandwidth of the latest switching ASICs.

Front-panel pluggable optical transceivers are chosen for data centers at present probably due to its ease of replacement and the existing standards allowing interchangeable devices from different manufacturers. However, they lead to the front-panel bottleneck [11] and limit the number of ports and bandwidth that can be provided per rack unit. In order to overcome this limitation and other important considerations, it is foreseen that optical transceivers allowing tighter integration with the switching ASIC will replace the front-panel pluggable devices [11, 38]. Indeed, as the data rate increases, signal distortion in the electrical transmission lines between optical modules and switching ASIC cannot be ignored. To suppress the signal distortion, optical modules must be located close to the ASIC and at high-density given the increase in ASIC bandwidth [39].

One option is the integration of On-Board Optics (OBO) devices. These devices have important advantages compared to front-panel transceivers: they have a larger port-count and bandwidth density, more energy-efficient electrical interfaces, and can be placed on the Printed Circuit Board (PCB) closer to the switching ASIC enabling reduced power consumption [39, 40]. However, they are not as easily replaced and they lack at present of published standards enabling interchangeable devices. There are already commercially available versions, such as the FIT Micropod and Minipod [41], Finisar BOA [42], or TE Connectivity Coolbit [43, 44]. As it will be shown in Chapter 3, we follow this approach to integrate switching ASIC and transceivers in our compact low-power electronic switch demonstrator.
Other approaches allow even tighter integration of ASIC and transceivers, placing the transceivers in the same substrate as the ASIC [45–47], or on top of the modules like in the POWER7-IH system [48].

Unfortunately, all of these approaches have to still overcome standards like the ones released for front-panel transceivers, making the adoption of such technologies difficult. In that sense, the Consortium for On-Board Optics (COBO) [49] is developing a standard for OBO, which will provide eight lanes at 50G using 25G symbol rate with PAM4 modulation using the IEEE CDAUI-8 electrical interface [50].

2.1.3 Electronic switches

An Electronic Switch (ES) performs the switching function in the electronic domain. The main component of these devices is the switching ASIC. It includes electronic buffers to store the packets received in the input ports, and also, processing units to decode the packet headers and forward the packets to the corresponding output ports according to the programmed routing tables.

Although electronic switches are widely deployed, and it is common to find them also in home networks, there are important differences between home and data center switches. The first distinction is that the interfaces in data center switches are typically optical thanks to the inclusion of optical transceivers, at least for long distances given the increasing data rates of the interfaces. The second difference is that the number of ports is significantly larger: e.g. state-of-the-art versions scale-up to 128 ports with 10G [51] and 25G [52] interfaces. More recently, switching ASICs with 50G interfaces (2 x 25G with NRZ modulation) [53] have been released, and with 100G (2 x 50G with PAM4 modulation) [54] interfaces have been announced. The limitation of the number of ports per ASIC despite the increasing bandwidth is sometimes referred as to the ASIC bottleneck [11].

Since the switching ASIC has electrical interfaces, it requires the co-integration of optical transceivers when it interacts with an optical transmission medium. ASIC and optical transceivers are typically connected through PCB high-speed differential traces. At present, industry favors front-panel pluggable transceivers because of its ease of replacement and the available standards enabling interchangeable modules from different manufacturers. This approach leads to devices occupying at least one rack unit (1 U) when packaging a single switching ASIC. The layout of a typical electronic data center switch is visualized in Fig. 2.1. The main component is a PCB, which integrates the switching ASIC, the required receptacles for the front-panel transceivers and the socket for the Central Processing Unit (CPU) control board. In addition, the rack unit includes a redundant power supply and hot-air extracting fans located in the back.

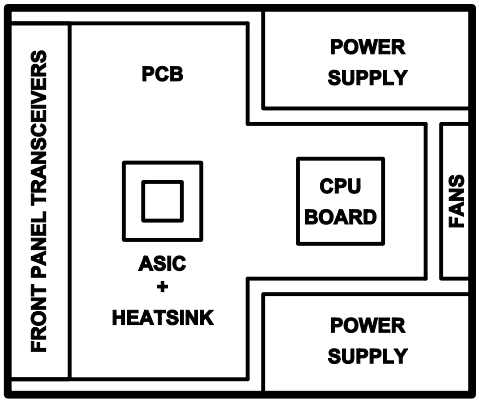

Fig. 2.1 Typical electronic switch based on front-panel pluggable transceivers.

A comparison of commercially available electronic switches [55] is reported in Table 2.1, with a number of relevant remarks. First, all the switches are packaged

with front-panel optical transceivers: SFP+ and QSFP+ are the chosen format factors for 10G interfaces; SFP28 and QSFP28 are selected for 25G interfaces. Second, the devices requiring one or two rack units integrate a single switching ASIC with the corresponding number of optical transceivers. The front-panel in a single rack unit is limited to a maximum of 36 QSFP - 4x10G (QSFP+) or 36 QSFP - 4x25G (QSFP28), which is sometimes referred to as the front-panel bottleneck [11, 38]. This forces the packaging of the BCM56970 ASIC in two rack units because it needs to accommodate 64 QSFP28 transceivers to provide access to the 128 50G ports. In Chapter 3 we suggest and demonstrate a solution to the ASIC and front-panel bottlenecks of electronic switches, based on the integration of On-Board Optics transceivers and the packaging of multiple ASICs per rack unit. Our solution, even being based on 10G devices, achieves 4 x 1.28 Tbps and 4 x 128 ports per rack unit. Third, solutions such as Facebook Sixpack and Facebook Backpack implement a large switch based on multiple switching ASICs: e.g. they integrate twelve switching ASICs to implement a device with four times the number of ports of a single switching ASIC. As these examples illustrate, and as it will be demonstrated in Chapter 3 for a Fat-Tree topology, this approach requires three times more switching ASICs and points out the benefits of scaling-out the number of commodity networking devices [56]. Note that 1 G4 is 13.85 inches high, which is approximately 8 U of 1.75 inches high.

Table 2.1 Commercial electronic switches.

Manufacturer / Product name	**Switching ASIC**	**BW (Tbps) // Ports per rack unit**	**Optical transceivers**	**Size**
Edgecore Networks AS6712-32X	BCM56850	1.28 // 128 @ 10G	32 QSFP+	1 U
Edgecore Networks AS7712-32X	BCM56950	3.2 // 128 @ 25G	32 QSFP28	1 U
Edgecore Networks AS7800-64X	BCM56970	3.2 // 128 @ 25G	64 QSFP28	2 U
Facebook Sixpack	12 BCM56850	2.2 // 85 @ 10G	128 QSFP+	7 U
Facebook Wedge 100	BCM56950	3.2 // 128 @25G	32 QSFP28	1 U
Facebook Backpack	12 BCM56950	4.8 // 64 @ 25G	128 QSFP28	1 G4 $\approx$ 8 U

2.1.4 Optical switches

An Optical Switch (OS) performs the switching function in the optical domain. Therefore, an important advantage of optical switches compared to electronic switches is that they do not require optical transceivers to perform the E/O and O/E conversion. This enables reduced power consumption and cost, and it is further investigated in Chapter 5. Another important advantage of OSs is that they are bit-rate and data-format agnostic, ideally suited to switch very high data-rate signals and multiple wavelengths. Again, this enables more power efficient devices since they do not require to store-and-forward every bit like ESs [57].

However, OSs face also important challenges, such as the lack of optical buffers [58]. Indeed, ESs rely on electronic buffers to store-and-forward the packets, enabling the ability to cope with contention. This important limitation of optical switches forces the implementation of some other techniques to deal with contention. A suggested approach is the addition of (fixed) fiber delay lines used as buffers. However, this adds significant complexity and may cause performance degradation since they cannot delay the packets for arbitrary amounts of time [58]. Another approach is to move the packet buffering out to the edge points of the network where the data can be buffered electronically [59, 60]. This works well if the switch fabric configuration can be efficiently controlled by a logically centralized traffic arbiter and scheduler [57]. Other approaches suggest the introduction of electronic buffers in the optical switches [61–63].

A number of optical switching techniques have been investigated for data center applications: optical circuit switching (OCS), optical burst switching (OBS), and optical packet switching (OPS). In an OCS-based solution, the connectivity path between the source and destination is established before sending the data. This is a time-consuming procedure [64], but once the circuit is established, the data is sent at maximum throughput with minimum delay. In an OBS-based solution, a burst control header is created and sent towards the destination. The header is processed electronically at the OBS-routers, which allocate the optical connection path for the duration of the data transfer [65]. In an OPS-based solution, one or more electrical data packets with similar attributes are aggregated in an optical packet and attached with an optical label indicating the destination. The optical packet switch processes the label and forwards the optical packet to the right output port [66].

Plenty of technologies have been suggested to implement the optical fabric. The resulting devices may be classified into *slow* optical switches and *fast* optical switches [67]. Slow optical switches are based on technologies such as Micro-Electro-Mechanical System (MEMS) [68, 69] or direct fiber-to-fiber alignment

switches [70], providing large port-count devices with reconfiguration times in the order of microseconds or milliseconds, respectively. Fast optical switches with faster reconfiguration times are typically based on Semiconductor Optical Amplifiers (SOAs) [57, 71, 72] or in a combination of Tunable Lasers (TLs), Tunable Wavelength Converters (TWCs), and Arrayed Waveguide Grating Routers (AWGRs) [59, 73–77]. SOA-based fast optical switches are normally based on broadcast-and-select networks [71, 72], although there are also other approaches such as the 16-port optical switch implemented with a Clos network of smaller 4-port SOA-based switching elements [57].

Finally, it is important to mention that although optical switching is being successfully deployed in traditional telecommunication networks, it is still finding difficulties to fully enter into data centers. This is mainly because the switching speed and the port count of the optical fabric do not fulfill the requirements for data center applications yet [73, 77]. Unfortunately, there are not fast optical switches with large port-count commercially available at present. However, recent advances in silicon photonics integration are promising [78–80].

2.2 Data center networks topologies

Data Center Network (DCN) topologies describe how to physically connect switches and servers. The selection of the data center topology has important implications in terms of scaling, cost, latency, throughput, fault tolerance, path diversity, power consumption, and many other relevant features [10]. Thus, the number of proposed architectures for data centers continuously grows [81].

DCN topologies are typically divided into direct and indirect topologies. Direct topologies connect servers to each switch. Indirect topologies connect the servers only to certain switches; the rest of the switches are interconnected among them. Classic examples of direct topologies are mesh, torus, and hypercube networks; a classic example of indirect topology is a tree [9].

Another manner to classify DCN topologies is in electronic, hybrid, and all-optical networks. Electronic DCNs are based exclusively on electronic switches, and up to present, they dominate data center deployments. Hybrid DCNs, integrating electronic and optical switches, and all-optical DCNs, including only optical switches, are attracting attention. They may overcome issues as the large power consumption of electronic switches or the cost of optical transceivers (dominating power consumption and cost of DCNs, as it will be shown in Chapter 5). This section presents a brief overview of a number of electronic and hybrid topologies.

2.2.1 Topologies based on electronic switches

There are plenty of topologies based on electronic switches. For instance, DCell [82], BCube [83], and MDCube [84] are server-centric topologies, where servers are equipped with multiple network ports and act not only as end hosts but also as relay nodes for traffic forwarding. Scafida [85] and JellyFish [86] topologies suggest a more random, asymmetric distribution of the network.

We will focus on this section on two topologies: Fat-Tree [14, 56, 87, 88] and HyperX [89].

Fat-Tree is a well-known topology, deployed in high-performance computers [90] and data centers (e.g. Google and Facebook) [91–93]. As it will be shown in Chapter 3 to Chapter 5, we select this topology has the driving thread to evaluate the scaling of data center networks.

Fat-Tree is an indirect topology with remarkable features and here we present only a number of them. First, it scales to any number of servers which is especially relevant in order to support the ever-increasing demands of bandwidth in data centers. It does so by interconnecting layers of switches with any number of ports. Other topologies limit the network size depending on the number of ports of the switches. Second, it provides full bisection bandwidth, i.e, if we partition the network in two bisections with half the servers, there are enough links connecting the bisections to communicate the servers at full link speed. This is also remarkable because other solutions are only able to connect a large number of servers by introducing oversubscription. Third, it has great path diversity which enables load balancing and fault resiliency: e.g. the Fat-Tree example of Fig. 2.2 implemented with two layers of 8-port switches provides four different paths between any pair of servers.

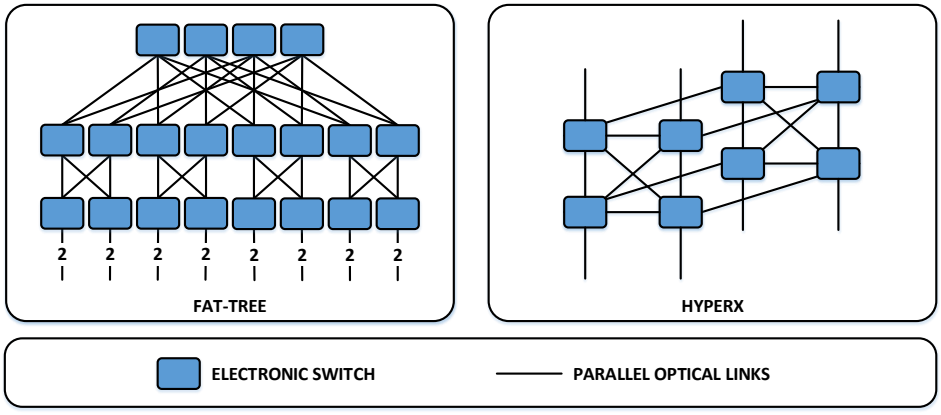

Fig. 2.2 Fat-Tree and HyperX topologies.

HyperX provides the analytic framework describing a number of direct topologies, where each switch is connected to all of its peers in each dimension and the number of switches in each dimension can be different. An example of a three-dimensional HyperX built with 4-port switches is shown in the right diagram of Fig. 2.2. The HyperX model describes using the same parameters a number of topologies, such as the one-dimensional fully connected ring, the hypercube, and the Flattened-Butterfly [94, 95]. Such analytic model is the inspiration for our Hybrid Fat-Tree model of Chapter 3, which describes a number of hybrid topologies exploring the introduction of optical switches and wavelength division multiplexing technologies in Fat-Tree. Although HyperX may provide, like Fat-Tree, full bisection bandwidth among the servers, it is limited in scaling by the number of ports of the switches.

2.2.2 Topologies based on electronic and optical switches

Due to the high cost and power consumption of data center networks based exclusively on electronic switches (and optical transceivers), alternative approaches suggest the combination of electronic and optical switching technologies in hybrid networks [67, 96–103]. Despite the advantages of optical switching, the combination of these technologies bring also important challenges, since the temporal and spatial heterogeneity in data center traffic require supporting dynamic switch scheduling decisions on aggressive time scales [104].

The hybrid architectures suggested differ in the choice of optical switching technologies, and also, in the way devices are interconnected. For instance, Helios [96] and HyPaC [97] architectures include slow optical circuit switches based on MEMS, as shown in Fig. 2.3.

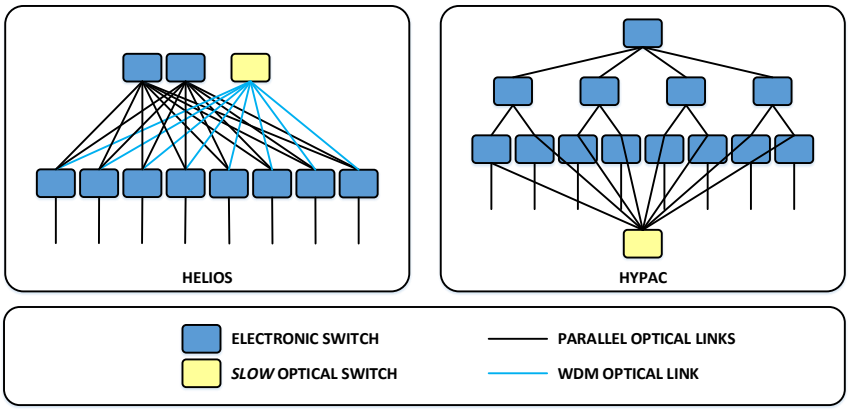

Fig. 2.3 Helios and HyPaC topologies.

Helios substitutes a number of the electronic switches at the core of the network by optical circuit switches; HyPaC connects the optically switched network directly to the racks, creating an alternative circuit switched path parallel to the packet switched network. A number of important cloud applications, including virtual machine migration, large data transfers, and MapReduce experience significant performance improvements while running on such a hybrid network with the potential for much lower cost, deployment complexity, and energy consumption than purely packet-switched networks [105].

Examples of other hybrid topologies such as Proteus, HOSA, and OpSquare are shown in Fig. 2.4. Proteus [98] includes a single MEMS-based Optical Circuit Switch (OCS) to connect all the racks in the network. It also exploits DWDM technologies and Wavelength Selective Switches (WSSs) in order to achieve dynamic bandwidth allocation in the high-capacity WDM links. Every rack switch is connected to the MEMS switch through a multiplexer and a WSS unit. For instance, a 64-port rack switch has 32 ports with 32 different wavelengths facing up. These wavelengths are multiplexed into a single fiber, which is divided into four groups by the WSS, and connected to four ports of the OCS. By using the dynamic configuration capabilities of the WSS, each one of the four links may be assigned a capacity between 0 and 32 times the link speed. Unfortunately, Proteus does not scale to a large number of servers since a single optical switch connects all the racks. Also, it relies on (still expensive) technologies such as DWDM and WSS, and includes oversubscription to reduce the cost of the network. HOSA [67] suggests exploiting slow and fast optical switches at the core of the network, interconnecting the racks. OpSquare [100–103] employs two parallel levels of fast optical switches, one for intra-cluster connectivity and the other one for inter-cluster connectivity.

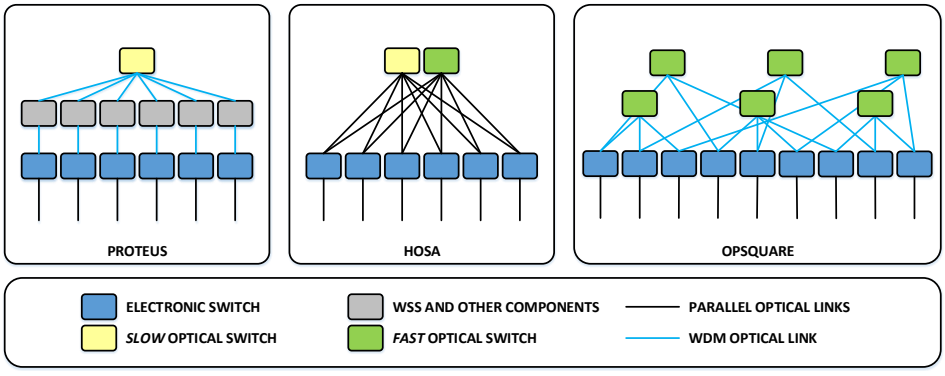

Fig. 2.4 Proteus, HOSA, and OpSquare topologies.

An important limitation of the mentioned hybrid architectures is that they do not provide an analytic model describing them in terms of devices. This makes more difficult the investigation of the scaling of these networks in terms of devices, power consumption and cost, and the comparison with other electronic topologies.

2.3 Summary

This chapter has introduced a number of basic concepts used in the remaining of this thesis. The basic architecture of a data center is briefly explained in terms of the system blocks. Since the focus of this work relies on data center networks, a summary of technologies available at present to deploy such networks is presented: fiber optics, optical transceivers, electronic switches, and optical switches. Finally, a brief survey of a number of electronic and hybrid topologies suggested in the literature is included, with special attention to Fat-Tree, HyperX, Helios, and HyPaC architectures.

As it will be shown, Chapter 3 presents a solution to improve the scaling of electronic switches by integrating On-Board Optics devices instead of the traditional front-panel optical transceivers, and by placing multiple switches per rack unit instead of only one. Chapter 4 introduces our analytic model describing electronic and hybrid topologies. The model, inspired by HyperX, explores the introduction of optical switches and wavelength-division multiplexing technologies into Fat-Tree like topologies, which are able to scale to any number of servers with full bisection bandwidth. Helios and HyPaC architectures are particular examples of our model. Chapter 5 employs the analytic model to investigate the scaling of these topologies in terms of devices, power consumption, and cost. Finally, Chapter 6 presents an experimental demonstration of the feasibility of integration of these technologies in a hybrid data center scenario by means of an SDN central controller.

Chapter 3

Electronic Switches with On-Board Optics

3.1 Introduction

At present, data center networks are mostly based on electronic switches. These switches perform the electronic switching function in the electronic domain, and for that reason, they require the integration of optical transceivers to perform the E/O and O/E conversion to interface the optical fiber transmission medium. Given the ever-increasing demands for cloud computing, these networks are forced to continuously expand, integrating a growing number of more powerful devices. The largest data centers integrate at present more than one hundred thousand servers which require thousands of racks space, tens of megawatts of power consumption, and tens of millions of dollars in deployment costs. Further scaling is challenging.

As discussed in Section 2.1.3, electronic switches suffer from two important bottlenecks [11, 38] in the switching ASIC and the rack-unit front panel. Regarding the first bottleneck, the switching ASICs of the electronic switches is limited at present to 128 ports. Even the newer versions with 256 Serializer-Deserializer (SerDes) still provide only 128 ports. This limitation has a great impact on the scaling of data centers because large port-count switches are very desirable to reduce the size and diameter of the network, i.e., large port-count switches enable flatter data centers. Regarding the second bottleneck, the 128-port switching ASICs are packaged at present with 32 optical transceivers that completely fill the front panel area of the rack unit (the newer version with 256 SerDes requires 64 optical transceivers and two rack units). Thus, the choice of pluggable optical transceivers limits the scaling of the rack unit because its number of ports and bandwidth is

constrained by the number and type of optical transceivers that can be fitted in the front panel area.

The choice of front panel transceivers has other problems associated. An important disadvantage is that long PCB high-speed traces are required to reach the front panel. This forces electrical interfaces in ASIC and transceivers to be able to cope with the associated losses, which increase with the growing data-rates. Thus, these devices must include additional circuits to overcome the impairments of the transmission medium, such as CDR, FFE, and/or DFE circuits [25], which traduces in extra complexity, power consumption, and cost. Another important problem is related to thermal management because the front panel optical transceivers are squeezed and stacked in the front panel area. This makes more difficult the air flow and jeopardizes the performance and reliability of the transceivers, highly dependent on temperature.

This chapter investigates theoretically and experimentally how to improve the scaling of the electronic switches, and consequently, the scaling of current data center networks.

First, in our theoretical analysis, we compile a set of four design principles for electronic switches by using the analytic model of FT topology as the driving thread. The first design principle remarks the relevance of scaling-up the port-count of electronic switching ASICs: the manufacturers should focus on increasing the number of ports of these devices and the switch designers should select the ASICs with the largest port-count available. The second design principle points out that it is more efficient to scale-out the number of commodity switches in the network based on a single switching ASIC than building larger port-count devices integrating multiple switching ASICs. Finally, the third and four design principles indicate that special attention should be taken to reduce the size and power consumption of these devices as much as possible because these parameters limit the scaling of data centers with constraints in space and power.

Following these design principles, the second part of the chapter presents an experimental prototype which overcomes the front panel and ASIC bottlenecks. The first key decision in the design process is the integration of OBO transceivers to overcome the front panel bottleneck. OBO devices have a compact size, with higher port and bandwidth density. They are placed on the PCB surrounding the ASIC. The result is a front panel area free of transceivers, and very compact devices with reduced power consumption and improved thermal behavior. Our prototype requires only one-fourth of the rack unit space, consumes by design less than 150 W, which is half of the corresponding power consumption of similar devices based on front panel pluggable transceivers, and operates the transceivers below 40° C. The second key decision in the design process is the placement of multiple compact

electronic switches per rack unit to overcome the ASIC bottleneck. In effect, we suggest it is time to start placing multiple switches per rack unit, inspired by similar approaches placing multiple servers per rack unit or multiple cores per processing unit [16, 17]. We demonstrate the feasibility of this approach with our packaging demonstrator integrating four switches in a single rack unit. The result is a rack unit with four times the number of ports and bandwidth density compared with similar devices based on front panel transceivers.

The remainder of this chapter is organized as follows. First, Section 3.2 presents an overview of the well-known FT architecture, including an example, the model parameters, and the analytic model. We extend the model including additional parameters and equations to compute the number of transceivers and fibers of the network. This model is used to investigate the scaling of the network and to infer four design principles in Section 3.3. Then, Section 3.4 reports the design, implementation, packaging, and characterization of our prototype. Finally, Section 3.5 concludes the chapter.

3.2 Fat-Tree topology

As discussed in Chapter 2, FT has two interesting features to build large-scale high-performance data center networks: it provides full bisection bandwidth and it is not limited in scaling by the number of ports of the switches. In this section, we present our analytic model of FT, which extends the traditional model by including two novel parameters and two novel equations. Our extended model allows computing the number of devices (i.e. switches, transceivers, and fibers), which in turns, enables also to calculate the power consumption and cost. It will be used as the baseline for our investigation of the scaling of electronic and hybrid networks in Chapter 5. An example of FT topology implemented with 3-layers of 8-port switches is shown in Fig. 3.1. The example connects 128 servers with 80 switches.

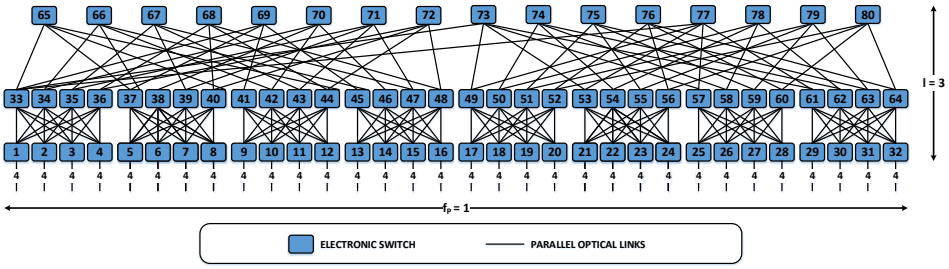

Fig. 3.1 FT network with 3 layers and 8-port switches.

The analytic description of the FT topology requires the definition of the model parameters summarized in Table 3.1. The parameters usually found in the literature [56, 89] are the number of ports in the switches, k, and the number of layers, l. For instance, the example previously shown in Fig. 3.1 has k = 8-port switches and l = 3 layers.

Table 3.1 Model parameters defining FT topology.

Symbol	**Description**
$k \in [2,\infty) \in \mathbb{N}$	number of ports per switch (k is power of 2)
$l \in [2,\infty) \in \mathbb{N}$	number of layers
$f_P = 1/2^n \; with$ $n \in [0, \log_2 k/2] \in \mathbb{N}$	fraction of the full network implemented
$f_{E\ NO\ WDM} = 1/2^n \; with$ $n \in [0, \log_2 k/2] \in \mathbb{N}$	reciprocal of the number of ports of NO WDM transceivers connecting electronic switches

In order to add generality and flexibility to the model, we added two novel parameters: the partition factor f_P and the $f_{E\ NO\ WDM}$ factor. f_P adjusts the size of the network in a discrete manner, ensuring that no resources are wasted and that links are evenly distributed. It allows to overcome the abrupt scaling of FT topologies with the number of layers; e.g. the network scales from 8192 with two layers of 128-port switches to 524888 servers with three layers. In addition, it ensures that the size adjustment of the topologies results in a valid solution. Otherwise, adjusting the network size to a certain number of servers could result in unused ports in the switches (wasting resources) or in non-homogeneous connections (making more difficult deployment and routing decisions). This factor is further explained in Section 4.3.1. The $f_{E\ NO\ WDM}$ factor represents the number of ports of the NO WDM transceivers. It is useful to compute the number of transceivers with the corresponding equation added in our extended model. The NO WDM transceivers are based on parallel optical channels and are thus an exercise in packaging together single channel transceivers into a combined packaged form factor. In that respect, the increase in the number of ports in a NO WDM transceiver has no impact on the number of fibers in the system. It may, however, impact the cost and power as there are small gains to be had by co-integrating multiple single channel transceivers into a single package.

The analytic model of FT includes Eq. (3.1), Eq. (3.2), Eq. (3.3), and Eq. (3.4). They compute the number of servers N_{FT}, switches S_{FT}, transceivers T_{FT}, and fibers F_{FT} in the network, respectively.

$$N_{FT} = 2 \cdot f_P \cdot (k/2)^l \tag{3.1}$$

$$S_{FT} = (2 \cdot l - 1) \cdot f_P \cdot (k/2)^{l-1} \tag{3.2}$$

Eq. (3.3) calculates the number of transceivers as a function of the model parameters. It can be understood as the number of transceivers with $1/f_{E\ NO\ WDM}$ ports required by k-port S_{FT} switches. It can also be expressed as $T_{FT} = S_{FT} \cdot k \cdot f_{E\ NO\ WDM}$, or $T_{FT} = (2 \cdot l - 1) \cdot f_{E\ NO\ WDM} \cdot N_{FT}$. The equation includes only the transceivers required by the switches and it does not include the 1-port N_{FT} transceivers needed by the servers.

$$T_{FT} = 2 \cdot (2 \cdot l - 1) \cdot f_{E\ NO\ WDM} \cdot f_P \cdot (k/2)^l \tag{3.3}$$

Eq. (3.4) obtains the number of fibers as a function of the model parameters. It can also be expressed as $F_{FT} = 2 \cdot l \cdot N_{FT}$, since every layer in FT requires $2 \cdot N_{FT}$ fibers.

$$F_{FT} = 4 \cdot l \cdot f_P \cdot (k/2)^l \tag{3.4}$$

The number of switches, transceivers, and fibers required to build 3-layer networks with 128-port switches is represented in Fig. 3.2. The curves are obtained using the equations of the analytic model. The markers of the curves represent a different value of the partition factor.

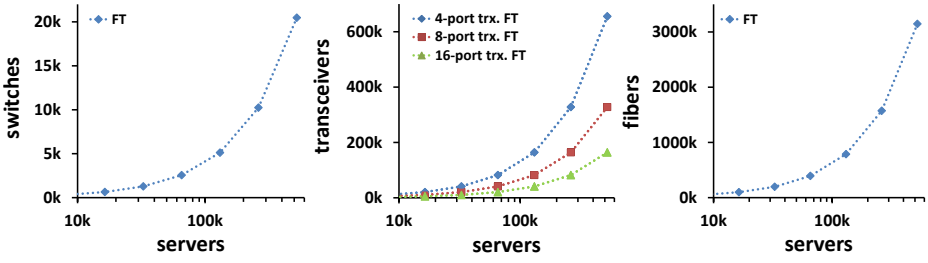

Fig. 3.2 Devices in FT network with 3 layers and 128-port switches.

The highest points in the curves represent the full network with f_P = 1 and 524288 servers; the second highest points represent a half of the network with partition factor f_P = 1/2 and 262144 servers; and so on. For instance, a 131072 servers network requires 5120 128-port switches, 163840 4-port NO WDM transceivers, and 786432 fibers.

3.3 Design principles for electronic data center switches

This section summarizes four design principles compiled to improve the scaling of data center switches, which is limited at present by the ASIC and front panel bottlenecks as discussed in Section 2.1.3 and Section 3.1. The presentation of each design principle starts with the introduction of the intuitive idea governing the principle. Then, the FT topology and its analytic model presented in Section 3.2 are employed as the driving thread to explore analytically and graphically the impact of each design principle on the scaling of large data center networks.

3.3.1 Principle 1: scale-up port count of switching ASIC

The basic idea behind the first principle is simple: *the more ports in the switching ASICs, the better*. These devices are the main piece of hardware required to implement electronic data center switches and their number of ports is probably the most important factor in the scaling of data center networks [106]. A large number of ports in the switching ASICs reduces the number of switches needed to implement the network. This implies additional reductions in latency, power consumption, and cost. Unfortunately, the effect of increasing the number of ports in the switching ASIC does not reduce the number of transceivers and fibers of the network. For instance, even if only one-half of the switches (with the double number of ports) is needed to interconnect a certain number of servers, the number of transceivers and fibers remains constant. This is due to the fact that these half number of switches with the double number of ports require the double number of transceivers and fibers.

Analytically, we can consider two networks with the same number of servers and layers, but built with switches of a different number of ports, i.e., k_1 and $k_2 = n \cdot k_1$. According to Eq. (3.1) to Eq. (3.4), the ratio of servers, switches, transceivers, and fibers in both networks is given by Eq. (3.5) to Eq. (3.8).

$$\frac{N_{FT}\Big|_{k=k_1}}{N_{FT}\Big|_{k=k_2=n\cdot k_1}} = \frac{f_{P1}}{f_{P2}} \cdot \left(\frac{k_1}{k_2}\right)^{l} = \frac{f_{P1}}{f_{P2}} \cdot \left(\frac{1}{n}\right)^{l} \tag{3.5}$$

$$\frac{S_{FT}\Big|_{k=k_1}}{S_{FT}\Big|_{k=k_2=n\cdot k_1}} = \frac{f_{P1}}{f_{P2}} \cdot \left(\frac{k_1}{k_2}\right)^{l-1} = \frac{f_{P1}}{f_{P2}} \cdot \left(\frac{1}{n}\right)^{l-1} \tag{3.6}$$

$$\frac{T_{FT}\Big|_{k=k_1}}{T_{FT}\Big|_{k=k_2=n\cdot k_1}} = \frac{f_{P1}}{f_{P2}} \cdot \left(\frac{k_1}{k_2}\right)^{l} = \frac{f_{P1}}{f_{P2}} \cdot \left(\frac{1}{n}\right)^{l} \tag{3.7}$$

$$\frac{F_{FT}\Big|_{k=k_1}}{F_{FT}\Big|_{k=k_2=n\cdot k_1}} = \frac{f_{P1}}{f_{P2}} \cdot \left(\frac{k_1}{k_2}\right)^{l} = \frac{f_{P1}}{f_{P2}} \cdot \left(\frac{1}{n}\right)^{l} \tag{3.8}$$

Given the fact that both networks have an equal number of servers, Eq. (3.5) is reduced to 1 ($N_{FT}\big|_{k=k_1} = N_{FT}\big|_{k=k_2=n\cdot k_1}$). Based on this assumption, Eq. (3.6) to Eq. (3.8) can be simplified to Eq. (3.9) to Eq. (3.11). While Eq. (3.9) means that n times more k_1-port switches are required to build an equal size network implemented with k_2-port switches, Eq. (3.10) and Eq. (3.11) show that the number of transceivers and fibers in both networks are equal.

$$\frac{S_{FT}\Big|_{k=k_1}}{S_{FT}\Big|_{k=k_2=n\cdot k_1}} = n \tag{3.9}$$

$$\frac{T_{FT}\Big|_{k=k_1}}{T_{FT}\Big|_{k=k_2=n\cdot k_1}} = 1 \tag{3.10}$$

$$\frac{F_{FT}\Big|_{k=k_1}}{F_{FT}\Big|_{k=k_2=n\cdot k_1}} = 1 \tag{3.11}$$

This conclusion is visualized graphically in Fig. 3.3, that illustrates the benefits of increasing the port-count in the switching ASIC in order to reduce the number of switches required. The curves represent three-layer networks implemented with

64-port, 128-port, and 256-port switches. For instance, a network with 65536 servers requires 5120 64-port switches, 2560 128-port switches, and 1280 256-port switches.

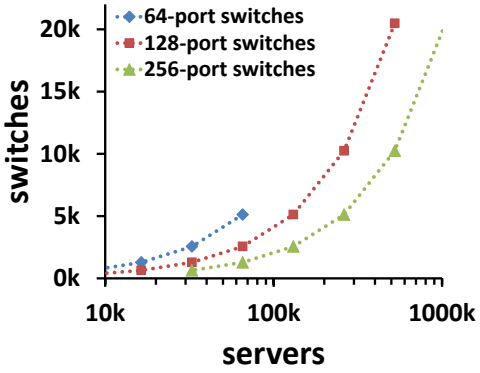

Fig. 3.3 Scale-up port count of switching ASIC.

3.3.2 Principle 2: scale-out number of commodity switches

The second design principle points out that it is more efficient to build the network with commodity switches based on a single switching ASIC than to build the network using larger port-count switches based on multiple switching ASICs [56]. Note that, while the second design principle deals with *scaling-out* the number of switches, the first design principle refers to *scaling-up* the number of ports of the switching ASICs.

According to the first design principle, 5120 128-port switches, 2560 256-port switches or 1280 512-port switches are required in FT to connect 131072 servers. Since only 128-port switching ASICs are available at present, the last two options require the implementation of 256-port switches integrating six 128-port switching ASICs or 512-port switches with twelve 128-port switching ASICs. Thus, the first option is more efficient because the last two options require the triple number of devices, i.e., 15360 128-port switching ASICs.

The next lines demonstrate analytically this conclusion. Switches and servers are related by Eq. (3.12), obtained from the ratio of Eq. (3.1) and Eq. (3.2).

$$S_{FT} = \frac{2 \cdot l - 1}{k} \cdot N_{FT} \tag{3.12}$$

Eq. (3.13) is obtained by particularizing Eq. (3.12) as indicated. It means that $3 \cdot n$ switching ASICs with k ports are required to implement an electronic switch with $n \cdot k$ ports: e.g. six 128-port switching ASICs are required to implement a

256-port switch and twelve 128-port switching ASICs are needed to implement a 512-port switch.

$$S_{FT}\Big|_{\substack{N_{FT}=n\cdot k\\ l=2}} = \left(\frac{2\cdot l-1}{k}\cdot N_{FT}\right)\Bigg|_{\substack{N_{FT}=n\cdot k\\ l=2}} = 3\cdot n \tag{3.13}$$

An important remark is that the number of switches computed with Eq. (3.2) is equal to the number of switching ASICs if and only if these switches are based on a single switching ASIC. In the case that multiple ASICs are used to implement a larger switch, the number of switching ASIC of k_1 ports required to build S'_{FT} switches with $n\cdot k_1$ ports is given by Eq. (3.14).

$$S'_{FT}\Big|_{k=k_2=n\cdot k_1} = 3\cdot n\cdot(2\cdot l-1)\cdot f_{P2}\cdot(k_2/2)^{l-1} \tag{3.14}$$

Thus, the ratio between the switching ASICs required by switches based on a single switching ASIC and multiple switching ASICs is given by Eq. (3.15).

$$\frac{S_{FT}\Big|_{k=k_1}}{S'_{FT}\Big|_{k=k_2=n\cdot k_1}} = \frac{1}{3\cdot n}\cdot\frac{f_{P1}}{f_{P2}}\cdot\left(\frac{k_1}{k_2}\right)^{l-1} = \frac{1}{3\cdot n}\cdot\frac{f_{P1}}{f_{P2}}\cdot\left(\frac{1}{n}\right)^{l-1} \tag{3.15}$$

Finally, Eq. (3.16) is obtained from Eq. (3.15) by assuming again that the networks implemented with both types of switches have the same size, which implies that Eq. (3.5) equals 1. It means that the triple number of switching ASICs are required to deploy a network of a certain size if large port-count switches implemented with multiple switching ASICs are employed.

$$\frac{S_{FT}\Big|_{k=k_1}}{S'_{FT}\Big|_{k=k_2=n\cdot k_1}} = \frac{1}{3} \tag{3.16}$$

The same conclusion is illustrated in Fig. 3.4, which compares the number of switching ASICs required to build 3-layer networks. The green curve represents the case requiring fewer switches, which is not possible to implement today because there are not 256-port switching ASICs available. The red and blue curves display the options feasible at present: the red curve visualizes the case where 128-port switches based on 128-port switching ASICs are deployed; the blue curve presents

the case where 256-port (or 512-port) switches are implemented with six (or twelve) switching ASICs.

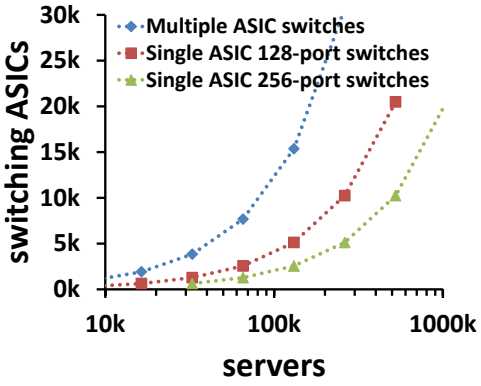

Fig. 3.4 Scale-out number of commodity switches.

3.3.3 Principle 3: shrink size of electronic switches

The third design principle means that it is relevant to build compact switches (and servers) to improve the scaling of data center networks. By compact switches we mean devices smaller than one rack unit (1 U). The current trend implements 1 U devices (switches and servers, although servers are already starting to be shrunk [16, 17]). Consecutive increases in the capacity of data centers by adding more devices of the same size imply that extra space is needed. As a consequence, the cost associated with the floor space, expensive when the data center is located close to the user to reduce latency, and with the length of the fibers also increases. We suggest here that, in order to keep scaling data centers constrained in space, it is mandatory to start shrinking the devices.

Let's denominate V_D to the volume required by a device (switch or server), and consider V_D = 1 U, 0.5 U, and 0.25 U devices in order to investigate the impact of compact devices in data centers constrained in space. Let's denominate V_{MAX} to the maximum space intended for devices in the data center, measured in rack units and sufficient to accommodate all switches and servers. Thus, V_{MAX} may be expressed according to Eq. (3.17) by making use of Eq. (3.12).

$$\begin{aligned} V_{MAX} &= (S_{FT\ MAX} + N_{FT\ MAX}) \cdot V_D = \\ &= \frac{2 \cdot l + k - 1}{2 \cdot l - 1} \cdot S_{FT\ MAX} \cdot V_D = \frac{2 \cdot l + k - 1}{k} \cdot N_{FT\ MAX} \cdot V_D \end{aligned} \tag{3.17}$$

Therefore, the maximum number of switches and servers that can be deployed in a data center of size V_{MAX} is given by Eq. (3.18) and Eq. (3.19). These equations imply that, by constraining the data center to a certain size, a number of solutions may not be valid because they require more than the available space. These equations may provide non-integer solutions and, in that case, they should be rounded down to the closest integer.

$$N_{FT\ MAX} = \frac{k}{2 \cdot l + k - 1} \cdot \frac{V_{MAX}}{V_D} \tag{3.18}$$

$$S_{FT\ MAX} = \frac{2 \cdot l - 1}{2 \cdot l + k - 1} \cdot \frac{V_{MAX}}{V_D} \tag{3.19}$$

A constraint of V_{MAX} = 200 kU is selected to investigate the impact of setting a space limitation in data center deployments. This value of V_{MAX} is chosen in order to set a value large enough to build a data center with more than one hundred thousand servers and assuming that this space should be kept constant for future scaling of the network. This constraint could be set to any other value, and in that case, the exact values of the following examples would change. Thus, the following results should be understood just as an example to illustrate the implications of such constraint; other data centers with different space constraints would lead to other values; however, the general conclusions are still valid regardless the chosen value.

Fig. 3.5 is a replica of the left graph in Fig. 3.2, but including the effect of imposing a space constraint. The vertical lines represent the maximum number of servers considering 1 U, 0.5 U, and 0.25 U devices, and are obtained with Eq. (3.18).

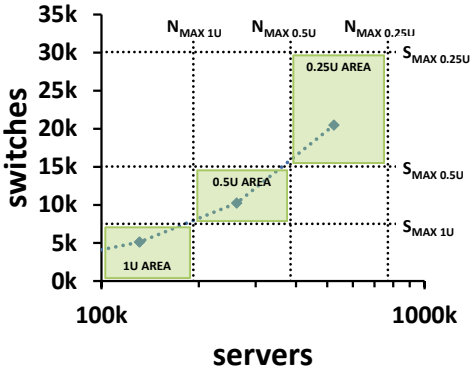

Fig. 3.5 Shrink size of electronic switches.

Analogously, the horizontal lines represent the maximum number of switches considering 1 U, 0.5 U, and 0.25 U devices, and are obtained with Eq. (3.19). Although the network scales up to 524288 servers, this is only feasible with 0.25 U devices; with 0.5 U devices the network scales up to 262144 servers, and with 1 U devices only up to 131072 servers.

Another method to reach the same conclusion is by using Eq. (3.20) to Eq. (3.23). The first two equations compute the number of racks required by servers and switches, assuming racks of capacity V_{RACK} deployed with a single type of device. The last two equations compute the maximum number of racks required for switches and servers. Note that the first two equations depend on the size of the devices V_D, whereas the last two equations do not. These equations may provide non-integer solutions and, in that case, they should be rounded up to the closest integer.

$$R_{SERVERS} = \frac{N_{FT}}{V_{RACK}/V_D} \tag{3.20}$$

$$R_{SWITCHES} = \frac{S_{FT}}{V_{RACK}/V_D} \tag{3.21}$$

$$R_{SERVERS\ MAX} = \frac{N_{FT\ MAX}}{V_{RACK}/V_D} = \frac{k}{2 \cdot l + k - 1} \cdot \frac{V_{MAX}}{V_{RACK}} \tag{3.22}$$

$$R_{SWITCHES\ MAX} = \frac{S_{FT\ MAX}}{V_{RACK}/V_D} = \frac{2 \cdot l - 1}{2 \cdot l + k - 1} \cdot \frac{V_{MAX}}{V_{RACK}} \tag{3.23}$$

These equations are used to generate the curves of Fig. 3.6, corresponding to a three-layer data center built with 128-port switches. The valid solutions with the number of racks needed by switches (left graph) and servers (right graph) are below the horizontal lines; above them, the solutions overtake the space constraint of V_{MAX} = 200 kU. For instance, it is possible to implement data centers with 131072 servers and 5120 switches with 1 U, 0.5 U, and 0.25 U devices. These solutions require 128 / 64 / 32 racks for 1 U / 0.5 U / 0.25 U switches and 3277 / 1639 / 820 racks for 1 U / 0.5 U / 0.25 U servers. If the data center scales up to 262144 servers, then only solutions with 0.5 U and 0.25 U devices exist (requiring 128 / 64 racks for 0.5 U / 0.25 U switches and 3277 / 1639 racks for 0.5 U / 0.25 U servers). Finally, data centers with 524288 servers are only feasible with 0.25 U devices, and require 128 racks for switches and 3277 racks for servers.

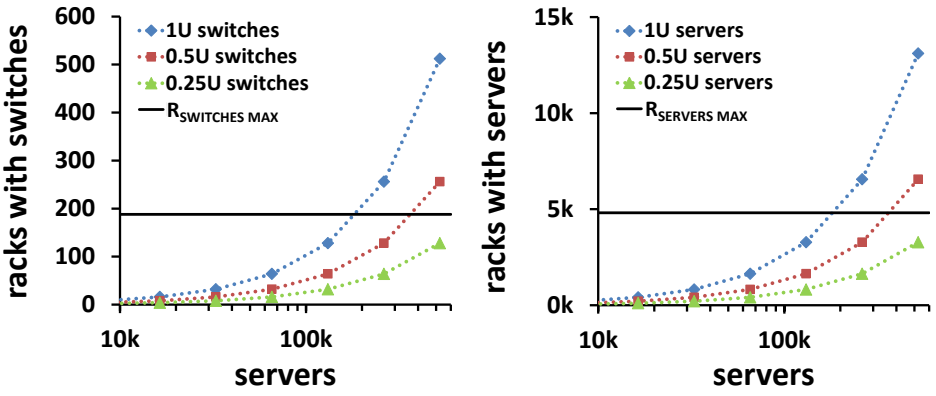

Fig. 3.6 Number of racks with switches and servers.

3.3.4 Principle 4: shrink power consumption of electronic switches

The fourth design principle points out the relevance of reducing the power consumption of electronic switches (and servers). In order to support the ever-increasing demands for cloud computing, more devices with improved features in bandwidth and computational power are interconnected. This implies a growing power consumption in data centers which is already in the order of tens of megawatts in large deployments. For instance, the challenging target for exascale systems is around 20 MW and a Microsoft's data center had an initial power source of up to 40 MW [7].

The large operational costs associated with such values make desirable to set a maximum power constraint to the data center to keep the consumption below a certain value. Let's denominate P_{MAX} to such power constraint and assume that P_{MAX} is distributed following [1]: 70% for servers (CPU + RAM + disks), 15% for cooling, 10% for the power distribution network, and 5% for the switches. Then, the maximum power available for all the switches and the servers of the data center is given by Eq. (3.24) and Eq. (3.25).

$$P_{SERVERS\ MAX} = 0.7 \cdot P_{MAX} \tag{3.24}$$

$$P_{SWITCHES\ MAX} = 0.05 \cdot P_{MAX} \tag{3.25}$$

Consequently, the maximum power available per switch and per server are given by Eq. (3.26) and Eq. (3.27).

$$P_{SERVER\ MAX} = \frac{P_{SERVERS\ MAX}}{N_{FT}} = \frac{0.7 \cdot P_{MAX}}{N_{FT}} \tag{3.26}$$

$$P_{SWITCH\ MAX} = \frac{P_{SWITCHES\ MAX}}{S_{FT}} = \frac{0.05 \cdot P_{MAX}}{S_{FT}} \tag{3.27}$$

Let's select P_{MAX} = 25 MW as an example to illustrate the implications of such constraint. Other data centers with different power constraints would lead to other values; however, the general conclusions are still valid regardless the chosen value. This means that from the 25 MW of our example, a total of 1.25 MW is available for switches and 17.5 MW is available for servers, as shown in the left diagram of Fig. 3.7. As a consequence, the constraint in power implies that consecutive increases in the size of the data center reduce the power available per device, as shown in the right diagram of Fig. 3.7. For instance, with 65536 servers and 2560 switches, each server has 267.0 W available and each switch has 488.3W. By duplicating the size of the network to 131072 servers and 5120 switches, the power available per server reduces to 133.5 W and per switch to 244.1 W. Therefore, it is possible to conclude that in order to keep scaling the network without requiring more power it is mandatory to reduce the power consumption of the devices. Note that, although the 17.5 MW available for all the servers is much larger than the 1.25 MW available for all the switches, the power available per server ends up being smaller than the power available per switch because of the much larger number of servers in the network.

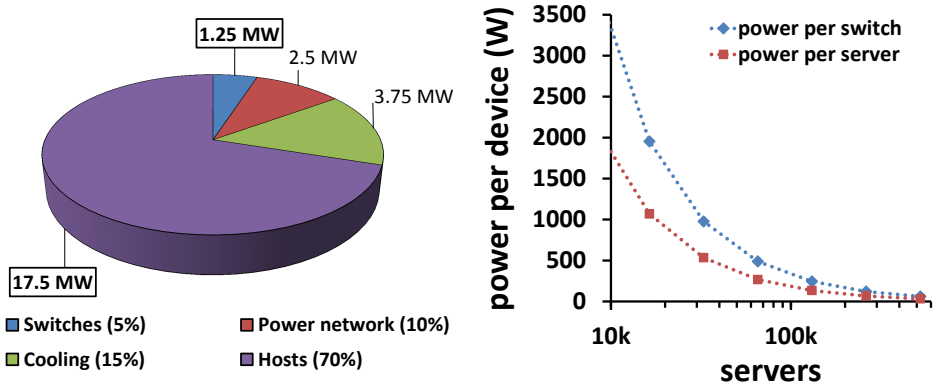

Fig. 3.7 Power distribution in a 25 MW data center and power per device.

Another manner to reach the same conclusion is by computing the total power required by servers and switches with Eq. (3.28) and Eq. (3.29). These equations

include the parameters P_{SERVER} and P_{SWITCH}, which model the power consumption per server and per switch, respectively.

$$P_{SERVERS} = P_{SERVER} \cdot N_{FT} \tag{3.28}$$

$$P_{SWITCHES} = P_{SWITCH} \cdot S_{FT} \tag{3.29}$$

The curves of Fig. 3.8 represent the total power consumption of switches and servers for different values of P_{SWITCH} (100 W, 200 W, 300 W, and 400 W) and P_{SERVER} (50 W, 100 W, 150 W, and 200 W). The horizontal lines are obtained with Eq. (3.24) and Eq. (3.25), and the curves are obtained with Eq. (3.28) and Eq. (3.29). The feasible solutions are below the horizontal lines. For instance. it is possible to implement data centers with 65536 servers and 2560 switches with all the values of P_{SERVER} and P_{SWITCH} under consideration. However, it is only possible to implement data centers with 131072 servers and 5120 switches with 50 W or 100 W servers, and 100 W or 200 W switches. Solutions with 262144 servers and 10240 switches require 50 W servers and 100 W switches. There are no solutions with 524288 servers and 20480 switches with the values of P_{SERVER} and P_{SWITCH} under consideration.

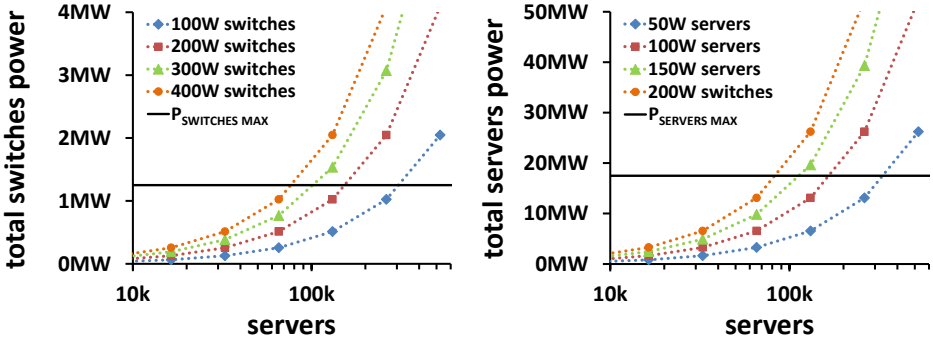

Fig. 3.8 Restriction of solutions due to power consumption constraint.

Finally, note that although both space and power constraints limit the valid solutions, one of them may be the limiting factor. For instance, power is the limiting factor in our examples: 0.25 U devices enable a solution with 524288 servers which is not feasible with 100 W switches and 50 W servers.

3.4 Electronic switch with On-Board Optics prototype

3.4.1 Introduction

We followed the design principles to implement an electronic switch prototype. According to the first design principle, we selected as switch designers an ASIC with the maximum number of ports among the commercially available versions, i.e., 128 ports. At the beginning of the project, the state-of-the-art ASIC provided 10G interfaces and 1.28 Tbps bandwidth. Following the second design principle, we implemented an electronic switch based on a single switching ASIC; we did not attempt to build a larger switch based on multiple switching ASICs. Finally, according to the third and fourth design principles, we designed the electronic switch to be as compact and low-power consumption as possible with technologies commercially available at present. The result of applying the design principles during the design phases was the prototype shown in Fig. 3.9. It has a size of 20 cm by 20 cm requiring only 0.25 U of rack space and consumes below 150 W of power. We packaged four of these switches in a rack unit, effectively multiplying the port and bandwidth density of the rack unit by four while only requiring the double amount of power compared to similar solutions based on front panel transceivers.

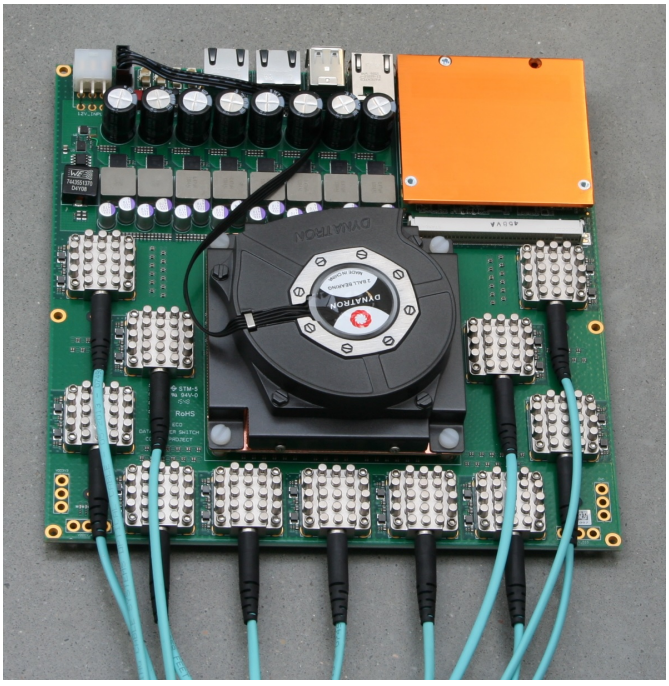

Fig. 3.9 Electronic switch with On-Board Optics.

The bottom section of the board includes the data plane: switching ASIC and transceivers. The switching ASIC, in the center, is covered by its fan in the picture. It is surrounded by the eleven OBO transceivers, placed as close to it as possible to reduce the size of the system and the length of the high-speed traces. Three transceivers are placed on the left, another three transceivers on the right, and the remaining five transceivers are located at the bottom. The top-left section includes the connectors at the edge and most of the power block. The large electrolytic bulk capacitors are placed at the 12 V input of the power block and at the output of every generated voltage (5 V, 3.3 V, and 1 V). The top-right section includes the CPU control board with an orange heat-sink. The space below the CPU control board is used to place the remaining electronics and it is better visualized in Fig. 3.12 when presenting the layer 1 layout.

There are two key design decisions that enable these results: the choice of OBO transceivers and the approach of generating all internal voltages from an external power supply.

OBO transceivers do not suffer the front panel bottleneck because they are not located in the front panel area; instead, they are placed on the PCBs occupying the whole rack unit area. Moreover, the selected version of OBO modules has higher port and bandwidth density compared to the wide-spread QSFP family form factor: it requires approximately half the space and provides three times more ports. Thus, whereas the front panel transceiver approach usually package 32 4-port transceivers in a rack unit, we package 44 12-port transceivers in the same space. In addition, these devices are more power efficient and easier to cool down being distributed in the whole rack unit area.

The generation of all internal voltages from an external 12 V power supply is inspired by OpenCompute [107], suggesting the introduction of 12 V power rails in the racks. The traditional approach requires two power supply units per device: e.g., 272384 power supply units for 131072 servers and 5120 switches. Besides, although integrating the power supply units in the rack unit may generate some of the internal voltages, such as the Advanced Technology Extended (ATX) standard generating 12 V, 5 V, and 3.3 V, they do not completely remove the necessity to generate voltages in the board. In effect, switching ASICs use lower and lower voltages to reduce the power consumption: low voltages such as 1 V and large currents such as 100 A mostly force the generation of these voltages close to the ASIC. In contrast, our approach has the advantage of providing extra flexibility to power distribution engineers to optimize the number, characteristics, and location of the power supply units. More importantly, our approach frees substantial space in the rack unit that may be used to accommodate additional devices, like our packaging prototype integrating four switches in the rack unit.

3.4.2 System overview

The design of an electronic data center switch with 128 10G ports is not a trivial task. The system is complex, integrating above 2000 components and requiring more than 250 differential pairs for high-speed signals. The selection and location of each component must be done carefully in order to reduce size and power as much as possible according to the third and four design principles. Our system is logically divided into the eight blocks summarized in Fig. 3.10. The main design challenges are the power and the switching ASIC blocks.

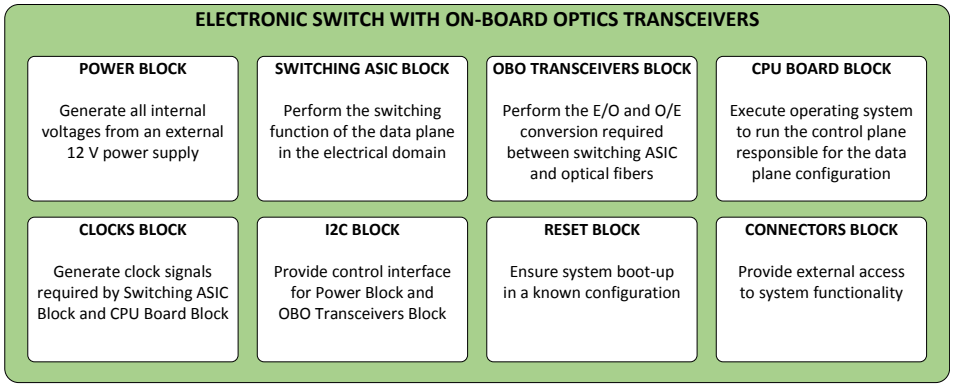

Fig. 3.10 Main blocks of the system.

1) Power Block

The electronic switch requires 12 V, 5 V, 3.3 V, and 1 V to operate: 12 V is used for the fans and to generate the rest of the voltages with a number of buck-converter controllers; 5 V is mainly used by the CPU board; 3.3 V is needed for most of the digital control logic of the electronic switch; 1 V is used by the switching ASIC. The block includes a power sequencer to ensure that the voltages are powered up and down in the right timing sequence and order, according to the switching ASIC specifications. It also integrates the required logic to completely switch off the data plane and leave the control plane running.

2) Switching ASIC Block

The more than 2400 pins of the switching ASIC block are distributed approximately in more than 1500 pins for power signals, more than 500 pins for high-speed signals, and the rest for control interfaces. Such a number of pins implies the need for a large number of layers in the PCB in order to access all the signals (20 layers in our case). Besides, each 10G signal requires a differential pair with 100 Ohms characteristic impedance.

3) On-Board Optics Transceivers Block

The switch integrates eleven 12-port transceivers providing a total of 132 ports, of which 128 ports are used and 4 are wasted. The optical side of each transceiver includes a Multi-fiber Push On (MPO) connector with 24 fibers required for the twelve 10G channels. The electrical side of each transceiver has 24 pins for the high-speed electrical interfaces, power pins, and Inter-Integrated Circuit (I2C) pins for the control and monitoring interface.

4) CPU Board Block

The electronic switch integrates a Mobile PCI Express Module (MXM) connector for a QA3 form factor CPU board. The selected model integrates an Intel Atom E3845 quad-core processor, 2 GB Double Data Rate type 3 (DDR3) DRAM, and 4 GB embedded Multi-Media Controller (eMMC) on-board flash. The Unix operating system runs the software controlling the data plane composed by switching ASIC and transceivers. The interface with the switching ASIC is a Peripheral Component Interconnect Express (PCIe) x2 interface and the interface with the transceivers and the power controllers is an I2C interface. The CPU board includes other interfaces such as a Gigabit Ethernet (GbE) interface for the control plane network, Universal Asynchronous Receiver-Transmitter (UART), and Universal Serial Bus (USB).

5) Clocks Block

This block generates a number of clock signals required by the switching ASIC to operate: three 25 MHz clocks for the digital core, a 100 MHz clock for the PCIe interface, and four 156.25 MHz clocks for the high-speed signals.

6) I2C Block

The CPU board uses the I2C interface to monitor and control the transceivers and the power controllers. Since all the transceivers have the same I2C address interface, this block includes an I2C hub and two I2C switches to interface each transceiver without collision. This block also includes other I2C devices such an Analog-to-Digital Converter (ADC) block to measure the voltages on the board, an Electrically Erasable Programmable Read-Only Memory (EEPROM) to store identification information of the switch, and a General Purpose Input-Output (GPIO) block to generate internal control signals.

7) Reset Block

This block generates the reset signals for the CPU, the switching ASIC, and the I2C block during power-up, hardware, and software reset events to ensure that the system operates in a known state.

8) Connectors block
This block integrates the connectors required to interface with the electronic switch: e.g. the 12 V input connector, the RJ45 connector, the USB connector, the case fans and the switching ASIC fan connectors.

3.4.3 Printed Circuit Board design

The PCB is 20 cm by 20 cm size, with 20 copper layers organized as shown in the layer stack-up of Fig. 3.11. The layers are distributed among low-speed signals, high-speed signals, and/or power planes.

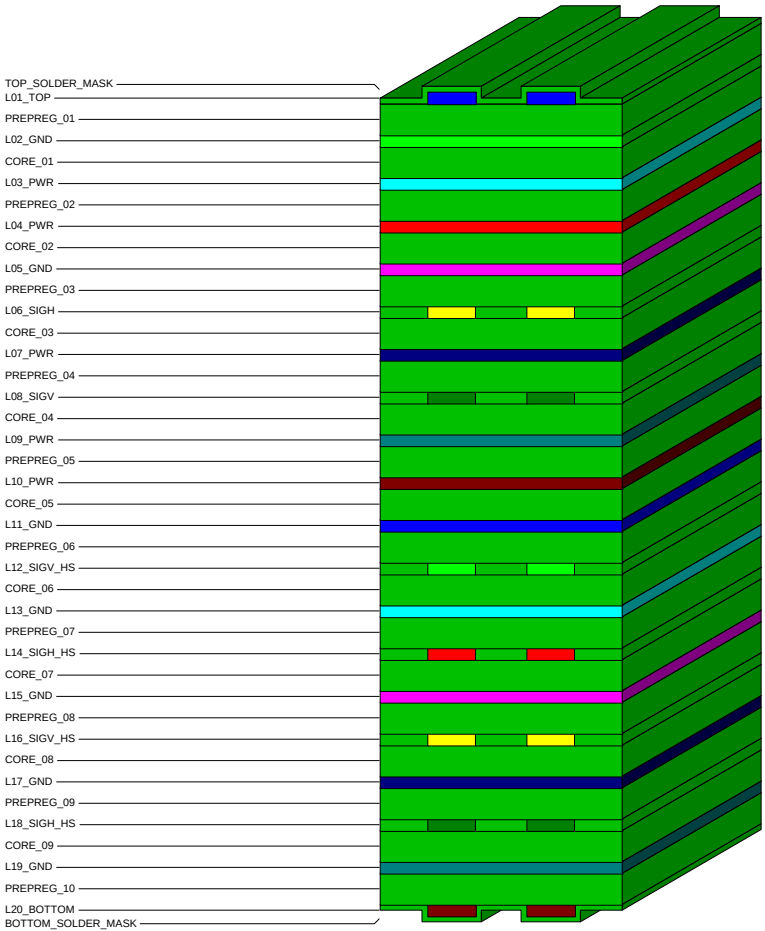

Fig. 3.11 PCB layer stack-up diagram.

The top and bottom layers (L1 = layer 1 and L20) contain the components: while most of them are on the top layer, a number of power decoupling capacitors (required by the switching ASIC) are placed on the bottom layer. L3 and L8 are used for low-speed signals such as I2C. L6, L12, L14, L16, and L18 are used for high-speed signals, such as clocks, PCIe, and 10G interfaces. L2, L5, L11, L13, L15, L17, and L19 are used for ground planes, which apart from providing the required reference planes for the components also isolate adjacent layers. Finally, L3, L4, L7, L9, and L10 are used for the power distribution planes.

The layout of three of the twenty layers is shown as an example: L1 with the components in Fig. 3.12, L4 with a number of the power planes, and L12 with a number of the high-speed 10G lines.

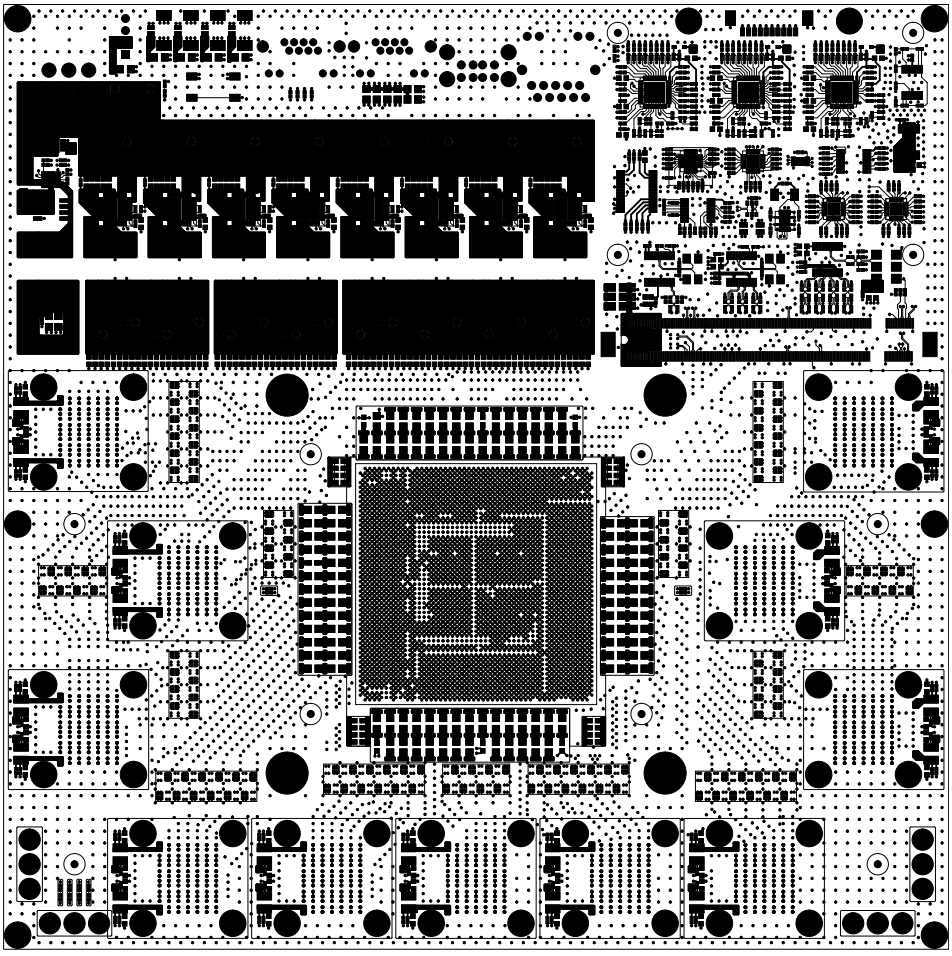

Fig. 3.12 PCB layer 1 routing.

The layout of L4, an example of a copper layer used for power distribution planes, is shown in Fig. 3.13. The power plane including the top-left corner has 12 V; the power plane including the top-right corner has 5 V; the largest power plane including both bottom corners has 1 V; finally, the small power planes in the center area power different blocks of the switching ASIC after additionally filtering the 1 V plane.

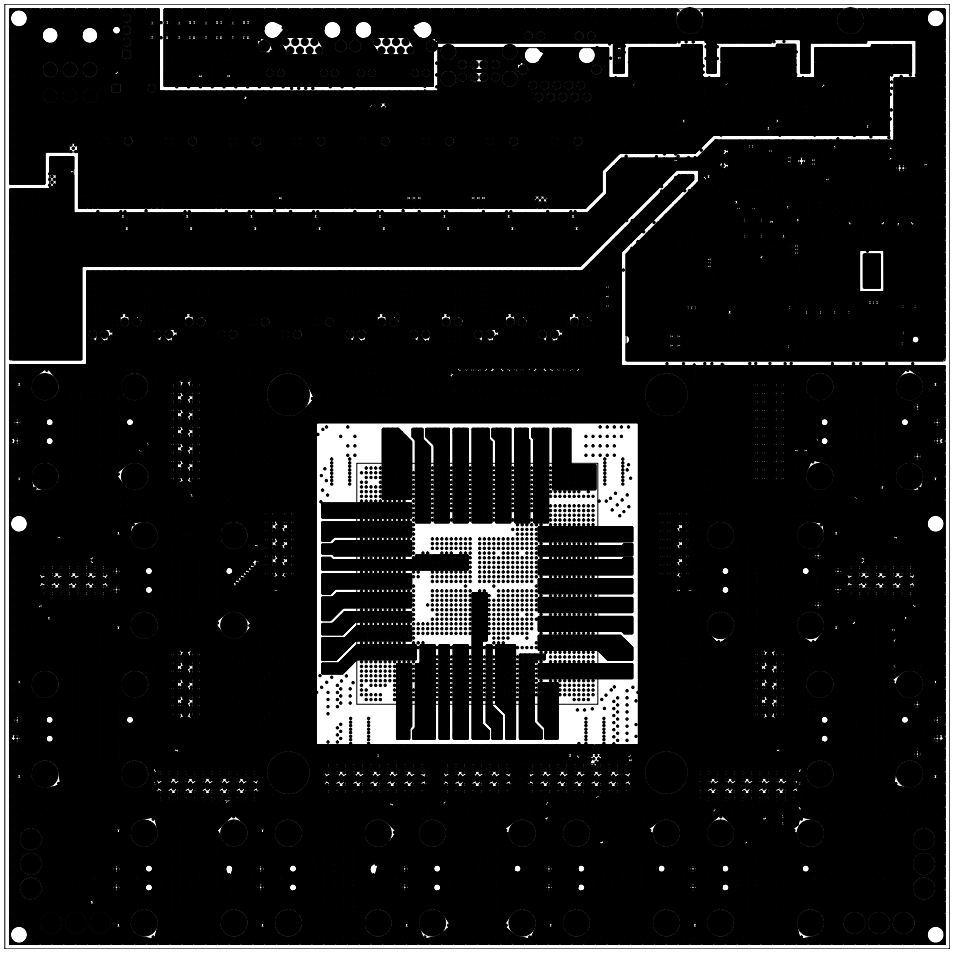

Fig. 3.13 PCB layer 4 routing.

The layout of L12, one of the four layers used for the 10G high-speed differential pairs, is shown in Fig. 3.14. It contains 64 of the 256 differential pairs required to route the 128 ports of the electronic switch. Each differential pair has a characteristic impedance of 100 Ohm and connects the corresponding port between ASIC and transceiver.

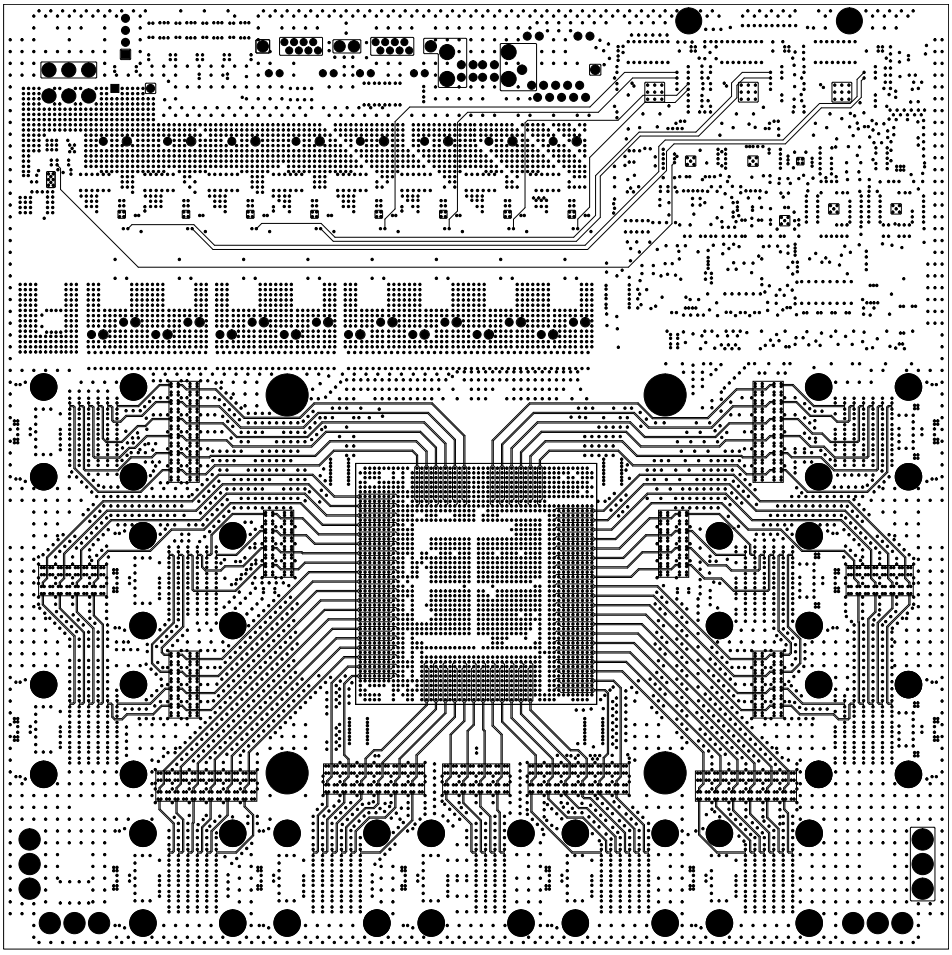

Fig. 3.14 PCB layer 12 routing.

3.4.4 Rack unit packaging

Each one of the compact boards requires only half of the rack width in a 19-inch rack unit. As a result, two boards are easily accommodated side-by-side in the rack unit assuming the redundant power supply is maintained. On top of that, the approach of generating all internal voltages from a 12 V power supply input allows removing the power supplies from the rack unit and fully populate the space available with four boards. The picture of our packaging prototype is shown in Fig. 3.15. The front panel, free from transceivers, accommodates three 2x8 MPO connectors arrays, resulting in a total of 48 slots to connect the 44 MPO connectors of our switches. While this design provides four times the bandwidth (4 x 1.28 Tbps) and port density (4 x 128 ports) compared to a switch implementation using

QSFP+ modules, there is still enough room in the front panel for air openings. This, together with the additional space in the back panel for extra fans improves the heat removal from the case.

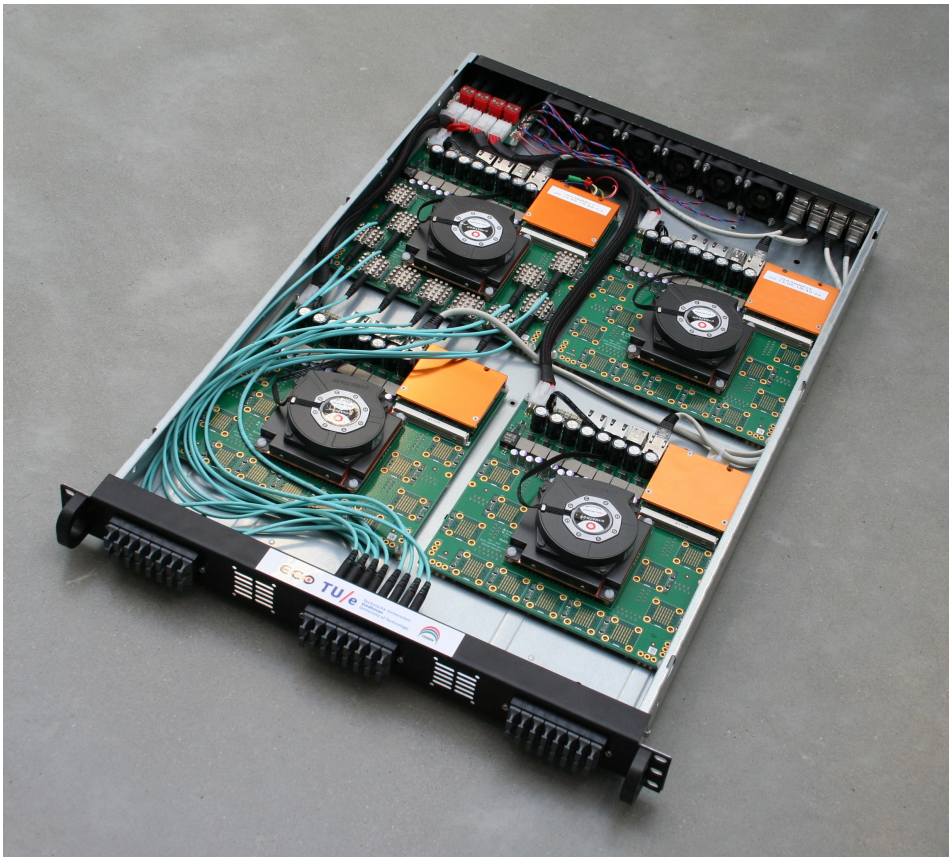

Fig. 3.15 Packaging of four compact switches in a single rack unit.

There are other differences compared with the traditional packaging of electronic data center switches previously shown in Fig. 2.1. One difference is that the rack unit includes a small power distribution board that multiplexes a single 12 V input into four 12 V outputs that power each one of the four boards. This points out the flexibility provided by decoupling the power supply units from the rack unit: in our case, a single (more powerful) unit is used to power all the boards. Another difference is that the rack unit includes four RJ45 connectors for the control plane network, one for every switch. Finally, the six fans of the case are shared among the four switches, with advantages in power consumption and cost.

The validity of this approach is prepared for future developments given two advances in technology already happening: first, higher port and bandwidth density transceivers and switching ASICs will be designed in the future; and second, tighter integration of these technologies is possible, e.g., with transceivers placed on top of the ASIC or co-integrated with the ASIC. Thus, it is possible to imagine a rack unit integrating a larger number of even more compact devices. The rack unit front panel can accommodate five of these MPO arrays, with a total of 80 MPO-24 connectors providing access for up to 960 ports. In case that the integration goes even further, up to 80 MPO-72 connectors could be placed in the front panel area providing access up to 2880 ports. With such a number of ports, the optical bandwidth of the rack unit could scale up to 28.8 Tbps with 10G ports, 72 Tbps with 25G ports, and beyond.

3.4.5 Prototype characterization

The prototype is characterized in terms of temperature and power consumption in order to experimentally validate the feasibility of the approach. By integrating four switches in a single rack unit, thermal and power considerations should be taken into careful consideration. Thermal management becomes more critical, especially because the performance and reliability of passively cooled optical transceivers are highly dependent on temperature. Similarly, power management becomes also more critical, since the rack unit includes multiple devices instead of only one.

A number of modifications are introduced in the thermal management system of the packaging experiment shown in Fig. 3.15 which result in an operating temperature of the transceivers around 40° C. The main modification consists of removing the power supplies from the rack case, getting rid of all the heat generated by them. In addition, without the power supplies, it is possible to install more fans in the back panel for extra forced cooling, like in our example with six fans. All this, combined with the fact that also the front panel has some free area for extra ventilation being not fully populated with fiber array connectors, explains the sufficient airflow and heat removal of the rack unit. Finally, another important advantage is that OBO modules are distributed all around the rack case, and not concentrated in the front panel, so they receive additional benefit from the improved thermal system of the case.

The thermal simulation results are shown in Fig. 3.16. Temperature and air velocity of the case are included in the simulation, that predicts a maximum temperature of 35.5° C.

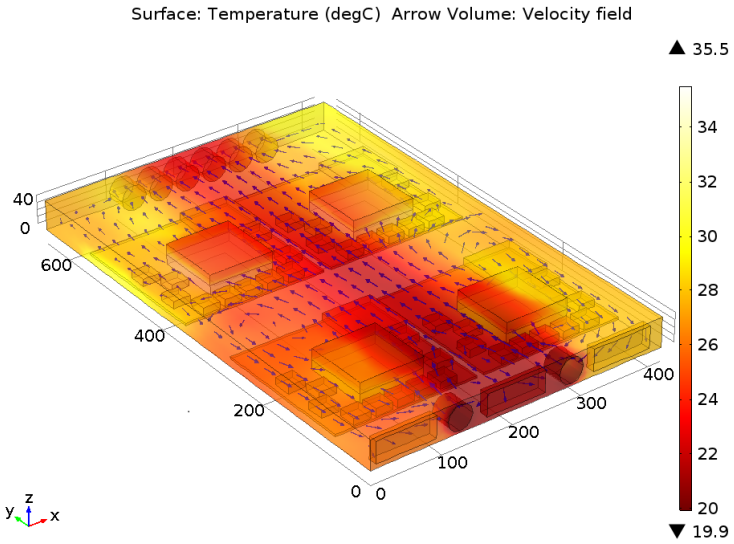

Fig. 3.16 Simulation of temperature and airflow of the rack unit.

The experimental results are obtained by direct reading the temperature sensors included in the transceivers of one of the boards and are summarized in Fig. 3.17. A set of experiments with two, four, and six fans configurations is carried out, concluding that every extra pair of fans reduces the thermal stress in the transceivers, not only by reducing their maximum operating temperature but also by reducing the temperature range variation.

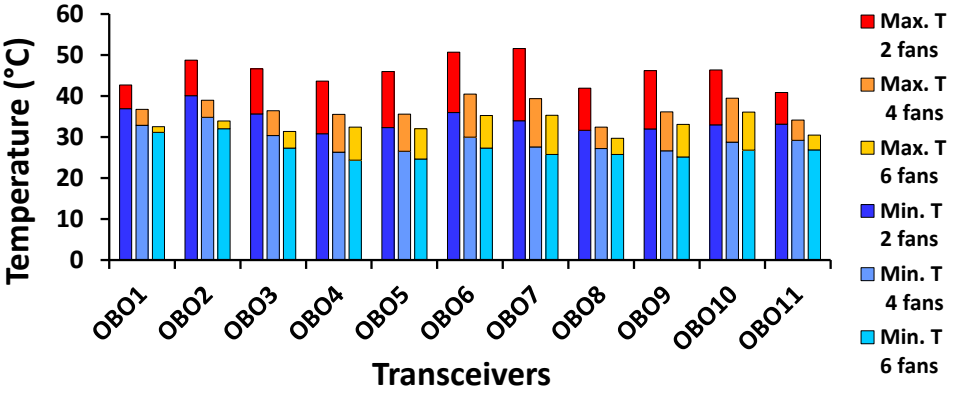

Fig. 3.17 Measured temperature of the On-Board Optics transceivers.

For instance, transceiver six has a maximum temperature of 50.7° C / 40.5° C / 35.2° C, a minimum temperature of 36° C / 30° C / 27.3° C with 2 / 4 / 6

fans, resulting in operating ranges of 14.7° C / 10.5° C / 7.9° C respectively. The measured maximum operating temperature of the transceivers is 51.6° C with 2 fans, 40.5° C with 4 fans, and 36° C with 6 fans. These values are slightly higher than the results obtained with the thermal simulation due to model approximations.

Regarding power consumption, the prototype switch shown in Fig. 3.9 is designed for a maximum power consumption of 150 W distributed approximately as follows: 100 W for the switching ASIC, 40 W for the OBO transceivers, and 10 W for the CPU control board.

The measured results for the power required by a single switch are shown in Fig. 3.18. The minimum power consumption, around 7 W, is achieved by turning off the data plane (ASIC and optical transceivers), and leaving the CPU control board on (*phase 1*). Potentially, since the switches are SDN-enabled, it is possible to power down a part of the system to greatly reduce the power consumption of the network. Once the ASIC and transceivers are switched on, there is a sharp rise in power consumption that stabilizes after a period (*phases 2* and *3*, respectively). Then, a set of tests transmitting packets with the ports loop-backed are executed (*phases 4, 5, 6* and *7*). With 24/48/72/96 ports enabled the power consumption increases to 83.3 W/89.7 W/97.9 W/105.6 W, respectively. The maximum load test with only 96 ports is due to limitations imposed by the normal operation mode of the ASIC. The gap between the maximum designed power consumption and the experimentally measured can be associated with the loading of the CPU, the setting of the ASIC in over-subscription mode, and the transceivers. Thus, we conclude that our OBO-based design can operate at a power consumption far below 150 W which is about half power consumption than a standard 1 U switch.

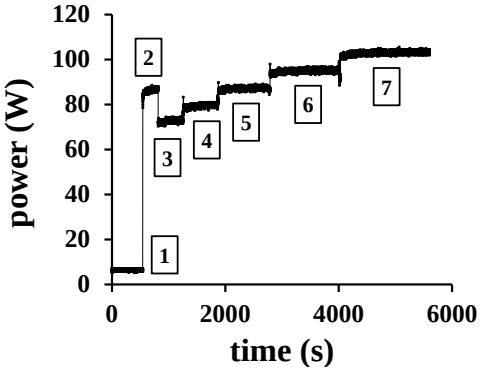

Fig. 3.18 Power consumption per switch including optical transceivers.

3.5 Summary

In this chapter, we have presented our analytic model of FT, which extends the traditional model by including two additional parameters and two extra equations. The f_P parameter allows partitioning the network in a discrete manner without wasting resources and ensuring even distribution of links; the $f_{E\ NO\ WDM}$ models the number of ports in the transceivers; the extra equations allow computing the number of transceivers and fibers in the network.

The analytic model of FT is used to infer four design principles for electronic data center switches in order to improve the scaling of data center networks. The first design principle points out the relevance of scaling-up the number of ports in the switching ASICs in order to reduce the number of switches in the network. The second design principle remarks that it is more efficient to build the network scaling-out the number of commodity switches integrating a single switching ASIC than using large port-count electronic switches implemented with multiple switching ASICs. Finally, the third and four design principles signal the importance of shrinking the size and power consumption of electronic switches in order to further scale data centers constrained in space and power.

An electronic switch prototype has been designed and implemented following these design principles. Commercially available technologies at present lead us to the integration of On-Board Optics transceivers, with more efficient behavior in terms of power consumption and temperature. The result is a compact prototype, that with a size of only 20 cm by 20 cm is probably one of the 128-ports switch most compact ever implemented. It overcomes the front panel bottleneck by the integration of OBO modules, and the ASIC bottleneck by allowing the packaging of four devices per rack unit owing to a power consumption below 150 W and the single voltage power supply. As a result, the port and bandwidth density of the rack unit is multiplied by four and only requiring about the double amount of power. The prototype is characterized in terms of temperature and power consumption, validating empirically the feasibility of the approach.

From all the above, we conclude that compact electronic switches integrating On-Board Optics transceivers improve the scaling of data center networks, and take the integration of electronic switching ASICs and optical transceivers a step further.

Chapter 4

Analytic Model of Electronic and Hybrid Data Center Networks

4.1 Introduction

Data center networks are based at present mostly on ESs. Chapter 3 has introduced a set of general design rules for ESs, the main building block of these networks. These design rules have been applied to the implementation of a low-power compact electronic switch. Our prototype has demonstrated the feasibility of overcoming the front-panel bottleneck and the ASIC bottleneck, which are limiting the scaling of these devices. The key design decisions are the substitution of front-panel optical transceivers by another type of transceivers allowing tighter integration with the switching ASIC, such as On-Board Optics transceivers, and the packaging of multiple switches per rack unit. As a result, our prototype scales by a factor four the rack unit port and bandwidth density, requiring half power consumption. At data center level, and assuming similar achievements with the servers, these results imply that the double number of servers can be connected requiring the same power consumption and half the space than solutions deployed at present.

The topology selected as the driving thread for the analysis of Chapter 3 has been Fat-Tree. This topology is a well-known and widespread architecture for building data center networks [14, 56, 87, 88, 91–93]. Among its features, two are especially relevant for our purposes of building large-scale high-performance data center networks. First, it scales to any number of servers by interconnecting switches of any number of ports (just by adding layers of switches). Second, it provides full bisection bandwidth among the servers, i.e. a partition with half

of the servers can communicate with the other half partition at full link speed [108]. However, Fat-Tree requires a large number of switches, transceivers, and fibers, which implies large power consumption and cost. For instance, a three-layer Fat-Tree topology with 131072 servers requires 5120 128-port switches, 163840 4-port transceivers, and 786432 fibers.

Such numbers have motivated research efforts to improve the scaling that may be classified into two main categories. One direction suggests topologies based on ESs [82–86, 89, 94]. However, they are usually limited in scaling by the number of ports of the switches or rely on over-subscription to increase the number of servers. The other direction suggests the introduction of OSs, thus, building hybrid data center networks [104, 105]. We explore this approach in Chapter 4 and Chapter 5 because OSs are promising technologies overcoming important limitations of ESs. OSs perform the switching function in the optical domain, consequently not requiring optical transceivers for the E/O and O/E conversion. This implies savings in devices, power consumption, and cost. In addition, OSs are transparent, both bit-rate and data-format agnostic. This is especially suitable for the introduction of WDM technologies, which are interesting to reduce the number of fibers and further exploit the capacity available in optical fibers.

We focus our attention on hybrid networks able to provide full bisection bandwidth and not limited in scaling by the number of ports of the switches, such as Helios [96] and HyPaC [97]. Helios is a Fat-Tree network where the core layer can be totally or partially substituted by OSs. As a result, all or some of the inter-cluster communication is carried through OSs. HyPaC splits the bandwidth at the rack level, so part of the intra-cluster and inter-cluster communication is carried through OSs.

An important limitation of these hybrid networks is that they do not include a complete set of analytic formulas that links and describes them. This makes it almost impossible to properly investigate the scaling of these networks in terms of devices (i.e. servers, switches, transceivers, and fibers), power consumption and cost. Moreover, a realistic comparison among different electronic and hybrid topologies is difficult.

In this chapter, we present a novel general analytic model describing electronic and hybrid Fat-Tree like topologies, not limited in scaling by the number of ports in the switches and providing full bisection bandwidth. The model includes a set of equations to accurately compute the number of servers, switches, transceivers, and fibers of the architectures. It also introduces a set of configuration parameters modeling relevant aspects of the architectures, which provide great flexibility for the design of different networks. For instance, the model includes configuration parameters to adjust the size of the network, to characterize the fraction of ESs

replaced by OSs, and to select among different types of transceivers depending on the desire to include WDM technologies on electronic and/or hybrid topologies. As a result, topologies such as Fat-Tree, Helios, and HyPaC are covered by this model.

Our analytic model is used in Chapter 5 to investigate the scaling of electronic and hybrid topologies in terms of devices, power consumption, and cost. It answers questions such as how many devices (i.e. electronic and optical switches, transceivers, and fibers) are required to build a network of a certain size?, how much power consumption and cost are required for electronic and hybrid data center networks? what devices dominate power consumption and cost of such networks?, are WDM technologies interesting for electronic and/or hybrid networks given the extra cost of the required transceivers?, what is the impact on the scaling in terms of devices, power consumption, and cost of increasing the number of wavelengths in networks using WDM technologies?

The remainder of this chapter is organized as follows. Section 4.2 briefly introduces the model parameters defining the four topologies. Section 4.3 explores in more detail the impact of each parameter. Section 4.4 introduces a set of equations to calculate the number of servers, switches, transceivers, and fibers for each topology. Section 4.5 presents a set of validity conditions for the combination of model parameters. Section 4.6 summarizes a number of relations among the equations. Section 4.7 visualizes a number of full examples. Finally, Section 4.8 concludes the chapter.

4.2 Model parameters

Examples of the four topologies included in our general model are shown in Fig. 4.1. We name the ES-based topologies FT and EFT, and the corresponding hybrid versions, HFT and EHFT. The examples are two-layer networks implemented with 128-port switches and connecting 2048 servers. FT and EFT are topologies based only on ESs. The difference between them is that EFT employs WDM in the core layer, whereas FT does not. HFT and EHFT are hybrid topologies including a combination of ESs and OSs. Both of them use WDM in the links going through OSs. The distinction between them is that EHFT brings WDM into play also between ESs in the core layer, in a similar manner as EFT does with FT.

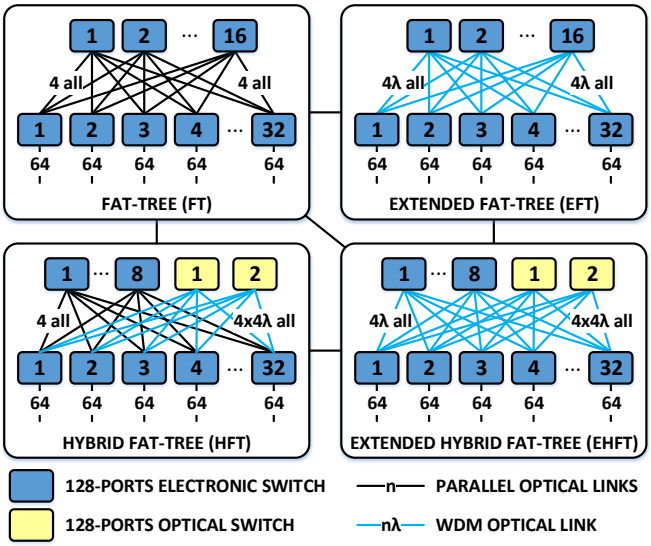

Fig. 4.1 Examples of FT, EFT, HFT, and EHFT topologies.

The parameters describing the four topologies are summarized in Table 4.1. The only topology requiring all the parameters is EHFT, the more general network of this model. The impact of each parameter is introduced briefly in the following lines and is further explored in Section 4.3.

The number of ports per switch (k), the number of layers of the network (l), and the partition factor (f_P) are common parameters to all topologies and they have been already introduced in Section 3.2.

The number of hybrid layers (l_H) and the optical factor (f_O) are only required to define the hybrid topologies, HFT and EHFT. l_H accounts for the number of layers substituting a number of ESs by OSs. The f_O factor describes the fraction of this hybrid layer(s) where ESs are replaced by OSs. For instance, in the hybrid examples previously shown in Fig. 4.1, half of the upper layer of ESs is substituted by OSs ($l_H = 1$, $f_O = 1/2$). Similarly, in the EHFT example of Fig. 4.2, half of the two upper layers of ESs is replaced by OSs ($l_H = 2$, $f_O = 1/2$). In case that $f_O = 1$, then the whole layer(s) is implemented with OSs.

Finally, the $f_{E\ NO\ WDM}$, $f_{E\ WDM}$, and $f_{O\ WDM}$ factors model the different types of transceivers in the networks: $f_{E\ NO\ WDM}$ for NO WDM optical transceivers connecting ESs, $f_{E\ WDM}$ for WDM transceivers connecting ESs, and $f_{O\ WDM}$ for WDM transceivers connecting ESs to OSs. EHFT has all three types of transceivers; EFT and HFT have two types of transceivers (modeled by $f_{E\ NO\ WDM}$ and by $f_{E\ WDM}$ or $f_{O\ WDM}$ respectively); FT has only one type of transceiver (represented by $f_{E\ NO\ WDM}$).

Table 4.1 Model parameters defining FT, EFT, HFT, and EHFT topologies.

Symbol	Description
$k \in [2,\infty) \in \mathbb{N}$	number of ports per switch (k is power of 2)
$l \in [2,\infty) \in \mathbb{N}$	number of layers
$l_H \in [1,l) \in \mathbb{N}$	number of hybrid layers
$f_P = 1/2^n$ *with* $n \in [0,\log_2 k/2] \in \mathbb{N}$	fraction of the full network implemented
$f_O = 1/2^n$ *with* $n \in [0,\log_2 k/2] \in \mathbb{N}$	fraction of the hybrid layers populated with optical switches
$f_{E\ NO\ WDM} = 1/2^n$ *with* $n \in [0,\log_2 k/2] \in \mathbb{N}$	reciprocal of the number of ports of NO WDM transceivers connecting electronic switches
$f_{E\ WDM} = 1/2^n$ *with* $n \in [0,\log_2 k/2] \in \mathbb{N}$	reciprocal of the number of ports of WDM transceivers connecting electronic switches
$f_{O\ WDM} = 1/2^n$ *with* $n \in [0,\log_2 k/2] \in \mathbb{N}$	reciprocal of the number of ports of WDM transceivers connecting elect. and optical switches

These factors are defined as the reciprocal of the number of ports of the transceivers. The limit values of these factors are 1, representing one port or wavelength, and $(k/2)^{-1}$, representing $k/2$ ports or wavelengths. For instance, by setting $f_{E\ NO\ WDM}$ = $f_{E\ WDM}$ = $f_{O\ WDM}$ = 1/4 in the examples of Fig. 4.1 all transceivers in the switches have four ports. Similarly, by setting $f_{E\ NO\ WDM}$ = $f_{E\ WDM}$ = $f_{O\ WDM}$ = 1/2 in the example of Fig. 4.2, all transceivers have two ports.

The parameters reported in Table 4.1 are visualized in Fig. 4.2 with a three-layer EHFT example. The network implemented with k = 8-port switches has l = 3 layers, from which l_H = 2 are hybrid. Only f_P = 1/2 of the full network, which would have 128 servers, is implemented. As a result, the network connects 64 servers. The optical factor is f_O = 1/2, and therefore, half of the hybrid layers substitutes ESs by OSs. All transceivers in the ESs have two ports ($f_{E\ NO\ WDM}$ = $f_{E\ WDM}$ = $f_{O\ WDM}$ = 1/2). The port distribution of the ESs is shown in detail on the left side diagrams, layer by layer. The edge switches integrate two types of transceivers: NO WDM transceivers connected to servers and aggregation switches, and WDM transceivers connected to OSs. The aggregation switches have also two types of transceivers, but the WDM transceivers are connected to the core ESs.

Finally, the core ESs have only WDM transceivers facing down to the aggregation switches.

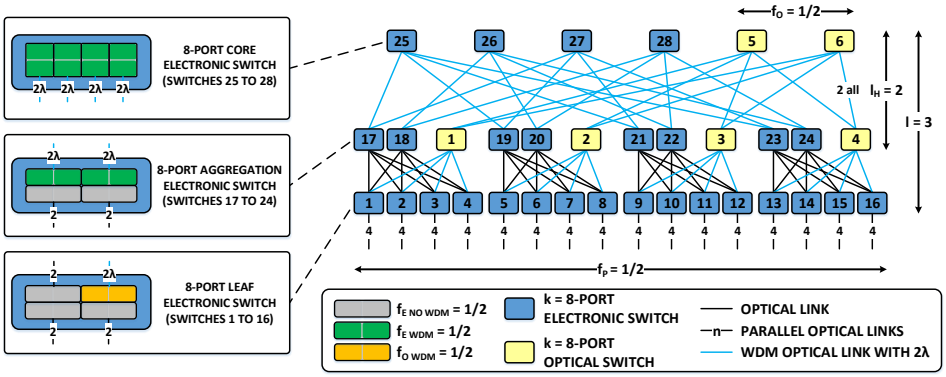

Fig. 4.2 Example of EHFT topology, visualizing all model parameters.

4.3 Impact of model parameters

4.3.1 Impact of partition factor f_P

The partition factor, f_P, allows adjusting the size of the networks to accommodate the desired number of servers. Without this factor, the networks scale abruptly with an increasing number of layers: e.g. a network built with 128-port switches scales from 8192 servers with two layers to 524288 servers with three layers.

This adjustment is possible only in a discrete manner to ensure two important aspects: a partition of the network uses all the port in the switches to avoid wasting resources, and the homogeneity of the network is maintained to avoid uneven distribution of links that difficult routing and deployment. In order to illustrate these aspects, three examples are shown in Fig. 4.3: a full network in the left diagram, a valid partition of the network in the center diagram, and an invalid partition of the network in the right diagram. The full two-layer FT is implemented with 8-port switches and connects 32 servers.

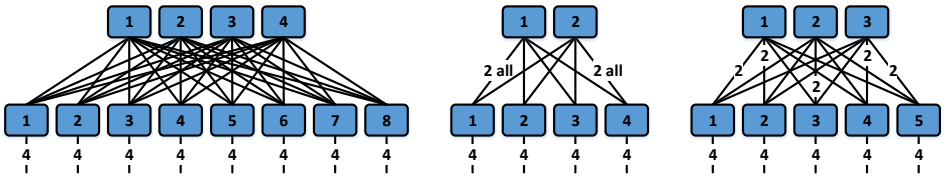

Fig. 4.3 Full two-layer network, valid partition, and invalid partition.

The valid partition has half switches and servers; all the ports in the switches are used and all the links are evenly distributed. The invalid partition has 20 servers; all ports in the edge switches are used but some ports in the core switches are not: e.g. core switches 1 and 3 waste one port, and core switch 2 wastes two ports; the distribution of the links is uneven.

In our model, the full network is represented by $f_P = 1$ and the smallest network size by $f_P = (k/2)^{-1}$. The effect of applying three different values of f_P in a HFT network is shown as example in Fig. 4.4, with parameters $l = 2$, $k = 128$, $f_O = 1$, and $f_{O\,WDM} = 1$. The left diagram is the full HFT with 8192 servers; the center diagram reduces the size by 8 with $f_P = 1/8$ and 1024 servers; the right diagram shows the smallest topology with $f_P = 1/64$ and 128 servers.

An important observation is that the width of the links in the core layer changes when applying f_P in order to maintain the constant bisection bandwidth. In this example, each edge electronic switch is connected to the optical core switch/es by 1 link $f_P = 1$, with 8 links with $f_P = 1/8$, and by 64 links with $f_P = 1/64$.

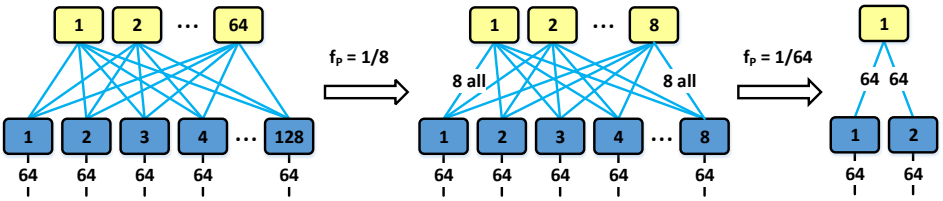

Fig. 4.4 Impact of f_P on a two-layer network.

The effect of f_P in a three-layer network built with 128-port switches is explored in Fig. 4.5.

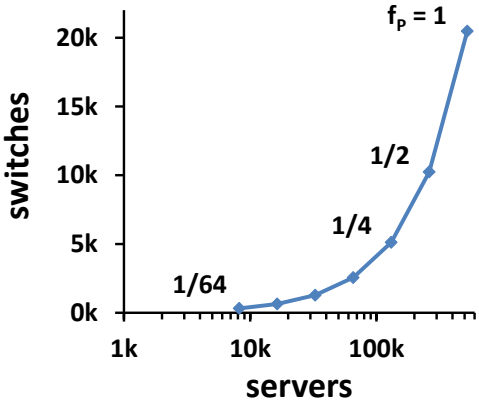

Fig. 4.5 Impact of f_P on a three-layer network.

The highest point in the curve corresponds to a full network with 524288 servers and 20480 switches; the second highest point represents a half of the network with half switches and servers; the lowest point corresponds to the minimum network size with 8192 servers and 320 switches.

4.3.2 Impact of number of hybrid layers l_H

The number of hybrid layers, l_H, is a parameter exclusive of hybrid topologies (i.e. HFT and EHFT). It defines the number of layers where ESs are replaced by OSs. In two-layer hybrid networks, there is by definition just one hybrid layer. In three-layer networks, this parameter may be set to be one or two. Examples of three-layer hybrid networks with one and two hybrid layers are shown in Fig. 4.6 (parameters $k = 128$, $f_P = 1$, and $f_O = 1$).

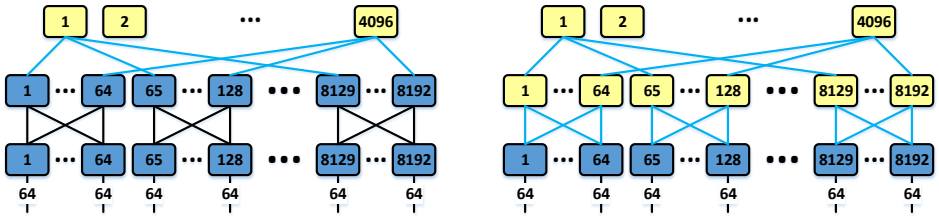

Fig. 4.6 Impact of l_H on three-layer networks.

4.3.3 Impact of optical factor f_O

The optical factor, f_O, is also a parameter exclusive of hybrid topologies. It allows adjusting the fraction of the l_H hybrid layers implemented with OSs. With $f_O = 1$ the whole layer is populated with OSs. Changes in f_O do not modify the total number of switches; it only changes the fraction of these switches that are OSs. As a result, the fraction of the l_H hybrid layers defined by f_O carries the communication through OSs. In two layers hybrid networks, with $l_H = 1$ by definition, at least part of the inter-rack communication is carried through OSs. In three layers hybrid topologies with $l_H = 1$ layer, at least a fraction of the inter-cluster communication is carried through OSs. In three layers hybrid topologies with $l_H = 2$ hybrid layers, at least a fraction of the intra-cluster and inter-cluster communication is carried through OSs.

Examples of HFT for three different values of f_O for a two-layer network built with 128-port switches are shown in Fig. 4.7. The diagram on the left has a fully

optical core layer where all the 64 switches are OSs; the diagram on the center represents the case where f_O = 1/8 and only 8 switches are OSs; the diagram on the right represents the limit case, where f_O = 1/64 and only one OS in the whole hybrid layer.

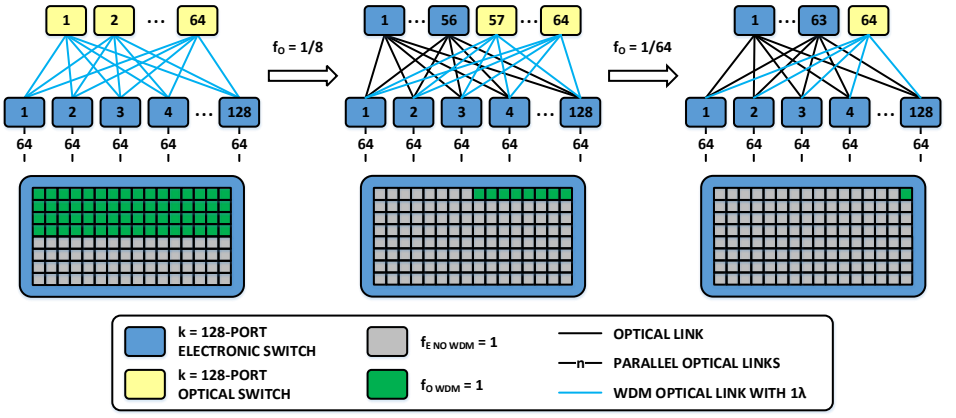

Fig. 4.7 Impact of f_O on a two-layer network.

Note that the total number of switches is equal to 192 in the three examples and equivalent to the number of switches in electronic topologies. The port distribution of the edge switches is represented in the lowest section of Fig. 4.7. In these examples, half of the ports are connected to the servers and the other half is connected to switches. From this half of the ports connected to switches, 64 ports are connected through OSs in the first case, 8 ports in the second case, and only 1 port in the last case.

The effect of f_O in three-layer hybrid networks implemented with 128-port switches is explored in Fig. 4.8. The graph on the left corresponds to networks with one hybrid layer and the graph on the right to networks with two hybrid layers.

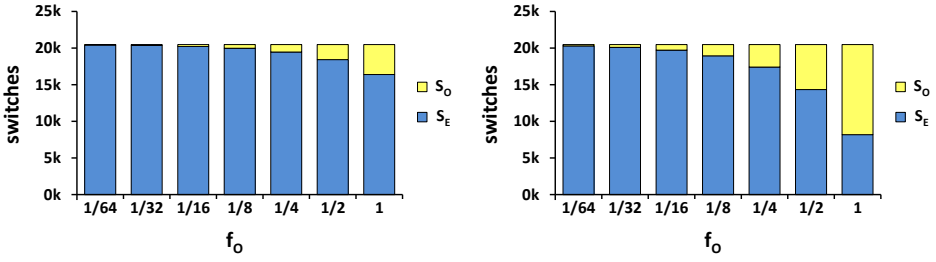

Fig. 4.8 Impact of f_O on three-layer networks with one and two hybrid layers.

The total number of switches remains constant in both cases at 20480 switches. This number is equal to the results obtained for electronic topologies, but in the case of hybrid architectures is distributed between ESs and OSs. f_O adjusts the fraction of these switches that are optical. f_O = 1 yields the maximum number of OSs since all the switches in the hybrid layer/s are optical (4096 and 12288 OSs, respectively).

4.3.4 Impact of $f_{E\ NO\ WDM}$, $f_{E\ WDM}$, and $f_{O\ WDM}$ factors

These factors model the different types of transceivers in the networks. The examples previously shown in Fig. 4.1 are reproduced again in the top section of Fig. 4.9, which includes the port distribution of the edge and core electronic switches in the bottom section. The 128-port switches have 32 4-port transceivers because in these examples the parameters are fixed to $f_{E\ NO\ WDM} = f_{E\ WDM} = f_{O\ WDM}$ = 1/4. However, every topology includes different type of transceivers.

FT has only NO WDM optical transceivers represented by gray boxes. EFT includes WDM transceivers at the core, and requires two types of transceivers represented by gray and green boxes: the edge switches have half of the ports connected with NO WDM transceivers to the servers and the other half with WDM transceivers connected to the core switches; the core switches have all WDM transceivers. HFT has two types of transceivers, represented by gray and orange boxes, and distributed as shown in the diagrams. Finally, EHFT has all three type of transceivers.

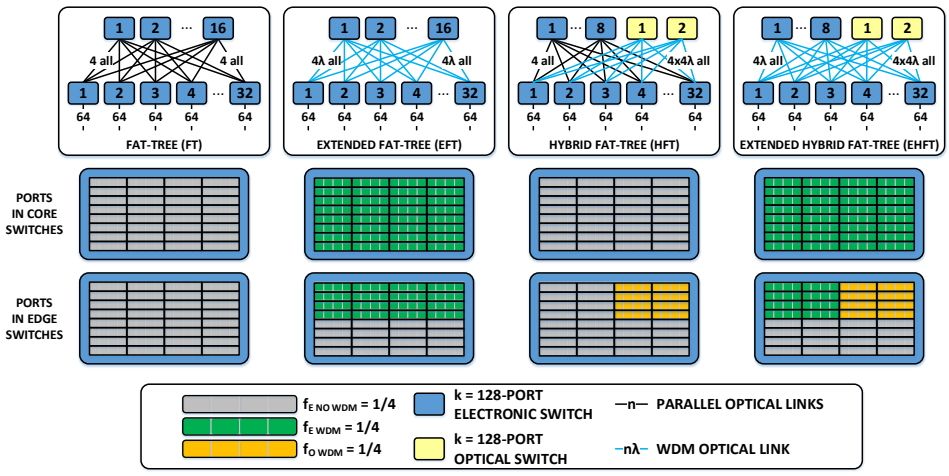

Fig. 4.9 Types of transceivers per topology.

The next two figures illustrate the impact of the $f_{O\,WDM}$ factor: it reduces the total number of switches by reducing the number of OSs required. Indeed, fewer OSs are required when more wavelengths are available (assuming that every optical port is able to switch the corresponding number of wavelengths). This feature is illustrated in Fig. 4.10 with three HFT topologies with one wavelength (left), 8 wavelengths (center), and 64 wavelengths (right). The diagrams at the bottom of the figure represent the port distribution of the edge switches when the WDM transceivers have 1, 8, and 64 ports respectively. In this example, the total number of switches reduces from 192 switches with one wavelength to 129 switches with 64 wavelengths thanks to the reduction of OSs; the number of ESs remains constant at 128 switches.

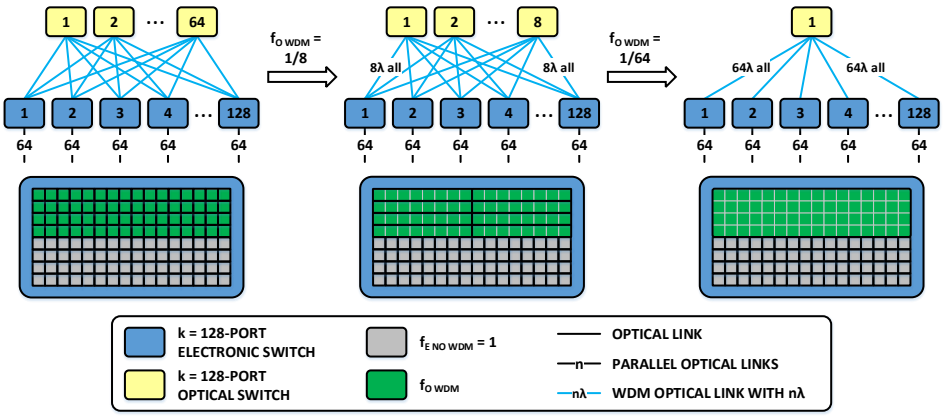

Fig. 4.10 Impact of $f_{O\,WDM}$ on a two-layer network.

The effect of $f_{O\,WDM}$ is explored in Fig. 4.11 corresponding to three-layer HFT with 131072 servers built with 128-port switches. The graph on the left studies HFT with one hybrid layer and the graph on the right with two hybrid layers.

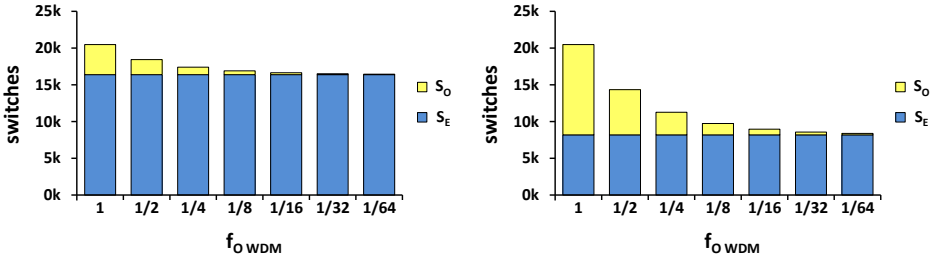

Fig. 4.11 Impact of $f_{O\,WDM}$ on a three-layer network.

The number of ESs remains constant in both graphs: 16384 ESs and 8192 ESs, respectively. The number of OSs reduces from 4096 to 64 OSs with one hybrid layer, and from 12288 OSs to 192 OSs with two hybrid layers.

The impact of each one of these factors on three-layer networks built with 128-port switches is evaluated in Fig. 4.12. By reducing $f_{E\ NO\ WDM}$, the number of NO WDM transceivers are reduced. In other words, by increasing the number of ports of the NO WDM transceivers, fewer transceivers are required. However, the number of switches remains constant since they are simply packaged with more integrated transceivers, and the number of fibers remains also constant since they are NO WDM transceivers. An example of the reduction in the number of transceivers for FT topology is shown in the left graph.

By reducing $f_{E\ WDM}$, not only fewer transceivers are required, but also fewer fibers thanks to the use of WDM. However, the number of switches remains constant because ESs require one port per wavelength. An example of the reduction in the number of fibers for EFT using WDM links between the aggregation and core ESs is shown in the center graph.

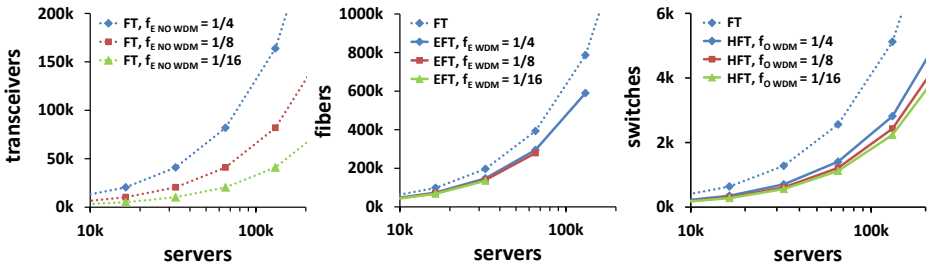

Fig. 4.12 Impact of $f_{E\ NO\ WDM}$, $f_{E\ WDM}$, and $f_{O\ WDM}$ on three-layer networks.

By reducing $f_{O\ WDM}$, fewer transceivers, fewer fibers, and fewer (optical) switches are required. The number of transceivers reduces by increasing the number of ports of the devices. The number of fibers decreases by the use of WDM. The number of switches scales down because, in contrast with ESs that require one port per wavelength, OSs are able to switch several wavelengths per port thanks to its transparency. An example of the reduction in the number of switches in HFT is shown in the right graph.

4.4 Model equations

This section presents the equations describing the number of servers (N), switches (S), transceivers (T) and fibers (F) for the four topologies. The equations are derived starting from the simplest case, FT, and finishing with the more general case, EHFT. Servers and switches are expressed in terms of the model parameters; transceivers and fibers are presented also in terms of the number of servers (N).

4.4.1 Servers

The four topologies share Eq. (4.1) for the number of servers. In addition to previous studies [56, 89], it includes the f_P factor for the reasons explained before. An example with 2048 servers for each topology has been previously shown in Fig. 4.1 (with parameters $k = 128$, $l = 2$, and $f_P = 1/4$).

$$N = 2 \cdot f_P \cdot (k/2)^l \tag{4.1}$$

4.4.2 Switches

The number of ESs in FT and EFT topologies is given by Eq. (4.2). Again, it includes the f_P factor in addition to previous literature studies.

$$S_{FT} = S_{EFT} = (2 \cdot l - 1) \cdot f_P \cdot (k/2)^{l-1} \tag{4.2}$$

The total number of switches in hybrid topologies is derived with Eq. (4.3). The first two terms provide the number of ESs and the last term the number of OSs. In more detail, the first term is equal to Eq. (4.2) and corresponds to the number of ESs in the network if all the switches were electronic. The second term subtracts part of these ESs. It can be visualized as a sub-tree of ESs with l_H layers, from which a fraction is selected with the f_O factor. The third term substitutes the removed sub-tree of ESs by fewer OSs thanks to the additional $f_{O\,WDM}$ factor included in this term. Indeed, the use of WDM in OSs effectively multiplies the number of ports of OSs. As a result, several ESs are replaced by fewer OSs, reducing the number of switches in our hybrid topologies. This is the main feature of OSs that we exploit to improve the scaling of our hybrid topologies. Examples of this advantage have already been shown in Fig. 4.1 and Fig. 4.2. For instance, in the hybrid topologies of Fig. 4.1, the $l = 2$ layers tree has 48 switches (first term). The $l_H = 1$ layer sub-tree has 16 switches from which half of them are

removed (f_O = 1/2, second term). Then, the 8 ESs switches removed are replaced by 2 OSs using four wavelengths per link ($f_{O\,WDM}$ = 1/4, third term).

$$S_{HFT} = S_{EHFT} = [(2 \cdot l - 1) - (2 \cdot l_H - 1) \cdot f_O + \\ + (2 \cdot l_H - 1) \cdot f_O \cdot f_{O\,WDM}] \cdot f_P \cdot (k/2)^{l-1} \tag{4.3}$$

4.4.3 Transceivers

Equation (4.4) determines the number of transceivers in a FT topology. This topology has only one type of transceiver, with $1/f_{E\,NO\,WDM}$ ports, interconnecting ESs without WDM. The equation can be written also as $T_{FT} = S_{FT} \cdot k \cdot f_{E\,NO\,WDM}$, and visualized as the number of $(1/f_{E\,NO\,WDM})$-port transceivers required by S_{FT} k-port switches.

$$T_{FT} = (2 \cdot l - 1) \cdot f_{E\,NO\,WDM} \cdot N \tag{4.4}$$

Equation (4.5) provides the number of transceivers in an EFT topology. EFT substitutes the NO WDM transceivers in the core layer of FT by WDM transceivers. Thus, it requires two types of transceivers, and consequently, the equation has two terms. Each layer of FT has N connections requiring $2 \cdot N$ 1-port transceivers since every connection needs two transceivers. EFT substitutes the $2 \cdot N \cdot f_{E\,NO\,WDM}$ $(1/f_{E\,NO\,WDM})$-port transceivers required by the core layer of FT by $2 \cdot N \cdot f_{E\,WDM}$ $(1/f_{E\,WDM})$-port WDM transceivers.

$$T_{EFT} = [(2 \cdot l - 3) \cdot f_{E\,NO\,WDM} + 2 \cdot f_{E\,WDM}] \cdot N \tag{4.5}$$

Equation (4.6) obtains the number of transceivers in a HFT topology. It has two types of transceivers and two terms. Since only the ESs require transceivers in a HFT topology and every ES integrates k 1-port transceivers, a total of $[(2 \cdot l - 1) - (2 \cdot l_H - 1) \cdot f_O] \cdot N$ 1-port transceivers are needed, according to Eq. (4.3). From all these transceivers, $f_O \cdot N$ transceivers connect ESs with OSs. The remaining $(2 \cdot l - 2 \cdot l_H \cdot f_O - 1) \cdot N$ transceivers connect ESs. Therefore, the first term of Eq. (4.6) provides the number of $(1/f_{E\,NO\,WDM})$-port transceivers connecting ESs and the second term gives the number of $(1/f_{O\,WDM})$-port transceivers connecting ESs and OSs.

$$T_{HFT} = [(2 \cdot l - 2 \cdot l_H \cdot f_O - 1) \cdot f_{E\ NO\ WDM} + f_O \cdot f_{O\ WDM}] \cdot N \tag{4.6}$$

Equation (4.7) calculates the number of transceivers in an EHFT topology. It is obtained following a reasoning similar to Eq. (4.6). However, in this case, the $(2 \cdot l - 2 \cdot l_H \cdot f_O - 1) \cdot N$ 1-port transceivers connecting ESs are distributed in another two terms. One term for the $2 \cdot (1 - f_O) \cdot N$ WDM transceivers connecting ESs in the core layer. The remaining $[(2 \cdot l - 3) - 2 \cdot (l_H - 1) \cdot f_O] \cdot N$ transceivers interconnect ESs without WDM. Thus, the first term of Eq. (4.7) gives the number of $(1/f_{E\ NO\ WDM})$-port transceivers connecting ESs; the second term expresses the number of $(1/f_{E\ WDM})$-port transceivers connecting ESs with WDM links in the core layer; the third term provides the number of $(1/f_{O\ WDM})$-port transceivers connecting ESs and OSs.

$$\begin{aligned} T_{EHFT} = [[(2 \cdot l - 3) - 2 \cdot (l_H - 1) \cdot f_O] \cdot f_{E\ NO\ WDM} + \\ + 2 \cdot (1 - f_O) \cdot f_{E\ WDM} + f_O \cdot f_{O\ WDM}] \cdot N \end{aligned} \tag{4.7}$$

4.4.4 Fibers

Equation (4.8) provides the number of fibers in a FT topology, where every layer requires $2 \cdot N$ fibers.

$$F_{FT} = 2 \cdot l \cdot N \tag{4.8}$$

Equation (4.9) determines the number of fibers in an EFT topology. It replaces $2 \cdot N$ fibers required by the core layer of FT by $2 \cdot N \cdot f_{E\ WDM}$ fibers. Thus, $l - 1$ layers have $2 \cdot N$ fibers and the last layer has $2 \cdot N \cdot f_{E\ WDM}$.

$$F_{EFT} = 2 \cdot (l - 1 + f_{E\ WDM}) \cdot N \tag{4.9}$$

Equation (4.10) obtains the number of fibers in a HFT topology. From all the $2 \cdot l \cdot N$ fibers required by the full tree (first term), $2 \cdot l_H \cdot f_O \cdot N$ fibers of the optical sub-tree are removed (second term) and replaced by $2 \cdot l_H \cdot f_O \cdot f_{O\ WDM} \cdot N$ WDM links (third term).

$$F_{HFT} = 2 \cdot [l - l_H \cdot f_O + l_H \cdot f_O \cdot f_{O\ WDM}] \cdot N \tag{4.10}$$

Equation (4.11) calculates the number of fibers in an EHFT topology. FT requires a total of $2 \cdot l \cdot N$ fibers, where every layer has $2 \cdot N$ fibers. EHFT substitutes $2 \cdot (1 - f_O) \cdot N$ of the core-layer fibers by $2 \cdot (1 - f_O) \cdot f_{E\,WDM} \cdot N$ WDM links between ESs (with $1/f_{E\,WDM}$ wavelengths per link). It also replaces $2 \cdot l_H \cdot f_O \cdot N$ links by $2 \cdot l_H \cdot f_O \cdot f_{O\,WDM} \cdot N$ WDM links between ESs and OSs (with $1/f_{O\,WDM}$ wavelengths per link).

$$\begin{aligned} F_{EHFT} = 2 \cdot [l - 1 - (l_H - 1) \cdot f_O + \\ + (1 - f_O) \cdot f_{E\,WDM} + l_H \cdot f_O \cdot f_{O\,WDM}] \cdot N \end{aligned} \tag{4.11}$$

4.5 Validity conditions

The factors f_P, f_O, $f_{E\,NO\,WDM}$, $f_{E\,WDM}$, and $f_{O\,WDM}$ have been equally defined in Table 4.1. Consequently, they have the same minimum ($f_{min} = 2/k$) and maximum ($f_{max} = 1$) individual values. However, they are interrelated and they must satisfy a number of conditions in order to provide valid solutions with the previous equations. These conditions are summarized mathematically in Table 4.2.

The conditions with the largest implications in the scaling of the networks are (1) and (1'), associated to EFT and EHFT respectively. These conditions mean that only one wavelength can be used in a full EFT/EHFT ($f_P = 1$), that two wavelengths can be used in half EFT/EHFT ($f_P = 1/2$), and so on. The main difference between them is that (1') only applies when the upper layer of EHFT is not fully optical, and there are WDM transceivers connecting ESs. This restriction comes by the fact that, in a full EFT/EHFT, every switch is connected to another switch by a single link, and thus, it is not possible to aggregate links between ESs with wavelength-division multiplexing. In half EFT/EHFT, the switches in the upper layer are connected with double links, and therefore, it is possible to bundle them in a single link with WDM.

Table 4.2 Validity conditions associated to model parameters.

Reference	Validity condition	Required by
(1)	$f_{E\,WDM} \geq f_P$	EFT
(1')	$f_{E\,WDM} \geq f_P \; if \; f_O \neq 1$	EHFT
(2)	$f_O \cdot f_P \geq (2/k)^{l-1}$	HFT / EHFT
(3)	$f_O \cdot f_{O\,WDM} \geq (2/k)$	HFT / EHFT
(4)	$f_O \cdot f_{E\,WDM} \geq (2/k)$	EHFT

In effect, in a half network, the aggregation/core switches are connected to only half of the core/aggregation switches compared to a full network. In order to keep the full bisection bandwidth, the core links in half network are doubled, so they can be bundled together by using transceivers with two wavelengths. For instance, four wavelengths may be used in the upper layer of the EFT/EHFT examples of Fig. 4.1 with f_P = 1/4. It is worth noting that optical switches do not suffer from the analogous condition $f_{O\,WDM} \geq f_P$.

The validity condition (2) must be fulfilled so Eq. (4.3) provides an integer number of electronic switches. Validity conditions (3) and (4) are imposed by physical limitations due to the number of ports of the switches.

The model parameters introduced in section 4.2 and the validity conditions introduced in this section are shown graphically in Fig. 4.13. EHFT is the more general topology, requiring all parameters to be defined and all validity conditions to be fulfilled.

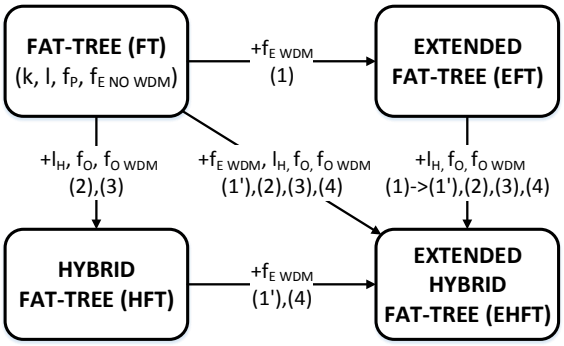

Fig. 4.13 Parameters and validity conditions required by each topology.

4.6 Equations relations

The equations of Section 4.4, describing the topologies in terms of switches, transceivers, and fibers are interrelated. Equation (4.12) to Eq. (4.20) present a number of these relations. With them, the number of devices of simpler topologies can be derived from the more general topologies expressions. For instance, according to Eq. (4.12), it is possible to obtain the number of switches in electronic topologies by making the optical factor zero in the expression giving the number of switches in the hybrid topologies. Intuitively, if none of the ESs is substituted by OSs in the hybrid topologies, then the number of switches is equivalent to the electronic topologies.

4.6.1 Relations between switches equations

$$S_{FT} = S_{EFT} = S_{HFT}|_{f_O=0} = S_{EHFT}|_{f_O=0} \tag{4.12}$$

4.6.2 Relations between transceivers equations

$$T_{FT} = T_{EFT}|_{f_{E\ WDM}=f_{E\ NO\ WDM}} = T_{HFT}|_{f_O=0} = T_{EHFT}\Big|_{\substack{f_O=0 \\ f_{E\ NO\ WDM}=f_{E\ WDM}}} \tag{4.13}$$

$$T_{HFT} = T_{EHFT}|_{f_{E\ WDM}=f_{E\ NO\ WDM}} \tag{4.14}$$

$$T_{EFT} = T_{EHFT}|_{f_O=0} \tag{4.15}$$

$$T_{HFT}|_{f_O=1} = T_{EHFT}|_{f_O=1} \tag{4.16}$$

4.6.3 Relations between fibers equations

$$F_{FT} = F_{EFT}|_{f_{E\ WDM}=1} = F_{HFT}|_{f_O=0} = F_{EHFT}\Big|_{\substack{f_O=0 \\ f_{E\ WDM}=1}} \tag{4.17}$$

$$F_{HFT} = F_{EHFT}|_{f_{E\ WDM}=1} \tag{4.18}$$

$$F_{EFT} = F_{EHFT}|_{f_O=0} \tag{4.19}$$

$$F_{EFT}|_{f_O=1} = F_{EHFT}|_{f_O=1} \tag{4.20}$$

4.7 Model examples

This section presents illustrative examples of FT, EFT, HFT, and EHFT topologies. All the architectures have three layers and interconnect half network (i.e. 64 servers) with 8-port switches. The low number of ports per switch and the small size of the networks are selected to enable diagrams including all switches and links. Analogous examples with 128-port switches are used in Chapter 5 to investigate the scaling of the topologies in terms of devices, power consumption, and cost.

Regarding hybrid topologies, four different cases are considered, namely A, B, C, and D. Cases A and B have one hybrid layer, and cases C and D have two hybrid layers. Cases A and C have half optical hybrid layer/s, and cases B and D have fully optical hybrid layer/s. The parameters associated with each case are: A: $f_O = 1/2$, $l_H = 1$; B: $f_O = 1$, $l_H = 1$; C: $f_O = 1/2$, $l_H = 2$; and D: $f_O = 1$, $l_H = 2$. These values are selected because they serve as indicators of the maximum savings achievable by these topologies; lower values of the optical factor f_O are possible, but they result in more moderate savings.

It is worth noting that the choice of replacing a number of ESs by OSs in the hybrid topologies has a natural consequence. It implies that the traffic routed before only through ESs is routed now (partially or totally) through OSs. In order to better illustrate the impact of the selected values for the configuration parameters on the topologies, the distribution of traffic in the different architectures is summarized in Table 4.3: *E* means traffic through ESs and *O* means traffic through OSs. In electronic topologies, FT and EFT, all the traffic goes through ESs. In hybrid topologies, the traffic distribution depends on the network type. In HFT-A / EHFT-A, half of the inter-cluster traffic is optical; the rest is electronic. In HFT-B / EHFT-B, all the inter-cluster traffic is optical; the rest is electronic. In HFT-C / EHFT-C, half the intra-cluster, and half the inter-cluster traffic is optical; the rest is electronic. Finally, in HFT-D / EHFT-D, all the intra-cluster and inter-cluster traffic is optical; only the intra-rack traffic is electronic.

Table 4.3 Traffic distribution in topologies under consideration.

Topology	**Number of hybrid layers**	**Intra-rack traffic**	**Intra-cluster traffic**	**Inter-cluster traffic**
FT / EFT	0	*E*	*E*	*E*
HFT-A / EHFT-A	1	*E*	*E*	*E* and *O*
HFT-B / EHFT-B	1	*E*	*E*	*O*
HFT-C / EHFT-C	2	*E*	*E* and *O*	*E* and *O*
HFT-D / EHFT-D	2	*E*	*O*	*O*

The diagrams represent electronic switches as blue boxes and optical switches as yellow boxes. Besides, the following convention is adopted regarding the visualization of NO WDM and WDM links. Black lines denote NO WDM links: a black line without number denotes a link with two fibers, a black line marked with number *2* denotes two parallel links with four fibers, and so on. Blue lines represent WDM links: a blue line without number denotes one wavelength and

two fibers, a blue line marked with 2λ represents two wavelengths and also two fibers, and so on; a blue line marked with $2x4\lambda$ means two parallel WDM links, each one of them with four wavelengths and two fibers. Finally, when all the links in the layer have the same characteristics, this is marked for clarity just in one (or two) of the links of the layer with prefix *all*. For instance, a black line marked with *8 all* means that all links in that layer have width 8 and 16 fibers.

4.7.1 FT and EFT examples

A FT example deploying half network (i.e. with only four of the eight pods of the full network) is shown in Fig. 4.14. Each pod connects 16 servers with 8 switches arranged in the first two layers. The core switches have two links (and four fibers) connected to an aggregation switch of every pod. These links require four fibers because WDM is not used.

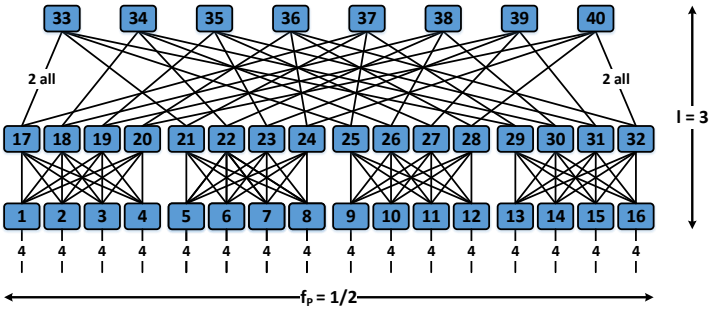

Fig. 4.14 FT example.

An example of EFT is shown in Fig. 4.15, which in contrast with FT includes WDM links in the core layer to save fibers.

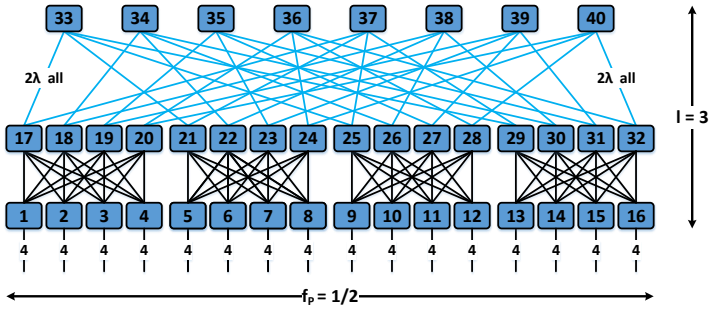

Fig. 4.15 EFT example.

As a result, the double links of the previously shown Fig. 4.14 requiring four fibers are reduced to two fibers by multiplexing two wavelengths per link. Note that the use of WDM in EFT is linked to the f_P factor, as mentioned in Section 4.5. The full network cannot use WDM because every core switch has connected just one port to an aggregation switch of every pod. (Properly speaking, the full network may use WDM with one wavelength per link, but it does not have an effect when the goal is to save fibers).

4.7.2 HFT and EHFT examples

Examples of HFT-A and EHFT-A topologies are shown in Fig. 4.16 and Fig. 4.17, respectively. Both networks have a single hybrid layer, replace half of the ESs in the hybrid layer by OSs, and use WDM links with two wavelengths to connect the OSs. The difference between them is that EHFT employs WDM links between the ESs in the core layer too, and therefore, requires fewer fibers than HFT. The restriction to two wavelengths in these examples is given by the validity conditions (i.e. condition (3) for HFT-A, and conditions (1'), (3), and (4) for EHFT-A).

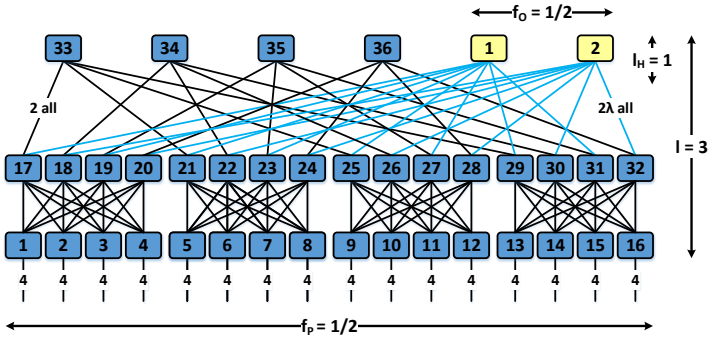

Fig. 4.16 HFT-A example.

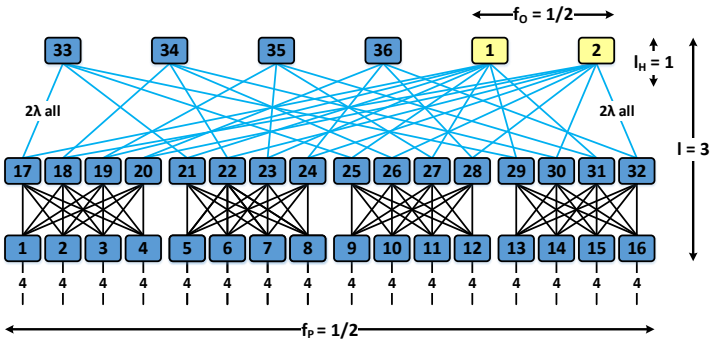

Fig. 4.17 EHFT-A example.

An example of HFT-B and EHFT-B, with a fully optical core layer, is shown in Fig. 4.18. In effect, the hybrid topologies are equivalent in cases B and D. Mathematically, $S_{HFT} = S_{EHFT}$, $T_{HFT}|_{f_O=1} = T_{EHFT}|_{f_O=1}$, and $F_{HFT}|_{f_O=1} = F_{EHFT}|_{f_O=1}$. Another way to reach the same conclusion is by realizing that the difference between HFT and EHFT is that the last one substitutes the links between the ESs in the core layer by WDM links, but in a fully optical core there are no ESs and both topologies are identical. Note that cases B and D are able to use four wavelengths in the links of the upper layer/s due to validity condition (3).

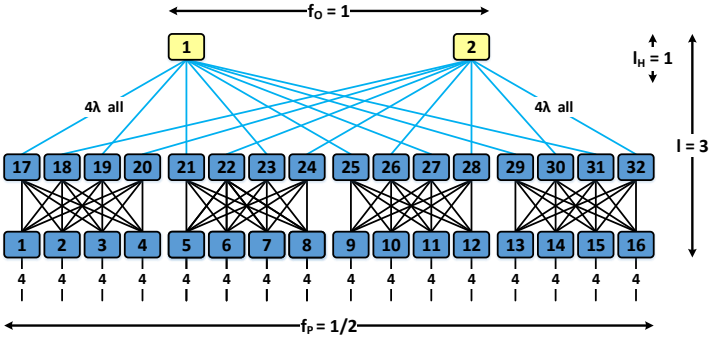

Fig. 4.18 HFT-B = EHFT-B example.

Examples of HFT-C and EHFT-C are represented in Fig. 4.19 and Fig. 4.20, respectively. Both of them have two hybrid layers where half of the ESs have been replaced by OSs. The difference between them is that EHFT-C includes also WDM among the ESs, in a similar manner to case A.

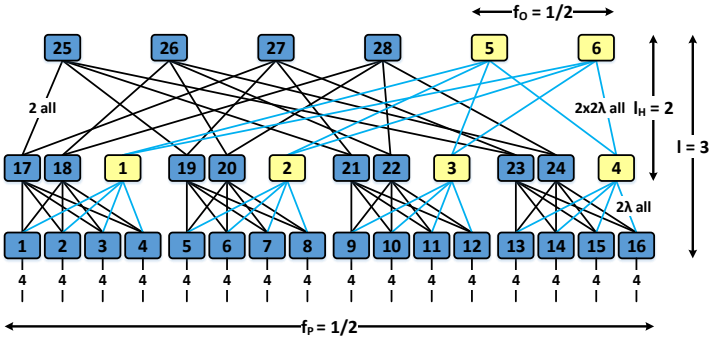

Fig. 4.19 HFT-C example.

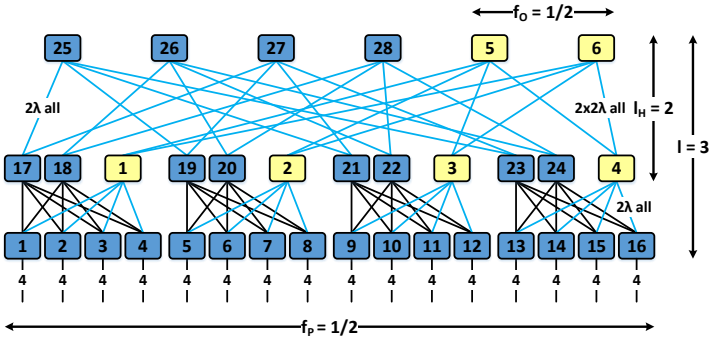

Fig. 4.20 EHFT-C example.

An example of HFT-D and EHFT-D, with two fully optical layers, is shown in Fig. 4.21. This is the minimum topology that can be achieved with this model (and these parameters). It is obtained by maximizing the number of hybrid layers, which must be fully optical, and the number of wavelengths (in this case, with f_O = 1, l_H = 2, and $f_{O\,WDM}$ = 1/4).

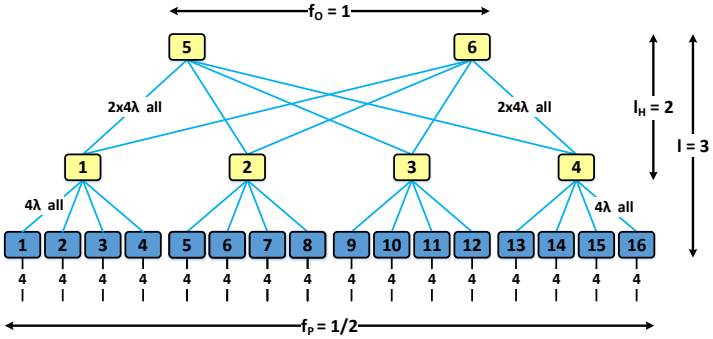

Fig. 4.21 HFT-D = EHFT-D example.

4.8 Summary

Fat-Tree topology is a well-known and widely deployed data center topology with two important features: it has full bisection bandwidth and it is not limited in scaling by the number of ports of the switches. However, being a topology purely based on electronic switches with (NO WDM) optical transceivers, it scales poorly in the number of devices, power consumption, and cost.

In this chapter, we have presented our analytic model describing electronic and hybrid topologies sharing these two important features and exploring the introduction of WDM and optical switching technologies to improve the scaling of the

networks. EFT includes wavelength division multiplexing in the core layer of FT in order to reduce the number of fibers. The hybrid topologies, HFT and EHFT, introduce also OSs in order to reduce the number of switches, transceivers, and fibers.

Our analytic model describes accurately these topologies by providing a set of equations to compute the number of servers, switches, transceivers, and fibers. As a result, it enables to investigate and compare the scaling of the topologies in terms of devices, power consumption, and cost, as it will be shown in Chapter 5.

The model parameters describing the topologies provide great flexibility to customize and design multiple architectures. A collection of examples is visualized, helping to illustrate the flexibility of the model, and providing the basis for the scaling study of Chapter 5. The number of layers (l), the number of ports in the switches (k), and the partition factor (f_P) allow sizing the network to the desired number of servers (in an appropriate manner without wasting resources). The number of hybrid layers (l_H) and the optical factor (f_O) allow selecting the fraction of optical switches replacing electronic switches in the hybrid networks. Finally, the transceivers factors ($f_{E\ NO\ WDM}$, $f_{E\ WDM}$, and $f_{O\ WDM}$) model different types of transceivers and enable computing the number of transceivers and fibers in the networks. Moreover, they also allow investigating the impact of introducing WDM technologies on electronic and hybrid networks, and the impact of the number of wavelengths in such transceivers.

Chapter 5

Scaling of Electronic and Hybrid Data Center Networks

5.1 Introduction

Chapter 4 has presented the analytic model of a total of four topologies: FT, EFT, HFT, and EHFT. These topologies explore the introduction of OSs and WDM technologies in Fat-Tree like topologies having two important features: full bi-section bandwidth and scalability in number of servers not limited by the number of ports of the switches. FT and EFT are architectures based exclusively on ESs; HFT and EHFT are hybrid topologies combining ESs and OSs. The difference between the electronic topologies is that FT does not utilize WDM technologies, whereas EFT does. The distinction between the hybrid topologies is that HFT uses only WDM technologies in the links connected to OSs whereas EHFT includes also WDM links between ESs.

The analytic model comprises a set of equations to compute the number of servers, switches, transceivers, and fibers for the topologies. The formulas are expressed in terms of the same configuration parameters, which provide great flexibility to implement a variety of architectures as shown in a number of examples.

This chapter explores and compares the scaling of these networks by using the novel analytic model. The scaling is investigated in two directions: first, by studying the scaling of the networks in terms of devices, i.e., in terms of switches, transceivers, and fibers; second, by investigating the scaling of the architectures in terms of power consumption and cost.

The first study solves questions such as: how many fibers are saved by introducing WDM technologies in FT?, how many switches, transceivers, and fibers

are saved by introducing OSs and WDM in FT? what values of the configuration parameters provide maximum savings of devices? what are the values of these maximum savings?, and what is the impact of the number of wavelengths per transceiver in these savings? We find out that wavelength multiplexing in EFT reduces the number of fibers by 25% with 4-port transceivers compared with FT. Minimum hybrid networks with 4-port transceivers reduce the number of switches by 45%, the transceivers by 60%, and the fibers by 50%. The maximum savings are achieved with the configuration parameters that maximize the fraction of OSs and wavelengths in the hybrid networks. The maximum theoretical savings in minimum hybrid networks with 128-port switches are obtained with 64-port transceivers, reducing by 59.06% the number of switches, by 97.50% the transceivers, and by 65.63% the fibers. Thus, most of the savings are already possible with four wavelengths per transceiver although higher port-density devices reduce the devices even further.

The second study provides answers to questions such as: how much cost is saved thanks to the savings in fibers of EFT? how much power consumption and cost are saved due to the reduction in the number of switches, transceivers, and fibers of HFT and EHFT? which devices are the main contributors to power consumption and cost? should the whole network be implemented with SM devices, or by a combination of MM and SM devices? We find out that EFT, despite reducing the number of fibers, increases cost above 10% due to the increased expense of WDM transceivers. The topologies more efficient in terms of power consumption and cost are HFT-D and EHFT-D, with two layers of optical switches, reducing power consumption by 55.6% and cost by 48.8%. Electronic switches are the main contributors in the system to the power consumption, whereas optical transceivers are the main contributors to the cost. Regarding the MM and SM study, the best choice in terms of power consumption and cost in hybrid networks depends on the number of hybrid layers: with one hybrid layer the MMS option is more efficient, in which only the hybrid layer is SM; with two hybrid layers the MSS option is preferred, in which both hybrid layers are SM.

5.2 Selected values of model parameters

The generality and flexibility of the model presented in Chapter 4 provide a large set of solutions depending on the chosen values of the configuration parameters. These parameters allow implementing different flavors of each one of the four architectures by adjusting the size, the type of optical transceivers, the fraction of optical switches, and other relevant characteristics.

In order to provide concrete results, this chapter considers only a subset of all possible theoretical solutions, summarized in Table 5.1. The number of ports of the switches is $k = 128$ in all cases. This is the maximum amount of ports in ESs, based on a single switching ASIC, commercially available at present: e.g. there are available versions with 128-ports at 10G [51], 25G [52], and recently announced versions at 50G [53], and 100G [54]. The number of layers is $l = 3$ in all cases. Although small data centers with only two layers and up to 8192 servers are also relevant, this study focuses on the scaling of larger data centers with up to 524288 servers. As a consequence of investigating three-layer networks, the number of hybrid layers l_H in hybrid topologies may be set to be one or two. Regarding the optical factor f_O, fully optical layer/s and half optical layer/s are examined. These values of f_O produce the largest savings in hybrid topologies so they are chosen to evaluate its scaling. The selected values of l_H and f_O result in a total of four different cases for hybrid topologies: A: $l_H = 1$, $f_O = 1/2$; B: $l_H = 1$, $f_O = 1$; C: $l_H = 2$, $f_O = 1/2$; and D: $l_H = 2$, $f_O = 1$. Thus, the three-layer networks differ in the number of optical switches included in the hybrid layer/s: type A networks have a single hybrid layer of switches, from which half are optical switches and the other half are electronic switches; type B networks have a single layer of optical switches; type C networks have two layers of switches, from which half are optical switches and the other half are electronic switches; type D networks have two layers of optical switches. As it will be shown, the maximum savings are achieved with case D. Regarding the transceivers, 4-port, 8-port, 16-port, 32-port, and 64-port devices are considered. 4-port transceivers are common at present, with standards like QSFP+ (4x10G) and QSFP28 (4x25G). 8-port transceivers will be soon available with the QSFP-DD standard (8 x 25G with NRZ modulation or 8x50G with PAM4 modulation), required by the latest ASIC developments.

Table 5.1 Selected values of model parameters.

Symbol	**Values**	**Units**
k	128	ports
l	3	layers
l_H	1, and 2	hybrid layers
f_O	1, and 1/2	—
f_P	1/4, 1/8, 1/16, 1/32, and 1/64	—
$f_{E\ NO\ WDM}$	1/4, 1/8, 1/16, 1/32, and 1/64	1/ports
$f_{E\ WDM}$	1/4, 1/8, 1/16, 1/32, and 1/64	1/ports
$f_{O\ WDM}$	1/4, 1/8, 1/16, 1/32, and 1/64	1/ports

16-port and 32-port transceivers are included to investigate the impact of a larger number of ports in these devices; 64-port transceivers are considered to evaluate the maximum savings that may be achieved. Finally, although the model provides the flexibility of choosing a different number of ports for the distinct types of transceivers, this study fixes the same number of ports for all transceivers. Therefore, $f_{E\ NO\ WDM} = f_{E\ WDM} = f_{O\ WDM}$ = 1/4, 1/8, 1/16, 1/32 or 1/64.

5.3 Scaling of topologies in terms of devices

This section explores the scaling of the four topologies in terms of devices, i.e., in terms of switches, transceivers, and fibers. First, the scaling of EFT, HFT, and EHFT is compared graphically against FT, one topology at a time. Then, the scaling of the four topologies is compared numerically for a 131072 servers network using FT as a baseline.

5.3.1 Scaling of FT and EFT topologies

The scaling of FT (dotted curves) and EFT is compared graphically in Fig. 5.1. The first important conclusion is that the number of switches and transceivers in both topologies is equivalent. This is graphically visualized in the overlapping curves of switches and transceivers, shown in the left and center graphs respectively. Mathematically, both topologies share the same equation for the switches, $S_{FT} = S_{EFT}$, and the number of transceivers is related by the expression $T_{EFT}|_{f_{E\ WDM}=f_{E\ NO\ WDM}} = T_{FT}$. The second important conclusion is that EFT saves fibers compared to FT. Graphically, the EFT curves are below the FT (dotted) curves in the right graph. Mathematically, $F_{EFT}|_{f_{E\ WDM}<1} < F_{FT}$. Intuitively, EFT introduces WDM in the core layer of FT, and therefore, achieves savings in fibers, but the number of switches and transceivers does not change.

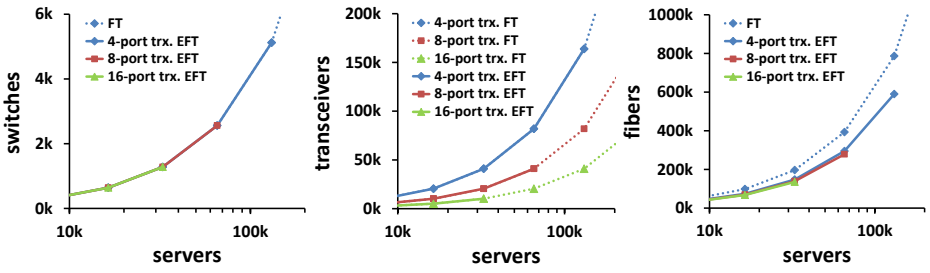

Fig. 5.1 Scaling in number of devices of FT and EFT.

An important remark is that the savings in fibers can only be achieved in the upper layer, as it has been discussed in Section 4.5. Hence, the impact of increasing the number of wavelengths in the transceivers to reduce the number of fibers of EFT is limited. In effect, the curves for fibers corresponding to 4/8/16-ports are very close to each other. Finally, as pointed out previously in the same section, these savings in fibers are possible if and only if the full network is not deployed. As a result, while FT can scale up to 524288 servers (not shown in the graphs), EFT can expand only up 131072 servers with 4-port transceivers, up to 65536 servers with 8-port transceivers, and up to 32768 servers with 16-port transceivers. For instance, in a 1/16 of the full network, with 32768 servers, the total number of fibers is 196608 in a FT without WDM; in EFT, the number of fibers reduces by 25.0% to 147456 fibers with 4-port transceivers, by 29.2% to 139268 fibers with 8-port transceivers, and by 31.25% to 135168 fibers with 16-port transceivers.

5.3.2 Scaling of FT and HFT topologies

The scaling of HFT-A, HFT-B, HFT-C, and HFT-D is explored graphically in Fig. 5.2, Fig. 5.3, Fig. 5.4, and Fig. 5.5, respectively. One important conclusion from the figures is that HFT accomplishes savings in switches (left graph), transceivers (center graph), and fibers (right graph). Indeed, the HFT curves are below the FT (dotted) curves in each graph. The reduction in switches is due to the fact that the number of ports of the OSs is effectively multiplied by the number of wavelengths used in the WDM links, and therefore, fewer OSs are required compared to ESs. The reduction in transceivers is due not only to the lower number of switches needed (implying fewer transceivers), but also because OSs do not necessitate transceivers. Finally, the decrease in fibers is justified by the introduction of WDM, and because fewer switches and transceivers require fewer fibers to be connected. Another important conclusion is that the scaling of HFT is not limited as EFT is. Thus, HFT scale as well as FT (in these examples up to 524288 servers).

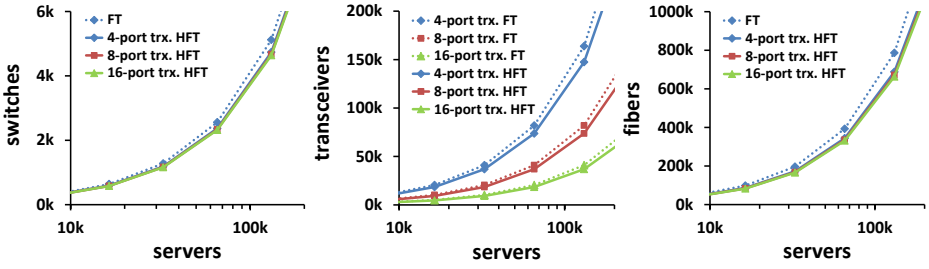

Fig. 5.2 Scaling in number of devices of FT and HFT-A.

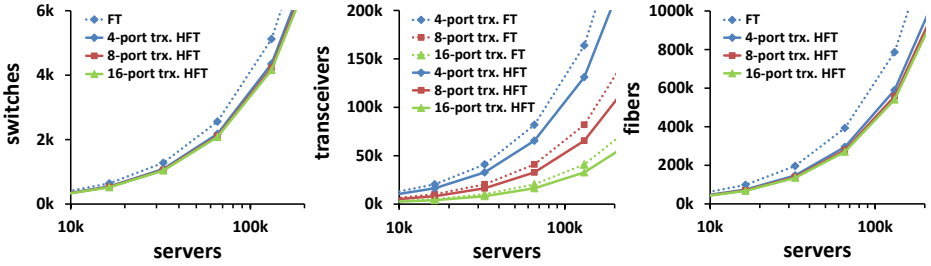

Fig. 5.3 Scaling in number of devices of FT and HFT-B = EHFT-B.

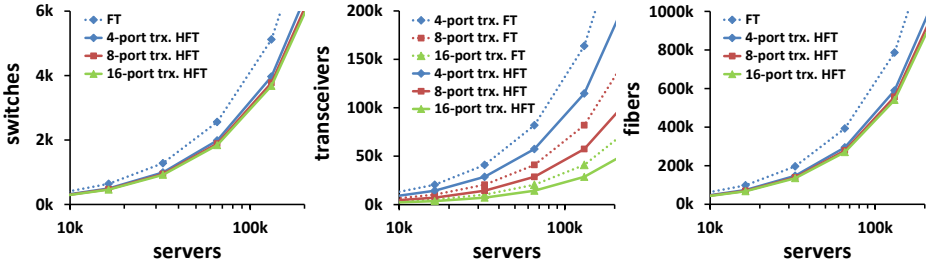

Fig. 5.4 Scaling in number of devices of FT and HFT-C.

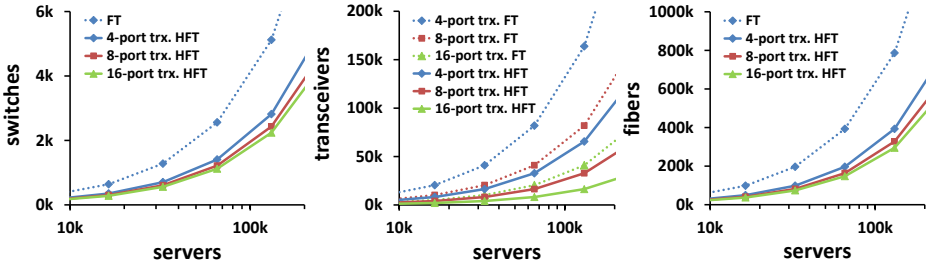

Fig. 5.5 Scaling in number of devices of FT and HFT-D = EHFT-D.

Finally, by comparing the gap between the curves, it is possible to infer that solutions A, B, C, and D require increasingly fewer devices. This is explained because the more ESs are substituted by OSs, the larger are the savings. Cases A and B replace one half/full layer of ESs, respectively. Cases C and D substitute two half/full layers of ESs, respectively. For every case, the larger the number of wavelengths, the larger the savings. Thus, from all the options considered, the one providing fewer switches, transceivers, and fibers is case D with 16-port transceivers.

5.3.3 Scaling of FT and EHFT topologies

As discussed before, HFT and EHFT are equivalent in cases B and D. Thus, the previously shown Fig. 5.3 and Fig. 5.5 of HFT are also valid for EHFT. However, cases A and C for EHFT, shown in Fig. 5.6 and Fig. 5.7, have an important difference in terms of scaling with respect to HFT.

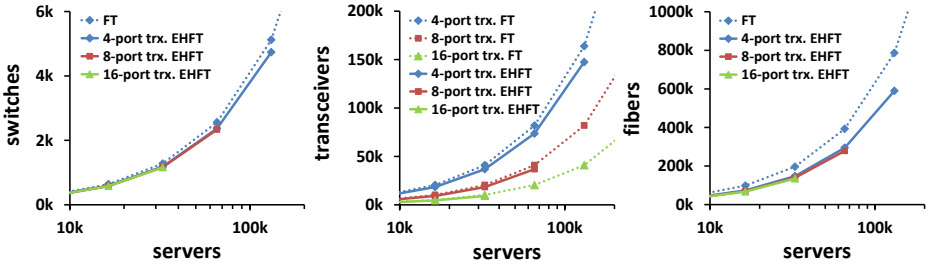

Fig. 5.6 Scaling in number of devices of FT and EHFT-A.

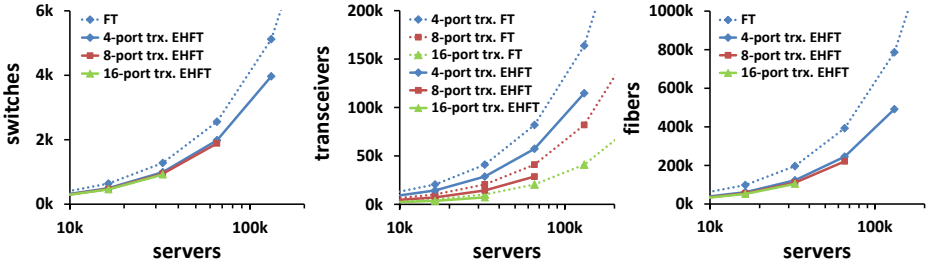

Fig. 5.7 Scaling in number of devices of FT and EHFT-C.

Indeed, EHFT is limited by the validity condition (1') in a similar manner to EFT being limited by validity condition (1). Thus, the network only scales up to 131072 servers with 4-port transceivers, up to 65536 servers with 8-port transceivers, and up to 32768 servers with 16-port transceivers. Apart from this, the conclusions are similar to HFT.

5.3.4 Scaling of FT, EFT, HFT, and EHFT topologies

A numerical comparison of the savings in devices achieved by EFT, HFT, and EHFT against FT in networks with 131072 servers is shown in Table 4.3. The baseline is a FT requiring 5120 128-port switches, 163840 4-port transceivers, and 786432 fibers. The hybrid topologies consider four cases: A: $l_H = 1$, $f_O = 1/2$; B:

$l_H = 1$, $f_O = 1$; C: $l_H = 2$, $f_O = 1/2$; D: $l_H = 2$, $f_O = 1$. A number of the solutions in the table correspond to the highest points of a number of the curves of Fig. 5.1 to Fig. 5.7. Similar conclusions to the graphical analysis are reported in the table.

EFT does not achieve savings in switches or transceivers but reduces by 25% the number of fibers with 4-port transceivers. EFT does not provide solutions with transceivers with more than four ports in networks of the considered size because of the validity condition (1).

HFT provides savings in switches, transceivers, and fibers in all cases. It is not limited in scaling like EFT or EHFT, and it provides solutions with 8-port, 16-port, 32-port, and (in cases B and D) 64-port transceivers. HFT has equal savings than EHFT with fully optical hybrid layers (cases B and D). HFT does not have solutions with 64-port transceivers in cases A and C due to validity condition (3).

Table 5.2 Savings (%) in devices of EFT, HFT, and EHFT.

Topology	**Savings in switches (%)**	**Savings in trx. (%)**	**Savings in fibers (%)**
4-port trx. EFT	0	0	25.00
4-port trx. HFT-A	7.50	10.00	12.50
8-port trx. HFT-A	8.75	55.00	14.58
16-port trx. HFT-A	9.38	77.50	15.63
32-port trx. HFT-A	9.69	88.75	16.15
4-port trx. EHFT-A	7.50	10.00	25.00
4-port trx. HFT-B/EHFT-B	15.00	20.00	25.00
8-port trx. HFT-B/EHFT-B	17.50	60.00	29.17
16-port trx. HFT-B/EHFT-B	18.75	80.00	31.25
32-port trx. HFT-B/EHFT-B	19.38	90.00	32.29
64-port trx. HFT-B/EHFT-B	19.69	95.00	32.81
4-port trx. HFT-C	22.50	30.00	25.00
8-port trx. HFT-C	26.25	65.00	29.17
16-port trx. HFT-C	28.13	82.50	31.25
32-port trx. HFT-C	29.06	91.25	32.29
4-port trx. EHFT-C	22.50	30.00	37.50
4-port trx. HFT-D/EHFT-D	45.00	60.00	50.00
8-port trx. HFT-D/EHFT-D	52.50	80.00	58.33
16-port trx. HFT-D/EHFT-D	56.25	90.00	62.50
32-port trx. HFT-D/EHFT-D	58.13	95.00	64.58
64-port trx. HFT-D/EHFT-D	59.06	97.50	65.63

EHFT also provides savings in devices in all cases. It reduces the fibers even further than HFT with half optical hybrid layers (cases A and C). However, in these cases, it does not provide solutions for transceivers having more than four ports because of validity condition (4).

In general, the more ESs are replaced by OSs, the larger the savings in the number of switches, transceivers, and fibers. Thus, with one hybrid layer best results are obtained with a fully optical core (case B). Similarly, with two hybrid layers, the reduction is maximized when both layers are fully optical (case D). With 4-port transceivers, the number of switches is reduced by 45%, the transceivers by 60%, and the fibers by 50%. These numbers can be further reduced when transceivers with higher port-density become available. For instance, by having 8-port transceivers, the number of switches reduces by 52.50%, the transceivers by 80%, and the fibers by 58.33%. The maximum savings achievable by HFT and EHFT with 128-ports switches require the availability of 64-port optical transceivers and reduce the number of switches by 59.06%, the transceivers by 97.50%, and the fibers by 65.63%.

5.4 Scaling of topologies in power consumption and cost

Section 5.3 has explored the scaling of FT, EFT, HFT, and EHFT topologies in terms of switches, transceivers, and fibers. This section investigates the scaling of these topologies in terms of power consumption and cost in a 25G real case scenario that can be deployed at present with commercially available technologies. The 25G real case scenario is a three-layer data center built with 128-port switches, 4-port transceivers, and 131072 servers (l = 3, k = 128, f_P = 1/4, $f_{E\ NO\ WDM} = f_{E\ WDM} = f_{O\ WDM}$ = 1/4). Switches and transceivers have 25G interfaces. Again, cases A, B, C, and D are considered for the hybrid topologies. Cases A and B have one hybrid layer, and cases C and D have two hybrid layers. Cases A and C have half optical hybrid layer/s, and cases B and D have fully optical hybrid layer/s. Besides, three different options are studied in terms of MM and SM devices: MMS, MSS, and SSS. MMS represents the situation where the first two layers are MM and the core layer is SM. In the MSS case, only the first layer is MM and the rest are SM. In the SSS case, all layers are SM. Finally, in a similar manner to the equations of section 4.4, the graphs do not include the transceivers required by the servers.

5.4.1 Assumptions

In order to evaluate the power consumption and cost of the topologies, two additional items are required: first, a more detailed distribution of switches, transceivers, and fibers; second, some assumptions about the power consumption and cost of each component in the system. For instance, once the number of electronic and optical switches is established, the partial contribution of each type of switch is obtained as the product of its number by the respective power consumption or cost of the device. The total power consumption and cost of the system is obtained as the sum of partial contributions.

Regarding the first point, the equations presented in Section 4.4 provide the total amount of switches, transceivers, and fibers for all topologies. For the power and cost analysis, these equations are partitioned into the individual contributions of each type of component. This allows assigning different power consumption or cost to these contributions. This partitioning is explained graphically in Fig. 5.8 with a number of three-layer networks built with 128-port switches. In terms of switches (left graph), the total number of switches is distributed between electronic and optical switches. All the switches in FT and EFT are electronic (first column). However, in HFT and EHFT topologies there are electronic and optical switches (second through to the last column). The total number of switches depends on the configuration factors, being minimum for case D, in which more ESs are replaced by OSs. In terms of transceivers (center graph), the total number of transceivers is partitioned into the contribution of each type. The example corresponds to a HFT-B topology in the MMS, MSS, and SSS cases. Although the total number of transceivers is the same for all three cases, the distribution of different types of transceivers changes in each case. Finally, in terms of fibers (right graph), the total number of fibers is split into MM and SM. The graph shows also the MMS, MSS, and SSS cases for a HFT-B topology. Note that the switches are color coded in red, the transceivers in green, and the fibers in blue. These color codes are maintained in the next section.

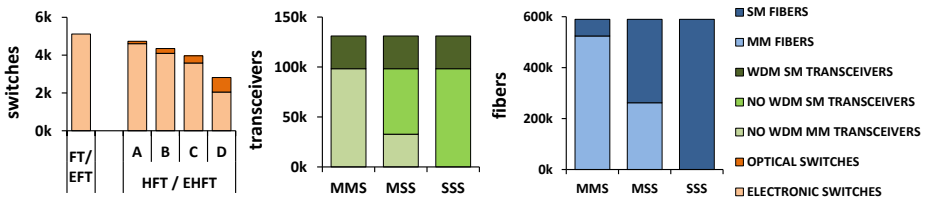

Fig. 5.8 A more detailed distribution of switches, transceivers and fibers.

In relation to the second point, the assumptions taken on the power consumption and cost of each device are summarized in Table 5.3. The values reported in the table should be understood as approximations to real values. Moreover, the numbers may fluctuate with technology evolution and with economy of scale (i.e. bulk purchasing discount) considerations.

Table 5.3 Additional parameters required for power and cost analysis.

Device	**Power (W)**	**Cost (US$)**
128-ports electronic switch	324	3840
128-ports optical switch	128	12800
4-port MMF NO WDM transceiver	3.5	280
4-port SMF NO WDM transceiver	3.5	600
4-port SMF WDM transceiver	3.5	1100
meter of MM fiber	—	0.9
meter of SM fiber	—	0.3
fiber length in layer 1	3 m	
fiber length in layer 2	100 m	
fiber length in layer 3	300 m	

Regarding the 128 x 25G ES, it is assumed a power consumption of 324 W. This value corresponds to the maximum power consumption estimated by [109]. It is distributed as follows: 200 W switching ASIC [52], 84 W fans, 30 W control plane processor, and 10 W for other components. The 112 W power consumption of the 32 QSFP28 (4x25G) transceivers is considered separately. The cost of the ES is calculated estimating US$30/port. Regarding OSs, only optical circuit switches with direct fiber-to-fiber alignment [70] or MEMS-based [68] are considered. Although our analytic model includes all types of OSs, these types of switches are the only ones available at present with at least 128 ports. Much larger amount of ports have been reported [69]. For these OSs, the power consumption considered is 1 W/port and the cost US$100/port. Three types of QSFP28 transceivers are considered, all of them consuming less than 3.5W according to [110]. The MM version without WDM has a 100GBASE-SR4 electrical interface with MPO connector. It reaches up to 100 m in OM4 MMF and it is the cheapest option. The SM transceiver without WDM has a 100G PSM4 electrical interface, an MPO connector, and it reaches up to 2 km in SMF. The more expensive option is the SM transceiver with WDM: it has a 100G CWDM4 interface, an LC connector, and it reaches up to 2 km in SMF. Finally, MMF is three times more expensive than SMF, and the maximum fiber length of every layer is 3 m, 100 m, and 300 m, respectively.

5.4.2 Scaling in power consumption and cost

The scaling of all topologies in terms of power consumption and cost is compared in Fig. 5.9, Fig. 5.10, and Fig. 5.11. The graphs investigate the MMS, MSS, and SSS cases respectively. All networks interconnect 131072 servers and are implemented with three layers of 128-port switches integrating 4-port transceivers. The legend is only included in the graphs of Fig. 5.9: switches contributions are color coded in red, transceivers contributions in green, and fibers contributions in blue.

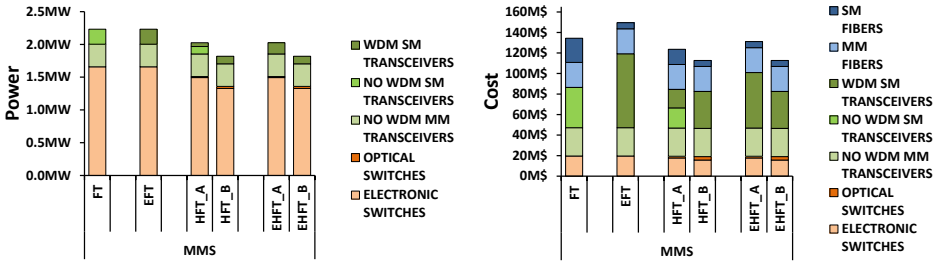

Fig. 5.9 Power consumption and cost per topology in MMS case.

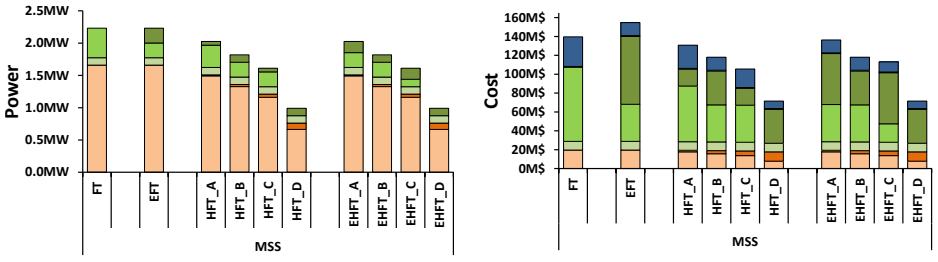

Fig. 5.10 Power consumption and cost per topology in MSS case.

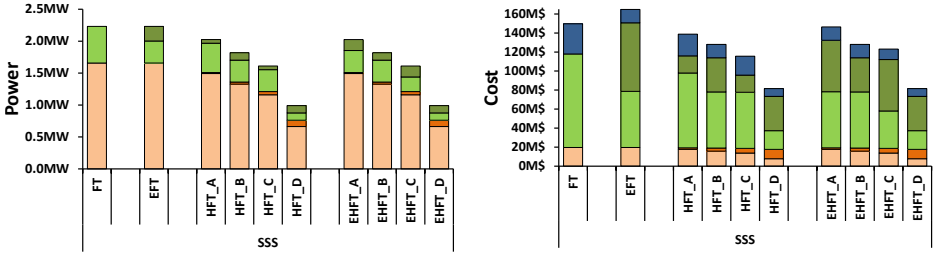

Fig. 5.11 Power consumption and cost per topology in SSS case.

The total power consumption is obtained as the sum of partial contributions of switches and transceivers. ESs clearly dominate over the contributions of OSs and transceivers. The topologies based only on ESs, FT and EFT, consume 2.2 MW in all cases. This is because they have the same number of switches, and according to Table 5.3, all different types of transceivers have equal power consumption. The hybrid topologies, HFT and EHFT, achieve equal savings in power in the respective cases for the same reasons. Comparing electronic and hybrid topologies, case A reduces power consumption by 9.3% to 2.0 MW, case B by 18.5% to 1.8 MW, case C by 27.8% to 1.6 MW, and case D by 55.6% to 991.2 kW. These reductions are explained because hybrid topologies require fewer switches and transceivers, as previously concluded in Section 5.3. Note that the MMS case does not have solutions in the C and D cases, with two hybrid layers, because OSs are SM devices.

The total cost graphs show the contributions of switches, transceivers, and fibers. Cost is dominated by transceivers whereas power consumption is dominated by switches. EFT turns out to be more expensive than FT: despite requiring fewer fibers than FT, the change of the transceivers in the core layer from NO WDM to WDM ends up increasing the costs. For instance, FT costs US$134.4 million, US$139.6 million, US$149.7 million in the MMS, MSS, SSS cases, respectively. EFT increases these costs by 11.2% to US$149.5 million in the MMS case, by 10.8% to US$154.7 million in the MSS case, and by 10.1% to US$164.7 million in the SSS case.

In contrast, hybrid topologies decrease cost in all cases. This reduction is mainly due to the savings in transceivers. In effect, although fewer switches are required by hybrid topologies, the additional cost of OSs ends up by practically canceling these savings. However, hybrid topologies also necessitate fewer transceivers and fibers, which reduces cost. HFT turns out to be the topology achieving larger savings in all cases. For instance, HFT reduces cost by 8.0% to US$123.6 million in case MMS-A, by 16.1% to US$112.8 million in case MMS-B, by 24.4% to US$105.6 million in case MSS-C, and by 48.8% to US$71.5 million in case MSS-D. EHFT achieves equal savings than HFT in cases B and D. However, it is more expensive than HFT in cases A and C because of the WDM transceivers contribution, despite needing fewer fibers.

The relative savings in percentage of power and cost of EFT, HFT, and EHFT topologies compared to their respective FT are summarized in Table 5.4. The baseline is a FT with parameters k = 128-port switches, l = 3 layers, f_P = 1/4, and $f_{E\ NO\ WDM} = 1/4$ (requiring a total of 5120 switches, 163840 4-port transceivers, and 786432 fibers). The conclusions are similar to the graphs.

Table 5.4 Savings (%) in power and cost of EFT, HFT, and EHFT.

Topology	Savings in power (%)			Savings in cost (%)		
	MMS	MSS	SSS	MMS	MSS	SSS
EFT	0.0	0.0	0.0	+11.2	+10.8	+10.1
HFT-A	-9.3	-9.3	-9.3	-8.0	-6.4	-7.2
HFT-B	-18.5	-18.5	-18.5	-16.1	-15.5	-14.5
HFT-C	—	-27.8	-27.8	—	-24.4	-22.8
HFT-D	—	-55.6	-55.6	—	-48.8	-45.5
EHFT-A	-9.3	-9.3	-9.3	-2.4	-2.3	-2.2
EHFT-B	-18.5	-18.5	-18.5	-16.1	-15.5	-14.5
EHFT-C	—	-27.8	-27.8	—	-19.0	-17.7
EHFT-D	—	-55.6	-55.6	—	-48.8	-45.5

5.5 Summary

This chapter has investigated the scaling of FT, EFT, HFT, and EHFT topologies by using the analytic model introduced in the previous chapter. The scaling is studied in terms of devices, power consumption, and cost.

In terms of devices, we conclude that 4-port WDM transceivers available at present reduce the number of fibers in EFT by 25%. Hybrid topologies, including both OSs and WDM technologies, reduce the number of switches by 45%, the transceivers by 60%, and the fibers by 50%. 4-port transceivers already achieve most of the savings, although larger savings are predicted when higher port-density transceivers become available. For instance, with 8-port transceivers, the savings are reduced by 56.25% in the number of switches, by 80% in transceivers, and by 62.50% in fibers. The maximum savings require the availability of 64-port transceivers, which predict reduction by 59.06% in the number of switches, by 97.50% in the number of transceivers, and by 65.63% in the number of fibers.

In terms of power consumption and cost, the real case scenario with 25G technologies available at present, concludes that HFT-D and EHFT-D are the topologies more efficient. They reduce the power consumption by 55.6% and the cost by 48.8%. EFT, despite reducing the number of fibers, turns out to be greater than 10% more expensive due to the extra cost of the WDM transceivers.

Finally, it is worth noting that the analytic model of Chapter 4 is valid for any type of OS. However, it assumes that ESs and OSs have an equal number of ports. As explained in Section 5.4.1, this assumption restricts the type of OS at present to OCSs. Since this type of switches generally provide pure spatial switching, the use

of WDM technologies implies that a group of wavelengths is switched together, and this may restrict the traffic patterns. Chapter 6 explores a solution to recover (at least part of) the lost granularity in the traffic patterns by taking advantage of the ESs already present in hybrid networks.

Chapter 6

Experimental Demonstrator of Hybrid Data Center Networks

6.1 Introduction

Chapter 4 has introduced the general analytic model governing electronic and hybrid Fat-Tree like topologies. A number of examples illustrated how the configuration parameters allow generating multiple architectures taking advantage of WDM and/or OSs. Chapter 5 has investigated the scaling of these networks in terms of devices, power consumption, and cost. It has concluded that large savings are possible in hybrid topologies.

These savings attained by our hybrid topologies rely on an important feature of OSs: they can switch several wavelengths per port, in contrast with ESs that switch only one wavelength per port. As a consequence, fewer OSs are required compared to ESs assuming they have an equal number of ports.

Although our analytic model of hybrid topologies of Chapter 4 is valid for any kind of OS, it also relies on the assumption that ESs and OSs have an equal number of ports. Our real case scenario study of Chapter 5 has pointed out that, at present, this condition is only fulfilled by slow OCSs (such as [68, 70]). Indeed, OCSs are able to switch groups of m wavelengths among n ports, effectively connecting $m \cdot n$ devices with an n port switch. This leads to a reduction in the number of switches, transceivers, and fibers. However, if these groups of m wavelengths are statically assigned to groups of m servers, the traffic patterns are limited and the high-capacity WDM links may be underutilized.

This problem is illustrated in Fig. 6.1. The diagram on the left interconnects 16 servers with a 16-port ES, which allow any communication pattern between the

servers. The diagram on the right also connects 16 servers but employing a 4-port OS and four wavelengths. The system statically assigns one wavelength to each server. As a result, the 16 servers are organized into four groups, each one of them with four servers and four wavelengths. Each one of these groups is connected to one of the OS ports. Note how the solution based on OS and WDM scales better, requiring only a 4-port OS (or similarly, 64 servers can be connected with a 16-port OS and four wavelengths).

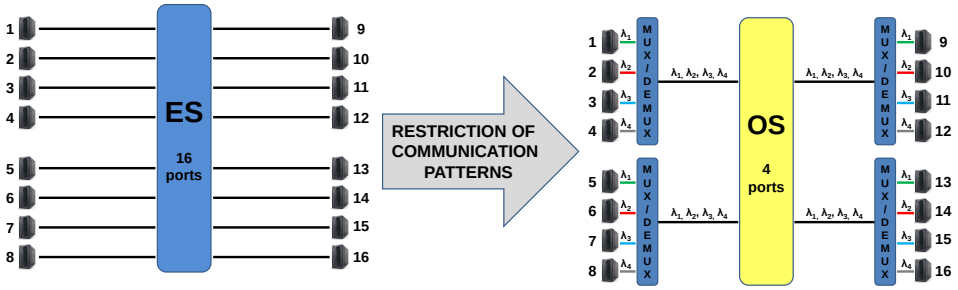

Fig. 6.1 Servers connected with ES or with OS and WDM.

However, the communication patterns are restricted if the OS provides exclusively spatial switching and switches together groups of four wavelengths. The first traffic restriction avoids the communication among servers belonging to the same group (intra-group limitation). Indeed, the servers inside each group have a different wavelength assigned, and thus, cannot communicate with other servers in the same group. As a consequence, the servers of one group must communicate with the servers of another group. The second traffic restriction imposes that the servers can only communicate with other peers having the same assigned wavelength (inter-group limitation). For instance, it is possible to establish connections between servers 1-2-3-4 and servers 5-6-7-8. However, it is not possible to connect servers 1-2-3-4 with servers 8-7-6-5 because they have different wavelengths, or with servers 9-10-15-16 because they belong to different groups.

The solution to these problems of pure spatial switching exploiting WDM links is the dynamic assignment of wavelengths to the links. Several approaches have been suggested, usually presenting a different implementation of a Wavelength Selective Switch (WSS). This adds the wavelength switching plane to the spatial switching plane and provides the granularity at wavelength level. For instance, a common alternative is to build a WSS based on broadcast-and-select networks and SOA arrays [71, 72]. Another option is to build wavelength switching cards with an

optical cross-connect and an array of lasers with fixed wavelengths [111]. Others require Arrayed Waveguide Gratings (AWGs), TWCs, and TLs [59, 73–77].

In this chapter, we demonstrate a different approach to dynamically assign wavelengths to the WDM links by exploiting ESs already present in hybrid data centers. Thus, our approach does not require the introduction of additional and costly WSSs, AWGs, TWCs, or TLs.

We name this technique E-WDM. It is an SDN control technique to dynamically assign the wavelengths of the high-capacity links in hybrid data centers exploiting WDM. It emulates wavelength switching by tight integration of ESs, OCSs, and servers by means of a central SDN controller. As a result, the utilization of the high capacity links is optimized for flexible traffic patterns, overcoming the intra-group and inter-group traffic limitations.

As a proof-of-concept, we built the ECO-IPI Hybrid Data Center demonstrator, which integrated a data center central controller, electronic switches, FOX, and servers. FOX is our SOA-based fast OS operating at 1310 nm. The number of ports of FOX is effectively duplicated by connecting two wavelengths to each port. It constitutes the core optical unit of our 10G ECO-IPI Hybrid Data Center demonstrator and provides 20G WDM links from side-to-side of the ECO-IPI Hybrid Data Center.

The remainder of this chapter is organized as follows. First, the architecture of FOX is presented, together with the configuration options and a visualization of the prototype. Then, a diagram and picture of the ECO-IPI Hybrid Data Center are visualized. Once the hardware is introduced, the software E-WDM technique is explained with a number of examples. Finally, a set of experiments with different traffic patterns demonstrates the feasibility of the approach.

6.2 FOX

FOX, our Fast Optical Circuit Switch, is based on the work done by [100–103]. The data plane, composed of an array of discrete SOA, is practically equivalent. The only difference is that the chosen SOAs have the center wavelength at 1310 nm instead of 1550 nm. This modification is done in order to be able to integrate servers in our hybrid data center setup and transmit traffic between them with WDM. The commercially available QSFP+ CWDM transceivers included in the ECO-IPI Hybrid Data Center demonstrator have wavelengths at 1270 nm, 1290 nm, 1310 nm, and 1330 nm.

The control plane, on the other hand, is substantially different. It operates the data plane as an OCS and not as an Optical Packet Switch (OPS). Our approach

results in a simpler and cheaper implementation of the device (and the system) because it does not need to include special label generation or label processing hardware. However, it also implies that the optical switch is not able to take any more local decisions and relies on a central SDN controller for every reconfiguration event.

A conceptual diagram of FOX is shown in Fig. 6.2. The device is logically divided into three main blocks: the power and cooling block, the data plane block, and the control plane block.

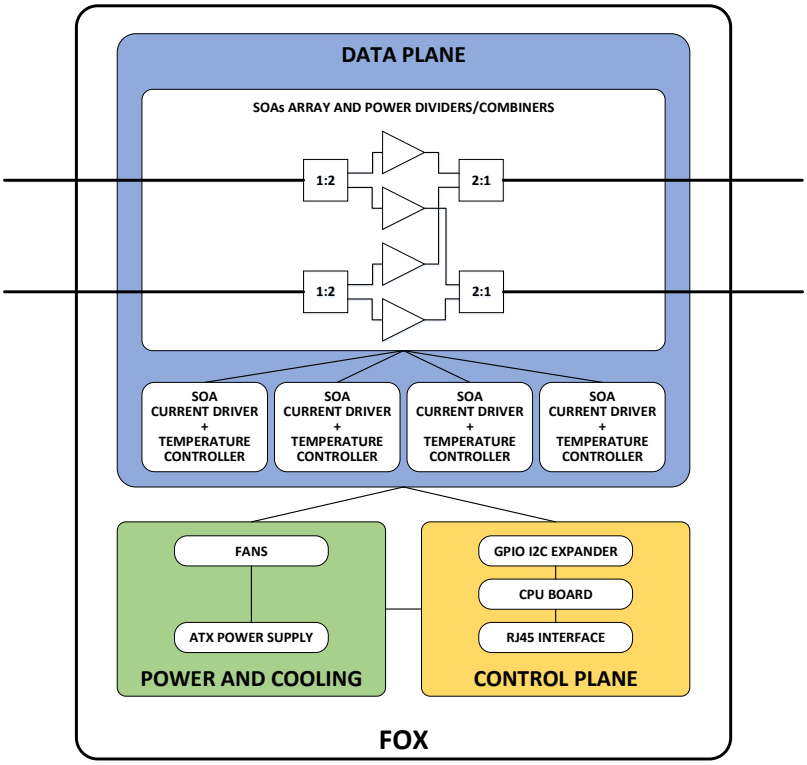

Fig. 6.2 FOX diagram.

The power and cooling block includes an ATX power supply unit that generates the internal voltages (3.3 V, 5 V, and 12 V) from a 220 AC input, and two fans. The data plane provides two optical ports. It consists of an array of four discrete 1310 nm SOAs and four optical power dividers/combiners. Each input port is split into two branches with an optical power divider. Each one of the resulting four branches is fed into one of the SOAs. The outputs of one of the SOAs of each input port are combined again with a power combiner.

By controlling each one of the four SOAs it is possible to decide to which output port is sent the corresponding input port. All possibilities are summarized in Table 6.1. For instance, it is possible to send the input port in1 (or in2) to output port out1, out2, out1 and out2, or to none of them. The configurations where both input ports are sent to the same output port must be avoided to prevent contention. This is represented in the table as "in1+in2 (!)". The worst case happens when all SOAs are active and there is contention in both output ports.

Table 6.1 FOX data plane configuration options.

SOA1	SOA2	SOA3	SOA4	OUT1	OUT2
off	off	off	off	—	—
off	off	off	on	—	in2
off	off	on	off	in2	—
off	off	on	on	in2	in2
off	on	off	off	—	in1
off	on	off	on	—	in1+in2 (!)
off	on	on	off	in2	in1
off	on	on	on	in2	in1+in2 (!)
on	off	off	off	in1	—
on	off	off	on	in1	in2
on	off	on	off	in1+in2 (!)	—
on	off	on	on	in1+in2 (!)	in2
on	on	off	off	in1	in1
on	on	off	on	in1	in1+in2 (!)
on	on	on	off	in1+in2 (!)	in1
on	on	on	on	in1+in2 (!)	in1+in2 (!)

The data plane also includes a current driver and a Thermo Electric Cooler (TEC) controller for each SOA. The current driver allows adjusting the current injected into each SOA, which in turns allows controlling the SOA optical gain. As a result, the optical signal can be amplified by setting adequate values to the SOA current driver and recover the lost optical power in the optical power dividers. An adequate current value is set during the assembly and integration of FOX in the ECO-IPI Hybrid Data Center demonstrator. Later on, it is only necessary to switch the current driver on and off to control the behavior of the SOA.

The TEC controller controls the behaviour of the TEC integrated into each discrete SOA. It cools down the SOA to ensure that it operates in a safe temperature

range, avoiding overheating that will decrease the performance, and eventually, damage the SOA.

The FOX control plane board integrates a Qseven form factor processing unit, integrating an Intel Atom E3845 quad-core processor, 2GB DDR3 DRAM, and 4GB eMMC on board flash. It executes a Unix operating system, that connects to the data center controller through an RJ45 interface using the Internet Protocol (IP) control plane network. Its main function is to command the data plane according to the instructions received from the data center central controller. In order to do that it is connected to the central controller using a socket; when it receives a new configuration command from the central controller, it uses the I2C interface to control a GPIO expander. This expander controls the status of the current driver associated to each SOA, and therefore, its state.

A picture of FOX is shown in Fig. 6.3. The device is packaged in a 19-inch rack unit with a transparent cover. The back plate of the unit gives access to the ATX power supply, two fans, and the RJ45 connector for the control plane network. The ATX power supply generates 12 V for both fans and 5 V / 3.3 V for the data and control plane boards. The front plate gives access to the switch ports. The data plane board, on the left, has the SOA array, the current drivers, and the temperature controllers. The potentiometers shown in the picture allow adjusting the optical gain of the SOAs by injecting the corresponding current. The control plane board, on the right, includes the central processing unit and the mentioned interfaces.

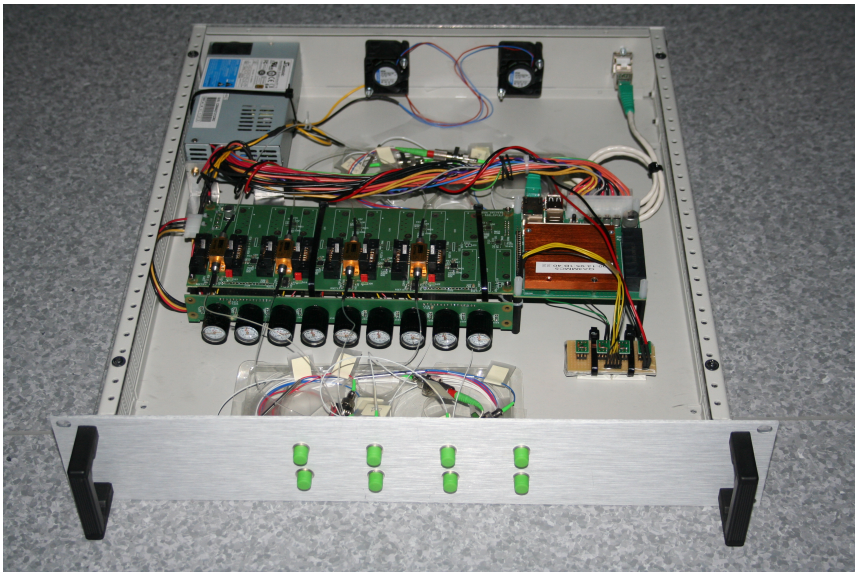

Fig. 6.3 FOX picture.

6.3 ECO-IPI hybrid data center demonstrator

Our ECO-IPI Hybrid Data Center demonstrator integrates a data center central controller, eight servers, six electronic switches, and FOX according to the diagram shown in Fig. 6.4.

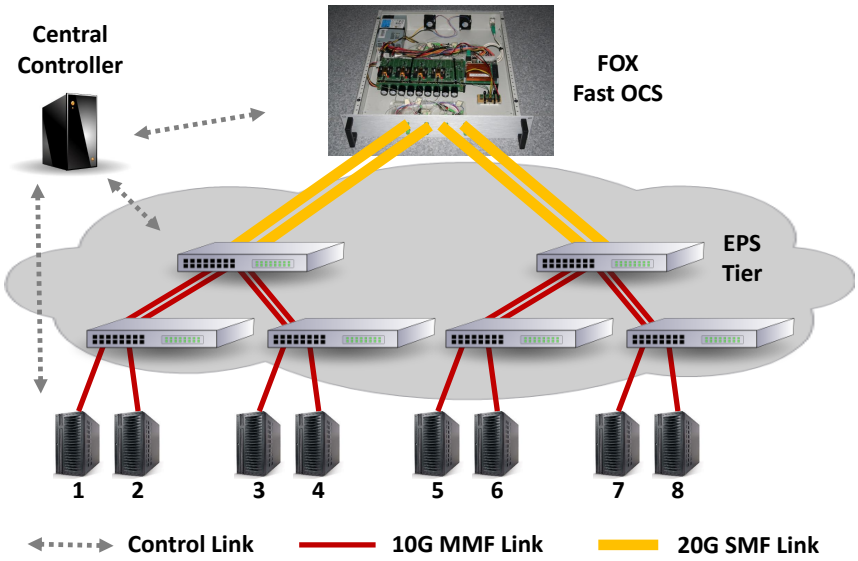

Fig. 6.4 ECO-IPI Hybrid Data Center demonstrator diagram.

The diagram represents the control network with dotted arrows and the data plane network with 10G-red and 20G-yellow links. The control plane network uses copper as transmission medium whereas the data plane network utilizes optical fiber. The 10G links of the data plane network are based on MMF and VCSEL-based MM transceivers; in contrast, the 20G links are based on SMF and SM CWDM transceivers. Both control and data plane networks are IP-based networks and send IP packets. The control plane network connects the data center controller to all servers and switches. The data center controller employs this network to orchestrate the behavior of switches and servers, setting paths and traffic by configuring the devices accordingly. The data plane network transports the data packets between servers, traversing the selected switches depending on the configured path.

The eight servers are (virtually) organized in four racks. Each rack has a top of the rack switch and two servers. The servers have 10G Network Interface Cards (NICs), with Small Form-factor Pluggable - 1x10G (SFP+) MM transceivers working at 850 nm. The edge switches have 72 10G ports accessed by six 120 Gb/s 12x Small Form-factor Pluggable (CXP) transceivers.

The second layer of electronic switches aggregates the traffic of two racks into clusters. As a result, the eight servers of the demonstrator are organized into two clusters with two racks per cluster and two servers per rack. The aggregation switches are also responsible for converting the single-wavelength 10G MM links of the edge switches to double-wavelength 20G SM WDM links. The wavelength conversion from 850 nm to 1310 nm is required because FOX and the CWDM transceivers are SM devices working at 1310 nm.

Finally, the third layer integrates FOX, our optical switch, which constitutes the optical core performing the inter-cluster connections. A more realistic scenario would include more devices (i.e. more servers, racks, electronic and optical switches); however, this modest setup is sufficient to demonstrate the feasibility of the integration of these devices and our E-WDM control technique. Our demonstrator is a fully centralized data center where the SDN central controller orchestrates the behavior of ESs, FOX, and servers.

A picture of the ECO-IPI Hybrid Data Center demonstrator is shown in Fig. 6.5. The left rack accommodates the electronic switches and the data center controller. The right rack includes the servers, a screen for local access to the setup, and (usually) FOX (which lies on the table in the picture).

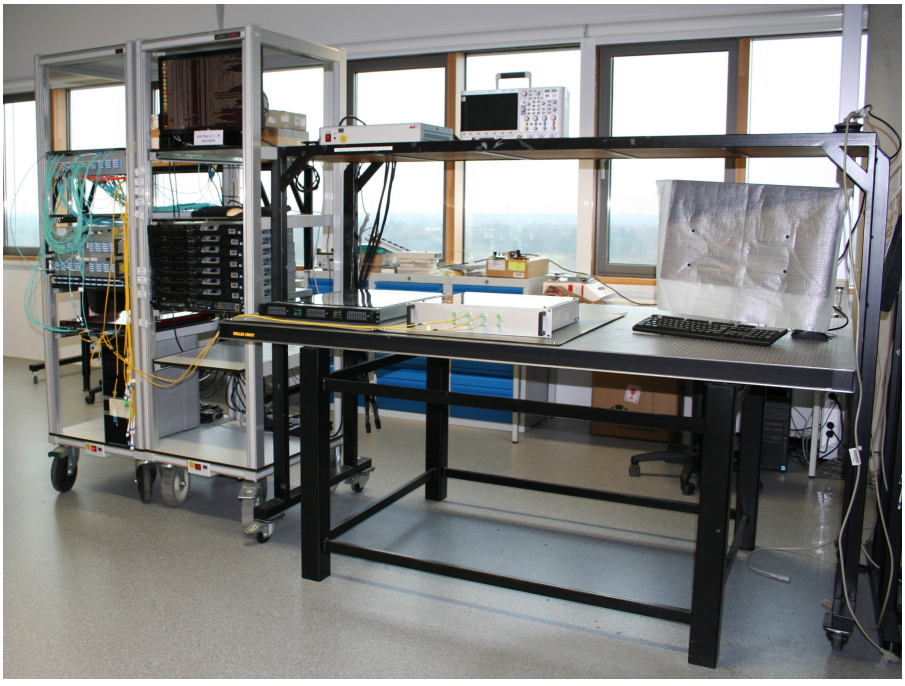

Fig. 6.5 ECO-IPI Hybrid Data Center demonstrator picture.

The data plane connections can be seen in the front section of the racks: blue colored fibers are MM and yellow/orange colored fibers are SM. The control plane connections cannot be seen in the picture because they are in the rear section of the racks, together with the power connections.

6.4 E-WDM technique

E-WDM is a software control technique to dynamically assign wavelengths to high-capacity links in hybrid networks. It is useful for hybrid networks where the optical switches switch together a group of wavelengths per port without providing granularity at wavelength level (i.e. optical switches providing pure spatial switching with WDM links). This is the case, for instance, of OCSs such as [68, 70] switching together several wavelengths per port. As we have seen in Chapter 5, this type of optical switches are the only ones available at present providing a large number of ports and the capacity to switch more than one wavelength per port. As concluded in Chapter 4 and Chapter 5, these two characteristics enable large savings in hybrid topologies. However, these technologies working together may limit the communication patterns as explained in Section 6.1.

The setup to demonstrate the E-WDM technique includes FOX. Our optical switch can switch two wavelengths per port, effectively duplicating the number of servers that can be connected. The two-wavelengths constraint is due to the relatively limited optical bandwidth of the SOAs, and also, to the wide 20 nm spacing of the CWDM channels of our transceivers. In case of having available DWDM transceivers, FOX could switch a larger number of wavelengths.

The main idea of E-WDM consists in installing different OpenFlow rules in the SDN-enabled electronic switches depending on the desired traffic patterns. In this way, certain wavelengths are dynamically assigned to the traffic between two servers. For instance, consider the case where certain output port of one of the aggregation switches is statically assigned wavelength λ_i. Depending on the OpenFlow rules installed on the electronic switches, it is possible to direct the traffic between two servers to that output port, dynamically assigning that wavelength to that traffic. Later on, the OpenFlow rules can be modified so that wavelength is reserved for the traffic between other two servers.

Examples of the E-WDM technique are illustrated in Fig. 6.6, Fig. 6.7, and Fig. 6.8. The eight servers are (logically) organized in four racks: rack A for servers 1-2, rack B for servers 3-4, rack C for servers 5-6, and rack D for servers 7-8. The FOX configuration is maintained without modification in the following examples.

A possible initial assignment of the wavelengths in the high-capacity links of FOX is shown in Fig. 6.6. Servers 1, 3, 5, and 7 are assigned wavelength λ_1, and servers 2, 4, 6, and 8 are assigned wavelength λ_2. This assignment is accomplished by setting the appropriate routing rules in the electronic switches, that direct the traffic to the corresponding port. In this way, the traffic is initially established between servers 1-5, 2-6, 3-7, and 4-8.

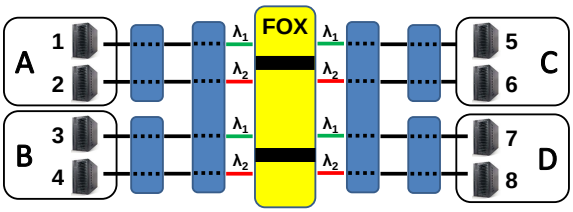

Fig. 6.6 E-WDM example: initial wavelength assignment.

Another case is illustrated in Fig. 6.7. It is desired to change the traffic pattern between the same racks in such a way that servers 1-2 and servers 3-4, connect with servers 6-5 and servers 8-7, respectively. In order to accomplish this goal, the wavelength assignment is modified by setting adequate routing rules in the electronic switches. Servers 2, 4, 5, and 7 are assigned wavelength λ_1, and servers 1, 3, 6 and 8 are assigned wavelength λ_2. In this manner, the traffic is established between servers 1-6, 2-5, 3-8, and 4-7.

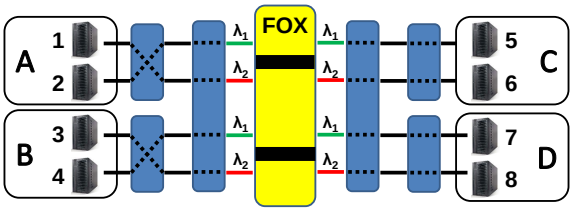

Fig. 6.7 E-WDM example: flexible intra-cluster communication.

In the above two examples, the traffic is established between all servers in one rack and all servers in another rack. With E-WDM, the communication patterns can be established between the desired pairs of servers. Another case is illustrated in Fig. 6.8, where it is desired to establish the traffic between the servers of one rack with the servers of two different racks. It sets the adequate routing rules to assign wavelength λ_1 to servers 1, 2, 5, and 7, and wavelength λ_2 to servers 3, 4, 6, and 8. As a result, the communication is established between servers 1-5, 2-7, 3-6, and 4-8.

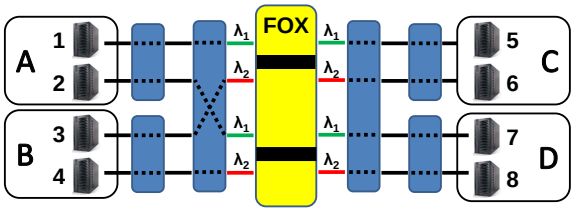

Fig. 6.8 E-WDM example: flexible inter-cluster communication.

An important remark is that all the previous examples have the same optical switch configuration. The different traffic patterns are achieved by changing exclusively the configuration of the electronic switches.

Moreover, the technique is very flexible and allows obtaining equivalent traffic patterns with different configurations in the electronic and/or optical switches because every switch in the path adds a degree of freedom. For instance, eight from all possible configurations of the electronic and/or optical switches achieving the same communication pattern as in Fig. 6.7 are shown in Fig. 6.9. It is possible to configure the input or output edge switches, the input or output aggregation switches, the optical switch, and/or a combination of them.

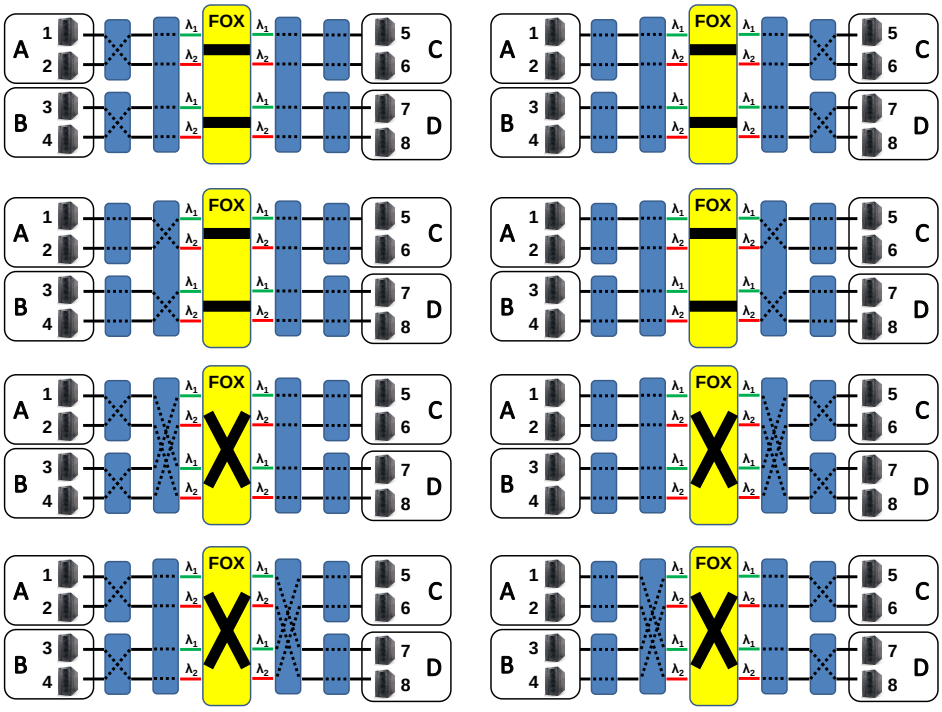

Fig. 6.9 Equivalent traffic patterns with different E-WDM configurations.

One of the challenges of E-WDM is to manage the electronic and optical switches, which are completely different technologies. In one hand, electronic switches are packet-based devices with buffers to store and forward the packets. On the other hand, our optical switch is a circuit-based device without buffers. The lack of buffers of optical switches and the possibility of contention for some of the configuration options makes imperative a proper management of the devices to avoid losing information during reconfiguration events. In our case, we follow a fully centralized approach represented conceptually in Fig. 6.10.

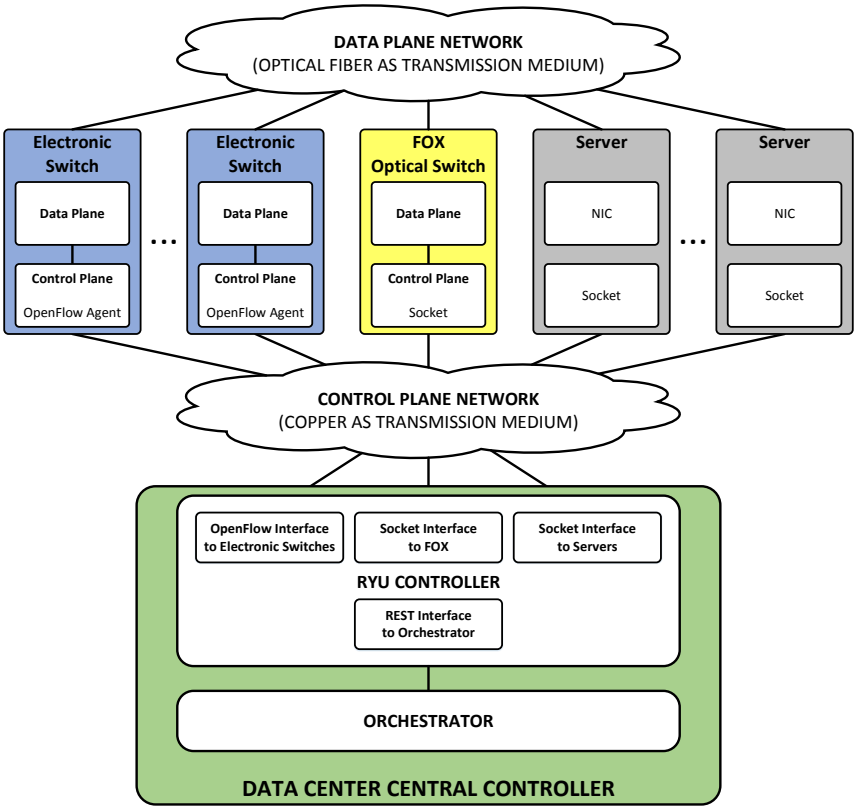

Fig. 6.10 Software architecture of the ECO-IPI Hybrid Data Center.

We take advantage of the unified control network in our system. Although the data plane of electronic and optical switches is radically different, the control plane is very similar. All switches in the network have the same control processing unit, running the same operating system, and therefore, with the same capabilities. The data center controller orchestrates the behavior of servers and switches to ensure that, for each desired traffic pattern, the switches are configured accordingly to

avoid contention and lose data. In other words, the desired path is configured in advance, and once it is set, the data is sent.

The data center controller is connected to all servers and switches through the copper-based control plane network. The orchestrator sends instructions to an SDN Ryu controller using a basic Representational State Transfer (REST) interface. The commands received through the interface are translated into OpenFlow messages for the electronic switches, and to socket messages for the optical switch and servers. Each electronic switch has an OpenFlow agent responsible for receiving the control messages from the orchestrator and transforming them into reconfiguration commands for the data plane switching ASIC. In addition, the socket opened with the optical switch allows sending byte streams to FOX. The control plane processor of FOX transforms these byte streams into I2C commands controlling the SOAs state. Regarding the servers, the socket connections allow creating processes in the servers from the data center controller and collect the corresponding data back.

For example, the data center controller can send some OpenFlow messages to the electronic switches to delete the corresponding flow tables and install the new rules for certain traffic pattern. Then, it can command the optical switch for the corresponding SOA configuration according to 6.1. Once the switches are properly configured and the paths are established, the orchestrator can command the servers to send traffic with *iperf*, or measure the Round Trip Time (RTT) with *ping*.

6.5 E-WDM experimental demonstration

We validate the feasibility of the E-WDM software control technique presented in Section 6.4 by using the ECO-IPI Hybrid Data Center demonstrator introduced in Section 6.3. FOX, deployed at the core layer of the demonstrator, switches together groups of multiple wavelengths per port.

Four different communication patterns among the eight servers are presented in Fig. 6.11. The diagrams only include the end-points of the traffic patterns. However, they should be understood in a similar way to Fig. 6.6, Fig. 6.7, Fig. 6.8, and Fig. 6.9. For instance, the first traffic pattern establishes connections between servers 1-5, 2-6, 3-7, and 4-8; a possible E-WDM configuration is shown in Fig. 6.6. Similarly, the second traffic pattern performs intra-group swapping to communicate servers 1-6, 2-5, 3-8, and 4-7; a possible E-WDM configuration is shown in Fig. 6.7.

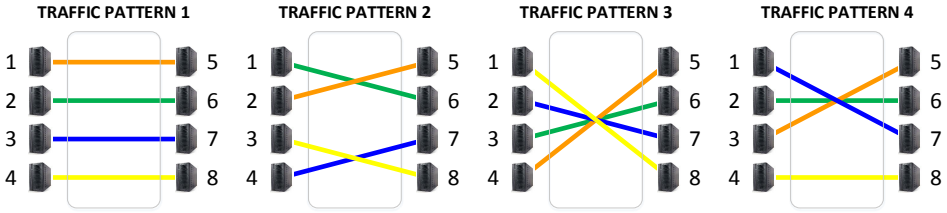

Fig. 6.11 Communication patterns for E-WDM experimental demonstration.

The bandwidth measurements at the servers are reported in Fig. 6.12. For every traffic pattern the throughput between servers is measured with *iperf* and, in all cases, achieve values close to the maximum available with the 10G links.

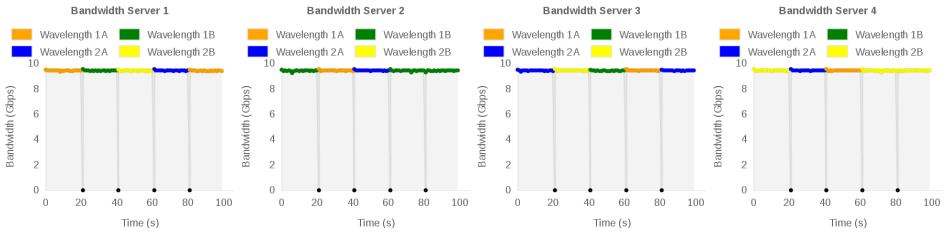

Fig. 6.12 Measured bandwidth during E-WDM experimental demonstration.

The reconfiguration times of electronic switches, FOX, and servers are shown in Fig. 6.13. The configuration of the electronic packet switches takes most of the reconfiguration time. The removal of previous flow rules requires 15 - 20 ms, and the insertion of new flows in the switches around 35 ms. Hence, important savings are achieved if these operations are minimized. The central controller reconfigures FOX in approximately 1.5 ms. The communication is carried through a permanently established socket, where most of the time is needed by the use of I2C as the interface to control the SOAs of FOX. Finally, the servers require less than 15 ms to stop and start the traffic.

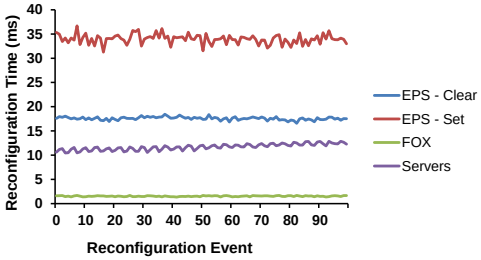

Fig. 6.13 Reconfiguration times of ESs, OSs, and servers.

6.6 Summary

This chapter has presented our optical switch FOX, our ECO-IPI Hybrid Data Center, and E-WDM, a software control technique to dynamically assign the wavelengths to high-capacity links. FOX is our fast SOA-based optical circuit switch, which effectively duplicates the number of servers connected by switching together two wavelengths per port. Our ECO-IPI Hybrid Data Center orchestrates servers, electronic switches, FOX and WDM technologies under the commands of the central controller. E-WDM is meant for hybrid data centers with high-capacity WDM links where the optical switches switch together groups of wavelengths with pure spatial switching (like slow OCSs or our FOX). It is a flexible technique that exploits the electronic switches already present in hybrid data centers to dynamically assign the wavelengths of the WDM links to the traffic of different servers. In this way, the utilization of such links is optimized for flexible communication patterns.

Chapter 7

Summary and Outlook

7.1 Summary

Data centers are the underlying infrastructure supporting the exponential growth of cloud services such as Google, Facebook, and Amazon. The ever-increasing bandwidth demands imply the need for interconnecting more and more servers, which supply the computational and storage capabilities of data centers. However, these complex systems with hundreds of thousands of servers in the largest deployments are expensive to deploy and maintain, making difficult further expansion.

This work has focused on the scaling of data center networks, which are basically a collection of switches providing connectivity among the servers. We have presented a number of solutions improving the scalability of such networks, i.e., solutions helping to interconnect more devices while keeping space, power, and/or cost constraints. We have followed both theoretical and experimental approaches. The first part of this work has investigated electronic switches, the main building block of data center networks deployed at present. The second part has focused on hybrid data center networks, a promising approach to solve the limitations of networks based exclusively on electronic switches by the integration of optical switches.

7.1.1 Compact electronic switches with On-Board Optics

Our theoretical analysis of electronic switches has studied a number of design factors influencing the scaling of such devices and the networks based on them. In order to illustrate the impact of these factors, Fat-Tree topology was selected as an example of data center network based exclusively on electronic switches.

The study has analyzed the impact on the scaling of: 1) the number of ports of the switches, 2) the integration of one or multiple switching ASICs per electronic switch, 3) the size of the devices, and 4) the power of the devices. We concluded that it is relevant to increase the number of ports of the switching ASIC as far as possible since it is the main factor influencing the number of switches and diameter of the networks. We also concluded that it is more efficient to implement the network based on devices including a single switching ASIC than implementing the network with devices including multiple switching ASIC. Finally, we have also illustrated the relevance of reducing size and power consumption of the devices in order to keep scaling data center networks without requiring extra space or power.

Our study showed that electronic switches at present suffer from two important bottlenecks: the switching ASIC bottleneck and the front-panel bottleneck. The ASIC bottleneck is limiting at present the number of ports of the switching ASICs to 128. The front panel bottleneck is limiting at present the number of ports and bandwidth per rack unit to 3.2 Tbps since only 32 QSFP28 optical transceivers fit in the face-plate.

We suggested (and demonstrated) that, in order to overcome these bottlenecks, it is time to: 1) increase the number of devices per rack unit by integrating multiple compact switches, 2) select another type than front-panel optical transceivers allowing tighter integration of switching ASIC and transceivers, such as On-Board Optics devices.

Our experimental demonstrator, based on commercially available components, is an electronic switch requiring only one-fourth of the rack unit space and half the power than similar electronic switches based on front-panel transceivers. The main two design decisions taken during its design and implementation were the integration of On-Board Optics transceivers and the generation of all internal voltages from a single external power supply. Our packaging demonstrator integrates four of these switches in a single rack unit providing four times the number of ports and bandwidth per rack unit compared to front-panel transceivers solutions. Assuming similar results with the servers, a network keeping the same power constraint duplicates the number of servers and requires only half the space.

7.1.2 Hybrid data center networks

Our study on hybrid data center networks has shown that, although important efforts have been done suggesting the introduction of optical switching technologies in data centers, most of them lack an analytic model describing the topologies. Another important limitation found in our research is the assumption that the

hybrid network contains a single optical switch, or similarly, that a single optical switch connects clusters with thousands of devices. This assumption is not very realistic given the large size of the data center networks under consideration, the number of ports in the optical switches, and the number of wavelengths in the transceivers available at present. With these important limitations, it is not possible to accurately evaluate the scaling of hybrid networks in terms of devices (such as switches, transceivers, and fibers), power consumption, and cost. Moreover, the comparison with topologies based on electronic switches is also difficult.

Our research effort in the second part of this work has covered these weaknesses, providing an analytic model accurately defining a number of hybrid topologies and assuming a more realistic scenario where the hybrid networks are deployed with a collection of electronic and optical switches.

The analytic model extends the well-known and widespread fat-tree model to include optical switches and wavelength multiplexing in hybrid fat-tree topologies. The model inherits the scalability and full bisection bandwidth features of fat-tree, and it accurately defines a number of configuration parameters modeling important factors such as the number of ports of the switches or the number of wavelengths per transceiver. With these configuration parameters, it is possible to describe a number of different architectures in a flexible manner. The number of electronic and optical switches, transceivers, and fibers required to build a network with a certain number of servers is computed with a set of equations based on these configuration parameters.

We have employed the analytic model to compare the scaling in devices, power consumption, and cost of electronic and hybrid data center networks. We concluded that, in effect, optical switches and wavelength division multiplexing technologies are promising solutions improving the scalability of data center networks. For instance, with four wavelengths per transceiver and optical switches available today, it is possible to reduce the number of switches by 45%, the number of transceivers by 60%, the number of fibers by 50%, the power consumption by 55.6%, and the cost by 48.8%. On top of that, our model predicts further savings when transceivers with higher port-density become available.

Another relevant insight is that power consumption is dominated by electronic switches whereas cost is dominated by optical transceivers. This points out the relevance of solutions such as the one presented in our electronic switch demonstrator, which integrates On-Board Optics transceivers and cuts by half the power consumption of electronic switches. Also, it signals the importance of including optical switching technologies within the data center networks, which perform the switching function in the optical domain and do not require optical transceivers (except at the edge of the network).

We also concluded that MM technologies, (i.e. MM transceivers and fibers), are mainly interesting for short reach interconnections because in this case the reduced cost of optical transceivers overcomes the larger cost of MM optical fiber. Larger connections are dominated by the cost of the fibers, and our analysis showed that SM technologies are more appropriate because the reduced cost of SM fiber overcomes the extra cost of SM transceivers. Besides, with increasing data rates and size of the data centers, MM technologies with limited reach do not seem a solution for the longest connections at least in the near future.

Finally, we experimentally demonstrated the feasibility of integrating all these technologies in our hybrid data center demonstrator. We implemented FOX, a fast optical circuit switch with an IP control plane and multiple-wavelength ports. FOX is deployed together with electronic switches, servers, WDM links, and a central controller in the ECO-IPI Hybrid Data Center demonstrator. The behaviour of the network is orchestrated by the central controller by means of software defined networking. E-WDM, a software control technique, demonstrated how to overcome the restrictions in communication patterns inherent to networks with pure spatial switching exploiting WDM technologies. The technique enables the dynamic assignment of traffic to different wavelengths in the high-capacity WDM links by leveraging electronic switches present in hybrid networks.

7.2 Outlook

Apart from demonstrating promising solutions improving the scalability of data center networks, our work also opens relevant directions for further research effort.

7.2.1 Future work in electronic switches

Regarding electronic switches, we have demonstrated with 10G devices that it is possible to multiply by four the port and bandwidth density per rack unit. During these years of investigation, switching ASIC and transceivers at 25G became available. It would be relevant to demonstrate that similar results are possible with 25G technologies, despite the increased power consumption introducing additional challenges. From our experience, this integration should be possible since these 25G devices have the same form factor as the 10G devices integrated into our demonstrator and On-Board Optics transceivers have excellent performance in terms of heat dissipation. This direction will also confirm if the shorter traces result of the tighter integration of On-Board Optics transceivers and the switching ASIC allow disabling the CDR circuits typically required at these data rates.

During these years, also a switching ASIC with 128 50G (2 x 25G) interfaces has become available. It requires at present two rack units in order to package 64 QSFP28 front-panel optical transceivers and it will be needed to wait until QSFP-DD transceivers become available to package them in a single rack unit. It would be relevant to demonstrate that it is possible to package such ASIC in a single rack unit with 25G On-Board Optics transceivers available at present. From our experience, it should be possible to package two of these devices per rack unit assuming that similar results to our demonstrator can be achieved at 25G.

Looking further ahead, we suggest here another two paths especially interesting to continue our research.

First, how can the number of compact switches per rack unit be pushed forward and how far? We have demonstrated that it is possible to integrate four devices per rack unit with On-Board Optics transceivers available at present. Tighter integration of On-Board Optics devices or other approaches such as On-Top-of-ASIC transceivers will help to achieve higher densities. In principle, a rack unit could scale up to 80 MPO-72 connectors with a total of 2880 ports, resulting in bandwidths per rack unit of 72 Tbps at 25G, 144 Tbps at 50G, and beyond.

Second, how can the current ASIC bottleneck, limited at present to 128 ports, be overcome to produce 256-port devices? In effect, ASICs are still limited to 128 logical ports despite having already 256 physical interfaces; e.g. 256 25G physical interfaces results in 128 logical 50G ports. To answer this question it is needed to change the role of switch designers to the role of ASIC designers. Probably, further improvements in microelectronic integration and design optimization of the switching ASIC with this goal in mind would be required.

7.2.2 Future work in hybrid data center networks

Regarding hybrid networks integrating optical switching and wavelength division multiplexing technologies, there are also many challenges still to be solved.

Regarding our analytic model, it would be relevant to work in two different directions: generalization to other topologies and performance analysis. Regarding the first point, our analytic model includes Fat-Tree like topologies and it would be interesting to extend it to include other hybrid topologies enabling a more general comparative study. The configuration parameters defined in our model, describing features such as the number of ports of the switches or the number of wavelengths per transceivers, are valid for other hybrid topologies. However, the derivation of new equations for other hybrid topologies based on these configuration parameters is still required. Regarding the second point, our model allows comparing

topologies in terms of devices, i.e., in terms of switches, transceivers and fibers. This enables the investigation of the scaling of the networks in terms of power consumption and cost. However, the comparison in terms of performance of such networks is still missing. Thus, it would be interesting to extend the model to include the performance analysis of the networks. This is not trivial task because the model should account, among other, for different types of optical switching technologies, traffic, and control network approaches. A number of attempts have already been carried out; however, they typically provide solutions restricted to a certain architecture and lack a general approach enabling the comparison of different topologies.

Regarding our real case scenario study, it considers exclusively slow optical circuit switches, since they are the only ones available at present with a large number of ports. The reconfiguration time of these type of switches implies important limitations in the traffic patterns and performance of the networks. Thus, it would be interesting to explore how to design and implement fast optical switches with a large number of ports. To our understanding, this is the main limiting factor stopping the introduction of optical switching technologies into data centers.

Finally, our hybrid data center demonstrator has pointed out that control networks play a critical role in such networks. We have demonstrated that the integration of different technologies such electronic packet switches and optical circuit switches is possible under the commands of a centralized SDN controller. However, further investigation to implement fast control networks will be key to ensure adequate performance and scalability of such networks.

References

[1] L. A. Barroso, J. Clidaras, and U. Hölzle, *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Second Edition*, ser. Synthesis Lectures on Computer Architecture. Morgan & Claypool Publishers, 2013. [Online]. Available: https://doi.org/10.2200/S00516ED2V01Y201306CAC024

[2] P. M. Mell and T. Grance, "SP 800-145. the NIST Definition of Cloud Computing," Gaithersburg, MD, United States, Tech. Rep., 2011. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

[3] Cisco Networks White Paper. (2016) Cisco Global Cloud Index: Forecast and Methodology, 2015–2020. [Online]. Available: https://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf

[4] M. F. Bari, R. Boutaba, R. Esteves, L. Z. Granville, M. Podlesny, M. G. Rabbani, Q. Zhang, and M. F. Zhani, "Data Center Network Virtualization: A Survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 2, pp. 909–928, 2013.

[5] H. Kim and N. Feamster, "Improving Network Management with Software Defined Networking," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 114–119, 2013.

[6] D. Cohen, F. Petrini, M. D. Day, M. Ben-Yehuda, S. W. Hunter, and U. Cummings, "Applying Amdahl's Other Law to the Data Center," *IBM J. Res. Dev.*, vol. 53, no. 5, pp. 683–694, 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1850670.1850675

[7] P. Kogge, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R. S. Williams, and K. Yelick, "ExaScale Computing Study: Technology Challenges in Achieving ExaScale Systems," DARPA IPTO, Air Force Research Labs, Tech. Rep., 2008. [Online]. Available: http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf

[8] A. Greenberg, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Towards a Next Generation Data Center Architecture: Scalability and Commoditization," in *Proceedings of the ACM Workshop on Programmable Routers for Extensible Services of Tomorrow*, ser. PRESTO '08. New York, NY, USA: ACM, 2008, pp. 57–62. [Online]. Available: http://doi.acm.org/10.1145/1397718.1397732

[9] D. Abts and J. Kim, *High Performance Datacenter Networks: Architectures, Algorithms, and Opportunities*, San Rafael, California, 2011. [Online]. Available: http://dx.doi.org/10.2200/S00341ED1V01Y201103CAC014

[10] A. Hammadi and L. Mhamdi, "Review: A Survey on Architectures and Energy Efficiency in Data Center Networks," *Comput. Commun.*, vol. 40, pp. 1–21, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.comcom.2013.11.005

[11] A. Ghiasi, "Large Data Centers Interconnect Bottlenecks," *Opt. Express*, vol. 23, no. 3, pp. 2085–2090, 2015. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-23-3-2085

[12] C. D. Patel, R. Sharma, C. E. Bash, and A. Beitelmal, "Thermal Considerations in Cooling Large Scale High Compute Density Data Centers," in *ITherm 2002. Eighth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (Cat. No.02CH37258)*, 2002, pp. 767–776.

[13] B. Heller, D. Erickson, N. McKeown, R. Griffith, I. Ganichev, S. Whyte, K. Zarifis, D. Moon, S. Shenker, and S. Stuart, "Ripcord: a Modular Platform for Data Center Networking," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 457–458, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=2043164.1851261

[14] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "VL2: A Scalable and Flexible Data Center Network," *Commun. ACM*, vol. 54, no. 3, pp. 95–104, 2011. [Online]. Available: http://doi.acm.org/10.1145/1897852.1897877

[15] A. Tavakoli, M. Casado, T. Koponen, and S. Shenker, "Applying NOX to the Datacenter." in *Proc. of Workshop on Hot Topics in Networks (HotNets-VIII)*, L. Subramanian, W. E. Leland, and R. Mahajan, Eds. ACM SIGCOMM, 2009. [Online]. Available: http://conferences.sigcomm.org/hotnets/2009/papers/hotnets2009-final103.pdf

[16] OpenCompute - Server. (2018) OpenCompute Server - Specs and Designs. [Online]. Available: http://www.opencompute.org/wiki/Server/SpecsAndDesigns

[17] R. Luijten, D. Pham, R. Clauberg, M. Cossale, H. N. Nguyen, and M. Pandya, "Energy-Efficient Microserver Based on a 12-core 1.8GHz 188K-CoreMark 28nm Bulk CMOS 64b SoC for Big-Data Applications

with 159GB/S/L Memory Bandwidth System Density," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, 2015, pp. 1–3.

[18] H. Liu, R. Urata, and A. Vahdat, *Optical Interconnects for Scale-Out Data Centers*. New York, NY: Springer New York, 2013, pp. 17–29. [Online]. Available: https://doi.org/10.1007/978-1-4614-4630-9_2

[19] H. Liu, C. F. Lam, and C. Johnson, "Scaling Optical Interconnects in Datacenter Networks Opportunities and Challenges for WDM," in *18th IEEE Symposium on High Performance Interconnects*, 2010, pp. 113–116.

[20] R. E. Freund, C.-A. Bunge, N. N. Ledentsov, D. Molin, and C. Caspar, "High-Speed Transmission in Multimode Fibers," *J. Lightwave Technol.*, vol. 28, no. 4, pp. 569–586, 2010. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-28-4-569

[21] E. Haglund, *VCSELs for High-Speed, Long-Reach, and Wavelength-Multiplexed Optical Interconnects*, ser. Doktorsavhandlingar vid Chalmers tekniska högskola. Ny serie, no:. Department of Microtechnology and Nanoscience, Photonics, Chalmers University of Technology,, 2015, 180.

[22] J. M. Castro, R. Pimpinella, B. Kose, and B. Lane, "Investigation of the Interaction of Modal and Chromatic Dispersion in VCSEL–MMF Channels," *J. Lightwave Technol.*, vol. 30, no. 15, pp. 2532–2541, 2012. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-30-15-2532

[23] J. A. Tatum, "Evolution of VCSELs," *Proc. SPIE: Vertical-Cavity Surface-Emitting Lasers XVIII*, vol. 9001, pp. 9001 – 9001 – 9, 2014. [Online]. Available: http://dx.doi.org/10.1117/12.2037387

[24] M. A. Taubenblatt, "Optical Interconnects for High-Performance Computing," *J. Lightwave Technol.*, vol. 30, no. 4, pp. 448–457, 2012. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-30-4-448

[25] J. F. Bulzacchelli, C. Menolfi, T. J. Beukema, D. W. Storaska, J. Hertle, D. R. Hanson, P. H. Hsieh, S. V. Rylov, D. Furrer, D. Gardellini, A. Prati, T. Morf, V. Sharma, R. Kelkar, H. A. Ainspan, W. R. Kelly, L. R. Chieco, G. A. Ritter, J. A. Sorice, J. D. Garlett, R. Callan, M. Brandli, P. Buchmann, M. Kossel, T. Toifl, and D. J. Friedman, "A 28-Gb/s 4-Tap FFE/15-Tap DFE Serial Link Transceiver in 32-nm SOI CMOS Technology," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 12, pp. 3232–3248, 2012.

[26] E. Haglund, A. Haglund, P. Westbergh, J. S. Gustavsson, B. Kogel, and A. Larsson, "25 Gbit/s Transmission over 500 m Multimode Fibre Using 850 nm VCSEL with Integrated Mode Filter," *Electronics Letters*, vol. 48, no. 9, pp. 517–519, 2012.

[27] A. Ghiasi, F. Tang, and S. Bhoja, "Measurements Results of 25.78 GBd Over OM3 with and without Equalization," *Presented to IEEE 802.3 Standards Committee*, 2012. [Online]. Available: http://www.ieee802.org/3/100GNGOPTX/public/mar12/plenary/ghiasi_01_0312_NG100GOPTX.pdf

[28] K. Szczerba, P. Westbergh, M. Karlsson, P. A. Andrekson, and A. Larsson, "60 Gbits error-free 4-PAM operation with 850 nm VCSEL," *Electronics Letters*, vol. 49, no. 15, pp. 953–955, 2013.

[29] B. G. Lee, D. M. Kuchta, F. E. Doany, C. L. Schow, P. Pepeljugoski, C. Baks, T. F. Taunay, B. Zhu, M. F. Yan, G. E. Oulundsen, D. S. Vaidya, W. Luo, and N. Li, "End-to-End Multicore Multimode Fiber Optic Link Operating up to 120 Gb/s," *J. Lightwave Technol.*, vol. 30, no. 6, pp. 886–892, 2012. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-30-6-886

[30] SWDM Alliance. (2018). [Online]. Available: http://www.swdm.org/

[31] J. A. Tatum, D. Gazula, L. A. Graham, J. K. Guenter, R. H. Johnson, J. King, C. Kocot, G. D. Landry, I. Lyubomirsky, A. N. MacInnes, E. M. Shaw, K. Balemarthy, R. Shubochkin, D. Vaidya, M. Yan, and F. Tang, "VCSEL-Based Interconnects for Current and Future Data Centers," *J. Lightwave Technol.*, vol. 33, no. 4, pp. 727–732, 2015. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-33-4-727

[32] D. Kuchta, T. Huynh, F. Doany, A. Rylyakov, C. Schow, P. Pepeljugoski, D. Gazula, E. Shaw, and J. Tatum, "A 4-λ, 40Gb/s/λ Bandwidth Extension of Multimode Fiber in the 850nm Range," in *Optical Fiber Communication Conference*. Optical Society of America, 2015, p. W1D.4. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2015-W1D.4

[33] R. Baca, P. Kolesar, J. Tatum, D. Gazula, E. Shaw, and T. Gray, "Advances in Multimode Fiber Transmission for the Data Center," in *Optical Fiber Communication Conference*. Optical Society of America, 2015, p. W2A.6. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2015-W2A.6

[34] I. Lyubomirsky, R. Motaghian, H. Daghighian, D. McMahon, S. Nelson, C. Kocot, J. A. Tatum, F. Achten, P. Sillard, D. Molin, and A. Amezcua-Correa, "100G SWDM4 Transmission over 300m Wideband MMF," in *2015 European Conference on Optical Communication (ECOC)*, 2015, pp. 1–3.

[35] C. Kocot, R. Motaghian, anna Tatarczak, I. Lyubomirsky, S. Hallstein, D. Askarov, H. Daghighian, S. Nelson, and J. Tatum, "SWDM Strategies to Extend Performance of VCSELs over MMF," in *Optical Fiber Communication Conference*. Optical Society of America, 2016, p. Tu2G.1. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2016-Tu2G.1

[36] QSFP-DD Multi-Source Agreement. (2018) QSFP-DD Specification. [Online]. Available: http://www.qsfp-dd.com/

[37] CFP Multi-Source Agreement. (2018) CFP Specification. [Online]. Available: http://www.cfp-msa.org/

[38] A. Ghiasi, "Is There a Need for On-Chip Photonic Integration for Large Data Warehouse Switches," in *The 9th International Conference on Group IV Photonics (GFP)*, 2012, pp. 27–29.

[39] T. Sugimoto, Y. Hashimoto, K. Yamamoto, M. Kurihara, M. Oda, J. Sakai, H. Ono, T. Akagawa, K. Yashiki, H. Hatayama, N. Suzuki, M. Tsuji, I. Ogura, H. Kouta, and K. Kurata, "12-Channel x 20-Gbps On-board Parallel Optical Modules Using Multi-Chip Visual Alignment Technique," in *2010 Proceedings 60th Electronic Components and Technology Conference (ECTC)*, 2010, pp. 256–262.

[40] J.-M. Verdiell, "Advances in Onboard Optical Interconnects: A New Generation of Miniature Optical Engines," *SAMTEC White Paper Presented in DesignCon*, 2013. [Online]. Available: http://suddendocs.samtec.com/notesandwhitepapers/designcon-west-2013-12-ta4_advancesinonboardopticalinterconnects.pdf

[41] FIT - Foxconn Interconnect Technology. (2018) MiniPOD/MicroPOD Embedded Parallel Optics. [Online]. Available: http://www.fit-foxconn.com/Product/SearchByFamily?topClassID=Fiber%20Optics&ProductClass=Fiber%20Optics&ProductFamily=Embedded%20Optical%20Modules

[42] Finisar. (2018) Finisar - Optical Engines. [Online]. Available: https://www.finisar.com/optical-engines

[43] J. Duis and T. Hultermans, "The Cool Future of Optics "CoolBit"," in *2014 The European Conference on Optical Communication (ECOC)*, 2014, pp. 1–3.

[44] TE Connectivity, "End-to-End Communications with Advanced Fiber Optic Technologies," *White Paper Submitted to OSA Enabled By Optics Event*, 2014. [Online]. Available: https://www.osa.org/osaorg/media/osa.media/CorporateGateway/EnabledByOptics/TEConnectivityCoolbitWhitepaper-OSAEnabledbyOptics.pdf

[45] C. L. Schow, F. E. Doany, A. V. Rylyakov, B. G. Lee, C. V. Jahnes, Y. H. Kwark, C. W. Baks, D. M. Kuchta, and J. A. Kash, "A 24-Channel, 300 Gb/s, 8.2 pJ/bit, Full-Duplex Fiber-Coupled Optical Transceiver Module Based on a Single "Holey" CMOS IC," *J. Lightwave Technol.*, vol. 29, no. 4, pp. 542–553, 2011. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-29-4-542

[46] F. E. Doany, B. G. Lee, C. L. Schow, C. K. Tsang, C. Baks, Y. Kwark, R. John, J. U. Knickerbocker, and J. A. Kash, "Terabit/s-Class 24-Channel Bidirectional Optical Transceiver Module Based on TSV Si Carrier for Board-Level Interconnects," in *2010 Proceedings 60th Electronic Components and Technology Conference (ECTC)*, 2010, pp. 58–65.

[47] F. E. Doany, C. L. Schow, B. G. Lee, R. A. Budd, C. W. Baks, C. K. Tsang, J. U. Knickerbocker, R. Dangel, B. Chan, H. Lin, C. Carver, J. Huang, J. Berry, D. Bajkowski, F. Libsch, and J. A. Kash, "Terabit/s-Class Optical PCB Links Incorporating 360-Gb/s Bidirectional 850 nm Parallel Optical Transceivers," *J. Lightwave Technol.*, vol. 30, no. 4, pp. 560–571, 2012. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-30-4-560

[48] A. Benner, D. M. Kuchta, P. K. Pepeljugoski, R. A. Budd, G. Hougham, B. V. Fasano, K. Marston, H. Bagheri, E. J. Seminaro, H. Xu, D. Meadowcroft, M. H. Fields, L. McColloch, M. Robinson, F. W. Miller, R. Kaneshiro, R. Granger, D. Childers, and E. Childers, "Optics for High-Performance Servers and Supercomputers," in *Optical Fiber Communication Conference*. Optical Society of America, 2010, p. OTuH1. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2010-OTuH1

[49] COBO - Consortium for On-Board Optics. (2018). [Online]. Available: http://onboardoptics.org/

[50] Gazettabyte. (2018) "COBO Targets Year-End to Complete Specification". [Online]. Available: http://www.gazettabyte.com/home/2017/8/29/cobo-targets-year-end-to-complete-specification.html

[51] Broadcom - TridentII. (2018) BCM56850 Series. [Online]. Available: https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm56850-series

[52] Broadcom - Tomahawk. (2018) BCM56960 Series. [Online]. Available: https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm56960-series

[53] Broadcom - TomahawkII. (2018) BCM56970 Series. [Online]. Available: https://www.broadcom.com/news/product-releases/broadcom-first-to-deliver-64-ports-of-100ge-with-tomahawk-ii-ethernet-switch

[54] Innovium. (2018) Teralynx Products. [Online]. Available: http://www.innovium.com/products/teralynx/

[55] OpenCompute - Networking. (2018) OpenCompute Networking - Specs and Designs. [Online]. Available: http://www.opencompute.org/wiki/Networking/SpecsAndDesigns

[56] M. Al-Fares, A. Loukissas, and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 63–74, 2008. [Online]. Available: http://doi.acm.org/10.1145/1402946.1402967

[57] A. Wonfor, H. Wang, R. V. Penty, and I. H. White, "Large Port Count High-Speed Optical Switch Fabric for Use Within Datacenters," *J. Opt. Commun. Netw.*, vol. 3, no. 8, pp. A32–A39, 2011. [Online]. Available: http://jocn.osa.org/abstract.cfm?URI=jocn-3-8-A32

[58] X. Ye, R. Proietti, Y. Yin, S. J. B. Yoo, and V. Akella, "Buffering and Flow Control in Optical Switches for High Performance Computing," *J. Opt. Commun. Netw.*, vol. 3, no. 8, pp. A59–A72, 2011. [Online]. Available: http://jocn.osa.org/abstract.cfm?URI=jocn-3-8-A59

[59] K. Xi, Y.-H. Kao, and H. J. Chao, *"A Petabit Bufferless Optical Switch for Data Center Networks"*. New York, NY: Springer New York, 2013, pp. 135–154. [Online]. Available: https://doi.org/10.1007/978-1-4614-4630-9_8

[60] H. Liu, F. Lu, A. Forencich, R. Kapoor, M. Tewari, G. M. Voelker, G. Papen, A. C. Snoeren, and G. Porter, "Circuit Switching Under the Radar with REACToR," in *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'14. Berkeley, CA, USA: USENIX Association, 2014, pp. 1–15. [Online]. Available: http://dl.acm.org/citation.cfm?id=2616448.2616450

[61] Z. Zhang and Y. Yang, "Performance Analysis of Optical Packet Switches Enhanced with Electronic Buffering," in *2009 IEEE International Symposium on Parallel Distributed Processing*, May 2009, pp. 1–9.

[62] L. Liu, Z. Zhang, and Y. Yang, "In-Order Packet Scheduling in Optical Switch With Wavelength Division Multiplexing and Electronic Buffer," *IEEE Transactions on Communications*, vol. 62, no. 6, pp. 1983–1994, 2014.

[63] L. Liu and Y. Yang, "Optimal Packet Scheduling in WDM Optical Switches With Output Buffer and Limited Wavelength Conversion," *J. Lightwave Technol.*, vol. 29, no. 9, pp. 1227–1238, 2011. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-29-9-1227

[64] G. S. Zervas, M. D. Leenheer, L. Sadeghioon, D. Klonidis, Y. Qin, R. Nejabati, D. Simeonidou, C. Develder, B. Dhoedt, and P. Demeester, "Multi-Granular Optical Cross-Connect: Design, Analysis, and Demonstration," *J. Opt. Commun. Netw.*, vol. 1, no. 1, pp. 69–84, 2009. [Online]. Available: http://jocn.osa.org/abstract.cfm?URI=jocn-1-1-69

[65] S. J. B. Yoo, "Optical Packet and Burst Switching Technologies for the Future Photonic Internet," *J. Lightwave Technol.*, vol. 24, no. 12, pp. 4468–4492, 2006. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-24-12-4468

[66] M. J. O'Mahony, D. Simeonidou, D. K. Hunter, and A. Tzanakaki, "The Application of Optical Packet Switching in Future Communication Networks," *Comm. Mag.*, vol. 39, no. 3, pp. 128–135, 2001. [Online]. Available: http://dx.doi.org/10.1109/35.910600

[67] M. Imran, M. Collier, P. Landais, and K. Katrinis, "HOSA: Hybrid Optical Switch Architecture for Data Center Networks," in *Proceedings of the 12th ACM International Conference on Computing Frontiers*, ser. CF '15. New York, NY, USA: ACM, 2015, pp. 27:1–27:8. [Online]. Available: http://doi.acm.org/10.1145/2742854.2742877

[68] Calient. (2018) Series S. [Online]. Available: http://www.calient.net/products/s-series-photonic-switch/

[69] J. Kim, C. J. Nuzman, B. Kumar, D. F. Lieuwen, J. S. Kraus, A. Weiss, C. P. Lichtenwalner, A. R. Papazian, R. E. Frahm, N. R. Basavanhally, D. A. Ramsey, V. A. Aksyuk, F. Pardo, M. E. Simon, V. Lifton, H. B. Chan, M. Haueis, A. Gasparyan, H. R. Shea, S. Arney, C. A. Bolle, P. R. Kolodner, R. Ryf, D. T. Neilson, and J. V. Gates, "1100 x 1100 Port MEMS-Based Optical Crossconnect with 4-dB Maximum Loss," *IEEE Photonics Technology Letters*, vol. 15, no. 11, pp. 1537–1539, 2003.

[70] Polatis. (2018) Series 7000. [Online]. Available: http://www.polatis.com/series-7000-384x384-port-software-controlled-optical-circuit-switch-sdn-enabled.asp

[71] T. Lin, K. A. Williams, R. V. Penty, I. H. White, and M. Glick, "Capacity Scaling in a Multihost Wavelength-Striped SOA-Based Switch Fabric," *J. Lightwave Technol.*, vol. 25, no. 3, pp. 655–663, 2007. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-25-3-655

[72] R. Hemenway, R. Grzybowski, C. Minkenberg, and R. Luijten, "Optical-Packet-Switched Interconnect for Supercomputer Applications," *J. Opt. Netw.*, vol. 3, no. 12, pp. 900–913, 2004. [Online]. Available: http://jon.osa.org/abstract.cfm?URI=jon-3-12-900

[73] K. i. Sato, H. Hasegawa, T. Niwa, and T. Watanabe, "A Large-Scale Wavelength Routing Optical Switch for Data Center Networks," *IEEE Communications Magazine*, vol. 51, no. 9, pp. 46–52, 2013.

[74] C.-T. Lea, "A Scalable AWGR-Based Optical Switch," *J. Lightwave Technol.*, vol. 33, no. 22, pp. 4612–4621, 2015. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-33-22-4612

[75] R. Proietti, Y. Yin, R. Yu, C. J. Nitta, V. Akella, C. Mineo, and S. J. B. Yoo, "Scalable Optical Interconnect Architecture Using AWGR-Based TONAK LION Switch with Limited Number of Wavelengths," *J. Lightwave Technol.*, vol. 31, no. 24, pp. 4087–4097, 2013. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-31-24-4087

[76] Z. Zhang and Y. Yang, "NEO: A Nonblocking Hybrid Switch Architecture for Large Scale Data Centers," in *2014 43rd International Conference on Parallel Processing*, 2014, pp. 510–519.

[77] X. Ye, Y. Yin, S. J. B. Yoo, P. Mejia, R. Proietti, and V. Akella, "DOS: A Scalable Optical Switch for Datacenters," in *Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, ser. ANCS '10. New York, NY, USA: ACM, 2010, pp. 24:1–24:12. [Online]. Available: http://doi.acm.org/10.1145/1872007.1872037

[78] L. Chen, A. Sohdi, J. E. Bowers, L. Theogarajan, J. Roth, and G. Fish, "Electronic and Photonic Integrated Circuits for Fast Data Center Optical Circuit Switches," *IEEE Communications Magazine*, vol. 51, no. 9, pp. 53–59, 2013.

[79] D. Nikolova, S. Rumley, D. Calhoun, Q. Li, R. Hendry, P. Samadi, and K. Bergman, "Scaling Silicon Photonic Switch Fabrics for Data Center Interconnection Networks," *Opt. Express*, vol. 23, no. 2, pp. 1159–1175, 2015. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-23-2-1159

[80] J. E. Bowers and E. Hall, "Silicon Photonic Switching for Data Center Applications," in *IEEE Winter Topicals 2011*, 2011, pp. 129–130.

[81] 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.

[82] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, "Dcell: A Scalable and Fault-tolerant Network Structure for Data Centers," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 75–86, 2008. [Online]. Available: http://doi.acm.org/10.1145/1402946.1402968

[83] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 63–74, 2009. [Online]. Available: http://doi.acm.org/10.1145/1594977.1592577

[84] H. Wu, G. Lu, D. Li, C. Guo, and Y. Zhang, "MDCube: A High Performance Network Structure for Modular Data Center Interconnection," in *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies*, ser. CoNEXT '09. New York, NY, USA: ACM, 2009, pp. 25–36. [Online]. Available: http://doi.acm.org/10.1145/1658939.1658943

[85] L. Gyarmati and T. A. Trinh, "Scafida: A Scale-free Network Inspired Data Center Architecture," *SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 5, pp. 4–12, 2010. [Online]. Available: http://doi.acm.org/10.1145/1880153.1880155

[86] A. Singla, C.-Y. Hong, L. Popa, and P. B. Godfrey, "Jellyfish: Networking Data Centers Randomly," in *Presented as part of the 9th USENIX*

Symposium on Networked Systems Design and Implementation (NSDI 12). San Jose, CA: USENIX, 2012, pp. 225–238. [Online]. Available: https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/singla

[87] C. E. Leiserson, "Fat-trees: Universal Networks for Hardware-efficient Supercomputing," *IEEE Trans. Comput.*, vol. 34, no. 10, pp. 892–901, 1985. [Online]. Available: http://dl.acm.org/citation.cfm?id=4492.4495

[88] R. Niranjan Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat, "PortLand: A Scalable Fault-tolerant Layer 2 Data Center Network Fabric," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 39–50, 2009. [Online]. Available: http://doi.acm.org/10.1145/1594977.1592575

[89] J. H. Ahn, N. Binkert, A. Davis, M. McLaren, and R. S. Schreiber, "HyperX: Topology, Routing, and Packaging of Efficient Large-scale Networks," in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, ser. SC '09. New York, NY, USA: ACM, 2009, pp. 41:1–41:11. [Online]. Available: http://doi.acm.org/10.1145/1654059.1654101

[90] S. Scott, D. Abts, J. Kim, and W. J. Dally, "The BlackWidow High-Radix Clos Network," *SIGARCH Comput. Archit. News*, vol. 34, no. 2, pp. 16–28, 2006. [Online]. Available: http://doi.acm.org/10.1145/1150019.1136488

[91] B. Lebiednik, A. Mangal, and N. Tiwari, "A Survey and Evaluation of Data Center Network Topologies," *CoRR*, vol. abs/1605.01701, 2016. [Online]. Available: http://arxiv.org/abs/1605.01701

[92] A. Andreyev. (2014) Introducing Data Center Fabric, the Next-Generation Facebook Data Center Network. [Online]. Available: https://code.facebook.com/posts/360346274145943/

[93] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart, and A. Vahdat, "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network," *SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 183–197, 2015. [Online]. Available: http://doi.acm.org/10.1145/2829988.2787508

[94] J. Kim, W. J. Dally, and D. Abts, "Flattened Butterfly: A Cost-efficient Topology for High-radix Networks," *SIGARCH Comput. Archit. News*, vol. 35, no. 2, pp. 126–137, 2007. [Online]. Available: http://doi.acm.org/10.1145/1273440.1250679

[95] J. Kim, W. J. Dally, and D. Abts, "Efficient Topologies for Large-Scale Cluster Networks," in *Optical Fiber Communication Conference*. Optical Society of America, 2010, p. OMV1. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2010-OMV1

[96] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. –, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=2043164.1851223

[97] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. E. Ng, M. Kozuch, and M. Ryan, "C-Through: Part-Time Optics in Data Centers," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. –, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=2043164.1851222

[98] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: A Topology Malleable Data Center Network," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, ser. Hotnets-IX. New York, NY, USA: ACM, 2010, pp. 8:1–8:6. [Online]. Available: http://doi.acm.org/10.1145/1868447.1868455

[99] M. Fiorani, S. Aleksic, and M. Casoni, "Hybrid Optical Switching for Data Center Networks," *JECE*, vol. 2014, pp. 1:1–1:1, 2014. [Online]. Available: http://dx.doi.org/10.1155/2014/139213

[100] F. Yan, W. Miao, O. Raz, and N. Calabretta, "OPSquare: A Flat DCN Architecture Based on Flow-Controlled Optical Packet Switches," *J. Opt. Commun. Netw.*, vol. 9, no. 4, pp. 291–303, 2017. [Online]. Available: http://jocn.osa.org/abstract.cfm?URI=jocn-9-4-291

[101] W. Miao, F. Yan, O. Raz, and N. Calabretta, "OPSquare: Assessment of a Novel Flat Optical Data Center Network Architecture under Realistic Data Center Traffic," in *Optical Fiber Communication Conference*. Optical Society of America, 2016, p. W1J.3. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2016-W1J.3

[102] N. Calabretta, F. Yan, and W. Miao, "OPSquare: Towards Petabit/s Optical Data Center Networks Based on Fast WDM Cross-Connect Switches and Optical Flow Control," in *Advanced Photonics 2017 (IPR, NOMA, Sensors, Networks, SPPCom, PS)*. Optical Society of America, 2017, p. NeW1B.3. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=Networks-2017-NeW1B.3

[103] W. Miao, F. Yan, and N. Calabretta, "Towards Petabit/s All-Optical Flat Data Center Networks Based on WDM Optical Cross-Connect Switches with Flow Control," *J. Lightwave Technol.*, vol. 34, no. 17, pp. 4066–4075, 2016. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-34-17-4066

[104] H. H. Bazzaz, M. Tewari, G. Wang, G. Porter, T. S. E. Ng, D. G. Andersen, M. Kaminsky, M. A. Kozuch, and A. Vahdat, "Switching the Optical Divide: Fundamental Challenges for Hybrid Electrical/Optical Datacenter Networks," in *Proceedings of the 2Nd ACM Symposium on Cloud Computing*, ser. SOCC '11. New York, NY, USA: ACM, 2011, pp. 30:1–30:8. [Online]. Available: http://doi.acm.org/10.1145/2038916.2038946

[105] G. Wang, D. G. Andersen, M. Kaminsky, M. Kozuch, T. S. E. Ng, K. Papagiannaki, M. Glick, and L. B. Mummert, "Your Data Center Is a Router: The Case for Reconfigurable Optical Circuit Switched Paths," in *Proc. of Workshop on Hot Topics in Networks (HotNets-VIII)*. ACM SIGCOMM, 2009. [Online]. Available: http://conferences.sigcomm.org/hotnets/2009/papers/hotnets2009-final52.pdf

[106] J. Kim, *High-radix Interconnection Networks*. Stanford, CA, USA: Stanford University, 2008, aAI3302840.

[107] OpenCompute. (2018). [Online]. Available: http://www.opencompute.org

[108] W. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003.

[109] Facebook. (2018) Facebook Wedge 100 - 32 x 100G. [Online]. Available: http://www.opencompute.org/wiki/Networking/SpecsAndDesigns#Facebook_Wedge_100_-_32x100G

[110] Finisar. (2018) Finisar - Optical Transceivers. [Online]. Available: https://www.finisar.com/optical-transceivers

[111] P. G. Raponi, N. Andriolli, I. Cerutti, and P. Castoldi, "Two-Step Scheduling Framework for Space-Wavelength Modular Optical Interconnection Networks," *IET Communications*, vol. 4, no. 18, pp. 2155–2165, 2010.

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description
A	Ampere
ADC	Analog-to-Digital Converter
ASIC	Application-Specific Integrated Circuit
ATX	Advanced Technology Extended
AWG	Arrayed Waveguide Grating
AWGR	Arrayed Waveguide Grating Router
CIFS	Common Internet File System
CDR	Clock Data Recovery
cm	Centimeter
COBO	Consortium for On-Board Optics
COSIGN	Combining Optics and SDN In next Generation data center Networks
CPU	Central Processing Unit
CRAC	Computer Room Air Conditioning
CWDM	Coarse Wavelength Division Multiplexing
CFP	100G Form-factor Pluggable
CXP	120 Gb/s 12x Small Form-factor Pluggable
dB	Decibel
DCN	Data Center Network
DDR3	Double Data Rate type 3
DFB	Distributed Feedback
DFE	Decision Forward Equalization
DFS	Distributed File System
DRAM	Dynamic Random Access Memory
DWDM	Dense Wavelength Division Multiplexing
EB	Exabyte

ECO	Electro-Optical Communications
EEPROM	Electrically Erasable Programmable Read-Only Memory
EFT	Extended Fat-Tree
EHFT	Extended Hybrid Fat-Tree
eMMC	embedded Multi-Media Controller
EPS	Electronic Packet Switch
ES	Electronic Switch
E/O	Electronic-to-Optical
FFE	Feed Forward Equalization
FT	Fat-Tree
GB	Gigabyte
GbE	Gigabit Ethernet
Gbps	Gigabits per second
GFS	Google File System
GPIO	General Purpose Input-Output
HFT	Hybrid Fat-Tree
I2C	Inter-Integrated Circuit
IaaS	Infrastructure as a Service
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
IP	Internet Protocol
IPI	Institute for Photonic Integration
km	Kilometer
kV	Kilovolt
LD	Laser Driver
LR	Long Reach
m	Meter
MCM	Multi-Chip Module
MEMS	Micro-Electro-Mechanical System
MHz	Megahertz
MM	Multi-Mode
MMF	Multi-Mode Fiber
MPO	Multi-fiber Push On
MPO-24	Multi-fiber Push On - 24 fibers
MPO-72	Multi-fiber Push On - 72 fibers

MXM	Mobile PCI Express Module
MW	Megawatt
NAS	Network-Attached Storage
NFS	Network File System
NFV	Network Function Virtualization
NIC	Network Interface Card
nm	nanometer
NRZ	Non-Return-to-Zero
OBO	On-Board Optics
OBS	Optical Burst Switch
OCS	Optical Circuit Switch
OPS	Optical Packet Switch
OS	Optical Switch
O/E	Optical-to-Electronic
PaaS	Platform as a Service
PAM4	Pulse-Amplitude Modulation - 4 levels
PC	Personal Computer
PCB	Printed Circuit Board
PCIe	Peripheral Component Interconnect Express
PD	Photodiode
PDU	Power Distribution Unit
QSFP	Quad Small Form-factor Pluggable - 4x1G
QSFP+	QSFP - 4x10G
QSFP28	QSFP - 4x25G
QSFP-DD	QSFP - Double Density
REST	Representational State Transfer
RTT	Round Trip Time
SaaS	Software as a Service
SDN	Software Defined Network
SerDes	Serializer-Deserializer
SFP	Small Form-factor Pluggable - 1x1G
SFP+	Small Form-factor Pluggable - 1x10G
SFP28	Small Form-factor Pluggable - 1x25G
SiP	System-in-Package
SM	Single-Mode

SMF	Single-Mode Fiber
SOA	Semiconductor Optical Amplifier
SR	Short Reach
SWDM	Shortwave Wavelength Division Multiplexing
Tbps	Terabits per second
TEC	Thermo Electric Cooler
TIA/LA	Transimpedance Amplifier / Limiting Amplifier
TL	Tunable Laser
TWC	Tunable Wavelength Converter
U	Rack Unit
UART	Universal Asynchronous Receiver-Transmitter
UPS	Uninterruptible Power Supply
USB	Universal Serial Bus
VCSEL	Vertical Surface Emitting Laser
V	Volt
W	Watt
WAN	Wide Area Network
WDM	Wavelength Division Multiplexing
WSS	Wavelength Selective Switch
ZB	Zettabyte

List of Publications

Journals

1. G. Guelbenzu de Villota, N. Calabretta, and O. Raz, "Hybrid Fat-Tree: Extending Fat-Tree to Exploit Optical Switches Transparency with WDM," in *Optical Fiber Technology (OFT)*, 2018.

2. G. Guelbenzu de Villota, N. Calabretta, and O. Raz, "Towards Standardization of Compact Data Center Switches," in *ICT Express*, vol. 3, no. 2, pp. 72-75, 2017.

3. C. Li, T. Li, G. Guelbenzu de Villota, B. Smalbrugge, R. Stabile, and O. Raz, "Chip Scale 12-channel 10 Gb/s Optical Transmitter and Receiver Subassemblies Based on Wet Etched Silicon Interposer," in *Journal of Lightwave Technology*, vol. 35, no. 15, pp. 3229 - 3236, 2017.

4. C. Li, R. Stabile, T. Li, B. Smalbrugge, G. Guelbenzu de Villota, and O. Raz, "Wet Etched 3-Level Silicon Interposer for 3 Dimensional Embedding and Connecting of Opto-electronic Die and CMOS IC," in *Transactions on Components, Packaging and Manufacturing Technology*, 2017.

5. H. J. S. Dorren, E. H. M. Wittebol, R. de Kluijver, G. Guelbenzu de Villota, P. Duan, and O. Raz, "Challenges for Optically Enabled High-radix Switches for Data Center Networks," in *Journal of Lightwave Technology*, vol. 33, no. 5, pp. 1117-1125, 2015.

Conferences

7. G. Guelbenzu de Villota, W. Miao, N. Calabretta, and O. Raz, "E-WDM: Wavelength Switching in Hybrid Networks Without Wavelength Selective Switches," in *Proceedings of the 22nd Annual Symposium of the IEEE Photonics Society Benelux Chapter*, Delft, The Netherlands, November 2017.

8. T. Li, S. Dorrestein, G. Guelbenzu de Villota, C. Li, P. Stabile, and O. Raz, "4810 Gbps Cost-effective FPC-based On-board Optical Transmitter with PGA Connector," in *IEEE 67th Electronic Components and Technology Conference (ECTC)*, pp. 1755-1760, Orlando, Florida, USA, May 2017.

9. G. Guelbenzu de Villota, W. Miao, Y. Ben-Itzhak, C. Caba, L. Schour, S. Vargaftik, K. van de Plassche, N. Calabretta, and O. Raz, "Combining Optics and SDN to Enable True Hybrid Integration of Electronic and Photonic Switching Solutions," in *IEEE Photonics Society Summer Topicals Meeting Series*, pp. 139-140, San Juan, Puerto Rico, July 2017.

10. G. Guelbenzu de Villota, C. Li, T. Li, E. H. M. Wittebol, N. Calabretta, and O. Raz, "On-board Optics and Compact switches for Mega-size Data Center Networks,"' in *Proceedings of the 21st Annual Symposium of the IEEE Photonics Society Benelux Chapter*, pp. 83-86, Gent, Belgium, November 2016.

11. O. Raz, G. Guelbenzu de Villota, T. Li, C. Li, W. Miao, F. Yan, H. J. S. Dorren, P. Stabile, and N. Calabretta, "Optical Solutions for the Challenges of Mega-size Data Center Networks," in *2016 Optical Fiber Communication Conference (OFC 2016)*, paper W1J.4, Anaheim, California, USA, March 2016.

12. G. Guelbenzu de Villota, N. Calabretta, and O. Raz, "Mid-board Optics as an Essential Building Block for Future Data Center Switches," in *Proceedings of the 20th Annual Symposium of the IEEE Photonics Benelux Chapter*, pp. 245-248, Brussels, Belgium, November 2015.

13. O. Raz, G. Guelbenzu de Villota, T. Li, and H. J. S. Dorren, "The Need for Low-cost Optical Transceivers in Future Data Center Networks," in *17th International Conference on Transparent Optical Networks (ICTON 2015)*, pp. 1-4, Budapest, Hungary, July 2015.

14. H. J. S. Dorren, G. Guelbenzu de Villota, and O. Raz, "Reality and Challenges of Photonics for DATACOM," in *Proceedings of the 40th European Conference on Optical Communication (ECOC 2014)*, p. We. 2.1.1, pp. 1-3, Cannes, France, September 2014.

15. G. Guelbenzu de Villota, O. Raz, and H. J. S. Dorren, "Impact of Emerging Technologies and Topology Selection in Large Scale Data Centers Design," in *Proceedings of the 19th Annual Symposium of the IEEE Photonics Society Benelux Chapter*, pp. 149-152, Enschede, The Netherlands, November 2014.

Acknowledgements

Last four years have been a memorable adventure. Pursuing the PhD degree has been very challenging, and it has required non-stopping dedication and unbreakable motivation to overcome all kind of difficulties. I would like to express my sincere gratitude to many people who made this success possible. It is only thanks to their help that this work has culminated in this manuscript.

First of all, I would like to thank my parents for their unconditional support and guidance all these years. Although they are not able to share with me these moments, I am sure they would be proud and celebrate this achievement with me.

I would like to thank prof. Harm Dorren, who suddenly passed away almost three years ago, for giving me the opportunity to pursue my PhD in the ECO group. His insightful suggestions and role model have pushed me to think in the overall picture, without getting lost in the details at first, and to work hard throughout my research work. I would like to thank prof. Ton Koonen for his support in the background, always available when needed, and for his valuable comments to improve this thesis. I would like to thank my daily supervisor dr. Oded Raz for his support, patience, and guidance from day one. I have been extremely fortunate to work under his supervision. His clever thinking quickly grasped the underlying scientific interest of all the ideas explored, including the ones dropped in the journey. I would like to thank dr. Nicola Calabretta for our interesting discussions. His excellent expertise in optical switching technologies has clearly enrichened the quality and impact of this work. I would like to thank all the committee members for reviewing, commenting, and approving this work: prof.dr. Lars Dittmann, dr. Salvatore Spadaro, dr. Bert Sadowski, and dr. Yaniv Ben-Itzhak.

I would like to thank José Hakkens, Jolanda Levering, Brigitta van Uitregt - Dekkers, and Yvonne van Bokhoven for their help to solve quickly and efficiently all the related paperwork issues. I would like to thank Femke Verheggen for her understanding in the difficult times. I would like to thank prof. Huug de Waardt, dr. Chigo Okonkwo, dr. Patty Stabile, dr. Georgios Exarchakos, dr. Eduward Tangdiongga, prof. Antonio Liotta, and prof. Sonia Heemstra for their support.

I would like to thank Frans Huijskens and Johan van Zantvoort for their consistent help with all practical issues in the workshop and laboratory. I would like to thank EPC for the support in all design, assembly, and mechanical issues. It has been a pleasure to work in their facilities surrounded by helpful and knowledgeable engineers. I would like to especially thank Erik Wittebol, Rob Kluijver, and Paul Beijer. I would like to thank Antwan Langeveld and Rob Sanders from Neways Technologies for the support during the manufacturing of the electronic switch prototype.

I would like to thank my friend and colleague Dr. María Torres Vega for her support during these years, always ready to listen, advice, and proofread the scientific drafts. I would like to thank Karel van de Plassche and Dennis Kofflard, students who contributed to this work. It has been really a pleasure to be part of the ECO group during these years and work with so many talented colleagues: Decebal Mocanu, Elena Mocanu, Michele Chincoli, Simone Cardarelli, Federico Forni, Jon Kjellman, Chenhui Li, Teng Li, Wang Miao, Fulong Yan, Miquel Caimari, Prometheus DasMahapatra, Nathaniel Groothof, Rianne Plantenga, Fernando Nuñez Serrano, Laura Martín González, Netsanet Tessema, Ketemaw Mekonnen, Roy van Uden, Vincent Sleiffer, Pinxiang Duan, Qing Wang, Hedde Bosman, Roshan Kotian, Haoshuo Chen, Joanne Oh, Robbert van der Linden, Kristif Prifti, Jim Zou, Fausto Gómez, Samina Subhani, Sjoerd van der Heide, Mahir Mohammed, and Chetan Belagal.

I would like to especially thank my friends Dr. Javier Quevedo and Dr. Patricia Aparicio, who housed me during the first months in the Netherlands and showed me the way to success through their example. I would like to thank my friends Willy, Jix, Juantxo, Paula, Marine, Nako, Guille, Fede, Ana, and Poche for being always there. I would like to thank my brother and sisters: Daniel, Eloísa, Leonor and Magdalena, and to my aunt Lolo.

Finally, I would like to thank my love, Mar, always close to me despite distances and difficulties. I am yours and you are mine. Forever.

About the Author

Gonzalo Guelbenzu de Villota was born on 23-02-1982 in Madrid, Spain. He graduated as Telecommunication Engineer in 2013 at the Technical University of Madrid (UPM) in Madrid, Spain. His main field of interest is the design and implementation of novel systems pushing the boundaries by the integration of state-of-the-art devices. During his studies, he worked in the development of embedded systems in a company based on Madrid and a research team investigating about hydrogen fuel cells. From 2013 he started a PhD project at Eindhoven University of Technology (TU/e) at Eindhoven, the Netherlands, of which the results are presented in this dissertation.