Optical Multidimensional Switching for Data Center Networks

Kamchevska, Valerija; Galili, Michael; Oxenløwe, Leif Katsuo; Berger, Michael Stübert

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Optical Multidimensional Switching for Data Center Networks

Ph.D. Thesis

Valerija Kamchevska

June 2nd, 2017

DTU Fotonik
Department of Photonics Engineering

DTU Fotonik
Department of Photonics Engineering
Technical University of Denmark
Ørsteds Plads 343
DK-2800 Kgs. Lyngby
Denmark

Preface

The work presented in this thesis was carried out as a part of my Ph.D. project in the period March 1st, 2014 to May 31st, 2017. The work took place at DTU Fotonik (Technical University of Denmark, Department of Photonics Engineering), Columbia University, New York, NY, USA and IBM T. J. Watson Research Center, Yorktown Heights, NY, USA.

The Ph.D. project was partially financed by the ECFP7 grant no. 619572, COSIGN and supervised by

- Michael Galili (main supervisor), Associate Professor, DTU Fotonik, Technical University of Denmark, Kgs. Lyngby, Denmark
- Leif K. Oxenløwe (co-supervisor), Professor, DTU Fotonik, Technical University of Denmark, Kgs. Lyngby, Denmark
- Michael S. Berger (co-supervisor), Associate Professor, DTU Fotonik, Technical University of Denmark, Kgs. Lyngby, Denmark

Abstract

Optical switches are known for the ability to provide high bandwidth connectivity at a relatively low power consumption and low latency. Several recent demonstrations on optical data center architectures confirm the potential for introducing all-optical switching within the data center, thus avoiding power hungry optical-electrical-optical conversions at each node. This Ph.D. thesis focuses precisely on the application of optical technologies in data center networks where optics is not only used for extending the reach, but more importantly the benefits of photonic devices are exploited for the purpose of deploying optical switching within the network.

First, the Hi-Ring data center architecture is proposed. It is based on optical multidimensional switching nodes that provide switching in hierarchically layered space, wavelength and time domain. The performance of the Hi-Ring architecture is evaluated experimentally and successful switching of both high capacity wavelength connections and time-shared subwavelength connections is demonstrated. Error-free performance is also achieved when transmitting 7 Tbit/s using multicore fiber, confirming the ability to scale the network.

Moreover, the limitations of previously proposed optical subwavelength switching technologies are discussed and a novel concept of optical time division multiplexed switching is proposed. A detailed elaboration of the envisioned scheme is given, with a special focus on the problem of synchronization. A novel synchronization algorithm for the Hi-Ring architecture is proposed and experimentally validated. Furthermore, software controlled switching in the data plane is experimentally demonstrated when the proposed algorithm is used for synchronization.

Finally, integration is discussed from two different perspectives: the first one referring to hardware-software integration where the data plane is integrated with a centralized control plane deploying a software defined controller, and the second one referring to on-chip integration of devices in

the data plane ultimately leading to integrated systems and networks on chip. Software controlled switching using an on-chip integrated fiber switch is demonstrated and enabling of additional network functionalities such as multicast and optical grooming is experimentally confirmed. Altogether this work demonstrates the potential of optical switching technologies and their implementation in future data center networks.

Resume

Optiske netværkskontakter, ofte kaldet optiske switche, er kendetegnet ved den store båndbredde og lille tidsforsinkelse på de etablerede optiske forbindelser kombineret med et beskedent effektforbrug. Adskillige nylige demonstrationer af optiske datacenterarkitekturer bekræfter potentialet af eksklusiv brug af optiske switche i datacentrene, hvorved de effektkrævende optisk-til-elektrisk-til-optisk konverteringer i hvert knudepunkt kan undgås. Denne PhD afhandling fokuserer netop på brugen af optiske teknologier i datacenternetværk, ikke blot for at øge rækkeviden men også for at høste de fordele fotoniske elementer tilbyder når de udnyttes til optisk switching i selve netværket.

Først i afhandlingen introduceres vores nye forslag til en datacenter-arkitektur, Hi-Ring. Denne arkitektur er baseret på optiske flerdimensionelle kontaktknudepunkter, der switcher i tre dimensioner, tid, rum og frekvens (bølgelængde), efter en hierarkisk opdeling. Funktionen af Hi-Ring arkitekturen demonstreres succesfuldt eksperimentelt, og ydeevnen evalueres i eksperimenter med switching af både højkapacitetsbølgelængde-forbindelser samt tidsdelte subbølgelængdeforbindelser. Fejlfri funktion opnås efter transmission af 7 Tbit/s over en flerkernefiber, hvilket understreger potentialet for at skalere netværket op yderligere.

Dernæst diskuteres begrænsninger af eksisterende forslag til optiske subbølgelængde switchteknologier, og et nyt koncept til optisk tidsmultiplekssswitching foreslås. Det foreslåede koncept uddybes i detaljer med særlig fokus på synkronisering. En ny synkroniseringsalgoritme til Hi-Ring arkitekturen foreslås og bekræftes eksperimentelt. Ved at benytte den foreslåede synkroniseringsalgoritme bliver softwarekontrolleret switching på dataplanet også eksperimentelt demonstreret.

Til sidst bliver integration diskuteret fra to forskellige perspektiver. Det første tager udgangspunkt i hardware-software integration hvor dataplanet er integreret med det centrale kontrolplan via en softwaredefineret kon-

trolenhed. Det andet perspektiv refererer til on-chip integration af enheder i dataplanet hvilket i yderste konsekvens fører til chipintegrerede systemer og netværk. Softwarekontrolleret switching baseret på en onchip integreret siliciumswitch demonstreres, hvorved ekstra netværksfunktionaliteter som multicast og optisk grooming eksperimentelt bekræftes. Det samlede arbejde i denne afhandling understreger potentialet af optiske switchteknologier og deres implementering i fremtidens datacenternetværk.

Acknowledgements

If I knew what I was doing, it wouldn't be called research. - Albert Einstein

The results presented in this thesis would not have been achieved if not for the direct or indirect contribution of many people to whom I am extremely grateful and I will start by apologizing if I may have forgotten some of you.

First and foremost I would like to thank my supervisors Assoc. Prof. Michael Galili, Prof. Leif K. Oxenløwe and Assoc. Prof. Michael S. Berger for their invaluable guidance throughout these years. Their constant support and extreme patience as well as friendliness have made this period a pleasant and enjoyable learning opportunity. I will always be grateful for introducing me to experimental work and guiding me through this journey.

I am grateful to my examiners Dr. Dominique Chiaroni (Bell Labs, Nokia, France), Dr. Benn Thomsen (University College London, UK) and Dr. Henrik Wessing (DTU Fotonik, Denmark) for taking the time to read this thesis and for their constructive feedback.

I would like to thank all former and present colleagues and friends in the High-Speed Optical Communication Group at DTU Fotonik for the great working environment that made the whole period much more fun. Thank you Francesco, Dragana, Valentina, Ashenafi, Feihong, Rameez for the much needed coffee breaks and for being great colleagues and office mates. A big thanks to Kjeld and Yunhong whose devices have made this work possible. Thanks to Anders for his help with the practical bureaucratic matters and translations in Danish.

I would like to thank the Technical University of Denmark and the ECFP7 project COSIGN for supporting this Ph.D. project. I am grateful to Otto Mønsted, Oticon and Valdemar Trane foundations for their financial

support which allowed me to attend several international conferences as well as provide support during my external research stay in US.

A special thanks to Keren Bergman and Laurent Schares for welcoming me at Columbia University and IBM T. J. Watson Research Center, respectively and for making my external research stay interesting and enjoyable.

To all my friends, thanks for all the time you dragged me out of the lab when I really needed a break. Last but definitely not least, many thanks to my parents and my sister for their continuous support unaffected by the distance and to my husband, Darko for his immense patience throughout these years.

Ph.D. Publications

The following publications have resulted from this Ph.D. project.

Articles in international peer-reviewed journals: (4)

J1 **V. Kamchevska**, Y. Ding, K. Dalgaard, M. Berger, L. K. Oxenløwe, and M. Galili, "On-Chip SDM Switching for Unicast, Multicast and Traffic Grooming in Data Center Networks," *IEEE Photonics Technology Letters*, vol. 29, no. 2, pp. 231-234, 2017.

J2 **V. Kamchevska**, V. Cristofori, F. Da Ros, B. Guo, C. Jackson, A. M. Fagertun, S. Ruepp, R. Nejabati, D. Simeonidou, L. Dittmann, M. Berger, L. K. Oxenløwe, and M. Galili, "Synchronization in a Random Length Ring Network for SDN-Controlled Optical TDM Switching," *Journal of Optical Communications and Networking*, vol. 9, no. 1, pp. A26-A34, 2017 [Invited].

J3 Y. Ding, **V. Kamchevska**, K. Dalgaard, F. Ye, R. Asif, S. Gross, M. Withford, M. Galili, T. Morioka, and L. K. Oxenløwe, "Reconfigurable SDM Switching Using Novel Silicon Photonic Integrated Circuit," *Scientific Reports*, vol. 6, p. 39058, 2016.

J4 **V. Kamchevska**, A. K. Medhin, F. Da Ros, F. Ye, R. Asif, A. M. Fagertun, S. Ruepp, M. Berger, L. Dittmann, T. Morioka, L. K. Oxenløwe, and M. Galili, "Experimental Demonstration of Multidimensional Switching Nodes for All-Optical Data Center Networks," *Journal of Lightwave Technologies*, vol. 34, no. 8, pp. 1837-1843, 2016 [Invited].

Contributions to international peer-reviewed conferences: (15)

C1 C. Jackson, K. Kondepu, A. Beldachi, Y. Ou, A. Pagès Cruz, F. Agraz, F. Moscatelli, W. Miao, **V. Kamchevska**, N. Calabretta, G. Landi, S. Spadaro, and D. Simeonidou, "COSIGN : A Complete SDN Enabled All-Optical Architecture for Data Centre Virtualisation with Time and Space Multiplexing," *European Conference and Exhibition on Optical Communications (ECOC)*, 2017 (accepted).

C2 Y. Ding, **V. Kamchevska**, K. Dalgaard, D. Bacco, K. Rottwitt, H. Hu, M. Galili, T. Morioka, and L. K. Oxenløwe, "Silicon Photonics for Multicore Fiber Communication," *Asia Communications and Photonics Conference (ACP)*, paper AF1G.1, 2016 [Invited].

C3 M. Galili, **V. Kamchevska**, Y. Ding, M. Berger, L. K. Oxenløwe, and L. Dittmann, "The Hi-Ring DCN Architecture," *Asia Communications and Photonics Conference (ACP)*, paper AS1E.1, 2016 [Invited].

C4 L. Dittmann, A. M. Fagertun, **V. Kamchevska**, M. Galili, L. K. Oxenløwe, S. Ruepp, and M. Berger, "A Roadmap for Evolving Towards Optical Intra-Data-Center Networks," *European Conference and Exhibition on Optical Communications (ECOC)*, paper M.2.F.1, 2016 [Invited].

C5 M. Galili, **V. Kamchevska**, Y. Ding, and L. K. Oxenløwe, "The Hi-Ring Architecture for Datacentre Networks," *International Conference on Transparent Optical Networks (ICTON)*, paper Th.B5.2, 2016 [Invited].

C6 Y. Ding, **V. Kamchevska**, K. Dalgaard, F. Ye, R. Asif, S. Gross, M. Withford, M. Galili, T. Morioka, and L. K. Oxenløwe, "Reconfigurable SDM Switching Using Novel Silicon Photonic Integrated Circuit," *Conference on Lasers and Electro-Optics (CLEO)*, paper STu1G.3, 2016.

C7 V. Cristofori, **V. Kamchevska**, Y. Ding, A. Shen, G. Duan, C. Peucheret, and L. K. Oxenløwe, "Error-free Dispersion-uncompensated Transmission at 20 Gb/s over SSMF using a Hybrid III-V/SOI DML with MRR Filtering," *Conference on Lasers and Electro-Optics (CLEO)*, paper STu1G.4, 2016.

C8 **V. Kamchevska**, V. Cristofori, F. Da Ros, B. Guo, C. Jackson, A. M. Fagertun, S. Ruepp, R. Nejabati, D. Simeonidou, L. Dittmann,

M. Berger, L. K. Oxenløwe, and M. Galili, "Synchronization Algorithm for SDN-controlled All-Optical TDM Switching in a Random Length Ring Network," *Optical Fiber Communication Conference and Exposition (OFC)*, paper Th3I.2, 2016.

C9 A. M. Fagertun, M. Berger, S. Ruepp, **V. Kamchevska**, M. Galili, L. K. Oxenløwe, and L. Dittmann, "Ring-based All-Optical Datacenter Networks," *European Conference and Exhibition on Optical Communications (ECOC)*, paper P.6.9, 2015.

C10 **V. Kamchevska**, A. K. Medhin, F. Da Ros, F. Ye, R. Asif, A. M. Fagertun, S. Ruepp, M. Berger, L. Dittmann, T. Morioka, L. K. Oxenløwe, and M. Galili, "Experimental Demonstration of Multidimensional Switching Nodes for All-Optical Data Center Networks," *European Conference and Exhibition on Optical Communications (ECOC)*, paper Tu.1.2.2, 2015.

C11 X. Chen, J. Regan, T. Durrant, Y. Shu, G. Saridis, G. Zervas, D. Simeonidou, **V. Kamchevska**, A. M. Fagertun, and S. Yu, "Monolithic InP-based Fast Optical Switch Module for Optical Networks of the Future," *International Conference on Photonics in Switching (PS)*, 2015.

C12 M. Galili, **V. Kamchevska**, A. M. Fagertun, S. Ruepp, M. Berger, L. K. Oxenløwe, and L. Dittmann, "COSIGN – Developing an Optical Software Controlled Data Plane for Future Large-Scale Datacenter Networks," *International Conference on Transparent Optical Networks (ICTON)*, paper Mo.D3.7, 2015 [Invited].

C13 A. K. Medhin, **V. Kamchevska**, H. Hu, M. Galili, and L. K. Oxenløwe, "Experimental Demonstration of Optical Switching of Tbit/s Data Packets for High Capacity Short-Range Networks," *Conference on Lasers and Electro-Optics (CLEO)*, paper SW3M.5, 2015 [Invited].

C14 A. K. Medhin, **V. Kamchevska**, M. Galili, and L. K. Oxenløwe, "1x4 Optical Packet Switching of Variable Length 640 Gbit/s Data Packets Using In-band Optical Notch-Filter Labeling," *European Conference and Exhibition on Optical Communications (ECOC)*, paper P.4.11, 2014.

Book chapters: (1)

B1 **V. Kamchevska**, Y. Ding, M. Berger, L. Dittmann, L. K. Oxenløwe, and M. Galili, "The Hi-Ring Architecture for Data Center Networks," *Optical Switching in Next Generation Data Centers*, F. Testa, L. Pavesi, Springer, 2017.

Contents

Chapter 1

Introduction

The development of different applications and technologies over the years has made Internet services such as web browsing, voice calls, email access etc. available today on devices like personal computers, tablets and smartphones. The rapid penetration of new services directly contributes to the growth of the overall Internet Protocol (IP) traffic. According to the Cisco visual networking index (VNI) forecast [1], the smartphone traffic will exceed the personal computer traffic by 2020, while traffic from wireless and mobile devices will account for two thirds of the total IP traffic by 2020. Additionally, the number of devices connected to the Internet will be three times higher than the global population in 2020. In order to provide support for various services, service providers have to enable non-interrupted and synchronized access to user account data independent of the underlying device used to access it.

As more information moves to the cloud, data centers (DCs) are faced with a challenge of maintaining the pace at which traffic grows. In order to cope with this growth, data center operators have been forced to expand their existing DCs in size and additionally build new ones. Over a period of seven years, data centers have grown in size by an amazing 173 % [2] and this trend is also likely to continue in the future. Moreover, projections on the growth of the data center construction market reveal an increase of 50 % by 2020 [3]. As shown in Fig. 1.1, by 2020, assuming a 27 % compound annual growth rate (CAGR), the global data center IP traffic is expected to reach 15.3 Zettabytes [4]. By expanding existing data centers and building new ones, the overall number of data centers in the world will continue to grow. According to the Cisco global cloud index (GCI) forecast [4], this will result in an astonishing number of 485 hyper-scale data centers around the

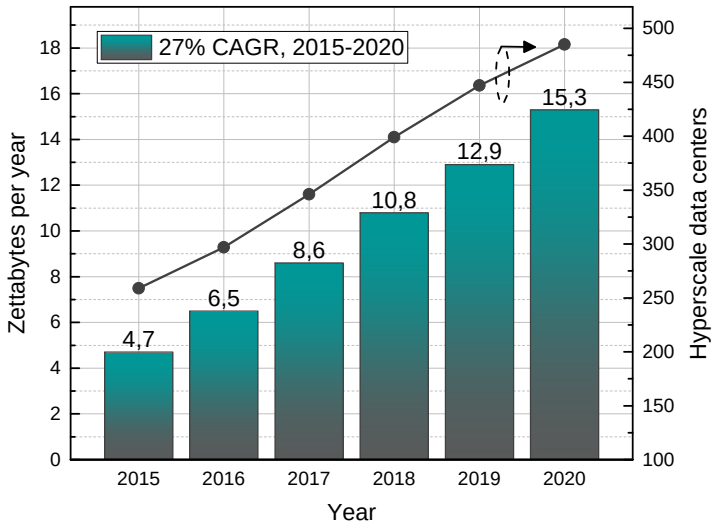


Figure 1.1: Forecast of the global data center IP traffic growth and the number of hyperscale data centers by Cisco GCI [4] for a 27 % CAGR.

world by 2020. Considering that small and medium size data centers are not even included in this estimate, it becomes clear that the overall data center industry will continue to grow exponentially over the next few years.

Industry research studies from companies like Gartner [5] and the International Data Corporation (IDC) [6] indicate that the average age of a data center is 7 to 9 years, respectively. Furthermore, the current refresh status of data center equipment ranges from 3 to 5 years [7]. This stems from the fact that with the development of new technologies it becomes cheaper to replace the old equipment with newer and more energy efficient one. Additionally, older equipment will have an increased failure rate and will experience longer downtime, that will directly translate to an increased cost of maintenance and support. The rapid refresh cycle also comes from the fact that the currently deployed equipment in data centers relies heavily on both server interfaces and electrical switching at fixed serial data rates. With traffic growth, there is a turning point after few years when instalment of new equipment that operates at higher serial rate allows for a more cost effective scaling. The issue of scalability is one of the key challenges that today's data centers face and has been the main motivation of a lot

of existing work. Similarly, the main motivation of this thesis is to address this and other issues as discussed below, that pose a challenge to the future development of data centers.

1.1 Description and scope of the thesis

This Ph.D. work is part of the European Commission project ECFP7 grant 619572 combining optics and software defined networking in next generation data center networks (COSIGN) [8] that started in 2014. The main goal of COSIGN is the integration of advanced optical hardware and software defined networking (SDN) for future all-optical data center networks (DCNs). Considering that optical technologies are known for providing high bandwidth and energy efficient operation, COSIGN aims at proposing novel DCN architectures that deploy optics in the data plane.

Optical networking has the benefit of seamless scaling to higher data rates, as both relatively slow optical circuit switches and fast optical switches can work independently of the bit rate and channel count. This holds the promise of smoother scalability and less frequent refresh cycles. However, optical switching does not provide support for functionalities that are inherent to electrical packet switching (EPS), hence a novel control plane has to be developed that will allow for smart and effective resource allocation and switch control. Within COSIGN, not only new devices and hardware have been developed, but also a lot of attention has been devoted to integrating hardware and software in such a way that the control plane is completely independent from the underlying hardware and each device in the data plane can be controlled from a common software defined control plane.

As part of COSIGN, this Ph.D. project has mainly focused on optical switching technologies for data center networks. The main contribution of the Ph.D. project is the proposed Hi-Ring DCN architecture, in which the main switching elements are all-optical multidimensional nodes that enable switching with different granularities. To allow for optical subwavelength switching, the concept of optical time division multiplexing (TDM) switching and scheduling has been developed and furthermore the problem of synchronization in the Hi-Ring architecture has been addressed. In addition, this thesis tries to tackle some of the fundamental principles behind integration of the data plane and control plane. Software and hardware integration is fundamental in providing enhanced network functionalities and optimized network resource utilization.

1.2 Structure of the thesis

The Ph.D. thesis is organized in five main chapters. Chapter 2 gives an overview of today's data centers. The currently deployed DCN architectures are briefly presented and their main challenges such as scalability, energy efficiency, resource utilization etc. are discussed. By establishing the need for new DCN architectures, existing research work on novel DCN architectures is reviewed. The presented architectures are analysed with focus on their advantages and shortcomings. At the end of the chapter, state-of-the-art optical switching technologies are revised and their possible role in future DCN architectures is discussed based on their main characteristics.

Chapter 3 introduces the Hi-Ring data center architecture. A thorough overview of the structure of the multidimensional switching nodes is given and each switching layer is discussed in details. Moreover, the network implications are analysed, emphasising the main benefits of deploying optical multidimensional switching in future data centers. In order to evaluate experimentally the performance of a small subset of the proposed architecture, an experimental prototype is introduced, consisting of a ring with three multidimensional nodes. The initial system performance results are presented and analysed and the chapter is concluded with a discussion on the main benefits and future improvements needed for these technologies to be deployed.

Chapter 4 gives an overview of existing optical subwavelength switching paradigms such as optical packet switching (OPS) and optical burst switching (OBS). Additionally, the main challenges behind deployment of any of these technologies, namely optical buffering and synchronization are discussed in details. Then, by emphasizing the importance of optical subwavelength switching especially to resource utilization in data centers, optical TDM switching is introduced as a switching layer in the multidimensional node structure from the Hi-Ring architecture. Moreover, the problem of synchronization is addressed and an algorithm is proposed allowing for synchronization of all nodes within the Hi-Ring network. Finally, the algorithm behaviour is validated experimentally and the data plane operation is evaluated in a dynamic switching scenario with automated switch synchronization.

Chapter 5 deals with the process of software and hardware integration. Decoupling the control plane from the data plane is extremely important for providing unified software control over any type of underlying hardware deployed in the data plane. Furthermore, this chapter highlights the im-

portance of integrating SDN control platforms with novel on-chip integrated devices, ultimately leading to fully SDN-controlled system on chip (SoC) and network on chip (NoC). Different examples of software controlled switching devices are presented and provisioning of enhanced network functionalities, such as multicast and incast is discussed and experimentally demonstrated.

Chapter 6 summarizes the work presented in this Ph.D. thesis, by making concluding remarks and giving a brief outlook on possible future developments.

Chapter 2

Data center networks

2.1 Introduction

This chapter aims at providing a brief introduction of today's data center organization as well as outlining existing work on novel architectures and technologies developed for future data center networks. In Section 2.2, current data center networks are reviewed with special focus on the main issues and challenges that data centers face, such as scalability, energy efficiency, resource utilization etc. Furthermore, the implications of different network solutions on scaling, power consumption and cost are discussed in details. In Section 2.3, an overview of the work on novel data center architectures is given. Data center networks based solely on electrical switching, hybrid architectures as well as all-optical architectures are presented. These novel architectures are analysed in terms of their performance and contrasted against the main requirements outlined in Section 2.2. Additionally, in Section 2.4, optical switching technologies used for optical circuit switching (OCS), optical wavelength switching and fast optical switching are reviewed. Based on their main characteristics, their role in future data center networks is discussed. At last, Section 2.5 summarizes the current state of data center networks and outlines the motivation for novel architectures and technologies that will overcome the challenges that data centers face today.

2.2 Data center networks: current status and challenges

The constant traffic growth has led to an inevitable data center expansion over the years. Today, relatively large data centers cover an area of around 30 000 m^2. Newly built data centers, such as one of Facebook's data centers that will become operational in 2019 [9] are planned to cover an area of 50 000 m^2. In addition, there are a number of data centers in the world which are more than ten times larger and with size that surpasses 400 000 m^2 [10]. Although there are some fundamental differences between the different types of data centers, i.e. small, mid-range and large data centers, most of the general data center considerations apply to any type. However, as the main goal of this Ph.D. project revolves around solutions for relatively large scale data centers, all future references to data centers will be made having in mind data centers which fall under the mid-range and large-scale data center categories.

Traffic characteristics

At the beginning of the Internet era, data centers were mainly used to provide services such as email or basic web search. Accessing remotely stored content, at a time when rather obscure amount of data was available online resulted in communication that follows the north-south traffic pattern. This means that the end user or the client will access some information that is stored on a server in the data center and the request will be served by responding accordingly.

However, as the amount of information available continued to grow and most notably services such as social networking emerged, data centers have gone through a tremendous change. According to Moore's law [11], the number of transistors in an integrated circuit doubles approximately every two years. With such technological development and the immense growth of data stored in data centers, it becomes impossible to contain an application in a single server. Thus, scaling up by building larger servers as traffic grows becomes infeasible.

Instead, the scale out approach has been adopted and virtually massive supercomputing units have been built by interconnecting distributed resources, similar as in high performance computing (HPC) environments. Therefore, unlike responding directly to a client request as in legacy data centers, an access machine has to consult other machines within the data

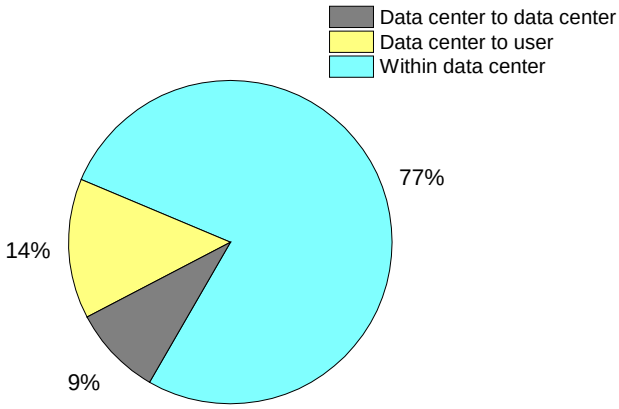


Figure 2.1: Global data center traffic by destination in 2020 according to Cisco [4].

center in order to be able to service the query [12]. This distributed computational framework also known as Map Reduce [13] allows tasks to be delegated to servers (*map phase*) and after completition, the access machine collects all the results and compiles a final response (*reduce phase*). Examples of implementation of this framework are the different distributed file systems widely used in data centers for storing information such as Google File System (GFS) [14] and Hadoop Distributed File System (HDFS) [15]. They operate in a way that files are divided into chunks of data (64 MB for GFS, 128 MB for HDFS), which are replicated and stored in different servers for reliability. Accessing the storage and providing the user with the requested content, results in a scatter-gather traffic pattern and generates a lot of server-to-server or east-west traffic.

According to a Cisco analysis [4], 77 % of the global data center traffic in 2020 will remain within the data center. This comes as a direct consequence of the distributed computing framework. As shown in Fig. 2.1, only 9 % of the overall traffic corresponds to traffic between data centers and is mainly due to data replication. The remaining 14 % is traffic between the data center and the end user, as a result of provisioning services such as email, web browsing, video streaming etc. This clearly indicates that most of the traffic in data centers is east-west, meaning within or between data centers, while only a small portion i.e. 14 % is north-south traffic or traffic between the end user and the data center. Moreover, it is important to note that even a small growth of the traffic from the end user generates an

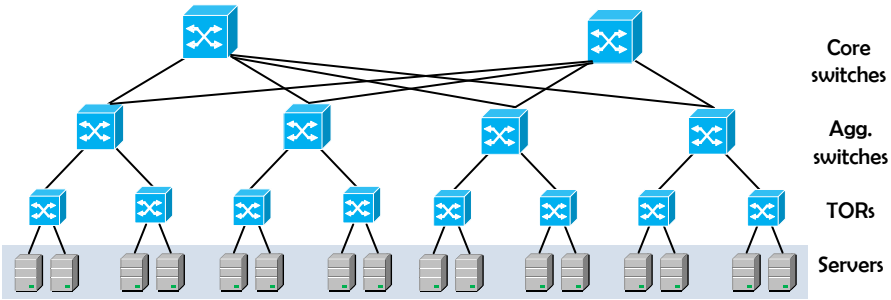


Figure 2.2: Standard three-tier tree data center architecture.

exponential growth of the traffic within the data center. Hence, it can be easily concluded that the traffic within the data center is the main driver of the data center expansion.

Organization

In a standard three-tier tree data center, electrical switches with different sizes are used to interconnect servers as shown in Fig. 2.2. Servers are placed in racks together with one or two top of rack (ToR) switches. The ToR switches are the smallest electrical switches and are used as an access point for the servers. Copper cables [16] are usually used for interconnection between the servers and the ToR switches. The interconnection between the ToR switches is established using several bigger aggregation switches often connected using short reach (SR) optics [16, 17]. These switches are in turn interconnected through few massive core switches using long reach (LR) optics [16, 17].

The rapid traffic growth has severely influenced this internal data center organization. The problem of scaling and building these enormous switches as well as maintaining and repairing them has shown to be detrimental in abandoning the scale up approach. Thus, attention has been focused towards modular scale out approaches that favour scalability. An example is shown in Fig. 2.3. The fat-tree data center architecture [18] is based on a Clos network [19] and incorporates many relatively small switches throughout the whole network. More precisely, there are k pods, each containing k switches. Each switch has k ports and a network built out of k switches supports $k^3/4$ hosts. The fat-tree architecture illustrated in Fig. 2.3 is an example of a Clos network with $k = 4$.

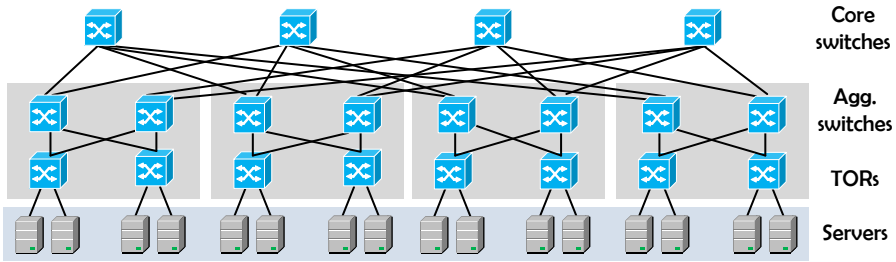


Figure 2.3: Fat-tree data center architecture.

The main advantage of this design is that all switching elements are relatively small and identical. Although the number of switches and connections in a fat-tree will be higher than in a standard tree architecture, the cost of all components will still be lower than the cost of a standard tree network with fewer, but larger and more expensive switches. Therefore, this architecture along with many other alterations of the Clos network are the most common ones in today's data centers around the world. Considering this as the state-of-the-art architecture, all further discussions on the challenges of data centers will refer to fat-tree networks.

Scalability

As previously mentioned, the fat-tree architecture has mainly been adopted due to the fact that it allows for significant cost savings when scaling the network compared to a standard tree network. However, besides cost, there are other metrics that are equally important and directly impacted by scaling the network, such as network performance. In order to observe the effect of scaling on performance, it is crucial to consider the different ways that a fat-tree network can be scaled.

In general, there are two ways of scaling a data center based on a fat-tree topology. The first approach is by replacing all the switches for others that have higher port count. However, it is important to note that the switch radix cannot be made arbitrarily high, as this goes against the whole idea behind the fat-tree topology in the first place. Additionally, replacing all the switches may not be the most cost effective method. Another approach is to keep the same switch port count and to gradually increase the number of switches at each tier. Although more feasible, this approach has a different shortcoming. Namely, as the number of servers grows rapidly in the data center, eventually the number of tiers has to be increased too.

Adding additional tiers affects the overall performance of the network as it implies an increased end-to-end latency. For latency-sensitive applications, like most of the applications in data centers, this may be detrimental. Thus, considering large-scale data centers, it becomes extremely difficult to scale the fat-tree network in a cost effective way without affecting the performance. Considering this, it becomes apparent that the deployed network topology in a large scale data center has to be scaled without impairing the performance.

Energy efficiency

The last decade has seen a rise in both data center facilities and their associated power consumption. The fast expansion compared to any other industry is especially important when taking into account sustainable energy as well as low carbon footprint. In 2010, the electricity consumption of data centers around the world accounted for 1.1 % to 1.5 % of the total electricity consumption [20] and currently this has increased to almost 3 % [21]. Moreover, the carbon footprint of data centers is around 0.3 % and they are the fastest growing contributor to the carbon footprint of the information and communication technologies (ICT) sector, which in total accounts for 2 % [22]. Thus, in order to sustain the future growth of the industry, it becomes crucial that the power consumed by data centers is reduced to the bare minimum.

Several recommendations and guidelines [23–25] have been published with the aim of achieving better energy efficiency in data centers. Metrics such as power usage effectiveness (PUE) and carbon usage effectiveness (CUE) are often used to identify how energy efficient and carbon efficient one data center is. A PUE value of 1 indicates that all the power is used for actual operation of the data center i.e. by equipment such as servers, switches etc. and no energy is used for redundant operations such as cooling and other types of overhead for running the facilities. A CUE value of 0 indicates that no carbon use is associated with the data center operation, making the data center carbon neutral.

Although data center operators have come a long way with respect of improving their PUE, there are still a lot of data centers that report values close to 2, which basicaly indicates that the facility consumes as much energy as the data center equipment. The biggest improvement can be seen by companies like Google and Facebook that have reported averaged twelve-month values for all of their facilities of 1.12 [26] and 1.09 [27], respectively.

Furthermore, all bigger data center operators have slowly started to shift to renewable energy sources such as solar or wind energy trying to improve their CUE. For example, Facebook has reported that in 2015, 35 % of the energy used to power their data centers came from renewable sources and the remaining was from coal, nuclear and natural gas [27]. In addition, their newly planned data center site [9] is expected to be powered 100 % by renewable sources and to operate with a PUE of 1.08.

Using renewable energy is one way to mitigate climate change. However there are also different measures that can be taken such as utilizing energy efficient equipment or maximizing the resource utilization, so that data center operators can reduce their overall power consumption. It has been estimated that the both servers and storage consume individually around 40 % of the total power, while the network portion is around 20 % [28]. Optimizations and further technological development, such as looking into blu-ray storage [29] aim to reduce the power consumption of the most hungry parts of the data center. However, a study from Microsoft [30] has shown that at low server utilization, the network share can be increased even up to 50 %, indicating that the power consumed by the network can affect significantly the total power consumed. Considering the huge impact that the network itself can have, it is very important to focus on how the network power consumption can be reduced.

As mentioned previously, data centers today rely on electrical switching and optical links. Thus, each optical link has to be terminated electrically at the switch interface, so that data can be switched in the electrical domain. Not only these frequent optical-electrical-optical (OEO) conversions contribute additionally to the power consumption of the network, but because of the limited serial rate, switching of high speed data streams, such as 40 Gbit/s or 100 Gbit/s is performed by switching several demultiplexed lower speed data streams. This poses a serious challenge for designing energy efficient data centers, as it implies that with traffic growth, an increased number of switching interfaces would be required, that will consume more power.

Cost

The overall cost of networking equipment in data centers can be classified as either capital expenditure (CAPEX) or operational expenditure (OPEX). Capital expenditure includes the initial cost for buying and installing the equipment. The operational expenditure is the cost incurred for the ongoing powering and running the equipment. Thus, cost efficiency can be achieved

by reducing both the capital and operational costs.

The main components of the capital investment regarding the network are the switches, optical transceivers and fiber links. Correspondingly, by reducing the power consumption of the used switches and transceivers, the operational costs can be minimized. There are different ways to minimize either one of these costs. Typically, data center operators have been applying some level of oversubscription. For standard tree topologies an oversubscription ratio higher than 1:2 is rather common and allows for some resource savings. This is usually achieved by providing lower available capacity on the ToR upstream facing links compared to the available ToR downstream facing bandwidth towards the servers. Although, this can result in improved utilization and ultimately lead to lower cost, in times of high traffic demand, some connections will be dropped and may experience longer delays. Thus, oversubscription can easily create bottlenecks and traffic congestion which impair the network performance.

Another way to minimize the cost is by lowering the direct equipment cost such as the cost of buying the switches, transceivers and fiber. Moreover, reduced power consumption of the used switches and transceivers can directly translate to lower operational costs. These reductions however are technology related and typically only gradual improvements are to be expected.

2.3 Novel data center architectures

Taking into account some of the aforementioned challenges that data centers face, several studies have investigated different approaches for novel data center architectures. These architectures can easily be classified as architectures based solely on electrical switching, hybrid architectures and architectures based on optical switching technologies. Since most of the architectures deploying some kind of optical switching often do so by replacing the electrical switches up-facing the ToR, the following considerations will apply onwards. Hybrid architectures are considered to be all architectures where both electrical and optical switching are present at aggregate/core level, while optical architectures are considered to be the ones that deploy only optical switching at aggregate/core level. Note that a data center architecture based on optical switching may have an electrical ToR switch according to this definition. The reason for this is that often these functionalities can be moved down to the server, allowing for the architecture to be all-optical.

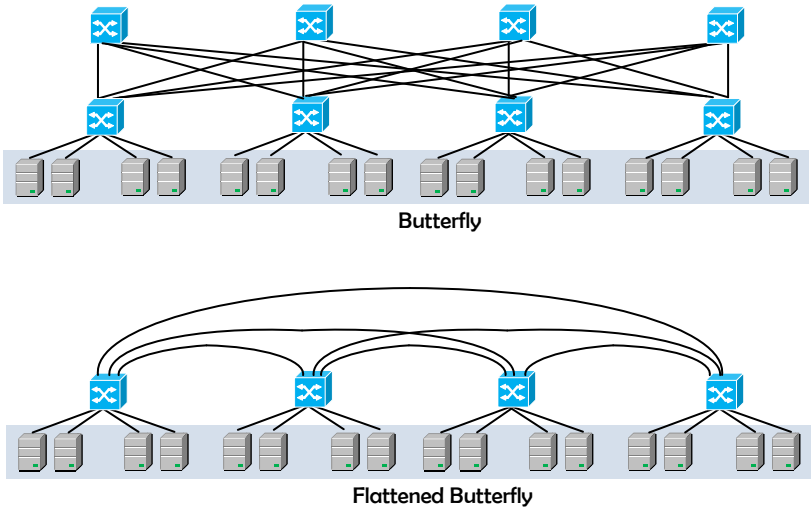


Figure 2.4: A 4-ary 2-fly butterfly and the corresponding 4-ary 2-flat flattened butterfly network.

Architectures based on electrical switching

There have been several different works on data center architectures based on EPS. Although most of them are different variations of a tree topology, such as the already mentioned fat-tree, there have also been several proposals on non-tree based topologies. In general, most frequently referred works relate to networks such as the flattened butterfly [31], dragonfly [32], DCell [33], BCube [34], or commercial architectures deployed by Google [35] and Facebook [36] etc. For the sake of brevity, attention is only focused on few of these architectures and their main advantages and disadvantages are briefly described.

The flattened butterfly uses high radix switches in order to reduce the number of tiers and generate a low-diameter network. In order to interconnect N servers, n dimensions are used with k switches in each dimension. Fig. 2.4 illustrates a standard butterfly network and a conversion to a 4-ary 2-flat flattened butterfly. The basic principle is that the number of tiers is flattened in a way that several switches are combined into one. This allows that the number of hops is reduced and intermediate stages are removed. As a result, the flattened butterfly scales better than a fat-tree network, allowing that a higher number of servers can be supported for the same switch radix. The reduced number of switches also leads to lower network

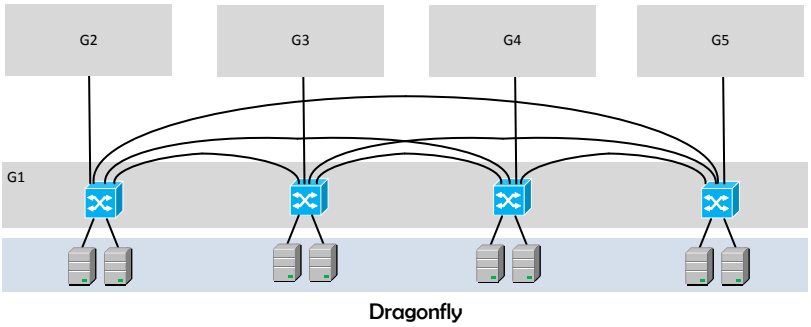


Figure 2.5: A dragonfly network with $g = 5$, $a = 4$, $p = 2$, $h = 1$.

cost and lower power consumption. However, due to the reduced number of paths and links between the switches, the flattened butterfly is a blocking network [37]. This results in a deteriorated performance for high traffic load, i.e. when all devices are transmitting and receiving at the same time. Such blocking behaviour can cause an unacceptable switching delay or packet loss.

The dragonfly network is a hierarchical three-level network that extends the flattened butterfly by increasing the effective radix of the switches to further reduce the cost and increase the scalability of the network. As the network fiber cost is dominated by the longer optical cables, the dragonfly reduces the number of global links by creating larger virtual switches. This is done by creating groups of switches and each group acts as a virtual high radix switch. There are g groups in the network and in each group there are a switches. A single switch connects to p terminals, $a - 1$ local switches within the group and has h global links towards switches in other groups. Thus, instead of having a switch radix of $k = p + h + a - 1$, a single group represents a virtual switch with radix $k' = a \times (p + h)$. Fig. 2.5 illustrates an example of a dragonfly network where the radix of the virtual switch ($k' = 12$) is double than the radix of the switches within the group ($k = 6$). Although, the dragonfly network offers several improvements in terms of cost and scalability, the fixed connectivity between groups is only suitable for a rather uniform traffic pattern. In the case of adversarial traffic or traffic that generates non-uniform load on the network, bottlenecks can be created [38]. In order to increase the throughput, non-minimal (indirect) routing [39, 40] can be applied, but this results in increased number of hops and thus higher latency [41,42].

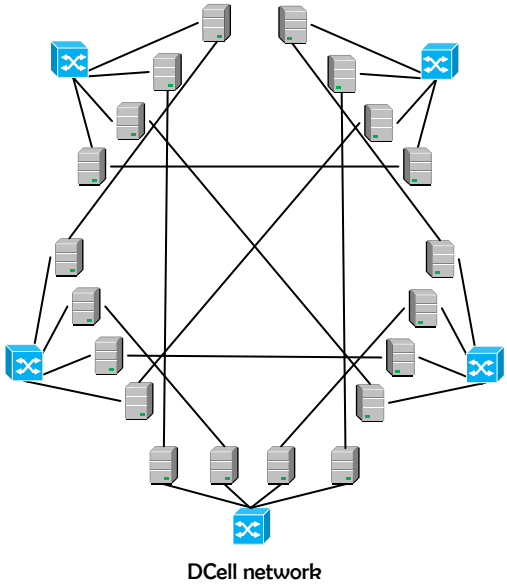
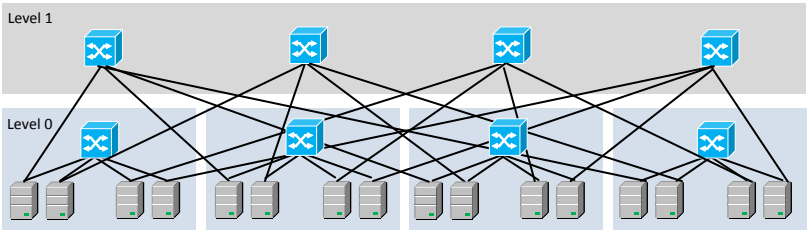


Figure 2.6: A level-1 DCell network.

DCell is a server-centric recursive architecture in which not only switches, but also servers themselves are used to switch traffic. Thus, a server is connected to a switch and additionally other servers. The basic building block of a DCell is called DCell_0 and it contains n servers and a switch. DCell_k is used to denote the level-k DCell. Fig. 2.6 illustrates an example of a DCell network with $n = 4$. A DCell_1 is constructed using $n+1$ basic DCells. Additionally, each DCell_0 connects to other DCell_0 using a single link. It can be seen that by increasing the level of the DCell, the number of servers increases double-exponentially, allowing the topology to scale very well and to be only limited by the number of ports on both the servers and the switches. However, DCell has issues with traffic congestion and latency especially as the network grows [43]. This is due to the fact that inter-DCell links in large networks are often congested and this results in packet loss. Moreover, direct routing is not inherent for this topology, so for high levels, the number of hops can increase dramatically.

Similarly to DCell, the BCube architecture is also a recursive network in which both servers and switches are used to switch traffic. This topology is specially designed for modular shipping container based data centers that can be portable and have a shorter deployment time. BCube_k is defined from BCube_0, which represents the basic building block and is



BCube

Figure 2.7: A BCube network.

equivalent to the basic DCell. The main difference lays in the process of constructing higher level topologies. A $BCube_k$ is constructed not only by using n $BCube_{k-1}$, but also by using n^k n-port switches. For example, a $BCube_1$ is constructed by using n $BCube_0$ and n additional n-port switches as shown in Fig. 2.7. Just like DCell, the number of levels is also dependent on the number of ports on the servers. However, although the number of servers in BCube grows exponentially with the levels, it grows much slower than DCell. Considering that this network is envisioned for modular data centers this may be acceptable, however its application for larger data center networks may be challenging.

Table 2.1 summarizes some of the main parameters characteristic for the different networks discussed such as the number of servers, switches and links supported for specific switch radix. It should be noted that notations may vary depending on the architecture, so please refer to the

Table 2.1: Summary of the main parameters of architectures based on electrical switching.

Topology / Resources	Fat-tree	Flattened butterfly	Dragonfly	DCell	BCube
Switch radix	k	$k' = n(k-1)+1$	$k = p+h+a-1$ $k' = a(p+h)$	n*	n*
No. of servers, N	$\frac{k^3}{4}$	k^n	gap	$\geq\left(n+\frac{1}{2}\right)^{2^k}-\frac{1}{2}$ $\leq(n+1)^{2^k}-1$	n^{k+1}
No. of switches	$\frac{5k^2}{4}$	$\frac{N}{k}$	ga	$\frac{N}{n}$	$n^k(k+1)$
No. of links	$\frac{3k^3}{4}$	$\frac{N}{2k}(k'-k)+N$	$\frac{ga(h+a-1+2p)}{2}$	$(\frac{k}{2}+1)N$	$n^{k+1}(k+1)$

* Only switch radix, but servers also used for switching.

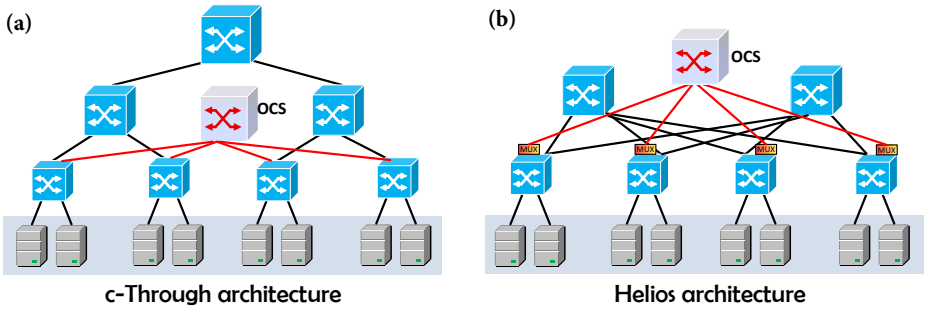


Figure 2.8: (a) A c-Through network and (b) a Helios network. (OCS - optical circuit switch).

text description for the different parameters used. In addition, some of the architectures like DCell or BCube also use servers for switching, so direct comparisons based on the switch radix and number of switches should be avoided, as the total number of switching interfaces in those cases will be a combination of the switch and server ports used for switching.

Hybrid architectures

The main motivation behind hybrid architectures is the introduction of optical switching in the data center in a way that it will either complement or exist in parallel with the already established network based on electrical packet switching. Among the many solutions, the c-Through [44] and the Helios [45] architectures are an example of a data center network in which optical circuit switching is gradually introduced.

The c-Through architecture proposes the use of optical circuit switching in parallel with the existing network based on electrical switching. As shown in Fig. 2.8 (a), an optical circuit switch is introduced in such a way that all the ToR switches are connected to both the electrical switches at aggregation level and to the optical switch. The goal is that, for high bandwidth connection between racks, the optical circuit switch can be used to offload the electrical packet switched network. This is done by monitoring the traffic at the host side.

The concept behind Helios is very similar to the c-Through proposal, with few main differences. First, the links in Helios are based on wavelength division multiplexing (WDM). A passive wavelength multiplexer at the ToR side is used to multiplex several wavelength channels as shown in Fig. 2.8 (b). These WDM signals can be regarded as links with high ca-

pacity. Another difference is that the Helios network envisions a flatter 2-layer architecture only composed of electrical ToR switches (also regarded as pod switches) and core switches that can be both electrical and optical. Similarly, the optical switches are used for high capacity long lived communication between racks.

The main advantage of both of these schemes is that they are based on components and technologies that are readily available and commercially mature. Additionally, they allow for incremental updates of the network. Introducing optics in the data center can provide increased capacity at a reduced cost and power consumption. Although these are crucial improvements that pave the way for optics in the data center, both proposals are based on optical circuit switching, which has a reconfiguration time in the millisecond range. This means that in order to compensate for the reconfiguration overhead, the optical switch should be used only for long lived connections. Thus, no changes and improvements are introduced to the electrical packet network that handles short lived connections.

Architectures based on optical switching

All-optical switching for data center networks has become quite attractive in recent years. Proposals such as the data center optical switch (DOS) architecture [46], Proteus [47], Petabit [48], a flat DC architecture [49], Lightness [50], throughput optimized photonically optimized embedded microprocessors system (TOPS) [51], hybrid optoelectronic packet router (HORP)-based torus [52], reconfigurable dragonfly [53], OPSquare [54], etc. are representatives of this type. Based on their similarities and whether they deploy only optical circuit switching, only optical packet switching, or combine both, a few of these architectures will be presented in details.

The reconfigurable dragonfly presented in [53] is an example of an all-optical data center architecture based solely on OCS. The dragonfly is a network topology in which the number of global links between groups is fixed. Thus, even when some of the links between groups are not used, they can not be allocated to groups that have high load. For this reason, the main goal of this proposal is to introduce optical switching for the purpose of physical rewiring in the data center in order to facilitate asymmetric link allocation for adversarial traffic patterns. The original dragonfly and the envisioned network are illustrated in Fig. 2.9 (a). Although, optical switching can provide significant performance improvements for adversarial traffic patterns compared to a standard dragonfly network, this proposal has the same drawback as the hybrid architectures. Namely, the reconfig-

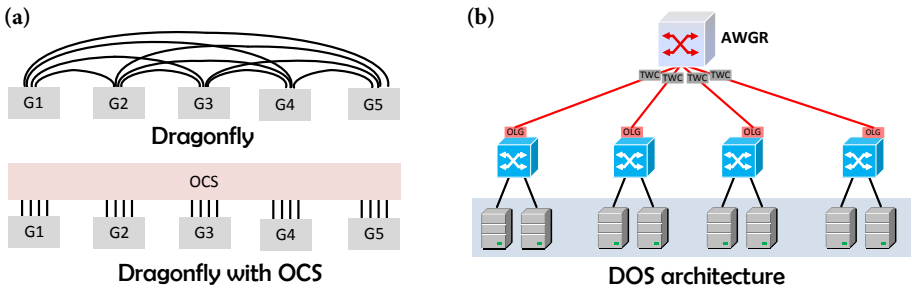


Figure 2.9: (a) A dragonfly network with fixed global links (top) and with reconfigurable global links (bottom). (b) The DOS architecture. (OCS - optical circuit switch, OLG - optical label generator, TWC - tunable wavelength converter, AWGR - arrayed waveguide grating router).

uration is only worth it if the data flows are long lived. Similar reasoning also applies to other all-optical data center architectures based on OCS such as Proteus [47].

The DOS architecture is an example of an all-optical architecture based solely on OPS. The main idea behind this proposal is to directly replace electrical packet switching with optical packet switching as shown in Fig. 2.9 (b). This is done by deploying an optical label generator (OLG) at the transmitter side, so that packet labels carrying the routing information can be inserted. A combination of tunable wavelength converters (TWCs) and arrayed waveguide grating router (AWGR) is used to perform the packet switching. Label extraction is performed at the input of the switch and based on the processed data, the TWC is configured, allowing for different routing through the passive AWGR. Congestions are solved by buffering in the electrical domain.

Similar proposals based on this concept, deploying the same or other technologies to perform the optical packet switching and solving contentions are Petabit [48], the flat DC architecture [49], a variation of the HORP-based torus [52], OPSquare [54], etc. For all of these works, similar considerations apply. Although the idea of replacing electrical packet switching with optical packet switching can bring advantages in terms of scaling and power consumption, there is one fundamental issue. Namely, even though optical buffering technologies have been researched for some time, there are still no mature and cost effective products. Thus, replacing electrical packet switching directly with optical packet switching is not straightforward. Moreover, buffering in the electrical domain as well as label process-

ing requires OEO conversion. These components together with the OLGs make the overall system cost relatively high.

At last, several architectures combine the use of OCS and OPS such as Lightness, [50], a variation of the HORP-based torus [52], etc. Considering the aforementioned considerations for each technology, although these solutions exploit the benefits of the two approaches, they are still limited by the unavailability of mature optical buffering techniques.

2.4 Optical switching technologies for data center networks

Since most of the different architectures already discussed are based on different optical switching technologies, this section will give an overview of their main characteristics. Based on this, their role in future data center networks will be studied.

Optical circuit switching technologies

Optical circuit switching is a mature technology that is deployed in metro/core and long-haul networks. The principle behing this type of switching is that first, a connection i.e. a circuit, has to be established and only then, data can flow between the two endpoints. The time to establish the connection depends on the reconfiguration time of the switch, which is often in the millisecond range. The switching can be based on different physical effects such as micro-electro-mechanical system (MEMS) based switching, piezoelectric beam-steering switching, or thermo-optic switching and commercial solutions of these technologies can easily be found [55–57]. Additionally, monolithically integrated silicon photonic MEMS switches with high radix have already been demonstrated [58].

Optical circuit switching offers protocol and bit rate independent operation. Additionally, several wavelength channels can be switched simultaneously, thus leading to energy efficient operation. This also means that, as the traffic grows, scaling by increasing the rate or the number of wavelength channels can be done without affecting the network infrastructure. Optical circuit switches often have very low insertion loss (maximum 1-2 dB) and excellent crosstalk performance (<-50 dB). Their port count can easily be scaled to 384x384 [59] and higher, with switch size ranging from chip scale to few rack units (RUs).

Optical wavelength switching technologies

Optical wavelength switching is also a mature and commercially deployed technology in metro/core and long-haul networks. The goal of this type of switching is to be able to access the individual channels from a WDM system where several wavelength channels can be transmitted simultaneously in the same fiber. The most commonly used switches from this type are liquid crystal on silicon (LCoS) based wavelength selective switches (WSSs) [60], which can be dynamically reconfigured.

The reconfiguration time of these types of switches is also in the millisecond range. Additionally, WSSs with radix as high as 1x93 have already been demonstrated [61]. WSSs often have insertion loss of few dB, relatively good crosstalk performance (<-35 dB) and size around one RU.

Fast optical switching technologies

Fast optical switches usually have reconfiguration time of few nanoseconds, that allows for good bandwidth utilization, as the channel is not used for transmitting data only during a small portion of the time. There are different types of switches with this characteristic, such as electro-optic lithium niobate ($LiNbO_3$) based switches [62, 63], semiconductor optical amplifier (SOA) based switches [11], lanthanum-modified lead zirconate titanate (PLZT) switches [64], etc.

They can operate with insertion loss of 0 dB for SOA based switches and up to 4-5 dB loss for $LiNbO_3$ and PLZT based switches. Their size can range from few millimetres for on-chip fully integrated solutions to few centimetres. Increasing their radix is rather challenging as their insertion loss also scales with the port count, or in the case of SOA based switches their operation is not noiseless. Moreover, port scaling is often achieved by adding intermediate stages in the switch matrix, which can affect the crosstalk performance.

2.5 Summary

In this chapter, we have provided a brief overview of the status in today's data centers and looked into their current organization. Based on the traffic distribution within the data center, we have discussed the main issues that data centers are facing, such as scaling, power consumption, cost, etc. Novel architectures based on electrical packet switching, hybrid architectures based on both electrical and optical switching, as well as all-optical

architectures have been presented and analysed with respect to how well do they satisfy the aforementioned requirements. At last, we have summarized state-of-the-art optical switching technologies, focusing on their main characteristics. By presenting the current issues and existing proposals and technologies, we have outlined the motivation for new data center architectures that will pave the way for deployment of optical switching in future data center networks.

Chapter 3

The Hi-Ring data center architecture

3.1 Introduction

This chapter introduces the Hi-Ring data center architecture. In Section 3.2, the main concept behind the proposed architecture is discussed, along with the motivation for deployment of optical switching technologies in data center networks in general. A detailed overview is given on the structure of the multidimensional switching nodes, describing thoroughly each switching layer. In addition, the main benefits of the Hi-Ring architecture as well as the overall network implications are discussed. In Section 3.3, the experimental setup for a lab prototype of a scaled-down version of the architecture is presented, outlining the different network scenarios investigated. Furthermore, the results of the experimental demonstration are presented and the system performance is analysed. Successful communication between three nodes in two different network scenarios is achieved with relatively low penalty, confirming the feasibility of the architecture. In addition, aggregated 7 Tbit/s throughput is achieved using a single multicore fiber, reaffirming the ability to provide support for tremendous capacity. At last, Section 3.4 summarizes the proposed data center network, the concept of multidimensional switching and the advantages of the Hi-Ring architecture. The experimental demonstration and results are outlined and based on the system performance, future work and improvements are discussed. The chapter is based on work published in [J4], [C10].

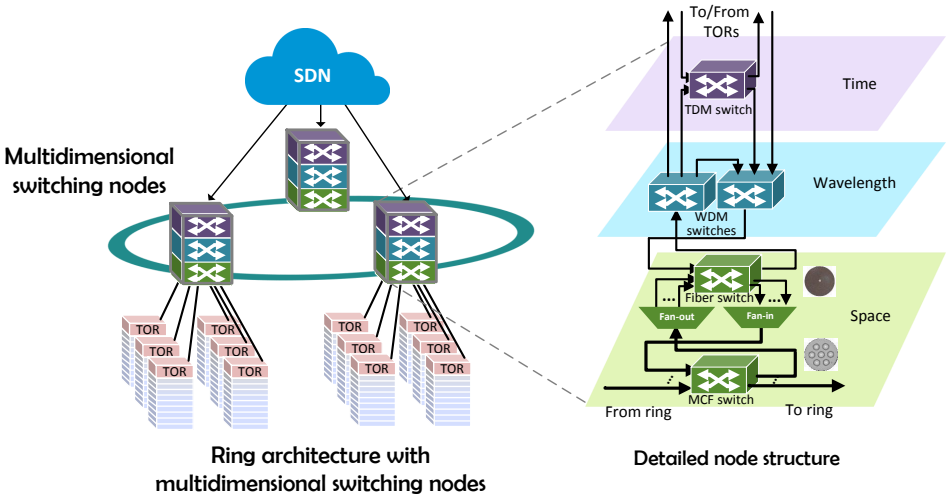


Figure 3.1: The Hi-Ring data center architecture. (ToR - top of rack switch, SDN - software defined networking, MCF - multicore fiber, WDM - wavelength division multiplexing, TDM - time division multiplexing.)

3.2 The Hi-Ring data center architecture

The proposed architecture is illustrated in Fig. 3.1. Servers are placed in racks and connected to ToR switches. Each ToR switch is connected to a multidimensional switching node. These nodes are composed of all-optical switches that can switch connections with different granularities, i.e. within the space, wavelength and time domains. Assuming that the add/drop ratio is asymmetric with respect to the pass-through traffic, a ring topology for interconnection of the nodes is envisioned.

The main motivation of the architecture is to deploy optical switching in data centers, as optical switches are known for providing high bandwidth at a relatively low power consumption and low latency. However, having previously discussed the different approaches to this matter, it is clear that any architecture deploying optical switching has to provide support for the different connectivity requests within the data center. Connections have commonly been classified as either 'elephant' flows i.e. large and long-lasting flows demanding greater bandwidth and 'mice' flows, namely the smaller, more dynamic flows requiring high connectivity. Thus, although optical circuit switching is suitable for 'elephant' flows, a viable alternative has to be found for serving 'mice' flows and replacing EPS.

In order to provide support for connections with different granularity, the proposed architecture introduces the use of full wavelengths and time slots (as a subwavelength entity), serving 'elephant' and 'mice' flows, respectively. Instead of trying to replace electrical packet switching directly with optical packet switching, the use of optical TDM switching is proposed. The concept behind TDM is that a single wavelength can be shared in time among few bursty connections that require bandwidth lower than a full wavelength. However, unlike packet switching, connections are circuit-oriented, meaning that resources have to be allocated in advance. By doing this, the need for optical buffering within the network is eliminated completely and buffering is pushed towards the network edge. For a detailed description of the envisioned scheme, refer to Chapter 4.

In addition, in order to support future growth and provide cabling consolidation, space division multiplexing (SDM) using multicore fibers (MCFs) is envisioned. Utilizing several different multiplexing technologies allows for flexibility in terms of the direction of scaling, i.e. by increasing the bit rate per channel, by adding more wavelengths, cores per fiber, multicore fibers etc. Furthermore, in order to provide optimal resource allocation and efficient bandwidth utilization, the deployment of a centralized SDN controller is crucial.

3.2.1 Miltidimensional node structure

The main building block of the Hi-Ring architecture is the multidimensional switching node. Three different multiplexing technologies are used as part of the architecture and thus reflect the switching levels in the node structure, i.e. space, wavelength and time-based switching. The node layout is depicted in Fig. 3.1. It can be seen that the three dimensions are hierarchically laid out, allowing that only connections that require switching with finer granularity proceed higher along the stack. Thus, bypass at lower levels results in both low switching penalty and low latency connectivity.

Switching in the space dimension is performed using two different types of switches, the multicore fiber switch and the fiber switch. The multicore fiber switch allows for switching of full multicore fibers, meaning that all cores are switched from the same input port to the same output port. A switch with this functionality has recently been demonstrated for the first time in [65]. As nodes are interconnected with multicore fibers, this switch allows that traffic from a specific multicore fiber can either be accessed or directed towards a specific destination node. This means that most of the ports on this switch are used for multicore fibers connecting the node

with its neighbouring nodes and a smaller portion of the input/output ports is dedicated for traffic originating or destined for the higher level switch within the node, i.e. the fiber switch. The spatial (de)multiplexing needed between the two levels is performed using designated fan-in and fan-out devices. Bypass of traffic at this level, results in negligible latency for the connections served through those physical resources. Moreover, the simultaneous switching of highly aggregated traffic leads to very low switching energy per bit, and in turn, low operational costs.

The fiber switch is a standard optical circuit switch that operates at a single fiber core granularity. By switching individual cores, a single core can be added or dropped at each node or rerouted to a different multicore fiber. Moreover, it is possible to use the two space switches in combination to change the core arrangement and perform bypass at a fiber core granularity, thereby providing additional flexibility in the network. By adding or dropping cores and performing switching at higher levels, the fiber switch allows for traffic to be repacked and rearranged and then added back to the ring. Most of the inputs/outputs of this switch are connected to the multicore fiber switch, with only a smaller portion of fiber cores connected to the higher level switches, i.e. the WDM switches.

The WDM switch is a reconfigurable wavelength selective switch. Using this switch, dynamic wavelength multiplexing and demultiplexing can be performed i.e. wavelengths can be added or dropped from fiber cores. This is important for two reasons. First, because full wavelengths are used for establishing long-lived connections that usually require a certain amount of bandwidth over longer period of time. The second reason is for accessing individual wavelengths that are used for establishing slotted connections, before proceeding with subwavelength switching. At each node in the Hi-Ring architecture, several 1xN WDM switches are deployed, enabling 1:N demultiplexing for dropped cores or N:1 multiplexing for cores on the add side. Thus, on both the add and drop side, a single fiber core is connected to a single WDM switch. The fibers carrying the individual wavelengths are connected either to the WDM ToR front-end ports allocated for long-lived connections, or to the ports of the higher level TDM switches. Additionally, it is possible to have bypass ports, i.e. directly connected ports between the demultiplexer and multiplexer, providing bypass at WDM level.

The TDM switch allows for optical subwavelength switching in the time dimension. As previously discussed, the main motivation behind TDM is to provide support for bursty connections by sharing a single wavelength and thus utilize the available bandwidth efficiently. The switching at this level

is achieved by using fast optical switches that switch in the nanosecond range, making the whole sharing concept, bandwidth efficient. The ports of the TDM switches are connected either to the WDM switches or to the ToR ports allocated for subwavelength connections. It is important to note that, ideally, the transmitter at the ToR is a wavelength tunable transmitter or as an alternative, a cheaper fixed wavelength transmitter can be used. In any case, a burst mode receiver (BMR) is required for reception of data in burst mode. Moreover, in order for transmission and switching to happen at the right time instance, a synchronization mechanism is required.

Node modelling and trade-offs

In order to choose a node structure with a given number of switches at each level, a specific case of a data center with N servers is considered. The goal is to demonstrate that the architecture is feasible in terms of practical deployment and equipment organization. Moreover, the trade-offs of having a lot of smaller multidimensional nodes vs. having a few bigger nodes are studied. The optimum node configuration is chosen by considering real implementation constraints, such as the signal integrity or the length of a row with server racks that connect to a single multidimensional node.

The following assumptions are made. A data center with 100 000 servers is considered as a reference case, although the same considerations will apply for any arbitrary number of servers. 24 servers are connected to a single ToR switch and there are 2 ToRs per server rack. The ratio of WDM and TDM ports is set to 50 %, meaning that half of the ToR ports are used for full wavelength connections, while half are shared in a time slotted manner. For this analysis, the port ratio is kept constant, however it is important to note that further optimization can be achieved by varying this parameter to better match the real traffic demand. 8x8 TDM switches, 1x16 WSSs, 384x384 fiber switches and 8x8 MCF switches are available for building the node. Eight ports of the WSSs are connected to the TDM switches and the remaining 8 ports are connected to the ToR switches.

The node is modelled based on the number of dropped fiber cores. An equivalent add and drop capacity is considered and either one of them can be used as a reference to indicate the number of ToRs or servers connected to the node. Thus, the number of dropped cores at each nodes dictates the different node configuration with different number of switches required at each level. Considering that more MCFs could be added between the nodes for increased flexibility without affecting the dimensioning of the node, only the minimum number of MCF switches at each node will be stated.

Table 3.1: Main parameters for modelling the Hi-Ring network in terms of the number of multidimensional nodes required and the structure of each node.

Dropped cores per node	Min. MCF switches	Fiber switches per node	WDM switches per node	TDM switches per node	Number of nodes	Servers per node	ToR switches per node	Length of rack row [m]
20	1	1	40	40	312.5	320	13.3	5.9
50	2	1	100	100	125	800	33.3	14.7
100	4	1	200	200	62.5	1600	66.7	29.3
170	7	1	340	340	36.8	2720	113.3	49.9
340	13	2	680	680	18.4	5440	226.7	99.7
480	18	3	960	960	13	7680	320	140.8
620	23	4	1240	1240	10.1	9920	413.3	181.9
690	25	4	1380	1380	9.1	11040	460	202.4
760	28	4	1520	1520	8.2	12160	506.7	222.9
830	30	5	1660	1660	7.5	13280	553.3	243.5

The parameters used for the different dimensioning cases are shown in Table 3.1. The number of nodes and the number of switches per node for each case are illustrated in Fig. 3.2. From Table 3.1, it can be seen that when varying the number of dropped cores, different node configurations with different amount of switches at each level are possible. Moreover, for the given node structure, different number of servers and ToRs will be served from a single node. Assuming 2 ToRs per rack and a rack width of around 0.4 m, for each case the length of the rack row will be different. As shown in Fig. 3.2, the lower the total number of nodes, the bigger the nodes will be, i.e. the number of switches at each level per node will be higher.

On one side, having a lot of small nodes with a relatively low number of ToR switches per node is not desirable, mainly because of maintaining a good signal integrity. In order to limit this to an acceptable value, the preferred number of nodes is set below 20. On the other side, assuming that a node is assigned to a single row of racks, then the number of racks per node will be limited by the calculated row length shown in Table 3.1. If we consider a row length of around 200 m to be the maximum acceptable, the minimum amount of nodes to support the given number of servers is around 9. Having a range of acceptable node configurations, demonstrates that the architecture is feasible in terms of practical deployment and the best node configuration can be chosen based on the specific network served.

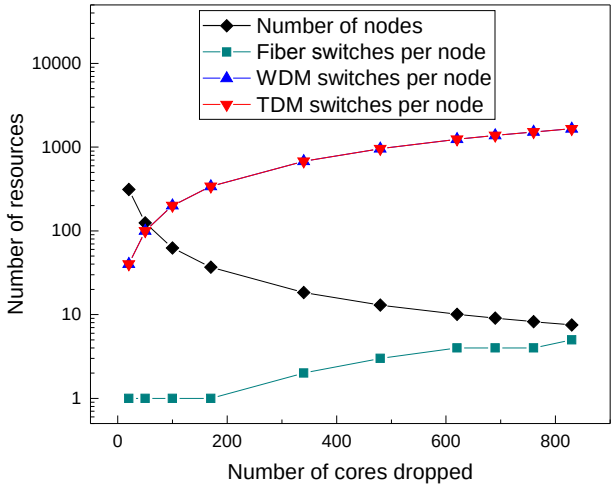


Figure 3.2: Resource dimensioning of a multidimensional node indicating the number of nodes and the number of switches per node for different node structures.

3.2.2 Network impact evaluation

As previously mentioned, the Hi-Ring architecture has several benefits. First, the hierarchical design as well as the use of SDM technologies allows for traffic aggregation which results in a relatively simplified physical topology. This means that a high number of servers can be supported with a reasonably low number of nodes and physical links, which directly translates to simplified cable management, easier maintenance and improved airflow.

Another important feature is the ability to perform bypass at different levels. Since the switching is performed optically and connections are pre-established, bypassing a node enables low latency communication, as the only delay experienced will be the physical propagation delay through the switch. Considering that many data flows are switched at the same time, the switching is performed with relatively low switching energy per bit. Moreover, intermediate nodes are not required to process all the incoming traffic at each level, but only at the lowest level needed. In this way, significant offloading is obtained and no additional components are required for processing at higher levels.

Unlike Ethernet switching, the proposed multidimensional nodes are fully bit rate independent and can easily upgrade to a higher serial rate

without the need to replace equipment. Thus, the same switch port count can be retained without having to resort to parallel solutions to support higher rate. This is important not only in terms of the capital investments savings, but also because the operational costs of the network can remain the same i.e. the same infrastructure can be used to provide higher throughput.

The knowledge of the traffic pattern in the data center can effectively be exploited in order to optimize the resource allocation through the SDN controller. Thus, application specific traffic patterns, long-term traffic patterns like 'peak of day' traffic demands, 'night and day' traffic pattern, seasonal patterns etc., can be taken into account. In addition, flexibility and adaptability can easily be achieved by dynamically reconfiguring the switches at each level.

There are two different approaches with respect to the actual physical structure of the node and the level of integration. One approach is to keep a modular node design where the different switching levels exist as different components that are interconnected with each other. The advantage of this approach is that it can provide simplified upgrades and component repairs as well as elastic migration towards new technologies. However, although individual switches at each level may provide greater flexibility when scaling, there are also arguments for integrating all or some of the components on a single platform such as silicon (Si). Integration is important in terms of achieving low footprint and low power consumption as well as addressing relevant commercial issues such as fabrication and packaging, cost, cooling approach etc. Moreover, as in most data center network proposals based on optical switching technologies, optical amplification is required to compensate for the insertion loss. Thus, integration can facilitate combining the switching devices and on-chip optical amplifiers.

Considering the definition of hybrid and all-optical architectures in Chapter 2, the Hi-Ring architecture can be classified as an all-optical architecture. Although in the current proposal the ToR switches are access point for the multidimensional switching nodes, in future, this functionality can easily be pushed down to the server level.

Having mentioned some of the main benefits offered by the Hi-Ring architecture, it is also important to reflect on areas that require future work. The overall network performance is one important issue to be addressed. This includes preserving the signal integrity, by carefully designing the network, limiting the maximum penalty experienced by connections, technological improvements with respect to insertion loss and crosstalk of the

Table 3.2: Overview of the cost and power consumption of the main network components used for the analysis. The specified values are an estimate from commercially available data and existing research studies available in [45, 55, 60, 64, 66–69].

Resource	Cost [$]	Power consumption [W]	Reference
Ethernet switch port	500	8.5	[45], [66], [67]
OCS port	500	0.14	[45], [55]
WDM switch port	900	1	[60]
TDM switch port	900	1	[64]
Transceiver, SR	65	1	[60], [68]
Transceiver, LR	130	1	[60], [68]
Transceiver, DWDM	350	1.5	[60], [68]
1m fiber (SMF, MMF)	0.2	/	[69]
1m fiber (MCF)	1	/	/

used switching devices, and of course, including optical amplification, if needed. Another important aspect is the control plane and the process of establishing connections. Unlike electrical packet switching, the nature of the connections is circuit-oriented, thus it is crucial to establish connections fast and to provide low latency communication.

Cost and power consumption analysis

As already discussed, the Hi-Ring architecture holds the promise to provide benefits in terms of cost and power consumption of the network. In order to evaluate these improvements, an inventory comparison has been made between a standard fat-tree network and a Hi-Ring network. The used values for the cost and power consumption of the different devices and resources in general are listed in Table 3.2. Note that the displayed values are estimates derived from commercially available product data sheets and existing research work.

Fat-tree

The following assumptions have been made regarding the fat-tree architecture. The number of servers supported is denoted as N. $2 \times N$ short reach transceivers are used for connecting the servers to the ToR switches, $2 \times N$ short reach transceivers are used for connecting the ToR switches to

the aggregation switches and $2 \times N$ long reach transceivers are used for connecting the aggregation switches to the core switches. The total transceiver cost, $C_{TX/RX}$ is defined as the sum of the cost of all transceivers or

$$C_{TX/RX} = 4 \times N \times C_{SR} + 2 \times N \times C_{LR} \tag{3.1}$$

with C_{SR} and C_{LR} indicating the cost of a short reach transceiver and the cost of a long reach transceiver, respectively. In connection with the short reach transceivers, multimode fibers (MMFs) are used, and single mode fibers (SMFs) are used with the long reach transceivers. The average length of the links is assumed to be 2 m for server to ToR, 50 m for ToR to aggregation switches and 500 m for the links between the aggregation and core switches. For simplicity, the same link length will be used for data centers with different number of servers. The overall link cost, C_{fiber} is calculated as the sum of the total cost for both MMF and SMF i.e.

$$C_{fiber} = 2 \times N \times 2 \times C_{MMF} + 2 \times N \times 50 \times C_{MMF} + 2 \times N \times 500 \times C_{SMF} \tag{3.2}$$

where C_{MMF} is the cost of MMF, and C_{SMF} is the cost of SMF. The Ethernet switches are assumed to operate at 10 Gbit/s and the total cost of the Ethernet switches, C_{sw} is calculated as the product of the total number of switch ports and the cost per port, C_{ETH}. The number of Ethernet switch ports in a fat-tree is $5 \times N$, thus

$$C_{sw} = 5 \times N \times C_{ETH} \tag{3.3}$$

The total cost of the fat-tree infrastructure can be expressed as the sum of the cost of transceivers, fiber and switches.

Hi-Ring

The following assumptions have been made regarding the Hi-Ring architecture. Similarly as for the fat-tree, the total cost of the Hi-Ring network can be divided into cost of transceivers, fiber and switches. Note that some of the components used in the Hi-Ring network are not even technologies that are mature and commercially available, but are rather in the research phase, so some of them will be omitted from the analysis. The number of servers supported is denoted as N. $2 \times N$ short reach transceivers are used for connecting the servers to the ToR switches and N dense wavelength division multiplexing (DWDM) transceivers are used for transmitting data

from the ToR switches towards the multidimensional switching nodes. The total transceiver cost, $C_{TX/RX}$ is defined as

$$C_{TX/RX} = 2 \times N \times C_{SR} + N \times C_{DWDM} \tag{3.4}$$

with C_{SR} and C_{DWDM} indicating the cost of a short reach transceiver and a DWDM transceiver, respectively. In connection with the short reach transceivers, MMF is used, SMF is used between the ToR switches and the nodes, and MCF is used between the nodes. Although some SMF will be used for connecting the switches within the node, at this point that cost is negligible with respect to the remaining fiber cost and will be disregarded. Moreover, another reason for dismissing this cost is the fact that no fiber will be used if the node is fully integrated. The length of the links is assumed to be 2 m for server to ToR, 50 m for ToR to a multidimensional switching node and 500 m for the links between the nodes. The total number of MCFs, x, depends on the number of nodes in each case. Similar as for the fat-tree, the same link length will be preserved for data centers with different number of servers. The overall link cost, C_{fiber} is calculated as the sum of the total cost of MMF, C_{MMF}, the cost of SMF, C_{SMF}, and the cost of MCF, C_{MCF} i.e.

$$C_{fiber} = 2 \times N \times 2 \times C_{MMF} + 2 \times N \times 50 \times C_{SMF} + x \times 500 \times C_{MCF} \tag{3.5}$$

For comparison purposes, the multidimensional nodes are assumed to operate at 10 Gbit/s and the total cost of a single node is calculated as the sum of the cost of the individual switches at the different layers. The cost of the switches at each level is calculated as the cost per port at the corresponding level, i.e. C_{SDM}, C_{WDM} and C_{TDM} multiplied with the number of ports at each level or N_{SDM}, N_{WDM} and N_{TDM}. Considering that the MCF switch [65] is not commercially available, an assumption will be made that its price will be equivalent to that of a fiber switch. Hence, the cost of SDM ports will include ports for both types of switches, the fiber switch and the MCF switch. The same price and power consumption will be used for both types. Thus, the total cost for switches in the Hi-Ring network can be defined as the cost of the switches comprising the multidimensional nodes and the cost of the ToRs, defined as the product of all ToR ports, and the cost per ToR port, C_{ETH}

$$C_{sw} = N_{SDM} \times C_{SDM} + N_{WDM} \times C_{WDM} + N_{TDM} \times C_{TDM} + 2 \times N \times C_{ETH} \tag{3.6}$$

Furthermore, it is important to note that supplementary components such as optical amplifiers or spatial (de)multiplexers may additionally contribute to the overall cost analysis of the Hi-Ring network. However, these components will be omitted from the current analysis, mainly because it is expected that those components will be integrated with other parts of the node such as the switches, and therefore will not contribute significantly to the overall cost. Additionally, the exact number of required devices can not be known in advance without detailed network planing and power budget estimation. At last, as it will be discussed in the following subsections, the cost of these extra components can be compensated by other means.

CAPEX comparison

Fig. 3.3 illustrates the CAPEX comparison for a fat-tree and a Hi-Ring network. It can be seen that the total cost of both networks is very similar. For a data center with 500 000 servers, the Hi-Ring network has a total cost that is higher than the fat-tree network for around 3 million USD.

The distribution of cost over the different components, i.e. switches, links and transceivers for both the fat-tree and the Hi-Ring network are shown in Fig. 3.4 (a) and Fig. 3.4 (b), respectively. In the Hi-Ring network, the cost of fiber and transceivers is lower compared to the fat-tree network. Although the price per unit of some of the transceivers and fibers used in the Hi-Ring network is higher than the ones used in the fat-tree network, the overall lower cost is mainly due to the fewer transceivers used and the reduced number of links.

The main cost contributor in both networks are the switches. The cost of switches in the Hi-Ring is higher compared to the cost of switches in the fat-tree network. This is mainly due to the fact that several different switching technologies are used in addition to the electrical ToRs. However, this capital investment can easily be justified, if it can be compensated with the annual or few year OPEX savings and a fast return of investment can be achieved.

OPEX comparison

In order to confirm that the Hi-Ring network has lower power consumption due to the transparent switching in the optical domain, the two architectures are compared based on the power consumption and annual

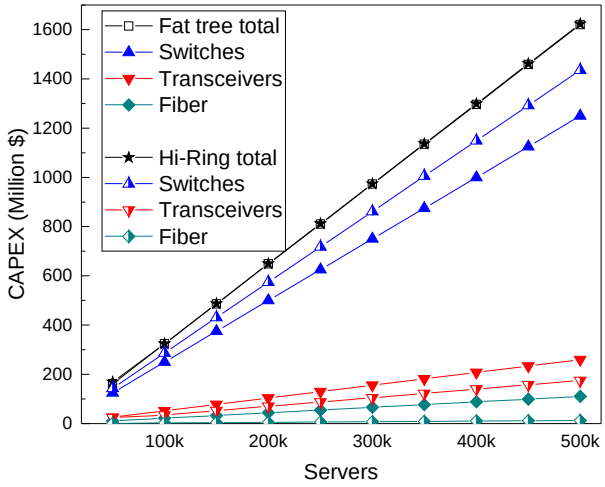


Figure 3.3: CAPEX comparison of a fat-tree and a Hi-Ring network.

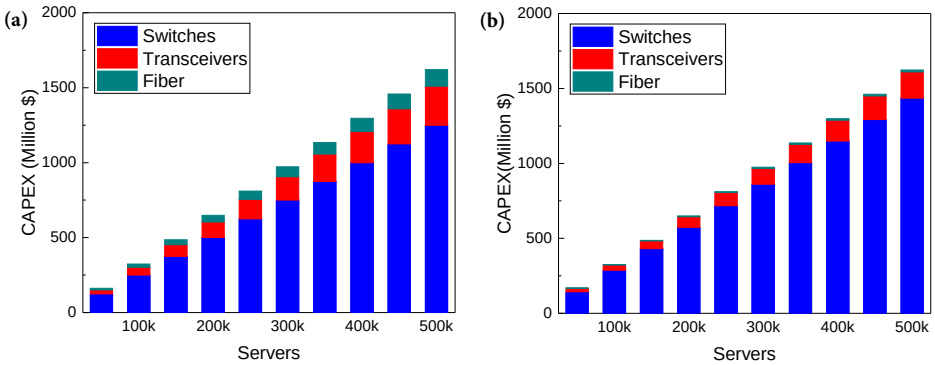


Figure 3.4: Total CAPEX distribution of (a) a fat-tree network and (b) a Hi-Ring network.

OPEX cost. The power consumption of both networks, P is calculated as the sum of the power consumed by switches, P_{sw} and transceivers, $P_{TX/RX}$ i.e.

$$P = P_{sw} + P_{TX/RX} \tag{3.7}$$

The annual OPEX cost is calculated as the cost of annual power consumption considering a price of 0.12 USD per kWh [70]. Fig. 3.5 (a) illustrates the total power consumption and Fig. 3.5 (b) shows the corresponding annual OPEX cost for both architectures. It can be seen that the

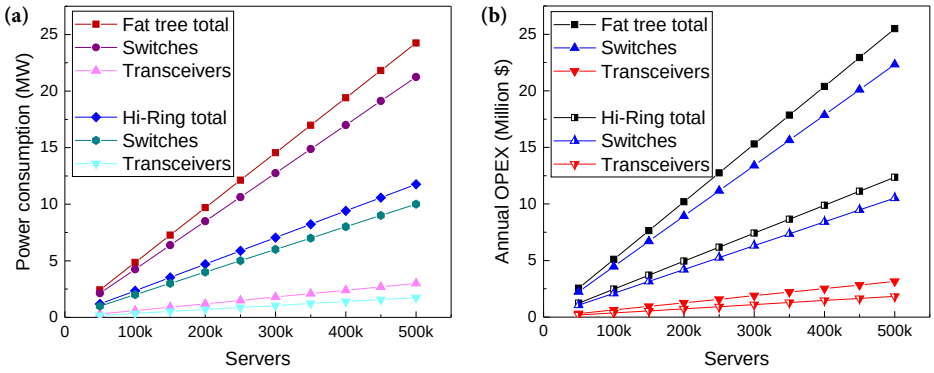


Figure 3.5: (a) Power consumption comparison of a fat-tree and a Hi-Ring network and (b) an annual OPEX comparison of a fat-tree and a Hi-Ring network.

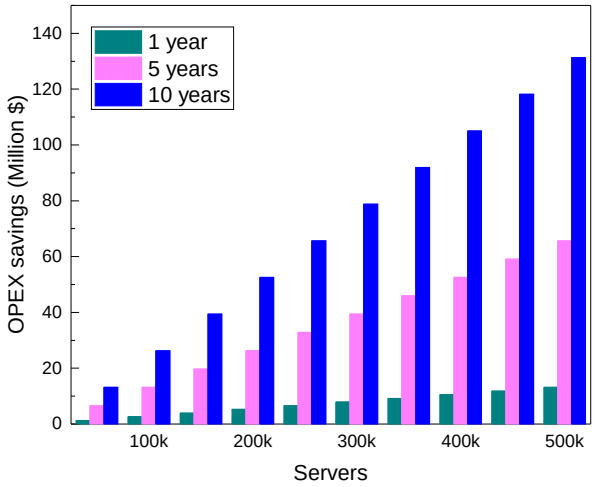


Figure 3.6: Total OPEX savings of deploying the Hi-Ring instead of the fat-tree network for 1, 5 and 10 years.

Hi-Ring consumes less power than the fat-tree which directly translates to lower annual OPEX cost.

Moreover, in order to see what is the time for return of investment, the annual OPEX savings have been used to estimate the OPEX savings for 1, 5 and 10 years. From Fig. 3.6, it is clear that, over a period of one year, the initial CAPEX cost difference can be compensated fully, and a

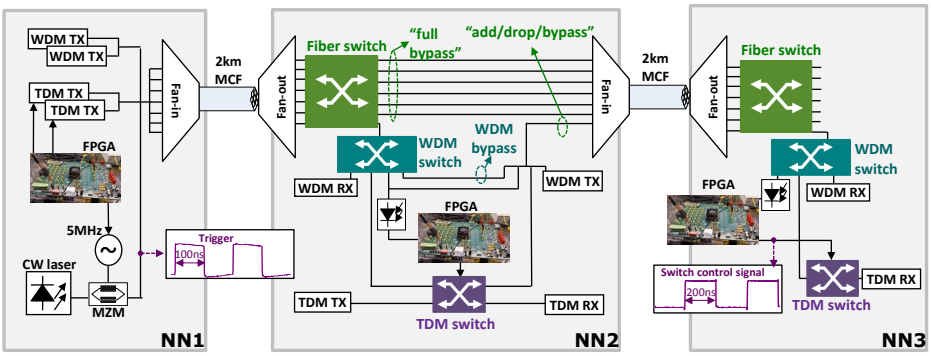


Figure 3.7: Experimental setup for the network prototype consisting of three nodes. (WDM - wavelength division multiplexing, TDM - time division multiplexing, TX - transmitter, RX - receiver, CW - continuous wave, MZM - Mach-Zehnder modulator, FPGA - field programmable gate array, MCF - multicore fiber).

significant amount of savings can be achieved over a period of 5 or 10 years. Considering this, it becomes clear that, even with the cost of additional components omitted in this analysis, the Hi-Ring architecture is a promising solution for an overall cost reduction in future data centers.

3.3 Experimental demonstration

Besides modelling and cost/power consumption analysis, it is also important to investigate the performance of the network experimentally. The experimental demonstration is significant in terms of identifying the parts that need future improvement and will make the architecture a viable alternative to existing solutions.

3.3.1 Experimental setup and scenarios

In order to assess the performance of the network experimentally, a small network prototype consisting of three nodes is built. The experimental setup is shown in Fig. 3.7. The nodes are interconnected with a single 2-km 7-core fiber on each of the two links. For the purpose of spatial multiplexing/demultiplexing of the traffic in the two MCFs, free space coupling devices and spliced fan-in/fan-out devices are used. Two types of connections are established, full wavelength connections and time slotted connections. The used wavelength channels are placed on a 100-GHz grid

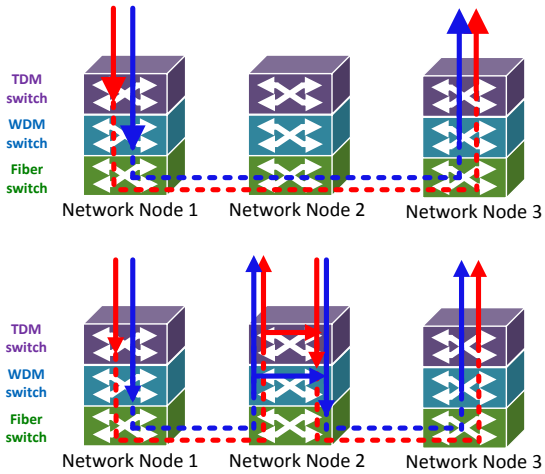


Figure 3.8: Investigated network scenarios. The 'full bypass' network scenario (top) and the 'add/drop/bypass' network scenario (bottom). (WDM - wavelength division multiplexing, TDM - time division multiplexing).

in the C-band and in both cases, 40 Gbit/s on off keying (OOK) modulated data is transmitted.

Initially, the system performance is evaluated when three data channels are used, two for full wavelength connections and a single shared wavelength for slotted operation. Subsequently, the number of channels is scaled to 25 (24 full wavelength connections and one wavelength channel used for establishing two TDM connections). The latter one, demonstrates that the network operation can easily be scaled to 1 Tbit/s/core capacity or 7 Tbit/s throughput using a single MCF.

At the first network node (NN), traffic is generated, aggregated and sent out through the MCF towards NN_2. The connections include two full wavelengths and one wavelength shared among two TDM transmitters in an alternating fashion, i.e. every second slot is allocated to the same transmitter. The used time slot width is 200 ns, out of which 190 ns are dedicated for transmission of useful data and 10 ns are used as a switching gap, which allows for switch reconfiguration between bursts of data that need to be directed towards different output ports.

The data bursts are generated using a field programmable gate array (FPGA), which coordinates the medium access of the two TDM transmitters. For the purpose of synchronous operation of the TDM transmitters and switches, a 5 MHz trigger is modulated on a separate wavelength and

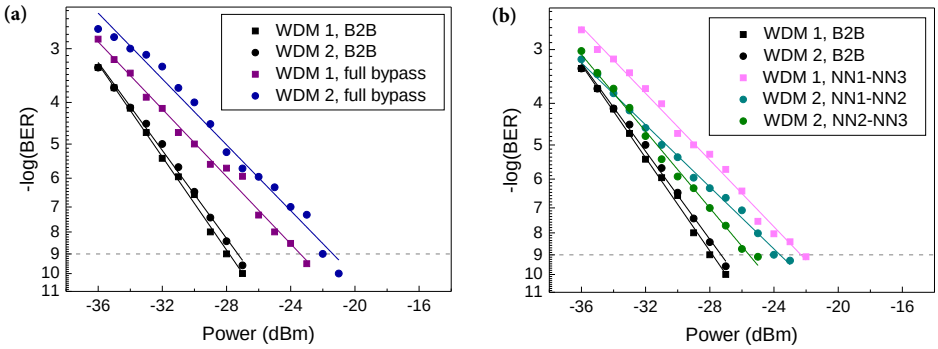


Figure 3.9: BER performance of the full wavelength connections in (a) the 'full bypass' network scenario and in (b) the 'add/drop/bypass' network scenario.

propagated along the network. Synchronization is enabled by detecting the trigger after propagation at each node and using it to determine the slot status. Considering that the trigger propagates along the same fiber that is used for data transmission, it is received with a delay that corresponds exactly to the propagation delay between the first node and any other node. This allows that the nodes can be aligned to the correct slot edge. Moreover, based on the recovered trigger and a counting process implemented on the FPGA, each node can keep track of the slot status, as time elapses. The delayed trigger indicates that the nodes will have a different slot status at a given moment in time, meaning that each slot can be uniquely identified by network elements such as path computation elements (PCEs) or an SDN controller.

NN_2 and NN_3 are composed of switches operating in three switching dimensions that are dynamically configured to demonstrate fully reconfigurable multidimensional switching. Because the nodes are connected with a single multicore fiber on each link, no MCF switch is used. A 48x48 beam-steering Polatis switch is used as a fiber switch at both NN_2 and NN_3, logically segmented into two 24x24 sections. A wavelength selective switch and a $LiNbO_3$ based 2x2 electro-optic switch is used at WDM and TDM level, respectively.

Two switching scenarios are considered as shown in Fig. 3.8. The first scenario is referred to as the 'full bypass' scenario, and it aims at investigating bypass of connections at a fiber core granularity using the space switch. Thus, all connections from NN_1 are destined to NN_3 and are by-

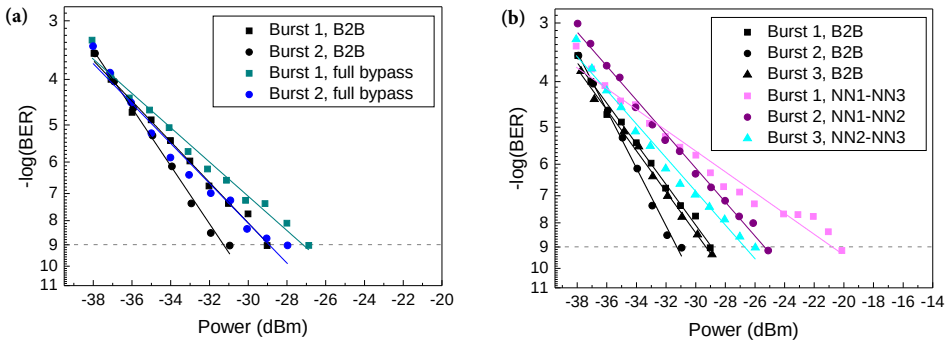


Figure 3.10: BER performance of the individual TDM connections in (a) the 'full bypass' network scenario and in (b) the 'add/drop/bypass' network scenario.

passed at the lowest level at NN_2. The second switching scenario i.e. the 'add/drop/bypass' scenario, aims at investigating the performance when traffic is repacked at intermediate nodes. Hence, the traffic generated and sent out from NN_1 is partially dropped at NN_2 (i.e. one full wavelength and one slotted connection) and the remaining is bypassed at higher levels and combined with traffic generated at NN_2. The burst added at NN_2 is generated based on the propagated trigger, thus the TDM synchronization is retained. At last, both the traffic originating from NN_1 and NN_2 is received at NN_3.

3.3.2 System performance

First, the performance of the system with three data channels is evaluated. The bit error rate (BER) results for the full wavelength connections in both switching scenarios i.e. the 'full bypass' and 'add/drop/bypass' scenario are shown in Fig. 3.9 (a) and Fig. 3.9 (b), respectively. For both cases, error-free performance (BER$<10^{-9}$) is achieved for all the connections. Moreover, the results confirm the preference of bypassing at lower levels in intermediate nodes, which results in lower penalty. However, as indicated by the performance of the first WDM channel, the penalty of bypassing at WDM instead of SDM level in intermediate nodes is relatively low, only 1 dB, which means that a connection could still use such a path in the lack of resources. By observing the performance of the second WDM channel, it can be seen that transmission over two spans of MCF results with additional penalty

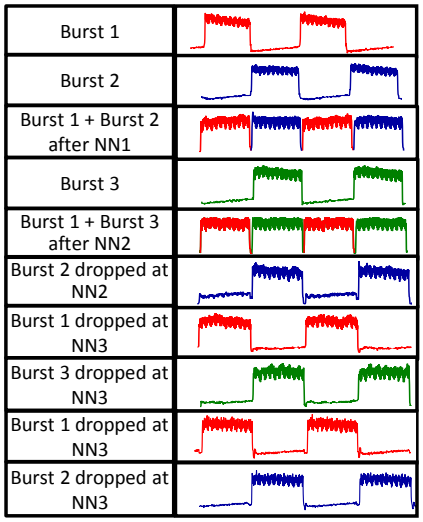


Figure 3.11: Time domain traces of the generated, switched and received bursts at each node. (NN - network node)

compared to a single span and this is mainly due to the impaired optical signal-to-noise ratio (OSNR) and degradation as a result of the crosstalk experienced from the spatial (de)multiplexers.

The BER results of the TDM connections in the 'full bypass' and 'add/drop/bypass' scenario are shown in Fig. 3.10 (a) and Fig. 3.10 (b), respectively. Similar to the WDM connections, all TDM connections have error-free performance (BER$<10^{-9}$). Again, bypass at lower levels in intermediate nodes is preferable, especially because of the limited suppression ratio of the used TDM switches that results in additional penalty due to crosstalk. However, this penalty can easily be minimized by using an algorithm for proper path allocation that takes this into account and limits the number of TDM switches in intermediate nodes along the path. The time domain traces of the bursts generated at NN_1 and NN_2 and dropped at either NN_2 or NN_3 in both scenarios are shown in Fig. 3.11.

Next, the system capacity is increased to 1 Tbit/s/core. Fig. 3.12 illustrates the spectra of all channels in all cores at the input of NN_3, observed through a 20-dB coupler. Only the 'full bypass' scenario is considered in this case. The measured receiver sensitivities (at BER$=10^{-9}$) of all 25 channels in a single core are shown in Fig. 3.13 (a). Moreover, in order to confirm that similar performance is achieved for all channels in the remaining cores, the receiver sensitivity of a representative channel is measured in

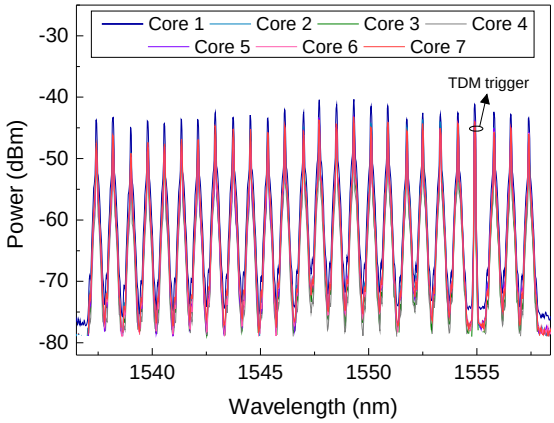


Figure 3.12: Spectra of all the channels in all the cores received at the input of NN_3.

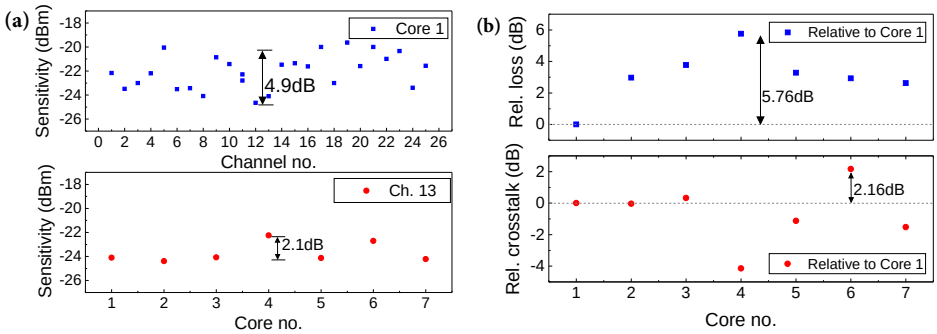


Figure 3.13: (a) Measured receiver sensitivity of all channels in one core (core 1) (top) and of a single representative channel (ch. 13) in all cores (bottom) and (b) Measured relative loss (top) and relative crosstalk (bottom) of the different fiber core pairs.

all the cores as shown in Fig. 3.13 (a). It can be seen that in one core the different channels have receiver sensitivity within 4.9 dB difference, while the receiver sensitivity of the single channel in the different cores is within 2.1 dB difference.

It is important to emphasize that due to the random choice of core pairs in the two different multicore fibers the total loss and crosstalk per fiber core pair were not optimized. Fig. 3.13 (b) shows the measured relative loss and crosstalk of each fiber core pair with respect to the loss and crosstalk of the pair denoted as core 1, on which the receiver sensitivity

of all channels was measured. It can be seen that both the maximum loss difference as well as the maximum crosstalk difference between any of the different pairs remains within a 6 dB margin. Moreover, it can be seen that two worst measured receiver sensitivities of a single channel in all cores (in cores 4 and 6), actually corresponds to the two cores in which the signals experience the highest insertion loss (core 4) and the worst crosstalk (core 6).

3.4 Summary

In this chapter, the Hi-Ring data center architecture was presented. The Hi-Ring architecture is based on multidimensional switching nodes that exploit optical switching in the space, wavelength and time domain. The envisioned switching dimensions allow for support of both long-lived connections i.e. 'elephant' flows, as well as short-lived bursty traffic or 'mice' flows. In addition, the deployment of novel types of optical fibers, i.e. multicore fibers holds the promise to carry a huge amount of traffic over a single fiber, simplifying the cable management and improving airflow.

The node modelling was discussed in details and feasible node structures were presented based on practical limitations. Moreover, the presented cost and power consumption analysis confirms the ability to achieve cost reductions compared to a standard fat-tree network. Although the Hi-Ring network has slightly higher CAPEX, significant savings can be obtained due to the lower OPEX costs.

In addition, the performance of the proposed architecture was experimentally validated on a small network prototype composed of three nodes. The presented results from the experimental demonstration indicate that successful communication with relatively low penalty can be achieved for different types of connections in different network scenarios. Transmitting 7 Tbit/s using a single multicore fiber and obtaining error-free performance implies that the Hi-Ring network can easily be scaled to provide support for immense capacity.

By outlining the main concept, advantages and areas that need improvement, as well as presenting results on the node modelling and system performance, the Hi-Ring network feasibility is reaffirmed. Hence, not only the motivation for optical switching in the data center is clear, but it becomes apparent that optical switching holds the promise to tackle the main challenges of today's data centers.

Chapter 4

Optical subwavelength switching and synchronization

4.1 Introduction

This chapter deals with the concept of optical subwavelength switching and the different aspects of its realization including control and synchronization. First, existing proposals on various types of optical subwavelength switching and their main differences are discussed. Moreover, the general synchronization requirements, as well as proposed synchronization methods are reviewed. In Section 4.2, the concept of optical TDM switching is presented and compared with other proposals. Unlike OPS paradigms, optical TDM switching represents a way to establish circuit-oriented connections with subwavelength granularity. In order to implement this type of switching in a Hi-Ring network, a new synchronization algorithm is developed. The algorithm allows for synchronization in a ring network irrespective of the ring length and operates by continuously estimating the propagation distance and deciding on a slot size. The behaviour of the implemented algorithm is validated experimentally as described in Section 4.3 and used for assisting the data plane operation in a simplified Hi-Ring network prototype. Successful synchronization and precise switching are demonstrated, confirming the feasibility of the synchronization approach. At last, in Section 4.4, the main concepts and results will be summarized. The chapter is based on work published in [J2], [C8].

4.1.1 Overview of optical subwavelength switching concepts

As already discussed, optical circuit switching is a technology used for providing connectivity with coarse wavelength granularity and slow switching speed, thus it is well suited for connections that require high bandwidth over longer periods of time. However, connections may often require only a portion of the full wavelength bandwidth and wavelengths can remain underutilized. In order to improve the bandwidth utilization, it is possible to share wavelengths in the time domain.

In general, any technology that facilitates establishing connections with a subwavelength granularity can be referred to as optical subwavelength switching technology. Considering that a wavelength is shared among several connections, the switching has to be done fast, so that as minimum bandwidth as possible is lost due to reconfiguration. For this reason, fast optical switches that can switch in the nanosecond range have been developed and researched over the years [11, 63, 71–73].

Depending on various characteristics, optical subwavelength switching technologies can easily be classified in different categories. One important feature is the resource reservation i.e. whether the connection is established before data is sent. Hence, one type is circuit-oriented subwavelength connections, where complete resource reservation precedes the transmission. The process of resource reservation can be either static, where a fixed schedule of transmissions is followed, or dynamic, meaning it is flexible and corresponds to the actual traffic demand [74, 75]. Basically, the only difference between this type of connections and conventional circuit-oriented communication is the granularity of the connection.

Another type is the so called packet-oriented communication, where data is sent without allocating resources. This definition will be used further on and it is important not to associate all subwavelength switching technologies with packet switching directly, as the type of connection differs from the connection granularity itself, i.e. packets or bursts of data will still be sent in both cases. The terms ‘packet’ and ‘burst’ may be used interchangeably to indicate a data chunk that can be transmitted in a different way throughout the network. However, ‘packet’ is most commonly associated with packet-oriented transmission, while ‘burst’ is frequently used to indicate some kind of circuit-oriented transmission.

Different proposals exist that fall into one of these two categories. Proposals on optical packet switching [76–80] are an example of a technology that is packet-oriented. Variations of optical circuit-based subwavelength

switching have also been proposed [81, 82]. Optical burst switching [83, 84] falls somewhere in between the two types, as data is sent without full path reservation, but there is a signalling message that is used to indicate the burst arrival. However, there is no guarantee and the burst may still be dropped if a contention occurs.

An important thing to note for circuit-oriented data transmission is that some initial delay will be experienced while the resources are reserved, but once the connection is established, the only delay experienced will be the actual propagation delay. In packet-oriented transmission, instead, the resources are not reserved in advance. On one hand, this results in lower delay at the transmitter side, as the transmitter can start sending as soon as there is data to send. On the other hand, this may result with longer delays or even packets being dropped in case there is a congestion within the network.

Thus, one may also distinguish between the two approaches based on the way they deal with contentions. In circuit-oriented communication, transmissions are contentionless, meaning that contentions are solved before transmitting and do not exists after transmission. In packet-oriented communication, contentions may exist and the probability of a contention to happen increases as the network load grows. Contentions can be solved by the use of different techniques [85], such as optical buffering [86–90], optoelectronic conversion followed by electrical buffering [46, 91, 92], wavelength conversion [93–96], deflection routing [97–100], etc.

Depending on the medium access, a distinction can be made between slotted and unslotted approaches or often refered to as synchronous and asynchronous aproaches, respectively [101]. When the medium access is slotted, a transmission can occur only at the beginning of a specifically chosen interval. On the contrary, an unslotted medium allows that transmissions can occur at any time the transmitter is ready to transmit. Existing studies [102] have shown that the network performance in terms of the blocking probability and the achievable throughput in slotted packet-oriented networks outperforms unslotted packet-oriented networks. Moreover, slotted networks are often associated with fixed size bursts, while switching of bursts with variable size is more common for unslotted networks.

The control of the switch can also be done in different ways. One way is by sending packet headers, either in-band or out-of-band with the data, that carry the control information. This is the typical way of routing optical packets in a packet-switched network. Another way is by using a time

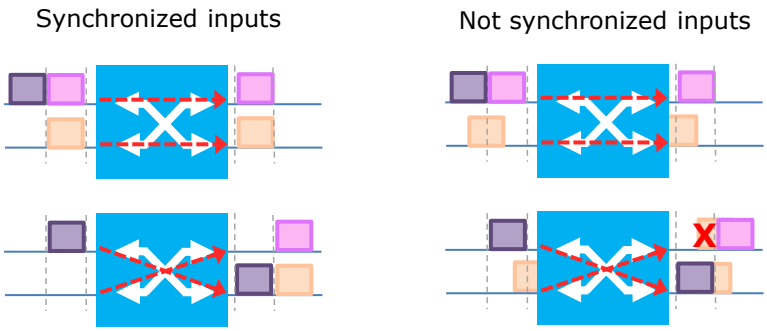


Figure 4.1: Switch operation in an optical subwavelength switched network, when the data arriving at the different inputs of the switch is aligned and not aligned.

reference and a lookup table filled with control information, distributed from a centralized controller to reconfigure the switches, without the need to send headers throughout the network. This, instead, is a common way to control the switches in a circuit-oriented network where transmissions are scheduled and the centralized controller has a full overview of the network. In both cases, when an NxN switch is used to perform the switching, the switch is reconfigured at discrete time intervals. Thus, it is necessary to perform alignment of the data arriving in the different inputs, so that simultaneous switching can occur [103]. This raises the question of synchronization which will be further discussed in the next subsection.

4.1.2 Overview of synchronization requirements

A synchronization mechanism is inevitable for subwavelength operation. The general problem of synchronization is illustrated in Fig. 4.1. It can be seen that the switch is reconfigured from a 'bar' state to a 'cross' state during two consecutive slots. If data arriving on the two inputs is aligned well to the slot boundaries, proper switching can occur. However, if data on the two inputs is not aligned, bursts may be directed to two outputs, losing some portion of the data, as the switch is reconfigured in the middle of the burst.

In order to prevent this, it is necessary to develop methods that will allow for alignment of data arriving on the different switch inputs. One way to achieve this is by the use of synchronizers [104–107]. This approach is common for unslotted networks, since transmissions can occur whenever

there is data to be sent and the delay between bursts arriving at the switch input can vary. In addition, considering that the propagation delay between switches can also be different, this method can also be applied in slotted networks in order to compensate for differences in the link lengths. A synchronizer can be built out of 2x2 fast optical switches and fiber delay lines (FDLs). The switches are connected in a sequence and the length of the FDLs between the switches follows a geometric progression [108] i.e. each FDL has a length of $slot \times (1/2)^x$, with $x = 1, 2, 3....$ This means that the length of the first FDL is such, that it will cause a delay of 1/2 slot duration, the second FDL will cause a delay of 1/4 slot duration, the third FDL will cause a delay of 1/8 slot duration, etc.

By dynamically reconfiguring the switches, a fiber length can be chosen, such that the data is delayed for the correct amount of time and arrives at the switch input aligned with data on the other inputs. A delay mismatch of a full slot can be successfully compensated by such schemes. By deploying synchronizers at each input of the fast optical switches, all inputs can be synchronized. However, the cascaded switching in the synchronizers may affect the signal integrity. The overall experienced insertion loss, as well as the limited suppression ratio of the switches can impair the performance.

A different solution is to use a method for global synchronization, by distributing a control signal along the network, often denoted as a clock or trigger. This allows that each node is provided with a common time reference and synchronous operation can take place. There are different ways to achieve this depending if the method accounts for the propagation delay between the nodes or not. One approach is to distribute a clock along the same fiber infrastructure that is used for data transmission, allowing that the clock experiences the same delay as the data. A second approach is to distribute the same time reference to all nodes, without accounting for the propagation delay between them. This can be done by using a local network for master/slave clock distribution [109, 110] or the Global Positioning System (GPS) [111]. It is important to note that when doing this, additional care has to be taken in adjusting the offset at each node by either deploying fibers with matched length, so that data is automatically aligned at the switch input or using synchronizers as mentioned previously. Although it is much simpler to deploy fibers with pre-engineered link length, this may pose severe practical limitations on the network planning, maintenance and flexibility. Commercial deployment of such a solution may be extremely challenging, considering that all link lengths in the network have to be strictly controlled.

At last, it is worth noting that a certain scheme can be better suited for a specific type of network i.e. packet-oriented or circuit-oriented, or within a specific network topology. For example, synchronization approaches that do not account for the propagation delay are rather independent from the network topology, as the clock distribution can happen through a different network infrastructure. On the contrary, approaches that do consider the propagation delay largely depend on the network topology, as this also defines the clock distribution network. Hence, synchronization in a star or tree network can be performed relatively easy, but other topologies may face additional limitations.

4.2 Optical TDM switching

Having presented an overview of the existing work with respect to both the general concepts of optical subwavelength switching and the problem of synchronization, the main concept behind optical TDM switching and its differences compared to existing approaches will be discussed. Moreover, the proposed algorithm for synchronization in the Hi-Ring network will be described in details.

4.2.1 Switching concept

Optical TDM switching is defined as a circuit-oriented subwavelength switching. Unlike OPS and OBS, optical TDM switching envisions contentionless communication where connections are established in advance and resources are allocated prior transmission. This also means that there is no need for optical buffering within the network i.e. buffering can be pushed towards the network edge, where data can be buffered in the electrical domain. Thus, the need for extra components and additional control overhead is eliminated, simplifying the network deployment and reducing the overall network cost.

The medium access is slotted and a single wavelength is shared by having time slots arranged in a periodic frame structure. A frame can contain an arbitrary number of time slots. A time slot is composed of a portion allocated for data transmission and a gap used to account for the switch reconfiguration time or additional inaccuracies, as it will be discussed further on. Initially, a fixed schedule of uniformly assigned slots exists. However, the schedule can be modified dynamically through an SDN controller, based on the traffic demand. This allows for a trade-off

between on one side, a simple, but rigid connectivity and on the other side, a complex, but completely dynamic connection establishment.

Unless otherwise indicated, it is possible that the current slot allocation can remain valid for the next frame, periodically repeating the schedule. Hence, the SDN controller is only required to configure the network initially and to convey information of future changes. In this way, the amount of control information that has to be exchanged between the nodes and the controller is minimized. Furthermore, in order to take into account the delay between the data plane and the control plane, the SDN controller sends control information for few frames at once, so that by the next update there is sufficient control information for configuring the switches. All the slots have the same duration and switching of bursts with equal size is performed throughout the whole network. However, it is possible to change dynamically the value of the slot duration, allowing for better adaptation to different traffic demands and contributing to the flexibility of the network.

Another important feature is the actual switch control and the time envisioned switching. Namely, the switches are configured based on two attributes: the current time slot at each node and the control information for that time slot that is stored in the lookup table edited by the SDN controller. Hence, it is important to note that, unlike in OPS, synchronization is not only needed to align the data arriving at the switch inputs, but also because each node has to keep track of the time.

This also means that there is no need to transmit the control information in the data plane using packet headers. Thus, better bandwidth utilization can be achieved and more bandwidth can be available for actual data transmission. In addition, there is no need to have bitwise synchronization, as header processing at each switch is eliminated. In this way, the control plane is completely decoupled from the data plane. Furthermore, the network is fully transparent to the traffic that is traversing it, and data processing occurs only at the end sides. Considering that no processing is done at any of the intermediate network nodes, the only delay experienced after connections are established is the physical propagation delay.

In summary, the proposed optical subwavelength switching is a technology much closer to existing optical circuit switched technologies rather than optical packet-switched technologies. The circuit-oriented connectivity allows for contentionless communication by a priori reserving the required network resources, thus eliminating the need of buffering within the network. However, in order to perform time based switching in the Hi-Ring network, it is required that a proper synchronization mechanism is imple-

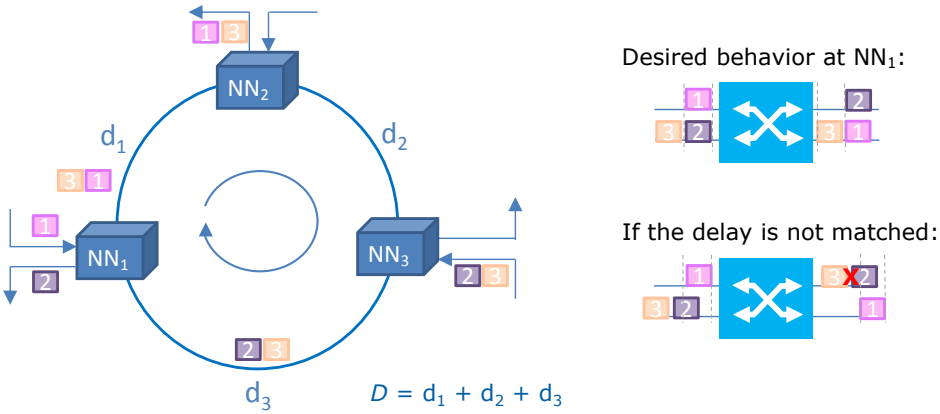


Figure 4.2: The problem of synchronization in a ring network with propagation delay along the ring, D and propagation delay of the individual links, d_1, d_2 and d_3. (NN - network node).

mented, not only to align the data arriving at the switch inputs, but also to be able to provide each node with a time reference that can be used for switching.

4.2.2 Synchronization in the Hi-Ring architecture

In order to provide synchronization among the nodes in the Hi-Ring network, it is necessary to consider the requirements coming from the network topology and the concept of switching itself. The Hi-Ring network is a ring network in which synchronization is required to align data at the input of all switches and to perform time based switching. However, there are additional limitations that stem from the network topology. These will be further discussed in this section and the proposed algorithm for synchronization in the Hi-Ring network will be presented.

Achieving synchronization in a ring network

The problem of achieving synchronization in a ring network is illustrated in Fig. 4.2. For simplicity, it is assumed that there are three nodes in the ring and only switching in the time domain is considered. Each node deploys an NxN fast optical switch to add and drop bursts, thus simultaneous add and drop can be performed only if the add and drop times at each node coincide.

The propagation delay experienced on each link is denoted as d_1, d_2 and d_3, while the total ring propagation delay is labelled as D. Furthermore the following connectivity is envisioned. At NN_1, burst 1 is added to the ring and at NN_2, the same burst is dropped. NN_3 adds two bursts to the ring, one of which has to be dropped at NN_1 and another at NN_2. It can be seen that, at NN_1, the arrival time of burst 2 has to match the add time of burst 1, so that during that slot, the switch can be configured in a 'bar' configuration and both bursts can be directed to the desired port. This will also mean that, during the next slot, the switch configuration will be changed to 'cross' and burst 3 can remain on the ring.

If the delay is not matched, then improper switching will occur and the switch will be reconfigured in the middle of the slot, which will result in data being lost. Thus, in order for successful add and drop operation to be performed at NN_1, the propagation delay between the two nodes has to be an integer multiple of the slot size. This condition also applies for any other pair of nodes in the ring. In general, if the propagation delay on all links in the ring is an integer multiple of the slot size, then the ring can be closed and proper switching can be performed at each node.

The issue of synchronization in a subwavelength switched ring network has commonly been resolved by using pre-engineered fiber links [109–111]. However, as previously discussed, this is a rather impractical solution as it requires that care is taken for the length of each fiber link. Basically, such solution indicates that each link in the network has to have a propagation delay that is an integer multiple of the used slot size.

Synchronization in a ring network independent of the fiber length has previously been proposed in [112]. The presented algorithm considers multi-ring networks and operates by separating the add and drop time at each node. This approach works well when tunable wavelength converters are used in combination with an AWGR, but can not be applied if an NxN fast optical switch performs the switching.

Synchronization algorithm for the Hi-Ring network

In order to circumvent these limitations and allow for synchronization in the Hi-Ring network (and in general, any ring network) irrespective of the propagation delay along the individual links and the ring itself, we have proposed a new synchronization algorithm. The algorithm works by assuming that a global clock is distributed through the same infrastructure that is used for data plane operation. This allows that the offset between the nodes is automatically compensated. For example, the slot boundaries

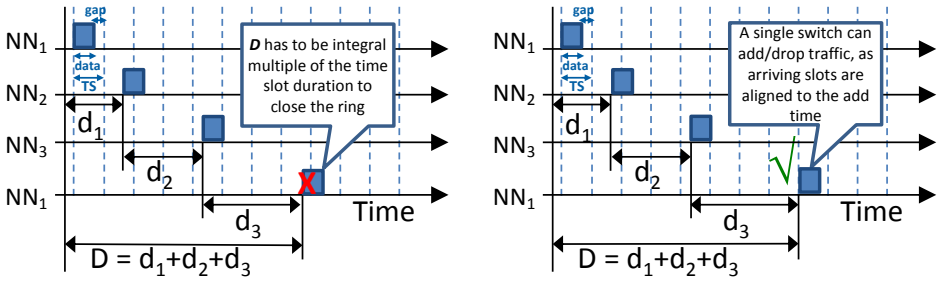


Figure 4.3: Achieving synchronization in a ring network when no requirements are posed on the length of the individual links, but rather only on the overall ring length. Note that the burst arrival time at NN_2 and NN_3 is off the time slot grid applicable to NN_1. (NN - network node).

at NN_1 together with the propagation delay to each node dictate the time at which switching is performed in the remaining nodes in the ring. Thus, if a slot starts at NN_1, the slot boundary at NN_2 will correspond to the slot boundary at NN_1 plus the delay d_1.

This feature is very important as it poses no requirements on the link length between the individual nodes in the Hi-Ring network. As shown in Fig. 4.3, the condition of having a propagation delay that is an integer multiple of the slot duration, does not apply to the individual link segments. If the individual links are observed on a time slot grid corresponding to NN_1, it can be seen that the arrival time of the burst at NN_2 and NN_3 does not have to be aligned to the time slot boundaries at NN_1. Instead the time slot grid at these two nodes will have some offset. This indicates that the links can have an arbitrary length that is not necessarily an integer multiple of the slot duration.

However, removing the individual link conditions results in a new condition in order to be able to close the ring, i.e. the total ring length has to be an integer multiple of the slot size. As illustrated in Fig. 4.3, this means that, if a burst of data is sent from NN_1, after propagation along the ring, this burst will arrive aligned to the slot boundaries of NN_1. Hence, the ring can successfully be closed.

In order to address the problem that the ring propagation delay should be an integer multiple of the time slot duration, a novel synchronization algorithm is proposed. The algorithm operates by first estimating the propagation delay along the ring. Considering that data centers are a closed and well-controlled environment, this can be done reliably and accurately.

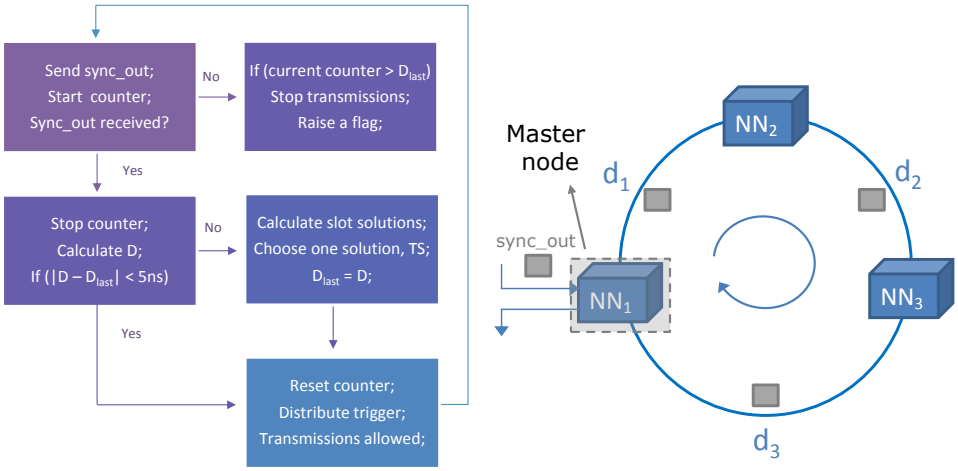


Figure 4.4: Diagram of the proposed synchronization algorithm and its implementation in the synchronization plane. (NN - network node).

Once the propagation delay is known, a slot size is chosen, such that the propagation delay is an integer multiple of the chosen slot size. A diagram of the algorithm flow and the implementation of the synchronization process in the Hi-Ring network is shown in Fig. 4.4.

One node in the ring, denoted as a master node is responsible for running the algorithm, which is implemented on an FPGA. The process of estimating the propagation delay along the ring starts by sending a signal, labelled as *sync_out* from the master node. Once this signal is sent out, a counter is started at the master node to measure the elapsed time. When the signal is received after propagation, the counter is stopped and the propagation delay along the ring, D is calculated. The accuracy with which this delay can be measured depends on the clock frequency used for counting. If a clock with frequency f and period $T = 1/f$ is used, then the accuracy of the measurement is defined by the clock period, T. For example, for the actual implementation of the algorithm, a 200 MHz clock is used, meaning that the propagation delay can be estimated with an accuracy of 5 ns.

The measurement inaccuracy is something that can easily be accounted for and compensated using the switching gap. The switching gap duration is mainly defined by the physical effect used for switching as well as the driving electronics. As long as any other inaccuracies are smaller or of the

same order as the main contributor of the switching gap, their impact to the overall system is not significant. In general, the switching time of fast optical switches is around 10 ns, thus any contribution to the switching gap can be considered negligible if it is lower or of the same order as this value.

Additionally, achieving a high precision propagation delay measurement is something that can be done by using higher frequency clock for counting. Considering that most FPGA boards [113,114] today deploy oscillators up to 1 GHz, this is a practically feasible solution. Although, more sophisticated equipment can be implemented in order to achieve high precision, reducing this contribution to the switching gap, will not affect the overall system much, as the gap will still be dominated by the rise and fall time of the optical switch.

Once the delay is calculated, the value is saved and assigned to the variable D_{last}. This allows that any further measurements can be compared to the last measured propagation delay. Then, the algorithm starts a search for slot durations such that the propagation delay is an integer multiple of the given slot size. The search is performed within predefined slot boundaries, i.e. the minimum, TS_{min} and the maximum, TS_{max} acceptable slot size. These values can be given as input parameters to the algorithm and can be modified each time before running the algorithm, if needed. Only the solutions within this range are considered feasible. The acceptable slot solutions are the ones that satisfy the following conditions

$$for\ (TS_{min} < TS_i < TS_{max})$$
$$if\ (mod(D, TS_i) == 0)$$
$$Save\ TS_i.$$

Depending on the range of acceptable slot size, for some cases it may be difficult to find a broad range of solutions or to find solutions at all. In order to avoid this, it is possible to make exceptions and accept solutions that would work for a propagation delay that is close to the exact delay measured, i.e. a delay that is a few clock periods shorter or longer than the measured one. Of course, by doing this, an extra inaccuracy is added and this also has to be compensated using the switching gap.

Besides when a solution cannot be found, accepting neighbouring slot solutions may also be beneficial in cases when there are acceptable solutions. This can be done for the purpose of expanding the space of acceptable solutions. The reason for doing this is to increase the flexibility of the system by allowing that a slot size is chosen, that is better matched to the

actual demand. Hence, the better the accuracy of the measurement process is, the more can the system flexibility be improved and a wider solution space can be provided for the cost of a relatively small contribution to the switching gap.

After finding the acceptable slot solutions, a single one can be chosen to be used in the data plane. It is assumed that the used frame will have a duration corresponding to the ring propagation delay and depending on the chosen slot solution, it will consist of a different number of slots. Each network node should keep entries for the slot configuration of at least one frame. For the first implementation of the algorithm, one solution is randomly chosen by the FPGA and used for establishing subwavelength connections in the data plane. A fixed switching gap of 25 ns is used irrespective of the chosen solution.

In a real network scenario, the slot solutions can be disclosed to an SDN controller and the decision of the slot size can be made by the controller based on different aspects, such as the real traffic demand. This allows that the network performance as a whole can be optimized by choosing a solution that can provide reduced blocking, lower latency, etc. The gap could also be adjusted such that the minimum required gap is implemented based on the chosen solution. Considering that the SDN controller has a global overview of the data plane operation as well as the solution space, it can also decide to dynamically change the slot solution for another one that is better suited. Thus, the algorithm can also provide adaptability to real traffic demands and facilitate network dynamicity and flexibility.

After deciding on a slot size, the SDN controller can inform the master node that is responsible for the synchronization of the network nodes as well as all transmitters. The master node is responsible of conveying this information to all nodes in the Hi-Ring network by distributing either a trigger or a clock signal. A trigger only initiates the counting process at each node based on a local clock, while a clock signal besides initiating the process, also provides the actual clock for counting. In both cases, the propagation delay is accounted for and the counting starts with a specific offset at each node. At this point, all nodes become synchronized, allowing that transmissions within the data plane can start.

The fact that an approach for distributing the clock through the same fiber infrastructure is taken, indicates that the propagation delay is automatically compensated. Hence, there is no need to implement offsets for when exactly the control signals for reconfiguration of the switches at the different nodes are applied, but rather nodes along the ring will receive a

delayed clock and will be automatically synchronized. This allows that a slot can be uniquely identified in the network, i.e. a burst of data sent in slot 1 at the first node will arrive at any other consecutive node when the local node status is slot 1. The SDN controller can also take advantage of this feature and use the slot identifier as a full connection identifier when reconfiguring the switches on the desired path.

In order to be able to provide automatic resynchronization, in case there are changes to the propagation delay of the ring, it is possible to implement a continuous measurement of the propagation delay. This can be done by sending the *sync_out* signal whenever it is received. The same procedure follows once the signal is received back. In addition, in order to avoid frequent reconfigurations of the slot size for small drifts, the measured propagation delay can be compared to the last measured and if the difference is within an acceptable range, the same slot size can be kept. For example, if the measured delay is one period off, i.e. one clock period less or more than the last measured delay (-5 ns or +5 ns), no change will be made. Of course, it is required that the implemented switching gap can account for this inaccuracy. If the difference is greater than the specified limit, a search for new slot solutions will begin again and the process will continue as initially described.

Considering that the *sync_out* signal is sent out only after it is received at the master node, in the worst case, a change of the propagation delay can be detected after two round trips. This is due to the fact that it takes one round trip to be able to detect the change. However, the speed of resynchronization can be improved in several different ways. A simple way is by increasing the frequency at which the *sync_out* signal is sent. Instead of sending it only when one round trip has been made, it is possible to send it few times per one cycle. For a relatively short period of the *sync_out* transmissions, the resynchronization time can be reduced to around one round trip.

It is also important to consider the environment where the algorithm will be applied and used. As the main purpose is to use it within the Hi-Ring network for data center applications, the ring length will be relatively short, i.e. in the order of tens of kilometers, in the worst case. Thus, a reaction time on the order of a single round trip is believed to be rather acceptable. Additionally, the main reason for variation of the propagation delay is the temperature dependence of the refractive index of the optical fiber. Standard optical fibers can experience delay variations on the order of 40 ps/°C/km [103, 115]. Considering that DCNs are a well controlled

environment [23], extreme temperature variations are not expected. Thus, even a fluctuation of 10 °C would result with only 4 ns variation on the propagation delay in a ring of 10 km, which can easily be accommodated in the switching gap.

The synchronization algorithm can also be used to provide synchronization in a network where several time shared wavelengths are used, as it is also envisioned in the Hi-Ring network. The typical fiber dispersion of a standard fiber is around 20 ps/nm/km. Therefore, considering a ring of 10 km, the delay variation for a 30 nm wavelength span is around 6 ns. Since this can be compensated in the switching gap, it is fair to claim that the proposed scheme is WDM compatible. Thus, simultaneous switching of bursts carried on different wavelengths can be achieved using the same control signal for switching. This simplifies the overall implementation and operation as no dispersion compensation is required before switching. Moreover, as the Hi-Ring envisions the use of MCFs, it is important to note that the synchronization algorithm is also SDM compatible, i.e. it can automatically provide joint synchronization for bursts in all cores of the MCFs.

At last, the synchronization algorithm can also provide additional functionalities with respect to failure detection within the network, by performing some additional processing. One way is to implement a condition that evaluates what actions should be taken if the *sync_out* signal is not received after propagation. An apparent indicator that something has changed in the network is the absence of the *sync_out* signal and a counter that exceeds the value of the last measured propagation delay. In that case, the master node can stop the transmissions in the data plane until resynchronization is achieved. Moreover, the master node can send a message to the SDN controller indicating that a possible node or link failure has occurred, thereby assisting protection and restoration schemes.

4.3 Experimental demonstration

In order to validate the behaviour of the proposed synchronization scheme, the algorithm is implemented on an FPGA. First, the propagation delay along the ring is changed and the ability to provide automatic synchronization is investigated. Then, the algorithm is applied to a simplified Hi-Ring network prototype and used to provide synchronization. After a slot size is chosen and synchronization is achieved, data bursts are transmitted, switched and received.

4.3.1 Algorithm validation

To confirm that the proposed algorithm can provide synchronization in a ring network, as well as automatic resynchronization, a simple validation test is performed. A ring network composed of three nodes, identical to the one shown in Fig. 4.4 is used. The nodes are connected using a single 7-core fiber. The fiber is 2 km long and a set of two cores are used on each link. One core is used to transmit the *sync_out* signal on each link and another core is used to distribute the trigger for synchronization and to transmit data.

NN_1 acts as a master node and runs the synchronization algorithm that is implemented on an FPGA. The *sync_out* signal is generated from the FPGA and used to modulate a continuous wave (CW) laser at 1549.72 nm. A 200 MHz clock is used for counting during the process of measuring the propagation delay along the ring, allowing that the delay can be measured with a precision of 5 ns.

For the purpose of validation and in order to investigate if the algorithm can detect different fiber lengths and act accordingly, the fiber length along the ring is varied on the link between NN_3 and NN_1. The detailed experimental setup of the synchronization plane is illustrated in Fig. 4.5. Initially, one reference fiber length is used to measure the propagation delay. Then, the fiber length is reduced in one case and increased in another, in order to see if the algorithm can detect the change.

The time domain traces of the received *sync_out* signal after propagation for the three different fiber lengths used are shown in Fig. 4.6 (a). It can be seen that the received signal has a different period depending on the measured propagation delay. Furthermore, the reference measured propagation delay either increases or decreases when the fiber length is changed.

Once the propagation delay along the ring is measured, the process of selection of the correct slot solution is verified. The acceptable range of slot solutions is defined within 100 ns and 1 μs. It is assumed that a transmitter connected to NN_1 will start sending data bursts in each slot after a slot size is chosen. By recording the time domain traces of the generated bursts at NN_1, the chosen slot size can be monitored and the behaviour of the algorithm can be verified.

Fig. 4.6 (b) illustrates the time domain traces of the generated bursts for the three different propagation delays measured. It can be seen that, for a propagation delay of 30.625 μs and 30.770 μs, the algorithm decides on a slot size for which the propagation delay is exactly an integer multiple

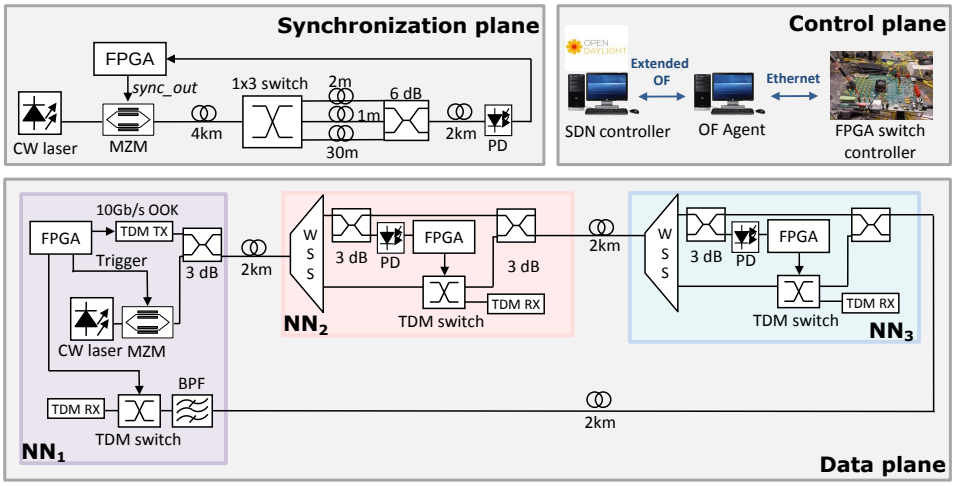


Figure 4.5: Experimental setup for validating the algorithm implementation and assessing the data plane performance when the proposed algorithm is used for providing synchronization. (NN - network node, CW - continuous wave, SDN - software defined networking, OF - open flow, WSS - wavelength selective switch, TDM - time division multiplexing, TX - transmitter, RX - receiver, OOK - on off keying, BPF - bandpass filter, PD - photo-diode, MZM - Mach-Zehnder modulator).

of the time slot size, i.e., 125 and 181 time slots of duration 245 ns and 170 ns, respectively.

For a propagation delay of 30.620 μs, the algorithm selects a slot size such that the ratio propagation delay vs. slot size is not an integer (157.026). This fiber length has been chosen on purpose, in order to investigate the algorithm behaviour for cases when a slot solution cannot be found. 157 slots with a duration of 195 ns are a correct solution for a propagation delay of 30.615 μs. As previously discussed, when a solution cannot be found (30 620 ns has no factors in the range of 100–1000 ns), the algorithm can consider neighbouring solutions, as it is this case. Thus, the chosen solution indeed conforms to the expected behaviour and the implementation of the algorithm complies with the described concept.

4.3.2 Data plane operation

After verifying that the algorithm performs accurate measurement of the propagation delay and correctly chooses an acceptable slot solution, the performance in the data plane is assessed. It is assumed that first the

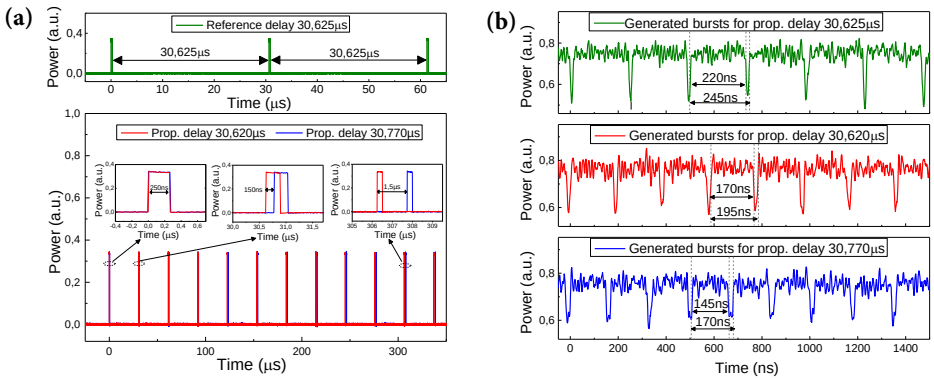


Figure 4.6: (a) Time domain traces of the received *sync_out* signal after propagation. (b) Time domain traces of the generated bursts for the different propagation delays measured.

algorithm runs and performs initial synchronization. Once the delay is measured and a slot size is chosen, the data plane operation can start. The same network is used, composed of three nodes connected in a ring. For simplicity, only the TDM level is implemented and the other switching levels from the multidimensional node structure are omitted. The detailed experimental setup, illustrating the implemented data plane, control plane and synchronization plane is shown in Fig. 4.5.

A single ring length is considered, with a propagation delay along the ring of 30.625 μs. The slot size for the data plane is the one chosen from the algorithm (245 ns) as discussed before. Three cores of the MCF are used for transmitting the *sync_out* signal and measuring the propagation delay along the ring. It is important to note that the propagation delay is continuously measured during the data plane operation. Other three cores are used for distributing a trigger for synchronization and transmitting the data bursts.

The *sync_out* signal is sent using the 1549.72 nm wavelength channel, while the trigger for synchronization and the data bursts use the 1550.12 nm and 1549.32 nm wavelength channels, respectively. NN_1 acts as a master nodes and after a slot size is chosen, it distributes the trigger along the network. At NN_2 and NN_3, the trigger and the data are separated using a WSS. Then, the trigger is split, allowing that one copy continues to propagate along the network. The second copy is used by the local node for synchronization. It is detected using a photo-diode (PD) and the corresponding output is passed to the FPGA.

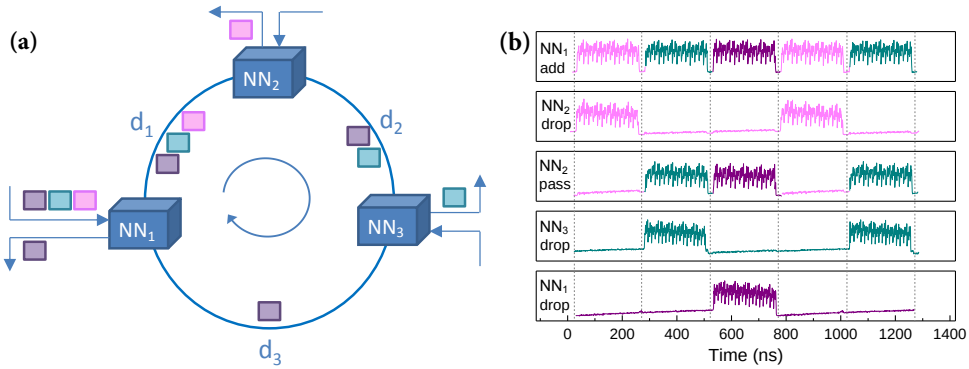


Figure 4.7: (a) Overview of the implemented data plane connections. (b) Time domain traces of the generated and switched bursts at the different nodes. (NN - network node).

Once synchronization is achieved, transmissions are allowed and a continuous stream of 10 Gbit/s OOK modulated data bursts are generated and added to the ring at NN_1. Fig. 4.7 (a) gives an overview of the envisioned data plane connectivity. One burst is dropped at each node, confirming that the ring can be closed successfully. At NN_1, only the data is filtered out using a band-pass filter (BPF) and the same signal that is used to control the transmissions is used to generate the control signal for the switch, confirming that the add and drop times coincide and the ring is perfectly synchronized.

Three 1x2 $LiNbO_3$ switches are used as TDM switches at each node. The switches are controlled by an Open Daylight (ODL) SDN controller [116]. The controller sends the control information regarding the slot configuration at each of the nodes to an Open Flow agent [117]. The agent, in turn, compiles an Ethernet frame with a predefined structure and sends it to the FPGA. The FPGA, based on the received information, generates digital signals for driving the three TDM switches. A subwavelength connection is established by configuring all the switches along the data path, i.e. connecting the relevant input switch port to the output switch port throughout the duration of a specific slot.

Although the same FPGA is used for enabling the synchronization and configuring all the different switches, the different operations are logically separated. For example, the switching at each node is performed based on the control information received from the SDN controller, but the slot status

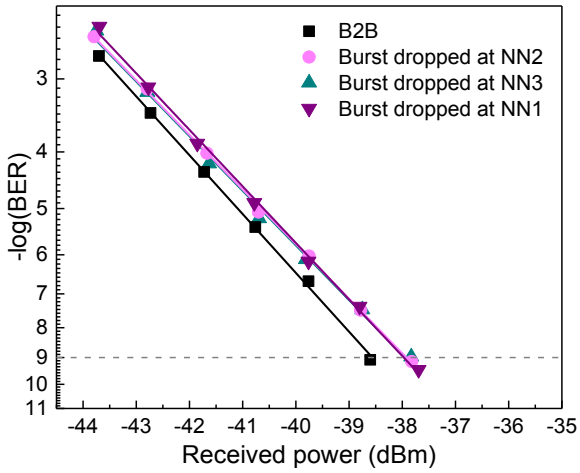


Figure 4.8: BER performance of the bursts dropped at each node. (NN-network node, B2B - back to back).

at each node depends on the received trigger. Thus, the control signals for the different switches are generated at different time slot grids, as it was previously discussed and illustrated in Fig. 4.3.

Fig. 4.7 (b) illustrates the recorded time domain traces of the generated and switched bursts. It can be seen that, at NN_1, bursts are generated in each slot with the correct slot duration. At NN_2, one burst is dropped, while the remaining data is directed towards NN_3. Similarly, at NN_3, one burst of the incoming data is dropped, while the remaining data is sent towards NN_1. Finally, at NN_1, the switch is configured in a bypass configuration by default, meaning that the burst will remain on the ring unless otherwise indicated. Hence, only by configuring the switch at NN_1 to drop the last burst, data can be received.

Fig. 4.8 shows the BER results of the bursts dropped at each node as a function of the received power. It can be seen that the penalty experienced by each burst dropped at different nodes is only 1 dB compared to the back-to-back BER performance. Considering that all nodes are synchronized and bursts are generated with the correct size confirms that the implemented synchronization algorithm operates as desired. Moreover, the proper switching at each node in the ring indicates that the algorithm can be practically used to facilitate the data plane operation in an optical subwavelength switched network.

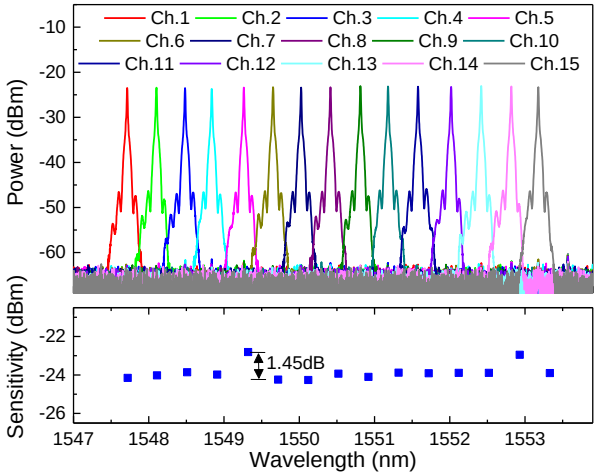


Figure 4.9: Spectra and receiver sensitivity of all channels on a single burst dropped at NN_2.

In order to demonstrate the scalability of the scheme and to confirm that it is WDM compatible, the number of wavelength channels carrying slotted data is increased to 15. The used channels are spaced on a 50 GHz grid and fall within the range of 1547.72 nm to 1553.33 nm, inclusive. Again, bursts are continuously generated in each slot at NN_1 and carry 10 Gbit/s OOK modulated data. The trigger is distributed using the wavelength channel at 1553.73 nm and the *sync_out* signal using the channel at 1547.32 nm. No dispersion compensation is done for the purpose of switching, and dispersion is compensated only at the receiver using dispersion compensation fiber (DCF).

Fig. 4.9 illustrates the spectra of all the wavelength channels dropped at NN_2 and the measured receiver sensitivity (BER=10^{-9}) of a single burst on all the wavelength channels. The individual bursts carried on all the wavelengths are switched using the same control signal for the switch, to confirm the compatibility of the scheme with WDM systems. It can be seen that all channels have similar performance (within a 1.45 dB margin). Compared to the single channel scenario, worse performance is observed i.e. higher receiver sensitivity is measured for each channel. This stems from the fact that due to power limitations of the used devices, the total optical power had to be kept constant as the number of channels was increased, hence amplification at a reduced power per channel resulted in OSNR degradation. However, the error-free performance of all channels with small margin con-

firms that the same control signal can be used to switch bursts on different wavelengths without observing synchronization errors. Considering that no dispersion compensation is performed prior switching indicates that the used switching gap is sufficient in order to compensate for possible offsets between the bursts that are carried on different wavelengths. Hence, the WDM compatibility of the scheme is undoubtedly confirmed.

Moreover, as the propagation delay is measured in a core different than the one used for data transmission, the ability to use the algorithm in systems deploying SDM technologies is verified. The presented results not only reaffirm the feasibility of the proposed synchronization algorithm, but also demonstrate its compatibility with both WDM and SDM systems, and the potential of optical subwavelength switching technologies for practical deployment and implementation.

4.4 Summary

In this chapter, an overview was given of the main concept behind optical subwavelength switching, as well as the different existing proposals for its implementation. The synchronization requirements in an optical subwavelength switched network were discussed in details and related work was reviewed. The concept of optical TDM switching was proposed and a thorough analysis of the synchronization requirements for the Hi-Ring network was presented.

Based on the main synchronization requirements outlined, a new algorithm was proposed, that allows for synchronization in a ring network with arbitrary link lengths and irrespective of the ring propagation delay. The algorithm operates by measuring the propagation delay along the ring and deciding on a slot size such that the propagation delay is an integer multiple of the chosen slot size. Besides providing support for synchronization, the algorithm can also be used to facilitate provisioning of additional network functionalities such as failure detection and restoration. Moreover, an SDN controller can dynamically make the decision of the slot size, allowing for adaptability to the current traffic demand.

The algorithm behaviour was validated experimentally on a ring network composed of three nodes. The propagation delay was measured accurately for three different fiber lengths and the correct slot solution has been chosen in each case. Furthermore, using the implemented algorithm, successful synchronization was achieved among the three nodes and data plane operation was experimentally demonstrated. Bursts with the right

slot size were generated and switched at each node, confirming that the algorithm operates properly.

At last, by using different fiber cores and switching bursts carried at different wavelength channels using the same control signal, the SDM and WDM compatibility of the algorithm is confirmed. Based on the verified behaviour and demonstrated performance, the algorithm, with no doubt, can be implemented easily in practice and can help pave the way toward commercially deployable optical subwavelength technologies.

Chapter 5

Software controlled on-chip integrated data plane

5.1 Introduction

This chapter focuses on two different aspects for future development of the Hi-Ring architecture. The first one is integration, including physical component integration, as well as software-hardware integration. Both of these features are extremely important for future-proof deployable technologies. The second aspect is provisioning of additional network functionalities in the Hi-Ring network. The use of optical switching for providing support to multicast and traffic grooming is considered. In Section 5.2, possible approaches to integration of the components of the multidimensional switching nodes will be reviewed. Existing proposals, as well as fabricated devices will be outlined. Moreover, the process of integration of the individual switches with an SDN controller will be elaborated and SDN - controlled operation will be demonstrated for different types of switches. In Section 5.3, provisioning of additional network functionalities, such as support for multicast and incast, will be discussed. The results from an experimental demonstration, using an on-chip integrated fiber switch with fan-in and fan-out devices, for the purpose of enabling unicast switching, multicast and incast in the Hi-Ring network will be presented, confirming the feasibility of the envisioned network scenarios. At last, in Section 5.4, the main concepts and results will be summarized. The chapter is based on work published in [J1], [J2], [C8].

5.2 Software controlled on-chip integrated multidimensional switching nodes

There are several important features that a future-proof data center network has to posses. One of them is the ability to physically integrate network components used in the data plane. On-chip, hardware component integration allows for building compact and cost effective systems, that are commercially attractive. Another crucial feature is the integration of hardware in the data plane with software in the control plane. Software defined networking has become increasingly popular over the last decade and it is paramount for any data center architecture. Having in mind these two features, existing efforts and future work on the Hi-Ring architecture and on-chip integration, as well as software-hardware integration will be discussed.

5.2.1 Approaches to hardware integrated mutidimensional switching nodes

As briefly mentioned in Chapter 3, there are different approaches to the actual physical structure of the multidimensional switching nodes. Adopting a modular design, by keeping all switching elements separate, may provide easier upgrade and debugging. However, in order for these technologies to become commercially attractive, integration is inevitable.

Integration is extremely important for achieving low footprint and compact devices that can be jointly powered and cooled, allowing for reduced power consumption. Moreover, this is the only way to tackle commercial issues such as fabrication and packaging, as well as improving the failure in time (FIT). Among different platforms, silicon photonics seems the most promising, mainly because of the mature complementary metal-oxide-semiconductor (CMOS) infrastructure and the ability to integrate photonic components with driving electronics. This yields low production cost and paves the way towards commercialization of optical technologies for data center application and deployment.

Although, full node integration on a single platform may be challenging due to the different nature of the individual components, the benefits of integration can be exploited even for partially integrated approaches, where only few node components are integrated on a single chip. One example is the silicon photonic integrated circuit (PIC) proposed in [J3] and [C6]. The PIC is composed of fan-in and fan-out devices for spatial

(de)multiplexing of a 7-core fiber and a 7x7 fiber core switch. This allows coupling a multicore fiber directly on-chip and all the operations like spatial (de)multiplexing and switching are taken care of by the chip. The switch is thermally controlled by configuring a heater in each of the 57 Mach-Zehnder interferometer (MZI) structures. The chip has been used for experimental demonstrations and integrated with an SDN controller, as it will be further discussed below. This device is a good example of a partial integration of the node components, which results with a low footprint chip (12 mm x 5 mm). Fabricating an array of these devices allows that a high amount of components and a subset of the node can be fabricated as a single on-chip integrated device with relatively small size.

Additionally, future work on node integration includes increasing the number of components that are integrated on a single platform. Integrated solutions composed of switches operating at the different layers of the multidimensional switching nodes, as well as integrated circuits (ICs) consisting of additional node components, such as optical amplifiers and spatial demultiplexers, have already been demonstrated using different technologies. A packaged and fiber-coupled gain-integrated Si photonic carrier with silicon nitride (SiN) waveguides and flip-chip attached SOA array has been deomonstrated in [118] and a grating coupler array on the silicon on insulator (SOI) platform for spatial (de)multiplexing has been fabricated and demonstrated in [119, 120].

With respect to switching devices, several works have presented integrated solutions of different types of optical switches, that have been fabricated and/or packaged. A silicon photonic MEMS switch with high port radix and submicrosecond switching time, suitable for a fiber switch operation, has already been demonstrated in [58, 121, 122]. An SOI fully integrated, fast reflective slot-blocker with nanosecond switching time, has been demonstrated in [123]. MZI based, nanosecond switching time, SOI switches with different switching matrices have also been demonstrated in [63, 71, 72, 124, 125]. An integrated silicon photonic circuit for wavelength selective switching has been fabricated in [126]. A cascade of micro ring resonators (MRRs) as wavelength selective elements has also been proposed to realize on-chip wavelength selective swtching [127].

This work indicates that integration of the different components of the multidimensional switching nodes is a feasible approach. Moreover, it confirms that integration holds the promise of fabricating market-ready deployable products.

5.2.2 Software controlled multidimensional switching nodes

Software defined networking is based on physical separation between the data plane and the control plane, allowing for control functions implemented in software to be decoupled from the actual hardware devices deployed in the data plane. The main motivation for a separate control plane is the flexibility, that enables rapid adaptation to innovation. Considering that the control functions are separated from the hardware platform, new services can be introduced easily. In addition, besides enabling control of programmable hardware, SDN offers great support for network function virtualization (NFV) [128]. NFV is a rapidly emerging network architecture concept that uses virtualization to create virtual instances of network node functions that can provide communication services.

Different communication interfaces, i.e. vendor specific or open source interfaces can be implemented between the control and data plane with the goal to support SDN. Open Flow [117] is an open source interface which enables that a common interface is established between the data plane and control plane. Hence, vendor specific protocols are completely eliminated and a single control plane can be used to control devices from different manufacturers. This reduces the cost and simplifies the operation significantly, as it removes the need of having several software controllers for different components. Furthermore, it enables that interoperability issues between the different controllers are completely eliminated.

There are two main approaches for implementation of the control plane, a centralized and a distributed approach. While a distributed control approach allows for locally optimized performance at a single node in the network, a centralized controller has a holistic view on the network, resulting in network-wise optimized throughput and improved resource utilization. Google's globally deployed wide area network (WAN), B4, that is used to interconnect Google's data centers around the world, is a great example of this. Using Open Flow and SDN with centralized traffic engineering (TE), nearly 100 % utilization of the network links in B4 has been achieved and reported [129].

Over the past few years, several different open SDN controllers have emerged, such as OpenDaylight [116], Floodlight [130], open network operating system (ONOS) [131], Ryu [132], etc. Moreover, besides controlling the network, control of other components, including storage and computing resources, as well as overall orchestration, has developed under the open source software platform for cloud computing, Open Stack [133]. This

Ethernet Frame for TDM Switch Configuration																																	Eth. Frame lines (Captured by Wireshark)
ADDR. / BYTE	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	
																			SN		Source MAC Address						Destination MAC Address					0
00	RESERVED																																10,20
01	RESERVED																																30,40
02	Port 1: Configuration for time slots 01 to 32																																50,60
03	Port 1: Configuration for time slots 33 to 64																																70,80
04	Port 1: Configuration for time slots 65 to 96																																90,a0
05	Port 2: Configuration for time slots 01 to 32																																b0,c0
06	Port 2: Configuration for time slots 33 to 64																																d0,e0
07	Port 2: Configuration for time slots 65 to 96																																f0,100
08																																	110.120

Figure 5.1: Structure of the Ethernet frame that is used between the SDN agent and the FPGA for configuring the TDM switches.

clearly illustrates the trend of decoupling the software control from the hardware components.

Compatibility with SDN paradigms is extremely important for commercial technologies and crucial for the deployment of any developing technology. Especially, new, on-chip fabricated devices have to demonstrate the ability to interconnect to an SDN entity, through which they can be remotely controlled and configured. Thus, as part of this Ph.D. project, attention has been focused not only on evaluating the performance of the data plane, but also striving to provide software controlled operation using a common SDN controller. The process of integrating the different hardware devices with a software-defined control plane will be further described in details.

5.2.3 Software controlled switching and data plane operation

Software controlled operation in the data plane can be realized by having a centralized SDN controller that communicates with SDN agents using Open Flow. One or several switches are most commonly controlled by a single SDN agent. The SDN agent receives the Open Flow messages from the SDN controller and is responsible for conveying the control information to a hardware, like an FPGA or application specific integrated circuit (ASIC), that can generate the control signals required to configure the switch. Ideally, the agent would run using the microprocessor on the same FPGA or ASIC. However, in order to simplify the implementation process, both the SDN controller and the agent run on a computer and an FPGA is used to interface with the different switches in the data plane.

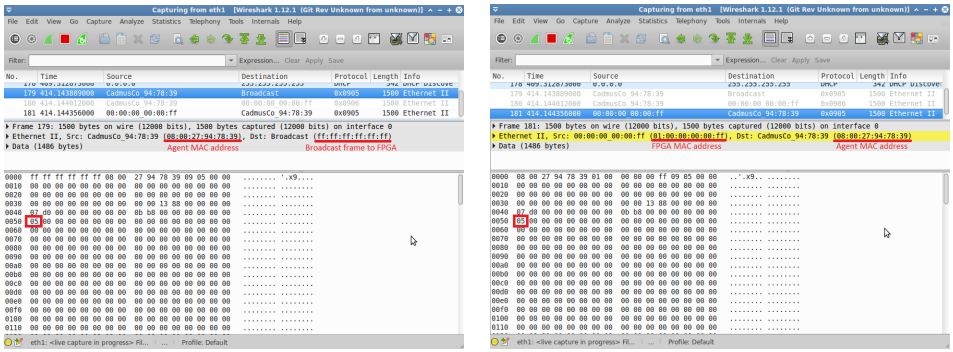


Figure 5.2: Wireshark snapshots of the exchanged Ethernet frames between the SDN agent and the FPGA. To the left, the frame sent by the agent to configure input port 1 to output port 2, during a single time slot. To the right, the frame sent back as a confirmation from the FPGA.

The used SDN controller is an Open Daylight controller. An FPGA is chosen instead of an ASIC, mainly because of its flexibility and the fact that it can be reprogrammed depending on the desired behaviour. The SDN agent and the FPGA communicate using raw Ethernet by exchanging Ethernet frames with a predefined structure. The control of the fast optical switches, as well as the on-chip integrated fiber switch will be further described in details.

Software controlled fast optical switching

SDN-controlled operation of the TDM switches is performed during the experimental demonstration described in Chapter 4. The FPGA that configures the switches communicates with the SDN agent. The structure of the Ethernet frame that is used between the SDN agent and the FPGA is illustrated in Fig. 5.1. Note that, for brevity, only the part with the used fields from the frame are shown, although the used frame length is fixed and set to 1500 Bytes.

It can be seen that besides the source and destination media access control (MAC) address, the only other remaining fields used are the sequence number (SN) and the slot configuration bytes. The SN is used to track the number of frames sent from the agent. When the FPGA receives a frame, it has to send a confirmation to the agent, thus it sets the correct source and destination addresses and replies by sending the same frame content

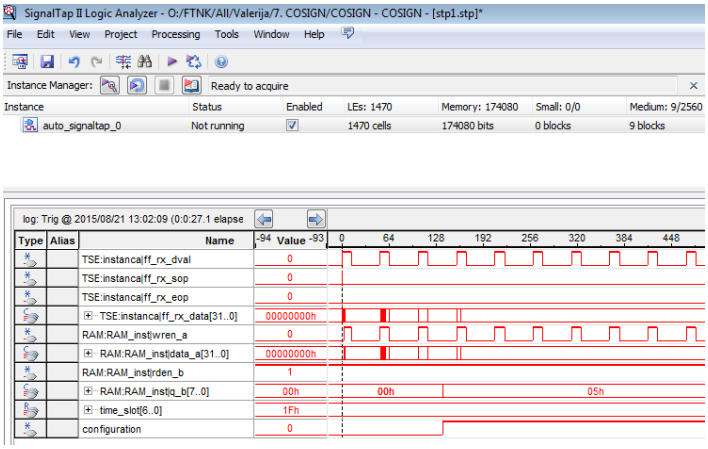


Figure 5.3: SignalTap snapshot of the the process of dynamic switch reconfiguration through the SDN controller. Once the Ethernet frame from the agent is received, the switch is reconfigured.

back to the agent. This allows that the agent can confirm what will be the exact switch configuration, which will be applied, and react accordingly, if errors have occurred.

The TDM switches are configured based on three distinct features, i.e. the input port, the time slot and the output port. The used agent allows for configuration of a 4x4 switch, however only a subset of the input and output ports are used. The switch input ports at the SDN controller are denoted as 0-3, while the output ports are denoted as 4-7. A time frame consisting of 96 time slots is defined and for each input port, the value of the output port is carried in a one byte field allocated for a single specific time slot. Hence, 96 Bytes indicate the configuration of input port 1, during 96 consecutive slots.

In order to verify that the switches are properly configured, Wireshark [134] is used to capture the frames exchanged between the agent and the FPGA, and SignalTap [135] is used to monitor the signals generated from the FPGA. The exchanged Wireshark frames are shown in Fig. 5.2 and the SignalTap snapshot is shown in Fig. 5.3. For simplicity, only reconfiguration of one switch is illustrated, although the same applies for the others.

As the default configuration of the switch is 'bar', the SDN controller is used to establish a connection from input port 1 (port label 0) to output port 2 (port label 5). Thus, as it can be seen in both frames, a value of '05' is carried in the first byte indicating the configuration for input port 1

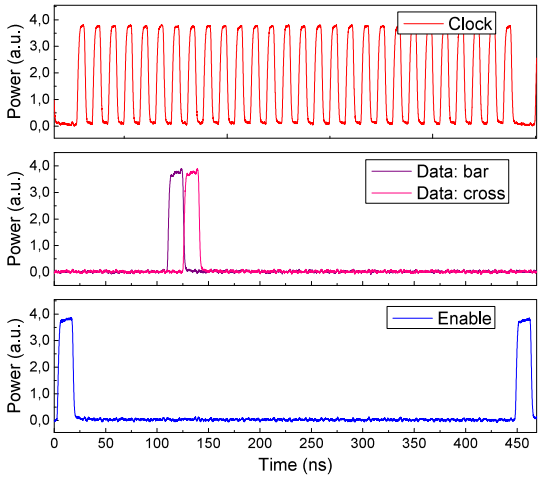


Figure 5.4: Time domain traces of the control signals for configuring the switch in 'bar' and 'cross' configuration. The signals are generated from the FPGA when instructed by the software controller.

during the first time slot. The same can be observed in Fig. 5.3, where as soon as the Ethernet frame by the agent is received, the value '05' is read from the random access memory (RAM) and the switch configuration signal, labelled 'configuration' is immediately set to high level. This confirms that successful SDN-controlled operation of the TDM switches can indeed be achieved, making the concept even more attractive for real commercial applications.

In addition to this work, software controlled operation of a fast optical switch has also been demonstrated using the switch presented in [71, 124, 125]. As the switch drivers are integrated on-chip, the control plane should provide only three driving signals, i.e. a 'clock', an 'enable' and a 'data' signal. This allows that the on-chip registers are set to the correct value and once the 'enable' signal is set to high level, the configuration stored in the registers is applied. A 2x2 switch was used, thus two possible configurations, a 'bar' and a 'cross' configuration are possible.

A Python script is used to control the switch and an FPGA is responsible for generating the three driving signals for the chip. The communication between the controller and the FPGA is established using the universal serial bus (USB) port. The oscilloscope traces of the driving signals at the output of the FPGA are used as an indicator of a successful software controlled operation. Fig. 5.4 illustrates the time domain traces of the FPGA

Ethernet Frame for Fiber Switch Configuration																																Eth. Frame lines (Captured by Wireshark)
ADDR. / BYTE	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	
																			SN	Source MAC Address						Destination MAC Address							0
00	RESERVED																																10,20
01	RESERVED																																30,40
02	RESERVED																																50,60
03	RESERVED																																70,80
04	RESERVED																																90,a0
05	RESERVED																																b0,c0
06	RESERVED																																d0,e0
07	RESERVED																																f0,100
08																										6	5	4	3	2	1	0	110.120

Figure 5.5: Structure of the Ethernet frame that is used between the SDN agent and the FPGA for configuring the on-chip fiber switch.

outputs, for static switch control and when both switching configurations are applied. It can be seen that, depending on the exact registers that have to be set in each case, the 'data' signal is set high either in the seventh or eighth clock cycle, indicating a 'bar' or 'cross' configuration, respectively.

Software controlled fiber switching

SDN-controlled operation has also been demonstrated using the on-chip fiber switch from [J3], [C6]. The 7x7 switch is envisioned to operate as a fiber switch, thus the required configuration data consist of a pair of input and output port that should be interconnected. Similarly, on the upstream side, the FPGA communicates with the SDN agent. However, on the downstream side, the FPGA is connected to a micro-controller that sets the correct voltage level for the individual heaters.

The structure of the Ethernet frame that is used between the SDN agent and the FPGA is illustrated in Fig. 5.5. Again, an Ethernet frame of 1500 Bytes is used, however only the part with the used fields from the frame are displayed. It can be seen that seven bytes are used (labelled from 0-6), one for each input port. Each byte carries the value of the output port to which the corresponding input should be connected. The values of the output ports on the SDN controller side are labelled from 7-13, or as displayed using hexadecimal notation, 07-0D.

In order to verify that the switch is properly configured, Wireshark is used again. The exchanged Wireshark frames are shown in Fig. 5.6. It is envisioned that the switch is by default in a 'bar' configuration, thus the SDN controller sets it in a 'cross' configuration, such that input port 1 (label 0) is connected to output port 7 (label 13), input port 2 (label 1)

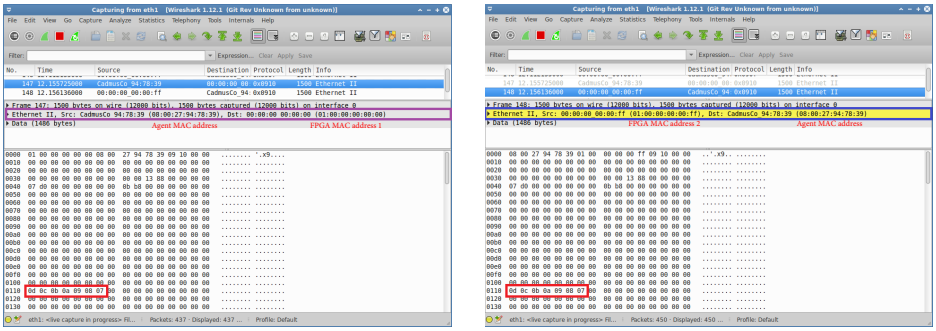


Figure 5.6: Wireshark snapshots of the exchanged Ethernet frames between the SDN agent and the FPGA. To the left, the frame sent by the agent to configure all input ports to a complementary output port. To the right, the frame sent back as a confirmation from the FPGA.

is connected to output port 6 (label 12), etc. It can be seen that the values of the corresponding output port are set correctly in the bytes for each individual input port of the frame sent by the agent. In order to distinguish between the frames for configuring a TDM and a fiber switch, the agent sends the frames for the fiber switch with a predefined MAC address.

A modified serial peripheral interface (SPI) protocol [136] is used for communication between the FPGA and the micro-controller. Thus, in order to observe the process of switch reconfiguration, the different signals are tapped from the micro-controller and monitored on an oscilloscope. The time domain traces of the tapped signals are shown on Fig. 5.7. Channel 1 (yellow label) is the clock signal sent from the FPGA and channel 2 (green label) is the transmitted data, also displayed in hexadecimal notation using the trace with blue label at the bottom. Channel 3 or the second trace from the top (blue label) shows the slave select (SS) signal, i.e. the cycle of master-slave communication and data transfer. Channel 4 or the first trace from the top (pink label) illustrates the output of one, randomly chosen digital to analog converter (DAC). Once the values for the DACs are received from the FPGA, the output voltage is set to a predefined value, in order to result in a specific configuration of that individual heater and the corresponding MZI.

It can be seen from Fig. 5.7, that one cycle of the SS signal, which basically defines the time to transfer the configuration data for all the used heaters, labelled as Δx is 84 μs. Note that this is dependent on the specific

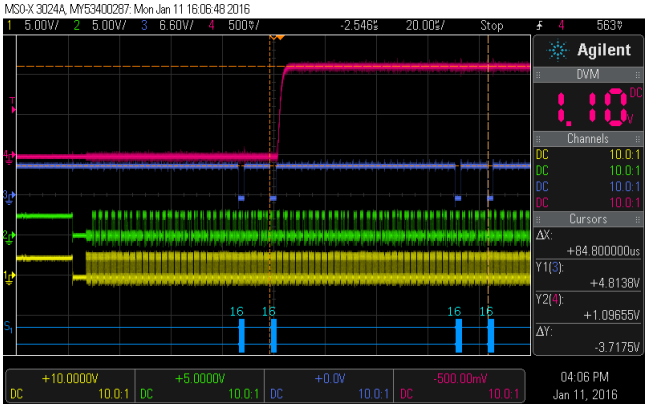


Figure 5.7: Time domain traces of the tapped signals from the micro-controller.

hardware available, as well as the existing limitations on the frequency of the used clock that is distributed from the FPGA to the micro-controller. It is expected that better performance can be achieved by using a more advanced micro-controller or by optimizing the communication protocol between the FPGA and the micro-controller. Considering that the average heater response time is around 30 μs, the rise and fall time of the switch is limited to a minimum of 30 μs, if all heaters are configured simultaneously. In any case, these results indicate that software control of the fabricated on-chip fiber switch is achievable, making the device attractive for real network deployment.

5.3 Enabling advanced network functionalities

As discussed in Chapter 2, in order to achieve sustainable data center growth, it is important to optimize the operational efficiency by either reducing the power consumption or obtaining a good bandwidth utilization. Exisiting work on optical networks has shown that switching in the optical domain can lead to several advantages, such as improved energy efficiency [137–139]. Power consumption savings ranging from 25 % - 40 % [140] have been demonstrated by preferring bypass at the optical level in IP over WDM networks.

Besides optical bypass or performing optical unicast (one input to one output) switching, different proposals exist on the use of optics in data centers in order to enable additional functionalities, currently performed at

higher levels such as IP [141–143]. Examples include, but are not limited to multicast communication (one node sends data to many nodes) or incast communication (one node receives data from many nodes). Note that the terms incast and traffic grooming can be used interchangeably. The ability to accomplish these operations in the optical domain by performing 1:N and N:1 optical switching yields several advantages.

For example, in the case of multicast communication, the same data has to be sent to N nodes. If optical multicast is not possible, N different connections have to be established and network resources have to be allocated along the full path for each individual connection. If optical multicast is possible, it is enough to send a single copy of the data and then propagate it along the network. At each of the N nodes, instead of performing unicast switching, the data will be split to two output ports (or more, if needed), allowing that each node receives a single copy, while another copy continues to propagate along the network until all nodes are reached. By replacing several unicast connections with one multicast connection, a reduced amount of network resources are used, resulting in lower network load and higher achievable network throughput. Furthermore, as the sender sends only one copy, energy efficient and low latency communication i.e. faster task execution time can be achieved, leading to improvement of the transmission time [144, 145].

There are several applications where the use of optical multicast and incast can be found useful, including services such as in-cluster updating [146], virtual machine (VM) migration [144, 147], data replication in distributed file systems such as GFS or HDFS [14, 15, 148] or distributed computational frameworks in general [13], as well as iterative machine learning [149].

5.3.1 SDM switching for unicast, multicast and traffic grooming in the Hi-Ring network

Similar as in other data center networks, optical multicast and incast can also be useful for the Hi-Ring network architecture. Besides the general motivation, there are special cases that arise as a result of the network topology and node structure, for which optical multicast and incast can bring additional advantages. These cases, as well as the process of implementation of these functionalities in the Hi-Ring architecture will be discussed below.

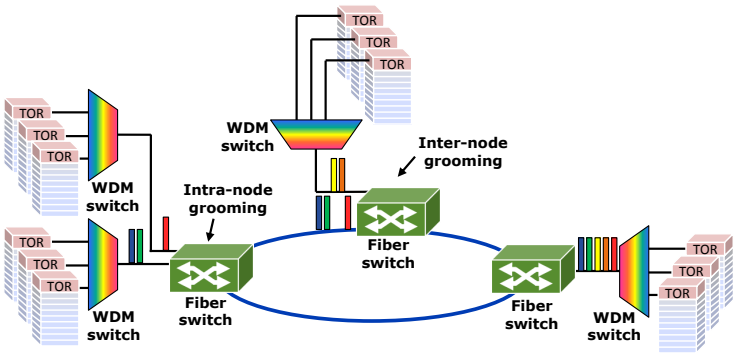


Figure 5.8: Inter-node and intra-node traffic grooming scenarios in the Hi-Ring network. (ToR - top of rack switch, WSS - wavelength selective switch).

Motivation for optical multicast and grooming in the Hi-Ring network

There are several reasons why optical multicast and grooming are important for the Hi-Ring network. First, it is important to consider the specific topology and the node structure. As already mentioned in Chapter 3, connections to and from the ToRs are either full wavelengths for bandwidth-hungry applications or time slots for bursty traffic. Thus, support for optical multicast and grooming should be provided for both types of connections. For simplicity, attention will be focussed only on optical multicast and grooming of full wavelength connections, although the same concept is applicable to time slotted connections.

The importance of optical traffic grooming is illustrated in Fig. 5.8. For brevity, the MCF switches have been omitted and only the fiber switches and wavelength switches at each node are displayed. Two different applications of optical grooming can be identified, inter-node and intra-node grooming, referring to grooming of traffic originating at different nodes or within the same node, respectively.

From the node modelling discussion, presented in Chapter 3, it is clear that a specific amount of switches are deployed at each level of the nodes. If it is assumed that a node has a single WSS for dropped traffic, then it can only receive simultaneously data carried in a single fiber core. Simultaneous reception of data carried in different fiber cores (i.e. coming from different WSSs within a single node or from WSSs located in different nodes), is not possible with a single WSS. The same reasoning applies for a node

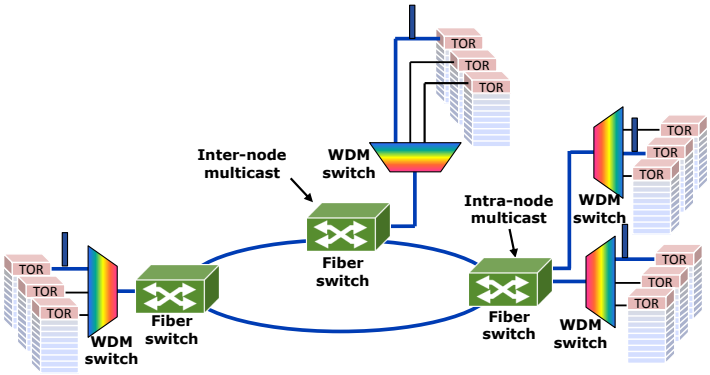


Figure 5.9: Inter-node and intra-node multicast scenarios in the Hi-Ring network. (ToR - top of rack switch, WSS - wavelength selective switch).

that deploys N switches, i.e. only N cores can be dropped at the same time. Thus, it is important that traffic is packed in a way that as many connections as possible can be established simultaneously and can coexist in the network.

As shown in Fig. 5.8, all connections that use wavelengths carried in different cores have to be established sequentially, even if they are destined to different destination ToR ports. Hence, even when the destination is free, an additional delay has to be experienced. In order to prevent this, traffic can be combined by using the fiber switch. This can be done both by combining traffic originating from different WSSs at one node, but destined to the same destination node (resulting in intra-node grooming), as shown at the first node on the left, or by grooming traffic between the two nodes (resulting in inter-node grooming), as shown at the node in the middle. At the node to the right, a single core can be dropped and all connections can be established simultaneously.

Alternatively, traffic grooming can also be done using the WSSs in the intermediate nodes, however this requires that some of the WSS ports are pure bypass ports, which may not be effective for a high hop count path as it will cause some extra delay. The fact that some of the ports are used for bypass would also mean that higher number of switches is required to serve the same server capacity, resulting in increased cost. Thus, being able to combine traffic using the fiber switch is a feasible alternative to reduce blocking and achieve higher network throughput in the Hi-Ring network.

The motivation for using multicast is similar as previously discussed for data center networks in general. As shown in Fig. 5.9, by using a single

wavelength to establish several connections at the same time, significant resource savings can be achieved. The obtained bandwidth efficiency is directly proportional to the multicast group size. Moreover, all connections are established simultaneously, allowing for the overall completion time to be shorter compared to the completion time of all the individual connections together, resulting in low latency connectivity. Similar as for optical traffic grooming, multicast can be performed inter-node, as well as intra-node among the different WSSs within a single node.

Implementation of optical multicast and grooming in the Hi-Ring network

In order to implement optical multicast and grooming in the Hi-Ring network, it is necessary that the used fiber switch can support power splliting and combining. For that purpose, the PIC composed of a fiber switch and fan-in and fan-out coupling devices [J3], [C6] has been chosen to be used in this occasion. Not only the switch perfectly suits this need, but also the chip itself is a great example of partially integrated multidimensional node components. This is an attractive feature allowing for the realization of compact devices that can perform different functions.

As shown in Fig. 5.10, the PIC combines two different components of the multidimensional switching nodes in the Hi-Ring architecture, the spatial (de)multiplexers and the fiber switch. In addition, the displayed switch matrix of the 7x7 switch shows how the different inputs and outputs can be connected. The different cores of the MCF are switched by controlling a heater in each of the 57 MZI structures. The use of MZI indicates that, besides unicast switching, power splitting and combining is also possible, with appropriate heater control.

Fig. 5.11 illustrates how the switch can be configured in a 1x7 multicast configuration. It is important to note that only the switching elements in a red box are the ones that need special control in order to operate in a power splitting mode, instead of a true switching mode. This means that establishing multicast/grooming connections can be done in a similar way to establishing unicast connections through the switch. For example, a 1x2 multicast from input port 1 to output port 1 and 2, requires that all MZIs are configured using the configuration settings for unicast connections from input 1 to outputs 1 and 2, and only a single MZI is configured to operate in power splitting mode. Similarly, a 1xN multicast requires modifying only the setting of N-1 MZIs. This means that the control of the switch is greatly simplified, as only few MZIs have to be configured with new settings, while

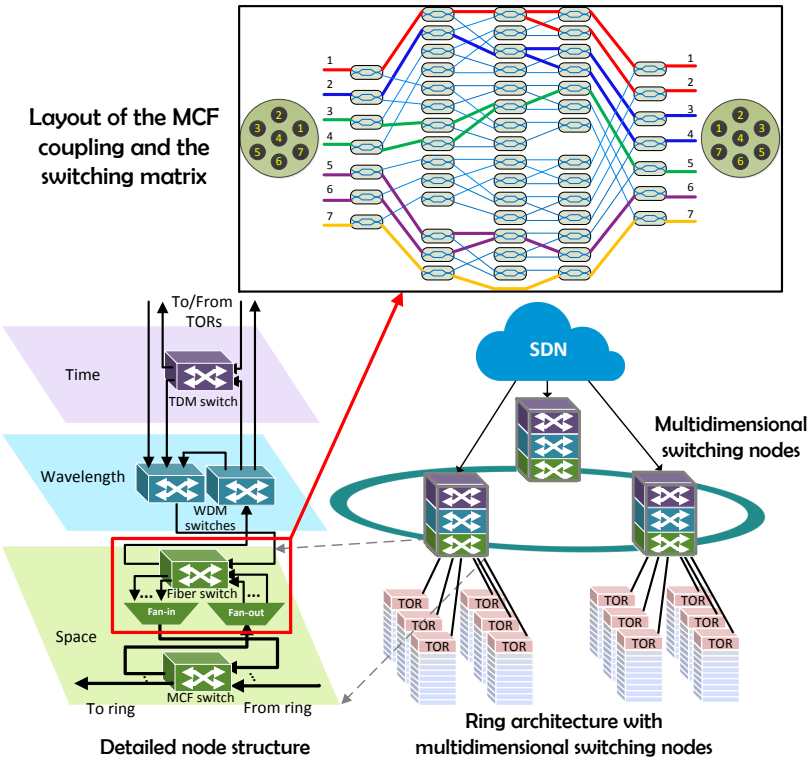


Figure 5.10: The proposed Hi-Ring architecture and the components included in the fabricated chip. The inset illustrates the chip layout and the switch matrix. (ToR - top of rack switch, WSS - wavelength selective switch, TDM - time division multiplexing, MCF - multicore fiber, SDN - software defined networking).

the remaining ones are configured in the same way as in unicast switching.

Moreover, multicast from any input port is possible similar to the illustrated example in Fig. 5.11. The multicast ratio can also be adjusted, i.e. the size of the multicast group can vary from two to seven outputs. Furthermore, any combination of output ports for a specific group size is possible. In order to perform grooming, the same considerations apply in reversed order. Additionally, the coupling ratio can be asymmetric for both multicast and grooming. This is very important as it allows for trade-offs, such as allocating more power to the inter-node terminated connections, resulting in similar inter/intra node performance.

Unlike other demonstrations of optical multicast [144,147] where optical

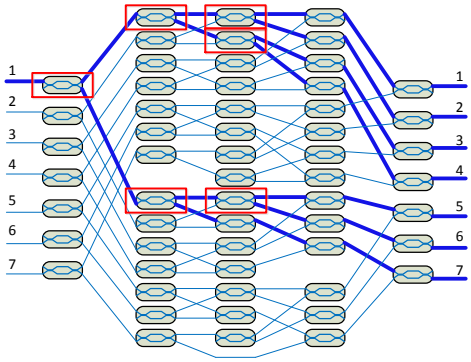


Figure 5.11: Configuration of the individual MZI elements in order to establish 1:7 multicast. Only the MZIs in a red box operate in power splitting mode.

splitters are required in addition to an optical circuit switch, the fabricated PIC can perform both multicast and grooming without the need of additional equipment. Moreover, switching in unicast and multicast/incast fashion can be done on-demand and provisioning of these additional network functionalities can be enabled by extending the existing support of SDN control.

5.3.2 Experimental demonstration

In order to confirm that the on-chip fiber switch can be used to perform optical multicast and traffic grooming, the system performance is experimentally investigated. Intra-node and inter-node scenarios are considered for both optical multicast and grooming. The receiver sensitivity is measured for multicast with different multicast group size and at different output port pairs. Moreover, multicast and grooming with different power ratios is investigated. At last, a combined case of simultaneous unicast, multicast and traffic grooming is evaluated, in order to confirm the system ability to deal with different types of switching scenarios.

Multicast and grooming characterization

Initially, a basic system characterization is performed to confirm that successful multicast and grooming can be performed using the PIC. The experimental setup is shown in Fig. 5.12 (a). For the different scenarios investigated for characterization purposes, similar considerations apply, with

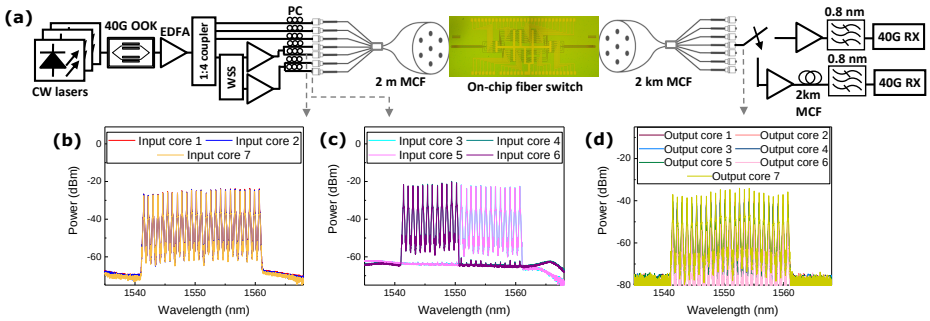


Figure 5.12: (a) Experimental setup. Spectra at the input of the MCF for (b) cores 1,2,7; (c) cores 3,4,5,6 and (d) spectra of all cores after switching. (CW-continuous wave, OOK-on off keying, WSS-wavelength selective switch, MCF-multicore fiber, RX-receiver, EDFA - Erbium-doped fiber amplifier, PC - polarization controller).

few small differences. When traffic grooming is investigated, the different wavelength channels are launched in the seven fiber cores. Grooming is performed using the fiber switch and the groomed traffic propagates in the 2 km MCF, as shown in Fig. 5.12 (a). When multicasting is performed, a single channel is launched in only one core propagating first in the 2 km MCF (reversed MCF coupling from Fig. 5.12 (a)). Thus, proper multicast can be performed using the switch (one copy sent and appropriately split at the desired node). Inter-node grooming and multicast is performed by adding extra 2 km MCF before the switch for grooming, and after the switch, when multicasting.

An important thing to note is the control of the switch. The configuration settings of all heaters for unicast switching have already been identified in a previous characterization [J3], [C6]. A micro-controller board and a feedback circuit are used to find the correct voltage setting for which different connections can be establish through the switch. Similarly, for the purpose of multicast and grooming, the micro-controller is used to find the desired voltage level for which the MZIs will perform power splitting or combination. A thorough search enables that the settings for all possible combinations of input and output ports, and ratios can be identified. These values can be stored, allowing for fast dynamic reconfiguration in future.

Initially, a single channel (1550.12 nm) is used, carrying 40 Gbit/s OOK modulated data. The channel is launched in one core of the 2-km MCF, that is coupled to the chip where multicast is performed. First, we investigate

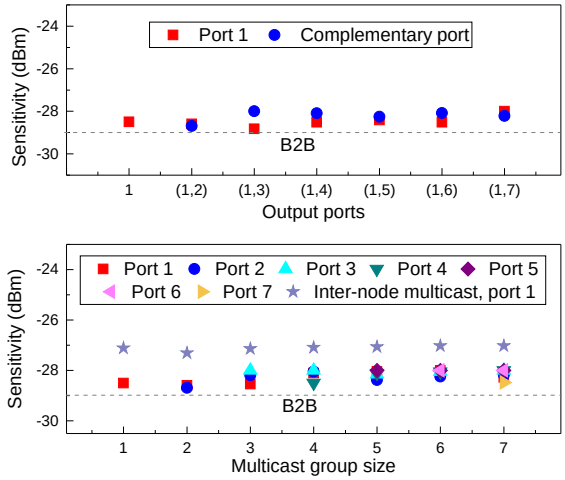


Figure 5.13: Receiver sensitivity (BER=10^{-9}) of single channel (1550.12 nm) for performing 1:2 multicast on different output ports (top) and for different multicast group size (bottom).

the ability to perform 1:2 multicast for six different combinations of output port pairs. Then, we verify that similar performance can be obtained for different size of the multicast group.

The measured receiver sensitivity (BER=10^{-9}), for both cases is shown in Fig. 5.13. It can be seen that for 1:2 multicast over different output ports, similar behaviour is observed. The minor variations in the performance are due to imperfections in the process of fine tuning the heaters for the different paths. It is expected that this can be avoided by further optimization of the control process. In any case, the similar performance is an indicator, that the switch can be used to provide flexible multicast along the full range of output ports. When the multicast group size is varied from 2 to 7 output ports, it can be seen that this results in a negligible penalty for both intra-node as well as inter-node multicast, confirming the ability to provide an on-demand multicast ratio without additional equipment.

In addition, the effect of the coupling ratio on the system performance is investigated, for both multicast and grooming scenarios. The measured receiver sensitivities (BER=10^{-9}) for both cases are shown in Fig. 5.14. The receiver sensitivity of a single channel (1550.12 nm), for a 1:2 multicast, is measured on a single fixed port (port 1) and for power splitting ratios ranging from 10 % to 89 %. It can be seen that, for both intra-node and inter-node multicast, the experienced penalty is negligible when the power

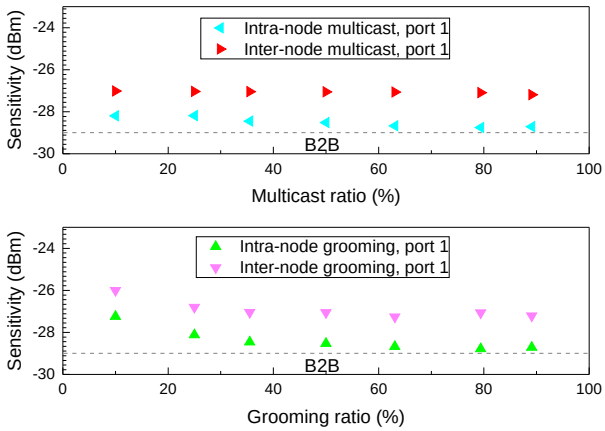


Figure 5.14: Receiver sensitivity (BER=10^{-9}) of a single channel (1550.12 nm) for performing 1:2 multicast (top) and 2:1 grooming (bottom) with different power ratios.

at the output port is reduced.

Similar behaviour is also observed for 2:1 grooming of two channels (1550.12 nm and 1550.92 nm) at two input ports. The performance of a single channel (1550.12 nm) is measured at the output of the switch (output port 1). It can be seen that the penalty of modifying the combining ratio is relatively small and satisfying performance can be achieved for a wide range of power ratios. This could be exploited in a real network scenario, where by identifying the optimal condition in different cases and using an asymmetric ratio, the overall performance can be improved.

Combined unicast, multicast and grooming scenario

In order to verify that all functionalities such as unicast, multicast and grooming can coexist and be simultaneously provided using the same switch, the following combined scenario is considered. On two input cores (input core 1 and input core 2), carrying 25 wavelength channels each, 1:2 multicast is performed. 2:1 grooming is performed on four input cores (input cores 3 and 4; input cores 5 and 6), where each groomed core carries 13 or 12 spectrally non-overlapping wavelength channels. At last, one input core (input core 7), carrying 25 wavelength channels is unicast switched.

The established paths through the switching matrix of the fiber switch are illustrated in Fig. 5.10. The experimental setup for the combined network scenario is shown in Fig. 5.12 (a). All wavelength channels carry

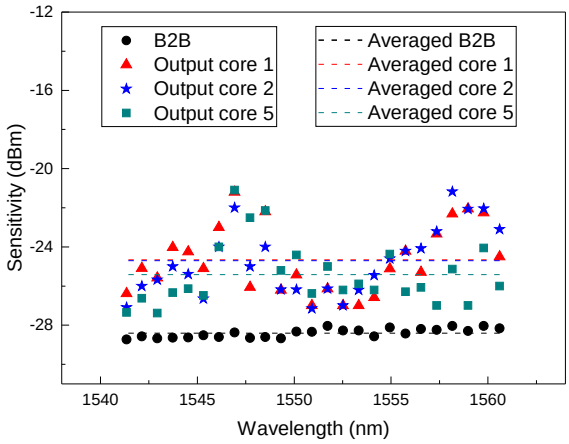


Figure 5.15: Receiver sensitivity (BER=10^{-9}) of all channels for three output cores (output cores 1 and 2, undergoing multicast and output core 5, undergoing grooming).

40 Gbit/s OOK modulated data. The spectra of all cores at the switch input and output, observed through a 20-dB coupler are shown in Fig. 5.12 (b-d). Due to the specific switching scenario considered, each core after switching has 25 wavelength channels transmitted in the 2-km MCF.

The performance of the system is evaluated by measuring the receiver sensitivities (BER=10^{-9}) on all channels for two output cores that undergo intra-node (output core 1) and inter-node (output core 2) multicast and for one output core (output core 5) that carries groomed traffic. Fig. 5.15 shows the measured receiver sensitivity of all channels for the three output cores. As it can be seen, error-free performance is achieved for all channels. The variations of the measured receiver sensitivity for the different channels are mainly due to the imperfect power equalization of the used lasers, as well as the wavelength dependent crosstalk of the switch, as a result of the polarization variation.

In addition, Fig. 5.16 illustrates the measured receiver sensitivity (BER=10^{-9}) of a single wavelength channel in all cores for the same switching configuration. It can be seen that BER$<10^{-9}$ is achieved for all cores. The observed variations in the measured receiver sensitivity are due to the different insertion loss and crosstalk of the MCF coupling devices, as well as the different crosstalk that is experienced in the specific switching configuration.

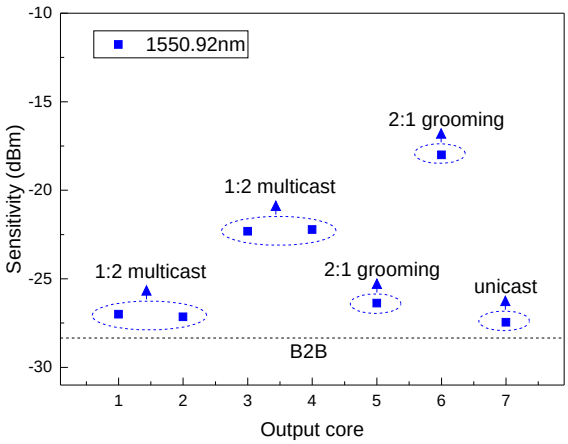


Figure 5.16: Receiver sensitivity (BER=10^{-9}) of a single channel (1550.92 nm) in all cores undergoing either unicast switching, multicast or grooming.

5.4 Summary

This chapter has addressed several important aspects of commercializing the technologies used with the multidimensional switching nodes, proposed as the main switching entity in the Hi-Ring architecture. First, the feasibility of on-chip integration is discussed and existing work is reviewed. Several examples of on-chip integrated switches and other components are mentioned, including some that have been used during experimental work described in this thesis.

Moreover, the process of integrating the hardware in the data plane with a software defined control plane has been presented, emphasising the main benefits of having a completely decoupled control plane. Several efforts on integrating different types of switches used in the data plane, including fast optical switches, as well as fiber switches, with software controllers have been presented. Successful integration and software controlled operation has been demonstrated.

At last, enabling advanced network functionalities has been discussed in details. The motivation behind using optical multicast and traffic grooming has been introduced, together with existing studies and implementation efforts. Moreover, the benefits of optical multicast and traffic grooming in the Hi-Ring network have been analysed in details. Using an on-chip fiber switch, the actual implementation of these functionalities has been

experimentally demonstrated.

Error-free performance (BER=10^{-9}) is achieved for 1:2 multicast at different output port pairs, as well as for multicast with different multicast group size (from 2 to 7) over a fixed range of output ports. In addition, it has been shown that it is feasible to perform both multicast and traffic grooming with different power ratios. Finally, error-free performance (BER=10^{-9}) has also been achieved in a combined network scenario deploying simultaneous switching of 5 Tbit/s in a unicast, multicast and incast fashion.

The presented analysis and experimental results confirm once again that both hardware component integration and hardware-software integration hold the promise to make the proposed Hi-Ring architecture and the concept of multidimensional switching, a commercially deployable solution. In addition, such future software controlled on-chip networks can easily support a range of additional network functionalities, like optical multicast and incast, allowing for optimized network performance.

Chapter 6

Conclusion

Over the last decade, data center traffic has grown tremendously, resulting in a lot of attention focussed on data center networks in general. In order to cope with this and achieve a sustainable growth, data center operators have been forced to look at ways to improve their current infrastructure. The use of novel technologies and alternative solutions for scaling future data center networks have become increasingly attractive and optics has emerged as a viable candidate.

The need of introducing optical switching in the data center has become prevalent from several reasons, the main reason being that optical switching holds the promise of energy-efficient and bit rate independent communication. This allows that data center networks can seamlessly scale and maintain the pace at which data center traffic grows. Different research proposals on hybrid and all-optical data center architectures have exploited optical switching for various applications in the data center and demonstrated that optics can be applied to solve many of the current challenges that data centers face today.

In this Ph.D. thesis, the concept of optical multidimensional switching for data center networks has been presented and a novel data center architecture, namely the Hi-Ring architecture, composed of multidimensional switching nodes has been proposed. The overall presented work aims to demonstrate that optical switching technologies are a promising solution for deployment in future data center networks. The presented concepts and results in the aforementioned chapters are summarized in the following section. Additionally, an overview is given on future work and the aspects worth further investigation are discussed in details.

6.1 Summary

In Chapter 2, an overview has been given of today's data centers. The current organization and challenges of the deployed data center architectures were discussed, outlining the motivation for new technologies and solutions. It has been shown that data centers today struggle to cope with the traffic growth from several aspects, including scaling and achieving energy-efficient operation. This has motivated the work behind several proposals on novel data center architectures based on different technologies. Data center architectures deploying electrical switching, optical switching and hybrid proposals have been reviewed and analysed in terms of their advantages and shortcomings. As the motivation to use optical switching in the data center starts to build up and becomes more apparent, state-of-the-art optical switching technologies have been revised, with the aim to determine their main features and based on that envision their possible role in future data center architectures.

In Chapter 3, the Hi-Ring data center architecture has been presented. The concept of optical multidimensional switching and the different switching technologies have been discussed in details. Modelling of a network composed of these nodes has shown that several feasible node implementations exist for data centers with different number of servers. Based on the modelled structure, the cost and power consumption of the Hi-Ring network has been compared to a fat-tree network. It has been shown that although the capital investment in the Hi-Ring network is higher than the fat-tree, significant OPEX savings can be achieved by using the Hi-Ring, that can compensate for the CAPEX difference. The feasibility of the Hi-Ring network has also been investigated in an experimental scenario. Switching of connections with different granularity has been performed successfully and moreover, error-free transmission of 7 Tbit/s between nodes has been demonstrated.

Chapter 4 has given an overview of existing optical subwavelength switching schemes. Optical subwavelength switching is crucial in shifting towards all-optical data center architectures and replacing electrical packet switching. The lack of optical buffering has shown detrimental in adopting optical packet switching paradigms and thus, the need for a new optical subwavelength technology is pressing. Optical TDM switching is presented as a viable alternative to both replace electrical packet switching and use the best of optics. However, deployment of any optical subwavelength technology is dependant on synchronization. The importance of synchronization

is elaborated in details and different proposals are revised. The problem of synchronization in the Hi-Ring network is revised and a new synchronization algorithm is proposed. The algorithm is experimentally verified and moreover, successful data plane operation is achieved when it is used to synchronize the SDN-controlled network nodes.

In Chapter 5 the integration and provisioning of advanced network functionalities have been discussed. First, on-chip hardware integration has been addressed and efforts towards node integration have been presented. Commercialization of any of the aforementioned solutions, strongly depends on hardware integration in order to reach a deployment phase. Low footprint, reduced cost and standardized fabrication and packaging are paramount. Another crucial characteristic is software control. Integration with an SDN platform and SDN compatibility are a must for any data center architecture. Several efforts on software controlled switching have been presented, confirming that the full multidimensional node could eventually be SDN controlled. Furthermore, the benefits of optical multicast and incast have been analysed and their importance to the overall performance in the Hi-Ring network has been highlighted. Provisioning of these functionalities in the Hi-Ring network using an on-chip integrated fiber switch has been experimentally demonstrated.

6.2 Future work

This thesis has addressed some of the challenges of current data centers and proposed concepts based on optical technologies, that aim to solve many of them. In several turns, the feasibility of these solutions has been confirmed either through analytical studies or experimental work. Successful switching of connections with different granularity, synchronized optical TDM switching, SDN-controlled operation and enabling advanced network functionalities such as multicast and incast are some among the main contributions of this thesis.

Although the proposals presented in this thesis have been verified and shown feasible, reaching the stage of commercial deployment still requires that some improvements are made, both from technological and network point of view. It has been shown that the CAPEX of the Hi-Ring network is higher than the fat-tree network. This directly indicates that the cost of the components of the multidimensional switching node has to be reduced. Integration is one way to achieve this, allowing that both the fibers within the node are eliminated and that joint powering and cooling is achieved,

that can reduce the overall cost. Silicon photonics has already been mentioned as a platform that holds the promise to facilitate the development of future SoC and NoC. Proposals of on-chip integrated devices including switches [58, 63, 71, 72, 121–126] and additional node components [119, 120] are some of the few, whose further development can bring optical switching technologies closer to the market.

From the experimental demonstrations on the Hi-Ring prototype and the TDM switching, it is clear that technological development is needed to improve the performance of the used optical switches. For a large data center network, it is extremely important to preserve the signal integrity. The insertion loss of all optical switches has to be optimized, so that the use of optical amplification along the network is minimized. For the fast optical switches, it is crucial that the crosstalk performance is satisfying, as significant degradation can occur otherwise. In addition, careful network planning is required in order to consider the physical impairments and incorporate them in the routing and resource allocation process. Several studies [75, 150] have already extended the work on the basic routing and wavelength assignment (RWA) problem to include either time slots or multicore fibers, and work of this kind is extremely important in order to be able to develop enhanced routing algorithms.

For the deployment of optical subwavelength switching, development of additional technologies, such as optical burst mode receivers is needed. It is desirable that the BMR have very fast locking time, allowing that a short preamble can be sent. The length of the preamble directly influences the bandwidth utilization of a channel that is time shared, similar like the switching gap of the fast optical switches. Reducing both of these, allows that most of the bandwidth is used for actual data transmission, making the sharing concept feasible. Additionally, BMRs operating at data rates higher than 10 Gbit/s, such as 25Gbit/s and higher are required in order to be able to support the current serial data rates in data centers. Existing research work on BMRs [151] has demonstrated locking time of several tens of nanoseconds. Future work on both system integration, as well as enhanced receiver performance is extremely important for adopting optical subwavelength switching.

Software controlled operation is another prerequisite for actual network deployment. It is necessary that all network elements are integrated with a chosen SDN controller and dynamical reconfiguration is enabled. Software controlled operation is crucial to maximize the resource utilization. A centralized controller with a holistic network overview can optimize the

overall network performance as shown in [129], by reducing the blocking and achieving higher throughput.

Acronyms

ASIC application specific integrated circuit

AWGR arrayed waveguide grating router

BER bit error rate

BMR burst mode receiver

BPF band-pass filter

B2B back to back

CAGR compound annual growth rate

CAPEX capital expenditure

CMOS complementary metal-oxide-semiconductor

COSIGN combining optics and software defined networking in next generation data center networks

CUE carbon usage effectiveness

CW continuous wave

DAC digital to analog converter

DC data center

DCF dispersion compensation fiber

DCN data center network

DOS data center optical switch

DWDM dense wavelength division multiplexing

EDFA Erbium-doped fiber amplifier

EPS electrical packet switching

FDL fiber delay line

FIT failure in time

FPGA field programmable gate array

GCI global cloud index

GFS Google File System

GPS Global Positioning System

HDFS Hadoop Distributed File System

HORP hybrid optoelectronic packet router

HPC high performance computing

IC integrated circuit

ICT information and communication technologies

IDC International Data Corporation

IP Internet Protocol

LCoS liquid crystal on silicon

$LiNbO_3$ lithium niobate

LR long reach

MAC media access control

MCF multicore fiber

MEMS micro-electro-mechanical system

MMF multimode fiber

MRR micro ring resonator

MZI Mach-Zehnder interferometer

MZM Mach-Zehnder modulator

NFV network function virtualization

NN network node

NoC network on chip

OBS optical burst switching

OCS optical circuit switching

ODL Open Daylight

OEO optical-electrical-optical

OF Open Flow

OLG optical label generator

ONOS open network operating system

OOK on off keying

OPEX operational expenditure

OPS optical packet switching

OSNR optical signal-to-noise ratio

PC personal computer

PC polarization controller

PCE path computation element

PD photo-diode

PIC photonic integrated circuit

PLC planar lightwave circuit

PLZT lanthanum-modified lead zirconate titanate

PUE power usage effectiveness

RAM random access memory

RWA routing and wavelength assignment

RU rack unit

RX receiver

SDM space division multiplexing

SDN software defined networking

Si silicon

SiN silicon nitride

SMF single mode fiber

SN sequence number

SOA semiconductor optical amplifier

SoC system on chip

SOI silicon on insulator

SPI serial peripheral interface

SR short reach

SS slave select

TDM time division multiplexing

TE traffic engineering

TOPS throughput optimized photonically optimized embedded microprocessors system

ToR top of rack

TWC tunable wavelength converter

TX transmitter

USB universal serial bus

USD United States dollar

VM virtual machine

VNI visual networking index

WAN wide area network

WDM wavelength division multiplexing

WSS wavelength selective switch

Bibliography

[1] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2015-2020," *Cisco Technical Report*, June 2016.

[2] F. T. Chong, M. J. R. Heck, P. Ranganathan, A. A. M. Saleh, and H. M. G. Wassel, "Data Center Energy Efficiency: Improving Energy Efficiency in Data Centers Beyond Technology Scaling," *IEEE Design & Test*, vol. 31, no. 1, pp. 93-104, 2014.

[3] C. L. Belady, "Projecting Annual New Datacenter Construction Market Size," *Microsoft Global Foundation Services Technical Report*, March 2011.

[4] Cisco, "Cisco Global Cloud Index: Forecast and Methodology, 2015-2020," *Cisco Technical Report*, Nov. 2016.

[5] Gartner, (www.gartner.com).

[6] IDC, (www.idc.com).

[7] IDC, "Why Upgrade Your Server Infrastructure Now?," *IDC White Paper*, July 2016.

[8] COSIGN, (www.fp7-cosign.eu).

[9] DR News, (www.dr.dk/nyheder/penge/facebook-vil-bygge-kaempe-datacenter-i-odense).

[10] A. Ghiasi, and R. Bacca, "Overview of Largest Data Centers," *IEEE 802.3bs Task Force Interim Meeting*, May 2014.

[11] G. E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, pp. 114-117, 1965.

[12] N. Farrington, and A. Andreyev, "Facebook's Data Center Network Architecture," *Optical Interconnects (OI) Conference*, paper TuB5, 2013.

[13] J. Dean, and S. Ghemawat, "MapReduce: Simplifed Data Processing on Large Clusters," *Symposium on Operating Systems Design and Implementation (OSDI)*, pp. 137–150, 2014.

[14] S. Ghemawat, H. Gobioff, and S. Leung, "The Google File System," *Symposium on Operating Systems Principles (SOSP)*, pp. 29–43, 2003.

[15] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," *Symposium on Mass Storage Systems and Technologies (MSST)*, pp. 1-10, 2010.

[16] IEEE 802.3 Ethernet Standard, (www.standards.ieee.org).

[17] Storage Networking Industry Association (SNIA), (www.snia.org).

[18] M. Al-Fares, A. Loukissas, and A. Vahdat, "A Scalable, Commodity Data Centre Network Architecture," *ACM SIGCOMM*, pp. 63-74, 2008.

[19] C. Clos, "A Study of Non-blocking Switching Networks," *Bell System Technical Journal*, vol. 32, no. 2, pp. 406-424, 1953.

[20] J. Koomey, "Growth in Data Center Electricity Use 2005 to 2010," August 2011.

[21] The Independent, (www.independent.co.uk/environment/global-warming-data-centres-to-consume-three-times-as-much-energy-in-next-decade-experts-warn-a6830086.html).

[22] The Climate Group, GeSI, "SMART 2020: Enabling the Low Carbon Economy in the Information Age," June 2008.

[23] ASHRAE, "2011 Thermal Guidelines for Data Processing Environments – Expanded Data Center Classes and Usage Guidance," May 2011.

[24] The Green Grid, (www.thegreengrid.org).

[25] Global e-Sustainability Initiative, (www.gesi.org).

[26] Google, (www.google.com/about/datacenters/efficiency/internal/).

[27] Facebook, (https://sustainability.fb.com/our-footprint/).

[28] GreenDataProject, "Where does power go?," June 2008.

[29] Facebook, (https://code.facebook.com/posts/1433093613662262/-under-the-hood-facebook-s-cold-storage-system-/).

[30] D. Abts, M. R. Marty, P. M. Wells, P. Klausler, and H. Liu, "Energy Proportional Datacenter Networks," *International Symposium on Computer Architecture (ISCA)*, pp. 338–347, 2010.

[31] J. Kim, W. J. Dally, and D. Abts, "Flattened Butterfly: A Cost-Efficient Topology for High-Radix Networks," *International Symposium on Computer Architecture (ISCA)*, pp. 126–137, 2007.

[32] J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-Driven, Highly-Scalable Dragonfly Topology," *International Symposium on Computer Architecture (ISCA)*, pp. 77–88, 2008.

[33] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, "DCell: A Scalable and Fault-Tolerant Network Structure for Data Centers," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 4, pp. 75–86, 2008.

[34] 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," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4, pp. 63–74, 2009.

[35] 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," *ACM SIGCOMM*, pp. 183-197, 2015.

[36] Facebook, (https://code.facebook.com/posts/360346274145943/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/).

[37] A. Tamarakuzhi, and J. A. Chandy, "2-Dilated Flattened Butterfly: A Nonblocking Switching Network," *IEEE International Conference on High Performance Switching and Routing (IEEE HPSR)*, pp. 153–158, 2010.

[38] A. Bhatele, N. Jain, Y. Livnat, V. Pascucci, and P. T. Bremer, "Analyzing Network Health and Congestion in Dragonfly-Based Supercomputers," *International Parallel and Distributed Processing Symposium (IDPDS)*, pp. 93-102, 2016.

[39] N. Jiang, J. Kim, and W. J. Dally, "Indirect Adaptive Routing on Large Scale Interconnection Networks," *ACM SIGARCH Computer Architecture News*, vol. 37, no. 3, pp. 220–231, 2009.

[40] N. Jain, A. Bhatele, X. Ni, N. J. Wright, and L. V. Kale, "Maximizing Throughput on a Dragonfly Network," *International Conference on High Performance Computing, Networking, Storage and Analysis*, pp. 336–347, 2014.

[41] B. Prisacari, G. Rodriguez, P. Heidelberger, D. Chen, C. Minkenberg, and T. Hoefler, "Efficient Task Placement and Routing of Nearest Neighbor Exchanges in Dragonfly Networks," *International Symposium on High-Performance Parallel and Distributed Computing*, pp. 129–140, 2014.

[42] A. Bhatele, K. Mohror, S. H. Langer, and K. E. Isaacs, "There goes the neighborhood: performance degradation due to nearby jobs," *International Conference on High Performance Computing, Networking, Storage and Analysis*, paper 41, 2013.

[43] K. Bilal, S. U. Khan, L. Zhang, H. Li, K. Hayat, S. A. Madani, N. Min-Allah, L. Wang, D. Chen, M. Iqbal, C. Z. Xu, and A. Y. Zomaya, "Quantitative Comparisons of the State of the Art Data Center Architectures," *Concurrency and Computation: Practice and Experience*, vol. 25, no. 12, pp. 1771-1783, 2013.

[44] 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," *ACM SIGCOMM*, pp. 327–338, 2010.

[45] 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," *ACM SIGCOMM*, pp. 339–350, 2010.

[46] X. Ye, Y. Yin, S. J. B. Yoo, P. Mejia, R. Proietti, and V. Akella, "DOS: A Scalable Optical Switch for Datacenters," *ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)*, pp. 1–12, 2010.

[47] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: A Topology Malleable Data Center Network," *ACM SIGCOMM Workshop on Hot Topics in Networks*, pp. 1–6, 2010.

[48] K. Xi, Y. H. Kao, and H. J. Chao, "A Petabit Bufferless Optical Switch for Data Center Networks," *Optical Interconnects for Future Data Center Networks*, C. Kachris, K. Bergman, and I. Tomkos, Eds. Springer New York, 2013, pp. 135–154.

[49] Z. Cao, R. Proietti, M. Clements, and S. J. B. Yoo, "Experimental Demonstration of Dynamic Flexible Bandwidth Optical Data Center Network with All-to-All Interconnectivity," *European Conference on Optical Communications (ECOC)*, paper PD.1.1, 2014.

[50] G. M. Saridis, S. Peng, Y. Yan, A. Aguado, B. Guo, M. Arslan, C. Jackson, W. Miao, N. Calabretta, F. Agraz, S. Spadaro, G. Bernini, N. Ciulli, G. Zervas, R. Nejabati, and D. Simeonidou, "Lightness: A Function-Virtualizable Software Defined Data Center Network With All-Optical Circuit/Packet Switching," *Journal of Lightwave Technologies*, vol. 34, no. 7, pp. 1618-1627, 2016.

[51] L. Schares, B. G. Lee, F. Checconi, R. Budd, A. Rylyakov, N. Dupuis, F. Petrini, C. L. Schow, P. Fuentes, O. Mattes, and C. Minkenberg, "A Throughput-Optimized Optical Network for Data-Intensive Computing," *IEEE Micro*, vol. 34, no. 5, pp. 52-63, 2014.

[52] K. Kitayama, Y. Huang, Y. Yoshida, R. Takahashi, T. Segawa, S. Ibrahim, T. Nakahara, Y. Suzaki, M. Hayashitani, Y. Hasegawa, Y. Mizukoshi, and A. Hiramatsu, "Torus-Topology Data Center Network Based on Optical Packet/Agile Circuit Switching with Intelligent Flow Management," *Journal of Lightwave Technologies*, vol. 33, no. 5, pp. 1063-1071, 2015.

[53] C. Minkenberg, G. Rodriguez, B. Prisacari, L. Schares, P. Heidelberger, D. Chen, and C. Stunkel, "Performance Benefits of Optical

Circuit Switches for Large-Scale Dragonfly Networks," *Optical Fiber Communications Conference (OFC)*, paper W3J.3, 2016.

[54] 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," *Optical Fiber Communications Conference (OFC)*, paper W1J.3, 2016.

[55] Calient, (www.calient.com).

[56] Hubner Suhner Polatis, (www.polatis.com).

[57] Glimerglass, (www.glimerglass.com).

[58] T. J. Seok, N. Quack, S. Han, W. Zhang, R. S. Muller, and M. C. Wu, "64x64 Low-Loss and Broadband Digital Silicon Photonic MEMS Switches," *European Conference on Optical Communications (ECOC)*, paper Tu.1.2.1, 2016.

[59] Hubner Suhner Polatis, (www.polatis.com/series-7000-384x384-port-software-controlled-optical-circuit-switch-sdn-enabled.asp).

[60] Finisar, (www.finisar.com).

[61] M. Iwama, M. Takahashi, Y. Uchida, M. Kimura, R. Kawahara, S. Matsushita, and T. Mukaihara, "Low Loss 1x93 Wavelength Selective Switch Using PLC-based Spot Size Converter," *European Conference on Optical Communications (ECOC)*, paper Mo.4.2.2, 2015.

[62] EOSPACE, (www.eospace.com).

[63] L. Qiao, W. Tang, and T. Chu, "16x16 Non-blocking Silicon Electro-optic Switch Based on Mach-Zehnder Interferometers," *Optical Fiber Communications Conference (OFC)*, paper Th1C.2, 2016.

[64] EpiPhotonics, (www.epiphotonics.com).

[65] H. C. H. Mulvad, A. Parker, B. King, D. Smith, M. Kovacs, S. Jain, J. R. Hayes, M. Petrovich, D. J. Richardson, and N. Parsons, "Beam-Steering All-Optical Switch for Multi-Core Fibers," *Optical Fiber Communications Conference (OFC)*, paper Tu2C.4, 2017.

[66] Cisco, (www.cisco.com).

[67] Arista, (www.arista.com).

[68] FiberStone, (www.fs.com).

[69] Fiber24, (www.fiber24.com).

[70] Ovoenergy, (www.ovoenergy.com/guides/energy-guides/average-electricity-prices-kwh.html).

[71] B. G. Lee, A. V. Rylyakov, W. M. J. Green, S. Assefa, C. W. Baks, R. Rimolo-Donadio, D. M. Kuchta, M. H. Khater, T. Barwicz, C. Reinholm, E. Kiewra, S. M. Shank, C. L. Schow, and Y. A. Vlasov, "Monolithic silicon integration of scaled photonic switch fabrics, CMOS logic, and device driver circuits," *Journal of Lightwave Technologies*, vol. 32, no. 4, pp. 743–751, 2014.

[72] M. Hai, P. Liao, M. Mir Shafiei, and O. Liboiron-Ladouceur, "MZI-based Non-blocking SOI Switches," *Asia Communications and Photonics Conference (ACP)*, paper ATh3A.147, 2014.

[73] M. A. Mestre, G. De Valicourt, P. Jenneve, H. Mardoyan, S. Bigo, and Y. Pointurier, "Optical Slot Switching-Based Datacenters With Elastic Burst-Mode Coherent Transponders," *European Conference on Optical Communications (ECOC)*, paper Th.2.2.3, 2014.

[74] I. Widjaja, and I. Saniee, "Simplified Layering and Flexible Bandwidth with TWIN," *ACM SIGCOMM*, pp. 13-20, 2004.

[75] G. Shan, G. Zhu, and D. Liu, "Study on the Problem of Routing, Wavelength and Time-Slot Assignment Toward Optical Time-Slot Switching Technology," *Photonic Network Communications*, vol. 22, no. 2, pp. 162-171, 2011.

[76] D. Blumenthal, "Photonic Packet Switching and Optical Label Swapping," *Optical Networks Magazine*, vol. 2, no. 6, pp. 54-65, 2001.

[77] M. J. O'Mahony, D. Simeonidou, D. K. Hunter, and A. Tzanakaki, "The Application of Optical Packet Switching in Future Communication Networks," *IEEE Communications Magazine*, vol. 39, no. 3, pp. 128-135, 2001.

[78] T. S. El-Bawab, and J. Shin, "Optical Packet Switching in Core Networks: Between Vision and Reality," *IEEE Communications Magazine*, vol. 40, no. 9, pp. 60-65, 2002.

[79] S. Yao, S. J. B. Yoo, B. Mukherjee, and S. Dixit, "All-Optical Packet Switching for Metropolitan Area Networks: Opportunities and Challenges," *IEEE Communications Magazine*, vol. 39, no. 3, pp. 142-148, 2001.

[80] D. Chiaroni, G. B. Santamaria, C. Simonneau, S. Etienne, J. Antona, S. Bigo, and J. Simsarian, "Packet OADMs for the Next Generation of Ring Networks," *Bell Labs Technical Journal*, vol. 14, no. 4, pp. 263-285, 2010.

[81] G. S. Zervas, J. Triay, N. Amaya, Y. Qin, C. Cervelló-Pastor, and D. Simeonidou, "Time Shared Optical Network (TSON): A Novel Metro Architecture for Flexible Multi-Granular Services," *Optics Express*, vol. 19, no. 26, pp. B509–B514, 2011.

[82] K. Hattori, M. Nakagawa, N. Kimishima, M. Katayama, A. Misawa, and A. Hiramatsu, "Optical Layer-2 Switch Network Based on WDM/TDM Nano-sec Wavelength Switching," *European Conference on Optical Communications (ECOC)*, paper We.3.D.5, 2012.

[83] C. Qiao, and M. Yoo, "Optical Burst Switching (OBS)-A New Paradigm for an Optical Internet," *Journal of High Speed Networks*, vol. 8, no. 1, pp. 69-84, 1999.

[84] I. Baldine, G. N. Rouskas, and D. Stevenson, "JumpStart: A Just-In-Time Signaling Architecture for WDM Burst-Switched Networks," *IEEE Communications*, vol. 40, no. 2, pp. 82-89, 2002.

[85] S. Yao, B. Mukherjee, S. J. B. Yoo, and S. Dixit, "A Unified Study of Contention-Resolution Schemes in Optical Packet-Switched Networks," *Journal of Lightwave Technologies*, vol. 21, no. 3, pp. 672-683, 2003.

[86] J. Bowers, E. Burmeister, and D. Blumenthal, "Optical Buffering and Switching for Optical Packet Switching," *Photonics in Switching (PS)*, 2006.

[87] Y. K. Yeo, J. Yu, and G. K. Chang, "A Dynamically Reconfigurable Folded-Path Time Delay Buffer for Optical Packet Switching," *IEEE Photonics Technology Letters*, vol. 16, no. 11, pp. 2559-2561, 2004.

[88] C. H. Chen, L. Johansson, V. Lal, M. Masanovic, D. Blumenthal, and L. Coldren, "Programmable Optical Buffering Using Fiber Bragg

Gratings Combined with a Widely-Tunable Wavelength Converter," *Optical Fiber Communications Conference (OFC)*, paper OWK4, 2005.

[89] A. Agrawal, L. Wang, Y. Su, and P. Kumar, "All-Optical Loadable and Erasable Storage Buffer Based on Parametric Nonlinearlity in Fiber," *Journal of Lightwave Technologies*, vol. 23, no. 7, pp. 2229-2238, 2005.

[90] A. Liu, C. Wu, Y. Gong, and P. Shum, "Dual-Loop Optical Buffer (DLOB) Based on a 3x3 Collinear Fiber Coupler," *IEEE Photonics Technology Letters*, vol. 16, no. 9, pp. 2129-2131, 2004.

[91] R. Luijten, W. E. Denzel, R. R. Grzybowski, and R. Hemenway, "Optical Interconnection Networks: The OSMOSIS Project," *17th Annual Meeting of the IEEE Lasers and Electro-Optics Society*, 2004.

[92] R. Hemenway, R. Grzybowski, C. Minkenberg, and R. Luijten, "Optical Packet- Switched Interconnect for Supercomputer Applications," *Journal of Optical Networking*, vol. 3, no. 12, pp. 900–913, 2004.

[93] J. Gripp, J. E. Simsarian, J. D. LeGrange, P. Bernasconi, and D. T. Neilson, "Photonic Terabit Routers: The IRIS Project," *Optical Fiber Communications Conference (OFC)*, paper OThP3, 2010.

[94] S. L. Danielsen, P. B. Hansen, and K. E. Stubkjaer, "Wavelength Conversion in Optical Packet Switching," *Journal of Lightwave Technologies*, vol. 16, no. 12, pp. 2095–2108, 1998.

[95] Y. Li, G. Xiao, and H. Ghafouri-Shiraz, "Fixed Wavelength Conversion for Contention Resolution in Optical Packet Switches," *Microwave and Optical Technology Letters*, vol. 41, no. 3, pp. 185-187, 2004.

[96] J. M. H. Elmirghani, and H. T. Mouftah, "All-Optical Wavelength Conversion: Technologies and Applications on DWDM Networks," *IEEE Communications Magazine*, vol. 38, no. 3, pp. 86-92, 2000.

[97] F. Borgonovo, L. Fratta, and J. A. Bannister, "On the Design of Optical Deflection Routing Networks," *INFOCOMM*, pp. 120-129, 1994.

[98] C. Hsu, T. Liu, and N. Huang, "Performance Analysis of Deflection Routing in Optical Burst-Switched Networks," *INFOCOMM*, pp. 66-73, 2002.

[99] X. Wang, H. Morikawa, and T. Aoyama, "Burst Optical Deflection Routing Protocol for Wavelength Routing WDM Networks," *OPTICOM*, pp. 257- 266, 2000.

[100] O. Pedrola, S. Rumley, D. Careglio, M. Klinkowski, P. Pedroso, J. Sole-Pareta, and C. Gaumier, "A Performance Survey on Deflection Routing Techniques for OBS Networks ," *International Conference on Transparent Optical Networks (ICTON)*, paper Mo.C3.1, 2009.

[101] D. Blumenthal, J. E. Bowers, L. Rau, H. Chou, S. Rangarajan, W. Wang, and H. N. Poulsen, "Optical Signal Processing for Optical Packet Switching Networks," *IEEE Communications Magazine*, vol. 41, no. 2, pp. 23-29, 2003.

[102] S. Yao, S. J. B. Yoo, and B. Mukherjee, "A Comparison Study Between Slotted and Unslotted All-Optical Packet-Switched Network with Priority-Based Routing," *Optical Fiber Communications Conference (OFC)*, paper TuK2-3, 2001.

[103] S. Yao, B. Mukherjee, and S. Dixit, "Advances in Photonic Packet Switching: An Overview," *IEEE Communications Magazine*, vol. 38, no. 2, pp. 84-94, 2000.

[104] N. Le Sauze, D. Chiaroni, O. Rofidal, and A. Dupas, "New Optical Packet Synchronizer for Optical Packet Routers," *Photonics in Switching (PS)*, pp. 57-59, 2001.

[105] A. Stavdas, A. Salis, A. Dupas, and D. Chiaroni, "All-Optical Packet Synchronizer for Slotted Core/Metropolitan Networks," *Journal of Optical Networking*, vol. 7, no. 1, pp. 88-93, 2008.

[106] J. Mack, H. Poulsen, and D. Blumenthal, "40 Gb/s Autonomous Optical Packet Synchronizer," *Optical Fiber Communications Conference (OFC)*, paper OTuD3, 2008.

[107] S. H. Chin, A. Franzan, D. Hunter, and I. Andanovic, "Synchronisation schemes for optical networks," *IEE Proceedings Optoelectronics*, vol. 147, no. 6, pp. 423–427, 2000.

[108] Wikipedia, (https://en.wikipedia.org/wiki/Geometric_progression).

[109] Y. Yan, G. S. Zervas, Y. Qin, B. R. Rofoee, and D. Simeonidou, "High Performance and Flexible FPGA-Based Time Shared Optical Network (TSON) Metro Node," *Optics Express*, vol. 19, no. 26, pp. B509–B514, 2011.

[110] B. R. Rofoee, G. S. Zervas, Y. Yan, D. Simeonidou, G. Bernini, G. Carrozzo, N. Ciulli, J. Levins, M. Basham, J. Dunne, M. Georgiades, A. Belovidov, L. Andreou, D. Sanchez, J. Aracil, V. Lopez, and J. P. Fernández-Palacios, "Demonstration of low latency Intra/Inter Data-Centre heterogeneous optical Sub-wavelength network using extended GMPLS-PCE control plane," *Optics Express*, vol. 21, no. 5, pp. 5463-5474, 2013.

[111] M. Baldi, M. Corra, G. Fontana, G. Marchetto, Y. Ofek, D. Severina, and O. Zadedyurina, "Scalable Fractional Lambda Switching: A Testbed," *Journal of Optical Communications and Networking*, vol. 3, no. 5, pp. 447-457, 2011.

[112] K. Hattori, M. Nakagawa, M. Katayama, and H. Ogawa, "Method for Synchronizing Timeslot of WDM/TDM Multi-Ring Network Independent of Fiber Delay," *OptoElectronics and Communication Conference and the Australian Conference on Optical Fibre Technology (OECC/ACOFT)*, pp. 227-229, 2014.

[113] Altera, (www.altera.com).

[114] Xilinx, (www.xilinx.com).

[115] A. Hartog, A. Conduit, and D. Payne, "Variation of Pulse Delay with Stress and Temperature in Jacketed and Unjacketed Optical Fibres," *Optical and Quantum Electronics*, vol. 11, no. 3, pp. 265–273, 1979.

[116] OpenDaylight, (www.opendaylight.org).

[117] Open Networking Foundation, (www.opennetworking.org/sdn-resources/openflow).

[118] L. Schares, T. N. Huynh, M. G. Wood, R. Budd, F. Doany, D. Kuchta, N. Dupuis, B. G. Lee, C. L. Schow, M. Moehrle, A. Sigmund, W. Rehbein, T. Y. Liow, L. W. Luo, and G. Q. Lo, "A Gain-Integrated Silicon Photonic Carrier with SOA-Array for Scalable Optical Switch

Fabrics," *Optical Fiber Communications Conference (OFC)*, paper Th3F.5, 2016.

[119] Y. Ding, C. Peucheret, H. Ou, and K. Yvind, "Fully Etched Apodized Grating Coupler on the SOI Platform with –0.58 dB Coupling Efficiency," *Optics Express*, vol. 39, no. 18, pp. 5348-5350, 2014.

[120] Y. Ding, F. Ye, C. Peucheret, H. Ou, Y. Miyamoto, and T. Morioka, "On-Chip Grating Coupler Array on the SOI Platform for Fan-In/Fan-Out of MCFs with Low Insertion Loss and Crosstalk," *Optics Express*, vol. 23, no. 3, pp. 3292-3298, 2015.

[121] T. J. Seok, N. Quack, S. Han, and M. C. Wu, "50x50 Digital Silicon Photonic Switches with MEMS-Actuated Adiabatic Couplers," *Optical Fiber Communications Conference (OFC)*, paper M2B.4, 2015.

[122] H. Y. Hwang, J. S. Lee, T. J. Seok, L. Carroll, M. C. Wu, and P. O'Brien, "Packaging of 50 x 50 MEMS-Actuated Silicon Photonics Switching Device," *IEEE Electronics Packaging Technology Conference (EPTC)*, pp. 245-249, 2016.

[123] G. de Valicourt, S. Chandrasekhar, J. H. Sinsky, C-M. Chang, Y. K. Chen, M. A. Mestre, Y. Pointurier, S. Bigo, J. -M. Fedeli, L. Bramerie, J.-C. Simon, L. Vivien, A. Shen, A. Le liepvre, and G. H. Duan, "Monolithic Integrated Reflective Polarization Diversity SOI-based Slot-Blocker for Fast Reconfigurable 128 Gb/s and 256 Gb/s Optical Networks," *European Conference on Optical Communications (ECOC)*, paper 0275, 2015.

[124] N. Dupuis, B. G. Lee, A. V. Rylyakov, D. M. Kuchta, C. W. Baks, J. S. Orcutt, D. M. Gill, W. M. J. Green, and C. L. Schow, "Design and Fabrication of Low-Insertion-Loss and Low-Crosstalk Broadband 2 x 2 Mach–Zehnder Silicon Photonic Switches," *Journal of Lightwave Technologies*, vol. 33, no. 17, pp. 3597-3606, 2015.

[125] N. Dupuis, B. G. Lee, A. V. Rylyakov, D. M. Kuchta, C. W. Baks, J. S. Orcutt, D. M. Gill, W. M. J. Green, and C. L. Schow, "Modeling and Characterization of a Nonblocking 4 x 4 Mach–Zehnder Silicon Photonic Switch Fabric," *Journal of Lightwave Technologies*, vol. 33, no. 20, pp. 4329-4337, 2015.

[126] S. Nakamura, S. Takahashi, M. Sakauchi, T. Hino, M. Yu, and G. Lo, "Wavelength Selective Switching with One-Chip Silicon Photonic Circuit Including 8 x 8 Matrix Switch," *Optical Fiber Communications Conference (OFC)*, paper OTuM2, 2011.

[127] D. H. Geuzebroek, E. J. Klein, H. Kelderman, F. S. Tan, D. J. W. Klunder, and A. Dressen, "Thermal Wavelength-selective Switch Based on Micro-ring Resonators," *European Conference on Optical Communications (ECOC)*, paper PLC II 4.2.5, 2002.

[128] Cisco, "Network Optimization Through Virtualization: Where, When, What, and How?," *Cisco White Paper*, June 2016.

[129] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu, J. Zolla, U. Hölzle, S. Stuart, and A. Vahdat, "B4: Experience with a Globally-Deployed Software Defined WAN," *ACM SIGCOMM*, pp. 3-14, 2013.

[130] Floodlight, (www.projectfloodlight.org/floodlight).

[131] Open Network Operating System (ONOS), (www.onosproject.org).

[132] Ryu, (https://osrg.github.io/ryu).

[133] Open Stack, (www.openstack.org).

[134] Wireshark, (www.wireshark.org).

[135] Altera SignalTap, (ftp://ftp.altera.com/up/pub/Altera_Material/12.1/Tutorials/Verilog/SignalTap.pdf).

[136] Wikipedia, (https://en.wikipedia.org/wiki/Serial_Peripheral_Interface_Bus).

[137] M. Z. Feng, K. Hinton, R. Ayre, and R. S. Tucker, "Energy Efficiency in Optical IP Networks with Multi-Layer Switching," *Optical Fiber Communications Conference (OFC)*, paper OWI2, 2011.

[138] Y. Zhang, P. Chowdhury, M. Tornatore, and B. Mukherjee, "Energy Efficiency in Telecom Optical Networks," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 4, pp. 441–458, 2010.

[139] K. Hinton, J. Baliga, M. Feng, R. Ayre, and R. S. Tucker, "Power Consumption and Energy Efficiency in the Internet," *IEEE Network*, vol. 25, no. 2, pp. 6-12, 2011.

[140] G. Shen, and R. S. Tucker, "Energy-minimized design for IP over WDM networks," *Journal of Optical Communications and Networking*, vol. 1, no. 1, pp. 176–186, 2009.

[141] P. Samadi, D. Calhoun, H. Wang, and K. Bergman, "Accelerating Cast Traffic Delivery in Data Centers Leveraging Physical Layer Optics and SDN," *International Conference on Optical Network Design and Modelling (ONDM)*, pp. 73-77, 2014.

[142] P. Samadi, V. Gupta, B. Birand, H. Wang, G. Zussman, and K. Bergman, "Accelerating Incast and Multicast Traffic Delivery for Data-Intensive Applications using Physical Layer Optics," *ACM SIGCOMM*, 2014.

[143] H. Wang, Y. Xia, K. Bergman, E. T. Ng., S. Sahu, and K. Sripanidkulachai, "Rethinking the Physical Layer of Data Center Networks of the Next Decade: Using Optics to Enable Efficient *-cast Connectivity," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 3, pp. 52-58, 2013.

[144] P. Samadi, J. Xu, and K. Bergman, "Experimental Demonstration of One-to-Many Virtual Machine Migration by Reliable Optical Multicast," *European Conference on Optical Communications (ECOC)*, paper We.3.5.7, 2015.

[145] P. Samadi, V. Gupta, B. Birand, H. Wang, R. Jensen, G. Zussman, and K. Bergman, "Software-Addressable Optical Accelerators for Data-Intensive Applications in Cluster-Computing Platforms," *European Conference on Optical Communications (ECOC)*, paper Th.2.2.2, 2014.

[146] Twitter, (https://blog.twitter.com/engineering/en_us/a/2010/murder-fast-datacenter-code-deploys-using-bittorrent.html).

[147] P. Samadi, V. Gupta, B. Birand, H. Wang, R. Jensen, G. Zussman, and K. Bergman, "Virtual Machine Migration over Optical Circuit Switching Network in a Converged Inter/Intra Data Center Architecture," *Optical Fiber Communications Conference (OFC)*, paper Th.4G.6, 2015.

[148] M. Wiesmann, F. Pedone, A. Schiper, B. Kemme, and G. Alonso, "Data replication techniques: a three parameter classification," *IEEE*

Symposium on Reliable Distributed Systems (SRDS), pp. 206-212, 2000.

[149] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time url spam filtering service," *IEEE Symposium on Security and Privacy (SP)*, pp. 447-462, 2011.

[150] A. Muhammad, G. Zervas, D. Simeonidou, and R. Forchheimer, "Routing, Spectrum and Core Allocation in Flexgrid SDM Networks with Multi-core Fibers," *International Conference on Optical Network Design and Modelling (ONDM)*, pp. 192-197, 2014.

[151] A. Rylyakov, J. E. Proesel, S. Rylov, B. G. Lee, J. F. Bulzacchelli, A. Ardey, B. Parker, M. Beakes, C. W. Baks, C. L. Schow, and M. Meghelli, "A 25 Gb/s Burst-Mode Receiver for Low Latency Photonic Switch Networks," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 12, 2015.