Data center optical networks: short- and long-term solutions

Miquel Angel Mestre Adrover

NOKIA Bell Labs

Spécialité : Electronique et communications

Ecole doctorale : Informatique, Télécommunications et Electronique de Paris

Présentée par

Miquel Angel Mestre Adrover

Pour obtenir le grade de
DOCTEUR DE TELECOM SUDPARIS

Data center optical networks: short- and long-term solutions

Réseaux optiques pour les centres de données: solutions à court et long terme

Soutenue le 21/10/2016

Devant le jury composé de:

Directeur de thèse
Prof. Badr-Eddine Benkelfat

Encadrants de thèse
Dr. Yann Frignac
Dr. Yvan Pointurier

Rapporteurs
Prof. Lars Dittmann
Prof. Christophe Peucheret

Examinateurs
Prof. Delphine Marris-Morini
Dr. Cédric Ware

N° NNT : 2016TELE0022

Acknowledgements

Many people say that writing the PhD thesis was one of the hardest moments in their life. Nevertheless, when you do it in a place like Bell Labs it becomes an amazing period. And this is not only because you have the means for pursuing the greatest challenges, but also because you are surrounded by the best researchers around the globe, which turn to be also wonderful people who won't hesitate in giving you a hand when you need it. Therefore I want to thank my colleagues and friends from Villarceaux, from whom I learned so much during the past three years, and with whom I have shared so many good moments, both inside and outside the lab. Specially I would like to thank Yvan for being an excellent supervisor (lucky me because all the others left), an exemplary boss and an amazing person. I cannot forget my lab/group mates, specially Haïk, Philippe, José and Nihel (I know you always wanted to be a lab rat); thank you guys for being always there! As well as the colleagues from the other group, with whom we "did not interact so much", but we kept doing great stuff together; specially with Rafael, Jeremie and Amir, who I bothered so many times. Finally I want to thank Sébastien for giving me the chance to work in such an amazing place.

Before switching to my mother tongue, I want to thank the many friends that I have met while moving from one place to another during these last years, you have all become a very important part of my life.

Now en mallorquí. Després de sa defensa, quan vaig tornar a cases, tothom que me trobava me deia, "Estàs d'enhorabona", "Això ja és molt important". I sa veritat és que jo també ho trob, però no per es doctorat. Estic d'enhorabona per sa gent que m'envolta: uns amics que sempre que tornes pareix que no has partit mai, una gran familia que t'omple de felicitat

i una dona que sents que t'estima com ningú. Això és lo realment important, i de lo que estic més orgullós. I per això vos vull donar ses gràcies a tots! Gràcies de veres per ser com sou! I especialment als meus pares, gràcies pes vostro suport incondicional que m'ha permès arribar fins aquí. I no te preocupis que no me podria oblidar mai de tu, que has sofrit ses penúries i alegries d'aquest doctorat com si fossis jo mateix. Cati, tu has estat lo millor que me podria haver passat mai durant aquest doctorat. I sí, com ja vaig dir aquell dia que me vaig quedar a mitges, gràcies per aguantar tant. No ha estat gens fàcil, però lo bo diven que se fa esperar, i noltros finalment hi hem arribat! Gràcies per ser-hi sempre, per celebrar amb jo es bons moments i sobretot per donar-me una aferrada durant es que no ho han estat tant. Gràcies per recolzar-me, per anirmar-me i sobretot per fer-me tant feliç. T'estim moltíssim videta!

Abstract

The spread of cloud services is driving a relentless increase of traffic demand in large-scale data centers, which is nearly doubling every year. After revealing the main trends driving such emerging traffic, and the technological evolution of data center networks, we present short- and long- term solutions for their physical intra-connection.

Today, rapidly-growing traffic in data centers highlights the urgent need for high-speed low-cost interfaces. Therefore, in the short-term we propose novel high-data rate optical transceivers enabling up to 200 Gb/s transmission, leveraging low-cost intensity-modulation and direct-detection schemes combined with advanced multi-level modulation formats and novel integrated high-bandwidth devices.

Notwithstanding, increasing data centers capacity while keeping today's multi-stage electronic switching topologies leads to highly complex networks composed of hundreds of thousands or even millions of networking components, which results into future unsustainable operational costs and exorbitant power consumption. In order to deal with such issues we propose the use of a burst optical slot switching (BOSS) ring-based flat architecture, which combines inner-ring transparency with high-data rate transponders performing statistical multiplexing on a microsecond scale. When comparing to current electronic switching networks, BOSS allows reducing the number of networking components (interfaces and cables) by more than a 100-factor, while halving the amount of energy consumed.

We investigate several technological approaches to best implement BOSS optical nodes and, accordingly, suggest modulation schemes that allow maximizing capacity and reach of such systems. We ultimately propose

the use of coherent-optical orthogonal frequency division multiplexing (CO-OFDM), which enables the use of low-cost components while increasing the average capacity (+30%) and reach (+40%) with respect to traditional Nyquist pulse-shaped quaternary-amplitude modulation (QAM) schemes; hence paving the way to highly scalable and sustainable data centers.

Contents

Chapter 1

Introduction

1.1 A hint of history on data centers

Data centers form today the brain that makes possible all cloud and web services extensively used around the globe. Such large facilities include a vast number of interconnected servers that store and process all information available in the world wide web and give rise to cloud/internet applications that we use in a day-by-day basis (e.g., cloud storage, video streaming, image and video sharing, social networks, etc.). Despite the relatively youth of such a digital interconnected world, the origin of data centers dates back from the early days of modern computing.

The story of data centers began in mid 1950 with the appearance of first commercial general purpose computers, called mainframes. Built by companies such as IBM, Remington Rand or General Electrics, mainframes were the first computers used by businesses to process data [1]. However, due to their high cost, even large corporations could typically afford only one system. With sizes of several square meters, mainframes were usually placed in a so-called computing room, and they were time-shared by multiple users performing different tasks. The first mainframes were managed through punch cards or paper tapes. Later teletypes, followed by terminals, could be attached to the mainframes which would interpret their commands through proprietary protocols [2].

Along the following decades, computing systems rapidly shrank in size. First the introduction of solid-state transistors, replacing traditional vacuum tubes, gave rise to minicomputers. DEC was one of the pioneering companies commercializing such computers in 1965. Thanks to their reduced cost and size, businesses could now afford buying several minicomputers. A few years later, Intel introduced the first microprocessor (1971). Electronic integration allowed building more compact and less costly computers well-suited for the end-user. Such machines were first called microcomputers and they evolved to the well-known personal computer (PC) used today. Being able to have a workstation per user, rapidly appeared again the concept of sharing resources; at this point through an early concept of local area network (LAN) [2, 3]. The first commercial networking system widely used to interconnect microcomputers was called Attached Resource Computer NETwork (ARCNET). Announced in 1977 by Datapoint Corporation, ARCNET was used to connect end-user workstations to shared storage and computing resources [4]. This way, users could rely on simple and relatively inexpensive terminals, while sharing the processing and storage capacity of more powerful machines (today called servers) typically placed in dedicated rooms. Nowadays, we call these rooms enterprise data centers.

The appearance of the internet along the late 80s and the concept of world wide web in 1990 gave rise to a new era of globe interconnection. With the dot.com bubble taking place during the late 90s would also emerge the data center business. Many companies, associations and even individuals were demanding a permanent presence on the internet. Along that period (1995-2000), the number of web sites existing in the world wide web increased from tens of thousands to more than ten millions [5]. Hence a large number of data center facilities opened to host the uncountable web sites and web services appearing in a day-basis. Since then, internet backbone traffic, and data center traffic demands have not stopped growing.

1.2 Global and intra data center traffic evolution

The evolution of the backbone network traffic since the appearance of the internet is shown in Fig. 1.1. This analysis was performed by Nokia Bell Labs Consulting, which identified the main trends inducing the traffic growth [3]. As depicted in the graph, prior to the 2000s, internet traf-

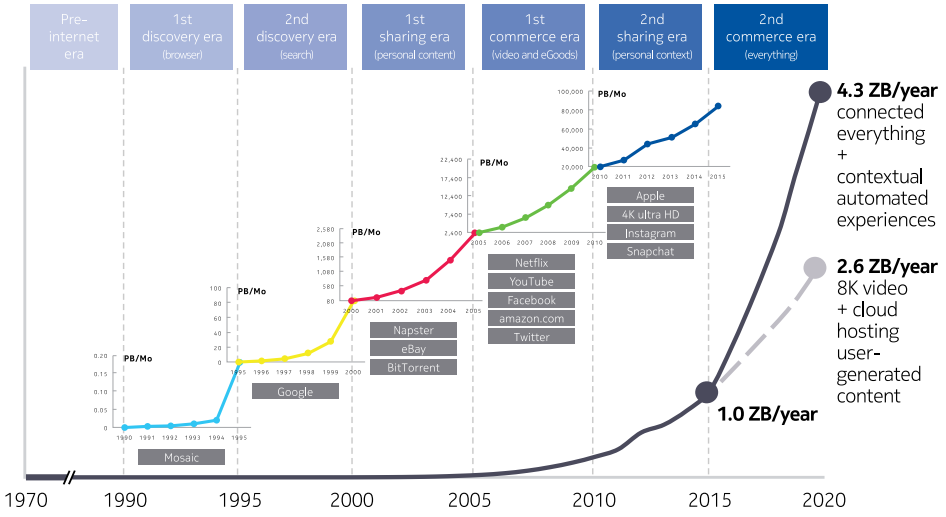


Figure 1.1: Backbone network traffic growth since the appearance of the Internet. From [3].

fic was governed by the web content. With the first browser released in 1993 (Mosaic) and the main web search engines (e.g., Altavista, Yahoo, Google) appearing during the following years, Internet traffic started growing at a fast pace, leading to 80 PBytes circulating globally each month (PB/Mo) when entering into the new millennium. In 1999 the concept of peer-to-peer file-sharing became popular with the appearance of Napster, which was mainly created to share music files between end-users. Many applications were developed following the peer-to-peer file-sharing concept (e.g., eDonkey, Bittorrent, Gnutella), which became extensively used by the internet users to exchange all kind of files (e.g., video, films, games, music, books), driving a rapid traffic increase, which quickly surpassed the 1-EB/Mo (1000 PB/Mo) around 2004. In February of the same year Facebook's website was launched, starting a new era of social networking and sharing of personal content. Video streaming became also popular around that period with the appearance of Youtube (2005) and later Netflix, which started the online streaming business in 2007. These novel trends, led to a factor ten traffic growth between 2005 and 2010, when 20 EB/Mo were transiting core networks.

Over the last few years, the smart-phone revolution has provided unlimited Internet access to everyone. Such technology combined with novel

(e.g., Instragram, Snapchat) and existent (Facebook, Twitter) social applications have led to a widely-spread "modus operandi": sharing everything at anytime and anywhere. Such vast sharing movement together with other coming trends (ulta-high-definition video or cloud storage), keep increasing backbone traffic, which surpasses today 1 ZB/year. Along the years to come another revolutionary technology is expected: the internet of things (IoT). A new era is about to start, during which not only everyone will be connected but also "everything". IoT will allow for a new automated world where machines will be efficiently managed from the cloud, giving rise to smart cars, buildings, hospitals and cities [3, 6]. Real-time extensive communications between a vast number of machines and cloud systems is supposed to generate an unprecedented traffic growth possibly exceeding 4 ZB/year in 2020 [3].

All the above mentioned applications and services, have been and are currently being hosted by data centers distributed around the globe. Hence, similarly to backbone networks, data centers have experienced an enormous traffic growth. Fig. 1.2 shows the global data center traffic evolution and

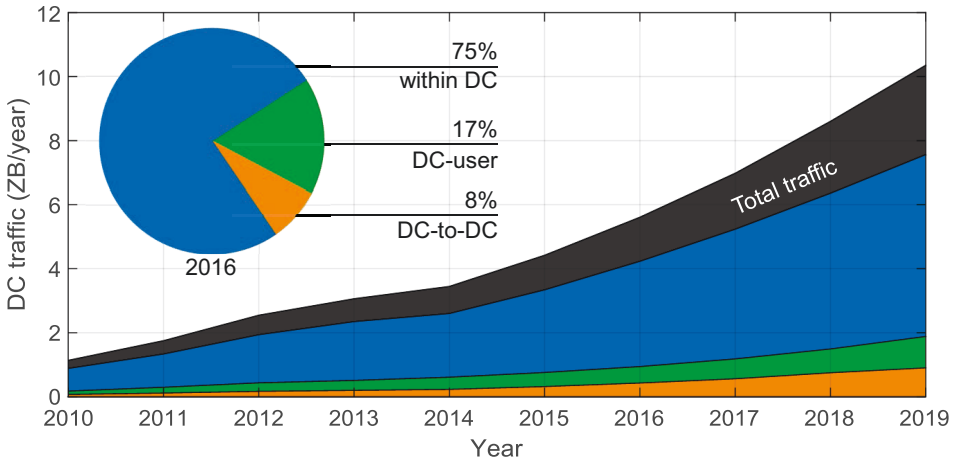


Figure 1.2: Data center traffic evolution and forecast between 2010 and 2019. Data extracted from Cisco global cloud index white papers [6–10].

data center (within DC in blue), 2) traffic exchanged between external users and data centers (DC-user in green) and 3) traffic exchanged between data centers (DC-to-DC in orange). The main graph depicts the evolution of the traffic for each type (described color map) and in total (dark gray). One can clearly depict the predominance of the internal data center traffic, accounting for 75% of the total traffic (see sector graph showing the percentage for the traffic forecast in 2016). This percentage spotlights the large amount of traffic exchanged between servers and/or storage units. The remaining traffic leaves the facility to communicate with users (17%) and other data centers (8%). Cisco global cloud index analysis shows that the total amount of traffic generated globally in data centers exhibits a ten factor growth (2010-2019) reaching more than 10 ZB/year in 2019.

The traffic analysis presented in Fig. 1.2 shows the data center's traffic trend in a global scale. Nonetheless, different kinds of data centers exhibit different trends. For instance, in its Global Cloud Index reports, Cisco distinguishes between two main kinds of data centers: traditional and cloud. The first kind hosts non-cloud traffic, which is typically attributed to research/university campus and enterprises using their own data center to provide the required resources to their users/employees. The traffic generated in such data centers is supposed to increase by a factor of 1.7 between 2010 (1.01 ZB/year) and 2019 (1.73 ZB/year) [6, 7]. On the other hand, traffic generated by cloud data centers is exhibiting a $\times$65-factor (from 0.13 to 8.62 ZB/year) growth during the same period [6, 7]. Such data centers provide different kinds of cloud services to end users and enterprises. Cloud services can be divided in three different models [11]:

- Software as a Service (SaaS): provide the capability of using services placed in a cloud infrastructure. Such services can be easily accessed from a web browser or a program interface. Some examples of SaaS are Google Drive, Dropbox or SalesForce.com.
- Platform as a Service (PaaS): provide the capability of developing applications typically in a proprietary platform that will run in a certain environment. Popular PaaS examples are Facebook Apps and Google App Engine.
- Infrastructure as a Service (IaaS): provide the user computing and storage resources to deploy and run freely their own software. One of the pioneers in such kind of services is Amazon, with their Amazon

EC2. This kind of services is today used for a large number of small and medium enterprises, which rent resources in cloud data centers to host their applications and services instead of building their own data centers.

Along this thesis we will focus on large scale cloud data centers, which can host tens or even hundreds of thousands of servers. Some of the major cloud data center are hosted by popular service providers such as Google, Microsoft, Facebook, Amazon, etc. Such large scale data centers may exhibit even faster growth. For instance, Google reported a $\times 50$ traffic growth between 2008 and 2014, which means that the traffic in their facilities is doubling every 12 to 15 months [12]. The unstoppable traffic growth forces largest data centers updating their infrastructure every few years. In the following sections we describe the transformations on networking architecture and the progress on both electrical and optical sub-systems that have led to modern data centers.

1.3 Towards modern data centers: topology and sub-systems

In order to process and store all data required to support the vast number of applications and services currently available in the cloud, data centers make use of a massive number of servers, which can be counted in hundreds or thousands in small/medium facilities, and in tens or hundreds of thousands in large-scale data centers. The inter-connection of such vast amount of servers is not trivial and hence it has been extensively studied along the past years, giving rise to many architectures (e.g. trees, Folded Clos, BCube, DCell, torus, etc.) [13]. Along this section we will describe two of the most popular topologies used in the past and in the present in large data centers: multi-tier trees and Folded Clos.

First data center generations usually adopted tree-like topologies, usually forming two or three switching levels using high-radix and expensive switches [14]. Fig. 1.3 shows a typical 3-tier multi-rooted tree-like data center network, which includes three levels (stages/tiers) of switching. The lower edge stage includes the so-called Top-of-Rack (ToR) switches, which interconnect 20-40 servers located in racks. To do so, ToRs typically contain

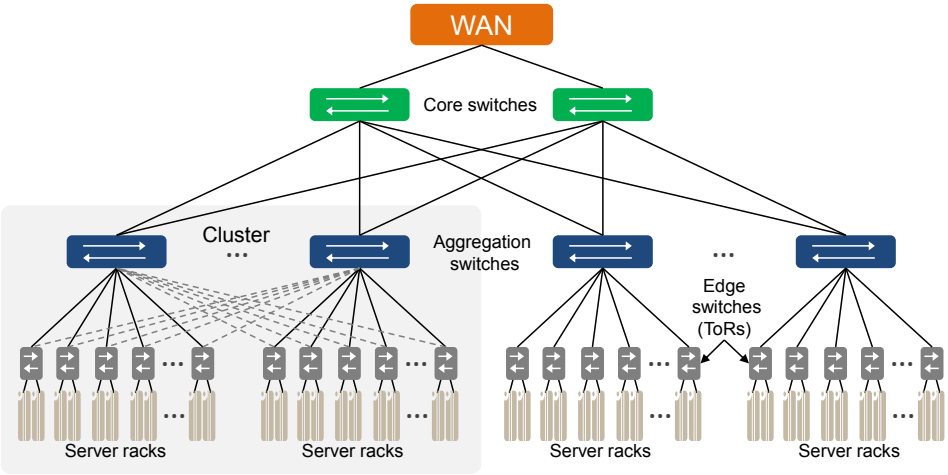


Figure 1.3: Traditional three-tier multi-root tree topology. Left-most: illustration of typical cluster implentation.

24 or 48 downlink ports with capacities ranging between 1 and 10 Gb/s. A large number of Top-of-Rack (ToR)s are then linked to one or several switches of the aggregation layer using several 10-Gb/s uplink ports. The aggregation stage usually contains large switches to interconnect a vast number of ToRs (up to 32-512 ports). In large data centers, sometimes aggregations switches are grouped together forming clusters, as shown in the left-most tree-branches in Fig. 1.3. This approach provides redundancy and enhanced switching capacity. Then, an extra core layer is used to exchange traffic from several aggregation switches or clusters and to connect to the external wide area network (WAN). Generally, in order to leverage the high cost of the large switches, the different switching levels are over-subscribed[1] by factors of 5:1 or higher [15]. For instance, earlier generations of Facebook data centers were built using a 3-tier 4-post clustered[2] network, exhibiting 10:1 and 4:1 over-subscription ratios at lower and higher tiers, respectively [16].

The above mentioned approach was extensively used in small-medium size campus and enterprise data centers, at which North-South[3] traffic was

[1]Over-subscription of x:1 means that the switching capacity of a certain stage can handle only a $1/x$ of the full capacity of the inferior stage

[2]N-post clusters include N aggregation switches and their linked ToRs and servers.

[3]North-South traffic refers to communication between servers and external clients (WAN).

predominant (95%) [17]. Despite the fact that several big players hosting today's largest cloud data centers also adopted such architecture in their early days [12, 16], they quickly realized that such approach would not support the rapid traffic growth that they were experiencing [12, 16]. First of all, in such architecture inter-cluster connections were commonly over-subscribed [15], hence limiting large-scale server-to-server communications (East-West traffic), which represents today 75% of the total traffic in a data center (shown above in Fig. 1.2). Reducing the over-subscription can be done by increasing the switching capacity (i.e., adding more switches and/or larger capacity switches). Nevertheless, due to the large cost of the high-end switches used by then, scaling-up capacity would lead to prohibitive costs [14]. Furthermore, another drawback of such approach is the low resiliency to failure, due to the "reduced number" of high-radix switches. For instance, in the Facebook data center mentioned above, each cluster contained four aggregation switches, and inter-cluster interconnection was performed via four core switches. Using this architecture an aggregation or core switch failure leads to a 25% reduction in intra- or inter-cluster capacity respectively. In order to overcome such pain points, many large-scale cloud data center have moved towards a better suited approach, which combines a Folded Clos topology with the utilization of a larger number of

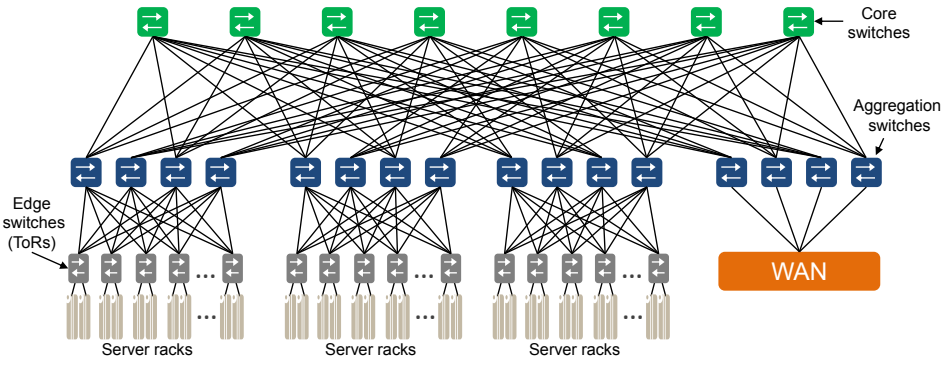


Figure 1.4: Modern Folded Clos topology.

commodity lower-radix switches profiting from a Folded Clos topology to create a highly interconnected network. For instance Google moved from a 4-post clustered structure (2004), using 512-port 1-Gb/s switches, to a Folded Clos topology (2005), building homemade aggregation and spine blocks of 32×10 Gb/s ports, which were based on 4-port 10-Gb/s silicon merchant chips (switches) [12]. The use of a Folded Clos topology allows for an arbitrary scaling of the size and bandwidth of a data center, through the addition of further switching stages [12]. Furthermore, the large number of links placed between the stages provides a large number of possible paths performing single server-to-server communication. Multi-path connectivity yelds to high failure resiliency [14]. The other main difference between the current and the latter architecture is the placement of the WAN external gateways. As mentioned above, in modern cloud data centers only 25% of the traffic enters/exits the facility. Hence, WAN gateways are now typically located in the lower levels of network [12, 18]. This way the remaining North-South traffic is converted to East-West, which allows optimizing the network for the latter one.

In order to cope with the rapidly growing bandwidth demand shown in the previous section, data centers have been upgrading their switches and servers towards higher capacities. Along the past years, several protocols have been used to perform data transmission within data centers (e.g. InfiniBand, Fibre Channel or Ethernet). Nevertheless, the industry has been envisioning for a long time the convergence into a single protocol capable to deal with the communications of all networking and computing elements in a data center. The emergence of high data-rate Ethernet, with capacities of 40 and 100 Gb/s, has made this protocol a potential candidate to displace the other ones and take the lead into a uni-protocolar data center network [19].

Fig. 1.5 shows the evolution and forecast of Ethernet interfaces since the 2000s [20, 21]. After the decline of Fast Ethernet (FE), reaching capacities of 100 Mb/s, soon Gigabit Ethernet (GE) governed server speeds in the early 2000s. Nevertheless, 10 GE started appearing in the top switching levels[4] after its standardization in 2002 [22]. The appearance of 10 GE LAN integrated on motherboards, drove a rapid adoption of 10 Gb/s data rate on a server level. 10 GE switch port shipments have exhibited 30%

[4]Google was already using 10 Gb/s links in 2004 to link their aggregation switches in a 4-post clustered topology [12].

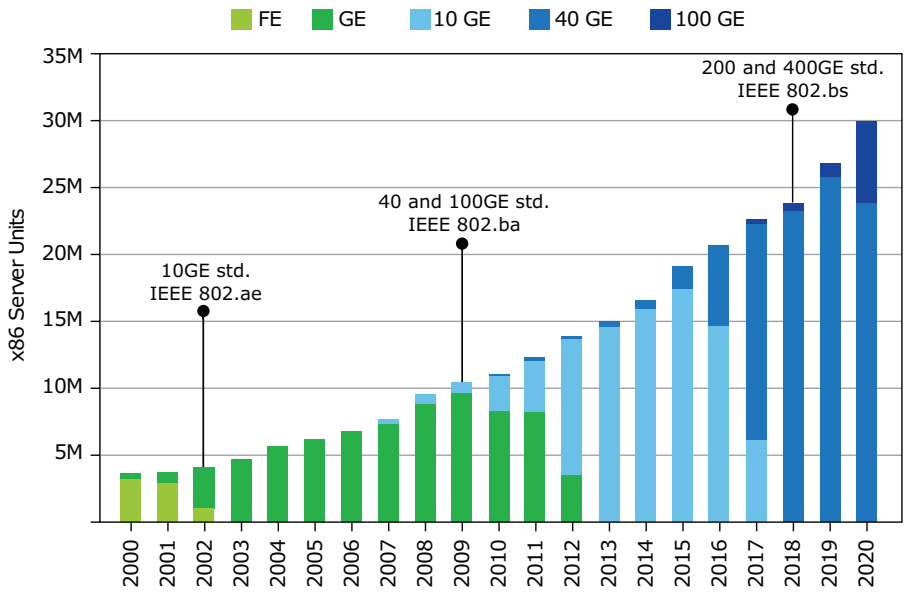


Figure 1.5: Server I/O data rate evolution and approval of Ethernet standards (From [20]).

growth between 2014 and 2015 (last quarters, 4Q), reaching 9.5 millions shipments (by 4Q2015) [23]. Nevertheless after its standardization in 2009, 40 GE, including 4×10-Gb/s lanes, quickly entered into the market. Google reported the introduction of 40 Gb/s links in both switching fabric and servers in 2012 [12]. In 2015, 40 GE market revenue denoted nearly a 50% increase over 2014 [23].

In the near term, 10 and 40 GE are expected to lead the Ethernet market, while vendors and service providers get ready for the jump to 100 GE, which is already about to happen. Changing to 100-GE produces a sudden disruption in what refers to the basic data rate, since the most popular 100 GE implementation carries 4 lanes at 25 Gb/s. Nevertheless, when 100 GE was standardized in 2009, there was no standard supporting single lane links at 25 Gb/s. Thus, in 2014 a group formed by the most powerful service providers and vendors (i.e., Google, Microsoft, Broadcom, Arista and Mellanox) created a consortium to promote the standardization of 25 GE [24], which standard was very quickly approved in June 2016 [25]. 25 Gb/s is expected to replace 10 Gb/s rates at all data center levels: starting from top tiers, using 100 GE (4×25-Gb/s) instead of 40 GE (4×10-

Gb/s), and down to the servers, supporting 25 GE instead of 10 GE [26]. Prompted by the relentless traffic growth, the IEEE is already defining new standards to support 50 Gb/s on a single lane and multipe-lane 200 and 400 GE, which are expected to be approved by 2018 [27, 28].

1.4 The role of optics in current data centers

During the early 2000s, most of the connections performed in a data center (i.e., server-switch, and switch-switch) were based on copper links due to their low price and power consumption. Nevertheless, copper induces large losses and distortions, which limit the transmission reach, e.g., 15 m for 10GBASE-CX4 (10 GE) [29]. Furthermore, the reach is further reduced when increasing the data rate. For instance, 40GBASE-CR4 (40 GE) and 100GBASE-CR4 (100 GE) standards exhibit limited reaches of 7 and 5 meters, respectively[5] [26]. Such distances are sufficient to connect servers to ToRs placed within the same rack. Nevertheless, when trying to interconnect the higher levels of the switching fabric, copper leads to tight restrictions when designing the data center physical layout. In order to overcome such limitations data centers profit from the extended reach of fiber optic systems, which can transmit more than 10, 40 or 100 Gb/s over hundreds of meters or tens of kilometers depending on the technology used[6]. Furthermore, fiber cables are much thinner and lighter than copper ones and can be further bent. Such properties are advantageous when designing intra- and inter-rack connectivity, at which usually a large number of cables are bundled together. Fiber bundles not only reduce the space required in cable trays but also diminish airflow blockage in racks, thus improving cooling efficiency [30]. These advantages together with the diminishing cost of optical systems, lead to a progressive replacement of copper links for optical ones, which are expected to account for more than 60% of data center links by 2017 [31].

For data center applications, optical transceivers are implemented in pluggable modules, which can be easily connected (plugged) into switch

[5] 10GBASE-T (10 GE) and 40GBASE-T (40 GE) can extend the reach of copper links to 100 and 30 meters, respectively, at the expense of higher cost, consumption and latency [26].

[6] Only accounting for non-coherent technology. More costly coherent technology allows transmitting over thousands of kilometers.

Table 1.1: Implementation technology, number of fibers, reach and power consumption of the most commonly used pluggable modules [22, 32, 33, 35–37].

	Type	Pluggable	Technology	Fibers	Reach	Consumption
10 G	SR	SFP+	1 x 10-Gb/s VCSEL	2 x MMF	400 m	<1 W
	LR	SFP+	1 x 10-Gb/s DFB	2 x SMF	10 km	<1 W
40 G	SR4	QSFP+	4 x 10-Gb/s VCSEL	8 x MMF	150 m	<1.5 W
	LR4	QSFP+	4 x 10-Gb/s DFB (WDM)	2 x SMF	10 km	<3.5 W
100 G	SR10	CFP/CFP2	10 x 10-Gb/s VCSEL	20 x MMF	150 m	CFP: <24 W CFP2: <12 W CFP4: <6 W QSFP28: <3.5 W
	SR4	CFP/CFP2/ CFP4/QSFP28	4 x 25-Gb/s VCSEL	8 x MMF	100 m	
	LR4	CFP/CFP2/ CFP4/QSFP28	4 x 25-Gb/s DFB (WDM)	2 x SMF	10 km	

racks and network interface controllers (NICs) placed in servers. Table 1.1 describes the technology and specifications for several frequently used optical pluggable modules commercially available today[7]. These modules are composed by the following elements: electrical on-board connectors to interface with the switch or server NIC, a cage for electromagnetic interference containment and the optical transceiver, which typically includes transmitter and receiver optical sub-assemblies (TOSA and ROSA). A TOSA incorporates: 1) one or multiple lasers and optionally modulators (when not using direct laser modulation), 2) the required electrical driving circuits and 3) the multiplexing (when transmitting several wavelengths) and coupling systems (e.g., micro-lenses) to output the light into the fiber(s). On the other hand a ROSA holds the optical systems used to couple out the light from the fiber(s) (and demultiplex the channels when required), one or multiple photodetectors and trans-impedance amplifiers (TIAs) [17].

As shown in Table 1.1, all Ethernet standards (10, 40 and 100 GE) include at least two types of modules: short reach (SR) and long reach (LR)[8] [22, 33]. SR modules are implemented using vertical cavity surface-emitting lasers (VCSELs), being such the most extensively used in today's data centers due to their extremely low cost [38]. Nevertheless, VCSELs need to be coupled to multi-mode fibers (MMFs) whose reach is limited to

[7]The implementation, reach and power consumption information have been extracted from the IEEE Ethernet standards [22, 32, 33] and from the catalog of several vendors (i.e., Finisar [34, 35], Arista [36] and Cisco [37]).

[8]The number after SR/LR indicate the number of duplex data lanes supported by the module.

few hundred meters (100-400 m when using the latest OM4 MMF) [32]. However, very large data centers require also cables of a few kilometers to link aggregation/core switches placed in different floors or buildings. Such long links are typically implemented with single-mode fiber (SMF) connecting two LR pluggable modules, which include directly modulated distributed feedback (DFB) lasers. Such modules are today more expensive but they can achieve up to 10-km transmission distances. Popular service providers such as Facebook and Google have revealed their preference towards the deployment of SMF for several reasons: 1) it has longer reach, which provides data center layout flexibility; 2) it can support very high data rates with negligible penalty, which allows the re-utilization of the same fiber plant after technology upgrade; and 3) it is cheaper than MMF (the fiber itself) [39].

Most manufacturers produce their modules following the design established in multi-source agreements (MSAs), which define electrical, mechanical and thermal specifications to ensure inter-operability but leave open their implementation to vendors [17]. Table 1.1 describes some of the commonly used form factors in data centers: small form factor (SFP), quad small form factor (QSFP) and C form factor (CFP). The first one, SFP, was originally supporting 1 GE (not shown in the table). Based on such design appeared the SFP+, which has the same form factor but allows transmitting full-duplex 10 Gb/s through a pair of connected fibers. SFP(+) have quite low power consumption (less than 1 W) and small footprint (8.5×13.4×56.5 mm), which allows placing up to 56 modules within a single 1U-rack[9].

Scaling to 40 Gb/s is achieved by increasing the number of 10-Gb/s lanes, which can be done in a slightly larger QSFP+ form factor, where Quad (Q) stands for four lanes (4×10 Gb/s in this case). SR QSFP+ modules require four pairs input/output fibers, which are arranged in fiber ribbons and connected to the modules through multi-fiber push on (MPO) connectors. On the other hand, in LR QSFP+, the four channels are multiplexed in the same fiber through different wavelengths using the coarse wavelength-division multiplexing (CWDM) technique. Hence LR modules just need one input/output fiber pair. Despite overall size (13.5×18.4×72.4 mm) and power consumption (less than 3.5 W) are higher than SFP+, the dimension and consumption per-bit are reduced when using QSFP+. Using

[9] 1U is the height of a single rack unit (44.45 mm).

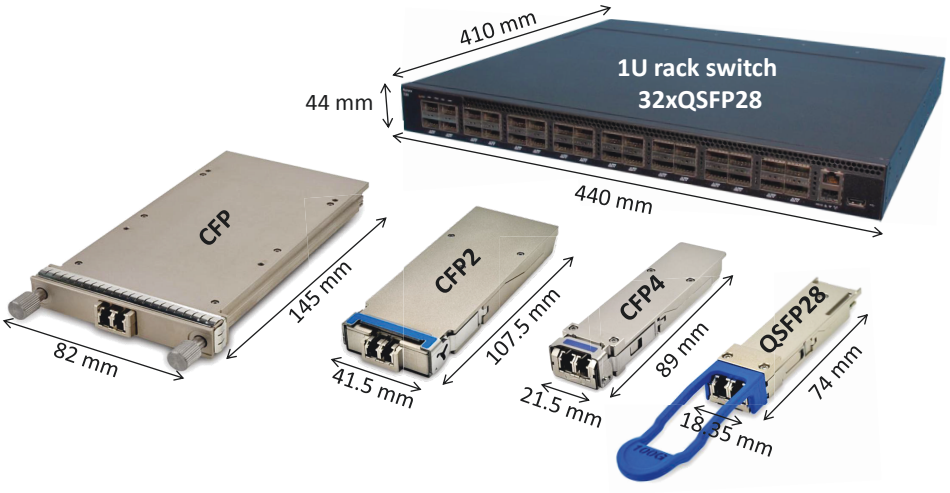


Figure 1.6: 100-GE pluggable form factors (From Finisar's product catalog [35]) and exemplary switch support 32×100-GE ports (From Netberg [40]).

the latter up to 36 pluggables can fit in a 1U-rack switch, which supports a total capacity of 1.44 Tb/s, being such 2.6 times higher than the capacity supported when using SFP+ (560 Gb/s). Furthermore, upgrading to QSPF+ results in a reduced per-bit power consumption, which is smaller than 0.1 W/Gb/s.

The jump to 100 Gb/s was initially challenging if such small form factors were to be kept. Consequently, larger modules were initially proposed based on the CFP MSA, at which C stands for hundred (in latin *centum*). As observed in Table 1.1, two main approaches were proposed for 100 GE: keeping the same basic speed while increasing the number of lanes (10×10 Gb/s), or increasing the data rate while keeping the four scaling factor (4×25 Gb/s). Although SR modules are implemented using both approaches, LR pluggables are based on the second approach, implemented using CWDM. Fig. 1.6 shows different generations of 100-GE pluggable modules and an exemplary 1U-rack switch. Clearly, the size of 100-GE pluggables have been reduced along the years, as well as the power consumption (see Table 1.1). The latest generation is implemented in QSFP28[10], which have the same dimensions and power consumption than the 40-Gb/s QSFP+, while carrying more than twice its capacity (100 Gb/s).

[10]The 28 suffix stands for 28 Gb/s, being such the maximum gross data rate supported by the module to account for possible forward error correction (FEC) overhead.

As mentioned above, the Ethernet Alliance is working on new standards supporting 200 and 400-Gb/s data rates. Accordingly, the main manufacturers are already investigating the implementation of such high data-rate interfaces while keeping low footprint and power consumption. In order to reach such speeds, more advanced and spectrally-efficient modulation formats (i.e., 4-level pulse amplitude modulation (PAM-4)) are envisioned to keep down the number of lanes while re-using today's bandwidth limited components [27, 28]. Current pluggables use the most basic 2-level pulse amplitude modulation (PAM-2) format, which transmits only 1 bit/symbol using two intensity levels of the transmitted light. Differently, PAM-4 makes use of four intensity levels, which can encode 2 bits in each symbol, thus doubling capacity when compared to PAM-2 while keeping the same symbolrate, and hence the same required bandwidth.

One of the most promising implementations for 200-Gb/s modules is based on 4× 50Gb/s in a QSFP56 form factor, which would have the dimensions of QSFP but increased lane data rate (up to 56 Gb/s to include possible FEC and protocol overhead). The 50-56Gb/s single lanes will be likely implemented using PAM-4 signals at 25-28 GBd. On the other hand 400-GE design is more controversial due to its challenging implementation. Several approaches are being proposed, all making use of different number of lanes to achieve a total 400-Gb/s data rate: 16×25 Gb/s (PAM-2), 8×50 Gb/s (PAM-4 at 25-28 GBd) and 4×100 Gb/s (PAM-4 at 50-56 GBd). The research community is today putting a lot of effort in demonstrating 100-Gb/s single-lane (single-carrier) interfaces, which would allow continuing with the 4-factor lane scaling currently used. The 100-GE on a single lane implemented in a SFP+ form factor is considered today the *Holy Grail* of data center pluggable modules due to its challenging development [32, 41].

1.5 Future challenges of large-scale data centers

1.5.1 Dissecting a large-scale data center

As described in the previous sections, in order to support the ever increasing bandwidth demand, large data centers rely on a Folded Clos networks based on an enormous number of electronic switches that provide full server connectivity. Main service providers chose such topology because it offers

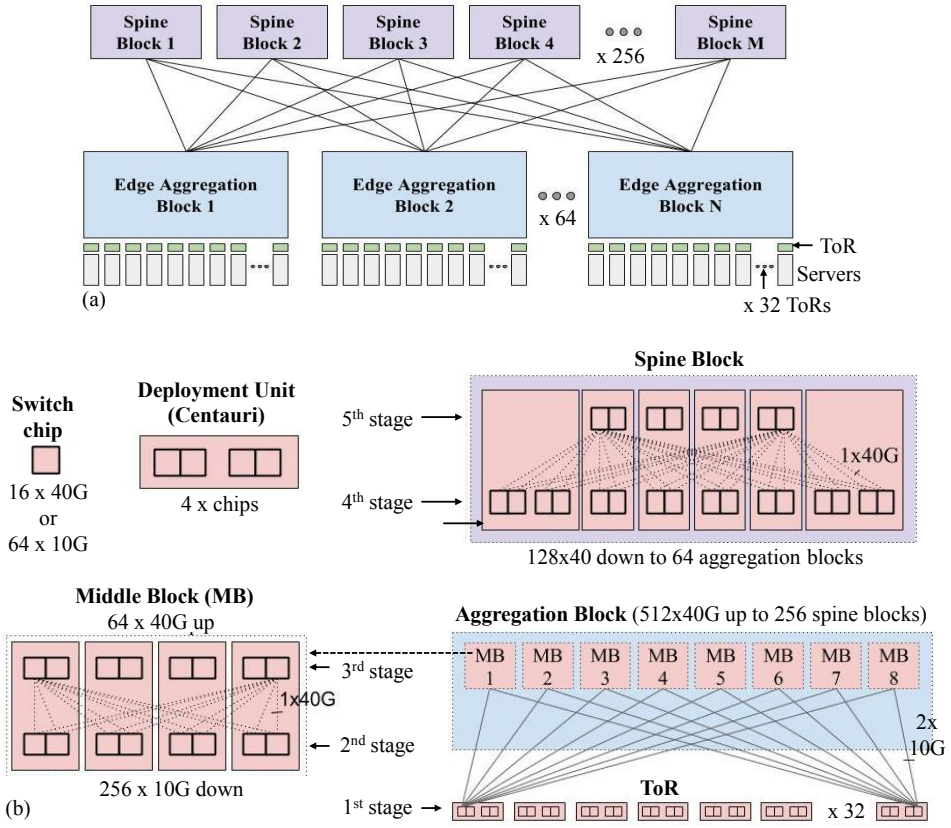


Figure 1.7: Schematic of the latest reported Google "Jupiter" data center network. From [12].

high scalability while using lower cost commodity switches. Nevertheless, large scalabiliy in such architecture is achieved by increasing the number of switching stages, and accordingly the number of switches and their number of ports and capacity. In order to get a better idea of a large Folded Clos network we show in Fig. 1.7 the latest reported Google data center configuration [12].

As depicted in Fig. 1.7(a), Google switching fabric is composed by N=64 aggregation blocks and M=256 spine blocks, which architecture is shown below in Fig. 1.7(b). All the switching fabric is based on a unique 16-port 40-Gb/s merchant silicon switch chip, four of which are used to built their deployment unit, the Centauri chassis, depicted in the top-left part of Fig. 1.7(b). Each port of the switch can be configured as 4×10-Gb/s,

which allows for higher-radix switches handling up to 64-port at 10 Gb/s or any mixed data rate configuration; hence giving the flexibility to build a fabric working under two data rates with the same chip. This way, the top stages of the fabric (spine blocks) use fully 40-Gb/s links, while the lowest one (ToRs) works at lower data rates (10 Gb/s), see right-most side of Fig. 1.7(b). The aggregation layer inter-connects the latter two, using 10-Gb/s downlinks and 40-Gb/s uplinks. Each aggregation block is composed by 8 middle blocks (MB), including each 16 switch chips connected by 40-Gb/s links in a two-stage Clos topology (shown in left-bottom side of Fig. 1.7(b). The spine blocks are also built in a two-stage (Folded) Clos topology, composed by 24 switches. Summing up all stages of the network we can account for 5 switching levels (ToR being the first switching stage), see Fig. 1.7 to better visualize the different stages. For each switching stage we describe in Table 1.2 the number of downlinks, uplinks (and their data

Table 1.2: Number of downlinks, uplinks (and their data rate) and switching (SW) chips used in the latest reported Google data center network. Data extracted from [12].

Blocks	Stages	Element	Per Chip	Per ToR* or MB**	Per Aggr. *or Spine**blocks	Total in DC
Servers	Servers	Uplinks	48x10G	192x10G*	6,144x10G*	393,216x10G
ToRs	1st Stage	Downlinks	48x10G	192x10G*	6,144x10G*	393,216x10G
		SW chips	1	4*	128*	8,192
		Uplinks	16x10G	64x10G*	2,048x10G*	131,072x10G
Aggre-gation blocks	2nd Stage	Downlinks	32x10G	256x10G**	2,048x10G*	131,072x10G
		SW chips	1	8**	64*	4,096
		Uplinks	8x40G	64x40G**	512x40G*	32,768x40G
	3rd Stage	Downlinks	8x40G	64x40G**	512x40G*	32,768x40G
		SW chips	1	8**	64*	4,096
		Uplinks	8x40G	64x40G**	512x40G*	32,768x40G
Spine blocks	4th Stage	Downlinks	8x40G	-	128x40G**	32,768x40G
		SW chips	1	-	16**	4,096
		Uplinks	8x40G	-	128x40G**	32,768x40G
	5th Stage	Downlinks	16x40G	-	128x40G**	32,768x40G
		SW chips	1	-	8	2,048

rate) and merchant silicon chips (switches) used in the so-called Jupiter data center network. The different columns show the quantity of elements included in each sub-block (e.g., uplinks, downlinks and chips per ToR, middle, aggregation and spine blocks) and in total in the whole data center.

In [12], Singh et al. specifically indicate that the ToRs' downlinks are configured in a 10-Gb/s mode (48 ports). Nevertheless, they also indicate in the paper that both 10 and 40-Gb/s servers are used in the network (without specifying the ratio). For simplicity and consistency with the specified 10-Gb/s downlinks placed in ToRs, we based our calculations on 10-Gb/s servers. Under such assumption, each ToR, including four switches, can host up to 192 servers using 10-Gb/s links. At the same time, each ToR is connected to all middle blocks through 64×10-Gb/s links (16 per chip). Accounting for all (64) aggregation blocks, each hosting 32 ToRs, the first switching stage includes more than 300,000 downlinks, over 100,000 uplinks and 8,000+ switches, all working at 10-Gb/s. Not surprisingly, the following second stage use the equivalent amount of 10-Gb/s downlinks (131,072) than uplinks in the first stage. Nonetheless, when climbing to upper stages the number of links can be divided by four thanks to the use of higher data-rate interfaces (40 Gb/s links). Hence all following upper stages are inter-connected through 32,768 40-Gb/s links data center-wide. The number of switches in the 2^{nd}, 3^{rd} and 4^{th} stages are 4,096, while the 5^{th} requires only the half amount of switches (2,048), because they fold into the fourth stage, which makes all switch ports available to be used as downlinks.

The enormous Google data center described above can handle up to 1.3 Pb/s of bisection bandwidth. Nevertheless, in [12] Google announced that their traffic demand is almost doubling every year. In order to keep increasing bandwidth at such a fast pace, large data centers and component vendors will need to overcome several barriers. In the following section we describe the main challenges appearing in the short and long term.

1.5.2 Data center challenges

1. Data center scale and capacity: In the near term, such large data centers will continue scaling up by increasing the speeds used and the number of ports integrated into single switch chips. As mentioned in the pre-

vious sections, the next expected step will be upgrading the whole system to the couple 25-100 Gb/s speeds (instead of 10-40 Gb/s). Nevertheless, using the same architecture, this change only yields a 2.5 fold increase of the overall capacity, which is not enough for large providers. Google for instance increased by a factor 6 their capacity when moving to the network described above [12], only three years after their previous upgrade. In order to keep increasing overall capacity while constraining the number of switching stages, switch vendors are producing chips supporting larger port counts. For instance, Broadcom and Mellanox have already released their latest 32×100-Gb/s Ethernet switches [42, 43]. Notwithstanding, largest data centers will soon require higher data-rate interfaces, with per-lane speeds larger than 25-Gb/s. Thus, the standardization of 50 GE, including 50-Gb/s lane-speed, 200-GE and 400 GE is already under progress. Component vendors will have to work hard to increase data rates towards 50 and 100-Gb/s/lane, while maintaining low cost and power consumption, which is very challenging.

Nevertheless, switch lane-speed and port-count cannot be increased arbitrarily. The ball grid array (BGA)[11] density limits the input/output chip bandwidth and hence the development of high-radix high-speed switches [44, 45]. Consequently, if cloud providers keep using the current topology, scaling overall data center bandwidth will require a further increase of switching levels, switches and interfaces; which leads to large network complexity and cost, huge power consumption and increased latency.

2. Large network complexity and cost: One of the main drawbacks of current Folded Clos architectures is the high complexity of the network. In Table 1.3 we present the total number of interfaces, switches and cables required in the Google data center described above. As depicted, to interconnect all their services Google uses tens of thousands of switches and millions of pluggables (interfaces) and cables. For simplicity we count all cables in the data center as fiber links[12]. Please notice that 40-Gb/s LR pluggables are connected through a pair of SMF fibers, while 40-Gb/s

[11] Packaging for integrated cicuits in which the electrical connections to the circuit are made through an array of tiny solder balls placed along the its surface.

[12] [12] indicates the use of fibers to interconnect the switching fabric above ToRs. Despite ToR-server connectivity is not specified, 10-G links are implemented with single fiber or copper cable-pairs, hence it does not affect the cable count.

Table 1.3: Total number of pluggable modules, switches and fibers placed in the latest reported Google data center supporting close to 400,000 10-Gb/s servers (estimated) [12], and their respective power consumption. Data extracted from Tables 1.1 and 1.2.

Device	Total count	Power per device	Total power
10G SFP+	1,048,576	1 W	1 MW
40G QSFP+	196,608	1.5 (SR) - 3.5 (LR) W	0.3 (SR) -0.7 (LR) MW
Switch chips	22,528	94 W	2.1 MW
Fibers (if 40G LR-SMF)	1,245,184	0 W	0 MW
Fibers (if 40G SR-MMF)	1,835,008	0 W	0 MW
		Total amount of power	**3.5-3.8 MW**

SR pluggables use four MMF pairs, bundled in ribbons. As the pluggable technology is not indicated in [12], we calculate the extreme cases, 100% SR (MMF) or 100% LR (SMF), to denote lower and upper bounds of cable count. The large number of devices and elements used in such large-scale switching fabric, requires enormous investments. For instance, by the end of 2011 Facebook owned more than $1 billion in network equipment [46]. Despite the cost of the equipment, one has to take also into account the enormous cost of development and operation of such complex infrastructure built upon millions of networking elements [47].

3. Vast power consumption: In 2014 U.S. data centers consumed as much as 2% of the whole country power consumption, which is equivalent to the consumption of 6.4 million average homes [48]. The interconnection network accounts for 10-20% of such consumption [49]. In order to give the reader a better idea of such consumtion, we show in Table 1.3, the calculated power consumption of the Google intra data center network, strictly accounting for data transport and switching contributions. We calculated again the lower and upper consumption bounds related to the use of SR or LR pluggable modules. The power consumption of the pluggable modules were extracted from Table 1.1, while switches consumption was taken from the specifications of the latest 16-port 100-Gb/s Mellanox switch [50], which consumption is similar to older 40-Gb/s switches [51, 52]. We can observe that both switches and pluggable modules strongly contribute to the energy consumption (pluggables: 1.3-1.7 MW, switches: 2.1 MW). In total the estimated power consumption only accounting for switching and

transporting data ranges between 3.5 and 4 MW. Nevertheless, there are other consuming sources related to the networking fabric, such as powering up the racks holding the switches and cooling and lighting systems. Some sources estimated in 2012 that major Google data centers consumed between 50 and 100 MW [53]. If we assume that 15% of such consumption is related to the networking fabric [49], we obtain 7.5-15 MW, which results into an astonishing yearly cost of $3.9-7.9 millions[13].

4. High latency: Having multiple levels of switching may drastically increase end-to-end latency. Latest 100-GE switches specify port-to-port latency of 300-400 ns [42, 43]. Accordingly, the overall latency reaches 2.7-3.6 µs when summing up the the total number of switches that an Ethernet packet has to traverse to perform any server-to-server connection in the data center (9 switches when traversing all stages) . Nevertheless, latency can further increase when taking into account packet dropping ocurring in the network. Packet dropping can occur for many reasons; however a large number of them are related to the switching fabric malfunctioning: fiber frame check sequence errors, switching ASIC defects, switch fabric flaw, switch software bug or the problem of network congestion [55]. In such large networks containing thousands of switches malfunctioning occurs frequently. In some cases switches send alarms to the management system to alert from failures; however sometimes it remain silent. Therefore, large data centers need to implement complex algorithms (e.g., *Pingmesh* [55]) to detect and recover from such failures. Some examples of silent packet dropping are the *packet black-hole*, which drops deterministically packets containing a certain "pattern", and the *silent random packet drop*, which more suddenly starts discarding packets in a more random way. *Packet black-hole* can be typically repaired by reloading the switch, while with *silent random packet drop* the switch needs to be replaced [55]. These switch failures can severely impact end-to-end latency. Therefore, further increasing the number of switching devices and stages will eventually lead to unsupportable latencies that might limit the deployment of expected real-time cloud applications, which present strong latency requirements.

In the near term, data centers will keep augmenting their capacity by upgrading switch and interface capacities and building larger multi-tier net-

[13]Calculation based on the industrial electricity cost in North Carolina in May 2016: 6.03 cents/kWh [54]. Cost = $kW \times 365(days/year) \times 24(hours/day) \times 6.03(cents/kWh)$

works. Nevertheless, in the end, increasing the data center bandwidth while maintaining current architecture may drastically worsen the above mentioned issues, which might critically deteriorate data center performance. Therefore, in a longer term, a reassessment of data center architecture is inevitable.

1.6 Thesis outline

In this first chapter we have described the data center evolution from the early days of modern computing up to modern large-scale cloud data centers. After a brief history of computing we have analyzed the traffic evolution since the early 90s. We have shown that internet and cloud services have induced a relentless bandwidth demand, which is still almost doubling every year in largest service provider facilities. Looking for a scalable and failure-robust solution, data centers have adopted a Folded Clos topology, including a large number of commodity electronic switches organized in a multi-tier architecture to provide full server-to-server connectivity. Subsequently, to keep up with the bandwidth demands, data center are fastly upgrading the capacity of their switching fabric: starting from 1 Gb/s, they moved to 10 Gb/s and 40 Gb/s switch ports and interfaces, such speeds being the most frequently used in today's data center networks. Nevertheless, since the appearance of 100-GE (most typically based on 4×25-Gb/s lanes), the 25-100 Gb/s couple is expected to replace the latter two.

In the previous section we have identified four challenges that need to be solved to keep increasing the capacity: 1) In the short term, in order to keep increasing data center capacity without drastically enlarging data center infrastructure, main service providers will require the development of novel switches, supporting larger number of ports and capacities, and accordingly, pluggable modules with lane data rates overcoming at each generation 25, 50 and 100 Gb/s barriers. Nevertheless, as described in the previous section, switch lane-speed and port-count cannot increase arbitrarily. Although the Folded Clos all-electronic switching topology offers theoretically arbitrary scalability, when scaling up towards the unprecedented capacities expected in the future, such architecture triggers other challenges that will need to be resolved in a longer term: 2) extremely high network complexity, comprising hundreds of thousands or even millions of network components; which induce high development, maintenance and op-

eration costs; 3) vast energy consumption, produced by the large amount of switches and interfaces required to perform server interconnection, which yearly costs millions of dollars to large service providers; 4) high end-to-end latency, originated by the many crossings through electronic switches placed in the large tier-count network; which may limit novel real-time cloud applications.

Along this thesis we propose optical solutions, from the physical layer perspective, that aim at addressing the above mentioned challenges present in both short and long term. Being the physical layer of optical communications the backbone of this work, we provide in Chapter 2 a background on optical communications. In that chapter we first provide an ample overview of optical networks in a global scale to later focus on more technical aspects, such as existing communication/switching techniques and the implementation of different kinds of transceivers that will be used in the following chapters.

In Chapter 3 we address the urgent need for high data-rate optical interfaces capable of achieving the specifications of the forthcoming Ethernet standards (Challenge 1). In order to cope with such demands, we propose and demonstrate several optical transceivers offering the expected 100-Gb/s single-lane capacity and beyond (reaching 200-Gb/s data rates). In order to achieve such high speeds we leverage advanced modulation formats (i.e., PAM-4 and PAM-8), while keeping the lowest possible cost using intensity-modulation and direct-detection transceivers. The required multi-level signals at such high-symbol rates (i.e., 56, 84 and 100 GBd) are generated by an integrated selector power digital-to-analog converter (DAC), which is able to double input symbol-rates, allowing to overcome the speed barrier of current electronics.

In order to overcome the latter introduced challenges (2: network complexity, 3: energy consumption and 4: latency), our team proposed in 2014 a network for data center intra-connection called Burst Optical Slot Switching (BOSS), which is extensively described in Chapter 4. Such proposal replaces the electronic switching fabric by a mesh of optical fiber rings, which inter-connect BOSS nodes organized in a torus topology, such that any-to-any BOSS node communication can take place with a single opto-electronic-optic conversion. The combination of high-speed transceivers (beyond 150 Gb/s), statistical slot multiplexing and a flattened inner-ring-transparent topology, offers a $\times$100-500 reduction in number of interfaces

and cables, hence diminishing network complexity and the development and operation cost (addressing challenge 2); a factor 2-3 of energy savings with respect to current all-electronic networks (challenge 3), and have the potential to diminish latency (challenge 4).

In Chapter 4 we explore several approaches for the physical implementation of the BOSS nodes, while investigating the technology-dependent impairments occurring when traversing a large number of nodes, which drives data center scalability. We asses the cascadability of such devices and evaluate the performance in terms of maximum reach and average overall capacity, while comparing several modulation schemes. Ultimately we propose coherent-optical orthogonal frequency-division multiplexing (CO-OFDM) as modulation approach, which spectral tailoring capabilities allows adapting to the tight filtering produced when cascading a large number of low-cost nodes. When comparing to the standard Nyquist pulse-shaped N-quadrature amplitude modulation (QAM) approach, our CO-OFDM transceiver provides 30% higher capacity (between 150-250 Gb/s per transceiver depending on the number of nodes) and 40% extended reach (more than 100 nodes).

Finally in Chapter 5 we conclude this thesis summarizing the presented work and discussing the main results while giving a hint of future perspective.

Chapter 2

Optical communication systems: a review

2.1 Introduction

Since their first trial demonstration in 1977, fiber optic systems have not stopped evolving, delivering every day higher capacities. First commercially available systems made use of intensity-modulation and direct-detection (IM-DD) schemes to transmit up to 45 Mb/s through 10-km (repeater-less) of MMF [56]. Just ten years later optical systems allowed transmitting up to 2.5 Gb/s capacities over 60-70 km [56]. This was enabled by the utilization of new lasers and detectors working at 1.55 µm (corresponding to the band with lowest fiber loss, typically 0.2 dB/km) and novel SMFs providing extended reach, together with the improvements in high speed electronics, which allowed a $\times 50$ increase in data rate [56]. After the invention of the erbium-doped fiber amplifier (EDFA), the concept of wavelength-division multiplexing (WDM) became very popular [57]. WDM permitted transmitting multiple wavelength channels within a single fiber, which allowed increasing fiber capacity above the Tb/s by the end of the 90s [57]. The latest optical revolution appeared with coherent technology in the 2000s. Different from IM-DD schemes, which only use the intensity of light to send data, coherent systems make use of the amplitude, phase and polarization of the light, hence increasing spectral efficiency (b/s/Hz). Today coherent transceivers are capable of transmitting 100 to 500 Gb/s per wavelength

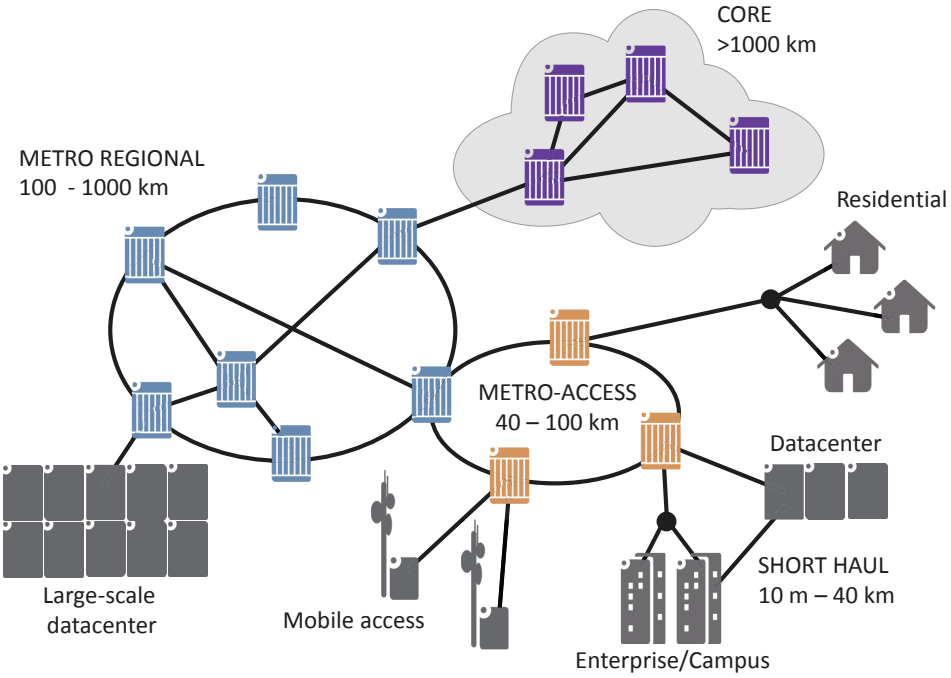


Figure 2.1: Network segments in a global telecommunication network.

channel over thousands/hundreds of kilometers [58].

Due to their high capacity and long reach, fiber optical systems have been extensively used to support worldwide core networks (also called backbone or long-haul networks). Nevertheless, led by the increasing Internet traffic demand and helped by the decreasing cost of optical systems, most telecommunications segments are adopting today fiber optics communication. Fig. 2.1 shows a diagram of a global telecommunications network, describing the diversity of network segments. Evidently, each segment adopts the technological solution that better fits its singular bandwidth/reach requirements and techno-economical constraints.

Core networks require huge and long data pipes capable of transporting world-wide all data traffic aggregated from the underlying segments. Being the network with fewest nodes, but requiring the highest capacity, core networks rely on high-end coherent transceivers to achieve large spectral efficiency and hence maximize fiber capacity. Such fibers carry typically tens of terabits per second through many wavelength channels each working at

100-200 Gb/s. The opposite occurs at the lowest segments of the network (i.e., access, mobile, data center, etc.), which do not require long reach (less than 40 km overall), but that are very sensitive to cost due to their proximity to the end-user. Hence such segments typically use lower cost IM-DD-based transceivers. The different segments in such layer present a large technological diversity due to their heterogeneous needs: Mb/s-Gb/s passive optical network (PON)s providing connectivity to residential areas, 1-10 Gb/s links for mobile backhauling or Enterprise/Campus data centers, and 10-100 Gb/s for large-scale data centers[1]. The intermediate metropolitan segment aggregates traffic from this heterogeneous ecosystem, thus traffic flows traveling through metropolitan networks is highly diverse, which makes their design challenging. Furthermore the traffic in such segments are rapidly increasing due to the creation of bandwidth-hungry cloud services. Consequently, most transceiver vendors are currently working on lowering the cost of coherent technology to offer a high-bandwidth solution for these networks, with capacities ranging between 100 and 400 Gb/s.

Along the following sections we provide the fundamental concepts on optical communication systems that will help the reader better understand the following chapters. First, in Section 2.2, we describe the different communication/switching techniques that can be used to exchange data between two nodes or more in an optical network. Later, we explain in Section 2.3 the different approaches and technologies that are used to implement optical transceivers, i.e., IM-DD and coherent technology, which will be extensively used in Chapter 3 and 4, respectively.

2.2 Communication/switching techniques

Different communication/switching techniques are being used today in the different network segments, depending on the kind of traffic flowing through the specific network. Along this subsection we will discuss three different switching techniques: optical circuit switching (OCS), electronic packet switching (EPS) and optical packet switching (OPS).

[1] 10-100 Gb/s IM-DD transceivers are used for intra data center networks, but external links might use coherent technology working at higher speeds.

2.2.1 Optical circuit switching

Optical circuit switching is the standard communication technique used nowadays in core networks. In such approach a fixed optical connection (circuit) is established between source and destination nodes; typically by assigning an end-to-end path through a certain wavelength, see Fig. 2.2, where color curves denote circuits established between two nodes. This way data can traverse many nodes transparently (i.e., without the need of opto-electro-optic (O/E/O) conversions), while being optically routed in reconfigurable optical add-drop multiplexer (ROADM) placed in the nodes. Each ROADM typically has as many wavelength-selective switches (WSSs) as input/output fibers, which allows redirecting any wavelength from any input fiber to any output fiber (or to the receivers). WSSs have slow reconfiguration time (millisecond to seconds), which is adequate for OCS networks, meant to be static or reconfigurable in the long term.

This approach is efficient if traffic between two end-to-end nodes is constant and the communication requires a full-wavelength capacity. This is usually the case of core networks, which aggregate a vast amount of traffic coming from the underlying networks, leading to averaged constant capacity needs. OCS is efficient in terms of power consumption and latency because light travels transparently from source to destination. Nevertheless, if this approach is used in environments with highly varying traffic, capacity is wasted.

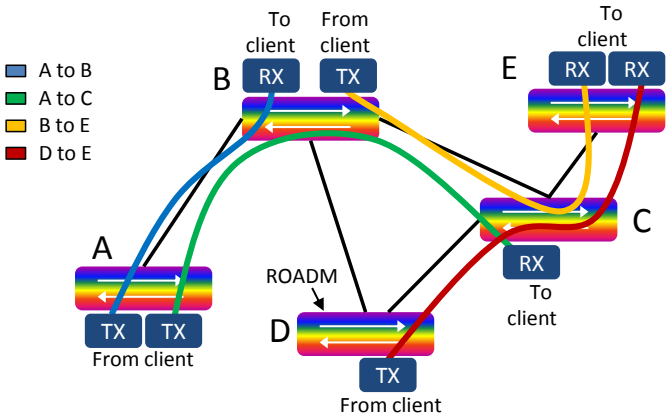


Figure 2.2: Optical circuit switching network.

2.2.2 Electronic packet switching

In other segments, such as data center networks, traffic between end-to-end nodes (servers in this case) is bursty and fastly varying, presenting connections requiring typically less than the channel/wavelength capacity. Hence using OCS results in large capacity wastes. Furthermore, the full interconnection of hundreds of thousands (N) of servers would require the establishment of N^2 circuit connections and hence the use of N^2 transceivers, which is impossible to implement given that the wavelength count extends to 80-96 when using the C-band. In such dynamic traffic environments, EPS (e.g., Ethernet) is used to flexibly adapt to traffic variations while optimizing equipment utilization. As shown in Fig. 2.3, in which each color indicates packets with same source and destination nodes, in EPS data going to different destinations (so belonging to different connections) can be sent through the same link multiplexed in short electrical packets (typically in the nano- or micro-second scale). Please note that using this approach the notion of transparency disappears. Depicted in Fig. 2.3, the electrically generated packets, are transmitted through "circuit-like" point-to-point optical links, which require transceivers at each fiber link termination. Hence, when traversing an intermediate node, the optical signal is first converted to the electrical domain (through a receiver), then each electrical packet is routed towards an output port in the electrical domain (electrical switch), and finally the electrical signal is reconverted to the optical domain (transmitter).

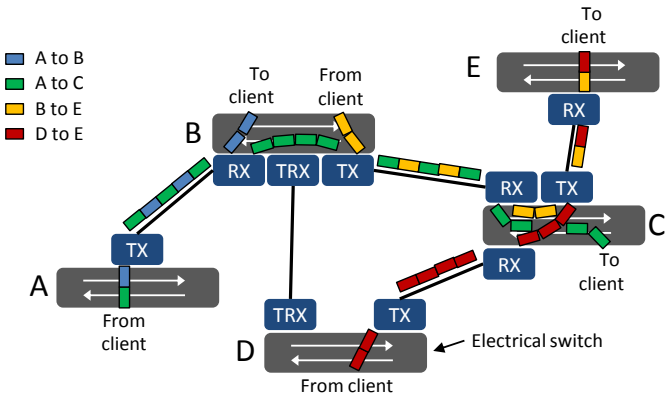


Figure 2.3: Electronic packet switching network.

A clear advantage of EPS is the capability of performing statistical multiplexing which enhances resource utilization (a single transceiver can be used to serve several connections). Nonetheless, EPS requires electronic switching (and accordingly O/E/O conversions) at each intermediate node, which not only increases the required energy consumption, but also the latency, due to the time spent at each node for data conversion (to electrical domain), processing, possibly queuing and re-conversion (to optical domain). Furthermore, upgrading the capacity in EPS networks typically requires changing most of the network components, including end-node (client) transceivers and intermediate node transceivers and electrical switches. On the other hand, in OCS networks, only end-node transceivers need upgrading, since data traverse transparently the ROADMs.

2.2.3 Optical packet switching

Aiming for simultaneous transparency (as in OCS) and statistical multiplexing (as in EPS), many research groups have proposed the use of OPS in highly dynamic networks, such as metropolitan and data center segments. Differently to EPS, where electrical packets are inserted in circuit-like point-to-point optical connections, optical links in OPS transport optical packet entities. As shown in Fig. 2.4, similarly to OCS optical packets are now routed in the optical domain. However, in OPS fast optical switches capable of re-directing the light in a packet time-scale (nanosecond switching) are required. Such optical switches can be for instance implemented in a broadcast and select architecture using semiconductor optical amplifiers (SOAs), electro-absorption modulators (EAMs) or variable optical attenuators (VOAs), between others, as optical gates.

As previously mentioned, OPS has the advantages of both latter approaches: statistical multiplexing with sub-wavelength granularity, which allows traffic adaptability and resource optimization, and transparency, which diminishes energy consumption, latency and facilitates capacity upgrade (only end-node transceivers need to be changed). Nevertheless, the main pain-point of OPS is the management of packet collisions. In a mesh-like topology, depicted in Fig. 2.4, each node has multiple input/output links. Hence, it is likely that two packets, arriving simultaneously from different inputs, require to be directed to the same output, see yellow and red packets in node C of Fig. 2.4. In this case, if packets coming from dif-

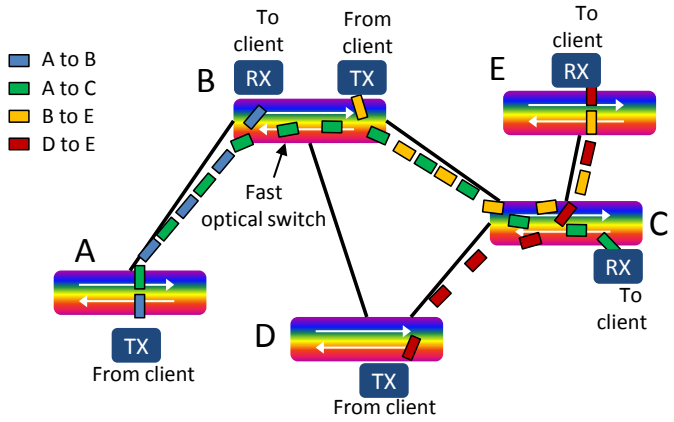


Figure 2.4: Optical packet switching networks when using a mesh-like topology.

ferent inputs are not synchronized or buffered, a packet collision may take place, which requires packet re-transmission hence leading to capacity waste and increased latency. In EPS systems, such events are managed through electronic buffering. Optical buffers can be implemented using fiber delay lines; however, they are very limited and tedious to manage. Another approach to avoid collisions would be the synchronization of all packets from their origin; nevertheless, in a mesh-like topology synchronization of a large number of nodes is impractical due to the large number of links with different lengths. Hence, most OPS mesh-based proposals typically run under relatively low network capacity loads (less than 50% [59, 60]) in order to diminish the probability of collisions. Nonetheless, in that case network resources need to be quite overdimensioned.

Such issue can be tackled by using a ring topology, described in Fig. 2.5. In a ring network, several nodes are connected through a single fiber ring and data is exchanged using multiplexed time slots. Each packet color depicted in Fig. 2.5 relates to a unique source and destination node. Dropping and adding packets out of/into the ring can be simply performed using optical couplers. In this case each receiver will select the packet under interest. In optical rings, each node has only one input/output fiber pair, hence no real packet switching occurs. However, as data traverse transparently through the nodes, an optical packet/slot blocker is required to erase the received packets, hence allowing for time-slot reuse, see blocked slots filled in white in Fig 2.5. The capacity in such network can be increased by using WDM as extra sharing dimension. This approach requires either

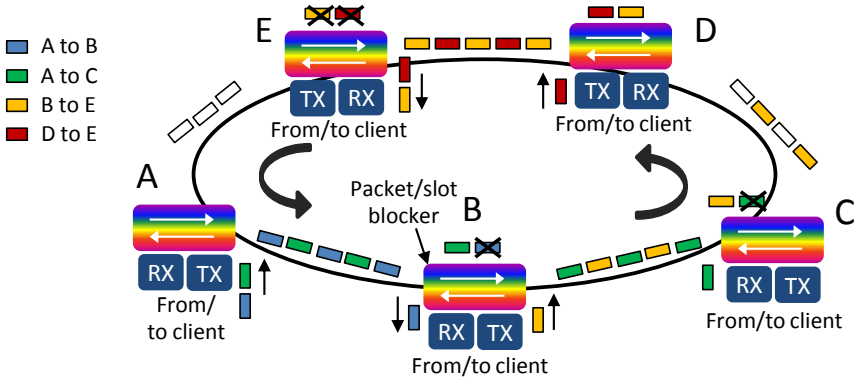


Figure 2.5: Optical packet switching networks when using a ring topology.

fast tunable transmitter or receivers (or both) to address the wavelength under interest.

In ring networks, packet synchronization can be easily achieved, since all packets travel through the same media (a single fiber ring). Therefore, collisions can be avoided and network can support high capacity loads close to 100%, which further improves resource utilization. The implementation of such networks will be further discussed in Chapter 4.

2.3 Intensity-Modulation Direct-detection (IM-DD) systems

As earlier described in Section 2.1, optical systems rely on different transmission approaches depending on the requirement of the specific network segments. Networks placed in the short-haul segments (residential and mobile access and data centers) typically profit from the low-cost of IM-DD transceivers, which transmit information using the light intensity. Along this section we describe first the different modulation approaches that we will use along the thesis and later the possible implementations of IM-DD transceivers.

2.3.1 IM-DD modulation techniques

Up until recently, IM-DD transceivers have relied on the most simple modulation format existing in optical communications: PAM-2. Such format encodes each bit $(0, 1)$ into a different intensity level (S_1, S_2), see constellation in Fig. 2.6(a). In IM-DD systems, the signals are typically represented by the eye-diagram, which superposes many samples of a given waveform on a single time-period (typically the duration of one or two symbols). The ideal PAM-2 eye-diagram is plotted in Fig. 2.6(a), depicting the two modulation intensity levels and the transitions between them. Below we also plot a more realistic PAM-2 eye-diagram, obtained simulating a PAM-2 signal while applying transmitter and receiver filters (Butterworth) with bandwidths (BW) equal to the baudrate (BR) of the signal (no noise was applied). Such modulation is preferred in low cost modules due to its simplicity of implementation, which allows lowering the cost. Hence PAM-2 is currently used in all pluggable modules implementing 1-, 10- and 25-GE. Nevertheless, when increasing the target data-rate of a single-lane (e.g., 50 or 100-Gb/s or beyond), such modulation format requires drastically increasing the bandwidth of optical and electrical components use to build the transceivers.

In order to keep increasing single lane data-rates, the Ethernet alliance proposes the use of more advanced modulation formats such as PAM-4 [27, 28]. As depicted in Fig. 2.6(b) this multi-level format makes use of four intensity-level symbols, each transmitting two bits of information, which allows doubling the capacity while using the same bandwidth, hence increasing the spectral efficiency. The ideal and simulated eye-diagram are also shown in Fig. 2.6(b), using the same configuration as for PAM-2. It can be easily depicted that the distance between symbols decreases by a 3-factor when comparing two PAM-2. The modulation order can be increased to an arbitrary number of levels (i.e., 8, 16, 32). In Fig. 2.6(c) we show the constellation and eye-diagrams of an 8-level pulse amplitude modulation (PAM-8) signal, which can transmit now up to three bits/symbol while diminishing the symbol distance by a 7-factor with respect to PAM-2.

Clearly, increasing the modulation order allows for higher spectral efficiency at the expense of limited inter-symbol interference (ISI). Hence, the higher the modulation order is, the more sensitive the modulation scheme is to noise or to transceiver bandwidth limitations. For instance, in the

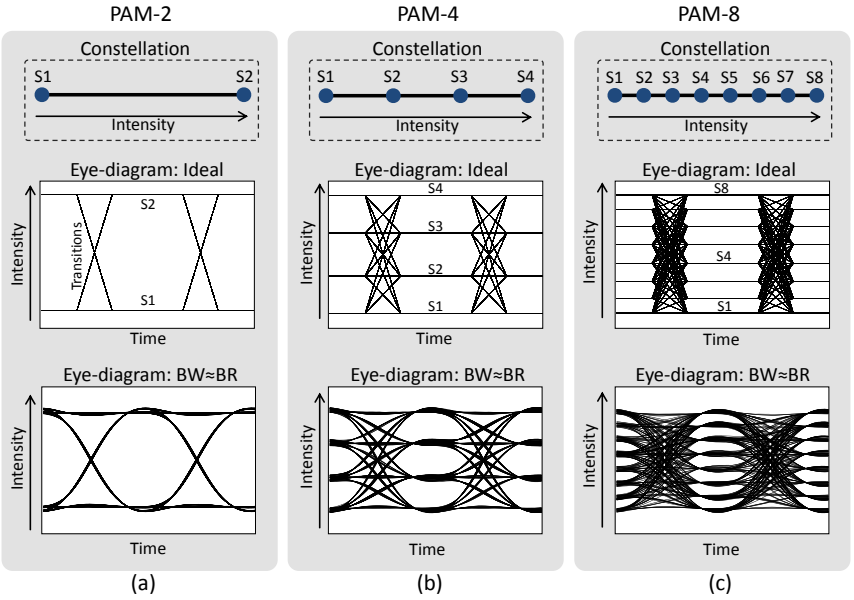


Figure 2.6: Generic block diagram of an intensity-modulation transmitter.

bottom noise-free eye-diagrams we can notice a small amount of ISI, arising from the limited bandwidth of the transceiver. We can clearly observe, especially for PAM-8 that ISI closes the eye (levels get thicker reducing the distance between them), which limits the performance. Please note that in this case we use very large end-to-end bandwidths. For instance, the represented eye-diagrams could illustrate 56-GBd signals with end-to-end bandwidths above 50-GHz. Nevertheless, current bandwidths of commercial opto-electronic devices remain between 30 and 40 GHz (to be described in the following subsections). Hence the eye-diagrams obtained with commercial available technology when implementing high baud-rate transceivers are even more closed. Nonetheless, performance can be improved by using equalization to compensate for bandwidth limitations.

Other modulation approaches are also being studied for their implementation in IM-DD. For instance multi-carrier and multi-band approaches such as discrete multitone (DMT) and carrierless amplitude/phase modulation (CAP) can be also beneficial in bandwidth stringed environments. Nevertheless, they require a more complex digital signal processing (DSP) and higher-resolution DACs and analog-to-digital converter (ADC)s, which makes them less popular for next generation low-cost pluggable modules.

2.3.2 IM-DD transmitter

A block diagram of a generic IM-DD transmitter is shown in Fig. 2.7. As depicted in the figure, an IM-DD transmitter can have two different implementations: directly modulated laser (DML) and externally modulated laser (EML). DML transmitters perform light intensity modulation by driving directly the gain section of the laser itself. This approach simplifies the transmitter module, which does not require an external modulator. Nevertheless, when modulating the gain section of the laser, the carrier density in that area changes and thus does also the refractive index of the material, which produce light phase variations. Such process leads to the so-called chirp effect, which induces spectral broadening and hence, reduces the reach due to chromatic dispersion.

On the other hand, EML transmitters, use an external modulator, which can provide chirp-free operation (depending on the type of modulator). Furthermore, external modulators typically offer higher bandwidth than directly modulated lasers. In both cases, the electrical signal is provided by a DAC, mainly when using multi-level modulation formats, typically followed by a driver amplifier, to provide enough modulation swing. The DSP in the transmitter side is almost negligible. Regular (non-pulse-shaped) pulse-amplitude modulation schemes, use only DSP to encode the data for latter forward er
symbols. Nevertl

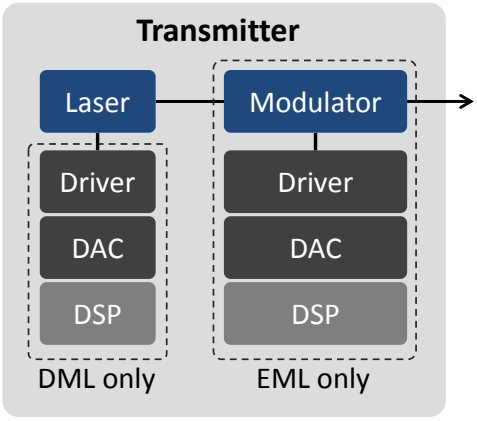


Figure 2.7: Generic block diagram of an IM-DD transmitter.

or Nyquist pulse-shaped pulse amplitude modulation require other further transmitter DSP blocks. These formats are out of the scope of this thesis, but the reader can refer to [61–64] for more information. Along the following paragraphs we briefly describe the different components that can be used in an optical transmitter.

Laser: As mentioned in the previous chapter we find mainly two kinds of semiconductor lasers in data center pluggable modules: VCSELs and DFB lasers. The first one, VCSEL, is the most commonly used nowadays due to its low cost and consumption. As shown in Fig. 2.8(a), the cavity of this laser is built vertically, placing the active layer between two Bragg reflectors (top and bottom). Hence light outputs the laser from the top surface of the chip, which makes on-wafer laser testing possible, hence reducing production cost, and easy coupling to MMF due to good mode matching and small nearly circular spots size, which reduces also packaging cost. These lasers are used to build the short reach (SR) pluggable modules described in Chapter 1, which are connected through MMFs allowing few-hundred meters transmission.

On the other hand, DFBs are edge-emitting lasers, which are built longitudinally along the wafer. The cavity consists of a distributed periodical grating, which acts as a reflector placed along the active layer, producing this way single-longitudinal mode operation, see Fig. 2.8(b). Although DFB lasers exhibit twice the cost of VCSELs [65], multiple DFBs can be densily integrated together with modulators and wavelength multiplexers. This way several wavelength channels can be generated by a single photonic chip, which can now be coupled to SMF, hence extending the reach to le

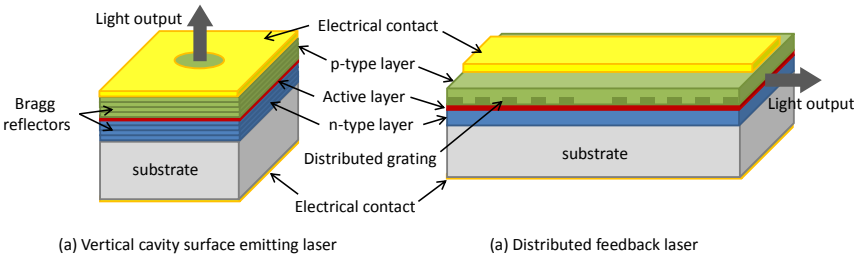


Figure 2.8: Exemplary illustrations of (a) a VCSEL and (b) a DFB laser.

modules, described in Chapter 1. Latest research works have reported modulation bandwidths of approximately 30 GHz, for both VCSELs [66] and DFBs [67], which allows for 56-64 Gb/s modulation [68].

Modulator: Larger bandwidths and extinction ratios can be typically achieved when using external modulators. We find two main kinds of modulators in today's optical commercial transmitters: EAM and electro-optic Mach-Zehnder modulator (MZM). In EAM the intensity of the input light is modulated by changing the absorption coefficient of a III-V semiconductor material (e.g., InPGaAsP structure for 1550 nm) by applying an external voltage/current, see transmittance-voltage curve in Fig. 2.9(a,bottom). Hence the structure is similar to the one of a semiconductor laser, however, the active layer is used to absorb instead of providing gain, as shown in Fig. 2.9(a,top). Such devices are very attractive for low-cost applications because they present small footprint, low driving voltages (<2 V), and they can be monolithically integrated with DFB lasers, providing tiny EML transmitters. Furthermore EAM can achieve modulation bandwidths beyond 50 GHz [69, 70]. However, they typically present high insertion losses of the order of 10-15 dB, which limits the output transmitter power to a few dBm, and induce chirp (as described above for lasers), both limiting the transmission reach[2]. In Chapter 3 we demonstrate a compact EML transmitter including a DFB laser and an EAM modulator capable of performing 112 Gb/s single-carrier transmission over 2 km.

On the other hand, as shown in Fig. 2.9(b), in MZM light is modulated through the interference created on a two-arm Mach-Zehnder interferometer. In such structure, the relative phase between the two arms is modulated by changing the refractive index of the waveguides through an electro-optic effect. This allows inducing any arbitrary interference which enables both amplitude (or intensity) and phase chirp-free modulation, see transmittance-voltage curve in Fig. 2.9(b). Hence, such structures can be used in both IM-DD and coherent transmitters (to be described in the following section). Traditionally, commercial MZM have been mostly fabricated in Lithium Niobate ($LiNbO_3$), offering more than 30-GHz of 3-dB

[2]Sometimes EML include an integrated SOA to increase the output power, and hence the reach in non-amplified links.

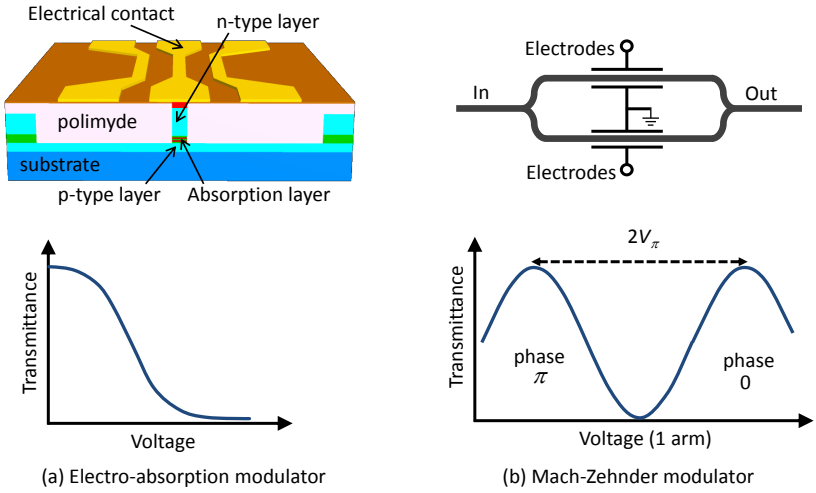


Figure 2.9: (a) Exemplary illustration of an EAM (From [79]). (b) Schematic of a MZM.

bandwidth[3] with driving voltages (V_π)[4] between 3 and 4 V. Nevertheless, novel photonic integration platforms have appeared, enabling the manufacturing of smaller and more cost-effective devices. The second integration platform to mature was Indium Phosphide (InP), which provides MZM with half the size of $LiNbO_3$ and reduced V_π (1.5-2.5 V), while allowing modulation bandwidths of the order of 40 GHz [72, 73]. More recently appeared the Silicon platform, which is promising for it potential for massive scale production and integration of photonics and electronics on the same substrate. Silicon-based MZM modulators have been reported today presenting bandwidths between 30 and 40 GHz [74–76]. Finally, the latest integration platform to appear is based on polymer materials, which have the potential for extremely large modulation bandwidths (up to 100 GHz) [77, 78].

Electrical generation: For a few years, high resolution DACs have been used in coherent technologies to generate all kinds of advanced modulation formats and pulse-shapes. Nevertheless, DACs will be also required now in IM-DD transmitters to be able to generate multi-level pulse am-

[3]Such modulators usually have soft frequency response decay, providing 6-dB bandwidth beyond 40 or 50 GHz [71].

[4]Voltage required to switch between minimum and maximum transmittance, see Fig. 2.9(b).

plitude modulation (PAM) formats and maybe in the future even more complex modulations schemes such as DMT. However, commercial DACs available in the market are one of the most limiting factors when targeting high-speed symbol-rates. Latest DAC generations can achieve quite high sampling rates; for instance Socionext/Fujitsu developed an 8-bit DAC capable of sampling at 92 GS/s (in CMOS), while Micram offers a 72-GS/s 6-bit resolution DAC built in SiGe. However their bandwidth is limited to 20-25 GHz [80–82]. Furthermore, their output swing is smaller than 1 V_{pp} (peak-to-peak voltage); hence, typically amplifier drivers need to be used to obtain enough swing to drive the modulators. On the other hand, InP DHBT technology can be used to obtain bandwidths above 50 GHz while outputting electrical signals above 4.5 V_{pp} [83], hence avoiding the use of external driver amplifiers. In Chapter 3 we use a 3-bit InP-DHBT selector power DAC fabricated in III-V Lab capable of achieving symbol rates as high as 100 GBd with sufficient output swing to directly drive a $LiNbO_3$-MZM modulator.

2.3.3 IM-DD receiver

The block diagram of the receiver is shown in Fig. 2.10(a). As depicted, in direct detection schemes receivers are implemented using a photodiode, which converts optical intensity into electrical signals, a TIA to electrically amplify them and an ADC to convert signals from analog to digital domain. Then a certain amount of DSP might be used to recover the signal.

Receiver front end: The most common photodiodes are PIN and avalanche photodiode (APD), where the latter includes an additional region in the conventional PIN semiconductor structure in which the number of conductible electrons is exponentially increased (avalanche region), hence providing gains one order of magnitude higher than conventional PIN photodiodes [84]. On the other hand, APDs exhibit larger cost and consumption, which together with PIN photodiodes' higher reliability and smaller footprint, makes the latter the most commonly used in data center pluggable modules. Subsequently, a TIA is required to amplify the detected signal. Today, very fast photodiodes can be found in the market; for instance Finisar sells a 100-GHz photodetector. Nevertheless, the bandwidth of commercially available PIN-TIA modules is limited to 40-50 GHz [85].

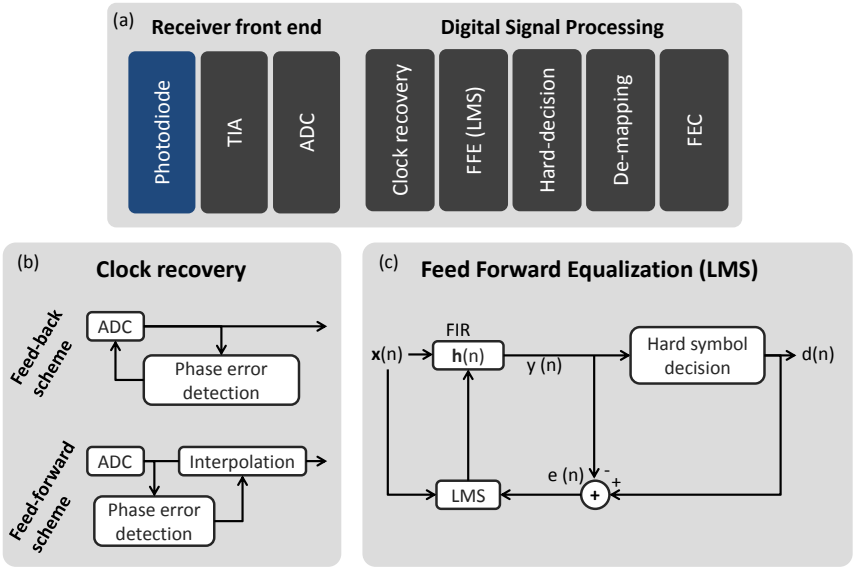


Figure 2.10: (a) Block diagram of a generic IM-DD receiver including a generic description of the DSP chain. (b) Illustration of two possible configurations to perform clock recovery. (c) Schematic of the feed-forward equalization and adaptation process.

Nonetheless, TIA providing bandwidths above 90 and 100 GHz have been reported recently in several research works [86, 87].

In what refers to ADCs, they evolve in commercial products similarly to DACs. 56, 92 and 100 GS/s ADCs have been reported, however they offer limited bandwidths that slightly surpass 20-GHz [88]. Typically, high-speed research demonstrations make use of large bandwidth oscilloscopes, capable today of sampling at 80 GS/s, 160 GS/s or even 200 GS/s, while exhibiting 3-dB bandwidths ranging between 30 and 100 GHz, achieved typically using digital bandwidth interleaving (see product catalogs of Lecroy, Keysight or Tecktronix for further information).

Digital signal processing: Although actual IM-DD receivers include few DSP functionalities, upcoming high-speed interfaces supporting multi-level modulation formats might require certain amount of digital signal processing to perform successful signal recovery due to their lower robustness to noise and distortions. The basic DSP blocks of an advanced IM-DD

receiver are shown in Fig. 2.10(a). The first block of the DSP chain is the clock recovery, which recovers the difference in sampling clock (frequency) between transmitter and receiver as well as the right symbol timing (phase). As shown in Fig. 2.10(b), we can find two generic configurations for clock recovery: a feedback scheme, in which the ADCs clock is re-addressed through a phase locked loop (PLL), aided by the output of a frequency/phase detection algorithm; and the feed-forward scheme, in which the output of the phase detection algorithm drives a subsequent interpolator. Several phase detection techniques can be used depending on the transceiver requirements. IM-DD systems typically use low complexity and fast schemes such as Alexander [89], Mueller & Mueller [90] or Hogge [91].

After clock recovery, feed-forward equalization (FFE) might be used to compensate for linear impairments (e.g., transceiver bandwidth limitations), to be extensively studied in Chapter 3. Fig. 2.10(c) shows a typical FFE block diagram using the least mean square (LMS) algorithm. In LMS, the signal is passed by a N-taps finite impulse response (FIR) filter $\mathbf{h}$, which is adaptively updated. As part of the adaptive process, symbol decision takes place obtaining $d(n)$, which is being used to calculate the error $e(n)$ obtained from the equalization process: $e(n) = d(n) - y(n) = d(n) - \mathbf{h}(n)^T\mathbf{x}(n)$. Filter adaptation $\mathbf{h}(n)$ typically uses the steepest descent technique which minimizes the mean square error $E\{|e(n)|^2\}$. Filter tap updating is done as follows:

$$\mathbf{h}(n+1) = \mathbf{h}(n) + \mu\mathbf{x}(n)e(n), \tag{2.1}$$

where $\mathbf{x}(n)e(n) = \nabla_{\mathbf{h}^T} E\{|e(n)|^2\}$, is the minimized cost function when using no error averaging, i.e., each $e(n)$ is used at each iteration to converge to the optimum filter. The cost function is scaled through μ, which defines the convergence step depending on the error variance, this way avoiding possible divergence behavior.

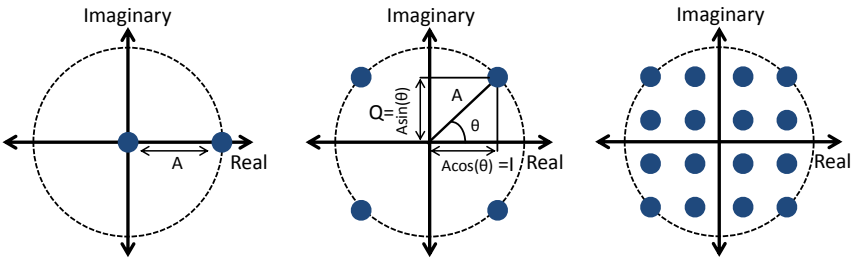
Once the signal is properly equalized hard symbol decision takes place followed by de-mapping, which translates symbols into bits with the equivalent mapping tables used in the transmitter, typically Gray mapping is applied. Finally FEC ensures error free transmission (usually considered as post-FEC bit error rate values below 10^{-15}). Constrained by low latency and power consumption requirements, the FEC scheme to be used in future 200 and 400-Gb/s Ethernet modules is still under debate. Various codes are suggested for the different modulation schemes. For instance, for PAM-2 modulation a low-complexity low-gain code such as KP4, with

pre-FEC limit bit error rate (BER) of $3 \cdot 10^{-4}$ might be sufficient to achieve error free operation. However, PAM-4 modulation might require schemes presenting larger coding gain. BCH and multi-level coding (MLC) schemes with coding gains above 8 dB appear to be suitable candidates, exhibiting pre-FEC BER requirements below 10^{-3} [92].

2.4 Coherent transceivers

As mentioned in the first section, coherent technology is typically used in long-haul systems. However, lowering the cost of such devices would made them very interesting in more cost sensitive but bandwidth-hungry environments such as data centers and metropolitan networks. In coherent systems data can be carried in both amplitude and phase of the optical signal. Fig. 2.11 illustrate some exemplary constellations plotted on a complex plane, including real and imaginary axes. Fig. 2.11(a) shows a PAM-2 constellation, evincing the fact that intensity modulation schemes require only one axis (e.g., real axis) of the complex plane.

On the other hand, in coherent systems two dimensions are used: amplitude and phase. In order to plot such kind of two-dimensional constellations we require the full complex plane. Fig. 2.11(b) shows a quadrature phase-shift keying (QPSK) constellation, in which information is coded along four equ ol S_n $_n$, whe a-



(a) PAM-2 (b) QPSK (c) 16-QAM

Figure 2.11: (a) PAM-2, (b) QPSK, and (c) 16-QAM constellations, plotted in the complex plane including real and imaginary axes.

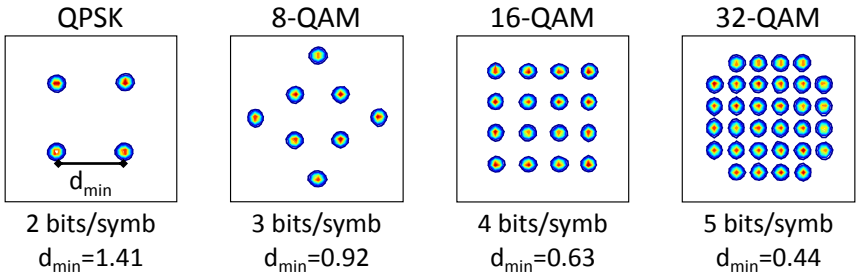


Figure 2.12: Experimentally generated constellations of (a) QPSK, 8-QAM, 16-QAM and 32-QAM signals. Theoretical number of bits per symbol and minimum symbol Euclidian distances (d_{min}) are indicated below for each format.

tion. Nevertheless, as illustrated in Fig. 2.11(b) we usually decompose the complex signal in the two orthogonal dimensions, i.e., in-phase (I) and quadrature (Q) components as follows:

$$S = A \cdot e^{j\theta} = A\cos(\theta) + j \cdot A\sin(\theta) = I + j \cdot Q. \tag{2.2}$$

In the previous modulation scheme (QPSK), data is encoded just onto the phase of the optical signal; however, amplitude can be also simultaneously used to transfer data leading to the family of QAM constellations. Fig. 2.11(c) represents a 16-QAM constellation, which makes use of 16 equally spaced-symbols placed at different phase-amplitude combinations onto the complex plane. Increasing the order of the constellations (number of symbols) allows for high spectral efficiency, i.e., more bits can be transmitted within a single time-symbol. For instance, 16-QAM constellations allow transmitting twice the number of bits per symbol (4) than QPSK (2 bits/symbol). In Fig. 2.12 we display the constellations for modulation formats that we use in Chapter 4, i.e., QPSK, 8-QAM, 16-QAM and 32-QAM, indicating below the number of bits per symbol that can carry each format and their minimum symbol Euclidean distance (d_{min}). We can clearly observe in Fig. 2.12 that d_{min} fastly decreases when increasing the symbol density (constellation order). It is important to notice that the closest the symbols, the lowest the robustness to noise and distortions. Therefore, channel impairments play a big role when choosing the modulation format to be used.

In the following subsections we describe the implementation of coherent

transmitters and receivers. We also detail the DSP required in coherent receivers to demodulate the signal.

2.4.1 Coherent transmitter

As depicted in Fig. 2.13, the layout of a coherent transmitter is quite similar to the one used in EML modules shown in the previous section. Nevertheless, in coherent systems one can transmit independent signals through the two orthogonal states of polarization: transverse electric (TE) and transverse magnetic (TM) modes, for which we will refer as X- and Y-polarization, respectively, along this section. This technique is known as polarization division multiplexing (PDM) which allows doubling capacity with respect to single-polarization systems. Furthermore, as described above, in each polarization we can now send data over the two dimensions of the complex field (I and Q), which makes a total of four available dimensions (I_x, Q_x, I_y, Q_y). Thus we now require four DACs and driver amplifiers to drive the four inputs of a dual-polarization (DP) I/Q-MZM, which is capable of modulating all four light components independently.

Laser: As in coherent systems the light phase is used for modulation, the transceivers require lasers with high frequency and phase stability, i.e., with low linewidth. DFB lasers described before, present linewidths in the order of a few MHz, which could be used for low order constellations such as QPSK bu
and avoid cycle s

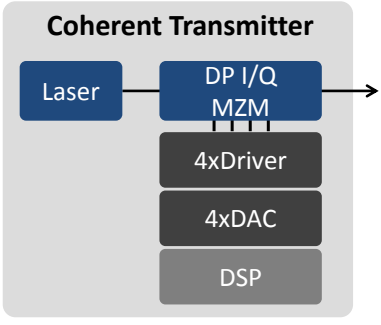


Figure 2.13: Block diagram of a coherent transmitter.

deployed. Thus lasers with large tunability are required, which DFB lacks[5]. High-end coherent transceivers typically use external-cavity lasers (ECLs), in which the gain medium is placed in between an external cavity built within a micro-optics assembly. ECLs provide low linewidth in the order of 100 kHz and full tunability over the C-band (40 nm). Nevertheless, widely tunable lasers with relatively low linewidth (sub-MHz) can also be built in monolithic integrated structures such as digital supermode distributed Bragg reflector (DS-DBR). Different from DFB lasers, in such structures the gain medium is placed between two Bragg reflectors, which usually contain multiple sections. Despite exhibiting larger linewidth than ECL lasers, they have reduced size and lower cost, which makes them interesting for lower cost coherent transceivers. Furthermore, they have the potential to provide fast tunability (100 ns), which becomes interesting when performing optical packet switching in WDM networks [93].

Modulator: The layout of DP-I/Q-MZM is shown in Fig. 2.14. The light of a laser source is split into two I/Q-MZM, each generating an independent signal for each polarization (see light gray areas). Each I/Q modulator contains two branches (I, Q), each performing amplitude modulation over one axis (one dimension). Then, one of the branches undergoes a 90° ($\pi/2$ rad) phase shift with respect to the other branch, hence providing the orthogonal I and Q axes after their combination. Finally both signals (polarizations) are combined using a polarization beam combiner (PBC).

A MZM typically supports a unique polarization mode (i.e., TE). Hence the laser light is singly polarized (TE mode), which polarization state is kept up to the modulator using polarization maintaining fiber. Hence both I/Q-MZMs work with TE signals until one of them is rotated to TM before entering the PBC. In some cases, the PBC is implemented such that one of the entries is automatically rotated, without the need for a previous rotator, see 2D-vertical grating coupler in [94].

Electrical generation: Current coherent transmitters use high-resolution (e.g., 8 bits) DACs typically integrated together with the DSP in the same ASIC in CMOS technology. Latest high-end coherent transceivers implement the state-of-the-art DACs described in the latter section (e.g.,

[5]DFB lasers typically offer a tunability of few nm only.

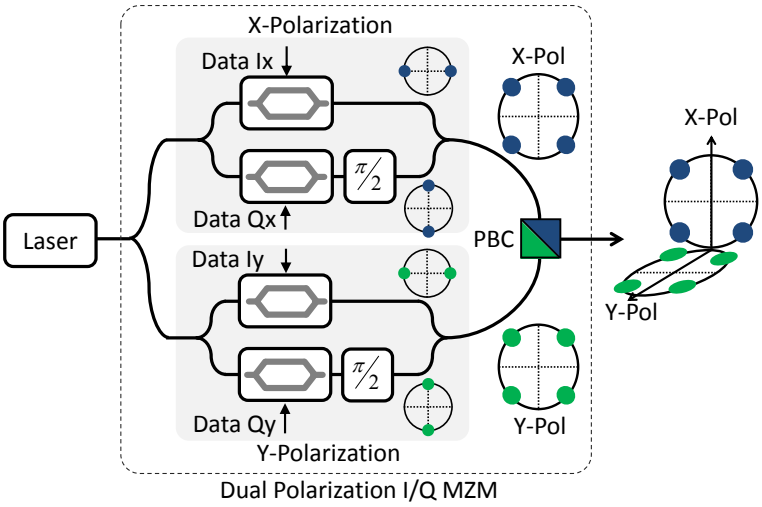


Figure 2.14: Layout of a DP-I/Q-MZM modulator.

92 GS/s DAC). Nevertheless, in this case four DACs are required. The output of each DAC is then amplified by a quad-driver circuit, including four amplifiers placed in the same chip.

2.4.2 Coherent receiver

As described in the previous section, in IM-DD systems a single photodiode is sufficient to detect the intensity of the signal, so a photodiode detects the modulus square of the signal field (intensity). Nevertheless, during that process, the phase information required in coherent reception is lost. Hence in coherent systems, the received optical signal is interfered with the light of a continuous-wave laser, called local oscillator (LO), working at a wavelength approximately equal to the transmitter laser. This process is called intradyne detection, which is used to coherently down-convert the received signal to baseband while capturing the whole complex field of the recovered optical signal.

The detailed layout of a coherent receiver is shown in Fig. 2.15. First the two polarizations of the received signal (S) are demultiplexed by a polarization beam splitter (PBS). As in the transmitter, the demultiplexed TM-

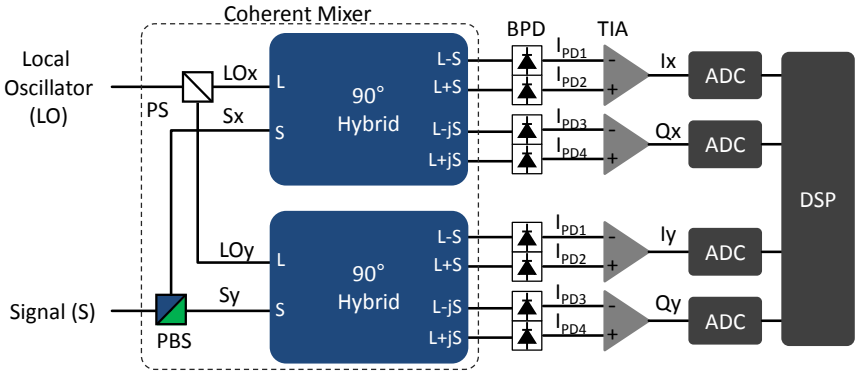


Figure 2.15: Block diagram of a coherent receiver.

polarization is rotated to TE[6]. This way the following blocks can be built supporting a single TE-polarization. Simultaneously, the light from the local oscillator (LO) traverses a power splitter (PS)[7], creating two beams (LO_x, LO_y) to interfere with signals coming from each polarization (S_x, S_y), respectively, in the following 90° hybrids.

From this point, each polarization undergoes the same process (see topmost and bottom-most parts of Fig. 2.15). The 90° hybrid creates four different interfering outputs: the local oscillator with the signal (L+S), the local oscillator with the signal phase-shifted by 90° (L+jS), and the local oscillator with the complementary signals of the two latter, (L-S) and (L-jS). The four interfering signals are detected by two pairs of balanced photodetectors (BPD).

Eq. (2.3) mathematically describes the input signals into the 90° hybrid (S and L):

$$\begin{aligned} L &= A_L \cdot e^{j(\omega_L t+\theta_L)} \\ S &= A_S \cdot e^{j(\omega_S t+\theta_S)} \end{aligned} \quad , \tag{2.3}$$

where A_i, ω_i and θ_i are the amplitude, angular frequency and phase of the local oscillator ($i = L$) and received signal ($i = S$). Using such definitions,

[6]The rotation can take place in the PBS when using 2D-vertical grating couplers.
[7]Note that in this case the local oscillator emits single TE-polarized light.

we can calculate the photodetected current of each output of the 90° hybrid:

$$\begin{cases} I_{PD\ 1} \propto |LO - S|^2 = |A_L|^2 + |A_S|^2 - 2A_L A_S \cos\left((\omega_S - \omega_L)t + \theta_S - \theta_L\right) \\ I_{PD\ 2} \propto |LO + S|^2 = |A_L|^2 + |A_S|^2 + 2A_L A_S \cos\left((\omega_S - \omega_L)t + \theta_S - \theta_L\right) \\ I_{PD\ 3} \propto |LO - jS|^2 = |A_L|^2 + |A_S|^2 - 2A_L A_S \sin\left((\omega_S - \omega_L)t + \theta_S - \theta_L\right) \\ I_{PD\ 4} \propto |LO + jS|^2 = |A_L|^2 + |A_S|^2 + 2A_L A_S \sin\left((\omega_S - \omega_L)t + \theta_S - \theta_L\right) \end{cases} \tag{2.4}$$

Each pair of photodiodes currents ($I_{PD\ 1}$, $I_{PD\ 2}$) and ($I_{PD\ 3}$, $I_{PD\ 4}$) are subtracted and amplified by the subsequent TIAs; leading to the corresponding in-phase (I) and quadrature (Q) components of the signal. If the pairs of photodetectors are well balanced, such differentiation allows removing the $|A_L|^2$ and $|A_S|^2$ components of the detection process, see the calculation as follows:

$$\begin{cases} I = I_{PD\ 2} - I_{PD\ 1} \propto A_L A_S \cos\left(\Delta\omega t + \theta_S - \theta_L\right) \\ Q = I_{PD\ 4} - I_{PD\ 3} \propto A_L A_S \sin\left(\Delta\omega t + \theta_S - \theta_L\right) \end{cases} . \tag{2.5}$$

Please note that the same process is performed for both polarizations. Finally the resulting I and Q components are digitized by four ADCs for later processing. Typically the latest generation of CMOS-based high-resolution ADCs, described in the previous section, are included in coherent receivers integrated in the same ASIC with the DACs and the DSP.

2.4.3 Digital Signal Processing

During this section we describe the DSP required in coherent receivers to be able to demodulate the received signals. Fig. 2.16(a) shows the block diagram of the DSP used in current coherent receivers, which work in circuit-mode. Nevertheless, in Chapter 4 we will perform coherent transmission in packet-mode, which requires the modification of some of the blocks in order to quickly adapt to the channel changes from burst to burst, and hence being able to recover short packets coming from different emitters. The blocks to be modified are shown in blue in Fig. 2.16(b). In the following subsection we first explain the functionality of each block in circuit-mode and later describe the novel blocks used for packet-mode demodulation.

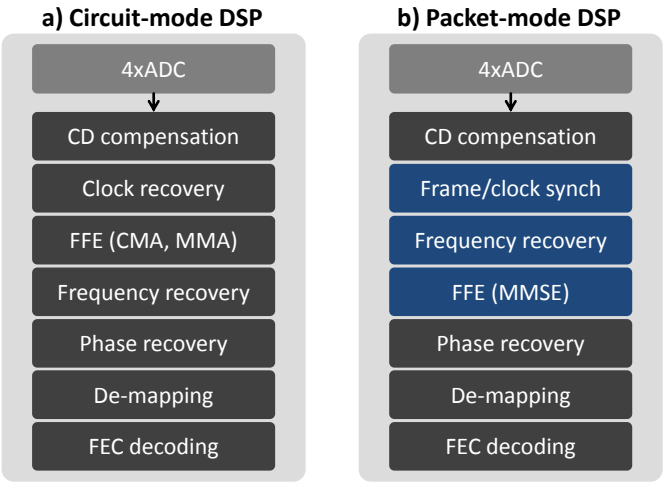


Figure 2.16: DSP blocks used in (a) a generic (circuit mode) receiver and (b) a packet-mode receiver.

2.4.3.1 Generic (circuit-mode) digital signal processing

Chromatic dispersion (CD) compensation: In optical fibers the group velocity of the propagating light depends on its wavelength/frequency. This effect is known as chromatic dispersion. Hence, in an optical modulated signal, each frequency component propagates at different group velocities, which induces a broadening of the temporal pulses (or symbols), hence generating ISI. The effect of the chromatic dispersion can be modeled through the following transfer function [95]:

$$G_{CD}(L,\omega) = e^{-j\frac{DL\lambda^2}{4\pi c}\omega^2}, \tag{2.6}$$

where ω is the angular frequency ($\omega = 2\pi f$), L and D are the length and dispersion coefficient of the fiber link, and λ and c are the wavelength and speed of the propagating light. The chromatic dispersion can be effectively compensated by applying a N-taps FIR filter with the following coefficients [95]:

$$h_k = \sqrt{\frac{jcT^2}{D\lambda^2 L}} e^{-j\frac{\pi cT^2}{DL\lambda^2}k^2}, \tag{2.7}$$

where $k \in \{\lfloor -N/2 \rfloor, \cdots, \lfloor N/2 \rfloor\}$ denotes the tap-sample index and $f_s = 1/T$ the sampling rate. The required filter length (N) increases with the fiber length, dispersion coefficient and baudrate. Please note that the

same filter is used for both polarizations. An upper bound for the required filter length can be obtained by supposing a constant dispersion over the frequency range $-0.5f_s \leq f \leq 0.5f_s$ through the following function [95]:

$$N = 2 \left\lfloor \frac{|D|\lambda^2 L}{2cT^2} \right\rfloor + 1. \tag{2.8}$$

This gives us an indications of the broadening of the impulse response as a function of the baudrate and the length of the fiber link. Using a 1550 nm laser, and standard SMF with D=17 ps/nm/km, a transmission of a 32.5 GBd signal over 1000 km requires more than 500 T/2 taps to recover from chromatic dispersion, which is usually done in the frequency domain. Nevertheless, in Chapter 4 we will use coherent transmission for short-reach intra-datacenter connections with maximum distances of few kilometers. The same calculation can be performed for 5-km links, resulting into a FIR filter with less than 10 T/2-spaced taps. Hence, in real data center networks, a CD compensation dedicated filter can be avoided, using then the following equalization (FFE) block to compensate for dispersion.

Contrarily to coherent receivers, IM-DD systems cannot effectively compensate for chromatic dispersion (digitally), because the phase is lost in the direct detection process. Hence, FFE used in IM-DD can just partially remove certain amount of CD-induced ISI. Furthermore, CD-induced ISI increases quadratically with the baudrate ($1/T$), which dramatically limits the reach of high-speed IM-DD systems when working in the C-band.

Clock recovery: As in IM-DD receivers, clock recovery aims at compensating for mismatch in sampling frequency between transmitters and receivers. As described, in Fig. 2.10, after sampling frequency/phase detection, the clock can be recovered in a feedback scheme, by updating the sampling rate of the ADCs through a PLL, or using a forward scheme, where clock recovery is performed by an interpolator. Coherent receivers use clock-frequency/phase detection algorithms that are insensitive to optical phase/frequency variations such as Gardner [96] or Godard [97].

Equalization (FFE): Along light propagation, the state of polarization slowly rotates in the transmission link. Hence, at the receiver one detects

a linear combination of both polarizations. The original polarization states can be recovered in 2×2 multi-input multi-output (MIMO) complex equalization process, which will also equalize the possible linear impairments occurring in the channel (e.g., ISI). The per-sample 2×2 equalization matrix can be observed in Eq. (2.9):

$$\begin{bmatrix} s_x(n) \\ s_y(n) \end{bmatrix} = \begin{bmatrix} \mathbf{h_{xx}}(n)^T & \mathbf{h_{yx}}(n)^T \\ \mathbf{h_{xy}}(n)^T & \mathbf{h_{yy}}(n)^T \end{bmatrix} \begin{bmatrix} r_x(n) \\ r_y(n) \end{bmatrix}. \tag{2.9}$$

where $[(r_x(n), r_y(n))]^T$ and $[(s_x(n), s_y(n))]^T$ are the n^{th} received (r) and equalized (s) samples for both X- and Y-polarizations [98], and $\mathbf{h_{ij}}$ are FIR equalizer filters, which lengths need to cover the channel impulse response.

As for IM-DD systems the adaptive process used to find the optimum filters is done through a gradient descent algorithm, which minimizes a certain cost function. In order to blindly adapt, coherent systems usually use phase insensitive cost functions such as constant-modulus algorithm (CMA) or multi-modulus algorithm (MMA) to recover constant amplitude (QPSK) or multi-level (QAM) signals respectively, which use the amplitude (intensity) of the signal as reference to calculate the error [98,99]. The cost functions of both algorithms are shown in Eq. (2.10):

$$\begin{aligned} J^i_{CMA} &= \left(1 - |s_i|^2\right)^2, \\ J^i_{MMA} &= \left(|\hat{a}|^2 - |s_i|^2\right)^2. \end{aligned} \tag{2.10}$$

where $i \in \{x, y\}$ refers to each polarization and s_i is the equalized signal described in Eq. (2.9). In J_{CMA} the normalized symbol intensity $|s_i|^2$ is subtracted to 1, which corresponds to the intensity radius of a power-normalized QPSK constellation. However, in multilevel signals, the constellation has different radius. In that case, J_{MMA} calculates the error taking as a reference the closest constellation intensity radius ($|\hat{a}|^2$) to the equalized signal $|s_i|^2$.

Carrier frequency and phase recovery: As described in Eq. (2.5), the process of intradyne reception induces a certain amount of frequency offset $\Delta\omega = \omega_S - \omega_L$, originated by the wavelength mismatch between transmitter (S) and local oscillator (L) lasers, and phase error θ_{ε_n} originated by the intrinsic phase variation taking place in both lasers, where n denotes the

n^{th} symbol of the received signal. Fig. 2.17 shows the process to perform frequency and carrier recovery.

First the frequency offset is estimated using for instance the N^{th}-power periodogram, in which first the signal is raised to the N^{th} power (typically 4) to partially cancel the signal modulation, and then the periodogram estimates the frequency offset $\Delta\omega$, which is subtracted from the signal: $s_1(n) = s_0(n) \cdot e^{-j\Delta\omega n}$. Then the remaining phase error θ_{ε_n} is tracked using a blind phase estimation algorithm. The complex conjugate of the estimated θ_{ε_n} is also applied to the signal to recover the original phase: $s_2(n) = s_1(n) \cdot e^{-j\theta_{\varepsilon_n}}$.

Multiple blind carrier phase estimators (CPE) have been proposed in the literature to recover the phase while using different modulation formats. Viterbi & Viterbi is one of the algorithms most extensively used for QPSK modulation due to its simplicity [100]. Such estimator raises the signal to the 4^{th} power to cancel the QPSK modulation, remaining then only the phase noise, which have two origins: intrinsic laser phase fluctuations, which typically follow a trackable random walk process (with time-variations much slower than the symbol-rate); and random white noise, produced in the other elements of the chain (e.g., EDFA). CPE estimators

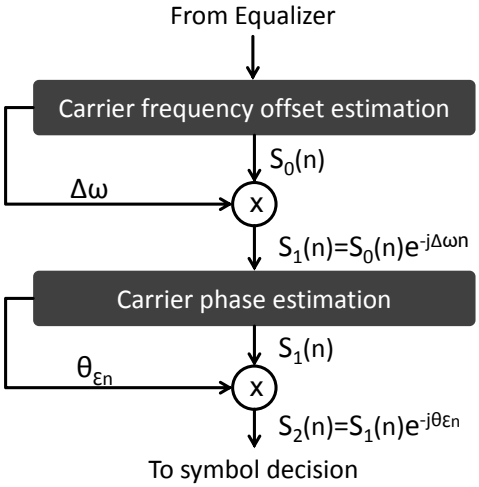


Figure 2.17: Carrier frequency offset and phase recovery blocks.

Despite the simplicity and robustness of this method, the 4^{th} power can only cancel QPSK modulation, for which this approach is limited to. For higher order modulations, such as 16-QAM, other methods transform the original constellation into a QPSK to apply later the Viterbi & Viterbi algorithm [101]. Generic algorithms capable to recover any constellation have been also proposed, for instance the blind phase search algorithm [102]. In such scheme, each symbol is multiplied by a set of test phases (typically 32 or 64, sub-dividing a symmetry quadrant of the transmitted constellation). Then, the algorithm performs the minimum Euclidean distance between the resulting tests and the original constellation. After averaging several time-symbols, the test giving the minimum Euclidean distance is chosen as phase error. In all above mentioned algorithms, the phase estimation can lead to ambiguities due to symmetry of the constellation. Such issue is typically addressed by regularly sending pilot symbols, which allows the receiver to resolve the ambiguities.

Forward error correction: After soft/hard de-mapping, the obtained bit-wise information enters into the FEC decoder. As shown in

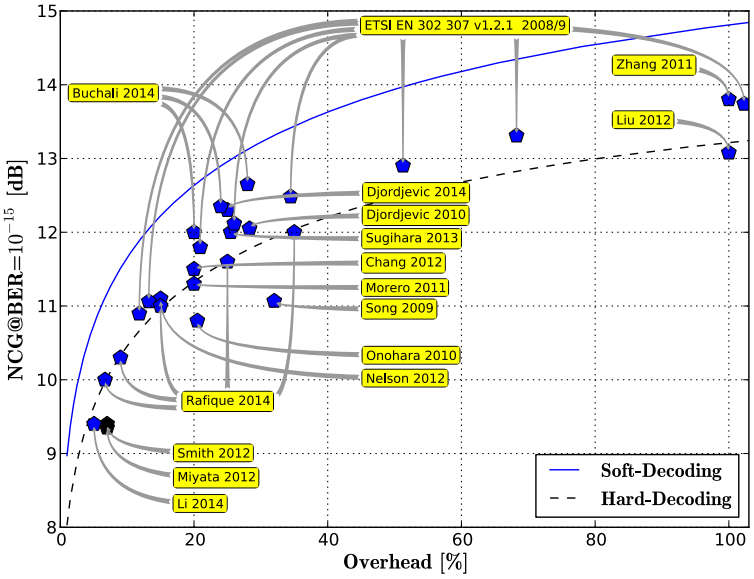


Figure 2.18: Net coding gain versus overhead of the state of the art FECs for high-speed optical communications. From [103].

Fig. 2.18 [103], a large number of FEC codes have been proposed for coherent applications, while implementing different schemes such as block-turbo codes (BTC) or low-density parity-check (LDPC). We can distinguish two main families: hard-decision (HD) and soft-decision (SD) FEC codes. As depicted in Fig. 2.18, SD-FEC typically provides 1.5 dB gain over HD-FEC schemes; nevertheless, such gain comes at the expense of implementation complexity. Despite their high complexity, current coherent transceivers implement SD-FEC schemes because they can achieve high coding gain (more than 11 dB) with supportable overhead ratios (≈20%). Neverthless, the FEC implementation will have an strong impact to the end-to-end latency. The longer the code-word used and the number of iterations performed along the decoding process, the larger the latency. FEC implementation is out of the scope of this thesis. Hence along our experiments we will use the pre-FEC BER limits of FEC codes extensively used in the literature to asses the performance of our transceivers.

2.4.3.2 Packet-mode DSP blocks

As shown in Fig. 2.16(b), there are only a few blocks that need to be updated to perform packet-mode operation: frame synchronization, carrier frequency offset and equalization.

Frame synchronization and frequency offset estimation: In packet-mode systems, a receiver can subsequently receive packets from different transmitters. However, in order to be able to correctly demodulate the signal, the beginning of each packet has to be detected at a symbol level. There are several approaches to perform frame synchronization. For instance, when empty gaps are inserted between optical packets, the power of the signal can be used to identify the beginning of the packet [104]. Nevertheless, such approach can lead to false packet detection in the presence of a noisy channel.

More robust approaches use training sequences to detect the beginning of a packet. The so-called data-aided schemes typically make use of a training sequence composed by several repeated sub-sequences whose polarity is varied depending on the method. For instance, Schmidl et al. [105], use a training sequence composed by two equal sub-sequences $[+A, +A]$,

while Shi et al. [106] include four equal sub-sequences presenting different polarities $[+B, +B, -B, +B]$. The correlation properties of such training sequences allow for robust timing estimators. In addition, the same schemes provide an accurate estimation of the frequency offset. In Chapter 4, we use the estimators proposed in [106], which provide fast and accurate timing and frequency offset estimation, while keeping low complexity.

Equalization: Blind adaptive equalization schemes such as the CMA and MMA require a certain converging period to optimize equalizer taps. The length of such period depends on the state of polarization of the received signal, on the amount of noise and on the modulation order. Large converging periods can be expected for noisy channels and high order modulations. For instance, a convergence study performed in [107], revealed that the CMA requires around 10,000 symbols to converge to nearly the optimum filter when transmitting a 16-QAM signal under certain conditions[8].

In packet-mode operation, each sequentially received packet can come from a different transmitter, hence each packet can be exposed to different impairments (different channel responses). Therefore FIR filters need be re-calculated on a per-packet basis. Accounting that packets have a length of the order of tens of thousands of symbols (32,000 symbols for one microsecond packet at 32-GBd), using a blind adaptation scheme (e.g., CMA or MMA) is unfeasible due to their large convergence time: one third of the packet symbols would be wasted. In this case data-aided schemes can be used to rapidly obtain the optimal filter. For instance using a training sequence as a reference, data-aided LMS can achieve the same performance (in terms of OSNR penalty) as the blind CMA with only few hundred symbols of convergence period (in the particular scenario described above [107]).

When using short microsecond-long packets, the channel can be considered non-variant over one packet, considering typical slow-rate polarization rotations (>millisecond). In this particular scenario, channel tracking (filters' continuous adaptation) can be avoided. Hence, for short packets, one can also use static equalization, in which first the channel is estimated using training sequences and the least square (LS) algorithm; and subsequently

[8]The study shows the evolution of the optical signal-to-noise ratio (OSNR) penalty (respect to theory) at BER= 10^{-3}, measured while different algorithms are under the adaptation process (converging), versus number of symbols, when using 13-taps T/2-spaced equalizers. The signal was a 112 Gb/s PDM-16-QAM.

the optimum filter can be obtained through the minimum mean square error (MMSE) estimator. Being such the fastest scheme for channel recovery, while requiring overheads smaller than 1% [107], we will use it in Chapter 4 to recover microsecond-scale packets.

Chapter 3

Short-term perspective: High data-rate IM-DD transceivers

3.1 Introduction

The revolution of cloud services is driving a huge bandwidth demand in datacenters, which is nearly doubling every year [12]. Such rapid traffic growth urgently requires a new generation of high-speed short reach optical transceivers operating at 100 Gb/s and beyond. However, the environment of data centers is extremely sensitive to cost, power consumption and footprint, requiring interfaces capable of large capacities, while guarantying low cost and fitting in tight interconnect slots. As described in Chapter 1, today's 100 GE (LR4) solutions generate 4×25 Gb/s wavelength-multiplexed channels to achieve 100 Gb/s transmission [33]. To cope with the demand for larger bandwidths, the IEEE is currently working on the standardization of 200 and 400-GE [28]. Proposed first 200 and 400-GE generations are based on 4 and 8 × 50 Gb/s wavelength-multiplexed channels, respectively, each transmitting 25-28 GBd PAM-4 signals, which are attractive since they are compatible with existing 100 GE building blocks in terms of bandwidth (driver, laser, photodiodes, TIA). Nevertheless, second generation 400-GE proposals follow the quad-lane trend of their predecessors

(100 and 200-GE), aiming for a transceiver including 4×100 Gb/s lanes each carrying (most probably) a PAM-4 56-GBd signals. Nevertheless this approach requires the development of novel components allowing modulation speeds above 50 GBd.

In order to achieve serial rate of 100 Gb/s with IM-DD transceivers, different modulation schemes have been proposed in the literature, ranging from the simple and robust on-off keying (OOK) [108], to the more spectrally-efficient duo-binary signaling [109], PAM [110–113], DMT modulation [61] or CAP [62]. Among such schemes, PAM signaling offers a good trade-off between spectral-efficiency and complexity. On one hand, PAM-4 signaling doubles spectral efficiency of OOK, hence it requires less opto-electronic bandwidth. On the other hand, multi-carrier formats such as DMT or CAP require more complex DSP and higher effective number of bits (ENOB) DAC and ADC than PAM-4. Furthermore, DACs and ADCs generally lack sufficient output swing to efficiently drive a modulator, so that an electrical amplifier (driver) must be inserted between the DAC and the modulator, which leads to higher cost, higher power consumption and bandwidth limitations. All these reasons make PAM today's most attractive scheme to cope with short reach 100-Gb/s interfaces [64].

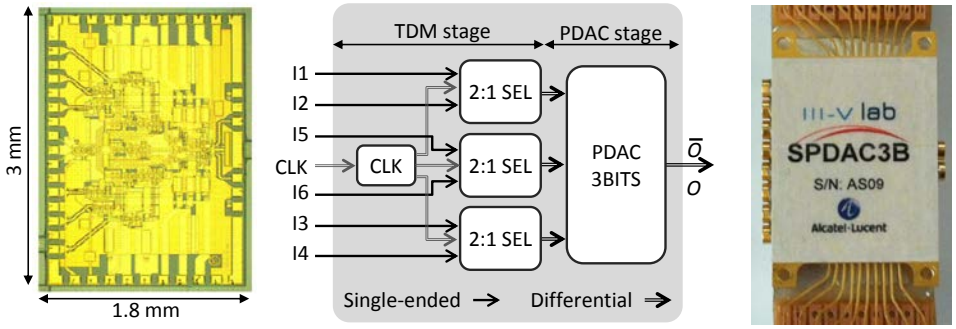
As target data rates surpass 100 Gb/s, the exploitation of light intensity as the only dimension for modulating results in severe opto-electrical bandwidth limitations, thereby aggravating the performance-complexity trade-off and making it truly challenging to keep up with the pace. Hence, several research experiments have proposed the use of PDM as means to increase capacity beyond 100 Gbit/s [114, 115]. Nevertheless PDM requires twice the number of modulators, four times the number of detecting devices, and complex digital signal processing in the receiver to demultiplex both polarizations. Consequently increasing data rate of single-polarization single-wavelength transceivers leads to more cost-effective solutions. One of the major barriers that slows down the increase of data rates above 100 Gb/s, is the relatively low 3-dB bandwidth of current commercial DACs (around 20 GHz). Latest research reports on IM-DD transceivers demonstrating throughput above 100 Gb/s make use of complex electrical setups which combine many state-of-the art electronic singular components to generate such high-speed electrical signals [63, 70], detailed in following subsections.

Along this chapter we present our solution for next generation datacenter optical interfaces. Our approach is based on three main pil-

lars: 1) intensity-modulation and direct-detection using single-polarization single-wavelength transmitters to keep cost low; 2) ultra-high symbol rate to increase data rate in a cost-effective way; and 3) multi-level PAM to double or triple capacity while improving spectral efficiency. In order to be able to overpass bandwidth limitations of commercial electronic components, and increase data rates above 100 Gb/s we use a selector power digital-to-analog converter (SP-DAC) fabricated in-house, which is capable of generating up to 8-levels PAM electrical signals at symbol rates as high as 100 GBd, refer to Section 3.2 for a detailed description of the component. Following sections present a series of experimental demonstrations realized along this thesis. We report first in Section 3.3 a 112-Gb/s (100 Gb/s net data rate) PAM-4 integrated transmitter including a DFB laser and a high-bandwidth EAM capable of working at 56 GBd [116, 117]. Being such the closest solution fulfilling today's market requirements, we have performed a complete performance analysis while studying the impact of several important parameters such as the receiver bandwidth and the equalization complexity. Then we present in Section 3.4 several results achieving data rates beyond 100 Gb/s. We first demonstrate 168-Gb/s (150 Gb/s net data rate) transmission using two different approaches: 1) increasing the modulation order to PAM-8 while keeping the symbol rate at 56 GBd [118, 119]; and 2) increasing the symbol rate up to 84 GBd, while keeping a simpler PAM-4 modulation. Finally we demonstrate a 200-Gb/s (178.5 Gb/s net data rate) transceiver by scaling up the symbol rate of a PAM-4 signal to 100 GBd [120].

3.2 Electrical generation: Selector Power DAC

As mentioned in the latter section, one of the main constraints found when willing to increase baud rates in optical systems comes from the limited bandwidth of electrical DACs. Multi-level signals with baud rates as high as 64 and 72 GBd have been demonstrated in coherent long-haul systems using commercially available high-resolution complementary metal-oxide-semiconductor (CMOS) DACs working at 88 and 72 GS/s [121, 122]. Nevertheless, both experiments had to make use of optical frequency predistortion (up to 15 dB in [122]) to compensate for the limited 3-dB electrical bandwidth of such DACs (< 20 GHz). Recently, a 100-GS/s CMOS DAC was also reported, however its 13-GHz bandwidth limited the sym-



a) SP-DAC chip b) SP-DAC schematic c) SP-DAC module

Figure 3.1: (a) SP-DAC chip microphotography, (b) functional schematic and (c) packaged module.

bol rate to 40 and 50 GBd, when using PAM-4 and PAM-8, respectively. Furthermore, CMOS technology usually does not provide suffiecient electrical swing ($< 1\ V_{pp}$) to efficiently drive an optical modulator, hence an electrical amplifier (driver) is also required. On the other hand InP DHBT techonology is known for its large electronic bandwidth (> 50 GHz) and its capacity for outputting high voltages ($> 4.5\ V_{pp}$) [83].

In all the experiments presented in this chapter we used a SP-DAC fabricated by InP DHBT technology by III-V Lab capable of outputting multi-level electrical signals at symbol rates as high as 100 GBd. The SP-DAC is a $1.8 \times 3\ mm^2$ integrated circuit composed of 151 transistors which combine high operation frequencies ($F_t/F_{max} > 300$ GHz) and high breakdown voltage (4.5 V) [123], see Fig. 3.1(a).

The functional schematic of the chip is depicted in Fig. 3.1(b). The SP-DAC comprises two main stages: a time division multiplexing (TDM) stage and a power DAC (PDAC) stage. In the first stage, six two-level data inputs (I1-I6) are temporally multiplexed by pairs through three selectors (2:1 SEL), each outputting a two-level signal at twice the inputs' data rate, see eye-diagram of an exemplary data input at 28 GBd in Fig. 3.2(a). The circuit is driven with one half-rate clock (CLK) to multiplex the input data and to align the power DAC inputs, see exemplary 28 GHz input clock in Fig. 3.2(b). Each selector supplies one of the three encoding bits of the following power DAC stage, which may operate in 1-, 2-, or 3-bit mode yielding PAM-2, PAM-4 and PAM-8 signals with up to 4-Vpp differential amplitude by simply setting the appropriate DC controls. Figs. 3.2(c,d and

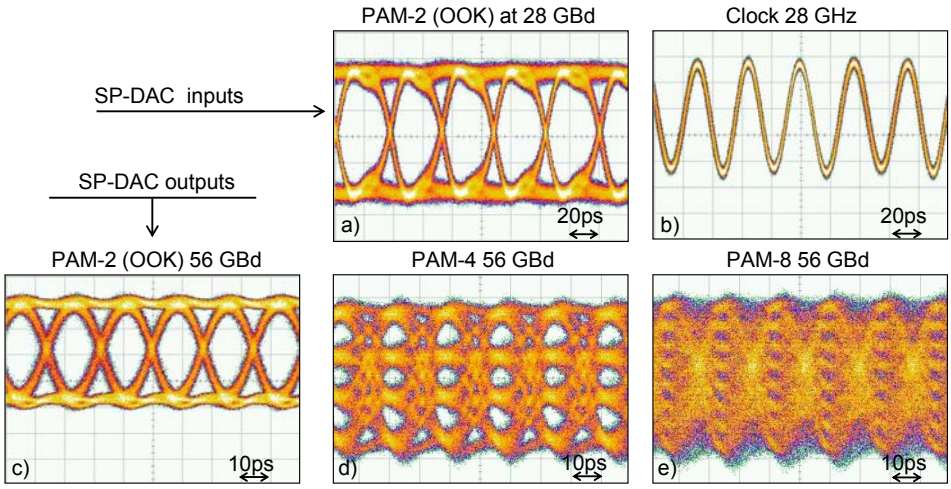


Figure 3.2: Exemplary traces of the SP-DAC inputs and outputs when generating 56-GBd multi-level signals. (a-b) SP-DAC inputs: 28-GBd PAM-2 data input eye-diagram and 2–GHz clock input, respectively. (c-e) SP-DAC outputs when working in different modes: 56-GBd PAM-2, PAM-4 and PAM-8, respectively.

e) show the eye-diagrams of the chip output when operating under different modes (PAM-2, PAM-4 and PAM-8 respectively) at 56 GBd.

As one can easily realize this chip requires input signals at only half the output symbol rate, thanks to the 2:1 selectors. The SP-DAC embeds the largest-speed electronics inside a single compact chip, which advantageously relaxes bandwidth constraints on all electronic wiring in the transmitter. A photograph of a packaged module with GPPO connectors can be seen in the inset of Fig. 3.1(c). The adjustable output amplitude of the module is large enough to efficiently drive the modulator, hence eliminating the need for any additional linear amplifier. Additionally, through the DC controls, the SP-DAC allows for amplitude level distortion and peaking (frequency pre-emphasis) of the output waveform which can be used to pre-compensate for possible modulator nonlinearities and for frequency response limitations.

3.3 112-Gb/s IM-DD optical transceiver

Being 10 and 40 Gb/s the interfaces most used today in large datacenters, next high-speed optical interfaces are expected to carry 100 and 400 Gb/s

to be able to cope with the fastly growing data traffic. The ultimate solution recommended today by IEEE P802.bs 400-Gb/s Ethernet Task Force [28] is based on four wavelength multiplexed channels, each carrying 112 Gb/s (100 Gb/s net data rate) by using PAM-4 modulation at 56 GBd [28]. Hence, the research community is putting lots of efforts in conceiving such transceivers. Several experiments have demonstrated PAM-4 112-Gb/s transmissions at 56 GBd, i.e., with silicon modulators [110] and InP EML [111]. In both experiments, (pre-) equalization is applied at transmitter (TX) and receiver (RX) sides, which requires a minimum of 7+27 taps (TX+RX equalizers) to achieve BER below $3.8 \cdot 10^{-3}$ (being such the 7%-overhead HD-FEC limit) [110], or 19+21 taps to reach $2 \cdot 10^{-4}$ BER [111]. Additionally, an EAM integrated together with a DFB laser transmitter was shown to provide an open 112-Gb/s PAM-4 eye-diagram without equalization, but no bit error rate measurement was reported at the date [112]. Finally, an experiment used PAM-4 over 2 km with a real time clock and data recovery (CDR) circuit, yielding BER as low as $2 \cdot 10^{-6}$. Nevertheless, this experiment was performed using a commercial MZM, which presents relatively large footprint [113].

Along this section we report a low-cost 56-GBd PAM-4 compact transmitter including in a single package a high-power 18-dBm DFB laser and a high-speed EAM modulator with up to 50-GHz bandwidth. Thanks to the large transmitter bandwidth we prove signal recovery without equalization at the expense of large bandwidth receiver. Subsequently we study the trade-off between receiver bandwidth and equalization complexity aiming at best possible sensitivity[1]. We demonstrate successful 112-Gb/s 2-km transmission, even when limiting receiver bandwidth down to 18 GHz or reducing equalizer length to only 3 taps [116, 117].

3.3.1 DFB-EAM transmitter

The transmitter module was designed and fabricated in III-V Labs. It includes a DFB laser co-packaged with a high-speed shallow ridge InGaAlAs EAM, see schematic and packaged module in Fig. 3.3(a-b). The light source

[1]Sensitivity: minimum signal optical power inserted into the receiver required to achieve BER below the FEC limit. The receiver sensitivity limits the power budget (acceptable loss) of the end-to-end link, including coupling and fiber losses, and inner-component losses (modulator and de/multiplexer in the case of a WDM transceiver).

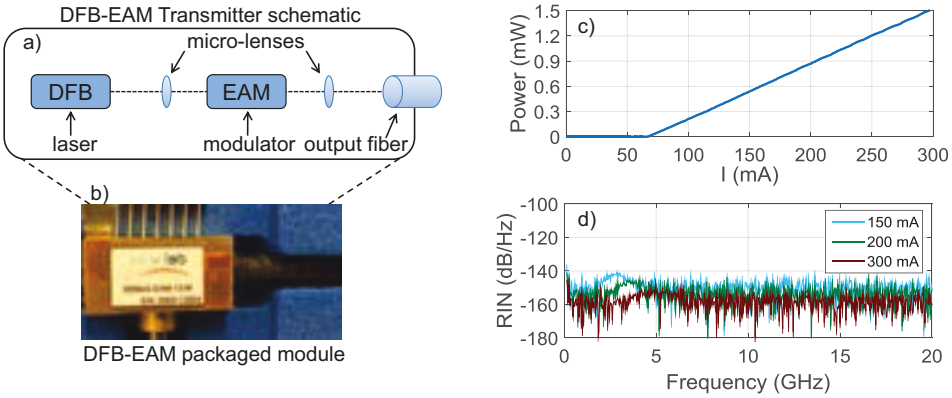


Figure 3.3: (a) integrated distributed feedback laser - electro-absorption modulator (DFB-EAM) transmitter schematic, (b) packaged module, (c) output power versus laser current (I) and (d) RIN measurement for several laser currents.

is a 3.5-µm width InP DFB shallow ridge laser with asymmetrical cladding and 6 quantum wells [124]. The device lases at 1545.8 nm and outputs up to 18 dBm of power when injecting a 300-mA current. Fig. 3.3(c) shows the transmitter module output power as a function of the laser driving current (I) at 25°C when the EAM is not biased. The laser exhibits a threshold current of 67 mA due to low optical confinement in the quantum well. However, a total optical power of 1.5 mW can be coupled out of the transmitter when driving the laser at 300 mA. The relative intensity noise (RIN) of this module is depicted in Fig. 3.3(c) for several laser driving currents. As can be observed, RIN values below -140 and -150 dB/Hz are obtained when driving the laser at 150 and 300 mA, respectively; which is within the range of commercially available high-performance DFB lasers[2] [126].

As shown in Fig. 3.3(a), the modulator is coupled to the DFB laser and to the output optical fiber with micro-lenses for maximum coupling efficiency. The EAM was designed over a 1.5-µm wide waveguide stripe. It includes an active section composed of 10 InGaAs/InGaAlAs tensile strained quantum wells [127] of 75-µm length, as defined by H+ implantation. In order to enlarge the EAM length and hence make its manipulation easier

[2]The RIN describes the intrinsec intensity noise (power fluctuations) of a laser, typically defined as the power spectral density of the intensity noise normalized to the average laser power. Such small RIN values barely impact perfomance, since overall main limitations will be goverened by limited end-to-end bandwidth and thermal noise induced at the receiver, specially for low input receiver powers [125].

and enhance packaging yield, the active section was placed in between two passive sections, making a total EAM length of 520 µm. These passive sections are made of an InGaAsP layer ($\lambda_g = 1.3$ µm) grown by gas source molecular beam epitaxy (GSMBE) and butt coupled to the active section to reduce propagation losses and to prevent early saturation of the device.

Using a broadband source, the EAM insertion loss (including absorption and coupling losses) was measured while changing the bias voltage as a function of the wavelength; see Fig. 3.4(a). When unbiased (0 V), the component presents insertion losses of 13.5 and 15.5 dB at 1600 and 1546 nm, respectively, including relatively high coupling losses of approximately 10 dB. The use of a passive Indium Galium Arsenide Phosphide (InGaAsP) layer limits the absorption in the passive section at low wavelength. Therefore, the modulator presents a large optical bandwidth with an extinction ratio above 10 dB over a large spectral range (1525-1595 nm). Fig. 3.4(b) depicts the extinction ratio (ER) of the transmitter for several laser currents at 25°C. It can be observed that, at a relatively low bias voltage of -2.5 V, the transmitter can achieve large ER values, ranging between 13 dB (100 mA) and 16.5 dB (300 mA). The increase of the extinction ratio when increasing laser current is probably due to self-heating of the modulator [128]. The transmitter frequency response was characterized for several laser currents with a vector network analyzer (VNA). The bias was adjusted from -0.9 V at 100 mA to 1.1 V at 300 mA to maximize the signal in the VNA. From the measurements reported in Fig. 3.4(c), we found

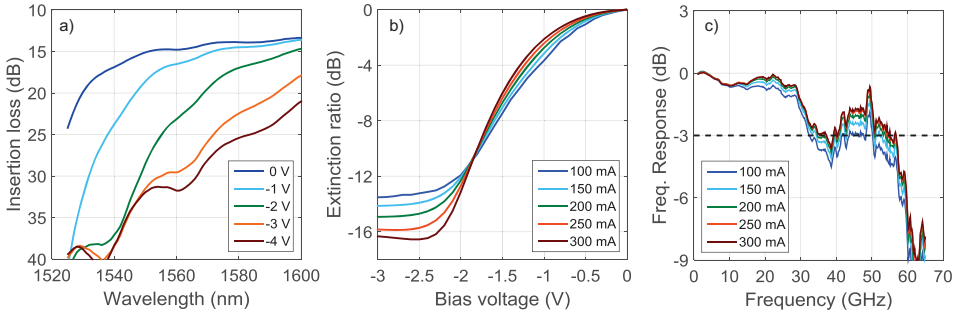


Figure 3.4: (a) EAM insertion loss (including absorption and coupling losses) versus wavelength for different modulator bias voltage. (b) Transmitter extinction ratio versus bias voltage and (c) frequency response for several laser currents.

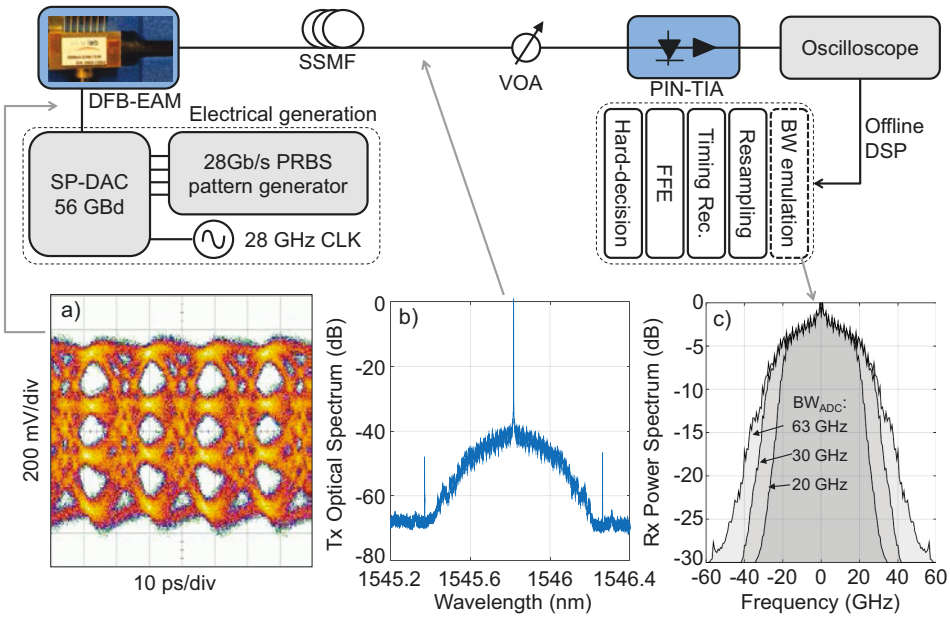


Figure 3.5: Top: Experimental setup. Bottom: (a) Eye-diagram of the electrical signal driving the transmitter module, (b) optical power spectrum of the PAM-4 56-GBd signal and (c) power spectrum of the received signal when emulating different ADC bandwidth.

receiver sensitivity. Then, the signal is detected by a 40-GHz bandwidth PIN photodetector with transimpedance amplifier (PIN-TIA). The amplified photocurrent is sampled at 160 GS/s by a 63-GHz bandwidth digital storage oscilloscope (DSO), and stored for offline processing.

One of the main issues encountered when working at large baud rates such as 56 GBd is the limited bandwidth of components. To evaluate the performance of our transmitter, we use a very high-bandwidth 63-GHz oscilloscope, such that the bandwidth limitations from the receiver are almost negligible. Then we artificially reduce the bandwidth of the oscilloscope, in order to emulate the impairments of practical components, e.g., ADC with more realistic bandwidth. In order to do so, we first apply a 6^{th}-order Butterworth filter with arbitrary bandwidth to the stored waveforms. In Fig. 3.5(c), we show the received signal power spectrum when emulating ADC bandwidths (BW_{ADC}) of 63 GHz (no filtering), 30 and 20-GHz. Note that the DC-component has been removed from the signal. Once the waveform is filtered we perform standard signal processing. First, the signal is resampled to twice the baud rate. Then, clock recovery [97] and FFE using N $T/2$-taps are performed. Finally, hard decision, followed by bit error counting, takes place. A detailed explanation of the digital signal processing is given in Chapter 2.

3.3.3 Back-to-back performance analysis

We first evaluate the performance of our transmitter without performing equalization at the receiver in a back-to-back (BtB) configuration. For such evaluation, we carry out hard decision directly after clock/timing recovery. Fig. 3.6(a) shows the BER as a function of the PIN-TIA input power P_{inRx} for several ADC bandwidths BW_{ADC}. It can be clearly observed that ADC bandwidth must be above 40 GHz in order to achieve BER below the limit of operation of $3.8 \cdot 10^{-3}$, established when considering 7%-overhead HD-FEC. BER improves when increasing BW_{ADC}, leading to a maximum measured performance of BER$= 10^{-3}$ at $P_{inRx} = -2$ dBm with 63-GHz ADC bandwidth; see measured eye-diagrams when using 33 and 63-GHz ADC bandwidths in Fig. 3.6(b-c). The receiver sensitivity, defined as the minimum P_{inRx} required to achieve a BER below the FEC limit, also improves from -3 to -4 dBm as we increase the bandwidth from 43 to 63 GHz. One could expect performance saturation when increasing the ADC bandwidth

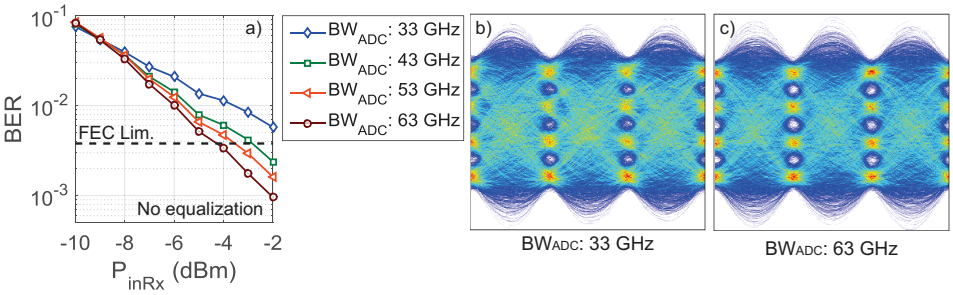


Figure 3.6: Performance in back-to-back configuration when using no equalization for different receiver bandwidths. BER versus receiver input power (a), and eye-diagrams when emulating 33-GHz (b) and 63-GHz (c) ADC bandwidths.

above the 40-GHz PIN-TIA 3-dB bandwidth. However, our calibration report of the PIN-TIA (not shown here) indicates that its frequency response does not decay drastically after 40 GHz, exhibiting 6-dB bandwidth above 50 GHz. As a result, BER is still improved when increasing the ADC bandwidth from 40 to 63 GHz. Still, the PIN-TIA remains the main bandwidth limitation and even better BERs can be expected when using PIN-TIAs with higher bandwidth.

One effective way to reduce the impact of the limited receiver bandwidth, while improving performance and receiver sensitivity is to introduce feed forward equalization at the receiver. Next we study the trade-off between receiver bandwidth and equalization complexity, while using the receiver sensitivity as performance metric. Each graph in Fig. 3.7(a-d) shows the BER versus receiver input power for ADC bandwidths ranging between 18 and 30 GHz for a certain number of equalizer taps: 3 (a), 7 (b), 15 (c) and 23 (d). As depicted in Fig. 3.7(a), the introduction of a low complexity equalizer with only 3 taps produces a great improvement when comparing to the non-equalized signal performances shown in Fig. 3.6(a). The simplest 3-taps equalization allows for a 40% decrease of ADC bandwidth requirement, which becomes 26 GHz (no-equalization: 40 GHz). Furthermore, with ADC bandwidth above 28 GHz, bit error rates below 10^{-4} can already be achieved. Fig. 3.7(e) depicts the measured receiver sensitivity as a function of the ADC bandwidth extracted from Figs. 3.7(a-d). For example, with 3-taps equalization the sensitivity is -5.5 dBm for a 26-GHz ADC bandwidth. Increasing the equalizer length to 7 taps offers 4 GHz extra ADC bandwidth reduction, and 2-2.5 dB better sensitivity. ADC bandwidth can be further reduced to 18 GHz, when using 15 taps; and sen-

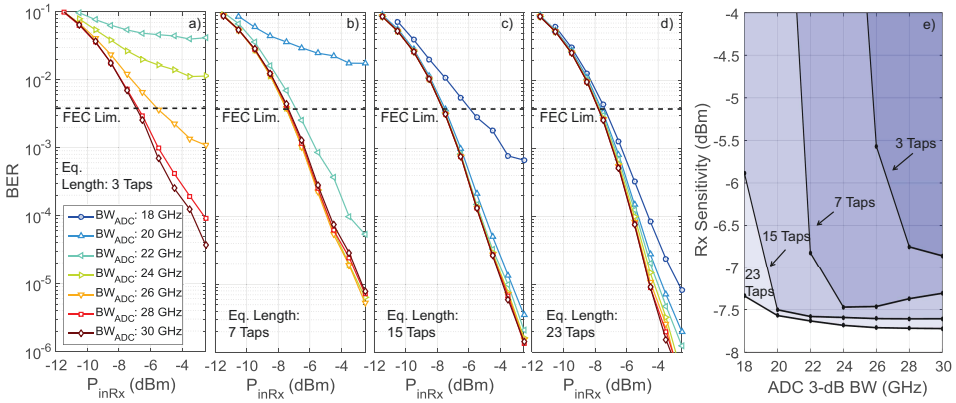


Figure 3.7: BER versus receiver input power for ADC bandwidths ranging between 18 and 30 GHz while applying feed forward equalization with (a) 3, (b) 7, (c) 15 and (d) 23 taps. (e) Receiver sensitivity as a function of ADC bandwidth (black lines) when using 3, 7, 15, and 23-taps equalizer. Areas above lines denote system operating range.

sitivity stabilizes between -7.3 and 7.7 dBm for all shown ADC bandwidths with a longer equalizer of 23 taps. These results clearly show the trade-off between receiver sensitivity, ADC bandwidth and required equalizer length. Moreover, increasing the number of taps also improves BER. Fig. 3.7(a-d) show that BER as low as 10^{-4}, 10^{-5} and 10^{-6} can even be achieved when using 3, 7 and 23-taps equalizers. Engineering the receiver for such low (pre-FEC) BERs can be interesting for low cost applications, because it alleviates the constraints on the FEC complexity while still guaranteeing error-free performance after FEC, i.e., a pre-FEC limit reduced to $2 \cdot 10^{-4}$ makes it possible to use lower complexity KP4 FEC [92, 129].

3.3.4 Transmission results

In this section we assess the performance of our 112-Gb/s transmitter after up to two kilometers propagation over standard single mode fiber. Fig. 3.8(a) shows the measured BER as a function of the input power into the receiver in back-to-back configuration and after 1.2 and 2 km transmission. Such measurement is taken under best equalization and ADC bandwidth conditions (23 taps and 30 GHz) shown in the previous subsection. We observe less than 1-dB sensitivity penalty with respect to BtB

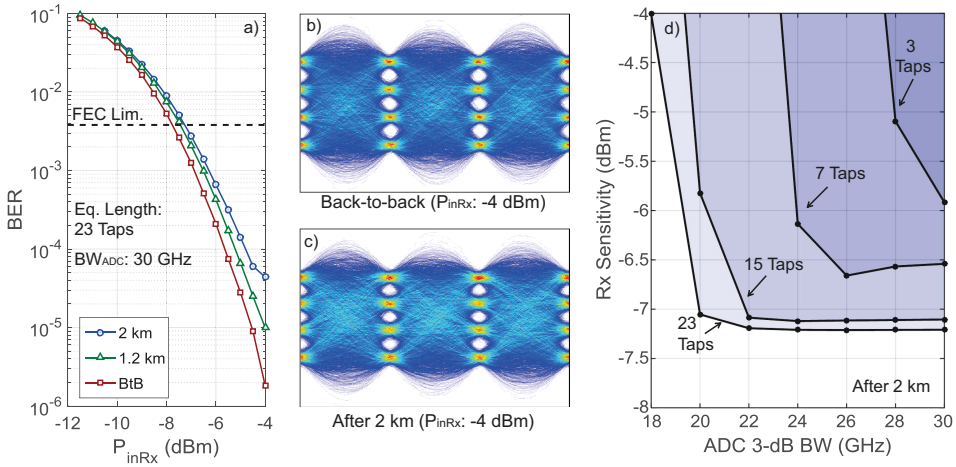


Figure 3.8: (a) BER versus receiver input power when using 30-GHz ADC bandwidth and 23-taps feed forward equalizer. (b) BtB and (c) 2-km eye-diagram constellations corresponding to $P_{inRx} = -4$ dBm in (a). (d) Receiver sensitivity as a function of ADC bandwidth (black lines) when using 3, 7, 15, and 23-taps FFE.

Along this section we have reported a low-cost compact 56-GBd PAM-4 transmitter including a DFB laser and high-speed EAM modulator exhibiting 50-GHz bandwidth and more than 13 dB extinction ratio. Such a module can successfully transmit 112 Gb/s over a distance of 2 km with relatively high residual margins of operation. Finally, an experimental study of the performance dependence on the receiver bandwidth and equalization length has revealed that our transmitter can be used within a large range of receiver conditions. Proper signal recovery can be achieved by trading sensitivity for low bandwidth (down to 18 GHz) or less complex digital signal processing (down to 3-taps feed forward equalization).

3.4 Beyond 100G: IM-DD ultra high-speed transceivers

As described in the previous section, many research teams have experimentally demonstrated 100-Gb/s IM-DD transceivers. However, not many have been able to surpass such data rate using single-wavelength and single-polarization approaches up to date. In 2015, Yamazaki et al. [63] reported a 160-Gb/s transmitter generating Nyquist PAM-4 signals at 80 GBd. In order to achieve such a high baud rate, they required a complex transmitter electrical setup comprising two digital-to-analog converters, a high-speed linear driver amplifier, and an analog multiplexer. Later, Kanazawa et al. [70] demonstrated in 2016 a 214-Gb/s (107-GBd) PAM-4 transmitter using again an electrical setup composed of multiple single components (high-speed multiplexers, attenuators and combiners). Both experimental demonstrations lacked a TIA, which is required to avoid extra optical amplification, but represents today a limiting factor in terms of bandwidth (40 to 50-GHz in high-end commercial TIAs) when transmitting such high baud rates.

We outline in this section our solutions to overcome the 100-Gb/s barrier using single-polarization single-wavelength IM-DD transceivers. We present two possible ways to increase capacity: 1) increasing the modulation order above PAM-4 and 2) increasing symbol rate beyong 56 GBd. In all the experiments we use the integrated SP-DAC to generate the electrical signals and off-the-shelf 40G electro-optical components (i.e., Mach-Zehnder modulator and PIN-TIA), see Section 3.4.1 for a detailed description of the

experimenta setup. First we demonstrate 168-Gb/s (150-Gb/s net data rate) optical transceiver using both approaches, by increasing the modulation format to PAM-8 while keeping 56-GBd baud rate [118], and by increasing the baud rate to 84 GBd, while keeping PAM-4 as a modulation format [120]. Both approaches are compared in Section 3.4.2. Then we achieve a symbol rate as high as 100 GBd using PAM-4 modulation, leading then to a 200-Gb/s (178.5-Gb/s net data rate) transceiver [120]. We describe such results later in Section 3.4.3.

3.4.1 Experimental setup: MZM-based transmitter

In order to further increase the modulation order and/or the symbol rate we had to change our transmitter. Despite the high bandwidth and small footprint of the previously used DFB-EAM, such device presents several drawbacks when working together with the SP-DAC:

1. Linearity: The DFB-EAM presents a certain degree of modulation non-linearity which cannot be totally compensated with our SP-DAC when generating a higher order modulation such as PAM-8. Hence, when using the DFB-EAM, the resultant 8 optical intensity levels are not perfectly equidistant, which induces performance penalties.
2. Single-drive: The DFB-EAM has a single driving input, which is not the optimum configuration for the SP-DAC. As will be shown below, a dual-drive modulator driven by both complementary outputs of the SP-DAC is preferred to enhance performance.
3. Chirp: EAM introduces a certain amount of chirp to the optical signals, which if it is not specifically designed to compensate a certain amount of chromatic dispersion, will further penalize system performance and reduce the overall reach.

Taking into account the above mentioned issues, we decided to use a commercial $LiNbO_3$ Mach-Zehnder modulator along the following experimental demonstrations. Such modulator was designed for 40G applications, and presents quite a large ≈30-GHz bandwidth with a slow decay-rate, showing 6-dB bandwidth above 50 GHz. Furthermore, such device presents a large linear modulation range, and no chirp. Finally, it can be

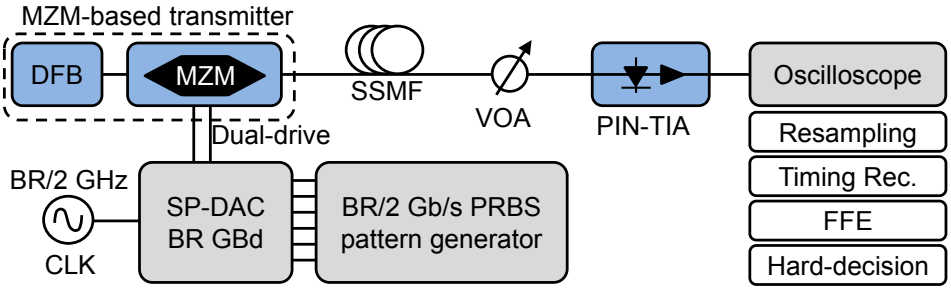


Figure 3.9: Experimental setup used for the optical transceviers generating data rates beyond 100 Gb/s.

driven in dual-mode. The main drawbacks of $LiNbO_3$ are their large footprint, which is nearly 10-cm long, and its difficulty to be integrated with other devices (e.g., lasers). However, as described in Chapter 2, integrable Mach-Zehnder modulators with small footprint are being fabricated today in novel platforms such as InP or Silicon/polymer, which recently start proving performances comparable to the more mature $LiNbO_3$ technology.

Our experimental setup is depicted in Fig. 3.9. When comparing to the setup used in the previous section, the main change lies in the transmitter side. We send light from a DFB laser at 1545 nm into a 30-GHz-bandwidth MZM, fed by the two complementary outputs of the SP-DAC. At the input of the SP-DAC, we inject up to six pseudo random bit sequences of $2^{15} - 1$ bits at half the output baud rate. Before entering the receiver, the signal is propagated through a section of SSMF with variable length. At the receiver side a VOA is used to vary the input power sent to a 40-GHz-bandwidth PIN-TIA. The received signal is then sampled by a digital storage oscilloscope and stored. We process the stored waveforms offline. First, we resample them to perform clock and timing recovery with T/2-spaced samples. Signals are then equalized using a T/2-spaced FFE. Finally, we perform hard-decision discrimination and count errors, to derive the BER.

The electrical setup used to generate and measure the electrical signals slightly changes for each experimental demonstration. Table 3.1 shows a detailed description of the parameters and settings used for each experiment described within this section. In order to validate our new setup, we first perform a PAM-4 56-GBd (112-Gb/s) experiment, as we did in the previous section, this time using the Mach-Zehnder modulator instead of the DFB-EAM transmitter. For such experiment, we used exactly the same electrical

Table 3.1: Electrical generation and detection parameters for all experiments of the current section.

In Sec.	Gb/s	Mod.	GBd	SP-DAC			Scope	
				Setup	$Data_{in}$ (Gb/s)	CLK_{in} (GHz)	GHz	GS/s
3.4.1	112	PAM-4	56	packaged	4×28	28	33	80
3.4.2	168	PAM-8	56	packaged	6×28	28	33	80
3.4.2	168	PAM-4	84	on-chip	4×42	42	62	160
3.4.3	200	PAM-4	100	on-chip	4×50	50	62	160

setup as described in Section 3.3.2. A fully-packaged SP-DAC module was fed with four 28-Gb/s pseudo-random bit sequence (PRBS) sequences of length $2^{15} - 1$ bits (two electrical inputs per encoding bit), together with a clock signal at also half the output data rate (i.e., 28 GHz). At the receiver site, a digital storage oscilloscope with 33-GHz bandwidth sampled the detected signals at 80 GS/s.

In order to characterize the performance of our MZM-based transmitter, we measured the BER versus input power into the receiver (P_{inRx}) for PAM-4 56-GBd signals in back-to-back (BtB) configuration, while varying the number of FFE taps. As depicted in Fig. 3.10(a), the 112-Gbit/s PAM-4 signal exhibits high performance, reaching BER below 10^{-6}, far from the limit of operation of the considered 7%-overhead FEC of $3.8 \cdot 10^{-3}$, see exemplary eye-diagram recovered after equalization in Fig. 3.10(c). The obtained receiver sensitivity, defined as the P_{inRx} yielding a BER at the FEC limit ($3.8 \cdot 10^{-3}$), is about -9 dBm for when using an equalizer with more than 11 taps. Already in BtB the current MZM-based transmitter leads to more than 1-dB sensitivity improvement when compared to the DFB-EAM integrated transmitter used in the previous section, which exhibited minimum sensitivity values of -7.7 dBm (see Fig. 3.7). Such improvement may be induced by the slightly better specifications of the MZM modulator (exhibiting high linearity and static extinction ratio above 20 dB). However, driving the modulator using both complementary SP-DAC outputs (dual-drive mode), also leads to better electrical performance (to be further explained below). As depicted in Fig. 3.10, a further decrease of the equalizer length induces signal deterioration. Nonetheless, successful reception can be achieved even when performing 3-tap equalization with a receiver sensitivity of -4 dBm.

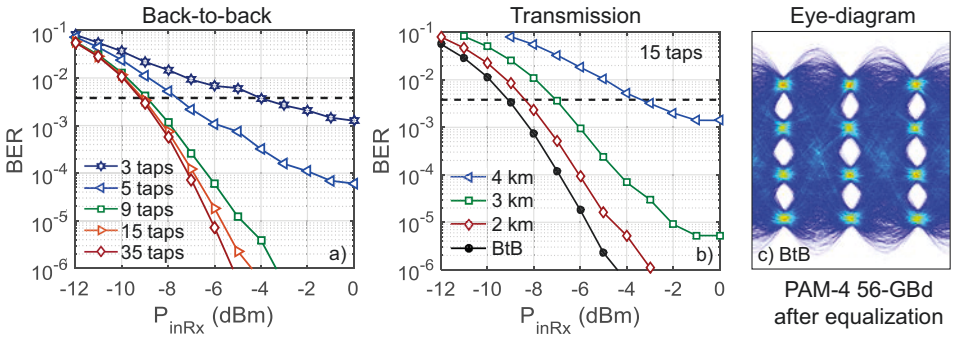


Figure 3.10: (a-b) BtB and transmission BER measured for a PAM-4 56-GBd signal as a function of the receiver input power (P_{inRx}) when using different equalizer lengths. (c) Exemplary PAM-4 56-GBd recovered eye-diagram after equalization when using 35-tap FFE in BtB ($P_{inRx} = -2$ dBm).

Transmission performance is evaluated in Fig. 3.10(b), when using 15-tap FFE, which guaranties lowest complexity while keeping performance close to the optimum. The 112-Gbit/s PAM-4 signal is impaired with a very small penalty after transmission over 2 km (no more than 0.6 dB compared to BtB), still allowing for BERs below 10^{-6}, and a sensitivity of 8.4 dBm. As expected, the performance degrades when transmission distance is increased as a result of chromatic dispersion. Nevertheless, we achieved 112-Gb/s PAM-4 transmission over 3.2 and 4 km, with receiver sensitivities of -7 and -3.5 dBm. Such results denote an extended reach by 2 km, when comparing to the transmission achieved when using the DFB-EAM transmitter in the previous section, which exhibited maximum reach of 2 km, see Fig. 3.8. The extended reach is enabled by the chirp-less operation of the MZM modulator.

Subsequently, in order to increase the data rate we either increase the modulation format (i.e., PAM-8) or the baud rate (i.e., 84 and 100 GBd). When increasing the modulation format to PAM-8, the same symbol rate was kept (i.e., 56 GBd), hence we simply injected two more electrical signals into the SP-DAC in order to generate the third encoding bit (six 28-Gb/s electrical signals in total). To do so, we added an extra pulse pattern generator (PPG) to generate the two new electrical SP-DAC input signals. However the rest of the setup remained unchanged. The eye-diagrams of the generated electrical PAM-4 and PAM-8 signals at 56 GBd, are shown in Fig. 3.11(a) and (b), respectively.

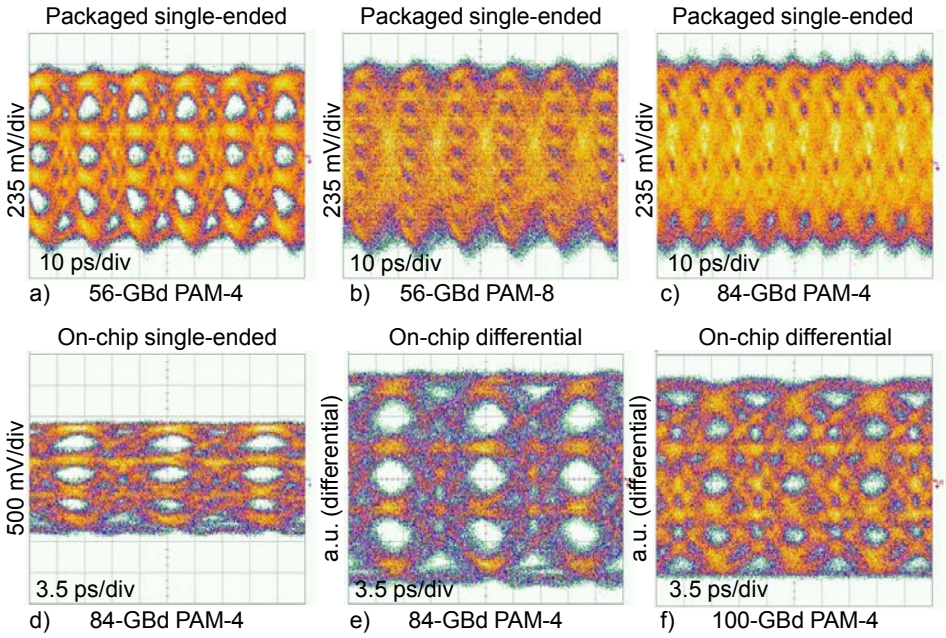


Figure 3.11: Eye-diagrams SP-DAC under different configurations. Single-ended measure of a packaged module generating: (a) 56-GBd PAM-4 and (b) PAM-8, and (c) 84-GBd PAM-4; on-chip measurements of the SP-DAC generating: (d) 84-GBd PAM-4 (single-ended measurement), (e and f) 84-GBd and 100-GBd PAM-4 (differential measurement).

On the other hand, increasing symbol rate to 84 and 100-GBd required a complete electrical setup upgrade. At the transmitter side, a state-of-the-art high-speed PPG was used to generate four 42-50 Gb/s electrical signals to feed the SP-DAC. The input clock was equally upgraded to 42 and 50 GHz to be able to generate the 84 and 100-GBd PAM-4 signals with the SP-DAC. Meanwhile, in the receiver site we used a 62-GHz bandwidth DSO capable of sampling at 160 GS/s. Despite the high-quality equipment used, the SP-DAC packaged module generated a closed eye-diagram even when working at 84 GBd, see Fig. 3.11(c). The main limitation was originated by the device packaging, which was not optimally designed to handle such high bandwidths. Hence, we decided to realize the experiments directly on-chip, using high-speed electrical probes to input and output the signals from the wafer (thus avoiding packaging limitations). Please note that such limitations can be overcome by optimizing the packaging or by co-packaging the SP-DAC with a modulator. Figs. 3.11(d-e) show single-ended (d) and

differential (e) eye-diagrams of the 84-GBd signals measured when the SP-DAC was manipulated on-chip. Differential eye-diagrams were obtained by subtracting both complementary SP-DAC outputs, after being simultaneously measured via two channels of a 70-GHz oscilloscope. Fig. 3.11(d), demonstrates that opened eye-diagrams can be achieved when avoiding the packaging; however, when working in single-ended mode, an asymmetry between top and bottom eyes is observed. This effect arises from different rising/falling times when working at very high symbol rates. As shown in Fig. 3.11(e), such effect can be compensated when working in differential mode, which exhibits the best performance. On-chip manipulation together with differential operation, allows generating symbols at baud rates as high as 100 GBd, see eye-diagram in Fig. 3.11(f).

In the following sections we show the experimental results achieving data rates above 100 Gb/s. First we compare in Section 3.4.2 the two above-mentioned approaches used to obtain 168-Gb/s data rates (PAM-8 56-GBd and PAM-4 84-GBd signaling). Finally we show in Section 3.4.3 the performance of our 200-Gb/s transmitter, obtained by transmitting PAM-4 signals at 100 GBd.

3.4.2 168 Gb/s IM-DD optical transceiver

Along this subsection we study the performance of our optical interface when transmitting 168 Gb/s. As mentioned below, we used two different methods to reach such data rate: PAM-8 56-GBd and PAM-4 84-GBd signaling. Fig. 3.12 shows a comparison between both methodologies. First, Figs. 3.12(a) and (b) depict the BER as a function of the receiver input power P_{inRx} for the PAM-8 56-GBd and PAM-4 84-GBd signals, respectively, in BtB configuration (a-b.1), after 1-km of SSMF (a-b.2) and beyond 1-km (a-b.3). All configurations are studied while using several equalizer lengths. The sensitivity is then plotted as a function of the equalizer length in Fig. 3.12(c) for all cases. Exemplary eye-diagrams of both approaches are depicted in Figs. 3.12(a-b.4).

PAM-4 84-GBd presents better overall performance than PAM-8 56-GBd. Depicted in Fig. 3.12(b.1), with PAM-4 at 84 GBd, we can achieve BERs close to 10^{-4} and sensitivities as low as -7.7 dBm when using a 35-tap equalizer; being such the maximum number of taps with which we obtain

significant performance gain with the equalization process. Decreasing the number of taps directly leads to performance degradation. The required minimum input power increases slowly when decreasing the equalizer length down to 9 taps, see Fig. 3.12(c). Below 9 taps, severe signal degradation occurs. Nevertheless, a 7-tap equalizer is still sufficient to successfully recover PAM-4 84-GBd signals, denoting a sensitivity of -5 dBm approximately.

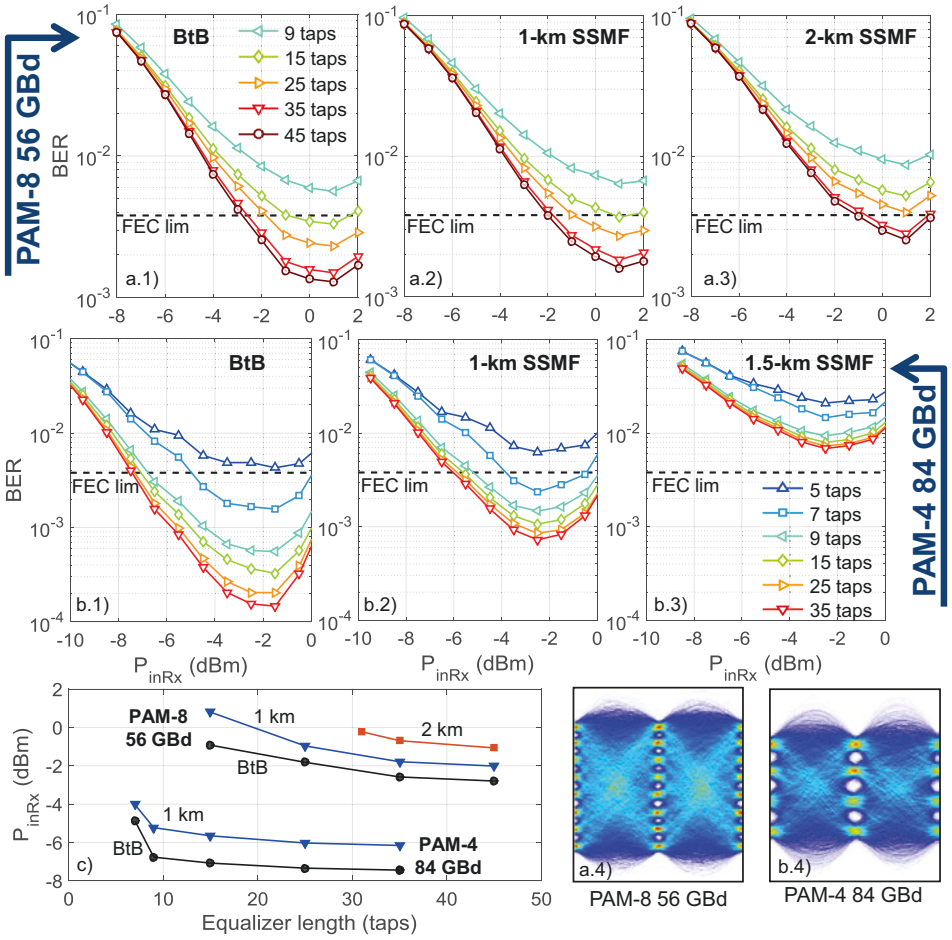


Figure 3.12: Performance evaluation (BER versus P_{inRx} for 168-Gb/s optical transceivers when using (a) PAM-8 56-GBd and (b) PAM-4 56-GBd signaling for the following cases: (a-b.1) BtB, (a-b.2) 1-km transmission, (a-b.3) and beyond 1-km. (c) Measured P_{inRx} as a function of the equalizer length for all cases shown in (a) and (b). (a-b.4) Exemplary eye-diagrams of the recovered signals after equalization in BtB.

Moving to PAM-8 56-GBd signaling causes more than 4-dB of sensitivity penalty w.r.t PAM-4 at 84-GBd, see Fig. 3.12(a.1), thus leading to a sensitivity of -3 dBm when using a FFE length of 45 taps (maximum equalizer length before gain saturation). The minimum equalizer length required to properly recover the PAM-8 56-GBd signal is 15 taps. Certainly such configuration requires 2-dB extra minimum power, see Fig. 3.12(c); which depicts the whole sensitivity trend.

Next we evaluate transmission performance of both approaches in Figs. 3.12(a-b.2-3). Overall, 1-dB and 2-dB extra penalties are observed for PAM-8 56-GBd and PAM-4 84-GBd signals, respectively. Both penalties arise from chromatic dispersion, which produces pulse-broadening in our signals, hence introducing a ISI that cannot be efficiently compensated by the equalizer in IM-DD schemes. As described in Chapter 2, the spread of the ISI (in number of symbols) increases quadratically with the baud rate. Thus, after 1-km transmission, PAM-4 84-GBd signals suffers twice the penalty than PAM-8 56-GBd. Also minimum BER further increases for the higher baud rate signal, which now gets closer to 10^{-3}. The required equalizer length does not change with respect to BtB in any of the modulation schemes; however, as mentioned above sensitivities are increased overall by 1 to 2 dB, see Fig. 3.12(c). Notwithstanding, after 1-km transmission PAM-4 84-GBd shows still lower sensitivity (4 dB) and lower BER than PAM-8 56-GBd. Such tendency changes when we further increase transmission distance. Fig. 3.12(b.3) clearly shows that PAM-4 84-GBd signals cannot be recovered after 1.5 km no matter the equalizer length, due to the critical impact of chromatic dispersion in such broad spectral signal. On the other hand, due to its relatively thinner spectrum, PAM-8 56-GBd signal can traverse up to 2-km transmission with just 1-dB extra sensitivity penalty.

This comparison shows the trade-off between performance, complexity and reach. PAM-4 84-GBd is the approach giving the best performance for up to 1-km links. On the other hand PAM-8 56-GBd signaling offers longer reach (up to 2-km distance). It is important to notice that the experiments were performed in the C-Band, at which chromatic dispersion is higher than at the O-Band when using standard single mode fibers. More than 10-km distance would be achieved working in the O-Band with PAM-4 84-GBd, with negligible penalty with respect to BtB. Nevertheless, in terms of electrical generation and reception PAM-8 56-GBd requires DACs and ADCs with higher number of bits, while PAM-4 84-GBd signaling

needs higher bandwidth and sampling rates. In what refers to equalizer length, it cannot be fairly compared with the results shown along this experiment, since different digital storage oscilloscopes were used for both approaches. All in all we have presented and experimentally validated two possible methods to achieve up to 168 Gb/s using an integrated SP-DAC and 40G off-the-shelf optical components.

3.4.3 200 Gb/s IM-DD optical transceiver

We evaluate in this subsection the performance of our 200-Gb/s IM-DD transceiver. In order to reach such high data rate we pushed further up the SP-DAC symbol rate to 100 GBd, while keeping the more robust PAM-4 modulation format. Along the following lines we first assess the performance of our transmitter while using the regular symbol-per-symbol hard-decision scheme, as in the previous sections. Then, in order to improve the performance we implement a N-symbol memory maximum-likelihood sequence detector (MLSD), to be described later.

The first set of results, obtained using the typical symbol-by-symbol hard-decision approach, is shown in Figs. 3.13(a-b.1) for BtB and 500-m transmission measurements. An exemplary eye-diagram of the received signal after equalization is depicted in Fig. 3.13(c). The limited bandwidth of the 40G optical components incurs severe signal degradation when generating 100-GBd signals, leading to approximately 3-dB of sensitivity penalty with respect to 84-GBd signals, refer to Fig. 3.12(b.1). For the best equalization length (35 taps) before complete gain saturation, we can denote a receiver sensitivity of -4.6 dBm and minimum BERs of the order of $2 \cdot 10^{-3}$. Furthermore, if we give up 1-dB extra sensitivity, 100-GBd PAM-4 signals can be recovered with a minimum equalizer length of 11 taps. As shown in Fig. 3.13(b.1), after 500 m the sensitivity is in general degraded by 2 dB due to chromatic dispersion pulse-broadening, drastically reducing the operation margins. A detailed analysis of the sensitivity as a function of the equalizer length can be observed in Figs. 3.13(a-b.4).

As aforementioned, the performance degradation of the 100-GBd PAM-4 is mainly caused the strong ISI stemming from the aggravation of low-pass filtering effects as the modulation bandwidth increases given the fixed

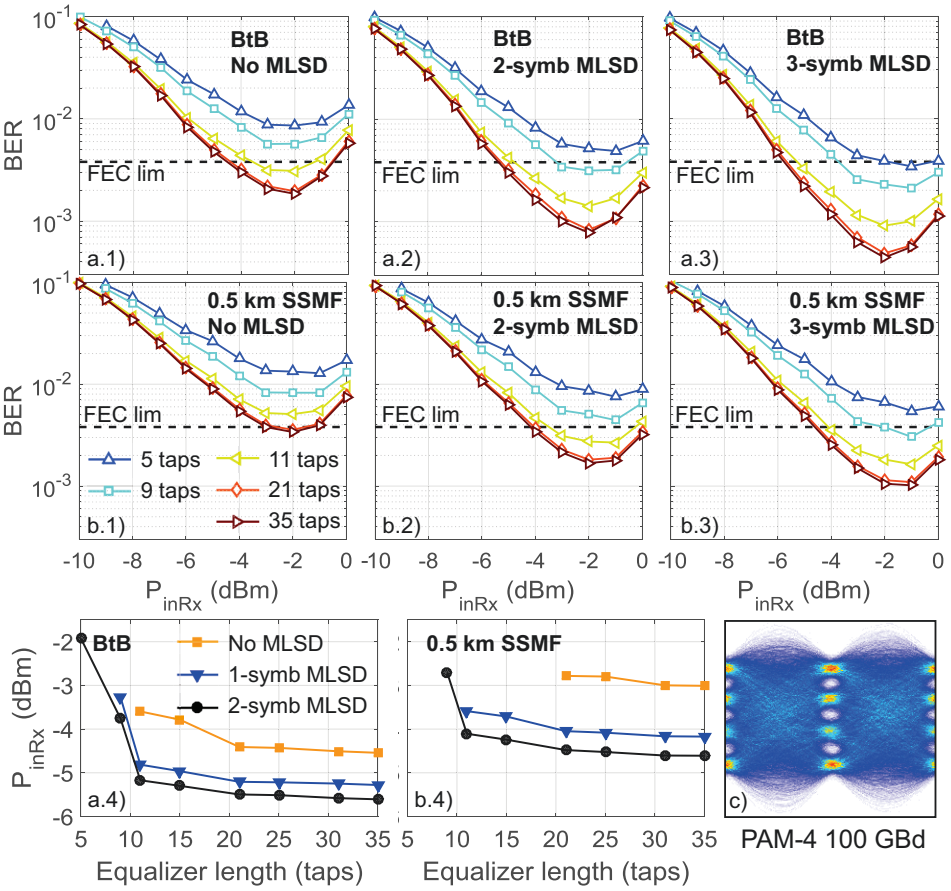


Figure 3.13: Performance evaluation (BER versus P_{inRx} for a 200-Gb/s PAM-4 100-GBd optical transceivers in (a) BtB configuration and (b) after 500 m when using different symbol decision schemes: (a-b.1) symbol-bt-symbol hard-decision, (a-b.2) 2-symbol and (a-b.3) 3-symbol MLSD . (a-b.4) Measured P_{inRx} as a function of the equalizer length for all cases shown in (a) and (b). (c) Exemplary eye-diagram of the recovered signals after equalization in BtB.

<40-GHz end-to-end system response[3]. In order to address this critical limitation without raising the equalizer complexity to unacceptable levels, we implement a 2-/3-symbol memory MLSD right after the regular FFE equalization stage. In this way, symbol detection becomes partially sensitive to time-domain correlations, which considerably increases the reliability of the demapper, even under imperfect equalization conditions.

[3]Given by the 35-GHz-bandwidth MZM and the 40-GHz-bandwidth PIN-TIA.

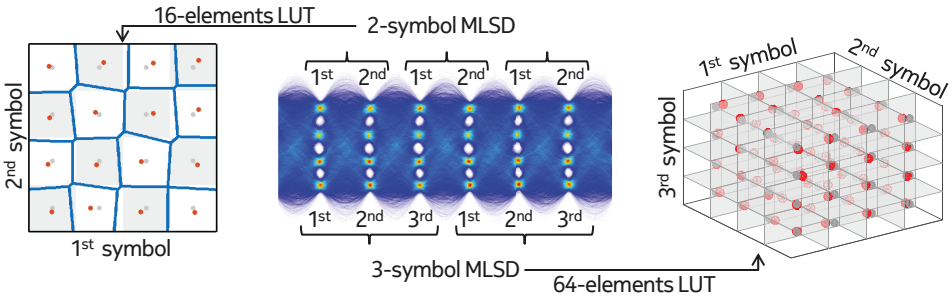


Figure 3.14: Visual description of MLSD process, showing a representation of calculated LUTs when using 2-symbols and 3-symbol MLSD.

In order to implement the MLSD, we initially train a static look-up-table (LUT) with 16 (4^2) and 64 (4^3) entries, standing for all possible PAM-4 class 2-symbol and 3-symbol combinations, respectively. Fig. 3.14 show visual representations of the 16- (left) and 64-elements (right) LUTs. As shown in the center of the figure, along the MLSD process the symbols are evaluated in pairs or trios. The 2-symbol LUT can be represented in a 2D-plane, by assigning the 1st and 2nd symbols of each pair to the x- and y-axis. This way, the LUT is equivalent to a 16-QAM constellation. Gray dots denote an ideal 16-elements constellation with no distortion, whose decision boundaries are represented as gray and white areas (such constellation is only used for visual comparison). On the other hand, red dots represent the 16-element LUT calculated to perform 2-symbol MLSD. Notice that the LUT accounts for time-invariant end-to-end signal distortions (e.g., ISI caused by band-limiting components), which shift each point from the ideal constellation. Blue lines denote the boundaries optimized to perform hard-decision taking into account the new reference constellation. Once the reference LUT is calculated, hard-decision takes place on symbol pairs (as opposed to symbol-by-symbol).

The 3-symbol MLSD is more complex since it requires a 64-elements LUT. Fig. 3.14(right) shows a visual representation of such a LUT, at which each symbol (1st, 2nd and 3rd) of each trio becomes an axis (x,y or z) of a 3D-space. Again the red dots represent the 64-elements LUT, which is plotted along with the ideal constellation (in gray). For better visual inspection we only plotted the boundaries of the ideal constellation. However, along the MLSD process, boundaries are adapted to the new constellation. In our experiments, we perform symbol decision by calculating the minimum

euclidean distance between the received pair (trio) and the reference LUT.

Figs. 3.13(a-b.2-4) shows the results obtained when using 2- and 3-symbol MLSD for both BtB and 500-m transmission. In back-to-back configuration, BERs below 10^{-3} are achieved for the first time using MLSD. 2-symbol MLSD allows for nearly 1-dB sensitivity gain, while 3-symbol MLSD contributes with extra 0.4-dB gain. After 500-m the MLSD gains are slightly higher than in BtB, see Figs. 3.13(a-b.4). This indicates that MLSD also accounts for a certain amount of signal distortions induced by chromatic dispersion. In our opinion, the small gain given by 3-symbol maximum likelyhood sequence detector (MLSD) with respect to the 2-symbol MLSD does not justify the high amount of extra complexity (decision between 64-elements compared to only 16-elements).

We have demonstrated a 100-GBd PAM-4 transmission over 500 m with up to 4-dB operation margin, obtained when using a feasible 2-symbol MLSD. Such detection scheme does not only improves performance but also allows reducing the FFE length to less than half the number of taps (11 taps) when comparing to the typical 1-symbol detection scheme (25 taps).

3.5 Summary

Along this chapter we have experimentally demonstrated several high-speed IM-DD transceivers capable of reaching the so-wanted 100-Gb/s per-lane capacity and beyond. In order to achieve such high data rates, we made use of advanced modulation formats such as PAM-4 and PAM-8 that allow doubling and tripling data rates of the more basic PAM-2 format, used today in data centers' optical interfaces. The very large baud-rate electrical signals are provided by a SP-DAC which accepts electrical inputs at half the output rate; hence keeping all high-speed electronic complexity in a compact integrated chip.

First we reported in Section 3.3 a compact EML transmitter, including a DFB laser and 50-GHz-bandwidth EAM. Such a module can successfully transmit 112 Gb/s over a distance of 2 km with relatively high residual margins of operation by using PAM-4 signals at 56 GBd. The large margins achieved by our transmitter allows working under a large range of receiver

configurations, supporting low-bandwidth receivers (down to 18 GHz) or low complexity equalization (down to 3 taps).

Later in Section 3.4 we have reported multiple solutions capable to overcome the 100-G barrier, all implemented with a MZM-based IM-DD transmitter. First we demonstrated 168-Gb/s data rate with two different approaches: PAM-4 84-GBd and PAM-8 56-GBd. The first approach (higher baud rate with lower modulation order) offers best performance in terms of BER and sensitivity, but with 1-km limited reach due to critical impact of chromatic dispersion. Conversely, the second approach (higher modulation order with lower baud rate), offers lower performance for distances below 1-km, but higher reach (up to 2-km transmission). Finally we have demonstrated 100-GBd PAM-4 transceiver emitting up to 200-Gb/s data rate over a distance of 500 m. A 2-symbol MLSD allowed detecting such extremely high-bandwidth signal with room for operation margins while employing commercially available 40G opto-electronic components, and less than 11 tap feed-forward equalization.

Chapter 4

Long-term solution: Burst optical slot switching ring-based datacenter

4.1 Introduction

Datacenters are becoming increasingly important and ubiquitous, ranging from large server farms dedicated to various tasks such as data processing, computing, data storage or the combination thereof, to small distributed server farms. As described in Chapter 1, main service providers host hundreds of thousands of servers to respond to the ever-growing traffic demand boosted today by novel cloud services. Servers communication is performed through EPS[1], which provides fast adaptability to traffic variations both in time and space, required due to highly dynamic inter-server traffic. Nevertheless, the EPS-based interconnection of such vast amount of machines requires an enormous network, including tens of thousands of electronic switches and millions of optical interfaces and cables, typically arranged in a Folded Clos topology, amply described in Chapter 1. Operating and managing such kind of networks is very challenging and costly. Furthermore, server-to-server communication typically requires traversing through multiple electronic switches, performing at each one conversion from opti-

[1] refer to Chapter 2 for a detailed explanation on EPS.

cal to electronic domain, electronic switching and re-conversion to optical domain. Such operations lead to large energy consumption and increased latency, refer to Chapter 1 for exemplary figures. In such architecture, traffic growth directly relates to an increase of networking components, including switches with higher port-count, interfaces and cables. Unsustainable cost that can be expected in the future together with the limited scalability of current architectures call for a network reassessment [44].

Cost and energy consumption can be reduced introducing optical transparency in the network, hence removing a substantial amount of opto-electronic (O/E) conversions. However, large datacenters can host so many networks elements that their full transparent interconnection by optical circuits is impractical, or impossible, owing to the finite wavelength count across the typical optical spectral width. Along the last few years, many novel schemes introducing partial or full optical transparency in datacenters networks have been reported.

A first group of proposals introduce OCS in traditional EPS networks. Such architectures make use of optical circuit switches, typically based on micro-electronic-mechanical systems (MEMS) technology to perform inter-rack high-bandwidth semi-static communication. These long-live interconnections can be reconfigured in the millisecond scale, and are used to alleviate the electronic switching fabric from long-steady inter-rack traffic, allowing for a reduction of the number of electronic switches and cables [130,131]. The main drawback of these solutions is the slow reconfiguration of current optical circuit switches, which cannot rapidly adapt to the fast traffic dynamics of current datacenters.

The second group of alternatives relies on the flexibility and granularity of optical packet switching, at which server racks are interconnected through a fully (or mostly) transparent optical network performing fast optical packet routing. Over the last few years many proposals have been revealed while using different kind of optical technologies and topologies to perform data center connectivity. Some of the proposals present optical interconnects based on array waveguide grating router (AWGR) and tunable wavelength converter (TWC). The main problem of such architecture is the limited scalability which depend on the number of wavelengths supported by the AWGR [132–134]. To increase the number of supported nodes (or server racks) the interconnects can be arranged typically in Clos topologies, being able then to leverage 100,000 servers [134]. Some of these

approaches require the optical use of shared electronic buffers to manage congestion when working under high-capacity loads, which leads to extra opto-electronic conversions. Another approach performs flat inter-cluster connectivity through both OPS and OCS interconnects [135]. In such proposal, the OPS node performs in-band optical label packet processing to manage routing and contention. Low priority packets are blocked while the high priority ones are forwarded. The fast optical switching is made by SOAs in broadcast and select switch connected to arrayed-waveguide grating (AWG)s, for wavelength multiplexing/routing. Yet another alternative solution proposes a flat torus topology at which servers racks exchange 4×lambda WDM packets using hybrid opto-electronic packet routers [59]. As in the previous case in-band labels are used for routing purposes. Despite the high scalability and resiliency to failure of torus topology, several processes need to be used to avoid contentions: deflection of packets through other routes, fiber delay lines and even shared electronic buffers. Most of the abovementioned OPS-based approaches, require the use of different processes to avoid contentions. Even so, in order to stay away from drastic packet losses, usually they need to work under relatively low capacity loads with maxima close to 50% [59, 60].

Along this chapter we describe our novel approach for intra-datacenter networking which is based on BOSS rings. Different from typical OPS systems, in BOSS rings, nodes exchange optical packets transparently in a synchronized manner using fixed-duration time slots multiplexed also in the wavelength domain. Slot synchronization avoids packet collisions, hence allowing for high capacity loads, close to 100% [136]. The use of BOSS rings was first proposed in metropolitan area networks to provide flexible optical transport with high-capacity and sub-wavelength granularity. In the context of datacenter intra-connection our team proposed a flat architecture to interconnect server racks through a mesh of BOSS rings organized in a torus topology, which provides high scalability and resiliency to failures. Such novel topology and its implementation are first discussed in Section 4.2. Later we evaluate the physical impairments occurring when the optical signals traverse a large cascade of BOSS nodes. First we study the accumulated nonlinear distortions induced by SOA-based optical gates in Section 4.3; and then we assess the frequency filtering produced when traversing many wavelength de/multiplexers in Section 4.4. We finally propose transponder designs that can adapt to such impairments and evaluate their performance in terms of reach and capacity.

4.2 BOSS ring-based intra-datacenter network: The concept

As described in Chapter 1, datacenters typically consist of end-host servers set within racks connected through a so called Folded Clos switching fabric, see Fig. 4.1(a). In such topology several switching stages provide full connectivity along server racks. To do so, O/E/O conversions take place at each switch in order to re-route the traffic in the electronic domain before being reconverted to the optical domain. Such process leads to several O/E/O conversions to perform rack-to-rack communication.

4.2.1 BOSS ring-based torus topology

We proposed in [137, 138] a novel topology for datacenter intraconnection based on burst optical slot switching (BOSS) rings. As shown in Fig. 4.1(b), in such architecture we remove the whole hierarchical electronic switching fabric observed in Fig. 4.1(a), and replace it with a flattened BOSS ring-based torus network, which is more adequate to prevalent East-West connectivity containing 75% of the whole datacenter traffic [6]. In this novel architecture we connect servers to BOSS nodes, which are interconnected through fiber rings forming a torus topology. Along each ring, data travels transparently (i.e., without O/E/O conversions) through all intermediate nodes (from source to destination), being encapsulated into wavelength- and time-multiplexed slots of fixed few-µs duration. In general, wavelength/time slots are not dedicated to any node in particular, but shared between all nodes and used upon request. This way we can flexibly adapt to fast varying datacenter traffic avoiding possible capacity waste, which may happen when pre-assigning resources.

For each fiber ring, slot-transmission is made synchronously for all wavelengths. Note that the synchronization of each fiber ring is independent from the others. Hence, inter-ring connection is performed electronically via the BOSS nodes. One needs to realize that full optical end-to-end connectivity is very challenging or leads to relatively poor capacity loads. On the one hand, full slot synchronization and scheduling over the many rings of a whole datacenter is impractical to manage. On the other hand, full asynchronous optical transmission requires the use of optical buffers

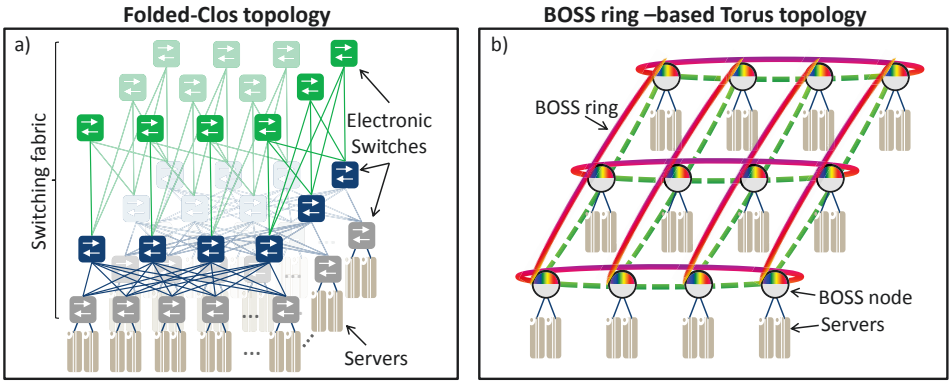


Figure 4.1: (a) Traditional Folded-Clos and (b) proposed BOSS ring-based data center topologies.

to partially avoid collisions. Furthermore, even when using optical buffers such networks are usually limited in terms of capacity load (e.g., less than 50%) [60]. Our approach ensures full inter-node datacenter connectivity with at most one intermediate O/E/O conversion (which takes place when changing rings), with respect to the three required (at most) in typical Folded Clos topologies (ToR-to-ToR, only accounting for intermediate O/E/O conversions taking place at the two higher levels of the switching fabric). In addition inner-ring synchronized slot-transmission avoids collisions, which leads to an enhanced capacity load of more than 90% [136], allowing for better resource optimization and hence diminishing the number of required network elements.

Along the following lines we list some of the benefits offered by BOSS ring-based torus networks:

- Flattened topology: we create a non-hierarchical architecture better suited to the prevalent East-West connectivity containing 75% of the whole datacenter traffic [6].
- Pay-as-you-grow architecture: Overall datacenter capacity can be easily upgraded by adding new rings to a previously deployed network. BOSS torus topology presents very high scalability, being capable of hosting millions of servers with low oversuscription ratios [138].
- Failure resiliency: Path diversity provided by the torus topology allows redirecting and distributing the traffic through other possible

paths in case of failure. It also allows isolating the affected area during a repair or an upgrade.

- O/E/O conversions reduction: Once packets are inserted in the BOSS ring, at most one O/E/O conversion is required to reach any other node in the datacenter. Folded Clos topology requires (at most) 3 O/E/O conversions for full connectivity between any ToR. Reducing O/E/O conversions means avoiding (de)encapsulation and queuing processes, which leads to lower end-to-end latency. The energy consumption of the switching fabric (above ToR) is also reduced by a factor $\times$2-3, when compared to traditional Folded Clos topologies [138].

- Networking elements reduction: The combination of reducing switching stages, allowing transparency and using high-data rate transponders allows for large reductions in networking elements. Between $\times$100 and $\times$500 fewer interfaces and cables are required, depending on the datacenter size and oversubscription ratio, when compared to traditional Folded Clos topology [138].

4.2.2 BOSS node functionality

As shown in Fig. 4.2, a BOSS node typically handles traffic traversing through several rings. Through each ring multiple wavelength data channels (colored packets) propagate along with a control channel (gray stream), which transports packets headers (including for instance source and destination nodes) and scheduling and networking management information. The control channel can be transmitted within the same fiber as the data channels, but through a dedicated wavelength or band. For instance, the O-Band can be used for control, while data travels along the C-Band. This way, only band (de)multiplexers are required in order to (drop)add the control channel, see Fig. 4.2. Contrarily to data channels, the control channel is detected and re-transmitted at each node. Note that a low-cost interface (e.g., 1-10 Gb/s) is sufficient to support the control channel traffic. Control is processed by the electronic board (in gray), which also handles inter-ring and servers' traffic; thus performing all layer 2 tasks such as packet buffering, (de)assembling and scheduling, accordingly with the information received through the control channel. Layer 2 design and implementation are out of the scope of this thesis.

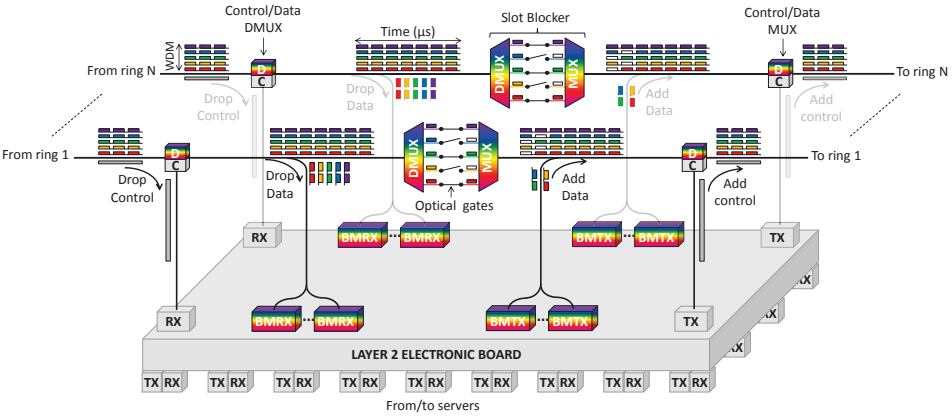


Figure 4.2: BOSS node schematic.

In what refers to the data channels, any optical technology (e.g., non-coherent or coherent) could be used by properly adapting the hardware in the BOSS node. However, we chose coherent technology to maximize the achievable data rate per wavelength. Coherent technology is extensively used today in long-haul systems due to its extremely high capacity. By using amplitude, phase and polarization light-dimensions, such technology is able to quadruple the spectral efficiency of IM-DD systems. Using high capacity interfaces allows reducing the number of transponders required per node. Furthermore, optimizing spectral efficiency increases the overall fiber capacity. Hence, the number of fibers required to transport inter-nodal traffic can be also reduced. Moreover a tunable laser can be used for colorless coherent detection [93], avoiding the need of optical filters (required in IM-DD systems when transmitting multiple wavelength channels). In coherent detection, wavelength channel selection is performed by tuning the local oscillator to the wavelength of the channel under interest. This way, the limited bandwidth of receiver's ADCs filters out the high-frequency beating produced between the local oscillator and neighboring channels.

Today, coherent systems are still expensive and hungry-consuming. Notwithstanding, leading transceiver manufacturers are putting lots of efforts in producing affordable coherent transponders in order to introduce such technology to more cost-sensitive sectors such as metropolitan or inter-datacenter networks [139,140]. In the near future, Silicon platform will most probably allow the co-integration of photonics and electronics, which will lead to huge savings in cost, footprint and power consumption [141].

BOSS nodes include three main hardware elements: burst-mode (BM) receiver (RX), slot blocker, and BM transmitter (TX). When entering the node, all data channels traverse a power splitter, which will spread all wavelengths towards the burst-mode receivers and the slot blocker. Using a splitter instead of a dropping device allows for transparent broadcast and/or multicast operations in the rings.

BM-RXs use fast tunable lasers to select through coherent detection the wavelength under interest on a per-slot basis. Fast laser tunability has been already demonstrated in Bell Labs while employing DS-DBR lasers driven by fast switching electronic boards made in-house. Such lasers exhibited tunning times lower than 100 ns, frequency offset of hundreds of MHz and linewidth inferior to 1 MHz, which is compatible with coherent technology [93]. Additionally, BM-RXs make use of fast adaptive algorithms capable of recovering the short (few microseconds) packets (to be explained in the following sections). Such packets are separated by guard intervals (≈100 ns long), allowing stabilization of devices (mainly lasers) after switching occurs.

The slot blocker has two main functions: 1) power channel equalization, to avoid loss/gain accumulation of particular channels when traversing the long node cascade; and 2) slot blocking, in order to erase wavelength-slots that have been already received for its latter reuse. One possible way of implementing such device is depicted in Fig. 4.2. In such scheme, a wavelength demultiplexer (DMUX) is used first to separate all wavelength channels. Then, each wavelength goes through an optical gate, which lets pass or erases certain slots. Wavelength slots containing "receive" packets are only erased if they can be immediately filled by a BM-TX placed in the same node; otherwise such "received" packets will continue propagating as "dummy" packets (considered as available slots). This way all wavelengths are continuously loaded (with no long power gaps), hence obtaining a circuit-like optical transmission and avoiding the use of specially designed burst-mode optical amplifiers. Finally a wavelength multiplexer (MUX) combines all channels.

Before leaving the node, BM-TXs insert novel data packets into the available (scheduled) emptied slots. Similarly to the dropping process, BM-TX data-packets are added through an optical splitter, hence allowing for colorless operation when using a fast-tunable laser.

BOSS ring-based networks can physically operate in several ways. Fig. 4.3 describes the two most differentiating options to transmit the data packets between nodes placed within the same ring. In Option 1, shown in Fig. 4.3(a), both transmitter and receiver contain fast tunable lasers, able to change wavelength in a slot basis. Furthermore, the slot blocker also includes fast optical gates capable to erase single packets. This way slots containing "received" or "dummy" packets can be emptied for its immediate reuse, which optimizes overall network capacity [142]. For such operation mode optical gates need to exhibit the following characteristics: 1) fast switching time to be able to switch between on and off states during the guard interval (typically less than 100 ns); 2) high extinction ratio (typically more than 20 dB), in order to avoid crosstalk induced from erased packets over new added packets; 3) negligible distortion (e.g., nonlinear distortions), which may limit performance when being accumulated after traversing over many nodes. Optical gates can be implemented with MZM [143], ring resonator structures [144],VOA based on p-i-n carrier-

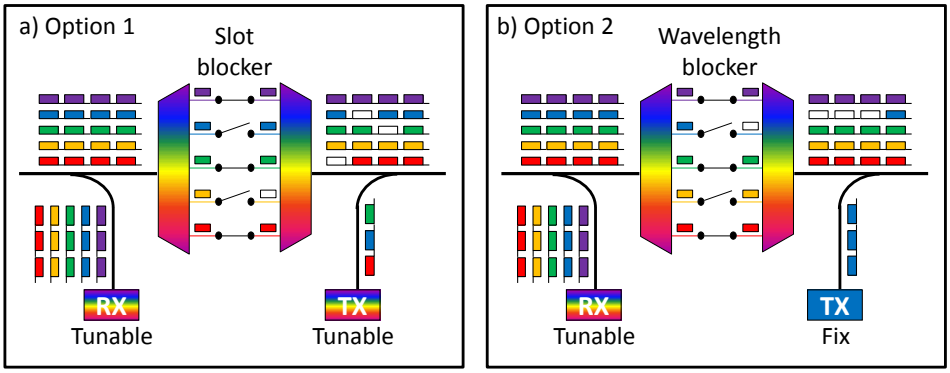


Figure 4.3: Operation options in BOSS networks.

select the pertinent slots. By dedicating a wavelength to each transmitter, each transmitted slot completes a ring round-trip before being blocked at its origin node. Hence, a fast slot blocker is not required with this operation mode. As shown in Fig. 4.3(b), we can now use a "wavelength blocker", which statically suppresses one ore several wavelengths, corresponding to those assigned to the transmitters placed in the same node. Note that the BM-TX and blocker do not require fast wavelength switching when using this option; nonetheless wavelength tunability (in a millisecond scale as in commercial systems) is desired to perform possible network reconfiguration. The wavelength blocker might be implemented with a WSS-like device or a lower-cost integrated slot blocker, described above.

It is evident that Option 2 leads to fiber capacity waste, which takes place when data slots complete the ring round-trip after being received. Comparing both approaches we can observe that Option 1 can reach capacity loads (≈95%) similar to Option 2 while using 35% fewer wavelengths, when assuming a uniform spatially distributed load [136]. However, from the practical point of view, the second option is much simpler and immediate to implement because both transmitter and wavelength blocker work in circuit-like mode. Hence commercial devices employed today in optical networks can be used for such elements. Furthermore, contrarily to Option 1, in which guard intervals are empty gaps to avoid possible inter-wavelength interference occurring when tuning the transmitter, in Option 2 guard intervals can be filled with special signals designed to ease data-packet recovery taking place at BM-RXs.

The decision of using one of both approaches depends on the required capacity and on the number of available wavelengths in the ring. Commercially deployed systems work typically with 50 or 100-GHz wavelength grids, which support usually 80 or 40 wavelengths, respectively in the C-Band. Hence, if the number of wavelengths is large enough to cover all transmitters placed in a ring, Option 1 does not lead to any gain in terms of capacity load. For instance, if we implement a BOSS ring including 40 nodes, including two transponders per node, a wavelength can be assigned to each transmitter (80 in total) without the need of wavelength sharing when using a 50-GHz grid. Nonetheless, if wavelength count is not sufficient to support the demand of all nodes, using Option 2 implies the addition of further rings with the corresponding blocking devices, transponders, etc. In that case, thanks to its optimized capacity, Option 1 allows reducing the number of network elements, hence diminishing cost.

The technology used when building the wavelength/slot blockers strongly impacts the overall capacity and node-count supported in BOSS rings. One needs to realize that the scalability (in terms of overall data-center capacity) of our BOSS solution strongly depends on the per-node capacity and on the number of nodes that can be placed into a single ring. Rings including a large number of high-capacity nodes (e.g., 50 to 100) are required to achieve large scalability. Hence, distinctly from long-haul networks, in which the reach is typically limited by the distance, in BOSS data center networks, the main limitation stems from the accumulation of distortions induced in large node cascades. Therefore it is important to identify node-related impairments, propose adequate modulation schemes and assess their reach (node-count) and capacity. Both D/MUXs and optical gates add distortions that, accumulated along a large node-cascade, imply severe signal degradation, limiting node-count, transponder capacity and wavelength count. In Section 4.3, we analyze the cascadability of SOA to be used as optical gates. We will observe that such devices limits the node-count due to the accumulation of nonlinear distortions when traversing many times through them. Later, in Section 4.4 we will study impact of traversing a large cascade of D/MUX while using different type of technologies (i.e., WSS and AWG) and accordingly we propose different transponder designs to enhance capacity and reach.

4.3 Node cascadability: Impact of SOA-based optical gates

Various devices have been proposed as key building blocks for fast optical gates, such as MZM [143], ring resonator structures [144], or VOA based on p-i-n carrier-injection structures [145]. However, optical gate specifications such as high extinction ratios above 20 dB and switching times between pass and block states below 30 ns are typically required. MZM or ring resonators exhibit relatively low extinction ratios (< 20 dB) [143,144], like the previous type of VOA when utilized for short switching time (less than 15 ns) [145]. Active devices such as SOAs are expected to offer high extinction ratios (up to 40 dB) and fast switching times along with a reduction of the insertion loss (due to their intrinsic amplification), thus avoiding the use of extra optical amplifiers. However, the SOA limits the signal reach due to the OSNR degradation by amplified spontaneous emission (ASE) and nonlinear distortions (mainly self-phase modulation) in the passing signal.

In this section, we experimentally evaluate the cascadability of SOAs while using advanced modulation formats QPSK and N-QAM. The capability of SOAs to amplify complex modulation formats has been numerically and experimentally investigated for one device [147] and for a cascade of few devices (up to 4) [148]. However, no large number of SOAs has been experimentally cascaded. Here, we establish a recirculating loop including one SOA-based optical gate and study the evolution of linear and nonlinear noise along a large cascade of SOAs. We evaluate the performance of several modulation formats and calculate the operation limits that guarantee the requirement of different FEC schemes for polarization division multiplexed QPSK, 8- and 16-QAM signals at 28 GBd [149].

4.3.1 Experimental setup

Fig. 4.4(a) shows the experimental setup used to evaluate a large cascade of SOAs. For such experiment we modulated the light of an external cavity laser using a dual-polarization I/Q modulator. Four digital-to-analog converters were used to drive the modulator, generating this way PDM signals at 28 GBd with diverse modulation formats: QPSK, 8- and 16-QAM. All signals are based on pseudo-random binary sequences of length $2^{15}-1$ bits. After optical amplification the light enters into a recirculating loop emulating the propagation through many optical nodes. A slot blocker is built with two liquid-crystal on Silicon (LCoS)-based wavelength selective switches working as wavelength (de)multiplexer and a commercial SOA as optical gate and amplifier with 11 dB optical gain (we set a 200-GHz grid to neglect filtering effects and focus on SOA impairments). Furthermore, we use a VOA to evaluate the performance for different power configurations at the input of the SOA, a 25-km span of SSMF is used to have stable loop operation and an erbium-doped fiber amplifier to make up for the extra losses of the recirculating loop. Finally we detect the signals using a coherent mixer and four pairs of balanced photodiodes connected to a 20-GHz bandwidth oscilloscope working at 40 GS/s. All measured waveforms were processed offline using standard DSP techniques detailed in Chapter 2.

The two switches observed in the setup work in a complementary manner and allow initializing the loop count. First, we close SW_{TX}, allowing the signals generated at the transmitter to enter into the loop. We call this operation "loop loading". Notice that when SW_{loop} is opened light does

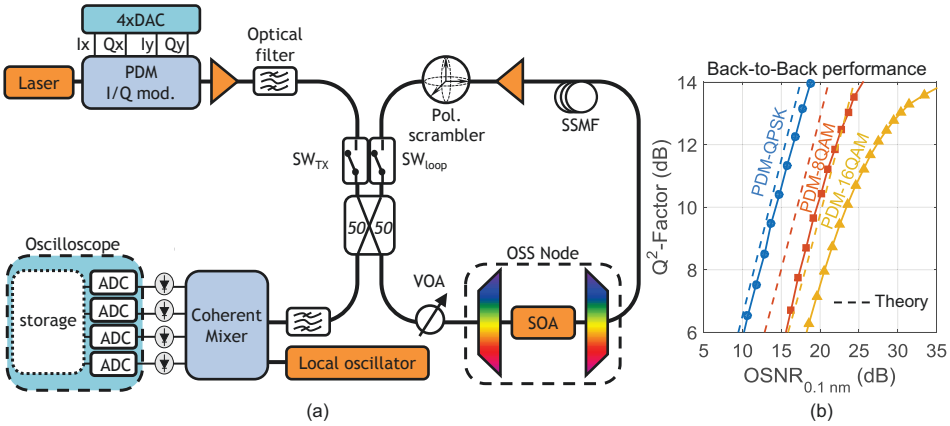


Figure 4.4: (a) Experimental setup to evaluate SOA cascadability. (b) BtB performance of our transmitter when using different modulation formats.

not recirculate. Once the loop is loaded, both switches change position, entering then into the looping mode, at which light from the transmitter is now blocked (SW_{TX} is open), and light within the loop recirculates over and over until we settle back in loop loading mode. Notice that the 50:50 coupler allows the receiver detecting the light at any moment, both during the loading (equivalent to BtB) and looping operations. A set of digital control signals (not shown in the setup for ease of visualization) are specially synchronized to trigger the switches and the oscilloscope. The DSO trigger needs to be calibrated with the roundtrip time of the loop in order to differentiate and capture the different passes through the loop.

The performance of our transponder in BtB configuration is depicted in Fig. 4.4(b), for the three modulation formats under evaluation. In such graph the Q^2-factor is calculated from measured BER as follows [150]:

$$Q^2 = 20\log_{10}(\sqrt{2}\cdot \mathrm{erfc}^{-1}(2\cdot BER)), \tag{4.1}$$

and then plotted as a function of the OSNR, calculated with 0.1-nm of resolution. As depicted, PDM-QPSK achieves performances very close to theory, denoting less than 1-dB of OSNR penalty. On the other hand, transmitter and receiver imperfections in terms of effective number of bits and bandwidth lead to further penalties when generating higher order modulation formats such as 8 or 16-QAM, which present OSNR penalties in the order of 2.5 dB with respect to theory.

4.3.2 SOA-cascade: Performance and noise analysis

Fig. 4.5 shows the performance of the different modulation formats when traversing the SOA cascade. In such graphs, each curve represents the Q^2-factor as a function of the power inserted into the SOA (P_{inSOA}) after an increasing number of passes through the device (N×SOA, see color scale and boxed numbering). We used a large range of P_{inSOA} (-24 to 0 dBm) and cascaded up to 50 times the SOA, leading to enough granularity to cover all possible configurations.

Each curve in Fig. 4.5 shows a maximum performance when setting an optimal input power of $P_{inSOA} \approx -18$ dBm (P_{opt}) for most cases, corresponding to the P_{inSOA} at which the SOA gain starts saturating. This input optical power separates the linear from the nonlinear regime, see exemplary constellations in Fig. 4.6 for visual assistance. On the one hand, in the linear regime ($P_{inSOA} < P_{opt}$), signals are mainly distorted due to ASE noise. In such regime lowering the input power means degrading OSNR, which is reflected by the isotropic growth of the noise observed in the constellations, see Fig. 4.6 ("Linear regime"). The increased noise produces the penalty in Q^2-factor observed in Fig. 4.5. On the other hand, in the nonlinear regime ($P_{inSOA} > P_{opt}$), distortions come mainly from the nonlinear phase noise induced by the SOA under saturation; which explains the non-isotropic evolution of the constellations, see Fig. 4.6 ("Nonlinear regime"). The further we increase the input power into the SOA, the higher the penalties due to nonlinear distortions. Of course the cascade results into an accumu-

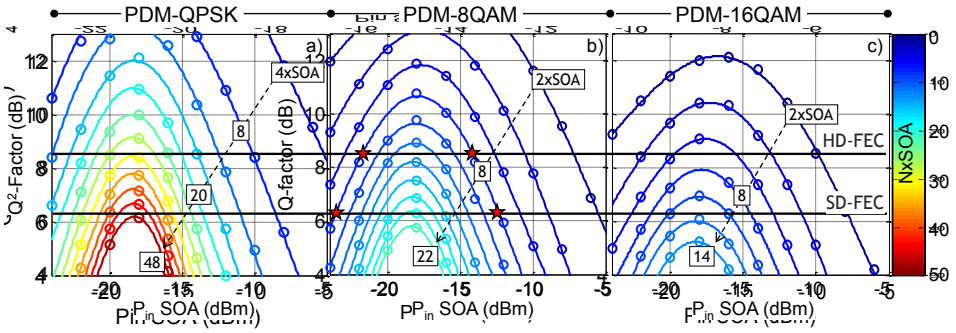


Figure 4.5: Q^2-factor versus SOA input power (P_{inSOA}) evaluated along the SOA cascade for all modulation formats: (a) PDM-QPSK, (b) PDM-8-QAM and (c) PDM-16-QAM.

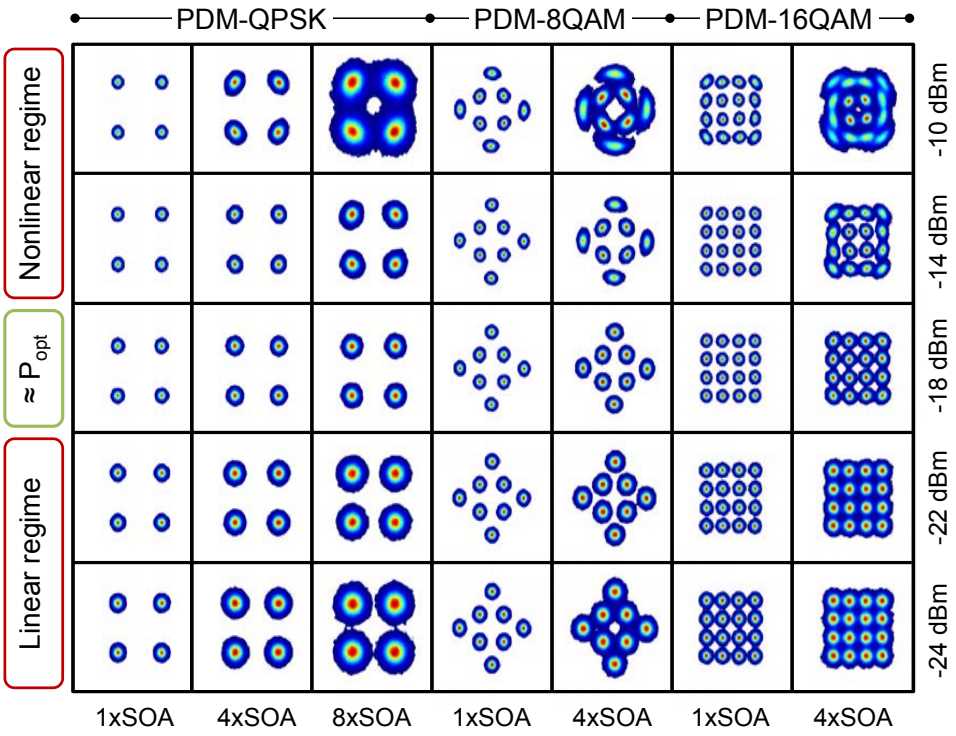


Figure 4.6: Exemplary recovered constellations for all modulation formats under different SOA input power conditions and after traverssing different number of devices.

lation of single-pass effects for both distortions (ASE and nonlinear noise), which further penalize signals, see Q^2-factor degradation along the cascade in Fig. 4.5.

In order to better understand such results we analyze the evolution of both phase and amplitude noise along the cascade of SOAs for all modulation formats. To do so we calculate the phase and amplitude variances (σ^2_{ph} and σ^2_{amp}, respectively) for several P_{inSOA}. Fig. 4.7(a) shows both σ^2_{ph} and σ^2_{amp} as a function of the number of cascaded nodes for a PDM-QPSK signal. Both variances increase as a function of the traversed SOAs but presenting different slopes, which depend on P_{inSOA}. The minimum slope is found for $P_{inSOA} = -18$ dBm ($\approx P_{opt}$), thus the optimum performance. That curve indicates that at the optimum input power, we start appreciating the nonlinear behavior: σ^2_{ph} is slightly higher than σ^2_{amp}. Such difference

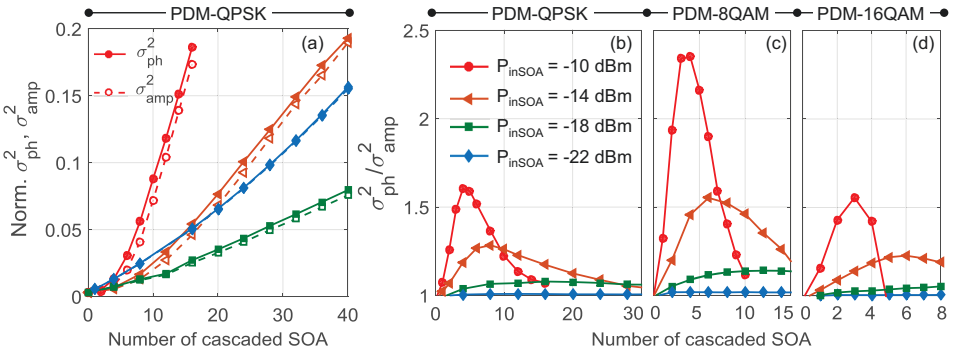


Figure 4.7: (a) Normalized amplitude and phase variances for the PDM-QPSK signal, and (b-d) phase-amplitude variance ratio for all formats along the SOA cascade.

increases while increasing the input power into the SOA as expected. On the contrary, at $P_{inSOA} = -22$ dBm (linear regime), both variances are superimposed, indicating perfect isotropic constellations.

To be able to observe in more detail the different evolution of σ^2_{ph} and σ^2_{amp}, we plot in Figs. 4.7(b, c and d) the ratio $\sigma^2_{ph}/\sigma^2_{amp}$ for the same input powers as in Fig. 4.7(a) for all modulation formats. All modulation formats, denote $\sigma^2_{ph}/\sigma^2_{amp} \approx 1$ at $P_{inSOA} = -22$ dBm, confirming the isotropic constellation along the whole cascade. As perceived in Fig. 4.7(a), $\sigma^2_{ph}/\sigma^2_{amp}$ increases when increasing P_{inSOA}: the higher P_{inSOA}, the faster $\sigma^2_{ph}/\sigma^2_{amp}$ increases. However, we can now observe better that after several SOAs, the $\sigma^2_{ph}/\sigma^2_{amp}$ starts decreasing again: the higher P_{inSOA}, the faster the decrease and the earlier it starts (i.e., after 6 SOAs for $P_{inSOA} = -14$ dBm, and after 4 SOAs for $P_{inSOA} = -10$ dBm). When comparing the different modulation formats, we notice that PDM-8-QAM induces higher nonlinear phase noise. This can be explained due to the larger amplitude difference between symbols in 8-QAM with respect to the other two, which further changes the instantaneous gain, inducing larger phase modulation.

We now evaluate the reach and the operation range in terms of P_{inSOA} for the different modulation formats. In order to do so, we choose two typical FEC schemes: HD-FEC with 7% overhead and SD-FEC with 20% overhead, which require Q^2-factors of 8.5 and 6.25 dB, respectively. To evaluate the operation range (defined as the range of P_{inSOA} that ensures correct FEC decoding), we calculate for each curve in Fig. 4.5 the P_{inSOA}

at which the Q^2-factor turns below the considered FEC limits (8.5 and 6.25 dB) for both linear and nonlinear regimes (see stars in Fig. 4.5(b) as examples). The results are shown in Fig. 4.8 for both FECs, where upper curves (NL lim.) indicate the maximum P_{inSOA} values to avoid critical nonlinear distortions, while lower curves (OSNR lim.) denote the minimum P_{inSOA} to guarantee a required OSNR, both as a function of the number of cascaded SOAs. We also plot the evolution of P_{opt} (empty markers), showing an initial decrease from -16 dBm to stabilize after 10 SOAs into a range between -18 to 18.5 dBm for all formats. We can observe that reaches of up to 44, 20 and 11 nodes are achieved with QPSK, 8- and 16-QAM signals, respectively, using SD-FEC (30, 12 and 6 SOAs using HD-FEC). The areas within both linear and nonlinear fitting curves indicate the margin of operation of each modulation format.

Nonlinear SOA distortions limit the number of nodes that can be supported in a BOSS ring. Using a HD-FEC maximum ring length of 30 nodes can be achieved when using 112-Gb/s PDM-QPSK signals (100 Gb/s net data rate when accounting for 7%-FEC an 5%-protocol overheads). The number of supported nodes can be increased to 44 when using SD-FEC schemes. However the effective data rate is reduced to 89.6% (20% FEC

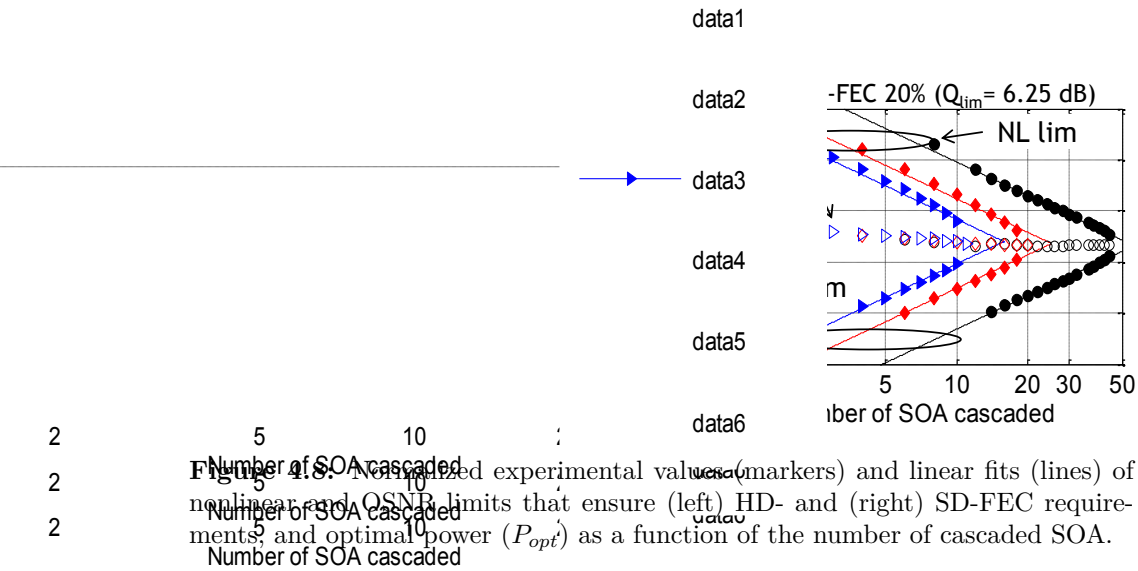


Figure 4.8: Normalized experimental values (markers) and linear fits (lines) of nonlinear and OSNR limits that ensure (left) HD- and (right) SD-FEC requirements, and optimal power (P_{opt}) as a function of the number of cascaded SOA.

cal gates. The main disadvantage of VOAs with respect to SOAs is the lack of optical amplification, which will have to be proportioned additionally through common erbium doped-fiber amplifiers. However, VOAs are absorbing devices which do not nonlinearly distort the signals. Furthermore, they have demonstrated extinction ratios of the order of 20 dB and fast switching times (lower than 15 ns), sufficient for slot suppression. In addition, VOAs can be monolithically integrated in Silicon together with D/MUXs, opening up the possibility of building integrated slot blockers without need of hybrid III/V on Silicon techniques (required when using SOAs) [94].

Using VOAs as optical gates, the main distortion expected when traversing through BOSS-ring networks, may stem from the filtering produced when passing through the de/multiplexers.Such distortions will be analyzed in the following subsections for different kinds of devices.

4.4 Node cascadability: Impact of (de)multiplexing devices

In large-scale datacenters, a single ring may connect a vast number of BOSS nodes. In such node-cascade conditions, it is important to study the accumulation of physical impairments. Assuming the use of linear optical gates, one the of main aggravating effects may arise from filtering induced by the wavelength de/multiplexer placed in each BOSS node. The accumulated filtering response along the cascade may severely filter the optical signals and lead to severe performance degradation. The evolution of the filtering response will mainly depend on the type of D/MUX device used and on the channel spacing. Many technologies have been used up to date to perform wavelength switching, blocking, multiplexing functions, e.g., approaches based on planar lightwave circuit (PLC), MEMS), liquid crystal (LC) or LCoS [151]. Along this section we study the filtering response evolution of various D/MUX devices. First we evaluate the response of high-end performance devices based on LCoS technology, using one of the latest generations of commercial WSS, which will give us a better understanding of the achievable reach when using today's most advanced technology. Then we analyze the filtering response evolution of flattop PLC-like AWG, whose lower cost make them attractive for data center applications.

Table 4.1: Measured and modeled filter bandwidth.

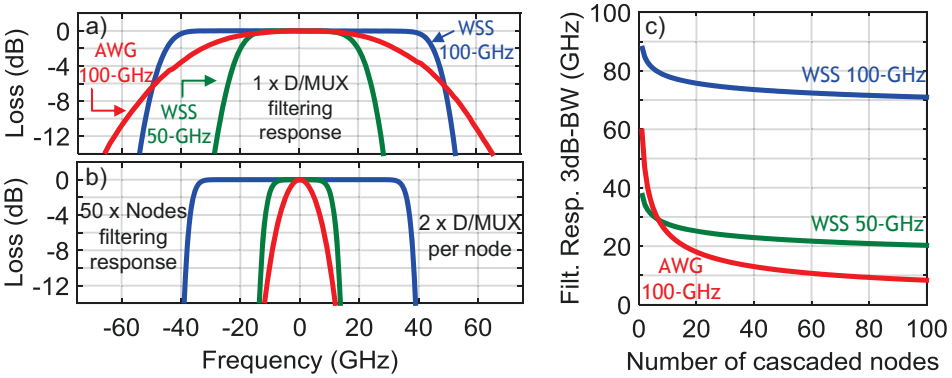


Figure 4.9: Filtering response evolution along a cascade of BOSS nodes while using as D/MUX device a WSS with 100-GHz and 50-GHz grids and flattop 100-GHz AWG. (a) Filtering response of a single D/MUX device, (b) filtering response after 50 BOSS nodes assuming 2 D/MUX per node and (c) evolution of the filtering response 3-dB bandwidth along the node cascade.

In Fig. 4.9(a), we show measured filtering responses of a commercial LCoS-based WSS working with 100-GHz and 50-GHz grid, and a commercial 100-GHz flattop AWG. Such filtering responses are first fitted using the convoluted-Gaussian model [152]. Please, refer to Table 4.1 for measured and modeled 3-dB and 1-dB bandwidths of all devices under study. Using the fitted responses (to avoid error accumulation due to measurement limitations) we calculate the aggregate filtering response obtained when cascading up to 100 BOSS nodes (two D/MUX per node). The evolution of the responses 3-dB bandwidth along the node cascade is depicted in Fig. 4.9(c); see exemplary filtering responses observed after 50 nodes in Fig. 4.9(b).

If we compare now WSS and AWG responses in Fig. 4.9(a), we observe that WSS technology provides much flatter response than lower-cost AWG-based multiplexers. Hence, as depicted in Fig. 4.9(c), AWG's fil-

tering response bandwidth decreases faster than the WSS's. Not even after 10 nodes, 100-GHz AWG's bandwidth becomes narrower than 50-GHz WSS's; see Fig. 4.9(b) for visual comparison after 50 nodes. Nevertheless, in cost-sensitive environments such as datacenters, AWG-based nodes are more likely to be deployed due to their lower cost and their possible integration with other components such as fast optical gates [145].

In the following sections, we first study the cascadabilty of WSS-based nodes (Section 4.4.1). To do so, we employ PDM N-QAM modulations schemes, traditionally used today in optical transport networks. Nevertheless, contrarily to traditional circuit networks, the detection algorithms are adapted to work in burst-mode operation [137, 138]. Subsequently, in Section 4.4.2 we evaluate the cascadability of AWG-based nodes. In order to deal with tightly filtered signals we propose then the use of PDM CO-OFDM signaling, which not only can greatly adapt to any kind of spectral impairment, but also exhibits inherent packet structure [153, 154]. Finally we compare both approaches in terms of reach and capacity.

4.4.1 High-end (de)multiplexers: N-QAM approach

In this subsection we evaluate the cascadability of WSS-based nodes with two different configurations: 100 and 50-GHz grids. The 50-GHz grid allows doubling the number of wavelength channels transmitted within the C-Band (typically 80 channels compared with 40 channels for the 100 GHz). However, reach might be compromised due to tight filtering. In order to better analyze such trade-off we evaluate the performance of several modulation formats while traversing a ring network with up to 100 nodes. The formats under study are PDM-QPSK, PDM-8-QAM, PDM-16-QAM and PDM-32-QAM working at 32.5 GBd, being such the typical schemes used today in coherent optical communications. Nonetheless, contrarily to today's circuit-mode operation, in our case the data is encapsulated in 2-µs slots separated by 100-ns guard-interval. Such interval is meant to amortize switching time of several devices such as lasers while tuning the wavelength or optical gates when blocking the slots. Certainly, also dedicated algorithms allowing fast-recovery of such short slots are implemented.

In order to experimentally study the node cascadability we build the setup depicted in Fig. 4.10(a). As shown in the figure, we pass the light

from an external cavity laser (ECL) into a PDM I/Q modulator, which is driven by the amplified outputs of four DACs generating 65-GS/s electrical signals. As shown in Fig. 4.10(top), such waveforms include a succession of 2-µs-long packets with the following composition: a 256-symbols header, including training sequences to perform data-aided frame synchronization and channel estimation (to be described in the following paragraphs), a 62080-symbols payload and a 100-ns gap. For homogeneity and noise robustness, the header is always modulated using QPSK. However, all modulation formats mentioned above are used for the data-payload. The modulated light is boosted and filtered before entering into a recirculating loop, with which we emulate the cascade of many BOSS nodes, as performed in the previous section. The loop incorporates two WSSs, which are configured with 100 and 50-GHz grids over different experiments. This way we can evaluate the degradation produced by high-end D/MUX (WSS) along the cascade for two possible grid configurations. The loop also includes a 3-km SSMF to ease loop operation (create clear transitions from one round trip to the next and provide enough time for signal acquisition) and two erbium-doped fiber amplifiers to make up for the loss induced by the span, node and the devices required for loop operation. Light from the loop is extracted and sent to a

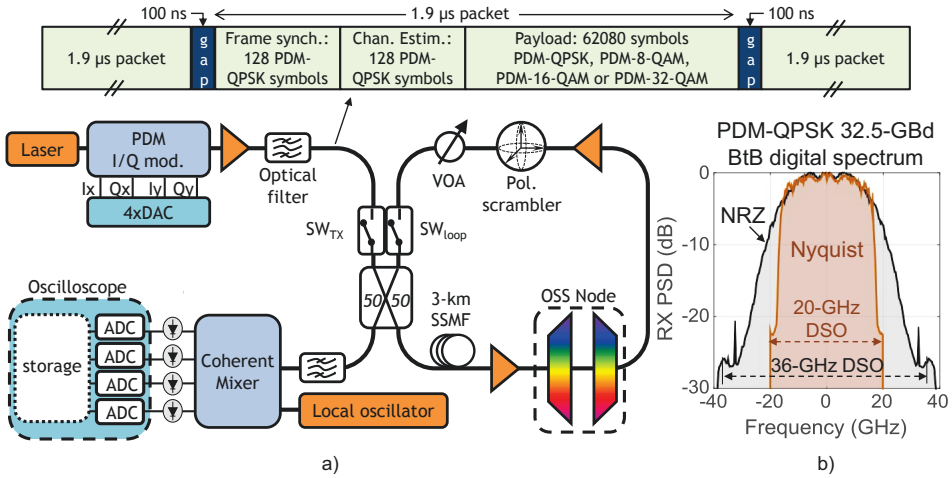


Figure 4.10: (a) Experimental setup and (b) received digital power spectrum measured in BtB configuration for non-return-to-zero (NRZ) and Nyquist-pulse-shaped PDM-QPSK signals at 32.5 GBd. Digital storage oscilloscopes (DSO) with 36-GHz and 20-GHz bandwidth were used for each measurement respectively. Top: Data frame structure used in the experiment.

coherent receiver, feeding the four ADCs of a digital storage oscilloscope.

The received packets are processed offline. For each node count, we digitally compensate for the chromatic dispersion accumulated along the loop. This block is required due to the distance accumulated when traversing more than 100 times over the 3-km long fiber span, which is needed for loop operation only. Notice that in a real datacenter, overall distances do not surpass a few kilometers; hence such process would not be required. For the packet recovery we use the first half (128 symbols) of the header to perform data-aided-packet synchronization, followed by carrier frequency offset compensation [106], and the second half to calculate the equalizer coefficients using data-aided MMSE [155]. Static equalization takes place in the time-domain with a zero-symbol convergence-delay, which is important due to the short length of the packets. After equalization, we perform carrier phase recovery using a maximum-likelihood blind phase search and post-equalization to mitigate possible transmitter/receiver impairments. Then hard-decision and bit-error count take place, refer to Chapter 2 for a more detailed explanation of the DSP blocks.

Along the experiments we studied the performance evolution of the different modulation formats while using two different kinds of signaling: NRZ and Nyquist pulse-shaping (NPS), for which we apply a root-raised cosine filter with roll-off factor $\beta = 0.1$ to all signals:

$$rrc[n] = \begin{cases} \frac{1}{\sqrt{T}}\left(1-\beta+4\frac{\beta}{\pi}\right) & \text{for } n=0 \\ \frac{\beta}{\sqrt{2T}}\left(\left(1+\frac{2}{\pi}\right)\sin\left(\frac{\pi}{4\beta}\right)+\left(1-\frac{2}{\pi}\right)\cos\left(\frac{\pi}{4\beta}\right)\right) & \text{for } n=\pm\frac{T}{2\beta} \\ \frac{1}{\sqrt{T}}\frac{\sin\left(\pi\frac{n}{T}(1-\beta)\right)+4\beta\frac{n}{T}\cos\left(\pi\frac{n}{T}(1-\beta)\right)}{\pi\frac{n}{T}\left(1-\left(4\beta\frac{n}{T}\right)^2\right)} & \text{otherwise} \end{cases}, \tag{4.2}$$

where n denotes samples and T is the symbol period. We show in Fig. 4.10(b) the received spectrum (digitally calculated from the oscilloscope traces) of both types of signaling in BtB configuration. As depicted, the main lobe of an NRZ signals extends up to $f = \pm BR = \pm 32.5$ GHz, being $BR = 1/T$ the symbol rate of the signal. On the other hand, when using Nyquist pulse-shaping, the spectral width of the main lobe can be reduced down to $f = \pm(1+\beta)BR/2 = \pm 17.875$ GHz, where $\beta = 0.1$. Along the experiments we used a 80 GS/s scope with 36-GHz of electrical bandwidth to measure the NRZ signals and a 20 GHz (40 GS/s) oscilloscope to measure Nyquist pulse-shaped (NPS) signals. Please notice that both

scopes have enough bandwidth to capture the main lobe of the respective signals, enabling hence the study of the accumulated filtering response in BOSS rings for both signaling schemes.

We first evaluate the performance of our transmitter in BtB configuration. The results are shown in Fig. 4.11, which depicts the Q^2-factor as a function of the OSNR for all modulation formats along with exemplary recovered constellations. Note that Q^2-factor was calculated from measured BER. For each modulation format, the theoretical curve is plotted in dashed lines by using the same color as its respective experimental curve. As depicted all transmitted signals show performances well above the SD-FEC limit of 6.25 dB, as achievable with an emulated 20% overhead. PDM-QPSK achieves performance very close to theory, presenting only 0.5 dB of OSNR penalty. Nonetheless, the penalty grows when increasing the modulation complexity, leading to ≈ 1.7-dB penalty for PDM-8-QAM and 16-QAM, and to ≈ 3.2-dB penalty for PDM-32-QAM. Such high penalties for PDM-32-QAM mainly stem from limited ENOB of DACs and ADCs

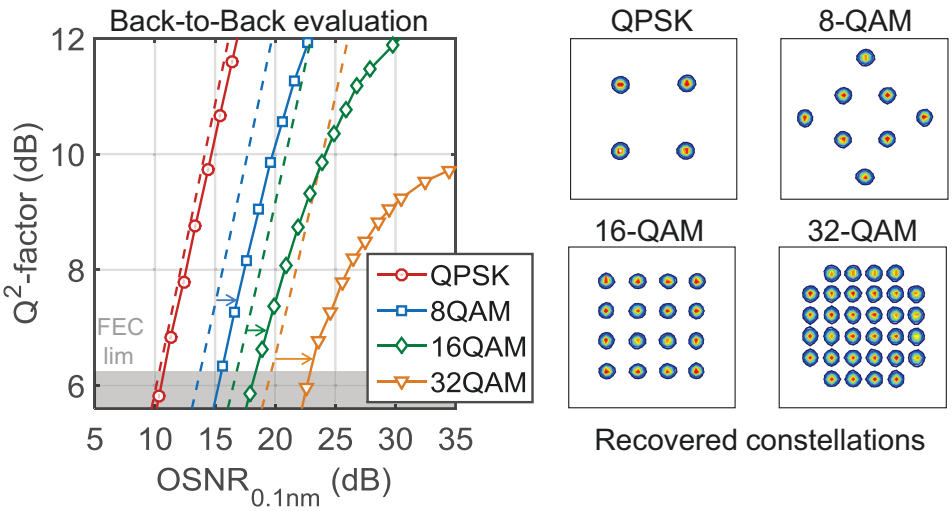


Figure 4.11: (left) Q^2-factor as a function of the OSNR for the NRZ signaling in BtB configuration for the evaluated modulation formats. (right) Exemplary recovered constellations.

specially challenging.

We now study the reach and performance of the different modulation formats when traversing a BOSS ring under different grid configurations and signaling conditions. We first focus on the transmission of NRZ signals along both 100 and 50-GHz grids, see Figs. 4.12(a-b.1-2). We show in Fig. 4.12(a.1) the received spectrum of a PDM-QPSK signal in BtB and after traversing 10 and 50 nodes when using 100-GHz WSSs. One can notice that the spectrum remains almost intact even after going through 50 nodes. This is not surprising if we recall the bandwidth evolution of 100-GHz-spaced WSS-based nodes, shown in Fig. 4.9(b). Note that the 3-dB bandwidth of the accumulated filtering response does not decrease below 70 GHz along the whole cascade. Hence, the 65-GHz-wide (32.5 GHz×2) main lobe of a NRZ signal suffers negligible spectral filtering. The slight asymmetric filtering observed after traversing 50 nodes has two main origins:

1. Amplifier slope: Even after slope compensation and optimization, all EDFAs exhibit a residual amount of frequency slope in their gain curve. One of the drawbacks of the recirculating loop is the lack of heterogeneity of the optical components, which produce the accumulation of their defects instead of averaging them out. In our case, traversing many times through the same gain curve makes the frequency gain slope to increase at every pass, which may lead to slightly asymmetric spectra. This effect is almost negligible in our experiment.

2. Signal frequency shift: In the recirculating loop we used an acousto-optic switch in order to change between looping and loading modes (explained in Section 4.3). This switch has very fast switching times, which allowed us diminishing the fiber length to only 3-km. Nevertheless such kind of switches also have a drawback: they induce a signal frequency shift. The switching process takes places thanks to a change of the effective refractive index induced by an acoustic wave. Along that process, due to Doppler effect, each pass produces a signal frequency shift equivalent to the frequency of the acoustic wave (i.e., 40 MHz in our case). Hence, after 50 nodes our signal is shifted by 2 GHz from its original frequency. This produces an accumulative detuning between the center of the filter response and the frequency of the propagating signal, which induces an asymmetric filtering. When

using a large filter such as the 100-GHz WSS, this effect is almost negligible, because the optical bandwidth of the filtering response is still larger than the signal bandwidth plus the detuning. However, we will observe along the following lines, that such effect becomes visible when using a WSS configured with the 50-GHz grid.

Fig. 4.12(b.1) shows the Q^2-factor of all transmitting modulation formats when using the NRZ signaling and the 100-GHz WSS. As the accumulated filtering response of 100-GHz WSS barely impacts our signal, the main signal impairment arises from OSNR degradation produced through the repeated optical amplification along the cascade. The digits placed next to each curve indicate the number of nodes that can be cascaded before the Q^2-factor drops below 6.25 dB, being such the SD-FEC limit imposed when using a 20% overhead. Thanks to its lower symbol density, PDM-QPSK

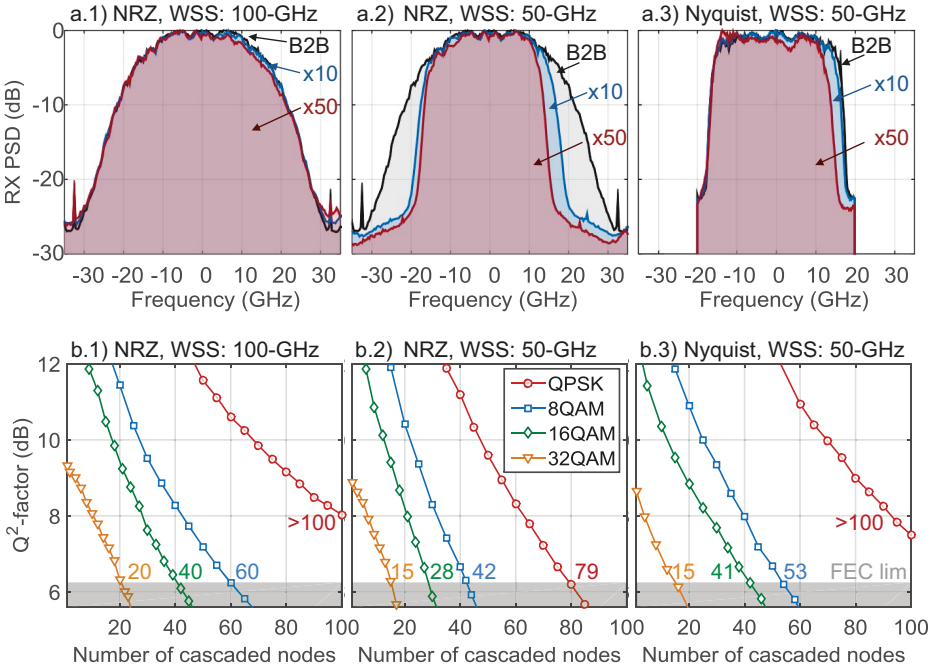


Figure 4.12: (a) Spectra of received signals after traversing 1,10 and 50 nodes. (b) Q^2-factor versus number of traversed nodes for all modulation formats under evaluation. (x.1-3) show different type of signaling and WSS configurations: NRZ signaling with 100- (a-b.1) and 50-GHz (a-b.2) grids and NPS signaling with a 50-GHz grid (a-b.3).

achieves the highest reach, allowing the transmission of 130 Gb/s (100 Gb/s net data rate) over more than 100 nodes. Net data rates are calculated assuming a 30% overhead, which includes SD-FEC (20%) training for packet recovery (<1%) and extra overhead to account for protocol and inter-packet gap. As expected, the maximum node count is reduced when increasing the modulation order due to their lower robustness to OSNR degradation. Nevertheless, up to 60, 40 and 20 nodes can be cascaded for 8-, 16-and 32-QAM PDM signals transmitting 150-, 200- and 250-Gb/s net data rates, respectively.

The same evaluation is performed while setting the WSS in the 50-GHz grid mode, see results in Fig. 4.12(b.2). Under this configuration PDM QPSK, 8-, 16- and 32-QAM can still achieve large reaches of 79, 42, 28 and 15 nodes. Nevertheless, these node counts denote that using a 50-GHz grid, leads to an overall reach reduction of 25-30% with respect to the 100-GHz grid (for NRZ signals). Such penalty can be better understood with the help of the spectra shown in Fig. 4.12(a.2). NRZ signals are severely cut already after 10 nodes. This filtering not only induces inter-symbol interference, which can be only partially removed through digital equalization, but also decreases further the OSNR due to the amount of signal power lost along the filtering. We can distinguish now the asymmetric filtering induced by the detuning accumulated through many passes over the acousto-optic switch (2 GHz after 50 nodes).

The severe filtering occurring when using the 50-GHz WSS can be partially avoided by using Nyquist signaling. As previously shown in Fig. 4.10, Nyquist signaling reduces by almost half the spectral width of the optical signals. This way they can better fit in tight filtering environments such as BOSS rings. Fig.4.12(a.3) shows the spectrum evolution of a Nyquist-pulsed shaped PDM-QPSK signal in BtB and after traversing 10 and 50 nodes, as in the previous cases. Using NPS signaling negligible filtering is observed after 10 nodes, and only a few GHz filtering after 50 nodes. As in the previous cases, the asymmetric spectrum is originated by the frequency shift produced by the acousto-optic switches. When comparing to NRZ, the softer filtering occurring to Nyquist signals translates into an enhanced reach, as depicted in Fig. 4.12. Thanks to Nyquist pulse-shaping we can almost recover the 20-25% of the reach lost with NRZ signaling when using the 50-GHz WSS.

In Table 4.2 we summarize the achievable number of nodes for each

modulation format under the different evaluated WSS configurations: NRZ signals traversing WSSs configured in 100-GHz and 50-GHz grids, and NPS signals traversing 50-GHz WSS-based nodes. For each configuration we calculate the average data rate (D_{ave}) as a function of the ring length, i.e., the number of nodes (N) placed into a ring. The calculation is described by Eq. (4.3), to be explained further, visually supported by Fig. 4.13. As observed in the figure each node-to-node connection has a maximum data rate D_i, which varies with the distance, i.e., number of hops (i), between source and destination. Maximum data rates are extracted from the results obtained in the experiments, see Table 4.2. For instance for NRZ signals and a 100-GHz-grid, maximum data rates are: $D_{1-20} = 250$ Gb/s, $D_{21-40} = 200$ Gb/s, $D_{41-60} = 150$ Gb/s, $D_{61->100} = 100$ Gb/s. For the calculation we assume uniform traffic distribution between nodes, i.e., the same amount of bits (B) in average is exchanged between any node in a ring. Due to different data rates, each node-to-node connection requires different transmission time (t_i), i.e., number of slots, to keep uniform traffic along the whole ring. For each node-to-node connection, t_i can be written as follows: $t_i = \frac{D_1}{D_i} t_1$, where D_1 and t_1 are the maximum data rate and time (number of slots) required to send B bits through 1-hop (direct) node-connections. As indicated in Eq. (4.3):

$$D_{ave} = \frac{\sum_{i=1}^{N-1} B}{\sum_{i=1}^{N-1} t_i} = \frac{(N-1)D_1 \cdot t_1}{\sum_{i=1}^{N} \frac{D_1}{D_i} \cdot t_1} = \frac{N-1}{\sum_{i=1}^{N} \frac{1}{D_i}}, \tag{4.3}$$

from the point of view of one particular node, we can calculate the average data rate as the sum of bits sent to all nodes ($N-1$ connections) divided by the sum of the required transmission time for each connection. After simplification, average data rate calculation is immediate.

The average capacity is depicted in Fig. 4.13 for all possible ring lengths and all evaluated configurations. We can easily notice in such plot that using NRZ signaling together with a 50-GHz grid leads to great capacity losses for any ring length surpassing 27 nodes. For 40- and 60-nodes rings, such configuration leads to 30 and 45 Gb/s lower capacities than using the 100-GHz grid. Furthermore ring length is limited to 80 nodes for this configuration. On the other hand, using NPS signals allows partially recovering such capacity loss, which now has a maximum value of 150 Gb/s (60 nodes) for any ring length. Nevertheless, very high average data rates can be achieved with both grid configurations, when using the right signaling NRZ (or NPS) for 100-GHz grids and NPS for 50-GHz grids. In both cases

Table 4.2: Maximum nodal reach for all evaluated configurations and modulation formats.

			Nodal reach		
Mod. format	Gross Data rate (Gb/s)	Net Data rate (Gb/s)	NRZ 100-GHz	NRZ 50-GHz	NPS 50-GHz
32-QAM	325	250	20	14	15
16-QAM	260	200	40	27	40
8-QAM	195	150	60	40	53
QPSK	130	100	>100	79	>100

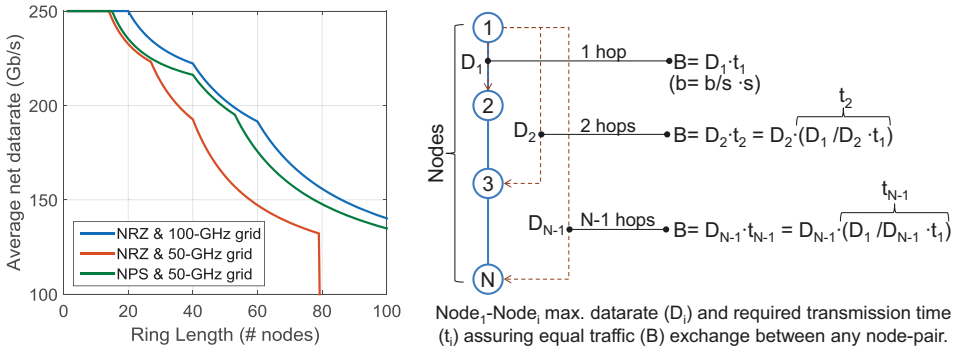


Figure 4.13: Average achievable net data rate per node as a function of the ring length (number of nodes in the ring).

average capacities greater than 200, 175 and (close to) 150 Gb/s (for NPS) can be obtained in 40-, 60- and 80-nodes rings. Furthermore ring length can be extended to more than 120 nodes (maximum measured length) allowing for capacities well above 100 Gb/s. Such results highlight that using a flexible N-QAM NPS transponder allows building very large BOSS-ring networks while assuring very high average capacities. Being able to build large rings is important in BOSS networks to provide high scalability while ensuring full-datacenter connectivity with a single O/E/O conversion, i.e., traversing through only two rings. Moreover, in the torus topology, the use of high-capacity transponders allows for the reduction of the amount of interfaces required to support nodal traffic.

4.4.2 Low-cost (de)multiplexers: CO-OFDM approach

Along the previous section we studied the cascadability of high-end WSS-based nodes. We showed in the introduction of Section 4.4, that such devices exhibit an extremely flat response, which allowed the very high cascadability of commonly used N-QAM signals. Nevertheless WSS are typically built in free-space optics using micro-optics and micro-mechanical elements. Large efforts are taking place today to reduce their footprint and cost, but they still remain rather expensive and relatively large, when compared to integrated technology. On the other hand AWG-based D/MUX can be densily integrated in Silicon technology with other devices, e.g., optical gates, allowing for fully integrated slot blockers with compact millimeter sizes [94]. Nevertheless, as mentioned above, the flatness of AWGs' filtering response is far lower than the one in WSS. In order to recall the reader such statement, we compare in Fig. 4.14 the filter responses of 50-GHz-grid WSS-based nodes (a), studied above, and 100-GHz-grid AWG-based nodes (b) after traversing over 1, 15, 50 and 100 nodes. In the background of each plot we depict a gray rectangle showing the width of NPS 32.5-GBd signals (used in the previous section). As depicted, the convolution of many WSS results into end-to-end filtering responses denoting very flat central

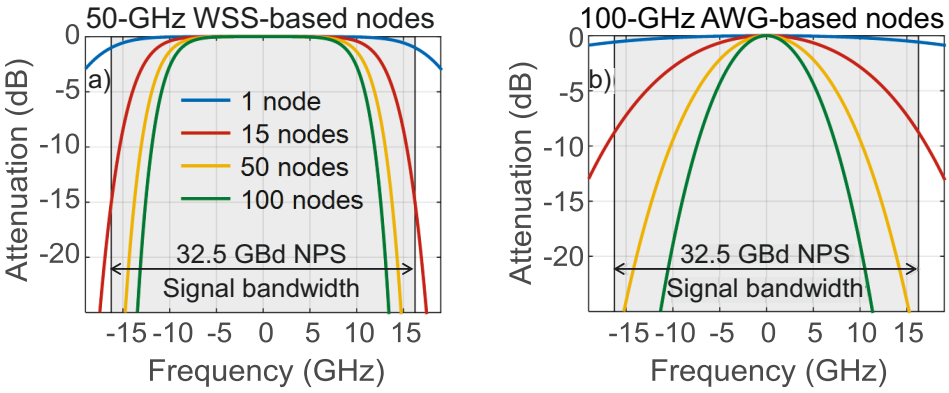


Figure 4.14: Filtering responses after 1, 15, 50 and 100 BOSS nodes, assuming 2 D/MUX devices per node, when using (a) 50-GHz-grid WSSs and (b) 100-GHz-grid AWGs as D/MUX devices.

In this section we propose the use of coherent-optical orthogonal frequency-division multiplexing (CO-OFDM) as modulation scheme for BOSS networks. Such scheme has intrinsically packet-like structure with inherent data-aided frequency-domain equalization and sub-gigahertz spectral shaping precision. CO-OFDM features are eagerly pursued in tight filtering filtering environments such as BOSS-ring networks when using low-cost D/MUX devices. In order to test the novel scheme we experimentally evaluate its performance and reach when traversing a large cascade of AWG-based nodes and compare it to the NPS N-QAM signaling.

4.4.2.1 CO-OFDM for BOSS networks

Different from single-carrier schemes typically used in optical communications, where data is encoded through a unique carrier with large symbol rate (i.e., 32.5 GBd), orthogonal frequency-division multiplexing (OFDM) is a multi-carrier modulation scheme. As observed in Fig. 4.15(a), data is sent through many (e.g., 256) digital low-symbol-rate (e.g., hundreds of MBd) sub-carriers (SC), which leads to a total data rate similar to single-carrier schemes. The subcarriers are closely spaced in the frequency domain by a frequency equivalent to their baud rate, see in Fig. 4.15(a) that this way, the main lobe of each subcarrier is placed between the lobes (in "zeros") of the others, leading to full sub-carrier orthogonality. Therefore, OFDM offers a reduced spectral width (equivalent to NPS signals) when compared to NRZ signaling with similar data rate.

As depicted in Fig. 4.15(a), when using OFDM, we can assign to each sub-carrier a different modulation format (e.g., any of the N-QAM formats used in the previous experiments); this process is called bit-loading, which allows for versatile rate and spectral adaptability. Its spectral tailoring capabilities made OFDM very popular in wireless and copper networks, where such scheme was used to combat multipath-frequency fading (wireless) and high-frequency filtering (copper networks). Such feature was not required in traditional optical networks, which have quite flat spectral responses. CO-OFDM has been previously proposed for long-haul optical communications systems; however, its implementation has not been successful due to three main concerns: 1) low tolerance to phase noise, 2) low robustness to non-linearities and 3) large overhead.

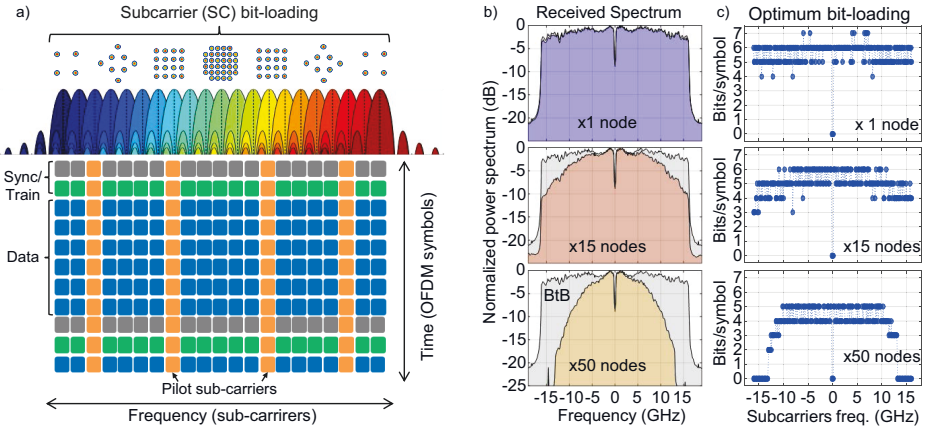


Figure 4.15: (a) Illustration of a OFDM spectrum together with a time-frequency representation of the OFDM structrure. (b) Experimentally measured filtering responses and (c) optimum calculated bit-loading when traversing 1, 15 and 50 AWG-based nodes.

Nevertheless, we will see that most of such concerns can be neglected in the particular scenario of datacenter networks. Being so, we can benefit from several interesting OFDM properties to enhance capacity and reach. Along the following paragraphs we first describe several OFDM features that make this scheme an interesting candidate for BOSS ring networks, and later we address the aforementioned OFDM concerns.

Resiliency to filtering effects: Filtering distortions are the dominant performance-degrading impairments in BOSS networks when using low-cost D/MUX. This becomes evident when exploring Fig. 4.15(b), which shows the spectra of received OFDM signals after traversing 1, 15 and 50 100-GHz-grid AWG-based nodes. The received spectrum in BtB is also plotted in gray to better visualize the spectral filtering impairing the signals (experimental setup is presented in the following subsections). When comparing to Fig. 4.14(b) we observe a good agreement between the theoretical filtering responses and the shapes of the received signals after traversing different number of nodes. It is evident that such tight filtering will severely impact transmission performance (to be shown in the following subsections). Hence, the capability to mitigate such impairments is an essential point in the implementation feasibility evaluation of considered proposals. Along this subsection we propose the use of CO-OFDM to spectrally adapt to

such tight filtering through a low-complexity non-iterative bit-loading, targeting average spectral efficiency (SE) maximization for a given reference bit-error rate [156]. The calculated bit-loading for 1-, 15, and 50-nodes transmission is shown in Fig. 4.15(c). One can clearly observe that the bit-loading process adapts to the spectral filtering by giving higher modulation formats to central (less filtered) sub-carriers, while assigning lower-density but more robust formats to the more impaired high-frequency sub-carriers. Sub-carrier nulling takes place for highly deteriorated sub-carriers, which allows increasing overall OSNR. Per-route dedicated signal tailoring offered by OFDM improves the resilience to filtering distortions caused by wavelength D/MUX cascading and/or wavelength detuning, in turn leading to enhanced receiver sensitivity and/or capacity maximization.

Robust equalization: In BOSS networks, every slot is treated as an independent information entity and, consequently, equalization is performed on each of them individually. With those slots lasting few µs, the equalization rate allows for quasi-continuous tracking of the channel response, sufficing to account for polarization changes considerably faster than expectable in short-reach applications [157]. On the other hand, such frequent channel estimation results in an overhead-driven degradation of the effective rate, evincing the need for robust and fast (both in terms of processing latency and required overhead) equalization methods.

In this regard, OFDM intrinsically has a packet-like structure. As shown in Fig. 4.15(a), data is always preceded by a set of symbols used for synchronization and channel estimation (training). Such structure is required in BOSS systems to perform fast packet synchronization and equalization. In addition, data-aided frequency-domain equalization has been long proved superior over the analogous time-domain alternative in terms of processing complexity ($O(L \log L)$) versus $\mathrm{O}(L^2)$, where L is the number of filter taps) [107]. Interestingly, unlike single-carrier modulation schemes, OFDM inherently implements data-aided frequency-domain equalization, meaning that no extra intermediate time-to-frequency and frequency-to-time transformations are required besides the necessary to perform demodulation. These facts clearly underscore the convenience of OFDM for BOSS networks from the equalization standpoint. Furthermore, because of the fine-granularity spectral adaptivity that OFDM features, further overhead reduction can be attained by assuming, not only that the channel remains static for the entire slot duration, but also that neighboring sub-

carriers present similar response. The latter allows for two dimensional noise-averaging (over both time and frequency) [158], thus reducing the time-duration of the overhead without performance degradation. Lastly, operating on training sequences ensures consistent equalization irrespective of the payload properties, thereby becoming remarkably useful in the considered high-heterogeneous multi-rate environment. This allows for using a universal and static overhead across the entire network in turn simplifying transceiver management and software development.

Frequency Pilot-Tone: As currently devised, wavelength-multiplexing in BOSS architectures relies on the fast tunability of the transponders, for whose implementation, fast-tunable lasers are preferred for their scalability. These lasers show residual post-switching frequency and phase instability [93] that requires inserting guard intervals (in the order of hundreds of nanoseconds) between consecutive slots, with the consequent reduction in effective rate. Robust carrier phase and frequency estimation becomes crucial, further emphasized by the strong sensitivity to phase noise that OFDM exhibits.

For that purpose, we propose using one time-slot-long unmodulated frequency pilot tone (FPT). Fig. 4.16(a) illustrates an OFDM spectrum including a FPT co-propagating along with the modulated sub-carriers. Surrounding the FPT, a few sub-carriers may be nulled in order to avoid interference coming from neighbouring modulated signals, which induces a certain amount of overhead ($\approx$1% in our case). Nevertheless, such overhead is an order of magnitude smaller than the one required by abovementioned traditional techniques. An experimentally measured CO-OFDM spectrum including the FPT can be observed in Fig. 4.16(b). This co-propagating tone continuously tracks the evolution of the phase fluctuations experienced by the electric field at the exact frequency where it is inserted. One can see in the zoomed spectrum shown in Fig. 4.16(c), the FPT including a certain amount of frequency offset and phase noise induced by both transmitter and receiver lasers. In the receiver, that information shall be used for data-independent correction of phase distortions common to all spectral components; see obtained phase noise compensation after FPT processing in Fig. 4.16(d). The associated receiver-side FPT processing is described along the following sections.

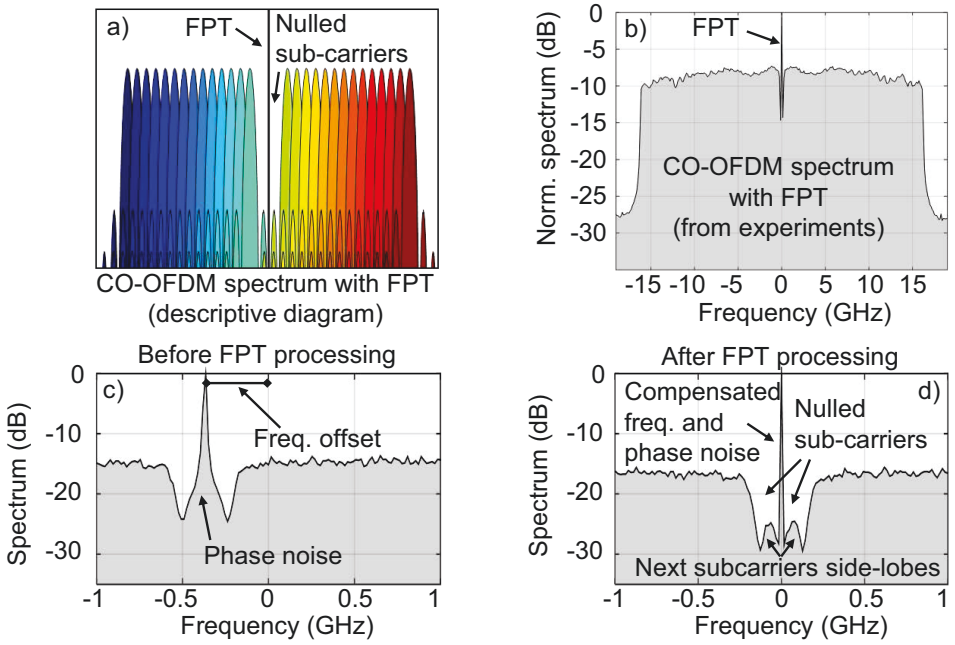


Figure 4.16: (a) Illustration of the spectrum of an OFDM signal including a frequency pilot tone (FPT) co-propagating with the sub-carriers, (b) spectrum of the experimentally generated CO-OFDM signal including a FPT, zoom on the CO-OFDM focusing on the FPT (c) before and (d) after FPT processing.

Addressing early OFDM concerns:

- CO-OFDM concern 1 - Low tolerance to phase noise: Laser phase noise induces inter-carrier interference (ICI) in OFDM signals, which severely impacts performance when phase noise is not properly compensated. As shown in Fig. 4.15(a), traditional OFDM schemes use several pilot sub-carriers to perform carrier phase estimation and recover from phase noise. However, if the lasers do not have extremely low linewidths (tens of kHz), phase noise is not constant along a whole OFDM symbol; hence the estimation is not correct, which leads to ICI [159]. Neverthess, different from traditional appraoches, we insert a FPT which co-propagates along with the data sub-carriers. This way, by simply filtering the FPT at the receiver, we can continuously track phase and frequency fluctuations at the ADC sampling-period precision (which period is hundreds of times smaller than the OFDM-symbol, depending on the number of sub-carriers). The effectiveness

and enhanced robustness of this technique when compared to other alternatives has already been demonstrated in similar scenarios [160].

- CO-OFDM concern 2 - Less tolerant to nonlinearities: OFDM signaling has higher peak-to-average power ratio (PAPR) than typical single-carrier approaches. Hence, such scheme suffers from a higher accumulation of nonlinear effects (e.g., self-phase modulation (SPM), cross-phase modulation (XPM)), which typically results into a reduced reach in long-haul transmissions [161]. Nevertheless, due to the limited size of a datacenter, BOSS rings will have a maximum extension of few kilometers. Under regular conditions in optical systems, nonlinear effects can be neglected in few kilometer-long links. Hence nonlinear impairments do not limit OFDM performance/reach in datacenter-size environments.

- CO-OFDM concern 3 - Large overhead (OH): Traditional OFDM schemes require a large amount of overhead which includes training and synchronization symbols (2-4% OH), pilot sub-carriers for phase noise recovery (10%) and cyclic prefix to avoid ICI induced by chromatic dispersion (10% in links with few thousand kilometers) [162]. Nevertheless, thanks to the short link length required in our scenario (few kilometers), the amount of cyclic prefix can be drastically reduced. In our case, we use a 2% cyclic prefix, which is large enough to accommodate for signal dispersion and possible polarization differential group delay. Such overhead is five times smaller than the one typically used for long-haul transmission. In this regard, the overhead reduction achieved by the two-dimensional averaging in the channel estimation process, the use of FPT instead of multiple pilot carriers for phase noise compensation and the small amount of required cyclic prefix lead to an overall overhead smaller than 4% (detailed in the following subsection), which is well leveraged with the increased capacity and reach offered by CO-OFDM in this scenario (to be shown in following sections).

4.4.2.2 OFDM generation and detection: DSP blocks

In this sub-section, we elaborate on the employed DSP for offline OFDM generation and detection. The description is supported by Fig. 4.17 for the ease of understanding, where the block diagrams of the transmitter's

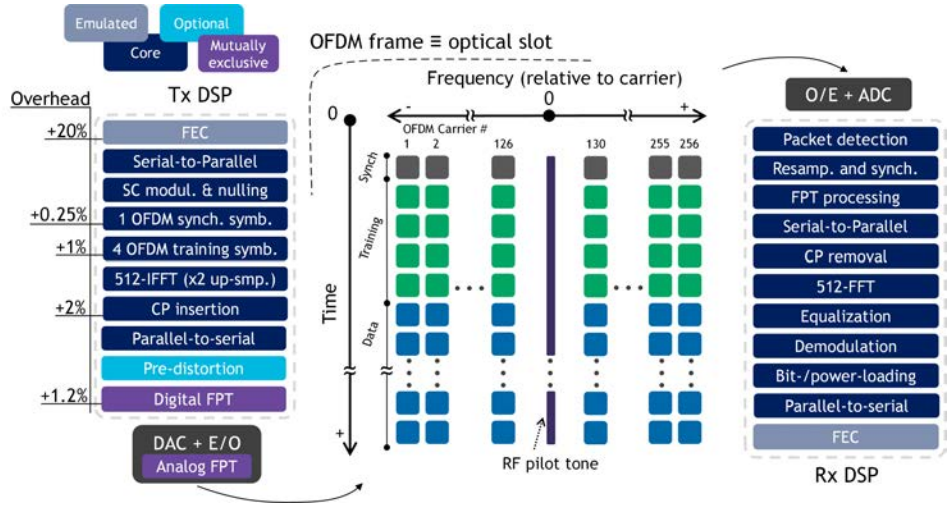


Figure 4.17: Block diagrams of the DSP employed in transmitter (left-most) and receiver (right-most) sides for PDM-OFDM transmission on the proposed BOSS ring network. In the middle, frequency-versus-time representation of one of the transmitted OFDM-based optical slots.

and receiver's DSP are shown (on the sides) together with a frequency-time representation of one of the transmitted OFDM-based optical slots (middle).

As shown in Fig. 4.17(left), after assuming correct channel encoding for FEC with 20%-overhead soft-decision code (6.25-dB Q^2-factor threshold), the transmitter input serial bit-stream is parallelized into 253 independent vectors, each corresponding to one information-carrying OFDM sub-carrier. After sub-carrier (SC) modulation according to a given bit-loading and power loading vectors, sub-carrier nulling is performed for interference minimization on the later inserted FPT. The unfilled sub-carriers correspond to direct current (DC) and 1 neighboring sub-carrier on each side (3 in total). Then, a total of 5 over-head (OH) OFDM symbols are then introduced: 1 for symbol synchronization based on the method described in [163] (0.25% OH[2]) and 2 correlated dual-polarization symbols (4 OFDM symbols in total) for channel estimation as described in [164] (1% OH). Following data-wise OFDM frame composition, Inverse Fast Fourier Transform (IFFT) is performed. The IFFT size equals 512 (oversampling factor

[2]The OFDM-related overhead is calculated in percentage with respect to the whole packet duration.

2), and a cyclic prefix (CP) of length 10 ($\approx$ 2% OH) was used to widely accommodate the low chromatic dispersion experienced within the datacenter. Finally, following parallel-to-serial conversion, $\approx$ 90% OFDM clipping factor and linear M-shape pre-distortion with 0.37 dB/GHz are applied for combating quantization noise and static transmitter band-limiting effects, respectively. These values are empirically calculated for optimum BtB BER performance. Slots of 3 µs are generated and loaded onto the DAC's memory with a guard-time period of 100 ns between consecutive slots.

Concerning FPT insertion, various approaches can be used including digital [165], where highly accurate center-frequency allocation is possible (readily allowing for radio frequency FPTs) at the expense of quantization noise; optical analog , where additional optical components are needed; and electrical analog [160], where the FPT is inserted by inducing certain level of DC component during electro-optical conversion, trading precise pilot-to-signal ratio (PSR) control and hardware minimization for power consumption and potentially causing non-linear signal distortions according to the modulator's response. In this investigation, the latter option was employed for implementation simplicity, where one single FPT was inserted in one polarization. Two main parameters need to be considered when implementing the FPT. Firstly the PSR, directly related to the accuracy of the phase estimation. By increasing the PSR, the influence of the measurement noise in the estimation becomes weaker, while excessively high PSR results in OSNR degradation for the information signal and quantization noise when digital FPT generation is employed. The optimum ratio depends on the measurement noise power, the phase noise variance, and the characteristics of the filter employed for FPT isolation [160]. Second important consideration is FPT neighboring sub-carrier nulling, governing the compromise between effective rate and the magnitude of the frequency excursions that the method is able to track. Given a fixed observation time, (e.g., one slot duration) time-varying phase fluctuations appear as FPT broadening in the frequency domain (see Fig. 4.16(c)); therefore, sufficient gap between the FPT and the closest information signal's frequency components need to be allocated so as not to incur in spectral overlapping, and thus the miscalculation of the phase fluctuations to be compensated for.

At the receiver side, resampling and coarse power-based packet detection are first carried out, followed by bulk digital CD compensation as pragmatic means of flexibly adapting the equivalent inter-node distance

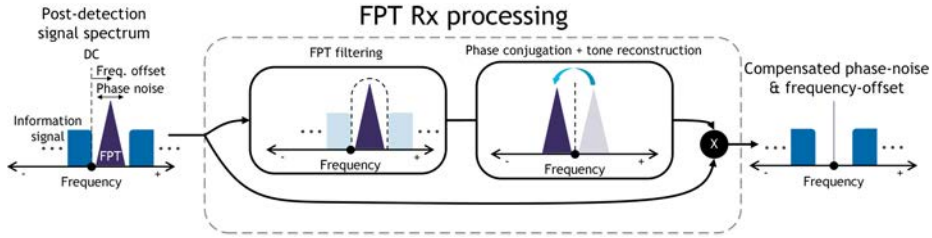


Figure 4.18: Conceptual schematic of FPT processing in the receiver side for phase-noise and frequency-offset compensation.

(set to 200 meters in the experimental demonstration). Note that this block would not be required in the final transponder implementation. Later, FPT processing is realized for blind frequency recovery and phase noise compensation. First we estimate the FPT center frequency through maximum peak detection in the Fourier domain. Then, as depicted in Fig. 4.18, the phase-distorted FPT is filtered using a ≈20-MHz bandwidth raised-cosine band-pass filter with roll-off factor of 1 centered at the FPT frequency. Subsequently, the complex conjugated exponential is reconstructed and applied sample-by-sample to the original frame. Timing synchronization and CP removal precede standard 512-Fast Fourier Transform (FFT) sub-carrier filtering. Afterwards, channel estimation using two-dimensional (time and frequency) measurement noise averaging is performed before channel equalization. The DSP chain ends with demodulation plus error counting, where the signal-to-noise ratio (SNR) per sub-carrier is estimated via error-vector magnitude. The SNR information can be sent to the transmitter through the control channel in order to update the bit- and power-loading.

4.4.2.3 Experimental setup

Fig. 4.19(a) shows the experimental setup used to generate, transmit and detect the PDM-CO-OFDM signals. Such setup is very similar to the one used in the previous section. Following the offline-DSP generation of the train of OFDM-based packets/slots described above, the signal vectors are loaded onto four 65 GS/s DACs. After amplification, the DACs' outputs modulate a 1552.52-nm ECL with ≈100-kHz linewidth with a PDM-I/Q modulator. The FPT is inserted through fine control of the modulator's DC bias in one of the polarizations, with empirically calculated optimum

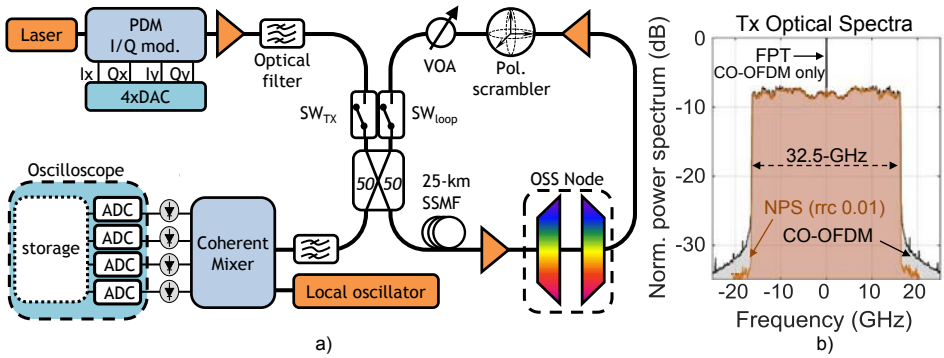


Figure 4.19: (a) Experimental setup schematic. (b) Optically measured transmitted spectra of CO-OFDM and NPS signals.

PSR of 15 dB.

The output from the optical modulator is pre-amplified and launched into a recirculating loop. If we recall the setup described in the previous section, in order to change between loading- and looping-mode of the recirculationg loop, we used two fast acousto-optic switches. However, such devices induce frequency shift of ≈40 MHz at each round trip to our signal; hence leading to up to 4 GHz of accumulated detuning after traversing 100 nodes. In order to avoid such effect, we used for this experiment electro-optic switches, which has slower switching time but no frequency shift. Due to the higher switching time we had to increase the fiber length to 25 km in order to ensure loop stability. The polarization scrambler and EDFAs for attenuation compensation were kept as in the previous setup. For this experiment we used two programmable optical filters with 1-GHz frequency resolution to emulate BOSS-node AWG-like (de)multiplexers. The filters feature 3-dB and 1-dB bandwidth at approximately 75% and 50% channel spacing (100 GHz in our implementation) respectively, corresponding to typical values of commercially available wideband arrayed waveguide gratings, whose filtering response was described at the beginning of Section 4.4. Note that in order to better focus on node cascadability, no switching operations are performed along this experiment. Laser switching and blocking functionalities are left for future work.

At the receiver side the signal is filtered and coherently mixed with an ECL with ≈100-kHz linewidth as local oscillator (LO) and detected. The outputs from four balanced photodiodes are digitized for offline processing

by a 40-GS/s, 20-GHz bandwidth DSO.

4.4.2.4 Results: Nyquist-N-QAM and CO-OFDM comparison

Along this section, we experimentally evaluate the node cascadability and wavelength detuning tolerance in a BOSS-ring network for PDM-CO-OFDM, and compare the results against the PDM N-QAM transponder. In order to enhance reach of the N-QAM transponder we used Nyquist pulse-shaping (NPS), as in the latter case shown in Section 4.4.1. For this experiment we pulse-shaped the N-QAM signals using a root-raised cosine with roll-off factor of 0.01. Such roll-off further reduces the spectral width of the signal to its baud rate (32.5 GBd), hence providing higher robustness to filtering. This way, NPS and CO-OFDM signals have similar spectrum shaping and width (≈32.5 GHz) which allows a fair comparison, see Fig. 4.19(b). Similar to the CO-OFDM approach, for NPS transmission we also generated 3-µs optical packets separated by 100-ns gaps. Packet encapsulation and recovery was performed as described in Sub-section 4.4.1.

Capacity and reach: In order to compare both CO-OFDM and NPS transponders, we assess the maximum bitrate that can be achieved as a function of the number of cascaded nodes. Such results are depicted in Fig. 4.20(a). In order to evaluate the maximum bitrate of the NPS transponder, we performed similarly as described in Section 4.4.1; this time using the new setup with 100-GHz AWG-based node. The reach for each format is again defined as the maximum node-count achieved while satisfying Q^2-factor$\geq$ 6.25 dB (the Q^2-factor is obtained from measured bit error rate). Orange circles in Fig. 4.20(a) denote the measured maximum reach and net data rate for each NPS case, which defines the staircase function describing the maximum net data rate versus number of nodes for the PDM-NPS transponder. For the net data rate calculation we substract 30% of overhead to the gross data rates, which includes 20% FEC overhead and a further 10% overhead to account for NPS packet recovery ($<$ 1%) and possible inter-slot gap. In contrast, CO-OFDM bitrate can be maximized for each node-count by adapting the bit-loading to the end-to-end spectral profile. Bit-loading adaptation was previously shown in Fig. 4.15(c). Such technique adapts the modulation format to the spectral deterioration, assigning lower modulation orders to most impaired sub-carriers.

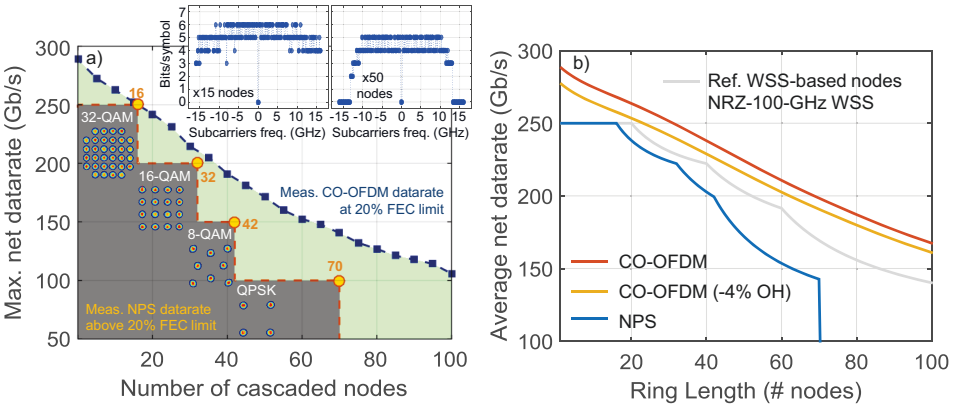


Figure 4.20: (a) Comparison of experimentally measured net data rate above FEC limit vs. number of cascaded nodes and (b) average net data rate vs. ring length for PDM-CO-OFDM and PDM-NPS. Insets in (a) show exemplary bit-loading settings for 15 and 50 nodes transmission.

Next to PDM-NPS maximum rates, blue squares in Fig. 4.20(a) delimit the net data rates achieved by PDM-CO-OFDM when optimizing bit-loading, while ensuring performance above FEC limit (also assuming 30% of total overhead). It is apparent that PDM-CO-OFDM's spectral adaptability allows for higher overall rates as filtering becomes the dominant degradation (>20 nodes). Along the first nodes, DAC quantization noise governs the penalty, reason for the observed performance equalization with respect to NPS (see node 15). Beyond ∼60 nodes, filtering response bandwidth reduction starts saturating [recall Fig. 4.9(c)]. Consequently, the effective signals' bandwidth stabilizes, causing a slight bitrate loss flattening [see Fig. 4.20(a)]. This explains the amount of increased reach when switching from PDM-8-QAM to PDM-QPSK (28 nodes). It is important to remember that OSNR degradation also contributes to bitrate decrease. This can be observed by comparing average bit-loading in central sub-carriers (not severally filtered), which changes from 5.5 bits/symbol to 4.5 bits/symbol between 15 and 50 nodes [see inset bit-loadings in Fig. 4.20(a)]. One has to notice that zeroing high frequency sub-carriers, enhances OSNR for all other sub-carriers, hence slowing down OSNR degradation along the cascade.

Fig. 4.20(b) shows the average net data rate achievable for NPS and CO-OFDM signals as a function of the ring length (in number of nodes), when assuming uniform throughput between any node-pair in the ring. The

average net data rate was calculated for each ring size using Eq. (4.3), described in Section 4.4.1. Both NPS (blue) and CO-OFDM (yellow) curves represent net data rates after the 30% overhead reduction. However, in orange we also plot the CO-OFDM average net data rate after further subtracting the 4% extra-overhead required for its implementation along this experiment[3]. In gray we plot as a reference curve the NPS average net data rate when traversing through 100-GHz WSS-based nodes (extracted from the previous section). We can observe that when using AWG-based nodes, for rings longer than 30 nodes, NPS signals average capacity drops rapidly with respect to WSS-based nodes, due to tight AWG accumulated filtering response. On the other hand, thanks to its high spectral adaptability, CO-OFDM surpasses NPS average data rates for any ring length, even when NPS signals traverse high-end WSS-based nodes. This fact demonstrates that CO-OFDM allows using low-cost devices with no extra capacity loss. Furthermore, when comparing both signaling types under the same conditions (traversing AWG-based nodes), CO-OFDM offers a 30% increased capacity for ring lengths above 60 nodes (approximately 200 Gb/s w.r.t. 150 Gb/s for a 60-node ring), and 40% enhanced reach (more than 100 nodes w.r.t. 70 nodes maximum reach).

Detuning tolerance: Commercially available high performance lasers and D/MUX specify frequency uncertainties of ± 1.5 and ± 2.5 GHz, respectively. Furthermore, low cost D/MUXs and fast-tunable lasers present even higher frequency variations. Such uncertainties, especially on the transmitter laser, lead to frequency mismatch between laser and D/MUXs' central frequencies, which can reach several GHz when performing fast tuning operations. In this sub-section, we shift the transmitter laser with respect to the D/MUX central frequency to characterize the resilience to frequency detuning of PDM-CO-OFDM and compare it to the PDM-NPS approach. In order to perform a fair comparison, we measure the performance in terms of

[3]Please notice that the OFDM-related overhead is calculated in percentage of time with respect to the whole packet duration. Hence, as we change the modulation format in each sub-carrier for each node-pair communication, when calculating the net data rate, the overhead (in terms of bits) should be calculated for each node-pair. For simplicity, in the curve we apply the same 4% overhead, independently on the node-count; which describes the worse-case scenario (the largest possible overhead), for which the same modulation format (e.g., QPSK) is used in all sub-carriers for both training and data symbols. Note that we typically use QPSK modulation for training symbols and higher-order modulation formats for data symbols, which in reality leads to bit-overheads lower than 4%.

Q^2-factor for two signals with the same net data rate. In the case of PDM-NPS, we use QPSK modulation format, while for CO-OFDM, bit-loading is designed to match PDM-QPSK ≈100 Gb/s data rate with maximum performance.

Fig. 4.21(a) depicts the measured Q^2-factor versus detuning for both approaches after 44 cascaded nodes. Such node-count represents the reach with highest power margin for PDM-NPS configuration without detuning. One can easily notice that even having higher margin at zero detuning, PDM-NPS performance rapidly drops, denoting 5-GHz of detuning tolerance, while PDM-CO-OFDM maintains high margins even after 9-GHz detuning. CO-OFDM detuning resiliency is due to its flexible adaptation. Fig. 4.21(b) shows the received CO-OFDM spectrum (blue) when applying 9-GHz detuning after 44 nodes (filter response in red). As observed, the upper ≈30% of the original back-to-back spectrum (in grey) is suppressed due to filtering and detuning. Nevertheless, bit loading optimization (blue dotted-dashed curve) allows transmitting the aimed data rate with margin for more than 4 GHz higher detuning than PDM-QPSK.

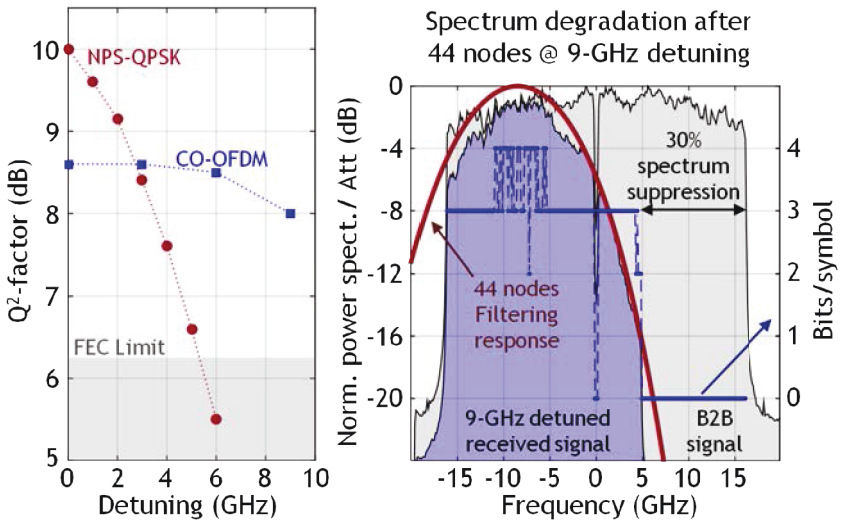


Figure 4.21: (a) Measured Q^2-factor vs. frequency detuning after 44 nodes for PDM-CO-OFDM and PDM-QPSK at 130 Gb/s. (b) Recovered PDM-CO-OFDM 9-GHz detuning power spectrum compared to BtB original spectrum, filtering response after 44 nodes, and bit-loading assigned.

In this section we have observed that over-filtering is the major degradation limiting the transmission reach AWG-based BOSS nodes. In this regard, due to its fine spectral adaptability, robust data-aided frequency-domain equalization, and inherent block-wise processing capabilities we have proposed CO-OFDM as a potential solution to be used in BOSS transponders for large-scale intra-datacenter networks. The reported experimental comparison between traditional NPS and our novel approach, demonstrates considerable gain in achievable average data rate (+30%), reach (+40%) and detuning tolerance (+80%), which allows the use of low-cost D/MUX devices with no extra capacity loss.

4.5 Connex work

Being kept outside of this document, the author of this thesis has been actively involved in the design and system testing of novel fully integrated slot blockers in Silicon platform. Such devices have been fabricated within the framework of the CELTIC+ SASER-SAVENET project, with the participation of Universite Rennes 1 (ENSSAT), Universite Paris-Sud, III-V Labs, CEA-LETI and Nokia Bell Labs. The project has led to the fabrication of complex devices including more than 30 integrated functional elements. Fig. 4.22 shows a schematic and a photography of a fully functional slot blocker supporting up to 16 100-GHz-spaced wavelength channels with dual-polarization diversity. As depicted in the figure, the slot blocker includes two 2D-vertical grating coupler (VGC) used to couple the light in/out to/from the device while de/multiplexing both polarizations. Then, two symmetric arms provide the fast packet blocking of each polarization. Each arm contains two 100-GHz spaced AWG, used to de/multiplex up to 16 wavelength channels, and 16 VOAs built with p-i-n junctions used as fast optical gates, which exhibit up to 20 dB of extinction ratio and less than 10 ns of switching response.

Silicon integrated platform provides large-scale integration and high-volume productions, which is expected to lower the cost and size of photonics systems. The potential of such technology is shown in this first prototype, which densely integrates more than 30 photonic functions in a compact chip (4.1×2.7 mm^2), paving the way to low-cost slot blockers for BOSS ring datacenter networks. For a more detailed information about the integrated slot blocker please refer to [94, 145, 166].

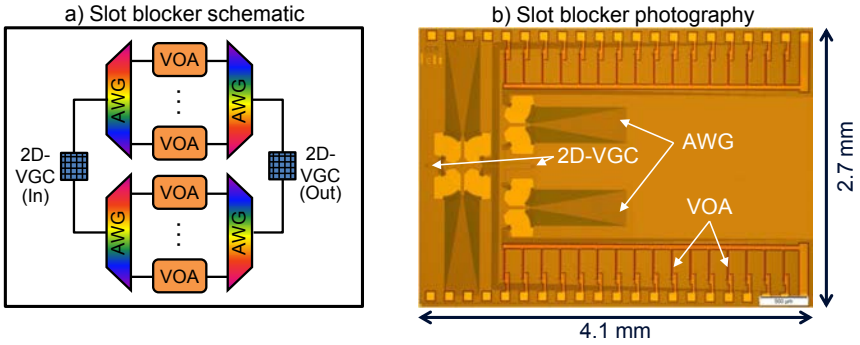


Figure 4.22: Monolithically integrated slot blocker: (a) schematic and (b) photography.

4.6 Summary

Along this chapter we have first introduced the BOSS ring-based topology for intra-datacenter networks. Combining transparent inner-ring transmission and high data rate transponders with slot-granularity statistical multiplexing, such architecture provides a highly scalable solution, capable of supporting millions of servers while reducing the number of networking component and energy consumption, when compared to traditional Folded Clos topology. The main goal of this chapter is the proposal of BOSS-node possible implementations and their evaluation in terms of cascadability (reach) and average capacity.

First we assessed the use of SOAs as optical gates in the BOSS-nodes, which are interesting for its unique amplification capacity. However, we have experimentally shown that nonlinear distortions induced by the SOAs limits the reach to approximately 40 nodes, which would lead to poor scalability. Hence, we move towards VOA-based optical gates, which do not introduce nonlinear distortions and can be integrated monolithically with other devices without the need for hybrid platforms.

Then we studied the cascadability of different kinds of D/MUX devices under several grid configurations. We demonstrated that when using 50- or 100-GHz-grid WSS-based nodes, flexible N-QAM (NRZ or NPS) transponders can attain very large node count and average net data rates above 200, 150 and 100 Gb/s for 40-, 70- and 100-node rings, respectively. Nevertheless, when moving to low-cost D/MUXs, such as 100-GHz-grid AWGs,

the aggressive accumulated filtering severally impacts performance of more spectrally efficient NPS N-QAM signals, which now lose more than 30% of average capacity and exhibit 70-nodes limited reach, when compared to WSS-based nodes. In order to overcome such limitations, we proposed the use of CO-OFDM which is capable to spectrally adapt to most deteriorating frequency channels through sub-carrier bit-loading optimization. We demonstrated that CO-OFDM can traverse AWG-based rings with superior average data rate than NPS or NRZ signals when going through high-end WSS-based nodes. When comparing both NPS and CO-OFDM in a low-cost environment, the latter provides a 30% increased average data rate, a 40% higher reach and a 80% further detuning tolerance than the former, thereby allowing the use of low-cost devices without incurring extra penalties.

Chapter 5

Thesis conclusions and perspectives

5.1 Thesis conclusions

Along this thesis we have analyzed the trends of current data centers and studied their needs to then propose optical physical layer solutions addressing both short and long term requirements. In Chapter 1 we have first identified the trends that have been pushing up Internet and data center traffic along the last decades, and later described the evolution of the data center infrastructure to cope with such traffic. Today, emerging cloud services are producing an enormous bandwidth demand in data centers which is increasing at a fast pace, almost doubling every year in largest service providers. In order to handle such amount of traffic, large facilities host hundreds of thousands of servers, which are typically interconnected through a complex Folded Clos network, which consist of several switching stages comprising each thousands of switches. Promoted by their large capacity and extended reach, optical systems have become the predominant technology used today for switch-to-switch and switch-to-server inter-connection.

In order to deal with the ever-increasing traffic, data centers are continuously upgrading their infrastructures. 10-Gb/s and 40-Gb/s (implemented with 4×10-Gb/s lanes) transceivers and switch ports are today governing data center facilities. However, the recently standardized 25-

and 100-Gb/s (4×25-Gb/s lanes) modules are expected to take over soon. Furthermore, the Ethernet Alliance is already working on the standarization of 200- and 400-Gb/s interfaces, which ultimate implementations are supposed to adopt the quad-lane trend, hence requiring 50 and 100-Gb/s data rates per lane, respectively. Such high data rate interfaces will be most likely deployed using advanced modulation formats (i.e., PAM-4), which allows doubling the data rate of currently used PAM-2 schemes, while using the same bandwidth. Hence 50-Gb/s lane-rates can be achieved with the limited bandwidth of today's 25G components. Nevertheless, when increasing to 100-Gb/s (per lane) while using PAM-4 modulation schemes, symbol rates must be increased to 50-56 GBd (56-GBd when including FEC overhead). The implementation of such interfaces is very challenging because they require the deployment of novel high-bandwidth components, in both electronic and photonic domains, while keeping in mind the strict data requirements such as cost, power consumption and footprint.

After reviewing the basic concepts on optical transmission systems in Chapter 2, we have addressed in Chapter 3 the urgent need for high-speed optical interfaces. Leveraging the low-cost of IM-DD schemes and the efficiency of advanced modulation formats such as PAM-4 and PAM-8, we have presented several approaches to achieve 100-Gb/s capacities and beyond. The high-speed electrical signals were generated with an integrated SP-DAC, which allows doubling the input symbol rate while producing up to 8-level pulse amplitude modulation. This way today's available 25-28G electronics can be directly inserted into the SP-DAC, producing in this case up to 56-GBd PAM-8 signals (168 Gb/s signals). Furthermore, this circuit outputs enough voltage to directly drive a modulator, hence avoiding the use of amplifier drivers, which simplifies the transmitter design.

In Chapter 3 we first demonstrated a 112-Gb/s (100 Gb/s net data rate) EML transceiver including a DFB and a 50-GHz bandwidth EAM co-packaged in a compact module. Such device is capable of successfully transmitting PAM-4 56-GBd signals over a distance of 2-km with a large range of possible receiver configurations, from low-bandwidth (down to 18 GHz) to low equalization-complexity (down to 3 taps). Later we scale up capacity by introducing more advanced modulation schemes and/or higher symbol-rates, this time using commercially available 40G components (MZM modulator in the transmitter and PIN-TIA in the receiver). We present two possible implementations for a 168-Gb/s (150-Gb/s net data rate) transceiver: PAM-4 84-GBd and PAM-8 56-GBd. PAM-4 84-GBd signaling exhibits the

best performance in terms of BER and sensitivity, offering a limited 1-km reach due to chromatic dispersion distortions. On the other hand, PAM-8 56-GBd, denoting lower perfomances below 1-km, present an extended 2-km reach, thanks to its thinner spectrum, which offers lower accumulation of chromatic dispersion distortions. Finally we reported a 100-GBd PAM-4 interface, capable of transmitting 200-Gb/s (178.5-Gb/s net data rate). We demonstrated that using a low-complex equalization (less than 11 taps) combined with a simple 2-symbol MLSD detection scheme, 500-m reach 200-Gb/s transceivers can be deployed using commercially available 40G components.

Notwithstanding, although current data centers scale up capacity by increasing port-count and lane-speeds of their switches, these two cannot increase arbitrarily due to fundamental limitations of the switch design, described in Chapter 1. Furthermore bandwidth demand is growing faster than per-switch capacities in large service providers. Therefore, large scale data centers increase their capacities by enlarging their networks (increasing the number of networks components and switching stages). We have shown in Chapter 1 that this approach leads to high operation and management costs, due to the high complexity of such vast networks, high power consumption, led by the large number of components performing electronic switching, and increased latency, given by the several switching stages that needs to surpass every single server-to-server communication.

In Chapter 4 we introduced a novel data center topology, based on burst optical slot switching (BOSS) rings, which aims at addressing the aforementioned issues. In such topology the electronic switching network is replaced by a mesh of BOSS nodes which are interconnected through optically transparent rings organized in a torus topology, such that any-to-any node connection can be performed with a single O/E/O conversion, which takes place when changing rings. Data exchanged between nodes is statistically multiplexed in both wavelength and time microsecond-long slots. The combination of high-data rate transponders, statistical multiplexing and optical transparency allows for a massive reduction of networking components ($\times$100-500 fewer interfaces and cables) and large savings in energy consumption ($\times$2-3) with respect to current electronic switching networks.

We tackled in Chapter 4 the implementation of BOSS nodes from a physical optical layer perspective. We proposed the use of coherent transponders aiming for highest possible capacity, and explored the best

modulation approaches given the possible impairments occurring in a large cascade of nodes (50-100 nodes), required to ensure scalability to very large data centers. We showed that the node impairments are extremely dependent on the technology used to build the slot blocker (device capable of blocking packets from any wavelength for its later reuse). We first disregarded the use of SOAs as optical gates for large node-counts. Despite their relevance for amplification, SOAs induce large nonlinear distortions in large cascades, limiting the reach and capacity. Hence, we use VOAs in later prototypes, which do not induce nonlinear distortions and can be integrated with other components, leading to monolithically integrated slot blockers in Silicon platform.

Then we evaluated the distortions introduced by different kinds of wavelength D/MUX and proposed accordingly appropriate modulation schemes. We have shown that, using flexible N-QAM transponders, we can achieve very large average data rates (above 100 or even 200 Gb/s) while supporting large node-count (100 and 50 nodes, respectively) when high-end (WSS-based) 100-GHz-grid D/MUXs are used. Nevertheless, when aiming for low-cost D/MUXs (AWG-based), the reach of N-QAM transponders is limited to 70 nodes due to the extremely the tight accumulated filtering response of such devices. Furthermore a 30% of average capacity is lost when comparing to the performance achieved with high-end D/MUXs, even when using the spectrally efficient Nyquist pulse-shaping (NPS). In order to overcome such limiting impairments, we proposed the use of CO-OFDM, which can optimize sub-carrier bit-loading to precisely adapt to any kind of spectral impairments. When comparing NPS-N-QAM and CO-OFDM traversing AWG-based nodes, the latter achieves 30% increased average data rate, 40% extended reach and 80% larger detuning tolerance. Furthermore, CO-OFDM proves even superior in low-cost environments than flexible N-QAM in high-end ones, which allows using low-cost D/MUX with no extra penalty.

5.2 Perspectives

It is clear that data centers are not ready today to move towards optical switching solutions, due to their lack of maturity, cost uncertainties and the requirement of a disruptive change of the totality of the data center infrastructure. Hence, along the following decade, data centers will keep

increasing their capacity by leveraging higher speeds interfaces and electrical switches appearing in the market. It is certain that the 50-Gb/s lane rates will appear soon in the market, which will be used to build first 200-Gb/s (4 lanes) then 400-Gb/s (8 lanes) interfaces. Nevertheless within a few years, improvements in both electronic and photonic components will allow for 100-Gb/s per lane, which might be then used in a quad-lane approach to build 400 Gb/s interfaces. Beyond that, 800-Gb/s and beyond 1-Tb/s interfaces are at the early research stage. While 800-Gb/s could be in principle realized with a single +200-GBd signal with PAM-4 modulation, as we demonstrated in this thesis, a practical implementation is very challenging. Hence, beyond 400-Gb/s high parallelization might be implemented at first, by using fiber ribbons or WDM. With massive production capabilities and the possibility of integrating electronics and photonics on a single chip, Silicon integration is expected to lower the cost of optical interfaces to unprecedented costs/bit. Silicon may allow for the integration of many lanes (i.e., 4, 8, 16 or more) on a single compact chip, which could be multiplexed in the wavelength domain requiring a single fiber/pair (hence diminishing packaging cost). Such approach could lead to 1.6 Tb/s capacities in a 16×100-Gb/s chip.

Nevertheless, as aforementioned, switching capacity per chip starts presenting critical limitations due to I/O chip density. Hence eventually, electronic switching will lead to a dead-end, which will open the doors to optical switching if technology is sufficiently mature. The work presented in this thesis is already a quite futuristic solution, which might not be implemented at first. We believe that the transition to optical switching should be done with a lower complexity solution. This means using IM-DD highly-integrated high-bandwidth Silicon transceivers, avoiding possibly the use of slot blockers.

The solution presented in this thesis results into further complexity but allows for much higher capacity. First the use of a slot blocker in a ring-based network allows for slot reuse, which increases fiber utilization. Nevertheless, slot blocker design needs to be improved. In the current architecture an optical gate is required per wavelength, which results into a large number of control pads (up to 160 when using the full C-Band (80 channels) with polarization diversity requiring 2 pads/channel), which not only increases packaging cost but also hinders chip designs. Furthermore, the use of AWG-like D/MUXs severly impairs the signals in large rings. Hence we are already working on an improved design, which requires much

fewer control pads while reducing filtering impairments.

The ultimate capacity can be achieved using coherent transponders which increase by a 4-fold the spectral efficiency of IM-DD transceivers. Silicon-based fully integrated coherent transceivers are being demonstrated today, which in the future might result into the lowest cost per bit. One needs to realize that the most consuming (more than 50%) block of a coherent DSP is the chromatic dispersion compensation, which is not required in data centers due to the short reach. Therefore, thinking that now optics evolution is more driven by data centers than by long-haul systems, and that future IM/DD systems will also include an ASIC to perform FEC en/decoding and possibly equalization, a short-reach low-cost coherent transceiver optimized for data center applications may appear. This would allow maximizing capacity per transceiver, per wavelength, and hence per fiber, which would lead to a drastic reduction on the number of networking components (transponders, slot blockers and fibers). Nonetheless, high advances in both electronic and photonic integration are required to effectively increase data center capacities while reducing the exorbitant cost and power consumption present today in data centers.

List of publications

International peer-reviewed journals: (6)

[J1] **M. A. Mestre**, J.-M. Estarán, P. Jennevé, H. Mardoyan, I. Tafur Monroy, D. Zibar and S. Bigo, "Novel coherent optical OFDM-based transponder for optical slot switched networks," in *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 8, pp. 1851-1858, April 2016 **(invited)**.

[J2] **M. A. Mestre**, H. Mardoyan , C. Caillaud, R. Rios-Müller, J. Renaudier, P. Jennevé, F. Blache, F. Pommereau, J. Decobert, F. Jorge, P. Charbonnier, A. Konczykowska, J.-Y. Dupuy, K. Mekhazni, J.-F. Paret, M. Faugeron, F. Mallecot, M. Achouche and S. Bigo, "Compact InP-based DFB-EAM enabling PAM-4 112 Gb/s transmission over 2km," in *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 7, pp. 1572-1578, April 2016 **(invited)**.

[J3] H. Mardoyan, **M. A. Mestre**, R. Rios-Müller, A. Konczykowska, J. Renaudier, F. Jorge, B. Duval, J-Y. Dupuy, A. Ghazisaeidi,P. Jennevé, M. Achouche, S. Bigo, "Single Carrier 168-Gbit/s Line-Rate PAM direct detection transmission using high-speed Selector Power DAC for Optical Interconnects," in *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 7, pp. 1593-1598, April 2016 **(invited)**.

[J4] G. de Valicourt, Y. Pointurier, **M. A. Mestre**, J.-M. Fédéli, K. Ribaud, L. Bramerie, E. Borgne, J.-C. Simon, L. Vivien, D. Marris-Morini, A. Shen, I. Ghorbel, G. H. Duan, S. Chandrasekhar, C.-M. Chang, S. Randel, Y. -K. Chen, "Monolithic integrated slot-blocker for high datarate coherent optical slot switched networks," in *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 8, pp. 1807-1814, April 2016 **(invited)**.

[J5] P. Kaspar, G. de Valicourt, R. Brenot, **M. A. Mestre**, P. Jen-

nevé, A. Accard, D. Make, F. Lelarge, G.-H. Duan, N. Pavarelli, M. Rensing, C. Eason, P. O'Brien, S. Olivier, S. Malhouitre, C. Kopp, C. Jany, S. Menezo, "Hybrid III-V/Silicon SOA in Optical Network Based on Advanced Modulation Formats," in *IEEE Photonics Technology Letters*, IEEE , vol. 27, no. 22, pp. 2383-2386, November 2015.

[J6] G. de Valicourt, H. Mardoyan, **M. A. Mestre**, P. Jennevé, J.-C. Antona, S. Bigo, O. Bertran-Pardo, C. Kazmierski, N. Chimot, F. Blache, "Monolithic Integrated InP Transmitter Using Switching of Prefixed Optical Phases," in *IEEE/OSA Journal of Lightwave Technology*, vol. 33, no. 3, pp. 663-669, February 2015 **(invited)**.

International peer-reviewed conferences: (16)

Postdeadline papers: (4)

[PDP1] **M. A. Mestre**, H. Mardoyan, A. Konczykowska, R. Rios-Müller, J. Renaudier, F. Jorge, B. Duval, J-Y. Dupuy, A. Ghazisaeidi, P. Jennevé, S. Bigo, "Direct Detection Transceiver at 150-Gbit/s Net Data Rate Using PAM 8 for Optical Interconnects," in Proc. *European Conference on Optical Communication (ECOC)*, September 2015, **Postdeadline paper** 2.4.

[PDP2] C. Caillaud, **M. A. Mestre**, F. Blache, F. Pommereau, J. Decobert, F. Jorge, P. Charbonnier, A. Konczykowska, J.-Y. Dupuy, H. Mardoyan, K. Mekhazni, J.-F. Paret, M. Faugeron, F. Mallecot and M. Achouche, "Low cost 112 Gb/s InP DFB-EAM for PAM-4 2 km Transmission," in Proc. *European Conference on Optical Communication (ECOC)*, September 2015, **Postdeadline paper** 1.5.

[PDP3] G. de Valicourt, **M. A. Mestre**, P. Jennevé, H. Mardoyan, J. C. Antona, S. Bigo, O. Bertran-Pardo, C. Kazmierski, N. Chimot, F. Blache, A. Garreau, "Ultra-Compact Monolithic Integrated InP Transmitter at 224 Gb/s with PDM-2ASK-2PSK modulation," in Proc. *Optical Fiber Communication Conference (OFC)*, March 2014, **Postdeadline paper** Th5C.3.

[PDP4] H. Mardoyan, O. Bertran-Pardo, P. Jennevé, G. de Valicourt, M.A. Mestre, S. Bigo, C. Kazmierski, N. Chimot, A.G. Steffan, J. Honecker, R. Zhang, P. Runge, A. Richter, C. Arellano, A. Ortega-Moñux and I. Molina-Fernandez, "PIC-to-PIC experiment at

130Gb/s Based on a Monolithic Transmitter Using Switching of Prefixed Optical Phases and a Monolithic Coherent Receiver," in Proc. *Optical Fiber Communication Conference (OFC)*, March 2014, **Postdeadline paper** Th5C.2.

Regular papers: (12)

[R1] **M. A. Mestre**, F. Jorge, H. Mardoyan, J.-M. Estarán, F. Blache, P. Angelini, A. Konczykowska, M. Riet, V. Nodjiadjim, J-Y. Dupuy, S. Bigo, "100-Gbaud PAM-4 Intensity-Modulation Direct-Detection Transceiver for Datacenter Interconnect," in Proc. *European Conference on Optical Communication (ECOC)*, September 2016, paper M.2.C.2 **(upgraded to invited)**.

[R2] R. Rios-Müller, J. Renaudier, **M. A. Mestre**, H. Mardoyan, A. Konczykowska, F. Jorge, B. Duval, J-Y. Dupuy, "Multi-Dimension Coded PAM4 Signaling for 100 Gb/s Short-Reach Transceivers", in Proc. *Optical Fiber Communication Conference (OFC)*, March 2016, paper Th1G.4.

[R3] J. Renaudier, R. Rios-Müller, **M. A. Mestre**, H. Mardoyan, A. Konczykowska, F. Jorge, B. Duval, J-Y. Dupuy, "Multi Rate IMDD Transceivers for Optical Interconnects Using Coded Modulation", in Proc. *Optical Fiber Communication Conference (OFC)*, March 2016, paper Tu2J.2.

[R4] J. Estaran, **M.A. Mestre**, P. Jennevé, H. Mardoyan, I. T. Monroy, D. Zibar, S. Bigo, "Coherent optical orthogonal frequency-division multiplexing for optical slot switched intra-datacenter networks", in Proc. *European Conference on Optical Communication (ECOC)*, September 2015, paper Tu.1.2.3 **(Highly scored paper)**.

[R5] Y. Pointurier, B. Uscumlic, **M.A. Mestre**,P. Jennevé, H. Mardoyan, A. Dupas, and S. Bigo, "Green Optical Slot Switching Torus for Mega-Datacenters," in Proc. *European Conference on Optical Communication (ECOC)*, September 2015, paper Tu.3.6.4.

[R6] G. de Valicourt, S. Chandrasekhar, J. H. Sinsky, C-M. Chang, Y. K. Chen, **M. A. Mestre**, Y. Pointurier, S. Bigo, J.-M. Fédéli, L. Bramerie, J.-C. Simon, L. Vivien, A. Shen, A. Le liepvre, G. H. Duan, "Monolithic Integrated Reflective Polarization Diversity SOI-based Slot-Blocker for Fast Reconfigurable 128Gb/s and 256 Gb/s Optical Networks," in Proc. *European Conference on Optical Commu-*

nication (ECOC), September 2015, paper Tu.3.5.4 **(Highly scored paper)**.

[R7] **M.A. Mestre**, P. Jennevé, H. Mardoyan, A. Ghazisaeidi, S. Bigo, G. de Valicourt, "Experimental evalution of SOA cascadability for optical packet switched networks," in Proc. *Optical Fiber Communication Conference (OFC)*, March 2015, paper Th2A.2.

[R8] H. Mardoyan, R. Rios-Müller, **M.A. Mestre**, P. Jennevé, L. Schmalen, A. Ghazisaeidi, P. Tran, S. Bigo, J. Renaudier, "Transmission of Single-Carrier Nyquist-Shaped 1-Tb/s Line-Rate Signal over 3,000 km," in Proc. *Optical Fiber Communication Conference (OFC)*, March 2015, paper W3G.2 **(Highly scored paper)**.

[R9] **M. A. Mestre**, G. de Valicourt, P. Jennevé, H. Mardoyan, S. Bigo, Y. Pointurier, "Optical Slot Switching-Based Datacenters With Elastic Burst-Mode Coherent Transponders," in Proc. *European Conference on Optical Communication (ECOC)*, September 2014, paper Th.2.2.3.

[R10] G. de Valicourt, **M. A. Mestre**, L. Bramerie, J.-C. Simon, E. Borgne, L. Vivien, E. Cassan, D. Marris-Morini, J.-M. Fédéli, P. Jennevé, H. Mardoyan, Y. Pointurier, A. Le Liepvre, G. H. Duan, A. Shen, S. Bigo, "Monolithic Integrated Silicon-based Slot-Blocker for Packet-Switched Networks," in Proc. *European Conference on Optical Communication (ECOC)*, September 2014, paper We.3.5.5 **(Highly scored paper)**.

[R11] G. de Valicourt, **M. A. Mestre**, J.-C. Antona, P. Jennevé, H. Mardoyan, S. Bigo, C. Kazmierski, N. Chimot, F. Blache , "Integrated Non-Quadrature Intensity Modulation Transmitter Based on Prefixed Optical Phases and intensity modulations," in Proc. *European Conference on Optical Communication (ECOC)*, September 2014, paper Tu4.4.3 **(Highly scored paper)**.

[R12] P. Jennevé, P. Ramantanis, J.-C. Antona, G. de Valicourt, **M. A. Mestre**, H. Mardoyan, S. Bigo, "Pitfalls of Error Estimation from Measured Non-Gaussian Nonlinear Noise Statistics over Dispersion-Unmanaged Systems," in Proc. *European Conference on Optical Communication (ECOC)*, September 2014, paper Mo.4.3.3.

National conferences: (2)

[N1] **M. A. Mestre**, G. de Valicourt, P. Jennevé, H. Mardoyan, S. Bigo, Y. Pointurier, "Centre de données a base de paquets optiques et de transpondeurs élastiques," in Proc. *Journées Nationales d'Optique Guidée (JNOG)*, October 2014, pp. 243-245.

[N2] P. Jennevé, P. Ramantanis, J.-C. Antona, G. de Valicourt, **M. A. Mestre**, H. Mardoyan, S. Bigo, "Caractérisation expérimentale et extension du modèle de bruit gaussien pour les systèmes monocanal non gérés en dispersion," in Proc. *Journées Nationales d'Optique Guidée (JNOG)*, October 2014, pp. 240-242.

Book chapters: (1)

[B1] G. de Valicourt, **M. A. Mestre**, N. D. Moroz, Y. Pointurier, "Semiconductor optical amplifiers for next generation of high data rate optical packet-switched networks," in "Some advanced functionalities of optical amplifiers," Intech, 2015. ISBN-978-953-51-2237-1.

Patent applications: (5)

[P1] **M. A. Mestre**, Y. Pointurier, G. de Valicourt,"Integrated slot blocker with reduced filtering effect. To be filed to European Patent Office.

[P2] B. Uscumlic, **M. A. Mestre**,"Virtual Fully transparent optical packet network. Filed to European Patent Office, August 2016.

[P3] P. Jennevé, **M. A. Mestre**, A. Morea,"Method for improving signal transmission quality in an optical network associated equipment. Filed to European Patent Office, June 2015.

[P4] **M. A. Mestre**, P. Jennevé, A. Morea,"Method for optimizing an optical network by analysis of statistical values. Filed to European Patent Office, May 2015.

[P5] **M. A. Mestre**, Y. Pointurier, G. de Valicourt,"Optical packet drop structure and associated node. Filed to European Patent Office, April 2015.

Acronyms

ADC analog-to-digital converter. 34, 39, 40, 48, 50, 58, 65, 66, 67, 66, 67, 68, 67, 68, 69, 78, 90, 104, 107, 118

APD avalanche photodiode. 39

ARCNET Attached Resource Computer NETwork. 1

ASE amplified spontaneous emission. 95, 98

AWG arrayed-waveguide grating. 86, 92, 94, 102, 103, 102, 103, 104, 112, 113, 114, 115, 123, 125, 127, 128, 129, 135

AWGR array waveguide grating router. 86

BER bit error rate. 41, 53, 55, 61, 66, 67, 68, 67, 68, 69, 72, 73, 76, 78, 79, 82, 83, 97, 107

BGA ball grid array. 19

BM burst-mode. 91, 92, 93, 94

BOSS Burst Optical Slot Switching. 23, 24, 87, 88, 89, 90, 91, 92, 94, 101, 102, 103, 102, 104, 106, 108, 110, 111, 113, 114, 115, 116, 117, 119, 120, 124, 127, 128, 133

BPD balanced photodetectors. 47

BtB back-to-back. 66, 68, 69, 73, 76, 78, 79, 82, 96, 97, 104, 106, 107, 108, 110, 115, 120, 127

BTC block-turbo codes. 53

CAP carrierless amplitude/phase modulation. 34, 35, 58

CD chromatic dispersion. 48, 50, 121

CDR clock and data recovery. 61

CFP C form factor. 13, 14

CMA constant-modulus algorithm. 51, 55

CMOS complementary metal-oxide-semiconductor. 59

CO-OFDM coherent-optical orthogonal frequency-division multiplexing. 24, 104, 113, 114, 115, 117, 119, 122, 124, 125, 126, 127, 129, 134

CP cyclic prefix. 120, 121

CWDM coarse wavelength-division multiplexing. 13, 14

D/MUX wavelength de/multiplexer. 92, 94, 101, 102, 104, 112, 113, 115, 126, 127, 129, 134, 135

DAC digital-to-analog converter. 23, 34, 35, 38, 40, 44, 45, 48, 58, 59, 60, 78, 104, 107, 120, 122, 124

DC direct current. 120, 121

DFB distributed feedback. 12, 36, 37, 44, 58, 61, 62, 63, 69, 72, 82, 132

DFB-EAM integrated distributed feedback laser - electro-absorption modulator. 63, 65, 71, 72, 73

DML directly modulated laser. 34

DMT discrete multitone. 34, 35, 38, 58

DP dual-polarization. 44, 45

DS-DBR digital supermode distributed Bragg reflector. 44, 92

DSO digital storage oscilloscope. 65, 74, 96, 123

DSP digital signal processing. 34, 35, 39, 40, 43, 44, 45, 48, 58, 96, 106, 119, 121, 136

EAM electro-absorption modulator. 30, 37, 38, 58, 61, 62, 63, 64, 69, 71, 82, 132

ECL external-cavity laser. 44, 122, 123

EDFA erbium-doped fiber amplifier. 25, 52, 108, 123

EML externally modulated laser. 34, 35, 37, 44, 61, 82, 132

ENOB effective number of bits. 58, 107

EPS electronic packet switching. 27, 29, 30, 85, 86

ER extinction ratio. 64

FE Fast Ethernet. 9

FEC forward error correction. 14, 15, 41, 53, 61, 62, 66, 67, 73, 95, 100, 101, 107, 109, 120, 124, 131

FFE feed-forward equalization. 41, 50, 66, 69, 72, 73, 76, 79, 82

FFT Fast Fourier Transform. 121

FIR finite impulse response. 41, 49, 50, 51, 55

FPT frequency pilot tone. 117, 118, 119, 120, 121, 122

GE Gigabit Ethernet. 9

GSMBE gas source molecular beam epitaxy. 63

HD hard-decision. 53, 61, 66, 100, 101

IaaS Infrastructure as a Service. 5

ICI inter-carrier interference. 118, 119

IFFT Inverse Fast Fourier Transform. 120

IM-DD intensity-modulation and direct-detection. 25, 26, 27, 26, 32, 34, 35, 37, 38, 40, 46, 50, 51, 58, 70, 78, 79, 82, 83, 90, 132, 135, 136

InGaAlAs Indium Galium Aluminium Arsenide. 62, 63

InGaAs Indium Galium Arsenide. 63

InGaAsP Indium Galium Arsenide Phosphide. 63, 64

InP Indium Phosphide. 37, 38, 61, 62, 71, 92

ISI inter-symbol interference. 33, 48, 50, 78, 79, 80

LAN local area network. 1, 9

LC liquid crystal. 102

LCoS liquid-crystal on Silicon. 96, 102

LDPC low-density parity-check. 53

$LiNbO_3$ Lithium Niobate. 37, 38, 71

LMS least mean square. 41, 55

LO local oscillator. 123

LR long reach. 12, 13, 14, 19, 20

LS least square. 55

LUT look-up-table. 81, 80, 81

MEMS micro-electronic-mechanical systems. 86, 102

MIMO multi-input multi-output. 50

MLSD maximum-likelihood sequence detector. 79, 81, 80, 81, 82, 83, 132

MLSD maximum likelyhood sequence detector. 82

MMA multi-modulus algorithm. 51, 55

MMF multi-mode fiber. 12, 19, 25, 36

MMSE minimum mean square error. 55, 106

MPO multi-fiber push on. 13

MSA multi-source agreement. 13, 14

MZM Mach-Zehnder modulator. 37, 38, 37, 38, 45, 61, 72, 73, 79, 83, 92, 95, 132

NIC network interface controller. 11

NPS Nyquist pulse-shaped. 106, 109, 110, 111, 112, 113, 114, 122, 124, 125, 126, 127, 129

NRZ non-return-to-zero. 104, 106, 108, 109, 110, 111, 114, 129

O/E/O opto-electro-optic. 27, 29, 87

OCS optical circuit switching. 27, 28, 29, 30, 86

OFDM orthogonal frequency-division multiplexing. 114, 115, 116, 117, 118, 119, 120, 119, 120, 122

OH over-head. 120

OOK on-off keying. 58

OPS optical packet switching. 27, 30, 86, 87

OSNR optical signal-to-noise ratio. 55, 95, 97, 98, 100, 101, 107, 109, 110, 115, 121, 124

PaaS Platform as a Service. 5

PAM pulse amplitude modulation. 38, 58

PAM-8 8-level pulse amplitude modulation. 33, 58, 59, 60, 61, 60, 70, 71, 73, 74, 76, 78, 82, 83, 132

PAM-2 2-level pulse amplitude modulation. 14, 15, 32, 33, 41, 42, 60, 61, 60, 82

PAM-4 4-level pulse amplitude modulation. 14, 15, 23, 33, 41, 57, 58, 59, 60, 61, 60, 61, 62, 65, 70, 72, 73, 74, 76, 78, 79, 80, 82, 83, 132

PAPR peak-to-average power ratio. 119

PBC polarization beam combiner. 45

PBS polarization beam splitter. 46

PC personal computer. 1

PDM polarization division multiplexing. 44, 55, 58, 96, 97, 98, 99, 100, 101, 104, 107, 108, 109, 110, 120, 122, 124, 126, 127

PIN-TIA PIN photodetector with transimpedance amplifier. 65, 66, 70, 72, 79, 132

PLC planar lightwave circuit. 102

PLL phase locked loop. 40, 50

PON passive optical network. 26

PPG pulse pattern generator. 74

PRBS pseudo-random bit sequence. 72

PS power splitter. 46

PSR pilot-to-signal ratio. 121, 122

QAM quadrature amplitude modulation. 24, 43, 51, 52, 55, 95, 96, 97, 98, 100, 104, 107, 109, 110, 111, 112, 113, 114, 124, 129, 134

QPSK quadrature phase-shift keying. 42, 43, 44, 51, 52, 95, 96, 97, 98, 100, 101, 104, 107, 108, 109, 110, 124, 126, 127

QSFP quad small form factor. 13, 14, 15

RIN relative intensity noise. 63, 62

ROADM reconfigurable optical add-drop multiplexer. 27, 29

RX receiver. 91, 92

SaaS Software as a Service. 5

SC sub-carrier. 120

SD soft-decision. 53, 100, 101, 107, 109

SE spectral efficiency. 115

SFP small form factor. 13

SMF single-mode fiber. 12, 19, 25, 36, 50

SNR signal-to-noise ratio. 121

SOA semiconductor optical amplifier. 30, 37, 86, 87, 92, 94, 95, 96, 97, 98, 99, 100, 99, 100, 101, 129, 133

SP-DAC selector power digital-to-analog converter. 58, 60, 61, 65, 70, 71, 72, 73, 74, 78, 79, 82, 132

SPM self-phase modulation. 119

SR short reach. 12, 13, 14, 19, 20

SSMF standard single mode fiber. 65, 72, 76, 96, 104

TDM time division multiplexing. 60

TE transverse electric. 44

TIA trans-impedance amplifier. 11, 39, 48, 57, 70

TM transverse magnetic. 44

ToR Top-of-Rack. 6, 11, 16, 90

TWC tunable wavelength converter. 86

TX transmitter. 91, 92, 93

VCSEL vertical cavity surface-emitting laser. 12, 36

VGC vertical grating coupler. 128

VNA vector network analyzer. 64

VOA variable optical attenuator. 30, 65, 72, 92, 95, 96, 101, 102, 128, 129, 133

WAN wide area network. 6, 8

WDM wavelength-division multiplexing. 25, 31, 44, 62, 86, 134

WSS wavelength-selective switch. 27, 93, 94, 102, 103, 102, 103, 104, 108, 109, 110, 112, 113, 125, 129

XPM cross-phase modulation. 119

Bibliography

[1] P. E. Ceruzzi, *A History of Modern Computing*. MIT press, 2015. [Online]. Available: http://english.360elib.com/datu/T/EM290636.pdf

[2] G. Lee, *Cloud Networking: Understanding Cloud-Based Data Center Networks*. Morgan Kaufmann, 2014.

[3] M. K. Weldon, *The Future X Network: A Bell Labs Perspective*. CRC Press, 2016.

[4] *The encyclopedia of information technology*. Atlantic, 2007.

[5] Netcraft, "Netcraft web server survey," 2009. [Online]. Available: http://news.netcraft.com/archives/2016/07/19/july-2016-web-server-survey.html

[6] Cisco, "Cisco Global Cloud Index: Forecast and Methodology, 2014-2019," 2015. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.html#wp9000816

[7] ——, "Cisco Global Cloud Index: Forecast and Methodology, 2010-2015," 2011. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.html#wp9000816

[8] ——, "Cisco Global Cloud Index: Forecast and Methodology, 2011-2016," 2012. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.html#wp9000816

[9] ——, "Cisco Global Cloud Index: Forecast and Methodology, 2012-2017," 2013. [Online]. Available:

http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.html#wp9000816

[10] ——, "Cisco Global Cloud Index: Forecast and Methodology, 2013-2018," 2013. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.html#wp9000816

[11] P. Mell and T. Grance, "SP 800-145, The NIST definition of cloud computing," 2011. [Online]. Available: http://dx.doi.org/10.6028/NIST.SP.800-145

[12] A. Singh, P. Germano, A. Kanagala, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart, A. Vahdat, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, and B. Felderman, "Jupiter rising: a decade of clos topologies and centralized control in Google's datacenter network," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4, pp. 183–197, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2785956.2787508

[13] A. Hammadi and L. Mhamdi, "A survey on architectures and energy efficiency in Data Center Networks," *Computer Communications*, vol. 40, pp. 1–21, 2014.

[14] M. Al-Fares, A. Loukissas, A. Vahdat, M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proc. Special Interest Group on Data Communication Conference (SIGCOMM)*, vol. 38, no. 4, 2008, pp. 63–74. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1402958.1402967

[15] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "VL2: a scalable and flexible data center network," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4, pp. 51–62, 2009. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1594977.1592576

[16] N. Farrington and A. Andreyev, "Facebook's data center network architecture," in *Optical Interconnects Conference*, 2013, pp. 49–50. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6552917

[17] C. DeCusatis, *Handbook of Fiber Optic Data Communication*. Academic Press, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124016736000064

[18] A. Andreyev, "Introducing data center fabric, the next-generation Facebook the next generation datacenter network," 2014. [Online]. Available: https://code.facebook.com/posts/360346274145943/introducing-data-center-fabric-the-next-generation

[19] C. Decusatis, "Optical interconnect networks for data communications," *IEEE/OSA Journal of Lightwave Technology*, vol. 32, no. 4, pp. 544–552, 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6584001

[20] Cisco, "The Future Is 40 Gigabit Ethernet," 2016. [Online]. Available: http://www.cisco.com/c/dam/en/us/products/collateral/switches/catalyst-6500-series-switches/white-paper-c11-737238.pdf

[21] Broadcom and Intel, "40G Ethernet Market Potential," *IEEE 802.3 HSSG April 2007 Interim Meeting*, 2007. [Online]. Available: http://www.ieee802.org/3/hssg/public/apr07/hays_01_0407.pdf

[22] IEEE, "IEEE Standard for Information technology– Local and metropolitan area networks– Specific requirements– Part 3: CSMA/CD Access Method and Physical Layer Specifications Amendment 5: Media Access Control Parameters, Physical Layers, and Management Paramet," pp. 1–302, 2010. [Online]. Available: http://ieeexplore.ieee.org/servlet/opac?punumber=5621023

[23] IDC, "Worldwide Ethernet Switch Market and Enterprise and Service Provider Router Market Post Positive Results for the Fourth Quarter and Full Year 2015," 2016. [Online]. Available: http://www.idc.com/getdoc.jsp?containerId=prUS41061316

[24] "25G Ethernet consortium," 2014. [Online]. Available: http://25gethernet.org

[25] IEEE, "IEEE P802.3by 25 Gb/s Ethernet Task Force," 2016. [Online]. Available: http://www.ieee802.org/3/by/index.html

[26] B. Smith, D. Chalupsky, and M. Nowell, "The 2016 Ethernet Roadmap," in *Open Server Summit (OSS)*, 2016, p. B103. [Online].

Available: http://www.openserversummit.com/English/Collaterals/Proceedings/2016/20160413_SB103_Kipp_panelists.pdf

[27] IEEE, "IEEE 802.3cd 50 Gb/s, 100 Gb/s, and 200 Gb/s Ethernet Task Force." [Online]. Available: http://www.ieee802.org/3/cd/index.html

[28] "IEEE P802.3bs 400 Gb/s Ethernet Task Force." [Online]. Available: http://www.ieee802.org/3/bs/index.html

[29] HP, "10 Gigabit Ethernet," Tech. Rep., 2006. [Online]. Available: http://www.hp.com/rnd/pdfs/10gig_cabling_technical_brief.pdf

[30] Finisar, "Cabling in the Data Center," 2016. [Online]. Available: https://www.finisar.com/markets/data-center/cabling-data-center

[31] L. Huff, "Data center optics, ECOC market focus," in *Proc. European Conference on Optical Communication (ECOC)*, 2013. [Online]. Available: http://www.ecocexhibition.com/sites/default/files/files/ECOCMarketFocusDataCenterOptics_DA.pdf

[32] J. D. Ambrosia and S. G. Kipp, "The 2015 Ethernet Roadmap," p. 15, 2015. [Online]. Available: http://www.ethernetalliance.org/wp-content/uploads/2015/03/Ethernet-Roadmap-2sides-Final-27April.pdf

[33] IEEE and ANSI, "IEEE Std 802.3ba Media Access Control Parameters, Physical Layers, and Management Parameters for 40 Gb/s and 100 Gb/s Operation," 2010. [Online]. Available: http://standards.ieee.org/findstds/standard/802.3ba-2010.html

[34] Finisar, "Product guide: Transceivers, Transponders and Active Optical Cables," Tech. Rep., 2015. [Online]. Available: https://www.finisar.com/sites/default/files/resources/finisar_optical_transceiver_product_guide_3_2015_web.pdf

[35] ——, "Optical Transceivers," 2016. [Online]. Available: https://www.finisar.com/optical-transceivers

[36] Arista, "Optical modules and cables," Tech. Rep., 2016. [Online]. Available: https://www.arista.com/assets/data/pdf/Datasheets/Transceiver-Data-Sheet.pdf

[37] Cisco, "Pluggable Optical Modules: Transceivers for the Cisco ONS Family," 2016. [Online]. Available: http://www.cisco.com/c/en/us/products/collateral/optical-networking/ons-15454-series-multiservice-provisioning-platforms/brochure_c02-452560.html

[38] J. A. Tatum, "The evolution of 850nm VCSELs from 10Gb/s to 25 and 56Gb/s," in *Proc. Optical Fiber Communication Conference (OFC)*, 2014, p. Th3C.1. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2014-Th3C.1

[39] A. Weissberger, "Hyper Scale Mega Data Centers: Time is Now for Fiber Optics to the Compute Server," 2016. [Online]. Available: http://community.comsoc.org/blogs/alanweissberger/hyper-scale-mega-data-centers-time-now-fiber-optics-compute-server

[40] Netberg, "Aurora 720 100G with ONIE," 2016. [Online]. Available: http://netbergtw.com/products/aurora-720/

[41] S. Kipp, "Ethernet Alliance 100GbE Challenges," 2014. [Online]. Available: http://www.ethernetalliance.org/wp-content/uploads/2013/04/Ethernet-Alliance-100GbE-Challenges-09-16-14.pdf

[42] Broadcom, "High-Density 25/100 Gigabit Ethernet StrataXGS® Tomahawk Ethernet Switch Series." [Online]. Available: https://www.broadcom.com/products/Switching/Data-Center/BCM56960-Series

[43] Mellanox, "SN2700: Spectrum-based 32-port 100GbE Open Ethernet Platform," 2016. [Online]. Available: http://www.mellanox.com/related-docs/prod_eth_switches/PB_SN2700.pdf

[44] H. J. S. Dorren, E. H. M. Wittebol, R. De Kluijver, G. Guelbenzu De Villota, P. Duan, and O. Raz, "Challenges for optically enabled high-radix switches for data center networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 33, no. 5, pp. 1117–1125, 2015.

[45] A. Ghiasi, "Is there a need for on-chip photonic integration for large data warehouse switches," in *Proc. 9th International Conference on Group IV Photonics (GFP)*, 2012, pp. 27–29.

[46] Datacenterknowledge, "How much Does Facebook Spend on Its Data Centers?" 2016. [Online]. Available: http://www.datacenterknowledge.com/the-facebook-data-center-faq-page-three/

[47] A. Vahdat, H. Liu, X. Zhao, and C. Johnson, "The Emerging Optical Data Center," in *Proc. Optical Fiber Communication Conference (OFC)*, 2011, p. OTuH2. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2011-OTuH2

[48] Y. Sverdlik, "Here's How Much Energy All US Data Centers Consume," 2016. [Online]. Available: http://www.datacenterknowledge.com/archives/2016/06/27/heres-how-much-energy-all-us-data-centers-consume/

[49] A. Greenberg, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Towards a next generation data center architecture: scalability and commoditization," in *Proc. ACM workshop on Programmable routers for extensible services of tomorrow*, 2008, pp. 57–62.

[50] Mellanox, "SN2100: Half-Width 16-port Non-blocking 100GbE Open Ethernet Switch System," Tech. Rep., 2016. [Online]. Available: https://www.mellanox.com/related-docs/prod_eth_switches/PB_SN2100.pdf

[51] ——, "SX1410: 48-port 10GbE + 12-port 40/56GbE SDN Switch System," Tech. Rep., 2016. [Online]. Available: https://www.mellanox.com/related-docs/prod_eth_switches/PB_SX1410.pdf

[52] ——, "SX6018: 18-port Non-blocking Managed 56Gb/s InfiniBand/VPI SDN Switch System," Tech. Rep., 2016. [Online]. Available: http://www.mellanox.com/related-docs/prod_ib_switch_systems/SX6018.pdf

[53] Rich Miller, "Google Data Center FAQ," 2012. [Online]. Available: http://www.datacenterknowledge.com/google-data-center-faq-part-2/

[54] U. S. Energy Information Administration, "Average Price of Electricity to Ultimate Customers by End-Use Sector," 2016. [Online]. Available: https://www.eia.gov/electricity/monthly/epm_table_grapher.cfm?t=epmt_5_6_a

[55] C. Guo, L. Yuan, D. Xiang, Y. Dang, R. Huang, D. Maltz, Z. Liu, V. Wang, B. Pang, H. Chen, and Others, "Pingmesh: A large-scale system for data center network latency measurement and analysis," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4, pp. 139–152, 2015.

[56] G. P. Agrawal, “Fiber-Optic Communication Systems,” 2002.

[57] A. Gnauck, R. W. Tkach, A. R. Chraplyvy, and T. Li, “High-Capacity Optical Transmission Systems,” *IEEE/OSA Journal of Lightwave Technology*, vol. 26, no. 9, pp. 1032–1045, 2008. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-26-9-1032

[58] Nokia, “PSE-2 super coherent technology,” 2016. [Online]. Available: https://networks.nokia.com/products/pse-2-super-coherent-technology

[59] K. I. Kitayama, Y. C. 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,” *IEEE/OSA Journal of Lightwave Technology*, vol. 33, no. 5, pp. 1063–1071, 2015.

[60] W. Miao, J. Luo, S. Di Lucente, H. Dorren, and N. Calabretta, “Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system,” *Optics Express*, vol. 22, no. 3, p. 2465, 2014. [Online]. Available: http://www.opticsinfobase.org/abstract.cfm?URI=oe-22-3-2465

[61] W. Yan, L. Li, B. Liu, H. Chen, Z. Tao, T. Tanaka, T. Takahara, J. C. Rasmussen, and T. Drenski, “80 km IM-DD transmission for 100 Gb/s per lane enabled by DMT and nonlinearity management,” in *Proc. Optical Fiber Communication Conference (OFC)*, 2014, p. M2I.4. [Online]. Available: https://www.osapublishing.org/abstract.cfm?uri=OFC-2014-M2I.4

[62] M. I. Olmedo, T. Zuo, J. B. Jensen, Q. Zhong, X. Xu, S. Popov, and I. T. Monroy, “Multiband carrierless amplitude phase modulation for high capacity optical data links,” *IEEE/OSA Journal of Lightwave Technology*, vol. 32, no. 4, pp. 798–804, 2014.

[63] H. Yamazaki, M. Nagatani, S. Kanazawa, H. Nosaka, T. Hashimoto, A. Sano, and Y. Miyamoto, “160-Gbps Nyquist PAM4 transmitter using a digital-preprocessed analog-multiplexed DAC,” in *Proc. European Conference on Optical Communication (ECOC)*. IEEE, 2015, p. PDP1013. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7341681

[64] L. Tao, Y. Ji, J. Liu, A. Tao Lau, N. Chi, and C. Lu, "Advanced modulation formats for short reach optical communication systems," *IEEE Network*, vol. 27, no. 6, pp. 6–13, 2013.

[65] D. Mahgerefteh, C. Thompson, C. Cole, G. Denoyer, T. Nguyen, I. Lyubomirsky, C. Kocot, and J. Tatum, "Techno-Economic Comparison of Silicon Photonics and Multimode VCSELs," *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 2, pp. 233–242, 2016. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-34-2-233

[66] E. Haglund, P. Westbergh, J. S. Gustavsson, E. P. Haglund, A. Larsson, M. Geen, and A. Joel, "30 GHz bandwidth 850 nm VCSEL with sub-100 fJ/bit energy dissipation at 25-50 Gbit/s," *Electronics Letters*, vol. 51, no. 14, pp. 1096–1098, 2015.

[67] Z. Zhang, J. liu, Y. Liu, J. Guo, H. Yuan, J. Bai, and N. zhu, "30-GHz directly modulation DFB laser with narrow linewidth," in *Proc. Asia Communications and Photonics Conference (ACP)*, 2015, p. AM1B.3. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=ACPC-2015-AM1B.3

[68] D. Kuchta, A. V. Rylyakov, C. L. Schow, J. Proesel, C. Baks, P. Westbergh, J. S. Gustavsson, and A. Larsson, "64Gb/s Transmission over 57m MMF using an NRZ Modulated 850nm VCSEL," in *Proc. Optical Fiber Communication Conference (OFC)*, 2014, p. Th3C.2. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2014-Th3C.2

[69] Neophotonics, "56G EML," 2016. [Online]. Available: https://www.neophotonics.com/product/56g-eml/

[70] S. Kanazawa, H. Yamazaki, Y. Nakanishi, T. Fujisawa, and K. Takahata, "Transmission of 214-Gbit/s 4-PAM signal using an ultra- broadband lumped-electrode EADFB laser module," in *Proc. Optical Fiber Communication Conference (OFC)*. Anaheim, C.A.: OSA, 2016, p. Th5B.3. [Online]. Available: https://www.osapublishing.org/abstract.cfm?URI=OFC-2016-Th5B.3

[71] Fujitsu, "40 Gb/s NRZ LiNbO external modulator," 2005. [Online]. Available: http://www.fujitsu.com/downloads/OPTCMP/lineup/40gln/40Glnnrz-catalog.pdf

[72] G. Letal, K. Prosyk, R. Millett, D. Macquistan, S. Paquet, O. Thibault-Maheu, J.-F. Gagné, P.-L. Fortin, R. Dowlatshahi, B. Rioux, T. Spring Thorpe, M. Hisko, R. Ma, and I. Woods, "Low Loss InP C-Band IQ Modulator with 40GHz Bandwidth and 1.5V Vpi," in *Proc. Optical Fiber Communication Conference (OFC)*, 2015, p. Th4E.3. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2015-Th4E.3

[73] M. Y. S. Sowailem, T. M. Hoang, M. Morsy-Osman, M. Chagnon, D. Patel, S. Paquet, C. Paquet, I. Woods, O. Liboiron-Ladouceur, and D. V. Plant, "400-G Single Carrier 500-km Transmission With an InP Dual Polarization IQ Modulator," *IEEE Photonics Technology Letters*, vol. 28, no. 11, pp. 1213–1216, 2016.

[74] A. Samani, M. Chagnon, D. Patel, V. Veerasubramanian, S. Ghosh, M. Osman, Q. Zhong, and D. V. Plant, "A Low-Voltage 35-GHz Silicon Photonic Modulator-Enabled 112-Gb/s Transmission System," *IEEE Photonics Journal*, vol. 7, no. 3, pp. 1–13, 2015.

[75] C. Doerr, L. Chen, D. Vermeulen, T. Nielsen, S. Azemati, S. Stulz, G. McBrien, X. M. Xu, B. Mikkelsen, M. Givehchi, C. Rasmussen, and S. Y. Park, "Single-chip silicon photonics 100-Gb/s coherent transceiver," in *Proc. Optical Fiber Communication Conference (OFC)*, 2014, p. TH5C.1.

[76] D. Patel, S. Ghosh, M. Chagnon, A. Samani, V. Veerasubramanian, M. Osman, and D. V. Plant, "Design, analysis, and transmission system performance of a 41 GHz silicon photonic modulator," *Optics Express*, vol. 23, no. 11, pp. 14 263–14 287, 2015. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-23-11-14263

[77] H. Huang, S. R. Nuccio, Y. Yue, J. Y. Yang, Y. Ren, C. Wei, G. Yu, R. Dinu, D. Parekh, C. J. Chang-Hasnain, and A. E. Willner, "Broadband Modulation Performance of 100-GHz EO Polymer MZMs," *IEEE/OSA Journal of Lightwave Technology*, vol. 30, no. 23, pp. 3647–3652, 2012.

[78] P. Groumas, Z. Zhang, V. Katopodis, A. Konczykowska, J. Y. Dupuy, A. Beretta, A. Dede, J. H. Choi, P. Harati, F. Jorge, V. Nodjiadjim, M. Riet, R. Dinu, G. Cangini, E. Miller, A. Vannucci, N. Keil, H. G. Bach, N. Grote, M. Spyropoulou, H. Avramopoulos, and

C. Kouloumentas, "Tunable 100 Gbaud Transmitter Based on Hybrid Polymer-to-Polymer Integration for Flexible Optical Interconnects," *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 2, pp. 407–418, 2016.

[79] M. Trajkovic, W. Yao, H. Debregeas, F. Blanche, K. A. Williams, and X. J. M. Leijtens, "High speed electroabsorption modulator in the generic photonic integration platform," *Symposium of the IEEE Photonics Society Benelux*, 2015.

[80] Fujitsu, "Fujitsu Oola/Rotta," 2016. [Online]. Available: http://www.fujitsu.com/cn/en/products/devices/semiconductor/fsp/asic/asic/ipmacro/networkingips/

[81] Socionext, "Socionext takes part in Record-breaking Transmission Field Trial of 38.4Tbps over 762 kilometers," 2015. [Online]. Available: https://www.socionext.com/en/pr/sn_pr20150630_01e.pdf

[82] Micram, "VEGA UltraFast Signal Converters," 2016. [Online]. Available: http://micram.net/products/vega-signal-converters/

[83] J. Godin, V. Nodjiadjim, M. Riet, P. Berdaguer, O. Drisse, E. Derouin, A. Konczykowska, J. Moulu, J. Y. Dupuy, F. Jorge, J. L. Gentner, A. Scavennec, T. Johansen, and V. Krozer, "Submicron InP DHBT technology for high-speed high-swing mixed-signal ICs," in *IEEE CSIC Symposium: GaAs ICs Celebrate 30 Years in Monterey*, 2008, pp. 1–4.

[84] O. Kharraz and D. Forsyth, "Performance comparisons between PIN and APD photodetectors for use in optical communication systems," *Optik-International Journal for Light and Electron Optics*, vol. 124, no. 13, pp. 1493–1498, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0030402612002859

[85] Finisar, "40 GHz Single High-speed photoreceiver," 2016. [Online]. Available: https://www.finisar.com/optical-components/xprv2021a

[86] K. Vasilakopoulos, S. P. Voinigescu, P. Schvan, P. Chevalier, and A. Cathelin, "A 92GHz bandwidth SiGe BiCMOS HBT TIA with less than 6dB noise figure," in *Bipolar/BiCMOS Circuits and Technology Meeting - BCTM, 2015 IEEE*, 2015, pp. 168–171.

[87] E. Bloch, H. C. Park, Z. Griffith, M. Urteaga, D. Ritter, and M. J. W. Rodwell, "A 107 GHz 55 dB-Ohm InP Broadband Transimpedance Amplifier IC for High-Speed Optical Communication Links," in *IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)*, 2013.

[88] L. Kull, T. Toifl, M. Schmatz, P. A. Francese, C. Menolfi, M. Braendli, M. Kossel, T. Morf, T. M. Andersen, and Y. Leblebici, "22.1 A 90GS/s 8b 667mW 64 x interleaved SAR ADC in 32nm digital SOI CMOS," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2014, pp. 378–379.

[89] J. D. H. Alexander, "Clock recovery from random binary signals," *IEEE Electronics Letters*, vol. 11, no. 22, pp. 541–542, 1975.

[90] K. Mueller and M. Muller, "Timing Recovery in Digital Synchronous Data Receivers," *IEEE Transactions on Communications*, vol. 24, no. 5, pp. 516–531, 1976.

[91] C. Hogge, "A self correcting clock recovery curcuit," *IEEE/OSA Journal of Lightwave Technology*, vol. 3, no. 6, pp. 1312–1314, 1985.

[92] Inphi and Broadcom, "FEC Codes for 400 Gbps 802.3bs." [Online]. Available: http://www.ieee802.org/3/bs/public/14_11/parthasarathy_3bs_01a_1114.pdf

[93] J. E. Simsarian, J. Gripp, S. Chandrasekhar, and P. Mitchell, "Fast-tuning coherent burst-mode receiver for metropolitan networks," *IEEE Photonics Technology Letters*, vol. 26, no. 8, pp. 813–816, 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6746025

[94] G. de Valicourt, Y. Pointurier, M. A. Mestre, S. Bigo, J. M. Fedeli, K. Ribaud, L. Bramerie, E. Borgne, J. C. Simon, L. Vivien, D. Marris-Morini, A. Shen, I. Ghorbel, G. H. Duan, S. Chandrasekhar, C. M. Chang, S. Randel, and Y. K. Chen, "Monolithic integrated slot-blocker for high datarate coherent optical slot switched networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 8, pp. 1807–1814, 2016.

[95] S. J. Savory, "Digital filters for coherent optical receivers," *Optics Express*, vol. 16, no. 2, pp. 804–817, 2008. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-16-2-804

[96] F. Gardner, "A BPSK/QPSK Timing-Error Detector for Sampled Receivers," *IEEE Transactions on Communications*, vol. 34, no. 5, pp. 423–429, 1986.

[97] D. Godard, "Passband Timing Recovery in an All-Digital Modem Receiver," *IEEE Transactions on Communications*, vol. 26, no. 5, pp. 517–523, 1978. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1094107

[98] S. J. Savory, "Digital Coherent Optical Receivers: Algorithms and Subsystems," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 16, no. 5, pp. 1164–1179, 2010.

[99] P. J. Winzer, A. H. Gnauck, C. R. Doerr, M. Magarini, and L. L. Buhl, "Spectrally Efficient Long-Haul Optical Networking Using 112-Gb/s Polarization-Multiplexed 16-QAM," *IEEE/OSA Journal of Lightwave Technology*, vol. 28, no. 4, pp. 547–556, 2010. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-28-4-547

[100] A. Viterbi, "Nonlinear estimation of PSK-modulated carrier phase with application to burst digital transmission," *IEEE Transactions on Information Theory*, vol. 29, no. 4, pp. 543–551, 1983.

[101] J. H. Ke, K. P. Zhong, Y. Gao, J. C. Cartledge, A. S. Karar, and M. A. Rezania, "Linewidth-Tolerant and Low-Complexity Two-Stage Carrier Phase Estimation for Dual-Polarization 16-QAM Coherent Optical Fiber Communications," *IEEE/OSA Journal of Lightwave Technology*, vol. 30, no. 24, pp. 3987–3992, 2012.

[102] T. Pfau, S. Hoffmann, and R. Noe, "Hardware-Efficient Coherent Digital Receiver Concept With Feedforward Carrier Recovery for M -QAM Constellations," *IEEE/OSA Journal of Lightwave Technology*, vol. 27, no. 8, pp. 989–999, 2009.

[103] D. A. Morero, M. A. Castrillón, A. Aguirre, M. R. Hueda, and O. E. Agazzi, "Design Tradeoffs and Challenges in Practical Coherent Optical Transceiver Implementations," *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 1, pp. 121–136, 2016.

[104] F. Vacondio, C. Simonneau, A. Voicila, E. Dutisseuil, J.-M. Tanguy, J.-C. Antona, G. Charlet, and S. Bigo, "Real time implementation of packet-by-packet polarization demultiplexing in

a 28 Gb/s burst mode coherent receiver," in *Proc. Optical Fiber Communication Conference (OFC)*. Optical Society of America, 2012, p. OM3H.6. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2012-OM3H.6

[105] T. M. Schmidl and D. C. Cox, "Robust frequency and timing synchronization for OFDM," *IEEE Transactions on Communications*, vol. 45, no. 12, pp. 1613–1621, 1997.

[106] K. Shi and E. Serpedin, "Coarse frame and carrier synchronization of OFDM systems: A new metric and comparison," *IEEE Transactions on Wireless Communications*, vol. 3, no. 4, pp. 1271–1284, 2004. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1310316

[107] M. Kuschnerov, M. Chouayakh, K. Piyawanno, B. Spinnler, E. De Man, P. Kainzmaier, M. S. Alfiad, A. Napoli, and B. Lankl, "Data-aided versus blind single-carrier coherent receivers," *IEEE Photonics Journal*, vol. 2, no. 3, pp. 387–403, 2010. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5447711

[108] W. Hartmann, M. Lauermann, S. Wolf, H. Zwickel, Y. Kutuvantavida, J. Luo, A. K. Y. Jen, W. Freude, and C. Koos, "100 Gbit/s OOK using a silicon-organic hybrid (SOH) modulator," in *Proc. European Conference on Optical Communication (ECOC)*, 2015, p. PDP.1.4. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7341678

[109] J. Lee, N. Kaneda, T. Pfau, A. Konczykowska, F. Jorge, J. Y. Dupuy, and Y. K. Chen, "Serial 103.125-Gb/s transmission over 1 km SSMF for low-cost, short-reach optical interconnects," in *Proc. Optical Fiber Communication Conference (OFC)*. San Francisco, L.A.: OSA, 2014, p. Th5A.5. [Online]. Available: https://www.osapublishing.org/abstract.cfm?uri=OFC-2014-Th5A.5

[110] M. Chagnon, M. Morsy-Osman, M. Poulin, C. Paquet, S. Lessard, and D. V. Plant, "Experimental parametric study of a silicon photonic modulator enabled 112-Gb/s PAM transmission system with a DAC and ADC," *IEEE/OSA Journal of Lightwave Technology*, vol. 33, no. 7, pp. 1380–1387, 2015.

[111] T. K. Chan and W. I. Way, "112 Gb/s PAM4 transmission over 40km SSMF using 1.3 Âţm gain-clamped semiconductor optical amplifier," in *Proc. Optical Fiber Communication Conference (OFC)*, 2015, p. Th3A.4. [Online]. Available: https://www.osapublishing.org/abstract.cfm?uri=OFC-2015-Th3A.4

[112] S. Kanazawa, T. Fujisawa, K. Takahata, T. Ito, Y. Ueda, W. Kobayashi, H. Ishii, and H. Sanjoh, "Flip-Chip Interconnection Lumped-Electrode EADFB Laser for 100-Gb/s/λ Transmitter," *IEEE Photonics Technology Letters*, vol. 27, no. 16, pp. 1699–1701, 2015. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7113803

[113] S. Shahramian, J. Lee, J. Weiner, R. Aroca, Y. Baeyens, N. Kaneda, and Y.-k. Chen, "A 112Gb/s 4-PAM Transceiver Chipset in 0.18µm SiGe BiCMOS Technology for Optical Communication Systems," in *IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)*. IEEE, oct 2015. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7314464

[114] M. Chagnon, M. Osman, D. Patel, V. Veerasubramanian, A. Samani, and D. Plant, "1 λ, 6 bits/symbol, 280 and 350 Gb/s Direct Detection Transceiver using Intensity Modulation, Polarization Multiplexing, and Inter-Polarization Phase Modulation," in *Proc. Optical Fiber Communication Conference (OFC)*, 2015, p. Th5B.2. [Online]. Available: http://www.osapublishing.org/abstract.cfm?uri=OFC-2015-Th5B.2

[115] M. Morsy-Osman, M. Chagnon, and D. V. Plant, "Four-dimensional modulation and stokes direct detection of polarization division multiplexed intensities, inter polarization phase and inter polarization differential phase," *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 7, pp. 1585–1592, 2016.

[116] M. A. Mestre, H. Mardoyan, C. Caillaud, R. Rios-Müller, J. Renaudier, P. Jennevé, F. Blache, F. Pommereau, J. Decobert, F. Jorge, P. Charbonnier, A. Konczykowska, J. Y. Dupuy, K. Mekhazni, J. F. Paret, M. Faugeron, F. Mallecot, M. Achouche, and S. Bigo, "Compact InP-Based DFB-EAM Enabling PAM-4 112 Gb/s Transmission over 2 km," *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 7, pp. 1572–1578, 2016.

[117] C. Caillaud, M. A. Mestre, F. Blache, F. Pommereau, J. Decobert, F. Jorge, P. Charbonnier, A. Konczykowska, J. Y. Dupuy, H. Mardoyan, K. Mekhazni, J. F. Paret, M. Faugeron, F. Mallecot, and M. Achouche, "Low cost 112 Gb/s InP DFB-EAM for PAM-4 2 km transmission," in *Proc. European Conference on Optical Communication (ECOC)*, 2015, p. PDP1.5. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7341679

[118] M. A. Mestre, H. Mardoyan, A. Konczykowska, J. Renaudier, and F. Jorge, "Direct Detection Transceiver at 150-Gbit / s Net Data Rate Using PAM 8 for Optical Interconnects," in *Proc. Optical Fiber Communication Conference (OFC)*, 2015, p. PDP2.4.

[119] H. Mardoyan, M. A. Mestre, S. Member, R. Rios-Müller, A. Konczykowska, J. Renaudier, F. Jorge, B. Duval, J.-y. Dupuy, A. Ghazisaeidi, and P. Jennevé, "Single Carrier 168-Gb/s Line-Rate PAM Direct Detection Transmission Using High-Speed Selector Power DAC for Optical Interconnects," *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 7, pp. 1593–1598, 2016.

[120] M. A. Mestre, F. Jorge, H. Mardoyan, J. M. Estaran, F. Blache, F. Angelini, A. Konczykowska, M. Riet, V. Nodjiadjim, J.-Y. Dupuy, and S. Bigo, "100-Gbaud PAM-4 Intensity-Modulation Direct-Detection Transceiver for Datacenter Interconnect," in *Proc. European Conference on Optical Communication (ECOC)*, 2016, p. M.2.C.2.

[121] R. Rios-Müller, J. Renaudier, P. Brindel, C. Simonneau, P. Tran, A. Ghazisaeidi, I. Fernandez, L. Schmalen, and G. Charlet, "Optimized spectrally efficient transceiver for 400-Gb/s single carrier transport," in *Proc. European Conference on Optical Communication (ECOC)*, 2014, p. PD.4.2. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6964270

[122] S. Randel, D. Pilori, S. Corteselli, G. Raybon, A. Adamiecki, A. Gnauck, S. Chandrasekhar, P. Winzer, L. Altenhain, A. Bielik, and R. Schmid, "All-electronic flexibly programmable 864-Gb/s single-carrier PDM-64-QAM," in *Proc. Optical Fiber Communication Conference (OFC)*, 2014, p. Th5C.8. [Online]. Available: https://www.osapublishing.org/abstract.cfm?uri=OFC-2014-Th5C.8

[123] J.-Y. Dupuy, M. Riet, A. Konczykowska, V. Nodjiadjim, F. Jorge, H. Aubry, and A. Adamiecki, "84 GBd (168 Gbit/s) PAM-4 3.7 Vpp power DAC in InP DHBT for short reach and long haul optical networks," *IEEE Electronics Letters*, vol. 51, no. 20, pp. 1591–1593, 2015. [Online]. Available: http://digital-library.theiet.org/content/journals/10.1049/el.2015.2316

[124] M. Faugeron, M. Tran, O. Parillaud, M. Chtioui, Y. Robert, E. Vinet, A. Enard, J. Jacquet, and F. Van Dijk, "High-power tunable dilute mode DFB laser with low RIN and narrow linewidth," *IEEE Photonics Technology Letters*, vol. 25, no. 1, pp. 7–10, 2013.

[125] K. Szczerba, P. Westbergh, J. Karout, J. S. Gustavsson, Å. Haglund, M. Karlsson, P. A. Andrekson, E. Agrell, and A. Larsson, "4-PAM for high-speed short-range optical communications," *Journal of Optical Communications and Networking*, vol. 4, no. 11, pp. 885–894, 2012.

[126] Yenista, "OSICS DFB DWDM: Distributed Feedback Laser," 2016. [Online]. Available: https://yenista.com/IMG/pdf/OSICS-DFB-DWDM_DS_2v2-5.pdf

[127] M. Le Pallec, J. Decobert, C. Kazmierski, A. Ramdane, N. El Dabdah, F. Blache, J.-G. Provost, J. Landreau, D. Carpentier, F. Barthe, and N. Lagay, "42 GHz bandwidth InGaAlAs/InP electro absorption modulator with a sub-volt modulation drive capability in a 50 nm spectral range," in *Proc. International Conference on Indium Phosphide and Related Materials (IPRM).*, 2004, pp. 577–580. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1442791

[128] T. Fujisawa, T. Yamanaka, T. Tadokoro, N. Fujiwara, M. Arai, W. Kobayashi, Y. Kawaguchi, K. Tsuzuki, and F. Kano, "Theoretical and experimental investigation of the incident-power-dependent extinction ratio of an electroabsorption modulator integrated with a distributed feedback laser," *IEEE Journal of Quantum Electronics*, vol. 47, no. 1, pp. 60–65, 2011. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5662957

[129] Broadcom, "FEC Structures for 400GbE Supporting Multi-PMDs," 2014. [Online]. Available: http://www.ieee802.org/3/bs/public/adhoc/logic/oct21_14/wangz_01_1014_logic.pdf

[130] 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 Computer Communication Review*, vol. 40, no. 4, pp. 339–350, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1851275.1851223

[131] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. Ng, M. Kozuch, and M. Ryan, "c-Through: Part-time optics in data centers," in *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4, 2011, pp. 327–338.

[132] X. Ye., Y. Yin., S.J.B.Yoo., P. Mejia., R.Proietti., and V. Akella., "DOS: A Scalable Optical Switch for Datacenters," in *Proc. ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, vol. 4, 2010.

[133] K. Xi, Y.-H. Kao, and H. J. Chao, "A petabit bufferless optical switch for data center networks," in *Optical Interconnects for Future Data Center Networks*. Springer, 2013, pp. 135–154.

[134] R. Proietti, Z. Cao, C. J. Nitta, Y. Li, and S. J. B. Yoo, "Interconnect Architecture Based on Arrayed Waveguide Grating Routers," *IEEE/OSA Journal of Lightwave Technology*, vol. 33, no. 4, pp. 911–920, 2015.

[135] S. Peng, D. Simeonidou, G. Zervas, R. Nejabati, Y. Yan, Y. Shu, S. Spadaro, J. Perello, F. Agraz, D. Careglio, H. Dorren, W. Miao, N. Calabretta, G. Bernini, N. Ciulli, J. C. Sancho, S. Iordache, Y. Becerra, M. Farreras, M. Biancani, A. Predieri, R. Proietti, Z. Cao, L. Liu, and S. J. B. Yoo, "A novel SDN enabled hybrid optical packet/circuit switched data centre network: The LIGHTNESS approach," in *Proc. European Conforence on Networks and Communications (EuCNC)*, 2014, pp. 1–5.

[136] N. Benzaoui and Y. Pointurier, "Impact of tunability and blocking fabric on optical slot switching ring performance," in *Proc. International Conference on Transparent Optical Networks (ICTON)*, 2016.

[137] M. A. Mestre, G. de Valicourt, P. Jennevé, H. Mardoyan, S. Bigo, and Y. Pointurier, "Optical Slot Switching-Based Datacenters With

Elastic Burst-Mode Coherent Transponders," in *Proc. European Conference on Optical Communication (ECOC)*, 2014, p. Th.2.2.3.

[138] Y. Pointurier, B. Uščumlić, M. A. Mestre, P. Jennevé, H. Mardoyan, A. Dupas, and S. Bigo, "Green Optical Slot Switching Torus for Mega-Datacenters," in *Proc. European Conference on Optical Communication (ECOC)*, 2015, p. Tu.3.6.4.

[139] Infinera, "Taking metro 100G to the next level," 2016. [Online]. Available: https://www.infinera.com/wp-content/uploads/2015/11/infinera-an-low-power-xtm-series-100g-metro-solution.pdf

[140] Acacia, "Metro," 2016. [Online]. Available: http://acacia-inc.com/applications/metro/

[141] D. Thomson, A. Zilkie, J. E. Bowers, T. Komljenovic, G. T. Reed, L. Vivien, D. Marris-Morini, E. Cassan, L. Virot, J.-M. Fédéli, and Others, "Roadmap on silicon photonics," *Journal of Optics*, vol. 18, no. 7, p. 73003, 2016.

[142] Y. Pointurier, G. de Valicourt, J. E. Simsarian, J. Gripp, and F. Vacondio, "High data rate coherent optical slot switched networks: a practical and technological perspective," *IEEE Communications Magazine*, vol. 53, no. 8, pp. 124–129, 2015.

[143] H.-W. Chen, J. D. Peters, and J. E. Bowers, "Forty Gb/s hybrid silicon Mach-Zehnder modulator with low chirp." *Optics express*, vol. 19, no. 2, pp. 1455–1460, 2011. [Online]. Available: https://www.osapublishing.org/oe/abstract.cfm?uri=oe-19-2-1455

[144] W. Zhang, L. Xu, Q. Li, H. L. R. Lira, M. Lipson, and K. Bergman, "Broadband silicon photonic packet-switching node for large-scale computing systems," *IEEE Photonics Technology Letters*, vol. 24, no. 8, pp. 688–690, apr 2012. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6145740

[145] G. de Valicourt, M. A. Mestre, L. Bramerie, J. Simon, E. Borgne, and L. Vivien, "Monolithic Integrated Silicon-based Slot-Blocker for Packet-Switched Networks," in *Proc. European Conference on Optical Communication (ECOC)*, 2014, p. We.3.5.5.

[146] G. de Valicourt, N. D. Moroz, P. Jennevé, F. Vacondio, G. H. Duan, C. Jany, A. Leliepvre, A. Accard, and J.-C. Antona,

"A next-generation optical packet-switching node based on hybrid III-V/silicon optical gates," *IEEE Photonics Technology Letters*, vol. 26, no. 7, pp. 678–681, 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6722911

[147] R. Bonk, G. Huber, T. Vallaitis, S. Koenig, R. Schmogrow, D. Hillerkuss, R. Brenot, F. Lelarge, G.-H. Duan, S. Sygletos, C. Koos, W. Freude, and J. Leuthold, "Linear semiconductor optical amplifiers for amplification of advanced modulation formats." *Optics Express*, vol. 20, no. 9, pp. 9657–72, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22535057

[148] H. Schmuck, R. Bonk, W. Poehlmann, C. Haslach, W. Kuebart, D. Karnick, J. Meyer, D. Fritzsche, E. Weis, J. Becker, W. Freude, and T. Pfeiffer, "Demonstration of an SOA-assisted open metro-access infrastructure for heterogeneous services," *Optics Express*, vol. 22, no. 1, pp. 737–748, 2014. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-22-1-737

[149] M. A. Mestre, P. Jennevé, H. Mardoyan, A. Ghazisaeidi, S. Bigo, and G. de Valicourt, "On the SOA cascadability and design rules for optical packet-switched networks," in *Proc. Optical Fiber Communication Conference (OFC)*. Optical Society of America, 2015, p. Th2A.2.

[150] E. Grellier and A. Bononi, "Quality parameter for coherent transmissions with Gaussian-distributed nonlinear noise," *Optics Express*, vol. 19, no. 13, pp. 12 781–12 788, 2011. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-19-13-12781

[151] T. A. Strasser and J. L. Wagener, "Wavelength-Selective Switches for ROADM Applications," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 16, no. 5, pp. 1150–1157, 2010.

[152] C. Pulikkaseril, L. A. Stewart, M. a. F. Roelens, G. W. Baxter, S. Poole, and S. Frisken, "Spectral modeling of channel band shapes in wavelength selective switches." *Optics Express*, vol. 19, no. 9, pp. 8458–70, 2011.

[153] M. A. Mestre, J. M. Estaran, P. Jennevé, H. Mardoyan, I. Tafur Monroy, D. Zibar, and S. Bigo, "Novel coherent optical OFDM-based transponder for optical slot switched networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 8, pp. 1851–1858, 2016.

[154] J. Estarán, M. A. Mestre, P. Jennevé, H. Mardoyan, and I. Tafur Monroy, "Coherent Optical Orthogonal Frequency-Division Multiplexing for Optical Slot Switched Intra-Datacenters Networks," in *Proc. European Conference on Optical Communication (ECOC)*, 2015, p. Tu.1.2.3.

[155] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, 2003. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1193803

[156] K. S. Al-Mawali, A. Z. Sadik, and Z. M. Hussain, "Simple Discrete Bit-loading for OFDM Systems in Power Line Communications," in *Proc. IEEE International Symposium on Power Line Communications and Its Applications (ISPLC)*, 2011, pp. 267–270. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5764405

[157] P. Krummrich and K. Kotten, "Extremely fast (microsecond timescale) polarization changes in high speed long haul WDM transmission systems," in *Proc. Optical Fiber Communication Conference (OFC)*, 2004, p. FI3.

[158] X. Liu and F. Buchali, "Intra-symbol frequency-domain averaging based channel estimation for coherent optical OFDM." *Optics Express*, vol. 16, no. 26, pp. 21 944–21 957, 2008. [Online]. Available: https://www.osapublishing.org/oe/abstract.cfm?uri=oe-16-26-21944

[159] S. Randel, S. Adhikari, and S. L. Jansen, "Analysis of RF-pilot-based phase noise compensation for coherent optical OFDM systems," *IEEE Photonics Technology Letters*, vol. 22, no. 17, pp. 1288–1290, 2010. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5491074

[160] S. L. Jansen, I. Morita, T. C. W. Schenk, and N. Takeda, "Coherent Optical 25.8-Gb / s OFDM Transmission Over 4160-km SSMF," *IEEE/OSA Journal of Lightwave Technology*, vol. 26, no. 1, pp. 6–15, 2008.

[161] G. Bosco, A. Carena, V. Curri, P. Poggiolini, and F. Forghieri, "Performance limits of Nyquist-WDM and CO-OFDM in high-speed PM-

QPSK systems," *IEEE Photonics Technology Letters*, vol. 22, no. 15, pp. 1129–1131, 2010.

[162] S. L. Jansen, I. Morita, K. Forozesh, S. Randel, D. Van Den Borne, and H. Tanaka, "Optical OFDM, a hype or is it for real?" in *Proc. European Conference on Optical Communication (ECOC)*, 2008, p. Mo.3.E.3. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4729133

[163] B. Park, H. Cheon, C. Kang, and D. Hong, "A novel timing estimation method for OFDM systems," *IEEE Communications Letters*, vol. 7, no. 5, pp. 239–241, 2003. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1200195

[164] X. Liu and F. Buchali, "A novel channel estimation method for PDM-OFDM enabling improved tolerance to WDM nonlinearity," in *Proc. Optical Fiber Communication Conference (OFC)*, 2009, p. OWW5. [Online]. Available: https://www.osapublishing.org/abstract.cfm?uri=OFC-2009-OWW5

[165] T. Kobayashi, A. Sano, A. Matsuura, Y. Miyamoto, and K. Ishihara, "Nonlinear tolerant spectrally-efficient transmission using PDM 64-QAM single carrier FDM with digital pilot-tone," *IEEE/OSA Journal of Lightwave Technology*, vol. 30, no. 24, pp. 3805–3815, 2012.

[166] G. de Valicourt, S. Chandrasekhar, J. H. Sinsky, C.-M. Chang, Y. K. Chen, and M. A. Mestre, "Monolithic Integrated Reflective Polarization Diversity SOI-based Slot-Blocker for Fast Reconfigurable 128 Gb/s and 256 Gb/s Optical Networks," in *Proc. European Conference on Optical Communication (ECOC)*, 2015, p. Tu.3.5.4.