Reliable and Cost Efficient Passive Optical Interconnects for Data Centers

Yuxin Cheng, Matteo Fiorani, Lena Wosinska, and Jiajia Chen

***Abstract*—To address the sustainability, scalability, and reliability problems that data centers are currently facing, we propose three passive optical interconnect (POI) architectures on top of the rack. The evaluation results show that all three architectures offer high reliability performance (connection availability for intra-rack interconnections higher than 99.999%) in a cost-efficient way.**

***Index Terms*—Optical communications, data center interconnects, reliability analysis, cost models.**

I. Introduction

THE growing popularity of cloud applications is drastically increasing the traffic volumes that data centers have to handle [1]. Consequently, the transmission capacity inside the data centers is rapidly growing. The majority of the servers today are equipped with 1 Gbps or 10 Gbps interfaces for communications, while in the future higher transmission rates are expected to be used (e.g., 40 Gbps or 100 Gbps per server) [2]. These trends lead to scalability problem for the network providing connectivity among severs inside the data center and toward the Internet.

Current data center interconnection networks include several tiers, such as edge, aggregation and core, and are based on electronic commodity switches. Scaling these networks to support very high transmission capacity may lead to dramatic increase in the total equipment cost and power consumption [3]. In this regard, optical communication has been considered as a promising technology for data center interconnects due to the ultra-high capacity that can be offered in cost- and energy-efficient way.

Several optical switching architectures have been recently proposed for data center networks [4]–[7]. They are based on either optical switches [4], [5] or passive optical components [6] and [7] at the aggregation and core tiers. However, in these architectures the edge tier is still based on electronic top-of-rack (ToR) switches, which limits the overall cost and energy savings [3]. Paper [8] has explored different possibilities for optical interconnection solutions at ToR and identified that using POI for inter-server communication (such as the architectures proposed in [6], [7]) can potentially offer significant energy saving while at the relatively low cost.

Meanwhile, the higher transmission rate, the larger the volume of traffic and number of cloud services can be affected in case of a failure in the network. The required availability of fault-tolerant data center infrastructure (including electrical power supply, storage and distribution facilities) should be higher than 99.995% [9]. Thus, the availability for any connection established within the data center needs to be even higher, since the communication system is only a part of the site infrastructure. Several topologies, e.g., fat-tree [10] and Quartz [11], have been proposed in order to improve the resiliency of large-scale data center networks. These topologies introduce redundancy in the aggregation and core tiers to increase reliability in the central part of the data center network. However, the edge tier is usually unprotected due to the high cost of introducing redundant ToR switches as well as due to the belief that edge tier can be self-healing (i.e., in case the connection to a certain server would be down, the task can be re-assigned and carried out by another server). Unfortunately, it may not be true in the scenario where the servers within the racks are highly loaded making it difficult to find resources to be allocated for a possible backup.

Therefore, the expected growth of traffic volume inside the data centers brings the need for highly reliable, yet cost and energy efficient, interconnection at the edge tier. In [3] and [8] several passive optical interconnects for the edge tier of data center networks have been presented showing that by replacing the electronic ToR switches with passive optical components it is possible to significantly reduce the overall energy consumption while offering high capacity interconnection.

On the other hand, such optical interconnects may lead to higher capacities needed in the aggregation and core tiers of the data center network because of the lack of statistical multiplexing in the optical domain. However, this problem can be mitigated by employing burst mode transceivers and a control protocol that is able to perform an efficient dynamic bandwidth allocation strategy [6], [7].

Despite of the growing importance of fault tolerance of interconnects at ToR, this aspect has not been studied yet. In this letter, we focus on highly reliable and cost efficient passive optical interconnects and propose three architectures for interconnections at ToR. They can be integrated with any topology supporting large-scale data center networks, e.g., fat-tree and Quartz. We also evaluate the proposed architectures in terms of connection availability and cost. Our results verify that ultra-high connection availability, i.e., close to 5 nines (99.999%), can be achieved with both 1 Gbps and 10 Gbps server interfaces. In addition, the cost of the proposed passive optical interconnects scale more efficiently with the server capacity compared to electronic commodity switches.

II. Reliable Passive Optical Interconnects

In this section, three passive optical interconnects (POIs) for the edge tier are presented. The first one is based on an arrayed waveguide grating (AWG), while the other two are based on a coupler which broadcasts the traffic sent by one input port to all the output ports. The proposed POIs are shown in Fig. 1. In all three schemes servers are equipped with optical network interfaces (ONIs) sending and receiving optical signals. The communication can be either intra-rack (shown as red solid lines) or to outside the rack (inter-rack or to/from the Internet, shown as blue dashed lines).

It should be noted that the dynamicity and programmability for POIs are provided by the ONIs of the servers. Therefore, all the presented POIs have the ONIs equipped with wavelength

Manuscript received June 18, 2015; revised August 10, 2015 and September 8, 2015; accepted September 8, 2015. Date of publication September 14, 2015; date of current version November 9, 2015. This work was supported by the Swedish Foundation for Strategic Research (SSF), Vetenskapsrådet, and Göran Gustafssons Stiftelse. The associate editor coordinating the review of this paper and approving it for publication was W. Fawaz.

The authors are with the Communication Systems Department (CoS), KTH Royal Institute of Technology, 16440 Stockholm, Sweden (e-mail: yuxinc@kth.se; fiorani@kth.se; wosinska@kth.se; jiajiac@kth.se).

Digital Object Identifier 10.1109/LCOMM.2015.2478474

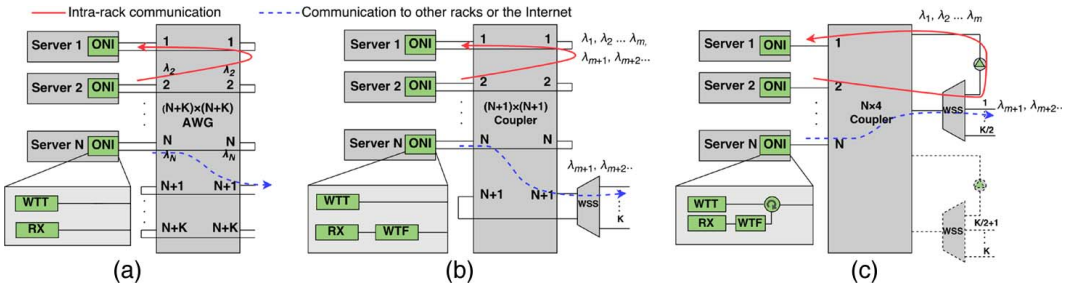

Fig. 1. (a) Scheme I: $(N+K)\times(N+K)$ AWG based POI, (b) Scheme II: $(N+1)\times(N+1)$ coupler based POI and (c) Scheme III: $N\times 4$ coupler based POI (WTT: wavelength tunable transmitter, AWG: arrayed waveguide grating, ONI: optical network interface, RX: receiver, WTF: wavelength tunable filter, WSS: wavelength selective switch).

From \ To		Within the rack				Outside of the rack		
		Server 1	Server 2	…	Server N	Interface 1	…	Interface K
Within the rack	Server 1		λ_2	…	λ_N	λ_{N+1}	…	λ_{N+K}
	Server 2	λ_2		…	λ_{N+1}	λ_{N+2}	…	λ_1
	…	…	…		…	…	…	…
	Server N	λ_N	λ_{N+1}	…		λ_{N+K}	…	λ_{N-1}
Outside of the rack	Interface 1	λ_{N+1}	λ_{N+2}	…	λ_{N+K}			
	…	…	…	…	…			
	Interface K	λ_{N+K}	λ_1	…	λ_{N-1}			

Fig. 2. Wavelength plan for AWG based POI.

tunability. It has been demonstrated in [6] and [7] that good network performance can be achieved with wavelength tuning speed in the magnitude of microseconds.

A. Scheme I: AWG Based POI

Fig. 1(a) shows the structure of the first type of POI, which is based on $(N+K)\times(N+K)$ AWG interconnecting the servers within a rack as well as providing connection to the switches at aggregation/core tier. Here, N is the number of servers and K is the number of the links between ToR and the aggregation/core tier. This scheme is inspired by the POI proposed in [4]. Each ONI has two fibers connected to the AWG input and output ports. For this type of POI, in total $N+K$ wavelengths are required. Thanks to the cyclic property of the AWG, a proper wavelength plan can be made (see Fig. 2 to set up a connection for any intra-rack or inter-rack communication without any conflicts in spectrum. Note that the grey fields in Fig. 2 represent no connection (i.e., there is no traffic passing through the POI destined to the same server or between two interfaces toward the aggregation/core tiers).

B. Scheme II: $(N+1)\times(N+1)$ Coupler Based POI

Fig. 1(b) shows the structure of the second proposed POI architecture. In Scheme II an $(N+1)\times(N+1)$ coupler is employed to interconnect N servers within the rack. Two ports of the coupler are connected to a wavelength selective switch (WSS) for inter-rack communications. The broadcast nature of the coupler provides higher flexibility in wavelength allocation than the AWG based scheme. In this scheme, the wavelengths can be dynamically assigned for the intra- and inter-rack communications, leading to high resource utilization. The wavelength tunable transmitters (WTTs) on the ONIs are able to use any available wavelength $(\lambda_1,\ldots\lambda_m)$ for intra-rack communications. The data is broadcast to all the output ports of the coupler. The ONI consists of wavelength tunable filter (WTF) and receiver (Rx) for receiving the signal. Due to the broadcast-and-select manner, the WTF is needed to select the wavelength assigned to a specific communication, while the signals on other wavelengths are dropped. For the traffic to outside of the rack, the WSS switches the corresponding wavelengths $(\lambda_{m+1},\lambda_{m+2}\ldots)$ and forwards the traffic to the aggregation and core tires. In this architecture, we assume the use of a $2\times K$ WSS, as the one demonstrated in [12]. Multiple interfaces $(K\geq 2)$ can be reserved to connect to the aggregation/core tier and support any topology (e.g., fat-tree and Quartz) for high scalability and resiliency.

C. Scheme III: $N\times 4$ Coupler Based POI

Fig. 1(c) shows the third proposed POI architecture, which enhances the reliability performance of the coupler based POI proposed in [3]. In Scheme III, ONIs at the servers are connected to N input ports of an $N\times 4$ coupler. By passing through a WSS, the wavelengths assigned for the intra-rack communications (i.e., $\lambda_1,\ldots\lambda_m$) are sent back to the coupler and then broadcasted to all the connected ONIs in the same rack. Like in Scheme II, the WTF is needed at the ONI to select the signal carried by the assigned wavelength. The wavelengths $(\lambda_{m+1},\lambda_{m+2}\ldots)$ are allocated for traffic sent to/received from the outside of the rack. Similar as the other schemes, this approach can also reserve several interfaces to connect to the aggregation/core tier for resiliency and scalability enhancement. From resiliency perspective, WSS is critical in this POI structure, as the traffic for both intra-rack communications and to the outside of the rack needs to pass this component. Considering the fact that WSS is an active component having obvious lower availability than the passive devices, backup is proposed by introducing an additional WSS.

III. Reliability and Cost Models

In this section we present an analytical model for reliability and cost evaluations of the three proposed POI architectures and the electronic ToR switch scheme.

A. Reliability Analysis

In this letter, we focus on ToR interconnect design and hence we perform the reliability performance analysis for the intra-rack communication. However, the same methodology can be applied to the core/aggregation tier.

Fig. 3 presents the reliability block diagrams (RBDs) for the intra-rack connections in the electronic switch based scheme and the three proposed POIs. RBD represents availability model of a system/connection, where series configuration corresponds to the case where all the connected blocks need to be available, while the parallel configuration means that at least one of the branches needs to be available. We can observe that in the three proposed POIs, the connections include several active and passive optical components. As a consequence, the connection availability (A) (i.e., the probability that the connection is operating) can be calculated by multiplying the availability of

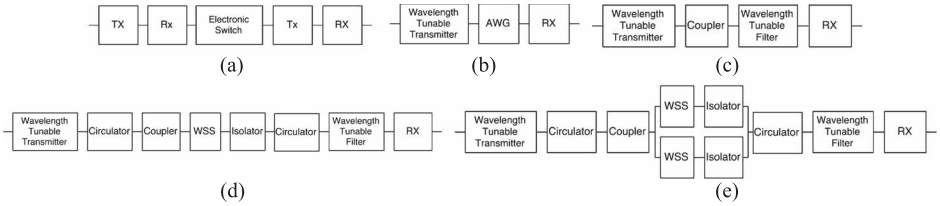

Fig. 3. Reliability block diagrams. (a) Electronic switch based scheme. (b) Scheme I. (c) Scheme II. (d) Scheme III without protection. (e) Scheme III with protection.

the cascaded components. We can then obtain the following formulas for availability of intra-rack connection in electronic switch based scheme (A_E), Scheme I ($A_{POI(I)}$), Scheme II ($A_{POI(II)}$), and Scheme III without protection ($A_{U-POI(III)}$):

$$A_E = A_{GTRX}^2 \times A_{ES}, \quad (1)$$

$$A_{POI(I)} = A_{TRX} \times A_{AWG}, \quad (2)$$

$$A_{POI(II)} = A_{TRX} \times A_{CP} \times A_{WTF}, \quad (3)$$

$$A_{U-POI(III)} = A_{TRX} \times A_{CL} \times A_{CP} \times A_{WSS} \times A_{IS} \times A_{WTF}. \quad (4)$$

Here, A_{TRX} represents the availability of the tunable transceiver (note that transmitter and receiver are embedded on the same board). Meanwhile, we denote A_{GTRX}, A_{ES}, A_{AWG}, A_{CL}, A_{CP}, A_{WSS}, A_{IS}, and A_{WTF} as the availability of the grey transceiver, electronic switch, AWG, circulator, coupler, WSS, isolator and WTF, respectively. The availability of each component can be obtained as the ratio between the mean lifetime and the mean time between failures (MTBF) [15]. In the protected Scheme III, the reliability is improved by the redundancy of WSS and isolator comparing to the unprotected one. The availability can be obtained according to the following formula:

$$A_{P-POI(III)} = \left(1 - (1 - A_{WSS} \times A_{IS})^2\right) \times A_{TRX} \times A_{CL} \times A_{CP} \times A_{WTF}. \quad (5)$$

B. Cost Analysis

To calculate the equipment cost for the three proposed POIs, we employ a similar approach as in [3]. We define the total cost of a POI (C_{POI})) as the sum of all the network components inside the rack. As a consequence, the total cost for electronic switch based scheme (C_E), Scheme I ($C_{POI(I)}$), Scheme II ($C_{POI(II)}$), and the unprotected Scheme III ($C_{U-POI(III)}$), can be calculated according to the following formulas:

$$C_E = 2 \times N \times C_{GTRX} + C_{ES}, \quad (6)$$

$$C_{POI(I)} = N \times C_{TRX} + C_{AWG}, \quad (7)$$

$$C_{POI(II)} = N \times (C_{TRX} + C_{WTF}) + C_{CP} + C_{WSS}, \quad (8)$$

$$C_{U-POI(III)} = N \times (C_{TRX} + C_{CL} + C_{WTF}) + C_{CP} + C_{WSS} + C_{IS}. \quad (9)$$

Here, N is the number of servers in the rack and C_{TRX} is the cost of a tunable optical transceiver. Moreover, C_{GTRX}, C_{ES}, C_{WTF}, C_{AWG}, C_{CP}, C_{WSS}, C_{CL}, and C_{IS} are the cost of grey transceiver, electronic switch, WTF, AWG, coupler, WSS, circulator and isolator, respectively. In the protected Scheme III, additional WSS and isolator are used inside the rack to improve resiliency. Accordingly, the total cost of the protected Scheme III ($C_{P-POI(III)}$) can be obtained through the following formula:

$$C_{P-POI(III)} = N \times (C_{TRX} + C_{CL} + C_{WTF}) + C_{CP} + 2 \times (C_{WSS} + C_{IS}). \quad (10)$$

IV. Numerical Results

In this section, we evaluate and compare the cost and reliability of the proposed POIs. We consider a conventional electronic ToR switch as benchmark. The overall results cover the cases of 1 Gbps and 10 Gbps transmission capacity per server.

TABLE I
MTBF and Cost of the Network Elements [8], [13], [14]

Components	MTBF[1]	Cost
1GE Electronic Switch	150 000 h	0.67 CU[2](port)
10GE Electronic Switch	150 000 h	3 CU
1Gbps Grey Transceiver	3 000 000 h	0.1 CU
10Gbps Grey Transceiver	600 000 h	0.5CU
1Gbps Tunable Transceiver	1 000 000 h	0.67CU
10Gbps Tunable Transceiver	500 000 h	1.3CU
WSS	300 000 h	8.3CU
AWG	4 000 000 h	0.1 CU(port)
Coupler	6 000 000 h	0.02 CU(port)
Isolator	12 000 000 h	0.3CU
Circulator	12 000 000 h	0.7 CU
Wavelength Tunable Filter	4 000 000 h	0.3 CU

1. Mean Time Between Failures
2. CU is the cost unit. 1 CU = 150 USD.

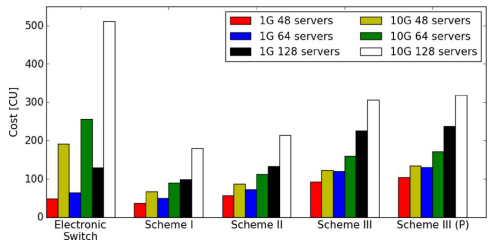

Fig. 4. Total cost for electronic ToR switch and the three proposed POI architectures.

Table I shows the MTBF and the cost values for the involved components [8], [13], [14], which are used to evaluate the reliability and the total cost of the proposed POI architectures. Fig. 4 shows the total cost of the proposed POIs and the conventional electronic ToR switch given the different number of total servers in a rack (48, 64 and 128) and 2 interfaces towards the aggregation/core tier. As reflected in the cost formulas, all the considered schemes show a linear increase in the total cost as a function of the number of servers. Considering the case with 1 Gbps per server, Scheme I and Scheme II show a similar total cost as the electronic ToR switch. On the other hand, the cost of Scheme III without protection is almost doubled due to the use of an additional circulator in the ONI. However, the circulator makes cabling easier since one server only has one fiber port for interconnection. The protected Scheme III shows a small increase in the total cost, since additional WSS and the isolator are needed for backup. For the case with 10 Gbps interface per server, the three POIs show great advantage in terms of cost comparing to electronic switches. The price of commercial electronic ToR switches operating at 10 Gbps is much more expensive than the price of electronic ToR switches operating at 1 Gbps. On the other hand, the increase of cost of three POIs from 1 Gbps to 10 Gbps is much lower than electronic switches. This is mainly due to the reason that for POIs the major cost increase is at transceivers side for higher data rate while the passive optical switch components remain the same. However, for solution

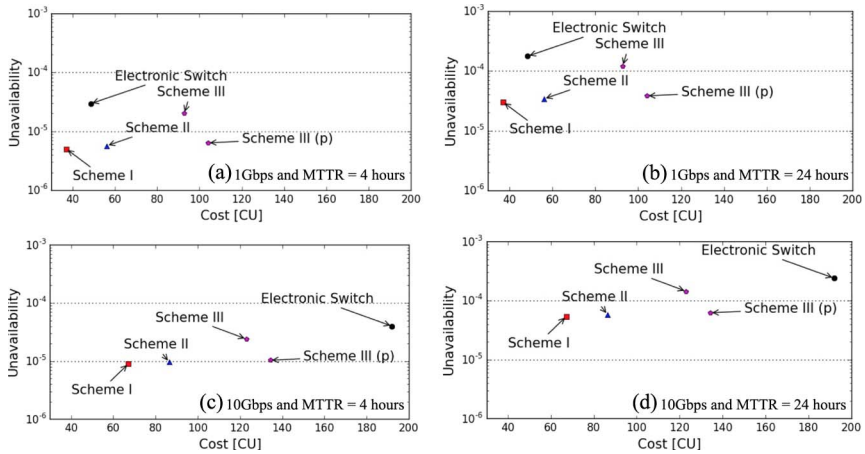

Fig. 5. Unavailability vs. total cost of three POIs for different MTTR values and server transmission capacities.

based on commodity switches, cost increase occurs at both transceivers at the servers and the switches for interconnection.

It is worth noting that, due to the lack of statistical multiplexing, the POIs may require more capacity, and thus higher costs, in the aggregation and core tiers. The cost results in [3] have shown that impact of the absence of statistical multiplexing in the edge tier on the overall data center network cost highly depends on the switching techniques adopted in core and aggregation tier. However, despite of the lack of additional statistical multiplexing in the edge tier, the cost reduction for the overall data center can still be achieved in many cases compared to the solution based on the conventional electronic switches [3].

Unavailability of a component or system is defined as the probability that it is failed and can be expressed as 1-A, where A denotes availability. Our calculation of the unavailability values of the three POIs is based on the MTBF of components presented in Table I. Regarding mean time to repair (MTTR), it is dependent on the maintenance policy adopted by the data center operator. We assume two values of MTTR (4 h and 24 h) reflecting different fault management policies in respect to the length of reparation time. Fig. 5 shows the unavailability of intra-rack connection versus total cost of 48 servers in a rack for the three considered optical intra-rack interconnects at 1 Gbps and 10 Gbps. It can be seen that Scheme I and Scheme II performs best, i.e., showing the lowest connection unavailability. In the case a fast reparation time (e.g., MTTR = 4 hours), the intra-rack connection availability for Scheme I and Scheme II can reach 5 nines (99.999%), meeting the reliability requirement for fault-tolerant site infrastructure, i.e., availability of 99.995% [9]. On the other hand, unprotected Scheme III supports similar level of connection availability as its electronic counterpart, while with the proposed redundancy of the WSS, its intra-rack connection availability can reach 5 nines. Even for a relatively long reparation time (e.g., MTTR = 24 hours), the availabilities up to 4 nines (99.99%) can be obtained for Scheme I, Scheme II and the protected Scheme III, which is much better than availability of ToR based on electronic switches. Advantages of POIs at 10 Gbps on reliability performance compared to the electronic interconnect is more obvious than at 1 Gbps. It is mainly because the reliability performance of the passive devices is not dependent on the data rate.

V. Conclusion

In this letter we proposed reliable optical interconnect architectures for the edge tier of data center interconnection networks. We have evaluated their cost and reliability performance and compared with the traditional commodity ToR switch. The results show that the proposed POIs outperform the electronic ToR switches in terms of both connection availability and cost. Compared to electronic commodity switches, the benefits of optical POIs are more obvious at the higher data rate. Therefore, our proposed POIs make it possible to reach the required connection availability of 99,995% and beyond at high data rates in the intra-data center networks.

References

[1] "Cisco global cloud index: Forecast and methodology, 2013–2018," Cisco, San Jose, CA, USA, White Paper, Sep. 2014.
[2] "Data center design considerations with 40 GbE and 100 GbE," Dell, Plano, TX, USA, White Paper, Aug. 2013.
[3] M. Fiorani, S. Aleksic, M. Casoni, L. Wosinska, and J. Chen, "Energy-efficient elastic optical interconnect architecture for data centers," *IEEE Commun. Lett.*, vol. 18, pp. 1531–1534, Sep. 2014.
[4] Y. Yawei *et al.*, "LIONS: An AWGR-based low-latency optical switch for high-performance computing and data centers," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 2, pp. 360–409, Mar./Apr. 2013.
[5] K. Chen *et al.*, "OSA: An optical switching architecture for data center networks with unprecedented flexibility," *IEEE/ACM Trans. Netw.*, vol. 22, no. 2, pp. 498–511, Apr. 2014.
[6] W. Ni *et al.*, "POXN: A new passive optical cross-connection network for low cost power efficient datacenters," *J. Lightw. Technol.*, vol. 32, no. 8, pp. 1482–1500, Apr. 2014.
[7] R. M. Indre, J. Pesic, and J. Roberts, "POPI: A passive optical Pod interconnect for high performance data centers," in *Proc. IEEE ONDM*, May 2014, pp. 84–89.
[8] J. Chen, Y. Gong, M. Fiorani, and S. Aleksic, "Optical interconnects at top of the rack for energy-efficient datacenters," *IEEE Commun. Mag.*, vol. 53, no. 8, pp. 140–148, Aug. 2015.
[9] "Data center site infrastructure tier standard: Topology," Uptime Inst., New York, NY, USA, 2010.
[10] R. N. Mysore *et al.*, "Portland: A scalable fault-tolerant layer 2 data center network fabric," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 39–50, Oct. 2009.
[11] Y. Liu, P. Gao, B. Wong, and S. Keshav, "Quartz: A new design element for low latency DCNs," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 283–294, Oct. 2014.
[12] K. Fontaine, R. Ryf, and D. T. Neilson, "N×M wavelength selective crossconnect with flexible passbands," in *Proc. IEEE OFC*, 2012, pp. 1–3.
[13] M. Mahloo *et al.*, "Toward reliable hybrid WDM/TDM passive optical networks," *IEEE Commun. Mag.*, vol. 52, pp. S14–S23, Feb. 2014.
[14] K. Grobe, M. Roppelt, A. Autenrieth, J. Elbers, and M. Eiselt, "Cost and energy consumption analysis of advanced WDM-PONs," *IEEE Commun. Mag.*, vol. 49, pp. S25–S32, Feb. 2011.
[15] J. Chen and L. Wosinska, "Analysis of protection schemes in pon compatible with smooth migration from TDM-PON to hybrid WDM/TDM PON," *OSA J. Opt. Netw.*, vol. 6, no. 5, pp. 514–526, May 2007.