Optical Interconnects for Cloud Computing Data Centers: Recent Advances and Future Challenges

Muhammad Imran[*][†]**and Saqib Haleem** [‡]

National Centre for Physics, Islamabad, Pakistan.
E-mails: muhammad.imran@ncp.edu.pk, saqib.haleem@ncp.edu.pk

It is widely argued that optical communication and networking technologies will play a significant role in future data centers. Although the optical technologies have made a significant advancements over the last few years towards providing a very high data transmission rate as well as increased flexibility and efficiency, an additional effort is needed to investigate suitable architectures and technologies for optical network within (intra) and outside (inter) data centers. This paper presents a brief overview on optical networks for data centers. Furthermore, the paper provides a qualitative categorization of the proposed schemes based on the type of optical switches. In the end, future research direction and opportunities of optical interconnect for data centers are discussed.
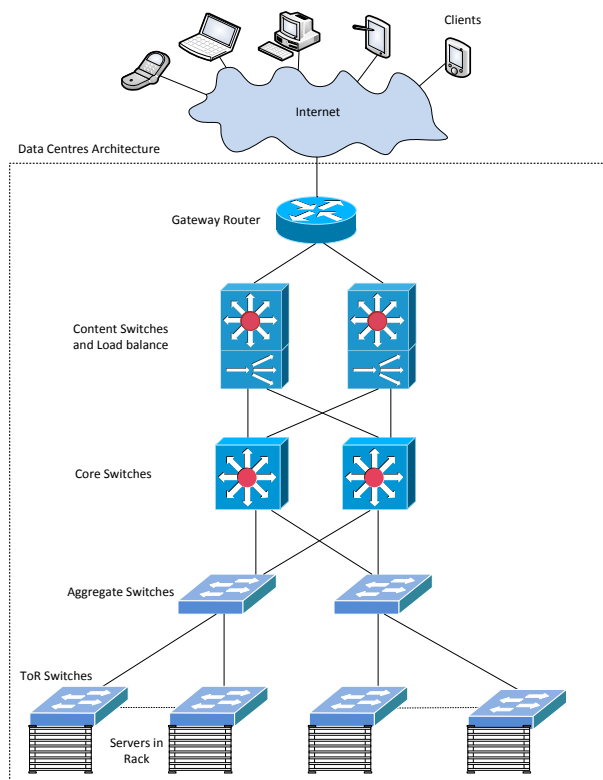
---

[*]Speaker.

[†]Dr. Imran works as a Senior Scientific Officer at National Centre for Physics (NCP), Islamabad, Pakistan.

[‡]Saqib works as a Manager IT at NCP.

Figure 1: Data Center Architecture.

1. Introduction

Internet traffic has been increasing significantly over the last few years due to the development of new applications such as social media and multimedia applications. These applications are based on cloud computing infrastructure which is deployed in data centers. Cloud computing has obtained a significant importance over the last few years because it has moved computation and storage away from end user devices to large-scale data centers [20].

The conventional design of the data center network (DCN) is based on a layered architecture as shown in Figure 1. It has various types of electrical switches at different layers. When a user request arrives from the internet in the data center, it first goes to the load balance and content switches which forward the request to the desire servers. To accomplish the request, the servers can communicate with other servers. For example, the application servers can communicate with the database servers to process the request. After accomplishing the request, the response is sent to the sender in the internet via the gateway routers.

1.1 Limitations of Traditional DCNs

Current data center network has some limitations. These limitations are high power consumption, increasing traffic within data centers, need of higher data rates, scalability, latency and performance etc. The electrical switches at various layers of a DCN and the transceivers which are

installed at each port of electrical switches are the major source of power consumption in DCN. Transceivers in electrical switches require optical-electrical-optical (O-E-O) conversion. For example, the power consumption of a port of an electrical switch is 3.7 W [2] and a 10G SFP+ transceiver's power consumption is 1 W [3] which is very high as compared to the power consumption of a port of the MEMS switch which does not require transcievers i.e. only 0.14 W [5]. Although the power consumption of the electrical switches and the transceivers over the last few years has been reduced significantly but it is still high as compared to the optical switches.

Increasing traffic within data centers is another significant challenge in future DCNs. According to Cisco cloud index, the movement of data center's traffic is shown in Figure 2. It can be seen that, major percentage of traffic will remain within data centers while only a very small percentage will go the external network. Small percentage of the traffic will also be interchanged between data centers. Due to this increasing traffic within data centers, higher data rates interconnects are required. Higher data rates interconnect are only feasible with optical fibers because the performance of electrical transmission lines is degraded with the increasing data rate. Scalability is also an important challenge in future DCNs. The future data centers must be large enough to host hundreds of thousands of servers. Another challenge is the latency of traffic within data center should be minimum. It is incurred by prorogation delay, queuing delay and processing delay at each node.
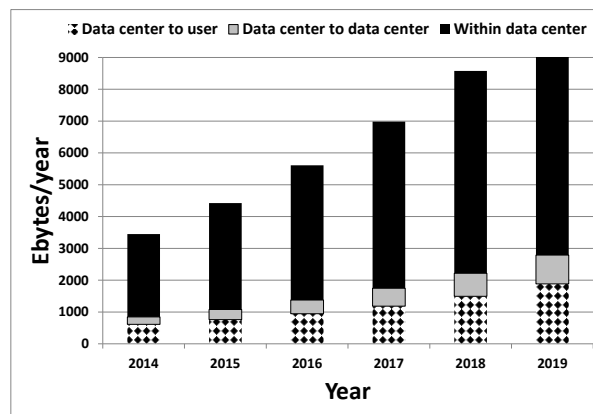

Figure 2: Traffic Movement in Data Centers from 2014 to 2019 [1]

An optical network within a data center can overcome challenges discussed above. For example, optical switches are power efficient as compared to electrical switches while cost per port of both types of switches are approximately the same [12]. Optical transceiver with optical fiber can fulfil the requirements of increasing traffic within data centers and higher data rates. In the near future, transceivers with high bandwidth will be available. Optical MEMS switches are also more scalable than electrical switches which together with optical fibers can meet the requirement of scalability as well.

This paper presents a brief overview on optical network data centers. We categorize various proposals of optical network according to the type of optical switches used. We also briefly discuss the various types of optical switches. Moreover, we also provide an insight on future directions and opportunities for optical interconnects. The remainder of this article is organized as follows. Section 2 briefly discuss different types of optical switches. In Section 3, we briefly discuss various

optical interconnects that exist in the literature. We provide future directions in section 4. In the end, we conclude in Section 5.

2. Optical Switches

We classify optical switches into two categories according to their switching time: (1) Slow optical switches; and (2) Fast optical switches. Slow optical switches are : (1) Micro electro mechanical system (MEMS) switch and (2) Wavelength selective switch. MEMS switches are based on mature technology and have been used in backbone optical network for many years. They have high port density, transceivers are not required due to all-optical switching, are power efficiency due to passive switching, have low insertion losses and crosstalk, are data rate independent, can provide bidirectional communication and are cheap. However, they have high switching time which is of the order of tens of milliseconds. These switches are mostly used in long-haul backbone optical network by using optical circuit switching for long-lived connections. MEMS switches with 320 ports (Calient) is commercially available while with as many as 1,000 ports being feasible [6].

Wavelength Selective Switch (WSS) is a device that can divide the incoming set of wavelengths to a different set of outgoing wavelengths. Each set of wavelength is destined to a specific output port. The reconfiguration time of WSSs has been reported in the range of a few milliseconds [4].

Fast optical switches are divided into two types: 1) Arrayed Waveguide Grating Router (AWGR), and 2) SOA-Based Switches. AWGR is a combination N $(1 \times N)$ arrayed waveguide gratings(AWGs) on the input side and N $(N \times 1)$ AWGs on the output side arranged in a cyclic way where N is the number of input/output ports. It can provide strictly non-blocking switching, if it is used with tunable wavelength converters (TWCs) or tunable lasers (TLs) at the input and fixed wavelength converters (FWCs) at the output [15]. SOA-based switches are mostly based on a broadcast-and-select architecture. First, the input signal is broadcast using a $(1 \times N)$ coupler. Each output of the $(1 \times N)$ coupler is attached to one of N SOAs per output port. The SOA is used as a gate element and also provides optical gain in order to make up for the losses.

3. Evolution of Optical Interconnects for DCN

In this section, we give a brief overview of various schemes for optical network in data centers that have been presented in the last few years. We classify these schemes into five types according to the type of optical switches. These are described below.

3.1 Architectures based on MEMS

Optical interconnect architectures that are based on optical MEMS switches are presented in studies [9, 11, 12, 24, 31]. A hybrid electrical/optical switch (Helios) architecture for data centers [12] uses a single optical MEMS switch and an array of electrical switches at the core while A hybrid packet/circuit switch (HyPaC) [31] also uses a single optical MEMS switch at the core but it has various layers of electrical switches i.e. access, aggregate and core layers. Main ideology of these two schemes is to use MEMS for long lasting traffic flows i.e., elephant flows and electrical switches for the short and bursty traffic. These schemes have limitation of the control plane

overhead and performance degradation due to large switching time of the MEMS switch in case of frequent configuration requests. These schemes are also not power efficient because they use electrical switches as well. The performance degradation for hotspots (ToR pairs with high traffic demand) is another major limitation in these techniques.

Other MEMS based architectures [9, 11, 24] do not have these limitation because they provide full bandwidth between any ToR pair. The optical switch architecture (OSA) [9] uses only a MEMS switch at the core and no electrical switches at the core. The main idea of this scheme is to connect ToR switches through the MEMS switch by a single hop that generate high inter-rack traffic, while bursty and low volume traffic are assigned to multi-hop connections. Multi-hopping technique avoid frequent reconfiguration of MEMS switch. However, it increases latency and power consumption for multi hops path. The OSA does not perform well in situations where most of the traffic flows last few milliseconds. All of the above three schemes are not scalable because they only use one MEMS switch. The reconfigurable Architecture [24] does not has the limitation of scalability. It uses a single stage core topology that has multiple MEMS switches at the core. This scheme also has all the benefits of previous three techniques. However, this scheme does not work well for bursty and mice flows. In this case, it will increase latency and power consumption due to multi-hopping.

A hybrid reconfigurable architecture [11] integrates positive aspects of above schemes into a single approach. This scheme is based on array of multiple MEMS switches at the core and an electrical network comprising electrical switches. Similar to OSA, and reconfigurable schemes, this scheme also uses multi-hopping technique. Traffic is shifted to electrical network during reconfiguration phase of MEMS switches. Like all other MEMS based architectures, the control plane of the HyDRA also has the limitation to support applications that have high traffic stability, i.e. workloads that last several seconds. This limitation is due to the large switching time of the MEMS switch and the control plane overhead.

3.2 Architectures based on SOAs

Optical interconnect architectures that are based on SOAs are presented in studies [8, 21–23, 30]. These architectures use SOAs as gate switch. SOAs not only act as gate switches that allows light to pass through or not but they also provide gain to boost signal strength. Data vortex [23] is based on a banyan-switch structure. It uses optical packet switching (OPS) and does not require optical buffers because it uses deflection routing technique. Due to the modular architecture, this design is scalable to thousand of nodes. However, latency might be high for packets that travel to multi-stage routes. Bidirectional architecture [30] is another scalable architecture based on SOAs that provides bidirectional communication. All other SOAs based design only provide unidirectional communication.

Space wavelength (SW) architecture [21] is another design that uses space and wavelength domain. It targets low latency and high throughput. Space time interconnection architecture (STIA) [22] is an enhanced version of SW architecture. It utilizes space, wavelength and time domain. STIA is more scalable than SW architecture. Recent work have shown scalability by arranging STIA in tree, folded clos or flattened butterfly topology [8]. All of the architectures based on SOAs have limitations of implementation complexity and high CAPEX cost.

3.3 Architectures based on AWGRs

Optical interconnect architectures that are based on AWGRs are presented in studies [7, 14, 19, 27–29, 32, 33]. Low-latency interconnect optical network switch (LIONS) architecture [33] uses AWGR in conjunction with tunable wavelength converters (TWCs) at the sender side and multiple receivers at each output port of AWGR. It uses AWGR's cyclic wavelength routing technique. The scalability of this architecture depends upon the number of ports of AWGR, its loopback buffer and tuning range of TWCs. The another variation of LIONS named as NACK-LIONS [29] removed loopback buffer and proposed all optical negative acknowledgement technique. All of the above architectures discussed in this section use a centralized control plane which limits the scalability. TOKEN-LIONS [27] is another variation of LIONS that uses a distributed control plane. The control plane in TOKEN-LIONS uses characteristics of saturation effect of reflective semiconductor amplifiers (RSOAs). TONAK-LION [28] is the advanced architecture of LIONS and all of its related designs. It makes use of the positive aspects of both NACK-LION and TOKEN-LION. Although the issue of scalability in LIONS and related has been addressed to some extent, but still the problem exists due to limitation of small number of ports of AWGR. Hi-LION is another design which is proposed recently to address this scalability issue [7].

Petabit optical switch [32] is another architecture that uses AWGR in three stage topology to address the limitation of scalability. Similarly to LIONS and related architecture, it uses TWCs at the sender. However, buffering is done only in advance, not in the middle. The connection for every packets is established first, then the packet is sent. In order to make this process fast, an efficient scheduling algorithm was proposed in this scheme. Although, this scheme has advantages of scalability, low latency and high throughput, but high CAPEX cost due to three stage architecture is the serious concern in the design. The Integrated router interconnected spectrally (IRIS) project [14] is also based on a three stage architecture. However, this design uses a different switch at each stage. A space switch is used in the first stage, a time switch in the second stage and again a space switch in the third stage is used. Similarly to Petabit, this design also has a high CAPEX cost due to three stage architecture.

Optical network based on OFDM with AWGR is presented in [19] which does not require tunable lasers (TLs) or TWCs. Instead of TLs or TWCs, this architecture uses OFDM modulator at the sender and a receiver with parallel signal detection (PSD) technique. In this technique, OFDM signal can be sent to many destination racks at the same time, and many ToR switches can send the signal to the same destination rack simultaneously. Fast switching, low latency and data rate granularity are the unique features of this scheme. However, this scheme is not scalable, is not power efficient, and has high CAPEX cost.

3.4 Architectures based on WSSs

Optical interconnect architectures that are based on WSSs are presented in studies [10, 26, 34]. Mordia architecture was presented by [26] and is based on a ring topology that uses OCS. Scalability is the major limitation of Mordia which is limited by the size of WSSs and the number of wavelengths. WaveCube architecture [10] is an advanced version of OSA but it uses WSSs instead of MEMS. This design is scalable due to the usage of topology. It uses the same concept

of multi-hopping as was used in OSA. Unlike Mordia, there is not a single point of failure and multiple paths exist between any pair of ToRs in WaveCube.

Optical pyramid data center network (OPMDC) [34] is a three tier architecture which comprises three types of WSS-based optical switching nodes in each tier. It uses circuit switching for elephant flows and packet switching for mice flows. Each tier comprises a group of pods. The traffic within a pod is routed horizontally in tier 1 while the traffic to other pods is routed first vertically in tier 2, then horizontally in that pod. These architectures have limitation of implementation complexity and scalability of the control plane.

3.5 Architectures based on Fast and Slow Optical Switches

Optical interconnect architectures that are based on fast and slow optical switches are presented in studies [13, 16–18, 25]. The LIGHTNESS is a hybrid design that uses fast and slow optical switches in the core [25]. Optical circuit switching with slow optical switches and optical packet switching with fast optical switches are used. Hybrid optical switching (HOS) [13] comprises three layers of fast and slow optical switches. It uses, packet, burst and circuit switching. The packet/bursts losses result in poor TCP performance. This is because of high packet latency and decrease in overall throughput. Other drawback is CAPEX cost of the interconnect. Hybrid optical switch architecture for DCNs named as HOSA has been proposed in [16, 17]. It uses a single stage of fast and slow optical switches instead of three stages as were in HOS. It uses optical burst switching.

HOSA shows low latency and high throughput and also achieves performance which is comparable to electrical DCNs. HOSA is highly scalable in terms of data plane. However, scalability of the control plane is also an open issue in this architecture. It is also power efficient. Fast optical switch architecture (FOSA) proposed in [18] is another architecture which is variation of the HOSA. In FOSA, only fast optical switches are considered. This design is also scalable to more than hundred thousand servers in terms of data plane. However, scalability of the control plane is also an open issue in this architecture.

4. Future Directions

There is an increasing trend of using virtualization and software-defined networking in data centers. These virtualization and SDN technologies were proposed for data centers that use electrical switches. However, optical technologies that have been used for many years in the backbone optical core network are now penetrating into data centers. So, virtualization and SDN in optical network for data centers will play a key role in the near future.

The other emerging areas in optical network for data centers is elastic optical network (EON). The deployment of EON in data centers can significantly lead to a more flexible operation, efficient use of resources and low power consumption. Thus, an EON for data centers can provide resource allocation dynamically according the requirements of applications.

There is an inherent advantage of using EON and SDN together in data centers because these technologies have potential compatibility with future optical network. The EON and SDN can play important role in upcoming future network for inter and intra data centers interconnect to efficiently manage and utilize resources.

5. Conclusion

In this paper, we presented brief survey of optical network for data centers that have proposed in recent years. Optical interconnects have gained significant attention in recent years because they can be used to address the challenges of traditional data centers networks.

We gave a brief overview of different types of optical switches that can be considered to design an optical network for data centers. A brief overview of different types of optical networks for DCNs was also given. We also gave directions for future work in optical interconnects for DCNs.

After discussing various architectures, it can be noticed that every architecture has some advantages and disadvantages. Generally, the architectures which are based on MEMS switches are more feasible to be deployed and used in the large scale data centers due to the availability of commercial off the shelf components. These designs are also cost efficient. The architectures which are based on other switches might be good in research in terms of the performance but they can not be deployed in the large scale data centers due to the unavailability of scalable optical devices. These are also the expensive solutions as compared to the MEMS based designs.

Acknowledgement

References

[1] Cisco global cloud index: Forecast and methodology, 2014–2019.
http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html (visited on 5/10/2016).

[2] Cisco Nexus 3064-X, 3064-T, and 3064-32T Switches.
http://www.cisco.com/c/en/us/products/collateral/switches/nexus-3000-series-switches/data_sheet_c78-651097.pdf (visited on 13/09/2018).

[3] Cisco S-Class 10GBASE SFP+ Modules Data Sheet.
http://www.cisco.com/c/en/us/products/collateral/interfaces-modules/transceiver-modules/datasheet-c78-733585.html (visited on 09/13/2018).

[4] FINISAR WSS datasheet & application note - datasheet archive.
http://www.datasheetarchive.com/FINISAR%20WSS-datasheet.html (visited on 5/10/2016).

[5] Photonic optical circuit switching | CALIENT technologies.
https://whitepapers.em360tech.com/wp-content/files_mf/1377248645CalientS320DataSheetMarch12013.pdf (visited on 13/09/2018).

[6] Photonic optical circuit switching | CALIENT technologies. http://www.calient.net/ (visited on 5/10/2016).

[7] Z. Cao, R. Proietti, and S. Yoo. Hi-LION: Hierarchical large-scale interconnection optical network with AWGRs [invited]. *Journal of Optical Communications and Networking*, 7(1):A97–A105, 2015.

[8] I. Cerutti, P. G. Raponi, N. Andriolli, P. Castoldi, and O. Liboiron-Ladouceur. Designing energy-efficient data center networks using space-time optical interconnection architectures. *Selected Topics in Quantum Electronics, IEEE Journal of*, 19(2):3700209–3700209, 2013.

[9] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen. OSA: An optical switching architecture for data center networks with unprecedented flexibility. *Networking, IEEE/ACM Transactions on*, 22(2):498–511, April 2014.

[10] K. Chen, X. Wen, X. Ma, Y. Chen, Y. Xia, C. Hu, and Q. Dong. Wavecube: A scalable, fault-tolerant, high-performance optical data center architecture. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 1903–1911. IEEE, 2015.

[11] K. Christodoulopoulos, D. Lugones, K. Katrinis, M. Ruffini, and D. O'Mahony. Performance evaluation of a hybrid optical/electrical interconnect. *Optical Communications and Networking, IEEE/OSA Journal of*, 7(3):193–204, March 2015.

[12] 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*, 41(4):339–350, 2011.

[13] M. Fiorani, S. Aleksic, and M. Casoni. Hybrid optical switching for data center networks. *Journal of Electrical and Computer Engineering*, 2014, 2014.

[14] J. Gripp, J. E. Simsarian, J. D. LeGrange, P. Bernasconi, and D. T. Neilson. Photonic terabit routers: the IRIS project. In *Optical Fiber Communication Conference*, 2010.

[15] Y. Hida, Y. Hibino, T. Kitoh, Y. Inoue, M. Itoh, T. Shibata, A. Sugita, and A. Himeno. 400-channel arrayed-waveguide grating with 25 GHz spacing using 1.5%-δ waveguides on 6-inch si wafer. *Electronics Letters*, 37(9):576–577, 2001.

[16] M. Imran, M. Collier, P. Landais, and K. Katrinis. HOSA: Hybrid optical switch architecture for data center networks. In *Proceedings of the 12th ACM International Conference on Computing Frontiers*, CF '15, pages 27:1–27:8, New York, NY, USA, 2015. ACM.

[17] M. Imran, M. Collier, P. Landais, and K. Katrinis. Performance evaluation of hybrid optical switch architecture for data center networks. *Optical Switching and Networking*, 21:1–15, 2016.

[18] M. Imran, M. Collier, P. Landais, and K. Katrinis. Software-defined optical burst switching for hpc and cloud computing data centers. *J. Opt. Commun. Netw.*, 8(8):610–620, Aug 2016.

[19] P. N. Ji, D. Qian, K. Kanonakis, C. Kachris, and I. Tomkos. Design and evaluation of a flexible-bandwidth OFDM-based intra-data center interconnect. *Selected Topics in Quantum Electronics, IEEE Journal of*, 19(2):3700310–3700310, 2013.

[20] B. Koley, V. Vusirikala, C. Lam, and V. Gill. 100GbE and beyond for warehouse scale computing. 2010.

[21] O. Liboiron-Ladouceur, I. Cerutti, P. G. Raponi, N. Andriolli, and P. Castoldi. Energy-efficient design of a scalable optical multiplane interconnection architecture. *Selected Topics in Quantum Electronics, IEEE Journal of*, 17(2):377–383, 2011.

[22] O. Liboiron-Ladouceur, P. G. Raponi, N. Andriolli, I. Cerutti, M. S. Hai, and P. Castoldi. A scalable space–time multi-plane optical interconnection network using energy-efficient enabling technologies [invited]. *Optical Communications and Networking, IEEE/OSA Journal of*, 3(8):A1–A11, 2011.

[23] O. Liboiron-Ladouceur, A. Shacham, B. A. Small, B. G. Lee, H. Wang, C. P. Lai, A. Biberman, and K. Bergman. The data vortex optical packet switched interconnection network. *Journal of Lightwave Technology*, 26(13):1777–1789, 2008.

[24] D. Lugones, K. Katrinis, G. Theodoropoulos, and M. Collier. A reconfigurable, regular-topology cluster/datacenter network using commodity optical switches. *Future Generation Computer Systems*, 30:78–89, 2014.

[25] J. Perelló, S. Spadaro, S. Ricciardi, D. Careglio, S. Peng, R. Nejabati, G. Zervas, D. Simeonidou, A. Predieri, M. Biancani, et al. All-optical packet/circuit switching-based data center network for enhanced scalability, latency, and throughput. *Network, IEEE*, 27(6):14–22, 2013.

[26] G. Porter, R. Strong, N. Farrington, A. Forencich, P. Chen-Sun, T. Rosing, Y. Fainman, G. Papen, and A. Vahdat. *Integrating microsecond circuit switching into the data center*, volume 43. ACM, 2013.

[27] R. Proietti, C. J. Nitta, Y. Yin, R. Yu, S. Yoo, and V. Akella. Scalable and distributed contention resolution in AWGR-based data center switches using RSOA-based optical mutual exclusion. *Selected Topics in Quantum Electronics, IEEE Journal of*, 19(2):3600111–3600111, 2013.

[28] R. Proietti, Y. Yin, R. Yu, C. J. Nitta, V. Akella, C. Mineo, and S. Yoo. Scalable optical interconnect architecture using AWGR-Based TONAK LION switch with limited number of wavelengths. *Journal of Lightwave Technology*, 31(24):4087–4097, 2013.

[29] R. Proietti, Y. Yin, R. Yu, X. Ye, C. Nitta, V. Akella, and S. B. Yoo. All-optical physical layer NACK in AWGR-based optical interconnects. *Photonics Technology Letters, IEEE*, 24(5):410–412, 2012.

[30] A. Shacham and K. Bergman. An experimental validation of a wavelength-striped, packet switched, optical interconnection network. *Journal of Lightwave Technology*, 27(7):841–850, 2009.

[31] 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*, volume 40, pages 327–338, 2010.

[32] K. Xia, Y.-H. Kaob, M. Yangb, and H. Chao. Petabit optical switch for data center networks. *Polytechnic Institute of New York University, New York, Tech. Rep*, 2010.

[33] Y. Yin, R. Proietti, X. Ye, C. J. Nitta, V. Akella, and S. Yoo. LIONS: an AWGR-based low-latency optical switch for high-performance computing and data centers. *Selected Topics in Quantum Electronics, IEEE Journal of*, 19(2):3600409–3600409, 2013.

[34] M. C. Yuang, P.-L. Tien, H.-Y. Chen, W.-Z. Ruan, T.-K. Hsu, S. Zhong, J. Zhu, Y. Chen, and J. Chen. Opmdc: Architecture design and implementation of a new optical pyramid data center network. *Journal of Lightwave Technology*, 33(10):2019–2031, 2015.