The Role of Integrated Photonics in Datacenter Networks

Madeleine Glick
College of Optical Sciences, University of Arizona, Tucson, AZ 85721, USA

ABSTRACT

Datacenter networks are not only larger but with new applications increasing the east-west traffic and the introduction of the spine leaf architecture there is an urgent need for high bandwidth, low cost, energy efficient interconnects. This paper will discuss the role integrated photonics can have in achieving datacenter requirements. We will review the state of the art and then focus on advances in optical switch fabrics and systems.

The optical switch is of particular interest from the integration point of view. Current MEMS and LCOS commercial solutions are relatively large with relatively slow reconfiguration times limiting their use in packet based datacenter networks. This has driven the research and development of more highly integrated silicon photonic switch fabrics, including micro ring, Mach-Zehnder and MEMS device designs each with its own energy, bandwidth and scalability, challenges and trade-offs. Micro rings show promise for their small footprint, however they require an energy efficient means to maintain wavelength and thermal control.

Latency requirements have been traditionally less stringent in datacenter networks compared to high performance computing applications, however with the increasing numbers of servers communicating within applications and the growing size of the warehouse datacenter, latency is becoming more critical. Although the transparent optical switch fabric itself has a minimal additional latency, we must also take account of any additional latency of the optically switched architecture. Proposed optically switched architectures will be reviewed.

Keywords: silicon photonics, optical interconnects, optical switch, data center

1. INTRODUCTION AND MOTIVATION

The staggering growth of Internet Protocol (IP) traffic and the data centers that service this traffic has become so commonly known that it is perhaps no longer surprising. The scale of the growth and the commercial requirements of low cost, power efficient, high bandwidth interconnection networks are creating new challenges for data center. Many see integrated photonics, in particular integrated silicon photonics, as the prime solution.

1.1 The scale of the requirements

The Cisco report "The Zettabyte Era"[1] gives fascinating insight into the scale of the traffic increases projected. It predicts annual global IP traffic to pass the zettabyte threshold by the end of 2016 and to reach over 2.2 zettabytes/year by 2020. Much of this increase in traffic is driven by video, with predictions that video traffic will be over 80% of all IP traffic by 2020. Or in other words, that for every second, one million minutes of video will cross the network!

At the same time as this tremendous growth of traffic into and out of the data center, the nature of the applications require that at least three quarters of the traffic stays internal to the data center. This traffic is called east – west traffic as opposed to north – south traffic that enters and exits the data center. The relative percentage of east – west traffic remains approximately the same as the traffic increases.[2]

Energy consumption becomes a related concern. Data centers currently account for approximately 2.5% of electricity consumption in the US.[3,4] From 1.5% of the total energy consumed in the US at a cost of $4.5B, the energy consumption of data centers has been predicted to triple by 2020.[5, 6, 7] With this concern, many companies such as Google have made extensive efforts to reduce energy consumption. For comparison's sake report they report that for a month's worth of the typical active Google user activity (25 searches, one hour of YouTube, Gmail and other services per day), Google emits the equivalent amount of CO_2 as driving a car for one mile.[8]

The capability of optics to achieve extremely high bandwidth data transmission[9] becomes the basis of using photonics as an enabling technology to solve data center bandwidth requirements. The challenge to photonics becomes to do this at commercially acceptable cost and energy consumption.

Next-Generation Optical Networks for Data Centers and Short-Reach Links IV,
edited by Atul K. Srivastava, Proc. of SPIE Vol. 10131, 1013104 · © 2017 SPIE
CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2255794

1.2 Data center architectures

Data centers have traditionally been built in a tree tiered network architecture as in Figure 1. This architecture worked well for small data centers when most of the traffic was between client and server (north – south traffic). Data center networks can now have 100,000's servers to connect. With the advent of virtualized servers and other applications with increased east west traffic, the three tier network began to have bandwidth bottlenecks. In addition the latency from server to server in the tiered network can vary depending on the path.

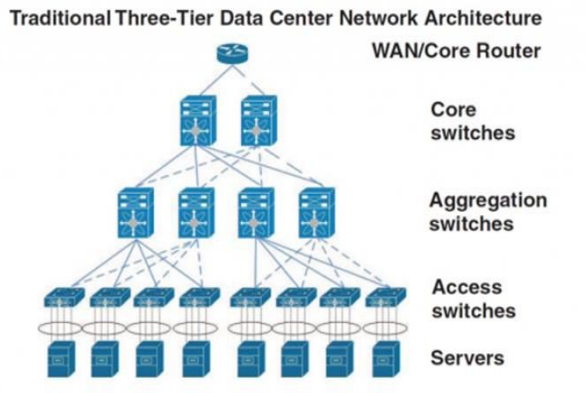

Figure 1. Schematic of the traditional three tier data center network architecture. http://www.fiber-optic-tutorial.com/how-to-connect-cisco-nexus-9396px-to-40g-network.html

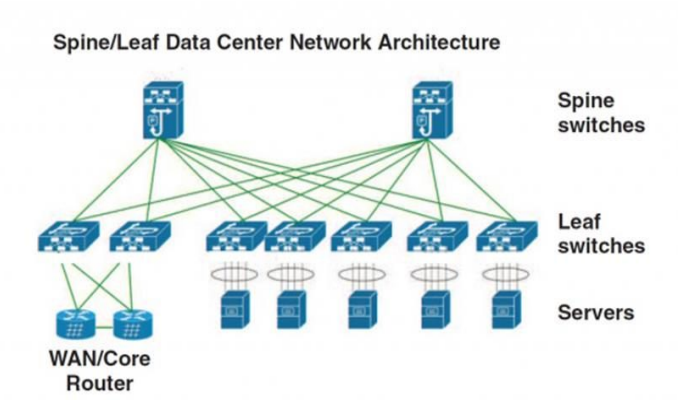

Figure 2. Schematic of the flat spine/leaf data center network architecture. http://www.fiber-optic-tutorial.com/how-to-connect-cisco-nexus-9396px-to-40g-network.html

The spine and leaf architecture (Figure 2) was developed to overcome the limitations of the traditional network. In the two tier Clos architecture, every lower tier switch in the leaf layer is connected to each of the top tier switches of the spine layer in a full mesh topology. The leaf switches connect to the servers. This architecture is more easily scaled for increasing traffic by adding spine switches. In addition, the latency can be reduced by avoiding one of the switching layers of the tiered network and is more uniform and predictable. The spine leaf network can however have considerably higher density of cabling especially for large scale networks.[10]

As we go forward to design photonics interconnects and device for the data center application it is critical to understand the system being designed for. We need to know the bandwidth required for links and the port count for switch fabrics.

2. OPTICS IN DATA CENTER ARCHITECTURES

2.1 Data Center rack to rack (specifically rack to rack issues)

Recent years have seen the emergence of novel data center proposals, taking advantage of both electrical and optical interconnects, often called hybrid networks.[11-17] Optical interconnects capable of ultra-high switching capacities, bitrate

transparency, and low power density, are promising candidates to meet the scale, footprint, and power density requirements of massive data centers. Purely optical interconnects, however, suffer from the lack of a viable, all-optical buffering technology and relatively low reconfiguration speeds based on commercially available optical micro-electro-mechanical (MEMS) switches. Thus, a dual or multi-fabric data center design that combines the advantages of both electrical and optical switching was initially proposed (Figure 3). Apart from indicating an approach to solve the bandwidth bottleneck of the traditional tiered network, the hybrid architecture is of interest because of its potential to significantly reduce cost and energy[13] primarily through the reduction of number of optical transceivers required in the system when incorporating the transparent optical switch fabric.

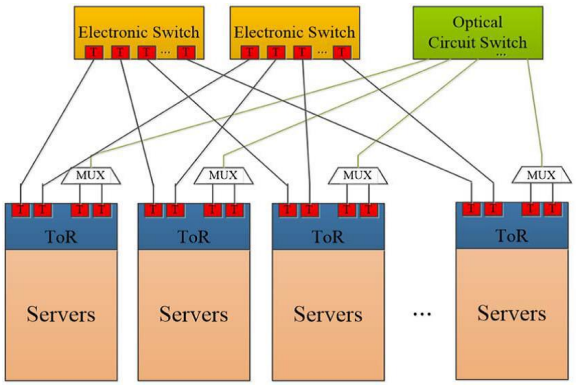

Figure 3. Schematic of the optically interconnected hybrid data center. (T, transceiver; ToR, top-of-rack switch; MUX, wavelength multiplexer).

The simple hybrid network initially proposed has several drawbacks. The reconfiguration time of the MEMS switch of approximately 30 ms is relatively long (on packet time scales). In addition as the number of servers in the data center gets larger and east-west traffic increases latency, becomes more of a critical metric in the data center than previously (especially as compared to high performance supercomputers). Higher fan out or port count than currently available from commercial optical switches is also desirable to achieve connectivity across the data center. Large data centers can have more than 1000 racks or TORs to connect. Another desirable attribute would be what can be called "seamlessness"[16] or a more complete capability of bandwidth reconfigurability. Table 1 summarizes much of the research on these issues to date. Other avenues to overcome these shortcomings have also been explored including architectures with tunable lasers and passive components,[18,19,20] software defined networking with distributed switch fabrics[21] and incorporating machine learning for improved resource utilization.[22, 23]

Table 1. Properties of reconfigurable interconnects. After Ref. 16.

	Enabling Technology	Seamless	Fan-out	Reconfiguration time
Helios[13], c-Thru[12], Proteus[44], Solstice[45]	Optical Circuit Switch	No	100-320	30 ms
Flyways[46], 3DBeam[47]	60GHz	No	~70	10 ms
Mordia[48]	Optical Circuit Switch	No	24	11 μs
Firefly[49]	Free Space Optics	Yes	10	20 ms
Projector [16]	Free Space Optics	Yes	18,432	12 μs

2.2 Integrated photonics

The networks discussed above focus on solving the traffic bottlenecks in the conventional tiered tree architecture, primarily by modifying the tree architecture itself using commercially available or near to available off-the-shelf equipment. Silicon photonics-based optical interconnects, leveraging the capacity and transparency of optics and

fabricated in high volume CMOS compatible foundries, form the foundation of a vision solving communications bottlenecks. For many years researchers have recognized that the relatively high cost of adoption of photonics in computer systems might be overcome if the photonic components could be made in fabrication environments compatible with silicon-based electronics.[24, 25] This reduction in cost is particularly critical to large data centers where cost efficiency is paramount. IBM, one of the groups at the forefront of silicon photonic research developed CMOS integrated silicon nanophotonics technology allowing monolithic integration of deeply scaled optical circuits into the FEOL of a standard sub-100 nm CMOS process and the demonstration of many devices with dense optical circuitry.[26] Roadmapping activities to understand and facilitate how technology, industry, and policy dynamics interact have been undertaken, with most recently additional government funding through AIM Photonics,[27] to enable progress in key areas.[28, 29] Most recently, the International Electronics Manufacturing Initiative (iNEMI) has, for the first time, included a section on photonics in the data center in their 2017 roadmap of the future technology requirements of the global electronics industry.[30]

Although the need for higher bandwidth links has led to a penetration of optics into the data center initially through the successful use of active optical cables, the roadmapping activities indicate that a significant improvement in performance can be achieved by more than simply replacing copper wires with optical fibers. The direction of data center development as described above clearly shows that the use of integrated photonics and silicon photonics for on-board high bandwidth transceivers and for high radix optical switches along with the fabrication and manufacturing technologies required to achieve performance at commercially competitive cost and power consumption metrics are key for photonics to improve data center performance.

A high radix, low power optical switch would be a significant enabler in the development of optically switched networks for the data center. High radix switches improve network performance, as increasing the radix reduces the number of switches required for a given system and reduces the number of hops or switches the packets must traverse from source to destination. This reduces cost, power, latency and component count. Early work using discrete switching elements to build up an optical switching fabric showed promise for optically switched architectures, however optical losses and power consumption were much too high for commercial applications.[31-34] It has been shown that for high performance computing a switch radix of at least 16 and preferably 32 or 64 is required for high performance interconnection topologies.[35] With the advent of silicon photonics capabilities, several groups have demonstrated significant progress in this active area of research.[36-39] In Ref 40 it is shown that with optimization of physical layer parameters with current state-of-the-art ring resonator devices, a high radix 128 × 128 silicon photonic single chip switch fabric with tolerable power penalty is feasible. Binkert et al.[41] explore power consumption considerations and show that by adopting optical interconnects and switches the energy per bit of a 100,000 port interconnection network can be reduced by a factor of 6 compared to an all electrical implementation.

There have been considerable advances in high bandwidth pluggable optical interconnects for the data center. One of the latest being the 100G silicon photonic transceivers from Intel.[42] Going forward however, in order to achieve the higher density bandwidth interconnects the optics should move "on-board" (Figure 4). Although this concept is not new, the nearer term data center requirements have provoked vendors to push on the technology to reduce cost by establishing the Consortium for On-Board Optics (COBO) led by Microsoft, that will define the standard for optical modules that can be mounted or socketed on a network switch or adapter motherboard.[43] The initial focus is on high density 400 GbE applications for initial implementation in 2018 with large cloud providers as the early adopters. COBO will not define the optical physical layer but certain requirements that must be met to be considered COBO compliant.

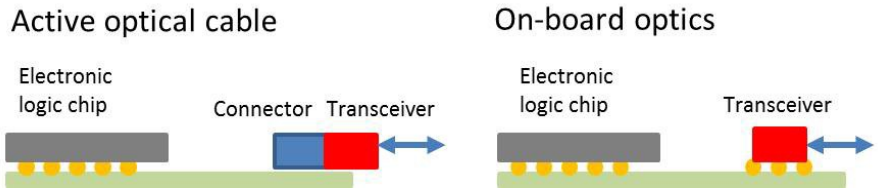

Figure 4. Schematic of active optical cable compared to on-board optics. For the active optical cable the transceiver is connected to the faceplate of the server by a connector whereas for the on-board optics the transceiver is connected to the board.

3. CONCLUSIONS

Integrated photonics, particularly in the form of high radix integrated silicon photonic optical switch fabrics and bandwidth dense on board transceivers are a promising solution to the challenge of increasing traffic requirements of the large scale data center. In addition to academic research activities the technology advances are being supported by government sponsored manufacturing initiatives such as the AIM Photonics and standards defining industry consortiums such as the Microsoft led Consortium for On Board Optics.

REFERENCES

[1] Cisco, "The Zettabyte Era" http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html

[2] Cisco "Global Cloud Index: 2013–2018" http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html

[3] Koomey, J. G., "Growth in data center electricity use 2005 to 2010", A report by Analytics Press, completed at the request of The New York Times. http://www.analyticspress.com/datacenters.html

[4] Balil, K., "Green Data Centers Networks: Challenges and Opportunities", Frontiers of Information Technology (FIT), 2013 International Conference on, DOI:10.1109/FIT.2013.49

[5] Church, Kenneth, Albert G. Greenberg, and James R. Hamilton. "On Delivering Embarrassingly Distributed Cloud Services." In HotNets, pp. 55-60. 2008.

[6] Glick, Madeleine. "Optical interconnects in next generation data centers: An end to end view." In Optical Interconnects for Future Data Center Networks, pp. 31-46. Springer New York, (2013).

[7] Kachris, Christoforos, and Ioannis Tomkos. "A roadmap on optical interconnects in data centre networks." In 2015 17th International Conference on Transparent Optical Networks (ICTON), pp. 1-3. IEEE, (2015).

[8] https://www.google.com/green/bigpicture/#beyondzero-datacenters

[9] Essiambre, René-Jean, Gerhard Kramer, Peter J. Winzer, Gerard J. Foschini, and Bernhard Goebel. "Capacity limits of optical fiber networks." Journal of Lightwave Technology 28, no. 4 (2010): 662-701.

[10] http://www.cisco.com/c/en/us/products/collateral/switches/nexus-7000-series-switches/white-paper-c11-737022.html

[11] Glick, M., D. G. Andersen, M. Kaminsky, and L. Mummert. "Dynamically reconfigurable optical links for high-bandwidth data center networks." In Optical Fiber Comm. Conf., p. OTuA3. OSA, (2009).

[12] Wang, Guohui, David G. Andersen, Michael Kaminsky, Michael Kozuch, T. S. Ng, Konstantina Papagiannaki, Madeleine Glick, and Lily Mummert. "Your data center is a router: The case for reconfigurable optical circuit switched paths." In Proc. of ACM HotNets-VIII, Oct. (2009).

[13] Farrington, N., 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 Comm. Rev. 40, no. 4 (2010): 339-350.

[14] Farrington, Nathan, Erik Rubow, and Amin Vahdat. "Data center switch architecture in the age of merchant silicon." In 2009 17th IEEE Symposium on High Performance Interconnects, pp. 93-102. IEEE, (2009).

[15] Schares, Laurent, Xiaolan J. Zhang, Rohit Wagle, Deepak Rajan, Philippe Selo, Shu-Ping Chang, James R. Giles et al. "A reconfigurable interconnect fabric with optical circuit switch and software optimizer for stream computing systems." In Optical Fiber Communication Conference, p. OTuA1. OSA, (2009).

[16] Ghobadi, Monia, Ratul Mahajan, Amar Phanishayee, Nikhil Devanur, Janardhan Kulkarni, Gireeja Ranade, Pierre-Alexandre Blanche, Houman Rastegarfar, Madeleine Glick, and Daniel Kilper. "Projector: Agile reconfigurable data center interconnect." In Proc. 2016 conference on ACM SIGCOMM 2016 Conference, pp. 216-229. ACM, (2016).

[17] Kachris, Christoforos, and Ioannis Tomkos. "A survey on optical interconnects for data centers." IEEE Communications Surveys & Tutorials 14, no. 4 (2012): 1021-1036.

[18] Funnell, Adam, Joshua Benjamin, Hitesh Ballani, Paolo Costa, Philip Watts, and Benn C. Thomsen. "High Port Count Hybrid Wavelength Switched TDMA (WS-TDMA) Optical Switch for Data Centers." In Optical Fiber Communication Conference, pp. Th2A-54. OSA, (2016).

[19] Grani, P., R. Proietti, S. Cheung, and S. Yoo. "Flat-Topology High-Throughput Compute Node with AWGR-based Optical-Interconnects." (2015).

[20] Ueda, Koh, Yojiro Mori, Hiroshi Hasegawa, Hiroyuki Matsuura, Kiyo Ishii, Haruhiko Kuwatsuka, Shu Namiki, and Ken-ichi Sato. "Demonstration of 720× 720 optical fast circuit switch for intra-datacenter networks." In SPIE OPTO, pp. 97750J-97750J. International Society for Optics and Photonics, (2016).
[21] Kanonakis, K., Y. Yin, P. Ji, and T. Wang. "SDN-Controlled Routing of Elephants and Mice over a Hybrid Optical/Electrical DCN Testbed." In Optical Fiber Comm. Conf., pp. Th4G-7. OSA, (2015).
[22] Viljoen, Nicolaas, Houman Rastegarfar, Mingwei Yang, John Wissinger, and Madeleine Glick. "Machine learning based adaptive flow classification for optically interconnected data centers." In Transparent Optical Networks (ICTON), 2016 18th International Conference on, pp. 1-4. IEEE, (2016).
[23] Rastegarfar, H., M. Glick, N. Viljoen, M. Yang, J. Wissinger, L. LaComb, and N. Peyghambarian. "TCP flow classification and bandwidth aggregation in optically interconnected data center networks." Journal of Optical Communications and Networking 8, no. 10 (2016): 777-786.
[24] Soref, Richard. "The past, present, and future of silicon photonics." IEEE Journal of selected topics in quantum electronics 12, no. 6 (2006): 1678-1687.
[25] Reed, Graham T. "Device physics: the optical age of silicon." Nature 427, no. 6975 (2004): 595-596.
[26] Vlasov, Yurii A. "Silicon CMOS-integrated nano-photonics for computer and data communications beyond 100G." IEEE Communications Magazine 50, no. 2 (2012): s67-s72.
[27] http://www.aimphotonics.com/
[28] Communications Technology Roadmap, https://mphotonics.mit.edu/ctr-documents
[29] https://aimphotonics.academy/roadmap/
[30] http://community.inemi.org/inemi-roadmap
[31] Duthie, P. J., and M. J. Wale. "16× 16 single chip optical switch array in Lithium Niobate." Electronics letters 27, no. 14 (1991): 1265-1266.
[32] Williams, K. A., G. Roberts, T. Lin, R. Penty, I. White, M. Glick, and D. McAuley. "Integrated optical 2× 2 switch for wavelength multiplexed interconnects." IEEE Jour. Selected topics in quantum elec. 11, no. 1 (2005): 78-85.
[33] Shacham, A., B. A. Small, O. Liboiron-Ladouceur, and K. Bergman. "A Fully Implemented 12 12 Data Vortex Optical Packet Switching Interconnection Network." Journal of Lightwave Technology 23, no. 10 (2005): 3066.
[34] Luijten, Ronald, Wolfgang Denzel, Richard Grzybowski, and Roe Hemenway. "Optical interconnection networks: The OSMOSIS project." In The 17th Annual Meeting of the IEEE Lasers and Electro-Optics Society.(2004).
[35] Rumley, Sébastien, Madeleine Glick, Raj Dutt, and Keren Bergman. "Impact of photonic switch radix on realizing optical interconnection networks for exascale systems." In IEEE Optical Interconnects Conference, pp. 98-99. 2014.
[36] Heck, M. J., J. F. Bauters, M. L. Davenport, J. K. Doylend, S. Jain, G. Kurczveil, S. Srinivasan, Y. Tang, and J. E. Bowers. "Hybrid silicon photonic integrated circuit technology." IEEE Jour. Selected Topics in Quantum Electronics 19, no. 4 (2013).
[37] Liu, Shiyun, Qixiang Cheng, Muhammad Ridwan Madarbux, Adrian Wonfor, Richard V. Penty, Ian H. White, and Philip M. Watts. "Low Latency Optical Switch for High Performance Computing With Minimized Processor Energy Load [Invited]." Journal of Optical Communications and Networking 7, no. 3 (2015): A498-A510.
[38] Lee, B. G., N. Dupuis, P. Pepeljugoski, L. Schares, R. Budd, J. R. Bickford, and C. L. Schow. "Silicon photonic switch fabrics in computer communications systems." Journal of Lightwave Technology 33, no. 4 (2015): 768-777.
[39] Tanizawa, K., K. Suzuki, M. Toyama, M. Ohtsuka, N. Yokoyama, K. Matsumaro, M. Seki et al. "Ultra-compact 32× 32 strictly-non-blocking Si-wire optical switch with fan-out LGA interposer." Optics express 23, no. 13 (2015): 17599-17606.
[40] Nikolova, D., S. Rumley, D. Calhoun, Q. Li, R. Hendry, P. Samadi, and K. Bergman. "Scaling silicon photonic switch fabrics for data center interconnection networks." Optics express 23, no. 2 (2015): 1159-1175.
[41] Binkert, N., A. Davis, N. Jouppi, M. McLaren, N. Muralimanohar, and J. Ho Ahn. "Optical high radix switch design." IEEE Micro 32, no. 3 (2012): 100-109.
[42] http://www.intel.com/content/www/us/en/architecture-and-technology/silicon-photonics/silicon-photonics-overview.html
[43] http://cobo.azurewebsites.net/
[44] Singla, A., A. Singh, and Y. Chen. “OSA: An optical switching architecture for data center networks with unprecedented flexibility.” NSDI’12, pages 239–252
[45] Liu, H., M. K. Mukerjee, C. Li, N. Feltman, G. Papen, S. Savage, S. Seshan, G. M. Voelker, D. G. Andersen, M. Kaminsky, G. Porter, and A. C. Snoeren. “Scheduling techniques for hybrid circuit/packet networks.” CoNext’15.
[46] Kandula, S., J. Padhye, and P. Bahl. “Flyways to de-congest data center networks.” HotNets’09, (2009).

[47] Zhou, X., Z. Zhang, Y. Zhu, Y. Li, S. Kumar, A. Vahdat, B. Y. Zhao, and H. Zheng. "Mirror mirror on the ceiling: flexible wireless links for data centers." ACM SIGCOMM Computer Comm. Review 42, no. 4 (2012): 443-454.
[48] Porter, G., 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. Vol. 43, no. 4. ACM, (2013).
[49] Hamedazimi, N., Z. Qazi, H. Gupta, V. Sekar, S. R. Das, J. P. Longtin, H. Shah, and A. Tanwer. "FireFly: a reconfigurable wireless data center fabric using free-space optics." ACM SIGCOMM Computer Communication Review 44, no. 4 (2015): 319-330.