Energy-efficient and Scalable Data Centers with Flexible Bandwidth SiPh All-to-All Fabrics

Energy-efficient and Scalable Data Centers with Flexible Bandwidth SiPh All-to-All Fabrics

Roberto Proietti*, Zhiyan Liu, Xian Xiao, Xiaoliang Chen, and S.J.Ben Yoo*

Department of Electrical and Computer Engineering, University of California, Davis, California 95616, USA

**rproietti@ucdavis.edu, sbyoo@ucdavis.edu*

Abstract: This paper presents a scalable and energy-efficient flexible-bandwidth optical interconnect architecture for data center networks. The proposed approach leverages silicon photonic reconfigurable all-to-all switch fabrics and a cognitive distributed control plane for optical reconfiguration.

1. Introduction

The rapid expansion and evolution of cloud-based services demand for data center and high-performance computing (HPC) architectures supporting high-bandwidth and low-latency communications among tens of thousands of servers. These systems are typically built with tree-based electronic packet switching (EPS) architectures (e.g., Fat Tree [1]) with point-to-point optical links for inter-rack communications. Fat Tree provides rich interconnectivity but suffers from high power consumption and end-to-end latency due to the cascaded switching stages and related optical to electrical to optical (O/E/O) conversions. As wavelength division multiplexing (WDM) transceivers are penetrating the datacom market (to sustain the ever-increasing switch port data rates - from 100G to 200G and beyond), and silicon-photonic (SiPh) technologies are becoming commercially viable, there are opportunities to leverage photonics beyond point-to-point communication by flattening the inter-rack network architecture with direct all-to-all wavelength routing and switching. Researchers have reported several hybrid optical/electrical switching architectures. These solutions leverage different switching paradigms (from slow and centralized optical circuit switching of elephant flows [2] to fast and distributed synchronous [3] or asynchronous packet switching [4]). A different optical switching paradigm involves adapting the inter-rack connectivity bandwidth (also referred to as optical reconfiguration for bandwidth steering) to the traffic patterns. This reconfiguration allows to remove link congestion due to hotspot traffic between specific rack pairs [5, 6]. The switching operation does not happen on a packet or flow basis but only when the traffic characteristic changes significantly and over time scales larger than hundreds of microseconds or milliseconds ([11]). This paper discusses our holistic approach to scalable and energy-efficient data center architectures leveraging a low-diameter rack-to-rack directly connected network with SiPh reconfigurable all-to-all fabrics. It is a Hyper-X-like optical interconnect architecture [7] that can provide scalable and flexible-bandwidth interconnections. Section 2 introduces the data plane architecture and enabling photonic technologies, touches upon the control plane architecture and algorithms for reconfiguration and routing, and it shows some performance analysis by simulations. Section 3 concludes the paper.

2. Flexible Bandwidth All-to-All Data Center Network Architecture

This section describes the proposed flexible-bandwidth photonic data center architecture, named Hyper-FleX-LION [8]. Fig. 1(a-b) shows Flex-LIONS, a silicon photonic (SiPh) bandwidth-reconfigurable all-to-all switching fabric at the core of the Hyper-FleX-LION. Flex-LIONS contains an arrayed waveguide grating router (AWGR), microring resonator (MRR) add-drop filters, and a Mach-Zehnder (MZ) switching network [9]. For an $N\times N$ Flex-LIONS, an $N\times N$ arrayed waveguide grating router (AWGR) provides all-to-all interconnection based on its wavelength-routing function. b microring resonator (MRR) drop filters at each input port of the AWGR are used to drop b wavelength division multiplexing (WDM) channels for bandwidth reconfiguration. The dropped channels are spatially switched and added to the desired output port by an $N\times N$ broadband Beneš Mach-Zehnder switch (MZS) network and b MRR add filters. In this way, the bandwidth between certain node pairs can be increased by a factor of b. As demonstrated in [9], Flex-LIONS can leverage multiple free spectral ranges (FSRs) to maintain the all-to-all interconnectivity before and after bandwidth reconfiguration. Before reconfiguration, all the WDM wavelengths in FSR0 and FSR1 of the AWGR are used for all-to-all communication (thanks to the cyclic feature of the AWGR). For bandwidth steering, only the wavelengths in FSR1 are switched by the MRR add-drop filters and the MZS network, while the wavelengths in FSR0 are untouched. In this case, the all-to-all interconnectivity is always maintained through FSR0, preventing unconnected nodes after reconfiguration. Fig. 1(b) shows the wavelength allocation before reconfiguration (both FSRs maintain the all-to-all connection). Fig. 1(c) depicts an example of wavelength allocation after reconfiguration.

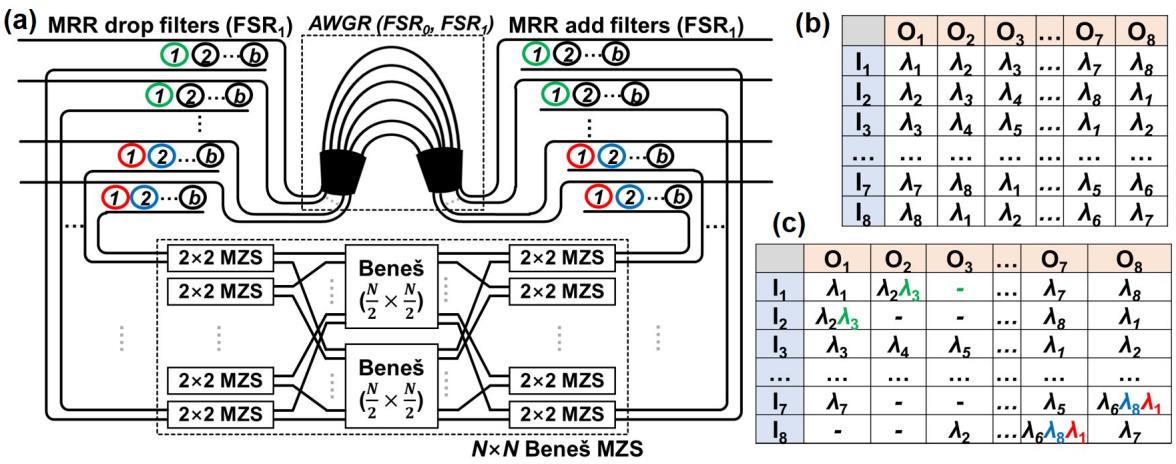

Fig. 1: (a) Block diagram of the Flex-LIONS device. Wavelength routing table (b) before and (c) after reconfiguration. The colored add/drop rings represent the rings that are tuned to the colored wavelengths in (c) to achieve bandwidth steering.

As shown in Fig. 2(a), it is possible to use Flex-LIONS to interconnect a group of P ToR switches (pod). By laying out these pods in rows and columns and using multiple Flex-LIONSs for inter-pod interconnection, we can effectively build a Hyper-FleX network (a reconfigurable Hyper-X [8]) with fixed hierarchical all-to-all connectivity on FSR0 and reconfigurability on FSR1. As discussed in [8], the architecture can scale to a larger number of nodes than a non-oversubscribed Fat Tree (for switch radix values higher than 128) while providing significant power savings (see Fig. 2(d)).

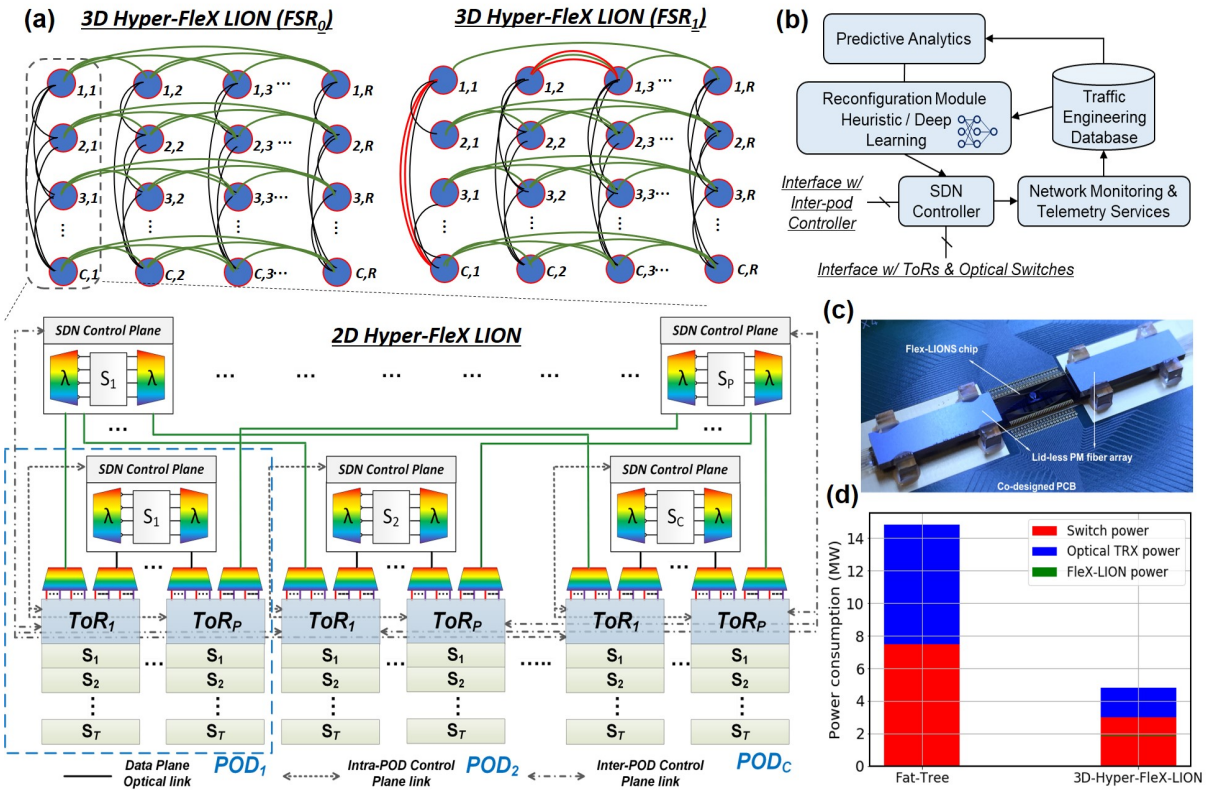

Fig. 2: (a) 3D-Hyper-FleX-LION interconnection with pods organized in rows and columns. The inset shows the implementation details of the inter-pod-level interconnection for a column or row. (b) SDN control plane architecture for each group of ToRs interconnected by one Flex-LIONS. (c) Picture of fabricated and packaged Flex-LIONS chip (courtesy of Optelligent, LLC). (d) Power consumption of 3D Hyper-FleX-LION and Fat Tree for a ToR switch radix = 128. [8]

Fig. 2(b) shows the architecture of the control plane (CP) that drives the reconfiguration operations. Note that the codesign of the data plane (hardware) and control and management plane (software and algorithms) is key for using any optical switching paradigm. The CP is centralized at the pod level but distributed between pods. For each pod, a software-defined networking (SDN) controller interfaces with the Flex-LIONS and the group of ToRs connected to it. This controller monitors the traffic demand for the ToR ports and can leverage predictive analytic [10] and a combination of heuristic [11] and deep learning tools to determine the reconfiguration policies

that best serve a specific traffic demand. Specifically, by leveraging the two FSRs described above, it is possible to design a routing, bandwidth, and topology assignment (RBTA) that can leverage FSR0 to offload the traffic from the FSR1 links that are involved in the reconfiguration operation. By codesigning the optical reconfiguration algorithm and packet-level routing in the ToRs (e.g., using weighted-cost multi-path routing), the RBTA algorithm can achieve seamless reconfiguration without loss of packets. Moreover, as FSR0 always guarantees minimum diameter connectivity, each pod in each dimension can be reconfigured independently without affecting the traffic coming from other pods, making the control scheme distributed at the inter-pod level.

Fig. 3 shows the average packet latency of Hyper-FleX-LION for two different traffic profiles with non-uniform distribution (i.e., Fill Boundary, Crystal Router [12]). We compared the results with a non-oversubscribed Fat Tree with a similar number of servers (i.e., 256). For the sake of comparison, we reported the performance of Hyper-FleX-LION without reconfiguration and with reconfiguration only inside the pods. The results show that the Hyper-FleX-LION network can provide latency and throughput comparable with Fat Tree while providing significant power savings (as shown in Fig. 2(d)).

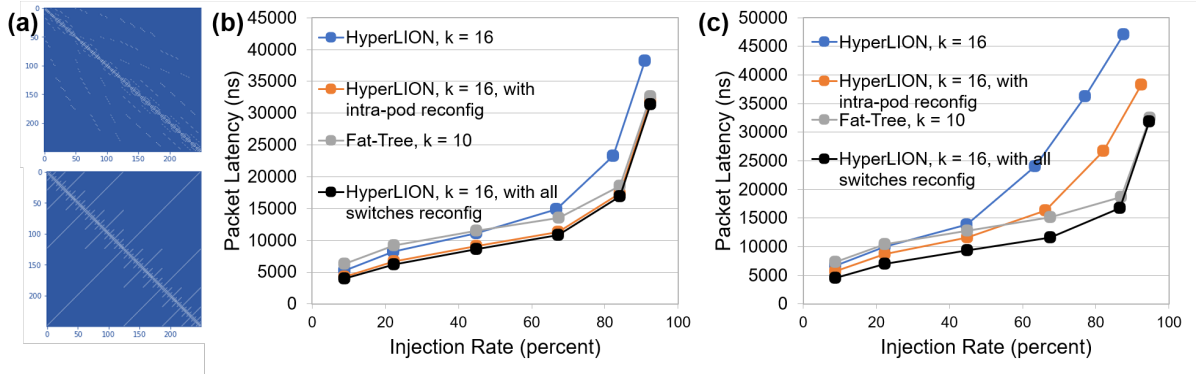

Fig. 3: (a) Traffic profiles used for the simulations [12]. (b-c) Average packet latency for Fat Tree and Hyper Flex-LION architectures for the traffic profiles of (a). All the simulations use equal-cost multi-path (ECMP) without flow splitting as the routing scheme.

3. Summary and Future Work

This paper summarizes our recent work on Hyper-FleX-LION, a flexible-bandwidth photonic interconnect architecture for cloud computing. Future studies will include testbed demonstrations of co-designed control and data planes leveraging RBTA algorithms and machine learning for traffic prediction and optimization of the reconfiguration strategies.

References

1. Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. 2008. A scalable, commodity data center network architecture. SIGCOMM Comput. Commun. Rev. 38, 4 (October 2008), 63–74.
2. N. Farrington et al., "Helios: a hybrid electrical/optical switch architecture for modular data centers," SIGCOMM Comput. Commun. Rev. 40, 4 (October 2010), 339–350.
3. H. Ballani, et al., "Sirius: A Flat Datacenter Network with Nanosecond Optical Switching," In Proceedings of ACM SIGCOMM '20.
4. X. Guo et al., "RDON: a rack-scale disaggregated data center network based on a distributed fast optical switch," in IEEE/OSA Journal of Optical Communications and Networking, vol. 12, no. 8, pp. 251-263, August 2020.
5. Z. Cao, R. Proietti, M. Clements and S. J. B. Yoo, "Experimental Demonstration of Flexible Bandwidth Optical Data Center Core Network With All-to-All Interconnectivity," in Journal of Lightwave Technology, vol. 33, no. 8, pp. 1578-1585, April 15, 2015.
6. M. Y. Teh et al., "Flexspander: augmenting expander networks in high-performance systems with optical bandwidth steering," IEEE/OSA Journal of Optical Communications and Networking, vol. 12, no. 4, pp. B44-B54, 2020."
7. J.H. Ahn et al., "HyperX: topology, routing, and packaging of efficient large-scale networks," In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis (SC '09). Association for Computing Machinery, New York, NY, USA, Article 41, 1–11.
8. G. Liu et al., "Architecture and performance studies of 3D-Hyper-FleX-LION for reconfigurable all-to-all HPC networks," in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '20). IEEE Press, Article 26, 1–16.
9. X. Xiao et al., "Multi-FSR Silicon Photonic Flex-LIONS Module for Bandwidth-Reconfigurable All-to-All Optical Interconnects," in Journal of Lightwave Technology, vol. 38, no. 12, pp. 3200-3208, June 15, 2020.
10. S. K. Singh et al., "Machine-learning-based prediction for resource (Re)allocation in optical data center networks," in IEEE/OSA Journal of Optical Communications and Networking, vol. 10, no. 10, pp. D12-D28, Oct. 2018.
11. X. Chen et al., "Machine-learning-aided cognitive reconfiguration for flexible-bandwidth HPC and data center networks [Invited]," in IEEE/OSA Journal of Optical Communications and Networking, vol. 13, no. 6, pp. C10-C20, June 2021..
12. "Characterization of the DOE Mini-apps", https://portal.nersc.gov/project/CAL/doe-miniapps.htm.

This work was supported in part by ARO award # W911NF1910470, DoD award # H98230-19-C-0209, NSF ECCS award # 1611560, and by DoE UAI consortium award # DE-SC0019582, DE-SC0019526, and DE-SC001969.