Optically Interconnected Data Center Architecture for Bandwidth Intensive Energy Efficient Networking

Howard Wang, *Student Member, IEEE*, Keren Bergman, *Fellow, IEEE*
Department of Electrical Engineering, Columbia University, 500 W. 120th St., New York, NY 10027, USA
Tel: 1 (212) 854 2768, Fax: 1 (212) 854 2900, e-mail: howard@ee.columbia.edu

ABSTRACT
The relentless rise of data-intensive cloud-based services continues to drive network performance requirements demanded by modern data centers. Scaling electronic packet-switched data center networks to provide bandwidths commensurate with traffic demands will either be prohibitively costly, overly complex, or result in unsustainable energy requirements. Network oversubscription, commonly used today to reduce costs, further exacerbates the performance bottleneck of communication-intensive applications. Recent network architectures based on optical circuit switching leverage the fundamentally higher capacity and energy efficiency of photonic technologies and can deliver bisection bandwidths comparable to fully provisioned packet-switched networks. However, in terms of connectivity, the high bandwidth optical paths are limited to one-to-one matching between racks within the network. We present an enhanced optically connected network architecture featuring advanced photonic functionalities to support a wider class of bandwidth-intensive traffic patterns characteristic of cloud computing systems. This proposed architecture can enable a rich set of photonic resources to be allocated on-demand to optimize communications between various applications within the data center. We construct a prototype of the proposed optical network architecture and demonstrate two unique functionalities, validating the physical layer feasibility of the system.
Keywords: optical network architecture, data center networks, reconfigurable optical network, optical multicasting

1. INTRODUCTION

The rapid ascendance of cloud computing has resulted in the creation of larger and more powerful data centers featuring substantial inter-node communications requirements, with some systems calling for petabits per second of aggregate bandwidth across hundreds of thousands of ports [1]. Unfortunately, conventional electronic switching technologies carry costs that scale super-linearly with bandwidth and port count, forcing data center designers to introduce oversubscription at the upper tiers of the network. Consequently, data-intensive computations commonly run on these systems become severely bottlenecked when inter-rack information exchange is required. Worse still, as these systems are typically designed to be utilized as a centralized pool of computational resources, traffic volatility arising from application heterogeneity further complicates optimization of the network through capacity engineering [2].

In addition to network performance bottlenecking, energy consumption has quickly become a dominant constraint in modern data center designs [3]. Contemporary electronic networks can consume hundreds of kilowatts of power, with this figure rising rapidly as larger and faster switches are installed. The staggering power densities associated with these electronic switches necessitate sophisticated cooling systems, further reducing overall data center energy efficiencies. Moreover, measurements on current data center deployments indicate significant wasted energy due to underutilization, with data-starved servers recording average utilization values of as low as 30% [4].

As a result, alleviating inter-rack communication bottlenecks has become a critical target in architecting next-generation data centers, accelerating both the execution of large-scale distributed applications and significantly reducing underutilization due to data starvation. While there have been notable architectural and algorithmic studies focused on improving data center network performance [5,6], they remain constrained by the limitations of electronic switching in terms of bandwidth and power consumption.

Optical interconnects hold distinct advantages over their copper counterparts—namely large bandwidth, bit-rate transparency, and distance immunity—thus representing a potentially disruptive solution for overcoming these constraints. Recent studies have explored the viability of augmenting existing oversubscribed hierarchical electronic packet-switched (EPS) networks with off-the-shelf circuit-switched MEMS-based optical switches [7,8]. These pioneering proposals have successfully demonstrated the potential of utilizing photonic technologies to ease bandwidth constraints within the context of data center traffic while achieving reduced complexity, component cost, and power in comparison to conventional fully electronic networks. However, in these systems, efficient utilization is achieved only by traffic patterns exhibiting pairwise communications over sufficiently long timescales. This can be attributed to the relatively slow switching speed of MEMS technologies in combination with the limited connectivity achievable by the singular space switch, i.e. one-to-one matchings between racks.

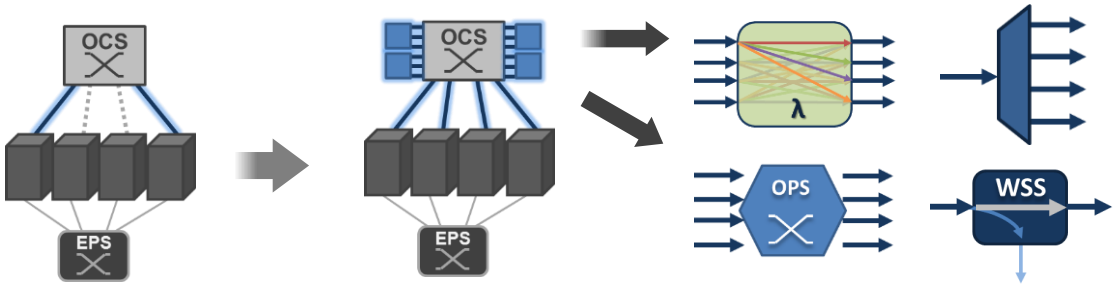

Fig. 1. A variety of optical functionalities can be used to enhance the optically connected data center network, increasing bandwidth granularity and connectivity, and yielding more efficient utilization of the photonic layer.

Recognizing that various subsets—as opposed to just pairs—of communicating racks may demand bandwidth with varying amounts of connectivity, an enhanced optically-connected network architecture enabling a diverse set of photonic functionalities can more effectively support a larger class of bandwidth-intensive traffic patterns often found within the data center. Each photonic subsystem is treated as a physical resource allocable on-demand, providing more optimal connectivity as demanded by various applications within the data center. A network prototype featuring four emulated end nodes connected to various optical functions through an optical space switch is constructed to evaluate the feasibility of this concept [9].

2. RECONFIGURABLE PHOTONIC RESOURCES

While network traffic is characteristically unpredictable due to application heterogeneity, communication patterns where only a few top-of-the-rack (ToR) switches are tightly coupled with long, extended data flows have been observed in production data centers [10]. Therefore, the utility of architectures based purely on point-to-point MEMS-switches are reliant on the inherent stability of traffic patterns within such systems. Nevertheless, further bandwidth flexibility remains a key target for future data center networks as applications require even higher capacities with increased interconnectivity demands. When architectures utilizing purely commercial MEMS-based switches are evaluated with richer communication patterns more representative of realistic applications, a significant proportion of bandwidth-intensive traffic is suboptimally mapped across the limited connectivity supported by the singular optical circuit switch, thus limiting tangible performance benefits and putting the practicality of such systems into question [2].

Although traffic heterogeneity is an unavoidable characteristic of data center systems, costly full-bandwidth all-to-all communication across the entire network remains unnecessary. As a result of multitenancy, it is rare for a single application to span an entire data center. Instead communications are typically limited to a multitude of logical subnets that vary in time as application execution and resource usage evolves across the system. For example, various management tasks represent classes of traffic that do not necessarily adhere to pairwise communications and require significant bandwidth between subsets of racks throughout the system [11]. Tasks such as virtual machine (VM) migration, provisioning, backup, and shuffling typically exhibit traffic patterns ranging from one-to-one, one-to-many, many-to-one, and many-to-many-type connectivity requirements, respectively.

In order to more effectively satisfy the diverse network requirements imposed by these bandwidth-intensive tasks, we have proposed a hybrid network concept wherein a reconfigurable pool of photonic functionalities are utilized to achieve a high degree of traffic adaptability. Advancements in photonic devices and switch architectures have enabled transmission capacities on the scale of terabits per second per link with high-radix switching, offering high-throughput interconnectivity for tens of thousands of nodes. As a result, the range of viable optical functionalities—such as switching speed, switching granularity, and connectivity granularity—represents a compelling design space in the context of inserting photonic technologies into what has been a traditionally electronically packet-switched cost-constrained environment. Such functionalities can be dynamically allocated on-demand to different subsets of communicating nodes across the system when and where they are required, thus providing a rich set of optical connectivity options. Furthermore, as the scale and specific traffic requirements of individual applications evolve, components can be combined to form more sophisticated functionalities when necessary. In practice, these functionalities can be attached to a subset of the ports of an optical circuit switch in the previously proposed architectures (Fig. 1). Resources can be managed by a central controller, which can either accept explicit requests for resources or allocate the resources based on demand estimation. Using this scheme, the superior capacities offered by photonics can be more efficiently utilized, allowing for more flexibility in demand ambiguity and higher granularity bandwidth allocation, potentially leading to a more simplified control plane. Furthermore, the modularity of this enhanced network architecture is well suited to the incremental nature of data center expansion. As the data center grows or its needs change, additional discrete functionalities can be added as needed by simply attaching or removing optical resources accordingly. Outside of occupying a small number of ports on the high radix space switch, the addition

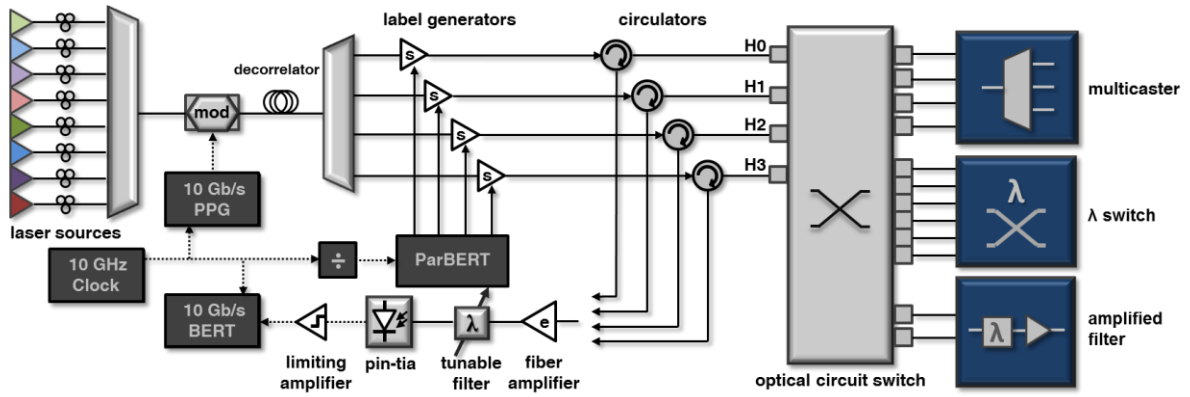

Fig. 2. The experimental setup emulating four nodes transmitting at up to 80 Gb/s utilizing a reconfigurable set of optical network functions.

of optical resources will not adversely affect the basic functionality of the system; every addition should represent an incremental improvement, minimizing initial costs and reducing the risk associated with implementation.

3. EXPERIMENTAL DEMONSTRATION

We have validated the feasibility of our proposed design through the construction of a network demonstrator featuring a number of on-demand photonic resources to be dynamically allocated to four 80 Gb/s-capable I/O ports via an optical circuit switch. Two unique photonic functionalities – optical multicasting and the formation of an optical local area network (OLAN) – were demonstrated in a number of configurations. The implemented experimental setup is depicted in Figure 2. Additional details regarding the experiment can be found in [9].

Each optical resource is implemented using a combination of passive and active optical components and connected to various ports of a Polatis 24×10 piezoelectric beam-steering optical space switch. The switch accepts SCPI commands via Ethernet from an external computer, which is used to configure the appropriate optical connections for a given topological configuration. Five different network configurations were demonstrated by rearrangeably interconnecting three photonic resources, delivering a rich set of connectivity options to the four emulated ports: H0, H1, H2, and H3. The following configurations are demonstrated: 1) optical multicasting from H0 to H1, H2, and H3 at 80 Gb/s per port (Fig. 3a); 2) multicasting from H3 to H0, H1, and H2; 3) OLAN generation between nodes H0, H1, and H2 at 60 Gb/s aggregate bandwidth per port (Fig.

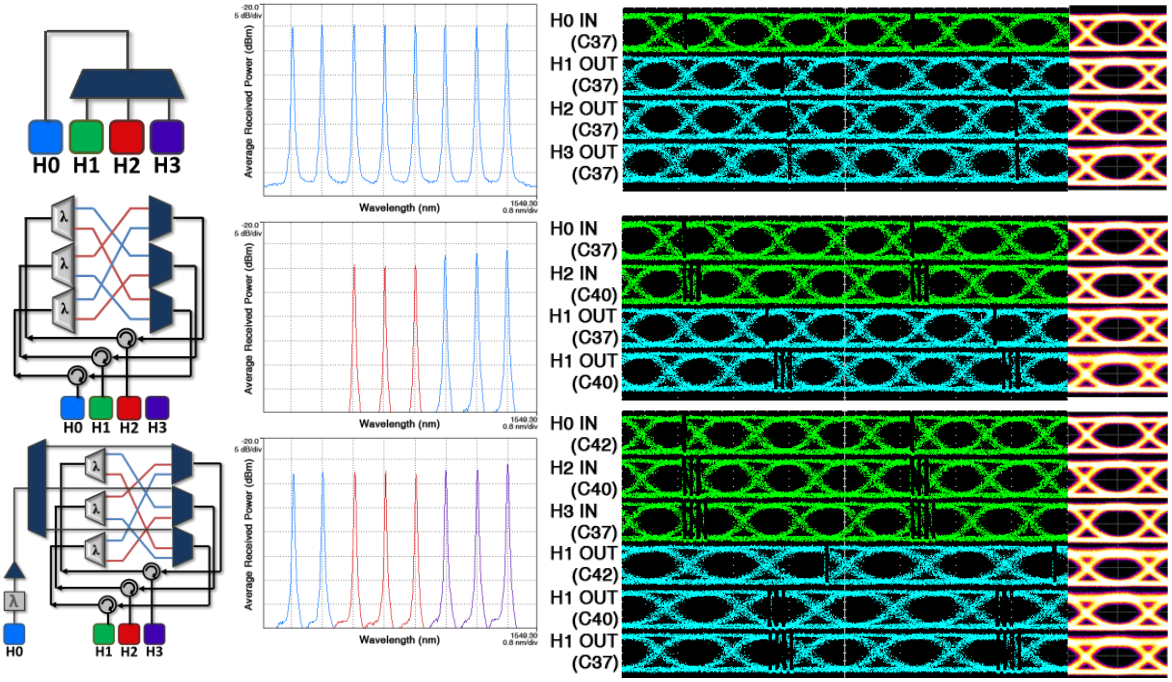

Fig. 3. (From left to right) The effective network topology, output spectrum at H1, and recovered electronic waveforms with associated optical eye diagrams at select input and output ports for a) an 80 Gb/s optical multicast from H0 to H1, H2, and H3; b) an OLAN between H0, H1, and H2 featuring simultaneous 30 Gb/s connections to each peer; and c) a combined multicast (60 Gb/s at H1, H2, and H3) and broadcast (20 Gb/s from H0). The multicaster in the last configuration utilizes an amplifier/filter resource to account for the extra loss in its path. Each stream is inscribed with a low speed (155 Mb/s) pattern denoting its origin port.

3b); 4) OLAN generation between H1, H2, and H3; and 5) 60 Gb/s OLAN generation between H1, H2, and H3 combined with a 20 Gb/s broadcast from H0 to each node in the OLAN (Fig. 3c). Figure 4 details three of the aforementioned configurations along with the associated spectra, recovered electrical waveforms, and optical eye diagrams. Bit-error rates are measured at each channel across all configurations, with all data recovered at bit-error rates better than 10^{-12}.

4. CONCLUSIONS

As scaling electronic packet-switched data center networks to provide bandwidths commensurate with traffic demands will either be prohibitively costly, overly complex, or result in unsustainable energy requirements. Leveraging the fundamentally higher capacity and energy efficiency of photonic technologies can potentially relieve network bottlenecks while achieving significant cost savings. However, current MEMS-based circuit-switched data center network designs provide limited capacity granularity and connectivity, resulting in ineffective utilization of the bandwidths offered by optics. We have proposed a new paradigm for optically connected data centers featuring dynamically reallocable photonic network resources. By providing a rich set of functionalities in the optical domain, support for a larger subset of bandwidth-intensive data center traffic can be realized.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support for this work from the NSF ERC on Integrated Access Networks (CIAN) (subaward Y503160).

REFERENCES

[1] A. Vahdat, M. Al-Fares, N. Farrington, R. N. Mysore, G. Porter, and S. Radhakrishnan, “Scale-out networking in the data center,” *IEEE Micro*, 30(4), 29-41 (2010).

[2] H. H. Bazzaz, M. Tewari, G. Wang, G. Porter, T. S. E Ng, D. G. Andersen, M. Kaminsky, M. A. Kozuch, A. Vahdat, “Switching the Optical Divide: Fundamental Challenges for Hybrid Electrical/Optical Datacenter Networks”, *In Proceedings of SOCC’11: ACM Symposium on Cloud Computing*, Cascais, Portugal, Oct. 2011.

[3] D. Abts, M. R. Marty, P. M. Wells, P. Klausler, and H. Liu, “Energy proportional datacenter networks,” in Proc. of 37th Annu. Int. Symp. Computer Architecture (ISCA ’10), 2010, pp. 338–347.

[4] D. Meisner, B. T. Gold, and T. F. Wenisch, “PowerNap: eliminating server idle power,” in Proc. of the 14th Int. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS ’09), 2009, pp. 205–216.

[5] M. Al-Fares et al. “A scalable, commodity data center network architecture.” *SIGCOMM Comput. Commun.* Rev., 38(4):63–74, 2008.

[6] A. Greenberg et al. “Vl2: a scalable and flexible data center network.*” SIGCOMM Comput. Commun.* Rev., 39(4):51–62, 2009.

[7] Farrington, N., Porter, G., Radhakrishnan, S., Bazzaz, H. H., Subramanya, V., Fainman, Y., Papen, G., and Vahdat, A. “Helios: a hybrid electrical/optical switch architecture for modular data centers.” *SIGCOMM Comput. Commun. Rev.* 40, 4 (Aug. 2010), 339-350.

[8] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. E. Ng, M. Kozuch, and M. Ryan. “c-Through: part-time optics in data centers.” *SIGCOMM Comput. Commun.* Rev. 40, 4 (August 2010), 327-338.

[9] H. Wang, C. Chen, K. Sripanidkulchai, S. Sahu, K. Bergman, "Dynamically Reconfigurable Photonic Resources for Optically Connected Data Center Networks," *Optical Fiber Communication Conference* (OFC) 2012 OTu1B.2 (Mar 2012).

[10] Benson, T., Anand, A., Akella, A., and Zhang, M. 2009. “Understanding data center traffic characteristics.” In Proceedings of the 1st ACM Workshop on Research on Enterprise Networking (Barcelona, Spain, August 21 - 21, 2009). WREN '09. ACM, New York, NY, 65-72.

[11] Soundararajan, V., and Anderson, J.M. 2010. “The impact of management operations on the virtualized datacenter.” In *Proceedings of the 37th annual international symposium on Computer architecture* (ISCA '10). ACM, New York, NY, USA, 326-337.