Dynamically Reconfigurable Photonic Resources for Optically Connected Data Center Networks

Howard Wang[1], Cathy Chen[1], Kunwadee Sripanidkulchai[2], Sambit Sahu[3], and Keren Bergman[1]
[1]*Department of Electrical Engineering, Columbia University, 500 West 120th Street, New York, New York 10027*
[2]*NECTEC, 112 Phahon Yothin Road, Klong Luang, Pathumthani 12120, Thailand*
[3]*IBM T.J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598*
Email: howard@ee.columbia.edu

Abstract: A modular, traffic-adaptable data center network architecture leveraging unique reconfigurable optical functionalities is presented. Optical multicasting and subnet formation is demonstrated between four ports at up to 80 Gb/s per port.

OCIS codes: (060.4250) Networks; (200.4650) Optical interconnects

1. Introduction

The relentless rise in popularity of data-intensive cloud-based services continues to raise the already unsustainable network performance requirements demanded by modern data centers. Moreover, application heterogeneity and multitenancy in cloud computing systems preclude the utility of any “one size fits all” network optimizations. Unfortunately, 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. To make matters worse, in order to maintain costs within reason, network oversubscription is common practice, further exacerbating the already severe performance bottlenecking of communication-intensive applications.

In an effort to overcome these challenges, researchers have proposed the insertion of high-radix MEMS-based optical circuit switches into the core of standard hierarchical electronic networks [1,2]. By leveraging the fundamentally higher capacity and energy efficiency of photonic technologies, bisection bandwidths comparable to fully provisioned packet-switched networks can be delivered while achieving significant cost savings. However, in terms of connectivity, the high bandwidth optical paths are limited to one-to-one matchings between racks within the network. Combined with the limited switching speed of MEMS technologies, efficient utilization is achieved only by traffic patterns exhibiting pairwise communications over sufficiently long timescales. As a result, a significant proportion of bandwidth-intensive traffic is suboptimally mapped across the limited connectivity supported by a singular optical circuit switch, thus limiting tangible performance benefits and putting the practicality of such systems into question [3].

In order to more effectively utilize the unique capabilities offered by optics, we propose 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. By recognizing the fact that various subsets – as opposed to just pairs – of communicating racks may demand high-bandwidths at varying levels of connectivity, our 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. A network prototype featuring various optical functions attached to an optical space switch is constructed. Four emulated end nodes, each communicating at up to 80 Gb/s, are connected in a variety of unique topologies, demonstrating the physical layer feasibility of this concept. Each high bandwidth data stream is received and tested for errors, ensuring bit-error rates of less than 10^{-12} for each configuration across all channels.

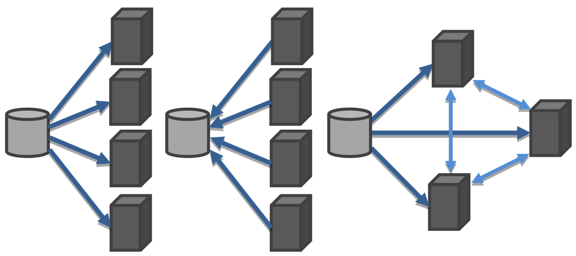

Fig. 1. Management tasks exhibit communication patterns that typically require simultaneous high-bandwidth connections between multiple compute and storage nodes within the data

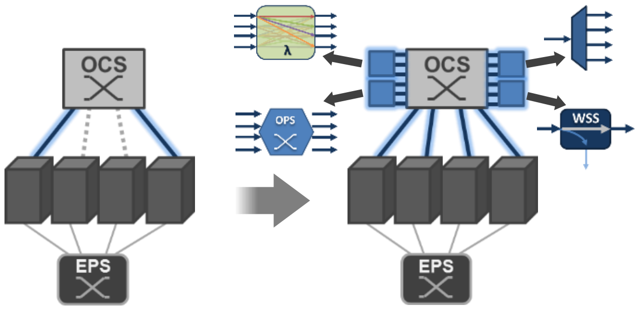

Fig. 2. A variety of optical functionalities can be used to enhance the optically connected data center network, increasing bandwidth granularity and connectivity, and yielding more efficient utilization of the photonic layer.

2. Reconfigurable Photonic Resources for Data Center Networks

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 [4]. 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 (Fig. 1).

In order to more effectively satisfy the diverse network requirements imposed by these bandwidth-intensive tasks, our proposed hybrid network concept utilizes a reconfigurable pool of advanced photonic functionalities to potentially realize extremely high levels of traffic adaptability. Advanced functionalities, such as optical multicasting/broadcasting, wavelength routing, and optical packet switching, 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. 2). 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 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 Setup and Evaluation

In order to validate the feasibility of our proposed design, we constructed a small-scale test bed 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) – are demonstrated in a number of configurations. The implemented experimental setup is depicted in Figure 3. Eight C-band 100 GHz-spaced CW distributed feedback laser sources ranging from 1542.94 nm (C43) to 1548.52 nm (C36) on the ITU grid are combined and simultaneously modulated with a 2^{15}-1 PRBS at 10 Gb/s by a $LiNbO_3$ modulator. The resulting WDM data stream is decorrelated by approximately 10 km of SMF-28 fiber and distributed to four SOA label generators, which serve to provide a unique low-speed modulation on each stream to facilitate source identification at the output of the network. Optical circulators are used at the switch input to multiplex counter-propagating data streams for bidirectional data transmission in the switch. Measurements on each channel are taken at the output ports of each circulator via a receiver chain consisting of an EDFA and tunable grating filter followed by an integrated 10 Gb/s PIN-TIA-LA assembly.

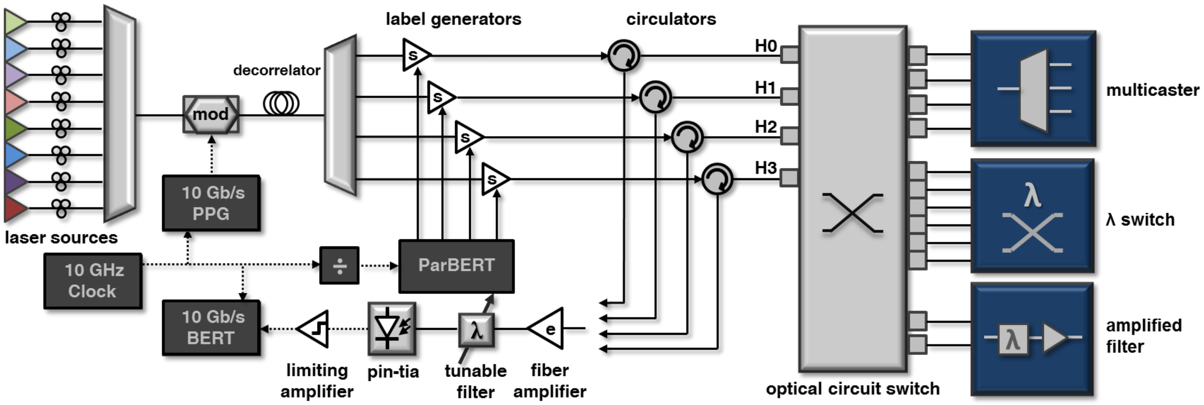

Fig. 3. The experimental setup emulating four nodes transmitting at up to 80 Gb/s utilizing a reconfigurable set of optical network functions.

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 are 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. 4a); 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. 4b); 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. 4c). 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}.

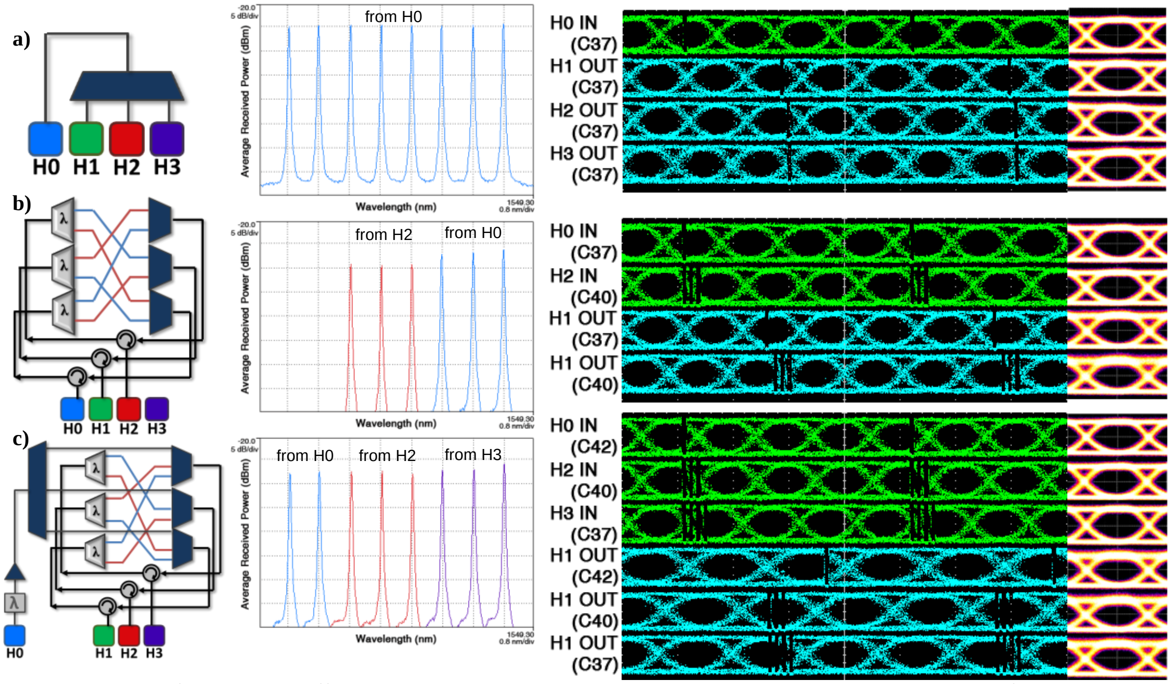

Fig. 4. (From left to right) The effective network topology, output spectrum at H1, and recovered electronic waveforms with associated optical eye diagrams at select input and output ports for a) an 80 Gb/s optical multicast from H0 to H1, H2, and H3; b) an OLAN between H0, H1, and H2 featuring simultaneous 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.

4. Conclusions

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 propose 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. We have experimentally demonstrated a variety of network configurations confirming the physical layer feasibility of this concept.

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

4. References

[1] 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

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

[3] Bazzaz, H.H., Tewari, M., Wang, G., Porter, G., Ng, T.S.E., Andersen, D.G., Kaminsky, M., Kozuch, M.A., and Vahdat, A. “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.

[4] 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.