A Bidirectional 2×2 Photonic Network Building-Block for High-Performance Data Centers

Howard Wang and Keren Bergman

Department of Electrical Engineering, Columbia University, 500 West 120th Street, New York, New York 10027
Author e-mail address: howard@ee.columbia.edu

Abstract: The design of a unique bidirectional 2×2 SOA-based optical switch is detailed and experimentally validated. Error-free operation and routing correctness is verified for 4×10Gb/s pseudorandom messages.
©2011 Optical Society of America
OCIS codes: (060.4250) Networks; (200.4650) Optical interconnects

1. Introduction

The tremendous growth in scale experienced by data centers in recent years can be attributed to the rapid ascendance in popularity of cloud-oriented media and services. Unfortunately, the communication-intensive workloads that are characteristic of these applications have already begun to overwhelm today's data centers due to the incommensurate bandwidths offered by the network. To further exacerbate the matter, the heterogeneous application classes typically supported in data centers have given rise to unpredictable traffic patterns within the network consisting of both bursty and long-lived message traffic [1,2]. While significant efforts in architectural innovation have been made with the aim of increasing overall network bisectional bandwidth [3,4], the proposed topologies come at the expense of power and complexity, ultimately being constrained by the underlying electronic switches.

By leveraging wavelength division multiplexing (WDM), bit-rate transparency, and a characteristically large bandwidth-distance product, optical technologies can enable the superlative capacities and large physical scales demanded by current and future data centers [5]. Recently, researchers have realized the potential of inserting circuit-switched optics into data centers, yielding significantly enhanced network performance with reduced power and cost over comparable electronic architectures [6,7]. However, the commercial MEMS-based switches employed in these networks are limited to millisecond-scale switching speeds. As a result, the performance improvement delivered by the high-capacity optics remains limited to only a subset of the traffic classes in the network [6,7]. Fortunately, a number of high-performance multi-stage network architectures have been realized utilizing photonic switch architectures based on semiconductor optical amplifiers (SOAs), which are capable of achieving ultra-high bandwidths at sub-nanosecond switching speeds [8].

Due to their ability to exploit locality, support simple routing protocols, and be rearrangeably non-blocking, fat tree topologies are nearly ubiquitous in high-performance computing systems and have recently garnered significant attention in data center networks. However, realizing the functionality of the requisite bidirectional switch using standard SOA-based architectures requires significant component cost in terms of SOA gating elements. In this work, we present a design for a truly bidirectional 2x2 photonic switch for tree-based data center network architectures. By leveraging the inherent bidirectional transparency of SOAs, we can achieve a minimal switch design utilizing just six SOA gate devices, representing a 63% percent savings in the number of devices required to implement photonic fat tree architectures. Functionality of the switch is demonstrated by establishing nanosecond-scale circuits while correctly routing short 40Gb/s WDM optical messages (4×10Gb/s) messages at a bit-error rates of less than 10^{-12}.
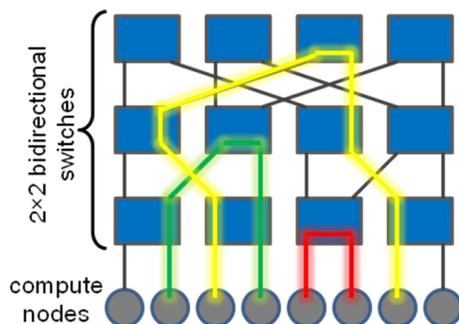

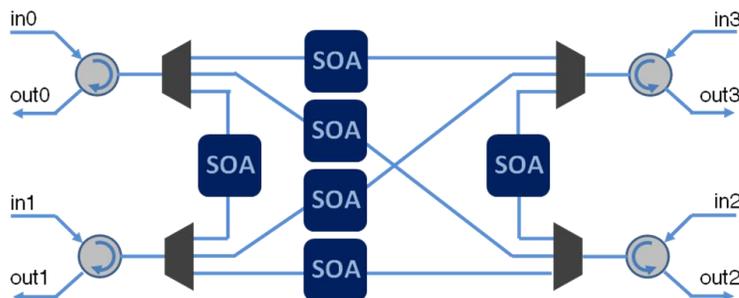Fig. 1. 2-ary 3-tree fat tree network topology interconnecting eight compute nodes.

Fig. 2. Proposed SOA-based wideband bidirectional 2×2 photonic switch.

2. Bidirectional Switching Node Design

A prototypical fat tree network (2-ary 3-tree) built using fixed arity switches is depicted in Fig. 1. The highlighted paths demonstrate the method with which communication paths are established between nodes in the network with varying levels of locality. Much like Banyan networks, k-ary n-trees support k^n processing nodes with n stages of k^{n-1} $k \times k$ switches, resulting in minimal network diameters. The simplest switch that can be used to construct a fat tree topology is a bidirectional 2×2 switch. However, utilizing a unidirectional broadcast-and-select photonic switch design would necessitate the use of a 4×4 switch in order to realize the equivalent functionality of a 2×2 bidirectional switch. The number of SOAs required in broadcast-and-select architectures scale as k^2 and, as a result, would require 16 SOAs to implement. Thus, they represent significant resources in terms of cost, footprint, and power, especially when scaling to larger network diameters.

We propose a unique switch design that exploits the bidirectional transparency of the SOA and the underlying optical medium resulting in a cost-effective and power-efficient architecture. Fig. 2 schematically illustrates the structure of the proposed photonic switch. Much like in conventional broadcast-and-select architectures, optical messages are passively split at the inputs and gated by SOAs connected to each corresponding output port. Each port is logically connected to each of the other ports by three SOAs (we assume messages will not need to be routed back to the originating port). However, by allowing them to operate bidirectionally, each SOA is, in effect, shared by two input ports. As a result, the entire node requires a total of only six SOAs to implement a fully bidirectional 2x2 switch, representing a significant advantage in terms of component cost, power, and footprint. The SOA gain is set so that the resultant power loss from the passive split is recovered within the node. Due to the broad gain bandwidth of the SOA gates, high-bandwidth wavelength-striped messages can be utilized to maximize the capacity of the switch. The switch operates in three distinct states – bar, cross, and u-turn – representing the possible source-destination pairs that can exist at any given time. Although simultaneous asymmetric message exchanges (i.e. message reception from one port while simultaneously transmitting to another port) are prohibited, we envision this switch to be used in a circuit-switched fashion in hierarchical tree topologies, thus forgoing the need for such exchanges. While this architecture can be straightforwardly modified to support wavelength-routed messages, we limit our exploration and demonstration in this work to fast circuit generation and message routing.

3. Experimental Demonstration and Results

An empirical demonstration and validation of the functionality of the proposed switch is performed via the experimental setup illustrated in Fig. 3. Four CW distributed feedback laser sources ranging from 1546.12 nm (C39) to 1548.52 nm (C36) 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 then decorrelated by approximately 10 km of SMF-28 fiber and distributed to four SOA gates generating packets of varying length separated by a minimum dead time of 6.4 ns. The packets are then injected into the switch via optical circulators, which serve to segregate the input and outputs of the switch. The circuits within the switch are set by an Agilent ParBERT system operating as a programmable pattern generator. The outputs of are then fed to an EDFA and tunable grating filter to isolate each channel to be received by a PIN-TIA receiver with integrated limiting amplifier.

The optical waveforms depicted in Fig. 4 demonstrate the correct operation and routing of the proposed 2×2 bidirectional switch. The switch is set in three circuit phases, corresponding to the three possible states of the switch. Each phase is 185.6ns in length with a dead time of 6.4 ns between subsequent phases. During each phase, a sequence of messages intended to fully exercise all possible transmission states is injected into the switch in a time-
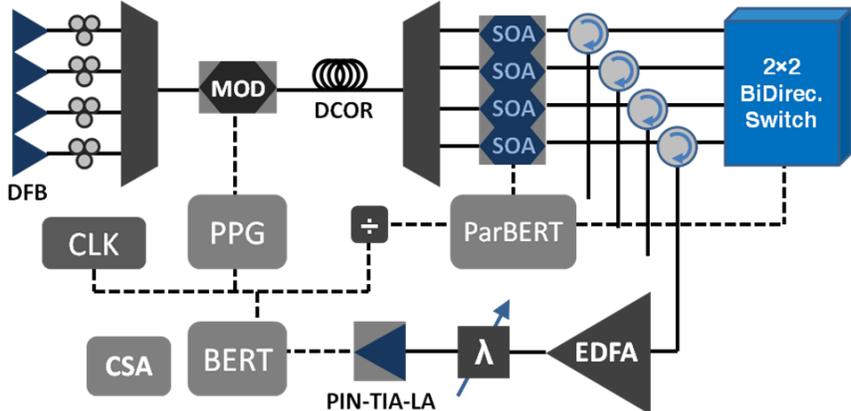

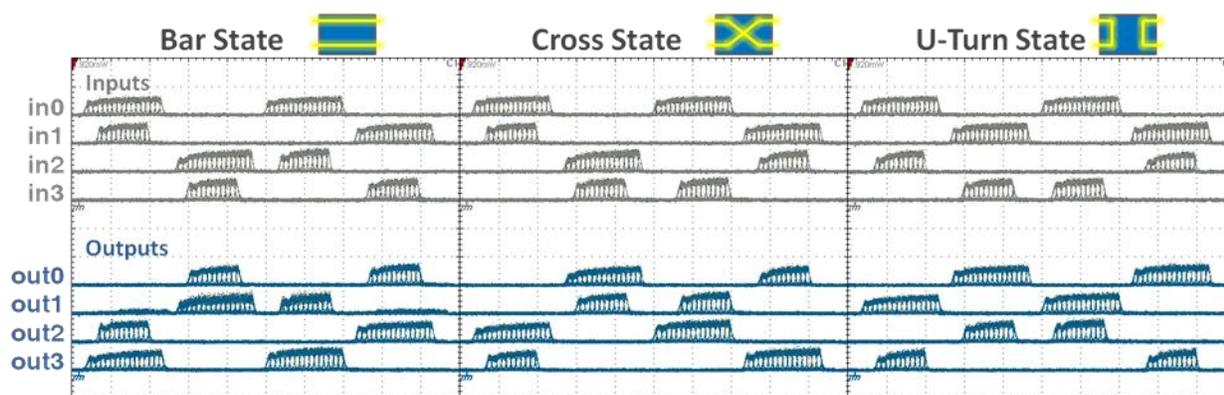Fig. 3. Schematic diagram representing the experimental setup built for evaluating the 2×2 bidirectional switch.

Fig. 4. Optical waveforms at the input and output of the switch for λ = 1546.12 nm.
Each state is fully exercised by the message patterns at the input of the switch.

slotted fashion. In order to facilitate the determination of message identities at the output of the switch one "short" message and one "long" message, each 25.6 ns and 38.4 ns in time, respectively, is transmitted at each time slot.

Furthermore, signal integrity is confirmed via the verification of error-free transmission across the switch. Bit error rates of less than 10^{-12} are confirmed across all four wavelengths for every message in each possible switch state, including, in particular, back-to-back counter-propagating messages. Open eye diagrams are observed at both the input and output of the switch for all four wavelengths (Fig. 5).
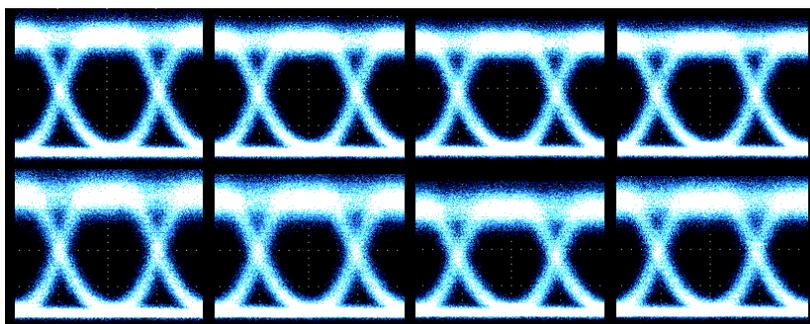

Fig. 5. 10 Gb/s optical eye diagrams at in0 (top) and out3 (bottom) for all four wavelengths (left to right: C36 to C39).

4. Conclusions

We presented and experimentally validated a cost-effective and power-efficient SOA-based 2×2 photonic switch architecture for use in fat tree architectures for next-generation high-performance data centers. By leveraging the bidirectional transparency of SOAs, we achieved a design utilizing a minimal number of active devices, thereby offering significant cost and power savings over conventional SOA-based designs.

5. References

[1] 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.
[2] S. Kandula, J. Padhye, P. Bahl, "Flyways To De-Congest Data Center Networks," In *Proc. ACM HotNets-VIII*, October 2009.
[3] Al-Fares, M., Loukissas, A., and Vahdat, A. 2008. "A scalable, commodity data center network architecture." in *Proc. of the ACM SIGCOMM 2008 Conf. on Data Communication* (Seattle, WA, USA, August 17 - 22, 2008). SIGCOMM '08. ACM, New York, NY, 63-74.
[4] Greenberg, A., Hamilton, J. R., Jain, N., Kandula, S., Kim, C., Lahiri, P., Maltz, D. A., Patel, P., and Sengupta, S. 2009. "VL2: a scalable and flexible data center network." *SIGCOMM Comput. Commun. Rev.* 39, 4 (Aug. 2009).
[5] Glick, M., "Optical Interconnects in Next Generation Data Centers: An End to End View," *HOTI '08. 16th IEEE Symposium on High Performance Interconnects, 2008*. vol., no., pp.178-181, 26-28 Aug 2008.
[6] Farrington, N., Porter, G., Radhakrishnan, S., Bazzaz, H. H., Subramanya, V., Fainman, Y., Papen, G., and Vahdat, A. 2010. "Helios: a hybrid electrical/optical switch architecture for modular data centers." *SIGCOMM Comput. Commun. Rev.* 40, 4 (Aug. 2010), 339-350
[7] Wang, G., Andersen, D. G., Kaminsky, M., Papagiannaki, K., Ng, T. E., Kozuch, M., and Ryan, M. 2010. "c-Through: part-time optics in data centers." In *Proceedings of the ACM SIGCOMM 2010 Conference on SIGCOMM* (New Delhi, India, August 30 - September 03, 2010). SIGCOMM '10. ACM, New York, NY, 327-338.
[8] A. Shacham, K. Bergman, "An Experimental Validation of a Wavelength-Striped, Packet Switched, Optical Interconnection Network," *Journal of Lightwave Technology* **27** (7) 841-850 (Apr 1, 2009).