A Multiport Microsecond Optical Circuit Switch for Data Center Networking

Nathan Farrington, *Member, IEEE*, Alex Forencich, *Student Member, IEEE*, George Porter, P.-C. Sun, Joseph E. Ford, *Member, IEEE*, Yeshaiahu Fainman, *Fellow, IEEE*, George C. Papen, *Member, IEEE*, and Amin Vahdat

***Abstract*— We experimentally evaluate the network-level switching time of a functional 23-host prototype hybrid optical circuit-switched/electrical packet-switched network for datacenters called Mordia (Microsecond Optical Research Datacenter Interconnect Architecture). This hybrid network uses a standard electrical packet switch and an optical circuit-switched architecture based on a wavelength-selective switch that has a measured mean port-to-port network reconfiguration time of 11.5 μs including the signal acquisition by the network interface card. Using multiple parallel rings, we show that this architecture can scale to support the large bisection bandwidth required for future datacenters.**

***Index Terms*— Networks, optical switching.**

I. Introduction

THE performance of large-scale data centers is limited by the internal network connectivity. Researchers have addressed this issue by developing hybrid network architectures that consist of a combination of electrical switching and optical switching. Optical circuit switching (OCS) in large-scale computing systems was initially proposed within the high-performance computing community [1]–[3]. More recently, OCS has been investigated for use in large-scale datacenters, which are designed around commodity computers [4]–[9].

The potential advantages of using an OCS within a datacenter include reduced power consumption and transparency with respect to the modulation format and the data rate. This transparency enables the same OCS to be used for multiple generations of optical interconnect and thus can reduce the cost of the network infrastructure.

Most existing approaches for using an OCS within a datacenter are based on finding large, stable flows and routing them over optical circuits, while forwarding the bulk of the traffic through the electrical packet-switched (EPS) network. These architectures are limited by the switching speed (10s of milliseconds) of commercially available MEMS-based optical circuit switching technology. In order to achieve reasonable network efficiency, this slow speed makes it necessary to form traffic aggregates from multiple end hosts that remain stable over a time interval that is long with respect to the switch speed. This efficiency is measured using the duty cycle that an optical circuit is active relative to the switching time when the circuit is being reconfigured. This kind of hybrid system is appropriate at the aggregation or higher levels of the network.

In this letter, we consider a different kind of optical circuit switch based on a wavelength-selective switch (WSS) instead of a 3D MEMS switch. The WSS has a switching speed on the order of ten microseconds. At this speed, which is approximately three orders of magnitude faster than the switching speed considered in the previous cited work, the data no longer needs to be aggregated for the OCS to have a reasonable duty cycle. Accordingly, a hybrid network can be designed to switch data at the Top-of-Rack (ToR) level within a datacenter instead of switching large aggregates of hosts at a higher level within the network. Here, we experimentally evaluate the network-level switching time of a functional 23-host prototype hybrid network for datacenters called Mordia (Microsecond Optical Research Datacenter Interconnect Architecture). This hybrid network uses an OCS architecture based on a custom interface to a commercial WSS and has a measured mean port-to-port network reconfiguration time of 11.5 μs. This time includes the physical switching time of the WSS and the acquisition time of the NIC.

II. System Design

The system-level block diagram of the Mordia hybrid network is shown in Fig. 1. Each computer host has a dual-port 10G Ethernet Network Interface Card (NIC) [11] with two SFP+ connections. One of these ports is connected to a

Manuscript received March 26, 2013; revised May 9, 2013; accepted June 4, 2013. Date of publication June 20, 2013; date of current version July 29, 2013. This work was supported in part by the NSF Center for Integrated Access Networks under Grant 0812072 and an NSF MRI under Grant 0923523. The construction of the prototype was funded in part by gifts from Google, Cisco Systems, and Corning.

N. Farrington was with the Department of Computer Science and Engineering, University of California, San Diego, CA 92093 USA. He is now with Facebook, San Mateo, CA 76584 USA (e-mail: nathan@nathanfarrington.com).

A. Forencich, J. E. Ford, Y. Fainman, and G. C. Papen are with the ECE Department, University of California, San Diego, CA 92093 USA (e-mail: alex@alexelectronics.com; jeford@ucsd.edu; fainman@eng.ucsd.edu; gpapen@ucsd.edu).

G. Porter is with the CSE Department, University of California, San Diego, CA 92093 USA (e-mail: gmporter@cs.ucsd.edu).

P.-C. Sun was with the Electrical and Computer Engineering Department, University of California, San Diego, CA 92093 USA. He is now with Cisco, London NW4 4BT, U.K. (e-mail: pcsun@eng.ucsd.edu).

A. Vahdat is with the CSE Department, University of California, San Diego, CA 92093 USA. Currently, he is on leave at Google, Moountain View, CA 94093 USA (e-mail: vahdat@cs.ucsd.edu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LPT.2013.2270462

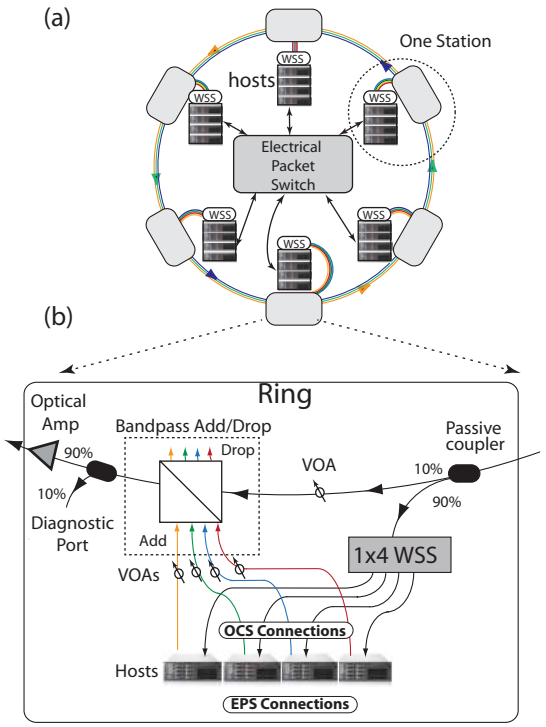

Fig. 1. (a) System-level block diagram of the Mordia network. (b) Components inside one station of the ring.

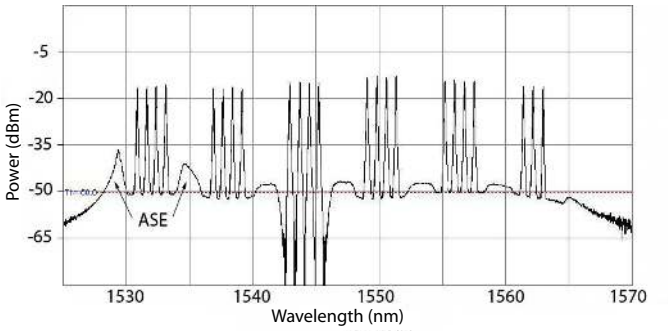

Fig. 2. Spectrum of 23 channels in the ring. Some amplified spontaneous emission (ASE) can been seen near the peak of the gain profile of the fiber amplifier.

standard 10G Ethernet EPS. The second port is connected to the prototype OCS that is based on a WSS. This switch is used to route a fixed wavelength channel from one port on one host to another port on one or more other hosts. The two networks are run in parallel producing a hybrid network.

The physical topology of Mordia is a unidirectional ring of N wavelengths in a single fiber and is shown in Fig. 1a. The logical topology of Mordia is a mesh and supports circuit unicast, circuit multicast, circuit broadcast, and loopback.

Each host is assigned its own fixed transmission wavelength using commercially available DWDM SFP+ modules (EOLP-1696-14XXN) on the ITU 100 GHz grid. The hardware is designed to support 24 hosts. Our initial experiments used either 22 or 23 hosts depending on the experiment because of difficulties in obtaining one of the fixed wavelength SFP+ modules.

Wavelengths are added or dropped from the ring at six stations as shown in Fig. 1b. Each station can support up to four hosts. The wavelengths for each host within a single station are spaced 100 GHz apart. The group of fixed transmission wavelengths assigned to the next station is offset by 400 GHz. On a 100 GHz channel grid, the channels are then "four-on", "four-off". The four wavelengths are injected into the ring using a bandpass add/drop filter. The optical power from each host is adjusted before injection using a variable optical attenuator (VOA). These components are shown in Fig. 1b.

The optical switching at each station is done outside the ring using a wavelength-selective switch (WSS) module, which is a customized 1×4-port Nistica Full Fledge 100 WSS module [12]. The customization consists of a high-speed interface and is discussed in the next section. At the input to each station, a passive, wavelength insensitive power splitter directs 90% of the power out of the ring and into the WSS. The remaining 10% of the signal stays in the ring. Because the splitter is passive, the input to the WSS module at each of the six stations contains all 23 wavelengths.

Each WSS module can route a controllable fraction of any of the individual wavelength channels to each of the four output ports. The resulting architecture is a "broadcast and select" network in which each fixed wavelength signal from each transmitter port can be routed to any receiver port. This kind of architecture requires the receiver to be wavelength insensitive over the range of transmission wavelengths.

The four wavelengths injected at each station travel once through the ring and are prevented from making more than one pass by the bandpass add/drop filter. This filter also injects the signal at that station into the ring. At each station, the signal in the ring is amplified using a 23 dBm booster optical amplifier (Optilab EDFA-B-23-M). All of the amplification is done inside the ring and all of the switching is done outside the ring. This design prevents transient power fluctuations in the optical amplifiers during a circuit reconfiguration.

The input level to each optical amplifier was set to minimize the optical noise. This power level was set using a combination of the 6 VOAs within the ring and the 23 VOAs outside the ring—one for each wavelength channel. A spectrum of the OCS at one station is shown in Fig. 2. By properly balancing the power incident to each amplifier within the ring, the OSNR for each channel is on the order of 30 dB, which can be seen in Figure 2.

The hardware was rack-mounted and was placed in a standard server room as is shown in Fig. 3.

III. Routing and the Control Plane

The Mordia network uses Time-Division Multiple Access (TDMA) [10]. The routing for the initial implementation used a round-robin schedule. For this schedule, the OCS capacity is divided equally among all hosts. The initial experiments used a data transmission window of 94.5 μs for each time slot in the schedule. The workload generated at each host emulated an all-to-all connection pattern.

The control plane that executes the schedule consists of a Field Programmable Gate Array (FPGA) and a 10G electrical

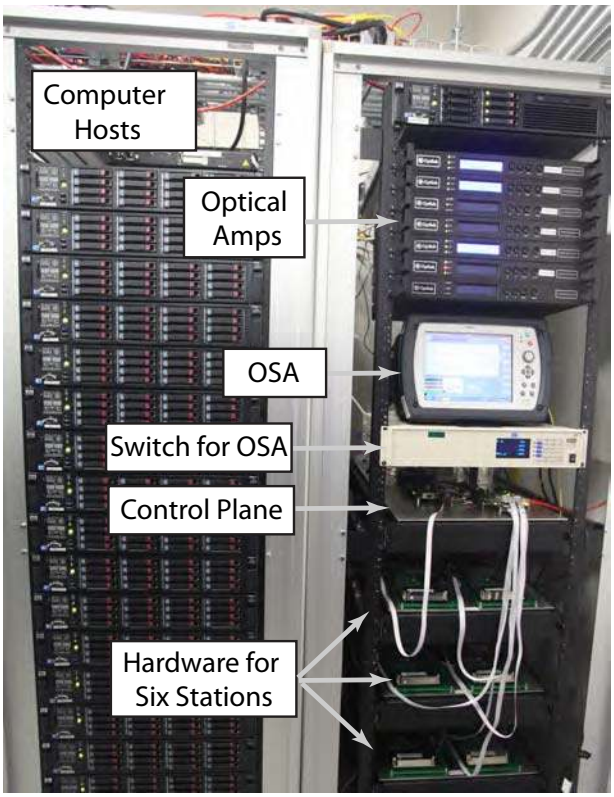

Fig. 3. Hardware for the Mordia prototype.

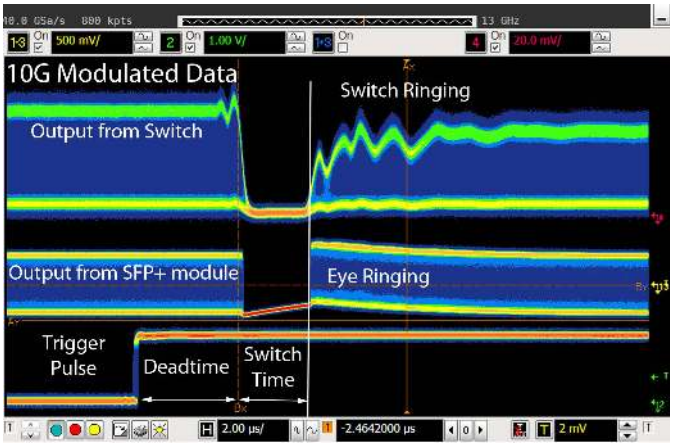

Fig. 4. Measured switch time.

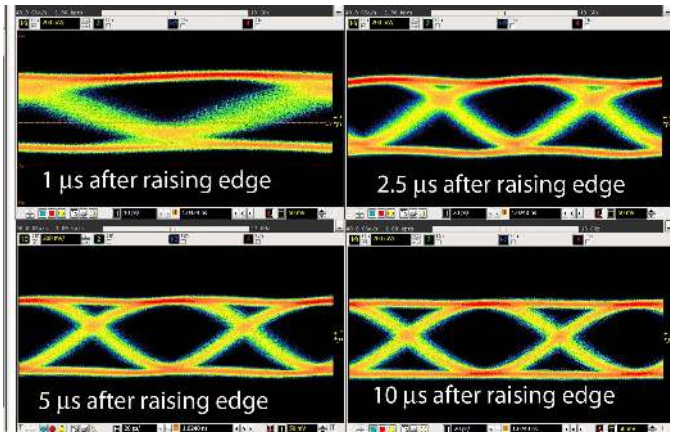

Fig. 5. Time evolution of the eye diagram after switching.

packet switch. The standard interface for controlling the WSS module is I^2C, which was too slow for our purpose. Therefore, a custom interface based on a SPI protocol was developed to enable high-speed switching of the WSS module using a digital trigger signal sent from the FPGA to each WSS module.

The FPGA synchronizes the hosts with the WSS by transmitting a broadcast synchronization packet over the electrical packet-switched network. Using the electrical packet-switched network for synchronization can introduce jitter because the arrival time of the synchronization packet to each end host depends on how the packet-switched network is loaded. For the characterization experiments of the OCS, the packet-based network was used only for control and no data was transmitted over it.

IV. Switching Speed

The overall switching speed can be decomposed into the physical switching speed of the WSS and the system-level switching speed as measured at the NIC for each end host.

The physical switching speed of the WSS module is shown Fig. 4. The lower trace is the trigger signal. The middle trace is the electrical 10 Gb/s data signal seen after the transceiver. The top trace is the 10 Gb/s optical data signal after the WSS module and before the transceiver. Starting from the lower trace, the switch begins to reconfigure after a delay of 3 μs relative to the rising edge of the trigger signal. The measured reconfiguration time is 2.25 μs following by ringing that lasts about 6-7 μs. To determine the effect of the ringing on the data, Fig. 5 shows the formation of the 10 Gb/s eye-diagram in the optical domain after the rising edge of output signal from the WSS module. Based on this data, we estimate that the physical-layer chip within the NIC card can lock between 5–10 μs after the raising edge of the optical signal.

The physical-level switching speed of the WSS module is not the speed at which packets are switched because of additional delays in the NIC caused by the aquisition time of the phase-lock loop (PLL). To measure the system-level switching speed for a packet, we used one host that continuously transmited blocks of minimum-sized Ethernet frames with each frame being 67.2ns long and each frame having a sequential frame number. Three other hosts captured all traces ignoring the synchronization packets. A total of 1,000,000 transmitted frames were collected and merged. During a switching event several frames are lost within each block. The duration of this event can be determined by measuring the gap in the sequence numbers of the frames that are correctly received with a temporal resolution that is equal to the duration of the frame. Multiplying the number of lost frames in each block by the duration of the frame yields the system-level switching time. Fig. 6 shows the resulting histogram using a total of 705 blocks of Ethernet frames. The distribution has a mean of 11.55 μs and a standard deviation of 2.36 μs. The variation in the switching time is a combination of the jitter in the arrival time of the packets and the ringing in the WSS shown in Fig.4, which can cause the PLL to lock at different times.

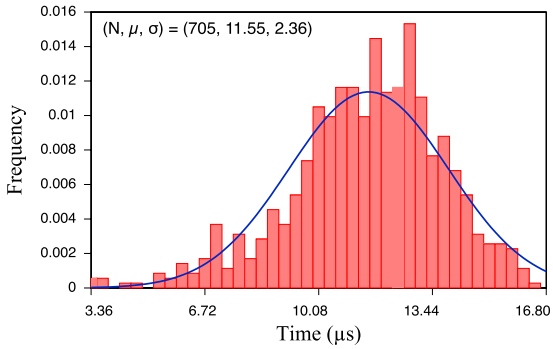

Fig. 6. Network switch time.

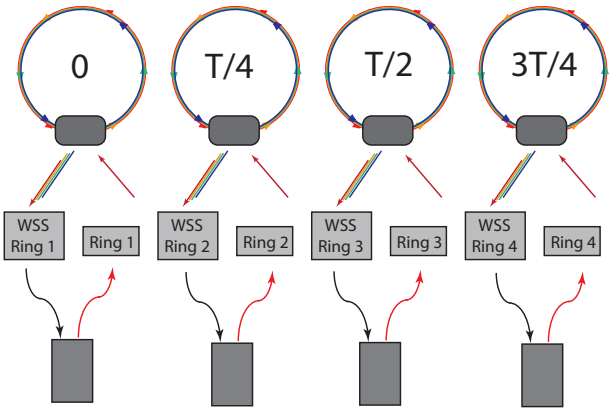

Fig. 7. Scaling the Mordia prototype using one ring for each port of a four-port host. The time when each ring is switched is offset in units of $T/4$ where T is the total reconfiguration time of a WSS.

Based on the characterization of the WSS module, we estimate the WSS switch time including the physical-layer chip to be about 9 μs with the delay in the other components in the system being about 2.5 μs. Therefore, the physical switch time of the OCS is comparable to the other delays in the hybrid network. Given that our data transmission window is 94.5 μs, the measured network reconfiguration time of 11.5 μs yields a duty cycle of 89.15%.

V. Scaling the Data Plane

The Nistica switch supports 80 channels on a 50 GHz grid. This basic configuration can be scaled to a larger capacity by connecting a single host to multiple rings using multiple ports per host. This is shown in Figure 7.

Consider a ring network that consists of M hosts, each of which have N output ports. Each of the M hosts is assigned a fixed wavelength, and each of the N output ports for one host is injected into a separate ring as shown in Figure 7. By adding additional rings, it is possible to support a higher port count per host. By adding additional wavelengths, it is possible to support more hosts in each ring.

The switch reconfiguration of each ring in a multi-ring design is staggered over time, so that only one ring is reconfigured at any instant in time. The switching of each ring in the network is offset by T/N where T is the switch reconfiguration time. The switching times for a four-ring/four-port configuration are shown in Figure 7. The offset switching time in each ring leads to lower latency because there is now an available circuit-switched network connection every T/N seconds.

VI. Conclusion

The 11.5 μs system-level switching time of the optical circuit switch in the Mordia network enables research into the development of hybrid switches at the Top-of-Rack (ToR) level within a datacenter network for which the hybrid switch is directly connected to an individual port on a host instead of being connected to large aggregate of data from many hosts.

The speed of such an optical switch leads to new research challenges and opportunities. At this switching speed, recent work [13] has shown that it is possible to schedule all of the traffic using only the optical circuit switch. When combined with a standard electrical packet switch, this provides significant design freedom in partitioning the traffic between the EPS and OCS. A key outstanding research challenge is the control of this kind of microsecond hybrid network so that the optical circuit switch is effectively used so that it can route a large fraction of the total traffic. This is an active research topic [14] and requires further study before such a hybrid architecture is viable at the Top-of-Rack level within a datacenter.

Acknowledgment

The authors would like to thank Nistica for providing the custom interface to the WSS and for technical support.

References

[1] K. Barker, *et al.*, "On the feasibility of optical circuit switching for high performance computing systems," in *Proc. ACM/IEEE Conf. Supercomput. 2005*, Nov., pp. 1–16.

[2] S. Kamil, D. Gunter, M. Lijewski, L. Oliker, and J. Shalf, "Reconfigurable hybrid interconnection for static and dynamic scientific applications," in *Proc. 4th Int. Conf. Comput. Frontiers, 2007*, pp. 183–194.

[3] L. Schares, *et al.*, "A reconfigurable interconnect fabric with optical circuit switch and software optimizer for stream computing systems," in *Proc. OFC*, Mar. 2009, pp. 1–3.

[4] M. Glick, D. G. Andersen, M. Kaminsky, and L. Mummert, "Dynamically reconfigurable optical links for high-bandwidth data center networks," in *Proc. OFC*, Mar. 2009, pp. 1–3.

[5] G. Wang, *et al.*, "Your data center is a router: The case for reconfigurable optical circuit switched paths," in *Proc. HotNets-8*, Oct. 2009.

[6] G. Wang, *et al.*, "c-Through: Part-time optics in data centers," in *Proc. ACM SIGCOMM 2010*, pp. 327–338.

[7] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: A topology malleable data center network," in *Proc. 9th ACM SIGCOMM Workshop Hotnets-9*, 2010, p. 8.

[8] N. Farrington, *et al.*, "Helios: A hybrid electrical/optical switch architecture for modular data centers," in *Proc. ACM SIGCOMM 2010*, pp. 339–350.

[9] N. Farrington, Y. Fainman, H. Liu, G. Papen, A. Vahdat, "Hardware requirements for optical circuit switched data center networks," in *Proc. OFC*, Mar. 2011, pp. 1–3.

[10] B. C. Vattikonda, G. Porter, A. Vahdat, and A. C. Snoeren, "Practical TDMA for datacenter ethernet," in *Proc. EuroSys 2012*, pp. 225–238.

[11] (2013). *Myricom Two-Port 10-Gigabit Ethernet Network Adapters* [Online]. Available: https://www.myricom.com/products/network-adapters/10g-pcie2-8b2-2s.html

[12] T. A. Strasser and J. L. Wagener, "Wavelength-selective switches for ROADM applications," *IEEE J. Sel. Topics Quantum Electron.*, vol. 16, no. 5, pp. 1150–1157, Sep./Oct. 2010.

[13] N. Farrington, G. Porter, Y. Fainman, G. Papen, and A. Vahdat, "Hunting mice with microsecond circuit switches," in *Proc. 11th ACM Workshop Hotnets-11*, 2012, pp. 115–120.

[14] G. Porter, *et al.* "Integrating microsecond circuit switching into the data center," in *Proc. SIGCOMM*, 2013, pp. 1–12.