A 10 μs Hybrid Optical-Circuit/Electrical-Packet Network for Datacenters

Nathan Farrington, Alex Forencich, Pang-Chen Sun, Shaya Fainman, Joe Ford, Amin Vahdat*, George Porter, and George Papen

*Departments of ECE and CSE, UC San Diego, 9500 Gilman Dr., La Jolla CA 92093; *on sabbatical at Google*

farrington@cs.ucsd.edu

Abstract: We built and evaluated a hybrid electrical-packet/optical-circuit network for datacenters using a 10 μs optical circuit switch using wavelength-selective switches based on binary MEMs. This network has the potential to support large-scale, dynamic datacenter workloads.

OCIS codes: 060.4250, 060.4253,060.6718.

1. Introduction

Modern datacenters are increasingly used as computing platforms for data-intensive applications that require bisection bandwidths which can easily exceed 10 Tb/s. Recently, we built and evaluated a hybrid network for a datacenter based on a combination of electrical-packet switching (EPS) and optical-circuit switching (OCS) [1]. We have also demonstrated practical performance using a TDMA protocol relying on an all-EPS environment that is amenable to circuit switching [2].

Here, we extend that work and experimentally evaluate a functional 24-node prototype hybrid network for datacenters called Mordia (Microsecond Optical Research Datacenter Interconnect Architecture). This hybrid network uses an OCS architecture based on a wavelength-selective switch (WSS) that has a measured mean host-to-host network reconfiguration time of 11.5 μs.

2. System Design

The system-level diagram of the Mordia hybrid network is shown in Fig. 1.

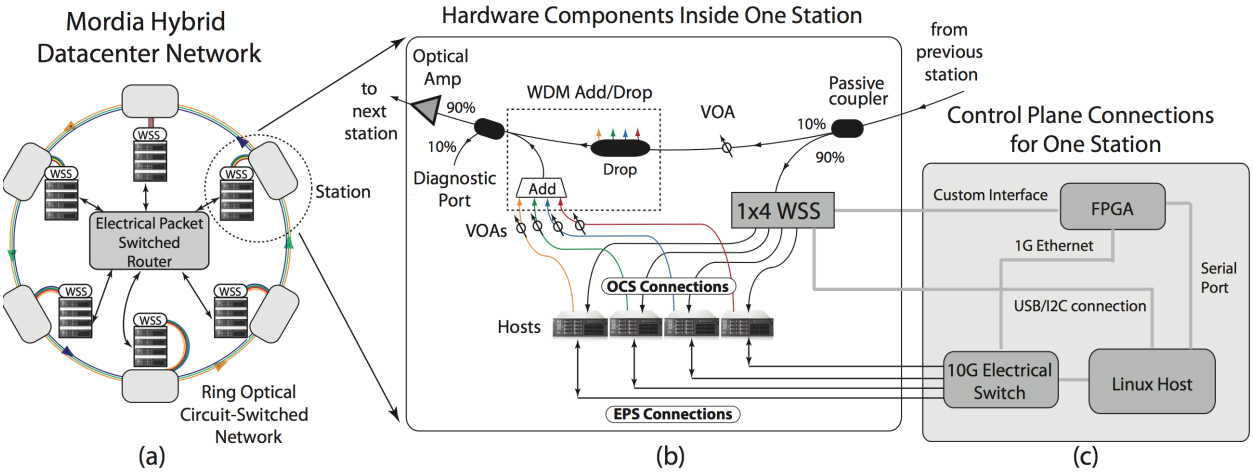

Fig. 1. System-level diagram of the Mordia network.

The initial configuration uses computer hosts with dual-port 10G Ethernet NICs with SFP+ connections. The network can also be used with top-of-rack switches. Each port of each host is connected to both a standard 10G Ethernet

EPS and a research OCS. The two networks are run in parallel producing a hybrid network. The physical architecture of the OCS is shown in Fig. 1a. It is a unidirectional ring of N individual wavelengths carried in a single fiber. Each host is assigned its own wavelength using commercially available DWDM SFP+ modules. Wavelengths are added or dropped from the OCS at six stations. At each station, four wavelengths are added to the ring from each host at that station as shown in in Fig. 1b. Each station has a 1×4-port Nistica Full Fledge 100 WSS based on TI's DLP binary MEMs technology [3]. A custom interface to this switch was developed to enable high-speed switching using a trigger signal. The input to each of the six WSS contains all 24 wavelengths. The WSS selects four of 24 wavelengths and routes one each to the four hosts at that station. Because any host can receive any wavelength, the logical topology is a mesh. Consequently, this OCS architecture supports circuit unicast, circuit multicast, circuit broadcast, and circuit loopback. The system is designed to support 24 hosts. Our initial experiments used either 22 or 23 hosts depending on the experiment.

Data Plane The physical components for one station are shown in Fig. 1b. A spectrum of the OCS in operation is shown in Fig. 2a.

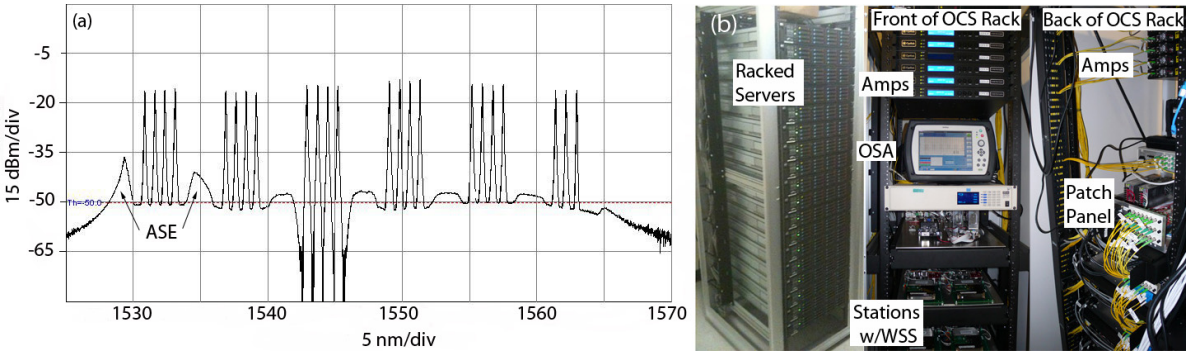

Fig. 2. (a) Spectrum of ring channel, (b) Servers and the racked OCS part of the hybrid network.

The input to each station consists of a passive 90/10 power splitter that directs 90% of the power signal out of the ring into the 1×4 WSS. The signal that remains in the ring is adjusted by a VOA. The four wavelengths that were injected into the ring at that station, which have traveled one round trip in the ring, are dropped by a filter to prevent lasing and interference. This filtering is not perfect and some ASE can be seen in Fig. 2a. The bypass wavelength channels are multiplexed with the channels injected into the ring at that station. The input powers are adjusted by VOAs to balance the total power in the ring. This wavelength-multiplexed signal is then amplified. There is an additional 90/10 power splitter inserted for diagnostics. The hardware is mounted in four rack-mounted sliding trays shown in Fig. 2b.

Control Plane The control plane shown in Fig. 1c consists of a Linux host to run non-real-time processes, a FPGA board to run real-time processes, the six WSS modules, and a 10G Ethernet switch. Mordia uses TDMA for the coordination of the hosts. Time is divided into fixed-length periods for data transmission and a time for circuit reconfiguration. The FPGA synchronizes the hosts and the WSS modules by transmitting a broadcast synchronization packet to all hosts over the EPS. For our initial experiments, we chose to use a simple round-robin policy where the OCS capacity is divided equally among all hosts. The initial experiments used a data transmission window of 94.5 μs. Given the measured reconfiguration time of 11.5 μs (see below), this yields a duty cycle of 89.15%.

3. Evaluation

Both the physical layer and the network layer were experimentally evaluated. The WSS was characterized using two wavelength channels. The right side of Fig. 3 shows the signal from one of the output ports of the Nistica WSS after it was triggered using the custom interface.

After a delay of 3 μs, the measured switch time is 2.25 μs following by ringing that last about 6-7 μs. To determine the effect of the ringing, the right side of Fig. 3 shows the formation of the eye-diagram after the raising edge of output signal. Based on this data, we estimate that the PHY chip on the NIC card can lock between $5-10$ μs after the raising edge.

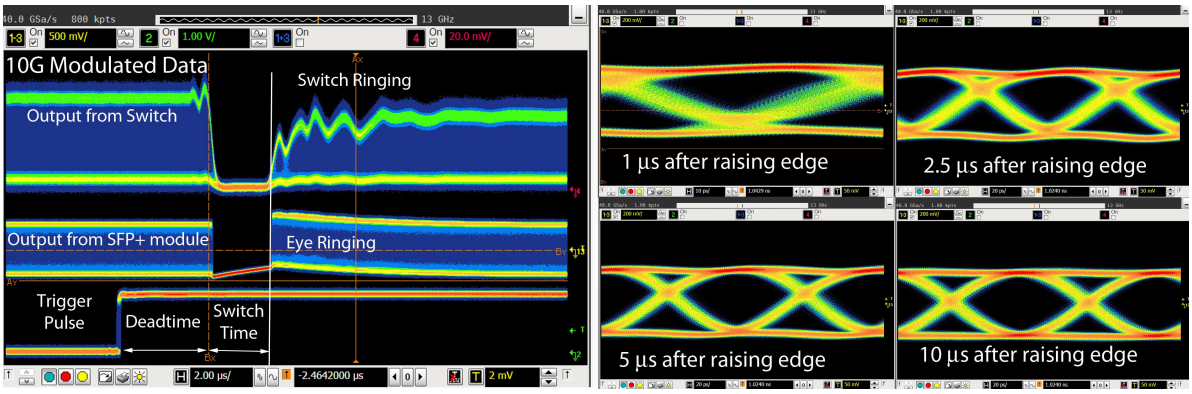

Fig. 3. Left: Measured switch time. Right: Time evolution of the eye diagram after switching.

Two system-level measurements were conducted. In the first experiment, there is one sender that continuously transmits minimum-sized Ethernet frames with each frame being 67.2ns long. Three devices capture all traces ignoring the synchronization packets. A total of 1,000,000 transmitted packets were collected and merged. We identified the network-level switch time from the temporal width of blocks of lost packets. Fig. 4 shows the resulting histogram using a total of 705 blocks. The distribution has a mean of 11.55 μs and a standard deviation of 2.36 μs (Fig. 4 reproduced from [4]). In the second experiment, we evaluate the sustained all-to-all UDP rate 5.7% less than the ideal duty rate.

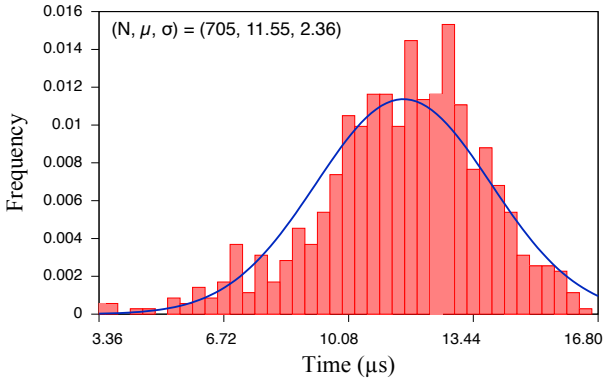

Fig. 4. Network switch time.

Based on a separate characterization of the NIC and the SFP+ module, we estimate the WSS switch time including the PHY chip to be 9.3 μs with the delay in the other components being 2.2 μs. The OCS switch time is thus comparable to the other delays in the hybrid network. This leads to a more balanced hybrid network that has the potential to support large-scale dynamic workloads.

References

1. N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat. Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers. In *ACM SIGCOMM '10*.
2. B. C. Vattikonda, G. Porter, A. Vahdat, and A. C. Snoeren. Practical TDMA for Datacenter Ethernet. *In EuroSys '12*.
3. T. A. Strasser and J. L. Wagener. Wavelength-Selective Switches for ROADM Applications. *IEEE Journal of Selected Topics in Quantum Electronics*, 16:1150–1157, 2010.
4. N. Farrington, R. Strong, A. Forencich, P-C. Sun, T. Rosing, Y. Fainman, J. Ford, G. Papen, G. Porter, A. Vahdat. MORDIA: A Data Center Network Architecture for Microsecond Circuit Switches. *submitted to NSDI'12*.