Optical/Electrical Hybrid Switching for Datacenter Communications

Nathan Farrington***, George Porter*, Alex Forencich*, Joseph Ford*,
Yeshaiahu Fainman*, Amin Vahdat**, and George Papen*
* University of California at San Diego, La Jolla CA, USA
** UCSD, on leave at Google, Mountain View CA, USA
*** Currently at Facebook, Menlo Park, USA

Abstract

We discuss optical/electrical hybrid switching for datacenters. Our current prototype uses an optical circuit switched architecture based on a wavelength-selective switch (WSS) that has a measured mean host-to-host network reconfiguration time of 11.5 µs.

I. Introduction

Hybrid networks for data centers are currently an active research topic. Here, we experimentally evaluate the network-level switching time of prototype hybrid network for datacenters called Mordia (Microsecond Optical Research Datacenter Interconnect Architecture). This hybrid network uses an optical circuit switched (OCS) architecture based on a wavelength-selective switch (WSS) that has a measured mean host-to-host network reconfiguration time of 11.5 µs.

II. System Design

The system-level diagram of the Mordia hybrid network is shown in Fig. 1. Each port of each host is connected to both a standard 10G Ethernet electrical packet switch (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 wavelengths 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, each of the four hosts adds a wavelength to the ring as shown in in Fig. 1b. Each station has a one-by-four port Nistica Full Fledge 100 WSS based on TI's DLP binary MEMs technology [1]. 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.

Each of the six nodes in the ring can support four ports for a total of 24 ports. Each port of the connected device transmits on a fixed wavelength. These four wavelengths are combined in a wavelength multiplexer shown in Figure 1b.

The wavelengths transmitted for each node are spaced 100 GHz apart. The next group of wavelengths transmitted for a different node are offset by 400 GHz or four channels. Each wavelength channel for each node then travels one round trip through the ring.

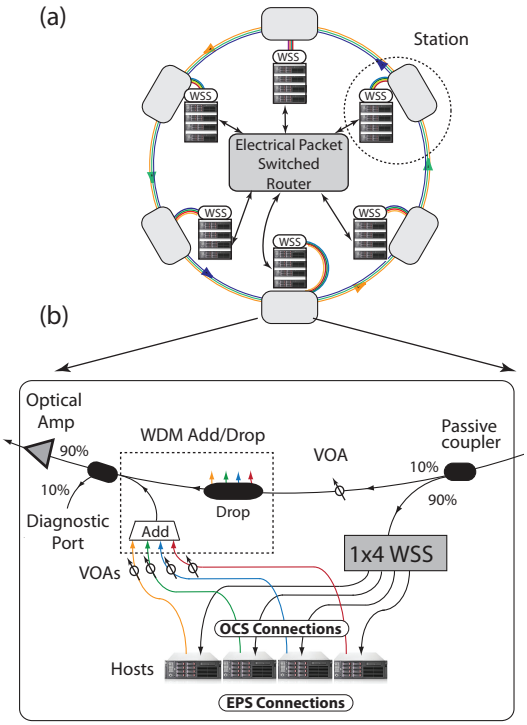

Fig. 1. (a)~System-level diagram of the Mordia network. ~(b)~ Components inside each station of the ring.

The input to each node consists of a passive power splitter. The power splitter directs 90% of the power in the wavelength-multiplexed signal out of the ring. This signal is input into the WSS.

The WSS is configured to route one of the wavelength channels to each of the four output ports. The 10% of the signal that is not directed to the WSS stays in the ring and its power is adjusted by another VOA. The four fixed wavelengths that were injected into the ring at that node, which have traveled a complete round trip passing all the other nodes, are then dropped by a filter preventing the signal from any port from traveling more than one round trip. The remaining ``bypass" wavelength channels that are not dropped by the filter are multiplexed with the

channels injected to the ring at that node as is shown in Fig.1b. This wavelength-multiplexed channel is then amplified. This architecture supports circuit unicast, circuit multicast, circuit broadcast, and circuit loopback. A spectrum of the channels is shown in Fig. 2.

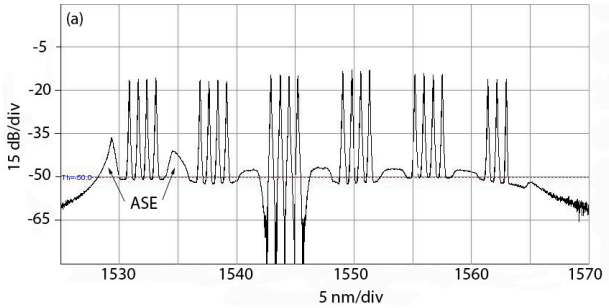

Fig. 2. Spectrum of the Mordia network.

A. Control Plane

The control plane 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.

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 schedule 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 network reconfiguration time of 11.5 μs (see below), this yields a duty cycle of 89.15%.

III. Switching Speed

The WSS was characterized using two wavelength channels. Figure 3 shows the signal from one of the output ports of the WSS after it was triggered.

After a delay of 3 μs, the measured switch time is 2.25 μs following by ringing that lasts about 6-7 μs. We estimate that the PHY chip on the NIC card can lock between 5-10 μs after the raising edge.

The physical-level switching speed is not the speed at which packets are switched because of a variety of factors. To measure this switching speed, we used one host that continuously transmits minimum-sized Ethernet frames with each frame being 67.2ns long. Three other hosts 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.

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.

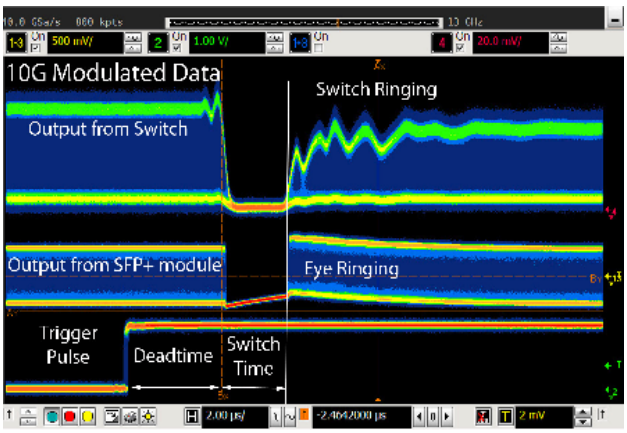

Fig. 3. Measured physical switch time.

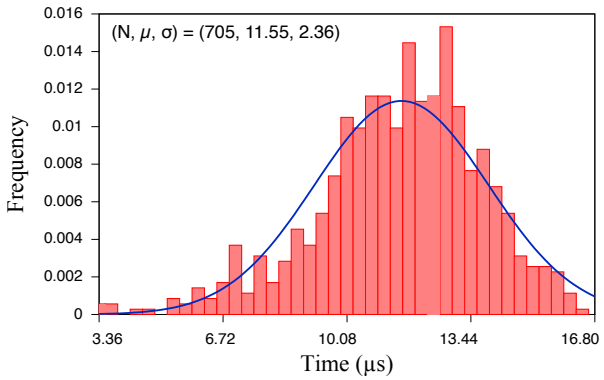

Fig. 4. Network switch time.

IV. Conclusions

In conclusion, we have constructed and demonstrated an optical circuit switch prototype that can switch in less than 10 μs. The questions of if and when this type of OCS technology will be deployed inside data centers depends upon the development of the technology that bridges this fast OCS to a data center along with further refinements in the underlying OCS technology.

Acknowledgments

This work is supported by the NSF Center for Integrated Access Networks (#0812072), gifts from Cisco Systems, Google, Inc, and Corning.

References

[1] T. A. Strasser and J. L. Wagener, "Wavelength-Selective Switches for ROADM Applications," *IEEE Journal of Selected Topics in Quantum Electronics* vol. 16, pp. 1150-1157, 2010.