A Demonstration of Ultra-Low-Latency Data Center Optical Circuit Switching

Nathan Farrington, George Porter, Pang-Chen Sun, Alex Forencich
Joseph Ford, Yeshaiahu Fainman, George Papen, Amin Vahdat

UC San Diego
http://mordia.net

ABSTRACT

We designed and constructed a 24×24-port optical circuit switch (OCS) prototype with a programming time of 68.5 µs, a switching time of 2.8 µs, and a receiver electronics initialization time of 8.7 µs [1]. We demonstrate the operation of this prototype switch in a data center testbed under various workloads.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Circuit-switching Networks

Keywords

data center networks, optical circuit switching

1. INTRODUCTION

Data center networks are crucial to the scalability and performance of data center applications, yet are often under provisioned due to their high CAPEX and OPEX [2]. Recent work combines both traditional electronic packet switches (EPS) as well as optical circuit switches (OCS) [3]–[5] to reduce cost. Unfortunately, the relatively slow OCS switching times of 12 ms make them suitable for a limited class of workloads such as rack-to-rack backup or virtual machine migration. We designed and constructed a 24×24-port optical circuit switch (OCS) prototype, called Mordia, for use in data center networks, with the specific goal of supporting the much more common class of workloads that exhibit all-to-all communication patterns, such as MapReduce and on-line webpage rendering using distributed memory caches.

2. DESIGN & IMPLEMENTATION

The prototype is separated into a data plane (Figure 1) and a control plane (Figure 2). The data plane is constructed as an optical ring of six stations. Each station contains all of the optical components to switch four input/output port pairs. We use fixed-wavelength laser transmitters and tunable wavelength-selective switch (WSS)-based receivers.

The control plane is responsible for programming the WSS modules and synchronizing with all devices connected to the OCS. The controller maintains a round-robin schedule of input-output port mappings. Every 80 µs, the controller programs the six WSS modules with the next input-output port mapping in the list. The round-robin schedule allows the OCS to support all-to-all communication patterns at very high speeds and provides throughput and latency fairness.

Both before and after reconfiguring the WSS modules, the controller broadcasts a synchronization packet to connected devices. This allows the connected devices to learn the current input-output port mapping and when it is safe to begin transmitting.

SIGCOMM'12, August 13–17, 2012, Helsinki, Finland.
ACM 978-1-4503-1419-0/12/06.

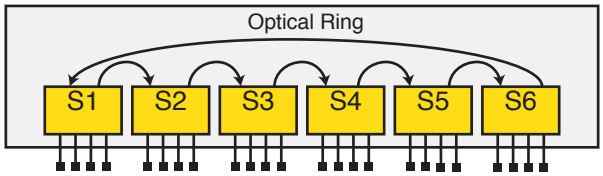

Figure 1. Data plane

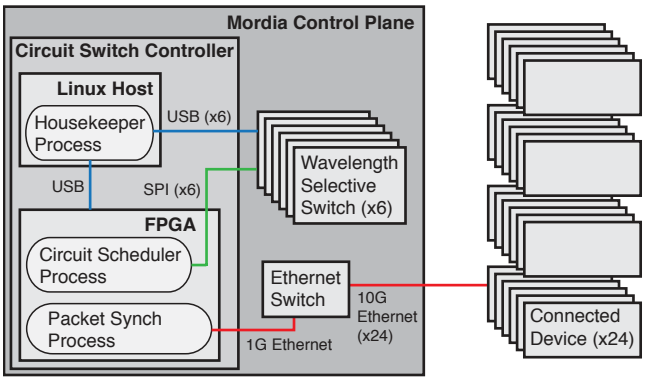

Figure 2. Control plane

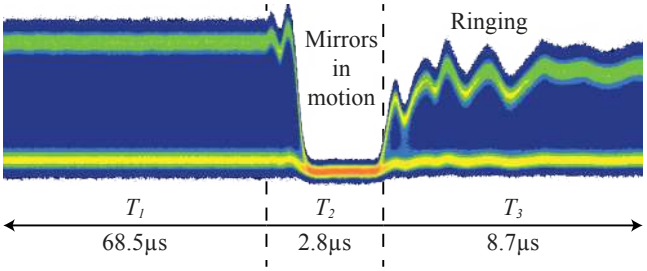

Figure 4. Measurements of switching time

Figure 3 shows that the implementation of the prototype occupies an entire datacenter rack. Each of the bottom three sliding trays contains the components for two stations. The FPGA placement is challenging because it must connect to all six WSS modules with short ribbon cables.

3. MEASUREMENTS

Figure 4 shows physical measurements of the OCS. We use the notation from [6]. The mirrors are only in motion for a small fraction of the total time. The majority of the loss-of-light time is due to ringing (T_3). Table 1 compares the switching speed of the Mordia prototype to Helios [3, 6]. Figure 5 shows measurements of the loss-of-light time ($T_2 + T_3$) as seen from the connected devices. The resulting minimum duty cycle is approximately 85%.

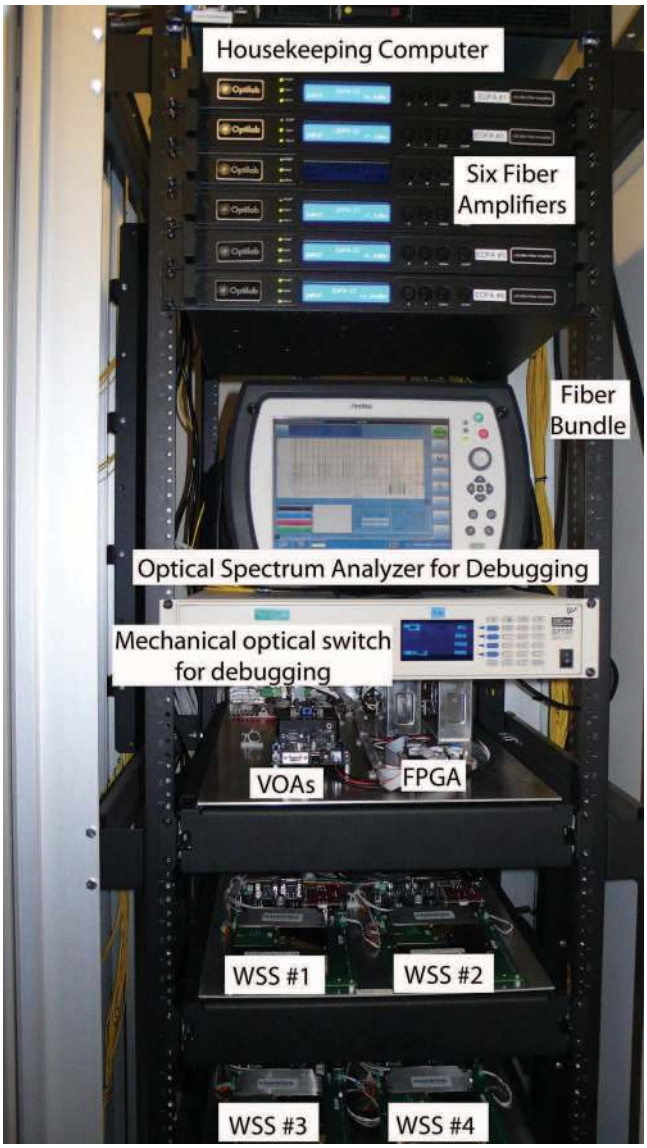

Figure 3. Rack containing Mordia OCS prototype

These measurements show that Mordia is three orders of magnitude faster than Helios and that the performance problems associated with the original Helios prototype have largely been solved in the Mordia prototype.

4. WORKLOADS

The demonstration consists of various workloads on a 24-node compute cluster and a visualization of the communication patterns in real time. Each workload can be run on either a traditional EPS or the Mordia OCS so as to compare the performance directly.

For example, one workload is a synthetic traffic generator called UDP Blaster where each host sends constant bitrate UDP traffic to each other host in the cluster. Good performance for this workload requires that the connected devices synchronize with the Mordia OCS so that they can transmit at the right time.

Another workload is TritonSort, the world's fastest large data sorting system [7]. TritonSort also has an all-to-all communication pattern, but uses TCP instead of UDP, and is rate limited by the disk I/O bandwidth. A successful demonstration of TritonSort on the

Table 1. Performance of Helios and Mordia

	Helios	Mordia	Speedup
T_1	5 ms	68.5 μs	73x
T_2	12 ms	2.8 μs	4,286x
T_3	15 ms	8.7 μs	1,724x

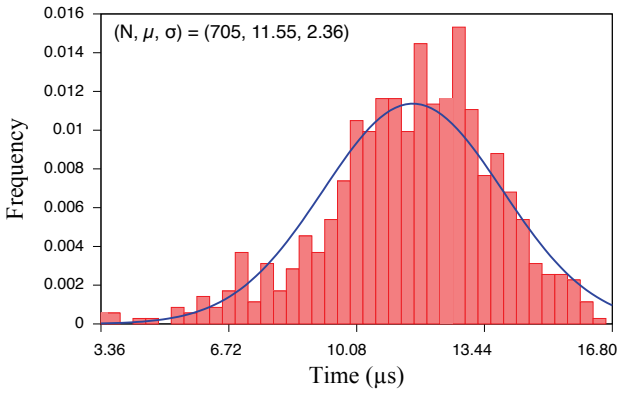

Figure 5. Measurements of loss-of-light time (T_2+T_3)

Mordia OCS prototype will may require careful analysis and possible modification of TCP.

5. ACKNOWLEDGMENTS

We would like to thank the NSF Center for Integrated Access Networks (#0812072) and gifts from Cisco Systems and Google, Inc. We would also like to acknowledge technical assistance from Mod Marathe at Cisco, Haw-Jyh Liaw of NetLogic, and Patrick Geoffray at Myricom, Inc.

REFERENCES

[1] Farrington, N., Porter, G., Sun, P.-C., Forencich, A., Ford, J., Fainman, Y., Papen, G., and Vahdat, A. The Design and Implementation of a Fast, Scalable Data Center Optical Circuit Switch. *Under review.*

[2] Al-Fares, M., Loukissas, A., and Vahdat, A. A Scalable, Commodity Data Center Network Architecture. In *SIGCOMM '08*.

[3] 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. In *SIGCOMM '10*.

[4] Wang, G., Andersen, D.G., Kaminsky, M., Papagiannaki, K., Ng, T.S.E., Kozuch, M., and Ryan, M. c-Through: Part-time Optics in Data Centers. In *SIGCOMM '10*.

[5] Chen, K., Singla, A., Singh, A., Ramachandran, K., Xu, L., Zhang, Y., Wen, X., and Chen, Y. OSA: An Optical Switching Architecture for Data Center Networks with Unprecedented Flexibility. In *NSDI '12*.

[6] Farrington, N., Fainman, Y., Liu, H., Papen, G., and Vahdat, A. Hardware Requirements for Optical Circuit Switched Data Center Networks. In *Optical Fiber Conference (OFC/NFOEC) '11*.

[7] Rasmussen, A., Porter, G., Conley, M., Madhyastha, H., Mysore, R.N., Pucher, A., Vahdat, A. TritonSort: A Balanced Large-Scale Sorting System. In *NSDI '11*.