Hardware Requirements for Optical Circuit Switched Data Center Networks

Nathan Farrington[1], Yeshaiahu Fainman[1], Hong Liu[2], George Papen[1], Amin Vahdat[1,2]
1 UC San Diego, 9500 Gilman Drive, M/C 0404, La Jolla, CA, 92093-0404, USA
2 Google, Inc., Mountain View, CA, USA
Author e-mail address: farrington@cs.ucsd.edu

Abstract: Based on measurements of a prototype, we identify hardware requirements for improving the performance of hybrid electrical-packet-switched/optical-circuit-switched data center networks.
OCIS codes: (200.4650) Optical interconnects; (200.6715) Switching; (060.4253) Networks, circuit-switched

1. Introduction

Several recent research efforts have applied optical circuit switching (OCS) to data center networks [1–3]. The Helios prototype (Fig. 1) promises massive scalability with large reductions in equipment cost, power consumption, and quantities of fiber [1] by strategically combining electrical packet switching and optical circuit switching. This paper identifies hardware requirements for optically circuit-switched data center networks with the goal of achieving nearly the same performance as pure electronic packet switched (EPS) networks. We focus on optical communications inside the data center, complementing earlier work [4] for communications between data centers.

In Fig.1, network traffic flows between pods of hosts (Pod 0-3) either by traversing a core EPS (Core 0) or a core OCS (Core 1-5). In general, stable traffic is routed through an OCS and bursty traffic is routed through an EPS. A real-time circuit scheduler measures the current traffic pattern and reconfigures the OCS so that circuits will be available for stable traffic [1]. Assuming sufficient utilization, optical switching is substantially cheaper than electrical packet switching, with higher capacity and lower energy per port. And since optical switches are bufferless and perform no O-E-O conversions, traffic traversing an OCS will also have lower latency.

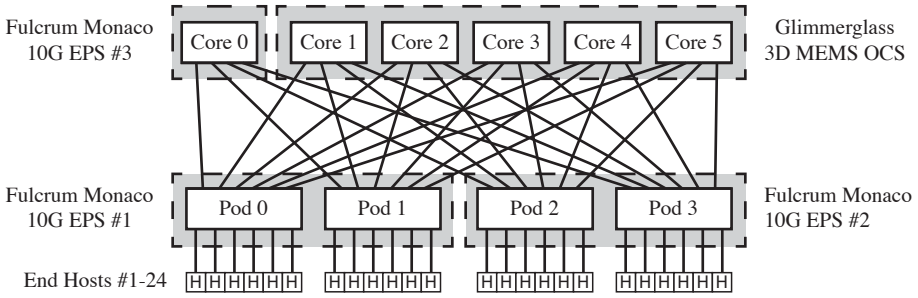

Fig. 1. The Helios prototype hybrid EPS/OCS data center network. We partitioned physical switches into multiple smaller switches to create a more balanced prototype.

2. Performance Measurements of the Helios Prototype Data Center Network

The time series in Fig. 2a shows measured throughput from the Helios hybrid EPS/OCS prototype for an adversarial traffic pattern (PStride [1]) and compares it with a pure EPS network. In this example, the traffic is stable for a period of 4s, and then changes abruptly such that no traffic passes through the existing circuits. The offered load remains constant but the source-destination pod pairs change. The dips in throughput occur precisely when the traffic pattern changes.

The Helios prototype used a 64 × 64 Glimmerglass 3D MEMS OCS. We found that OCS reconfiguration is divided into two consecutive time periods (Fig. 2b): command processing (T_1 = 5ms) and mirror reconfiguration (T_2 = 12ms). In addition, receiver electronics initialization (T_3) takes 15ms and begins after OCS mirror reconfiguration. The dips in Fig. 2a are caused by the $T_1 + T_2 + T_3$ = 32ms period when network traffic stops flowing over the existing circuits and is forced over the single core EPS; throughput recovers when the new circuits are established. If T_1, T_2, and T_3 can be reduced, then the performance of a hybrid EPS/OCS network can approach that of a pure EPS network, even for bursty traffic.

During T_1, the OCS receives a reconfiguration command message over a 1G Ethernet TCP/IP port and processes this message in software on an embedded CPU running Linux. Although the existing circuits remain established during this time, T_1 reduces network throughput by delaying the time between when a change in the traffic pattern is detected and when the mirrors are actually reconfigured. T_1 itself can be reduced by using a faster CPU and by

streamlining the software. By simplifying the protocol to be UDP/IP-based, it should be possible to reduce T_1 to approximately 10µs with an FPGA implementation.

During T_2, an embedded microcontroller moves a subset of the input and output mirrors to establish new circuits. Minimizing T_2 is critical because no network traffic can flow over the affected circuits during this time. T_2 can be reduced further by using smaller mirrors with less mass, a smaller turning range, and fewer turning steps, but this will increase optical loss and reduce maximum port count [5]. For example, in a 100-port 3D MEMS OCS with a 600µm diameter mirror size, Yamamoto, et al. [6] achieved a reconfiguration time of 1.5 ms, or 3ms when the switching is fully stable. Texas Instruments DLP technology can reconfigure in 15µs, but uses mirrors that are approximately 5µm in diameter and can only move between two fixed positions, which could lead to unacceptable optical loss and insufficient port counts if used as a 3D MEMS OCS [7].

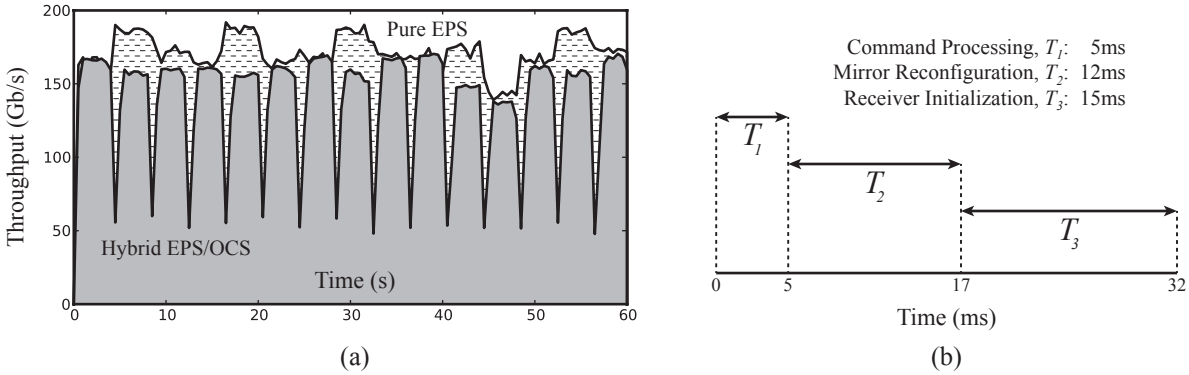

Fig. 2. (a) Performance of Helios prototype EPS/OCS data center network on an adversarial traffic pattern. The pure EPS network has higher instantaneous throughput due to a software artifact in the Helios prototype and can be ignored [1]. The drop in throughput corresponds to all traffic being forced over the single core EPS. (b) Measurements of OCS command processing time, OCS mirror reconfiguration time, and receiver electronics initialization time.

During T_3, the circuits are established, but the receiver electronics are still being initialized after a loss of signal during T_2. Minimizing T_3 is also critical because no network traffic can flow over the affected circuits until initialization is complete. T_3 is actually the maximum initialization time among all circuits in the receive direction of the data path (Fig. 3), specifically the transimpedance amplifier (TIA), the variable gain amplifier (VGA), the feed-forward equalizer and decision-feedback equalizer (FFE/DFE), and the clock/data recovery unit (CDR).

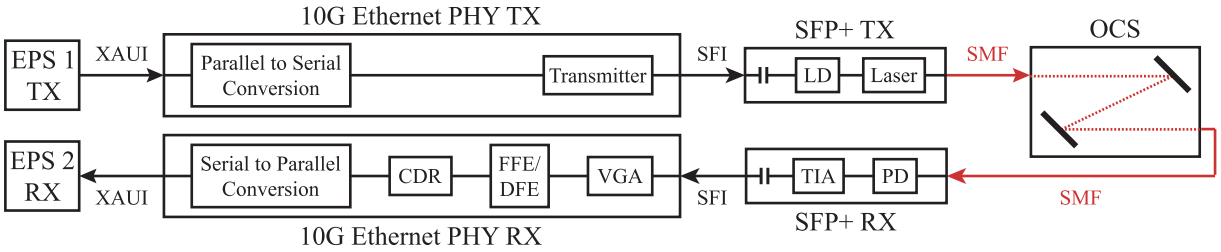

Fig. 3. Data path from the transmit port of EPS pod switch 1, through an OCS, to the receive port of EPS pod switch 2. The electronics on the receive path experience loss of signal during an OCS reconfiguration.

We used a NetLogic AEL2005 PHY in the Helios prototype. The FFE/DFE and VGA have initialization times of 600ms and 15ms, respectively. These two components are responsible for electronic dispersion compensation (EDC), which is not needed for our prototype because we employ optics with a limiting interface and only short runs of single-mode fiber (SMF). We safely disabled the FFE/DFE for the measurements in Fig. 2 but left the VGA enabled. CDR units initialize quickly, with a typical locking time of 50ns. After disabling the VGA and using smaller DC blocking caps, the primary bottleneck would be the continuous TIA, with typical initialization times of 2µs to 80µs [8]. This could be reduced by using a burst-mode TIA such as designed for 10G EPON. A burst-mode TIA has been demonstrated with an initialization time of less than 200ns [9].

3. How small do T_1, T_2, and T_3 need to be for good network performance?

For an adversarial traffic pattern, throughput drops to zero during the entire $T_1 + T_2 + T_3 = 32$ms period. Equation 1 gives the throughput ratio as a function of T_1, T_2, T_3, and S (period of traffic stability). Throughput is zero when $S \leq T_1 + T_2 + T_3$. The additional loss in throughput caused by the performance of the real-time circuit scheduler is outside the scope of this paper. Fig. 4 shows plots of (1).

$$\frac{S - T_1 - T_2 - T_3}{S} \quad (1)$$

From Fig. 4, we can see that the current Helios prototype performs nearly as well as a pure EPS network when *S* is at least 500ms, as the plot with the yellow triangle shows. The plot with the green square shows the performance for the case when the command processing time and mirror reconfiguration time are both reduced to 1ms each, and when EDC is fully disabled in the PHY. The plot with the blue diamond might be close to the maximum achievable performance, with a command processing period and mirror reconfiguration period of 10µs and 100µs, respectively, and when the receive path is optimized for burst-mode operation. In this example, a Helios network would achieve 75% of the throughput of an EPS network if the pod-to-pod traffic demand is stable for 500µs, i.e. if a pod transmits exactly 5 megabytes at a time to another pod. Close to 100% throughput can be achieved when transferring 50 megabytes of data at a time.

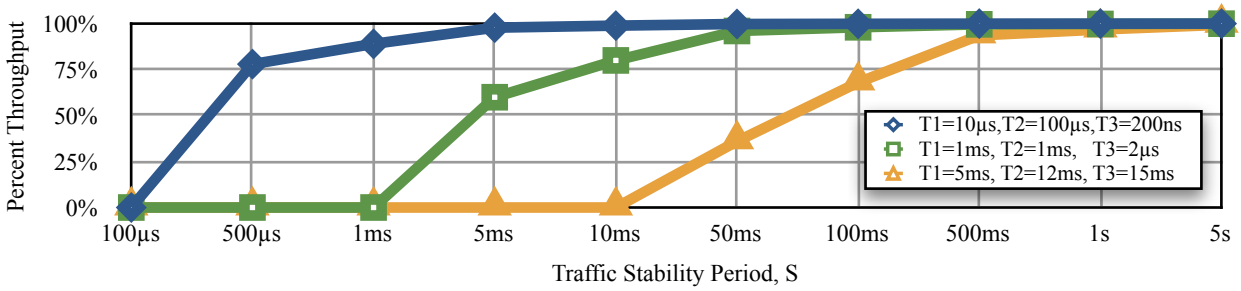

Fig. 4. Throughput compared to a pure EPS as traffic stability period, *S*, changes for different values of T_1, T_2, and T_3.

4. Conclusion

Based on measurements[1] of a hybrid EPS/OCS data center network, this paper identifies the electrical and mechanical bottlenecks limiting performance parity with pure electronic packet switching. Key concerns include the behavior of the 3D MEMS-based optical circuit switch and the initialization time of the electronics on the receiver data path. We analyzed the performance bottlenecks to find realistic lower bounds, and gave a formula for finding the performance speedup as these bottlenecks are removed.

5. Acknowledgements

The authors acknowledge the support of the Multiscale Systems Center, HP's Open Innovation Office, the UCSD Center for Networked Systems, and also support from the National Science Foundation through CIAN NSF ERC under grant #EEC-0812072 and an MRI grant CNS-0923523.

6. References

[1] N. Farrington, G. Porter, S. Radhakrishnan, 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, pp. 339–350.

[2] G. Wang, D. Andersen, M. Kaminsky, K. Papagiannaki, T. S. E. Ng, M. Kozuch, and M. Ryan, "c-Through: part-time optics in data centers," in ACM SIGCOMM '10, pp. 327-338.

[3] L. Schares, X.J. Zhang, R. Wagle, D. Rajan, P. Selo, S.P. Chang, J. Giles, K. Hildrum, D. Kuchta, J. Wolf, and E. Schenfeld, "A reconfigurable interconnect fabric with optical circuit switch and software optimizer for stream computing systems," in OFC/NFOEC'09, paper OTuA1.

[4] C.F. Lam, H. Liu, B. Koley, X. Zhao, V. Kamalov, and V. Gill, "Fiber optic communication technologies: what's needed for datacenter network operations," in IEEE Communications Magazine, (Jul. 2010), pp. 32–39.

[5] P.B. Chu, S.-S. Lee, and S. Park, "MEMS: the path to large optical crossconnects", in IEEE Comm. Mag., (Mar. 2002), pp. 80-87.

[6] T. Yamamoto, J. Yamaguchi, N. Takeuchi, A. Shimizu, E. Higurashi, R. Sawada, and Y. Uenishi, "A three-dimensional MEMS optical switching module having 100 input and 100 output ports', in IEEE Photonics Technology Letters, (Oct. 2003), pp. 1360-1362.

[7] L. Yoder, W. Duncan, E. M. Koontz, J. So, T. Bartlett, B. Lee, B. Sawyers, D. A. Powell, and P. Rancuret. "DLP technology: applications in optical networking," in Proc. of SPIE Vol. 4457, (2001), pp. 54-61.

[8] "MAX3798: 1.0625Gbps to 10.32Gbps, integrated, low-power SFP+ limiting amplifier and VCSEL driver", (Maxim Integrated Products, 2008), available at http://datasheets.maxim-ic.com/en/ds/MAX3798.pdf.

[9] Y. Ohtomo, H. Kamitsuna, H. Katsurai, K. Nishimura, M. Nogawa, M. Nakamura, S. Nishihara, T. Kurosaki, T. Ito, and A. Okada, "High-speed circuit technology for 10-Gb/s optical burst-mode transmission", in OFC/NFOEC '10, paper OWX1.

[1] All measurements of the Helios prototype were performed at UC San Diego.