Optical Circuit Granularity Impact in TCP-Dominant Hybrid Data Center Networks

Houman Rastegarfar*, Kamran Keykhosravi*, Krzysztof Szczerba*, Erik Agrell*, Lloyd LaComb†,
and Madeleine Glick†
*Department of Signals and Systems, Chalmers University of Technology, 412 96 Gothenburg, Sweden
Email: {houman.rastegarfar,kamrank,krzysztof.szczerba,agrell}@chalmers.se
†College of Optical Sciences, The University of Arizona,Tucson, AZ 85721, USA
Email: {llacomb,mglick}@optics.arizona.edu

Abstract—**Hybrid networking, based on electronic packet switching and optical circuit switching, has been proposed to resolve the existing switching bottlenecks in data centers in an energy-efficient and cost-effective fashion. We consider the problem of resource provisioning in hybrid data centers in terms of optical circuit switching capacity and granularity. The number of fibers connected to server racks, the number of wavelengths per fiber, and the ratio of capacity provided by the optical circuit-switched portion of the network to that of the electronic packet-switched portion are crucial design parameters to be optimized during the data center planning phase. These parameters in conjunction with the additive-increase, multiplicative-decrease (AIMD) congestion control mechanism of the Transmission Control Protocol (TCP) pose a significant impact on data center network performance. In this paper, we examine the combined impact of optical bandwidth settings and TCP dynamics using event-driven simulations. Our analysis reveals the strong dependence of overall network throughput on channel capacity (i.e., the bit rate per wavelength channel) and points to the advantages of optical bandwidth consolidation employing higher-order modulation formats.**

I. Introduction

To overcome the electronic switching bottlenecks, the research community is examining the viability of optical switching in data centers [1]–[6]. Optical interconnects allow for ultra-high switching capacities, bit-rate transparency, and low power density and are promising candidates to meet the scale and footprint requirements of next-generation data centers. Unlike electronic switches that support contention resolution anywhere in the network, optical switches suffer from the lack of viable all-optical buffers. Besides, high-port count optical switches can suffer from slow reconfiguration speeds. For instance, optical micro-electro-mechanical (MEMS) switches require 10's of milliseconds to establish new connections. Due to these shortcomings, hybrid data center solutions are interesting as they bring together the advantages of both optical and electrical interconnects and enable performance to be enhanced without resorting to expensive, power-hungry, and full-bisection bandwidth electrical interconnects.

In a hybrid data center network, flow scheduling is a crucial issue and optical and electrical fabrics should work synergistically to accommodate traffic with differing performance requirements [7]–[9]. An under-provisioned electrical network provides all-to-all connectivity among computing nodes, enabling the transport of short-lived, delay-sensitive flows (i.e., mice) and control messages across the network. In addition, an optical circuit switching fabric is provisioned to enable point-to-point, high-bandwidth connectivity by accommodating long-lived, bulk data transfers (i.e, elephant flows). From a traffic point of view, data centers comprise a very large number mice and a much smaller number of elephants. While the number of elephants is significantly smaller than the number of mice, the majority of bytes are carried in elephant flows [10], [11]. In a hybrid data center fabric, mice and elephants are assumed to be best serviced by electrical and optical interconnects, respectively. Resource sharing could also provide additional benefits [6].

In this paper, we examine a crucial planning problem for hybrid data centers in terms of optical network bandwidth capacity and granularity. While each server rack in a hybrid data center should connect to both optical and electrical networks, the capacity of each connection poses a significant impact. To quantify this, let's assume the number of (bidirectional) electrical and optical links interfacing a top-of-rack (ToR) switch to be N_e and N_o, respectively. An optical link is assumed to carry W wavelength-division-multiplexed (WDM) signals. We denote the capacity (in bits per second) of ToR electrical links as R_e and the capacity per wavelength in an optical link as R_w. We define *capacity ratio, CR,* as the ratio of ToR optical capacity to ToR electrical capacity. That is,

$$CR = (N_o \times W \times R_w)/(N_e \times R_e). \tag{1}$$

A value of $CR = 0$ corresponds to a purely electrical data center network and $CR = 1$ corresponds to a hybrid data center with equal electrical and optical capacity. We would like to examine how the data center performance scales with CR in the presence of Transmission Control Protocol (TCP) flows. Besides, the granularity of optical bandwidth (i.e., the number of wavelengths per fiber and wavelength capacity) are important design problems. To examine the provisioned optical capacity and granularity impact, we conduct discrete-event simulations that model the additive-increase, multiplicative-decrease (AIMD) feedback congestion control mechanism of TCP [12]. The major contribution of this work is the study of a variety of optical bandwidth settings in the presence of TCP dynamics within a hybrid data center network.

This work has been inspired by recent demonstrations on multilevel modulation formats for short-reach communications, making it feasible to tweak channel capacities as appropriate. Binary modulation (i.e., on-off keying) is commonly used in data center interconnects because of the cost and power constraints, but it offers a low spectral efficiency. The need for short-reach optical interconnects operating at 100 Gbps and above has drawn significant research and development efforts [13]. Data center network throughput can be increased with more spectrally efficient modulation formats, such as M-level pulse amplitude modulation (PAM) which is interesting for optical links with intensity modulation and direct detection (IM/DD). Multilevel PAM has been investigated for both 850 nm and 1550 nm optical interconnects [14]–[17] and holds promise for high-capacity optical channels in next-generation data centers.

The rest of this paper is organized as follows. In Section II, we introduce the hybrid data center architecture and control cycle. In Section III, we detail our analysis framework and examine the impact of capacity ratio on data center network performance. Finally, we summarize and conclude in Section V.

II. Hybrid Data Center Model and Operation

Fig. 1 depicts the architecture of a hybrid data center network. Without loss of generality, we assume that the electrical and optical networks are non-blocking and model each as a single switch. All server racks within the data center are connected to both electrical and optical networks with N_e and N_o links, respectively. According to the example in Fig. 1, $N_e = 2$, $N_o = 1$, and $W = 2$ since two wavelengths are multiplexed onto optical fibers.

We consider the hybrid data center operation to be governed by control cycles [2], [3]. Each control cycle includes four major tasks: 1) measuring current traffic demands, 2) estimating traffic for the newly started cycle, 3) calculating the optimal optical network topology, and 4) reconfiguring the network as required. Fig. 2 depicts the scheduling tasks within a control cycle, including mandatory (steps 1-3) and secondary (steps 1-6) loops. The secondary loop is executed only when the circuit-switched network requires reconfiguration. The control cycle should be long enough to compensate for scheduling and reconfiguration overheads.

A control cycle starts by measuring the number of elephant flows (detected per an appropriate classification algorithm) each rack has destined to other racks. Afterwards, the traffic estimation routine starts to determine the natural max-min fair bandwidth share of flows. TCP's AIMD dynamics try to achieve such an allocation. The algorithm proposed in [19] is used to estimate the traffic demands between rack pairs. Based on the estimation, the scheduler greedily calculates a maximal matching between racks considering the traffic demands and the number of optical ports per rack. The greedy nature arises due to the fact that in each iteration, circuit(s) will be set up between two nodes that have the highest estimated traffic demand. While an optical circuit is being established, it cannot be used for data transfer. If a circuit has to be torn down, the flows using that circuit will be migrated to the electrical portion of the network. Once the reconfiguration is complete, all flows that can use optical circuits will try to exploit them for an enhanced share of bandwidth.

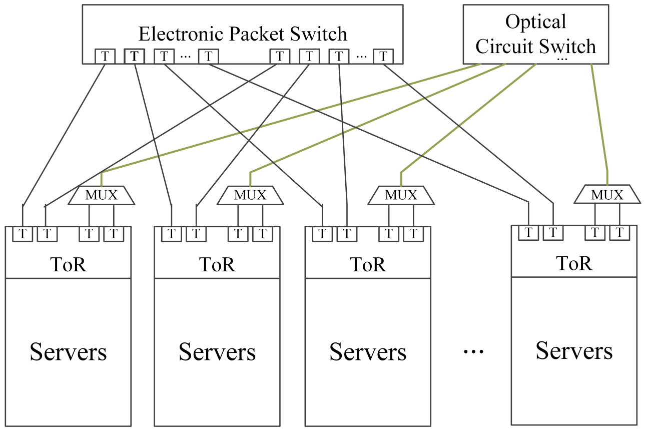

Fig. 1. Hybrid data center network architecture (T: transceiver, ToR: top-of-rack switch, MUX: wavelength multiplexer).

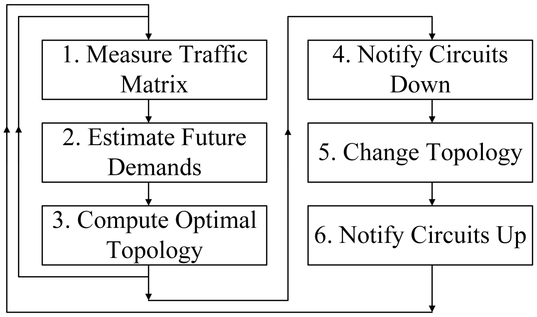
1. Measure Traffic Matrix
2. Estimate Future Demands
3. Compute Optimal Topology
4. Notify Circuits Down
5. Change Topology
6. Notify Circuits Up

Fig. 2. Control cycle components in a hybrid data center network.

Fig. 3 illustrates the data center scheduler pseudocode that we have implemented in our event-driven network simulator. In this figure, N_r denotes the number of server racks. $ControlCycles$ is the number of considered scheduling cycles (as depicted in Fig. 2). IT indicates the current scheduling cycle. CS is the number of time slots per control cycle (including reconfiguration). We consider each time slot to be equal to an average data center network round-trip time (RTT). T points to the current time slot. ***M**easurement, **E**stimation and **M**atching **T**ime, **MEMT***, is the number of time slots required for traffic measurement, estimation, and running a maximal matching algorithm. Finally, RT denotes the time required for reconfiguring the optical network hardware.

Besides circuit scheduling, the data center scheduler has to perform some regular tasks during every time slot. These have been defined in functions $Arrival()$ and $Service()$. The former handles new flow arrivals (and reroutes detected elephants if necessary) and the latter services all flows that exist within the data center. To handle arrivals, the scheduler considers the newly arrived flows during a time slot and assigns them to links. In assigning a flow to links, the scheduler selects the least occupied links in order to perform load balancing and

```
for $IT = 1 : ControlCycles$
  Calculate $ElephantCount(N_r, N_r)$
  $EstimatedTraffic \leftarrow Estimate(ElephantCount)$
  $MatchingMatrix \leftarrow Match(EstimatedTraffic)$
  for $T = 1 : MEMT$
    $time \leftarrow (IT - 1) \times CS + T$
    $Arrival(); Service()$
  if $MatchingMatrix$ implies changes
    $MigrateToElectrical(AffectedCircuits)$
    for $T = MEMT + 1 : MEMT + (RT/RTT)$
      $time \leftarrow (IT - 1) \times CS + T$
      $Arrival(); Service()$
    $FiberAssignment(NewCircuits)$
    $MigrateToOptical()$
    for $T = MEMT + (RT/RTT) + 1 : CS$
      $time \leftarrow (IT - 1) \times CS + T$
      $Arrival(); Service()$
 else
    for $T = MEMT + 1 : CS$
      $time \leftarrow (IT - 1) \times CS + T$
      $Arrival(); Service()$
```

Fig. 3. Pseudocode of the hybrid data center simulator.

reduce the risk of congestion.

The operation of the $Service()$ function in each time slot involves detecting all links that reach the congestion point, performing TCP congestion control, updating the send window per flow, and sending the desired segments through the network according to flows' send window size. When considering congestion for optical links, each channel is treated separately as a flow is assigned to a single wavelength. While several TCP variants can be considered [18], we examine TCP Tahoe and model congestion control in line with the following rules.

1) A slow start threshold (*ssthresh*) regulates the flow transmission rate in two distinct regimes. Below this threshold the flow congestion (send) window size (or equivalently its transmission rate) is doubled per RTT (slow-start phase). Once the send window size exceeds *ssthresh*, congestion avoidance will be in place and the flow rate is incremented linearly per RTT.
2) When a flow's path is saturated, congestion occurs and the flow congestion window will collapse to one segment worth of bytes and its status will be switched to slow start. *ssthresh* for the new iteration will be updated to one half of the maximum window size that the flow could achieve at the congestion point.

III. Simulation Results

We implemented a flow-level, time-slotted network simulator (in MATLAB) based on Fig. 3 to study the impact of optical bandwidth capacity and granularity [9]. Flow arrivals in our simulations are governed by a Poisson process. Each server generates on the average 20 new flows per second [20]. Due to traffic locality in data centers [10], [11], we assume that 75% of the flows remain within the rack. A flow's source is uniformly picked from the set of existing racks. As per destination, we consider moderate hot-spot communications. We assume that 1/4 of the racks are hot-spot nodes and that 3/4 of the outgoing flows are routed to these nodes. For flow size, we consider a rounded Pareto distribution. The flow size in bytes is calculated as

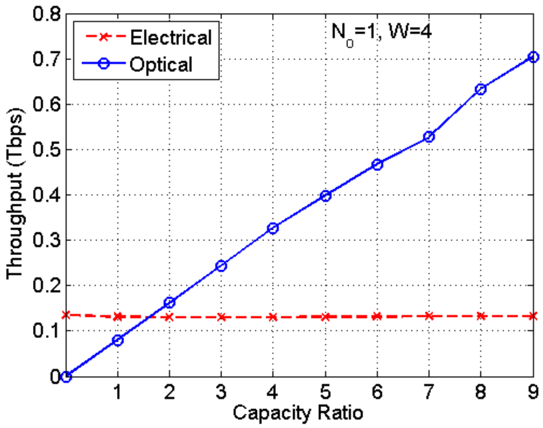

Fig. 4. Electrical and optical network throughput versus capacity ratio for $N_o = 1$ and $W = 4$.

$$L = \left\lfloor \frac{\delta}{U^{1/\alpha}} \right\rfloor \tag{2}$$

where U is a random variable uniformly distributed on (0,1) and $\lfloor \cdot \rfloor$ is the floor function. δ is the scale factor and denotes the minimum flow size and α is the tail index. Based on [10], we consider $\delta = 100$ B and $\alpha = 1/3$ which leads to significant variability in flow size (infinite mean and variance). Once L is determined, we divide it by the TCP maximum segment size (MSS =1500 B) to determine the number of segments a flow contains. Our simulator considers MSS as the data unit.

We simulate a data center with 32 racks of 40 servers each for 60 control cycles. We report the results collected during the latter 40 cycles (averaged over five simulation runs for each data point). Each control cycle comprises 10,000 RTTs where RTT=100 μs [19]. 10 percent of the control cycle duration is associated with circuit scheduling and reconfiguration overheads. Hardware reconfiguration time, RT, is 25 ms. Flows larger than 100 MB are considered as elephants with a detection rate of 0.9. Finally, *ssthresh* is set to 64 KB.

Each rack in Fig. 1 is connected to the electronic packet switching network via a 10 Gbps link ($N_e = 1$ and $R_e =$ 10 Gbps). However, we assume the capacity between a rack and the optical circuit-switched network to be variable and governed by capacity ratio as in (1). Please note that the values we pick for R_w do not necessarily reflect a scaled version of a practical channel capacity. The values are solely enforced to study network performance comparing the relative capacities of electrical and optical networks.

Fig. 4 depicts the throughput for the electrical and optical portions of the data center network considering $N_o = 1$ and $W = 4$. While the electrical network throughput remains relatively unchanged (transferring mice and elephants that

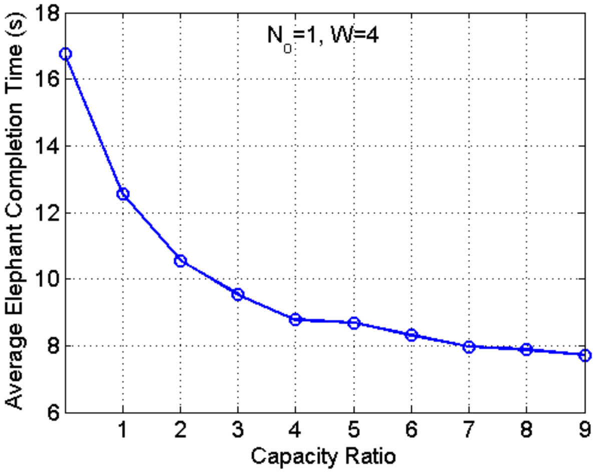

Fig. 5. Average elephant flow completion time versus capacity ratio for $N_o = 1$ and $W = 4$.

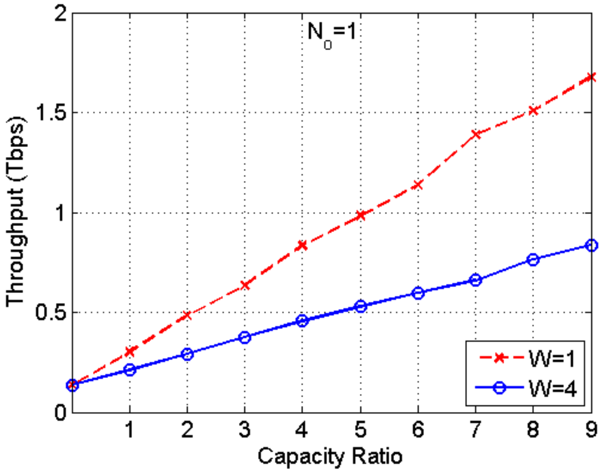

Fig. 6. Impact of channel bandwidth on network throughput for $N_o = 1$.

cannot be assigned to optical circuits), the optical network throughput grows linearly with capacity ratio. Compared to a purely electrical network, a capacity ratio of 9 (corresponding to 90% of the network resources being optical) leads to a 5-fold increase in overall network throughput.

Due to the significant size of elephant flows in our traffic files, in Fig. 5 we present the average elephant flow completion time (averaged over the number of elephant flows whose transmission is completed by the end of the simulation) versus capacity ratio. A maximum of 53.9% improvement is observed. However, the majority of improvement happens with low to moderate capacity ratios. For example, $CR = 1$ (corresponding to equal capacity in electrical and optical networks) translates to 24.9% decrease in completion time compared to a purely electrical network.

A key observation in our analysis is the significance of spectrally efficient modulation schemes in hybrid data centers. Fig. 6 depicts overall network throughput versus capacity ratio, comparing $W = 1$ with $W = 4$. $W = 1$ implies channels with four times higher capacity compared to $W = 4$. At $CR = 1$, the use of one wavelength per fiber translates to a 42.4% increase in total network throughput compared to when four wavelengths of the same aggregate capacity are used. The consequence of wavelength bandwidth consolidation at $CR = 9$ is a 100.5% increase in network throughput. Observe that we are not suggesting the use of one wavelength per fiber, but instead we call for employing higher capacity channels.

We also consider the impact of channel capacity consolidation on job completion time. We define the elephant completion ratio as the average number of (steady-state) elephants that are completed by the end of the simulation run divided by the total number of elephants that arrive during the steady-state period. Fig. 7 depicts this parameter versus capacity ratio for $W = 1$ and $W = 4$. The differences are not as significant as the case of throughput since elephants are quite large and the majority of them require more time than the simulation time span to be serviced. With regard to the traffic and simulation time settings, for $CR = 9$, the choice of $W = 1$ results in 6.1 % higher completion ratio compared with $W = 4$.

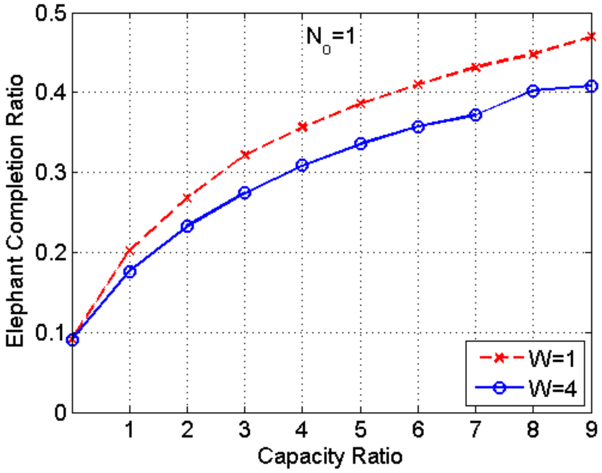

Fig. 7. Ratio of completed elephant flows versus capacity ratio for $W = 1, 4$ and $N_o = 1$.

The significant improvement in network throughput due to using wavelengths of higher capacity (i.e., fatter optical pipes) can be attributed to smaller congestion rates and statistical multiplexing. A high-capacity channel, carrying a mix of uncorrelated traffic flows, can result in less frequent congestion points due to the AIMD behavior of TCP and enhance the effective flow transmission rates. This important finding entails that channel bandwidth settings in a hybrid data center need to be carefully tweaked for an optimal performance. Pulse-amplitude modulation schemes are deemed strong candidates for increasing channel spectral efficiency in a cost- and power-efficient fashion.

Spatial parallelism; i.e., the use of multiple optical ports per rack, can be employed to enable point-to-multipoint connectivity in data centers. If a source rack has bulk transfers intended for several destinations racks, spatial parallelism can help to overcome circuit reconfiguration overheads for concurrent data

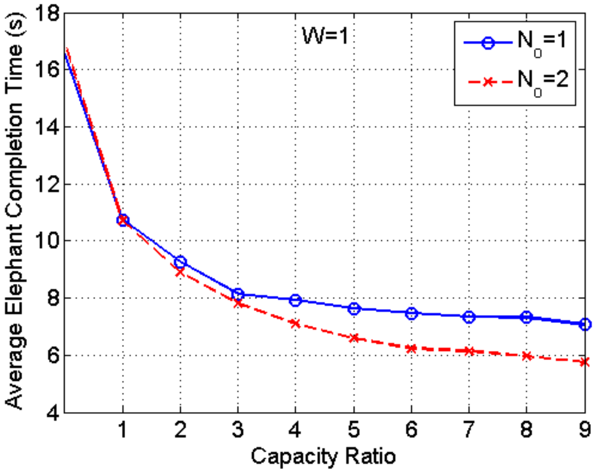

Fig. 8. Impact of spatial parallelism on elephant completion time for $W = 1$.

transmission. However, the number of optical switch ports is limited. For instance, a commercial optical MEMS switch may only have 320 ports [2], [5]. Provisioning multiple fiber ports per rack in a data center with hundreds to thousands of racks calls for a multistage circuit switching architecture that could be challenging from scheduling and physical-layer points of view. Hence, a single optical port per rack (with one or more channels) seems to be a reasonable design choice.

We examined the network performance for $N_o = 1$ and $N_o = 2$ to understand the potential impact of spatial parallelism in hybrid data centers. We could not observe a discernible change in network throughput. Fig. 8 reports average elephant completion time versus capacity ratio for $W = 1$. For moderate and large capacity ratios, the completion times are slightly lower for $N_o = 2$ which can be attributed to more concurrent circuit establishments. Please note that due to the definition of CR in (1), an increase in N_o is compensated by a decrease in channel capacity which according to our analysis is not suitable for performance. The results presented here pertain to a specific traffic model that we have adopted. A different traffic pattern such as one with more non-uniformity may result in a stronger spatial parallelism impact.

IV. Conclusion

Hybrid switching is attractive for the emerging cloud data centers. With hybrid switching, an optical circuit-switched network can complement an oversubscribed electronic packet-switched network to improve application performance with significant savings in data center capital and operational expenditures. However, the absolute capacity of the optical fabric, the number of fibers per rack, and the number of wavelengths per fiber pose a profound impact on the overall network performance. In this paper, we simulated a hybrid data center network, including TCP dynamics, to examine the interplay between optical-to-electrical network capacity (i.e., capacity ratio) and optical circuit switching granularity.

Our analysis revealed a linear dependence of overall network throughput on capacity ratio. For the examined traffic model, elephant completion time was heavily affected by small to moderate capacity ratios. Smaller number of wavelengths per fiber with higher capacity per channel resulted in a significantly larger network throughput, encouraging the use of higher-order modulation formats tailored for short-reach communications. Doubling the number of optical connections per rack in our analysis did not reflect remarkable changes in performance. Although this behavior heavily depends on the mix of traffic flows, the provisioning of multiple fibers per node does not seem to be a viable solution due to limitations on switch port count. Future work should study the physical layer implications of multilevel modulation formats in hybrid data center networks and examine the impact of more traffic patterns such as multicast traffic.

References

[1] C. Kachris, K. Kanonakis, and I. Tomkos, “Optical interconnection networks in data centers: recent trends and future challenges,” *IEEE Commun. Mag.*, vol. 51, no. 9, pp. 39–45, Sep. 2013.

[2] N. Farrington *et al.*, “Helios: a hybrid electrical/optical switch architecture for modular data centers,” in *Proc. ACM SIGCOMM*, Aug./Sep. 2010, pp. 339–350.

[3] G. Wang *et al.*, “c-Through: part-time optics in data centers,” in *Proc. ACM SIGCOMM*, Aug./Sep. 2010, pp. 327–338.

[4] G. Porter *et al.*, “Integrating microsecond circuit switching into the data center,” in *Proc. ACM SIGCOMM*, Aug. 2013, pp. 447–458.

[5] P. Samadi, V. Gupta, J. Xu, H. Wang, G. Zussman, and K. Bergman, “Optical multicast system for data center networks,” *Optics Express*, vol. 23, no. 17, pp. 22 162–22 180, Aug. 2015.

[6] K. Kanonakis, Y. Yin, P. Ji, and T. Wang, “SDN-controlled routing of elephants and mice over a hybrid optical/electrical DCN testbed,” in *Proc. OFC*, Mar. 2015, paper Th4G.7.

[7] H. H. Bazzaz *et al.*, “Switching the optical divide: fundamental challenges for hybrid electrical/optical datacenter networks,” in *Proc. 2nd ACM Symposium on Cloud Computing (SOCC)*, Oct. 2011, paper 30.

[8] H. Liu *et al.*, “Scheduling techniques for hybrid circuit/packet networks,” in *Proc. 11th International Conference on emerging Networking EXperiments and Technologies (CoNEXT’15)*, Dec. 2015, pp. 339–350.

[9] H. Rastegarfar, M. Glick, N. Viljoen, M. Yang, J. Wissinger, L. LaComb, and N. Peyghambarian, “TCP flow classification and bandwidth aggregation in optically interconnected data center networks,” *IEEE J. Opt. Commun. Netw.*, vol. 8, no. 10, pp. 777–786, Oct. 2016.

[10] A. Greenberg *et al.*, “VL2: a scalable and flexible data center network,” in *Proc. ACM SIGCOMM*, vol. 39, no. 4, Aug. 2009, pp. 51–62.

[11] T. Benson, A. Akella, and D. A. Maltz, “Network traffic characteristics of data centers in the wild,” in *Proc. 10th ACM SIGCOMM Conference on Internet Measurement*, Nov. 2010, pp. 267–280.

[12] “TCP Congestion Control,” [Online]. Available: https://tools.ietf.org/html/rfc5681, Jul. 2016.

[13] C. Cole, “Beyond 100G client optics,” *IEEE Commun. Mag.*, vol. 50, no. 2, pp. s58–s66, Feb. 2012.

[14] K. Szczerba, P. Westbergh, M. Karlsson, P. A. Andrekson, and A. Larsson, “60 Gbits error-free 4-PAM operation with 850 nm VCSEL,” *Electron. Lett.*, vol. 49, no. 15, pp. 953–955, Jul. 2013.

[15] ——, “70 Gbps 4-PAM and 56 Gbps 8-PAM using an 850 nm VCSEL,” *J. Lightw. Technol.*, vol. 33, no. 7, pp. 1395–1401, Apr. 2015.

[16] J. Lee *et al.*, “Demonstration of 112-Gbit/s optical transmission using 56 GBaud PAM-4 driver and clock-and-data recovery ICs,” in *Proc. ECOC*, Sep. 2015, paper 0604.

[17] S. Kanazawa *et al.*, “Transmission of 214-Gbit/s 4-PAM signal using an ultra-broadband lumped-electrode EADFB laser module,” in *Proc. OFC*, Mar. 2016, paper Th5B.3.

[18] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan, “Data center TCP (DCTCP),” in *ACM SIGCOMM*, Aug.-Sep. 2010, pp. 63–74.

[19] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, “Hedera: dynamic flow scheduling for data center networks.” in *Proc. ACM/USENIX NSDI*, Apr. 2010, paper 19.

[20] A. R. Curtis, W. Kim, and P. Yalagandula, “Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection,” in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 1629–1637.