OSA: An Optical Switching Architecture for Data Center Networks With Unprecedented Flexibility

Kai Chen, Ankit Singla, Atul Singh, Kishore Ramachandran, Lei Xu, Yueping Zhang, *Member, IEEE*, Xitao Wen, and Yan Chen

***Abstract*—A detailed examination of evolving traffic characteristics, operator requirements, and network technology trends suggests a move away from nonblocking interconnects in data center networks (DCNs). As a result, recent efforts have advocated oversubscribed networks with the capability to adapt to traffic requirements on-demand. In this paper, we present the design, implementation, and evaluation of OSA, a novel Optical Switching Architecture for DCNs. Leveraging runtime reconfigurable optical devices, OSA dynamically changes its topology and link capacities, thereby achieving unprecedented flexibility to adapt to dynamic traffic patterns. Extensive analytical simulations using both real and synthetic traffic patterns demonstrate that OSA can deliver high bisection bandwidth (60%–100% of the nonblocking architecture). Implementation and evaluation of a small-scale functional prototype further demonstrate the feasibility of OSA.**

***Index Terms*—Data center networks (DCNs), optical networking technology, switching architecture.**

I. Introduction

MANY online services, such as those offered by Amazon, Google, Facebook, and eBay, are powered by massive data centers hosting hundreds of thousands of servers. The network interconnect of the data center plays a key role in the performance and scalability of these services. As the number of hosted applications and the amount of traffic grow, the industry is looking for larger server-pools, higher bit-rate network interconnects, and smarter workload placement approaches to satisfy the demand. To meet these goals, a careful examination of traffic characteristics, operator requirements, and network technology trends is critical.

Traffic Characteristics: Several recent data center network (DCN) proposals attempt to provide uniformly high capacity between all servers [1]–[4]. Given that it is not known *a priori* which servers will require high-speed connectivity, for a static, electrical network, this appears to be the only way to prevent localized bottlenecks. However, for many real scenarios, such a network may not be fully utilized at all times. For instance, measurement on a 1500-server Microsoft production DCN reveals that only a few Top-of-Racks (ToRs) are hot, and most of their traffic goes to a few other ToRs [5]. Likewise, an analysis of high-performance computing applications shows that the bulk of interprocessor traffic is degree-bounded and slowly changing [6]. Thus, even for a few thousand servers, uniformly high-capacity networks appear to be an overkill. As the size of the network grows, this weighs on the cost, power, and wiring complexity of such networks.

Dealing With the Oversubscribed Networks: Achieving high performance for data center services is challenging with oversubscribed networks. One approach is to use intelligent workload placement algorithms to allocate network-bound service components to physical hosts with high bandwidth connectivity [7], e.g., placing these components on the same rack. Such workloads exist in practice: dynamic creation and deletion of VM instances in Amazon's EC2 or periodic backup services running between an EC2 (compute) instance and an S3 (storage) bucket. An alternate approach is to flexibly allocate more network bandwidth to service components with heavy communications. If the network could "shape-shift" in such fashion, this could considerably simplify the workload placement problem.

Higher Bit Rates: There is an increasing trend toward deploying 10 GigE NICs at the end-hosts. In fact, Google already has 10 GigE deployments and is pushing the industry for 40/100 GigE [8]–[10]. Deploying servers with 10 GigE naturally requires much higher capacity at the aggregation layers of the network. Unfortunately, traditional copper-wire 10 GigE links are not viable for distances over 10 m [11] due to the high electrical loss at higher data rate, necessitating the need to look for alternative technologies.

The optical networking technology is well suited to meet the above challenges. Optical network elements support on-demand connectivity and capacity where required in the network, thus permitting the construction of thin but flexible interconnects for large server pools. Optical links can support higher bit rates over longer distances using less power than copper cables. Moreover, optical switches run cooler than electrical ones [12], resulting in lower heat dissipation and cheaper cooling cost. The long-term advantage of optics in DCNs has been noted in the industry [12], [13].

Recent efforts in c-Through [14] and Helios [11] provide a promising direction to exploit the optical networking technology (e.g., one-hop high-capacity optical circuits) for building DCNs. Following this trailblazing research, we present OSA, a novel Optical Switching Architecture for DCNs. OSA achieves high flexibility by leveraging and extending the techniques devised by previous works and further combining them

Manuscript received August 15, 2012; revised January 02, 2013; accepted March 05, 2013; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor S. Sengupta.

K. Chen is with the Hong Kong University of Science and Technology, Kowloon, Hong Kong (e-mail: kaichen@cse.ust.hk).

A. Singla is with University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: singla2@illinois.edu).

A. Singh, K. Ramachandran, L. Xu, and Y. Zhang are with NEC Labs, Princeton, NJ 08540 USA (e-mail: atuls@nec-labs.com; kishore@nec-labs.com; leixu@nec-labs.com; yueping@nec-labs.com).

X. Wen and Y. Chen are with Northwestern University, Evanston, IL 60208 USA (e-mail: xwen@northwestern.edu; ychen@northwestern.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNET.2013.2253120

with novel techniques of its own. Similar to the previous works, OSA leverages reconfigurability of optical devices to dynamically set up one-hop optical circuits. Then, OSA employs the novel hop-by-hop stitching of multiple optical links to provide overall connectivity for mice flows and bursty communications and to handle workloads involving high fan-in/out hotspots [15] that the existing one-hop electrical/optical architectures cannot address efficiently via their optical interconnects. Furthermore, OSA dynamically adjusts the capacities on the optical links to satisfy changing traffic demand at a finer granularity. Additionally, to make efficient use of expensive optical ports, OSA introduces the circulator (Section II-B), a bidirectionality-enabling component for simultaneous transmission in both directions over a circuit, which potentially doubles the usage of optical switch ports.

Overall, the highlights of this paper are as follows.

Flexible DCN Architecture: Given a number N of ToR switches and a design-time-fixed parameter k, OSA can assume *any* k-regular topology over the N ToRs. To illustrate how many options this gives us, consider that for just $N = 20$, there are over 12 billion (nonisomorphic) connected 4-regular graphs [16]. In addition, OSA allows the capacity of each edge in this k-regular topology to be varied from a few to a few hundred gigabits per second on demand. Evaluation results in Section V-B.2 suggest up to 150% and 50% performance improvement brought by the flexible topology and flexible link capacity, respectively.

Analysis of OSA-2560: We evaluate a particular instance of container-size OSA architecture, OSA-2560 ($N = 80, k = 4$), with 2560 servers via extensive simulations and analysis. Our evaluation results (Section V-B) suggest that it can deliver high bisection bandwidth that is 60%–100% of the nonblocking network and outperform the hybrid structures by 80%–250% for both real and synthetic traffic patterns. Our analysis (Section III-C) shows that OSA has better performance/power and performance/wiring complexity ratios than either FatTree [1] or a traditional 2:1 oversubscribed network with the same number of servers. Compared to the hybrid structures, OSA achieves better performance with similar cost and slightly less power consumption. We believe that for DCNs that expect skewed traffic demands, OSA provides a compelling tradeoff between the cost, power, complexity, and performance.

OSA Prototype Implementation: We build a small-scale 8-rack OSA prototype with real optical devices and server-emulated ToRs. Through this testbed, we evaluate the performance of OSA with all software and hardware overheads. We find that OSA can quickly adapt the topology and link capacities to the changing traffic patterns, and our results show that it achieves nearly 60% of the nonblocking bandwidth in all-to-all communications. We further examine the impact of OSA on transferring bulk data and mice flows. We also measure the device characteristics of the optical equipment, evaluate the impact of multihop optical-electrical-optical (O-E-O) conversion, and discuss our experience building and evaluating the OSA prototype.

OSA, in its current form, has limitations. Small flows, especially those latency-sensitive ones, may experience nontrivial penalty due to the network reconfiguration latency (∼10 ms). While the fraction of such affected flows is small (Section VII), we propose multiple avenues to solve this problem. The second challenge is how to scale OSA from container-size to larger DCNs consisting of tens to hundreds of thousands of servers. This requires nontrivial efforts in both architecture design and management and is left as part of our ongoing investigation. In this paper, we describe OSA that is designed to interconnect a few thousands of servers in a container.

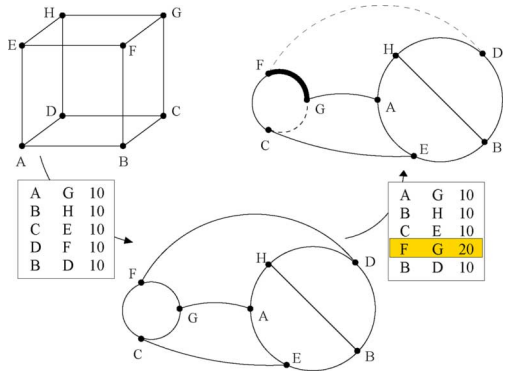


Fig. 1. OSA adapts the topology and link capacities to the changing traffic.

Roadmap: In Section II, we discuss the idea of OSA's unprecedented flexibility, followed by the background on optical technologies for OSA. Then, we describe OSA architecture (Section III) and its algorithm design (Section IV) in response to traffic patterns. In Sections V and VI, we evaluate OSA via extensive simulations and implementation, respectively. We discuss some design issues and related work in Section VII before concluding in Section VIII.

II. Motivation and Background

We first use a motivating example to show what kind of flexibility OSA can deliver. Then, we introduce the optical networking technologies that make OSA possible.

A. Motivating Example

We discuss the utility of a flexible network using a simple hypothetical example in Fig. 1. On the left is a hypercube connecting 8 ToRs using 10G links. The traffic demand is shown in the bottom left of Fig. 1. For this demand, no matter what routing paths are used on this hypercube, at least one link will be congested. One way to tackle this congestion is to reconnect the ToRs using a different topology (Fig. 1, bottom center). In the new topology, all the communicating ToR pairs are directly connected, and their demand can be perfectly satisfied.

Now, suppose the traffic demand changes (Fig. 1, bottom right) with a new (highlighted) entry replacing an old one. If no adjustment is made, at least one link will face congestion. With the shortest path routing, $F \leftrightarrow G$ will be that link. In this scenario, one way to avoid congestion is to increase the capacity of $F \leftrightarrow G$ to 20G at the expense of decreasing the capacities of $F \leftrightarrow D$ and $G \leftrightarrow C$ to 0. Note that in all these cases, the node degree remains the same (i.e., 3), and the node capacity is no more than 30G.

As above, OSA's flexibility lies in its flexible topology and flexible link capacity. In the absence of such flexibility, the above example would require additional links and capacities to handle both traffic patterns. More generally, a large variety of traffic patterns would necessitate the nonblocking network

construction. OSA, with its high flexibility, can avoid such nonblocking construction while still providing equivalent performance for various traffic patterns.

B. Optical Networking Technologies

We now discuss the optical networking technologies that enable the above flexibility.

1) Wavelength Division Multiplexing (WDM): Within the C-band (conventional band) and with 100 GHz DWDM channel spacing, typically 40 or more wavelength channels can be transmitted over a single optical fiber. For the purpose of our architecture, each wavelength is rate-limited by the electrical port to which it connects.

2) Wavelength Selective Switch (WSS): A WSS is typically a $1 \times k$ switch, consisting of one common port and k wavelength ports. It partitions (runtime-configurable within a few milliseconds) the set of wavelengths coming in through the common port among the N wavelength ports. For example, if the common port receives 80 wavelengths, it can route wavelengths 1–20 on port 1, 30–40 on port 2, etc.

3) Optical Switching Matrix (OSM): Most OSM modules are a bipartite $N \times N$ switching matrix where any input port can connect to any of the output ports. Microelectromechanical system (MEMS) is the most popular OSM technology and achieves reconfigurable (at 10 ms [17]) one-to-one circuit by mechanically adjusting micro mirrors. A few hundred ports are common for commercial products, and >1000 for research prototypes [18]. The current commercially available OSM modules are typically oblivious to the wavelengths carried across it. We use MEMS and OSM interchangeably.

4) Optical Circulators: Circulators enable bidirectional optical transmission over a fiber, allowing more efficient use of the ports of optical switches. An optical circulator is a three-port device: One port is a shared fiber or switching port, and the other two ports serve as send and receive ports.

5) Optical Transceivers: Optical transceivers can be of two types: coarse WDM (CWDM) and dense WDM (DWDM). We use DWDM-based transceivers in OSA, which support higher bit rates and more wavelength channels in a single piece of fiber compared to CWDM.

III. OSA Architecture

In this section, we introduce how OSA architecture is built from the above-described optical networking technologies. Our current design is intended for container-size DCNs.

A. Building Blocks

Flexible Topology: OSA achieves the flexible topology by exploiting the reconfigurability of the MEMS. If we start by connecting each of N ToRs to one port on an N-port MEMS, each ToR can only communicate with one other ToR at any instant given the MEMS's bipartite port-matching, leaving the ToR-level graph disconnected. If we connect N/k ToRs to k ports each at the MEMS, each ToR can communicate with k other ToRs simultaneously. Here, $k > 1$ is the degree of the ToR, not its port count, in the ToR graph. The MEMS configuration determines which set of ToRs are connected. OSA must ensure that the ToR graph is connected when configuring the MEMS.

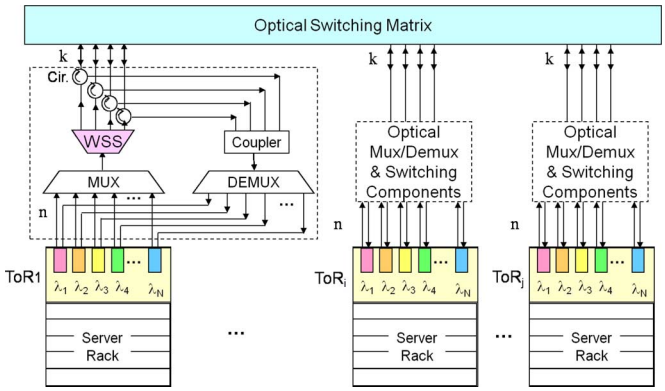


Fig. 2. Overall OSA architecture; detailed structure is shown only for ToR_1 for clarity.

Given a ToR topology connected by the MEMS optical circuits, we use hop-by-hop stitching of such circuits to achieve network-wide connectivity. To reach remote ToRs that are not directly connected, a ToR uses one of its k connections. This first-hop ToR receives the transmission over fiber, converts it to electrical signals, reads the packet header, and routes it toward the destination. At each hop, the packet experiences the conversion from optics to electronics and then back to optics (O-E-O) and the switching at the ToR. Note that at any port, the aggregate transit, incoming, and outgoing traffic cannot exceed the port's capacity in each direction. Therefore, the high-volume connections must use a minimal number of hops. OSA should manage the topology to adhere to this requirement. Evaluation in Section VI quantifies the overhead (both O-E-O and switching) of the hop-by-hop routing.

Flexible Link Capacity: In OSA, Each ToR connects to k other ToRs. If each link has a fixed capacity, multiple links may be required for this ToR to communicate with another ToR at a rate higher than a single link can support. To overcome this problem, OSA combines the capability of optical fibers to carry multiple wavelengths at the same time (WDM) with the dynamic reconfigurability of the WSS. Consequently, each ToR is connected to the MEMS through a multiplexer (MUX) and a WSS unit (Fig. 2).

Specifically, suppose ToR A wants to communicate with ToR B using w times the line speed of a single port. The ToR will use w ports, each associated with a unique wavelength, to serve this request. The WDM enables these w wavelengths, together with the rest from this ToR, to be multiplexed into one optical fiber that feeds the WSS. The WSS splits these w wavelengths to the appropriate MEMS port that has a circuit to ToR B (doing likewise for the rest $k-1$ groups of wavelengths). Thus, a $w\times$ (*line-speed*) capacity circuit is set up from A to B at runtime. By varying the value of w for each MEMS circuit, OSA achieves the dynamic link capacity.

We note that a fiber cannot carry two channels over the same wavelength in the same direction. Moreover, to enable a pair of ToRs to communicate using all available wavelengths, we require that each ToR port (facing the optical interconnect) is associated with a wavelength that is unique across the ports of a ToR. This wavelength–port association is a design time decision, and the same set of wavelengths is recycled across ToRs.

The same wavelength is used to receive traffic as well: Each port thus sends and receives traffic at one fixed wavelength. This allows all wavelengths at a source ToR to be multiplexed and delivered, after demultiplexing, to individual ports at the destination ToRs.

Efficient Port Usage: To make full use of the MEMS ports, we require that each circuit over the MEMS be bidirectional. For this purpose, we use optical circulators between the ToR and the MEMS ports. A circulator connects the send channel of the transceiver from a ToR to the MEMS (after the channel has passed through the MUX and WSS). It simultaneously delivers the incoming traffic toward a ToR from the MEMS (through the coupler and DEMUX) to this ToR. Note that even though the MEMS links are bidirectional, the capacities of the two directions are independent of each other.

B. Putting It All Together: OSA-2560

Fig. 2 illustrates the general OSA architecture. We now discuss one specific instantiation—OSA-2560 with $N = 80$ ToRs, $W = 32$ wavelengths, and ToR degree $k = 4$ using a 320-port MEMS to support 2560 servers.

Each ToR is a commodity electrical switch with 64 10-GigE ports [19]: 32 of these ports are connected to servers, while the remaining are connected to the optical interconnect. Each port facing the optical interconnect has a transceiver associated with a fixed and unique wavelength for sending and receiving data. The transceiver uses separate fibers to connect to the send and receive infrastructures.

The send fiber from the transceiver from each of the 32 ports at a ToR is connected to an optical MUX. The MUX feeds a 1×4 WSS. The WSS splits the set of 32 wavelengths it sees into four groups, each group being transmitted on its own fiber. These fibers are connected to the MEMS via circulators to enable bidirectional communications. The four receive fibers from four circulators are connected to a power coupler (similar to a multiplexer, but simpler), which combines their wavelengths onto one fiber. This fiber feeds a demultiplexer (DEMUX), which assign each incoming wavelength to its associated port on the ToR.

We point out two key properties of the above interconnect. First, each ToR can communicate simultaneously with any four other ToRs. This implies that the MEMS configuration allows us to construct all possible 4-regular graphs among ToRs. Second, through WSS configuration, the capacity of each of these four links can be varied in $\{0, 10, 20, \ldots, 320\}$ Gb/s. The MEMS and WSS configurations are decided by a central OSA manager. The manager estimates the traffic demand, calculates the appropriate configurations, and pushes them to the MEMS, WSS units, and ToRs. This requires direct, out-of-band connections between the OSA manager and those components. Note that our employment of such a central OSA manager is inspired by many recent works [2], [3], [11], [14], [20] in the context of DCNs given the fact that a DCN is usually owned and operated by a single organization.

Furthermore, we choose $k = 4$ for container-size DCNs because it is a tradeoff between the network size and performance. A larger k value can enable one ToR to connect to more other ToRs simultaneously, thus achieving higher performance. However, given a 320-port MEMS, it also means that fewer ToRs $(320/k)$ can be supported. Our experiments with $k = 1, 2, 4, 8$ indicate that $k = 4$ can deliver considerable bisection bandwidth between thousands of servers.

TABLE I
COST (USD) AND POWER (WATT) PER PORT FOR DIFFERENT ELEMENTS. WE REFER SOME OF THE VALUES FROM HELIOS [11]

Element	$	W	Element	$	W
ToR (10G port)	.2K	3	(DE)MUX	3K	0
MEMS	.5K	0.24	Coupler	.1K	0
WSS	1K	1	Circulator	.2K	0
Transceiver (Gray)	.2K	1			
Transceiver (CWDM)	.4K	1	-	-	-
Transceiver (DWDM)	.8K	3.5			

TABLE II
COST, POWER, WIRING (# OF INTER-TOR LINKS) AND PERFORMANCE FOR DIFFERENT NETWORKS TO SUPPORT 2560 SERVERS WITH 10G PORTS. (‡FOR TRAFFIC PATTERNS WE EVALUATE IN SECTION V)

Architecture	Cost ($)	Power (KW)	Wiring	% of non-blocking
Traditional	2.6M	25.6	1280	50%
Hybrid	4.5M	29.0	480	20%–50%‡
OSA	4.4M	27.1	320	60%–100%‡
FatTree	5.1M	51.2	5120	100%

C. Analysis

Table I lists the cost and power usage of different network elements. Table II is the comparison among the traditional network, hybrid structure, OSA, and FatTree.

Traditional Oversubscribed Network: To connect 2560 servers using a two-tiered 2:1 oversubscribed structure, we use 80 48×10 G port ToR switches and 80 16×10 G port aggregation switches. Each ToR switch has 32 ports connect to servers, and the remaining 16 ports connect to aggregation switches, which results in a 2:1 oversubscription ratio. Note that we picked the 2:1 oversubscription because, for all the traffic patterns we studied in Section V, OSA delivers network bisection bandwidth that is at least 60% of the nonblocking network. Thus, a 2:1 oversubscribed traditional network (50% of the nonblocking) is a conservative comparison point. This structure costs $2.6 M and consumes 25.6 kW. The number of cross-ToR fibers required is 1280. The bisection bandwidth provided is 50% of the nonblocking network. However, for skewed traffic demands, it is desirable to allocate high fraction of this capacity to more demanding flows and achieve better cost/performance tradeoff.

Simplified Model of Hybrid Structure: Helios [11] and c-Through [14] are two well-known hybrid electrical/optical structures. The hybrid structure model we used here and in Section V is an abstract model that captures key aspects of both. In this model, each ToR connects to both an electrical network and an optical network. The electrical network is a two- or three-tiered tree with a certain oversubscription ratio (8:1 for Table II). In the optical part, each ToR has only one optical link to one other ToR, but this link is of unlimited capacity.[1] This hybrid structure costs $4.5 M, consumes 29 kW, and has

[1]We note that, while each ToR connects to one other ToR in c-Through, one Pod can connect to one or multiple other Pods at the expense of consuming more MEMS ports in Helios.

480 inter-ToR long fibers—160 above the MUX in the optical part and 320 above the ToRs in the electrical part.

OSA: The total cost is approximately $4.4 M, with a power consumption of 27.1 kW. ToRs and transceivers are responsible for a large portion of the cost and power budget. Compared to the traditional network, the additional cost is mainly due to transceivers, (DE)MUX, and WSS units. OSA uses DWDM transceivers that are more expensive than Gray transceivers used in the traditional network. The number of inter-ToR links required by OSA is only 320—the lowest of all these structures. OSA's cost is similar as the hybrid structure, but is more expensive than the traditional structure. However, it can dynamically adjust the bandwidth allocated to demanding flows. For all the traffic demands we evaluated in Section V, this enables OSA to achieve 60%–100% of the nonblocking bandwidth. We note that the cost of optics is expected to fall significantly with commoditization and production volume. Much of these benefits have already been reaped for the electrical technology. There is also scope for packaging multiple components on a chip—the 32 transceivers and the MUX could be packaged into one chip. This will reduce power, cost, as well as the number of fibers.

FatTree: The cost and power of FatTree mainly depend on switch port density: A FatTree topology with p-port switches can connect $p^3/4$ hosts with a total number of $5p^3/4$ ports. Note that for 10G port electrical switches, optical transceivers for remote connections is a necessity. To connect 2560 servers, FatTree costs $5.1 M. The power consumption is 51.2 kW. Especially, the wiring complexity for FatTree is the highest—the number of links above the ToR layer is 5120. FatTree is more expensive and consumes more power because it is designed to provide nonblocking connectivity and is also highly fault-tolerant. Our intention is not to perform a head-to-head comparison to FatTree, but to illustrate the cost/power/performance tradeoff of building a nonblocking network architecture.

Summary: For data center deployments where the skewed traffic demands are expected, we believe that OSA is a better alternative than either FatTree or the traditional oversubscribed networks: FatTree incurs higher cost, power, and wiring complexity, while the traditional architectures are inflexible and cannot assign spare bandwidth to demanding flows on the fly. Compared to the hybrid structure, OSA can achieve better performance with similar cost and power.

IV. Design

In this section, we present OSA network optimization in detail. Our goal is to compute the optimal topology and link capacities such that the network bisection bandwidth is maximized for a given traffic demand. Estimating traffic demand is not our main focus of this paper, and we assume this can be achieved using a similar way as Helios [11], c-Through [14], or Flyways [15]. For optimization, we need to find: 1) a MEMS configuration to adjust the topology to localize high traffic volumes; 2) routes between ToRs to achieve high throughput, low latency, or avoid congestion; and 3) a configuration for each WSS to provision the capacities of its outgoing links.

In the following, we first present a mathematical formulation for optimization. Considering its complexity, we then introduce an approximation solution.

A. Mixed-Integer Linear Programming Problem Formulation

Given: A traffic demand D between ToRs—d_{ij} is the desired bandwidth from ToR_i to ToR_j.

Variables: We use four sets of variables: $l_{ij} = 1$ if ToR_i is connected to ToR_j through MEMS and 0 otherwise; $w_{ijt} = 1$ if l_{ij} carries wavelength λ_t in the $i \rightarrow j$ direction, and 0 otherwise; v_{ijt} is the traffic volume carried by wavelength λ_t along $i \rightarrow j$; a traffic-served matrix $S - s_{ij}$ is the bandwidth achieved fromToR_i to ToR_j. For the last two sets of variables, s_{ij} have end-to-end meaning, while v_{ijt} have hop-to-hop significance. For all the variables, $t \in \{1, 2, \ldots, w\}$; $i, j \in \{1, 2, \ldots, N\}$, $i \neq j$; l_{ij} are the only variables for which $l_{ij} = l_{ji}$, and all the other variables are directional.

Objective: To achieve the optimal network bisection bandwidth, we maximize the traffic served

$$\text{Maximize} \sum_{i,j} s_{ij}. \tag{1}$$

Constraints: If the number of outgoing ports of the WSS is k, thenToR_i is connected to exactly k other ToRs

$$\forall i : \sum_{j} l_{ij} = k. \tag{2}$$

A wavelength λ_t can only be used between two ToRs if they are directly connected via MEMS

$$\forall i, j, t : w_{ijt} \leq l_{ij}. \tag{3}$$

To avoid wavelength contention, ToR_i can only receive/send λ_t from/to at most one ToR

$$\forall i, t : \sum_{j} w_{jit} \leq 1; \sum_{j} w_{ijt} \leq 1. \tag{4}$$

Traffic carried by λ_t between two ToRs is limited by ToR port capacity (C_{port}) and wavelength capacity (C_λ)

$$\forall i, j, t : v_{ijt} \leq \min\{C_{\text{port}}, C_\lambda \times w_{ijt}\}. \tag{5}$$

The outgoing transit traffic is equal to the incoming transit traffic at ToR_i

$$\forall i : \sum_{j,t} v_{ijt} - \sum_{j} s_{ij} = \sum_{j,t} v_{jit} - \sum_{j} s_{ji}. \tag{6}$$

Finally, the traffic served is bounded by the demand

$$\forall i, j : s_{ij} \leq d_{ij}. \tag{7}$$

The above mixed-integer linear program (MILP) can be viewed as a maximum multicommodity flow problem with degree bounds, further generalized to allow constrained choices in link capacities. While several variants of the degree-bounded

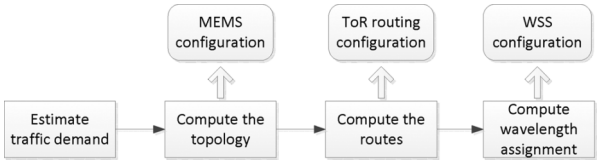


Fig. 3. Steps in the OSA control algorithm.

subgraph and maximum flow problems have known polynomial-time algorithms, the trivial combination of two is NP-hard [21].

B. Solution

As introduced above, in our approximation solution, we decompose the problem into three sequential subparts as shown in Fig. 3, i.e., computing the topology, computing the routing, and computing the wavelength assignment. Similar as Helios [11], in this paper, we adopt the traffic demand estimation method introduced in Hedera [22], which is based on the max-min fair bandwidth allocation for TCP flows in an ideal nonblocking network.

1) Compute the Topology: We localize high-volume communicating ToR pairs over direct MEMS circuit links. This is accomplished by using a weighted b-matching [23], where b represents the number of ToRs that a ToR connects to via MEMS ($b = k = 4$ in OSA-2560). In the ToR graph, we assign the edge-weight between two ToRs as the estimated demand between them, and then cast the problem of localizing high-volume ToR connections to b-matching. The weighted b-matching is a graph theoretic problem for which polynomial-time algorithm exists [23]. We implement it using multiple perfect matchings, for which public library is available [24].

The b-matching graph above is not necessarily a connected graph. Fortunately, connectivity is easy to achieve via the edge-exchange operation [25]. First, we find all the connected components. If the graph is not connected, we select two edges $a \rightarrow b$ and $c \rightarrow d$ with lowest weights in different connected components and connect them via replacing links $a \rightarrow b$ and $c \rightarrow d$ with links $a \rightarrow c$ and $b \rightarrow d$. We make sure that the links removed are not themselves cuts in the graph. The output of step 2 is used to tell the MEMS about how to configure the new topology.

2) Compute the Routes: Once we have connectivity, the MEMS configuration is known. We proceed to compute the routes using any of the standard routing schemes such as the shortest path routing or low congestion routing. Note that some of the routes are single-hop MEMS connections, while the others are multihop ones. For simplicity, we use the shortest path routing in this paper. However, our framework can be readily applied to other routing schemes. The output of step 3 is used by the ToRs to configure their routing tables.

3) Compute the Wavelength Assignment: Given the traffic demand and routes among ToRs, we compute the capacity desired on each ToR link in order to serve the traffic demand on this link.

With the desired capacity demand on each link, we need to provision a corresponding amount of wavelengths to serve the demand. However, wavelength assignment is not arbitrary: Due to the contention, a wavelength can only be assigned to a ToR at most once. Given this constraint, we reduce the problem to be the edge-coloring problem on a multigraph. We represent our ToR-level graph as a multigraph. Multiple edges correspond to the number of wavelengths between two nodes, and we assume each wavelength has a unique color. Thus, a feasible wavelength assignment is equivalent to an assignment from the colors to the edges of the multigraph so that no two adjacent edges have the same color—exactly the edge-coloring problem [26]. The edge-coloring is a known problem, and fast heuristics are known [27]. Libraries implementing this are publicly available.

We also require at least one wavelength to be assigned to each edge on the physical topology. This guarantees an available path between any pair of ToRs, which may be required for mice/bursty flows. The output of step 4 is used by the WSS to assign wavelengths.

All the above steps are handled by the OSA manager. Specifically, the OSA manager interacts with the MEMS, WSS units, and ToRs to control the topology, link capacities, and routing, respectively. We note that our decomposition heuristic is not optimal and there is room to improve. However, it provides satisfactory gains as we will see.

V. Simulation

In this section, we evaluate OSA-2560 via analytical simulations. We start with the simulation methodology, and then present the results.

A. Simulation Methodology

Simulation Goals: Since our testbed only has 8 ToRs (Section VI), to evaluate the performance of OSA at its intended scale, we conduct analytical simulations of the network bisection bandwidth of OSA-2560 under various traffic patterns. Our results in this section are essentially computations of the expected bisection bandwidth in the steady state, ignoring software and hardware overheads that are considered in our testbed experiments in Section VI. We compare OSA to a nonblocking network, a hybrid network with varied oversubscription ratios in the electrical part, and a 2:1 oversubscribed traditional network.

Communication Patterns: We use the following real measurement traces and synthetic traffic data to evaluate the performance of OSA in the presence of changing communication patterns and traffic demands.

1) Mapreduce-Demand: We collected real traffic matrices in a production data center with around 400 servers, which mainly runs Mapreduce applications.[2] We compute the network demands by averaging the traffic over 30-s periods. For each demand, we identify the communication pattern by filtering out mice flows and focusing on the elephant ones. We duplicate these traffic demands onto OSA-2560 using spatial replication.

2) Measurement-Based: Recent measurements [15], [28] reveal several data center traffic characteristics. One important feature is that the hotspot ToR links are often associated with a high fan-in (or fan-out), and most of the traffic (80%) are within the rack, resulting in a highly skewed distribution. We

[2]The name of the production data center company is anonymized.

synthesize this kind of traffic patterns by randomly choosing 12 hotspots out of 80 ToRs, with each one connecting to 6–10 other randomly chosen ToRs, respectively. We intentionally assume all the traffic exits the rack in order to create more intensive communications.

3) ToR-Level Shifting: We index the ToR switches from 0 to 79 and shift traffic round by round. Initially, all the servers in ToR i talk to all the servers in ToRs $(i \pm 1)$ mod 80 and $(i \pm 2)$ mod 80. Then, we shift these communications to servers in the next ToR after each round.

4) Server-Level Shifting: We index the servers from 0 to 2559. We start with server i talking to four other servers: $(i \pm 32)$ mod 2560 and $(i \pm 64)$ mod 2560. With 32 servers in a rack, initially, this implies that each rack communicates with four other racks. In successive rounds, server i talks to $(i \pm (32 + s))$ mod 2560 and $(i \pm (64 + s))$ mod 2560 $(s = 4, 8, 12, \ldots)$. This implies that each rack communicates with six racks in most rounds, with traffic spread across these six connections increasing and decreasing periodically.

5) Random Shifting: In each round, each server in ToR i talks to servers in up to 10 randomly selected ToRs. In this pattern, many ToRs may simultaneously talk to one ToR, creating hotspots and communication bottlenecks.

6) Increasing Destinations: We gradually increase the number of destinations for each ToR from 4 to 79 (i.e., all-to-all communications) to further investigate the impact of traffic spread on OSA performance.

Evaluation Metrics: First, we measure the network bisection bandwidth provided by OSA for each communication pattern. Then, we quantify the impact of the flexible topology and flexible link capacity of OSA architecture respectively. Finally, we measure the time cost of the control algorithm described in Section IV-B. The experiments are conducted on a Dell Optiplex machine with Intel 2.33 GHz dual-core CPU and 4 GB memory.

Hybrid Structure Model: We simulate the hybrid structure model introduced in Section III-C, which captures the key features of c-Through and Helios. To optimize the network to the traffic demand, we run the maximum weighted matching to determine which optical circuits to establish. Then, we calculate how much of the remaining demand can be satisfied by the electrical network at best.

Traditional 2:1 Oversubscribed Network: We also simulate a 2:1 oversubscribed electrical network whose details were described earlier in Section III-C.

B. Evaluation Results

1) Performance of OSA: In this experiment, the topology and link capacities are adaptively adjusted to the current traffic pattern. As soon as traffic pattern changes, the network reconfigures its topology instantaneously. In practice, the performance of OSA would be also impacted by the time taken to estimate the traffic demand, the time taken by the algorithms to identify the appropriate topology, and the reconfiguration time of the optical devices. Experimental results from our prototype will encompass all these overheads (Section VI).

Fig. 4 shows the average network bisection bandwidth over 100 instances of each traffic pattern obtained by different DCN

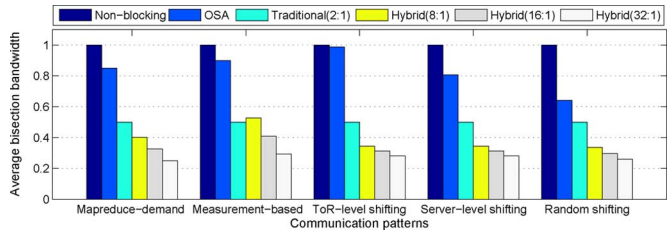


Fig. 4. Average network bisection bandwidth (normalized) achieved for different communication patterns.

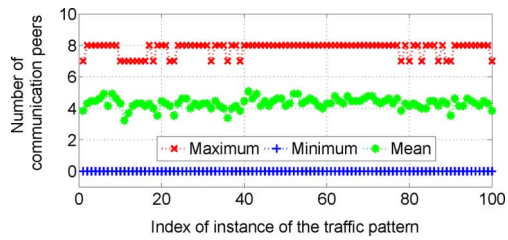


Fig. 5. Number of ToRs each ToR communicates with in every instance of the Mapreduce-demand pattern.

structures. Note that all the results are normalized by the bisection bandwidth of the nonblocking scenario. We make the following observations.

First, we find that OSA delivers high bisection bandwidth (60%–100% of the nonblocking network) for both the real and synthetic traffic patterns. Under the Mapreduce-demand, OSA can provide over 80% of the nonblocking bandwidth. This is because OSA adaptively changes its topology and link capacities according to the present traffic pattern. In our simulation setting, we choose 4-regular graph for OSA—that is why we are still 20% from the nonblocking given the communication distribution shown in Fig. 5. Because some ToRs talk to more than four (up to eight) other ToRs, OSA cannot assign direct circuits to feed all these communications. The multihop routing possibly causes congestion on the intermediate switches, leading to performance degradation. From the figure, we find that OSA delivers higher bandwidth (90% of the nonblocking) for the measurement-based pattern because it has relatively less hotspots compared to the previous one.

Second, when each ToR communicates with four other ToRs (in the ToR-level shifting pattern), OSA achieves bisection bandwidth nearly identical to that of the nonblocking network. This result is not surprising given that OSA allows a 4-regular graph and hence provides four optical circuits at each ToR to perfectly support the demand. Note that the traditional 2:1 oversubscribed network delivers 50% of the nonblocking for all the traffic patterns.

Third, in our results (not shown here due to lack of space), we observe that the bisection bandwidth achieved by OSA oscillates periodically from approximately 60% to 100% (with the average at 80%) of the nonblocking for the server-level shifting pattern. This is because each ToR would periodically communicate with four and six other ToRs in such traffic pattern. We further observe that the bisection bandwidth obtained by OSA in the random shifting pattern is the worst—60% of the nonblocking. This is expected since the number of peers each ToR

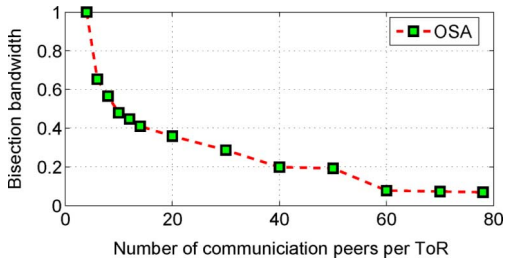


Fig. 6. Network bisection bandwidth with an increasing number of peers with whom each ToR communicates.

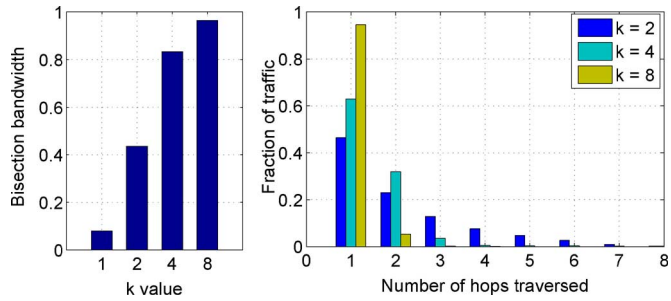


Fig. 7. Performance of OSA with *(left)* varied k values and *(right)* the number of hops traversed by traffic.

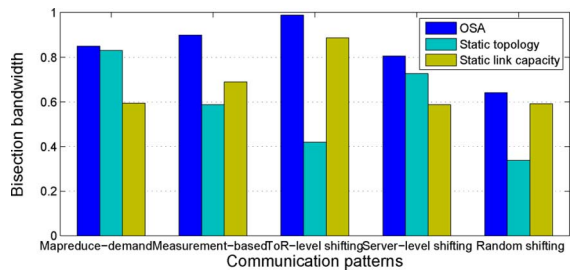


Fig. 8. Effect of flexible topology and flexible link capacity.

communicates with is larger than the other two shifting patterns. Specifically, for the ToR-level shifting, a ToR talks to four other peers. For the server-level shifting, a ToR communicates with four to six peers. For the random shifting pattern, a ToR communicates with 5–20 peers. As discussed above, when the number of communication peers for a ToR is larger than four, some flows will necessarily use multihop paths causing performance degradation. Concretely, most paths are direct for the ToR-level shifting, most paths are direct or 2 hops for the server-level shifting, and most paths are increased to 2–6 hops for the random shifting. Those flows passing through multiple hops would contend for the available bandwidth at the intermediate switches, limiting the maximal achievable throughput.

Next, we present the bisection bandwidth achieved by OSA with an increasing number of inter-ToR communications. As it moves gradually to the all-to-all communication (Fig. 6), as expected, the network bisection bandwidth drops due to the extensive bandwidth contention at the ToRs. Note that the traditional 2:1 oversubscribed network would continue to perform at 50% of nonblocking. This result is presented only for comparison purposes since OSA is not designed for the uniform all-to-all communication.

Furthermore, we note that OSA outperforms the hybrid model by 80%–250% in our evaluation. This is not a surprising result because the hybrid model only has a perfect matching between ToRs in the optical part. This means that one ToR is able to talk to one other ToR at a time. We increase oversubscription ratios in the electrical part from 32:1 to 8:1 and see only incremental improvement due to the oversubscribed network. In contrast, in OSA-2560, we have a 4-regular graph, meaning one ToR can directly communicate with four other ToRs simultaneously. Furthermore, OSA also dynamically adapts its link capacities to the traffic demand. The higher flexibility of OSA leads to its better performance.

In Fig. 7, we inspect the performance delivered by OSA with varied k values (left) and the number of hops traversed by the traffic (right) using the Mapreduce-demand. We assume that there are always 80 ToRs. It is evident from the left figure that with a larger k value, the network bisection bandwidth delivered is higher. However, the larger k value also necessitates more MEMS ports in order to support the same number of ToRs and servers. Note that $k = 2$, where we see low performance, is exactly equivalent to the optical part of the hybrid structure. From the right figure, we find that, for our case of OSA-2560 (i.e., $k = 4$), the vast majority of traffic only traverses less than 3 hops—over 60% of traffic goes one hop, and over 30% of traffic goes two hops. We also find that with a small k value, a considerable portion of traffic needs to traverse multiple hops to reach the destinations. When k increases, more traffic will go fewer hops, indicating better network performance. Though not shown, the similar trends hold for the remaining traffic patterns.

2) Effect of Flexible Topology and Link Capacity: We quantify the effect of the flexible topology and flexible link capacity respectively. For this purpose, in the first experiment we randomly select a fixed topology (e.g., the one generated by the first instance of a traffic pattern) and only adjust the link capacity according to the current traffic pattern. In the second experiment, we hypothetically assume each link has eight fixed wavelengths assigned (thus static link capacity) and only adjust the topology based on the current traffic pattern. Fig. 8 shows the bisection bandwidth of both scenarios and the original OSA. Comparing the static topology scenario to OSA, we observe up to $\frac{100\%-40\%}{40\%} = 150\%$ improvement due to the flexible topology in case of the ToR-level shifting pattern. Comparing the static link capacity scenario to OSA, we observe up to $\frac{90\%-60\%}{60\%} = 50\%$ improvement because of the flexible link capacity in case of the measurement-based traffic pattern. These results suggest that the flexible topology and link capacity are essential to improve the performance of OSA.

3) Time Cost of Control Algorithm: We measure the time cost of the OSA control algorithm as described in Section IV-B. We run our current software implementation with 50 randomly selected traffic patterns that we used above and compute the average value for each step. As shown in Table III, the total time is 290 ms. We observe that out of the four steps, the traffic demand estimation is dominant (161 ms). The reason is that the algorithm for this step is based on the number of servers, while the rest are based on the number of ToRs. Note that our demand estimation algorithm is adopted directly from Hedera [22], which has recently been shown to consume less than 100 ms for large

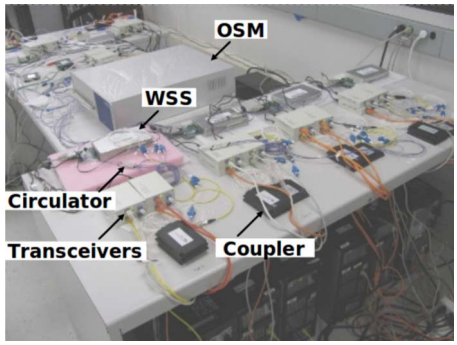


Fig. 9. OSA testbed.

TABLE III
TIME CONSUMPTION OF THE CONTROL ALGORITHM

Control algorithm	Time (ms)
Estimate traffic demand	161
Compute the topology	48
Compute the routes	41
Compute the wavelength assignment	40
Total	290

data centers via parallelization over multiple cores or machines. This means there is a large room to speed up with advanced technologies.

Though most of the remaining steps take only tens of milliseconds, we still believe optimizations are possible throughout the control software to make it more responsive even for larger networks. For instance, b-matchings for 1024 nodes could be computed in as few as 250 ms in the year 2000 with contemporary hardware [23]. It is also likely that better-performing, faster heuristics can be built based on more accurate models of the traffic.

VI. IMPLEMENTATION

We have built a small-scale OSA prototype with real optical devices (Fig. 9). We first introduce our testbed setup, and then present our experiments over it.

A. Testbed Setup

Our testbed connects 32 end-hosts, uniformly distributed in eight racks. To reduce the cost, we configure eight Dell Optiplex servers to emulate 32 end-hosts. Each server acts as a virtual rack of end-hosts (V-Rack), running four virtual-machines (VMs) to emulate four end-hosts.

We now do not have programmable ToR switches, so we use high-end servers to emulate ToRs. We have four Dell PowerEdge servers, each equipped with an Intel 2.4 GHz quad-core CPU, 8 GB DRAM, and 12 $\times$ 1 GigE NICs. On each such server, we deploy two VMs, giving us a total of eight virtual ToRs (V-ToRs). Each V-ToR binds to six NICs: One is connected to one V-Rack, one is used for a control connection to the OSA manager, and the remaining four are used as uplinks to reach other V-ToRs via optical elements.

On top of each V-ToR is a 1 $\times$ 4 CoAdna WSS, a coupler, a circulator, a 1 $\times$ 4 MUX and DEMUX pair, and four transceivers [which are packaged into a media converted (MC) unit]. As in Fig. 2, each ToR uplink is connected to a transceiver, with the send-fiber of the transceiver connected through the MUX,

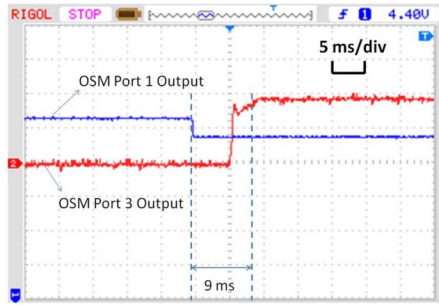


Fig. 10. Switching time of our OSM.

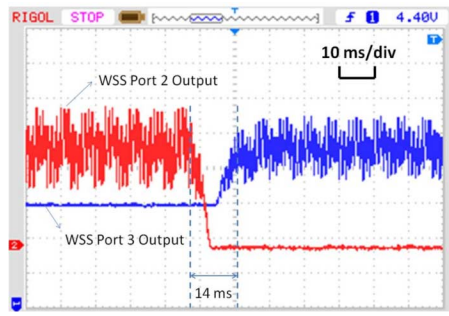


Fig. 11. Switching time of our WSS.

the WSS and the circulator to the OSM, and the receive-fiber connected to the same circulator through the coupler and the DEMUX. We use a 1 Polatis series-1000 OSM (a piezoelectric switch) with 32 ports, which allows a 16 $\times$ 16 bipartite interconnect. (Each V-ToR has two uplinks connected to each of these two sets of 16 ports.) We use four wavelengths: 1545.32, 1544.53, 1543.73, and 1542.94 nm, corresponding to channel 40, 41, 42, and 43 of ITU grid with 100 GHz channel spacing.

Furthermore, in our testbed, the OSA manager is a separate Linux server and talks to the OSM and ToRs via Ethernet ports, and to the WSS units via RS-232 serial ports.

B. Understanding the Optical Devices

Two critical optical devices in OSA are OSM and WSS. A common concern for them is the reconfiguration overhead. To measure the overhead, Fig. 10 shows the output power level on two ports of the OSM over time during a reconfiguration event. We see a clear transition period, i.e., from the high $\rightarrow$ low output power level shift on one port, to the low $\rightarrow$ high output power level shift on the other port. We observe that the switching delay of our OSM is 9 ms, consistent with [11] and [14].

Next, we measure the reconfiguration time of the WSS by switching a wavelength channel between two output ports. As shown in Fig. 11, this transition period is around 14 ms. However, the OSA manager can perform the reconfiguration of OSM and WSS in parallel to reduce the total time of reconfiguration.

C. Understanding the O-E-O Conversion

To measure the impact of O-E-O conversion, we specially connect four servers as in Fig. 12 (left). Two servers in the middle are configured as routers and equipped with optical media converters. We create a routing loop by configuring the IP forwarding tables of the routers. In each router,

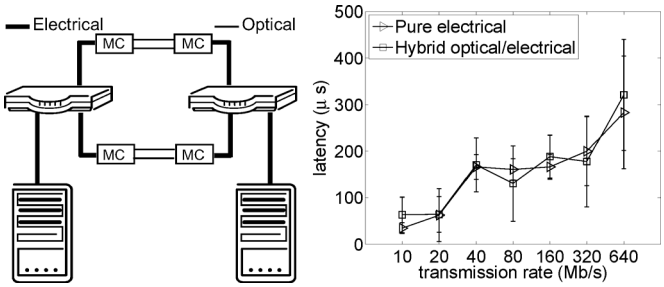


Fig. 12. Impact of O-E-O conversion.

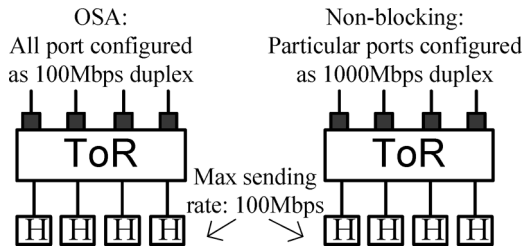


Fig. 13. Make a nonblocking network from OSA.

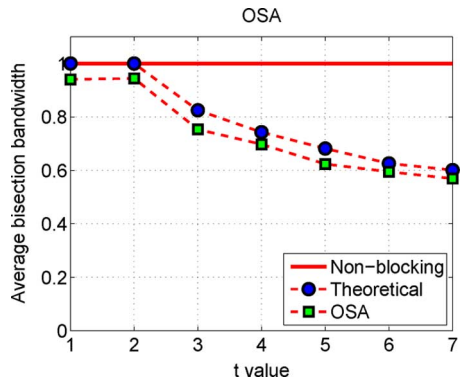


Fig. 14. Average bisection bandwidth of OSA.

we deploy a `netfilter` kernel module and utilize the `NF_IP_PRE_ROUTING` hook to intercept all IP packets. We record the time lag between the instant when the packets first arrive in the network and when their time to live (TTL) expires. This way, we are able to measure the multihop latency for O-E-O conversion and compare it to the baseline where all the servers are directly connected using only electrical devices. Results in Fig. 12 (right) compare the average one-hop switching latency for both the hybrid optical/electrical and pure electrical architectures under different traffic loads. It is evident from the figure that the O-E-O conversion does not incur noticeable (the maximum deviation in the absolute value and standard deviation is 38 and 58 μs, respectively), if any, additional switching latency, demonstrating the feasibility of O-E-O employed by OSA.

D. OSA System Performance

We conduct two sets of experiments: One is for original OSA, and the other is for OSA with static topology. We use synthetic traffic patterns similar to Section V-A. More specifically, traffic is described by parameters (t, r): Servers in ToR i $(i = 0 \ldots 7)$ send traffic to servers in t ToRs, i.e., $[i + r, i + r + 1, \ldots, i + r + (t-1)]$. We change t from 1 to 7 to generate different traffic loads ($t = 7$ means all-to-all communication). For each t, we vary r from 1 to 7.

Our goal is to compare the achieved bisection bandwidth of OSA against that of a nonblocking network as the traffic spreads out (with increasing t) and to measure the effect of topology reconfiguration. Note that varying r with a fixed t does not produce fundamentally different traffic distributions, as it merely permutes which ToRs talk with which other ToRs, thus necessitating a change of topology without a change in traffic load or spread.

In our testbed, the server NICs support 10, 100, and 1000 Mb/s full-duplex modes. In all our experiments, we limit the maximum sending rate of each flow to be 100 Mb/s. This enables us to emulate a nonblocking network for comparison (Fig. 13): For OSA, all the uplink ports of ToRs are set at 100 Mb/s, while for the nonblocking, we increase some particular uplink ports to be 1000 Mb/s to satisfy the traffic demands we simulate.

Results of OSA: Fig. 14 shows the average bisection bandwidth of OSA with changing traffic $(t = 1 \ldots 7)$. For each t, r steps 1 through 7 every 20 s. The network topology is dynamically reconfigured according to the current traffic demand. The results are along expected lines. We observe that the achieved bisection bandwidth of OSA is within 95% of the nonblocking network when t is 1 or 2. This is because when $t = 1$, each ToR talks with two other ToRs, and when $t = 2$, each ToR talks with four other ToRs. Given that our topology is a 4-regular graph, OSA assigns direct links to each pair of communicating ToRs for efficient communication. For $t > 2$, the performance of OSA decreases, along similar lines as in the simulation (Section V). A careful reader will notice that the performance of our testbed under the all-to-all communication is 58% of the nonblocking, much higher than that in our simulation results. The reason is simple: Our testbed has eightToRs, each having a degree 4, while our simulations used a sparse graph with 80 ToRs, each having a degree 4. Our intention of the testbed results is to demonstrate the feasibility of OSA rather than to show the performance achieved in a real deployment.

Next, Fig. 15 shows the impact of optical device reconfigurability on the end-to-end throughput between two hosts. We observe that the performance drops during reconfiguration, but quickly resumes after it finishes.

Finally, we also present the theoretical bisection bandwidth achievable in our testbed that ignores the overhead of reconfiguration, software routing, and TCP/IP protocol, etc. We observe that the gap between the theoretically achievable values and OSA is within 5%–7%, suggesting that our prototype implementation is efficient.

Results of OSA With a Static Topology: We randomly select a topology and run the same experiments as above. We present the results in Fig. 16. Given the small diameter of our topology, the static topology OSA still achieves satisfactory performance. For example, in the worst case of all-to-all traffic (i.e., $t = 7$), static

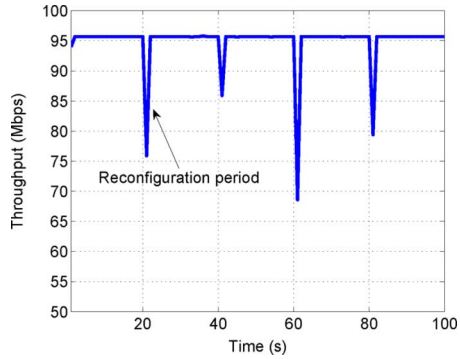


Fig. 15. Throughput of a flow in the presence of reconfigurations.

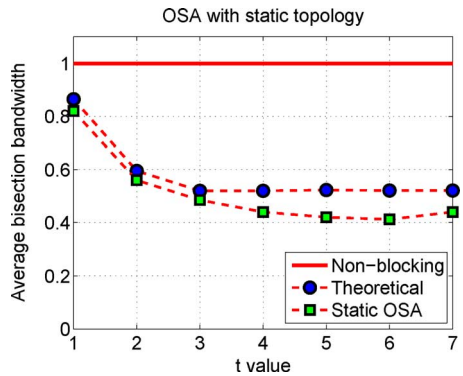


Fig. 16. Average bisection bandwidth of OSA with a static topology.

OSA achieves more than 40% of the nonblocking network's bisection bandwidth. Since all the paths are 1 or 2 hops long, even the randomly selected topology performs satisfactorily.

For different t values, we find that the performance of OSA on the static topology is lower than that on the dynamic topology by 10%–40%. This is because the topology is not optimized for the current traffic pattern. We expect that on a larger network where OSA topology is sparse (e.g., the one we used in Section V), this performance gap will become more pronounced, highlighting the need for a dynamically optimized network for better performance.

E. Bulk Data Transfer

We study how the network reconfiguration and multihop routing affect the bulk data transfer, i.e., elephant flows.

Impact of Network Reconfiguration: We periodically reconfigure the network and observe the completion time of transferring a chunk of data (a 100-MB file transferred using `scp`) during the reconfiguration events. We present the mean value of 100 trials. Fig. 17 shows our results and the baseline performance where no reconfiguration takes place. The stability time is defined as the lifetime for a single static topology, after which the network is reconfigured. We notice that the completion time increases in the presence of reconfigurations. After analyzing the network trace using `tcpdump`, we observed that the round-trip time (RTT) and accordingly the initial retransmission time out (RTO) values in data centers are very small (submillisecond level), while network reconfiguration requires tens of milliseconds. As a consequence, each reconfiguration almost always triggers RTO events, after which TCP waits for

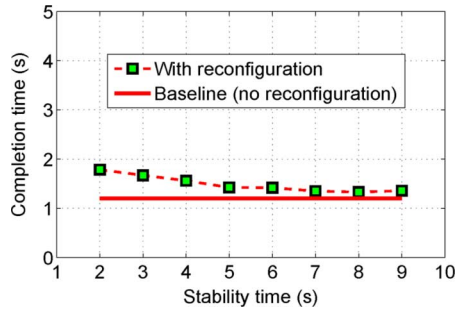


Fig. 17. Impact of topology reconfiguration on bulk data transfer.

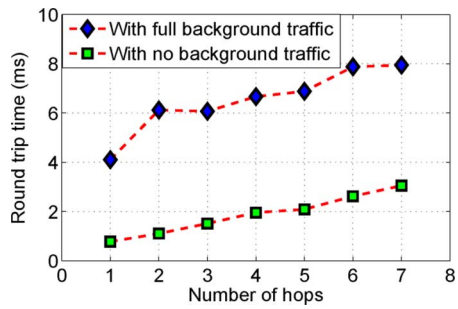


Fig. 18. Impact of multihop routing on bulk data transfer.

200 ms (Linux default RTO value) before the next retransmission, thereby degrading throughput and increasing latency. Recent work [29]–[31] has pointed out TCP's RTO issues in data centers and proposed to reduce it to the microsecond level by employing fine-grained timers. We expect TCP's performance in OSA under network reconfiguration to significantly improve once these changes are adopted. We also note from the figure that the completion time decreases as the stability time increases—larger stability period results in fewer network state changes and thus fewer RTO events during the course of data transfer.

Impact of Multihop Routing: Our prototype topology is a low-diameter network due to an 8-node 4-regular graph. In order to evaluate the impact of multihop routing on bulk data transfer, we rearrange our eight ToRs in a line topology with a larger diameter. In Fig. 18, we measure the completion time of data transfer (transferring a 100-MB file using `scp`) with increased hops. Specifically, we consider two cases: 1) the network is free of background traffic; 2) all the links are saturated by background elephant TCP flows. From the figure, we find that in both cases the completion time is relatively consistent regardless of the number of hops. These results imply that the influence of multihop O-E-O conversion during data transfer on our testbed is small, which is coherent with our observation in Section VI-C. We also observe a nearly constant gap between the two curves, which is due to the different link utilization in the two cases.

F. Mice Flow Transfer

After inspecting the performance of bulk data transfer, we further check the impact of multihop routing on transferring mice flows. For this purpose, we use `ping` to emulate latency sensitive flows and evaluate its performance with and without background traffic as above. Fig. 19 shows the average RTT of

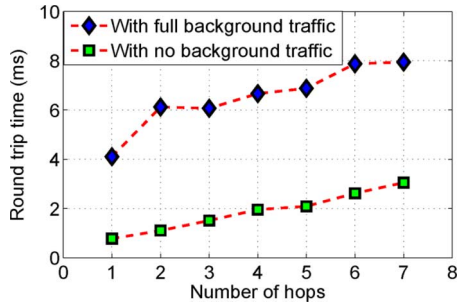


Fig. 19. Impact of multihop routing on simulated mice flows.

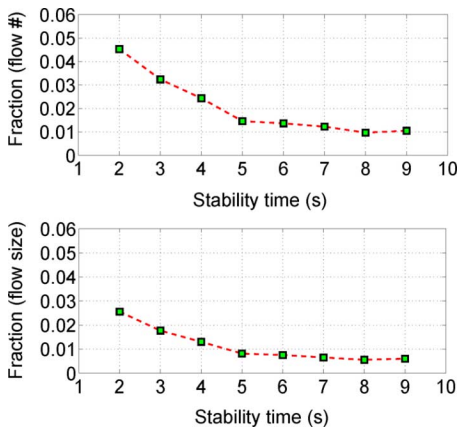


Fig. 20. Potentially affected mice flows during network reconfiguration.

100 `ping` packets with varying path lengths. As expected, the RTT increases with more hops: 1 ms without background traffic and 2 ms with full background traffic, respectively, after 7 hops. These results suggest that the hop-by-hop stitching of optical links is a feasible approach to provide the overall connectivity. We note that network reconfiguration may have nontrivial impact on the latency-sensitive flows transfer since it happens on the order of 10 ms. We further discuss options to handle such issues in Section VII.

VII. Discussion and Related Work

A. Mice Flow During Reconfiguration

OSA ensures that all the ToRs are in a connected graph and uses the hop-by-hop stitching of existing circuits to provide overall network connectivity. However, during the network reconfiguration, a pair of ToRs may be temporarily disconnected for around 10 ms. While this can be largely tolerated by latency-insensitive applications such as Mapreduce or Dryad, it would affect those operating with latency-sensitive mice flows like Dynamo [32].

In Fig. 20, we estimate, in the worst case, how many mice flows (in terms of flow count and size) can be potentially affected due to the reconfiguration. We use the production data center traffic from Section V-A and use 10 MB to differentiate the elephant flows from the small ones. We find that for this particular dataset, when the stability time varies from 9 to 2 s, there are 1% to 4.5% of the mice flows that can be affected during the reconfigurations. This implies that as the network experiences more frequent reconfigurations, a larger fraction of mice flows may get affected. We next discuss two possible options to handle this issue.

Our basic idea is to reserve a static, connected channel in OSA network. To do so, we can reserve a small number of wavelengths and MEMS/WSS ports that are never reconfigured, and mice flows are sent over them. Such a channel can be simply a spanning tree or other connected topologies. Given the topology of the channel is controlled by the MEMS, we can arrange it in a low-diameter manner so that the transmission of mice flows is optimized. However, this approach consumes expensive MEMS/WSS ports, which otherwise can be better utilized for other applications or at stable time.

An alternative approach to building the channel without using MEMS/WSS ports is directly connecting all the ToRs together to form a ring or a star network. For the ring, we can reserve two ports on each ToR and directly connect them iteratively. In case of OSA-2560 with 80 ToRs, the diameter is 40 hops. To reduce the path length, it is possible to reserve more ports on each ToR and connect them structurally using DHT techniques [33], e.g., the diameter is expected to be 3–4 hops with high probability for 80 ToRs if we reserve four ports on each ToR. Another option is to employ one additional central electrical switch—each ToR uses one port to connect to the central switch. Note that, in Helios or c-Through, the electrical switches (usually forming a tree or even a multiroot tree) are used for overall connectivity among all the Pods/ToRs. In OSA, the all-to-all connectivity is maintained by optical components. A comprehensive evaluation and comparison of these solutions is part of our ongoing work.

B. OSA Applicability Versus Traffic Properties

For the all-to-all traffic, the nonoversubscribed network is indeed more appreciated. However, such workloads are neither reflected in our dataset nor in the measurements elsewhere [2], [15], [28]. Our flexible OSA architecture would work best when the traffic pattern is skewed and stable on the order of seconds. It has been noted in [5] over the measurements of a 1500-server production DCN that “only a few ToRs are hot and most of their traffic goes to a few other ToRs.” Another study [2], also on a 1500-server production DCN, shows that more than 90% of bytes are in elephant flows. Regarding the traffic stability, a similarly sized study [34] shows that 60% of ToR-pairs see less than 20% change in traffic demands for between 1.6 to 2.2 s on average. Despite these, we expect that OSA may exhibit undesirable performance degradation if the traffic pattern is highly dynamic, in which case any topology adaptation mechanism may be unsuitable as the situation changes instantaneously. In practice, the infrastructure manager should choose the proper sensitivity of OSA according to the operational considerations.

C. Scalability

The current OSA design focuses on the container-size DCNs. To scale, we may confront several challenges. The first one is the MEMS’s port density. While the 1000-port MEMS is theoretically feasible, the largest MEMS as of today has 320 ports. One natural way to increase the port density is via interconnecting multiple small MEMS switches. However, this poses additional requirement for fast, coordinated circuit switching.

Second, larger network size necessitates more control and management overhead. In our OSA-2560 with 80 ToRs, all the intelligences, e.g., the network optimization and routing, are handled by the OSA manager. How to handle such tasks in a larger DCN with thousands of ToRs is an open question especially when the network environment is dynamic. Furthermore, circuit visit delay [14] is another issue to notice when scaling. We are considering all these challenges in our continuous effort designing a scalable optical DCN.

D. Closely Related Work

OSA's design goals are closely related to those of c-Through [14] and Helios [11]. In both approaches, flows requiring high bandwidth are dynamically provisioned on optical circuits, while a parallel electrical network is used to provide overall connectivity. OSA differs from these prior proposals in its degree of flexibility and its architecture. Both Helios and c-Through achieve some topology flexibility via a limited number of single-hop optical links. In their optical parts, one ToR only connects to one other ToR at a time. While it can connect to different ToRs at different time, the switching latency would be around 10 ms. On the contrary, in OSA, one ToR can connect to multiple ToRs simultaneously at a time, and multihop connection exists between any pair of remote ToRs through the hop-by-hop circuit stitching. Furthermore, OSA allows the link capacities to be adjusted on the fly. Unlike these existing hybrid architectures, OSA avoids using electrical components other than the ToR switches.

OSA is more comparable to c-Through than Helios because its current target is interrack DCNs with a few thousand servers, unlike Helios' intercontainer mega-DCN scale. Qualitatively, OSA provides more flexibility than either Helios or c-Through and is able to serve a larger space of skewed traffic demands with performance similar to that of the nonblocking network. We present a coarse quantitative comparison to an abstract hybrid architecture model in Section V, showing that OSA achieves significantly higher bisection bandwidth.

Recently, Kandula *et al.* [5], [15] proposed to dynamically configure 60-GHz short-distance multi-Gigabit wireless links between ToRs to provide additional bandwidth for hotspots. Optical and wireless interconnects provide different tradeoffs. For example, wired optical interconnects can deliver much more bandwidth at lower power consumption over long distance, while the wireless has lower costs and is easier to deploy though the management and interference are challenging issues to deal with.

VIII. Conclusion

In this paper, we have presented OSA, a novel Optical Switching Architecture for DCNs. OSA is highly flexible because it can adapt its topology as well as link capacities to different traffic patterns. We have evaluated OSA via extensive simulations and prototype implementation. Our results suggest that OSA can deliver high bisection bandwidth (60%–100% of the nonblocking network) for a series of traffic patterns. Our implementation and evaluation with the OSA prototype further demonstrate its feasibility.

Acknowledgment

The authors thank CoAdna Photonics and Polatis for the equipment support of WSS and OSM respectively, G. Liao for helpful discussions on WSS properties, and D. DePaolo for help during the testbed setup. The authors thank their NSDI shepherd D. Andersen, the anonymous IEEE/ACM Transactions on Networking and NSDI reviewers, and C. Guo for their valuable comments.

References

[1] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proc. SIGCOMM*, 2008, pp. 63–74.

[2] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "VL2: A scalable and flexible data center network," in *Proc. ACM SIGCOMM*, 2009, pp. 51–62.

[3] R. N. Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat, "PortLand: A scalable fault-tolerant layer 2 data center network fabric," in *Proc. ACM SIGCOMM*, 2009, pp. 39–50.

[4] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "BCube: A high performance, server-centric network architecture for modular data centers," in *Proc. ACM SIGCOMM*, 2009, pp. 63–74.

[5] S. Kandula, J. Padhye, and P. Bahl, "Flyways to de-congest data center networks," in *Proc. ACM HotNets*, 2009.

[6] K. Barker, A. Benner, R. Hoare, A. Hoisie, A. K. Jones, D. K. Kerbyson, D. Li, R. Melhem, R. Rajamony, E. Schenfeld, S. Shao, C. Stunkel, and P. Walker, "On the feasibility of optical circuit switching for high performance computing systems," in *Proc. SC*, 2005, p. 16.

[7] J. Hamilton, "Data center networks are in my way," 2009 [Online]. Available: http://mvdirona.com/jrh/TalksAndPapers/JamesHamilton_CleanSlateCTO2009.pdf

[8] H. Liu, C. F. Lam, and C. Johnson, "Scaling optical interconnects in datacenter networks opportunities and challenges for WDM," in *Proc. IEEE Symp. High Perform. Interconnects*, 2010, pp. 113–116.

[9] C. Lam, H. Liu, B. Koley, X. Zhao, V. Kamalov, and V. Gill, "Fiber optic communication technologies: What's needed for datacenter network operations," 2010.

[10] J. Rath, "Google eyes 'optical express' for its network," 2010 [Online]. Available: http://www.datacenterknowledge.com/archives/2010/05/24/google-eyes-optical-express-for-its-network/

[11] 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 *Proc. ACM SIGCOMM*, 2010, pp. 339–350.

[12] Communications Industry Researchers, Charlottesville, VA, USA, "40G Ethernet—Closer than ever to an all-optical network," 2012 [Online]. Available: http://cir-inc.com/resources/40-100GigE.pdf

[13] K. Patel, "The imminent reality of 40 and 100 Gigabit Ethernet," 2010 [Online]. Available: http://www.nxtbook.com/nxtbooks/bicsi/news_20100910/index.php?startid=39

[14] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. E. Ng, M. Kozuch, and M. Ryan, "c-Through: Part-time optics in data centers," in *Proc. ACM SIGCOMM*, 2010, pp. 327–338.

[15] D. Halperin, S. Kandula, J. Padhye, P. Bahl, and D. Wetherall, "Augmenting data center networks with multi-gigabit wireless links," in *Proc. ACM SIGCOMM*, 2011, pp. 38–49.

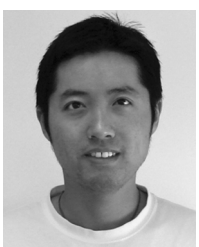
[16] M. Meringer, "Regular graphs," 2009 [Online]. Available: http://www.mathe2.uni-bayreuth.de/markus/reggraphs.html

[17] T. Truex, A. A. Bent, and N. W. Hagood, "Beam steering optical switch fabric utilizing piezoelectric actuation technology," in *Proc. NFOEC*, 2003.

[18] J. Kim, C. J. Nuzman, B. Kumar, D. F. Lieuwen, J. S. Kraus, A. Weiss, C. P. Lichtenwalner, A. R. Papazian, R. E. Frahm, N. R. Basavanhally, D. A. Ramsey, V. A. Aksyuk, F. Pardo, M.-E. E. Simon, V. A. Lifton, H. B. Chan, M. Haueis, A. S. Gasparyan, H. R. Shea, S. C. Arney, C. A. Bolle, P. R. Kolodner, R. R. Ryf, D. T. Neilson, and J. V. Gates II, "1100 × 1100 port MEMS-based optical crossconnect with 4-dB maximum loss," *IEEE Photon. Technol. Lett.*, vol. 15, no. 11, pp. 1537–1539, Nov. 2003.

[19] Broadcom, Irvine, CA, USA, "BCM56840 series enables mass deployment of 10 GbE in the data center," 2013 [Online]. Available: http://www.broadcom.com/products/features/BCM56840.php

[20] K. Chen, C. Guo, H. Wu, J. Yuan, Z. Feng, Y. Chen, S. Lu, and W. Wu, "Generic and automatic address configuration for data center networks," in *Proc. SIGCOMM*, 2010, pp. 39–50.
[21] E. Akcali and A. Ungor, "Approximation algorithms for degree-constrained bipartite network flow," in *Proc. ISCIS*, 2003, pp. 163–170.
[22] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: Dynamic flow scheduling for data center networks," in *Proc. NSDI*, 2010, p. 19.
[23] M. Müller-Hannemann and A. Schwartz, "Implementing weighted bmatching algorithms: Insights from a computational study," *J. Exp. Algor.*, vol. 5, p. 8, 2000.
[24] "LEMON-Library," 2011 [Online]. Available: http://lemon.cs.elte.hu
[25] K. Obraczka and P. Danzig, "Finding low-diameter, low edge-cost, networks USC, Los Angeles, CA, USA, , Tech. Rep., 1997.
[26] "Edge-coloring," 2013 [Online]. Available: http://en.wikipedia.org/wiki/Edge_coloring
[27] J. Misra and D. Gries, "A constructive proof of Vizing's theorem," *Inf. Process. Lett.*, vol. 41, no. 3, pp. 131–133, 1992.
[28] T. Benson, A. Akella, and D. Maltz, "Network traffic characteristics of data centers in the wild," in *Proc. IMC*, 2010, pp. 267–280.
[29] Y. Chen, R. Griffith, J. Liu, R. H. Katz, and A. D. Joseph, "Understanding TCP incast throughput collapse in datacenter networks," in *Proc. ACM WREN*, 2009, pp. 73–82.
[30] V. Vasudevan, A. Phanishayee, H. Shah, E. Krevat, D. G. Andersen, G. R. Ganger, G. A. Gibson, and B. Mueller, "Safe and effective fine-grained TCP retransmissions for datacenter communication," in *Proc. ACM SIGCOMM*, 2009, pp. 303–314.
[31] H. Wu, Z. Feng, C. Guo, and Y. Zhang, "ICTCP: Incast congestion control for TCP," in *Proc. ACM CoNEXT*, 2010, p. 13.
[32] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, and P. Vosshall, "Dynamo: Amazon's highly available key-value store," in *Proc. SOSP*, 2007, pp. 205–220.
[33] A. Rowstron and P. Druschel, "Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems," in *Proc. Middleware*, 2001, pp. 329–350.
[34] T. Benson, A. Anand, A. Akella, and M. Zhang, "The case for fine-grained traffic engineering in data centers," in *Proc. USENIX INM/WREN*, 2010, p. 2.



Kai Chen received the Ph.D. degree in computer science from Northwestern University, Evanston, IL, USA, in 2012.

He is an Assistant Professor with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong. His research interests include networked systems design and analysis, data center networks, and cloud computing.

Ankit Singla received the Bachelor's degree in computer science and engineering from the Indian Institute of Technology (IIT), Bombay, India, in 2008, and is currently pursuing the Ph.D. degree in computer science at the University of Illinois at Urbana–Champaign, Urbana, IL, USA.

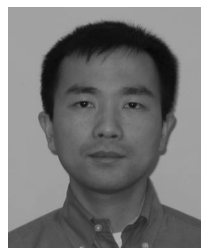
Mr. Singla was a recipient of Google's 2012 Ph.D. Fellowship in data center networking.

Atul Singh received the Ph.D. degree in computer science from Rice University, Houston, TX, USA, in 2009.

He then joined NEC Research Labs, Princeton, NJ, USA. Currently, he is a founder of an e-commerce startup. His main areas of research includes distributed systems, software fault-tolerance, and cloud computing.

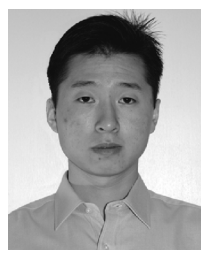
Kishore Ramachandran received the B.E. degree in electronics and telecommunications engineering from the University of Mumbai, Mumbai, India, in 2000, the M.S. degree in computer and systems engineering from Rensselaer Polytechnic Institute (RPI), Troy, NY, USA, in 2002, and the Ph.D. degree in computer engineering from Rutgers University, New Brunswick, NJ, USA, in 2009.

His interests broadly include design and implementation of networked computing systems. Past work includes designing, prototyping, and exploring the applicability of gigabit wireless networks to multiple domains (including data-centers) and building the world's largest, open-access, indoor wireless testbed (ORBIT) that was awarded the NSF's 4th Annual Schwarzkopf Award for Technological Innovation in 2008.



Lei Xu received the B.S. degree in geophysics from Peking University, Beijing, China, in 1997, the M.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2000, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2005.

He was with NEC Labs America, Princeton, NJ, USA, from 2004 to 2012, and is currently pursuing an entrepreneurship adventure. He has published more than 80 papers in journals and technical conferences, covering various topics in optical switching, networking, transmission, and devices. His research interests include advanced modulation and coding schemes for high-speed optical communications, innovative optical devices, and optical data center networks.



Yueping Zhang (S'03–M'08) received the B.S. degree from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2001, and the Ph.D. degree from Texas A&M University, College Station, TX, USA, in 2008, both in computer science.

He has been a Research Staff Member With NEC Labs America, Princeton, NJ, USA. His research interests include data center networking, software defined networking, and Internet congestion control.

Xitao Wen received the B.S. degree in computer science from Peking University, Beijing, China, in 2010, and is currently pursuing the Ph.D. degree in computer science at Northwestern University, Evanston, IL, USA.

His research interests span the area of networking and security in networked systems, with a current focus on software-defined network security and data center networks.

Yan Chen received the Ph.D. degree in computer science from the University of California, Berkeley, CA, USA, in 2003.

He is an Associate Professor with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA. Based on Google Scholar, his papers have been cited over 3600 times. His research interests include network security and measurement and diagnosis for large-scale networks and distributed systems.

Dr. Chen won the Department of Energy (DoE) Early CAREER Award in 2005, the Department of Defense (DoD) Young Investigator Award in 2007, and the Microsoft Trustworthy Computing Awards in 2004 and 2005 with his colleagues.