Rethinking the Physical Layer of Data Center Networks of the Next Decade: Using Optics to Enable Efficient *-Cast Connectivity

Howard Wang
Columbia University
howard@ee.columbia.edu

Yiting Xia
Rice University
yiting.xia@rice.edu

Keren Bergman
Columbia University
bergman@ee.columbia.edu

T. S. Eugene Ng
Rice University
eugeneng@rice.edu

Sambit Sahu
IBM T.J. Watson Research Center
sambits@us.ibm.com

Kunwadee Sripanidkulchai
NECTEC Thailand
kunwadee@nectec.or.th

ABSTRACT

Not only do big data applications impose heavy bandwidth demands, they also have diverse communication patterns (denoted as *-cast) that mix together unicast, multicast, incast, and all-to-all-cast. Effectively supporting such traffic demands remains an open problem in data center networking. We propose an unconventional approach that leverages physical layer photonic technologies to build custom communication devices for accelerating each *-cast pattern, and integrates such devices into an application-driven, dynamically configurable photonics accelerated data center network. We present preliminary results from a multicast case study to highlight the potential benefits of this approach.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*circuit-switching networks, network topology, packet-switching networks*

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

Data Center Networks, Optical Networks, Hybrid Networks, Photonics, Multicast, Unicast, Incast, All-to-All-Cast

1. INTRODUCTION

Current big data applications have diverse communication patterns [7]. Due to massive traffic volumes, these communication patterns lay increasing burden on data center networks. Measurement data in [14] shows that a typical group data delivery can involve up to 1000 servers with over 500MB of source traffic volume. [8] claims that the shuffle phase (all-to-all data exchange) on average accounts for 33% of the running time of Hadoop jobs.

Such examples are indicative of a clear need for flexible network capacity that cannot be met by the oversubscribed networks pervasive in today's production data centers. It is commonly believed that nothing less than full-bisection bandwidth networks are required in order to support such rich, high-volume traffic. While numerous non-blocking electronic packet-switch-based network designs have been proposed [2, 12, 13], only a handful of institutions have the necessary scale and wherewithal to support the significant cost, power, and wiring complexity required to build and maintain such systems.

The goal of this paper is to explore the role of the physical layer in supporting these diverse patterns *beyond* recent proposals that advocate the use of optical point-to-point circuits for traffic acceleration [11, 20, 6, 21, 10].

We classify the various communication patterns into four elementary *-cast categories. **Unicast** is standard point-to-point transmission of traffic flows, such as Virtual Machine (VM) migration, general-purpose data backup, and stream data processing. **Multicast** is a data transfer from a single sender to a group of receivers. Typical applications include distributing or updating software to servers within a cluster, replicating data to several servers in a distributed file system to improve reliability, and provisioning an OS image to a large number of VMs. **Incast** occurs when data are aggregated between servers in a many-to-one manner. For example, a MapReduce reducer needs to collect intermediate results from all the mappers for the reduce-phase computation, and operations in parallel database systems require merging data from many tables. **All-to-all-cast** delivers traffic among a set of nodes. This pattern is common in MapReduce shuffle where mappers and reducers concurrently exchange data. It also appears in high-performance computing tasks such as MPI FFT that iteratively retrieve intermediate results from other nodes.

Our key insight is that there are unique photonic technologies—such as passive directional couplers and wavelength combiners—that can provide physical layer capabilities that align well with each *-cast pattern, while maintaining the energy and capacity advantages enabled by circuit-switched optics. By treating modules of such devices as so-called "gadgets", the goal is to allocate specialized combinations of gadgets to satisfy these *-cast communication demands as they evolve in the network.

In this paper, we propose a new network architecture in which a library of function-specific photonic gadgets are dynamically used to accelerate each *-cast pattern. These *-cast-accelerating gadgets are integrated into a reconfigurable optical fabric such that they can be flexibly connected to

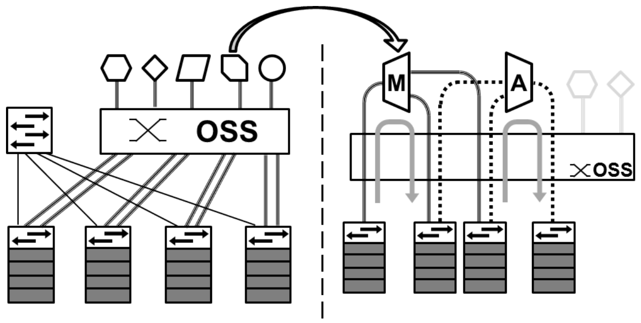

Figure 1: Photonic gadget-based network architecture (left). Using the OSS as a connectivity substrate to deliver multicast and incast (right).

nodes across the system. Such an optical fabric can be realized by a high-radix optical space switch. This architecture is qualitatively different from previous proposals that use optical circuit switches in the data center. Instead of serving merely as a server-rack to server-rack traffic carrier, the optical circuit switch in our architecture routes traffic to and from gadgets, thus enabling the flexible and dynamic use of photonic gadgets in various combinations to support complex traffic patterns. As these gadgets are all-optical, communication across the optical network is "single-hop" and, as a result, the capacity advantages afforded by the optical medium are maintained. Furthermore, as many of these gadgets can be realized using passive optical devices, such rich connectivity can be delivered with little to no additional power consumption. As a result, our gadget-based approach can potentially provide effective support for *-cast traffic with significantly less cost and complexity than a non-blocking network offering comparable performance.

We describe the architecture, photonic capabilities, and advantages of our design in §2. In §3, we present multicast as a case study to illustrate the early promise of our proposed architecture. Finally, we conclude in §4.

2. NETWORK ARCHITECTURE

A schematic representation of our network architecture is depicted in Figure 1. Our design consists of a hybrid aggregation layer composed of electronic packet switches and an optical network subsystem connected to top-of-rack (ToR) switches. The optical component of our design begins with fixed-wavelength transceivers at each ToR switch generating sets of non-overlapping channels. Each set is multiplexed together to form a wavelength-division multiplexed (WDM) signal and connected to each port of a high-radix optical space switch (OSS). While fully operational as a stand-alone point-to-point circuit switch, the primary purpose of the OSS is to serve as a reconfigurable connectivity substrate to provide agile system-wide connectivity to a library of advanced optical components.

In this way, we regard each optical component as a "gadget" (represented as a shape on the left-hand side of Figure 1) to be "attached" to the relevant ToR switches through the OSS. On the right-hand side of Figure 1, we depict a case where photonic multicaster and aggregator gadgets are connected through the OSS to four racks (with arrows indicating the direction of traffic flow). Gadgets and how they are assembled to form a given network can be managed by a central controller, which can either accept explicit requests for gadgets from applications or provision them based on demand estimation. By dynamically matching the unique performance profile of each photonic gadget to a given communication pattern at run-time, the resulting network architecture can effectively support the heterogeneous and ever-evolving *-cast traffic characteristic of data centers.

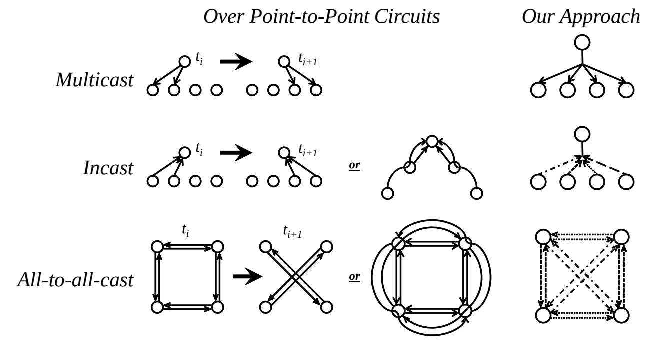

Figure 2: Comparison of point-to-point circuits (limited to two optical ports per rack) and our approach in handling multicast, incast, and all-to-all-cast. t_i and t_{i+1} denote two consecutive reconfiguration periods; each line style represents a wavelength.

2.1 Photonics-Enabled *-Cast

In addition to the inherent advantages in energy and raw capacity of photonic transmission, the defining capabilities of the optical medium—i.e., WDM and low transmission loss—can be leveraged in unique ways to provide a wide spectrum of capabilities, all while maintaining bit-rate transparency. Table 1 [17] provides a brief survey of various optical technologies and their characteristics. As we describe below, our gadget-based design can utilize such technologies to realize a broad spectrum of physical layer functionalities to enable intrinsic support for these data-intensive *-cast patterns beyond what is achievable by point-to-point circuits alone.

Unicast: Point-to-point unicast is straightforwardly achieved in photonics using optical space switches. In the simplest case, unicast connectivity can be realized using the MEMS-based OSS component of our architecture—a so-called ∅-gadget, exactly as proposed in [11, 20]. However, we can envision using combinations of faster space switches in tandem with MEMS to support different classes of unicast connectivity depending on an application's traffic requirements. For example, fast OSSs can serve to aggregate shorter messages from a small set of racks and forward them to singular optical circuit, which can be allocated to any other rack within the network. A switch design such as Mordia [10], which uses a fast OSS to enable fine-grained traffic scheduling, can also be treated as a gadget in our design. Essentially, our architecture enables us to accommodate a wide variety of OSS technologies—featuring different trade-offs between switching speeds and port-counts—as one of many physical layer gadgets in our design.

Multicast: In order to provide multicast communications over point-to-point optical circuits, an optical link must be established between the sender and each receiver. As illustrated at the top-left of Figure 2, when optical ports are lim-

Table 1: Optical Technologies and Properties

	Switching Time (s)	BW per port	Port Count	Power per port (mW)
Space Switches				
MEMS (2-D & 3-D)	10^{-3}	$>$Tb/s	10^2-10^4	10^2
Mach-Zehnder	10^{-10}	$>$Tb/s	$\sim 10^1$	10^0
PLZT	10^{-8}	$>$Tb/s	$\sim 10^1$	10^2-10^4
SOA	10^{-9}	$>$Tb/s	$\sim 10^2$	10^3
Ring Resonator	10^{-10}	$>$Tb/s	$\sim 10^1$	10^0
Multicasters				
Directional Coupler	n/a	$>$Tb/s	$\sim 10^1$	0
Nonlinear/Parametric	n/a	$\sim$Gb/s	$\sim 10^1$	10^3
λ Muxes/Demuxes				
Fabry-Perot	n/a	$\sim$Gb/s	$\sim 10^1$	0
Diffraction Grating	n/a	$\sim$Gb/s	$\sim 10^1$	0
Mach-Zehnder (AWG)	n/a	$\sim$Gb/s	$\sim 10^2$	0
Composite Devices				
Optical Packet Switch	10^{-9}	$>$Tb/s	10^1-10^3	10^2-10^4
λ-Selective Switch	10^{-6}	$\sim$Gb/s	$\sim 10^2$	10^1

ited, it takes several reconfigurations to serve all the flows, with performance severely bottlenecked due to circuit visit delay [20, 11]. However, by leveraging the inherently low loss and high bandwidth-distance product of photonics, directional couplers—one of the most basic technologies utilized in optical interconnection networks—can be used to achieve physical layer data-rate-agnostic data duplication. By splitting the power of multiwavelength optical signals to multiple ports, the physical layer multicast connectivity depicted at the top of the right-most column of Figure 2 can be realized. Furthermore, by combining gadgets consisting of directional couplers of various sizes and utilizing high-sensitivity optical transceivers [17], trees capable of up to 1,000-way multicasts are realizable.

Incast: Similar to the multicast case, and as illustrated by the middle of the left-most column of Figure 2, servicing an incast over optical circuits would require the limited number of optical links at the receiver's rack to be configured and reconfigured between each sender until the flow is completed. Alternatively, multi-hop routing [6, 21], as illustrated in the second column of Figure 2, can be used to provide incast functionality over a fixed topology constructed from optical circuits to avoid reconfigurations. However, such an approach requires the flows to traverse intermediate nodes, which results in significant load and complexity at these nodes. Moreover, by requiring packet switches for routing, flows are forced to experience numerous optical-to-electrical-to-optical conversions, essentially forfeiting the energy and throughput advantages offered by optics.

Fortunately, the ability to manipulate optical signals in the wavelength domain provides an additional dimension of granularity and control beyond the capabilities of an OSS. To enable physical layer incast, passive wavelength manipulators can be used to separate and recombine WDM channels originating from various racks into a single optical port, while the remaining channels of each sender's WDM signal can be utilized as part of another communication pattern. The static nature of the wavelength manipulator ensures that data streams do not collide in the wavelength domain. At the destination node, the N-channel signal is demultiplexed and individually received by N corresponding receivers. As a result, single-hop, single-configuration incast, as depicted in the middle right of Figure 2, can be attained through more efficient utilization of the bandwidth offered by WDM than can be achieved by space switching alone.

All-to-All-Cast: All-to-all-cast can consist of either multiple unicasts or can be a composite of both unicast and multicast primitives. Like incast, utilizing only point-to-point circuits to support all-to-all-cast will necessitate either costly reconfigurations or inefficient multi-hop routing (bottom of the first and second columns in Figure 2). However, our gadget-based architecture can likewise utilize a combination of the aforementioned technologies to efficiently support all-to-all-cast-type patterns. For example, arrayed waveguide gratings (AWGs), which can split and recombine multiwavelength optical signals on different physical links, can implement a multi-port passive wavelength router to partition what was originally a single WDM unicast link into multiple unicast channels supporting the pattern shown at the bottom of the right-most column of Figure 2. Alternatively, multiple multicast gadgets can be combined with incast aggregation gadgets to construct a super-gadget supporting unicast and multicast composite patterns.

2.2 Advantages of a *-Cast-Enabled Architecture

Advantages over recent circuit-based proposals.

Relying solely on point-to-point optical circuits for traffic off-loading has serious limitations. Obviously, multicast, incast, and all-to-all-cast communications can be translated into unrelated unicast flows and served individually by circuits [11, 20]. However, as illustrated in Figure 2, this requires an optical link to be set up between each single-flow rack pair. When optical ports are limited, it takes several reconfiguration iterations to serve all the flows. These reconfigurations greatly reduce the effectiveness of the optical network by introducing significant circuit visit delay. As proposed by [6] and [21], these delays can be avoided by performing multi-hop routing over static topologies constructed from optical circuits (second column in Figure 2). However, multi-hop routing introduces many of its own inefficiencies. First, by routing traffic across intermediate nodes, significant load and complexity are introduced at the associated racks. In addition, the computational complexity of determining an appropriate topology can further degrade the performance of the system. Most critically, however, a multi-hop approach requires traffic to be converted from optical to electrical and back to optical at each hop, negating the power and bandwidth advantages of the photonic medium.

Advantages from combining optical technologies.

A key advantage of this architecture is that it enables the synergistic utilization of the relative advantages of each photonic technology to account for the shortcomings of its counterparts. For instance, while achieving a high degree of reachability, MEMS suffers from extremely slow switching times. Meanwhile, purely wavelength-routed networks suffer from capacity scalability, while optical broadcasters and nanosecond-scale switches suffer from limited port counts. By allocating a gadget delivering connectivity to more than just two nodes, the MEMS switch can amortize the effect of its slow switching speed. Likewise, the high-radix MEMS switch affords photonic technologies with limited scalability to gain the agility necessary to reach every node in the network. Thus, our design provides a framework for an optically-enhanced hybrid network that can enable the use of these advanced technologies at the scale of the data center. As a result, we can achieve our goal of providing a physical layer that is intrinsically compatible with the rich set of data-intensive *-cast patterns.

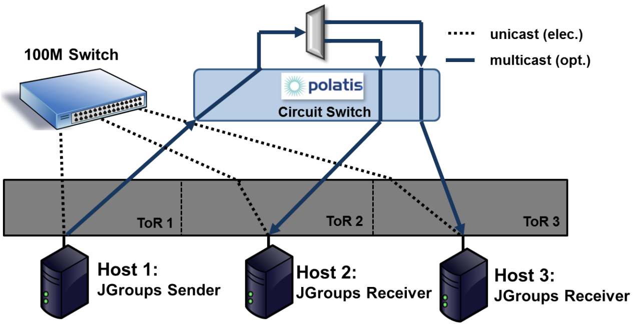

Figure 3: Experimental Testbed

Modularity advantages.

The modularity of this enhanced network architecture is well suited to the incremental nature of data center expansion. As the data center grows, its needs change, or further advancements in photonic technologies are made, additional discrete functionalities can be added as needed by simply attaching or removing optical gadgets. The number of ports occupied by gadgets can be as few or as many as desired, representing a design parameter that can be varied depending on various considerations—including cost and the relative *-cast traffic requirements of a given system. Each gadget represents an incremental addition to the capabilities of the system, minimizing the cost and risk associated with its implementation, while preserving the basic functionality of the optical circuit switch. As the scale and specific traffic requirements of individual applications continue to evolve, we can envision utilizing these photonic gadgets as basic building blocks to construct even more sophisticated optical capabilities on-the-fly. Thus, for the marginal cost and minimal (if any) power increase associated with each photonic gadget, the resulting network architecture can potentially rival the performance of a comparable electronic network in terms of both throughput and traffic agility.

Energy advantages.

To illustrate the relative energy advantage of our architecture over comparable non-blocking electronic networks, we consider the following comparison. A 50-W, 320-port optical space switch can enable any 160 pairs of end points to communicate simultaneously at line rate. Today's 10–40-Gb/s optical transceivers can draw anywhere from 1–3.5 W per port depending on technology (i.e., single- or multi-mode). Considering that adding optical gadgets adds little to no additional energy consumption, at 10 Gb/s this system would consume less than 370 W. Even when considering a worst-case 40-Gb/s single-mode optical system, the same MEMS-based solution would only draw approximately 1.17 kW.

In contrast, 160 pairs of end points communicating at a 10 Gb/s with commodity 48-port 10 Gb/s packet switches would require 21 switches and 320 cables arranged, for example, in a Clos topology drawing over 7 kW. Even by using a single state-of-the-art Arista 7500-class switch—which boasts a power consumption of approximately 10 W per 10-Gb/s port—such a system will still consume at least 3.2 kW.

Finally, the data-rate transparency of optics represents a fundamental advantage over any electronic packet switch, with the photonic fabric potentially supporting data-rates of 100G and beyond without additional power or modification.

Table 2: Throughput Comparison (Mb/s)

	Electronic	Optical
TCP	803	804
UDP	825	831
JGroups	810	832

3. MULTICASTING: A CASE STUDY

Typically implemented at the link or network layer, multicasting can be realized as a straightforward physical layer operation using photonic gadgets [18]. Therefore, we choose it as an initial case study to explore a number of challenges to feasibility facing our proposed architecture from the physical layer optics all the way up through the control plane.

3.1 End-to-End Testbed

The physical layer viability of our architecture has been evaluated in [22]. To demonstrate its end-to-end implementability, we constructed the small-scale multicast-enabled prototype as depicted in Figure 3. A 1-Gb Ethernet switch running Open vSwitch is logically partitioned into three distinct segments, modeling the functionality of three separate ToR switches. Uplink ports on each ToR are connected to both a commodity 100-Mb Ethernet switch and a Polatis optical space switch. At a subset of the OSS's ports, we attached a 1×3 optical splitter, which serves as our optical multicast gadget. Finally, the OSS is configured to map the input and two of the outputs of the multicast gadget to each of our three ToRs.

In order to successfully segregate unicast traffic bound for the electronic packet switch from multicast traffic intended to be delivered optically, we utilize OpenFlow to appropriately demultiplex messages at the ToRs. At the sender's ToR, traffic is forwarded to either the electrical packet switch or optical fabric via a rule matching on a flow's destination IP address. The unidirectional requirement of the optical multicaster is met by inserting OpenFlow rules at the receivers' ToRs ensuring that no traffic is propagated back through the output ports of the multicaster, forcing all traffic to be transmitted through the packet switch.

At each end host, we implement a simple reliable multicast application using JGroups [1] to evaluate the end-to-end performance of our system. Like IP multicast, JGroups utilizes UDP to transmit packets to a specified multicast address, but also detects and retransmits dropped packets to achieve reliability. As a result, a mix of both multicast and unicast traffic is generated and simultaneously switched across both the electronic packet switch and the optical multicast gadget network.

The performance of our optical multicast-enabled system is evaluated in comparison to a baseline configuration consisting of a 1-Gb/s packet switch in place of the optical network. We measure the average thoughput of TCP and UDP unicast using iperf along with JGroups' performance across both systems, obtaining the results summarized in Table 2. These measurements represent single trials over a 10-second interval, corresponding to approximately 1 GB of transmitted data. By successfully providing reliable multicast at the physical layer, we verify that our implementation is not only viable, but can perform as good as or better than an electronic packet-switched solution.

3.2 Simulation

Besides the proof-of-concept testbed experiments, we evaluate the benefits of the multicast-accelerating system through simulations. First, we demonstrate that the control algorithm can efficiently allocate optical circuits to nearly maximize network-wide traffic volume. Then, we show that our system can greatly accelerate traffic flow completion times when coupled to a 4:1 over-subscribed packet-switched network, and that the achieved performance is comparable to that of a non-blocking packet-switched network.

3.2.1 Control Algorithm Analysis

Finding the optimal circuit configuration when the traffic demand is a mix of unicast and multicast can be formulated as a Weighted k-Set Packing Problem. The ToRs form weighted sets, where a set contains the unicast rack pairs or the racks in a multicast group. The weight is the total traffic volume within the set. The maximum set size k is the number of ports on the optical multicaster gadget. We seek a maximum weight sub-collection of disjoint sets and allocate circuits for them. This problem is NP-hard [3], but many approximation algorithms exist [16, 4, 3, 5].

We use the simple greedy approach in [4] as the control algorithm and test its effectiveness through simulations. We take optimality as the metric, which is defined as the proportion of the total traffic volume obtained by the greedy-based control algorithm over that of the optimal solution. The simulations are performed on 200 ToR switches (40 servers per rack) in a multicast-capable hybrid network under a mixture of unicast and multicast traffic. Unicast traffic demands are generated according to traffic measurements in [15] using the methodology described in [19]. We pick concurrent virtual machine (VM) image provisioning as the multicast application and use 700MB—the typical size of a Ubuntu image—as the flow size. In each experiment, we generate up to 5 multicast groups, each involving 3 to 100 racks uniformly chosen across the 200 racks. While the greedy algorithm can handle a large number of multicast groups, the optimal solution solver's high complexity prevents it from computing complicated inputs such as heavy multicast traffic. Thus, we limit the number of multicast groups under 5 to ensure the solver computes in a reasonable amount of time. Experimental results show the algorithm consistently achieves 95% optimality for over 90% of the cases with over 70% reaching the optimum value. We also observe little change in optimality with increasing number of multicast groups, indicating the algorithm's ability to accommodate larger scale multicast.

A key to the agility of the optical network is the time required by the control algorithm to compute an optical path configuration. We run the algorithm on various network scales to evaluate its time cost. The settings are similar to the previous experiment, except that we fix the multicast traffic to 20 groups. The computation runs on one core of an Intel Xeon 2.83GHz processor with 4GB memory. Table 3 shows the average computation time of 100 experiment runs for each topology size. The algorithm is able to compute rapidly, taking less than 200ms for 1000 racks. [11] and [20] use Edmonds' maximum weighted matching algorithm [9] to compute the optical network configuration, which gives optimal solutions within polynomial time. However, Edmonds' algorithm is only suitable for unicast traffic. The approximation algorithm needed to support mixed unicast and multicast traffic demands runs 10× faster than the Edmonds' algorithm, indicating the optical network can reconfigure rapidly even in very large data centers.

Table 3: Average computation time (ms)

# Racks	200	400	600	800	1000
Time	5.9	26.9	65.6	123.4	199.7

3.2.2 Performance Comparison

We simulate a multicast-accelerating network of 60 racks, each with 40 servers, comparing its performance against the oversubscribed packet-switched network and the non-blocking structure of the same scale. We simulate a two-tiered 4:1 oversubscribed network using 60 ToR switches and a single aggregation switch. Each ToR switch has 40 1-Gb/s ports connected to the servers and a 10-Gb/s port connected to the aggregation switch. The multicast-accelerating network is built upon this structure, with each ToR switch having 4 additional optical ports connected to a 300-port space switch, leaving 60 ports for connectivity to 10 6-port multicast gadgets. The non-blocking architecture is similar to the oversubscribed network except that the upstream links from ToR switches are assigned 40 Gb/s of bandwidth to create a non-congested network core. All architectures support IP multicast, but we ignore the complications of state management and packet loss. We seek to stress the network by synthetic communication patterns adapted from [11] and [6]. For unicast traffic, all servers in rack i initially talk to all servers in rack $(i \pm 1)$ mod 60, $(i \pm 2)$ mod 60, and $(i \pm 3)$ mod 60, we then shift the communications to the next rack after each round. The multicast traffic comes from 1 to 10 multicast groups, each consisting of 6 consecutive racks starting from the end of the last group. Initially, the first group begins with the first rack, and we shift the starting point round-by-round. In a multicast group, the first server in the first rack is the sender and all the other servers are recipients. All the unicast and multicast flows are 10MB. To measure performance, we compute the flow completion time based on the flow's max-min fair share bandwidth. For multicast flows, the flow's max-min fair share is determined by the most congested link on the multicast tree. We assume the circuit reconfiguration delay is 10ms and the reconfiguration interval is 1s. The control algorithm computation time is measured at run time.

Figure 4 shows the average multicast and unicast flow completion time over 10 instances of each number of multicast groups obtained by different network architectures. We observe that the proposed architecture accelerates multicast delivery of the oversubscribed network and the no-blocking structure by 5.4× and 1.6× respectively. Since there are sufficient optical links and multicast gadgets, our approach can fully offload the multicast traffic to the optical network. This successfully circumvents the congested network core and thus improves the oversubscribed network considerably. Interestingly, the proposed architecture even outperforms the non-blocking structure. This is because in the proposed architecture, unicast flows have low rates due to congestion in the slow packet switched network. From the multicast sender's point of view, the unicast traffic originating from it cannot fully utilize the sender's link bandwidth. This residual bandwidth becomes fully used by the multicast flow

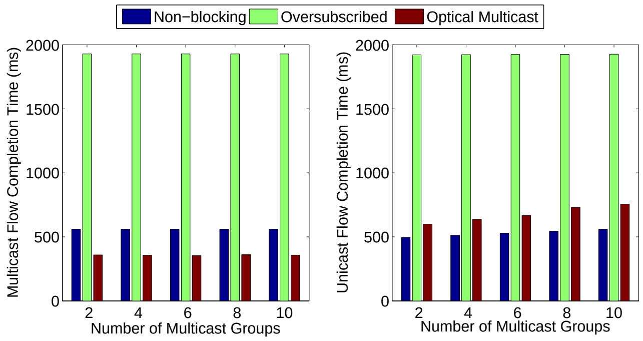

Figure 4: Average multicast (left) and unicast (right) flow completion time of various architectures

when accelerated optically. On the other hand, in the non-blocking network, unicast flows achieve much higher rates, leaving a smaller fair share for the multicast flow. We also observe that the proposed architecture can reduce the unicast flow completion time of the oversubscribed network by 60.8% to 68.8%, though it's still 22.7% to 35% higher than that of the non-blocking structure. With the increasing number of groups, the difference between our architecture and the non-blocking structure enlarges slightly, because the multicast flows occupy the optical ports that could otherwise be used to accelerate unicast flows. Given this minor degradation, our architecture shows great scalability in face of heavy multicast traffic. In all, the proposed architecture has performance comparable to the non-blocking structure. Recall the cost and energy effectiveness, it can provide reasonable performance benefits in practical data centers. We will further explore the system performance under diverse communication patterns and unfavorable conditions in our future work.

4. CONCLUSIONS

Our gadget-based *-cast-accelerating optical network has numerous advantages. Compared to previous proposals that advocate the limited use of point-to-point optical circuits in the data center, our architecture is able to combine the capabilities of a wide range of optical technologies to natively support diverse application communication patterns in the optical domain (i.e. at the physical layer). Our architecture is highly modular—different optical functional units can be connected to the OSS fabric and become instantly usable, making it easy to incrementally deploy such a network and expand it's capabilities. Compared to non-blocking packet-switched networks, our architecture, which simply adds an optical aspect to an existing oversubscribed data center network at the server rack level or above, is far less complex or costly to construct and maintain, has far lower energy requirements, and yet is able to dramatically speed up diverse application communication patterns. The concrete multicast case study we presented shows early promise that this architecture is feasible both in the physical layer and in the control plane. Much research still remains. We are currently exploring the design of control algorithms that can handle all *-cast functionalities, the way in which network and application control planes can interact, and the design issues at the transport layer to effectively leverage combined optical and electrical network resources.

Acknowledgements

This research was sponsored by the NSF under CNS-1018807 and CNS-1162270, by the CIAN NSF ERC (subaward Y503160), by the U.S. Department of Energy under grant DE-SC0005114, by an Alfred P. Sloan Research Fellowship, an IBM Faculty Award, and by Microsoft Corp.

5. REFERENCES

[1] JGroups - A Toolkit for Reliable Multicast Communication, http://www.jgroups.org/.

[2] M. Al-Fares, A. Loukissas, and A. Vahdat. A scalable, commodity data center network architecture. In *SIGCOMM '08*, page 63, New York, New York, USA, 2008.

[3] E. M. Arkin and R. Hassin. On Local Search for Weighted k-Set Packing. *Mathematics of Operations Research*, 23(3):640–648, Aug. 1998.

[4] A. Borodin. CSC2420 - Fall 2010 - Lecture 5, www.cs.toronto.edu/˜bor/2420f10/L5.pdf, 2010.

[5] B. Chandra and M. M. Halldórsson. Greedy Local Improvement and Weighted Set Packing Approximation. *Journal of Algorithms*, 39(2):223–240, May 2001.

[6] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen. OSA: An Optical Switching Architecture for Data Center Networks with Unprecedented Flexibility. In *NSDI '12*, San Joes, CA, USA, April 2012.

[7] M. Chowdhury and I. Stoica. Coflow: A Networking Abstraction for Cluster Applications. In *Hotnets 12*, pages 31–36, Seattle, WA, USA, Oct. 2012.

[8] M. Chowdhury, M. Zaharia, J. Ma, M. I. Jordan, and I. Stoica. Managing Data Transfers in Computer Clusters with Orchestra. In *SIGCOMM '11*, pages 98–109, Toronto, Canada, Aug. 2011.

[9] J. Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17:449–467, Jan. 1965.

[10] N. Farrington, G. Porter, Y. Fainman, G. Papen, and A. Vahdat. Hunting Mice with Microsecond Circuit Switches. In *ACM HotNets*, Redmond, WA, USA, oct 2012.

[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 *SIGCOMM '10*, page 339, New Delhi, India, Aug. 2010.

[12] 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 *SIGCOMM '09*, page 51, New York, New York, USA, 2009.

[13] 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 *SIGCOMM '09*, page 63, New York, New York, USA, 2009.

[14] C. Guo, Y. Xiong, and Y. Zhang. Datacast: A Scalable and Efficient Group Data Delivery Service for Data Centers. In *ACM CoNEXT 2012*, Nice, France, Dec. 2011.

[15] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken. The Nature of Data Center Traffic: Measurements and Analysis. In *IMC '09*, pages 202–208, Chicago, Illinois, USA, Nov. 2009.

[16] V. T. Paschos. A Survey of Approximately Optimal Solutions to Some Covering and Packing Problems. *ACM Computing Surveys*, 29(2):171–209, June 1997.

[17] R. Ramaswami, K. Sivarajan, and G. H. Sasaki. *Optical Networks: A Practical Perspective*. Morgan Kaufmann, 3rd edition, 2009.

[18] G. Rouskas. Optical layer multicast: rationale, building blocks, and challenges. *Network, IEEE*, 17(1):60 – 65, jan/feb 2003.

[19] B. Stephens. *Designing Scalable Networks for Future Large Datacenters*. M.s. thesis, Rice University.

[20] 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 *SIGCOMM '10*, page 327, New Delhi, India, Aug. 2010.

[21] G. Wang, T. S. E. Ng, and A. Shaikh. Programming Your Network at Run-time for Big Data Applications. In *HotSDN '12*, pages 103–108, Helsinki, Finland, Aug. 2012.

[22] H. Wang, C. Chen, K. Sripanidkulchai, S. Sahu, and K. Bergman. Dynamically Reconfigurable Photonic Resources for Optically Connected Data Center Networks. In *Optical Fiber Communication Conference*, page OTu1B.2, 2012.