Accelerating Cast Traffic Delivery in Data Centers Leveraging Physical Layer Optics and SDN

P. Samadi, D. Calhoun, H. Wang, and K. Bergman, *Fellow, IEEE*

***Abstract*—The rising popularity of cloud computing and big data applications has led to massive volumes of rack-to-rack traffic featuring diverse communication patterns denoted as *-cast that combines unicast, multicast, incast and all-to-all cast. Effective support of these traffic patterns in data center networks is still an open problem. We propose a hybrid (optical and electrical) approach that leverages physical layer optics to accelerate traffic delivery for each pattern. Our design consists of an application-driven control plane compatible with software defined networking (SDN) to dynamically configure the optics. We present the network architecture and control plane design and results on the multicast case.**

***Index Terms*—Optical Interconnections, Optical Switches, Network Topology**

I. Introduction

As cloud computing and big data applications continue to grow in scope and scale, they introduce increasing inter-rack traffic demands within the data center. Moreover, the heterogeneity of these applications results in highly diverse and transitory communication patterns. Not surprisingly, such applications [1-3] place considerable burden on the underlying interconnection network, which becomes a major bottleneck.

In [4, 5], it was proposed to alleviate this burden by offloading high-volume traffic to optical circuit-switched network composed of MEMS-based optical space switches for stand-alone point-to-point bulk transfers. These switches can support significantly higher bandwidth than networks composed purely of electronic packet switches. However, by virtue of the solely point-to-point nature and millisecond-scale switching time of optical space switches, these optical networks may become ineffective in the presence of other communication patterns that include transmitter and receiver groups or point-to-point or point-to-multipoint iterative data delivery in the millisecond range [2, 3]. Efficient support of such richer patterns necessitates a more innovative and disruptive approach.

Fortunately, MEMS-based optical space switches represent just one class of a wide range of photonic devices. Other devices such as optical splitters, wavelength selective switches (WSS), and arrayed waveguide gratings (AWG) provide capabilities ranging from passive wavelength routing to broadband nanosecond scale switching. These capabilities can potentially be utilized to intrinsically support a rich set of communication patterns at the physical layer and to construct highly efficient optical data center networks. These communication patterns can be classified to four categories based on the *-cast type [6]: **unicast,** data transmission from a single sender to a single receiver, **multicast,** from a single sender to multiple receivers, **incast,** from multiple senders to a single receiver, and **all-to-all-cast** data transmission among a set of nodes in high performance computing applications.

In this work, we present a hybrid optical network leveraging a library of function-specific photonic devices to individually accelerate and support *-cast-based traffic patterns. Each module will be integrated into a reconfigurable optical fabric such that they can be dynamically connected to racks across the data center. Using this fabric, individual or specialized combinations of these modules are allocated to satisfy communication demands as they arise in the network. Practically speaking, such an optical fabric can be realized by a high-radix optical space switch. However, instead of serving as a traffic carrier, the space switch in our architecture forms a connectivity fabric that routes traffic to and from various modules, thus enabling the flexible and dynamic run-time configuration of photonic devices to support complex traffic patterns.

The allocation of the photonic devices is dictated by requirements defined by the higher layers. These requirements can either be imposed by the application/service through explicit requests to the control plane or implicitly informed through the measurement of various performance metrics (e.g., flow counters, queue occupancies, etc.) from the data link layer up through the transport layer. Given this information from the higher layers, a network control algorithm will dynamically configure the underlying optical network to provide capabilities where they are needed most.

II. Network Architecture

In this section, we discuss the dynamic reconfigurable network architecture and the proposed photonic devices to support different traffic patterns. A schematic representation of the proposed network architecture is depicted in Fig. 1(a). Our design consists of a hybrid aggregation layer of an optical circuit switching network and an electrical packet switching network that connects the Top-of-Rack (ToR) switches. The

P. Samadi, D. Calhoun, H. Wang and K. Bergman are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: ps2805@columbia.edu, dmc2202@columbia.edu, howard@ee.columbia.edu, bergman@ee.columbia.edu).

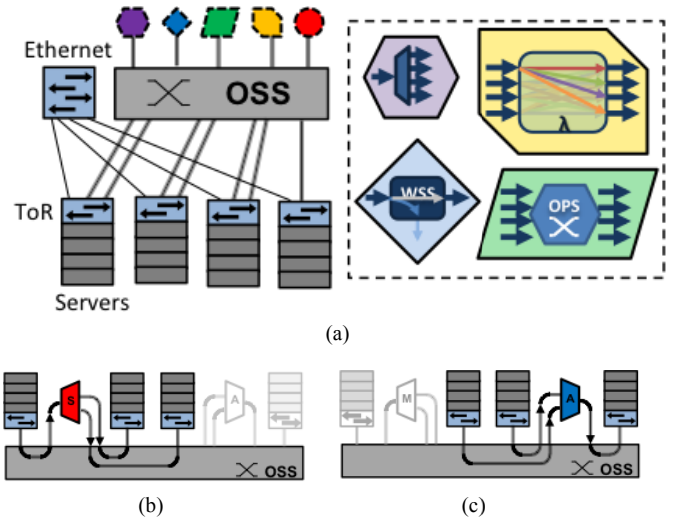


Fig. 1. (a) Data center network architecture in hybrid architecture (electrical packet switching (Ethernet) and optical circuit switching) featuring dynamic photonic devices (possible multi-colored units shown: an optical splitter, Wavelength Selective Switch (WSS), Arrayed Waveguide Gratings (AWG) or an optical sub-system such as Optical Packet Switching (OPS) network). The Optical Space Switch (OSS) serves as a connectivity substrate enabling support for patterns such as (b) multicast (S: Splitter) and (c) incast (A: AWG).

main contribution of this architecture comparing to the other proposed hybrid architectures [3, 4] is that the optical circuit switches provide a dynamic substrate on which traffic is routed to and from specialized photonic devices to accelerate the traffic delivery based on the cast type rather than acting as a point-to-point traffic carrier.

Multicast: Although multicast traffic delivery is not implemented in the network layer in current data centers, we have found various application/services in data centers that require multicast traffic delivery. Data dissemination such as virtual machine provisioning [7] or in-cluster software updating [8] in data center is one category of these applications. Data replication in distributed file systems [9] and parallel database relational join operations [10] is the other category. Moreover, the broadcast phase of iterative machine learning tasks requires multicast type traffic delivery [11]. By leveraging the inherently low loss and high bandwidth-distance product of photonics, one of the most basic technologies utilized in optical interconnection networks can be used to achieve physical layer data-rate-agnostic data duplication. Combination of directional couplers of various sizes as modules connected to the OSS form trees realizing n-way multicast through the passive splitting of broadband optical signals from a single physical port to a multitude of outputs [6]. Fig. 1(b) demonstrates 1:2 multicast leveraging an optical splitter connecting three ToR switches.

Unicast: Point-to-point unicast is straightforwardly supported in photonics using optical space switches. As proposed in [4, 5] point-to-point connectivity is realized by configuration of the MEMS-based OSS in our architecture. However these types of switches provide low speed reconfiguration time of 10s of milliseconds. There exist a wide range of optical space switch technologies, with advantages and tradeoffs typically centered on port-count vs. switching speed. For example in Mordia [12], WSS with lower port counts but faster switching speed is leveraged to provide microsecond switching. That can be used as a photonic device in our system to aggregate shorter messages from small set of racks and forward them to singular optical circuit that can be allocated to any other rack within the network.

Incast: This traffic pattern occurs when data is aggregated between servers in a many-to-one manner. The motivation for investigating this traffic pattern is found in a MapReduce reducer [13], which requires collection of intermediate results from all the mapper for the reduce-phase computation. Incast can be supported leveraging wavelength manipulation of optical signals that provides an additional dimension of granularity and control. Optical modules consisting of passive wavelength multiplexers and demultiplexers can be used to route and aggregate wavelengths from various sources to a single destination. Efficient utilization of the bandwidth offered by wavelength division multiplexing (WDM) to achieve zero-energy, single-hop, and single-configuration incast traffic delivery is still required. An alternative approach is leveraging high-speed optical space switches with sufficiently low switching times and time-division multiplexing.

All-to-all-cast: All-to-all cast is traffic delivery among a set of nodes. This pattern is common in MapReduce shuffle [13] where mappers and reducers concurrently exchange data. Similarly to incast, all-to-all-cast consist of multiple unicasts or a composite of both the unicast and multicast primitives. As such, our dynamic architecture can similarly utilize a combination of the aforementioned technologies to efficiently support all-to-all-cast-type patterns. For example, arrayed waveguide gratings (AWGs) implement a multiport passive wavelength router to support a multiple unicast-like pattern. Alternatively, combination of multiple multicast modules and incast aggregators construct a super-module supporting unicast and multicast composite patterns.

Leveraging optical networks for *-cast based data delivery has fundamental advantages in both energy and capacity compared to electronic networks. For example in multicasting, by duplicating data in the links using passive optical splitters, it avoids the high cost and complexity associated with the need for intermediate packet based multicast capable core switches. The inherent packet and data rate transparency of photonics also obviates the need for costly conversions between the electronic and optical domains. This design decouples the power consumption of the photonic fabric from data rate, thus providing built-in support for speeds beyond 40 Gb/s without any modification to the fabric.

III. Control Plane

In designing the control plane, either the application explicitly requests for an optical device or based on the network measurements, the controller is implicitly notified to use the optical network. In [5], the traffic demand is estimated by observing end-host buffer occupancy at run- time. In [6], the application/service requests photonic devices based on the traffic pattern and demand. In this approach, a control plane is

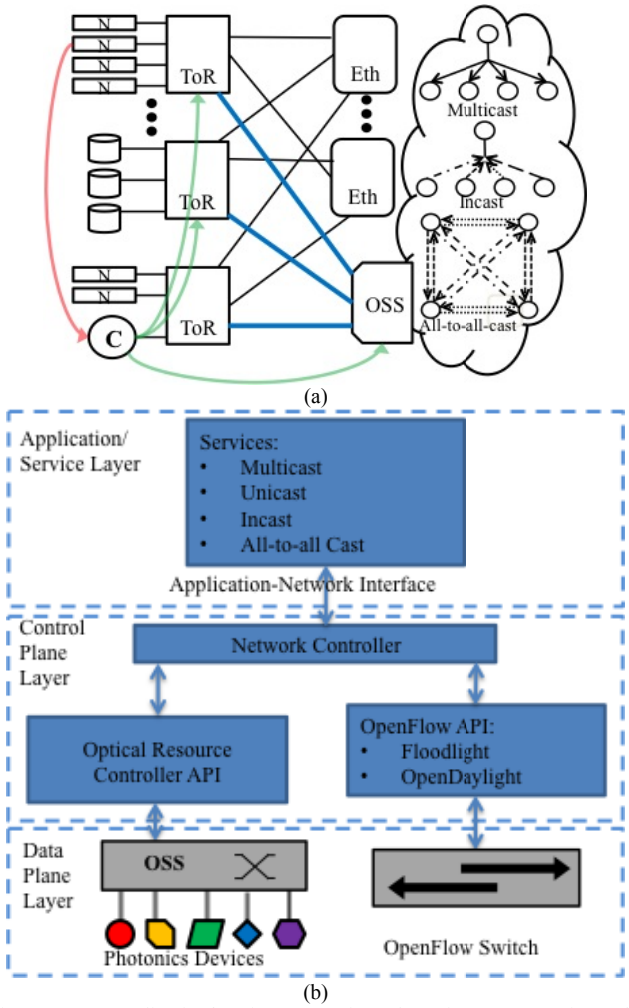


Fig. 2. (a) Application/service requesting photonic resources from the controller based on the traffic -*cast pattern, (b) Layers diagram composed of the Application/Service, Control Plane, and Data Plane.

required to dynamically allocate the photonic devices and manage the traffic between the optical circuit switched network and the electrical packet switched network. Fig. 2(a) shows the role of the control plane in our network architecture. Management of optical devices by the control plane requires investigating methods of abstracting the optical device functionalities to the high layers for flow control and arbitration of physical layer services.

Fig. 2(b) shows the layer block diagram of our network layers that consists of the Application/Service, the Control Plane, and the Data Plane layers. Our approach in designing the control plane is to provide application-defined networking; the application/service directly interacts with the network controller via requests/calls for network services. Network services are arbitrated by concurrent control of an optical resource controller and OpenFlow-enabled electrical switch, which provides both optical resources and network paths, simultaneously. This network controller interface removes any knowledge of actual physical layer components—whether they are optical or electrical—allowing application/service demands to be seamlessly allocated to specific photonic devices based on the traffic pattern (unicast, multicast, incast, all-to-all-cast). The status of traffic-specific resources are reported by the network controller upon request; if a particular connection is ready and available, the network controller informs the application/service to start data transmission.

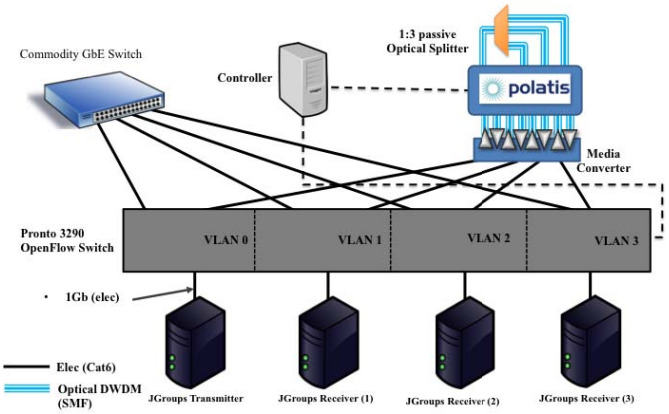
Network Controller: A key part of the Control Plane is the algorithm to maximize the network-wide throughput and to compute the corresponding circuit allocations. For the optical multicast-enabled network, finding the optimal circuit configuration when the traffic demand is a mix of unicast and multicast can be formulated as a Weighted k-Set Packing problem [14]. While NP-hard, but many solutions based on approximation algorithms exist [14-16]. Our initial studies explore one particular approximation [16] in order to analyze the performance of the greedy algorithm [6], shedding light on the feasibility of our architecture with respect to the control plane.

Additionally, APIs are required to first manage the flow table of the ToR switches and also the optical hardware resources in the physical layer. Currently we manage the OpenFlow switch by the OpenVSwitch [17] using the command-line, however in the development of the control plane, we are using the open source APIs such as Floodlight [18] and OpenDaylight [19] to control the ToR switches through the control plane. For the optical hardware resources, we have developed a Java-based API to manage the connectivity of the OSS fabric via the TL1 protocol.

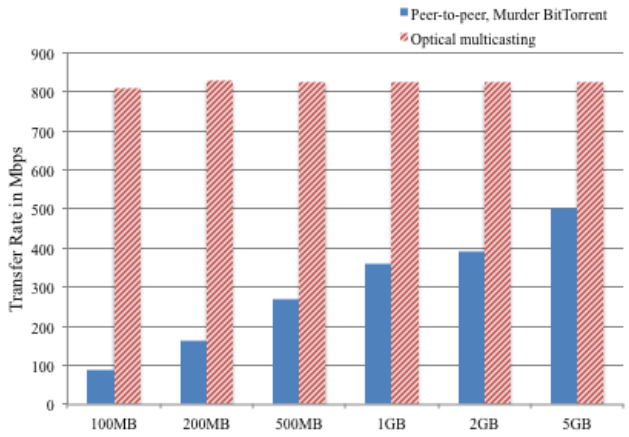
IV. Results and Discussion

Fig. 3(a) shows our datacenter test-bed architecture to evaluate an end-to-end feasibility of multicast traffic delivery. The test-bed consists of 4-nodes each with a Gigabit Ethernet network interface card (NIC), connected to a Pronto OpenFlow-enabled Gigabit Ethernet switch. The switch is partitioned into four logical ToRs, with Gigabit Ethernet uplink and downlink ports. The 4 ToRs are aggregated by a commodity 8-port Gigabit Ethernet switch and also a Polatis piezoelectric beam-steering optical space switch which acts as a connectivity substrate of photonic devices.

The photonic device used to enable physical layer multicasting was a 1:3 passive optical splitter that was attached to the subset of the optical switch's ports. The end-to-end performance of our architecture was evaluated through the implementation of a reliable multicast application at the end hosts using JGroups [20], a toolkit enabling reliable multicast communication. In our experiment, one node was configured as the sender and three nodes configured to join the group as receiver. By appropriately configuring the optical space switch and setting the necessary OpenFlow rules at the ToR switches, all the multicast traffic generated by the sender was sent to the input port of the 1:3 optical splitter and the three output ports were connected to the receiver nodes. Any back-propagating traffic originated from the receiver nodes to the sender were routed through the GbE network isolated from the optical splitter. Through continuous runs of the application, we measured JGroups' throughput performance across two cases: over just an electronic packet switch and through our hybrid system. We observed the effective saturation of the sender's



(a)



(b)

Fig. 3. (a) End-to-end experimental test-bed to evaluate multicast traffic delivery, (b) comparison of transfer rate for different data size between our approach and Murder

output interface in both cases, demonstrating the viability and functionality of our architecture.

In order to evaluate the performance of our architecture in multicast traffic delivery, we looked for existing efficient group data delivery methods for large data transmission in data centers. We found out in current data center architectures, multicast traffic is transmitted either in sequence of unicasts or more advanced peer-to-peer software solutions such as BitTorrent [21]. Sequence of unicast transmissions is definitely much less efficient than optical multicasting thus we compared our system with BitTorrent as a peer-to-peer software solution. There are several implementations of BitTorrent with various performances; Twitter Murder [8] is one of the most efficient implementations, using BitTorrent to distribute files to a large amount of servers within a production environment. The efficiency improvements in Murder compared to regular BitTorrent are: 1) Shorter Timeouts, 2) Encryption disabled, 3) Distributed Hash Table (DHT) disabled, 4) Upload from memory, and 5) Universal Plug and Play (UPnP) disabled. We compared multicast data transmission of different file sizes among four nodes by—first optical multicasting using JGroups protocol and second peer-to-peer using Murder—and measured the transfer rate. As shown in Fig. 3(b), the transfer rate of the optical multicasting using JGroups protocol is above 800 Mbps in all the cases. This represents over 80% link capacity usage (GbE NIC on the hosts and ToR switch ports). Murder seems to perform more efficiently with larger file sizes with the maximum of 50% of link capacity (500 Mbps) for a 5GB file size. In our experiments, we did not extend the file size above 5GB since we believe group data delivery of larger data sizes is far from reality in data centers applications [7-11].

Our end-to-end experiment was evaluated without the implementation of the control plane. The optical space switch and the OpenFlow switch were configured manually for the four-node multicast experiment. We have started the development of the control plane based on the description on section II. We believe our design of the control plane can be viewed as the first step towards designing the Software Defined Networks (SDNs) abstractions for the optical networks in datacenters.

There are still many open questions for investigation in this research work: **Interaction of the optical and electrical network:** As the optical network represents a limited resource, it may not be able to serve all heavy-loaded traffic demands in the system. Although the electrical network is generally slow, some schemes can be used to help with big data transmission. For example for multicast traffic delivery, BitTorrent as in Murder can act as a complement of the optical multicast solution. **Physical Layer Abstraction:** In conventional packet switches and routers, switching and routing of the various traffic patterns are typically delivered at L2 and L3. However, by enabling physical layer data routing, we break the assumptions made by traditional layered network protocols. To address these conflicts, modifications or abstractions to these protocols are unavoidable in order to define well-behaved interactions between the existing electronic network infrastructure and new advanced photonic technologies. **Control Algorithm:** In order to make intelligent scheduling decisions, the controller has to obtain information about the correlated flows. [4] and [5] both rely on application-agnostic demand estimation and traffic inference. Meanwhile, [22] proposed the direct interaction of the network controller with individual application controllers, which keep track of each application's task placement and traffic volume. The former precludes any application modifications, but can produce suboptimal results [23]. The latter will require the explicit specification of communication requirements from the application, but can yield more globally optimal solutions. For example, a series of light-weighted incast flows may be more favorable than a bulk unicast transfer, because the incast node is collecting data for a subsequent computation. Exploration of both methodologies in the context of our optically enhanced architecture to determine whether one, the other, or a combination of both techniques will yield the optimal results.

V. Conclusion

In this work we presented a hybrid network architecture to accelerate cast-based traffic delivery in data centers. Our design is based on the physical layer photonic technology and the SDN concept. This design benefits from leveraging the OSS as a platform to connect passive and active optical

devices. Modularity, data rate transparency and energy efficiency are the main advantages of our method. Moreover, management of the network layer based on the application/service layer requirements can be viewed as the first step towards designing Software Defined Networks (SDNs) for the optical (physical) layer. It has the potential to significantly enhance the performance of optical networks and to transform the way control and management is performed. We are now finalizing our first implementation of the control plane; however there are many open questions yet to be answered.

VI. ACKNOWLEDGEMENTS:

We would like to thank T. S. Eugene Ng. and Yiting Xia at Rice University for their ideas on the design of the control plane. We would also like to thank Polatis for providing the optical space switch.

REFERENCES

[1] M. Chowdhury and I. Stoica, "Coflow: A networking abstraction for cluster applications," in *Proc. ACM Hotnets'12*, Oct. 2012.

[2] C. Guo, Y. Xiong, and Y. Zhang, "Datacast: A scalable and efficient group data delivery service for data centers," in *Proc. ACM CoNEXT'12*, Dec. 2011.

[3] M. Chowdhury, M. Zaharia, J. Ma, M. I. Jordan, and I. Stoica, "Managing data transfers in computer clusters with orchestra," in *Proc. ACM SIGCOMM '11*, Aug. 2011.

[4] 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'10*, Aug. 2010.

[5] 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'10*, Aug. 2010.

[6] H. Wang, X. Xia, K. Bergman, T. S. Eugene Ng, S. Sahu, and K. Sripanidkulchai, "Rethinking the physical layer of data center networks of the next decade: Using optics to enable efficient *-cast connectivity," *ACM SIGCOMM Computer Commun. Review*, vol. 43, no. 3, July 2013.

[7] D. Li, M. Xu, M. C. Zhao, C. Guo, and Y. Wu, "Reliable data center multicast," in *Prof. IEEE INFOCOM*, Apr. 2011.

[8] Twitter, "Murder: Fast data center code deployment using bittorrent, "http://engineering.twitter.com/2010/07/murder-fast-datacenter-code-deploys.html, 2011, network Working Group.

[9] M. Wiesmann, F. Pedone, A. Schiper, B. Kemme, and G. Alonso, "Data replication techniques: a three parameter classification," in *Proc. SRDS*, 2000.

[10] W. Mach and E. Schikuta, "Parallel database join operations in heterogeneous grids," in *Proc. PD- CAT*, 2007.

[11] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time url spam filtering service," in *Prof. IEEE Symp. Security and Privacy (SP)*, 2011.

[12] G. Porter, R. Strong, N. Farrington, A. Forencich, P. Chen-sun, T. Rosing, Y. Fainman, G/ Papen, and A. Vahdat, "Integrating Microsecond Circuit Switching Into the Data Center", SIGCOMM'13, Aug. 12-16, 2013, Hing Kong, China.

[13] J. Dean, and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Google Inc., *Communications Of The ACM*, Vol. 51, No. 1, January 2008.

[14] E. M. Arkin and R. Hassin, "On local search for weighted k-set packing," *Math. Oper. Res.*, vol. 23, no. 3, pp. 640–648, Aug. 1998.

[15] V.T.Paschos,"Asurveyofapproximatelyoptimalsolutionstosomecoveringandpackingproblems," *ACM Comput. Surv.*, vol. 29, no. 2, pp. 171–209, Jun. 1997.

[16] A. Borodin, "CSC2420 - Fall 2010 - Lecture 5," pp. 1–5, 2010. [Online], Available: www.cs.toronto.edu/~bor/2420f10/L5.pdf

[17] An Open Virtual Swithc, http://openvswitch.org/.

[18] Project Floodlight, Open Source Software for Building Sofware-Defined Networks, http://www.projectfloodlight.org/floodlight/.

[19] OpenDaylight, Linux Foundation Collaborative Projects, http://www.opendaylight.org/.

[20] "JGroups - a toolkit for reliable multicast communication." [Online]. Available: http: //www.jgroups.org/.

[21] BitTorrent, http://www.bittorrent.com/.

[22] G. Wang, T. S. E. Ng, and A. Shaikh, "Programming your network at run-time for big data applications," in *Proc. ACM HotSDN'12*, Aug. 2012.

[23] H. H. Bazzaz, M. Tewari, G. Wang, G. Porter, T. S. E. Ng, D. G. Andersen, M. Kaminsky, M. A. Kozuch, and A. Vahdat, "Switching the optical divide: Fundamental challenges for hybrid electrical optical datacenter networks," in *Proc. ACM SOCC'11*, Oct. 2011.