Accelerating Incast and Multicast Traffic Delivery for Data-intensive Applications using Physical Layer Optics

Payman Samadi, Varun Gupta, Berk Birand, Howard Wang, Gil Zussman, Keren Bergman
Department of Electrical Engineering, Columbia University, New York, NY, 10027
[ps2805@, vg2297@, berk@ee, howard@ee., gil@ee., bergman@ee.]columbia.edu*

ABSTRACT

We present a control plane architecture to accelerate multicast and incast traffic delivery for data-intensive applications in cluster-computing interconnection networks. The architecture is experimentally examined by enabling physical layer optical multicasting on-demand for the application layer to achieve non-blocking performance.

Categories and Subject Descriptors: C.2.1 [Computer-Communication Networks]: Network Architecture and Design Network communications, Circuit-switching networks
Keywords: Hybrid Data Center Networks; Optics; Incast; Multicast.

1. INTRODUCTION

With the enormous increase in the generation and complexity of Big Data, new opportunities and challenges related to its storage and processing on cluster-computing platforms arise. The interconnection networks of such platforms are generally over-subscribed, due to the switching cost and cabling complexity. Designing these interconnection networks based on the traffic patterns of the applications running on clusters can improve the overall performance of the system. In general, these applications use distributed file systems for storage and MapReduce type of algorithms for data processing. In analyzing these applications and the interconnection network, we found that large flows with traffic patterns that include multiple nodes result in heavy network congestion.

For example, in GoogleFileSystem, when storing and requesting data, blocks of 64MB are sent from one node to multiple nodes (point-to-multipoint: Multicast) or one node receives data from multiple nodes (many-to-one: Incast). Similar processes exist in HadoopFileSystem and in the shuffle stage of the MapReduce algorithm. Additionally, in other data center applications such as virtual machine provisioning or in-cluster software updating, there is a multicast of 300-800MB among 100-1000 nodes. Moreover, in parallel database join operation, there is multicast of several 100MB. Furthermore, in the Broadcast phase of Orchestra controlled by Spark [1], 300MB is multicast on 100 iterations. In current platforms, these patterns are managed either by sequence of unicast transmissions or software solutions such as peer-to-peer methods. These methods are naturally inefficient since multiple copies of the same data are transmitted on the network. For instance, the Orchestra system transmits 12 copies of the same data.

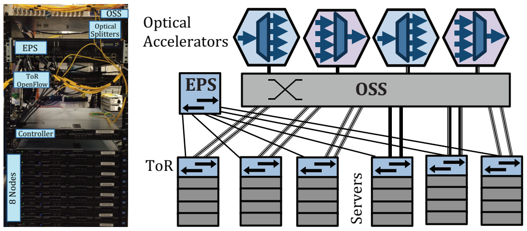


Figure 1: Optical accelerators in the hybrid arch.

Traditionally, electrical packet switching (EPS) networks in Fat-tree or multi-tier architecture are used for the interconnection networks of cluster-computing platforms. Recently, hybrid network architecture that combine EPS and optical circuit switching have been proposed to offload larger flows to the optical network [2, 3]. The switching of the optical network is implemented by optical space switches (OSS). The OSS, that performs point-to-point circuit switching can also serve as a substrate to connect optical modules as accelerators. For example, passive optical splitters can be used to provide a significantly more efficient multicasting [4]. In this method, data is optically replicated in the physical layer and transmitted by setting up a multicast tree between the source and the destinations. This operation sends only one copy of data and all the nodes receive the data simultaneously at the line-rate (non-blocking performance). Incast traffic delivery can also be optically accelerated using passive optical combiners and time-sharing the circuit between the senders. In both cases, the OSS that has slow switching time (25ms) is reconfigured once and this results in lower latency compared to the point-to-point architectures. These accelerators can potentially offload heavy data from the over-subscribed packet switching network and make them smaller and cheaper. The passive nature of these accelerators, can also reduce the networking power consumption. The main challenge in utilizing optical accelerator is the integration of

*This work was supported in part by CIAN NSF ERC under grant EEC-0812072.

SIGCOMM'14, August 17–22, 2014, Chicago, IL, USA.
ACM 978-1-4503-2836-4/14/08.
http://dx.doi.org/10.1145/2619239.2631436.

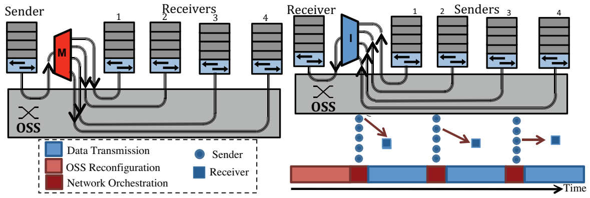
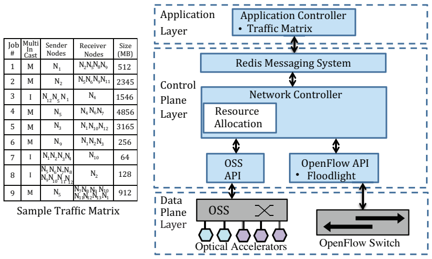


Figure 2: Left: Optical multicasting by passive optical splitter, Right: Optical incast using passive optical combiner and the time-sharing orchestration.



Job #	Multi In Cast	Sender Nodes	Receiver Nodes	Size (MB)
1	M	N_1	$N_2N_5N_8N_9$	512
2	M	N_2	$N_5N_6N_9N_{11}$	2345
3	I	$N_{12}N_5N_1$	N_4	1546
4	M	N_5	$N_4N_6N_7$	4856
5	M	N_3	$N_1N_{10}N_{12}$	3165
6	M	N_9	$N_1N_2N_3$	256
7	I	$N_1N_2N_3N_4$	N_{10}	64
8	I	$N_5N_6N_7N_8$ $N_9N_{10}N_{11}N_{12}$	N_2	128
9	M	N_5	$N_7N_8N_9N_{10}$ $N_{11}N_{12}N_{13}N_1$	912

Sample Traffic Matrix

Figure 3: Network Orchestration Architecture.

optics with current network architectures due to the complexities in configuring optical devices and the circuit-based nature of optics. Cross-layer architectures can potentially overcome these complexities and provide optical functionalities more seamlessly to the application layer.

In this work, we present a control plane architecture that accelerates multicast and incast data delivery using passive optical modules. The architecture is experimentally examined on our 10G hybrid cluster-computing testbed with a demonstration of physical layer optical multicasting. The implemented control plane employs a messaging system, a resource allocation algorithm, and APIs to control the optical and electrical switching network.

2. IMPLEMENTATION AND RESULT

Fig. 1 shows the hybrid network architecture with optical accelerators. Top-of-Rack (ToR) switches are aggregated by an electrical packet switching network and an optical circuit switching network. An OSS is used to provide switching in the optical network and also as a substrate to connect optical accelerators. Physical layer optical multicast can be addressed using passive optical splitters that transparently duplicate the optical signal. Fig. 2 (left) shows the hardware configuration to generate multiple copies of data at the line rate. The incast traffic can be addressed by passive optical combiner with an orchestration system running on the packet switching network (Fig. 2 (right)). Optical splitters and combiners are data rate transparent, passive, and commercially available in high port counts. Also the OSS can be configured in a way to cascade multiple modules. Fig. 3 demonstrates the network orchestration architecture consisting of the Application, Control, and Data Plane layers. The Control Plane has a central network controller that manages the network and includes a resource allocation algorithm. The network controller communicates with the Data Plane layer via southbound APIs including Floodlight for the OpenFlow switches and OSS API, which is a python-based API developed in-house. For the northbound API of the Control Plane, we have developed a pub/sub messaging system using Redis. Byte size messages are transmitted on the EPS network between the nodes and the central controllers. We observed the latency by sending 200 messages of 20 bytes among nodes and measured average latency of 300 μs.

Table 1: Completion time of 50 multicast jobs (seconds) Proposed optical Multicast and non-blocking Internet Protocol Multicast

	Group Size	
	7	4
Proposed	226.179	187.348
Non-blocking	218.517	191.680

For optical multicasting, the general algorithm of the network controller is as followings: The application controller submits the traffic matrix of the jobs that require optical resources to the network controller via the Redis messaging system. The traffic matrix includes the source, destinations, and the size of the multicast job. The resource allocation algorithm tries to schedule the optical splitters across the multicast requests by maximizing the obtained throughput. We model this problem as an Integer Program (IP) and solve using the GLPK solver that implements a branch-and-bound method. Using the solution of IP, the network controller generates the network configuration for the next job. Then, the ToR switches are configured using Floodlight, and the OSS connections are applied using our API. For optical multicasting, the network controller first notifies the receivers using the Redis messaging system. Then, the senders are notified to begin the transmission. Each receiver sends a message to the controller, notifying the completion of the job. Once the controller receives these messages from all the receivers for a given job, it updates the traffic matrix and reruns the algorithm. We evaluated the performance of our architecture for multicasting in comparison with Internet Protocol (IP) multicast over a non-blocking EPS network on our 8-node hybrid cluster-computing test-bed (Fig. 1). Two sets of 50 multicast jobs (500MB-5GB) with the maximum group size of 4 and 7 were generated. The former involves maximum half of the nodes and one splitter. The latter can involve the whole network and requires cascading of splitters. Table 1 shows the completion time in seconds. Our proposed optical multicast performs similar to the ideal fully non-blocking network. This means that our architecture provides line rate non-blocking multicast between the nodes that is not possible in practical interconnection networks. Additionally, it can potentially decongest the over-subscribed EPS network by offloading the multi/in-cast traffic to the optical network.

3. REFERENCES

[1] M. Chowdhury, M. Zaharia, J. Ma, M. I. Jordan, and I. Stoica. Managing data transfers in computer clusters with orchestra. In *ACM SIGCOMM*, 2011.

[2] 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 *ACM SIGCOMM*, 2010.

[3] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. E. Ng, M. Kozuch, and M. Ryan. c-through: Part-time optics in data centers. In *ACM SIGCOMM*, 2010.

[4] H. Wang, Y. Xia, K. Bergman, T. E. 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. *SIGCOMM Com. Comm. Rev.*, July 2013.