An Optical Programmable Network Architecture Supporting Iterative Multicast for Data-intensive Applications

P. Samadi[1], H. Wang[1], D. Calhoun[1], Y. Xia[2], K. Sripanidkulchai[3], T. S. Eugene Ng[2], K. Bergman[1]
[1]*Department of Electrical Engineering, Columbia University, 500 West 120th Street, New York, New York 10027*
[2]*Department of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005*
[3]*NECTEC, 112 Phahon Yothin Road, Klong Luang, Pathumthani 12120, Thailand*
E-mail: ps2805@columbia.edu

Abstract—We present an optical programmable network architecture to enable agile and efficient iterative multicasting for cluster computing framework. Support for multiple multicast groups and dynamic group reassignment are experimentally demonstrated.

I. Introduction:

The continuous growth in the scale and scope of data-intensive cluster computing applications has forged new opportunities and challenges in designing data center networks. Hybrid (electronic packet switching and optical circuit switching along with an intelligent control system) architectures can provide substantial bandwidth for inter-rack communications at low cost, energy consumption, and cabling complexity [1,2]. In previous work, we presented the idea of programmable network architecture featuring advanced photonic functionalities to accelerate bandwidth-intensive traffic patterns in data center networks [3]. We addressed multicast by adding passive optical splitters to the optical space switch (OSS) substrate and allowing the network controller to dynamically reconfigure connectivity. Due to the long switching latency of the OSS (tens of ms) and the disjoint multicast trees (single wavelength), our system targeted sparse long-lasting multicast traffic, such as in-cluster software updates, distributed file system data replication, and OS provisioning to many virtual machines.

However, many big data analysis tasks also feature multicast communication. Monarch [4], which is used for identifying spam links on Twitter converges after 100 iterations and performs a large multicast (300MB) per iteration—30% of the iteration time on a 30-node cluster. Similarly, Netflix iteratively multicasts roughly 385MB of data (45% of the iteration time at 60 nodes) in a collaborative filtering (CF) job to predict users' rating of unwatched movies. These iterative cluster computing tasks require multicast among a set of nodes, possibly slightly changing sources and destinations from iteration to iteration. In addition, since multicast is a typical communication pattern in cluster computing frameworks, there may be a large number of multicast groups coexisting in the network, calling for efficient usage of the optical resources to build denser tree structures. The state of the art technology to multicast in cluster computing platform relies on peer-to-peer solutions that cannot use the bandwidth efficiently. For instance, even the most recently proposed Orchestra system [5] transmits 12 copies of the same data, which is much less efficient than our optical multicast network [3] that always sends one copy (Link stress of 12 vs. 1). Although our previous optical multicast proposal offers efficient ultrafast transmission, it suffers from slow switching of the OSS (300MB at 40Gbps: 60ms, same order of OSS switching speed) and inflexible tree structures as a disjoint tree is needed for each multicast group.

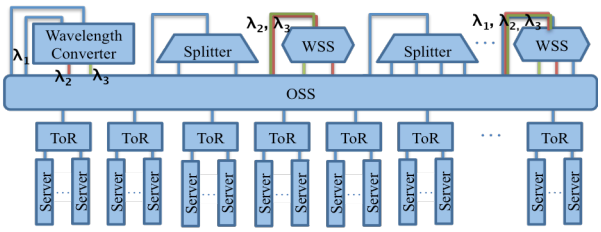

Figure 1: Programmable optical network architecture with optical components and subsystems to generate multicast groups.

In this paper, we present a programmable optical system design to support iterative multicast traffic at a fine timescale, leveraging wavelength conversion to support multiple multicast groups, wavelength selective switches (WSS) to achieve fast reconfiguration, and Software Defined Networking (SDN) to intelligently place WSSs and splitters to form efficient and flexible data distribution trees.

II. Optical Iterative Multicasting:

Fig. 1 and 2(a) shows our proposed architecture that can effectively support the stringent communication requirements of iterative multicast by extending our programmable network architecture [3,6] with wavelength-selective capabilities and wavelength conversion. By utilizing WDM and replacing a subset of the tree's internal nodes with WSS, a single physical tree topology can support multiple multicast groups. Given a k -ary multicast tree of height H —which supports $N = k^{H+1}$ receivers—we can efficiently define wavelength-specific multicast groups at a subset of the receivers through the intelligent placement of WSSs at heights $h_i < H$, yielding multicast groups of size $n_i = k^{H+1-h_i}$ using a single WSS. Moreover, the switching capabilities and faster switching performance of WSSs provide the additional degree of freedom and performance necessary to effectively support the dynamism of iterative multicasts. A centralized controller

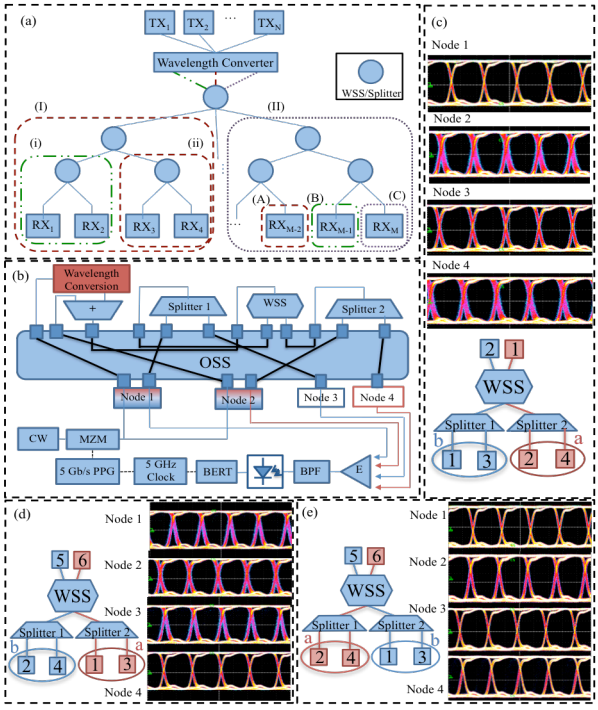

Figure 2: Demonstration of agile and efficient reconfiguration in multicasting groups in a tree architecture with N transmitters and M receivers, (b) Experimental setup of programmable optical network architecture supporting multiple multicast groups and fast switching, (c), Multicast data delivery on the same tree, (d) and (e) Fine time scale multicast group assignment.

configures the OSS and WSS to build the optical network and multicast groups.

As an example, in an architecture with N transmitter and M receiver nodes (transmitters and receivers can be Top-of-Rack (ToR) switches), based on the application, receiver nodes are clustered in multicast groups at different heights of the tree (Fig. 2(a), e.g. groups (A), (B) and (C) are representing 3 groups at the edge of the tree, groups (i) and (ii) are two lower branches with 4 nodes and groups (I) and (II) are at top of the tree with greater number of nodes. Assume in an iterative multicast application TX_1 and TX_2 are transmitting to multicast groups (i) and (ii) respectively. Leveraging WDM data delivery to both groups in the same tree is possible. Furthermore, changing the destination of transmitters between groups (i) and (ii) or assigning new members to the multicast group, i.e. (I) and (II), is achievable in µs-range using the WSS.

Our programmable optical network architecture is scalable and flexible for future developments due to the modularity and data rate transparency of the optical components. It is compatible with SDN in a sense that a centralized controller that is separated from the data plane could manage the network configuration. For iterative multicasting applications, comparing to the peer-to-peer solution, this architecture enables more efficient data delivery and has apparent advantages: 1) supporting several multicast groups within the same tree leveraging WDM; 2) dynamic multicast group member reassignment in µs range; 3) Supporting cheaper electronic commodity for the ToR switches by introduction of WDM in the optical physical layer through wavelength conversion.

III. Experiments and results:

We designed two experiments to address iterative multicasting applications. Due to equipment limitations, we designed our experiments to effectively demonstrate a portion of the network architecture in Fig. 2(a). In all experiments, an optical on-off keying signal was generated using 2^{31}-1 PRBS at 5 Gbps by modulating a C36 CW laser with a commercial LiNbO3 modulator. At the receiver side, the signal was amplified using an EDFA and filtered by a tunable band-pass filter. An integrated PIN-TIA-LA was used to generate the RF signal to the BERT system. We confirmed error-free (1E-10) signal transmission in all experiments.

In the first experiment, we present data delivery to two multicast groups within the same tree ((i) and (ii) in Fig. 2(a)). The setup and the measurement results are shown in Fig. 2 (b) and (c). In this experiment, node 1 is the transmitter in multicast group (a) and is a receiver in multicast group (b) and vice versa for node 2. Since there are two multicast groups, a second wavelength is introduced in the physical layer. A wavelength converter as a subsystem is connected to the OSS, and the data from node 1 is converted from C36 to C39. All-optical wavelength converters are yet to be commercially available; however, researchers have demonstrated several stable designs using nonlinear optics [7]. Since we are focusing on the application of wavelength conversion in data transporting rather than the process itself, we simply performed optical/electrical/optical (O/E/O) conversion with a C39 laser. The converted signals are set to have higher cross points.

In the second experiment, the idea of fine timescale multicast group member reassignment without OSS reconfiguration is examined. Fig. 2 (d) demonstrates multicast group (a) where node 6 is the transmitter and nodes 1 and 3 are the receivers and multicast group (b) where node 5 is the transmitter and nodes 2 and 4 are the receivers. The group members of multicast group (a) and (b) are switched by reconfiguring the WSS at higher speed than the OSS. The wavelength conversion is O/E/O with C39 laser. The eye diagrams of all 4 nodes in both setups are shown in Fig. 2 (d) and (e). The higher jitter in the O/E/O converted channels is due to the integrated limiting amplifier in the O/E conversion that introduces jitter for amplitude changes.

References:

[1] N. Farrington, *et al,* SIGCOMM Rev. 40, 4 (Aug. 2010), 339-350.
[2] G. Wang, *et al.*, SIGCOMM Rev. 40, 4 (August 2010), 327-338.
[3] H. Wang, *et al.*, SIGCOMM Rev. 43 (3) (Jul 2013).
[4] K. Thomas, *et al.*, IEEE Sym. on Sec. & Pri., pp. 447-462, 2011.
[5] M. Chowdhury, *et al.*, SIGCOM Rev. 41 (Aug. 2011), 98-109.
[6] H. Wang, *et al.*, Opt. Fib. Comm. (OFC), pp. 1-3, March 2012.
[7] A. Malacarne, *et al.*, JLT, V. 31, N. 11, 2013.