Scalable Interconnection Scheme for Data Center Multicast Applications

Scalable Interconnection Scheme for Data Center Multicast Applications

Kamran Keykhosravi
Dept. of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden
kamrank@chalmers.se

Houman Rastegarfar
College of Optical Sciences
The University of Arizona
Tucson, AZ, USA
houman@optics.arizona.edu

Erik Agrell
Dept. of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden
agrell@chalmers.se

***Abstract*—We propose a modular star-coupler-based switch architecture along with a scalable multicast scheduling algorithm to enable all-optical multicasting among data center nodes. With broadcast domain partitioning in a 126-port switch, we achieve up to 24% improvement in the maximum achievable throughput.**

***Index Terms*—Data center, optical multicasting, scheduling, star coupler, switch architecture.**

I. Introduction

Optical interconnects have received much attention for supporting a diverse set of data center traffic patterns in a flexible and bit-rate-transparent fashion. Today, a significant portion of data center jobs depend on the simultaneous transmission of the same information copy to a multitude of recipients (i.e., multicasting). However, the existing electronically switched deployments fall short of an efficient mechanism for handling this requirement. As such, recently several efforts have been devoted to developing data center multicast solutions directly in the optical domain [1], [2]. Due to their energy efficiency gains, multicast-enabled switches are also being proposed to replace the electronic top-of-rack switches [3].

The key enabler of traffic scheduling flexibility in the optical domain is the passive optical coupler. A $1 \times N$ power splitter allows the traffic of one sender to be delivered to all N receivers. An $N \times N$ star coupler, on the other hand, makes it feasible for N nodes to simultaneously participate in unicast, multicast, or broadcast sessions, provided that they are equipped with wavelength-tunable transceivers. The limited port count of star couplers as well as the tuning range of optical transceivers and filters make scalable optical multicasting a significant challenge [4]. Efforts have been made to overcome these shortfalls by interconnecting couplers in a nontransparent fashion [5]; however, performing frequent optoelectronic conversions is not desirable within data centers.

In this paper, we propose an all-optical multicast switch architecture for large-scale data center applications. Specifically, using K star couplers with port count N, our proposed solution can interconnect $K \times (N - K + 1)$ nodes. The multicast structure benefits from interconnecting disjoint broadcast domains, and can be used for either packet or circuit switching. Without loss of generality, here we focus on the packet-switching scheduling problem to quantify the delay–throughput performance. An optical packet comprises a burst of IP packets that are destined to the same place and is long enough to compensate for the hardware reconfiguration and scheduling overheads.

II. Modular Multicast Interconnect Architecture

To illustrate our star-coupler interconnection mechanism, we give an example with two $N \times N$ couplers that can lead to a multicast structure supporting $2(N-1)$ computing nodes. Fig 1(a) depicts the architecture of the interconnected switch. It is assumed that all servers (or computing nodes) are connected to a software-defined networking (SDN) controller, which performs the scheduling tasks. The first $N-1$ input/output ports of each coupler are connected to the servers and the last ports are employed to interconnect the two couplers using static filters (SFs) that act as optical gates. An SF could flexibly be realized through programming a wavelength-selective switch (WSS) [2]. Each server is equipped with a tunable transmitter and a tunable receiver and performs input buffering to resolve contentions. We assume that the transceivers are able to tune to W different wavelengths. The set of W wavelengths is partitioned into three disjoint subsets, namely WS1, WS2, and WS3. The wavelengths in WS1 are used for handling the traffic internal to each star coupler, whereas the wavelengths in WS2 (WS3) are used to carry the interdomain traffic from coupler 1 (2) to coupler 2 (1). As an example, server 1 is able to multicast its traffic to any subset of the servers connected to coupler 1 (coupler 2) using a single wavelength in WS1 (WS2). Deploying SFs in the interconnection scheme ensures that no collisions take place between the intra- and interdomain traffic demands.

It is straightforward to generalize the aforementioned scheme for interconnecting $K > 2$ star couplers of size $N \times N$ to support $K \times (N - K + 1)$ nodes in an all-optical multicast structure. To this end, the last $K-1$ input (output) ports of each coupler are connected to the output (input) ports of the remaining $K-1$ couplers via $2(K-1)$ optical fibers that are gated by SFs. In order to ensure disjoint broadcast domains in terms of interference, in each coupler the sets of wavelengths used for intradomain and interdomain communication

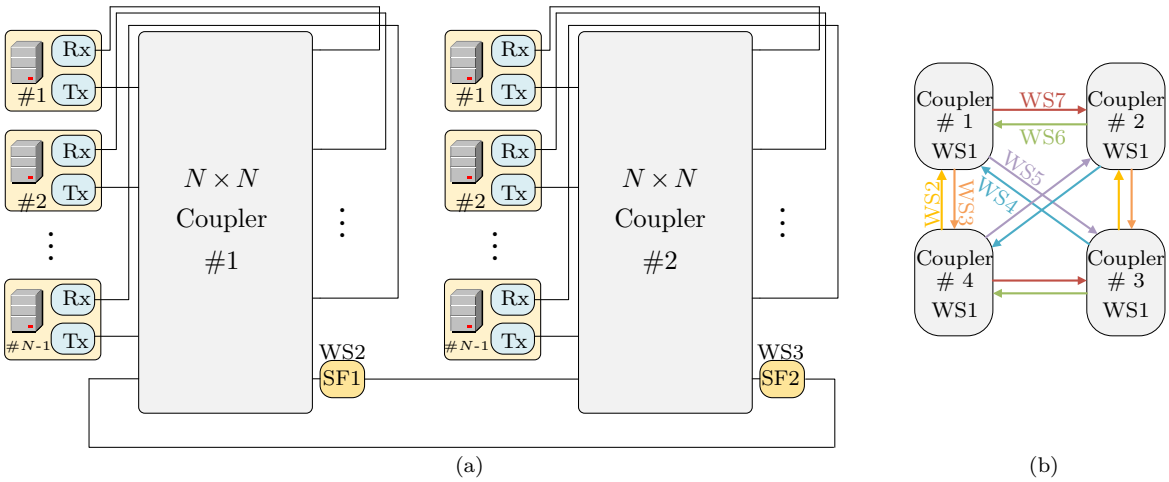

Fig. 1: Multicast switching structure deploying multiple couplers. The set of available wavelengths is partitioned into multiple subsets denoted by WSi. (a) The architecture of a $2(N-1) \times 2(N-1)$ switch based on two $N \times N$ star couplers. (b) The schematic of a switch constructed by interconnecting four star couplers along with the wavelength allocation pattern.

should be disjoint. Hence, we partition the set of available wavelengths into $2K-1$ subsets WSi, $1 \leq i \leq 2K-1$. The intradomain traffic is always serviced by the wavelengths in WS1, and each SF only passes the wavelengths in one subset WSi, $2 \leq i \leq 2K-1$, such that all fibers connected to any arbitrary coupler carry different sets of wavelengths. Fig. 1(b) illustrates this wavelength partitioning strategy in an interconnection among four couplers, where the 7 wavelength subsets are highlighted using different colors. In general, by casting the wavelength partitioning problem into a graph edge coloring problem, it can be shown that for even K's although each WSi, $2 \leq i \leq 2K-1$, is utilized multiple times, no two fibers with the same color are connected to the same coupler (for odd K's the number of subsets should be $2K+1$).

III. Multidomain Scheduling Algorithm

To schedule the optical interconnect based on K star couplers, we consider each server (input node) to have K queues, each buffering the traffic destined to a distinct coupler. A multicast packet whose destination set is spread over $1 \leq K' \leq K$ couplers will be copied K' times and will be placed in each of those K' queues. Considering this buffering strategy, we develop a modified version of the round-robin (RR) multicast scheduling algorithm in [4, Sec. III-B]. The scheduler starts from the queue of a node that is indicated by an RR pointer. Then, it sequentially examines every queue of every node and schedules the head-of-line (HOL) packets that can be transmitted to *all* of their destinations without contending with the already scheduled packets. Next, the scheduler examines all HOL packets and those that can be transmitted to a part of their destination set without contention get scheduled. Depending on the source and destination(s) of a scheduled packet, the corresponding transmitter and receiver(s) are tuned to a wavelength belonging to the desired wavelength subset (using first-fit wavelength assignment).

IV. Delay Performance Evaluation

We study the delay performance of the optical switch in Fig. 1(a) based on the interconnection of two 64×64 star couplers (i.e., a 126×126 switch). We conduct Monte Carlo simulations for different numbers of available wavelengths (W), multicast degrees, and traffic patterns. We consider two types of traffic, Bernoulli and geometric. In the former, during each time slot, a packet is generated in each node with probability $0 \leq p \leq 1$ and a set of destinations are picked uniformly at random. In the geometric traffic pattern, bursts of optical packets with the same destination set are generated. The length of each burst follows a geometric distribution, i.e., a discrete exponential distribution (with mean 16). The destination set of a packet is chosen uniformly from the set of all nodes excluding the transmitter itself. We assume that the fan-out, i.e., the cardinality of the destination set of a packet, follows a truncated geometric distribution with mean $\bar{F}$. A total number of 100,000 time slots were simulated, the second half of which contribute to the presented results. In order to speed up the simulations, we set the maximum queue length to 500 packets, which is large enough as long as queues are stable. We distribute the available wavelengths between WS1, WS2, and WS3 as uniformly as possible. Specifically, with $W = 16$, 32, and 64 available wavelengths, 6, 10, and 22 wavelengths are respectively assigned to WS1, and the remaining wavelengths are equally distributed between WS2 and WS3.

Fig. 2 depicts the average delay (i.e., the average number of time slots a packet spends in the buffer prior to transmission) versus the (normalized) throughput (i.e, the average number of packets transmitted in one time slot per node). The maximum achievable throughput increases with an increase in the number of wavelengths or the average fan-out. As the average fan-out increases, multicast scheduling allows for a larger number of

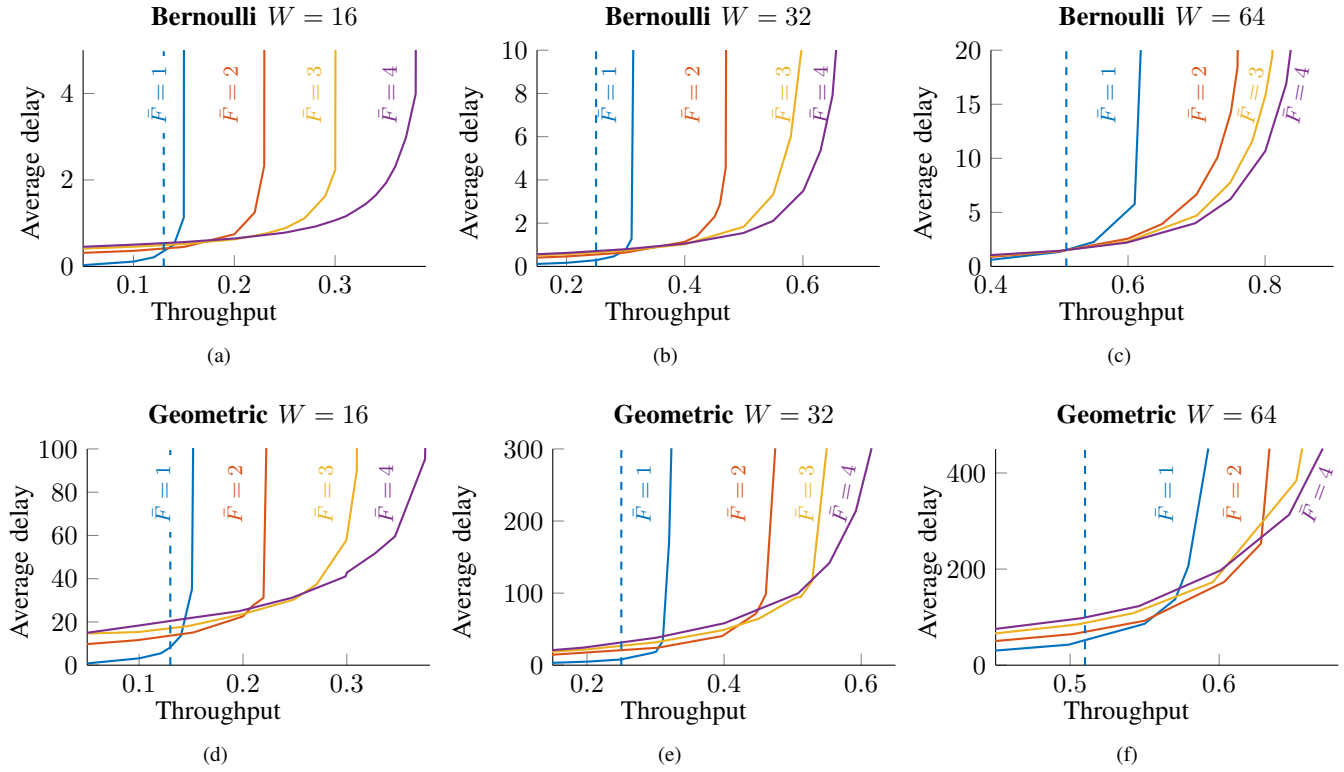

Fig. 2: The delay performance of a 126×126 switch constructed by interconnecting two 64×64 couplers as in Fig. 1(a).The dashed lines indicate the upper bound on the achievable throughput under unicast traffic ($\bar{F} = 1$) within a 126-port unpartitioned switch.

packet copies to be disseminated with a single wavelength, translating to an increased throughput. With $W = 16$, the maximum throughput under Bernoulli traffic increases from 0.15 for $\bar{F} = 1$ to 0.38 for $\bar{F} = 4$, corresponding to a 153% increase. The same trend is observed in Fig. 2(b) and Fig. 2(c) with larger numbers of wavelengths. With $\bar{F} = 4$ and 64 wavelengths, a maximum throughput of 0.83 is achieved, which is 2.8 times the maximum throughput with $\bar{F} = 4$ and $W = 16$. The switch exhibits a similar behavior under the geometric traffic. However, the delays are increased as queues build up fast under bursty traffic.

Compared with an equally sized single-coupler switch (based on a 128×128 star coupler with 126 used ports), one can notice the advantage of broadcast domain partitioning in our design. With a single broadcast domain of N ports, W wavelengths, and an average fan-out of $\bar{F}$, the upper bound of the throughput is calculated as $W\bar{F}/N$, which can only be realized in the absence of output port contentions. With $\bar{F} = 1$ and $W = 16$, this upper bound is approximately 0.13, which is less than the maximum achievable throughput in our design (i.e., 0.15). The same observation can be made with $W = 32$ ($0.31 > 0.25$) and $W = 64$ ($0.62 > 0.51$). In other words, the partitioning technique enables a throughput improvement of 15% for $W = 16$, 24% for $W = 32$, and 21% for $W = 64$. For $\bar{F} = 2$, the maximum achievable throughput of our design is close to the upper bound. For example, with $W = 32$ a throughput of 0.47 is achieved, whereas the theoretical upper bound is 0.51. For higher values of $\bar{F}$, the upper bound becomes loose as it disregards contentions.

V. Conclusion

We proposed a novel multicasting mechanism based on the transparent interconnection of star couplers, leading to broadcast domain partitioning. Due to partially reusing the spectral resources, the performance of our design can exceed that of an equally sized single-coupler switch. Besides, an increase in the number of wavelengths and the fan-out can further improve the throughput performance. To reduce the delays under bursty traffic, a load balancing stage can be employed at the switch input.

References

[1] P. Samadi, V. Gupta, J. Xu, H. Wang, G. Zussman, and K. Bergman, "Optical multicast system for data center networks," *Opt. Express*, vol. 23, no. 17, pp. 22 162–22 180, Aug. 2015.

[2] H. Rastegarfar, K. Keykhosravi, E. Agrell, and N. Peyghambarian, "Wavelength reuse for scalable multicasting: a cross-layer perspective," in *Proc. Optical Fiber Communication Conf. (OFC)*, San Diego, CA, USA, Mar. 2018, paper W2A.20.

[3] J. Chen, Y. Gong, M. Fiorani, and S. Aleksic, "Optical interconnects at the top of the rack for energy-efficient data centers," *IEEE Commun. Mag.*, vol. 53, no. 8, pp. 140–148, Aug. 2015.

[4] K. Keykhosravi, H. Rastegarfar, and E. Agrell, "Multicast scheduling of wavelength-tunable, multiqueue optical data center switches," *J. Opt. Commun. Netw.*, vol. 10, no. 4, pp. 353–364, Apr. 2018.

[5] Q. Li *et al.*, "Scaling star-coupler-based optical networks for avionics applications," *J. Opt. Commun. Netw.*, vol. 5, no. 9, pp. 945–956, Sep. 2013.