On-Chip SDM Switching for Unicast, Multicast and Traffic Grooming in Data Center Networks

Kamchevska, Valerija; Ding, Yunhong; Dalgaard, Kjeld; Berger, Michael Stübert; Oxenløwe, Leif Katsuo; Galili, Michael

Published in:
IEEE Photonics Technology Letters

Link to article, DOI:
10.1109/LPT.2016.2636866

Publication date:
2017

Document Version
Peer reviewed version

On-Chip SDM Switching for Unicast, Multicast and Traffic Grooming in Data Center Networks

Valerija Kamchevska, *Student Member, IEEE, Member, OSA*, Yunhong Ding, Kjeld Dalgaard, Michael Berger, Leif K. Oxenløwe and Michael Galili

***Abstract*—This paper reports on the use of a novel photonic integrated circuit that facilitates multicast and grooming in an optical data center architecture. The circuit allows for on-chip spatial multiplexing and demultiplexing as well as fiber core switching. Using this device, we experimentally verify that multicast and/or grooming can be successfully performed along the full range of output ports, for different group size and different power ratio. Moreover, we experimentally demonstrate SDM transmission and 5 Tbit/s switching using the on-chip fiber switch with integrated fan-in/fan-out devices and achieve error-free performance (BER≤10^{-9}) for a network scenario including simultaneous unicast/multicast switching and traffic grooming.**

***Index Terms*—Data centers, switching, traffic grooming, multicast communication.**

I. Introduction

Traffic in data center networks (DCNs) is growing at a fast pace [1] and gaining increasing attention over the past few years. For modern data center networks that run cloud-based applications, bandwidth utilization and energy efficiency become crucial for sustainable growth. Optical DCN architectures have been proposed on several occasions to alleviate network power consumption, as well as to address resource utilization proportionality and reduce network load when possible, by for example adopting optical switching for unicast, multicast and incast i.e. optical grooming [2]. This can be found useful for distributed file systems widely used in DCNs for storing data such as Google File System (GFS) [3] or Hadoop Distributed File System (HDFS) [4] where files are divided into chunks of data (64 MB for GFS, 128 MB for HDFS) that are replicated and stored in different servers for reliability. Other examples include MapReduce based distributed computational frameworks [5] which can exploit optical multicast and incast for both the *map* and *reduce* phase where the input data is distributed among different machines and the derived values are merged into a smaller set of data.

Manuscript received June 29, 2016. This work was supported by the ECFP7 Grant 619572, COSIGN.

V. Kamchevska, Y. Ding, K. Dalgaard, L. K. Oxenløwe and M. Galili are with the High-Speed Optical Communications Group, Department of Photonics Engineering, Technical University of Denmark, Kongens Lyngby 2800, Denmark (e-mail: vaka@fotonik.dtu.dk; yudin@fotonik.dtu.dk; kdal@fotonik.dtu.dk; lkox@fotonik.dtu.dk; mgal@fotonik).

M. Berger is with the Networks Technology and Service Platforms Group, Department of Photonics Engineering, Technical University of Denmark, Kongens Lyngby 2800, Denmark (e-mail: msbe@fotonik.dtu.dk).

With the number of servers increasing, the benefits of performing these operations optically become even more important. For example, replacing several unicast connections with one multicast connection allows for lower bandwidth usage which effectively leads to reduced network congestion and higher throughput. Moreover, as the sender would have to send only one copy, energy efficient and low latency communication i.e. faster task execution time can be achieved.

Previously, we proposed a DCN architecture based on multidimensional switching nodes connected in a ring [6]. In this letter, we focus our attention on enabling additional network functionalities such as providing support for multicast and optical grooming. In order to facilitate this, we use a novel photonic integrated circuit (PIC) composed of a switch matrix and fan-in/fan-out devices for coupling to a multicore fiber (MCF). The PIC can inherently provide support for switching at fiber core granularity [7, 8]. Here, we investigate the system performance when using this switch to perform multicast of traffic destined to servers connected to different wavelength selective switches (WSSs) within or between nodes (intra and inter-node) as well as grooming of traffic originating from servers connected to different intra and inter-node WSSs. We experimentally demonstrate BER performance ≤10^{-9} of a single channel for 1:2 multicast at different output ports as well as multicast with multicast group size ranging from 2 to 7. Negligible penalty is observed when varying the power ratio for 1:2 multicast and 2:1 grooming. Moreover, BER≤10^{-9} is achieved for simultaneous switching of 5 Tbit/s in a combined unicast, multicast and grooming scenario.

The remainder of this letter is organized as follows: in Section II we give an overview of the proposed architecture, the fabricated PIC and the concepts behind multicast and grooming. In Section III, we experimentally characterize the performance of a single channel for multicast at different ports and with different multicast group size, as well as multicast and grooming with different power ratio. Moreover, we experimentally verify the performance in a combined unicast, multicast and grooming scenario. At last, in Section IV we summarize the presented work and make concluding remarks.

II. Multicast and Grooming in Optical Ring DCN

The proposed architecture [6] is shown in Fig.1. Servers connected to Top-of-Rack (TOR) switches are interconnected through an optical ring network of multidimensional switching nodes. Each node enables switching at four granularities i.e.

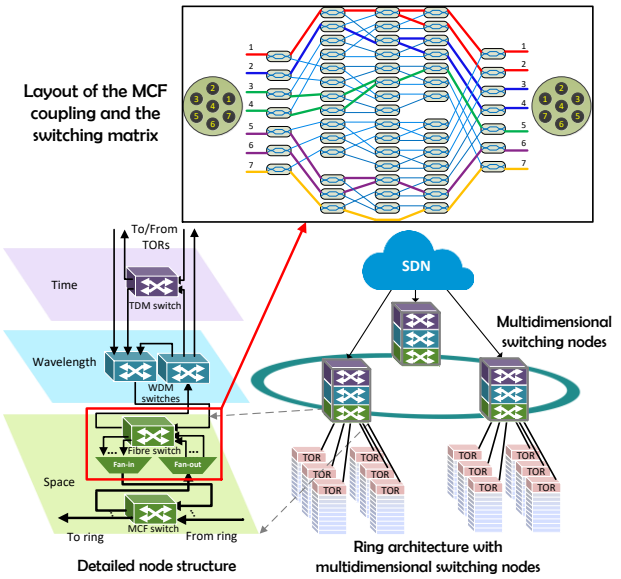

Fig. 1. Proposed architecture and chip layout.

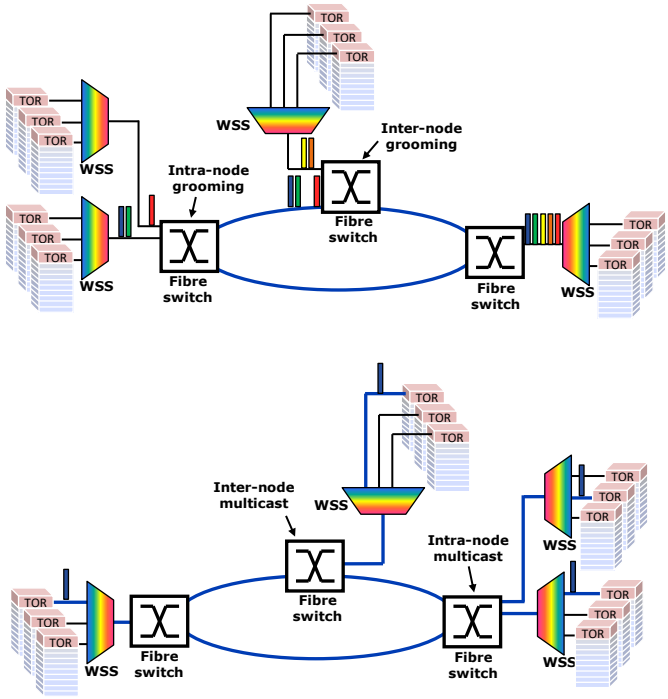

Fig. 2. Inter and intra-node grooming (top) and multicast (bottom) scenarios.

full MCF, single core, wavelength and subwavelength time slot granularity. As shown in Fig.1, the switching in the different domains is done using optical circuit switches (MCF and fiber switch) in the space domain, WSSs (WDM switches) in the wavelength domain, and fast optical switches (TDM switches) in the time domain. As connections to/from TORs are full wavelength for bandwidth-hungry applications or time slots for bursty traffic, we consider the way to provide support for multicast and grooming of these connections. Multicast or grooming of traffic destined to or originating from TOR ports connected to the same WSS or TDM switch can easily be provided by power splitting. For example, multicast among TOR ports connected to one WSS can be done by power splitting among the desired WSS output ports (intra-WSS multicast). However, as there may be several WSS and TDM switches per node, performing multicast and grooming among TOR ports that are connected to different WSS or TDM switches can no longer be done in the same way. For this reason, we envision the use of the fiber switch to enable these functionalities among TOR ports connected to different higher layer switches. For example, as illustrated in Fig.2 (top), using the fiber switch, traffic grooming can be done among TOR ports connected to different WSSs within the leftmost node in the ring. In addition, multicast among TOR ports connected to different WSSs can be performed using the fiber switch at the rightmost node in Fig.2 (bottom).

We use a silicon PIC shown in Fig.1 (inset) that acts as fiber switch with integrated grating couplers for MCF coupling. The 12mm x 5mm PIC is fabricated on a SOI platform with top silicon thickness of 250 nm. The measured insertion loss and crosstalk for C-band [7, 8] are less than 8 dB and -30 dB, respectively. The 7x7 switch is thermally controlled with switching time of ~30 μs and 13 mW power consumption per heater. The cores of the MCF are switched by controlling a heater in each of the 57 Mach-Zehnder interferometric (MZI) structures. The use of MZI allows not only for switching, but also power splitting and combining with appropriate heater control. Unlike previous demonstrations of optical multicast where optical splitters are required in addition to the optical circuit switch, the fabricated PIC can perform both multicast

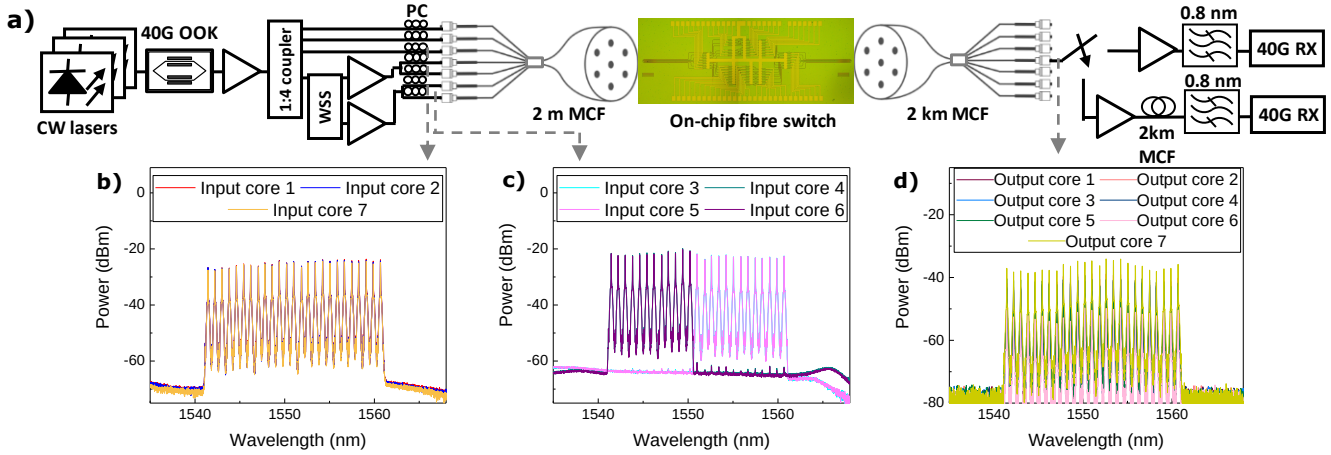

Fig. 3. a) Experimental setup. Spectra at the input of the MCF for b) cores 1,2,7; c) cores 3,4,5,6 and d) spectra of all cores after switching.

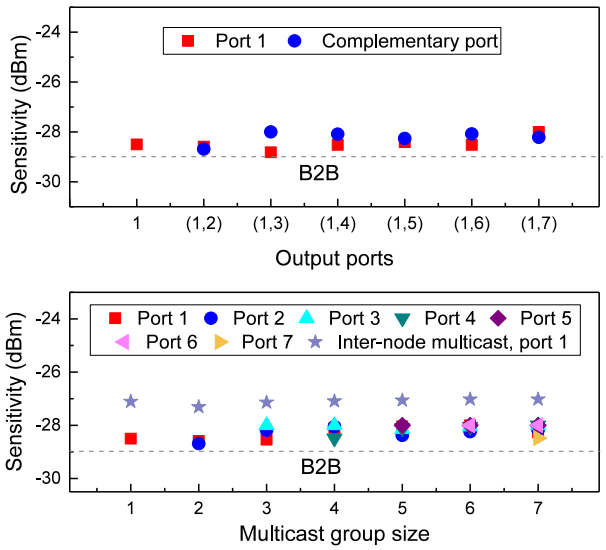

Fig. 4. Receiver sensitivity of single channel (1550.12nm) for 1:2 multicast on different output ports (top) and different multicast group size (bottom).

and grooming without the need of additional equipment.

Moreover, switching in unicast and multicast fashion and grooming can be done on-demand. To verify that the switch can support these features we focus on intra and inter-node grooming and multicast of wavelength connections as shown in Fig. 2, although the same is applicable to TDM connections.

Grooming is important as in each node the drop ports and the number of switches deployed at every layer will be limited. If we assume for simplicity that a node has only one WSS for dropped traffic then in order to receive data originating from different WSSs within a node or from WSSs in different nodes simultaneously, traffic needs to be packed so that as many connections as possible can be established. Without grooming, all connections among different WSSs illustrated in Fig.2 (top) would have to be established sequentially even if they have different destination TOR ports. Thus, only traffic originating from TORs under a single WSS can coexist. Alternatively, traffic can be repacked using the WSSs in intermediate nodes. However, due to the limited number of WSS bypass ports this is not effective for high hop count and can lead to extra delay. On the multicast side, using one wavelength to establish few connections at the same time allows for bandwidth efficiency that is directly proportional to the multicast group size and enables low latency connectivity. Furthermore, the coupling ratio can be asymmetric for both multicast and grooming. This allows for crucial trade-offs, such as allocating more power to the inter-node terminated connections resulting in similar inter/intra node performance.

An important thing to note is the control of the switch for performing multicast and grooming. Like unicast switching, where a microcontroller board was used to find the optimum settings of each MZI for a given configuration, a thorough search allows that the correct settings for multicast/grooming with different port arrangements and ratios are identified. After initial characterization, the values are saved, allowing for fast dynamic reconfiguration. Moreover, the control for establishing a unicast vs. a multicast/grooming connection is

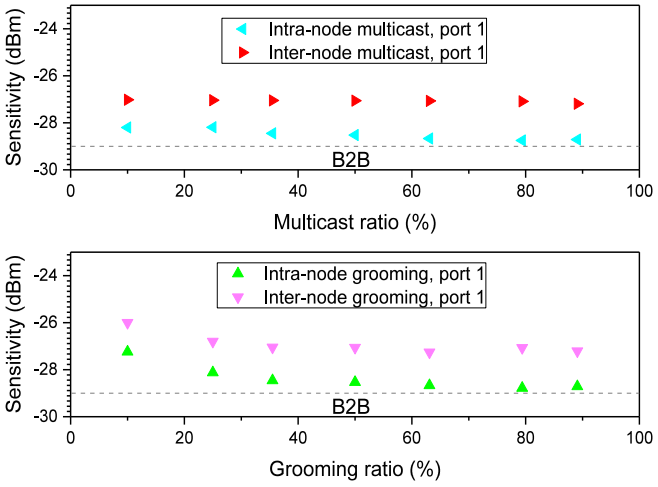

Fig. 5. Receiver sensitivity of a single channel (1550.12nm) for 1:2 multicast (top) and 2:1 grooming (bottom) with different power ratios.

rather similar. For example, a 1x2 multicast from input port 1 to output ports 1 and 2 requires that the configuration settings for the unicast connections from input 1 to outputs 1 and 2 are applied to all MZI but one, that is configured to act as a splitter. Similarly, a *1xN* multicast requires modifying only the setting of *N-1* MZI. With this, the control scheme is narrowed down to applying new settings for only few MZI and resembles the control for unicast switching otherwise.

III. Experimental Demonstration

A. Multicast and Grooming Characterization

In order to confirm that successful multicast and grooming can be performed using the PIC, we initially investigate the performance of the system for a few different cases. The characterization setup is similar to the experimental setup shown in Fig.3, that is used for the combined scenario. When grooming, different channels are launched in the seven different cores and the groomed traffic after the switch propagates in the 2km MCF (as shown in Fig.3). When multicasting, a single channel is launched in only one core propagating first in the 2km MCF (reversing the setup in Fig.3), allowing for proper multicast (one copy sent and appropriately split at the desired node). Inter-node grooming and multicast is done by adding extra 2km MCF before the switch when grooming and after the switch when multicasting.

Initially, we use a single channel (1550.12nm) that carries 40 Gbit/s OOK modulated data. The channel is launched in one core of a 2-km MCF that is coupled to the chip where multicast is performed using the switch. First, we investigate the ability to perform 1:2 multicast for six combinations of output port pairs and then we verify that similar performance is obtained for different multicast group size. The measured receiver sensitivity for both cases is shown in Fig.4. Similar behavior is observed for 1:2 multicast over different output ports confirming that the switch can provide flexible multicast along the full range of output ports. We believe that the minor variations are due to imperfections in fine tuning the heaters for the different paths and can be avoided by further control optimization. Varying the multicast group size from 2 to 7, results in negligible penalty for both intra and inter-node

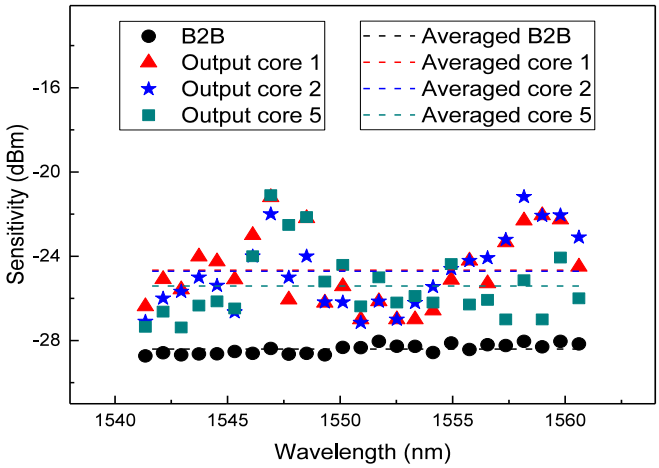

Fig. 6. Receiver sensitivity of all channels for three output cores.

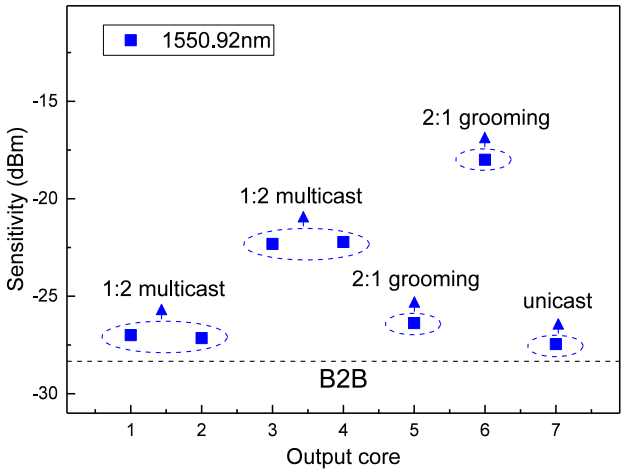

Fig. 7. Receiver sensitivity of a single channel (1550.92nm) in all cores.

multicast confirming the ability to provide an on-demand multicast ratio without additional equipment.

In addition, we investigate the effect of the coupling ratio on the system performance for both multicast and grooming. The measured receiver sensitivities for both cases are shown in Fig.5. For 1:2 multicast of a single channel (1550.12nm), the receiver sensitivity is measured on a single fixed port (port 1) for power splitting ratios ranging from 10% to 89%. It can be seen that for both intra and inter-node multicast, negligible penalty is observed when reducing the output power on port 1. Similar results are observed for 2:1 grooming of two channels (1550.12nm and 1550.92nm) at the two input ports. The performance of a single channel (1550.12nm) is measured at the output of the switch (port 1). It can be seen that the penalty of modifying the combining ratio is relatively small. However, considering a network, these results allow for improved performance by identifying the optimal ratio in different cases.

B. Combined Unicast, Multicast and Grooming Scenario

In order to verify that all functionalities such as unicast, multicast and grooming can simultaneously be provided using the same switch, we consider the following scenario: 1:2 multicast is performed on two input cores (1 and 2) carrying 25 channels each; 2:1 grooming is performed on 4 input cores (3 and 4; 5 and 6) where each of the 2 groomed cores carries 13 or 12 spectrally non-overlapping channels; and at last, one core (input core 7) carrying 25 channels is unicast switched. The paths in the switching matrix are illustrated in Fig.1a and the experimental setup is shown in Fig.2a. All channels carry 40 Gbit/s OOK modulated data. The spectra of each core at the input and at the output of the MCF are shown in Fig.2b/2c, and in Fig.2d, respectively. Due to the specific switching scenario considered, each core after switching has 25 channels transmitted in the 2-km MCF.

The performance of the system is assessed by measuring the receiver sensitivity on all channels for two output cores that undergo intra-node (output core 1) and inter-node (output core 2) multicast and for one output core (output core 5) that carries groomed traffic. Fig.6 shows the measured receiver sensitivity. All channels have similar performance with some experiencing greater penalty because of imperfect power equalization as well as weak wavelength dependent crosstalk in the switch, due to polarization variation. In addition, Fig.7 shows the measured receiver sensitivity of one channel in all cores for the same switching configuration. BER$\leq 10^{-9}$ is achieved in all cores although the different insertion loss and crosstalk of the MCF coupling devices as well as the experienced crosstalk in the switch contribute to variations.

IV. Conclusion

We propose on-chip fiber switching to facilitate traffic grooming and multicast in an optical DCN. Uniform behavior along the full range of output ports and excellent performance with negligible degradation when increasing the group size is achieved for a single channel multicast. Asymmetric power ratio for 1:2 multicast and 2:1 grooming also results with relatively small penalty for low signal power. Moreover, BER$\leq 10^{-9}$ is achieved for channels in different cores when simultaneous unicast and multicast switching as well as traffic grooming is performed with 5 Tbit/s throughput. Based on these results, we are confident that this novel and compact photonic integrated circuit with combined fan-in/fan-out and switching functionalities can be used to provide support not only for unicast switching, but also for multicast and grooming in future SDM-enabled optical data center networks.

References

[1] A. Singh *et al.*, "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network," Proc. SIGCOMM, pp.183-197 (2015)

[2] P. Samadi *et al.*, "Accelerating cast traffic delivery in data centers leveraging physical layer optics and SDN" Proc. ONDM, pp.73-77 (2014)

[3] S. Ghemawat *et al.*, "The Google file system," Proc. Symp. Oper. Syst. Principles, pp.29–43 (2003)

[4] K. Shvachko *et al.*, "The Hadoop Distributed File System," Proc. MSST, pp.1-10 (2010)

[5] J. Dean *et al.*, "MapReduce: Simplified data processing on large clusters," Proc. OSDI, pp.137–150 (2004)

[6] V. Kamchevska *et al.*, "Experimental Demonstration of Multidimensional Switching Nodes for All-Optical Data Center Networks," J. Lightwave Technol., vol. **34**, no. 8, pp.1837-1843 (2016)

[7] Y. Ding *et al.*, "Experimental Demonstration of 7 Tb/s Switching Using Novel Silicon Photonic Integrated Circuit," Proc. CLEO, Stu1G.3 (2016)

[8] Y. Ding *et al.*, "Reconfigurable SDM Switching Using Novel Silicon Photonic Integrated Circuit," arXiv:1608.05645 (2016)