DTU Library

Experimental Demonstration of Multidimensional Switching Nodes for All-Optical Data Center Networks

Kamchevska, Valerija; Medhin, Ashenafi Kiros; Da Ros, Francesco; Ye, Feihong; Asif, Rameez; Fagertun, Anna Manolova; Ruepp, Sarah Renée; Berger, Michael Stübert; Dittmann, Lars; Morioka, Toshio

Total number of authors:
12

Published in:
Journal of Lightwave Technology

Link to article, DOI:
10.1109/JLT.2016.2518863

Publication date:
2016

Document Version
Peer reviewed version

Citation (APA):
Kamchevska, V., Medhin, A. K., Da Ros, F., Ye, F., Asif, R., Fagertun, A. M., Ruepp, S. R., Berger, M. S., Dittmann, L., Morioka, T., Oxenløwe, L. K., & Galili, M. (2016). Experimental Demonstration of Multidimensional Switching Nodes for All-Optical Data Center Networks. *Journal of Lightwave Technology*, *34*(8), 1837-1843. https://doi.org/10.1109/JLT.2016.2518863

Experimental Demonstration of Multidimensional Switching Nodes for All-Optical Data Center Networks

Valerija Kamchevska, *Student Member, OSA*, Ashenafi Kiros Medhin, *Student Member, OSA*, Francesco Da Ros, *Member, OSA*, Feihong Ye, *Student Member, OSA*, Rameez Asif, Anna Manolova Fagertun, *Member, IEEE*, Sarah Ruepp, Michael Berger, Lars Dittmann, Toshio Morioka, *Member, IEEE, Fellow, OSA*, Leif Katsuo Oxenløwe, *Member, OSA*, and Michael Galili, *Member, IEEE*

***Abstract*—This paper reports on a novel ring-based data center architecture composed of multidimensional switching nodes. The nodes are interconnected with multicore fibers and can provide switching in three different physical, hierarchically overlaid dimensions (space, wavelength and time). The proposed architecture allows for scaling in different dimensions while at the same time providing support for connections with different granularity. The ring topology reduces the number of different physical links required, leading to simplified cabling and easier link management, while optical bypass holds the prospect of low latency and low power consumption. The performance of the multidimensional switching nodes has been investigated in an experimental demonstration comprising three network nodes connected with multicore fibers. Both high capacity wavelength connections and time-shared subwavelength connections have been established for connecting different nodes by switching in different physical dimensions. Error-free performance (BER<10^{-9}) has been achieved for all the connections with various granularity in all the investigated switching scenarios. The scalability of the system has been studied by increasing the transmission capacity to 1 Tbit/s/core equivalent to 7 Tbit/s total throughput in a single 7-core multicore fiber. The error-free performance (BER<10^{-9}) for all the connections confirms that the proposed architecture can meet the existing demands in data centers and accommodate the future traffic growth.**

***Index Terms*—Data center networks, optical switching, space division multiplexing, wavelength division multiplexing, time division multiplexing.**

This work was supported by the ECFP7 grant no. 619572, COSIGN and Innovations Fonden grant 0603-00514B, E-SPACE. We would like to thank OFS for fabricating and providing the spliced fan-in/fan-out multicore fiber and Polatis for providing the 48x48 beam-steering fiber switch.

V. Kamchevska, A. K. Medhin, F. Da Ros, F. Ye, R. Asif, T. Morioka, L. K. Oxenløwe, and M. Galili are with the High-Speed Optical Communications Group at the Department of Photonics Engineering, Technical University of Denmark, Ørsteds Plads Byg. 343, 2800 Kgs. Lyngby, Denmark.

A. M. Fagertun, S. Ruepp, M. Berger, and L. Dittmann are with the Networks Technology and Service Platforms Group at the Department of Photonics Engineering, Technical University of Denmark, Ørsteds Plads Byg. 343, 2800 Kgs. Lyngby, Denmark.

I. Introduction

The continuous growth of Internet traffic has huge implications on existing networks. In order to be able to sustain this growth, network operators are challenged by the scalability of the deployed technologies. Besides traditional telecom networks, an even more severe impact of the cloud computing expansion can be seen in data centers, where traffic is growing at an annual rate higher than 20% [1]. The implications of this growth have made data center operators and standardization organizations move at a rapid pace and making decisions that favor scalability. With respect to the fiber infrastructure, data centers are already adopting single mode fiber (SMF). As the fiber is intrinsically considered an infrastructure which remains in the data center network (DCN) while networking equipment such as transceivers and switches are upgraded, SMF is a more future-proof choice compared to multimode fiber due to its improved transmission reach and wavelength division multiplexing (WDM) support. Current DCNs already deploy 100G links and standardization of higher rates is on its way. Thus, scalable switching technologies that can easily support increased switching capacity without significant increase in power consumption, latency and cost are highly desirable.

Optical switches are known for providing high bandwidth at a relatively low power consumption and low latency, and have been proposed for intra-data center connectivity in several scenarios. However, the various applications running in DCNs result in the presence of two types of flows i.e. large and long-lasting flows demanding greater bandwidth also called ‘elephant’ flows as well as the smaller, more dynamic flows requiring high connectivity often denoted as ‘mice’ flows. The distinct traffic characteristics of these flows dictate that the network should provide support for connections with different granularity. Additionally, considering the future expansion of DCNs, the need for utilizing different multiplexing technologies in order to support scaling to immense capacity becomes more than apparent. The ability to scale the link capacity in different dimensions (by increasing the bit rate per channel, by adding more wavelengths or cores per fiber etc.), while at the same time handling all the connections with different granularity offers a scalable solution to address the future growth. Furthermore, in order to provide flexible resource allocation and efficient bandwidth utilization, dynamic assignment schemes have to be envisioned, making the network fully programmable and effectively controlled by software defined networking (SDN) paradigms.

In [2], we presented for the first time a novel ring based DCN architecture addressing the aforementioned issues by utilizing all-optical hierarchically layered multidimensional switching nodes for intra data center application and in [3] we studied the benefits in terms of blocking and resource utilization in a ring topology compared to a standard fat tree architecture. In this paper, we elaborate on the node structure as well as the requirements at each level, emphasizing the benefits of switching in three different physical dimensions (space, wavelength and time) in a hierarchical fashion. In addition, we discuss the results of the experimental demonstration and the performance of all the connections with different granularity switched between three multidimensional nodes, including a 1 Tbit/s/core or 7 Tbit/s throughput per multicore fiber error-free (BER<10^{-9}) communication between nodes connected with 2-km long multicore fibers.

The remainder of this paper is organized as follows: In Section II we give a brief overview of the current status in data centers and the future requirements. Based on the identified issues, we review some of the architectures already proposed and discuss the potential of different technologies to address some of the main challenges. In Section III, we present the concept of multidimensional switching and elaborate on the node structure. We propose a novel architecture based on these nodes and experimentally demonstrate error-free performance of connections with different granularity being switched between three multidimensional nodes. At last, in Section IV we summarize the proposed architecture supported by the experimental results and conclude the paper.

II. Overview of Existing Data Center Network Architectures And Technologies

A. DCN: Current State, Issues and Requirements

Current data centers are mainly based on Ethernet switching [4]. Even though Ethernet switching can inherently provide support for packet based traffic and is a mature technology, there are several issues that have to be taken into account when considering the future growth in data centers. From the architecture point of view, scalability is the first limiting factor. From 1998 to 2005, data centers grew in size by an amazing 173% [5] and considering that the Internet traffic is constantly growing [1], this trend is likely to continue. In order to support this growth, data centers should be able to scale the network as the number of servers increases. In a DCN based on a fat tree topology this can be done by either replacing the switches for others that have higher port count or by increasing the number of switches at each tier. The first approach is limited by the fact that the switch radix cannot be arbitrarily high and additionally, replacing all the switches may not be the most cost effective method. The latter one, on the other hand, leads to an increased number of switches at each tier and with the number of servers growing rapidly in large-scale DCNs the higher number of tiers becomes inevitable. However, adding additional tiers increases the end-to-end latency, and for applications that are latency-sensitive, this may have a detrimental effect on their performance.

Another issue is the oversubscription. As standard tree topologies often have oversubscription ratios higher than 1:2, links are shared among servers over time such that the upstream Top-of-Rack (TOR) bandwidth available is much lower than the downstream TOR bandwidth facing the servers. This of course limits the available capacity, creates bottlenecks and traffic congestion.

Furthermore, due to the switching being done in the electrical domain and the frequent optical-electrical-optical (OEO) conversions, the power consumption of DCNs based on Ethernet switching has been a hot topic for discussion over the years. In [6], Abts *et al*. have shown that at 100% utilization of the servers, the network consumes 12% of the overall DCN power consumption, however if the servers are operating at 15% utilization, the network share becomes 50% of the total power. Considering that the total power consumed by DCNs in the world in 2010 accounted for 1.1% to 1.5% of the total electricity use [7], it becomes crucial that the power consumed by the network is reduced to the bare minimum.

B. Overview of Proposed Architectures and Enabling Technologies

Several architectures have been proposed aiming to address some of the aforementioned issues in DCNs based on Ethernet switching. The benefit from introducing optics in data centers was first recognized in hybrid Ethernet-optical solutions like "Helios" and "c-Through" [8], [9]. These architectures identified the need to switch long-lived connections in a more energy-efficient manner with low latency and proposed that optical circuit switching (OCS) be used for that purpose. OCS has recently started to see commercial deployment in data centers, mainly due to the fact that besides support for real-time data it is very well suited for scheduled tasks such as data migration and storage backup, replication and virtual machine (VM) migration [10], which inherently require large amount of bandwidth over longer period of time.

Additionally, there have been several all-optical architectures [11-13] that proposed the use of optics for switching short-lived bursty data. However, unlike OCS, the main issue with optical packet switching (OPS) is the need for optical buffering. In order to perform packet switching in the optical domain, buffering has to be done in the network in case of congestion. Different optical buffering technologies or alternatively mixed approaches in which data is O/E converted, buffered in the electrical domain and retransmitted require extra components resulting in higher component cost and additional control overhead. Consequently, OPS has not been commercially deployed so far. This issue may alternatively be solved by technologies such as circuit-based time division multiplexed (TDM) optical switching similar to what has been proposed in [14], [15] in which the buffering is done at the network edges. As buffering in the electrical

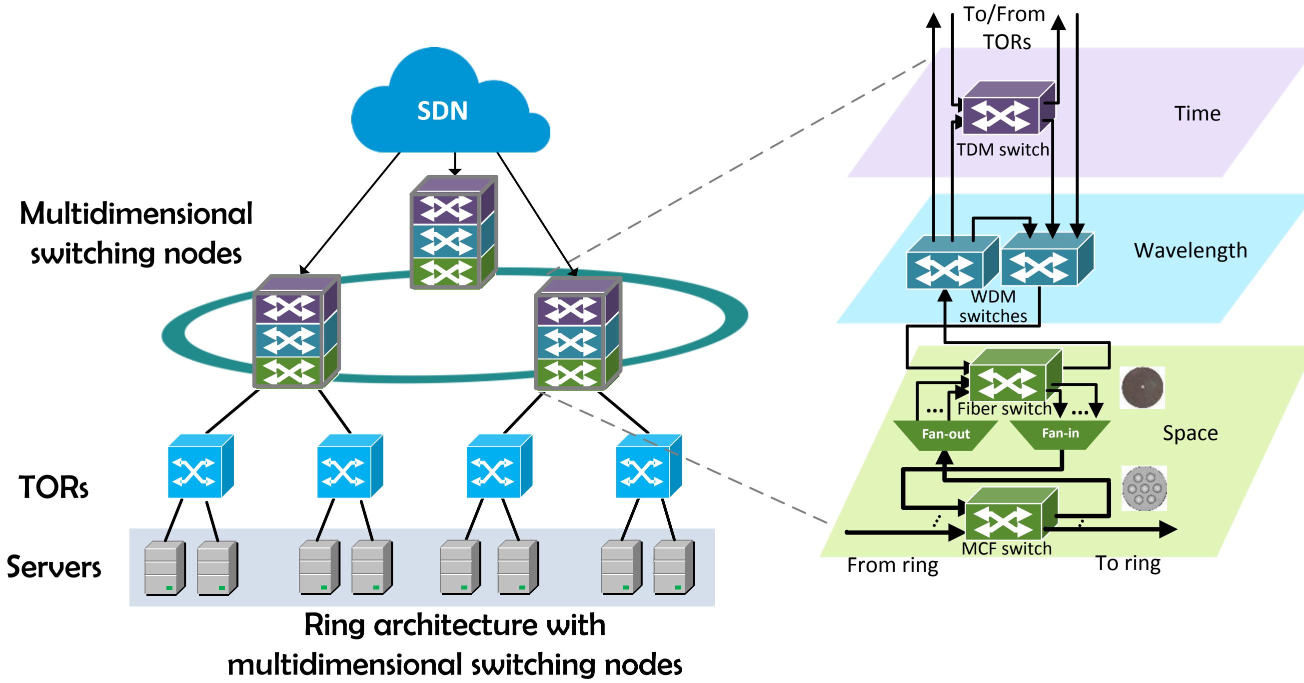

Fig. 1. Proposed ring architecture based on multidimensional switching nodes. The detailed node structure shows the switches deployed in each dimension.

domain is readily available, and there is no need for extra buffering in the network, this technology seems promising. Thus, utilizing all-optical technologies for DCNs may enable overcoming some of the drawbacks of Ethernet-based networks such as power consumption, latency and most importantly, scalability.

III. Proposed DCN Architecture

A. *Multidimensional Switching Node Structure*

Multiplexing technologies such as TDM, WDM and space division multiplexing (SDM) have been researched considerably over the years, as they allow for increased capacity and better bandwidth utilization. However building complex systems in which more than one of these technologies is deployed imposes the need for switching devices that would be able to operate with connections that have different granularities and switch in different dimensions.

The proposed multidimensional nodes are shown in Fig.1. They are all-optical hierarchically layered nodes that can switch in three different dimensions (space, wavelength and time) and are based on a reconfigurable optical add drop multiplexer (ROADM) structure. The hierarchical layout of the switches in the different dimensions enables that connections can be either bypassed at a lower level whenever they are destined to other nodes or to be processed at a higher level in case switching with finer granularity is required. Due to the inherently large amount of bypass traffic at each node, a ring topology for interconnection of the multidimensional switching nodes has been chosen as shown in Fig. 1.

Switching in the space dimension is performed using two different types of switches, one operating at a multicore fiber (MCF) granularity, and another one operating at a single fiber core granularity. At the lowest level switching of whole multicore fibers is performed. The switch at this level, i.e. the *MCF switch*, takes multicore fibers at the input and switches simultaneously all the cores of a particular fiber to a single multicore fiber output. Most of the inputs and outputs of this switch are connected to multicore fibers that link the specific node with the two neighboring nodes in the ring. A smaller portion of the inputs/outputs is instead connected to the higher level switch within the node, i.e. the *fiber switch* after passing through designated fan-in/fan-out devices. If the traffic carried in one multicore fiber is not intended for the node through which it is passing, then bypass at this level enables negligible delay for the connections being served through those physical resources. Furthermore, by simultaneously switching highly aggregated data i.e. large number of connections at this level, the switching energy per bit is expected to be radically low, ultimately leading to reduced operational costs.

The *fiber switch* allows for switching of individual cores. Using this switch it is possible to reshuffle cores within the dropped multicore fibers. Most of the inputs/outputs of this switch are linked to the *MCF switch*; while a smaller portion of fiber cores are connected to the switches at the higher level, i.e. the *WDM switches*. In this way, it is possible to use the two space switches in combination to change the core arrangement and perform bypass at a fiber core granularity, thereby providing additional flexibility in the network. Moreover, by adding and dropping cores and performing switching at higher levels, the *fiber switch* allows for traffic to be repacked and added to the ring.

The *WDM switch* is a reconfigurable wavelength switch which allows for dynamic wavelength multiplexing and demultiplexing. The main reason for exploiting WDM is the need to provide support for long-lived connections that require a certain amount of bandwidth to be allocated over longer period of time. In the proposed architecture, we envision the use of several of those switches where a single switch is allocated per dropped/added core. Therefore, on the drop side, a single core is connected to this switch and the output ports are going to a WDM receiver on the TOR frontend, or to a *TDM switch* or are being bypassed to another *WDM switch* that performs multiplexing. Correspondingly, on the add side, a *WDM switch* has inputs coming from *WDM switches* that perform demultiplexing, *TDM switch* outputs and WDM transmitters on the TOR frontend side. Hence, WDM connections can be established by utilizing the switches that operate in the space and wavelength dimension only.

The *TDM switch* allows for optical subwavelength switching in the time dimension. The main motivation behind TDM is the improved bandwidth utilization through efficient resource allocation with subwavelength granularity. In order for a TDM connection to be established, the switches at each level should be properly configured. On the input side, this switch is connected to either an output of a *WDM switch* that performs demultiplexing or to a tunable TDM transmitter on the TOR frontend side. TDM switching in the optical domain offers similar functionality as OPS, however it pushes the buffering towards the network edges, thus eliminating the need for buffering in the network. In order to be able to switch at the right time instances, all TDM transmitters and switches in the network have to be synchronized. Considering that DCNs are a closed and tightly controlled environment, we assume that synchronization can easily be achieved.

B. Benefits of the proposed architecture

The proposed architecture has several advantages. As most of the traffic inside a DCN, i.e. around 75% is east-west traffic [1], it is crucial that high connectivity in a scalable way is provided for interconnecting the servers. The use of several multiplexing dimensions and technologies provides inherent support for scaling in different directions, while at the same time allowing connections with different granularity to be established. The hierarchical node structure results in highly aggregated traffic on the actual physical links between the nodes ensuring that a relatively simple physical topology i.e. reasonable number of nodes and physical links is retained even when the number of servers increases. In addition, the ring topology results in reduced number of links compared to a fat tree topology, thus providing the same connectivity with simplified link management. Initial inventory comparisons in [3] reveal that for supporting the same number of TORs, 45.8% fewer links and 50% fewer nodes are required for ring interconnection compared to a fat-tree topology. Furthermore, simulation results in [3] show 40%-99% improvement in connection request blocking and 3%-17% improvement in resource utilization compared to fat-tree topology-based DCN for varied available capacity and inter/intra cluster traffic ratio.

Another important feature is that bypass at lower levels enables significant offloading of the nodes, as bypass traffic does not require higher level processing and therefore does not occupy additional resources. As a result, since the bypass is done optically and paths are pre-established, connections experience only the physical propagation delay through the switch. Thus, in contrast to Ethernet switching, low latency can be achieved through optical bypass, while simultaneous switching of high capacity data such as whole fibers reduces the switching energy/bit, leading to energy efficient operation.

Unlike electrical switching, the multidimensional switching node is bit rate independent, meaning that higher bit rate operation per channel does not require any physical configuration upgrade. Changes in channel bandwidth and flex grid operation can be supported by simple reconfiguration of the WDM switch. In addition, in order to maximize the link utilization and optimize the resource allocation, connections can share a single wavelength in the time domain by using optical slots to transmit bursts of data. With precise synchronization among the transmitters and TDM switches, controlled transmissions and pre-established connections, TDM eliminates the need for buffering at the network nodes.

In order to optimize the resource allocation, known traffic patterns such as application specific traffic patterns as well as more long-term traffic patterns like “peak of day” traffic demands, “night and day” traffic pattern, seasonal patterns, etc. could effectively be exploited. In addition, flexibility and adaptability can easily be achieved by dynamically making changes in the allocation process by using an SDN controller.

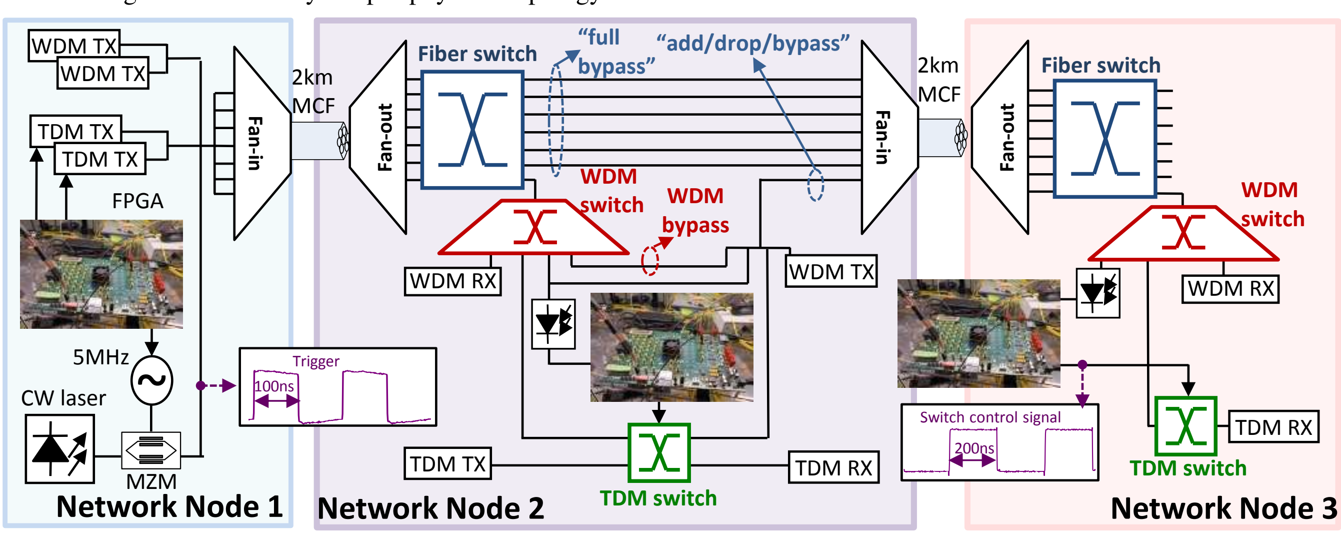

Fig. 2. Experimental setup for switching between three multidimensional nodes.

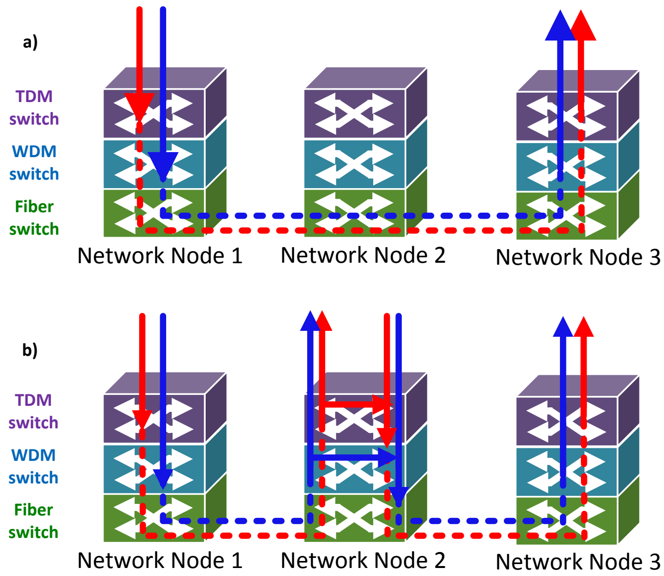

Fig. 3. Switching scenarios: a) full bypass and b) add/drop/bypass scenario.

With respect to the actual physical structure of the node, different deployments are possible. Although integration between any two or more levels is feasible, allowing for joint switching, individual switches at each level may have the advantage of higher flexibility when scaling the nodes. Moreover, physical connections can be assumed to TOR switches as in this work, or in a more long-term view, one could assume that the optical network nodes are the access point for the servers themselves, yielding a truly all-optical architecture. Pushing the electrical domain towards the end sides by introducing optical switching however implies that in order to compensate for loss, optical amplification becomes a necessity.

C. Experimental demonstration: setup and switching scenarios

The setup for the experimental demonstration is shown in Fig. 2. Three nodes interconnected with two links of 2-km 7-core fibers, are used to evaluate the system performance. Free space coupling devices and spliced fan-in/fan-out devices are used for spatial multiplexing/demultiplexing of the traffic in the two MCFs, respectively. At the first network node, NN1, generated traffic with both WDM and TDM granularity is aggregated and sent out towards NN2. The WDM channels are carrying 40 Gbit/s OOK modulated data each and are placed on a 100 GHz grid in the C-band. One of the WDM channels is assigned to carry two TDM channels to establish two different connections in an alternating fashion, by allocating every second time slot to one connection. The time slot width is 200 ns, out of which 190 ns are dedicated as a slot for transmission of useful data and 10 ns are used as a switching gap between bursts of data. The two bursts carry 40 Gbit/s data and are generated individually by using an FPGA to keep track of the slot status and the slots in which data transmission has been allowed for each TDM transmitter.

In addition, in order to enable synchronous operation of the TDM switches in the different nods, a 5 MHz trigger is modulated on a separate wavelength and propagated along

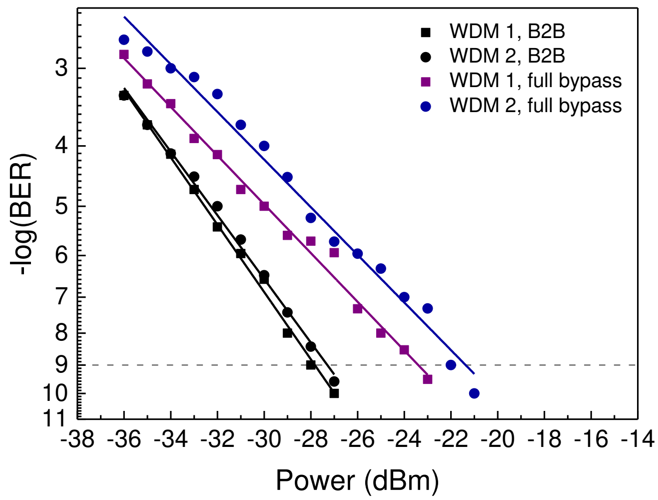

Fig. 4. BER performance of the WDM connections in the 'full bypass' scenario.

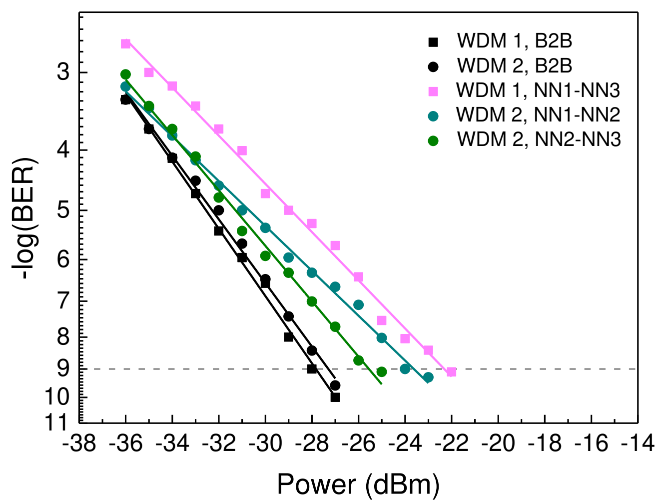

Fig. 5. BER performance of the WDM connections in the 'add/drop/bypass' scenario.

with the data channels. This signal is detected at each of the following nodes in the ring and used for synchronizing the TDM switches and transmitters. As the trigger is propagated through the same fiber as the TDM data it is received with a delay that corresponds to the propagation delay between the first node and any other node. The nodes are thus inherently synchronized to the right trigger count. At each node, the recovered TDM trigger is passed to an FPGA counter that is used to keep track of the slot status. Due to the delayed trigger, the nodes will have different slot status at a given moment in time, which however allows for slots to be uniquely identified by network elements such as path computation elements (PCE) or an SDN controller when connections are established. Hence, the controller may address a specific connection for example for configuration purposes by using the slot identifier for which the corresponding burst will be present at all the nodes on the predetermined path.

NN2 and NN3 are composed of switches operating in all three dimensions that are dynamically configured to demonstrate fully reconfigurable multidimensional switching. At the lowest level, a 48x48 beam-steering Polatis switch is

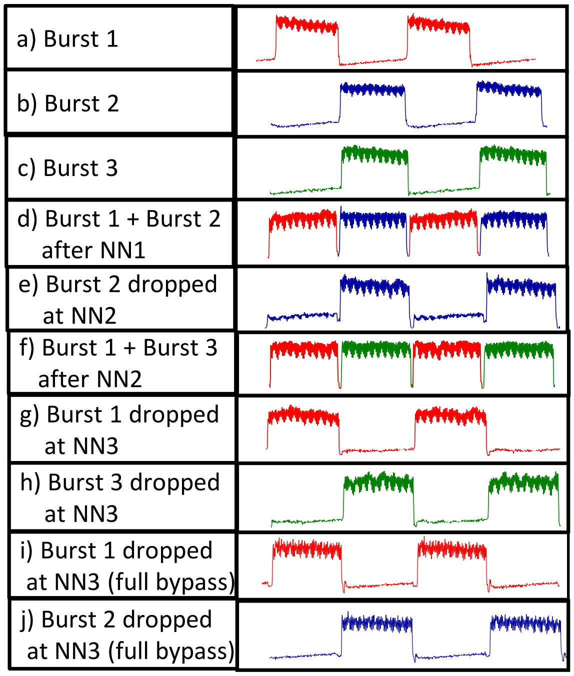

Fig. 6. Time domain traces of the bursts before and after switching.

used at both NN2 and NN3, logically segmented into two 24x24 sections. At the WDM and TDM levels, wavelength selective switches (WSS) and $LiNbO_3$ based 2x2 electro-optic switches are used, respectively.

Two switching scenarios are considered as shown in Fig. 3. In the first one, referred to as the ‘full bypass’ scenario, all the connections from NN1 are destined to NN3, so at NN2 all the connections are bypassed at a fiber core granularity using the space switch. In the second switching scenario i.e. the ‘add/drop/bypass’ scenario, the traffic generated and sent out from NN1 is partially dropped at NN2, partially bypassed at WDM and TDM level and in addition the dropped traffic is replaced with new connections being added at NN2. Then, the remaining traffic from NN1 and the newly added from NN2 are recombined and sent to NN3 where all the connections are terminated. The burst added at NN2 is generated by an FPGA control signal based on the propagated trigger, thus retaining the TDM synchronization.

The system performance was evaluated initially in a system with three data channels (two WDM connections and a single wavelength used to establish two TDM connections). Subsequently, the number of channels was scaled to 25 (24 WDM and one wavelength channel carrying two TDM connections) in order to demonstrate seamless scaling to 1 Tbit/s/core capacity or 7 Tbit/s throughput in a single MCF.

D. System performance

First, the performance of the system with three data channels is evaluated. The BER results for the WDM connections in the ‘full bypass’ and ‘add/drop/bypass’ scenario are shown in Fig. 4 and Fig. 5, respectively. In both cases, error-free performance (BER<10^{-9}) is achieved for all the connections. It can be seen that the experimental results confirm the preference of bypassing at lower levels in intermediate nodes, although the penalty of bypassing at WDM instead of SDM

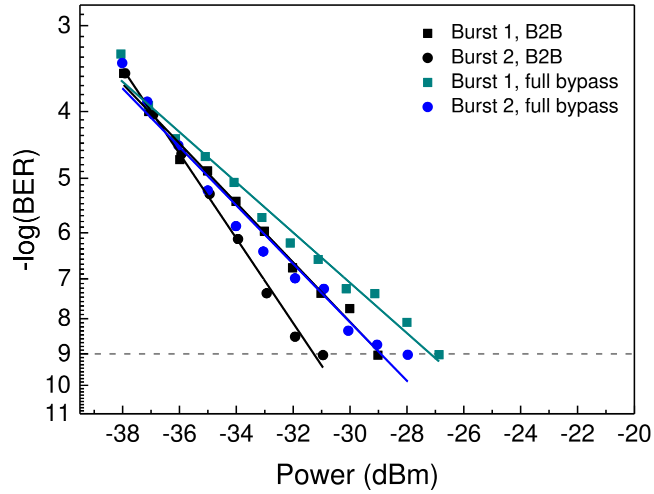

Fig. 7. BER performance of the TDM connections in the ‘full bypass’ scenario.

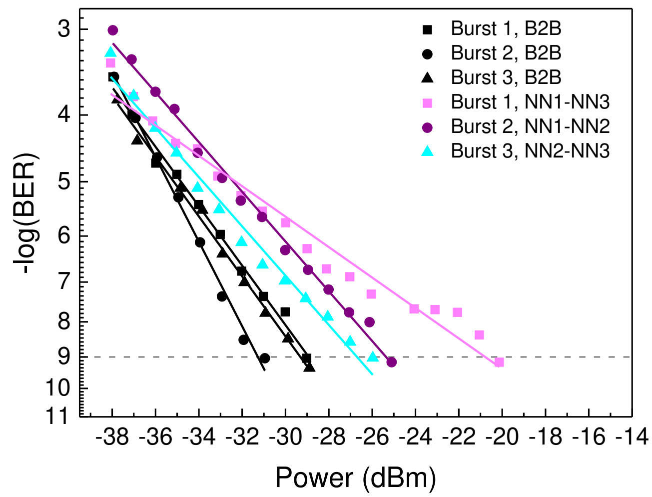

Fig. 8. BER performance of the TDM connections in the ‘add/drop/bypass’ scenario.

level is only 1 dB. This could be found useful for time critical connections that in the lack of resources, instead of being delayed may be assigned a path with slightly higher penalty.

Fig. 6 shows the time domain traces of the bursts generated and switched in the different switching scenarios. The BER results for the TDM connections in the ‘full bypass’ and ‘add/drop/bypass’ scenario are shown in Fig. 7 and Fig. 8, respectively. Again, all the connections have error-free performance (BER<10^{-9}). Similarly to the WDM case, bypass at lower levels in intermediate nodes is again preferable especially due to the limited suppression ratio of the TDM switches that results in additional penalty as a result of crosstalk. However, this penalty can easily be avoided or minimized by proper path allocation in which the number of intermediate TDM switches is limited.

Fig. 9 shows the spectra of all channels in all the cores taken at the input of NN3 when the system capacity has been increased to 1 Tbit/s/core. In this case the system performance is evaluated only in the ‘full bypass’ scenario. The measured receiver sensitivities (at BER=10^{-9}) of all 25 channels in a single core are shown in Fig. 10 (top). Moreover, in order to

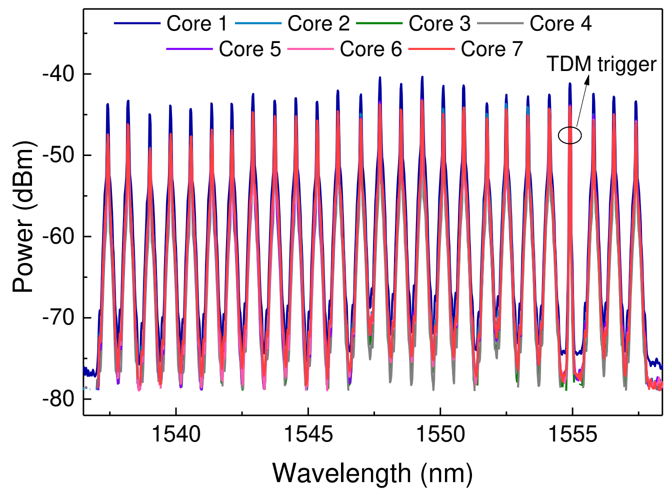

Fig. 9. Spectra of all the channels in all the cores received at NN3.

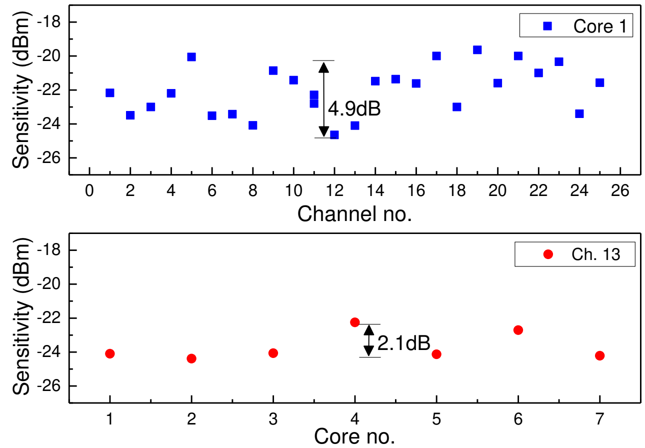

Fig. 10. Receiver sensitivity of all the channels measured in a single core (top) and receiver sensitivity of a single channel measured in all cores (bottom).

confirm that similar performance is achieved for all the channels in the remaining cores, the receiver sensitivity of a representative channel is measured in all the cores as shown in Fig. 10 (bottom). It can be seen that in one core the different channels have receiver sensitivity within 4.9 dB difference, while the receiver sensitivity of the single channel in the different cores is within 2.1 dB difference. It is important to emphasize that due to the random choice of core pairs in the two different multicore fibers the total loss and crosstalk per fiber core pair were not optimized. Fig. 11 shows the relative loss and crosstalk of each fiber core pair with respect to the loss and crosstalk of the pair denoted as core 1, on which the receiver sensitivity of all channels was measured. It can be seen that both the maximum loss difference as well as the maximum crosstalk difference between any of the different pairs remains within a 6 dB margin.

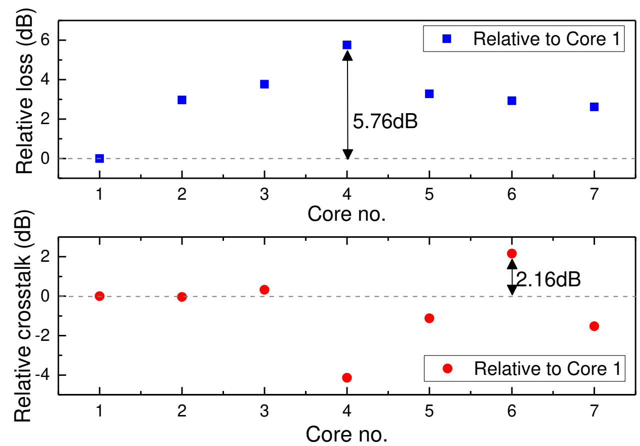

Fig. 11. Loss difference (top) and crosstalk difference (bottom) between all core pairs relative to pair core 1.

IV. Conclusion

We have presented a novel ring-based intra data center architecture that enables optical switching in several dimensions (space, wavelength and time). The proposed solution is highly attractive with respect to scalability, as it involves the use of different multiplexing technologies and network nodes that have the ability to switch connections with different granularity. The hierarchical node structure combined with the ring topology can accommodate large connectivity demands with reduced number of links and simplified cabling. Initial performance results indicate successful switching operation for all the connections in the several switching scenarios investigated. Achieving error-free performance (BER<10^{-9}) for all the channels for a scaled capacity of 1 Tbit/s/core confirms the system feasibility.

References

[1] Cisco Global Cloud Index: Forecast and Methodology, 2013–2018.

[2] V. Kamchevska *et al.*, "Experimental Demonstration of Multidimensional Switching Nodes for All-Optical Data Centre Networs," in *Proc*. ECOC 2015, Tu.1.2.2, Valencia, Spain, 2015.

[3] A. M. Fagertun *et al*., "Ring-based All-Optical Datacenter Networks," *in Proc*. ECOC 2015, P.6.9, Valencia, Spain, 2015.

[4] M. Al-Fares *et al.*, "A scalable, commodity data centre network architecture," *in Proc.* ACM SIGCOMM, pp. 63-74, 2008.

[5] F. T. Chong *et al*., "Data Center Energy Efficiency:Improving Energy Efficiency in Data Centers Beyond Technology Scaling," in IEEE Design & Test, vol.31, no.1, pp.93-104, Feb. 2014.

[6] D. Abts, M. R. Marty, P. M. Wells, P. Klausler, and H. Liu, "Energy proportional datacenter networks," in ISCA 2010, pp. 338–347, 2010.

[7] J. Koomey , "Growth in data center electricity use 2005 to 2010," Oakland, CA: Analytics Press, 2011.

[8] N. Farrington *et al*., "Helios: A hybrid electrical/optical switch architecture for modular data centers," *in Proc.* ACM SIGCOMM, pp. 339-350, 2010.

[9] G. Wang *et al*., "c-Through: part-time optics in data centers," *in Proc*. ACM SIGCOMM, pp. 327-338, 2010.

[10] P. Samadi, J. Xu, K. Bergman, "Experimental Demonstration of One-to-Many Virtual Machine Migration by Reliable Optical Multicast," *in Proc*. ECOC 2015, We.3.5.7, Valencia, Spain, 2015.

[11] K. Xi, Y.H. Kao, and H. J. Chao, "A Petabit bufferless optical switch for data center networks," in Optical Interconnects for Future Data Center Networks, C. Kachris, K. Bergman, and I. Tomkos, Eds. Springer New York, 2013, pp. 135–154.

[12] X. Ye *et al.*, "DOS—A scalable optical switch for datacenters," *in Proc.* ACM/IEEE Symp. Archit. Network. Commun. Syst., 2010, pp. 1–12.

[13] Z. Cao *et al.*, "Experimental demonstration of dynamic flexible bandwidth optical data center network with all-to-all interconnectivity," *in Proc*. ECOC 2014, PD.1.1, Cannes, France, 2014.

[14] G. S. Zervas *et al.*, "Time Shared Optical Network (TSON): a novel metro architecture for flexible multi-granular services," Opt. Express vol. 19, no. 26, pp. B509–B514, 2011.

[15] I. Widjaja and I. Saniee: "Simplified layering and flexible bandwidth with TWIN", *in Proc.* SIGCOMM, New York, USA, pp. 13-20, 2004.