A Programmable Si-Photonic Node for SDN-enabled Bloom Filter Forwarding in Disaggregated Data Centers

M. Moralis-Pegios[*a], N. Terzenidis[a], C. Vagionas[a], S. Pitris[a], E. Chatzianagnostou[a], A. Brimont[b], A. Zanzi[b], P. Sanchis[b], J. Marti[b], J. Kraft[c], K. Rochracher[c], S. Dorrestein[d], M. Bogdan[e], T. Tekin[e], D. Syrivelis[f], L. Tassiulas[g], A. Miliou[a], N. Pleros[a], K. Vyrsokinos[h]

[a]School of Informatics, Aristotle University of Thessaloniki, 54624, Greece,
[b]Nanophotonics Technology Center, UPVLC, Camino de Vera s/n, 46022, Valencia, Spain
[c]Austriamicrosystems AG, Unterpremstäten, Austria
[d]TE Connectivity, Rietveldenweg 32, 5222AR, s'-Hertogenbosch, Netherlands
[e]Fraunhofer Institute for Reliability and Microintegration, Gustav-Meyer-Allee 25, Germany
[f]Department of Electrical and Computer Engineering, University of Thessaly, Volos, Greece
[g]School of Engineering and Applied Science, Yale University, USA
[h]School of Physics, Aristotle University of Thessaloniki, 54624, Greece

[*]mmoralis@csd.auth.gr

ABSTRACT

Programmable switching nodes supporting Software-Defined Networking (SDN) over optical interconnecting technologies arise as a key enabling technology for future disaggregated Data Center (DC) environments. The SDN-enabling roadmap of intra-DC optical solutions is already a reality for rack-to-rack interconnects, with recent research reporting on interesting applications of programmable silicon photonic switching fabrics addressing board-to-board and even on-board applications. In this perspective, simplified information addressing schemes like Bloom filter (BF)-based labels emerge as a highly promising solution for ensuring rapid switch reconfiguration, following quickly the changes enforced in network size, network topology or even in content location. The benefits of BF-based forwarding have been so far successfully demonstrated in the Information-Centric Network (ICN) paradigm, while theoretical studies have also revealed the energy consumption and speed advantages when applied in DCs. In this paper we present for the first time a programmable 4x4 Silicon Photonic switch that supports SDN through the use of BF-labeled router ports. Our scheme significantly simplifies packet forwarding as it negates the need for large forwarding tables, allowing for its remote control through modifications in the assigned BF labels. We demonstrate 1x4 switch operation controlling the Si-Pho switch by a Stratix V FPGA module, which is responsible for processing the packet ID and correlating its destination with the appropriate BF-labeled outgoing port. DAC- and amplifier-less control of the carrier-injection Si-Pho switches is demonstrated, revealing successful switching of 10Gb/s data packets with BF-based forwarding information changes taking place at a time-scale that equals the duration of four consecutive packets.

Keywords: Bloom Filter, Silicon Photonic Switch Fabric, FPGA, Information Centric Network, Disaggregated Data Centers

1. INTRODUCTION

Disaggregation of resources is currently being promoted as a cost- and energy-efficient paradigm in DCs for supporting their continuously increasing size and performance requirements. Treating, however, computing, memory and storage as a pool of remotely located resources brings some new challenges to the DC networking equipment that has to ensure programmability and reconfigurability in the fastest possible way [1]. Disaggregation imposes also additional requirements in the routing of the data as new Virtual Machines (VMs) are established and existing VMs are migrating continuously in the DC network, irrespective of their physical location. In this respect, SDN over optical interconnect technologies arises as a key enabling technology for future DCs towards flexible network topologies that can handle the growing demand for bandwidth as well as the continuously changing DC topology and size [2,3].

The realization of VMs is an efficient way to keep stable the number of servers in DCs constant[4] with their number exploding now to millions in mega DCs. The drawback of this new trend is that the size of the look up tables in the routers is exploding as well with direct effect in their power consumption and requirement for enhanced memory speed. One of the promising solutions proposed so far to alleviate this problem is on the use of Bloom filter labeling schemes[5]. These are comprised of probabilistic data structures that can provenly offer memory, power and speed advantages over traditional look-up tables in modern DCs[6]. Bloom filter-based addressing has been favored in Information-Centric Networking as an efficient structure for denoting destination by means of the file content and not of the device. Recently also the scheme has been adapted also towards yielding significant benefits in flexible DC networking.

The programmable control plane of the switch requires, however, also an underlying interconnect infrastructure that can offer Tb/s-scale bandwidth at improved power- and cost-efficient envelopes compared to current electronics[6] throughout the whole hierarchical chain of the DC network. As optics are moving deeper into the DC hierarchy to board-to-board and even on-board environments, the use of Si-photonics switch technology is gaining momentum as the preferable solution due to their compatibility with the CMOS fabrication processes. In this context, various Si-based 2x2 switching elements have been demonstrated employing Mach Zehnder Interferometers [8-12], micro-ring resonator[13,14], ring-assisted MZIs [15,16] and micro-electromechanical systems [17].

Even though the latest research in Si photonics switch fabrics demonstrates promising results in terms of port count, insertion loss, speed and power consumption, their credentials at system-level flexibility and programmability have only recently started to be investigated. Recent works on photonic hardware-software co-development have been demonstrated focusing on optical memories [18], or optical switching [19-23]. Programmability has been shown in Si-based switching matrices offering reconfigurability to one out of the several MZIs included in the matrix, the main target being the power distribution for multicasting purposes[23].

In this paper, we demonstrate a programmable Si photonic node performing BF-based forwarding and being formed by a Si-based 4x4 electro-optic carrier injection SM directly controlled by an amplifier-less and DAC-less FPGA board. The FPGA provides the BF labels assigned to the SM's port for the rapid reconfiguration of the DC network following updates in the DC as devices can be added, removed or simply change physical location without necessitating the update of large look-up tables. The header of the packets injected into the SM are processed in order to check if they are members of the BF label set and if the result is positive, the FPGA generates the corresponding control signals driving the right 2x2 switching elements. The proposed work demonstrates the highest number of FPGA-controlled MZI elements among the state-of-the-art programmable Si-photonic switches, reporting on three out of the six MZIs as FPGA-controllable elements. The control signals are revised simultaneously with the change of the lookup table that takes into account the updated BFs, emulating a change in the network topology. The proposed BF labeling of the switch's ports is demonstrated in a 1x4 scenario with packets at 10Gb/s line rate successfully routed to the desired output port.

2. CONCEPT

Modern DCs may have thousands or even millions of end-host MAC addresses assigned to active Virtual Machines, creating a constant need to maintain and update constantly large lookup tables. Moreover, the rapid increase of link speeds, requires even-faster and hence, more expensive and more power "hungry" memory for the forwarding tables. These tight requirements mandate that new methods should be developed for the representation of the forwarding tables in each node, such as Bloom Filters.

A BF is a hash-based method for representing a set of elements to support membership queries. They were invented by Bloom[24] and, among other network applications [4], have been proposed as an efficient way of labelling packets or destinations, simplifying header processing to a few basic logic operations and avoiding the use of complex look-up functions [6]. Their employment has been mainly promoted so far in Information-Centric networking schemes, where files and data, instead of devices, are designated with certain destination addresses [25,26]. In DC network environments, BF-based forwarding has been already shown to significantly reduce the memory size and cost requirements and to improve packet forwarding performance, yielding scalable infrastructures where devices can be added, removed or simply change physical location without necessitating the update of large look-up tables [6].

The BF labels are created by allocating to a vector *v* of *m* bits, initially set to 0, and then choose *k* independent hash functions each with range {1,…,m}. For each element that we want to group the bits at positions h_1 (a),…h_k(a) in *v*

are set to 1. To query, if an element *b* is a member of this BF we check the bits at positions $h_1(b),\ldots,h_k(b)$. If any of them is 0 then *b* is certainly not a member of the group. Otherwise, we can conjecture that *b* belongs to this group, even though there is a possibility that this is a wrong assumption, called a "false positive". For n numbers of elements, the probability of this false positive occurring show good congruence with the approximation of $P = (1-(1-\frac{1}{m})^{nk})^k$. In this respect choosing k and m according to network requirements can modify the probability of false positive.

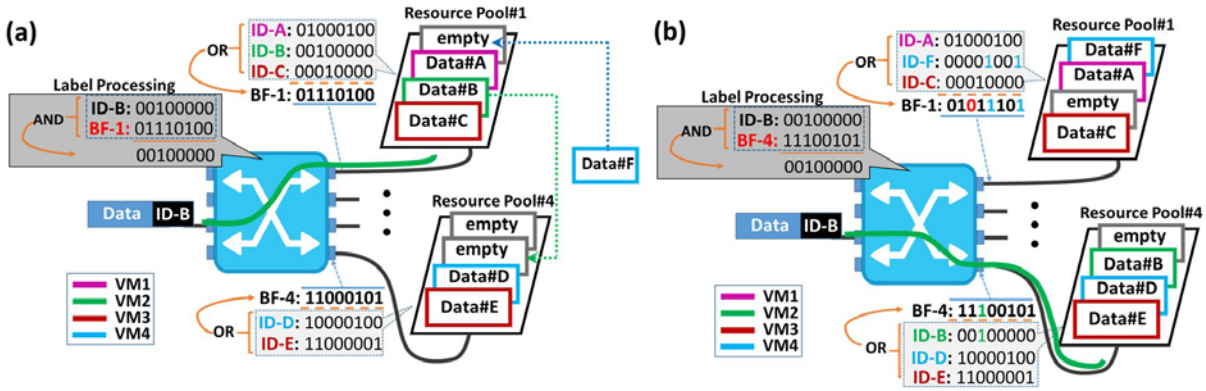

Figure 1. Schematic illustration of the concept (a) Data#B residing at Resource Pool#1 and (b) with Data#B having migrated to Resource Pool#1 and Data#F having been added i Resource Pool#1.

The proposed BF labelling concept for DCs is schematically illustrated in Figure 1 where data content, and not physical location, is associated with a designation address following the Information-Centric networking paradigm. Figure 1 (a) shows a switch with 4 output ports, where every port connects to a physical resource pool, which can encompass computing, memory and/or storage modules usually termed as bricks, in the DC environment. Data sections including specific content are labelled with an ID irrespective of their physical location. The length of the ID, and as a consequence the length of the resulting bloom filter, is chosen so that the possibility of false positives approaches effectively 0. The switch port connecting to the set of data residing on the same physical resource pool (i.e. a tray) is addressed by a BF stemming from a bit-wise OR operation between the IDs of the data sections. For example, BF-1 originating from the OR operation between the IDs of Data-A and Data-C of the Resource Pool#1 becomes the address of switch port#1. Whenever an incoming packet destined for a certain Data segment enters the switch, the packet header carrying the destination address ID is logically ANDed with the four BFs corresponding to the four switch outputs. In case the result of this AND operation is identical to the packet header, it means that the packet header is most probably a member of the BF and the packet is forwarded to the respective switch port. Figure 1 (a) depicts the case where the incoming packet wants to access the Data-B segment, showing also that the subsequent VM reconfiguration will require the movement of Data-B from Resource pool#1 to Resource pool#4 as well as the entry of a new Data Set (Data-F) into an empty segment of the Resource Pool#1. Once the VM reconfiguration is completed, the new BFs corresponding to the switch ports are calculated and are fed to the switch as its new forwarding table information so that the incoming packets can continue to be directed to the desired data segment. Figure 1(b) depicts the example of a new packet entering the switch and destined again for Data-B. Its ID is now being processed and is found to be a member of the new BF corresponding to switch port#4. So, it becomes again successfully switched to Data-B that are now located at Resource Pool#4. This approach is in principle compatible with the virtual addresses assigned to memory pages in current computing schemes, where pages can be distributed among several memory modules.

3. EXPERIMENTAL SETUP AND RESULTS

The 4x4 Si-Photonic switch used for the experimental demonstration was fabricated at AustriaMicroSystems AG foundry using standard CMOS processes in 200mm wafers with 220nm x 400nm Si waveguides on top of 2um SiO_2. The slab height is 150nm. The switch is composed from six MZI switches arranged in a Benes architecture, as illustrated in Figure 2 (a), providing non-blocking operation during packet forwarding. The chip is mounted on a high speed Printed Circuit Board (PCB) that fans out the pads from the phase shifters to SMA connectors placed at the edge of the PCB.

Figure 2 (b) depicts a photo of the packaged device. Two electrical phase shifters (PS) are employed per interferometer, accounting for twelve in total.

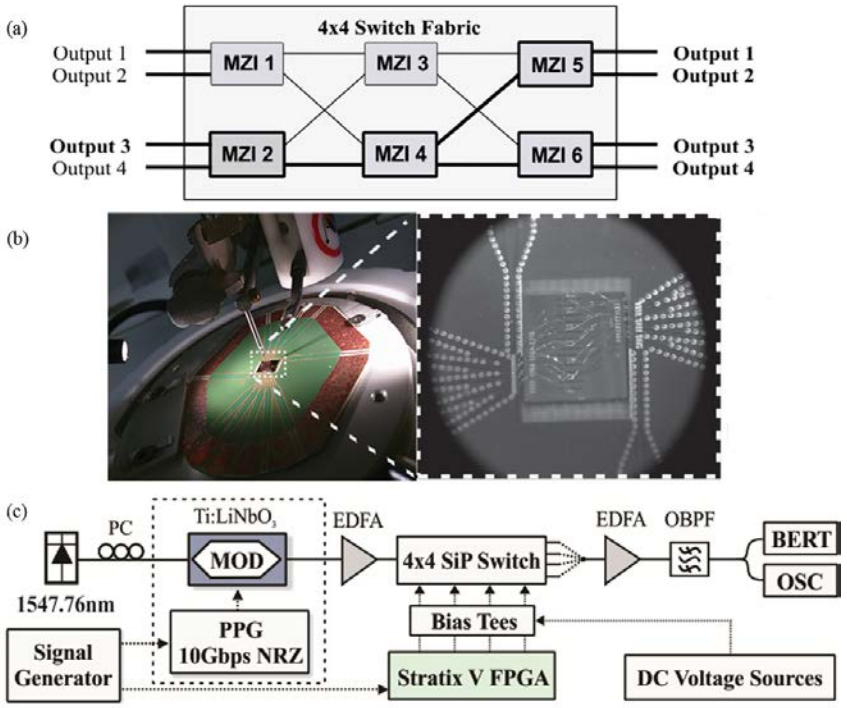

Figure 2. (a) The 4x4 Benes switch architecture depicting in bold the switches that were used in the 1x4 programmable node demonstration, (b) Packaged SiP chip (c) Experimental setup used for experimental validation of the proposed network scenario.

The concept described in detail in Section 2 was verified with the experimental setup illustrated in Figure 2 (c). A CW laser beam at 1547.76 nm was modulated via a $LiNbO_3$ modulator driven by a PPG so as to produce 200-bitlong optical packets encoded with 2^7-1 NRZ data pattern at 10 Gb/s. The successive packets were placed apart with a guard band of 7.3 nsec in order to avoid imperfect switching during the switch ON/OFF times. The packet stream was subsequently amplified and was inserted to the switch via its input port#3. After exiting the switch, the optical packets were again amplified and filtered out with a 0.8 nm bandpass filter prior being recorded at a digital sampling oscilloscope and evaluated with a BER tester. The switch chip was controlled by a Stratix V FPGA (Field Programmable Gate Array) board that provides control signals with a peak-to-peak voltage of ~1.1 V. The six MZIs were biased at the optimum operating point with an array of power supplies. The mixing of the RF and DC electrical signals was performed with six Bias-T that were connected to the SMA connectors leading to the upper phase shifter of each MZI. In the electrical signal path there were 1dB losses from the Bias-T and an additional 1dB from the RF cables and the connectors. After losses the electrical pulsed signal applied to the phase shifters had a peak power of ~0.9V that was marginal adequate for pi phase shift. The DC bias was set-for all MZIs at the operating point where the highest possible extinction ratio between the two output ports was observed.

In the demonstrated scenario MZI #4, #5 and #6 out of the six were controlled by respective FPGA outputs, while MZI#2 was DC biased to perform at its OFF state. The headers of the packets are assumed to be also in the same line rate of 10Gb/s and to arrive at the FPGA over a dedicated and independent electrical control channel. However, the concept is fully compatible with transmission schemes of optical header information, either when using dedicated optical address/control channels[22-26] or when carrying the optical header at the front-end of the optical packet payload [27], provided of course that optoelectronic conversion takes place prior launching the header in the FPGA. Rows 6,7 and 8 illustrate the respective control signals produced by the FPGA for driving MZIs 4-6 (CTR_MZI4, CTR_MZI5 CTR_MZI6) after processing the packet header information and the port ID. During Phase #1, each of the four packets is routed to a different output port based on its header information and the respective Port ID. At the end of Phase #1, the four BFs corresponding to the respective port IDs get a new label, resembling the case of a change in the network

topology and the new location of data blocks. This results into a different routing path for each packet even though every packet retains its header during both Phases. For example, the packet with header '11001001' is routed to Output port 1 during Phase 1 and to Output Port 2 during Phase 2. Figure 4 indicates proper routing functionality and switch re-configurability for all packets during both Phases. The ON/OFF suppression ratio between packets at the output of the SM was measured to be 7 dB, mainly owing to the sub-optimal switching obtained at MZI#6.

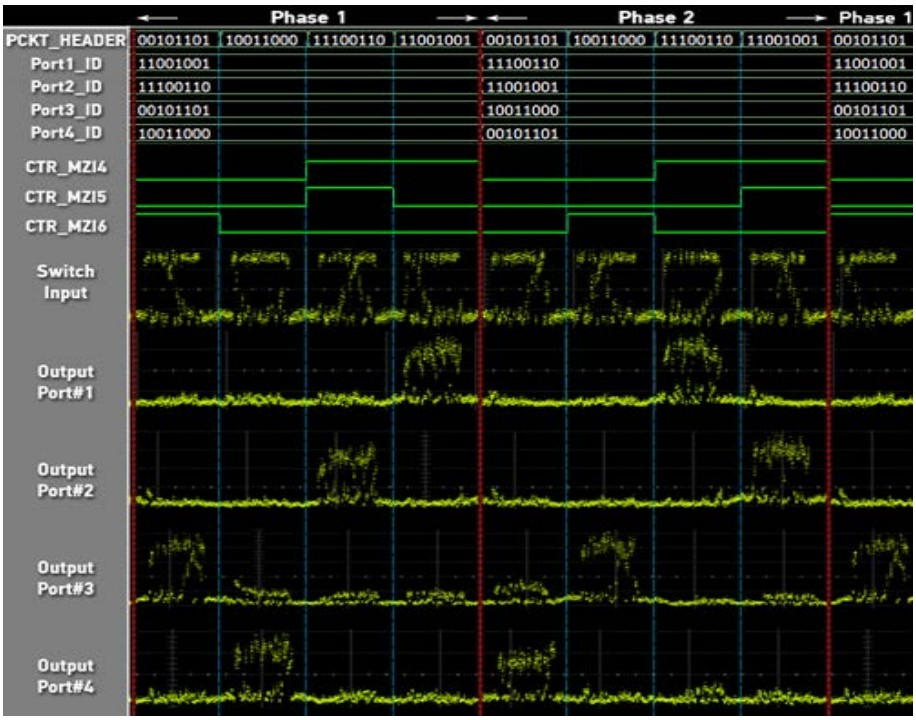

Figure 4. Experimental packet trace results

BER measurements were performed additionally for the evaluation of the signal quality through the switch matrix in packet forward mode with the BF concept. The 10 Gb/s packets were inserted in port 3 of the SM and routed to all output ports. The BER measurements illustrated in Figure 5 revealed an error floor at 10^{-4}, which mainly stems from the limitations on the maximum current drawing enforced by the Pseudo Current Mode Logic (PCML) interface of the FPGA output transmitter channels. The PCML interface constrained the current to values lower than 10mA, which translates to 9mW of maximum electrical power applied to each MZI. This power ensures optimal switching only for MZI#5, according to Table I, while all other switches required higher power levels for altering their state. However, the on-chip integration of thermal heaters for biasing the MZIs would reduce these requirements and could eventually render the 9mA current value sufficient for a complete pi phase shift and as such for improved BER performance.

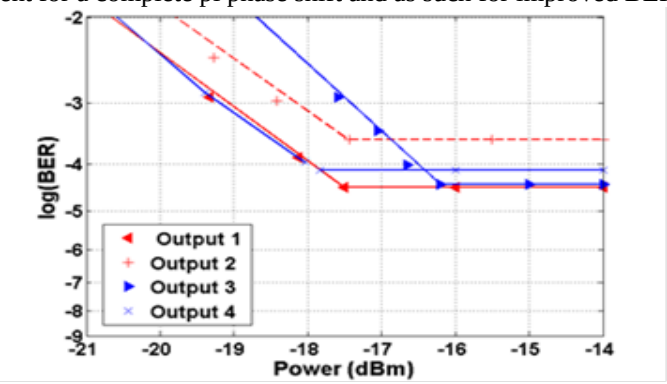

Figure 5. BER measurements

4. CONCLUSION

We have experimentally demonstrated a programmable Si photonic node formed by the direct control of a Si based 4x4 electro-optic switching matrix from a high speed FPGA board. The FPGA is performing the BF checking of the incoming packets headers and provides the corresponding electrical control signals required to drive the right 2x2 switching elements in the SM. For the first time, the FPGA board controls three out of the six MZIs in the 4x4 SM directly for successful routing of 10 Gb/s packets to the desired output port in a 1x4 packet forwarding. This work can be in principle extended in 32x32 or even higher radix Si photonic matrixes for disaggregated DC networks of larger dimensions, where BF based forwarding is expected to provider even higher performance benefits.

5. ACKNOWLEDGMENTS

This work has been supported in part by the European Commission through FP7-ICT PHOXTROT (contract no. 318240).

6. REFERENCES

[1] K. Lim, et al., "Disaggregated memory for expansion and sharing in blade servers." ACM SIGARCH Computer Architecture News. Vol. 37. No. 3. ACM (2009).

[2] W. Miao, et al., "SDN-enabled OPS with QoS guarantee for reconfigurable virtual data center networks." J. of Optical Communic. and Network, vol. 7, no. 7, pp. 634-643 (2015).

[3] P. Jokela, et al., "LIPSIN: line speed publish/subscribe inter-networking." ACM SIGCOMM Comp. Communicat. Review, vol. 39, no. 4, pp. 195-206 (2009).

[4] http://www.datacenterknowledge.com/archives/2016/06/27/heres-how-much-energy-all-us-data-centers-consume

[5] A. Broder and M. Mitzenmacher, “Network applications of Bloom filters: A survey,” Internet Mathematics, vol. 1, no. 4, pp. 485–509 (2005).

[6] M. Yu, A. Fabrikant, and J. Rexford. "BUFFALO: Bloom filter forwarding architecture for large organizations." Presented at the 5th ACM CoNEXT, Rome, Italy, Dec. (2009).

[7] D. AB. Miller, "Rationale and challenges for optical interconnects to electronic chips." Proceedings of the IEEE 88.6, pp. 728-749 (2000).

[8] Lu, et al., "16× 16 non-blocking silicon optical switch based on electro-optic Mach-Zehnder interferometers." Opt. Express, vol. 24, no. 9, pp. 9295-9307, (2016).

[9] B. G. Lee, et al., "Monolithic silicon integration of scaled photonic switch fabrics, CMOS logic, and device driver circuits." J. of Lightwav. Technol. Vol. 32, no. 4, pp. 743-751, (2014).

[10] J. Xing, et al., "Nonblocking 4× 4 silicon electro-optic switch matrix with push–pull drive." Opt. Lett., vol 38, no. 19, pp. 3926-3929 (2013).

[11] M. Yang, et al. "Non-blocking 4x4 electro-optic silicon switch for on-chip photonic networks." Opt. Express vol. 19, no. 1, pp. 47-54 (2011).

[12] K. Suzuki, et al. "Ultra-compact 8× 8 strictly-non-blocking Si-wire PILOSS switch." Opt. Express, vol. 22, no. 4, pp. 3887-3894, (2014).

[13] R. Ji, J. Xu, and L. Yang. "Five-port optical router based on microring switches for photonic networks-on-chip." Photon. Technol. Lett., vol. 25, no .5, pp. 492-495 (2013).

[14] Sherwood-Droz, Nicolás, et al. "Optical 4× 4 hitless silicon router for optical networks-on-chip (NoC)." Opt. Express vol. 16, no. 20, pp.≈ 15915-15922 (2008).

[15] L. Lu, et al., "Low-power 2× 2 silicon electro-optic switches based on double-ring assisted Mach–Zehnder interferometers." Opt. Lett., vol. 39, no. 6, pp. 1633-1636 (2013).

[16] L. Lu, et al. "Silicon Optical Switches Based on Double-Ring-Assisted Mach–Zehnder Interferometers." Photon. Technol. Lett., vol. 27, no.23, pp. 2457-2460 (2013).

[17] T. J. Seok, et al., "Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers." Optica, vol. 3, no. 1, pp. 64-70 (2016).

[18] T. Shiraishi, et al. "A reconfigurable and redundant optically-connected memory system using a silicon photonic switch." Presented at OFC 2014, San Francisco, CA, USA, Th2A.10

[19] Y. Xiong, et al. "Towards a Fast Centralized Controller for Integrated Silicon Photonic Multistage MZI-based Switches." Presented at OFC 2016, Los Angeles, CA, USA, W1J.2

[20] D. Calhoun, et al., "Dynamic reconfiguration of silicon photonic circuit switched interconnection networks." Presented at IEEE High Perform. Extreme Comput. Conf. 2014, Waltham, MA, USA

[21] F. Lou, et al., "Towards a centralized controller for silicon photonic MZI-based interconnects." 2015, presented at IEEE Optical Interconnects Conference (OI), IEEE, 2015, San Diego, CA, USA

[22] D. Calhoun, et al., "Programmable wavelength locking and routing in a silicon-photonic interconnection network implementation." Presented at OFC2015, Los Angeles, CA, USA, Tu2H.3

[23] C.P. Chen, et al., "Programmable Dynamically-Controlled Silicon Photonic Switch Fabric," J. Lightwav. Technol., vol. 34, no. 12, pp. 2952-2958 (2016).

[24] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors." Communications of the ACM, vol. 13, no. 7, pp. 422-426 (1970).

[25] W. You, et al., "Dipit: A distributed bloom-filter based pit table for ccn nodes." Presented at ICCCN 2012, Munich, Germany

[26] D. Syrivelis, et al., "Pursuing a software defined information-centric network." 2012 European Workshop on Software Defined Networking. IEEE (2012).