Optically-enabled Bloom Filter Label Forwarding using a Silicon Photonic switching matrix

N. Terzenidis, M. Moralis-Pegios, C. Vagionas, S. Pitris, E. Chatzianagnostou, P. Maniotis, D. Syrivelis, L. Tassiulas, A. Miliou, N. Pleros, K. Vyrsokinos

***Abstract*—Simplified forwarding schemes relying on Bloom Filter (BF)-based labels emerge as a promising approach for coping with the substantial increase in lookup table memory requirements associated with the growing number of end-hosts in DataCenters. In this paper, we present for the first time the successful implementation of a BF-label forwarding scheme over a silicon photonic switch fabric and we demonstrate its functionality with 10Gb/s data packets that carry BF-encoded labels. The optically enabled BF-label forwarding setup utilizes a Si-based 4x4 electro-optic switch directly controlled by an amplifier-less and Digital-to-Analog-Converter (DAC)-less high speed Field Programmable Gate Array (FPGA) board. The FPGA is responsible for extracting the BF-label from the incoming packets and for carrying out the BF-based forwarding function, determining the appropriate switching state towards conveying incoming packets to the desired output. The use of BF-label forwarding allows for rapid switch reconfiguration avoiding the need for large look-up table updates as the network topology changes and devices are added, removed or simply change physical location. Successful operation for 10 Gb/s data packets has been obtained for a 1x4 routing layout.**

***Index Terms*—Optical switching devices, Optical communications, Silicon Photonics, Mach-Zehnder Interferometers, Bloom filter, Router forwarding, Data Centers, Dissagregated Data Centers.**

I. Introduction

Current DataCenter (DC) architectures rely on the use of server blades as their basic building blocks, with every blade carrying all types of resources like processor, memory and often also storage modules. However, the significant heterogeneity in resource usage per machine and workload[1, 2] leads to considerable resource underutilization, severely impacting cost- end energy-efficiency. This has brought forward the concepts of resource disaggregation and rack-scale computing [3-5], investing in a finer granularity and higher modularity DC architecture in order to gain significant cost and energy savings [6, 7]. This approach aims at the use of computing, memory and storage as distinct pools of homogeneous resources that are synergized into a DC via interconnecting the different pools together [8, 9].

This puts, however, additional pressure to the already congested DC networking equipment. Interconnect links have to ensure high-bandwidth, low-latency and low-energy communication in order to sustain performance despite the increased physical distances between resource blocks [5, 8, 10]. At the same time, the transition to finer-granularity resources enforces a significant increase in end-hosts and respective network addresses. With modern mega DCs comprising already tens or even hundreds of thousands end-hosts, this perspective will require a huge amount of memory space reserved by the routing look-up table in case traditional forwarding and routing schemes based on the hard-coded MAC host-address are employed [11-13]. Moreover, this memory space has to allow for efficient routing look-up table reconfiguration and upgrade once end-hosts are changing location or new hosts are added, as it is typically the case in Virtual Machine migration and network maintenance or extension, respectively [11].

The solution for high-bandwidth, low-latency and low-energy interconnection is currently being sought in the area of optics [14-19], while the memory space challenge can be addressed by simplified forwarding schemes relying on Bloom filter labels [11, 13, 20]. Advances in silicon photonics (Si-Pho) have rendered Si-based switch fabrics as a highly promising candidate for on-board and board-to-board connectivity, offering fast response times [21-

25] and a low-cost integrated perspective due to their CMOS compatible fabrication processes. Si-pho switch fabrics have emerged also recently in multi-casting routing schemes [24] and in programmable setups, exploiting photonic hardware-software co-development efforts [24, 26-31] in order to gradually enrich their portfolio for supporting the increased dynamicity and reconfigurability required in DC environments.

At the same time, the use of Bloom-filter (BF)-based forwarding schemes has been shown to offer successful intra-datacenter routing with reduced memory requirements [11, 20] and enhanced multicasting credentials [13] even in large DC enterprises. BFs comprise probabilistic data structures that encode a set of elements into a single string, so that BF-based forwarding encodes either the set of end-host addresses serviced by the next hop[11] or the forwarding identifiers of a certain delivery tree [20, 32]. This simplifies the look-up process into a membership check procedure and allows for storing a complete set of end-host addresses into a single memory entry, facilitating also look-up table updates once network topology changes are taking place. BF-label forwarding in large DC organizations has been shown to offer a reduction of at least 65% in fast memory requirements and a performance boost of 10% compared to a standard Ethernet packet forwarding scheme [11]. However, BF-label forwarding has so far never co-operated with an underlying photonic switch fabric towards realizing optically enabled solutions with efficient and reconfigurable intra-DC forwarding. In our recent work [33], the credentials of a Si-Pho switching platform to respond successfully on look-up table changes were experimentally verified, using, however, BF-encoded destination addresses that were statically pre-assigned in the FPGA and not carried by the individual optical packets, as would be required in a real use case scenario.

In this paper, we extend our previous work [33] and we demonstrate the complete operation of a BF-label forwarding optically-enabled node using a programmable Si-pho switching matrix and 10Gb/s incoming optical packets that carry their BF-encoded destination address as in-packet label. This optically-enabled BF-label-based forwarding scheme separates the optical BF-encoded address from its packet payload at the FPGA entry prior comparing the BF-encoded address with the look-up table entries and forwarding the optical packet to the correct switch output port. The proof-of-concept demonstration was performed using a Si-based 4x4 electro-optic carrier injection switching matrix directly controlled by an amplifier-less and DAC-less FPGA board, which mimics the role of an actual control plane. Optical packet headers are received by the FPGA prior carrying out their membership checking in the look-up table implemented at the FPGA. The look-up table includes the BF strings assigned to all switching matrix outgoing ports and can get updated as end-host devices are added, removed or simply change physical location, enforcing a simultaneous switch matrix reconfiguration for directing the incoming packets to their desired outgoing port. Our optically-enabled BF-based forwarding scheme is demonstrated in a 1x4 switching scenario with 10Gb/s data packets that are successfully routed to the desired output port even when the look-up table content is updated every four packets in order to reflect network topology changes.

This paper is organized as follows: Section II describes the main principles of the Bloom filter structures, the concept of our BF-labelling scheme and its utilization in disaggregated DC networks. Section III presents the experimental setup, while the respective results are described in Section IV. Section V concludes the paper.

II. Concept

Bloom-Filter is a hash-based method for representing a set of elements to support membership queries. They were invented by Bloom [34] and, among other network applications [35], 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 [11]. Their employment has been promoted so far in Information-Centric networking [36, 37] and more recently also in DC network environments [11, 20, 35, 38], where they revealed significant reduction in memory size and cost requirements [11].

The BF labels are created by allocating 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, *a* 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 $h_1(b)…h_k(b)$. 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 $P = (1 - \left(1 - \frac{1}{m}\right)^{nk})^k$. In this respect choosing *k* and *m* according to network requirements can minimize the probability of false positive [34].

An example for better understanding the use of Bloom Filters is illustrated in Fig. 1, where an 8-bit vector along with 3 hash functions h_1, h_2 and h_3 are considered. Fig. 1(a) depicts the construction process of a Bloom filter containing three different data elements in two stages. Initially, the representation for each element is generated using the hash functions and subsequently the three elements are encoded in a BF using a bitwise-OR operation. Fig. 1(b) illustrates the inverse operation where a query for a single element (Data#B) is performed in the BF produced in Fig. 1(a). A bitwise-AND operation is performed between the Bloom filter and the element to be checked, with the result of this operation defining whether the element is a member of the Bloom Filter. In case the result is identical to the element's representation then the element is *possibly* a member of the Bloom Filter, otherwise the element is *definitely* not a member.

The proposed BF labelling concept for DCs is schematically illustrated in Fig. 2 where each resource is associated with a designation address or *ID* following a traditional packet switching scheme. Fig. 2(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. The length of the ID, and therefore the length of the resulting bloom filter, has to be chosen so that the possibility of false positives approaches effectively 0. The switch port connecting to the set of

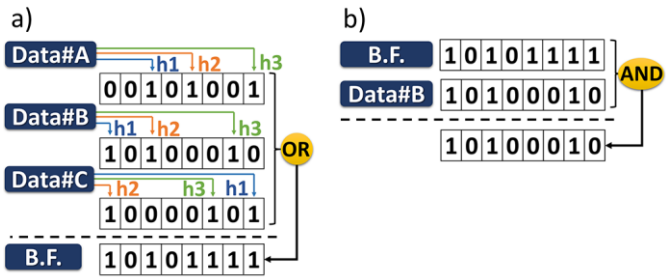

Fig. 1: (a) Construction process of a Bloom filter containing three different data elements, (b) Inverse operation where a query for a single element (DATA#B) is performed on the BF

resources residing on the same physical resource pool (i.e. a tray or a rack incorporating several processor or memory or storage nodes) is addressed by a BF stemming from a bit-wise OR operation between the IDs of these resources. For example, BF-1 originating from the OR operation between the IDs of Res-A, Res-B and Res-C of the Resource Pool#1 becomes the address of switch port#1. Whenever an incoming packet destined for a certain Resource 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. Fig. 2(a) depicts the case where the incoming packet wants to access the Res-B segment. Its ID is being processed and is found to be a member of the BF corresponding to switch port#1. A subsequent network topology reconfiguration that will require the movement of Res-B from Resource pool#1 to Resource pool#4 as well as the entry of a new Resource (Res-F) into an empty segment of the Resource Pool#1 is presented

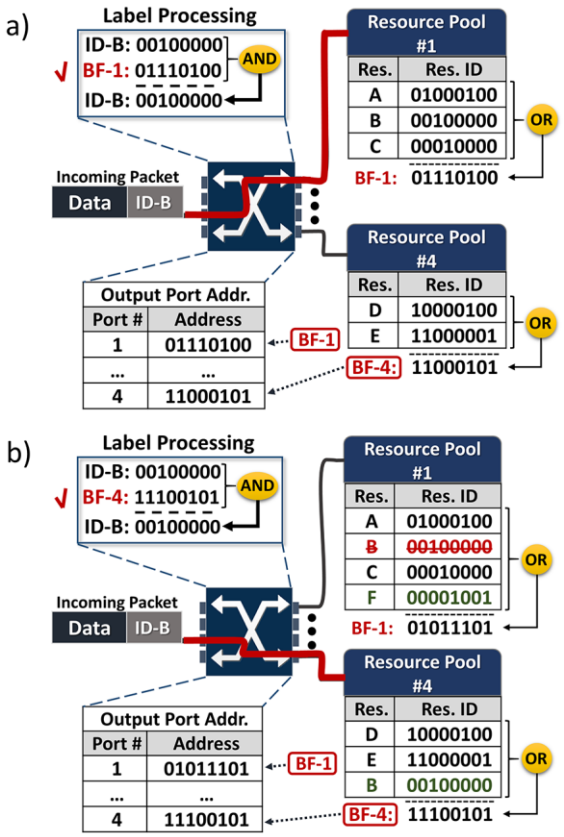

Fig. 2: Schematic illustration of the concept (a) with Data#B residing at Resource Pool#1 and (b) with Data#B having migrated to Resource Pool#1 and Data#F having been added in Resource Pool#1.

in Fig. 2(b). Once the topology 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 resource. Fig. 2(b) also depicts the example of a new packet entering the switch and destined again for Res-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 Res-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.

III. EXPERIMENTAL SETUP

The concept described in detail in Section II was verified with the experimental setup illustrated in Fig. 3. A CW laser beam at 1547.76 nm was modulated via a $LiNbO_3$ modulator driven by a Pulse Pattern Generator (PPG) so as to produce 200-bit

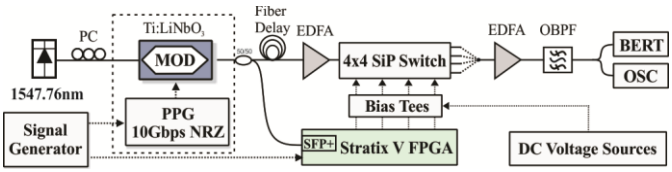

Fig. 3: Experimental setup used for experimental validation of the proposed network scenario.

long optical packets at 10 Gb/s. The packets were comprised by predefined 8-bit headers and 127bit payload encoded with 2^7-1 NRZ data pattern. Successive packets were placed apart with a guard band of 6.5 nsec in order to avoid imperfect switching during the switch ON/OFF times. An initialization sequence, equivalent to 2 consecutive K28.5 8b/10b codewords, is sent to provide word alignment for the FPGA. A 56-bit pattern of alternating 0 and 1 bits, equivalent to the Ethernet preamble, precedes the data packets to establish receiver clock synchronization at the beginning of the demonstration, while the continuous data transmission realized subsequently retains synchronization throughout the experiment. The guard-band duration of 6.5ns (65bits @ 10Gb/s) were selected so as to be within the maximum running length of 66 bits that is supported by our SFP+ module in case 64/66b encoding of 10G Ethernet is used.

The packet stream was split by a 50/50 coupler into a fiber destined to the input port#3 of the 4x4 switch and a fiber destined to the FPGA. The optical stream destined to the FPGA was received by a commercial SFP+ module, where the information was converted to its electronic format prior performing the necessary deserialization, word-alignment and synchronization functions by the FPGA transceiver circuit. The 32-bit de-serialized data stream was subsequently forwarded to the BF-label processing functions in order to generate the appropriate control signals for the switch. The required label bits were extracted from the data packet and compared with the BF-labels of the switch output ports according to the procedure described in Section II. Every label bit is compared with the respective bit of the BF-labels assigned to the switch ports that are stored in a look-up table. In the event of a network topology reconfiguration, only this look-up table has to be updated with the new BF-labels. A separate look-up

table contains the switch configuration settings for every input-output port combination. When a packet label matches the BF-label of a specific output port, the appropriate MZI control signals for successful forwarding are derived from this look-up table and are activated prior being serialized by the FPGA transceiver circuit and forwarded to the corresponding MZIs. The latency introduced by the FPGA following the data reception and until generating the switch control signals was measured to be approximately 452 ns, with 444ns originating from the transceiver circuit and its constituent SerDes, word-alignment and synchronization functions and just 8ns or 2 cycles originating from the BF-label processing operations. To counterbalance the FPGA BF-label forwarding latency, the optical packet stream was delayed in a fiber delay line prior entering the integrated switch. After exiting the switch, the optical packets were amplified and filtered in a 0.8 nm bandpass filter prior being recorded at a digital sampling oscilloscope and evaluated with a BER tester.

The 4x4 Si-pho 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 Fig. 4(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.

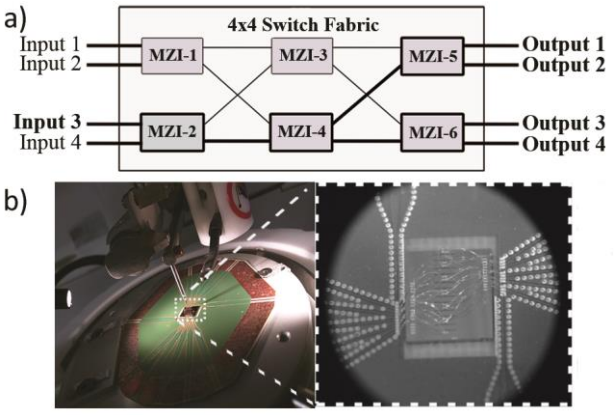

Fig. 4: (a) The 4x4 Benes switch architecture depicting in bold the switches that were used in the 1x4 programmable node demonstration, (b) Packaged Silicon photonic chip

Due to space limitations and the small pitch of the pads, the wire bonds from the phase shifters were in close proximity to the grating couplers, prohibiting this way the use of a fiber-array for accessing simultaneously more than one input/output ports. Fig. 4(b) depicts a photo of the packaged device, showing in more detail the very dense wiring connecting the MZIs to the PCB pads. Two electrical phase shifters (PS) are employed per interferometer, accounting for twelve in total.

The electro-optical PSs of the MZIs are 1.25mm long vertical PN diodes. Switching is realized by means of forward carrier injection so as to avoid the high V_π requirements of the carrier depletion mechanism and to allow for the direct PS connection to the FPGA control ports without the intervention of amplifiers in the transmission line. Due to non-optimal fabrication process, the MZIs presented variation in the power required for setting in the BAR or CROSS state with the results presented in Table I. The measurements reveal a V_π value for the six MZIs in the range of 0.9-1.1 V. The power consumption in the two extreme cases of having all phase shifters turned on and turned off are 115mW and 5.6mW, respectively. Due to imperfect p-n junction fabrication that resulted to almost 9dB insertion losses for every p-n phase element, the optical losses were found to be in the range of 48-52 dB depending on the input-output port pair. The cross-talk value of an individual MZI was measured to be in the range of 10dB-18dB. Moreover, the transfer function of the 4x4 switch has been measured by injecting light in port #3, while switching to all output ports was realized via biasing the phase shifters from a DC source. The results have already been reported in [39], with crosstalk ranging from 8dB to 12dB for 1535nm-1550nm operation.

The characterization of the switch in the time domain was performed by driving MZI#6 with RF signals, while the other MZIs were DC biased. The PN junctions were forward biased with two driving

TABLE I: Optimum Biasing Conditions for each one of the SM's MZIs in CROSS or BAR state

	MZI-1	MZI-2	MZI-3	MZI-4	MZI-5	MZI-6
CROSS	1.08V (21mA)	0V	1.03V (17mA)	0.98V (11mA)	0.89V (6mA)	1.18V (36mA)
BAR	0.89V (4mA)	1.02V (16mA)	0V	0V	0V	0.9V (7mA)

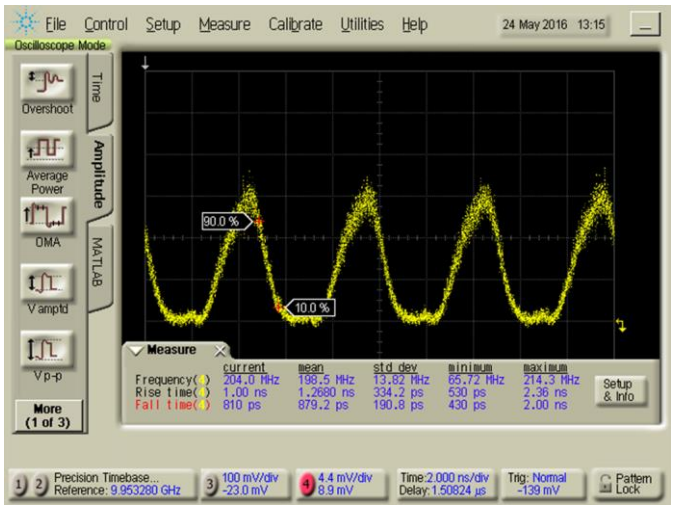

Fig. 5: Optical trace when light is injected in port #3 and a 200MHZ electrical clock signal is used to modulate a single MZI switch in the optical path. Push -pull configuration is used.

signals of $1V_{p\text{-}p}$ coming from a PPG for push pull operation of the MZI. The PPG was running at 200MHz with 50% duty cycle pulses. Fig. 5 illustrates the recorded optical trace revealing rise and fall times of 0.8ns and 1.2ns, respectively. In the case of the single-drive MZI operation, the respective response times were higher with a rising time of 2.5ns and a fall time of 3ns.

The switch chip was controlled by a Stratix V FPGA board that provides up to 12.5 Gb/s control signals with maximum voltage for the high level 1.23 V and 0.15 V for the low one resulting a total 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 devices that were connected to the SMA connectors leading to the upper phase shifter of each MZI. The electrical pulsed signal applied to the phase shifters had a peak power of ~0.9V that was marginally adequate for obtaining a π 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 (CROSS) state.

IV. RESULTS

Fig. 6 illustrates the oscilloscope traces obtained for two different operation phases (Phase #1 and #2), with every phase lasting for the duration of four data

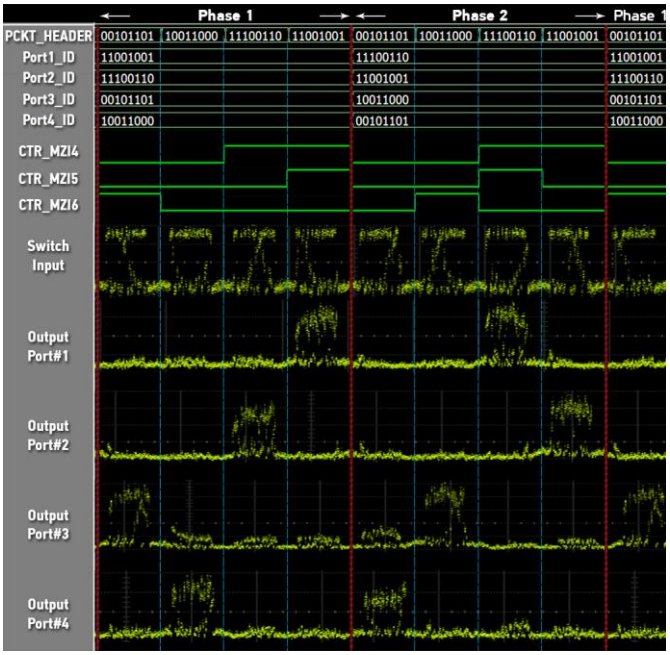

Fig. 6: Experimental packet trace results.

packets and corresponding to a different combination of BF-based addresses (IDs) for the four switch output ports. The first five rows in Fig. 6 show the packet headers and the four output port IDs (Port1_ID, Port2_ID, Port3_ID, Port4_ID). Rows 6,7 and 8 illustrate the respective control signals produced by the FPGA for driving MZI #4, #5 and #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 resources. This results into a different routing path for each packet even though every packet retains its header during both Phases. For example, the first packet with header '00101101' is routed to Output port 3 during Phase 1 and to Output Port 4 during Phase 2. Fig. 6 indicates proper routing functionality and switch re-configurability for all packets during both Phases.

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 switching matrix and routed to all output ports. The BER measurements illustrated in Fig. 7 revealed

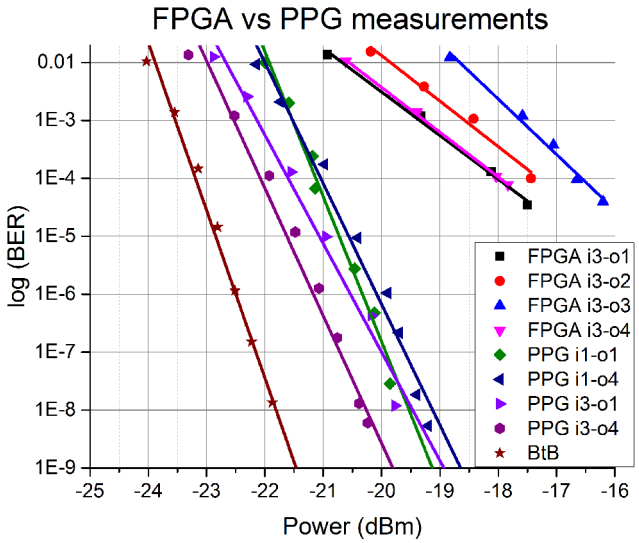

Fig. 7: BER measurements of the proposed signal using the FPGA vs PPG for driving the phase shifter, using a BtB as a reference point. In the label i stands for input and o stands for output

an error-rate 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.

The PCML interface and the associated current drawing-based origin of the 10^{-4} error-rate was verified by replacing the FPGA board with a PPG capable to deliver programmable patterns up to 12.5 Gb/s with voltage levels up to 2.5Vpp and without any limitations in current draining. Two (MZI#4, MZI#6) out of the three switches in the path i1/i3 to o1/o4 were controlled by the PPG, while MZI#2 was biased with a constant DC signal. The two electrical control signals had a voltage swing of 1V and 1.1V, respectively, and the BER performance is presented in Fig. 7. The enhanced ER of the switches from the unlimited current draining led to error-free operation up to 10^{-9}, with a power penalty ranging between 1.2dB-2.5dB. This indicates that the use of DAC or amplification at the FPGA output can yield error-free operation at the expense of increased power consumption. The use of driving circuits to interface the FPGA with the MZIs yielding both optimal voltage and current driving levels has already been reported in [40] and can result in an extra power consumption of at least 1.2 W [41] for the whole 4x4 switch with 6 MZIs. As such, the capability to perform also in DAC- and amplifier-less mode allows for deciding upon the operational scheme based on the application-enforced signal quality and energy consumption requirements.

The ON/OFF suppression ratio between packets at the output of the switching matrix was measured to be 7 dB, mainly owing to the sub-optimal switching obtained at MZI#6. Given the fact that the 4x4 switching matrix is capable of providing cross-talk values up to 12dB when driven by a DC electrical source and operating under optimal driving conditions [39], the utilization of a current amplification stage for optimally biasing the MZIs could in principle improve crosstalk up to 5dB at the cost of extra power consumption. Improving cross-talk performance of the switch beyond this point, would involve modifications to the switch design. A number of different techniques have been already demonstrated towards this direction such as the use of a dilated Benes architecture [42] that could result in lower crosstalk, but would increase the number of stages, as well as the losses. Other approaches comprise the use of SOA's in an MZI branch, that could offer lossless operation along with lower crosstalk as in [43], or the modification of the 2x2 MZI basic building block to offer lower crosstalk as in [44].

V. Discussion

The strict network requirements imposed by disaggregated DC architectures have initiated the discussion for alternative forwarding schemes that can ensure low latency, high bandwidth, rapid reconfiguration and reduced memory capacity for the look-up tables. Through our experimental evaluation, we have demonstrated an optically enabled Bloom Filter-based forwarding scheme, where BF-label forwarding is carried out in an FPGA controller that co-operates directly with an underlying programmable Si-Pho 4x4 switching matrix. This 4x4 switch–and-forwarding implementation could be in principle applied in intra- and inter-tray disaggregated Data Center network architectures as on- or off-board switch, respectively, where each tray incorporates a number of compute, memory or storage resources and the board-level

waveguide or fiber-link lengths are well-known and rather small, i.e. not higher than a few meters. The combined demonstration of the Si-based switching matrix along with its BF-label forwarding plane confirms the feasibility of transferring the proven advantages of BF-label forwarding in terms of rapid reconfiguration and memory requirements' reduction [11] onto a silicon-based high-bandwidth, low-latency and low energy optical interconnect. Moreover, this 4x4 switch-and-forwarding plane can be easily expanded towards a complete routing node by relying on well-known concepts for supporting synchronization, contention resolution and scheduling.

Synchronization between multiple input packets can be ensured by operating the proposed scheme in a synchronous network where time is divided into discrete-time packet-slots of constant duration [45], by taking into account the fiber lengths and the respective propagation times from the transmitting and receiving nodes. Packet transmissions can start only at the beginning of packet-slots, while every transmitting node is properly aligned to the slots taking into account the propagation delay of the fiber link connecting to the 4x4 switching matrix [46]. This procedure is essential as the nodes are located at different distances from the switch. An extra guard-band at every slot can be added to compensate for delay variations (originating, for example, from temperature variations etc.) and ensure simultaneous handling of all routing requests, during the same slot, at the FPGA controller. These concepts have already been investigated in optical network architectures targeted for PON deployments [46, 47] as well as for Data Center switching [48]. An alternative synchronization scheme that could be employed, where knowledge of the fiber-link lengths is not necessary, could rely on the use of optical synchronizers [49-51] as a discrete optical unit externally connected at the front-end of the proposed switching matrix.

Contention resolution can be effectively offered by relying on edge buffering schemes with packet drops and subsequent retransmissions, as has been already proposed in optical switching schemes for Data Center applications in order to obtain low-latency values [48, 52, 53]. Edge buffering schemes incorporate electronic buffers on the source and destination nodes, with the switching nodes relying on buffer-less architectures that are capable of dropping packets in case of collisions. The transmitting nodes store a copy of their sent packets in their buffer until an acknowledgment (ACK) is received by the destination node. In case no ACK is received within a predefined timeframe, that has to be slightly higher than the maximum round-trip-time of the network, or in case a signaling message denoting the packet drop by an intermediate node is received, the transmitting node retransmits the packet through the network. This contention resolution strategy is inherently compatible with our scheme and can be easily adapted by adding an ON/OFF optical switch stage per input port that will perform the drop operations when needed. The FPGA data plane can be modified to operate as the controller of the ON/OFF switch stage just by introducing an extra process for this purpose that decides upon which packets will be dropped in case of path or outgoing port collision. An indicative scheme regarding the ON/OFF switching stage to be employed at the four incoming ports of the proposed 4x4 switch could rely on the addition of one Si-based MZI at every input port. In this way, the FPGA control plane will process in parallel the BF-labels of all four incoming packets deciding upon the requested switching paths that have to be established and, in case of colliding paths or colliding outgoing ports, will drive these additional MZIs either in pass or in drop switching mode.

As far as scheduling is concerned, this process has to fulfill two main objectives: a) to decide which packet or packets have to be dropped in case of multiple packets contending for the same output, and b) to reconfigure the Benes switching matrix established paths in order to satisfy the connectivity requests, if this is possible. The scheduling policy and the decision about the packets to be dropped is typically made upon the network and traffic requirements and there are many solutions that have been proposed in the literature [54-60]. A simple scheme that could be also applied in our proposed BF-label forwarding implementation could include Round-Robin or Strict packet priority [60], where the scheduler serves packets according to the priority assigned to each one. Regarding the reconfiguration of the Benes switching matrix, various scheduling algorithms have been proposed in the literature and could be in principle also applied to our scheme, like

the looping algorithm [61] and variations that improved its performance [62-64].

Finally, the expansion of the concept towards an integrated version of switch-and-forwarding node should take into account the latency introduced by the BF-label processing that directly reflects on the necessary optical delay that has to be introduced at the front-end of the switch. The FPGA transceiver latency that includes the SerDes, word alignment and synchronization functions and accounts for the biggest part of the overall latency, depends obviously on the FPGA model and the respective PHY IP block offered by the FPGA board manufacturer. Custom transceiver designs can be stripped down to include only the necessary blocks, in order to reduce the latency imposed, and have been reported to offer latency values of just 57.9ns [30]. An alternative scheme for latency reduction can be the adoption of parallel label transmission techniques, like WDM labelling [65] or RF-tone in-band labelling [66], where label bits arrive already as parallel streams at the FPGA inputs negating the need for SerDes-based BF-label extraction and parallelization functions at the expense of additional optical hardware at the FPGA front-end. Parallel encoding of the packet label is realized in such schemes by exploiting the use of separate wavelengths or RF tones for each bit of the label. This way, the large amount of latency introduced by the serialization/deserialization of the BF-encoded label can be eliminated, leaving only a small amount of 8nsec FPGA processing functions that has to be compensated by means of the optical delay at the switch front-end. The latter can be easily mitigated using either a small piece of 1.6meters of optical fiber or even integrated photonic delay line solutions in case on-board application of the switch is targeted. Towards this direction, different implementations based on Si_3N_4 [67], Silica-on-Silicon [68] and the μm-scale SOI platform [69] have been shown to accommodate compact on-chip spiral waveguides with delays even higher than 10ns and could be assembled on-board for FPGA processing delay compensation purposes.

VI. Conclusion

We have experimentally demonstrated for the first time, to the best of our knowledge, an optically enabled BF-label forwarding node operating with 10Gb/s optical BF-encoded destination labels assigned to 10Gb/s optical packets. The BF-based forwarding node demonstrates the interaction between an FPGA control plane and a Si-based 4x4 electro-optic switching matrix, with the FPGA performing BF-label extraction, membership checking and control signal generation. Successful 1x4 routing of the incoming optical packets has been demonstrated even for look-up table changes emerging every four packets, indicating that the proposed scheme can respond to network topology changes experienced at very small-time scales. The use of BF-encoded forwarding requires a number of look-up table entries that equals the number of switch outgoing ports, suggesting significant savings in router look-up table memory requirements. Even-though the employed Si-Pho switch experienced rather high-losses due to p-n junction fabrication errors, the remarkably low losses and high-port counts reported by state-of-the-art Si-pho switches [70-74] indicates that optically enabled BF-based forwarding can be accomplished even for high-radix Si-based switch fabrics.

Acknowledgment

This research has been partially supported by the EC FP7-ICT project PhoxTrot (contract number 318240), the EC H2020 project dRedBox (contract number 687632) and the EC H2020 project L3MATRIX (contract number 688544).

References

[1] S. Di, D. Kondo, and F. Cappello, "Characterizing Cloud Applications on a Google Data Center," in *42nd International Conference on Parallel Processing (ICPP)*, Lyon, France, 2013, pp. 468-473.

[2] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch, "Heterogeneity and dynamicity of clouds at scale: Google trace analysis," in *Proceedings of the Third ACM Symposium on Cloud Computing*, San Jose, California, 2012, pp. 1-13.

[3] Facebook. (1-10-2017). *The Open Compute server architecture specications.* Available: www.opencompute.org

[4] Intel. (1-10-2017). *Rack Scale Architecture.* Available: http://www.intel.com/content/www/us/en/architecture-and-technology/rack-scale-design-overview.html

[5] J. Weiss, R. Dangel, J. Hofrichter, F. Horst, D. Jubin, N. Meier, *et al.*, "Optical interconnects for disaggregated resources in future datacenters," in *European Conference on Optical Communication (ECOC)*, Cannes, France, 2014, pp. 1-3.

[6] H. M. M. Ali, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy efficient disaggregated servers for future data centers," in *20th European Conference on Networks and Optical Communications - (NOC)*, London, UK, 2015, pp. 1-6.

[7] A. D. Papaioannou, R. Nejabati, and D. Simeonidou, "The Benefits of a Disaggregated Data Centre: A Resource Allocation Approach," in *IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, 2016, pp. 1-7.

[8] K. Katrinis, D. Syrivelis, D. Pnevmatikatos, G. Zervas, D. Theodoropoulos, I. Koutsopoulos, *et al.*, "Rack-scale disaggregated cloud data centers: The dReDBox project vision," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Dresden, Germany, 2016, pp. 690-695.

[9] K. Lim, J. Chang, T. Mudge, P. Ranganathan, S. K. Reinhardt, and T. F. Wenisch, "Disaggregated memory for expansion and sharing in blade servers," *SIGARCH Comput. Archit. News*, vol. 37, pp. 267-278, 2009.

[10] S. Han, N. Egi, A. Panda, S. Ratnasamy, G. Shi, and S. Shenker, "Network support for resource disaggregation in next-generation datacenters," in *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*, College Park, MD, USA, 2013, pp. 1-7.

[11] M. Yu, A. Fabrikant, and J. Rexford, "BUFFALO: bloom filter forwarding architecture for large organizations," in *Proceedings of the 5th international conference on Emerging networking experiments and technologies*, Rome, Italy, 2009, pp. 313-324.

[12] A. B. M. Carlos, C. E. Rothenberg, and F. M. Maurício, "In-packet Bloom filter based data center networking with distributed OpenFlow controllers," in *IEEE Globecom Workshops*, Miami, FL, USA, 2010, pp. 584-588.

[13] D. Li, Y. Li, J. Wu, S. Su, and J. Yu, "ESM: Efficient and Scalable Data Center Multicast Routing," *IEEE/ACM Transactions on Networking*, vol. 20, pp. 944-955, 2012.

[14] Y. Matsuoka, N. Kohmu, Y. Lee, H. Arimoto, T. Takai, N. Chujo, *et al.*, "A Compact 310-Gb/s Optical Transceiver for High-Density Optical Interconnects," *IEEE Photonics Technology Letters*, vol. 29, pp. 334-337, 2017.

[15] K. Nagashima, N. Nishimura, A. Izawa, T. Kise, and H. Nasu, "28-Gb/s × 24-channel CDR-integrated VCSEL-based transceiver module for high-density optical interconnects," in *Optical Fiber Communications Conference and Exhibition (OFC)*, Anaheim, CA, USA, 2016, pp. 1-3.

[16] S. Bernabé, K. Rida, G. Parès, O. Castany, D. Fowler, C. Kopp, *et al.*, "On-Board Silicon Photonics-Based Transceivers With 1-Tb/s Capacity," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 6, pp. 1018-1025, 2016.

[17] K. Hasharoni, S. Benjamin, A. Geron, G. Katz, S. Stepanov, N. Margalit, *et al.*, "A high end routing platform for core and edge applications based on chip to chip optical interconnect," in *Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)*, Anaheim, CA, USA, 2013, pp. 1-3.

[18] R. Pitwon, K. Wang, and A. Worrall, "Converged photonic data storage and switch platform for exascale disaggregated data centers," in *Proceedings of SPIE Photonics West Vol. 10109*, San Francisco, CA, USA, 2017, pp. 101090Y-101090Y-9.

[19] Y. Yan, G. M. Saridis, Y. Shu, B. R. Rofoee, S. Yan, M. Arslan, *et al.*, "All-Optical Programmable Disaggregated Data Centre Network Realized by FPGA-Based Switch and Interface Card," *Journal of Lightwave Technology*, vol. 34, pp. 1925-1932, 2016.

[20] C. E. Rothenberg, C. Macapuna, F. Verdi, M. Magalhães, and A. Zahemszky, "Data center networking with in-packet Bloom filters," in *Proc. SBRC*, Gramado, Brazil, 2010, pp. 553-566.

[21] L. Qiao, W. Tang, and T. Chu, "Ultra-large-scale silicon optical switches," in *IEEE 13th International Conference on Group IV Photonics (GFP)*, Shanghai, China, 2016, pp. 1-2.

[22] T. J. Seok, N. Quack, S. Han, R. S. Muller, and M. C. Wu, "Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers," *Optica*, vol. 3, pp. 64-70, 2016/01/20 2016.

[23] K. Suzuki, K. Tanizawa, T. Matsukawa, G. Cong, S.-H. Kim, S. Suda, *et al.*, "Ultra-compact 8 × 8 strictly-non-blocking Si-wire PILOSS switch," *Optics Express*, vol. 22, pp. 3887-3894, 2014/02/24 2014.

[24] C. P. Chen, X. Zhu, Y. Liu, K. Wen, M. S. Chik, T. Baehr-Jones, *et al.*, "Programmable Dynamically-Controlled Silicon Photonic Switch Fabric," *Journal of Lightwave Technology*, vol. 34, pp. 2952-2958, 2016.

[25] N. Dupuis, A. V. Rylyakov, C. L. Schow, D. M. Kuchta, C. W. Baks, J. S. Orcutt, *et al.*, "Nanosecond-Scale Mach-Zehnder-Based CMOS Photonic Switch Fabrics," *Journal of Lightwave Technology*, vol. 35, pp. 615-623, 2017.

[26] D. Nikolova, D. M. Calhoun, Y. Liu, S. Rumley, A. Novack, T. Baehr-Jones, *et al.*, "Modular architecture for fully non-blocking silicon photonic switch fabric," *Microsystems & Nanoengineering*, vol. 3, p. 16071, 01/16/online 2017.

[27] T. Shiraishi, Q. Li, Y. Liu, X. Zhu, K. Padmaraju, R. Ding, *et al.*, "A Reconfigurable and Redundant Optically-Connected Memory System using a Silicon Photonic Switch," in *Optical Fiber Communication Conference*, San Francisco, California, 2014, p. Th2A.10.

[28] Y. Xiong, F. Gohring de Magalhães, B. Radi, G. Nicolescu, F. Hessel, and O. Liboiron-ladouceur, "Towards a Fast Centralized Controller for Integrated Silicon Photonic Multistage MZI-based Switches," in *Optical Fiber Communication Conference*, Anaheim, California, 2016, p. W1J.2.

[29] D. Calhoun, K. Wen, X. Zhu, S. Rumley, L. Luo, Y. Liu, *et al.*, "Dynamic reconfiguration of silicon photonic circuit switched interconnection networks," in *Proc. IEEE High Perform. Extreme Comput. Conf*, Waltham, MA, USA, 2014.

[30] F. Lou, M. M. P. Fard, P. Liao, M. S. Hai, R. Priti, Y. Huangfu, *et al.*, "Towards a centralized controller for silicon photonic MZI-based interconnects," in *IEEE Optical Interconnects Conference (OI)*, San Diego, CA, USA., 2015, pp. 146-147.

[31] D. M. Calhoun, Q. Li, C. Browning, N. C. Abrams, Y. Liu, R. Ding, *et al.*, "Programmable wavelength locking and routing in a silicon-photonic interconnection network implementation," in *2015 Optical Fiber Communications Conference and Exhibition (OFC)*, 2015, pp. 1-3.

[32] M. Antikainen, T. Aura, and M. Särelä, "Denial-of-Service Attacks in Bloom-Filter-Based Forwarding," *IEEE/ACM Transactions on Networking*, vol. 22, pp. 1463-1476, 2014.

[33] M. Moralis-Pegios, N. Terzenidis, C. Vagionas, S. Pitris, E. Chatzianagnostou, A. Brimont, *et al.*, "A programmable Si-photonic node for SDN-enabled Bloom filter forwarding in disaggregated data centers," in *Proceedings of SPIE Photonics West Vol. 10109*, San Francisco, CA, USA, 2017, pp. 101090X-101090X-7.

[34] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, pp. 422-426, 1970.

[35] A. Broder and M. Mitzenmacher, "Network applications of bloom filters: A survey," *Internet mathematics*, vol. 1, pp. 485-509, 2004.

[36] D. Syrivelis, G. Parisis, D. Trossen, P. Flegkas, V. Sourlas, T. Korakis, *et al.*, "Pursuing a Software Defined Information-centric Network," in *European Workshop on Software Defined Networking*, Darmstadt, Germany, 2012, pp. 103-108.

[37] W. You, B. Mathieu, P. Truong, J. F. Peltier, and G. Simon, "DiPIT: A Distributed Bloom-Filter Based PIT Table for CCN Nodes," in *21st International Conference on Computer Communications and Networks (ICCCN)*, München, Germany 2012, pp. 1-7.

[38] Y. Du, G. He, and D. Yu, "Efficient Hashing Technique Based on Bloom Filter for High-Speed Network," in *8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, China, 2016, pp. 58-63.

[39] K. Vyrsokinos, M. Moralis-Pegios, C. Vagionas, A. Brimont, A. Zanzi, P. Sanchis, *et al.*, "Single Mode Optical Interconnects for future data centers," in *2016 18th International Conference on Transparent Optical Networks (ICTON)*, 2016, pp. 1-4.

[40] Y. Xiong, F. G. d. Magalhães, G. Nicolescu, F. Hessel, and O. Liboiron-Ladouceur, "Co-design of a low-latency centralized controller for silicon photonic multistage MZI-based switches," in *2017 Optical Fiber Communications Conference and Exhibition (OFC)*, 2017, pp. 1-3.

[41] T. I. Incorporated. (1/10/2017). *LM7171 Datasheet*. Available: http://www.ti.com/product/LM7171/datasheet

[42] D. Celo, D. J. Goodwill, J. Jia, P. Dumais, Z. Chunshu, Z. Fei, *et al.*, "32×32 silicon photonic switch," in *2016 21st OptoElectronics and Communications Conference (OECC) held jointly with 2016 International Conference on Photonics in Switching (PS)*, 2016, pp. 1-3.

[43] Q. Cheng, A. Wonfor, J. L. Wei, R. V. Penty, and I. H. White, "Demonstration of the feasibility of large-port-count optical

switching using a hybrid Mach-Zehnder interferometer-semiconductor optical amplifier switch module in a recirculating loop," *Opt Lett*, vol. 39, pp. 5244-7, Sep 15 2014.

[44] N. Dupuis, A. V. Rylyakov, C. L. Schow, D. M. Kuchta, C. W. Baks, J. S. Orcutt, *et al.*, "Ultralow crosstalk nanosecond-scale nested 2 × 2 Mach–Zehnder silicon photonic switch," *Optics Letters*, vol. 41, pp. 3002-3005, 2016/07/01 2016.

[45] G. C. Sankaran and K. M. Sivalingam, "Time synchronization mechanisms for an optically groomed data center network," in *2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC)*, 2016, pp. 1-8.

[46] Y. Luo, F. J. Effenberger, and N. Ansari, "Time synchronization over ethernet passive optical networks," *IEEE Communications Magazine*, vol. 50, pp. 136-142, 2012.

[47] G. Kramer, B. Mukherjee, and G. Pesavento, "IPACT a dynamic protocol for an Ethernet PON (EPON)," *Comm. Mag.*, vol. 40, pp. 74-80, 2002.

[48] S. D. Lucente, N. Calabretta, J. A. C. Resing, and H. J. S. Dorren, "Scaling low-latency optical packet switches to a thousand ports," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 4, pp. A17-A28, 2012.

[49] H. Hu, E. Palushani, J. L. Areal, M. Galili, A. Clausen, M. S. Berger, *et al.*, "Optical Frame Synchronizer for 10 G Ethernet Packets Aiming at 1 Tb/s OTDM Ethernet," in *National Fiber Optic Engineers Conference*, San Diego, California, 2010, p. JWA46.

[50] S. Battia, M. Zghal, N. Boudrigaa, and T. Hallb, "An all-optical synchronizer for switching node using single-sideband modulator and fiber Bragg gratings," in *2009 IEEE Symposium on Computers and Communications*, 2009, pp. 736-741.

[51] J. P. Mack, H. N. Poulsen, and D. J. Blumenthal, "Variable Length Optical Packet Synchronizer," *IEEE Photonics Technology Letters*, vol. 20, pp. 1252-1254, 2008.

[52] C. Xianhui, D. K. Hunter, and I. D. Henning, "Switched optical star-topology network with edge electronic buffering and centralized control," in *2008 Third International Conference on Communications and Networking in China*, 2008, pp. 243-246.

[53] H. J. Chao and K. Xi, "Bufferless Optical Clos Switches for Data Centers," in *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2011*, Los Angeles, California, 2011, p. OTuG2.

[54] M. Alizadeh, S. Yang, M. Sharif, S. Katti, N. McKeown, B. Prabhakar, *et al.*, "pFabric: minimal near-optimal datacenter transport," presented at the Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM, Hong Kong, China, 2013.

[55] P. Goyal, H. M. Vin, and C. Haichen, "Start-time fair queueing: a scheduling algorithm for integrated services packet switching networks," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 690-704, 1997.

[56] R. Mittal, R. Agarwal, S. Ratnasamy, and S. Shenker, "Universal packet scheduling," presented at the Proceedings of the 13th Usenix Conference on Networked Systems Design and Implementation, Santa Clara, CA, 2016.

[57] H. Sariowan, R. L. Cruz, and G. C. Polyzos, "SCED: a generalized scheduling policy for guaranteeing quality-of-service," *IEEE/ACM Transactions on Networking*, vol. 7, pp. 669-684, 1999.

[58] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," *SIGCOMM Comput. Commun. Rev.*, vol. 19, pp. 1-12, 1989.

[59] G. Gibb, G. Varghese, M. Horowitz, and N. McKeown, "Design principles for packet parsers," in *Architectures for Networking and Communications Systems*, 2013, pp. 13-24.

[60] Y.-L. C. a. Z. T. Tsung-Yu Tsai, *Introduction to Packet Scheduling Algorithms for Communication Networks*: InTech, 2010.

[61] D. C. Opferman and N. T. Tsao-wu, "On a class of rearrangeable switching networks part I: Control algorithm," *The Bell System Technical Journal*, vol. 50, pp. 1579-1600, 1971.

[62] S. Andresen, "The Looping Algorithm Extended to Base 2t Rearrangeable Switching Networks," *IEEE Transactions on Communications*, vol. 25, pp. 1057-1063, 1977.

[63] L. Kyungsook Yoon, "A New Benes Network Control Algorithm," *IEEE Transactions on Computers*, vol. C-36, pp. 768-772, 1987.

[64] Z. Jing and H. Gu, "A partially adaptive routing algorithm for Benes network on chip," in *2009 2nd IEEE International Conference on Computer Science and Information Technology*, 2009, pp. 614-618.

[65] N. Wada, H. Harai, W. Chujo, and F. Kubota, "80G to 10G bit/s variable rate photonic packet routing based on multi-wavelength label switch," in *Proceedings 27th European Conference on Optical Communication (Cat. No.01TH8551)*, 2001, pp. 308-309 vol.3.

[66] J. Luo, H. J. S. Dorren, and N. Calabretta, "Optical RF Tone In-Band Labeling for Large-Scale and Low-Latency Optical Packet Switches," *Journal of Lightwave Technology*, vol. 30, pp. 2637-2645, 2012.

[67] H. Lee, T. Chen, J. Li, O. Painter, and K. J. Vahala, "Ultra-low-loss optical delay line on a silicon chip," vol. 3, p. 867, 05/29/online 2012.

[68] J. D. LeGrange, J. E. Simsarian, P. Bernasconi, L. Buhl, J. Gripp, and D. T. Neilson, "Demonstration of an Integrated Buffer for an All-Optical Packet Router," *IEEE Photonics Technology Letters*, vol. 21, pp. 781-783, 2009.

[69] M. Moralis-Pegios, G. Mourgias-Alexandris, N. Terzenidis, M. Cherchi, M. Harjanne, T. Aalto, *et al.*, "Optical Buffering and Time-Slot Interchanger with integrated Si-based delay lines," in *Advanced Photonics 2017 (IPR, NOMA, Sensors, Networks, SPPCom, PS)*, New Orleans, Louisiana, 2017, p. PTu1D.2.

[70] L. Qiao, W. Tang, and T. Chu, "16×16 Non-blocking Silicon Electro-optic Switch Based on Mach-Zehnder Interferometers," in *Optical Fiber Communication Conference*, Anaheim, California, 2016, p. Th1C.2.

[71] N. Dupuis, B. G. Lee, A. V. Rylyakov, D. M. Kuchta, C. W. Baks, J. S. Orcutt, *et al.*, "Modeling and Characterization of a Nonblocking 4x4 Mach-Zehnder Silicon Photonic Switch Fabric," *Journal of Lightwave Technology*, vol. 33, pp. 4329-4337, 2015.

[72] L. Qiao, W. Tang, and T. Chu, "32 x 32 silicon electro-optic switch with built-in monitors and balanced-status units," *Sci Rep*, vol. 7, p. 42306, Feb 09 2017.

[73] I. White, M. Ding, A. Wonfor, Q. Cheng, and R. V. Penty, "High Port Court Switch Architectures for Data Center Applications," in *Advanced Photonics 2017 (IPR, NOMA, Sensors, Networks, SPPCom, PS)*, New Orleans, Louisiana, 2017, p. NeW1B.2.

[74] R. Stabile, A. Albores-Mejia, and K. A. Williams, "Monolithic active-passive 16x16 optoelectronic switch," *Optics Letters*, vol. 37, pp. 4666-4668, 2012/11/15 2012.