Flex-LIONS: a Silicon Photonic Bandwidth-Reconfigurable Optical Switch Fabric

IEICE TRANS. COMMUN., VOL.E103–B, NO.11 NOVEMBER 2020

INVITED PAPER *Joint Special Section on Opto-electronics and Communications for Future Optical Network*

# Flex-LIONS: A Silicon Photonic Bandwidth-Reconfigurable Optical Switch Fabric

**Roberto PROIETTI**[†a)], **Xian XIAO**[†], **Marjan FARIBORZ**[†], **Pouya FOTOUHI**[†], **Yu ZHANG**[†], *and* **S. J. Ben YOO**[†], *Nonmembers*

**SUMMARY** This paper summarizes our recent studies on architecture, photonic integration, system validation and networking performance analysis of a flexible low-latency interconnect optical network switch (Flex-LIONS) for datacenter and high-performance computing (HPC) applications. Flex-LIONS leverages the all-to-all wavelength routing property in arrayed waveguide grating routers (AWGRs) combined with microring resonator (MRR)-based add/drop filtering and multi-wavelength spatial switching to enable topology and bandwidth reconfigurability to adapt the interconnection to different traffic profiles. By exploiting the multiple free spectral ranges of AWGRs, it is also possible to provide reconfiguration while maintaining minimum-diameter all-to-all interconnectivity. We report experimental results on the design, fabrication, and system testing of $8 \times 8$ silicon photonic (SiPh) Flex-LIONS chips demonstrating error-free all-to-all communication and reconfiguration exploiting different free spectral ranges ($FSR_0$ and $FSR_1$, respectively). After reconfiguration in $FSR_1$, the bandwidth between the selected pair of nodes is increased from 50 Gb/s to 125 Gb/s while an all interconnectivity at 25 Gb/s is maintained using $FSR_0$. Finally, we investigate the use of Flex-LIONS in two different networking scenarios. First, networking simulations for a 256-node datacenter inter-rack communication scenario show the potential latency and energy benefits when using Flex-LIONS for optical reconfiguration based on different traffic profiles (a legacy fat-tree architecture is used for comparison). Second, we demonstrate the benefits of leveraging two FSRs in an 8-node 64-core computing system to provide reconfiguration for the hotspot nodes while maintaining minimum-diameter all-to-all interconnectivity.

***key words:*** *arrayed waveguide grating router, optical interconnections, optical switches, photonic integrated circuits, silicon photonics*

## 1. Introduction

Applications running in modern high-performance computing (HPC) and cloud data center systems adopting heterogeneous processing [1], [2] are typically non-uniformly distributed between compute nodes and exhibit communication patterns with temporal and spatial bursts [3]–[5]. Therefore, being able to reconfigure the interconnection bandwidth between compute nodes to adapt their interconnection to spatial and temporal traffic variations could significantly improve the performance in terms of application execution time, network throughput and latency, and overall energy efficiency.

At the physical layer, optical interconnects are becoming the dominant communication technology in HPC and datacenters driven by the ever-increasing bandwidth scaling pushed by the widespread adoption of the cloud and emerging applications. In the past few years, we have witnessed a number of different integrated reconfigurable wavelength routing and space switching solutions that allow redefining the connectivity in both spectral and spatial domains, dynamically [6]–[15] Among these works, in [7] we proposed and demonstrated SiPh Flex-LIONS (silicon photonic flexible low-latency interconnect optical network switch), a reconfigurable photonic interconnect architecture that leverages both wavelength routing and spatial switching to spatially and temporarily steer and increase the communication bandwidth between specific node pairs (a comparison between Flex-LIONS and other existing approaches is presented in Sect. 2.2).

While the reconfiguration discussed above aims at resolving the congestion in the hotspot links, one limitation of existing solutions, including the work in [7]–[9] is the fact that the reconfiguration operation reduces the connectivity between the other nodes in the network. This could lead to a significant increase in the latency for the traffic that is not part of the hotspot links due to the additional number of hops required to reach the destination nodes. To this aim, this paper extends the work presented in [7]–[9] by presenting a modified version of the architecture to exploit multiple free spectral ranges (FSRs) of the arrayed waveguide grating router (AWGR) at the core of Flex-LIONS. The first FSR guarantees a minimum-diameter all-to-all topology among the $N$ connected nodes, while the second FSR can be freely used to boost the bandwidth between specific node pairs. System experiments using a Silicon Photonic Flex-LIONS (SiPh Flex-LIONS) device fabricated at University of California, Davis (UC Davis) demonstrate the use of Flex-LIONS with two FSRs. Additional results presented in this paper include details regarding the fabrication of different Flex-LIONS chips using two different architectures for the spatial switch component, as well as an extended section discussing the potential benefits in terms of networking performance results when using Flex-LIONS.

The remainder of this paper is organized as follows: Section 2 introduces the working principle of Flex-LIONS with multiple FSRs. Section 3 discusses the fabrication of Flex-LIONS chips using a spatial switch based on a microring resonator (MRR) matrix (as in [8]) or a Beneš Mach-Zehnder switch (MZS) network [9]–[11]. Section 4 reports an experimental demonstration of the Multi-FSR Flex-LIONS principle using a fabricated 8-port SiPh integrated

Manuscript received January 8, 2020.
Manuscript revised April 3, 2020.
Manuscript publicized May 14, 2020.

[†]The authors are with the Electrical and Computer Engineering Department, University of California, Davis, 95616, USA.
a) E-mail: rproietti@ucdavis.edu

DOI: 10.1587/transcom.2019OBI0004

chip. Section 5 analyzes the potential benefits of optical reconfiguration by means of network-level simulations while discussing the control plane challenges associated with the scheduling of reconfiguration. Section 6 concludes this paper summarizing the main contributions of this work and discussing existing challenges and future work related to the physical layer and control plane scalability aspects.

## 2. Flex-LIONS Architecture

### 2.1 Working Principle

Figure 1(a) illustrates the Flex-LIONS architecture (with $N = 8$) exploiting two FSRs ($FSR_0$ and $FSR_1$). An $N$-port AWGR can provide all-to-all connectivity for each of the two FSRs. Figure 1(b) shows the optical spectrum of a silicon photonic AWGR chip fabricated at U C Davis and UC Berkeley with $FSR_0 = FSR_1 = 12.8$ nm.

There are $b$ MRR add-drop filters at each AWGR input/output port. These MRRs can work on either $FSR_0$ or $FSR_1$ and are used for dropping/adding certain wavelength channels at different AWGR input/output ports. By tuning the MRR add-drop filters, $b$ of the $N$ wavelengths from input port $i$ can be dropped and then routed to the desired output port $j$ by an $N \times N$ multi-wavelength switch (e.g. a strictly non-blocking multi-wavelength MRR-based crossbar switch as reported in [7] or a Benes MZS architecture). In this way, the bandwidth between input port $i$ and output port $j$ is effectively increased by adding up to $b$ wavelengths.

The key benefit of exploiting two FSRs is twofold: (1) as shown in Fig. 1(c), it is possible to provide bandwidth steering using one FSR (e.g. $FSR_1$) while maintaining a basic all-to-all connectivity with minimum network diameter among the $N$ nodes; (2) reconfiguration on $FSR_1$ can be done without restrictions on exceeding the maximum number of reconfigured links that would isolate one or multiple nodes from the others.

Figure 1(d) shows the implementation of the multi-wavelength switch as an MRR crossbar switch formed by a matrix of $N^2$ MRRs. Here, the FSR of the multi-wavelength MRR is designed to match with the AWGR channel spacing so that all the $b$ wavelength channels dropped by the MRR drop filters can be simultaneously routed to the desired output port by tuning the desired multi-wavelength MRR in the crossbar [12]. The insertion loss of multi-wavelength MRR crossbar switch is mainly decided by the drop loss of single MRR which can be relatively low with optimized design.

Figure 1(e) shows the implementation of the multi-wavelength switch as a Beneš MZS network switch, which is a multistage switch network with $N \log_2 N - N/2$ $2 \times 2$ MZS as building blocks. The Beneš topology (rearrangeable nonblocking) is highly popular as it requires the minimum number of switching elements among all the multistage network topologies [11]. Since the MZS is wide-band, a Beneš MZS network switch can spatially switch all the wavelength channels simultaneously. The number of cascaded MRRs on the path of the reconfigured channels in a Flex-LIONS with the Beneš MZS network is two while that of a Flex-LIONS with multi-wavelength MRR crossbar is three, so that the bandwidth-narrowing effect is reduced.

### 2.2 Comparison with Other Approaches

Table 1 compares Flex-LIONS with various state-of-the-art wavelength-and-space selective reconfigurable switching fabrics, including indium phosphide (InP) AWGRs + semiconductor optical amplifier (SOA) gates [13], silicon (Si) echelle gratings + (micro-electro-mechanical system) MEMS arrays [14], and multi-wavelength selective crossbar [15]. In particular, the comparison study takes into account the port count, on-chip loss, and the number of switching elements. Here we assume $b = N$ and consider the worst-case on-chip loss for all architectures to make the comparison fair.

It can be seen that Ref. [13] architecture has the problem of high on-chip insertion loss due to a large number of power splitters. Although the SOA gates can be used to compensate for such high loss, the low energy efficiency prevents Ref. [13] architecture to scale up to high radix. Reference [14] architecture suffers not only from the high number of switching elements ($N^3$) but also from high on-chip insertion loss since the number of waveguide crossings increases by $\sim N^2$, while the number of waveguide crossings in Flex-LIONS increases by $\sim N$. Reference [15] architecture also has the issue of a high number of switching elements which makes the control plane more complex and limits the scalability. Compared with Flex-LIONS with multi-wavelength MRR crossbar, Flex-LIONS with the Beneš MZS network exhibits a lower number of switching elements and reduced bandwidth-narrowing effect at the expense of higher on-chip insertion loss and a more complex control mechanism due to the rearrangeable non-blocking nature of this solution.

## 3. Silicon Photonic Integration

The SiPh Flex-LIONS devices are designed and fabricated on a multi-layer platform. The bottom 220-nm-thick Si layer contains the MRR add-drop filters and multi-wavelength spatial switching fabrics. The low-loss and low-crosstalk AWGRs are on the 200-nm-thick silicon nitride (SiN) layer which is 600 nm above the Si layer. The SiN layer vertically interfaces with the Si layer through inverse-tapered evanescent couplers. On top of the silicon oxide cladding are the 400-nm-thick Ti heater layer and 800-nm-thick Au contact metal layer for thermo-optical (TO) tuning.

The radii of the MRR add-drop filters and multi-wavelength MRR are fabrication-calibrated to be 4.75 μm and 63 μm corresponding to the FSRs of 19 nm and 1.6 nm, respectively. The gap between the bus waveguides and the MRRs is 300 nm and 450 nm to minimize the insertion loss for dropping. Spiral resistive heaters along the MRR add-drop filter waveguide are designed to increase the TO tuning efficiency. The $2 \times 2$ MZS contains two $2 \times 2$ multimode

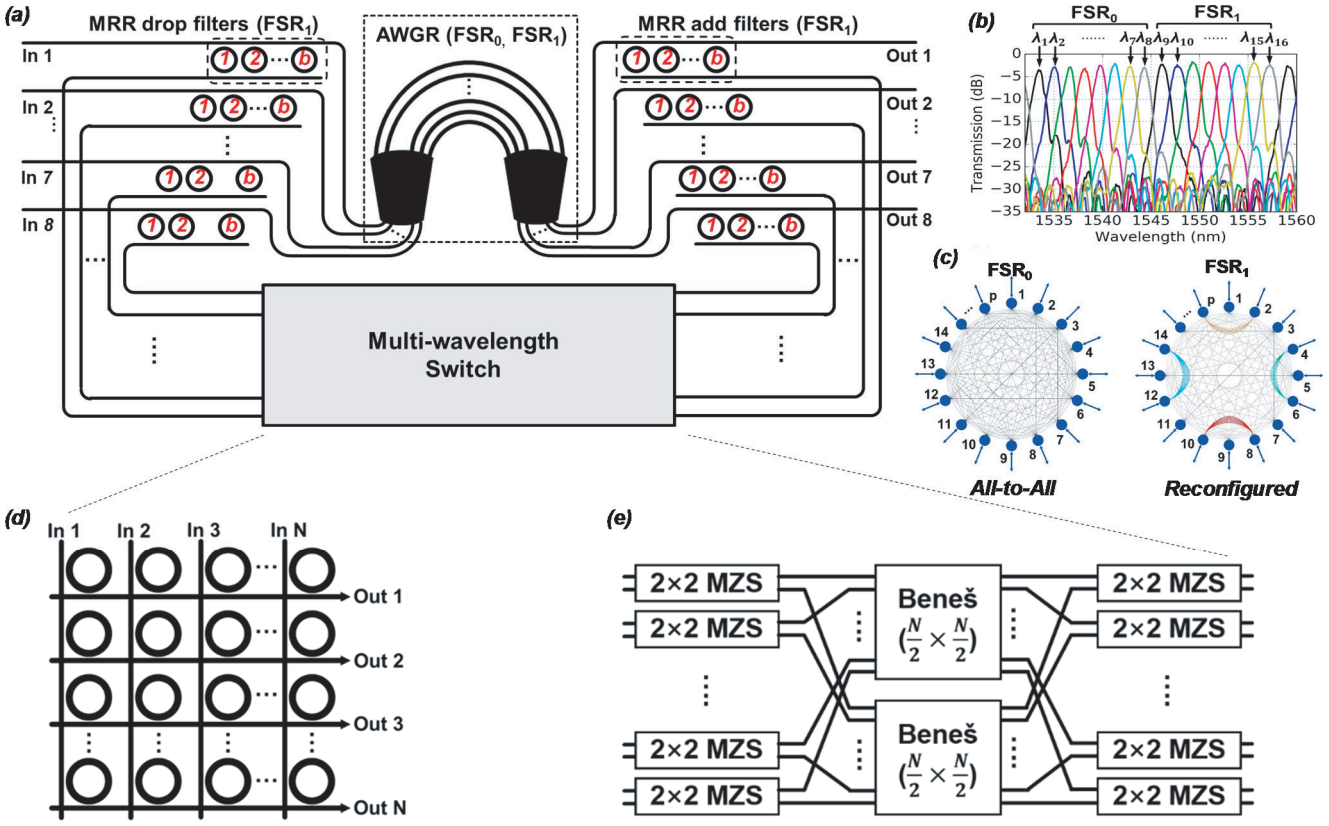



**Fig. 1** (a) Two-FSR Flex-LIONS architecture with AWGR MRR add-drop filters and multi-wavelength switch. (b) AWGR spectrum showing two FSRs. (c) Network topology diagrams showing all-to-all interconnection with $FSR_0$ and reconfiguration for bandwidth steering using $FSR_1$ between four node pairs. (d) Schematic of multi-wavelength MRR crossbar switch. (e) Schematic of Beneš MZS network switching fabric.

**Table 1** Comparison with the state-of-the-art bandwidth-reconfigurable switching fabrics.

| Architecture | Port Count | On-chip Loss (dB) | Number of Switching Elements* | On-chip Loss* (dB) | Reference |
|---|---|---|---|---|---|
| InP AWGRs +SOA gates | 4×4 | 23.7 | $2N^2$ | $(N\text{-}1)\times0.5$ $+\log_2N\times7+8.5$ | [13] |
| Si echelle gratings +MEMS arrays | 8×8 | 16 | $N^3$ | $N\times0.18+N(N\text{-}1)\times0.034+12.6$ | [14] |
| Multi-Wavelength Selective Crossbar | 8×4 | 14 | $N^3$ | $(N\text{-}1)\times1.2+4.7$ | [15] |
| Flex-LIONS with Multi-wavelength MRR Crossbar | 8×8 | 6 | $3N^2$ | $(2N+5)\times0.1+(2N\text{-}2)\times0.09+3.5$ | [7] [8] |
| Flex-LIONS with Beneš MZS Network | 8×8 | 8.4 | $2N^2+N\log_2N\text{-}N/2$ | $4+(N\text{-}1)\times0.16+N\times0.5$ | [9] |

* For $N\times N$ scale

interference (MMI) couplers and two 500-μm-long arms. In order to achieve minimum TO tuning power, heaters are placed on both arms of the MZS. The width of the Ti heaters is 1 μm.

The Flex-LIONS chips with multi-wavelength MRR crossbar and Beneš MZS network were fabricated using the micro and nanoscale fabrication facilities at UC Davis and UC Berkeley. Figure 2(a) and (b) show the micro-

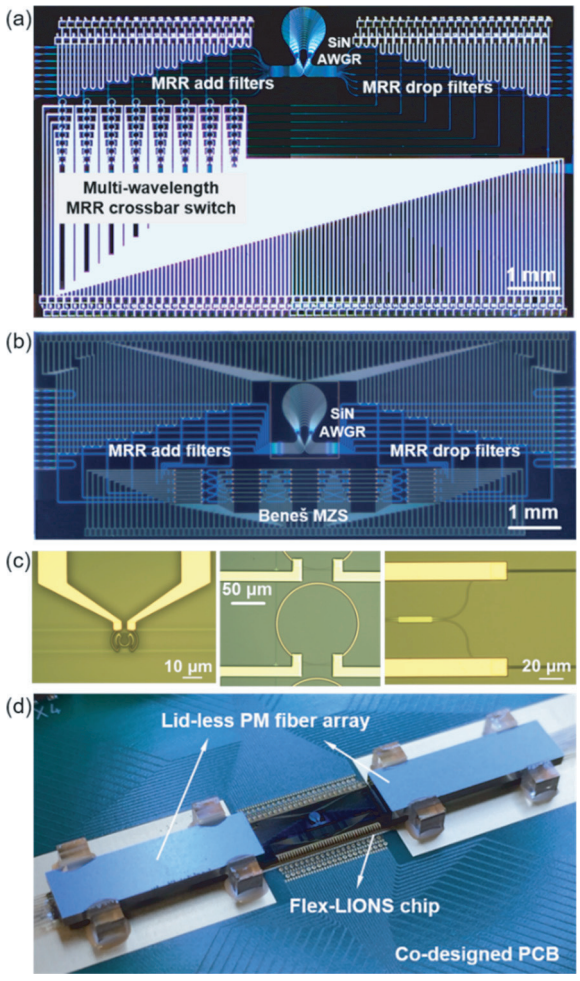



**Fig. 2** (a) Microscope image of the fabricated 8 × 8 SiPh Flex-LIONS ($N = 8$, $b = 3$) chip with multi-wavelength MRR crossbar. (b) Microscope image of the fabricated 8×8 SiPh Flex-LIONS ($N = 8$, $b = 3$) chip with the Beneš MZS network. (c) Microscope image of the MRR add-drop filter, the multi-wavelength MRR switch, and part of the 2 × 2 MZS. (d) Photograph of the integrated Flex-LIONS module (courtesy of Optelligent, LLC).

scope images of the fabricated SiPh Flex-LIONS ($N = 8$, $b = 3$) chips. Figure 2(c) shows the microscope image of the switching elements including the MRR add-drop filter, the multi-wavelength MRR switch, and the 2 × 2 MZS.

The fabricated chip was wire-bonded to a co-designed printed circuit board (PCB) for electrical fan-out. Two lid-less 16-channel 127-μm-pitch polarization-maintaining (PM) fiber arrays were attached to the input and output of the chip using index-matching UV (ultraviolet curing) epoxy. The coupling loss from the PM fiber array to the chip after packaging is 4.7–5.7 dB/facet. Figure 2(d) shows the photograph of the integrated Flex-LIONS module.

## 4. System Experiments

Figure 3(a) shows the experimental setup we used to demonstrate the multi-FSR operating principle discussed above. Eight distributed feedback (DFB) lasers on a 200 GHz spacing wavelength division multiplexing (WDM) grid are wavelength-multiplexed and modulated at 25 Gb/s using a Mach Zehnder (MZ) modulator driven by a $2^{11} - 1$ pseudo random bit sequence (PRBS) signal generated by a high-speed digital to analog converter (DAC). This WDM signal is coupled in/out of a SiPh Flex-LIONS device using lensed fibers. A real-time error analyzer (EA) performed bit error rate (BER) measurements as a function of the receiver input power measured by the built-in optical power monitor of a variable optical attenuator (VOA). The Flex-LIONS chip is wire-bonded on a PCB and driven by a multi-channel DAC controller producing driving signals for tuning the MRR add-drop filters as well as the multi-wavelength crossbar switch. $FSR_0$ and $FSR_1$ [see Fig. 1(b)] are used for all-to-all communication and bandwidth steering, respectively. Figure 3(b) shows the BER measurements for the reconfigured lambdas in $FSR_1$ as well as lambdas for all-to-all in $FSR_0$. In particular, $\lambda_1, \lambda_2 \ldots, \lambda_8$ belong to $FSR_1$ (for reconfiguration), while $\lambda_1$-FSR, $\lambda_2$-FSR…, $\lambda_8$-FSR are for $FSR_0$ (for maintaining all-to-all). Before reconfiguration, $\lambda_1$, $\lambda_3$, and $\lambda_5$ ($\lambda_1$-FSR, $\lambda_3$-FSR, and $\lambda_5$-FSR) from input 4 were used to connect with outputs 1, 3, 5 respectively. After reconfiguration, $\lambda_1$, $\lambda_3$, and $\lambda_5$ from input 4 are reconfigured to output port 8 while $\lambda_1$-FSR, $\lambda_3$-FSR, and $\lambda_5$-FSR are still used to connect with output 1, 3, 5 respectively. After reconfiguration, the bandwidth between input 4 and output 8 is effectively increased by 2.5 × (from 50 Gb/s to 125 Gb/s). The dashed red line represents the optical back to back curve. The power penalty is mainly caused by in-band coherent crosstalk in the AWGR.

## 5. Networking Performance Studies

### 5.1 256-Node Datacenter Inter-Rack Case Study

In this section, we evaluate the use of Flex-LIONS for implementing the rack-level interconnect in a 256-node datacenter with 16 top-of-rack switches. We compare it with a tree-based topology (see Table 2) where Flex-LIONS is replaced by the second level of switches forming a two-level Fat-Tree architecture with an oversubscription factor equal to two. To compare the Flex-LIONS approach with the most aggressive baseline, we modeled the power consumption and latency of the switches based on state-of-the-art commercially-available datacenter switches, which consume 95 W power and offer a 100 ns switch traversal latency [16]. We considered two transceiver technologies in our study: (a) Intel's SiPh transceivers which consume 35 pJ/bit with 100G line rate [17] (this represents an advanced commercially-available technology), and (b) research-grade tightly-integrated electronic-photonic co-designed transceivers that can consume as little as 2 pJ/bit in a 65 nm technology [18]. Table 3 lists the parameters we used for the power modeling of the SiPh components (transceivers and Flex-LIONS). While we modeled Flex-LIONS only with SiPh transceivers, we modeled the legacy Fat Tree topology with both transceiver types (a) and (b) described above. These comparisons allow us to reveal

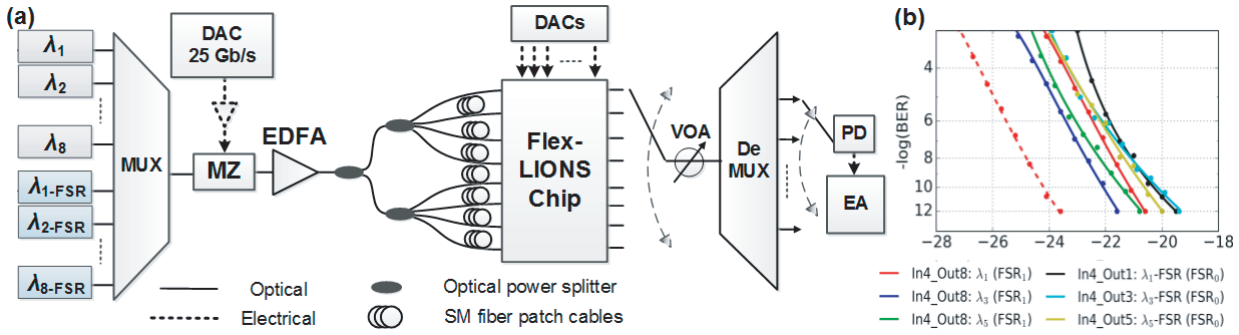



**Fig. 3** Experimental setup for demonstrating the concept of multi-FSR Flex-LIONS with two FSRs. (De)Mux: (De)Multiplexer; MZ: Mach Zehnder modulator; DAC: digital to analog converter; EDFA: erbium-doped fiber amplifier; VOA: variable optical attenuator; PD: photodetector; EA: error analyzer.

**Table 2** Resource requirements of the rack-scale networks.

| | *Fat-Tree* | *Flex-LIONS* |
|---|---|---|
| # Switches | 24 | 16 |
| # Transceivers | 640 | 512 |
| Link data rate | 100G | 100G |
| Bisection bandwidth | 12.8Tb | 25.6Tb |

**Table 3** SiPh technology parameters.

| *Parameter* | *Value* |
|---|---|
| Laser Efficiency | 14% |
| Waveguide Loss | 2.2 dB/cm |
| Receiver Sensitivity | -17.7 dBm |
| Add-drop Filter Loss | 1.5 dB |
| Coupler Loss | 1 dB |
| Photodiode | 0.1 dB |
| Modulator Loss | 1 dB |
| Power Margin | 3 dB |

the power savings of Flex-LIONS in comparison to legacy topologies with the same transceiver technology and to illustrate the total power saving potential that Flex-LION provides compared to state-of-the-art topologies (i.e., Fat Tree) with state-of-the-art commercially available transceivers.

We used gem5 [19] with Garnet2.0 [20] for detailed performance simulations. We evaluate the network under a range of traditional synthetic traffic patterns (uniform random, bit complement, tornado, shuffle). We modeled the network reconfiguration by analyzing the link utilizations in the network for each traffic pattern and subsequently assigned link bandwidth based on the utilization rate of the previous run. It should be noted that this study aims only at stressing all corner cases of the topologies assuming specific traffic patterns and ideal scheduling of reconfiguration, as it would be possible to monitor the traffic profiles and schedule the reconfiguration instantaneously based on the traffic characteristics. In reality, especially in multi-tenant datacenter scenarios, the traffic is related to multiple applications and exhibit temporal and spatial variations that can be difficult to monitor or predict in order to promptly schedule the optical reconfiguration operations, especially when considering this aspect at large scales typical of current datacenters with thousands of racks. This problem is still an open research challenge that is common to any existing optical switching and reconfiguration architecture presented in literature. While these critical architectural and control-plane aspects go beyond the scope of this paper, we will briefly discuss some potential approaches in the final section of this paper (Sect. 6).

Given the premise above, we studied Flex-LIONS with different degrees of reconfigurability to expose the benefits and drawbacks of providing different levels of flexibility (i.e., $b$ number of MRRs per input-output port which determines the link bandwidth enhancement factor). Flex-LIONS Full denotes full reconfigurability (i.e., each wavelength available to a sender can be re-assigned to any desired destinations), Flex-LIONS Half denotes half of the wavelengths can be reassigned, and Flex-LIONS Quarter denotes a quarter of the wavelengths can be re-assigned. In addition, to study the impact reconfiguration can have, we include a simple all-to-all network without reconfiguration capability into our study (LIONS).

Figure 4 shows the performance results for the different synthetic traffic patterns for varied offered network loads. Given that the routing algorithms always choose the shortest path and that there is only one shortest path in an all-to-all network, LIONS without network configuration performs poorly for each traffic pattern aside from uniform random where traffic is evenly distributed across all links. Flexibility in bandwidth reconfiguration is therefore key if the traffic does not follow this corner case. For the different Flex-LIONS reconfiguration capabilities, we observe that the more flexibility in the bandwidth assignment is available, the higher the total accepted traffic gets, which is in line with our hypothesis that fine adjusting the bandwidth to links based on link utilization results in performance gains. Compared to Fat Tree 2:1–a more light-weight implementation of the full Fat Tree– Flex-LIONS nearly doubles the total bandwidth and can compete even for lower levels of reconfiguration.

Figure 5 illustrates the TPW (maximum sustained throughput per Watt) for the different network designs. Energy is reported based on the 65 nm TRX technology and commercially available TRX ('comm'). Flex-LIONS outperforms all other designs significantly on all traffic patterns. Fat tree 2:1 is the closest competitor but only exhibits

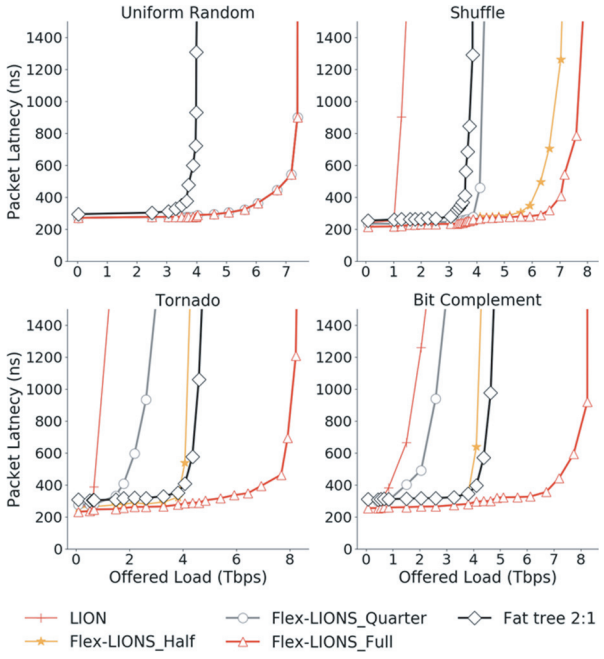



**Fig. 4** Average packet latency (ns) vs. offered load (Tbps) for different synthetic traffic patterns.

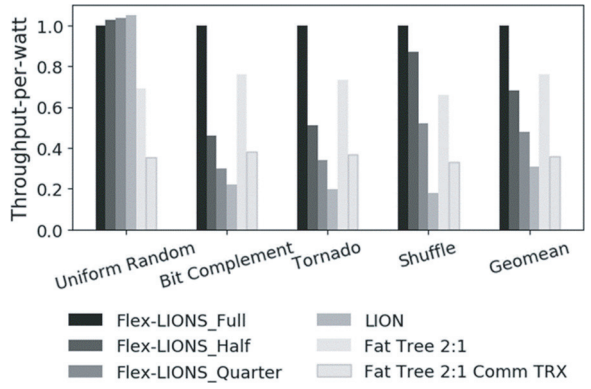



**Fig. 5** Throughput-per-Watt (measured with maximum sustained throughput divided by power consumption) for the different network designs normalized to Flex-LIONS.

0.7× of Flex-LIONS TPW. Though supporting much less maximum throughput, Fat Tree 2:1 is actually more power-efficient than a full Fat Tree as it saves a lot of power through implementing fewer switches and transceivers. In fact, our TPW results reveal that Fat Tree 2:1 would actually be more power-efficient than Flex-LION with reduced bandwidth reconfiguration capability, which wastes much of their bandwidth on barely utilized links.

In fact, unless traffic is perfectly uniformly distributed (in which case Flex-LIONS without reconfigurability capability is the most power-efficient as it provides the same throughput at reduced power), power efficiency in Flex-LIONS is always the best with maximum flexibility in reconfiguration. However, such traffic patterns are very uncommon in both HPC and data center networks and are therefore, of low practical relevance. The network bandwidth reconfigurability of Flex-LIONS is thus not only beneficial in terms of power efficiency but also necessary to compete with state-of-the-art networks.

### 5.2 8-Node Multi-Core Computing System Case Study

The simulation studies reported in this section aim at showing the specific benefits of using multiple FSRs (i.e. two FSRs) in Flex-LIONS. For this scenario, we simulated a network consisting of eight compute nodes, each containing four cores generating traffic according to certain traffic distributions that could be representative of a certain application whose traffic changes temporally and spatially [21], [22]. We defined a traffic distribution among the eight nodes characterized by three phases. In Phase 1, the traffic among the eight nodes is uniform random. In Phase 2, the traffic is composed of a uniform random component plus hotspot traffic between node pairs 0-1 and 4-5. In the third phase, we added two more hotspots between node pairs 2-3 and 6-7. Note that, while careful scheduling of reconfiguration discussed above is still necessary, in such small-scale scenario with a single application running, the problem of scheduling the reconfiguration can be much simpler as it is possible to profile applications and characterize their spatial and temporal traffic behavior [21]–[23].

By using Multi-FSR Flex-LIONS, each phase of the workload can be assigned to a network topology based on the traffic pattern and location of the hotspot links using one FSR, while the second FSR maintains minimum-diameter all-to-all connectivity. We have chosen three topologies based on the different phases of our synthetic traffic. The reconfiguration is implemented using $FSR_1$ while $FSR_0$ implements the basic minimum-diameter all-to-all connectivity. We compared the results of using Multi-FSR Flex-LIONS with regular Flex-LIONS and static all-to-all interconnection (called LIONS).

In the first phase, we assigned an all-to-all topology for both Flex-LIONS and Multi-FSR Flex-LIONS. Since the traffic is evenly distributed between the nodes, having all-to-all connectivity with minimum network diameter represents the optimum solution (for a fair comparison, the bit-rate per lambda in Flex-LIONS is twice the bit-rate per lambda in Multi-FSR Flex-LIONS since in the latter there are two lambdas between each node pair). In the second phase of the proposed workload, we have two hotspot links between nodes 0-1 and nodes 4-5. In all the other links, the traffic is uniform random with lower injection rates compared to the hotspot (80%~ lower). In this phase, for the Flex-LIONS-based topology, it is necessary to remove the links between nodes 0-5 and nodes 1-4 to steer their bandwidth to the hotspot links. Therefore, in this phase, the communication between nodes 0-5 and nodes 1-4 would take place through an additional hop. Differently, when using Multi-FSR Flex-LIONS, the reconfiguration happens on one FSR (e.g. $FSR_1$) while $FSR_0$ still guarantees the shortest path between nodes 0-5 and nodes 1-4. The same discussion ap-

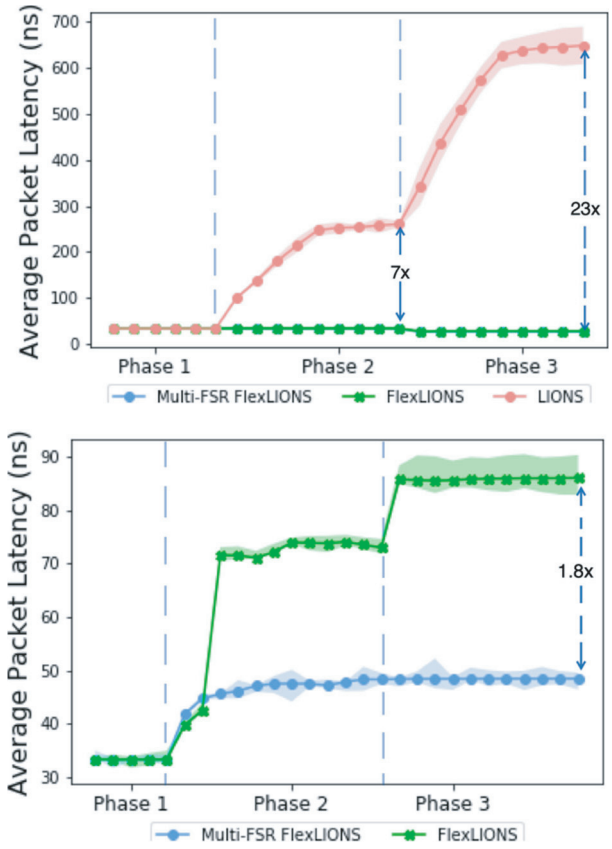



**Fig. 6** Simulation results: (top) the average packet latency comparison between different phases of the application; (bottom) impact of Flex-LIONS and Multi-FSR Flex-LIONS on the background uniform random traffic. Shaded areas represent the minimum and maximum latency measured for each data point.

plies to the third phase, with the only difference that since the number of hotspots doubles, it doubles also the number of node pairs losing the direct connectivity (and requiring one additional hop) when using standard Flex-LIONS.

As shown in Fig. 6 (Top), both Flex-LIONS and Multi-FSR Flex-LIONS perform similarly in terms of average packet latency across the three phases– this is simply because the background uniform random traffic for this simulation scenario has a low injection rate and therefore the number of packets that need to take an additional hop in Flex-LIONS case (because of the deleted links) is limited. However, it is evident the advantage of both architectures against a static full-mesh (all-to-all) named here LIONS. Based on the simulations shown in Fig. 6 (Top), reconfigurability can improve the average packet latency by up to 7× in phase 2, and up to 23× in phase 3. The ratio between the hotspot links to all-to-all links increases at each phase. When using LIONS, at each phase there are more buffers getting congested and therefore the average packet latency increases rapidly at the beginning of each phase. There is also a steady state in all phases where all of the buffers for the hotspot links are fully congested and further packet injection would incur in packet loss (so the average packet latency reported is only for the packets delivered).

The interesting advantage of Multi-FSR Flex-LIONS can be seen actually in Fig. 6 (Bottom), which shows the average packet latency for the background uniform random traffic only. In phase 1, since there is no reconfiguration, the average packet latency is the same for both Flex-LIONS architectures. In phase 2, since two links are removed for Flex-LIONs, there is a certain number of packets that will need to go through more hops to get to their destinations. Even more packets need to take the extra hop for phase 3. Vice versa, with Multi-FSR Flex-LIONS, none of the packets need to take an extra hop. As a consequence, the latency in phase 3 is significantly lower. In conclusion, Multi-FSR Flex-LIONS improves the average packet latency of the background traffic by up to 1.8× over Flex-LIONS in phase 3 by maintaining minimum-diameter all-to-all connectivity through one of the two FSRs.

## 6. Conclusions

This paper presents the architecture, device, system, and networking performance results for Multi-FSR Flex-LIONS, a silicon photonic fabric for reconfigurable optical interconnection in datacenter and HPC network applications. It allows bandwidth steering for hotspot links while guaranteeing single-hop communication for all the other links. Device design, fabrication, and system testing experiments demonstrate error-free bandwidth reconfiguration from 50 Gb/s to 125 Gb/s between selected node pairs. After reconfiguration in $FSR_1$, error-free all-to-all optical communication is maintained through $FSR_0$ with a worst-case crosstalk penalty of ∼ 5 dB.

Through networking simulations with different synthetic traffic patterns, we investigated the impact of the degrees of reconfigurability on the latency and energy efficiency benefits when compared to a legacy Fat-Tree approach. We also demonstrated the benefits of the multi-FSR approach to limiting the impact of reconfiguration on the traffic with uniform random distribution.

As mentioned above, while optical reconfiguration for bandwidth steering, and more in general optical switching, can find application in intra-node, inter-node within a rack and inter-rack communications, the scheduling of reconfiguration for any optical reconfiguration and switching approach is still an open challenge. This is mainly related to the fact that optical switching and its lack of practical optical buffering solutions requires a centralized approach to schedule reconfiguration and avoid contention. This is especially challenging for scenarios where the traffic is generated by multiple concurrent applications (like in datacenters) and over a very large-scale network. Possible solutions currently under investigation include the use of optical packet switching approaches with optical flow control and distributed control planes [24] as well as the application of emerging machine learning techniques for traffic prediction and topology matching [25], [26].

At the physical layer, an important aspect to consider

is the scalability to large radix, which is mainly limited by crosstalk, loss, and the number of wavelengths. A promising approach is using a Thin-CLOS Flex-LIONS architecture as reported in [8]. All of the above aspects will be the objectives of our future studies.

## Acknowledgments

This work was supported by ARO award # W911NF1910470, DoD award # H98230-19-C-0209 and NSF ECCS award # 1611560. The authors would like to acknowledge Yi-Chun Ling from the Next Generation Network System (NGNS) Laboratory at UC Davis, Paul Gaudette, and David C. Scott from Optelligent, LLC for device packaging.

### References

[1] M.J. Schulte, M. Ignatowski, G.H. Loh, B.M. Beckmann, W.C. Brantley, S. Gurumurthi, N. Jayasena, I. Paul, S.K. Reinhardt, and G. Rodgers, "Achieving exascale capabilities through heterogeneous computing," IEEE Micro, vol.35, no.4, pp.26–36, 2015.

[2] S. Mittal and J.S. Vetter, "A survey of CPU-GPU heterogeneous computing techniques," ACM Comput. Surv., vol.47, no.4, pp.69:1–69:35, July 2015.

[3] A. Roy, H. Zeng, J. Bagga, G. Porter, and A.C. Snoeren, "Inside the social network's (datacenter) network," SIGCOMM Comput. Commun. Rev., vol.45, no.4, pp.123–137, Aug. 2015.

[4] Q. Zhang, V. Liu, H. Zeng, and A. Krishnamurthy, "High-resolution measurement of data center microbursts," Proc. 2017 Internet Measurement Conference, pp.78–85, 2017.

[5] K. Wen, P. Samadi, S. Rumley, C.P. Chen, Y. Shen, M. Bahadori, K. Bergman, and J. Wilke, "Flexfly: Enabling a reconfigurable dragonfly through silicon photonics," International Conference for High Performance Computing, Networking, Storage and Analysis, SC, pp.166–177, 2017.

[6] H. Guan, S. Rumley, K. Wen, D. Donofrio, J. Shalf, and K. Bergman, "Reconfigurable silicon photonic interconnect for many-core architecture," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol.10524 LNCS, pp.89–97, 2017.

[7] X. Xiao, R. Proietti, S. Werner, P. Fotouhi, and S.J. B. Yoo, "Flex-LIONS: A scalable silicon photonic bandwidth-reconfigurable optical switch fabric," 2019 24th OptoElectronics and Communications Conference (OECC) and 2019 International Conference on Photonics in Switching and Computing (PSC), pp.1–3, 2019.

[8] X. Xiao, R. Proietti, G. Liu, H. Lu, P. Fotouhi, S. Werner, Y. Zhang, and S.J.B. Yoo, "Silicon photonic flex-LIONS for bandwidth-reconfigurable optical interconnects," IEEE J. Sel. Top. Quantum Electron., vol.26, no.2, pp.1–10, 2020.

[9] X. Xiao, R. Proietti, G. Liu, H. Lu, Y.-C. Ling, Y. Zhang, and S.J.B. Yoo, "Integrated SiPh flex-LIONS module for all-to-all optical interconnects with bandwidth steering," 2020 Optical Fiber Communications Conference and Exhibition (OFC), pp.1–3, 2020.

[10] S. Zhao, L. Lu, L. Zhou, D. Li, Z. Guo, and J. Chen, "16 × 16 silicon Mach–Zehnder interferometer switch actuated with waveguide microheaters," Photonics Res., vol.4, no.5, p.202, Oct. 2016.

[11] N. Dupuis and B.G. Lee, "Impact of topology on the scalability of Mach–Zehnder-based multistage silicon photonic switch networks," J. Light. Technol., vol.36, no.3, pp.763–772, 2018.

[12] B.G. Lee, A. Biberman, P. Dong, M. Lipson, and K. Bergman, "All-optical comb switch for multiwavelength message routing in silicon photonic networks," IEEE Photon. Technol. Lett., vol.20, no.10, pp.767–769, 2008.

[13] R. Stabile, A. Rohit, and K.A. Williams, "Monolithically integrated 8 × 8 space and wavelength selective cross-connect," J. Light. Technol., vol.32, no.2, pp.201–207, 2014.

[14] T.J. Seok, J. Luo, Z. Huang, K. Kwon, J. Henriksson, J. Jacobs, L. Ochikubo, R.S. Muller, and M.C. Wu, "MEMS-actuated 8 × 8 silicon photonic wavelength-selective switches with 8 wavelength channels," 2018 Conference on Lasers and Electro-Optics (CLEO), pp.1–2, 2018.

[15] A.S.P. Khope, M. Saeidi, R. Yu, X. Wu, A.M. Netherton, Y. Liu, Z. Zhang, Y. Xia, G. Fleeman, A. Spott, S. Pinna, C. Schow, R. Helkey, L. Theogarajan, R.C. Alferness, A.A.M. Saleh, and J.E. Bowers, "Multi-wavelength selective crossbar switch," Opt. Express, vol.27, no.4, p.5203, Feb. 2019.

[16] "Intel® Omni-Path Edge Switch 100 Series," [Online]. Available: https://www.intel.com/content/www/us/en/products/network-io/high-performance-fabrics/omni-path-edge-switch-100-series.html [Accessed: 31-Mar-2020].

[17] "Intel® Silicon Photonics 100G CWDM4 QSFP28 Optical Transceiver," [Online]. Available: https://www.intel.com/content/www/us/en/architecture-and-technology/silicon-photonics/optical-transceiver-100g-cwdm4-qsfp28-extended-temperature-brief.html [Accessed: 31-Dec-2019].

[18] H. Li, R. Ding, T. Baehr-Jones, M. Fiorentino, M. Hochberg, S. Palermo, P.Y. Chiang, Z. Xuan, A. Titriku, C. Li, K. Yu, B. Wang, A. Shafik, N. Qi, and Y. Liu, "A 25 Gb/s, 4.4 V-swing, AC-coupled ring modulator-based WDM transmitter with wavelength stabilization in 65 nm CMOS," IEEE J. Solid-State Circuits, vol.50, no.12, pp.3145–3159, Dec. 2015.

[19] N. Binkert, B. Beckmann, G. Black, S.K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D.R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M.D. Hill, and D.A. Wood, "The gem5 simulator," ACM SIGARCH Comput. Archit. News, vol.39, no.2, pp.1–7, Aug. 2011.

[20] N. Agarwal, T. Krishna, L.S. Peh, and N.K. Jha, "GARNET: A detailed on-chip network model inside a full-system simulator," ISPASS 2009 - International Symposium on Performance Analysis of Systems and Software, pp.33–42, 2009.

[21] T. Sherwood, S. Sair, and B. Calder, "Phase tracking and prediction," Proc. 30th Annual International Symposium on Computer Architecture, pp.336–347, 2003.

[22] A. Georges, D. Buytaert, L. Eeckhout, and K. De Bosschere, "Method-level phase behavior in Java workloads," Proc. 19th Annual ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications, pp.270–287, 2004.

[23] Y. Shen, S. Rumley, K. Wen, Z. Zhu, A. Gazman, and K. Bergman, "Accelerating of high performance data centers using silicon photonic switch-enabled bandwidth steering," European Conference on Optical Communication, ECOC, vol.2018-September, 2018.

[24] F. Yan, X. Xue, and N. Calabretta, "HiFOST: A scalable and low-latency hybrid data center network architecture based on flow-controlled fast optical switches," J. Opt. Commun. Netw., vol.10, no.7, pp.B1–B14, July 2018.

[25] S. Salman, C. Streiffer, H. Chen, T. Benson, and A. Kadav, "DeepConf: Automating data center network topologies management with machine learning," Proc. 2018 Workshop on Network Meets AI & ML, pp.8–14, 2018.

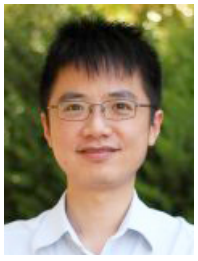
[26] R. Proietti, Y. Shang, X. Xiao, X. Chen, Y. Zhang, and S. Ben Yoo, "Self-driving reconfigurable silicon photonic interconnects (flex-LIONS) with deep reinforcement learning," Supercomputing, Poster 118, 2019.

**Roberto Proietti** received the M.S. degree in telecommunications engineering from the University of Pisa, Italy, in 2004, and the Ph.D. degree in electrical engineering from Scuola Superiore Sant'Anna, Pisa, Italy, in 2009. He is currently a Project Scientist with the Next Generation Networking Systems Laboratory, University of California, Davis, CA, USA. His research interests include optical switching technologies and architectures for supercomputing and datacenter applications, high spectral efficiency transmission systems for core, metro and access networks, and machine-learning aided autonomous cognitive elastic optical networks.

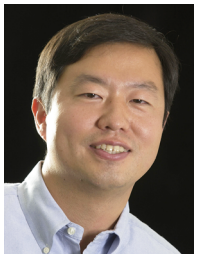
**Xian Xiao** received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 2012 and 2015. He has been working toward the Ph.D. degree in electrical and computer engineering at the University of California, Davis, CA, USA, since 2015. He was a research intern with Nokia Bell Labs in the summer of 2016 and 2017, with Lawrence Berkeley National Laboratory from 2017 to 2018, and with Hewlett-Packard Labs in the summer of 2018. His current research interest includes silicon photonics, optical interconnects, 2.5D/3D photonic integration, neuromorphic computing.

**Marjan Fariborz** received the B.S. degree from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran and M.S. degree from Lehigh University, PA, USA in 2014 and 2018 respectively. She has been working on her Ph.D. degree in Electrical and Computer Engineering at the University of California Davis since 2018. Her current research interest is in the area of computer architecture and networks, high-performance computing, and utilizing optical interconnects in state-of-the-art computing systems.

**Pouya Fotouhi** received the B.Sc. degree in electrical engineering from the University of Isfahan, Iran, and M.Sc. degree in computer engineering from the University of Delaware, Newark, DE, USA, in 2017. He is currently working toward the Ph.D. degree in computer engineering with the Department of Electrical and Computer Engineering at the University of California, Davis, CA, USA. His research interests include optical interconnects, flat memory systems, and heterogeneous computing.



**Yu Zhang** received the B.S. degree in optoelectronics from the Huazhong University of Science and Technology, Wuhan, China, in 2010, and the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology, Hong Kong, in 2016. He is currently an Assistant Project Scientist in electrical and computer engineering with the University of California, Davis, CA, USA. His current research interests include 3-D integrated silicon photonic optical phased array, hybrid silicon laser, amplifier, and phase modulators and silicon photonic integrated circuit for scalable high-performance computing systems.



**S. J. Ben Yoo** received the B.S. degree in electrical engineering with distinction, the M.S. degree in electrical engineering, and the Ph.D. degree in electrical engineering with a minor in physics, all from Stanford University, Stanford, CA, USA, in 1984, 1986, and 1991, respectively. He currently serves as a Distinguished Professor of electrical engineering at UC Davis. His research at UC Davis includes 2-D/3-D photonic integration for future computing, communication, imaging, and navigation systems, micro/nano systems integration, and the future Internet. Prior to joining UC Davis in 1999, he was a Senior Research Scientist at Bellcore, leading technical efforts in integrated photonics, optical networking, and systems integration. His research activities at Bellcore included the next-generation Internet, reconfigurable multiwavelength optical networks (MONET), wavelength interchanging cross-connects, wavelength converters, vertical-cavity lasers, and high-speed modulators. He led the MONET testbed experimentation efforts, and participated in ATD/MONET systems integration and a number of standardization activities. Prior to joining Bellcore in 1991, he conducted research on nonlinear optical processes in quantum wells, a four-wave-mixing study of relaxation mechanisms in dye molecules, and ultrafast diffusion-driven photodetectors at Stanford University. He is a Fellow of OSA and NIAC and a recipient of the DARPA Award for Sustained Excellence (1997), the Bellcore CEO Award (1998), the Mid-Career Research Faculty Award (2004 UC Davis), and the Senior Research Faculty Award (2011 UC Davis).