SwInt: A Non-Blocking Switch-Based Silicon Photonic Interposer Network for 2.5D Machine Learning Accelerators

Ebadollah Taheri, *Student Member, IEEE,* Mohammad Amin Mahdian, *Student Member, IEEE,* Sudeep Pasricha, *Fellow, IEEE,* and Mahdi Nikdast, *Senior Member, IEEE*

***Abstract*—The surging demand for machine learning (ML) applications has emphasized the pressing need for efficient ML accelerators capable of addressing the computational and energy demands of increasingly complex ML models. However, the conventional monolithic design of large-scale ML accelerators on a single chip often entails prohibitively high fabrication costs. To address this challenge, this paper proposes a 2.5D chiplet-based architecture based on a silicon photonic interposer, called SwInt, to enable high bandwidth, low latency, and energy-efficient data movement on the interposer, for ML applications. Existing silicon photonic interposer implementations suffer from high power consumption attributed to their inefficient network designs, primarily relying on bus-based communication. Bus-based communication is not scalable, as it suffers from high power consumption of the optical laser due to cumulative losses on the readers and writers when the bandwidth per waveguide (i.e., wavelength division multiplexing degree) increases or the number of processing elements in ML accelerators scales up. SwInt incorporates a novel switch-based network designed using Mach-Zehnder Interferometer (MZI)-based switch cells for offering scalable interposer communication and reducing power consumption. The designed switch architecture avoids blocking using an efficient design, while minimizing the number of stages to offer a low-loss switch. Furthermore, the MZI switch cells are designed with a dividing state, enabling energy-efficient broadcast communication over the interposer and supporting broadcasting demand in ML accelerators. Additionally, we optimized and fabricated silicon photonic devices, Microring Resonators (MRRs) and MZIs, which are integral components of our network architecture. Our analysis shows that SwInt achieves, on average, 62% and 64% improvement in power consumption under, respectively, unicast and broadcast communication, resulting in 59.7% energy-efficiency improvement compared to the state-of-the-art silicon photonic interposers specifically designed for ML accelerators.**

***Index Terms*—2.5D Networks, ML accelerators, chiplet Systems, Energy-Efficiency, Interposer network.**

I. Introduction

The rapid growth in demand for machine learning (ML) applications has highlighted the critical need for ML accelerators capable of efficiently performing the ever-growing computations required by ML models. These accelerators play a pivotal role in achieving energy efficiency and high-performance computing within the ML domain [1]. Therefore, there has been a surge of interest in the development of ML accelerator architectures that can meet these escalating demands [2].

The authors are with the Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, USA.
E-mail: {ebad.taheri,amin.mahdian,sudeep,mahdi.nikdast}@colostate.edu

Nevertheless, the conventional monolithic design of large-scale ML accelerators on a single chip often leads to prohibitively high fabrication costs, primarily due to low-fabrication yields [3]. Addressing this challenge requires innovative approaches, and one such promising strategy is the adoption of a 2.5D chiplet-based architecture [2], [4]–[7]. In this paradigm, the large-scale accelerator is disintegrated into multiple smaller chiplets, and interconnected through an interposer [8]. These chiplets are linked to the interposer using microbump technology, with the interposer network serving as the vital communication backbone among them.

In the context of 2.5D ML accelerators, a common approach involves the utilization of Multiply-and-Accumulate (MAC) chiplets, alongside a Global Buffer (GLB) chiplet, as seen in state-of-the-art designs [1]–[3], [5]. The GLB chiplet plays a crucial role in storing the weights and activations required for extensive MAC operations. However, efficiently managing data exchange between the GLB chiplet and the MAC chiplets poses a significant challenge to interposer network design, mainly due to the substantial volume and time criticality of data involved.

There are three types of interposers: passive, active, and SiPh interposers. The first two are metallic/electronic interposers, while the latter is an optical interposer designed to support high bandwidth and low latency communication. Passive interposers [9] consist of simple substrates with only wiring layers, offering simpler connectivity between chiplets with lower power consumption but with limited capabilities for long-distance communication. On the other hand, active interposers [10], [11] incorporate complex functions, enabling diverse chiplet communication despite yield and thermal challenges. However, both passive and active electronic interposers suffer from low bandwidth and high latency, especially when scaling up the system [12], [13]. Therefore, electronic interposers are not suitable candidates for large-scale future ML accelerators where high bandwidth is required due to communication-intensive applications, and high latency might be imposed due to the large-scale system and one-to-many and many-to-one communication patterns. SiPh interposers not only offer the high bandwidth and low latency requirements for ML accelerators but also support broadcast communication, where one source can send the same data to all destinations and vice versa, in an energy-efficient manner [4], [5]. However, SiPh interposers designed for ML accelerators [4], [5] employ bus-based communication, which is not energy efficient when

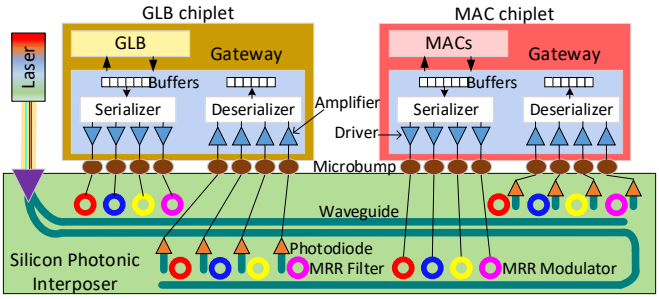

Figure 1. A bidirectional silicon photonic link with four wavelengths to provide optical communication between the GLB chiplet and a MAC chiplet.

supporting a large-scale system.

To overcome this challenge, this paper proposes the integration of switch based SiPh interposers, offering attributes such as low latency, high bandwidth, and energy-efficient communication between the GLB chiplet and the MAC chiplets. While SiPh interposers have exhibited the potential to enhance communication efficiency, recent implementations [1], [4], [5] have suffered from heightened power consumption, primarily stemming from inefficient network design. In response to this issue, this paper introduces SwInt, a non-blocking Switch-based SiPh Interposer for 2.5D ML Accelerators. The main contributions of this paper are as follows:

- We introduce a Butterfly-based interposer network topology aimed at minimizing switch stages, thereby enhancing the energy efficiency of 2.5D ML accelerators.
- Our design incorporates novel techniques to resolve conflicts within the switch, effectively eliminating blocking and ensuring seamless communication.
- Leveraging innovative approaches, we present a broadcast architecture characterized by minimal laser power overhead, optimizing energy efficiency.
- Through optimization of silicon photonic devices, we enhance the energy efficiency of SwInt's switch, crucial for supporting our proposed broadcast architecture.
- We conduct a comprehensive exploration and comparative analysis of switch-based versus bus-based architectures, culminating in the proposal of a hybrid architecture that combines the strengths of both paradigms.

The organization of this paper is structured as follows: In the upcoming section, we review the background and related work. Following that, in Section III, we present our proposed architecture. Section IV involves a comprehensive design space exploration of the switch architecture and the MZI switch cell. Our simulation results are showcased in Section V. Additionally, in Section VI, we explore a hybrid architecture of SwInt with switch and bus. Finally, Section VII concludes the paper and highlights the significance of our contributions.

II. Background and Related Work

A. Silicon Photonic Communication

In Fig. 1, we illustrate a bidirectional SiPh link setup. The GLB chiplet utilizes modulators to modulate electrical signals on optical ones generated by a laser. These modulators can

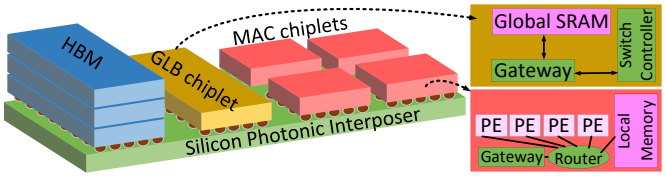

Figure 2. Chiplet system architecture considered in SwInt.

be designed based on microring resonators (MRRs), tuned to resonance at desired wavelengths [14], [15]. The modulated signal is transmitted through a dedicated bus waveguide (depicted in dark green) to the MAC chiplet. The MAC chiplet retrieves the optical signal from the bus waveguide using optical filters, which are custom-designed with MRRs tuned to filter the modulated wavelength. The same process occurs when the MAC chiplet communicates with the GLB chiplet. However, this specific bus-based setup, recognized as a single-writer single-reader (SWSR) configuration, lacks scalability. As the number of writers and readers increases, substantial power is required from the laser source to compensate for optical losses incurred during transmission, including through loss, waveguide crossing losses, and bending losses. To improve scalability, optical switches have been used to devise more scalable network architectures. The switche cell typically feature four ports and can establish either a Bar or Cross connection between two input and output ports (see our designed MZI switch later in Fig. 8). In the Bar state, each input is linked to a symmetrical output port, whereas in the Cross-state, as the name implies, each port connects to the crossing output port. By connecting these four-port switch cells in a specific topology, large-scale switches can be designed to offer high bandwidth while accommodating a greater number of inputs and outputs.

B. Silicon Photonic Interposers

SiPh interposers have received attention in the context of emerging many-core architectures. Fotouhi et al. [16] introduced an interposer design based on Arrayed-Waveguide Grating Routers (AWGRs) to address the latency issues associated with electronic interposers. AWGRs, functioning as passive optical devices that route data by wavelength, offer cost-effectiveness and higher bandwidth compared to their electronic counterparts. However, it is worth noting that achieving high bandwidth with AWGR-based interposers necessitates a significant number of wavelengths, which can result in less power efficiency. PROWAVES [17] and ReSiPI [12] utilize bus-based communication with a single-writer-multiple-reader (SWMR) protocol for inter-chiplet communication while dynamically managing bandwidth. However, as we will show in Section III, such bus-based communication is power efficient only for small-scale systems with a small number of writers and readers. Another work, FLUMEN [18], introduces in-network computing within photonic interposer networks, combining communication and computation. While it enables parallel linear computation during low network loads, this approach may not be ideal for ML applications due to their high data communication demands on the interposer network.

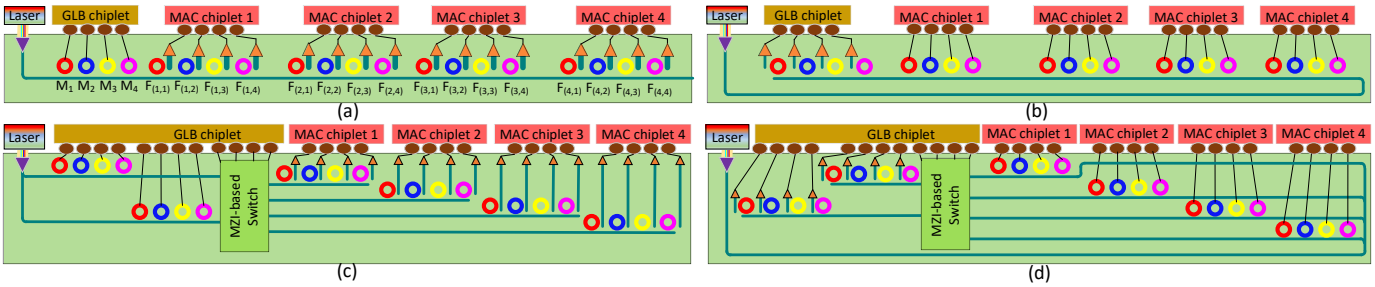

Figure 3. (a) GLB to MAC chiplets communication via a bus-based approach, (b) MAC chiplets to GLB communication via a bus-based approach, (c) SwInt facilitating GLB to MAC chiplets communication using a switch-based interposer, and (d) SwInt enabling MAC chiplets to GLB communication.

SPRINT [5] is another SiPh interposer network designed to facilitate inter-chiplet communication for 2.5D Convolutional Neural Network (CNN) accelerators. It relies on point-to-point SiPh links to establish high-bandwidth, low-latency communication channels between the GLB chiplet and individual MAC chiplets. However, SPRINT's scalability is hindered by the need for separate optical links for each receiver, potentially leading to inefficiencies. While it offers dynamic reconfiguration for broadcast communication, this feature comes at the cost of high laser power tuning and increased latency. SPACX [4] and ASCEND [1], propose SiPh interposer networks tailored for neural network accelerators, that address SPRINT's scalability limitations. Although these networks achieve improved scalability, the proposed solutions introduce challenges related to laser power scalability due to waveguide sharing. To compensate for losses incurred when optical signals traverse multiple receivers, higher optical power is required, resulting in increased energy consumption. In Section III-B, we delve into the energy consumption implications of SiPh bus-based communication in large-scale ML accelerators and motivate our SiPh interposer network architecture.

To improve bus-based architectures, in [19], we introduced TRINE as a silicon photonic interposer network based on SiPh switches, aimed at facilitating energy-efficient ML acceleration. However, TRINE utilizes multiple sub-networks with tree topologies to handle traffic between the GLB chiplet and the MAC chiplets, and vice versa. Although the idea of employing multiple sub-networks is straightforward, it suffers from limitations in flexibility and efficiency of communication. The inherent constraints of the tree topology result in restricted bandwidth between GLB and MAC chiplets, and the utilization of several sub-networks further constrains flexibility as communication is confined within local sub-networks. Additionally, TRINE is incapable of efficiently supporting broadcast communication.

III. Proposed Interposer Network: SwInt

A. Architecture of the 2.5D Accelerator in SwInt

The 2.5D accelerator architecture depicted in Fig. 2 serves as the basis for SwInt. It consists of multiple MAC chiplets, a GLB chiplet, and an HBM main memory. This design aligns with prior research efforts [3]–[5], where an SRAM-based GLB is strategically employed to store the weights and activations required by all MAC chiplets. Each chiplet features a gateway responsible for data storage and forwarding between the GLB and the individual chiplets. These gateways play a pivotal role in controlling the modulator and filter MRRs on the interposer, to facilitate efficient optical communication. The MAC chiplets comprise several processing elements (PEs) interconnected through routers, with each PE housing a MAC unit capable of performing a set of parallel MAC operations. While our primary focus is to design an interposer network for low-latency and energy-efficient communication within this architecture, our network's design principles can be applied to other 2.5D ML accelerators without a loss of generality.

B. Motivation for Switch-based Network Architectures

As previously mentioned, state-of-the-art SiPh interposer networks have been predominantly designed around bus-based communication, a popular choice for small-scale Network-on-Chip (NoC) configurations. However, numerous limitations become exacerbated in the context of large-scale interposer networks. While bus-based communication offers a smaller physical footprint, it exhibits energy inefficiency and lacks the necessary flexibility required for larger systems. For example, in Fig. 3(a), a bus network with SWMR necessitates the optical signal to traverse multiple readers to reach chiplet 4 (the worst-case scenario for defining the required laser power intensity). A similar situation arises in the multiple-writers single-reader (MWSR) case, as depicted in Fig. 3(b), where the writer of chiplet 4 must pass through several MRRs to reach the GLB. In contrast, our switch-based design, illustrated in Fig. 3(c) and Fig. 3(d), addresses these concerns. In the following, we discuss the challenges with bus-based networks in more detail and advocate for a scalable switch-based interposer network.

1) Bandwidth and Adaptation: Bus-based communication struggles to offer high bandwidth efficiently. One potential solution involves increasing the degree of wavelength-division multiplexing (WDM), enabling the transmission of a larger number of wavelengths over the same waveguide. However, this approach can introduce significant challenges, including higher optical power losses and crosstalk noise. As more wavelengths are transmitted in proximity, there is an increased likelihood of signal leakage to nearby filters with close resonant wavelengths, resulting in optical power loss and crosstalk.

In our architecture, MZI switches are employed due to their broadband capabilities and high bandwidth support for WDM, while adiabatic microrings are used as modulators and

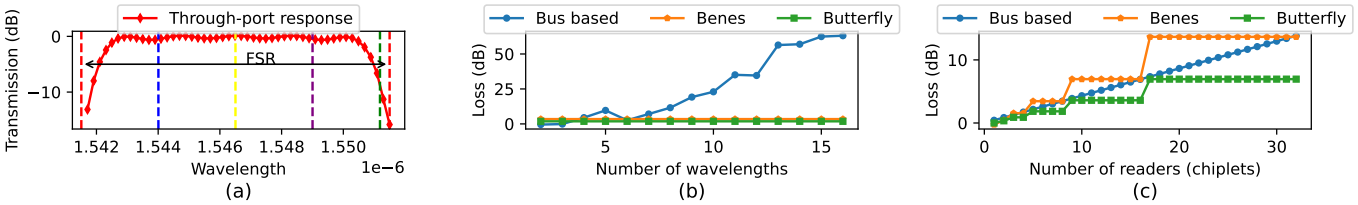

Figure 4. (a) MRR's Through-port response and optical loss under different wavelengths. (b-c) power loss of filters in bus-based network compared to power loss of Benes and Butterfly networks. Assumptions: switch loss is considered; while other loss mechanisms, such as the impact of extra splitters in the switch-based network, couplers, and laser efficiency are not included.

filters for their high modulation speed, which helps mitigate resonance shifts caused by fabrication variations. Our system assumes a modulation frequency of 12 GHz, whereas non-volatile microrings, such as those discussed in B. Tossoun et al. [20], typically offer a modulation frequency of 1 GHz, leading to increased communication latency.

2) Power Inefficiency: In Fig. 4(a), we present the frequency response of an MRR filter on the Through port. The y-axes depicts the transmission of the Through port relative to the input port. The resonant wavelength is indicated by the red lines. In an ideal scenario, 0-dB loss should occur on the Through port for wavelengths other than the resonant one (blue, yellow, and purple). In this scenario with a small number of wavelengths (i.e., four), all the three wavelengths, which are not the intended resonant ones, exhibit nearly 0-dB loss. However, as additional wavelengths are introduced to expand bandwidth, the resonance frequencies of these new wavelengths become closer to that of the red wavelength, potentially leading to power losses. For instance, the green wavelength experiences significant loss in the case of using 32 wavelengths on the same waveguide. This concern exponentially worsens as the number of reader chiplets increases. However, as discussed next, our switch-based approach remains unaffected because the signal does not traverse numerous readers or writers.

We assess the bus-based network alongside two switch-based counterparts: Benes and Butterfly topologies. In our evaluation of the switch-based networks, we exclusively account for losses stemming from the switch itself, encompassing waveguide crossings and MZI switch cells. Conversely, for the bus-based network, we solely factor in the Through losses of MRRs. This delineation reflects the primary distinctions between these two communication paradigms. In Fig. 4(b), we investigate scenarios involving 6 reader chiplets while varying the number of wavelengths from 1 to 16. In Fig. 4(c), we maintain 6 wavelengths and alter the number of readers from 1 to 32. Our findings indicate that switch-based networks employing Benes and Butterfly topologies offer substantially enhanced scalability when increasing the number of wavelengths to achieve high bandwidth. Moreover, a switch-based network employing Butterfly topology demonstrate scalability when increasing the number of readers, even with a relatively low bandwidth (i.e., 6 wavelengths). Notably, the Butterfly topology exhibits lower power loss due to its minimized number of switch stages when compared to Benes. However, this power-saving advantage comes at the cost of increased blocking. In contrast, as we will show later, SwInt addresses this issue by minimizing blocking while retaining the same number of stages. Note that in our loss assessments for Fig. 4(b)-(c), all switch losses, including insertion, crosstalk in both Bar and Cross states, waveguide crossing, and waveguide bends, were considered. Additionally, all microring losses, including insertion and crosstalk in both through and drop states, were accounted for. However, losses from couplers, splitters, and laser efficiency were not included.

3) Multicast and Broadcast Challenges: Prior work [1], [4] assumed that broadcast operations in bus-based SiPh networks offer minimal energy overhead, but such networks present their own set of challenges when used in large-scale systems. Typically, passive filters are considered [1], [12], [17] and in both communication modes, unicast and broadcast, worst-case power loss is considered. However, tuning the laser power for the worst-case scenario (i.e., broadcast) imposes severe energy inefficiency on the laser when unicast mode is used. Considering active filters imposes latency and power overhead to tune filters accurately for all communication modes and routing scenarios. In SwInt, unicast and broadcast operations can be efficiently managed, while minimizing the laser power consumption. Even if considering active filters, in prior work [1], [4], [5], the laser power for broadcast and unicast is considered the same instead of dynamic laser power management. On the other hand, in SwInt, dynamic power management is achieved via dynamic laser power tuning at the MZI splitter, which will be discussed in more detail in Section III-E.

Furthermore, broadcasting in bus-based architectures presents challenges that result in high power consumption and area overhead. As illustrated in the example depicted in Fig. 3(a), M_i sends data to $F_{(i,j)}$, where i denotes the reader (chiplet), and j represents the wavelength. For clarity, we focus on communication involving the red wavelength and consider two scenarios: 1) unicast, where M_1 modulates data for the first chiplet ($F_{(1,1)}$); and 2) broadcast, where all the chiplets receive the data.

To efficiently utilize laser power, each filter must be precisely tuned to deliver a specific portion of the optical signal to the photodetector, with the remainder passed on to other readers. In the unicast scenario, $F_{(1,1)}$ is simply tuned to capture the entire signal with the red wavelength. However, in the broadcast scenario, $F_{(1,1)}$ should be tuned to receive only $1/4$ of the signal, and the rest of the signal should be passed to the other readers. Thus, the filter should be tuned to receive $1/4$ of its input and pass $3/4$.

$F_{(2,1)}$, $F_{(3,1)}$, and $F_{(4,1)}$ should be tuned to receive $1/3$, $1/2$, and $1/1$ of their input, respectively. Regardless of area and controlling overheads, achieving exact tuning of the filters might not be possible due to process and thermal variations [21]. Therefore, laser power should be increased to compensate for such issues, which further results in energy inefficiency.

A similar concept applies to the multiple-writers single-reader (MWSR) protocol, as shown in Fig. 3(b). In our switch-based approach, multicast and broadcast operations can be efficiently managed, while minimizing laser power consumption. Therefore, as shown in our power analysis in Section V-B, SwInt offers higher power efficiency when used in broadcast mode.

C. SwInt Switch Topology

In this paper, our proposed switch topology is based on the Butterfly topology. The Butterfly-based topology of SwInt is the ideal choice for our large-scale ML accelerators, because of its minimized number of stages to improve the footprint and power consumption of the laser. When considering the number of receivers as N, the stages in the Butterfly topology can be calculated as:

$$NS_{But} = \lceil \log_2 N \rceil \tag{1}$$

In contrast, topologies like Benes, with more switch stages, increase costs and power consumption. Additionally, the adaptability of Butterfly topology compared to a traditional bus-based network, though adding routing complexity, provides multiple routing paths between the GLB chiplet and the MAC chiplets, crucial for optimizing ML accelerator performance in dynamic workloads.

However, the original Butterfly switch may experience blocking for specific input-output request combinations. To address this, we optimize the switch topology to effectively reduce instances of blocking. The blocking insances can be improved by adding extra stages. However, adding an extra stage to the switch introduces losses, as each stage requires the signal to pass through an additional MZI, incurring inherent losses within the MZI components. This underscores our commitment to minimizing the number of stages to mitigate such losses and maintain overall efficiency.

In addition to the inherent loss within the switching elements (i.e., MZI), the number of waveguide crossings increases proportionally to the number of stages. The total number of waveguide crossings in the Butterfly topology can be calculated as follows:

$$C_{But} = \sum_{n=1}^{NS_{But}} 2^n + (2^n - 1) \tag{2}$$

For instance, the butterfly switch shown in Fig. 5(b) is a 3-stage switch with 10 crossings. Two crossings are between the first and the second stage, and eight crossings are between the second and the third stage.

Therefore, the worst-case loss of the switch from input ports to any output port is given by:

$$L_{But} = NS_{But} \cdot \max(L_{MZI_{Cross}}, L_{MZI_{Bar}}) + C_{But} \cdot L_{WG_{Cross}} \tag{3}$$

Here, $L_{MZI_{Cross}}$ represents the loss from an input port of MZI to the Cross power of MZI, while $L_{MZI_{Bar}}$ denotes the loss from an input port of MZI to the Bar power of MZI. Additionally, $L_{WG_{Cross}}$ represents the loss of waveguide crossing. A flattened butterfly to improve the crossings is also presented in [22].

Now, let's examine the key distinctions between the SwInt topology and the original Butterfly topology.

Our proposed switch topology, shown in Fig. 5(c) and Fig. 5(d), is based on the Butterfly topology. As previously mentioned, the Butterfly-based SwInt topology stands out as the optimal solution for ML accelerators, primarily because of its laser power efficiency, which is achieved through its reduced stage count. While the standard Butterfly topology, with its typical number of inputs and outputs, may encounter blocking issues, we optimize SwInt to minimize such blocking, thereby enhancing the overall performance. Blocking occurs when competing requests for the same switch cell result in conflicts, hindering the immediate fulfillment of those requests. The number of inputs is inherently smaller than that of a typical Butterfly switch, which reduces the occurrence of blocking scenarios. This is primarily due to the communication pattern within the chiplet-based ML accelerator, which is one-to-many (from the GLB to the MAC chiplets) and many-to-one. Please note that another switch is employed to facilitate communication from MAC chiplets to the GLB chiplet, utilizing the same topology but with inputs and outputs swapped. Due to brevity, we do not delve into the details of this additional switch in this discussion. As shown in Fig. 5(e), the Benes topology offers non-blocking communication at the cost of a larger number of stages and power loss.

To ensure the efficient utilization of bandwidth offered by the GLB chiplet, we align the number of switch inputs (communication lines of the GLB chiplet) with the bandwidth of the GLB chiplet. Let L_{GLB} represent the number of communication lines at the GLB chiplet (switch inputs). This is defined by:

$$L_{GLB} = \lceil \frac{B_{GLB}}{F_M \times N_w} \rceil. \tag{4}$$

Here, B_{GLB} denotes the bandwidth of the GLB chiplet, while F_M and N_W represent the SiPh modulation frequency and the number of wavelengths, respectively. By matching L_{GLB} to the GLB chiplet bandwidth, our design ensures optimal utilization of the communication lines, thereby maintaining efficient data transfer within the system. For instance, considering a 100 GB/s (800×10^9 bits per second) bandwidth of GLB chiplet [2], modulation frequency of 12 GHz [12], and considering 16 wavelengths, L_{GLB} will be 5 lines.

It is important to note that, while our architecture removes certain switches compared to a typical Butterfly/Benes topology, this reduction in input ports is compensated by optimizing resource usage. This strategic allocation ensures that bandwidth remains optimal and the system's communication requirements are adequately met. By addressing potential concerns about the impact of architectural modifications on bandwidth, we emphasize our commitment to maintaining high-performance data transfer.

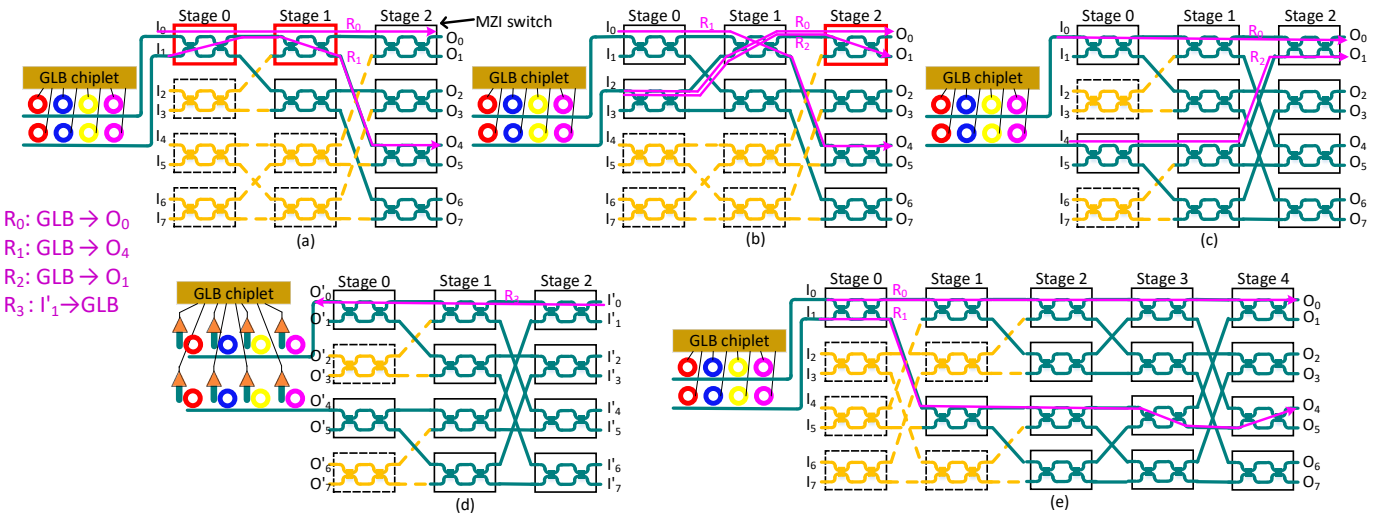

Figure 5. An example of (a-b) Butterfly topology with suboptimal input selection, (c) SwInt network for GLB to MAC chiplets communication, (d) SwInt network for MAC to GLB chiplets communication, and (e) Benes topology. Orange switches are removed compared to a typical Butterfly/Benes topology. Red switches suffer from blocking.

However, the bandwidth between the GLB and MAC chiplets, and vice versa, may be constrained by the maximum bandwidth of the GLB or the chiplet-to-interposer bandwidth. Factors such as the large area and cost of microbumps influence this limitation. To optimize SwInt's performance, we fine-tune the number of switch inputs and outputs based on these bandwidth constraints.

B_{GLB} represents the upper limit imposed by the system configuration and hardware capabilities. By considering the maximum bandwidth of the GLB and the chiplet-to-interposer bandwidth, we adjust our design parameters to ensure efficient data transfer and maintain optimal performance within the given hardware constraints.

There are two factors considered in the SwInt topology to minimize blocking scenarios in comparison with the original Butterfly: 1) input swapping and 2) input selection. The *input swapping* essentially explores various input combinations to establish connections to outputs without causing blockages. This process incurs no performance overhead as all input lines connect to the same source. In Section III-D, we will delve into our offline routing algorithm, which seeks to minimize blocking by evaluating different input swapping options. The *input selection* technique identifies the input combination that minimizes blocking. Fig. 5 shows examples of how SwInt reduces blocking with better input selection. In Fig. 5(a), blocking occurs when considering R_0 and R_1 due to conflicts in switch cells (red switch cells), while in Fig. 5(b), blocking arises when routing R_0 and R_2. However, in Fig. 5(c), SwInt's *input selection* technique is employed, rendering this topology example non-blocking. Our designed algorithm, shown in Algorithm 1, automatically selects the optimized inputs for various combinations of switch sizes and input numbers. Please note that in Fig. 5(d), our switch network supports communication from MAC chiplets to GLB chiplets, akin to the GLB-to-MAC communication but with fewer output links to GLB chiplets. Similarly, Algorithm 2 shows selecting the optimized output of the MAC-to-GLB network. The size of these two networks may vary depending on the bandwidth requirements dictated by the accelerator dataflow.

Algorithm 1 Input selection to generate GLB-MAC SwInt topology

Con_list_i: List of connected inputs to the GLB
L_{GLB_i}: Optical communication line from GLB
L_{SW}: Communication lines of Butterfly = Number of gateways
$L \leftarrow 0$
while Length of $Con\_list_i < L_{GLB_i}$ **do**
  for i in range L_{SW} **do**
    if i not in Con_list_i & length of $Con\_list_i < L_{GLB_i}$ **then**
      if $i\%(L_{SW}/(2^L)) == 0$ **then**
        Add i to Con_list_i
      end if
    end if
    $L \leftarrow L + 1$
  end for
end while

Algorithm 2 Output selection to generate MAC-GLB SwInt topology

Con_list_o: List of connected outputs to the GLB
L_{GLB_o}: Optical communication line to GLB
L_{SW}: Communication lines of Butterfly = Number of gateways
$L \leftarrow 0$
while Length of $Con\_list_o < L_{GLB_o}$ **do**
  for o in range L_{SW} **do**
    if o not in Con_list_o & length of $Con\_list_o < L_{GLB_o}$ **then**
      if $o\%(L_{SW}/(2^L)) == 0$ **then**
        Add o to Con_list_o
      end if
    end if
    $L \leftarrow L + 1$
  end for
end while

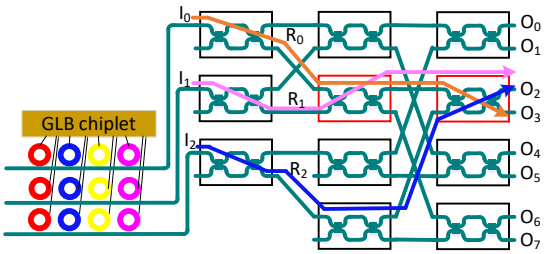

Figure 6. Illustration of input swapping technique. Conflict resolution example using I_0 and I_2 instead of I_0 and I_1.

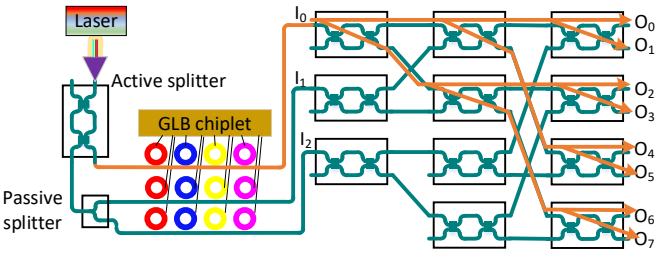

Figure 7. An example of a broadcast in SwInt, achieved using the Divide state in the MZI switch.

D. Routing Strategies and Optimization in SWInt

Switch configuration for routing from inputs to outputs can be accomplished through online or offline methods. The online configuration offers flexibility to adapt to various applications but can become complex, especially for large-scale switches, even though the butterfly switch uses straightforward routing. Fortunately, ML applications typically feature predictable and predefined dataflow communication patterns. In SwInt, we have chosen to implement offline routing analysis, where we optimize switch configurations in advance and store them in a configuration table. During runtime, the switch configuration controller employs this table to configure the switch according to the dataflow. It is worth noting that we plan to explore the development of an online configuration controller for SwInt in future work, extending its versatility to support a broader range of applications. Our offline routing algorithms explore potential combinations of simultaneous outputs within the communication tasks of accelerator dataflow. We refer to Fig. 6 as an example to explain the input swapping technique during the read operation (either fetching input activations or weights). When MAC chiplets are reading data from the GLB chiplet, switch conflicts might arise, leading to blocking of requests. For instance, consider a scenario where two MAC chiplets connected to O_2 and O_3 simultaneously attempt to read from the GLB. If I_0 and I_1 are used to transfer data from the GLB to the MAC chiplets, conflicts within the switch cells can result in blocking. However, by using I_0 and I_2 instead, the data can be sent without any conflict, through R_0 and R_2 paths, therefore avoiding the blocking scenario through the input swapping. Once a non-blocking configuration is successfully identified, considering the input swapping technique, it is recorded in the configuration table. In cases where the algorithm cannot find a non-blocking configuration for a particular combination, it is marked for further processing, with outputs being handled one at a time, although our evaluation in Section V-B shows no blocking even with a high GLB chiplet bandwidth.

E. Multicast and Broadcast Capabilities in SWInt

To enable multicast and broadcast functionality within our SiPh switch, we introduce a new state for our switch cell (our device design of the MZI is elaborated in Section IV). In addition to the Cross and Bar states, traditionally used in optical switches for unicast communication, this new state, called Divide state, allows the input signal to be split between Cross and Bar outputs. This Divide state empowers the switch to direct data to multiple intended outputs concurrently. As depicted in Fig. 7, several switches can be configured in a tree-like pattern, efficiently splitting input data to all the designated destinations. However, to support the receivers adequately, the laser power of the input line needs to be increased. To achieve this, we incorporate an MZI at the input of the first line, serving as an active splitter. This active splitter directs the entire input power of the laser to this line, ensuring sufficient power for all intended receivers. Proper tuning of the laser power is also essential to achieve optimized power in this multicast and broadcast configuration. Algorithm 3 outlines the procedure for managing the SwInt interposer network. The splitting ratio in the split state is:

$$R_S = \frac{1}{L_{GLB} - 1}, \tag{5}$$

For instance, considering L_{GLB} to be 5 lines, a splitting ratio of 0.25 should be used.

Algorithm 3 SwInt network management procedure

```
for All communication tasks in the dataflow do
  if Broadcast or multicast then
    Laser power ← Multicast power (∝ number of readers)
    Switches of the broadcast path ← Divide state (50:50)
    Tunable splitter ← Split state (R_S)
  else
    Laser power ← Unicast power
    Tunable splitter ← Bar state
    Switch is configured according to the configuration table
  end if
end for
```

Our design allocates the switch exclusively to broadcast operations, eliminating the possibility of unicast transmissions during broadcast states. Therefore, all switch cells along the tree path from input to output are in the split state, as illustrated in Fig. 7, effectively preventing any blocking. This approach contrasts with broadcasting via multiple unicasts, which may either introduce latency overhead to circumvent blocking or heighten the likelihood of blocking.

IV. Design and Evaluation of SwInt's Devices

A. MZS Design and Optimization

MZI-based switches (MZS) are one of the most well-known and extensively used devices for optical switching. An MZI switch consists of two 3-dB couplers that are connected to

each other by two waveguide arms. By tuning the phase of either of the arms, the MZI can operate as a switch to route either of the two inputs to either of the two outputs. The 2×2 couplers in the MZI structure are the main source of wavelength dependency and can be realized through multi-mode interference (MMI) couplers since they have higher fabrication tolerance compared to directional couplers and can provide a more uniform 50:50 splitting ratio over the C-band(1.53 $\sim$ 1.56 μm) the fundamentals of MMI operation further discussed in details in [23]. Fig. 8(a) showcases our MZS structure. In this design, by doping silicon to form p-n junctions and applying a voltage across the junction, carriers (electrons and holes) can be either injected into or depleted from the silicon waveguide region. This process changes the carrier concentration in the silicon, which in turn modifies its refractive index through the plasma dispersion effect causing the light traveling through the injected region to experience a different optical path length compared to the light in the non-injected arm. This difference in optical path length leads to a phase shift, altering the interference pattern when the light waves are recombined at the output, and thereby modulating the output signal. Fig. 8(b,c) demonstrates an adiabatic micro-ring resonator that is discussed in the following section.

Fig. 9(a-c) shows three different states of the MZS. In Fig. 9(a), the MZS is in the default Cross state, and that light is switched to the opposite output port. Fig. 9(b) shows the Bar state of the switch where light exits through the same output port as it would by inducing a π phase shift. Fig. 9(c) shows a 50/50 dividing state where the light is evenly split between the two output ports, achieved by precise interference via inducing a $\pi/2$ phase shift. The intricate manipulation and control over these states are managed by optical network controllers [24].

MMI coupler operations are based on multimode interference, where, at specific lengths (L_{MMI}) and widths (W_{MMI}) of the MMI, the optical power can be equally divided between the two outputs [23]. Unlike traditional design methods that proceed from known design parameters to performance outcomes, we employed an inverse design approach for the MMI coupler. This method involves specifying the desired outcome—in our case, the maximization of the figure of merit (FOM)—and using computational algorithms to determine the optimal physical parameters (e.g., L_{MMI}, W_{MMI}) to achieve that outcome. Inverse design is crucial for our project because it enables the exploration of a broader design space, potentially uncovering innovative configurations that traditional methods might overlook. This approach is particularly beneficial for enhancing the performance of the MMI coupler, as it systematically identifies the design parameters that optimize the FOM, defined by:

$$FOM = -(|(T_{Cross} - 0.5)| + |(T_{Bar} - 0.5)|) \quad (6)$$

using a particle swarm optimization (PSO) technique and using Lumerical FDTD simulation where the T_{Cross} and T_{Bar} are the transmission of the Cross and Bar outputs of the MMI coupler. For better mode matching between the input waveguides and the MMI modes, the input waveguides are connected to the MMI structure using tapers, and their geometry is defined by:

$$w(x) = \alpha(L - x)^m + w_2. \quad (7)$$

In this equation, m determines the curvature of the taper and $\alpha = (w_1 - w_2)/L^m$. Here, w_1 and w_2 denote the widths of the waveguides at the start and end of the taper section, respectively, while L represents the taper's length. The optimized dimension for the 2×2 MMI is calculated to be: $L_{MMI} = 17.1\mu$m, $W_{MMI} = 2.2\mu$m, $L_{Taper} = 4.37\mu$m where L and W representing the length and width (see Fig. 8(a)), and the optimum values for parameters in Equation 7 are $m = 1.75$, $w_1 = 500$ nm, $w_2 = 1.08\mu$m , $gap = 680$ nm. We used Lumerical FDTD and DEVICE simulation tools to design and optimize our optical devices. Simulations showed a worst-case transmission loss of a single MMI coupler with these specifications to be 0.02 dB with a power splitting imbalance of less than 0.14 dB over the entire C-band. The designed MZI based on this MMI coupler can achieve a worst-case low loss of 0.12 dB in the Cross state and 0.5 dB in the Bar state using the electro-optical (E-O) tuning method [21]. The lengths of the MZI arms are considered to be 200 μm and the V_π is calculated to be 1.3V with a switching time of $5.7ns$.

To validate our design, we fabricated our shape-optimized MMI couplers through Applied Nanotools Inc. (ANT) foundry utilizing their 220 nm silicon-on-insulator (SOI) technology. The fabrication process was meticulously planned and executed to ensure that the MMI couplers adhered to our precise design specifications, optimizing their performance capabilities. The collected fabrication results from the MMI couplers show a wide band response over the C-band and the measured insertion loss for the central wavelength of operation is calculated to be 0.17 dB. The measured response of the fabricated device is shown in Fig. 10(a) over the C-band central wavelength. It can be observed that the best performance of the MMI is achieved around 1550 nm where the device is optimized and the least power imbalance happens at the same wavelength range. However, the shape optimization of the MMI could provide an outstanding imbalance between the two channels over a wide range of wavelengths. Further comparative analysis among the various fabricated devices was performed to ascertain the consistency and reliability of our MMI designs. A focal point of this evaluation was the determination of the maximum tolerance for power imbalance between the two outputs of the MMI. In our MZS, the MMI couplers are designed to provide evenly distributed outputs. However, any imbalance between these outputs can introduce crosstalk, potentially affecting the performance of the device. Our experimental findings indicated an average maximum imbalance tolerance of <0.5 dB as depicted in Fig.10(b), underscoring the robustness and precision of our design in maintaining uniform output levels under fabrication variations.

This crosstalk resulting from the power imbalance of our optimized MMI is more pronounced as the systems get larger. However, since the number of stages does not increase linearly and our device optimizations have minimized the imbalance, the total loss due to imbalance does not introduce a large overhead to the laser power. In Fig. 11, we show the imbalance

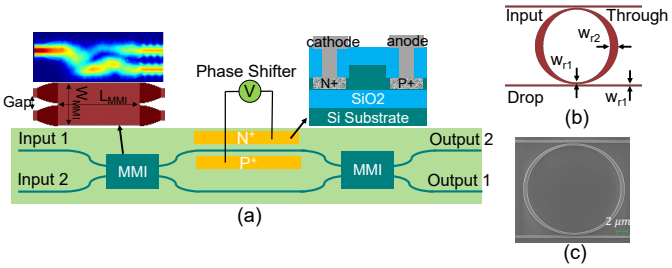

Figure 8. (a) An MZS including the MMI couplers and PN junction phase shifter. (b) an adiabatic MRR ($w_{r2} > w_{r1}$), and (c) SEM image of our fabricated Adevice.

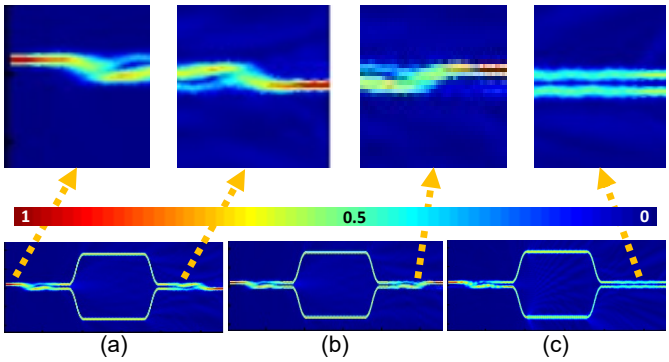

Figure 9. (a) Field distribution of Cross-state (b) Bar-state, and (c) Divide-state.

effect of the "Divide" state on the scalability of broadcasting. In this figure, the total loss imposed by imbalance is compared as the size of the switch increases, and it can be seen that it does not increase linearly. Moreover, the introduction of the 'Divide' state in the MZI switch, allowing for simultaneous transmission to multiple outputs, enhances broadcasting capabilities in an energy-efficient manner as discussed in section V-B

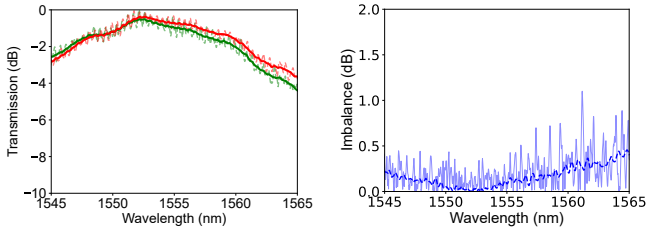

Figure 10. (a) The response of the fabricated MMI. (b) The power imbalance between the two outputs of the MMI.

B. Micro Ring Resonator Design and Optimization

MRRs find extensive application in diverse fields, including filtering, switching, and modulation. Nevertheless, conventional MRR designs are susceptible to process variations, which can unexpectedly shift their resonance frequencies. Addressing these deviations necessitates corrective measures that, unfortunately, come at the expense of increased power

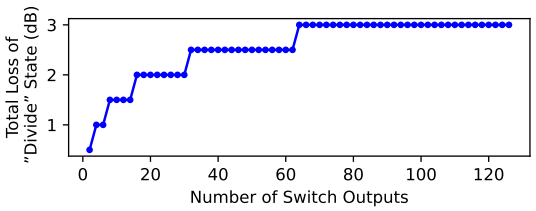

Figure 11. Power imbalance effect of the "Divide" state on the scalability of broadcasting.

consumption and more complex system designs. To overcome these challenges, [25] has designed adiabatic structures to improve the MRRs tolerance toward fabrication errors resulting in smaller resonance wavelength shift per ring ($\Delta\lambda_R$) (see Fig. 8(b)). Adiabatic microring resonators (AMRs) offer significant advantages over traditional microring resonators due to their unique design and operational principles. By incorporating an interior wall design, AMRs effectively cut off higher-order modes that are conventional designs. This design feature ensures that only the lowest-order radial mode propagates, eliminating the adverse effects of higher-order radial modes [26]. Besides, AMRs demonstrate a wide, uncorrupted FSR that is beneficial for applications requiring a broad optical bandwidth.

To enhance our understanding and validation of AMRs, we conducted a comprehensive analysis that included both simulations and experimental fabrication. The device, as depicted in Fig. 8(c), revealed a drop loss of 0.5 dB, attributable to additional losses encountered during testing, whereas our simulations predicted a slightly lower loss of 0.3 dB. The dimensions of the waveguide widths denoted as w_{r1} and w_{r2} and illustrated in Fig. 8(b), were carefully chosen to be 500 nm and 850 nm, respectively. Moreover, the design incorporated a waveguide-ring gap of 100 nm, and the ring itself featured a radius of 10μm. This strategic configuration resulted in a significant enhancement in power efficiency within the switching interface (SwInt) per MRR, demonstrating a 50% reduction in the average wavelength shift ($\Delta\lambda_R$). This strategic optimization notably influences the tuning power required for adjusting the resonant frequencies of the rings, leading to enhanced energy efficiency during device calibration [24]. Tuning in this context refers to the precise adjustment of the resonant frequencies of the AMRs to match specific wavelengths, optimizing the device's performance. Furthermore, our AMR design achieved a FSR of 9.9 nm, complemented by a channel spacing of 0.62 nm across 16 wavelengths, as detailed in Section V-A. This meticulous design ensured that the average Through-port loss per AMR was maintained at a remarkably low level of 0.02 dB, thereby illustrating the efficiency and effectiveness of our AMR in optical signal processing applications.

V. SwInt's Architecture Evaluation

A. Simulation Setup

We conduct a comprehensive comparison involving SwInt against SiPh interposer networks tailored for ML acceler-

Table I
SIMULATION SETUP.

Modulation freq.	12 GHz [12]	MAC chiplets	8
Wavelengths	16 [17]	MACs per Gateway	4
PD sensitivity	9 dBm [27]	PEs per chiplet	16 [2]
Bending loss (90°)	0.01 dB [28]	Data resolution	8 bits [2]
Coupler loss	4.55 dB [29]	Vector MAC Width	8 [2]
Y-splitter loss	0.2 dB [4]	Vector MACs (per PE)	8
Laser efficiency	10%	Weight(per PE)	32 KiB [2]
Gateway freq.	2 GHz	Input buffer(per PE)	8 KiB [2]
Gateways	4 (per chiplet)	Output buffer(per PE)	3 KiB [2]

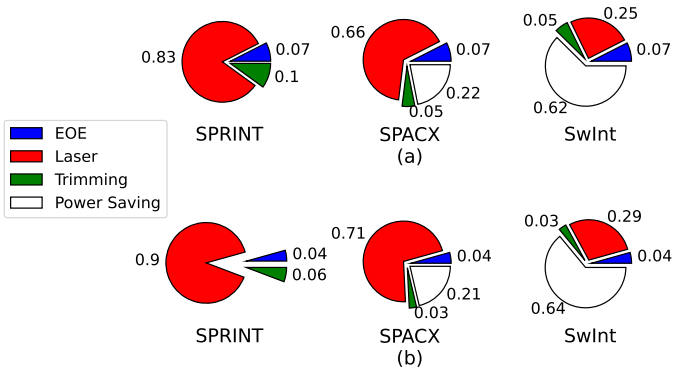

Figure 12. Power analysis: (a) Unicast state, and (b) Broadcast state. White portion in SPACX and SwInt shows power saving.

ation: SPRINT [5] and SPACX [4]. Our analysis encompasses a range of neural network models, employing Tensorflow 2.13.0 alongside Qkeras for model representation. We evaluate six DNN models for the ImageNet dataset, including DenseNet121, EfficientNetB0, LeNet5, MobileNetV2, ResNet50, and VGG16. To ensure consistency, we adhere to a weight stationary dataflow paradigm, where weights remain within vector MAC registers, facilitating reuse across iterations [2]. Our power modeling aligns with the approach described in [30] for laser power estimation and assumed thermo-optic tuning [28] for trimming power considerations. We considered the power model and parameters used in [17] for Electrical-to-Optical (E-O) and Optical-to-Electrical (O-E) conversions. Our latency and energy modelings also align with the analysis principles of MAESTRO [31].

Our chiplet design closely resembles that of Simba [2], featuring sixteen MAC PEs per chiplet. However, in contrast to Simba's utilization of an electronic network-on-chip (NoC), we assume four gateways per chiplet. Four PEs are connected to a router and the router is connected to a gateway, facilitating communication through the SiPh interposer network.

Furthermore, we opt for SRAM technology with size 13 MiB for the Global Buffer (GLB) in line with [3]. Additionally, we set a maximum chiplet-interposer bandwidth of 100 GB/s per chiplet, a constraint determined by the microbump area [2]. For more details on our simulation setup and modeling parameters, please refer to the summary provided in Table I. Please note that the parameters of the devices extracted from our evaluation in Section IV are not re-reported in the Table.

B. Simulation Results

In Fig. 12, we analyze SwInt's power consumption in comparison to other SiPh interposer networks. SwInt demonstrates a significant improvement in power efficiency compared to SPRINT [5] and SPACX [4], primarily attributable to its remarkable reduction in laser power requirements. The lower laser power intensity requirement is due to the fact that the optical signal generated from the laser does not pass through many writers/readers. SPRINT relies on individual photonic links for each Reader, resulting in the incorporation of numerous MRR modulators. Consequently, SPRINT consumes more power for trimming processes compared to SPACX, while SwInt maintains a small trimming power since the number of MRRs are minimized to match the switch bandwidth with the system bandwidth (see equation 4). As mentioned earlier, we employ thermo-optic tuning for trimming, leveraging its broader tuning range [28]. Trimming serves to address thermal and process variations. Each MR filter/modulator undergoes analysis to ensure frequency resonance accuracy. If any deviation occurs due to process or thermal factors, tuning circuits apply voltage to heaters to adjust the MR frequency and align it with the expected frequency. In architectures such as SPACX, SPRINT, and SwInt, trimming power correlates with the number of filters and modulators. Consequently, both SPACX and SwInt encounter similar trimming power requirements, whereas SPRINT necessitates additional trimming power due to its greater number of modulators and filters supporting SWSR communication.

As depicted in Fig. 12(b), in the broadcast state, bus-based architectures (e.g., SPRINT and SPACX) consume more laser power due to inefficient broadcasting. As discussed in Section III-B, achieving broadcast in bus-based architectures requires precise tuning of filters to receive a portion of the signal and pass the rest to other receivers. However, accurate tuning of filters is affected by process variations of MRRs. Consequently, laser power must be increased to compensate for this variability, resulting in power inefficiency in broadcasting for bus-based architectures. To ensure a fair comparison with bus-based architectures, we assume an accurate tuning with 99% accuracy according to [32]. In this analysis, we also considered the imbalanced signal division in the MZI divide state of SwInt. This resulted in a 0.5 dB loss for the worst-case imbalance between the two MZI outputs at each stage, as determined by our device-level analysis of the fabricated MMIs. The white portions in Fig. 12 represent power savings, where SwInt demonstrates significant power improvement of 62% in the unicast state and 64% in the broadcast state compared to the SPRINT architecture.

As shown in Fig. 13(a) SwInt introduces only a minor increase in latency when compared to SPRINT, primarily due to the switch reconfiguration delay. We use the first two letters of the DNN models in the figure. On average, the latency is 7.3% higher than SPRINT and 74.2% lower than SPACX. Also, we estimated the footprint of the SwInt switch, according to our MZS design, which is $0.8mm^2$ (13.3% of our chiplet size).

In Fig. 13(b), we present our energy efficiency results. On

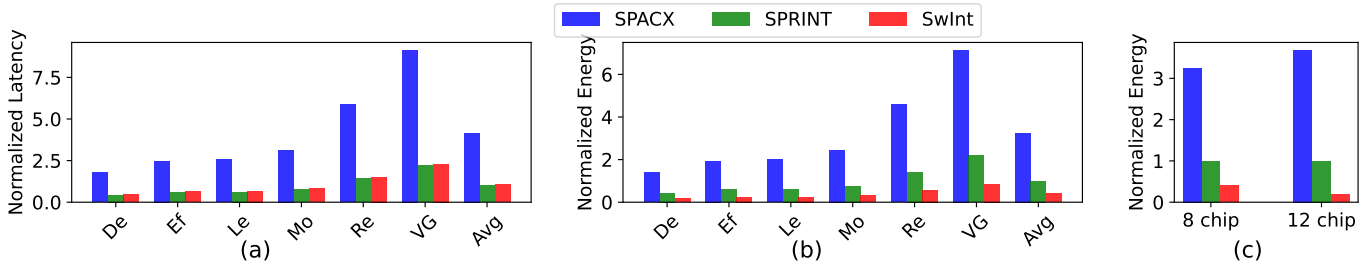

Figure 13. (a) Latency analysis, (b) energy analysis (the acronym represents models' name listed in Section V-A), and (c) scalability analysis. The results are normalized to average case (Avg) in SPRINT

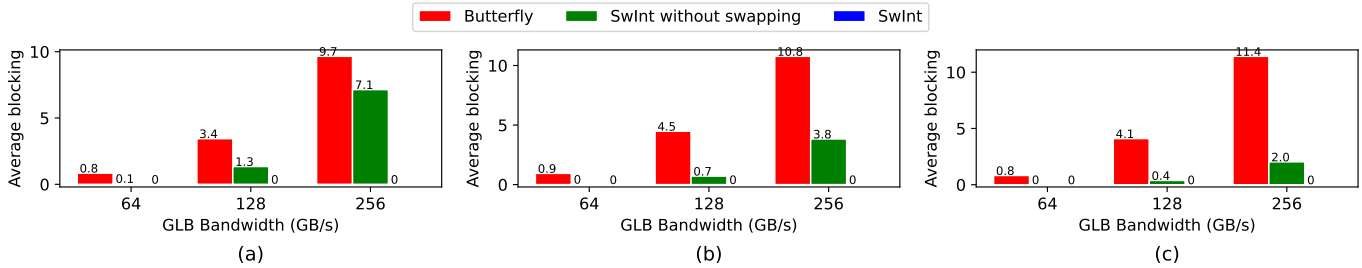

Figure 14. Analyse blocking instances of Butterfly in comparison with SwInt. (a) 4 chiplets, (b) 8 chiplets, and (c) 12 chiplets.

average, SwInt achieves an impressive 87.6% reduction in energy consumption compared to SPACX and a substantial 59.7% reduction compared to SPRINT. This significant energy efficiency can be attributed to SwInt's streamlined network architecture. SwInt significantly improves power consumption while introducing only a minimal latency, resulting in minimized energy consumption. Additionally, its efficient broadcast approach, combined with dynamic laser power management, further enhances energy efficiency. SwInt optimizes network topology ensuring that the interposer network's offered bandwidth aligns with the system's bandwidth requirements. Scalability analysis is also presented in Fig. 13(c), demonstrating a substantial increase in energy efficiency for SwInt as the number of chiplets scales from 8 to 12. This finding underscores the scalability of SwInt's design for future large-scale accelerators.

In Fig. 14, we compare the blocking instances between Butterfly and SwInt. This analysis includes the average number of blocking instances with varying GLB chiplet bandwidth and the number of MAC chiplets. In the experimental setup, we aimed to provide a comprehensive analysis of blocking in our switch application, considering various communication scenarios. While specific dataflows or models may result in lower blocking, we addressed the most pessimistic scenario for robustness. To achieve this, we conducted an exhaustive search across approximately 1000 random communication patterns, calculating the average blocking observed. This approach allows us to present a generalized view of the switch's performance, avoiding potential under-representation of worst-case scenarios.

As can be observed, SwInt (which uses both input selection and input swapping) offers a significant improvement over Butterfly (no blocking in all cases), effectively removing blocking instances. In the SwInt without swapping scenario, where only input selection is used, blocking instances still occur, but they are notably reduced compared to Butterfly. SwInt without swapping is suitable when aiming to use online routing for switch configuration, reducing complexity. However, in our study, we focus on offline routing, where the swapping technique does not introduce any additional overhead to the configuration process.

Although SPRINT and SPACX do not require switches and do not impose the area overhead of switches on the interposer, as we showed they are not scalable in terms of energy efficiency. Based on our designed MZS, we estimated the size of SwInt's switches to be 1.2 mm^2 each. Considering one switch for GLB-to-MAC communication and one for MAC-to-GLB communication, the total area is 2.4 mm^2. We also estimated the size of the interposer to be 192 mm^2, so the switch area is small compared to the interposer (1.25% of the interposer area).

VI. SwInt with Bus Integration

A. Switch-based and bus-based architectures

In advocating for the adoption of a switch-based network architecture in the preceding sections, it is essential to acknowledge that the choice between switch-based and bus-based architectures is context-dependent. While our proposal emphasizes the advantages of a switch-based network for efficient silicon photonic interposers, with increased Wavelength WDM degrees and a large number of MAC chiplets, it is prudent to recognize scenarios where a bus-based may offer more practical solutions.

In small-scale systems, the simplicity and lower overhead of bus-based communication system might outweigh the complexities associated with a switch-based alternative. The re-

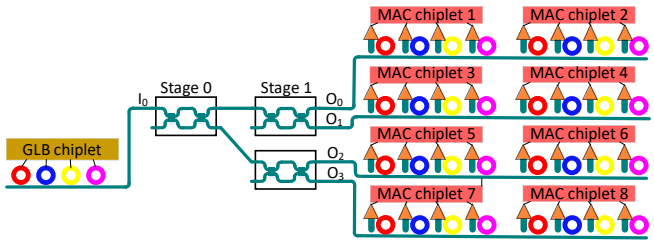

Figure 15. SwInt with Bus, with 2 readers at each output ($B = 2$).

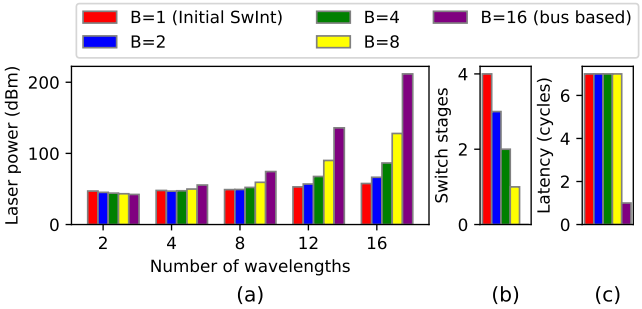

Figure 16. Exploring optimal B (number of readers at each output of SwInt) in SwInt with Bus design under different numbers of wavelengths.

duced hardware requirements and streamlined communication paths can make bus-based architectures a more cost-effective and straightforward solution for such small systems.

Similarly, in instances where the WDM degree is limited, and the overall network complexity is constrained, a bus-based architecture may be a viable choice. The simplicity of bus-based communication can be advantageous when dealing with scenarios where the benefits of a switch-based network may not be fully realized.

B. SwInt with Bus Integration: Design Trade-offs

As shown in Fig. 15, a hybrid architecture can be used for a small-scale network or with a limited WDM degree, where a bus is employed at each port of the switch. This strategic augmentation provides enhanced flexibility, allowing for multiple readers at the output of GLB to MAC switch or several writers at the input of the MAC to GLB switch. In this case, the number of switch stages is reduced. Considering B as the number of readers on each switch output, the number of switch stages is

$$NS_{But_B} = \lceil \log_2 N/B \rceil \tag{8}$$

A smaller number of switch stages results in smaller switch loss (see Equation 3). However, since there are more filters on each output of the switch, losses on microrings increase, which increases laser power. Therefore, depending on the number of wavelengths and number of readers, the optimum architecture may require a different number of readers on each output of the switch. In Fig. 16(a), we compare laser power for different numbers of wavelengths with varying numbers of readers on each switch output. Since laser power dominates the total power of the system (see Fig. 12), minimizing it aids in reducing the overall power consumption. In this analysis, the total number of readers is sixteen. For more than four wavelengths, one reader at each output is the optimum architecture (initial SwInt design). However, in the case of four wavelengths, two readers on each output are optimum, while for the two-wavelength case, having sixteen readers on each output, i.e., removing the switch and having a bus architecture, is preferable. As a result, using a bus on switch outputs is efficient only when the number of wavelengths is small, whereas when the number of wavelengths exceeds four, a pure switch architecture is recommended.

However, reducing the number of readers on switch outputs increases switch size, which can potentially increase area overhead and reconfiguration latency. Therefore, in Fig. 16(b), we explore the number of switch stages under different numbers of readers on each switch output. As can be seen, increasing the number of readers on each output (B) decreases the number of switch stages. However, considering the latency and area overhead, the advantage of having a larger switch under a high-bandwidth network with a large number of wavelengths is beneficial. The reason is that, as mentioned in the last section, the switch size is small compared to the size of the interposer. Moreover, since the switch cells can be configured in parallel, the switch size does not significantly affect the reconfiguration time. We further evaluated the configuration controller of SwInt's switches by implementing it in Verilog and analyzing it using Cadence Genus under 15 nm technology. Our analysis indicates that for all the switch sizes shown in the figure, the configuration decision can be made in a single cycle at 2 GHz frequency. Therefore, as demonstrated in Fig. 16(c), the switch does not impose a high latency compared to its power efficiency benefits.

VII. Conclusion

This paper proposed an innovative SiPh interposer network architecture for large-scale ML accelerators. We developed an energy efficient Butterfly-based interposer network topology to minimize switch stages while preventing blocking. Our architecture improved energy efficiency by 59.7% on average, by utilizing a low loss switch with an efficient broadcast technique. SwInt's MZI-based switch also significantly reduces laser power consumption. We introduced a new "divide" state for our fabricated MZIs alongside existing Cross and Bar states to facilitate our energy-efficient broadcast communication designed to minimize laser power usage. Additionally, we employ active splitters and dynamic laser power management to tune laser power for both unicast and broadcast modes, effectively reducing energy consumption. These innovations offer a significant step forward in addressing the growing demand for efficient ML accelerators, emphasizing the importance of interposer communication in chiplet-based ML accelerators.

Acknowledgment

This work was supported by the National Science Foundation (NSF) under grant number CNS-2046226.

REFERENCES

[1] Y. Li, K. Wang, H. Zheng, A. Louri, and A. Karanth, "ASCEND: A scalable and energy-efficient deep neural network accelerator with photonic interconnects," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 7, pp. 2730–2741, 2022.

[2] Y. S. Shao, J. Clemons, R. Venkatesan, B. Zimmer, M. Fojtik, N. Jiang, B. Keller, A. Klinefelter, N. Pinckney, P. Raina *et al.*, "SIMBA: Scaling deep-learning inference with multi-chip-module-based architecture," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 14–27.

[3] R. Guirado, H. Kwon, S. Abadal, E. Alarcón, and T. Krishna, "Dataflow-architecture co-design for 2.5 d dnn accelerators using wireless network-on-package," in *2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 2021, pp. 806–812.

[4] Y. Li, A. Louri, and A. Karanth, "SPACX: Silicon photonics-based scalable chiplet accelerator for dnn inference," in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2022, pp. 831–845.

[5] ——, "SPRINT: a high-performance, energy-efficient, and scalable chiplet-based accelerator with photonic interconnects for CNN inference," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 10, pp. 2332–2345, 2021.

[6] A. Kannan, N. E. Jerger, and G. H. Loh, "Enabling interposer-based disintegration of multi-core processors," in *Proceedings of the 48th international symposium on Microarchitecture*, 2015, pp. 546–558.

[7] E. Taheri, S. Pasricha, and M. Nikdast, "Red: A reliable and deadlock-free routing for 2.5 d chiplet-based interposer networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024.

[8] J. Yin, Z. Lin, O. Kayiran, M. Poremba, M. S. B. Altaf, N. E. Jerger, and G. H. Loh, "Modular routing design for chiplet-based systems," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2018, pp. 726–738.

[9] P. Ehrett, T. Austin, and V. Bertacco, "Sipterposer: A fault-tolerant substrate for flexible system-in-package design," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019, pp. 510–515.

[10] P. Vivet, E. Guthmuller, Y. Thonnart, G. Pillonnet, C. Fuguet, I. Miro-Panades, G. Moritz, J. Durupt, C. Bernard, D. Varreau *et al.*, "Intact: A 96-core processor with six chiplets 3d-stacked on an active interposer with distributed interconnects and integrated power management," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 79–97, 2020.

[11] E. Taheri, S. Pasricha, and M. Nikdast, "DeFT: A deadlock-free and fault-tolerant routing algorithm for 2.5 d chiplet networks," in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2022, pp. 1047–1052.

[12] ——, "ReSiPI: A reconfigurable silicon-photonic 2.5 d chiplet network with pcms for energy-efficient interposer communication," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, 2022, pp. 1–9.

[13] F. Sunny, E. Taheri, M. Nikdast, and S. Pasricha, "Machine learning accelerators in 2.5 d chiplet platforms with silicon photonics," in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2023, pp. 1–6.

[14] A. Mirza, F. Sunny, P. Walsh, K. Hassan, S. Pasricha, and M. Nikdast, "Silicon photonic microring resonators: A comprehensive design-space exploration and optimization under fabrication-process variations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 10, pp. 3359–3372, 2021.

[15] A. Mirza, S. M. Avari, E. Taheri, S. Pasricha, and M. Nikdast, "Opportunities for cross-layer design in high-performance computing systems with integrated silicon photonic networks," in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2020, pp. 1622–1627.

[16] P. Fotouhi, S. Werner, R. Proietti, X. Xiao, and S. B. Yoo, "Enabling scalable disintegrated computing systems with awgr-based 2.5 d interconnection networks," *Journal of Optical Communications and Networking*, vol. 11, no. 7, pp. 333–346, 2019.

[17] A. Narayan, Y. Thonnart, P. Vivet, and A. K. Coskun, "PROWAVES: Proactive runtime wavelength selection for energy-efficient photonic nocs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 10, pp. 2156–2169, 2020.

[18] K. Shiflett, A. Karanth, R. Bunescu, and A. Louri, "Flumen: Dynamic processing in the photonic interconnect," in *Proceedings of the 50th Annual International Symposium on Computer Architecture*, 2023, pp. 1–13.

[19] E. Taheri, M. A. Mahdian, S. Pasricha, and M. Nikdast, "Trine: A tree-based silicon photonic interposer network for energy-efficient 2.5 d machine learning acceleration," in *Proceedings of the 16th International Workshop on Network on Chip Architectures*, 2023, pp. 15–20.

[20] B. Tossoun, D. Liang, S. Cheung, Z. Fang, X. Sheng, J. P. Strachan, and R. G. Beausoleil, "High-speed and energy-efficient non-volatile silicon photonic memory based on heterogeneously integrated memresonator," *Nature Communications*, vol. 15, no. 1, p. 551, 2024.

[21] S. Pasricha and M. Nikdast, "A survey of silicon photonics for energy-efficient manycore computing," *IEEE Design & Test*, vol. 37, no. 4, pp. 60–81, 2020.

[22] J. Kim, J. Balfour, and W. Dally, "Flattened butterfly topology for on-chip networks," in *40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007)*. IEEE, 2007, pp. 172–182.

[23] M. A. Mahdian and M. Nikoufard, "Thz multimode interference power divider based on groove gap waveguide configuration," *IEEE Transactions on Nanotechnology*, vol. 21, pp. 259–265, 2022.

[24] M. A. Mahdian, E. Taheri, and M. Nikdast, "Pars: A power-aware and reliable control plane for silicon photonic switch fabrics," in *2023 International Conference on Photonics in Switching and Computing (PSC)*. IEEE, 2023, pp. 1–3.

[25] J. C. Mikkelsen, W. D. Sacher, and J. K. Poon, "Adiabatically widened silicon microrings for improved variation tolerance," *Optics Express*, vol. 22, no. 8, pp. 9659–9666, 2014.

[26] M. R. Watts, "Adiabatic microring resonators," *Optics Letters*, vol. 35, no. 19, pp. 3231–3233, 2010.

[27] C. Li, R. Bai, A. Shafik, E. Z. Tabasy, B. Wang, G. Tang, C. Ma, C.-H. Chen, Z. Peng, M. Fiorentino *et al.*, "Silicon photonic transceiver circuits with microring resonator bias-based wavelength stabilization in 65 nm cmos," *IEEE journal of solid-state circuits*, vol. 49, no. 6, pp. 1419–1436, 2014.

[28] F. P. Sunny, A. Mirza, I. Thakkar, M. Nikdast, and S. Pasricha, "ARXON: A framework for approximate communication over photonic networks-on-chip," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 29, no. 6, pp. 1206–1219, 2021.

[29] M. Antelius, K. B. Gylfason, and H. Sohlström, "An apodized soi waveguide-to-fiber surface grating coupler for single lithography silicon photonics," *Optics express*, vol. 19, no. 4, pp. 3592–3598, 2011.

[30] Y. Ye, J. Xu, X. Wu, W. Zhang, W. Liu, and M. Nikdast, "A torus-based hierarchical optical-electronic network-on-chip for multiprocessor system-on-chip," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 8, no. 1, pp. 1–26, 2012.

[31] H. Kwon, P. Chatarasi, V. Sarkar, T. Krishna, M. Pellauer, and A. Parashar, "Maestro: A data-centric approach to understand reuse, performance, and hardware cost of dnn mappings," *IEEE micro*, vol. 40, no. 3, pp. 20–29, 2020.

[32] E. Peter, A. Thomas, A. Dhawan, and S. R. Sarangi, "Active microring based tunable optical power splitters," *Optics Communications*, vol. 359, pp. 311–315, 2016.

Ebadollah Taheri (Student Member, IEEE) earned his M.Sc. degree in Electronic Engineering from Iran University of Science and Technology in 2016. Currently, he is a Ph.D. candidate in Computer Engineering with the Department of Electrical and Computer Engineering at Colorado State University, USA. His research primarily revolves around Reliable and High-Performance Computer Architecture, with a specific focus on On-Chip Interconnection Networks. He is a student member of the IEEE.

Mohammad Amin Mahdian (Student Member, IEEE), earned his M.Sc. degree in Electronic Engineering from Kashan University in 2018. Currently, he is a Ph.D. student in Electrical Engineering with the Department of Electrical and Computer Engineering at Colorado State University, USA. His research focuses on Silicon Photonic integrated devices for communication systems, optical hardware security, and Optical interposer design for machine learning accelerators. He is a student member of the IEEE.

Sudeep Pasricha (Fellow, IEEE), received his Ph.D. in Computer Science from the University of California, Irvine in 2008. He is currently a Walter Scott Jr. College of Engineering Professor in the Department of Electrical and Computer Engineering, at Colorado State University. His research focuses on the design of innovative software algorithms, hardware architectures, and hardware-software co-design techniques for energy-efficient, fault-tolerant, real-time, and secure computing. He has co-authored seven books and published more than 300 research articles in peer-reviewed journals and conferences that have received 17 Best Paper Awards and Nominations at various IEEE and ACM conferences. He has served as General Chair and Program Committee Chair for multiple IEEE and ACM conferences and served in the Editorial board of multiple IEEE and ACM journals. He is a Fellow of the IEEE, Fellow of the AAIA, Distinguished Member of the ACM, and an ACM Distinguished Speaker.

Mahdi Nikdast (Senior Member, IEEE) is an Associate Professor and an Endowed Rockwell-Anderson Professor in the Department of Electrical and Computer Engineering at the Colorado State University (CSU), Fort Collins, CO. He received his Ph.D. in Electronic and Computer Engineering from The Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2014. From 2014 to 2017, he was a Postdoctoral fellow at McGill University and Polytechnique Montreal, Canada. He is the director of the Electronic-PhotoniC System Design (ECSyD) Laboratory at CSU. His primary research goals are focused on the design and development of photonic systems-on-chip and next-generation data-communication, computing, and sensing systems employing integrated photonics while emphasizing energy efficiency and robustness. Prof. Nikdast and his students have received multiple Best Paper awards for their work in the area of integrated photonics and design for manufacturability. He is a senior member of the IEEE.