TRINE: A Tree-Based Silicon Photonic Interposer Network for Energy-Efficient 2.5D Machine Learning Acceleration

Ebadollah Taheri, Mohammad Amin Mahdian, Sudeep Pasricha, and Mahdi Nikdast
Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, USA

ABSTRACT

2.5D chiplet systems have showcased low manufacturing costs and modular designs for machine learning (ML) acceleration. Nevertheless, communication challenges arise from chiplet interconnectivity and high-bandwidth demands among chiplets. To address these challenges, we present TRINE, a novel tree-based silicon photonic interposer network for energy-efficient ML acceleration. Leveraging silicon photonics and broadband optical switching, TRINE enables efficient inter-chiplet communication with reduced latency and improved energy efficiency. Considering several ML workloads, our simulation results demonstrate significant improvements in the average energy efficiency by 61.7% and 40% when comparing TRINE with two recently proposed silicon photonic interposer networks. By overcoming communication limitations in 2.5D ML accelerators, this work is a promising step towards advancing 2.5D photonic-based ML accelerator design.

CCS CONCEPTS

• **Hardware → Photonic and optical interconnect**; **Network on chip**.

1 INTRODUCTION

The computation power required for machine learning (ML) acceleration has been experiencing a relentless surge as the demand for advanced ML applications continues to rise [1, 2]. In response to such an ever-increasing demand, chiplet systems have emerged as a promising solution, offering improvements in reusability, scalability, and manufacturing costs [3, 4]. A critical aspect of chiplet systems lies in their inter-chiplet communication, necessitating a high bandwidth interposer network to realize seamless data exchange among different chiplets [5]. Traditional metallic interconnects, though widely used, suffer from limitations such as low bandwidth and high latency. To address these challenges, silicon photonic (SiPh) networks have emerged as a compelling alternative for interposer communication due to their distance-independent latency and high bandwidth per link [6]. Recent research has shown significant energy and latency improvements by integrating SiPh into the design of chiplet systems for ML hardware accelerators [5, 7, 8].

Despite several advantages, current SiPh-based chiplet systems primarily rely on point-to-point communication [5, 7, 8], employing waveguide configurations such as single-writer-single-reader (SWSR) or single-writer-multiple-readers (SWMR) principles. However, these configurations encounter severe energy overhead and fabrication challenges when the system scales up. In state-of-the-art 2.5D accelerators, there are two main communication flows on the interposer: 1) multiply-accumulate processing elements (MAC PEs) that read inputs or weights of a layer from a global buffer (GLB), and 2) MAC PEs that write the output of a layer to a GLB. This one-to-many and many-to-one communication motivates the design of a SiPh interposer network that can accommodate such patterns in ML hardware accelerators (see Fig. 1).

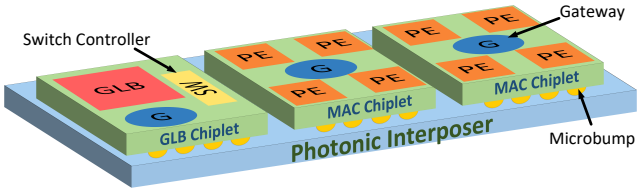

Figure 1: Chiplet-based ML accelerator with SiPh interposer (GLB: Global Buffer, PE: Processing Element).

In this paper, we propose TRINE, a novel SiPh-based interposer network that employs several sub-tree-based networks within the interposer, aiming at maximizing the bandwidth between GLB and MAC PEs while enhancing communication efficiency by supporting energy-efficient data exchange among chiplets. Central to our approach is the adoption of a high-bandwidth SiPh switch to enable a high degree of wavelength-division multiplexing (WDM) with minimal power loss. This seamless integration fosters unparalleled connectivity throughout the interposer network, facilitating efficient data transfer and alleviating bottlenecks. Through extensive simulation-based evaluations, we demonstrate the efficacy of our scalable design which achieves a remarkable 72.8% reduction in latency compared to a recently proposed SiPh interposer [5] and a significant improvement in energy consumption of 61.7% and 40% compared to state-of-the-art SiPh interposers in [5] and [7], respectively. This breakthrough in inter-chiplet communication represents a crucial step forward in the development of energy-efficient ML accelerators, empowering them to fulfill the ever-growing demands of modern ML applications.

The organization of the rest of this paper is as follows. In Section 2, we present some background on chiplet-based ML accelerators and SiPh. Moreover, we discuss related work on SiPh interposers for ML accelerators. In Section 3, we delve into the details of our proposed tree-based SiPh interposer network architecture. In Section 4, we present the performance evaluations that showcase the benefits our design brings to chiplet systems for ML accelerators. Finally, in Section 5, we conclude the paper.

2 BACKGROUND AND RELATED WORK

2.1 Chiplet-based ML Accelerators

Fig. 1 presents an example of a baseline chiplet-based ML accelerator considered in this paper, where three chiplets are integrated

NoCArc 2023, October 28, 2023, Toronto, Canada

ACM ISBN 978-1-4503-XXXX-X/18/06. https://doi.org/10.1145/3610396.3618091

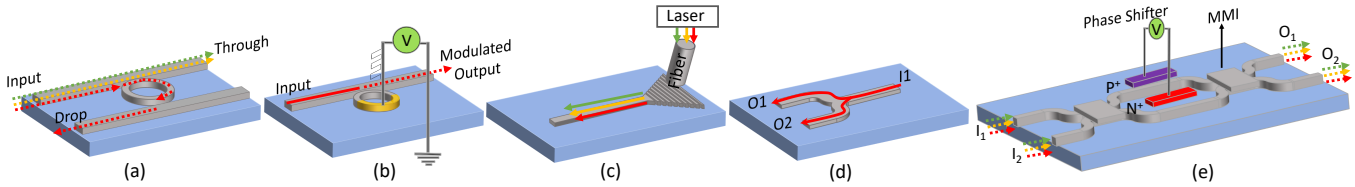

Figure 2: SiPh devices: (a) MRR add-drop filter, (b) MRR modulator, (c) grating coupler, (d) Y-splitter, and (e) the MZS-based switching element for TRINE's tree-based network (MMI: Multi-Mode Interferometer).

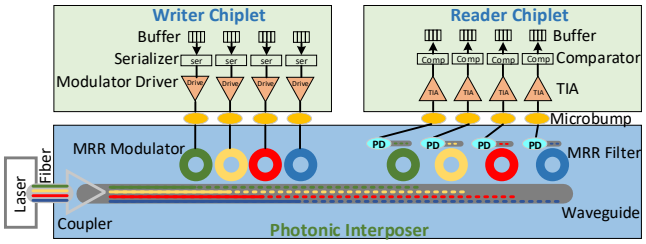

Figure 3: A SiPh link placed on the interposer to send data from a Writer chiplet to a Reader chiplet (PD: photodetector).

on a SiPh interposer. Our network architecture is designed to be seamlessly applicable to any 2.5D accelerator, and while various integration approaches are available [8], the emphasis here lies on addressing high-bandwidth requirements for inter-chiplet communication. In this architecture, there are two types of chiplets: GLB chiplets and MAC chiplets. A MAC chiplet is further divided into several MAC Processing Elements (PEs), each equipped with multiple MAC units capable of parallel MAC operations. For the MAC chiplet, we adopt the chiplet design proposed in [3]. The GLB chiplet feeds new parameters to the MAC PEs placed on the MAC chiplets, necessitating a high-bandwidth communication on the interposer. This has motivated us to design a SiPh interposer network for chiplet-based ML accelerators.

The accelerator employs gateways on both the GLB and MAC chiplets to facilitate efficient inter-chiplet communication, as shown in Fig. 1. These gateways are capable of receiving, storing, and forwarding data between the chiplets and the SiPh interposer, which is the key concept providing high-bandwidth communication within the accelerator. Microbump technology is also considered to provide a connection between a chiplet and the photonic interposer. In the next subsection, we discuss how SiPh is used to provide baseline communication on the interposer.

2.2 Silicon Photonics

Fig. 2 shows some of the most common SiPh devices that are also used throughout this paper. On the SiPh interposer, the connections are established using silicon waveguides that function as the optical counterparts of wires in electronic circuits. Silicon waveguides are characterized by propagation losses, which can range from 0.1 to 3 dB/cm depending on the waveguide structure, caused by surface roughness or scattering loss.

To couple light from an off-chip laser to the photonic interposer, there are primarily two methods. One involves the use of edge couplers, that are placed on the edge of the photonic interposer, while the other is based on grating couplers, as shown in Fig. 2. In the grating coupler method, the input light is directed vertically onto the surface of the grating, allowing the light to couple to the on-chip waveguide (and vice versa for an output light). Note that (de)coupling the light introduces some losses to the system due to the limited efficiency of coupling methods [9, 10].

Microring Resonators (MRRs) in Figs. 2(a) and 2(b) are widely used for wavelength-selective filtering/switching/modulation based on coupling specific wavelengths to the ring from the input waveguide, based on the MRR's physical parameters. These wavelengths are then redirected to the drop waveguide, acting as a filter/switch (e.g., the red signal in Fig. 2(a) matches with the MRR's resonant wavelength, and hence is dropped). However, this process imposes some loss to the optical signal traversing the MRR. Signals dropped to an MRR experience higher losses (drop loss) compared to those passing the MRR to the through port [11]. In an MRR modulator, the wavelength selectivity of the MRR is used to modulate electronic data on different wavelengths, as illustrated in Fig. 2(b). By applying an external electrical signal to the MRR modulator, the MRR's resonant response, which impacts the wavelengths that couple from the input waveguide to the ring, can be tuned, modulating the output light following some modulation scheme (e.g., ON-OFF keying) [12].

In photonic networks, symmetric power dividers (3-dB couplers) like Y-splitters [5], Multi-Mode Interferometer (MMI) couplers [13], and Directional Couplers (DCs) [14] are commonly used. In Fig. 2(d), a 3-dB Y-splitter is depicted, dividing the incoming optical signal into two output paths (or acting as a combiner from the other end) while introducing some optical losses due to such splitting. The extent of such losses depends on the Y-splitter's design, fabrication quality, and the wavelength of operation. Therefore, it is critical to consider signal losses due to power splitters when designing interposer networks.

In a photonic communication on a photonic interposer, as shown in Fig. 3, different wavelengths of light are placed on bus waveguides where different modulators (i.e., MRRs in this example) controlled by a Writer chiplet modulate buffered data from the memory on the optical signals passing the MRRs. Modulated data is transferred over the waveguide links to the Reader filters where it is filtered through optical filters and converted to electrical current by photodetectors. After transferring the current to the Reader chiplet and converting it to electrical voltage using transimpedance amplifiers (TIAs), data is stored in the receiver buffers.

2.3 Prior Related Work

SPRINT [7] is a SiPh interposer designed to facilitate inter-chiplet communication for 2.5D convolutional neural network (CNN) accelerators. The basic network architecture of SPRINT consists of point-to-point SiPh links, allowing the GLB to communicate with

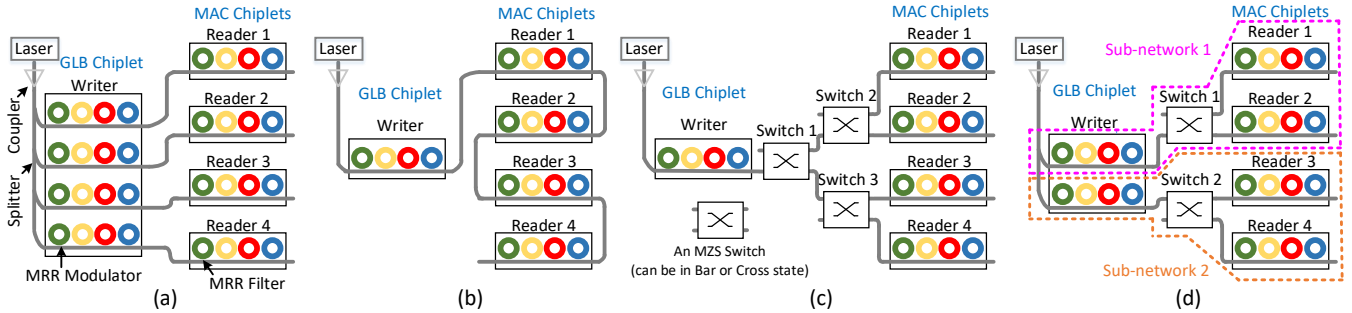

Figure 4: Network architectures and configurations for a SiPh interposer: (a) Point-to-point topology with SWSR paradigm, (b) Point-to-point topology with SWMR paradigm, (c) Tree topology, and (d) proposed TRINE network.

each MAC chiplet individually. While this dedicated link design ensures high-bandwidth and low-latency communication, it becomes less scalable as the GLB requires a separate optical link (e.g., one splitter, one waveguide, and a group of modulators) for each receiver (e.g., a gateway of a MAC chiplet). In Section 3.1, we further compare the point-to-point communication with our TRINE architecture in detail. SPRINT's architecture adopts a SWSR paradigm, but it can be dynamically reconfigured to a SWMR paradigm to facilitate broadcast communication. This enables the GLB to send the reusable data (e.g., inputs of a DNN layer) to all the chiplets simultaneously, hence providing an efficient communication. However, the broadcast is not continuously active, and the laser power needs to be tuned at a high rate—e.g., the worst-case scenario of power loss in the SiPh interposer network—to support the broadcast using the SWMR paradigm. Consequently, the high power consumption of the laser imposes significant energy overhead on the system. Additionally, the reconfiguration process introduces latency due to controlling and tuning the switches, accompanied by additional energy consumption to compensate for the power losses in these switches.

SPACX [5] is another SiPh interposer designed specifically for neural network accelerators. SPACX addresses the scalability limitations observed in SPRINT by adopting the SWMR paradigm for communication. Unlike SPRINT, SPACX achieves better scalability as the number of optical links at the sender (GLB) is independent of the number of receivers (MAC chiplets). This design improvement enables SPACX to efficiently handle a larger number of chiplets. However, while SPACX improves on some scalability aspects, it introduces a new challenge in terms of laser power scalability. The data transmission from GLB to the MAC chiplets involves sharing waveguides. When an optical signal from the laser crosses multiple receivers, the through losses of MRR filters on these receivers accumulate. Consequently, the laser must generate higher optical power to compensate for these losses, leading to increased energy consumption and reduced overall energy efficiency.

In this paper, TRINE is introduced as a novel approach for inter-chiplet communication in 2.5D ML accelerators. TRINE employs tree-based connections between a sender and a group of receivers to minimize signal attenuation and improves laser power efficiency. It employs multiple tree-based sub-networks to enhance inter-chiplet bandwidth through parallel data communications. TRINE also utilizes a broadband switch to support high bandwidth communication, making it a scalable and energy-efficient solution.

3 TRINE: PROPOSED PHOTONIC INTERPOSER

3.1 TRINE Network Architecture

TRINE uses a tree-based architecture that allows data to be exchanged directly without traversing multiple receivers, thus minimizing signal attenuation and reducing the laser power penalty to compensate for losses. Compared to a single tree-based architecture, TRINE utilizes multiple tree-based sub-networks to facilitate parallel and high bandwidth communication between the GLB and the MAC chiplets. Note that, for brevity, we are only discussing sending data from GLB to the MAC chiplets in this section. The process of sending data from MAC chiplets to the GLB is performed following the same concept and without any loss of generality.

To motivate our TRINE network architecture, Fig. 4 shows several SiPh network architectures compared with TRINE architecture. In the SWSR design, shown in Fig. 4(a), each chiplet requires a separate photonic link, including one waveguide and a group of modulators, from the GLB. While this approach allows for direct communication between the GLB and individual MAC chiplets, it becomes a scalability challenge as the number of MAC chiplets (receivers) increases. Additionally, the presence of multiple modulator MRRs connected to the GLB chiplet leads to higher power consumption during MRR tuning. Furthermore, the SWSR design may lead to the inefficient utilization of network resources, as several photonic links remain unused, when GLB bandwidth or chiplet-interposr bandwidth is smaller than the modulation bandwidth, causing inefficiencies in communication and resource wastage.

In contrast to SWSR, the SWMR design shown in Fig. 4(b) offers greater scalability by using a single row of MRRs at the GLB, reducing the need for a separate photonic link for each Reader. Additionally, SWMR allows for broadcast communication, enabling the GLB to send the same data to all the MAC chiplets simultaneously. However, there are challenges in large systems due to significant through-loss accumulation of the MRR filters, resulting in higher power consumption. The broadcast feature, while advantageous, demands higher laser power even when not utilized constantly, leading to energy wastage during unicast communications.

Fig. 4(c) shows a SiPh tree-based architecture. The tree-based architecture employs a tree switch, making it a scalable design. Compared to SWMR, the through loss of filters is improved, resulting in better power efficiency. As the number of switch stages increases at a lower order than the number of filters in SWMR, though the switch cells impose some loss, the power consumption of this architecture is more scalable than SWMR. The tree-based

architecture can be particularly advantageous in terms of power consumption when the number of chiplets in the system grows.

To elaborate on the power efficiency of Tree architecture in comparison with SWMR, we modeled the laser power of both architectures. The laser power should be larger than the sensitivity of the photodetector when received at the Reader. The optical signal received at the photodetector includes the losses introduced by different SiPh devices on the path from the laser to the photodetector of the receiver. Therefore, the minimum laser power is:

$$P_{laser} \geq P_{loss} + S_{PD}, \quad (1)$$

where P_{loss} represents the sum of losses from the SiPh devices along the path between the laser and the receiver photodetector. Also, S_{PD} denotes the sensitivity of the photodetector. Note that P_{loss} includes the through loss by passing the MRR filters. Therefore, as the number of MRR filters increases on the path, the losses add up, and the laser power must be increased to compensate for such losses. Thus, SWMR is unscalable to include a large number of Readers due to the high power losses on the SiPh path. On the other hand, the losses on the filters of the tree architecture do not significantly increase with large numbers of wavelengths or Readers.

We illustrate a comparison of laser power between these two architectures in Fig. 5. The figure demonstrates the power improvement of the tree-based architecture in comparison to SWMR. Notably, the Tree architecture shows a significant enhancement in laser power when both the number of wavelengths and/or the number of gateways scale up. However, as the Writer can send data to only one Reader at a time, the maximum network bandwidth achieved from the GLB is limited.

TRINE, shown in Fig. 4(d), addresses the limitations of the Tree architecture by utilizing several sub-networks with tree topology to improve the network bandwidth connected to the GLB. We break a large tree network into several smaller trees and we call each small tree a sub-network. Each sub-network handles communication with a group of Readers. We show an example in Fig. 4(d) where the large tree network in Fig. 4(c) is broken into two sub-networks; sub-network 1 handles communication to Reader 1 and Reader 2 while sub-network 2 is used to communicate with Reader 3 and Reader 4. The number of sub-networks can be adjusted based on the GLB bandwidth, making it a flexible and efficient solution. We discuss the optimized number of sub-networks in TRINE in Section 3.3. TRINE stands out with its improved scalability, power efficiency, and flexibility in managing communication bandwidth compared to SWSR, SWMR, and the conventional tree-based architecture. By eliminating the need for a separate sender hardware for each chiplet, TRINE offers better scalability than SWSR and SWMR. Additionally, it efficiently manages through losses, resulting in improved power efficiency in large systems. All such advantages make TRINE a promising network candidate for inter-chiplet communication in 2.5D ML hardware accelerators, providing significant advancements in communication bandwidth and power efficiency. However, TRINE requires a high-bandwidth broadband switching element to support a large number of wavelengths and a high degree of WDM, which is discussed next.

3.2 Switching Element Design

A Mach–Zehnder Interferometer (MZI)-based switching element (MZS), shown in Fig. 2(e), switches the input light between the two

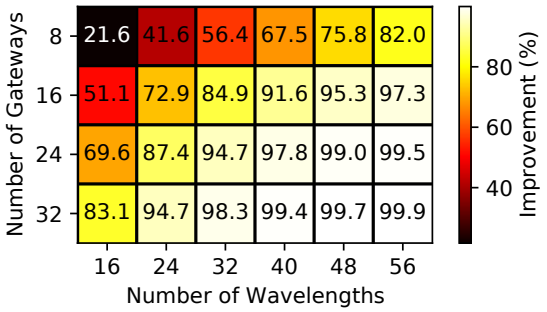

Figure 5: Laser power improvement of tree-based architecture (see Fig. 4(c)) compared to SWMR (see Fig. 4(b)) considering different number of gateways and wavelengths.

outputs based on the phase difference between the splitted signals on its arms. Such a phase difference can be induced using a phase shifter (a fast phase shifter is considered in our work). The MZS can be in two states. In the Bar state, all wavelengths on I_1 and I_2 are guided to O_1 and O_2, respectively. While in the Cross state, all wavelengths on I_1 and I_2 are guided to O_2 and O_1, respectively.

Common methods for implementing phase shifters in SiPh include Electro- and Thermo-optic (EO and TO) tuning. TO tuning is based on the use of micoheaters to leverage the Silicon TO coefficient (1.86×10^{-4} / K [15]). In carrier injection phase shifters (i.e., EO tuning), a P-I-N diode is created by P and N doped regions on the sides of the waveguide, and when the diode is biased (forward or reversed), it enables the free-carrier plasma dispersion (FCD) effect [15]. EO tuning comes with the benefit of faster tuning time (in the range of nanoseconds) at a cost of higher optical loss compared to the TO tuning mechanism. Therefore, in the MZS in Fig. 2(e), the insertion loss when the MZS is in the bar state is higher (by ≈1 dB), when the EO phase shifter needs to induce a π phase shift.

As the data is modulated on different wavelengths, also known as channels, having a broadband switching element is necessary for the operation of TRINE. Moreover, sufficient channel spacing is necessary because decreasing it imposes limitations in the design of MRR filters since the slightest shift in the resonance of an MRR, e.g. caused by thermal variations, causes inter-channel crosstalk. The MZS used for TRINE modeling has a bandwidth of 30 nm [16] that can cover the entire optical C-band and does not impose aggressive channel spacing. Therefore, employing the MZS to make the tree-based architecture of TRINE does not limit the bandwidth of communication between pairs of Writer-Reader. We also consider the MZS of TRINE under the GLB chiplet where the configuration of MZS is managed by a controller on the GLB chiplet (see Fig. 1). We use table-based routing [17] in our controller to achieve a fast routing decision and high-speed controller.

3.3 Number of Sub-networks and MZS in TRINE

The number of sub-networks in TRINE is influenced by the bandwidth-power trade-off in the network. Decreasing the number of sub-networks can introduce bandwidth overhead while improving power efficiency, as the laser needs to support a smaller number of SiPh communication links. On the other hand, a larger number of sub-networks can potentially be more bandwidth-efficient but lead to increased laser power overhead. Nevertheless, the bandwidth between GLB to MAC chiplets and from MAC chiplets to GLB may be

limited by the maximum bandwidth of GLB or the maximum chiplet-to-interposer bandwidth, dictated by factors such as the large area and cost of microbumps. To optimize TRINE's performance, we fine-tune the number of sub-networks based on the maximum bandwidth of GLB and the maximum chiplet-to-interposer bandwidth. We define this maximum bandwidth as the limitation imposed by the system configuration and hardware.

Let us assume the maximum bandwidth that can be achieved by the system hardware from a Writer's perspective is B_w. Then, we adjust the bandwidth of the network to operate at B_w to achieve the maximum performance while avoiding unnecessary power consumption. With this definition, the number of SiPh communication links (number of rows in the MRR group of the Writer) is:

$$N_{PLink} = \frac{B_w}{N_{lambda} \times F_{modulation}}, \tag{2}$$

where N_{lambda} is the number of wavelengths and $F_{modulation}$ is frequency of data modulation in the SiPh interposer. Based on the number of SiPh communication links, we can calculate the number of sub-networks in TRINE to save power while offering the maximum bandwidth:

$$N_{sub-networks} = 2^{\lceil \log_2 N_{PLink} \rceil}. \tag{3}$$

In addition to the smaller number of SiPh links, the number of MZS and switch stages in TRINE is significantly smaller than the Tree architecture when the network scales up. The number of stages in TRINE is:

$$N_{S_{TRINE}} = N_{S_{tree}} - \lceil \log_2 N_{sub-networks} \rceil. \tag{4}$$

Accordingly, the number of MZS in TRINE is:

$$N_{MZS_{TRINE}} = N_{sub-networks} \times (2^{N_{S_{TRINE}}} - 1). \tag{5}$$

4 EVALUATION RESULTS AND DISCUSSIONS

4.1 Simulation Setup

We compare TRINE with the state-of-the-art SiPh interposers designed for ML acceleration: SPRINT [7] and SPACX [5]. We use Tensorflow 2.13.0 along with Qkeras to model various neural network models. We use six DNN models under the ImageNet dataset for our analysis: DenseNet121, ResNet50, LeNet5, VGG16, MobileNetV2, and EfficientNetB0. We assume weight stationary for dataflow where the weights stay within the vector MAC registers to be reused throughout the iterations [3]. We consider the power modeling used in [24] for laser power and the power model employed in [22] for trimming power. Moreover, for the power of electronic circuitry to perform Electrical-to-Optical (E-O) and Optical-to-Electrical (O-E) conversions, we consider the power model and parameters used in [19]. We also adopt the chiplet design of Simba [3], featuring sixteen MAC PEs per chiplet. However, instead of the electronic network-on-chip (NoC) used in Simba, we assume four gateways per chiplet, where each PE is directly connected to the gateway to communicate through the photonic interposer network. We also considered SRAM technology for GLB [25] and assumed a maximum chiplet-interposer bandwidth of 100 GB/s per chiplet [3], which is limited by the microbump area. More details of our simulation setup and modeling are summarized in Table. 1.

Table 1: Simulation parameters and setup (E-O: Electrical-to-Optical, O-E: Optical-to-Electrical).

Photonic interposer	Modulation frequency	12 GHz [18]
	Number of wavelengths	16 [19]
	Channel spacing	50 GHz [20]
	MZS bandwidth	30 nm [16]
	MZS switching time	5.7 ns [16]
	MZS Bar-state loss	1.44 dB [16]
	MZS Cross-state loss	0.44 dB [16]
	Photodetector sensitivity	9 dBm [21]
	MRR through loss	0.02 dB[22]
	MRR drop loss	0.7 dB[22]
	Waveguide propagation loss	1 dB/cm[22]
	Bending loss	0.01 dB/90°[22]
	Grating coupler loss	4.55 dB [9]
	Y-splitter loss	0.2 dB [5]
	Laser efficiency	10% [23]
	E-O and O-E power modeling	based on [19]
Package	Number of MAC chiplets	8
	Number of MACs per gateway	4 [8]
	Max chiplet-interposer bandwidth	100 GBs/chiplet [3]
	Number of gateways	32 (for MAC PEs)
MAC Chiplet	Number of PEs per chiplet	16 [3]
	Gateways per chiplet	4
	Data resolution	8 bits [3]
	Vector MAC Width	8 [3]
	Number of Vector MACs per PE	8 [3]
	Weight buffer size of PE	32 KiB [3]
	Input buffer size of PE	8 KiB [3]
	Output buffer size of PE	3 KiB [3]
	Frequency	2 GHz [8]

4.2 Evaluation Results

Fig. 6 compares TRINE's performance with other SiPh interposer networks for ML acceleration. Fig. 6(a) shows the power analysis, demonstrating that TRINE exhibits a significant improvement compared to SPRINT [7] and SPACX [5]. Both Tree and TRINE architectures demonstrate over 50% power improvement when compared to SPRINT, primarily due to the remarkable reduction in laser power. Notably, SPRINT employs higher power consumption on the laser to support broadcast, even when the broadcast is not required throughout the network operation.

As SPRINT has a separate photonic link at the Writer corresponding to each Reader, it has many MRR modulators, so it consumes higher power on trimming in comparison with SPACX. Similarly, TRINE imposes trimming and laser power overhead compared to Tree topology due to having more photonic links at the Writer to support higher bandwidth. However, this helps TRINE to offer significant latency improvement, as shown in Fig. 6(b). Tree architecture shows a latency overhead compared to SPACX, since the MZS switch cells require time to be configured when forwarding data to different Readers. Similarly, TRINE exhibits only a minor latency overhead compared to SPIRINT, which is imposed by the switch reconfiguration delay (i.e., 5.7 ns). We also show energy results in Fig. 6(c). On average, TRINE archives 61.7% energy improvement compared to SPACX and 40% improvement compared to SPRINT,

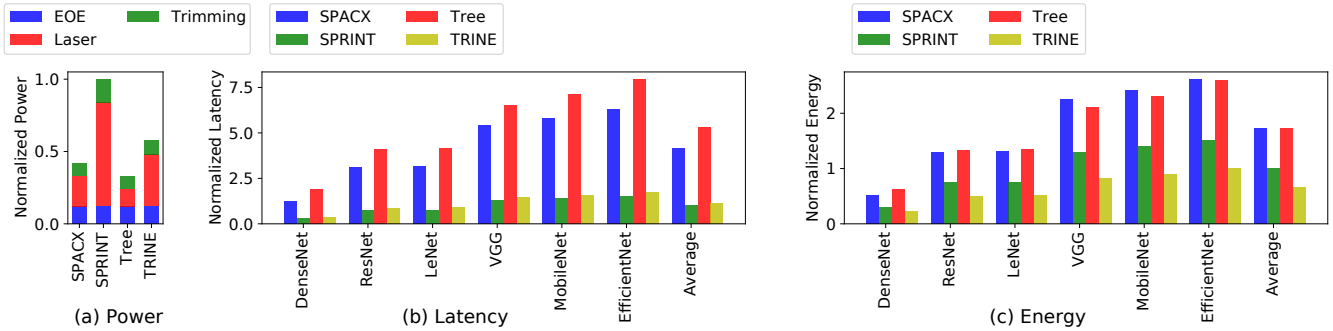

Figure 6: Network performance of TRINE compared to SPACX [5] and SPRINT [7]: (a) Network power, (b) Normalized network latency, and (c) Normalized network energy (normalized to "Average" case in SPRINT).

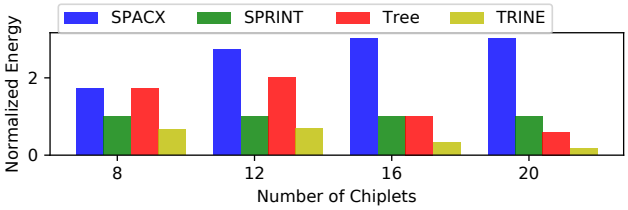

Figure 7: Scalability analysis: Normalized energy with respect to the number of chiplets.

due to its efficient network architecture. Moreover, TRINE achives 61% energy improvement compared to Tree architecture, because it can offer high and efficient bandwidth using parallel communication in our sub-networks. As we discussed in Section 3.3, due to the optimized sub-networks of TRINE, the offered bandwidth of our interposer network is matched with the system bandwidth, which results in energy optimization compared to the Tree architecture.

Based on the MZS in [16], we estimated the size of TRINE's switches to be 1.05 mm^2, which is 17.5% of Smiba's chiplet [3] and less than 1% of Smiba's package. TRINE achieves around 60% reduction in area overhead compared to the Tree architecture in our case study, as TRINE has less number of switching stages. We also implemented the configuration controller of TRINE's switches in Verilog and analyzed the area using Cadence Genus under 15 nm technology, which incurred negligible overhead (i.e., 204 μm^2). Moreover, we analyzed the scalability of TRINE in comparison with other SiPh interposers. Fig. 7 shows the normalized average energy consumption across all six DNN models with varying numbers of chiplets. As shown, when the system scales up, energy consumption significantly improves compared to both SPRINT and SPACX. In the case of 20 chiplets, the energy improvement in TRINE is 82.2%.

5 CONCLUSION

This paper presented TRINE, a SiPh interposer network for 2.5D ML hardware accelerators. We use a tree-based network architecture between the GLB and MAC chiplets to support energy-efficient communication. Compared to a conventional tree-based architecture, TRINE employs several tree sub-networks to enable high-bandwidth communication between GLB and MAC chiplets. Our simulation shows that, on average, TRINE improves energy consumption by at least 40% compared to state-of-the-art SiPh interposers designed for ML accelerators. TRINE's high energy-efficiency and bandwidth performance helps pave the way for enabling high-performance and cost-effective ML accelerators.

6 ACKNOWLEDGMENT

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

REFERENCES

[1] Ananda Samajdar et al. A systematic methodology for characterizing scalability of dnn accelerators using scale-sim. In *ISPASS*, 2020.
[2] Febin P Sunny et al. A survey on silicon photonics for deep learning. *ACM JETC*, 2021.
[3] Yakun Sophia Shao et al. SIMBA: Scaling deep-learning inference with multi-chip-module-based architecture. In *MICRO*, 2019.
[4] Ebadollah Taheri et al. DeFT: A deadlock-free and fault-tolerant routing algorithm for 2.5 d chiplet networks. In *DATE*, 2022.
[5] Yuan Li et al. SPACX: Silicon photonics-based scalable chiplet accelerator for dnn inference. In *HPCA*, 2022.
[6] Asif Mirza et al. Opportunities for cross-layer design in high-performance computing systems with integrated silicon photonic networks. In *DATE*, 2020.
[7] Yuan Li et al. SPRINT: a high-performance, energy-efficient, and scalable chiplet-based accelerator with photonic interconnects for cnn inference. *IEEE TPDS*, 2021.
[8] Febin Sunny et al. Machine learning accelerators in 2.5 d chiplet platforms with silicon photonics. In *DATE*, 2023.
[9] Mikael Antelius et al. An apodized soi waveguide-to-fiber surface grating coupler for single lithography silicon photonics. *Optics express*, 2011.
[10] Xin Mu et al. Edge couplers in silicon photonic integrated circuits: A review. *Applied Sciences*, 2020.
[11] Wim Bogaerts et al. Silicon microring resonators. *Laser & Photonics Reviews*, 2012.
[12] Guoliang Li et al. Ring resonator modulators in silicon for interchip photonic links. *IEEE J. Sel. Top*, 2013.
[13] Mohammad Amin Mahdian et al. Thz multimode interference power divider based on groove gap waveguide configuration. *TNANO*, 2022.
[14] Zeqin Lu et al. Broadband silicon photonic directional coupler using asymmetric-waveguide based phase control. *Optics express*, 2015.
[15] Benjamin G Lee et al. Silicon photonic switch fabrics: Technology and architecture. *JLT*, 2018.
[16] Liangjun Lu et al. 16× 16 non-blocking silicon optical switch based on electro-optic mach-zehnder interferometers. *Optics express*, 2016.
[17] Mohammad Amin Mahdian et al. Pars: A power-aware and reliable control plane for silicon photonic switch fabrics. In *PSC*, 2023.
[18] Ebadollah Taheri et al. ReSiPI: A reconfigurable silicon-photonic 2.5 d chiplet network with pcms for energy-efficient interposer communication. In *ICCAD*, 2022.
[19] Aditya Narayan et al. PROWAVES: Proactive runtime wavelength selection for energy-efficient photonic nocs. *IEEE TCAD*, 2020.
[20] Noam Ophir et al. Silicon photonic microring links for high-bandwidth-density, low-power chip i/o. *IEEE Micro*, 2013.
[21] Cheng Li et al. Silicon photonic transceiver circuits with microring resonator bias-based wavelength stabilization in 65 nm cmos. *IEEE JSSC*, 2014.
[22] Febin Sunny et al. ARXON: A framework for approximate communication over photonic networks-on-chip. *IEEE TVLSI*, 2021.
[23] Javad Rahimi Vaskasi et al. High wall-plug efficiency and narrow linewidth iii-v-on-silicon c-band dfb laser diodes. *Optics Express*, 2022.
[24] Yaoyao Ye et al. A torus-based hierarchical optical-electronic network-on-chip for multiprocessor system-on-chip. *ACM JETC*, 2012.
[25] Robert Guirado et al. Dataflow-architecture co-design for 2.5 d dnn accelerators using wireless network-on-package. In *ASP-DAC*, 2021.