Waferscale Silicon Photonics Systems: A Cost-Benefit Analysis and Optimization

Robert Bao† *UIUC* Zongrui Cai† *UIUC* Shuangliang Chen *UIUC* Ajay Joshi *Lightmatter* Darius Bunandar *Lightmatter* Rakesh Kumar *UIUC*

2025 IEEE/ACM International Conference On Computer Aided Design (ICCAD) | 979-8-3315-1560-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCAD66269.2025.11240761

***Abstract*—Silicon photonics holds considerable promise for reducing long reach communication overheads in future computing systems. Similarly, waferscale integration promises dramatic improvements in performance and energy efficiency for scale out systems, but suffers from the long reach limitations of electrical interconnects. No prior work has looked at the performance benefits of silicon photonics over electrical interconnects to address the long reach challenges of waferscale integration, or at the overheads of silicon photonics for such systems across multiple implementations. In this work, we study a tile-based silicon photonics waferscale system for different implementations of waveguide networks and topologies, and across multiple applications and number of tiles. We find that the performance benefits of using silicon photonics instead of electrical interconnects at waferscale are highly application-dependent - benefits primarily come from reduced communication latency. The power and area overheads of implementation are high, especially for high connectivity topologies and when reconfigurability is considered. Some implementations are infeasible - the microring resonator maximum power limits are exceeded for these implementations. Custom waveguide networks address the problem and limit the overheads when supporting high connectivity topologies and reconfigurability. Overall, this is the first paper to analyze the performance benefits of silicon photonics vs electrical interconnects at waferscale and optimize the implementation overheads of waferscale silicon photonics systems.**

I. Introduction

As artificial intelligence, high performance computing, and datacenter applications proliferate, there is an increased appetite for disruptive technology-based approaches that have the potential to significantly improve performance and cost efficiency at scale.

One such approach is silicon photonics (Si-Ph). Silicon photonics combines the performance, scalability, and cost advantages of silicon-based circuits with semiconductor lasers to create photonic integrated circuits on silicon microchips. These circuits can be used either for communication or computation.

A large body of work has explored the use of silicon photonics for communication on chip-scale systems [1], [2]. Using light running along *waveguides* for communication between components of a chip is enticing since it can potentially outperform electrical links in terms of bandwidth, latency, and energy efficiency. However, several challenges prevented the eventual commercial adoption of on-chip silicon photonics links. At short distances, silicon photonics links do not demonstrate significantly better energy and latency characteristics compared to electrical links. In fact, their characteristics may be worse at link lengths shorter than 500μm due to the additional Optical-Electrical-Optical conversion required. In addition, the high design complexity of additional silicon photonics components further disincentivized commercial adoption in on-chip settings.

Another disruptive technology-based approach to significantly increase the performance and energy efficiency of some of today's most lucrative and challenging computing applications is waferscale integration [3], [4]. Waferscale integration dramatically reduces the cost of communication between components by integrating them on the same wafer. On-wafer interconnects can be one-two orders of magnitude more energy efficient than the off-package communication interconnects that they replace [3].

Unfortunately, the energy and latency cost of interconnection is high for a waferscale system for long reach (>10mm) interconnects. The existing on-wafer electrical interconnects demonstrate up to 10 mm of reach before the signal needs to be retimed/repeated. This significantly limits the choice of physical topology in an electrical waferscale system. For example, a 2D-Mesh topology is used in almost all existing systems [5], [6] in order to keep the electrical links short. Long links are possible in electrical systems, but that would require using retimers, incurring additional costs in terms of *energy-per-bit* and *latency*. In addition, although the *bandwidth density* of the electrical waferscale system is much higher than that of off-chip links, it is still a significant limitation for hardware that demands ultra-high connectivity, such as a waferscale network switch [7] or hardware that supports high connectivity topologies such as all-to-all.

Technology	Si-IF [3]	InFO-SoW [4]	CoWoS-L	UCIe-S [8]	UCIe-A [8]	Silicon Photonics [9]
I/O Pitch (μm)	2-10	80-150	3-10	100 - 130	25-55	2-10
Interconnect Wire Pitch (μm)	10	20	2-10	100 - 130	25-55	4
Maximum Sizes/Dies	Full Wafer	Full Wafer	28cm^2	Inter-chiplet	Inter-chiplet	Full Wafer
Inter-die Distance (mm)	0.5	5-30	5	$\leq$25	$\leq$2	800
BW Density (Gbps/mm/layer)	640-2000	1000-3200	5000	$\leq$224	$\leq$1317	179200
Energy/b (pJ/bit)	0.06-4	1.5-3	0.25	0.5	0.25	0.5-1 [1]
Latency (ns)	0.03-0.2	0.03-0.2	10-15	2	2	2

TABLE I: Technologies that can enable chiplet-based waferscale integration, vs silicon photonics).

Naturally, a question arises: *what if we replaced the electrical links on a waferscale system with optical links?* Optical interconnects provide long reach connectivity at high efficiency (Table I). We estimate that the energy efficiency of an optical link that connects two corners of a wafer would be at least an order of magnitude better than that of an electrical link [3], [9]. They also provide two orders of magnitude higher bandwidth density than electrical interconnects (Table I).

Several recent works [10], [11] have recognized the potential importance of waferscale silicon photonics, but have not performed the comparison to electrical waferscale architectures. We assert that the choice between implementing waferscale architectures using electrical or silicon photonic

†These authors contributed to the work equally.

interconnects is not obvious. The decades of research and development into electrical interconnects have made them competitive for short reach connectivity on a waferscale system vs silicon photonic technologies in terms of both latency and bandwidth. Even for long-reach connectivity on a waferscale system, electrical interconnects are competitive in terms of bandwidth due to high SerDes overhead associated with connecting chips to the optical links, albeit at the expense of latency (Section IV). So, application-level benefits may be possible for silicon photonics vs electrical interconnects only for applications that benefit significantly from reduced communication latency (Section IV).

Additionally, to the best of our knowledge, no prior work has attempted to understand the feasibility and overheads of waferscale silicon photonics when carefully considering different implementation details: choice of waveguide network (WGN), topology of interconnection, number of tiles, and reconfigurability. Overall overhead from factors like optical loss from waveguides or other optical components would be much higher for a waferscale system vs chip-scale systems due to longer physical waveguides and longer logical links.

We study a tiled silicon photonics waferscale system for different numbers of tiles, WGNs, and network topologies. We estimate the power overhead of implementing silicon photonics in each case. We pay close attention to the power that individual waveguides must carry for each implementation - silicon microring resonators (MRRs) experience degraded performance beyond a maximum power. We consider both generic (hardware configuration agnostic) and custom WGNs. We study both implementations – ones which have been designed to support a specific network topology and ones where reconfigurability is supported - i.e., circuit switching can be used to configure the system for different network topologies.

This paper makes the following contributions:

- We perform the first characterization of application-level performance benefits of a silicon photonics based waferscale systems versus a system based on conventional electrical interconnections and traditional server rack-based systems. We show that a 6.6x speedup can be achieved by switching to optical connections at waferscale for certain workloads. Some workloads see little benefit from silicon photonics vs electrical interconnects.
- We perform the first characterization of the power and area overheads of implementing silicon photonics at waferscale across different implementations. This characterization was performed with and without reconfigurability support. We show that the overhead of supporting reconfigurability is small.
- We show that it is not possible to build some topologies (hypercube, all-to-all), at least naively, at waferscale using silicon photonics because the maximum power limit for silicon MRRs is violated.
- We propose the use of custom waveguide networks which provide significant reduction in overall power and enable feasible implementation of even high connectivity topologies. They can also support a large number of tiles can be supported, especially in conjunction with grid routing.

II. Background

Silicon photonic communication systems generally consist of four primary elements: a laser to generate light, a mechanism to encode data onto that light, a medium to transmit the signal, and a receiver to convert it back into an electrical form. The laser source, which can be integrated on the chip or located externally, is often positioned off-chip to minimize thermal inefficiencies. To encode information onto the optical signal, microring modulators (MRMs) are commonly used [12]. These modulators work by tuning their resonant frequency—typically through thermal or electrical control—to match specific wavelengths. When resonance occurs, the modulator absorbs the passing laser light, representing a binary ”0”. When off-resonance, the light continues through the system, representing a ”1”.

Once modulated, the light travels through optical waveguides to its destination. These waveguides function similarly to electrical wires but with the potential for higher bandwidth at a similar pitch using dense wavelength division multiplexing (DWDM). DWDM enables the simultaneous transmission of multiple data streams within a single waveguide, with each stream occupying a different wavelength. Developments in multi-layer silicon photonics have allowed the integration of multiple silicon (Si) and silicon nitride (SiN) waveguide layers. While SiN waveguides have a much lower loss (0.1dB/cm compared to 0.5dB/cm), they are unable to support active devices like MRMs due to their lack of Pockels effect or carrier-plasma dispersion effect [13].

At the receiving end, MRRs serve as selective filters that extract specific wavelengths from the waveguide. These filtered signals are directed to drop ports connected to photodetectors, which transform the incoming light pulses into electrical signals that can be interpreted by standard electronic components.

The arrangement of the basic devices above describes a single-writer single-reader (SWSR) optical link. Alternatives to SWSR include single-writer multiple-reader (SWMR) and multiple-writer single-reader (MWSR) buses. SWMR buses introduce optical splitters which divide optical power along two paths; one which continues along the bus and one which diverts power to a receiver. This provides a simple way to support a broadcasting communication pattern while using just one waveguide as shown in Figure 1. However, SWMR links introduce untenable losses when broadcasting to a large number of tiles. MWSR buses require an arbitration mechanism, such as token passing, to determine which writer is allowed to use the bus, reducing performance in order to share 1 waveguide. Thus, we consider point-to-point SWSR links in this work.

Optical circuits are able to be switched, just like conventional circuits. The mechanisms for switching include MRRs, Mach-Zehnder Interferometers (MZIs), Micro-Electro-Mechanical Systems (MEMS), and Arrayed Waveguide Grating Routers (AWGR). MRRs can be used to steer light

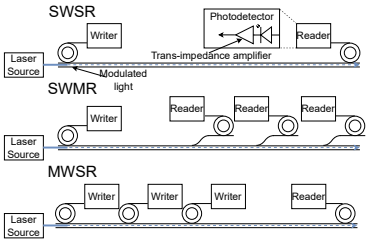

Fig. 1: SWSR, SWMR, and MWSR bus types.

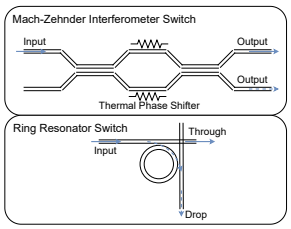

Fig. 2: MZI and MRR used as a 1x2 switch.

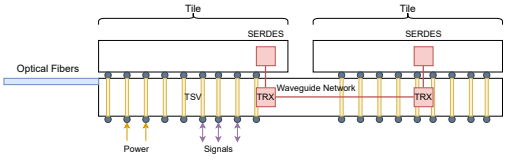

Fig. 3: Chip on wafer packaging of tiles on top of a photonic substrate to build a waferscale silicon photonics system.

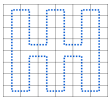

Fig. 4: A serpentine ring WGN.

onto a different waveguide using charge injection or thermal tuning as shown in Figure 2. Mach-Zehnder Interferometers use directional couplers and thermo-optical effects to change the phases of light and direct the incoming light to 1 of 2 output waveguides. While MRRs are physically smaller than MZIs, they can only switch 1 wavelength whereas MZIs switch all wavelengths. Similar to MRRs, AWGRs are used for wavelength routing, which limits the achievable bandwidth compared to MZI-based implementations. MEMS switches use an array of movable micro-mirrors to circuit switch all wavelengths of light, but they have worse latency characteristics than MZIs due to their mechanical nature. Thus, we focus on MZI-based switching throughout this work.

There are physical limitations of silicon photonics devices that dictate system design. As the optical power on the waveguide on which the MRM operates increases, nonlinear effects from two-photon absorption and free carrier absorption introduce unwanted resonant frequency shifts [14]–[16]. There is also a limit to the number of wavelengths that can be supported since MRRs are not perfect filters and waveguides can only optimally carry a certain range of wavelengths. In addition to these hard limitations, the optical devices incur optical signal loss due to scattering or absorption. For example, waveguides incur loss that is proportional to their length. Other devices like ring resonators, MZIs, passive waveguide crossings, and photodetectors have a fixed loss when light passes through them.

III. An Example Chiplet-based Waferscale Silicon Photonics System

An example chiplet-based waferscale silicon photonics system (Figure 3) consists of a multi-layer photonic substrate that lies underneath tiles which are Known-Good-Dies (KGDs) of any kind, bonded using chip-on-wafer packaging. These tiles communicate using transceiver banks (referred to as banks) and a physical network of waveguides. Physical WGNs in previous works [17], [18] have often been laid out as serpentine ring structures (Figure 4), which we also assume in this example. A ring WGN layout allows all tiles to be connected using one cluster of physical waveguides.

The WGN uses dedicated banks, each with multiple transmitters and receivers, to form pairs of communicating tiles. The tiles are bonded to the substrate using microbumps. Each tile has a set of serializer/deserializer (SerDes) modules which are each electrically connected to their respective banks

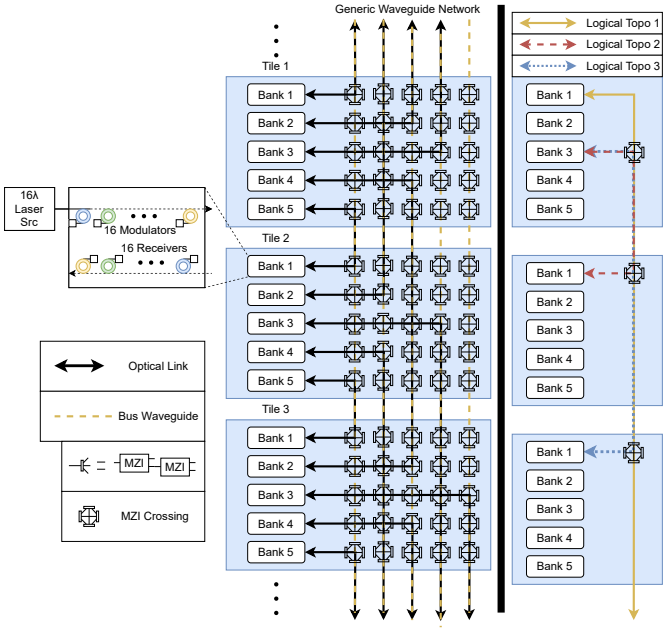

Fig. 5: Left: Banks supporting 16 wavelength communication and generic WGN. Right: Using MZIs for reconfigurability.

in the optical substrate. Each bank (Figure 5) contains a laser source, MRMs (one per wavelength) to facilitate the transmission of data, and MRR filters and optical receivers (one per wavelength) to receive incoming data. Each receiver contains a photodetector and amplifier circuitry. Since each resonator is tuned to a different wavelength, their operation does not interfere with others. The ring resonator modulators and filters must also be constantly tuned (using thermal or voltage tuning) to operate on the correct wavelength [2]. All active devices within the banks are located on the Si layer. The ring WGN is implemented in a single SiN layer for lower losses over long distances. Transitions between the layers use adiabatic tapers [19] as shown in Figure 6.

For a tile to send data to another tile (for example tile 1 to tile 2 shown in Figure 5), the corresponding SerDes module

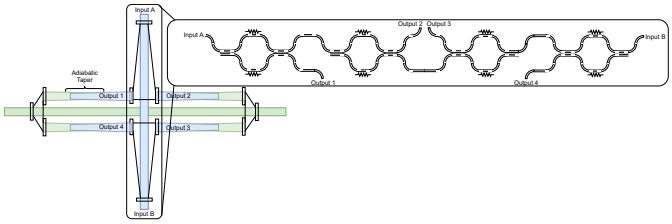

Fig. 6: MZI crossing

sends the electrical signal to its bank (bank 5 on tile 1), which then uses that signal to modulate the ring resonators. The data is encoded onto the different wavelengths of light and is sent along the WGN to its destination (bank 1 on tile 2). At the destination bank, the ring resonator filters pass the light to the receivers. The receivers send the signals to the destination SerDes and the data transfer is complete.

The WGN is composed of many point-to-point links between pairs of tiles. A logical network topology can be constructed between the tiles by identifying the links that need to be made between the tiles to support the topology and then routing the links along the ring WGN. To allow any bank to attach to any bus waveguide, MZI crossings are placed at each intersection between the waveguides coming out of the bank and the bus waveguides. An MZI crossing is shown in Figure 6. It requires 8 MZIs in order to allow light to couple between the layers in any direction. This allows the interconnect to be *generic*. We define a generic WGN as one which is capable of supporting *up to* a desired connectivity *and* is able to connect any bank to any bus waveguide. This allows one optical substrate to support a variety of use cases (logical topologies) through programming the onboard MZIs.

IV. Quantifying Application-Level Benefits

First, we explore if the silicon photonic waferscale system provides application-level performance benefits over electrical waferscale and conventional systems.

Although silicon photonics provides a theoretically higher bandwidth density in terms of TB/s/mm, the achievable bandwidth out of a given area may actually be lower than what is possible electrically, depending on the size of the tile (Figure 7). Take, for example, a 24mmx36mm reticle-limit sized chip. The perimeter of the chip would support $120mm \times 1.35TB/s/mm = 162TB/s$ using UCIe 1.0 electrical interconnects [8]. However, a silicon photonic interconnect requires area-expensive SerDes modules to pump high data rates through its waveguides. Assuming SerDes to provide $116GB/s/mm^2$, the tile can support $100TB/s$ when filling the *entire* tile with SerDes, a great deal lower than what is theoretically possible given DWDM bandwidth densities ($120mm \times 16\lambda \times \frac{112Gbps}{\lambda} \times \frac{1}{4um} = 6720TB/s$). Clearly, the SerDes is a severe bottleneck to the actual bandwidth achievable by silicon photonics. To add to the concern, the SerDes must compete for area within the chip itself.

Also, many applications have inherent locality in their communications. Examples include applications which have nearest-neighbor/grid communication patterns [20], those that employ the common ring-allreduce algorithm, and any embarrassingly parallel applications. These classes of problems would likely benefit very little from switching to a silicon photonic interconnect over electrical interconnects. High bandwidth mesh interconnects like those demonstrated in Cerebras' WSE [5] and Tesla's Dojo [6] would likely be sufficient.

We study three use cases: HPC applications, LLM inference, and Key-Value store (KVS). For each application, we compare 3 types of systems (Table II): a conventional system with

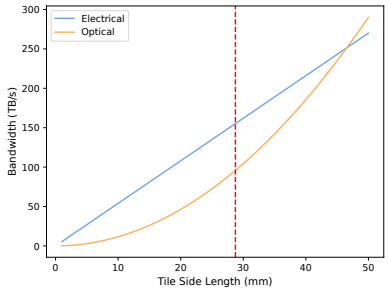

Fig. 7: Bandwidth achievable out of a single square tile when scaling tile size for electrical and silicon photonic interconnects.

compute units in separate server racks, an electrical waferscale system with compute units connected using a 2D mesh topology on wafer, and an optical waferscale system which connects compute units in an all-to-all (A2A) topology. On the waferscale systems, compute units sit on "tiles". We assume a 210mm x 210mm system cut from a 300mm diameter wafer. The interconnect latency for the conventional system is assumed to be 2us, while the electrical waferscale system has a 30ns latency per hop between adjacent tiles and the optical waferscale system has a constant 20ns latency between connected tiles. For the optical A2A system, we provision each tile to have N-1 optical banks (N is the number of tiles) to limit the area overhead of SerDes, which consume 45mm^2 when N=48. We assume the electrical mesh to support 9 TB/s between any two connected tiles after Tesla's Dojo system. The conventional system has a bandwidth of 900GB/s, similar to NVLink.

Common Parameters	
Number of Compute Tiles	32
Number of Cores per Tile	8
Number of Memory Tiles	16
Memory Capacity per Tile	216 GB
Memory Bandwidth per Tile	7200 GB/s
Memory Access Delay	300 ns
Core	3 Ghz, 2-issue wide
L1	Per-Core, 32 KB, 8-Way
L2	Per-Tile, 128 KB, 4-Way
Optical A2A Parameters	
Bandwidth Between Tiles (A2A)	112 GB/s
Latency Between Tiles	20 ns
Electrical Mesh Parameters	
Bandwidth Between Tiles	9000 GB/s
Latency Between Tiles	30 ns
Conventional Cluster Parameters	
Bandwidth Between CPU/GPU	900 GB/s
Latency Between CPU/GPU	2 μs

TABLE II: Architecture Parameters

Kernel	
FFT	-p256 -m16
Radix	-p256 -n1048576
LU	-p256 -n512
Application	
Ocean	-p256 -n258
Radiosity	-p 256 -ae 5000 -bf 0.1 -en 0.05 -room -batch
Raytrace	-p256 -m64 inputs/car.env
Volrend	256 inputs/head 8

TABLE III: SPLASH-3 Benchmark Parameters

For LLM inference, we simulate the three systems as GPU clusters using a modified version of LLM-Analysis [21]. Each compute unit is a GPU, modeled after an Nvidia H200. The systems each have 24 such GPUs running in tensor parallelism. Each GPU is connected to 6 HBM3e stacks providing 216GB of capacity and 7.2TB/s of bandwidth. Llama 3.1 405b using full context length and a 50/50 split of input/output tokens is used as the workload. The results are shown in Figure 9.

Going from conventional interconnects to waferscale integration dramatically reduces the latency of each inference request by reducing the time spent on the all-reduce communications needed for tensor parallelism, reducing end-to-end latency by 6.2x. Moving to an optical substrate provides limited benefit (1.09x decreased latency over electrical), as the

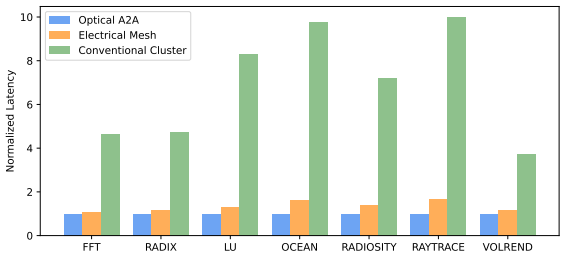

Fig. 8: Normalized runtime for each SPLASH-3 benchmark on all systems.

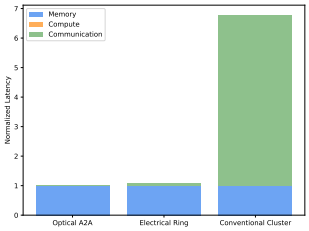

Fig. 9: Normalized latency of LLM inference.

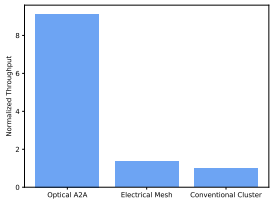

Fig. 10: Normalized throughput of KVS.

communication pattern of the all-reduce is a ring, demanding only adjacent connectivity which electrical links can provide. The A2A topology of the optical system allows all-reduce to be performed as a reduce-scatter then all-gather, reducing the number of communication hops from 24 to 2. However, the portion of runtime taken up by communication is already a small fraction (10%) of total runtime once the systems are moved on-wafer.

To simulate an HPC system, we consider a NUMA system with 32 compute tiles and 16 memory tiles. Each memory tile has 6 HBM3e chips. Each chip provides 1.2TB/s of memory bandwidth and 36GB of capacity. Each memory tile is connected to 2 compute tiles. Each compute tile is the home node to the memory of 3 HBM3e chips. For compute units to access remote memory, the request and data must take an indirect route with an additional hop. We use SST [22] to model the systems. We use SPLASH-3 [23] as our benchmark suite. Figure 8 shows the performance results. The waferscale systems clearly outperform the conventional system, showing an average of 5x speedup. Moving from the electrical interconnect to the optical interconnect brings another 1.34x speedup on average, showing that high connectivity topologies are worth pursuing using silicon photonics to achieve the best performance.

For benchmarks such as *ocean* and *raytrace*, which have a high fraction of read/write accesses to shared memory locations, the silicon photonic systems see a larger benefit. In addition to accesses to shared data, *radiosity* and *raytrace* have nonuniform communication patterns which benefit from the low hop counts in the silicon photonics systems. Benchmarks with a high computation to communication ratio such as *volrend* see smaller performance benefits. While *FFT* has all to all communication patterns, the traffic generated scales only logarithmically with the number of processors (traffic reaches a ceiling as number of processors scales up) and, as such, does not benefit from better overall latency characteristics of optical communication.

We evaluate a KVS workload on the systems, each now with 64 CPUs. Each CPU is connected to its own HBM3e stack and NIC. The conventional system uses a Concurrent Read Exclusive Write (CREW) key-to-core mapping strategy to localize writes, decreasing reliance on inter-node communication. The waferscale systems use a Concurrent Read Concurrent Write (CRCW) to minimize load imbalance, resulting in better performance than using CREW. The key distribution is assumed to be Zipfian with a 0.99 skewness factor. Strong consistency is assumed for the per-key consistency model. Since CRCW introduces high network traffic to achieve load balancing, the high connectivity optical network allows the CPUs to communicate without requiring expensive hops across a mesh. The performance benefit going from a conventional system to an electrical waferscale system is 1.37x as shown in Figure 10. Moving to optical waferscale brings another 6.67x. For use cases which generate high network traffic across many components, like load balancing across many CPUs, silicon photonics is extremely advantageous.

Overall, the results show that the performance benefits of using silicon photonics instead of electrical interconnects at waferscale are highly application-dependent - benefits primarily come from reduced communication latency.

V. Understanding Implementation Overheads

A. Power and Area Estimation

To calculate the laser power ($P_{laser}[\mathrm{mW}]$) needed for each link, we start with the photodetector sensitivity ($S[\mathrm{dBm}]$), then add the losses from each component: $L_c[\mathrm{dB}]$ for bank component losses (sum of ring resonator modulator, filter, and photodetector losses, which are the same for all links), $L_l[\mathrm{dB}]$ for waveguide length loss, $L_x[\mathrm{dB}]$ for waveguide under/overpass losses, and $L_{mzi}[\mathrm{dB}]$ for MZI crossing losses. The total optical power for a single photonic link is the sum of laser power, resonator tuning power, modulator power, MZI power, and photodetector/receiver power.

$$P_{laser} = 10^{\frac{S+L_c+L_l+L_x+L_{mzi}}{10}}$$

We implement a 2D routing algorithm (Crossing-aware channel routing), similar to [24], to map logical topologies onto the serpentine ring WGN. The algorithm takes as input a list of tuples. Each tuple represents a link between two tiles. It then creates the link by routing a path between two banks (one on each tile) on the WGN, keeping track of the length of the link, number of waveguide crossings, and number of MZI crossings. We use this algorithm as the basis to calculate the total optical power and the worst case power for a waveguide. The *worst case power* for a given WGN is the maximum laser power for a link across all links in the system.

The system faces an issue with microring modulators [25]. As the optical power on the waveguide on which the modulator

operates increases, nonlinear effects from two-photon absorption and free carrier absorption introduce unwanted resonant frequency shifts [14]–[16]. Due to this, today's systems limit laser power to the single digit mW per waveguide (Lightmatter's Passage prototype [9] has 5mW power at each TX, [15] quotes a maximum power of 1.5mW). We choose an aggressive limit of 35mW, above which the Q factor of the MRMs degrades significantly as shown by [16].

The area overhead of the active device layer in the optical communication system is estimated as the sum of MRM, MRR, MZI, and photodetector area. The area consumed by the waveguides is calculated as pitch$\times$length. Since the WGN is located in the SiN layer(s) below the active device layer (Si), we report the overall area overhead of optical communication to be that of the layer which takes the most area.

Table IV shows the parameters used for our estimations.

Si Waveguide Loss [2]	0.5dB/cm
SiN Waveguide Loss [26]	0.1dB/cm
SiN Waveguide Under/Overpass Loss [27]	0.0034dB
Vertical Coupling Loss [19]	0.1dB
Ring Resonator Tuning Power [2]	0.1mW
Ring Resonator Modulator Energy [2]	35fJ/bit
Ring Resonator Modulator Insertion Loss [18]	0.5dB
Ring Resonator Through Loss [28]	0.05dB
Ring Resonator Filter Drop Loss [2]	1.5dB
MZI Power [9]	1mW
MZI Loss [9]	0.08dB
Photodetector Energy [29]	0.17pJ/bit
Photodetector Loss [18]	0.1dB
Photodetector Sensitivity [30]	-17.4dBm
Optical Link Data Rate per Wavelength [9]	112Gbps
Bank/SerDes Area [9]	3.44mm^2
Waveguide Pitch [9]	4µm
Wavelengths per Bank [9]	16

TABLE IV: Optical device parameters used in this work.

B. Overhead Analysis

We estimate the total power (Figure 11), worst case link power (Figure 12), and area overhead (Figure 13) of implementing several topologies on the generic ring. We find that for low connectivity topologies like mesh and torus, worst case power is under the 35mW threshold we set. The total power for these cases is dominated by the active components (MZIs, modulators, PDs) as these are needed to support an optical link regardless of the losses which may appear along that link. The sparse connectivity means that there are few crossings and therefore the loss is low along all waveguides.

The power from active components increases linearly with the number of connections that must be made, but the contribution of laser power increases super-linearly as seen in Figure 11. Although the total power is high (around 2000W to support an A2A topology), this would be an insignificant portion of the system's total power. If we assume each tile consumes 700W (H200 TDP), the interconnect would be only 10.5% of total system power.

The major concern here is worst case power. For higher connectivity topologies (e.g., A2A and hypercube), a larger number of MZI crossings must be introduced into the WGN to enable routing over the long distances, raising the laser power for the links. These longer links also increase the loss due to distance. However, this is not a significant factor due to the low loss of SiN waveguides. Both for hypercube and A2A, at least one link breaches the 35mW limit.

Area is not an issue here with even A2A requiring less than $44100\mathrm{mm}^2$. Since the WGN is implemented on a separate layer from the active devices, we only show the larger of the two (which happens to be the area of active devices, the SiN WGN requires $640\mathrm{mm}^2$ at most).

VI. Custom Waveguide Networks

Seeing as the generic WGN is infeasible for high connectivity topologies, we implement a topology-customized ring WGN. A *custom* WGN replaces MZI crossings, which previously allowed connections from any bank to any bus waveguide, with static waveguide bends and couplers attached to specific bus waveguides to achieve a given connectivity between tiles. Figure 15 shows the ring WGN when customized to support either a static ring topology or an A2A topology. Switching to a custom WGN reduces the crossing loss and area overhead incurred by MZI crossings at the cost of reduced flexibility. The desired topology must be known before fabricating the custom WGN and cannot be changed afterward.

The benefits of customization are shown in Figures 16, 17, and 18. The overall power savings from switching to a custom WGN are from 45.8% for mesh up to 93.4% for A2A. The removal of MZI tuning power accounts for 99.7% and 6% of the total power reduction for mesh and A2A respectively. The reduction of laser power due to the lack of MZIs on the WGN accounts for 0.3% and 94% of the total power reduction for mesh and A2A respectively. This is because MZI tuning power being much more significant than laser power for the mesh and vice versa for A2A.

The benefit of customization in terms of worst case power is minimal for low connectivity topologies since such topologies have low laser power and loss already. However, custom WGNs provide significant worst case power benefits for higher connectivity topologies. For an A2A topology, the custom WGN replaces 45dB of MZI crossing loss with 0.4dB of waveguide under/overpass loss, while other components of loss (MRM, MRR, length, coupler) remain constant. This reduction in loss along the optical links transforms hypercube and A2A from infeasible to feasible topologies.

Additionally, the area overhead of MZI crossings is eliminated, reducing overall area by up to 84.5% for A2A.

VII. Changing the Waveguide Network Layout

Figure 17 shows that a major contributor to loss is the waveguide length. To address this, we also study a grid WGN layout. This network uses X-Y routing to take shorter paths between connected tiles, reducing loss over the length of the waveguides. Figure 22 shows the connectivity to one example tile in a custom grid WGN.

Figures 19, 20, and 21 show the impact of a grid WGN layout. The shorter waveguide lengths in the grid WGN reduce length loss (by up to 40%) and thus laser power. However, most of the power is incurred by active optical components (99.7% for A2A), rather than laser power (0.3% for A2A) in the custom WGNs as shown in Figures 16 and 19, so the

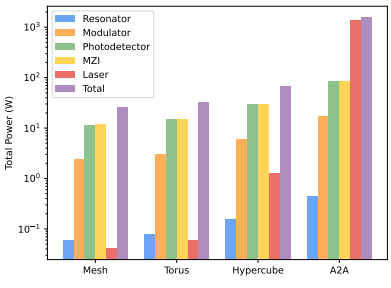

Fig. 11: Breakdown of total power for various logical topologies on a generic ring WGN.

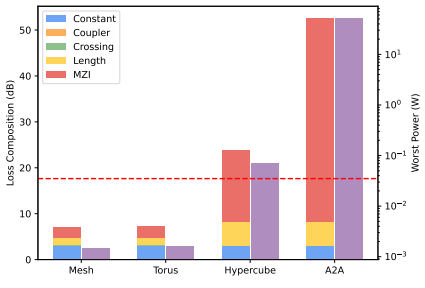

Fig. 12: Loss breakdown (left axis) and power on the worst case (right axis) link on a generic ring WGN. Dotted red line indicates the 35mW worst case power limit. Worst case powers shown in purple.

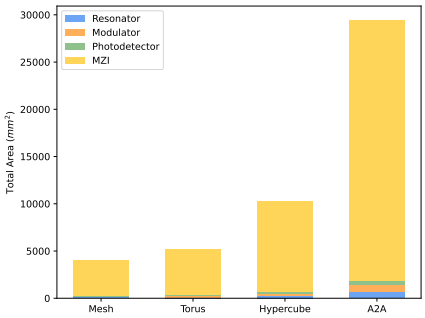

Fig. 13: Area overhead for different topologies on a generic ring WGN.

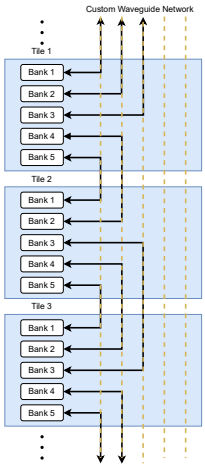

Fig. 14: Ring topology on custom ring WGN.

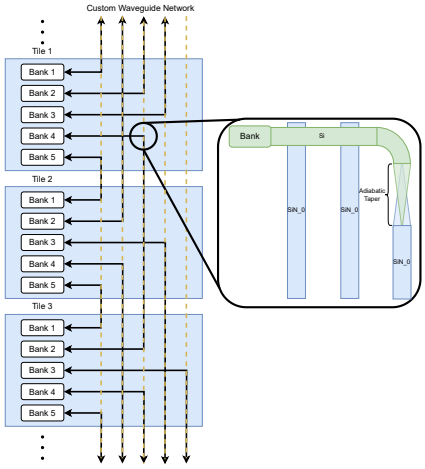

Fig. 15: A2A topology on custom ring WGN. Custom ring WGN replaces MZI crossings with waveguide bends and vertical couplers.

difference in power consumption between the grid and ring WGNs is only 0.1%. Worst case power is not an issue when using either a custom ring or grid WGN.

However, the benefit of a grid becomes clear (Figure 23) as the number of tiles in the system is increased (by making each tile smaller). While the length of the ring increases with the number of tiles (since the serpentine must pass through every tile), the X-Y routing of the grid ensures that waveguide lengths remain unchanged. As a result, the gap between the worst case powers on a grid and ring grows as the number of tiles increases. In fact, while ring-based systems become infeasible at a larger number of tiles, the grid-based systems remain feasible.

VIII. Reconfigurability

Several previous works identify network reconfigurability (the ability to redistribute bandwidth by switching topologies) as beneficial for application-level performance [31]. We explore supporting this feature on the optical waferscale interconnect. On a ring WGN, one can switch to a different logical network topology with lower connectivity seamlessly (to save power, for example), by turning off a subset of banks (and possibly SerDes modules). For example, banks 2, 3, and 4 in Figure 5 can be turned off in each tile to form a ring topology. However, this only works when switching between a topology and a subset of that topology (like A2A and mesh). Banks can be turned off in a similar manner on a grid WGN to achieve the same effect with the same topology limitations. When we want to preserve aggregate bandwidth (keep all banks active), we must use optical circuit switching which is implemented using MZI crossings.

On the right side of Figure 5, we show three logical topologies which contend for the same bus waveguide on a ring WGN. MZI crossings are added along the bus waveguide to guide light to the right destinations depending on the topology. To implement reconfigurability on a grid WGN, all necessary links to support the three architectures at full bandwidth are routed. MZIs are then placed near each bank to switch its connection to the waveguide necessary to support the desired topology.

We explore the effect on power, worst case power, and area when supporting reconfigurability between an A2A, hypercube, and torus topologies (while maintaining approximately equal aggregate bandwidth). The "1xA2A" topology supports 112 GB/s per link, while the "3xHypercube" supports 336 GB/s per link, and the "6xTorus" supports 672 GB/s per link. The maximum aggregate bandwidth (number of links $\times$ BW per link) for the hypercube and torus is 64.5 TB/s and 61.8 TB/s for A2A. Results for the ring WGN are shown in the rightmost bars of Figures 16, 17, and 18. The overheads of adding this feature to a grid WGN are shown in the rightmost bars of Figures 19, 20, and 21.

Almost all the additional power comes from the extra banks and their components which are needed to accommodate the slightly higher aggregate bandwidth, lower connectivity topologies. The additional MZIs needed to support reconfigurability add MZI crossing loss, slightly increasing worst case powers as seen in Figures 17 and 20, but remain under the 35mW limit. Overall, the overhead of supporting reconfigurability is minimal, suggesting that this may be an avenue for silicon photonics to achieve even better application-level

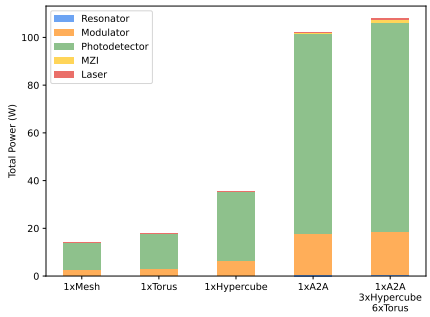

Fig. 16: Breakdown of total power for various logical topologies on a custom ring WGN.

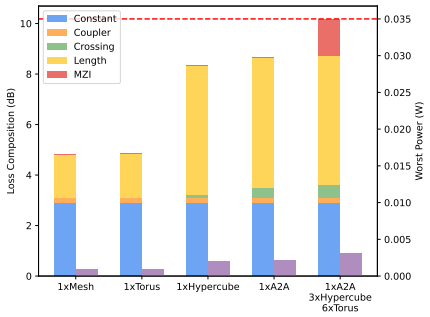

Fig. 17: Loss breakdown (left axis) and power on the worst case (right axis) for various logical topologies on a custom ring WGN. Worst case powers shown in purple.

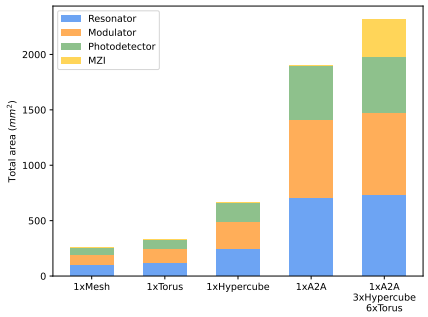

Fig. 18: Area overhead for different topologies on a custom ring WGN.

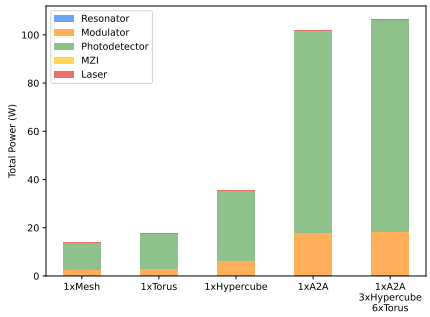

Fig. 19: Breakdown of total power for various logical topologies on a custom grid WGN.

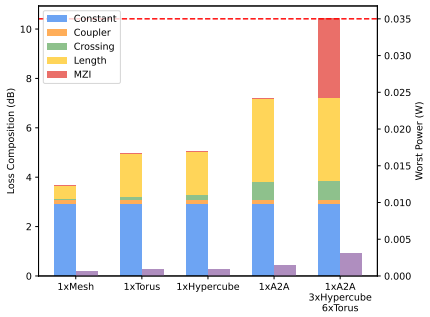

Fig. 20: Loss breakdown (left axis) and power on the worst case (right axis) for various logical topologies on a custom grid WGN. Worst case powers shown in purple.

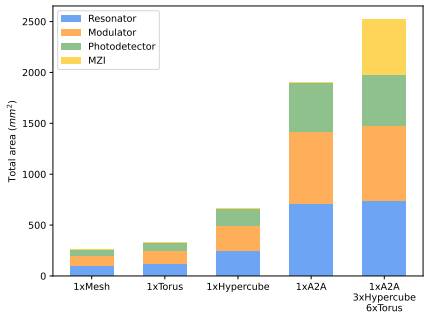

Fig. 21: Area overhead for different topologies on a custom grid WGN.

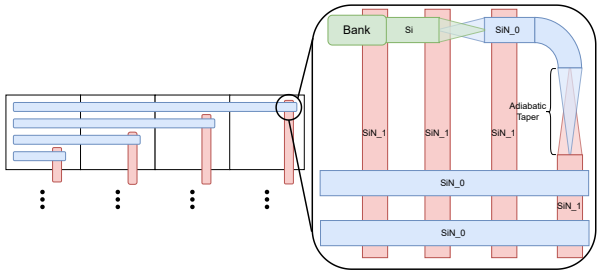

Fig. 22: Connectivity from one tile in a custom grid WGN. Inset: multi-layer silicon photonic system with inter-layer couplers (adiabatic tapers).

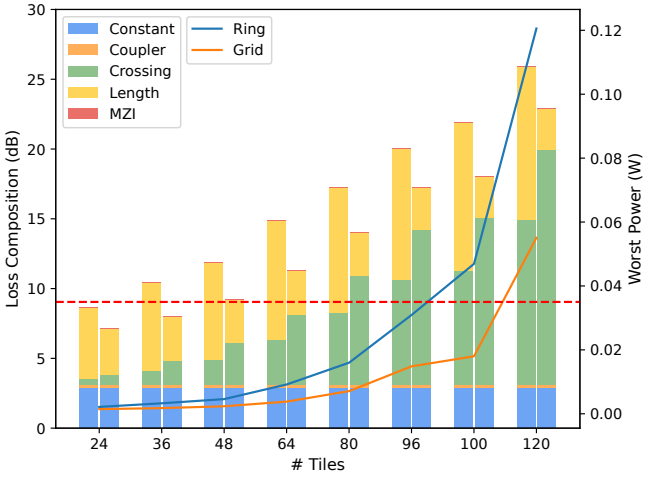

Fig. 23: Worst case power scaling when implementing A2A topology on an increasing number of tiles. Bars correspond to left axis, lines correspond to right axis. Left bar in each cluster corresponds to ring WGN, right bar corresponds to grid WGN.

performance.

IX. Conclusion

No prior work has quantified the performance benefits of employing silicon photonics instead of electrical interconnects with waferscale integration, or the the overheads of implementing waferscale silicon photonics systems. We studied a tile-based silicon photonics waferscale system for different implementations of waveguide networks and topologies, and across multiple applications and numbers of tiles. We found that the benefits from using silicon photonics at waferscale are highly application-dependent, with benefits primarily derived from reduced communication latency. The power and area overheads are high, with worst case power making some high connectivity topologies infeasible. Our analysis showed that custom topology-specific WGNs can mitigate these issues and increase scalability, especially when used in conjunction with grid routing. Reconfigurability can be supported at small overheads for added flexibility. Overall, this is the first work to carefully study the performance benefits of silicon photonics over electrical interconnects at waferscale and characterize the overheads of implementing such systems.

References

[1] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary, "Firefly: illuminating future network-on-chip with nanophotonics," in *Proceedings of the 36th annual international symposium on Computer*

architecture, ser. ISCA '09. New York, NY, USA: Association for Computing Machinery, Jun. 2009, pp. 429–440. [Online]. Available: https://dl.acm.org/doi/10.1145/1555754.1555808

[2] P. Koka, M. O. McCracken, H. Schwetman, X. Zheng, R. Ho, and A. V. Krishnamoorthy, "Silicon-photonic network architectures for scalable, power-efficient multi-chip systems," in *Proceedings of the 37th annual international symposium on Computer architecture*, ser. ISCA '10. New York, NY, USA: Association for Computing Machinery, Jun. 2010, pp. 117–128. [Online]. Available: https://dl.acm.org/doi/10.1145/1815961.1815977

[3] S. S. Iyer, S. Jangam, and B. Vaisband, "Silicon interconnect fabric: A versatile heterogeneous integration platform for ai systems," *IBM Journal of Research and Development*, vol. 63, no. 6, pp. 5:1–5:16, 2019.

[4] S.-R. Chun, T.-H. Kuo, H.-Y. Tsai, C.-S. Liu, C.-T. Wang, J.-S. Hsieh, T.-S. Lin, T. Ku, and D. Yu, "Info_sow (system-on-wafer) for high performance computing," in *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, 2020, pp. 1–6.

[5] G. Lauterbach, "The path to successful wafer-scale integration: The cerebras story," *IEEE Micro*, vol. 41, no. 6, pp. 52–57, 2021.

[6] E. Talpes, D. Williams, and D. D. Sarma, "Dojo: The microarchitecture of tesla's exa-scale computer," in *2022 IEEE Hot Chips 34 Symposium (HCS)*, 2022, pp. 1–28.

[7] S. Chen, S. Pal, and R. Kumar, "Waferscale network switches," in *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, Jun. 2024, p. 215–229. [Online]. Available: https://ieeexplore.ieee.org/document/10609578

[8] D. Das Sharma, G. Pasdast, Z. Qian, and K. Aygun, "Universal chiplet interconnect express (ucie): An open industry standard for innovations with chiplets at package level," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 12, no. 9, p. 1423–1431, Sep. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9893865/

[9] N. C. Harris, D. Bunandar, A. Joshi, A. Basumallik, and R. Turner, "Passage: A wafer-scale programmable photonic communication substrate," in *2022 IEEE Hot Chips 34 Symposium (HCS)*, 2022, pp. 1–26.

[10] Y. Safari, R. Mohammadrezaee, D. Al Saleh, and B. Vaisband, "Hybrid interconnect infrastructure for inter-chiplet communication in wafer-scale systems," in *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*, May 2024, p. 2229–2236. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10565187

[11] S. Zhang, Z. Zhang, M. Naderan-Tahan, H. SeyyedAghaei, X. Wang, H. Li, S. Qin, D. Colle, G. Torfs, M. Pickavet *et al.*, "Photonic network-on-wafer for multichiplet gpus," *IEEE Micro*, vol. 43, no. 2, pp. 86–95, 2023.

[12] G. Li, A. V. Krishnamoorthy, I. Shubin, J. Yao, Y. Luo, H. Thacker, X. Zheng, K. Raj, and J. E. Cunningham, "Ring resonator modulators in silicon for interchip photonic links," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, no. 6, p. 95–113, Nov. 2013. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6615926

[13] C. Xiang, W. Jin, and J. E. Bowers, "Silicon nitride passive and active photonic integrated circuits: trends and prospects," *Photonics Research*, vol. 10, no. 6, p. A82, Jun. 2022. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=prj-10-6-A82

[14] M. De Cea, A. H. Atabaki, and R. J. Ram, "Power handling of silicon microring modulators," *Optics Express*, vol. 27, no. 17, p. 24274, Aug. 2019. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=oe-27-17-24274

[15] K. Preston, N. Sherwood-Droz, J. S. Levy, and M. Lipson, "Performance guidelines for wdm interconnects based on silicon microring resonators," in *CLEO: 2011 - Laser Science to Photonic Applications*, May 2011, p. 1–2. [Online]. Available: https://ieeexplore.ieee.org/document/5950503/

[16] M. Novarese, S. R. Garcia, S. Cucco, D. Adams, J. Bovington, and M. Gioannini, "Study of nonlinear effects and self-heating in a silicon microring resonator including a shockley-read-hall model for carrier recombination," *Optics Express*, vol. 30, no. 9, p. 14341, Apr. 2022. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=oe-30-9-14341

[17] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn, "Corona: System Implications of Emerging Nanophotonic Technology," *SIGARCH Comput. Archit. News*, vol. 36, no. 3, pp. 153–164, Jun. 2008. [Online]. Available: https://dl.acm.org/doi/10.1145/1394608.1382135

[18] Y. Demir, Y. Pan, S. Song, N. Hardavellas, J. Kim, and G. Memik, "Galaxy: a high-performance energy-efficient multi-chip architecture using photonic interconnects," in *Proceedings of the 28th ACM international conference on Supercomputing*, ser. ICS '14. New York, NY, USA: Association for Computing Machinery, Jun. 2014, pp. 303–312. [Online]. Available: https://dl.acm.org/doi/10.1145/2597652.2597664

[19] W. D. Sacher, Y. Huang, G.-Q. Lo, and J. K. S. Poon, "Multilayer Silicon Nitride-on-Silicon Integrated Photonic Platforms and Devices," *Journal of Lightwave Technology*, vol. 33, no. 4, pp. 901–910, Feb. 2015. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7014292

[20] P. G. Raponi, F. Petrini, R. Walkup, and F. Checconi, "Characterization of the communication patterns of scientific applications on blue gene/p," in *2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum*, May 2011, p. 1017–1024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6008951

[21] C. Li, "Llm-analysis: Latency and memory analysis of transformer models for training and inference," https://github.com/cli99/llm-analysis, 2023.

[22] A. F. Rodrigues, K. S. Hemmert, B. W. Barrett, C. Kersey, R. Oldfield, M. Weston, R. Risen, J. Cook, P. Rosenfeld, E. Cooper-Balis, and B. Jacob, "The structural simulation toolkit," *SIGMETRICS Perform. Eval. Rev.*, vol. 38, no. 4, p. 37–42, Mar. 2011. [Online]. Available: https://dl.acm.org/doi/10.1145/1964218.1964225

[23] C. Sakalis, C. Leonardsson, S. Kaxiras, and A. Ros, "Splash-3: A properly synchronized benchmark suite for contemporary research," in *2016 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, Apr. 2016, p. 101–111. [Online]. Available: https://ieeexplore.ieee.org/document/7482078

[24] C. Condrat, P. Kalla, and S. Blair, "Crossing-aware channel routing for photonic waveguides," in *2013 IEEE 56th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug. 2013, p. 649–652. [Online]. Available: https://ieeexplore.ieee.org/document/6674732/

[25] V. S. P. Karempudi, J. Bashir, and I. G. Thakkar, "An analysis of various design pathways towards multi-terabit photonic on-interposer interconnects," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 20, no. 2, p. 1–34, Apr. 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3635031

[26] D. J. Blumenthal, R. Heideman, D. Geuzebroek, A. Leinse, and C. Roeloffzen, "Silicon nitride in silicon photonics," *Proceedings of the IEEE*, vol. 106, no. 12, p. 2209–2231, Dec. 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8472140

[27] W. D. Sacher, J. C. Mikkelsen, Y. Huang, J. C. C. Mak, Z. Yong, X. Luo, Y. Li, P. Dumais, J. Jiang, D. Goodwill, E. Bernier, P. G.-Q. Lo, and J. K. S. Poon, "Monolithically integrated multilayer silicon nitride-on-silicon waveguide platforms for 3-d photonic circuits and devices," *Proceedings of the IEEE*, vol. 106, no. 12, p. 2232–2245, Dec. 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8452165

[28] C. Li, F. Jiang, S. Chen, J. Zhang, Y. Liu, Y. Fu, and J. Xu, "Accelerating Cache Coherence in Manycore Processor through Silicon Photonic Chiplet," in *2022 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, Oct. 2022, pp. 1–9, iSSN: 1558-2434. [Online]. Available: https://ieeexplore.ieee.org/document/10069774/

[29] Y. Xiang, H. Cao, C. Liu, J. Guo, and D. Dai, "High-speed waveguide Ge/Si avalanche photodiode with a gain-bandwidth product of 615 GHz," *Optica*, vol. 9, no. 7, p. 762, Jul. 2022. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=optica-9-7-762

[30] "Mics lab — 3d-integrated high-sensitivity optical receiver." [Online]. Available: https://www.mics.caltech.edu/3d-integrated-high-sensitivity-optical-receiver/

[31] N. P. Jouppi, G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, S. Subramanian, A. Swing, B. Towles, C. Young, X. Zhou, Z. Zhou, and D. Patterson, "Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings," no. arXiv:2304.01433, Apr. 2023, arXiv:2304.01433. [Online]. Available: http://arxiv.org/abs/2304.01433