This article has been accepted for publication in IEEE Micro. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MM.2023.3237927



# Photonic Network-on-Wafer for Multi-Chiplet GPUs

Shiqing Zhang[†], Ziyue Zhang[†‡], Mahmood Naderan-Tahan[†], Hossein SeyyedAghaei[†], Xin Wang[†‡], He Li[†‡], Senbiao Qin[†‡], Didier Colle[‡†], Guy Torfs[†‡], Mario Pickavet[†‡], Johan Bauwelinck[†‡], Günther Roelkens[†‡], Lieven Eeckhout[†]
[†]Ghent University, Belgium [‡]Imec, Belgium

**Abstract**—This paper introduces the *Photonic Network-on-Wafer (NoW) GPU architecture* to overcome fundamental limitations in electrical interconnect scaling by implementing the inter-GPU network in a wafer-scale optical interposer. We argue that the photonic-NoW GPU is a scalable architecture, delivering significant performance benefits in a power-efficient manner.

## 1 Introduction

In the last decade, advancements in Graphics Processing Units (GPUs) have been propelling major developments in artificial intelligence (AI), high-performance computing (HPC), and data analytics. Continuing this trend in any of these domains requires the ability to continuously scale GPU performance. Until recently, GPU performance has been scaled by increasing the number of Streaming Multiprocessors (SMs) across generations. This was made possible by leveraging Moore's Law and using the maximum possible transistor count in the most advanced chip technology node. Unfortunately, transistor scaling is slowing down and is likely to eventually stop. In addition, manufacturing issues further constrain the maximum die size as modern-day GPUs are approaching the reticle limit (around 800 $mm^2$). Moreover, very large dies lead to yield issues, rendering the cost of large monolithic GPUs to undesirable levels.

The solution to GPU performance scaling is to connect multiple physical GPUs together while providing the abstraction of a single logical GPU to software. One approach is to connect multiple GPUs on a Printed Circuit Board (PCB). Scaling GPU workloads on these multi-GPU systems is hard because of the limited inter-GPU bandwidth offered. On-package interconnects, e.g., through interposer technology, provide higher bandwidth and lower latency than off-package interconnects, providing a promising direction to scale GPU performance to a handful GPUs [1]. Wafer-scale integration goes one step further by bonding pre-manufactured dies on a silicon wafer, providing a pathway towards a wafer-scale GPU with tens of GPUs [2]. Unfortunately, providing high bandwidth density at low power consumption over long distances is fundamentally challenging using electrical interconnects, constraining GPU scaling using electrical interposer technology.

In this paper, we propose the photonic Network-on-Wafer (NoW) GPU architecture in which pre-manufactured and pre-tested GPU dies and memory chips are mounted on a wafer-level interposer that connects the GPU chips through a photonic network layer, while connecting each GPU die with its local memory stack electrically, as illustrated in Figure 1. The key asset of the photonic-NoW GPU architecture is the ability to achieve high bandwidth density at low power over relatively long, wafer-scale distances (up to tens of centimeters). The goal of this paper is to present the vision of the photonic-NoW GPU architecture, and argue for its potential and feasibility based on a preliminary quantitative and qualitative evaluation. More specifically, our preliminary simulation results indicate that GPU applications benefit from increased inter-chip bandwidth and that bandwidth sensitivity increases with system size, supporting the case for a photonic wafer-scale inter-GPU interconnect. We further argue that manufacturing a photonic-NoW appears to be technically feasible in the near future, making it a promising direction to scale GPU performance. This paper further highlights research and design opportunities and challenges in the context of the photonic-NoW GPU paradigm.

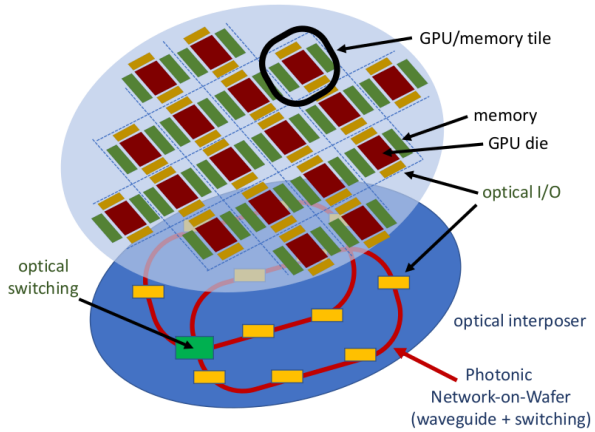



Fig. 1: Photonic-NoW GPU architecture. *A high-bandwidth, low-energy photonic network connects the GPU tiles across a wafer.*

We believe that this work is in line with where industry is heading. Nvidia's NVLink and NVSwitch technology provide high-bandwidth electrical interconnects within and across server nodes.[1] Ayar Labs and Nvidia recently announced to explore high-bandwidth, yet low-power optical-based interconnects to develop scale-out multi-GPU architectures.[2] Cerebras Systems developed a wafer-scale AI accelerator in which the cores are connected through an (electrical) on-wafer interconnect.[3] Lightmatter very recently announced Passage, a photonic wafer-scale interconnect that ties chiplets with silicon photonics and co-packaged optics; while conceptually similar to our proposal, unfortunately, not many details are provided.[4]

1. https://www.nvidia.com/en-us/data-center/nvlink/
2. https://ayarlabs.com/ayar-labs-to-accelerate-development-and-application-of-optical-interconnects-in-artificial-intelligence-machine-learning-architectures-with-nvidia/
3. https://www.cerebras.net/product-chip/
4. https://lightmatter.co/products/passage/

This article has been accepted for publication in IEEE Micro. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MM.2023.3237927

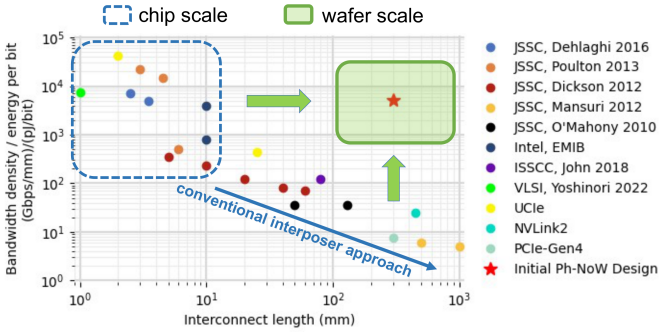



Fig. 2: Bandwidth density per energy per bit as a function of interconnect length. *Current state-of-the-art electrical interconnects cannot achieve high-bandwidth density and low energy over long distances.*

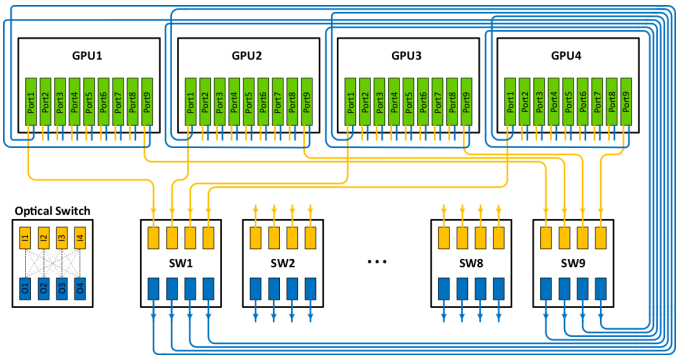



Fig. 3: A photonic-NoW 4-GPU system. *High-radix optical switches provide all-to-all connectivity; multiple ports per GPU and multiple wavelengths per port enable achieving high bandwidth.*

## 2 Motivation

GPU systems are bandwidth-hungry: keeping thousands of concurrent threads fed with data requires high off-chip bandwidth. While individual GPU chips need high-bandwidth interconnects to their local high-bandwidth memory stack, multi-GPU systems in addition need a high-bandwidth inter-GPU interconnection network for accessing data in remote memory stacks. Providing a high-bandwidth network is challenging, particularly when considering a wafer-scale GPU architecture, for three reasons. For one, we need high bandwidth density, or high bit rate per cross-sectional area of the interconnect. Second, the energy consumed per bit needs to be affordable. Third, we need interconnects over fairly long distances, up to tens of centimeters in wafer-scale GPUs.

Figure 2 summarizes the current state-of-the-art in electrical interconnect technology in terms of these three goals: bandwidth density, energy per bit, and distance of communication. The vertical axis reports bandwidth density (Gbps/mm) per energy per bit (pJ/bit), while the horizontal axis shows interconnect length (mm). It is clear from this graph that current electrical interconnects achieve high bandwidth density at affordable energy per bit at the chip scale (i.e., for short distances of less than one centimeter). However, this high figure of merit plummets when considering interconnects in the tens of centimeter range which we need for a wafer-scale network. More specifically, based on our survey of the recent literature, we find current technology to provide less than 100 Gbps/mm per pJ/bit for interconnect lengths in the 10+ cm range. Ideally, to support a high-bandwidth wafer-scale interconnect, we would like to achieve one or two orders of magnitude higher bandwidth density per energy per bit.

We thus need to revert to other technologies than electrical interconnects to provide a high-bandwidth, low-energy wafer-scale interconnection network. We argue that a photonic Network-on-Wafer could be this technology for the following four reasons. First, a photonic layer can achieve much higher bandwidth density compared to electrical interconnects. Electrical impedance-controlled high-speed transmission lines on chip (or on a silicon interposer) are typically shielded and separated by a distance of 100–200 $\mu$m to avoid crosstalk (on an organic interposer this increases to approximately 500 $\mu$m), while optical waveguides in silicon can be spaced on a pitch of 25 $\mu$m or less. Second, wavelength multiplexing can be used to further increase the bandwidth density which is not possible using electrical interconnects. Third, the power consumption of optical interconnects is quasi-independent of the interconnect length when low-loss waveguide technology is used, making it the preferred solution for wafer-scale interconnects over a few tens of centimeters. Fourth, optical interconnects can cross each other in the same layer, thereby significantly simplifying the design, in contrast to electrical interconnects which need to be routed to different metal layers.

## 3 Photonic-NoW GPU Architecture

We propose the photonic-NoW GPU architecture, as previously illustrated in Figure 1. A so-called *GPU tile* groups a GPU chiplet with its local memory stack. Because of the close proximity (less than 1 cm), this GPU-memory interconnect can be realized using traditional electrical interposer technology. Interconnecting the different GPU tiles on the other hand, involves interconnects on the order of tens of centimeters. As aforementioned, we rely on photonic waveguides to provide high-bandwidth, low-power interconnects between different GPU tiles across the wafer. Each GPU tile thus contains electro-optical transceivers to convert bits from the electrical domain to the optical domain, and back. Optical switches route these bits through the waveguides from source to destination. Wavelength division multiplexing (WDM) increases the bandwidth achieved per waveguide.

The photonic NoW is a circuit-switched network in which all switching happens in the optical domain, i.e., no back-and-forth conversion between the electrical and optical domains on the path from sender to receiver. The reason for opting for a circuit-switched network is because optical switches cannot buffer packets as in the electrical domain. We currently consider high-radix optical switches[5] where the radix is determined by the number of GPUs in the system that we want to connect to each other through a single-hop connection. The different ports per GPU connect to the different switches through a waveguide. Wavelength routing is deployed in the optical switches in which a wavelength determines the destination GPU in the system. Each GPU maintains a routing table to keep track of the network port and wavelength to reach a particular destination GPU. The routing tables are configured such that there is no conflict in any of the switches.

Figure 3 illustrates this topology for a 4-GPU system with 9 switches and 9 network ports per GPU. A single high-radix switch provides all-to-all connectivity. To achieve high bandwidth, we envision 16 wavelengths per network port and waveguide providing 32 Gbps of unidirectional bandwidth

5. The radix of a switch is defined by the number of I/O ports.

This article has been accepted for publication in IEEE Micro. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MM.2023.3237927

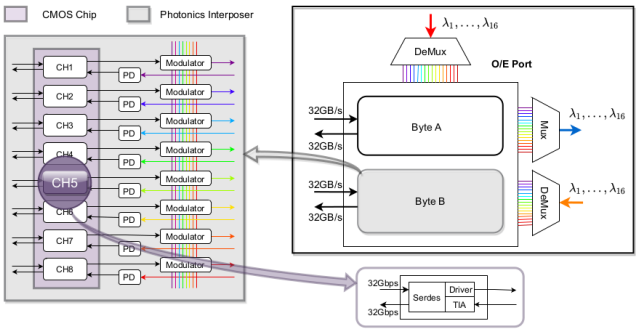



Fig. 4: High-speed optical transceiver. *A 32 Gbps bitstream per port is (de)modulated to (from) 16 wavelengths.*

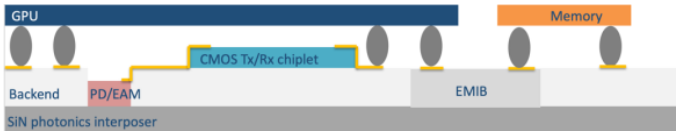



Fig. 5: Cross-section of the photonic-NoW GPU. *GPU chips are connected to their local memory stack through electrical interposer interconnects; different GPU chips across the wafer are connected through waveguides in the wafer-scale photonic network.*

per wavelength. To further increase bandwidth, we provide multiple network ports. For example, 9 network ports provide 576 GBps unidirectional or 1.152 TBps bidirectional bandwidth per GPU. A waveguide connects a GPU's network port to an I/O port of the optical switch. The target network bandwidth determines the number of optical switches and the number of network ports per GPU. Section 4 further motivates this photonic NoW design in light of a variety of network-level design trade-offs.

The interface between the electrical and optical domains relies on high-speed optical transceiver chiplets with dedicated electronic circuitry. The key electronic functions in the optical link are the modulator driver, the low-noise transimpedance amplifier (TIA) and the clock-and-data recovery (CDR) circuits to modulate outgoing traffic to one of the 16 wavelengths and to demodulate incoming traffic, as illustrated in Figure 4. Key challenges for the CMOS circuit design are the high-speed operation at low power consumption, the electro-optic co-design and interfacing, and the high-density multi-channel integration. More details are provided in Section 5.

The cross-section of the photonic-NoW GPU architecture in Figure 5 illustrates its physical implementation. GPU chips are connected to their local memory chips through traditional electrical interposer technology (e.g., EMIB). The optical transceiver chiplets discussed above provide connectivity to the optical layer in which the photonic NoW is implemented (waveguide routing and switching). Lasers are sourced outside the wafer. Power is provided using through silicon vias (TSVs). The optical waveguide layer consists of ultra-low loss SiN waveguide circuits for waveguide routing and optical switching. Active components such as high-speed photodiodes (PD), electro-absorption modulators (EAM) and driver electronics are integrated on the optical waveguide layer through micro-transfer printing as will be elaborated in Section 6. Optical amplifiers could be integrated as well to overcome the insertion loss of the photonic switches. However, they need to be positioned as far as possible from the GPU tiles to prevent efficiency degradation due to high operating temperatures.

| Technology | Footprint | Achievable radix | Switching time | Wavelength routing | Power |
|---|---|---|---|---|---|
| MEMS | large | very high | $\sim$1 $\mu$s [3] | no | low |
| MZI | medium | high | $\sim$10 ns [4] | no | high |
| push-pull MR MZI | medium | medium | $\sim$10 $\mu$s [5] $\sim$10 ns feasible* | yes | high |
| AWGR | small | high | N/A | yes | low |

TABLE 1: High-radix optical switches: network characteristics. *assumes electro-optical switching as in [4].

## 4 Photonic Network-on-Wafer

While photonic networks have been explored at the chip level, there are at least two key differences between a photonic Network-on-Chip (NoC) versus a photonic Network-on-Wafer. For one, the area footprint is much larger for an NoW (on the order of $10^4$–$10^5$ mm$^2$) compared to an NoC ($10^2$–$10^3$ mm$^2$). This implies that a photonic NoW can rely on large-footprint electronic and photonic devices that were out of reach for photonic NoCs. Second, a photonic NoW needs to provide (much) higher bandwidth between the network nodes. The network nodes in a chip-level network are individual cores or cache banks, in contrast to the full-fledged GPUs in a wafer-scale network which need much higher bandwidth.

Our proposed NoW architecture relies on high-radix optical switches, as discussed in the previous section. The key benefits of a high-radix optical switch include high bandwidth per network node, low hop count for routing, and (relatively) simple network topologies. An optical switch has the potential to provide a higher radix compared to their electrical counterparts, and thus connect more GPUs through single-hop connections. If the number of GPUs were to exceed the switch's radix, multi-hop communication would be needed to fully connect the network.

Different optical switch designs have been proposed with varying properties in terms of area footprint, achievable radix, switching time (i.e., the time it takes to reconfigure the input-output connectivity in the switch), support for wavelength routing, and power consumption. We now discuss existing optical switch designs and their key properties, see also Table 1.

1) **MEMS:** Optical switches based on microelectromechanical systems (MEMS) feature a high radix (100s of ports) and are energy-efficient. On the flip side, MEMs switches suffer from a relatively low switching speed. The size of a MEMS crossbar scales quadratically with its radix. Most MEMS optical switches do not support wavelength routing. Kwon et al. [3] demonstrated a $128 \times 128$ MEMS switch with a $\sim$1 $\mu$s switching time.
2) **MZI:** Mach-Zehnder-Interferometer (MZI)-based optical switches feature a high switching speed, but their achievable radix is lower compared to MEMS due to attenuation and cross-talk limitations. A $64 \times 64$ integrated MZI optical switch has been demonstrated by Qiao et al. [4] with a $\sim$10 ns switching time. Huang et al. [5] demonstrated that by using push-pull over-coupling micro-rings, MZI switches can support wavelength routing, however, the achievable radix is somewhat lower (10s of ports) due to increased attenuation loss. MZI switches consume more power than MEMS switches.
3) **AWGR:** The Arrayed Waveguide Grating Router (AWGR) provides static all-to-all connections between its input and output ports through wavelength routing but is non-switchable. AWGR incurs a small footprint and is power-efficient, however, the biggest limitation is that AWGR is non-switchable and may therefore

be inefficient for unbalanced NoW traffic. An $8 \times 8$ O-band AWGR for on-chip communication has been demonstrated by Pitris et al. [6].

The photonic NoW design assumed in this work aligns with the push-pull micro-ring MZI switches, i.e., we assume that the switches support wavelength routing and that if workloads exhibit time-varying bandwidth demands between different GPUs in the system, the switch can be reconfigured as such.

## 5 High-Speed Optical Transceiver

The high-speed electronic and electro-optical transceiver circuits translate the data between the digital, analog and optical domains for optimum transmission and reception. The very high density of optical waveguides and the use of multiple wavelengths in the photonic NoW provide a massive amount of parallel interconnects so that basic binary NRZ modulation at a reasonable rate (e.g., 32 Gbps) can be adopted for high energy efficiency (minimal pJ/b) [7]. This avoids more complex multi-level modulation techniques such as PAM-4, which is now widely adopted by the datacom interconnect industry. Implementing PAM-4 would bring significant disadvantages, such as higher circuit complexity, reduced noise margins, and the need for forward error correction (FEC). NRZ modulation is most efficient when sufficient bandwidth is available, which is certainly the case for the very compact (so low-capacitance, $\sim$10 fF) silicon-photonic EAM modulators and PDs considered here (and which have already been demonstrated for up to 100 Gbps NRZ [8]).

The transceiver's chiplet size and power consumption will be mainly determined by the key electronic functions, consisting of the modulator driver, the low-noise transimpedance amplifier (TIA) and the serializer-deserializer (serdes) circuits including clock-and-data recovery (CDR). Advanced CMOS technology is preferred to co-integrate all these analog and mixed-signal transceiver circuits in a very compact transceiver chiplet. Large circuits such as complex equalizer circuits, T-coils or peaking inductors are avoided by the use of moderate bit rates per wavelength and by dense co-integration through micro-transfer printing (see next section). Micro-transfer printing saves significant area as it avoids the use of flip-chip pads (with e.g. 40 $\mu$m diameter) and the associated parasitics (two pads plus bumps) as micro-transfer printing can provide interconnections on pads of only 10 by 10 $\mu$m.

We believe it is feasible to achieve a figure of merit (bandwidth density per energy per bit) that enables a wafer-scale photonic interconnect. More specifically, based on recent results by Guermandi et al. [9], we make the following estimates. Bandwidth density can be computed as the product of the bit rate per wavelength times the number of wavelengths per waveguide divided by the waveguide pitch:

$$\textit{bandwidth density} = \frac{\textit{bit rate per wavelength} \times \textit{no. wavelengths}}{\textit{waveguide pitch}}.$$

Assuming a TRx bit rate of 32 Gbps NRZ per wavelength and 16 wavelengths per waveguide, and a waveguide pitch of 25 $\mu$m at the edge of the TRx chiplet (assuming 25 mm shore line for optical I/O and a maximum of 8000 TRx chiplets of 200 $\mu$m by 400 $\mu$m underneath a single GPU tile), we obtain a total bandwidth density of 20,480 GBps/mm. Accounting for 4 pJ/b energy consumption (this includes the laser, optical switching and amplification), we arrive at the predicted figure of merit of 5,120 Gbps/mm per pJ/b, as indicated by the red star in the target area in Figure 2.

## 6 Optical Layer

The optical layer has to provide the optical connectivity between the different GPU tiles. As already elaborated before, as the size of the multi-GPU system increases, so does the interconnection distance between the GPUs. This requires the implementation of low-loss waveguide technology. SiN waveguide circuits allow realizing low-loss (dB/m-level) waveguides, while at the same time keeping the bend radius small (50 micron), allowing for compact routing. It is however a purely passive platform, except for heaters that can be implemented to enable optical switching in tens of micro-seconds. For the optical transceivers, as well as for the in-line optical amplification, non-native opto-electronic components need to be integrated on the SiN photonic interposer. A very efficient way to realize this is the use of micro-transfer printing technology [10], in which the non-native opto-electronic components (semiconductor optical amplifiers, photodiodes, modulators) are fabricated on their native (III-V) substrate, after which the devices are released from their substrate and transferred in a massively parallel way onto the SiN interposer. The devices that are transferred are only 10s of micron wide and a few micron thick, and can be placed with sub-micron precision on the SiN interposer. Once printed, the devices are electrically connected to the interposer back-end stack using a metal redistribution layer. Besides opto-electronic components, also ultra-thin CMOS transceiver circuits and high-density electrical interposer chiplets (for interconnecting the GPU to memory) can be integrated using the same micro-transfer printing technology.

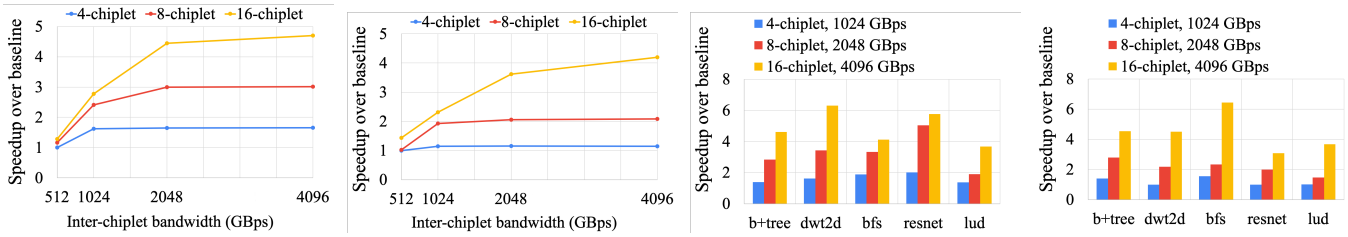
## 7 Experimental Setup

We use simulation to conduct a preliminary performance evaluation of the proposed photonic-NoW GPU architecture. We extended the GPGPU-Sim simulator [11] to model a multi-tile architecture. Each tile consists of a GPU chip along with a high-bandwidth (2 TBps) 4 GB memory stack. The GPU consists of 64 SMs and features a 4 MB last-level cache (LLC), which is configured as a memory-side cache. An on-chip 4 Tbps crossbar interconnection network connects the SMs to the LLC. The photonic NoW is modeled using BookSim 2.0 [12], and is integrated with GPGPU-Sim to model the entire system. We further assume first-touch page allocation, and consider both round-robin and distributed CTA schedulers to optimize data locality within each tile [1]. We consider a diverse set of benchmarks taken from the AI and HPC application domains, namely b+tree, dwt2d, bfs and lud from Rodinia, and ssd-resnet34 from the MLPerf inference benchmark suite. To make full use of the resources provided for increasing chiplet count, we carefully scale the input sets to provide enough threads.

We simulate 4, 8 and 16-chiplet systems. In the 4-chiplet system, we consider 8 optical switches. We further assume 16 wavelengths per port/waveguide with 32 Gbps unidirectional bandwidth, providing a total of 512 GBps unidirectional bandwidth (or 1024 GBps bidirectional bandwidth) per GPU chiplet. This baseline bandwidth configuration corresponds (roughly) to what Nvidia's NVLink provides in its fourth generation, in terms of per-GPU bandwidth, namely 900 GBps bidirectional bandwidth. We explore GPU performance's sensitivity to inter-chip network bandwidth (512, 1024, 2048 and 4096 GBps) by varying the number of ports per GPU and the number of switches proportionally. For example, a 1024 GBps bandwidth configuration requires 16 ports per GPU and 16 optical switches to provide 1024 GBps unidirectional bandwidth per GPU.

We further assume that the latency across the photonic NoW equals 6 ns: 2 ns for electrical-to-optical conversion, 2 ns for

This article has been accepted for publication in IEEE Micro. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MM.2023.3237927




(a) Harmonic mean with RR (b) Harmonic mean with DS (c) Benchmarks with RR (d) Benchmarks with DS

Fig. 6: Normalized performance (IPC): (a) the harmonic mean with the RR scheduler, (b) the harmonic mean with the DS scheduler, (c) individual benchmarks with the RR scheduler and (d) individual benchmarks with the DS scheduler. *GPU applications benefit from increased inter-chip bandwidth (see (a), (b), (c) and (d)), bandwidth sensitivity increases with system size (see (a) and (b)), and substantial speedup is obtained with higher chiplet count if balanced inter-chiplet bandwidth is provided (see (c) and (d)).*

optical transmission (assuming a ∼30 cm waveguide length), and 2 ns for optical-to-electrical conversion.

## 8 Preliminary Evaluation

Round-robin (RR) and distributed CTA scheduling (DS) are common policies used in multi-GPU systems. We find that the optimum CTA scheduling policy varies across benchmarks, number of chiplets, and inter-chiplet bandwidth. We hence evaluate bandwidth sensitivity for both RR and DS.

Figures 6a and 6b report average normalized performance (harmonic mean IPC or number of instructions executed per cycle across all benchmarks) for the 4, 8 and 16-chiplet systems under RR and DS, respectively. Four bandwidth configurations are considered, and the results are normalized to our baseline configuration, a 4-chiplet GPU with 512 GBps unidirectional inter-chiplet bandwidth per chiplet. There are two important conclusions to be taken from these results. First, performance improves significantly as we increase the number of chiplets and inter-chiplet bandwidth beyond the baseline. We note a $4.70\times$ and $4.19\times$ performance improvement under RR and DS, respectively, for 16 chiplets with 4096 GBps unidirectional inter-chiplet bandwidth per chiplet. In other words, important AI and HPC workloads benefit from increased chiplet count and inter-chiplet bandwidth. Second, when comparing the 16-chiplet performance results against the 4 and 8-chiplet results, we note that the performance improvement increases with increasing inter-chiplet bandwidth. For RR scheduling, the 4096 GBps configuration yields $1.66\times$, $3.01\times$ and $4.7\times$ higher performance for 4, 8 and 16 chiplets, respectively. Similarly, for DS scheduling, the 4096 GBps configuration yields $1.14\times$, $2.09\times$ and $4.19\times$ higher performance for 4, 8 and 16 chiplets, respectively. This suggests that increasing inter-chiplet bandwidth is more critical for increased system size. The reason is that the effective GPU-to-GPU bandwidth decreases as system size increases assuming fixed per-GPU off-chiplet bandwidth.

Figures 6c and 6d report speedup for individual benchmarks over the baseline configuration. Per-benchmark results are shown for three balanced configurations in which we simultaneously scale inter-chiplet bandwidth and system size, namely 1024, 2048 and 4096 GBps unidirectional bandwidth for the 4, 8 and 16-chiplet systems, respectively. We note speedups as high as $6.3\times$ for dwt2d under RR and $6.45\times$ for bfs using DS for the 4096 GBps 16-chiplet configuration. The key observation here is that substantial speedups are obtained at high chiplet count if inter-chiplet bandwidth increases commensurably. These results demonstrate that there is ample room for improving performance through a high-bandwidth photonic inter-chiplet network-on-wafer as we scale the number of GPUs.

## 9 Summary and Future Research Directions

This paper proposed the photonic-NoW GPU architecture and argued that it is promising and technically feasible paradigm to scale GPU performance. The photonic-NoW GPU as described and evaluated in this paper is just a first attempt to explore the large design space. We believe that the photonic-NoW GPU paradigm opens up a wide arena of potential topics for future research, across different layers in the system stack.

At the architecture level, the high inter-chip bandwidth offered through the photonic-NoW might change how best to organize and manage the memory hierarchy on a per-workload basis. In particular, a high-bandwidth inter-chiplet network renders remote accesses relatively cheap, which provides opportunities to adaptively reconfigure the memory hierarchy to cache data locally (to maximize effective bandwidth) versus remotely (to maximize the effective cache capacity).

At the network level, network reconfiguration methods will be explored for our proposed photonic NoW architecture. Suitable control mechanisms and algorithms can be designed for adapting the network's bandwidth distribution among GPUs to meet up their time-varying bandwidth demands. Moreover, when the number of GPUs exceeds the achievable radix of the optical switches, multi-hop networks will need to be designed.

For the electro-optical transceiver design, high efficiency and compact driver, TIA, and serdes circuits have been demonstrated in the literature. However, significant performance gains can still be found by customizing the designs for this particular application (data format, optical link budget, floorplan) and technology platform that combines high-speed photonics with very low optical losses and very low electrical parasitics thanks to micro-transfer printing. Furthermore, to efficiently support wavelength routing and dynamic bandwidth configuration in the network, fast wake-up and fast-locking burst-mode CDR circuits also need to be investigated to save power when no data transfer takes place.

To implement the photonic layer, further development of the micro-transfer printing technology is of key importance. This includes both the development of the transfer printing of III-V opto-electronic components, as well as electronic ICs and interposers. Proof-of-principle demonstrations of the micro-transfer printing of such components have already been made, but scaling up the technology to wafer-scale photonic interposers still needs to be demonstrated.

## ACKNOWLEDGEMENTS

We thank the guest editors and reviewers for their valuable feedback. This work is supported by UGent project BOF21-GOA-014.

## BIOGRAPHIES

**Shiqing Zhang** is a doctoral student at Ghent University, Belgium. Her research interests include GPGPU system design, computer architecture modeling and optimization. She received a master's degree in computer science from the National University of Defense Technology, China in 2018. Contact her at Shiqing.Zhang@UGent.be.

**Ziyue Zhang** is a doctoral student in the Fixed Internet Architectures & Optical Networks (FARON) research group at Ghent University-imec, Belgium. His research interest is in optical network design and network algorithm design. He received a master's degree in photonics from Ghent University in 2021. Contact him at Ziyue.Zhang@UGent.be.

**Mahmood Naderan-Tahan** is a post-doctoral researcher at Ghent university, Belgium. He previously was a lecturer at Shahid Chamran University of Ahvaz, Iran from 2016 to 2020. His research interest is in the field of computer architecture, GPU acceleration and performance benchmarking. He received a Ph.D. degree from Sharif University of Technology, Iran in 2016. Contact him at Mahmood.Naderan@UGent.be.

**Hossein SeyyedAghaei** is a doctoral student at Ghent University, Belgium. His research interests include GPGPU system design, scale-down simulation and optimization. He received a master's degree in computer engineering from Tehran University, Iran in 2016. Contact him at SeyyedHossein.SeyyedAghaeiRezaei@UGent.be.

**Xin Wang** is a doctoral student in the IDLab Design Group at Ghent University-imec, Belgium. His research interest is in high-speed mixed-signal integrated circuit design for (opto-)electronic communication systems. He received a master's degree in microelectronics from Fudan University, China in 2013. Contact him at Xin.Wang@UGent.be.

**He Li** is a doctoral student at Ghent University and imec, Belgium. His research interests include photonic integrated circuits and optical interconnects. He received a master's degree in material engineering from Nanjing University, China in 2018. Contact him at He.Li@UGent.be.

**Senbiao Qin** is a doctoral student at Ghent University and imec, Belgium. His research interests include photonic integrated circuits and optical interconnects. He received a master's degree in optical engineering from Huazhong University of Science and Technology, China in 2021. Contact him at Senbiao.Qin@UGent.be.

**Didier Colle** is a Full Professor at Ghent University and imec, Belgium. His research interests include fixed Internet architectures and optical networks, Green-ICT, design of network algorithms, and techno-economic studies. He received a Ph.D. degree from Ghent University in 2002. Contact him at Didier.Colle@UGent.be.

**Guy Torfs** is an Associate Professor at Ghent University and imec, Belgium. His research focuses on high-speed (opto-)electronic integrated circuits. He received a Ph.D. degree in applied sciences and electronics from Ghent University in 2012. Contact him at Guy.Torfs@UGent.be.

**Mario Pickavet** is a Senior Full Professor at Ghent University and imec, Belgium. His research interests include optical networking, green ICT and algorithm design for complex networking problems. He received a Ph.D. degree in electrical engineering from Ghent University in 1999. Contact him at Mario.Pickavet@UGent.be.

**Johan Bauwelinck** is an Associate Professor at Ghent University and imec, Belgium. His research focuses on high-speed (opto-)electronic integrated circuits for optical interconnects and sensing. He received a Ph.D. degree in electrical engineering from Ghent University in 2005. Contact him at Johan.Bauwelinck@UGent.be.

**Günther Roelkens** is a Full Professor at Ghent University and imec, Belgium. His research interests include photonic integrated circuits and in particular heterogeneous photonic/electronic ICs. He received a Ph.D. degree in electrical engineering from Ghent University in 2007. Contact him at Gunther.Roelkens@UGent.be.

**Lieven Eeckhout** is a Senior Full Professor at Ghent University, Belgium. His research interests include computer architecture performance analysis and modeling, and CPU/GPU microarchitecture and resource management. He received a Ph.D. degree in computer science engineering from Ghent University in 2002. He is an IEEE and ACM Fellow. Contact him at Lieven.Eeckhout@UGent.be.

## REFERENCES

[1] A. Arunkumar, E. Bolotin, B. Cho, U. Milic, E. Ebrahimi, O. Villa, A. Jaleel, C.-J. Wu, and D. Nellans. MCM-GPU: Multi-chip-module GPUs for continued performance scalability. In *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, pages 320–332, 2017.

[2] S. Pal, D. Petrisko, M. Tomei, P. Gupta, S. S. Iyer, and R. Kumar. Architecting waferscale processors: A GPU case study. In *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 250–263, 2019.

[3] K. Kwon, T. J. Seok, J. Henriksson, J. Luo, L. Ochikubo, J. Jacobs, R. S. Muller, and M. C. Wu. 128× 128 silicon photonic mems switch with scalable row/column addressing. In *CLEO: Science and Innovations*, pages SF1A–4. Optical Society of America, 2018.

[4] L. Qiao, W. Tang, and T. Chu. Ultra-large-scale silicon optical switches. In *2016 IEEE 13th International Conference on Group IV Photonics (GFP)*, pages 1–2. IEEE, 2016.

[5] Y. Huang, Q. Cheng, A. Rizzo, and K. Bergman. Push—pull microring-assisted space-and-wavelength selective switch. *Optics Letters*, 45(10):2696–2699, 2020.

[6] S. Pitris, G. Dabos, C. Mitsolidou, T. Alexoudi, P. De Heyn, J. Van Campenhout, R. Broeke, G. T. Kanellos, and N. Pleros. Silicon photonic 8× 8 cyclic arrayed waveguide grating router for o-band on-chip communication. *Optics Express*, 26(5):6276–6284, 2018.

[7] M. Raj, Y. Frans, P.-C. Chiang, S. L. Chaitanya Ambatipudi, D. Mahashin, P. De Heyn, S. Balakrishnan, J. Van Campenhout, J. Grayson, M. Epitaux, and K. Chang. Design of a 50-gb/s hybrid integrated si-photonic optical link in 16-nm finfet. *IEEE Journal of Solid-State Circuits*, 55(4):1086–1095, 2020.

[8] J. Verbist, M. Verplaetse, S. A. Srinivasan, J. Van Kerrebrouck, P. De Heyn, P. Absil, T. De Keulenaer, R. Pierco, A. Vyncke, G. Torfs, X. Yin, G. Roelkens, J. Van Campenhout, and J. Bauwelinck. Real-time 100 Gb/s NRZ and EDB transmission with a GeSi electroabsorption modulator for short-reach optical interconnects. *Journal of Lightwave Technology*, 36(1):90–96, 2018.

[9] D. Guermandi, L. Bogaerts, M. Rakowski, Y. Ban, P. De Heyn, N. Pantano, P. Bex, J. De Coster, J. He, A. Phommahaxay, S. Balakrishnan, C. Demeurisse, J. Bertheau, K. J. Rebibis, P. Nolmans, X. Sun, A. Srinivasan, S. Van Huylenbroeck, S. Lardenois, A. Miller, P. Absil, P. Verheyen, D. Velenis, M. Pantouvaki, and J. Van Campenhout. TSV-assisted hybrid FinFET CMOS – silicon photonics technology for high density optical I/O. In *45th European Conference on Optical Communication (ECOC 2019)*, pages 1–4, 2019.

[10] J. Zhang, G. Muliuk, J. Juvert, S. Kumari, B. Haq, C. O. de Beeck, B. Kuyken, G. Morthier, D. V. Thourhout, R. Baets, G. Lepage, P. Verheyen, J. V. Campenhout, A. Gocalinska, J. O'Callaghan, E. Pelucchi,

B. Corbett, A. Trindade, and G. Roelkens. III-V-on-Si photonic integrated circuits realized using micro-transfer-printing. *APL Photonics*, 4(11):10, 2019.

[11] A. Bakhoda, G. L. Yuan, W. W. Fung, H. Wong, and T. M. Aamodt. Analyzing CUDA Workloads Using a Detailed GPU Simulator. In *Proceedings of the International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 163–174. IEEE, 2009.

[12] N. Jiang, D. U. Becker, G. Michelogiannakis, J. Balfour, B. Towles, J. Kim, and W. J. Dally. A detailed and flexible cycle-accurate network-on-chip simulator. In *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 86–96, 2013.