Title

Scalable High Performance Memory Subsystem with Optical Interconnects

Author
Fotouhi, Pouya

Publication Date
2021

Scalable High Performance Memory Subsystem with Optical Interconnects

By

Pouya Fotouhi

Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

S.J. Ben Yoo, Chair

Venkatesh Akella

Jason Lowe-Power

Committee in Charge

2021

Contents

List of Figures

List of Tables

Abstract

Scalable High Performance Memory Subsystem with Optical Interconnects

Data movement has become a limiting factor in terms of performance, power consumption, and scalability of high-performance compute nodes with increasing numbers of processor and memory systems. Optical interconnects enabled by Silicon Photonics could not only overcome this limitation but also change the way we think about system architectures and memory hierarchies. This dissertation aims to introduce and evaluate scalable high performance computing architectures based on optical interconnects. This dissertation presents the motivation and background, architecture design, and evaluation results for the following case studies:

Investigating the design challenges in large-scale many-core processors, the impact of interconnection fabric on the overall system performance and power consumption, and how Silicon Photonics can alleviate system constraints.

Studying off-chip memory networks capable of providing HPC compute nodes with terabytes of memory capacity by interconnecting several 3D stacked DRAM modules through a packet-switched network interface. Replacing legacy interconnects with sophisticated optical networks could significantly reduce memory access time and energy - a largely unexplored research area.

Addressing the scaling limitations in chiplet-based systems, in particular, large inter-chiplet non-uniform latencies, distance-related energy overheads, and limited Input-Output (IO) bandwidth, and exploiting the properties of optical interconnects to propose a scalable uniform memory architecture.

Rethinking the architecture of state-of-the-art high-throughput accelerators, the impact of memory access latency variations on the overall performance and system design, and the key challenges in scaling memory and compute capacity in these systems. A new architecture is proposed to reduce the contention within the memory system with the help of a partitioned memory controller and an all-to-all passive optical interconnect that is amenable for a 2.5D based implementation using off-the-shelf memory modules.

ACKNOWLEDGMENTS

I would like to thank Professor Yoo for giving me the opportunity to be part of his dynamic research team at UC Davis. This multi-dimensional learning experience would benefit my career without any doubts.

I owe my gratitude to my committee members Professor Akella and Professor Lowe-Power. Professor Akella was exactly the advisor I needed to close the gap between the architectural ideas and discussions at the technology level. Looking back at my interactions with Jason, I cannot imagine a better mentor for this chapter of my life. I am truly grateful for having the opportunity to work with Jason as any aspect of my technical character is influenced and improved by him.

I would like to thank my colleagues in UC Davis, Dr. Roberto Proietti, Dr. Sebastian Werner, Xian Xiao, and Marjan Fariborz for fruitful discussions and collaborations over the years.

Finally, I would like to thank my family for their support along the way. To my parents, Farideh and Fereidoun, thank you for your ever-present support in every step. To my wife, Elnaz, thank you for being there in all the ups and downs. Without your consistent encouragement I could not achieve my academic goals.

Chapter 1

Introduction

Modern High-Performance Computing (HPC) systems exploit increasing numbers of interconnected heterogeneous processor and memory systems to achieve peta-FLOP performance goals [1, 2]. For example, exascale computing initiatives such as Frontier and El Capitan announced by the US government recently are expected to have thousands of CPU and GPU nodes to meet their performance targets. Meanwhile, we are on the threshold of an exponentially-growing data-driven transformation of the entire economy. Running data-intensive HPC applications on heterogeneous systems in a distributed fashion ties the overall performance of the system to the cost and efficiency of data movements.

Moreover, growing data sets in modern workloads are driving the need for higher processing power and memory bandwidth/capacity in HPC systems. Unfortunately, the slowing down of Moore's law reduces the ability to attain higher processing power and memory capacity within a single compute node (”scale-up”) for each generation of silicon technology, forcing system designers to add compute nodes (”scale-out”) to satisfy performance demands. Scaling-out, however, leads to distributed memory architectures with compute nodes operating in different address spaces, and requires explicit synchronization through software (e.g. message passing). Again, this approach results in significant performance and energy overheads for data movement between compute nodes - a key challenge in current HPC systems. Scaling-up shared memory architectures within a single address space, on the other hand, allows hardware-managed coherence which is significantly faster and allows programmers to focus on what matters for parallel speed-ups rather than synchronization. Moreover, the cost of providing hardware coherence (i.e., traffic, storage cost for tracking sharers, latency, and energy) is generally considered to scale gracefully with the core count for hierarchies in modern systems [3]. Given the technological limitations of electrical interconnects in terms of link bandwidth, link charge time, attainable radix, and energy consumption, the interconnect fabric has become a bottleneck in HPC systems [4].

Silicon Photonics (SiPh) offers high-bandwidth distance-independent links within a lower power envelope and potentially enabling flatter topologies for higher performance. Moreover, SiPh devices can exploit Wavelength-Division Multiplexing (WDM) to perform wavelength-selective routing which allows one node to be connected to multiple other nodes through a single optical IO pin (addressing them on different wavelengths), enabling high-radix low-diameter networks with high bandwidth density.

While SiPh manufacturing process is not as mature and largely adopted as standard CMOS, SiPh manufacturing and photonic integrated circuit design kits (PDKs) have seen significant growth and investment in the past ten years, now allowing low-cost SiPh integration [5]. Among several SiPh fabrics, the Arrayed Waveguide Grating Router (AWGR) is a passive SiPhs fabric with compact layout which offers scalable all-to-all connectivity through use of wavelength routing. Recent advances in the fabrication process of AWGRs now enable their integration with significantly reduced footprint ($1mm^2$), crosstalk (<-38dB), and loss (<2dB), making AWGRs a favorable candidate for energy-efficient all-to-all connectivity within HPC systems.

This dissertation explores the design space of AWGR-based interconnects for both processor and memory systems, and comparing them to the state-of-the-art SiPhs fabrics and aggressive electrical baselines under contemporary workloads.

The organization of this dissertation is as follows. Chapter 2 presents a brief background discussion on SiPhs as the major enabling technology used throughout this dissertation. Chapter 3 evaluates the use of AWGR-based SiPhs Network-on-Chip to achieve energy-efficient all-to-all connectivity in large-scale interposer-based HPC systems. Chapter 4 studies off-chip memory networks capable of providing tera-bytes of memory capacity by optically interconnecting several 3D stacked DRAM modules through a packet-switched network interface. Chapter 5 investigates the scaling limitations in chiplet-based systems, in particular, large inter-chiplet non-uniform memory architecture (NUMA) latencies, distance-related energy overheads, and limited IO bandwidth, and exploiting the properties of optical interconnects to propose a scalable uniform memory architecture (S-UMA) that overcomes all NUMA-related performance challenges. Chapter 6 discusses the architecture of state-of-the-art high-throughput accelerators, the impact of memory access latency variations on overall performance and system design, and key challenges in scaling memory and compute capacity in these systems. This chapter proposes a novel high-throughput accelerator architecture which aims to reduce the contention within the memory system with the help of a partitioned memory controller and an all-to-all passive optical interconnect. At last, Chapter 7 presents a summary of the work presented in this dissertation and illustrates future perspectives.

Chapter 2

Background

This dissertation investigates different components in HPC systems in search for performance and power improvements. This chapter presents a brief discussion on SiPhs as the major enabling technology used throughout this dissertation.

Figure 2.1 depicts a reference optical SiPhs link. A laser functions as a light source to provide the optical medium, confined within a waveguide, for data transmission, on which data is transmitted through modulators and received through filters and photodetectors. Both modulators and filters are based on microring resonators (MRs), each of which tuned to one particular wavelength.

2.1 Microring Resonators

Microring resonators are designed and fabricated to resonate only with specific individual wavelengths. The resonance optical frequencies are repeated at intervals known as the Free Spectral Range (FSR). That is, a microring with f as original resonance frequency will also resonate at $f \pm k \times FSR$ where $k \in \mathbb{Z}$. This frequency specific response enables implementation of MRs as filters. This filtering functionality can be obtained using passive MRs with a fixed resonance frequency assigned during the design process, or through use of active MRs designed to *tune* their resonance frequency as the amount of current in their base layer changes. The latter class of MRs, with tunable resonance frequencies, are ideal candidates for modulation.

In order to understand the significance of MRs to SiPhs interconnects, it is useful to discuss how data transmission is performed in an optical link. Figure 2.1 depicts an example optical

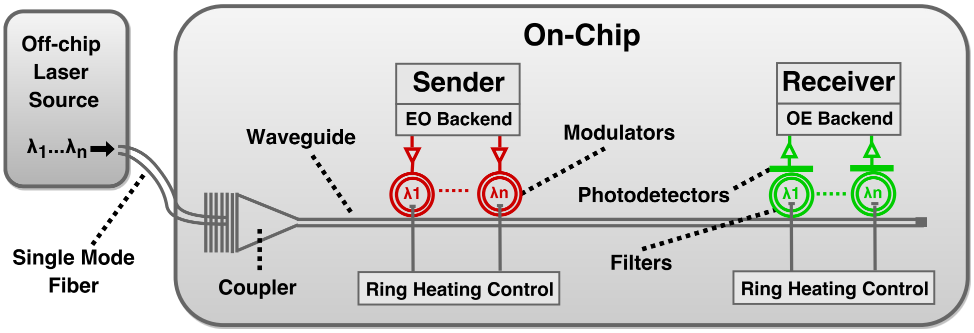
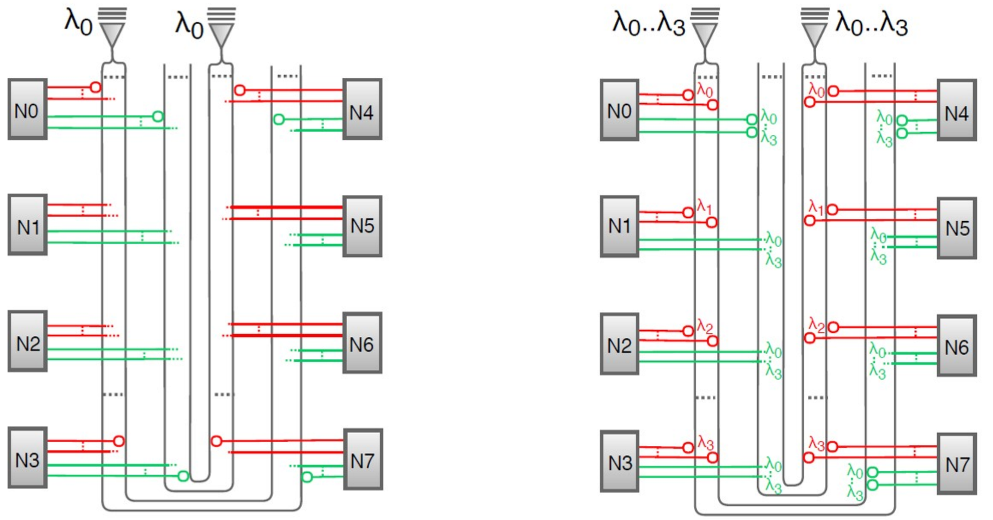


Figure 2.1: Reference SiPhs Optical Link

link with all components needed to transmit data between a source-destination pair. Light generated from a laser is confined inside an optical fiber which is then coupled into an on-chip waveguide. Modulators encode bits onto the optical medium (electrical-to-optical (EO) conversion), and filter-photodetector pairs extract the optical signal, performing optical-to-electrical (OE) conversion.

Optical signals can consist of multiple frequencies ($f_0 - f_n$) on which data can be transmitted in parallel - a technique commonly referred to as Wavelength-Division Multiplexing (WDM). In order to exploit WDM, one modulator and MR filter per wavelength is needed at the sender and receiver, respectively. Given that MRs are resonant devices, their resonance wavelength depends on device geometry, dimensions, the ambient temperature, and variation thereof can cause the resonance frequency to deviate from their design values, effectively causing malfunctioning. While fabrication yield can be mitigated by MR trimming, protecting MRs from on-chip temperature variations requires integrated heaters ensuring thermo-optical control of each individual MR.

Aside from ensuring a correct behavior, integrated heaters can also be used deliberately to dynamically turn on/off MR filters and implementing wavelength-selective switching devices by “dropping” wavelengths from one waveguide to another. Changing the ambient temperature of a MR with heaters so that its resonance wavelength shifts beyond the free spectral range of all wavelengths on a link effectively allows to dynamically turn off (and on) a MR. Several previous studies leverage this approach to implement path setup and tear down functionality of circuit-switched optical networks based on wavelength-selective routing [6–8].



(a) Single-Writer Single-Reader (SWSR) Buses

(b) Multiple-Writers Single-Reader (MWSR) Buses

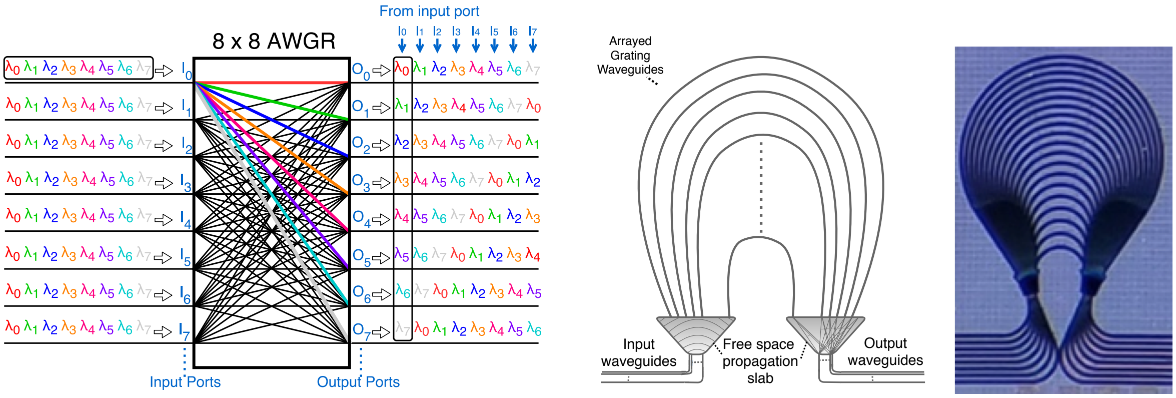
Figure 2.2: Single Reader Optical Buses Implemented using Microrings with (a) Single Writer and (b) Multiple Writers [9].

2.1.1 Optical Buses with MRs

Microrings, in combination with WDM, can be used to implement optical buses. The reference example in Figure 2.1 connects one source-destination pair, and it is referred to as Single-Writer-Single-Reader (SWSR) optical bus. An implementation of a 4×4 SWSR bus is shown in Figure 2.2a. Alternatively, WDM can also be leveraged to connect multiple source-destination pairs with just one waveguide. By equipping each destination with MR filters tuned to different wavelengths a Single-Writer-Multiple-Reader (SWMR) bus can be implemented. Similarly, as shown in Figure 2.2b, a Multiple-Writer-Single-Reader (MWSR) can be implemented using MR modulators tuned to different wavelengths at sender side.

2.1.2 Broadband Ring Resonators

MRs are promising SiPh devices for implementing filtering and modulation functionalities. However, since MRs are designed to resonate at a narrow range of wavelengths, MR based solutions often require enormous amount of MRs to be fabricated. In contrast, Broadband Ring Resonators (BRRs) [6][10] can be tuned at a wide range of wavelengths and have been proposed with a circuit switched approach in which path setup/teardown phases are necessary prior to data



(a) Switching Matrix (b) Schematic (c) Layout

Figure 2.3: The Arrayed Waveguide Grating Router (AWGR). (a) AWGR Structure. (b) Switching Functionality. (c) Realized Physical Layout of a SiN AWGR

transmission. That is, BRRs must be dynamically tuned to a certain wavelength range to route data to the correct destination. Although promising, with the current technology in hand, BRRs have been found to be less practical due to the latency overheads caused by the setup/teardown and tuning phases and large area footprint [10].

2.2 Arrayed Waveguide Grating Router

An AWGR functions as a wavelength-multiplexer and demultiplexer, with the wavelength routing pattern as shown in Figure 2.3a. All wavelengths ($\lambda_0..\lambda_7$) inside a waveguide entering the AWGR on a given input port are evenly distributed over all the output ports of the AWGR (one wavelength to one unique output port), effectively implementing a wavelength demultiplexer function (as shown for input 0 in Figure 2.3a). At the same time, different inputs can communicate to the same output on different wavelengths, effectively implementing a multiplexer function at each AWGR output port. One intriguing property of AWGRs is that multiple signals on the same wavelengths can traverse the AWGR without interfering with each other out as long as they enter the device on different input ports. This means that multiple input waveguides can be connected to an AWGR (like in Figure 2.3a), whose wavelengths are evenly distributed to the output ports (depending on the input port), effectively forming an all-to-all, non-blocking switching fabric–with just one device and without the need for MR filters or a large number of waveguides.

Figure 2.3b shows the schematic of an AWGR device. Wavelengths inside the input waveguides enter the free-space propagation region and subsequently traverse the grating waveguides, which have a constant length increment (δL). Each wavelength undergoes a constant change of phase attributed to the constant length increment in the grating waveguides. Light diffracted from each waveguide of the grating interferes constructively and refocuses at the output waveguides, with the output channels being wavelength dependent on the array phase shift. Figure 2.3c is a screen-shot of the physical implementation of a fabricated Silicon Nitride (SiN) 8×8 AWGR, which is just $1mm \times 0.8mm$ in size.

AWGRs have already found application in the telecom industry allowing for years of fabrication know-how with high-yield manufacturing [11]. Several previous studies for on-chip implementations dismissed the use of AWGRs due to their large area footprint (in the range of multiple mm^2); however, those footprints mostly refer to AWGRs fabricated with Silica. Recently demonstrated Silicon Nitride (SiN) based AWGRs have an area footprint of less than $1mm^2$ [12], a completely reasonable footprint for the architectures investigated in this proposal.

Another interesting property of AWGRs is that they are symmetric and can be used bidirectionally, i.e. it offers its properties in both directions (wavelengths entering the AWGR from the left or right are multiplexed the same way). This provides an interesting interconnection fabric (Figure 2.4) that can provide all-to-all connectivity between several nodes with just one device - highly efficient from a layout point-of-view. In particular, this is actually the ideal connectivity pattern for memory-to-processor communication, which will be discussed in detail in Chapter 4.

The efficient wavelength-distribution mechanism of AWGRs coupled with its small area footprint and low losses make it an ideal candidate for global all-to-all connectivity, especially as input/output waveguides can be directly routed to the senders/receivers, thereby minimizing path lengths and waveguide crossing. SiPh typically leverages Dense Wavelength-Division Multiplexing (DWDM) signals to increase bit-parallelism and, in turn, link bandwidth, within a single waveguide. The wavelength routing attributes of AWGRs introduced in this section shows that AWGRs can only distribute a single wavelength between each input-/output-port pair, which prevents multi-wavelength communication between nodes, and thereby limits to-

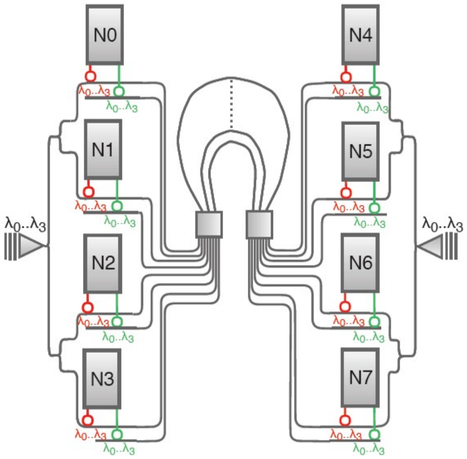


Figure 2.4: Bidirectional AWGR

tal node-to-node bandwidth inside a single AWGR to the modulation rate (i.e., data rate per wavelength).

2.2.1 Achieving High Node-to-node Bandwidth in AWGRs

As discussed in the previous section, AWGR is an ideal candidate for implementing all-to-all topologies. While this an ideal topology from a performance point-of-view, it distributes the output bandwidth of a given node between all the destinations it has. Thus, point-to-point bandwidth limitations could be a serious drawback of topologies enabled by AWGRs. However, three recent technological key advances now enable efficient bandwidth scaling inside AWGRs, making it a favorable design choice.

2.2.1.1 PAM4 Modulation

The limitation of single-wavelength communication between nodes would have been a serious concern in previous studies which mostly assume On/Off keying (OOK) modulation (1 bit per symbol), a modulation rate of 10Gb/s, and DWDM levels between 16-64 wavelengths (shown to provide the highest energy efficiency [13]). To satisfy bandwidth demands without DWDM, significantly higher modulation rates than 10Gb/s would be necessary. While higher modulation rates are not significantly detrimental to the energy efficiency of the photonic components (referred to as *bit-rate transparency* in the photonics community [4]), clock generation/recovery and driver and SERDES circuitry consume more energy at higher data rates.

One way of increasing the data rate is using advanced modulation techniques that increase the data rate by encoding multiple bits into one symbol. Although technologically feasible, the required transceiver circuitry for such modulation techniques was shown to consume too much energy (~3pJ/bit [14]). Fortunately, Moazeni et al. [15] recently demonstrated a new PAM4 transceiver (2 bits per symbol) on a 45nm platform which only requires a 'spoked' MR (and driver circuitry) with just 5μm in radius and 0.197pJ/bit to convert two electrical input bits into a PAM4 signal at 20Gb/s modulation rate–effectively enabling a data rate of 40Gb/s per wavelength (four times higher than 10Gb/s OOK) with very high energy efficiency and compact layout.

Although this PAM4 transceiver is a big step towards efficient AWGR-based optical interconnects, single-wavelength communication at 40Gb/s bandwidth between source-destination pairs is still significantly lower than in current multi-lane electrical counterparts (e.g., the typical assumption of 128-bit wide links at 2GHz provides 256Gb/s bandwidth). Luckily, the following advances boost the AWGR bandwidth even further.

2.2.1.2 Spatial-division Multiplexing with AWGRs

Spatial-Division Multiplexing (SDM) can be used to increase bandwidth by adding links (or in this case, AWGRs) to the optical interconnect architecture. Although the AWGR is the only device necessary to provide all-to-all connectivity, implementing multiple AWGRs aside each other is ultimately limited by the footprint of AWGRs ($1mm^2$) and losses incurred by more complex wiring and waveguide crossings (leading to higher laser power). The more complex physical layout can be detrimental to the overall power efficiency and bandwidth density and would require additional fabrication efforts (e.g., for tapering waveguide crossings to reduce loss [16]).

Fortunately, both the footprint and the layout concerns are overcome by recent demonstrations of AWGRs implemented on separate SiPhs layers [17]. This stacked AWGR approach not only removes any area/footprint concerns, but also eliminates waveguide crossings altogether and allows to physically place the AWGR between the nodes such that path lengths, and, in turn losses, are minimized, leading to reduced output power requirements at the laser source. This enables the use of SDM of AWGRs to increase bandwidth without negative impacts on laser power or area.

2.2.1.3 Bit-parallelism in AWGRs

One interesting feature of AWGRs is that the wavelength routing is cyclic with the period, called FSR, which means that an output port j can be reached by an input port i using wavelength $\lambda_{ij} + k \times FSR$ and $k \in \mathbb{Z}$. This cycling behavior enables each input port to communicate with each output port using multiple wavelengths (DWDM), referred to as *multi-FSR AWGRs*. Although limited by the crosstalk inside the AWGR and the wavelength range of the laser, this bit parallelism does not need to be very high to provide sufficient bandwidth when combined with modulation rates of up to 40Gb/s (and possibly SDM). Although it has been known to be theoretically possible, only until recently, Grani et al. actually successfully demonstrated the feasibility of AWGRs with bit-parallelism by leveraging multiple FSRs [18].

With all these recent advancements, a high-bandwidth all-to-all network can be constructed with just a single AWGR. For instance, the bisection bandwidth of an 8×8 AWGR with 32Gb/s modulation rate and a bit-parallelism of 2 is $8 \times 8 \times 32 \times 2 = 4096Gb/s(4Tb/s)$, which equals the bisection bandwidth of an 8×8 2D Mesh with 128-bit wide links at 2GHz, and could even be improved further by implementing two AWGR atop each other without any impact on area footprint or layout.

2.2.2 AWGRs vs. State-of-the-art SiPhs Interconnects

Some of the main benefits of AWGRs, such as short paths with fewer waveguide crossings, compact layout and fabrication with high device maturity, have already been discussed. While being significantly easier to physically implement than topologies consisting of just basic optical buses with MRs, the main benefit over wavelength-routed interconnects is that AWGRs are not resonance devices and thus do not require heating[1] (as opposed to MRs). This is a major benefit as less on-chip power means that more of the power budget can be dedicated to the compute and memory. In addition, if an off-chip laser is used, the only power consumption consumed on the interconnect fabric is dynamic power (and some leakage power of the EO/OE circuitry), which is known to be relatively low compared to electrical interconnects, and distance-independent. Therefore, AWGRs can be a major enabler for a high-bandwidth, low-power interconnection

[1] The refractive index of AWGR changes with the ambient temperature, therefore some heat-control mechanism is required. However, given AWGR is a single device to control its temperature, the heat control problem is less eminent to solve compared to similar topologies implemented with a large number of MRs.

fabric for processor-to-processor and processor-to-memory communication as will be discussed in details in Chapter 3 and Chapter 4.

Chapter 3

Photonic Interconnects for Large-scale 2.5D Integrated Systems

Commonly called “2.5D” integration [19] exploits an interposer to tightly integrate processor and memory dies side-by-side within the same package, eliminating the large parasitic from off-package communication, greatly increases in-package memory bandwidth while largely avoiding thermal challenges associated with 3D stacking [20]. In particular, *processor disintegration* [21] represents a promising approach to decrease the overall cost of 2.5D integrated systems by leveraging the higher manufacturing yield of small many-core dies compared to larger ones. For instance, instead of implementing one 64-core processor die, four 16-core processor dies integrated aside each other on an interposer can provide similar processing power at relatively higher manufacturing yield of the smaller dies and, in turn, lower overall system cost [22]. Several commercially-available products already benefit from 2.5D integration [23][24] and future systems can be expected to further exploit the memory bandwidth and cost benefits of 2.5D integration with disintegrated processors by integrating increasing numbers of dies into the same package.

Recent studies have shown that 2.5D integrated systems, by nature, put significant strain on the network-on-chip (NoC) by exhibiting high communication traffic [25]. In addition, shrinking power budgets, large physical distances, and poor technology scaling of electrical interconnects make the design of energy-efficient high-bandwidth NoCs extremely challenging. Moreover, current commercially-available systems were shown to suffer from high communication latency overheads between processors on the interposer which significantly degrade system performance [26][27].

Low-diameter topologies can potentially reduce the latency but are prohibitively expensive due to the energy consumption of electrical interconnects for interconnecting chips over large distances. This limitation could prevent 2.5D integrated systems to scale to larger number of processing dies in the future.

Silicon photonics (SiPh)–enabling optical communication on chip–features ideal physical properties to overcome these challenges, i.e., almost distance-independent energy consumption and high bandwidth density through dense wavelength-division multiplexing (DWDM)[4]. These advantages over electrical interconnects allow designers to exploit SiPh to design NoCs with 'flatter' low-diameter topologies and capitalize on their performance metrics. Moreover, their distance-independent energy consumption allows adjusting the spacing between dies on interposers to larger/varying physical distances, which was recently found to provide significant performance improvements by overcoming the 'Dark Silicon' problem caused by thermal challenges [28].

Unfortunately, enabling global all-to-all connectivity with SiPh comes with its own challenges. State-of-the-art SiPh all-to-all fabrics proposed to date are either based on optical buses [6] or wavelength-routed photonic NoCs (WRPNoCs)[29]. While bus-based designs quickly become impractical and cost-inefficient due to either large numbers of waveguides or wavelengths, WRPNoCs are based on microring resonators (MRs) to perform wavelength-selective routing, causing power overheads for thermo-optical control and a challenging physical layout. The ideal wavelength routing fabric would provide all-to-all connectivity without excessive need for waveguides, wavelengths, and MR heating, while enabling a compact physical implementation.

The Arrayed Waveguide Grating Router (AWGR) enable scalable, low-loss wavelength routing between all input and output ports by utilizing N wavelengths and N input and output waveguides in support of an all-to-all $N \times N$ interconnection. Recent fabrication advances in CMOS-compatible Silicon Nitride (SiN) AWGRs enable footprints of ~1mm^2 [12] for 16×16 AWGR, which is very compact as the only fabric needed for routing (as opposed to 256 MRs in WRPNoCs [10]). Moreover, four recent key demonstrations make AWGRs a viable candidate as a high-bandwidth interconnect. First, recently demonstrated sub-pJ Pulse-Amplitude Modulation (PAM4) transceivers at 40Gbps data rate offer high-data-rate low-energy communication on a single wavelength [30]. Second, *bit-parallel* AWGRs capable of routing multiple wavelengths to the same output port by exploiting the AWGR's cyclic routing properties now enable low levels of DWDM inside AWGRs, too [18]. Third, recent demonstrations of AWGR fabricated on an SiPh interposer show low crosstalk at scale [31]. Fourth, multiple AWGRs can now be fabricated atop each other with negligible inter-layer crosstalk, thereby eliminating any layout and area concerns of multiple AWGRs inside a NoC [17]. Based on these advancements, AWGRs could be a major enabler for energy-efficient, high-bandwidth, and scalable all-to-all connectivity in 2.5D integrated system and therefore deserve a detailed analysis of their potentials and shortcomings.

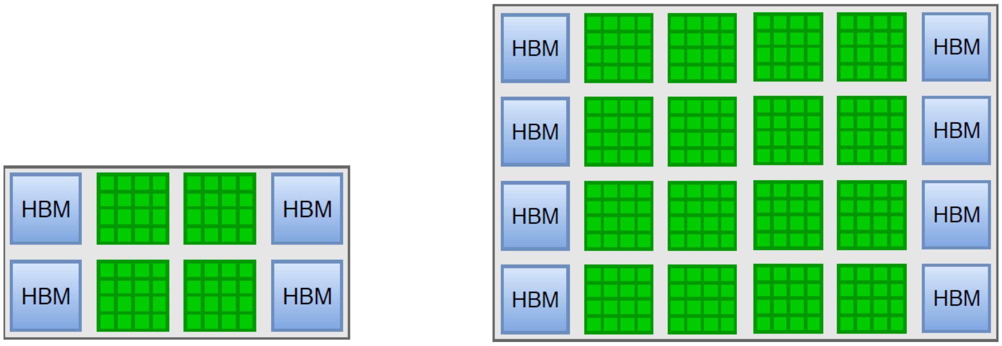
This chapter conducts the following:

- A scalability study of large-scale 2.5D integrated computing systems showing AWGR-based interconnection networks as a promising and suitable solution in terms of latency, bandwidth, and energy per bit.

- An exploration of different AWGR-enabled topologies, as well as their use cases and suitability to solve the challenges of interposer-based large-scale systems with up to 256 cores.

- An extensive power and performance evaluation of AWGR-based networks and a comparison to state-of-the-art interconnects with up to 256 cores (16 processor dies).

Simulation results show that AWGR-based topologies can offer an average speed-up of at least 1.2× (64 cores) and 1.25× (256 cores) compared to the closest electrical competitor for a range of PARSEC3.0/SPLASH-2x workloads with at least 1.32× lower power, and more than 2× reductions in average packet latency across different synthetic workloads with up to 3× sustained bandwidth at 256 nodes. These results suggest that the AWGR could be a key enabler for scaling future 2.5D integrated systems both in terms of performance and power by providing a low-latency, scalable, and lower-power interconnect.



(a) 64 Cores

(b) 256 Cores

Figure 3.1: Example interposer-based systems integrating (a) 64- and (b) 256-core processors composed of 16-core processor dies alongside 3D-stacked DRAMs (in this example, high-bandwidth memories (HBMs) [33]).

3.1 2.5D Integrated Systems

3.1.1 2.5D Integration: Opportunities and Challenges

Increasing interposer sizes offer many opportunities for future large-scale systems inside a single package and enable higher numbers of processor and memory dies to be tightly integrated side-by-side on an interposer. Figure 3.1 depicts example floorplans of 2.5D integrated systems with a large many-core processor (64 and 256 cores) disintegrated into smaller 16-core dies and 3D-stacked DRAMs distributed along the two opposite edges of the chip (as commonly found in literature and commercial designs [21, 23, 24])[1]. Such systems could enable over 1000 processor cores with hundreds of GB memory capacity tightly integrated in the same package and thereby be a key enabler for future high-performance chips operating at high energy efficiency.

Several studies have explored the design space of 2.5D integrated systems [25], interconnection networks extended to the interposer to increase bandwidth [20], and processor disintegration to lower cost through improved overall manufacturing yield [21], and have made compelling cases for enabling exascale systems [32]. Nevertheless, Loh et al. [25] identified numerous design challenges, many of which are yet to be solved.

First, the trend towards growing numbers of high-bandwidth memory (HBM) stacks inside

[1] Although the processor dies in this example are many-core processors, heterogeneous integration of various different computing and memory chips (such as GPUs, FPGAs, non-volatile memories, etc.) have also been considered an attractive solution for future systems and would equally benefit from the contributions of this approach) [32]

the same package, more channels per HBM, and wider DRAM buses to increase memory bandwidth leads to higher bandwidth demands on the NoC. This will make the implementation of electrical NoCs within acceptable power envelopes extremely challenging - especially in combination with the large distances imposed by interconnecting several dies on an interposer.

Second, the NoC's clock network must deal with die-to-die-to-interposer process variations, possibly even with different technology generations of different dies or heterogeneous integration of multiple different dies. Loh et al. [25] proposed to decompose the NoC into smaller, independent clock domains to have easier timing and to support dynamic voltage and frequency scaling (DVFS). This indicate that topologies should ideally support clustering or be hierarchical.

Thirdly, large distances (e.g., AMD's FURY is $1011mm^2$ [24]) and routing between dies increases link latency, suggesting that disintegration comes with a performance-cost trade-off. Routing electrical signals over such distances at satisfactory speed can only be attained with power-consuming repeater circuitry, resulting in more of the power budget being dedicated to the NoC and less to the compute (assuming a system operating under a power cap) [25]. Electrical NoCs tailored to interposer-based systems were shown to be more efficient than conventional NoCs for monolithic chips [20], but cannot fully overcome these limitations. Especially for larger-scale systems with hundreds of cores implemented with tens of processor dies, the interconnection network represents a major obstacle to power efficiency.

3.1.2 Using Silicon Photonics To Overcome Design Challenges

Recent studies have shown how the performance, power, and scalability benefits of integrated SiPh interconnects in interposer-based systems become increasingly evident (compared to their electrical counterparts) with growing number of dies and physical distances [34]. These benefits are mainly enabled by the physical properties of optical communication, which offer distance-independence in terms of energy and latency and provide high bandwidth links with better scalability.

The energy-efficient high-bandwidth interconnects offered by SiPh provide sufficient bisection bandwidth in the NoC to support core clustering, which, in turn, allows practical and efficient DVFS control by grouping clustered cores into separate clock domains. Besides, the discussed physical properties of SiPh allow to implement flatter topologies (and even all-to-all connectivity) in NoCs with much higher scalability than electrical interconnects. This can be leveraged to offer very low latencies even for large physical distances (like in the 256-core example in Figure 3.1), effectively giving the illusion of moving cores 'closer together'.

While the scientific literature is replete with proposals utilizing SiPh fabrics to construct NoCs [10][35][36], they do not study the utilization of AWGRs in interposer-based systems which provide highly scalable and energy-efficient all-to-all connectivity with just a single passive device. More importantly, significant technological improvements of AWGRs [31, 37] in the last years (in terms of footprint, loss, and crosstalk) make them superior to state-of-the-art SiPh fabrics. AWGR-based NoCs eliminate the need for on-chip heating power for thermo-optical control in the switching fabric, thereby largely overcoming one of the most important concerns of optical interconnects at the chip level. In addition, in combination with an off-chip laser, AWGR-based NoCs significantly reduce on-chip power consumption without performance degradation, leaving more of the power budget (constrained by thermal design point in HPC systems) to the compute and memories. The following section discusses the benefits of SiPh, AWGRs, and the topologies they enable in more detail.

3.2 AWGR-enabled Networks

The unique wavelength routing of AWGRs opens up many opportunities and a new design space to be explored. As we will see in this section, the structure of the AWGR, its placement of input and output ports, and all-to-all connectivity pattern are ideal for global all-to-all implementations in NoCs. In particular, bipartite graphs and all-to-all networks can be efficiently implemented with AWGRs–both of which providing flat, low-diameter topologies capable of enabling low-latency communication not attainable with electrical interconnects at high energy efficiency and compact physical implementation. This section first discusses how AWGRs can enable bipartite graphs and all-to-all topologies, followed by a discussion on enabling multi-wavelength high-bandwidth communication with AWGRs as the switching fabric and a comparison of AWGR to alternative SiPh interconnection fabrics.

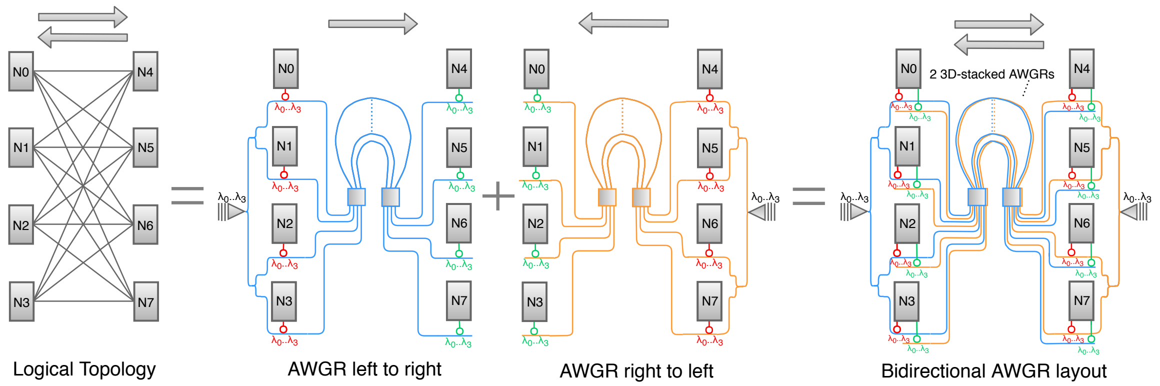


Figure 3.2: Bipartite graph constructed out of two unidirectional AWGRs fabricated atop each other on separate SiPh layers. AWGR stacking enables a bipartite graph with short point-to-point links and without waveguide crossings. Note that the same can be obtained using a single $2N \times 2N$ AWGR used bidirectionally, though leading to higher crosstalk.

3.2.1 Bipartite Graphs with AWGRs

In principle, AWGRs are *bidirectional*, i.e., light can traverse an AWGR in both directions without interference (and with the same wavelength routing pattern), effectively forming a bidirectional all-to-all switching fabric with just a single device (this logical topology is shown in Figure 3.2 on the left).

Two design options to implement bipartite graphs exist: 1) utilizing two AWGRs unidirectionally or 2) utilizing a single AWGR with bidirectional operation. Both enable a compact, low-loss all-to-all fabric with short and direct links between each source-destination pair and without any waveguide crossings, but have their own set of benefits and trade-offs, which will be discussed in the following for the example 4×4 bipartite graphs in Figure 3.2.

Constructing a bipartite graph with two separate $4{\times}4$ AWGRs–one for each direction–is easily feasible in interposer-based systems, whose size ($>1000mm^2$ is a well-established size [21]) can conveniently accommodate several AWGRs (few mm^2); however, recent demonstrations of 3D-stacked AWGRs on separate SiPh layers show that AWGRs can be integrated vertically with negligible inter-layer crosstalk and loss [17] (more details in Section 2.2.1.2), thereby taking up the horizontal real estate of just a single AWGR. Figure 3.2 illustrates how two AWGRs in opposite directions stacked atop each other provide a compact implementation of a bipartite graph.

Utilizing a single AWGR and exploiting its symmetric, bidirectional wavelength routing operation requires an 8×8 AWGR to provide 4×4 bidirectional all-to-all connectivity (each node needs a separate input and output port to avoid filtering out its own signals). Therefore, the final layout would look exactly like the stacked AWGRs shown on the right in Figure 3.2, just that instead of two SiPh layers and 4×4 AWGRs, only one 8×8 AWGR on a single layer is used (in general, with such an implementation, an N×N bipartite graph needs a 2N×2N AWGR).

While both approaches enable a compact all-to-all switching fabric, each entails its own set of benefits and trade-offs, and several aspects of AWGRs should be considered when constructing all-to-all connectivty between the input/output ports. The loss inside AWGRs is mainly caused by the free-space propagation region and is relatively independent of the port count (e.g. the loss inside an 8×8 and an 16×16 AWGR is very similar [12]), meaning that a doubling of the port count to construct a bipartite graph using a single AWGR does not increase the loss inside an AWGR noticeably.

However, the footprint of an AWGR increases with the port count and utilizing two AWGRs with half the port count will result in a more compact implementation by adding an extra layer during the fabrication. In addition, a design with two separate AWGRs versus one AWGR will require smaller wavelength range (N×channel spacing of the AWGR compared to 2N×channel spacing of the AWGR).

This work follows the Space Division Multiplexing (SDM) approach combined with stacking, and utilizes 2-AWGRs for implementing a bidirectional all-to-all fabric. This approach results in reduced footprint, allows designing AWGR with wider channel spacing, and thereby provides a more scalable fabric.

3.2.2 All-to-all Connectivity with AWGR

Although the bidirectional all-to-all fabric discussed in the previous section utilizes the AWGR in the most efficient manner in terms of footprint, layout, and loss, a true all-to-all fabric connecting all nodes directly with each other offers the ideal from a performance point-of-view (offers 1) a diameter of one which minimizes zero load latency and 2) maximum path diversity for load balancing) and could simplify the programming of many-core processors by enabling uniform memory/cache access. AWGRs provide an efficient implementation of such a fabric when connecting each input/output port to each sender/receiver, respectively.

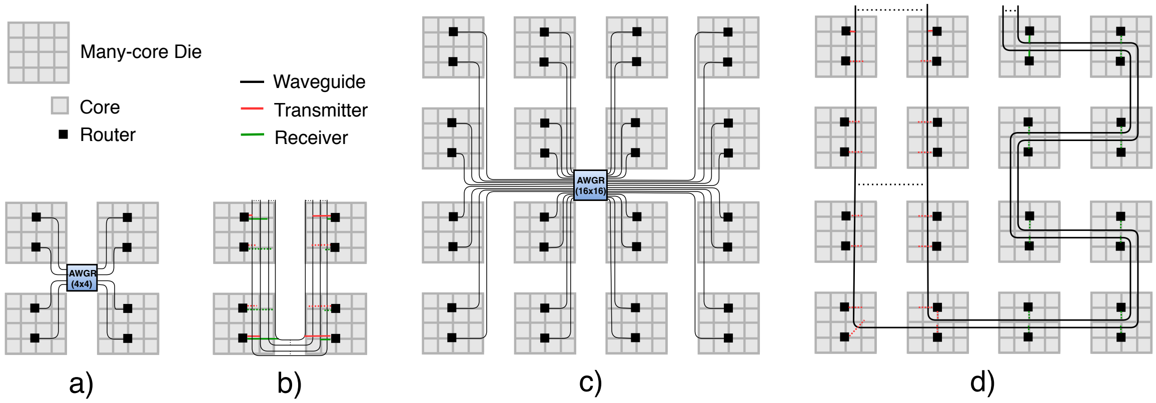


Figure 3.3: Bidirectional all-to-all NoC layout with optical buses and AWGRs for 64 (a) and b)) and 256 (c) and d)) cores with a clustering of 8 cores at each router (note that, for illustration purposes, d) only shows one side (left) of the bipartite graph. The same waveguides are needed to connect the nodes on the right to those on the left).

The all-to-all utilization scenario of AWGRs, however, causes significantly higher crosstalk compared to the bidirectional use as more signals on the same wavelength are traversing the AWGR (for supporting the same number of nodes in the NoC as the bipartite graph). Moreover, the number of wavelengths required to provide all-to-all connectivity inside the AWGR equals the number of ports (and, in turn, nodes in the NoC). A 64 × 64 AWGR would thus require 64 wavelengths for routing which enter the AWGR in each input port and impose crosstalk upon each other inside the AWGR. In fact, all-to-all connectivity with a single AWGR for port counts higher than 32 was shown to be challenging with SiN AWGRs (the material providing the lowest footprint and loss) due to excess crosstalk and require multiple AWGRs [38] to keep both crosstalk and laser power at feasible and practical levels. Therefore, this chapter focuses on the bipartite topology enabled by AWGRs.

3.2.3 AWGRs vs. State-of-the-art SiPh Fabrics

Figure 3.3 compares the physical implementation of a global all-to-all interconnect constructed with AWGRs to bus-based designs (e.g. SWSR, MWSR, SWMR - the layout of each would be the same) for 64 (a, b) and 256 (c, d) cores in a realistic example target system which is like the disintegrated processor design placed on an interposer discussed in Section 3.1. With 16 cores per die, 8 of which are clustered at one router. The red and green lines indicate that nodes need to place MRs to modulate and filter signals adjacent to these waveguides to enable optical communication (as introduced in Figure 2.1).

3.2.3.1 AWGRs vs. SiPh Buses

The bus-based crossbars have a U-shaped layout, which has widely been used in recent literature [35][39][40] as it allows for a crossbar implementation with a straight-forward layout and without waveguide crossings. The U-shape of the waveguides leads to longer waveguides and, in turn, path losses; however, direct links between all sender-receiver pairs would lead to a very challenging layout and introduce a large number of waveguide crossings, making the U-shaped layout the most efficient. The AWGR-based crossbar allows for direct links without imposing waveguide crossings. These benefits become more important as the system scales to a larger number of nodes (Figure 3.3 c and d): while the AWGR still provides short links and a compact layout, in bus-based designs waveguides must be routed in an S-shaped fashion to be in close proximity to the nodes (otherwise, modulators and receivers would have to be driven over mm distances), not only causing a more complicated layout, but also higher waveguide losses.

Aside from these benefits of AWGRs, there are a number of additional challenges of contemporary SiPh switching fabrics that can be overcome by AWGRs. The number of waveguides in crossbars consisting of SWSRs grows quadratically with the number of nodes, which is area-inefficient and complicates layout. SWMRs or MWSRs overcome these issues by requiring only one waveguide per sender or receiver, respectively; however, assigning waveguides to senders/receivers complicates the physical implementation as more nodes are added to the NoC (e.g., in the SWMR case, each receiver must place MRs at each of the senders waveguides to filter out signals). Waveguide pitches, MR radii, and spacing between components are in the range of ~5μm [40]. This results in designs in which MRs are placed fairly far away (could be $> 100\mu$m) from the actual nodes, complicating placement of driver and heating circuitry, causing non-negligible energy consumption on the interconnect, and limiting scalability.

3.2.3.2 AWGRs vs. WRPNoCs

WRPNoCs (not shown in Figure 3.3) overcome the layout issue as each node only needs one waveguide for sending and receiving, respectively. MR filters are strategically placed between waveguides to route wavelengths through the network to the correct destinations [10, 29, 41]. A sender merely has to modulate its data on the correct wavelengths to ensure that its data packet will arrive at the destination. WRPNoCs require fewer and shorter waveguides to create a crossbar than buses but rely on MRs for routing which consumes heating power. Moreover, MRs are typically distributed across the chip (depending on which layout provides the lowest losses), which complicates layout as heating circuitry must be co-located. Numerous studies dedicated just for investigating efficient WRPNoC layouts underline this issue (i.a. [29, 42]).

Besides, the number of MRs in WRPNoCs has poor scalability as the number of nodes in a NoC increases although numerous studies with advanced topologies aimed to decrease the number of MRs for switching (thousands of MRs are needed for switching for NoC sizes > 32 nodes) [7, 10, 29]. This leads to significant on-chip power for thermo-optical control of MRs, which makes them less practical than bus-based designs.

Using an AWGR alleviates all of the aforementioned issues. First, one input and output waveguide per node is required which allows placing all of the transceiver circuitry close to the nodes, thus simplifying layout. Second, wavelength routing does not rely on MRs and AWGRs do not require on-chip heating (refractive index changes caused by temperature variations can either be controlled by off-chip thermo-electric coolers (TECs) or can be avoided altogether with athermal AWGRs [43]), thus completely eliminating heating circuitry and power for routing. Also, as mentioned above, an AWGR-based crossbar does not exhibit any waveguide crossings, which lead to higher losses in WRPNoCs [29] and can only be avoided by U-shaped layouts in bus-based designs. Finally, AWGRs can be used bidirectionally or used unidirectionally stacked atop each other which allows constructing a bidirectional all-to-all fabric using just one/two passive component(s) consuming no power.

Given these benefits over state-of-the-art SiPh fabrics, AWGRs represent a promising candidate to enable low-power, low-latency, high-bandwidth, and scalable interconnection between processor dies in large-scale 2.5D integrated systems.

3.3 Methodology

The goal of this study is to investigate the benefits and drawbacks of AWGR-based NoC architectures and to reveal which interconnection fabric–both electrical and photonic–provides the best scalability for large-scale 2.5D integrated systems. Simulations are based on the con-

Table 3.1: Target System Configuration (layout as in Fig. 3.1)

Parameter	Description
Cores	64 and 256 cores, 16-core dies; x86 out-of-order; 2GHz
Caches	Private 32kB L1I/D and 256kB L2 per core; MSI coherence
Memory	8GB HBM2.0 per die; 1024-bit 1GHz interface
Dimensions	2mm tile width/length; 2mm spacing between dies
NoC	Routers : 128-bit at 2GHz; 5 flit deep buffers; 2 cycle traversal
	Electrical links: 128-bit at 2GHz; 1 cycle traversal
	Optical links: 64-bit at 2GHz; 1 cycle traversal
	6 virtual channels per port with virtual cut-through switching

figuration listed in Table 3.1, and assume a target architecture like the disintegrated processor in Figure 3.1 interconnected as shown in Figure 3.3. Each die has 16-cores, i.e., the 64- and 256-core configurations have 4 and 16 dies placed on the interposer. This work assumes that all interconnection fabrics are exclusively routed on the interposer. With processor dies of $\sim 74mm^2$ [21] and HBM dies of $\sim 42mm^2$ [44], with a $200\mu m$ spacing for die placement [25], the total interposer areas for 64- and 256-core configurations are $\sim 360mm^2$ and $\sim 1500mm^2$ respectively.

3.3.1 Experimental Setup

Simulation study are done through Sniper [45] with high-performance applications from the SPLASH-2x and PARSEC3.0 [46] benchmark suites, covering workloads of various different communication profiles. In addition, Garnet2.0 [47] inside gem5 [48] was utilized for performance simulation with synthetic traffic. Power and latency of the CMOS circuitry (i.e., electrical links, routers, and EO/OE backends) were modeled with DSENT [13] and a 22nm technology node. Laser power was modeled based on the formula by Li et al. [40] with 20% laser efficiency [49], -18dBm receiver sensitivity [50], 1dB coupler loss [51], 0.2dB splitter loss, 0.027dB/mm waveguide propagation loss, 0.01dB MR through loss, 0.5dB MR drop loss, and 0.12dB waveguide crossing loss [10][49]. The switching fabric includes 1.4dB, 1.5dB, and

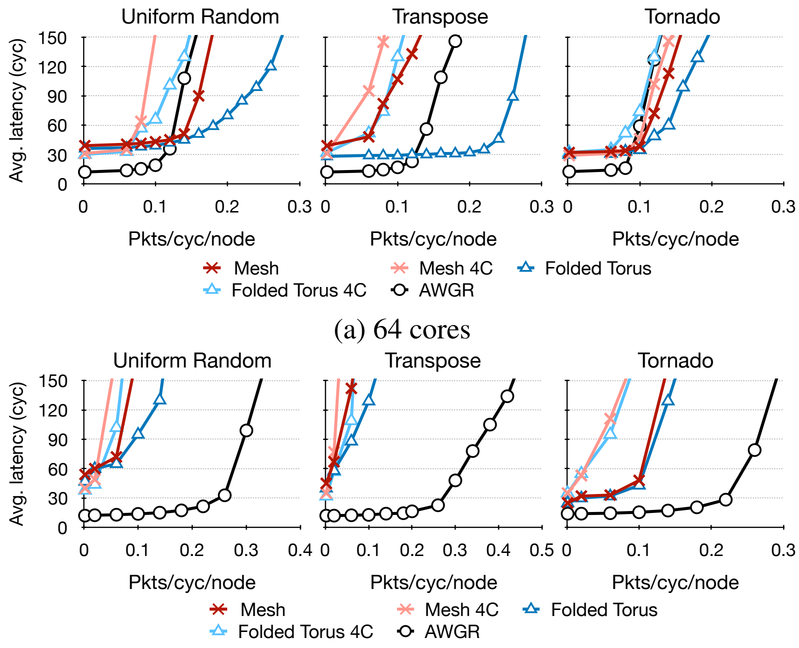
1.8dB loss for a 4 × 4, 8 × 8 and 16 × 16 SiN AWGR with -27dB, -24dB, and -20dB crosstalk, respectively [12]. The power model assumes $20\mu W/MR$ for thermo-optical control of MRs, and 11ps/mm signal propagation of light in silicon. This proposal relies on off-chip static WDM lasers with 8 and 32 unique wavelengths in the 64- and 256-core cases, respectively.

3.3.2 NoCs Under Investigation

The vast majority of previously proposed NoCs make use of SiPh interconnects with optical buses, i.e., SWSR, SWMR, or MWSR (some prominent examples ATAC [52], Firefly [53], Meteor [54], Corona [55]. Therefore, this chapter compare the bipartite graph use case ('AWGR') of AWGRs to implementations with SWSR buses. MWSR and SWMR buses assign subsets of wavelengths to each destination on one waveguide, which would require hundreds of different wavelengths for a bipartite graph supporting more than 64 cores, which would be an unrealistic design consideration. Therefore, some sort of SDM is necessary to obtain a feasible design, and since area constraints are not critical on the interposer, this chapter aims to compare AWGRs with SWSR buses[2]. This chapter compares the proposed NoC with aggressive electrical baselines including 2D Mesh ('Mesh'), 2D Mesh with a clustering of 4 ('Mesh4C'), 2D Folded Torus ('FoldedTorus'), and 2D Folded Torus with a clustering of 4 ('FoldedTorus4C')–all of which utilizing XY routing–to identify the benefits of SiPh in large-scale systems.

The all-to-all use case of AWGRs in this study is ommited as analysis has shown that an optical all-to-all NoC imposes impractical laser power overheads for core counts larger than 64 and crosstalk that might render their implementation infeasible for the current state of SiPh technology. Bidirectional AWGR NoC connects the cores of the target system as shown in Figure 3.3: 8 cores are clustered at each router, resulting in a 4 × 4 and 16 × 16 bipartite graph for 64 and 256 cores, respectively. The AWGRs implementing these graphs are assumed to be stacked atop each other, with one AWGR for each direction. To support a 64-bit wide link (at 2GHz), this proposal utilizes 32Gbps PAM4 signals, a bit-parallelism inside the AWGR of 2, and a SDM level of two AWGRs stacked on top each other (leading to a stacking of four AWGR in total for the entire NoC). For the SWSR implementation, four wavelengths at 32Gbps PAM4

[2] Note that a more extensive comparison between the different SiPh interconnection fabrics in terms of loss, power consumption, etc. is provided in [56].



(a) 64 cores

(b) 256 cores

Figure 3.4: Average packet latency (cyc) vs. injection rate (pkts/cyc/node) for synthetic workloads

on each waveguide is assumed.

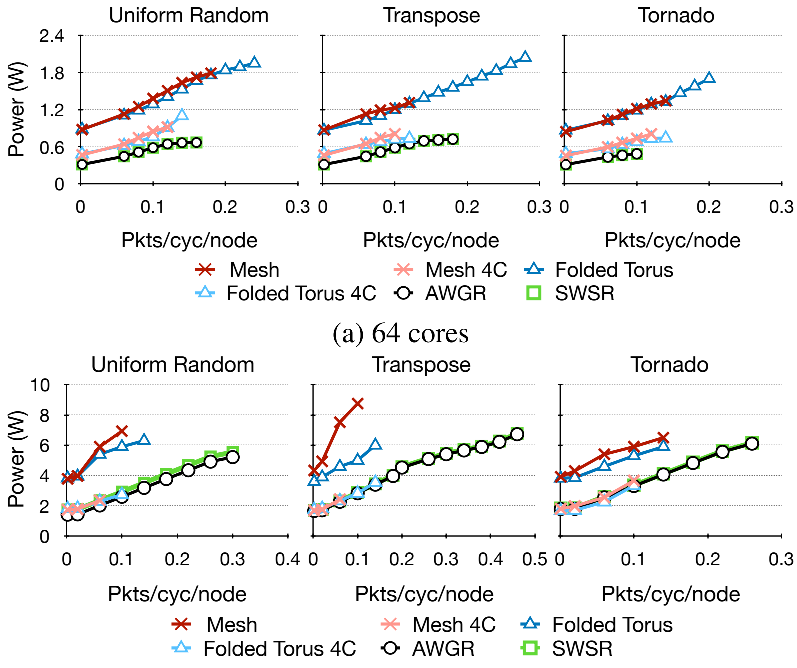
3.4 Evaluation Results

3.4.1 Synthetic Traffic

3.4.1.1 Performance Results

Figure 3.4 shows the latency results of the NoCs under investigation for varying injection rates with uniform random, transpose, and tornado traffic to stress different corner cases of the topologies (sources compute destination nodes based on the synthetic traffic model by Dally et al. [57]). Each core in the system injects packets into the system with an increasing injection rate and packet sizes varying from 8 bytes to 72 bytes based on Garnet's pseudo cache coherence model [47]. The figures reporting latency do not show the bipartite graph implementation with SWSR as it has the same performance results as the AWGR.

Our AWGR-based topology reduces packet latency by more than 2× prior to reaching network saturation compared to all alternative NoCs for both network sizes and all workloads. From a throughput point-of-view, only the folded torus topology can sustain noticeably higher



(a) 64 cores

(b) 256 cores

Figure 3.5: Power consumption (W) vs. injection rate (packets/cycle/node) for synthetic workloads

throughput than the AWGR for 64 cores. For 256 cores, the AWGR-based topology dominates all other NoCs in terms of throughput, attributed to the high bisection bandwidth of the global crossbar and fewer number of hops which combined lead to less network congestion.

3.4.1.2 Power Results

Figure 3.5 plots the power consumption vs. injection rate, which allows to identify whether the high network loads can be sustained with satisfactory power consumption. The power results include the entire network power, i.e., leakage, dynamic, MR heating, and off-chip laser power.

Not only does the AWGR-based topology offer much lower latency and sustains higher network loads, but also does so with less power consumption. Only the clustered versions of the electrical NoCs can compete with the AWGR, mainly due to the high leakage power overheads and high dynamic power imposed by larger number of hops in the non-clustered NoCs. Compared to the crossbar implementation with SWSR, AWGR-based topologies offer sightly less power consumption, which comes from the lower losses (and, in turn, lower laser power) in the AWGR fabric provided by shorter waveguides.

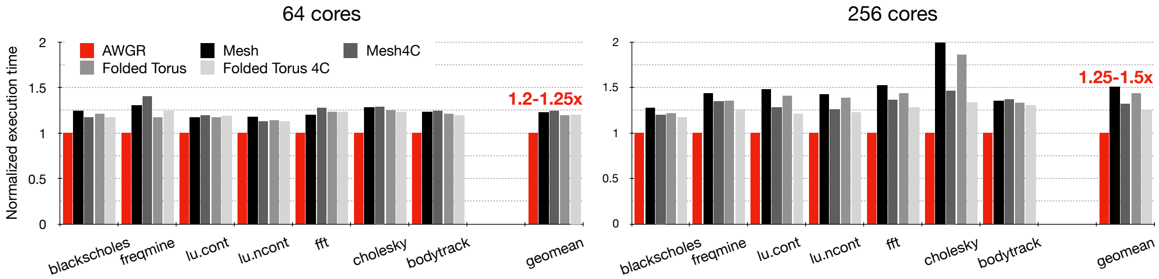


Figure 3.6: Application execution time normalized to AWGR-based topology

3.4.2 Application Traffic

3.4.2.1 Performance Results

Figure 3.6 shows the application execution time normalized to the AWGR topology for 64 and 256 cores. For both cases, the AWGR-based topology reduces execution time of each of the simulated applications. The flat topology enabled by SiPh and the AWGR fabric offers a significantly reduced application execution time for both 64 and 256 cores. Generally, the higher the degree of data sharing in the application (and, in turn, on-chip traffic), the bigger the performance gains of the AWGR topologies, implying that applications exhibiting higher on-chip traffic profiles than those from the SPLASH2.x/PARSEC3.0 benchmark suites might benefit from AWGR-based interconnects even more.

3.4.2.2 Power Results

Figure 3.7 shows the power breakdown of all topologies for 64 and 256 cores, respectively. Breakdowns for each application are omitted for brevity, considering that we have not observed significant variations across different workloads. The AWGR based topologies require the lowest power consumption out of all topologies for both cases, confirming the supreme scalability and energy efficiency of AWGR-based interconnects.

Leakage power is known to dominate the power budget for NoCs with buffers and virtual channels for technology nodes of 22nm and lower [58] (power gating techniques can almost halve leakage power [59], but cannot fully overcome these overheads). Deploying a high-bandwidth low-loss SiPh fabric like AWGRs allows to cluster more nodes at each router without performance drawbacks, allowing for much fewer routers in total and, in turn, less leakage power (despite the fact their routers have higher radix). Dynamic power plays an increasingly smaller role as the system size increases, which is likely due to the fairly low NoC utilization

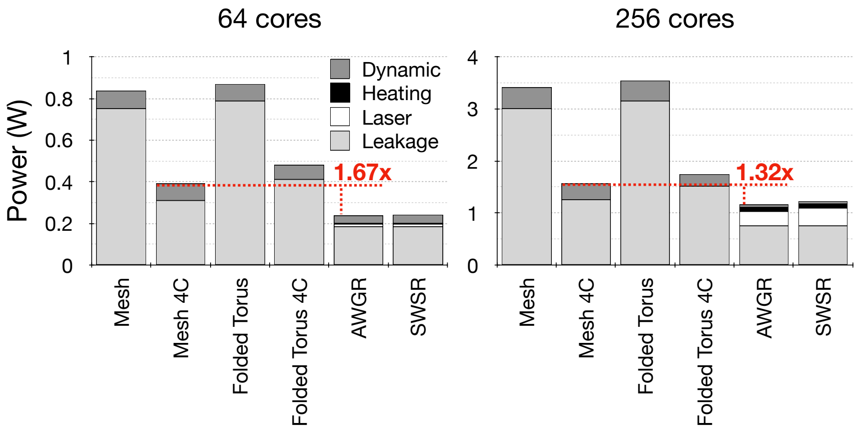


Figure 3.7: Power breakdown for 64 and 256 cores

characteristics of the SPLASH-2x/PARSEC-3.0 workloads and their relatively small data sets (compared to the total size of the on-chip caches in system configuration). Multi-programmed workloads, highly virtualized systems, and applications with higher cache miss rates, data sharing, or data sets would probably benefit from the AWGR even more as it offers lower dynamic power due to its low-diameter topology and distance-independent energy consumption.

The AWGR and SWSR have very similar power consumption for 64 nodes; however for 256 cores, the waveguide length of the bus based design and the number of waveguides needed (scaling quadratically with the number of nodes in SWSR crossbar) leads to significant waveguide propagation and splitter loss, and in turn to higher laser power requirements compared to the AWGR-based solution which offers short direct links between source-destination pairs.

Figure 3.8 plots the energy-delay-product (EDP) of the considered NoCs, workloads, and system sizes to put the performance speed-up into perspective with power consumption. In general, AWGR offer by far the most energy-efficient design. The EDP benefits compared to a SWSR bus are lower mostly because both networks provide the same performance and thus the same application execution, de-emphasizing the power reductions of the AWGR compared to the SWSR. Compared to the electrical baselines, however, AWGR improves power efficiency by at least 1.67× for both network sizes.

3.4.3 Discussion

Simulation results revealed that the low diameter of global bipartite graphs can have a large impact on packet latency, execution time and energy efficiency of applications in interposer-based

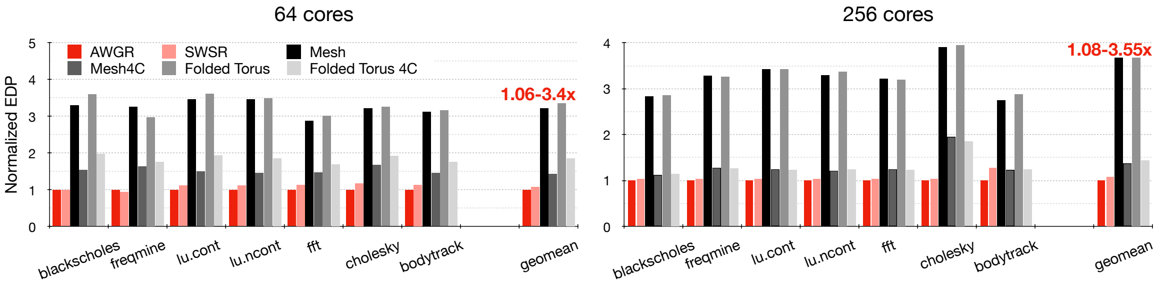


Figure 3.8: Energy-delay-product normalized to AWGR-based topology

systems. The low network diameter reduces network latency by more than 2× for low network loads, which makes them ideal for large-scale interposer-based systems executing latency-critical applications. This low latency also allows to make easier estimates on the quality of service, and makes large-scale systems easier to program as memory accesses are much less likely to have large latency differences (as it is the case in electrical NoCs).

AWGRs not only provide better performance and power metrics, but also represent a scalable and compact wavelength routing platform that allows for a simple, straight-forward physical layout. Rather than imposing large overheads in the number of waveguides or a complicated physical layout with MR-based switching fabrics, the AWGR's unique wavelength routing mechanism might be a key enabler for practical future SiPh on-chip interconnects.

This proposal requires 4×4 and 16×16 AWGRs for 64- and 256-core configurations respectively. In terms of scalability, a system with 1024 cores would require 64×64 AWGRs. Currently, there are, to the best of our knowledge, no demonstrated 64×64 SiN AWGRs to be found in literature. The loss inside AWGRs is relatively independent of the port count, and the main challenge for AWGRs with high port counts would be the crosstalk. However, there has been successful demonstrations of techniques to use multiple smaller AWGRs (in terms of port count) to provide the same functionality at lower crosstalk [38]. Also, AWGRs with much higher port counts have already been demonstrated in Si [60], albeit with considerably larger footprint ($176mm^2$ compared to $1mm^2$). This area might be negligible for a system with 512 dies (each ~ $74mm^2$), but the interposer size/cost and crosstalk of the AWGR should be considered.

Moreover, the footprint overhead of this proposal is insignificant. Each processor die should accommodate the coupler ($2\mu m^2$ [51]), MR ($25\mu m^2$ [30]), and backend circuitry for EO/OE conversion ($930\mu m2$ calculated using DSENT [13]) for each link. Thus, the total area occupied by optics for 64- and 256-core designs are 3828μm^2 (+0.005%) and 15312μm^2 (+0.0%2) respectively. With processor die size of $\sim$ 74mm^2 [21] and 1mm^2 for AWGRs, the aggregate overhead is 0.021% and 0.082% for 64- and 256-core configurations.

SiPh evolves quickly, and fabrication advances create new opportunities for NoC architectures. For instance, compelling demonstrations of on-chip lasers enable low-latency/energy adaptive laser control which can save large amounts of laser power [61][62][63]. The AWGR-based topologies proposed in this chapter could be efficiently combined with adaptive lasers to further improve power efficiency. Although many challenges regarding stabilization mechanisms and laser control are needed, this could represent a great opportunity for power savings.

All in all, simulation results confirm that SiPh is, in general, an excellent candidate for overcoming the interconnect bottleneck in large-scale interposer-based systems, which would enable more of the power budget to be dedicated to the processor and memory dies. Using AWGRs further supplements SiPh by offering a switching fabric that allows for direct links between source-destination pairs without imposing any waveguide crossings and their associated losses and additional fabrication steps. All these attributes make AWGR a key enabling technology for future computing systems leveraging tight integration in the same package to meet performance goals at high energy efficiency.

3.5 Conclusion

The study presented in this chapter is published in the Journal of Optical Communications and Networking (JOCN) [64]. This chapter investigated the use of AWGRs inside NoCs for interposer-based disintegrated processors to address the power, performance, and scalability drawbacks of electrical NoCs in large-scale systems, studied AWGR-based NoC topologies, and compared them to state-of-the-art SiPh interconnects and aggressive electrical baselines. Simulation results show that AWGRs provide significant performance speed-up, power reductions, and better scalability compared to the state of the art while enabling a practical physical implementation of low-diameter interconnection networks. AWGRs could be a key enabler of future scaling of 2.5D integrated systems with low communication latency, which could be of high impact for current and the future of computing systems that leverage tight integration.

Chapter 4

Optically-Interconnected Memory Networks

Growing data sets in big data, machine learning, and HPC workloads drive the need for large memory capacity. Working sets in data centers and high-performance computing systems are continuously increasing in size, and machine learning relies on very large and diverse data sets to train more powerful models and fuel further advancements in the field.

Although conventional DDR memories are currently still the predominantly utilized memory technology in commercial products, increasing memory capacity with DDRs on dual-inline memory modules (DIMMs) is typically limited to a few DIMMs per channel as it requires to trade-off memory bandwidth for capacity[1]. This trade-off is unacceptable for future HPC systems since the notorious ‘memory wall’ is already encumbering performance scaling of processors. Alternatively, a higher number of memory channels could provide higher memory capacity, but the total pin count on processor packages have merely doubled every six years since 1996, and the vast majority of pins are dedicated to power and ground, leaving very few pins to memory channels (often called ‘pin wall’) [66][67]. Even if more pins could be dedicated to memory channels, this typically entails a costly impact on packaging [65].

Due to these technological limitations, 3D stacked DRAM–offering larger memory bandwidth and energy efficiency–has made its way into commercial products in the form of JEDEC’s

[1] Adding DIMMs to a channel increases the electrical load on a channel, which in turn often translates to lower operating frequency [65]

high-bandwidth memory (HBM) [68][69] (e.g. NVIDIA's Volta GPUs [23]) or Micron's hybrid memory cube (HMC) [70] (a customized version of which can be found in Intel's Knights Landing [71]), both of which are projected to have a global growth rate of 33% until 2023 [72].

HMCs are a particularly interesting approach to attain high, tera-byte scale memory capacities as they implement an abstract high-speed serial interface and a logic layer underneath the stacked DRAM implementing both memory controllers and a switching fabric. This allows creating a network of HMCs (often referred to as *memory networks* [65]) and, in turn, an efficient way of scaling memory capacity. To provide parallel access to the memory (i.e. higher throughput), an HMC's stacked DRAM is divided into several *vaults* (4-32 based on the configuration of the HMC [70]), each of which controlled by a memory controller on the base logic layer.

The switch on the logic layer of an HMC must provide connectivity between each input port (4-8, based on configuration [70]) and each memory controller at high speed (vault access rates are 80Gbps, input links need SERDES (SERializer/DESerializer circuitry). It runs at up to 25 Gb/s per lane [70]) over relatively large distances (Micron's HMC2.0 die is $34mm \times 34mm$[70]), and effectively implements a crossbar fabric. All of these attributes cause the switch to consume non-negligible power consumption and could potentially lead to a performance bottleneck. While CMOS circuitry in HMCs will scale power with more advanced technology nodes, electrical interconnects are notoriously limited in scalability. In fact, Zhan et al. [73] revealed that the logic layer consumes 67% of the energy in HMCs. Most of the energy consumption is in the switch and SERDES interface, and the minority in the actual DRAM layers.

As SiPh interconnects offer low-power, high-bandwidth, and distance-independent communication, they would represent an ideal candidate for an efficient on-chip crossbar switch implementation in HMCs as will be illustrated later in this dissertation. In particular, this dissertation advocates the use of SiPh AWGRs - a device that has seen significant technological improvements in recent years - as the interconnect fabric offering an efficient all-to-all interconnection network with multiple advantages over alternative photonic interconnects. In addition, optics can be used to perform wavelength routing which can be exploited to provide direct communication between the processor and the vaults in an MC, thereby avoiding electrical switch traversal and its latency and energy overheads. The properties of optics and AWGRs in com-

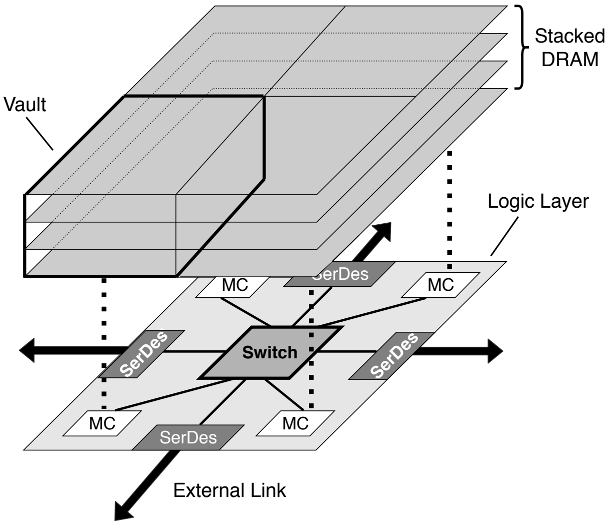


Figure 4.1: Structure of a Memory Cube

bination with a memory interface in the processor that can communicate directly with vaults provides both lower latency and energy, and a more efficient way to scale memory capacity.

4.1 Hybrid Memory Cubes

The design of Micron's HMC has inspired the adoption of 3D-stacked DRAM on top of a logic layer to build high-capacity computing nodes in many research studies [65][74]. Note that from this point on this approach is referred to as MC (for 'Memory Cube') as the general approach is adopted but not necessarily all the implementation detail of the HMC commercial design.

Figure 4.1 illustrates an example MC with four memory controllers for each DRAM partition ('vault') and four input ports. The number of vaults and input ports is configurable and can currently be anywhere between 4-32 and 2-8, respectively, effectively turning the switch on the logic layer into an intra-MC network, typically implemented as a high-radix switch.

While memory controllers are traditionally placed on the host processor to interface the DRAM, in MCs, they are 'outsourced' and placed on the logic layer–right underneath the vaults for close proximity and, in turn, power efficiency. The host processor has a memory interface (MI) through which memory requests are sent as network packets to the MC, which will be routed through the switch and to the corresponding memory (vault) controller through an abstract, packet-switched network interface. This approach not only provides an efficient architecture for processor-vault communications, but also supports efficient scaling of memory capacity which can be done by simply connecting several MCs into memory networks, possibly in different topologies [65][75].

The main competition of MCs are currently HBMs, both utilizing 3D-stacked DRAM; however, while MCs use a narrow, high-speed abstract interface for a packet-switched network, HBMs have a very wide, low-clocked IO interface based on low-level hardware signals and do not support an efficient way of extending system memory by chaining modules. Besides, the total number of HBMs directly connected to a host processor is limited by the pin count of the processor [66], severely limiting the total achievable memory capacity.

4.1.1 Design Challenges

Although this packet-switched network approach with an on-die switching fabric allows scaling memory capacity efficiently, recent simulation results revealed that switch traversals cause significant latency overheads and increase application execution time [65]. From a power perspective, high-radix switches–as the one on the logic layer inside an MC–are prone to high power consumption. Configurations with a large number of vaults and input links (e.g., 32 and 8, respectively) would put particular strain on the switch. In addition, memory cubes tend to be relatively large in size (a 4-link HMC has package dimensions of $34mm \times 34mm$ [76]) and vaults are evenly distributed across the chip, meaning that the switch needs to connect nodes at distances that result in considerable energy consumption on the interconnect. Given the poor scalability of electrical interconnects in terms of energy consumption compared to transistors, this issue is likely to become increasingly severe for more advanced technology nodes with smaller feature sizes.

In summary, a scalable, low-power, high-bandwidth switching fabric will sooner or later become essential to MCs to scale memory bandwidth and capacity within acceptable power envelopes. SiPh offer a particularly interesting design option for that purpose as optical communication overcomes many of the shortcomings of electrical interconnects, enable efficient crossbar fabrics, and could, therefore, be ideal to satisfy all the demands of the MC switch.

4.2 Silicon Photonics for Processor-to-MC Communication

Connecting the host processor to MCs with SiPh interconnects was shown to provide significant improvements over conventional electrical interconnects in terms of memory bandwidth, bandwidth-per-pin, and power consumption [74][77]. Those studies, however, still assume an electrical switch inside the MCs. Intra-MC networks (i.e., the switch on the logic layer of an MC) have, thus far, only been studied by Zhan et al. [73], who propose a unified electrical NoC architecture for both on- and off-chip traffic.

This study propose to re-architect the logic layer of an MC by using SiPh to connect a host processor directly to the vaults inside an MC through the help of wavelength-division multiplexing and a wavelength-selective switching fabric on the logic layer inside the MC. The following sections first present a summary of the related work, then introduce the AWGR-based proposal precluded by a discussion on its benefits over alternative SiPh switching fabrics, and followed by a comparison to conventional, electrical interconnects of currently deployed MCs.

4.3 Related Work

Re-architecting DRAM memory systems using SiPh has first been addressed by Beamer et al.[78], who proposed to utilize optical interconnects both for processor-to-DRAM links, as well as inside the memory for connecting DRAM cells. The higher pin bandwidth and low energy of optical links was shown to yield 10× lower power consumption compared to an electrical counterpart while enabling performance speed-up–laying the foundation for numerous studies in the years to come.

With the emergence of novel 3D-stacked memory technologies such as HBM and HMC, several studies were conducted to evaluate the role that SiPh could play in interconnecting such systems, particularly as the memory wall problem has only been postponed rather than solved by these technologies. Grani et al. [79] studied the use of AWGRs for interconnecting HBM modules with processor chiplets in a 2.5D integrated system. The AWGR was leveraged to provide all-to-all connectivity between all HBM and processors, offering an interesting approach of highly-efficient terabyte-scale computing nodes. However, the very nature of optics and the HBM interface seem rather impractical to be combined as HBMs have a low-frequency, wide

IO interface with 1024-bits at 1GHz, while optics are known for high-speed serial links. Both a large number of SERDES as well as a very large number of wavelengths were evaluated to interconnect HBMs optically, however, these approaches either lead to high energy consumption in the SERDES circuitry or large amounts of laser power due to high number of wavelengths, respectively.

HMC is based on a high-speed serial interface and thus more suitable to be interconnected with optics. MOCA [74] proposes a design that replaces the electrical links and SERDES circuitry with optical links and EO/OE interfaces to connect MCs with the processor, and report 3.4× higher energy efficiency with a 2.6× speed-up in execution time; however, MOCA does not change the architecture of the on-MC switch and thus still imposes the latency and energy overheads of electrical switch traversal and routing. Similar approaches as MOCA, just in different system configurations, have also underlined the benefits of using SiPh for optically-interconnecting memories with processors, some of which even with integrated laser sources and system demonstrations [80][81][82][83]. All these studies confirm the energy and performance benefits of optical processor-to-DRAM interconnects.

This dissertation proposes a new architectural approach that connects the processor directly to on-MC vaults through a compact, low-energy switching fabric (i.e. AWGR). This approach improves access latency and energy, reduces packet size, exploits the high-bandwidth density of optics, and could form the basis for systems with several MCs whose vaults could be access directly. Not only offering an approach to overcome the memory wall, but also enabling efficient extensions of system memory capacity.

4.4 Proposed Solution

Figure 4.2 shows the target system, which connects four memory cubes to a 64-core processor, with relevant SiPh technologies discussed in Chapter 2. The MI modules on the processor can directly communication with each individual vault inside the connected MC using WDM and an on-MC AWGR which distributes the wavelength(s) to the destination vault. Rather than having to encode the destination vault ID in the packet header to allow an electrical on-MC switch to perform routing, the MI modulates the memory request on the wavelength assigned to the vault

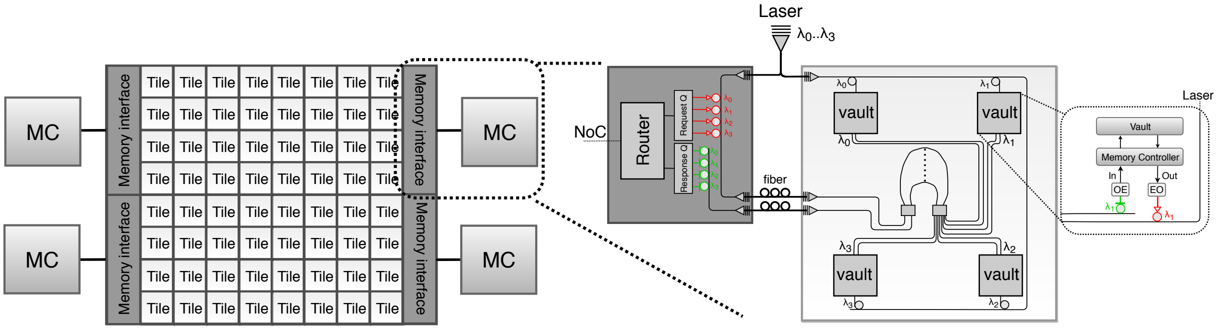
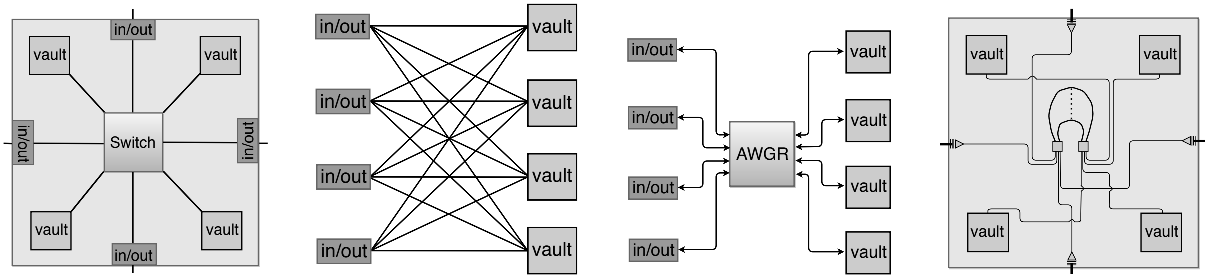


Figure 4.2: Target System. Memory interfaces on the host processor chip can directly communicate with vaults inside the MCs by modulating requests on the wavelengths assigned to the vaults. Similarly, vaults use their assigned wavelengths to modulate and send their responses back to the memory interface.

that holds the requested data, which will then automatically be forwarded to the vault by the AWGR. Similarly, each vault responses by modulating its packet on the wavelength assigned to it. Therefore, each MI must implement SiPh modulators and filters for each wavelength (per vault) to transmit and receive packets. Not only does this approach eliminate the need for encoding the vault ID and for traversing an electrical switch, it also allows the MI to communicate with each vault directly and independently, effectively boosting the memory bandwidth. Note that while in Figure 4.2, one off-chip laser provides the wavelength for both the MI and MC, alternative approaches are also feasible (e.g., one laser for each vault, or even on-chip lasers co-located with the vaults). In general, off-chip lasers are easier to thermally control, have higher maturity and efficiency, are easier to test or replace, and keep heat dissipated by the laser outside of the chip, leaving more of the power budget to compute and memory. However, making on-chip lasers more efficient and reliable is receiving much attention as they would allow for fast adaptive laser source mechanisms that have the potential to significantly reduce laser power consumption if the on-chip temperature is kept relatively moderate ($< 60^{\circ}$) [61]. For the current state of the SiPh technology, however, the former case is the more realistic one in the near future, especially due to the high chip temperature ($> 100^{\circ}$).

While Figure 4.2 illustrated the implementation of an AWGR inside an MC for one input/output port, Figure 4.3 provides a more detailed and general depiction of the interconnection requirements of an intra-MC NoC and how an AWGR inherently provides such a connectivity pattern–for any number of input/output ports.



(a) MC Switch

(b) Connectivity

(c) Connectivity-AWGR

(d) AWGR inside MC

Figure 4.3: Implementing the interconnect fabric inside a memory cube with an AWGR

Figure 4.3a shows the basic layout of an MC in which the input ports are connected to the switch, which, in turn, is responsible to route data packets to the vault in which the address of the packet resides. Similarly, responses from the vault must be routed through the switch to the output port, which means that each input port must have a connection to each vault, and each vault must have a connection to each output port with a connectivity pattern shown in Figure 4.3b.

The wavelength-distribution functionality of the AWGR and its bidirectional behavior make it ideal for this connectivity pattern, which can be enabled by simply connecting all input and output ports on one side, and all vaults on the other side of the AWGR (as shown in Figure 4.3c). Figure 4.3d shows the final physical implementation of such an arrangement, with the AWGR placed in the middle of the chip, which offers short, direct links between all nodes and a negligible number of waveguide crossings.

Despite its benefits, AWGRs are not ideal for point to point multi-wavelength transmission - a crucial mechanism to increase bandwidth in optical interconnects - due to their inherent switching matrix that distributes *one* wavelength to each output port. The only two options are either to use an AWGR with a higher port count and assign multiple ports to one node, or to use multiple AWGRs in parallel to attain the bit-level parallelism of WDM. Although the former approach one requires only one AWGR, it also needs a higher number of wavelengths (one for each port) which directly increases laser power consumption.

Utilizing multiple AWGR seems concerning with regard to area footprint at first, however, especially inside an MC, these concerns are uncritical: First, with an area footprint of $1mm^2$, multiple AWGRs can conveniently fit on a MC die. Second, AWGRs do not need to be placed next to each other; in fact, fabricating AWGRs on different SiPh layers on top of each other can be done efficiently. This reduces the area footprint to roughly one AWGR in x and y dimension, and since the AWGR is a SiPh device that does not require heating, adding multiple devices is no concern with regard to the thermal design power. Finally, the bit-level parallelism required to access vaults in a MC is manageable with relatively low levels of WDM (e.g. Micron's HMC has a vault data rate of 80 Gb/s [70]). With recent advances of energy-efficient PAM4 modulators that were demonstrated to require just 0.685 pJ/bit at an effective data rate of 40 Gb/s in a 45nm technology node [15], merely two wavelengths (and in turn two AWGRs) are required to satisfy an MCs bandwidth demands.

4.4.0.1 Benefits of Optical Intra- and Inter-MC Communication

The main benefits of using optical interconnects for processor-DRAM communications have been revealed by both simulation studies [74] as well as through the implementation of a many-core chip fabricated with SiPh [84]: high-speed low-energy optical communication without the need for repeater circuitry, high bandwidth density with lower pin counts, and direct inter-chip communication. Therefore, the need for an electrical switch inside an MC is further eliminated by directly connecting memory interface modules with the on-MC vaults through WDM and AWGRs. This not only erases the need for switch traversal, which should reduce both latency and energy, but also shrinks the packet header size since fields such as vault ID are not needed anymore, further reducing bandwidth requirements and energy consumption. This architectural approach is expected to improve the state of the art of MC design in terms of latency and energy, and form the basis for future high-capacity memory modules that can be accessed with low latency and energy overheads.

4.5 Evaluation

4.5.1 Methodology

This section compares performance and power consumption of the optically-interconnected MC (OMC) target design in Figure 4.2 with a legacy electrically-interconnected implementation (EMC) for 4, 8, and 16 vault configurations of the MC. In addition, different SiPh interconnection fabrics, i.e. SWSRs and MWSRs, are compared with the AWGR implementation introduced in the previous section to find the most efficient SiPh fabric for processor-vault-communications.

The target system shown in Figure 4.2 has four memory channels, MCs with 8Gb capacity, 64 tiles with a 32kB private L1(I/D) and 256kB L2 cache. Tiles are connected through a regular 2D mesh with 128-bit flits and two cycle hop latency at a clock frequency of 2 GHz. For power estimations of the processor-to-MC interconnect DSENT [13] with a 22nm technology node is used. For different parameters in SiPh components, 20 μW for MR heating power, 20% laser efficiency, 1dB coupler loss, 0.5dB MR-drop loss, 0.01dB MR-through loss, 0.1 dB/mm waveguide propagation loss, 0.1dB power splitter loss, and -21dBm receiver sensitivity [85][10][49] is assumed with an electrical link traversal between the processor and MC consuming 5 pJ/bit and DRAM Read/Write 12 pJ/bit [86]. Electrical link traversal take 2.5ns over 4 inches of electrical strip [36], SERDES traversal of 2ns, on-MC electrical switch traversal of 2ns, and vault access times of tRDC=12ns, tCL=6ns, tRP=14ns, tRAS=33ns [65]. For EMC, 16-bit links running at a frequency of 15Gbps is implemented (according to HMC2.0 [86]). For the OMC, 80 Gb/s links to and from each vault on the MC [86] is implemented, with an optical signal with 2 wavelengths at 20Gbps with PAM-4 modulation (resulting in $2 \times 2 \times 20 = 80$ Gb/s).

Simulation experiments are based on sniper simulator [45] with a range of high-performance workloads from the PARSEC [87] and SPLASH-2 [88] benchmark suites with *sim-large* input sets.

4.5.2 Benchmarking Results

Figure 4.4 shows the average DRAM access latency throughout the execution of different applications.

DRAM access latency was measured from the time a memory request is issued by the MI on the host processor side until the response is received from the MC. OMC reduces the average DRAM access latency for every workload, and tends to become more efficient as the number of vaults is increased. This further underlines that link and on-MC switch traversal has a noticeable impact on the total DRAM access latency, and that direct optical connections to the vaults through integrated SiPh links largely reduce this latency. Figure 4.5 shows how these

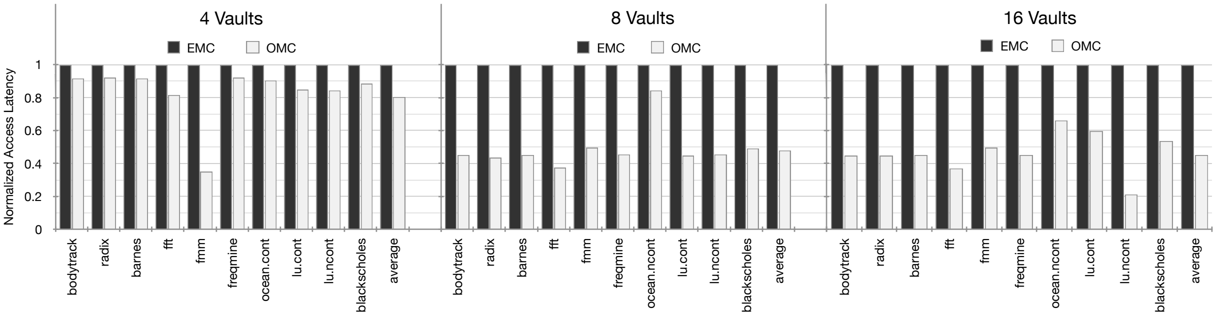


Figure 4.4: DRAM Access Latency

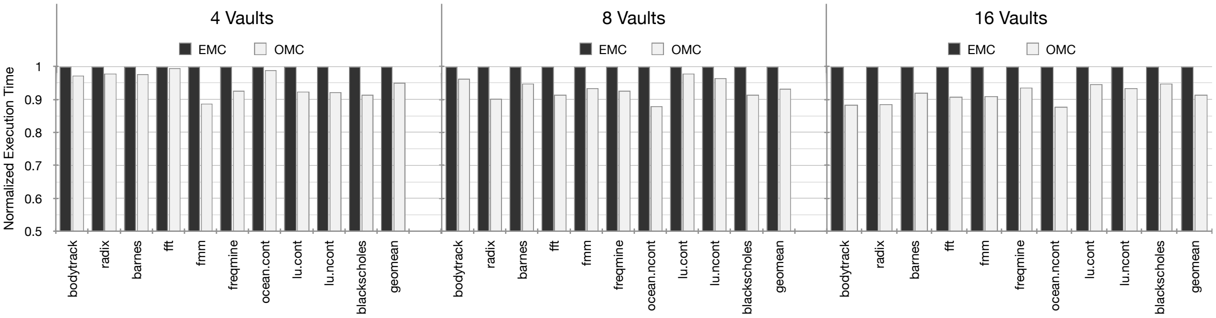


Figure 4.5: Application Execution Time

performance benefits translate to the total application execution time for these HPC workloads. Benchmarking results show speed-up gained for every application, and on average a speed-up of 5%, 7%, and 9% for 4, 8, and 16 vaults, respectively. The reason why the large reductions in DRAM access latency do not translate into similar application speed-up likely stems from the fact that HPC workloads from the PARSEC and SPLASH-2 benchmarks do not stress the off-chip bandwidth too much. For other workloads, such as GPU workloads, the latency and bandwidth benefits attained through the OMC might lead to much more significant overall performance gains.

Figure 4.6 shows the simulations results of the average DRAM access energy consumption, which includes processor-to-DRAM link traversal (including EO (electrical-to-optical) and OE (optical-to-electrical) backends and SERDES), on-MC switch traversal, laser and MR heating power for OMC, and DRAM read/write.

Exploiting distance-independent low-energy SiPh interconnects in combination with the elimination of switch traversal results in significant reductions in DRAM access energy for each application, and on average at least by 40%. This is highly encouraging and can have a

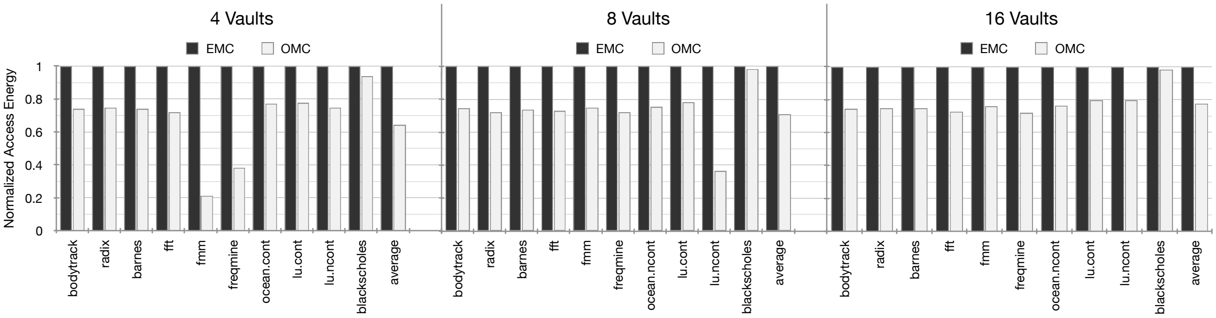


Figure 4.6: DRAM access energy of OMC in relation to EMC

considerable impact on the overall system efficiency, particularly for memory-intensive workloads or application domains in which DRAM energy takes up the majority of the power budget. In server or data center environments in which processor chips with larger number of nodes and memory channels are deployed, these energy savings can be decisive, especially as they come with superior performance.

4.5.3 AWGR vs. Alternative SiPh Interconnects

The previous section has shown that OMCs offer superior energy consumption and performance compared to legacy electrically-interconnected MCs; however, AWGRs are not the only SiPh switching fabric candidate to enable direct processor-vault communication. Therefore, the power consumption and SiPh properties of two alternative switching fabrics that deserve legitimate consideration for this task - SWSR, SWMR, and MWSR buses - are discussed. Note that wavelength-routed optical NoCs (WRONoCs) based on MR switches and wavelength-selective routing could also be a candidate, however, a dedicated switching topology for the fairly low number of nodes in processor-to-MC networks is unnecessary as buses or AWGRs can easily provide the required connectivity without suffering from issues regarding scalability and power consumption. In addition, as opposed to buses and AWGRs, WRONoCs need MRs for switching, all of which must be heated, resulting in additional MR heating power on the MC die–an undesired side effect as thermal behaviors and heat dissipation is already a sensible factor in 3D-stacked chips. Therefore, WRONoCs are not considered in this study.

Table 4.1 lists the maximum optical path losses (IL_{max}), the required number of wavelengths, the laser power for each processor-MC-network, and the total laser power consumption for the target system. SWMR buses have one sender and multiple receivers on one waveguide,

		Assumption: 0.1 dB/mm waveguide loss					
		Processor-to-MC			MC-to-Processor		
	V	AWGR	SWSR	SWMR	AWGR	SWSR	MWSR
ILmax (dB)	4	19.13	25.26	18.1	25.5	25.6	19.1
	8	19.9	28.4	18.9	26.1	29.9	20.7
	16	20.8	33.1	19.9	27	33.1	21.8
Num. λ	4	4	2	8	4	2	8
	8	8	2	16	8	2	16
	16	16	2	32	16	2	32
Laser Power (mW)	4	2.6	5.33	5.13	11	11.4	10.45
	8	5.2	11.9	14.8	22.3	15.7	20.9
	16	10.2	22.23	19.4	33.6	32.2	23.8

	V	AWGR	SWSR	SWMR/MWSR
Laser Power (mW) Total (Normalized)	4	13.6 (1)	16.7 (1.23)	15.6 (1.15)
	8	27.5 (1)	27.6 (1.01)	35.7 (1.3)
	16	43.8 (1)	54.4 (1.24)	43.2 (0.99)

		Assumption: 0.027 dB/mm waveguide loss					
		Processor-to-MC			MC-to-Processor		
	V	AWGR	SWSR	SWMR	AWGR	SWSR	MWSR
ILmax (dB)	4	18.4	19.3	13.7	16.4	21.6	12.8
	8	18.9	22.5	14.1	16.9	22.9	13.9
	16	19.5	25.3	14.7	17.4	26.1	14.1
Num. λ	4	4	2	8	4	2	8
	8	8	2	16	8	2	16
	16	16	2	32	16	2	32
Laser Power (mW)	4	2.22	2.27	2.44	1.34	1.35	1.49
	8	4.47	4.12	4.5	2.78	3.1	2.8
	16	5.4	6.41	6.6	4.8	5.44	4.74

	V	AWGR	SWSR	SWMR/MWSR
Laser Power (mW) Total (Normalized)	4	3.56 (1)	3.62 (1.02)	3.93 (1.1)
	8	7.25 (1)	7.22 (1)	7.3 (1.01)
	16	10.2 (1)	11.85 (1.16)	11.34 (1.05)

Table 4.1: SiPh interconnect fabric comparisons for total optical path losses (IL_{max}), number of wavelengths (Num. λ) and laser power (LP). Parameter V represents the number of vaults.

and connects the memory interface (MI) to the four vaults. Similarly, each vault is connected to a waveguide which connects them to the MI on the processor. Therefore, assuming two wavelengths at 40Gbps, each SWMR/MWSR bus needs ($numberofvaults \times 2$)-wavelengths to implement a non-blocking crossbar. In the SWSR bus case, each MI and vaults communicate on a dedicate waveguide, thus requiring only two different wavelengths at the laser source. The AWGR, as discussed in the previous sections, needs one wavelength for each output port (i.e. vault), and a DWDM level of two is implemented with two stacked AWGR, leading to a total of four different wavelengths at the laser.

Throughout this study, it has been realized that the main benefit of AWGR compared to point-to=point optical bus-based designs is that it provides short, direct links between all nodes without additional losses incurred by waveguide crossings, etc., whereas buses need a U-shaped layout to allow all nodes to connect to it, leading to higher path lengths and optical losses. Since the AWGR itself also incurs losses, its advantage in terms of power consumption depends on the path lengths needed by bus-based designs and waveguide propagation loss (WG loss) per mm. While the former depends on the size of the MC and the distance and number of the vaults, the latter is technology dependent with recent studies assuming a wide range of different WG loss values (mainly due to the large number of different waveguide designs, materials, and fabrication optimizations). To obtain insightful results, the AWGR implementation is compared with both the most aggressive WG loss, as well as moderate WG loss. Therefore, 0.1 dB/mm WG loss (moderate) [10] and 0.027 dB/mm WG loss [89] (the lowest reported value reported in the scientific literature to date) are considered for power calculations. It should be noted that the grating waveguides inside the AWGR only have a very small impact on the total AWGR loss, and the vast majority of the loss (~90%) inside an AWGR comes from the free space propagation slab. Therefore, a comparison between AWGRs and aggressive, varying WG loss values is still fair, although leading to a slightly pessimistic assumption of the AWGR loss.

In both WG loss cases, the AWGR does not impose any noticeable laser power overheads for all number of vault configuration. In fact, in most cases, laser power is saved, up to 30% compared to the SWMR case for 8 vaults. Lower WG loss has a more significant impact on the bus-based designs as the majority of the loss is WG loss. The laser power savings of the

AWGR observed for the moderate WG loss case are therefore lower in the aggressive WG loss case. Nevertheless, even in the most aggressive case, the AWGR is still the most power-efficient design for any number of vaults.

Finally, MR heating power is the same in all approaches since they are only used for modulation and detection, the same number of which is needed in each interconnection fabric. For two-wavelength DWDM, two modulators and receivers are needed for each source-destination pair, i.e. 2×(1 + number of vaults). Besides, inside the MC, 2 additional MRs are needed for light distribution at each vault. For every memory channel in the target design, this leads to 40, 80, and 160 MRs for 4, 8, 16 number of vaults, respectively. With a total of 4 memory channels and 20 μW/MR heating power, this leads to 3.2mW, 6.4mW, and 12.8mW, total power consumption respectively. Depending on the assumed WG loss and number of vaults, heating is within 30-50% of the total SiPh power consumption. In absolute terms, however, both laser and MR heating power are very low as novel PAM-4 modulation is leveraged to reduce the number of MRs and wavelengths, and utilize AWGRs - a switching fabric that does not need heating.

4.6 Conclusion and Future Work

The hypothesis assumed earlier that avoiding electrical switch traversals inside MCs by exploiting SiPh links with wavelength routing to save energy and latency was confirmed by the simulation results. In particular, enabling direct communication between the host processor and the vaults by modulating data on the correct wavelengths eliminates the need for routing decisions inside the MC switch, reduces packet sizes, and provides overall higher bandwidth. In addition, both laser and MR heating power are at low levels even for 16 vaults inside an MC, which might pave the way for large memory modules with a high number of vaults, all of which can be accessed directly with low-latency and low-energy links. Alternatively, as shown in Figure 4.3, the wavelength-switching AWGR can be used to connect multiple MCs together optically to extend memory capacity, which would allow the host processor to directly address vaults in different MCs by modulating data on the corresponding wavelengths. Both of these opportunities will be part of the future work.

Although simulation results already underline the benefits of the OMC implementation, further improvements can be obtained by combining this approach with different SiPh technologies. For instance, when observing the memory access patterns of HPC workloads such as PARSEC and SPLASH-2, one can notice that the memory is accessed in short, bursty phases with high bandwidth demands and that the average memory utilization throughout the entire application execution time is actually low–especially when assuming systems with abundant on-chip caches that can easily house the entire application data sets. Therefore, adaptive lasers that can be switched on/off based on the current communication demands could be very efficient to lower the laser power consumption. A caveat with this approach is, however, that the laser is on-chip as controlling an off-chip lasers entails considerable energy and latency overheads. Despite the currently low manufacturing yield of on-chip lasers and their high susceptibility of thermal variations, technological advances would surely make a strong case for using adaptive on-chip lasers for such memory access patterns.

The study presented in this chapter is published in the international symposium on memory systems (MEMSYS) [90]. Based on presented results, other, more memory-intense applications such as GPU workloads would greatly benefit from the OMC architecture as the reductions in latency and energy of a memory access gains more significance the higher the memory utilization becomes. In fact, previous studies have made the case that the low energy, low latency, and high bandwidth attributes of optical processor-to-DRAM links could be exploited to reduce the amount of on-chip caches and dedicate more of the on-chip real estate to the compute [79]. This would change the way we approach the design of future computer architectures.

By exploiting SiPh for integrated optical links, a compelling case for the suitability of AWGRs as a layout-efficient, mature, and highly energy-efficient interconnection fabric between the processor and the vaults inside the MC is made. Wavelength-routing through the AWGR allows the processor to directly communicate with the vaults without having to route packets through the on-MC electrical switch. Simulation results for HPC workloads show that both DRAM access latency and energy can be reduced significantly through this approach. In addition, memory networks with higher number of vaults and implementations with several MCs per memory channel could efficiently be supported by this approach, thereby largely alleviating the memory wall problem.

Chapter 5

Scalable Chiplet-based Uniform Memory Architectures with Silicon Photonics

Growing data sets in modern workloads are driving the need for higher processing power and memory bandwidth/capacity in HPC systems. Unfortunately, the slowing down of Moore's law reduces the ability to attain higher processing power and memory capacity within a single compute node ("scale-up") for each generation of silicon technology, forcing system designers to add compute nodes (”scale-out”) to satisfy performance demands. Scaling-out, however, leads to distributed memory architectures with compute nodes operating in different address spaces and explicitly communicating and managing coherence through software (e.g. message passing), thereby causing significant performance and energy overheads for data movement between compute nodes–a key challenge in current HPC systems. Scaling-up shared memory architectures within a single address space, on the other hand, allows hardware-managed coherence which is significantly faster and allows programmers to focus on what matters for parallel speed-ups rather than communication and synchronization. Moreover, the cost of providing hardware coherence (i.e., traffic, storage cost for tracking sharers, latency, and energy) is generally considered to scale gracefully with the core count for hierarchies in modern systems [3].

One significant bottleneck to further scaling-up processing power within tight power envelopes are the growing monolithic silicon development and manufacturing costs which have seen a 7× increase from 28nm to 7nm and have lead companies to increasingly rely on breaking monolithic chips into smaller "chiplets" [91]. Utilizing several smaller chiplets assembled using advanced packaging technologies instead of one large monolithic chip reduces costs by exploiting the higher yield of smaller dies at low performance and energy overheads through tight integration in the same package [21]. In addition, it allows freedom of mixing and matching the most cost- and power-efficient process nodes for chiplets, particular for those with harder-to-shrink or purpose-built components, thereby representing a highly promising technique to support the trend of increasingly heterogeneous computing systems.

Looking ahead, we can expect future systems to further exploit the cost benefits of chiplets to scale-up performance at acceptable power and cost by increasing the total number of chiplets within a single package–a trend observed in several recent commercial designs [27, 92, 93]; however, to enable further scaling of chiplet-based systems, several key challenges must be addressed:

1. Interconnection Challenge. The dependency of energy consumption on interconnect length coupled with centimeter-scale chiplet sizes only allows interconnecting adjacent chiplets (on a planar layout) without excessive crosstalk and energy overheads. This leads to low-radix/high-diameter topologies with high average hop counts in which each inter-chiplet hop imposes tens of nanoseconds latency [26]. Given these latency overheads, inter-chiplet communication in general, but cache coherence traffic (often requiring multi-hop coherence protocols) in particular, can now significantly degrade system performance and thereby limit scalability.

2. NUMA Challenge. Current designs with low numbers of chiplets are already designed as non-uniform memory architectures (NUMA) exhibiting significant variances in access latency to different addresses [26]. NUMA systems are notoriously difficult to program, making it extremely challenging for programmers to extract performance. Energy and performance limitations on the interconnect exacerbates this issue by making scalable chiplet-based UMAs difficult to attain with acceptable energy efficiency. Recent solutions aiming to make chiplet-based system designs more unified by implementing a central IO chip through which all traffic passes through are temporary remedies [92]. However, the bottleneck for performance and power will ultimately be the switching fabric on the IO die.

3. Disintegration Limits. The interconnect challenge limit not only the scalability of chiplet-based systems, but also how much processor disintegration (*i.e.*, the process of breaking one large many-core processor into smaller ones) can be exploited. As a result, designers will be forced to opt for few large chiplets rather than many small ones due to unacceptable latency overheads on the interconnect, despite it incurring much more manufacturing costs and limiting the freedom of choosing the most suitable process node.

4. Packaging Challenge. Accommodating larger numbers of chiplets in the same package requires larger substrates. Silicon interposers offer high IO density but are too expensive for the system sizes in current HPC nodes which mainly use less expensive organic substrates with lower IO densities [94]. However, IO density is crucial to satisfy future bandwidth demands of inter-chiplet links. Silicon bridges [95] integrated into organic substrates connect the edges of tightly-coupled chiplets with very high IO density, but can only connect physically-adjacent chiplets. Clearly, interconnects with high IO density and energy-efficient signaling over long distances integrated into a cost-efficient organic package substrate are in high demand.

Integrated optical links enabled by Silicon Photonics (SiPhs) provide low latency, high bandwidth density through wavelength-division multiplexing (WDM), and distance-independent energy consumption, and can now be integrated on organic package substrates [94], making them an attractive technology for inter-chiplet connectivity. Moreover, SiPh devices can exploit WDM to perform wavelength-selective routing which allows a chiplet to connect to multiple other chiplets through a single optical IO pin (addressing them on different wavelengths), enabling high-radix low-diameter networks. Optical networks provide sufficient scalability in terms of crosstalk and power consumption to enable point-to-point connectivity between up to 32 nodes with bisection bandwidths matching current chiplet-based systems [56]–a design option infeasible with electrical interconnects due to large amounts of wiring, distance-dependent energy consumption, and IO pin requirements.

This chapter proposes a novel chiplet-based scalable UMA (**S-UMA**) exploiting SiPh interconnects to solve all above mentioned challenges, and makes the following architectural **contributions**:

- A scalable and compact SiPh point-to-point interconnection fabric integrated on an organic package enabling a chiplet-based uniform memory architecture with distance independent energy consumption and latency and low pin IO requirements by exploiting

WDM and wavelength-selective routing.

- Dis- and re-integration of large LLCs from processor to separate chiplets containing LLCs, directory and memory controllers to enable lower leakage power of optimized processes for SRAM, cost reductions through higher manufacturing yield of smaller chiplets, and more flexibility in implementing heterogeneous memory technologies with less integration complexity. The high-bandwidth SiPh point-to-point interconnection between processor and LLC chiplets amortize any off-chiplet LLC access latency overheads.

- Integration of SiPh transceivers (TRXs) fabricated on separate dies rather than monolithically integrated on the chiplets to enable the freedom of choosing the most appropriate processes, prevent the reduced yield of SiPh components to decrease the yield of processor chiplets, and to remove potential area concerns of μm-scale SiPh components. Silicon bridges tightly connect TRXs and chiplets with very high IO density at energy and latency similar to on-chip wires, causing negligible overheads.

While a few previous works have studied SiPh for chip-to-chip communication on-board [36, 96] or interposers [56], this study is, to the best of our knowledge, the first to evaluate the energy and performance benefits of a point-to-point SiPh interconnect and its ability to provide a scalable chiplet-based UMA system integrated on an organic package substrate. Evaluation results show that S-UMA provides a speed-up of 23% and reduces network power consumption by 30% compared to state-of-the-art chiplet-based NUMA and provide a cost-efficient, practical way for scalable chiplet-based UMA systems. Therefore, S-UMA could be a key enabler with long-term impact for the future of scaling-up processing power of HPC nodes that exploit the benefits of chiplet-based systems: high energy efficiency through tight, heterogeneous integration and reduced cost through processor disintegration.

5.1 Chiplet-based Systems: Challenges and Opportunities

Chiplet-based systems integrate and interconnect several–possibly heterogeneous–processor and/or memory dies in the same package. Figure 5.1 illustrates an example layout of such a system with CPUs, GPUs, FPGAs, and an IO die for connecting chiplets with each other and

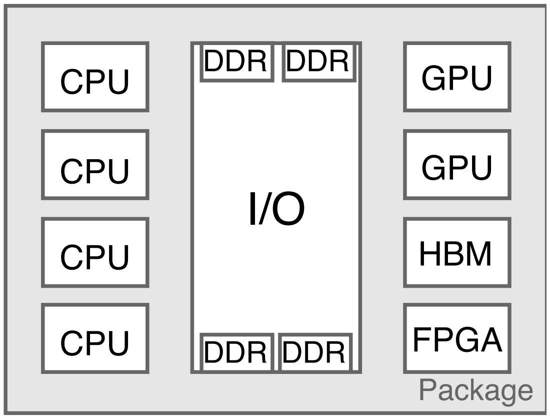


Figure 5.1: An example chiplet-based system with heterogeneous processor dies integrated in the same package

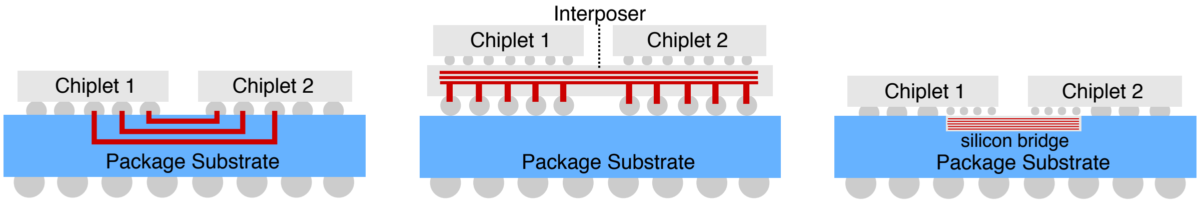
to DDR memory interfaces. This chapter reviews and analyzes the state-of-the-art packaging and integration technologies as they are a key design factor with significant implications on manufacturing cost, interconnect density and energy, and thereby on the scalability and future outlook of chiplet-based systems.

5.1.1 Packaging and Interconnect Technologies

Various chiplet-based system architectures from different vendors have emerged in recent years based on different packaging technologies, each with different benefits and trade-offs regarding cost, performance, energy, and scalability, as well as implications on the memory subsystem, interconnection network, and overall system design. Figure 5.2 illustrates the state-of-the-art techniques deployed in current commercial designs: 1) Multi-Chip Modules (MCMs) [27, 92, 97–99]; 2) 2.5D integration with a silicon interposer [20–25, 100]; and 3) Silicon Bridges (like in Intel's Embedded Multi-Die Interconnect Bridge (EMIB) technology) [95, 101]. Table 5.1 provides a summary and comparison of each integration technique.

5.1.1.1 Multi-Chip Modules (MCMs)

MCMs (Figure 5.2a) mount and connect chiplets with high-density interconnects (HDIs) on the package substrate using wire-bond or flip-chip technology [94]. MCMs typically utilize organic package substrates as these are not manufactured in the foundry (as opposed to Silicon interposers) and therefore much cheaper. In addition, no further processing steps are needed (e.g., 2.5D integration needs additional processing step for the vertical interconnects), making



(a) Multi-Chip Module (b) Silicon Interposer (c) Silicon Bridge

Figure 5.2: State-of-the-art Integration Technologies for Chiplet-to-chiplet Interconnection

MCMs the cheapest option from both a material and processing cost perspective, and especially attractive for systems of larger scale. For instance, packages of MCMs deployed in current HPC nodes can be around $10cm \times 10cm$ in size [92, 97, 102]).

Challenges. Wire-bond or flip-chip interconnects offer relatively low IO pin densities, thereby restricting off-chip(let) bandwidth. High IO density, however, is crucial as chips are facing a ”pin wall” where the vast majority of pins is dedicated to power/ground, leaving very few pins to satisfy off-chip communication demands [66]. Although current electrical interconnect technologies on organic substrates (20Gb/s operation per IO pin at 0.54pJ/bit over 4.5mm at 28nm [103]) appear to satisfy the bandwidth demands of current systems [27], higher pin data rates are difficult to attain due to excessive crosstalk of electrical signaling.

In addition, high IO pitches can also restrict the number of chiplets a chiplet can be connected to. From a network perspective, this leads to low-radix chiplets requiring networks with high diameters and average hop counts. Inter-chiplet hop latencies have a large impact on system performance (more than 30ns per hop in current systems [26, 27]) and lead to complex NUMA systems with high latencies variations. Topologies can exhibit lower diameters with the same radix by connecting distant nodes; however, energy grows linearly with distance for electrical links, making this approach infeasible for the dimensions in chiplet-based systems.

Therefore, although attractive from a cost perspective, the bandwidth and distance related challenges of MCMs severely limit their ability to satisfy the performance demands of future chiplet-based systems of larger scale.

5.1.1.2 2.5D Integration with Silicon (Si) Interposers

2.5D integration (Figure 5.2b) places an additional silicon die on top of the package substrate, and the chiplets on top of the interposer. Chiplets connect to each other and to the package substrate through the interposer with through-silicon vias (TSVs) and μbumps. Interposers can be passive (interconnects only) or active (interconnects and logic) [104]. The main benefit of 2.5D integration is the substantially higher interconnection density compared to MCMs [105–107], either allowing for higher maximum bandwidth or for lower energy per bit by reducing the data rates of the IO transceivers.

Challenges. 2.5D integration with Si interposers overcomes the challenges of MCMs by offering higher IO pin density through smaller pitches of μbumps and TSVs [108–110]; however, Si is significantly more expensive than organic substrates and 2.5D integration requires additional (and more complex) processing steps. Material costs could be somewhat amortized through high-volume manufacturing, but the very-large size needed for large-scale chiplet-based systems in the HPC domain would still make Si economically unreasonable.

Although solving the IO density issues of MCMs, 2.5D integration cannot overcome the limitations imposed by interconnect length. High IO density can enable higher-radix switches on the chiplets (i.e., connect each chiplet to more other chiplets), thereby reducing network diameter; however, chiplets are relatively large ($\sim 1cm^2$) and are laid out on 2D planar floorplan, meaning that connecting to chiplets that are not directly adjacent requires to route electrical interconnects over cm-scale distances, which will require repeater and buffer circuitry that lead to very high energy per bit, especially for high-speed links. Acceptable energy can thus only be provided on links connecting to adjacent chiplets, thereby also limiting the radix and its impact on the network diameter.

5.1.1.3 Silicon bridges

Si bridges, like Intel's EMIB technology (Figure 5.2c), aim to solve the limitations of both MCM (poor interconnection density) and 2.5D integration (high cost for Si interposer) by embedding small and thin (less than $75\mu m$ [95]) Si chips (”bridges”) with (currently) four metal layers into an organic package substrate to interconnect the edges of adjacent chiplets. Si bridges offer very high IO density with latency and energy metrics similar to on-chip wires (by integrating fine-pitched “back end of line” (BEOL) interconnects) and enable short interconnects through tight packaging with just 100 μm between chiplets. Si bridges thereby offer a more scalable solution by combining the low material costs of organic substrates with the high

	MCM	2.5D Integration	Silicon Bridge	SiPh MCM
Materials	Organic substrate	Si interposer	Organic substrate	Organic substrate
Material Cost	$	$$$	$$	$
Pin BW[1]	20Gbps	28Gbps	28Gbps	160-640Gbps[2]
Pin Pitch[3]	6μm	2μm	2μm	5μm
pJ/bit/Gbps[4]	0.027 (4.5mm)	0.0114 (3.5mm)	<0.035 (1mm)	0.017 (several cm)

Table 5.1: Summary and properties of state-of-the-art integration techniques. Note that these values vary depending on the technology node, interconnect length, and optimized integration approaches; however, general trends and physical limitations stated in this table hold true nevertheless.

IO bandwidth density of Si interposers.

Challenges. Just like Si interposers and MCMs, Si bridges utilize electrical interconnects and thus impose the same distance-related energy limitations, and thereby the same network radix/diameter problem. Consequently, Si bridges alone cannot overcome the NUMA, interconnect, and scalability challenges in chiplet-based computing systems. Systems with one large chiplet and several (much) smaller chiplets could exploit the high IO density of Si bridges to directly connect the large chiplet to each small chiplet, but inter-chiplet traffic would likely be bottlenecked by the crossbar on the large chiplet, limiting the scalability of this approach. In fact, the vast majority of current designs utilize the batch processing and design re-use benefits processor disintegration (integrating smaller, replicated processor chips) [21, 24, 27, 97, 100], benefits that could be exploited even more in systems of larger scale.

5.1.2 Packaging: Implications on System Design

In summary, although each integration technology comes with its own benefits and trade-offs (as listed in Table 5.1), the following fundamental challenges limiting all of these state-of-the-art integration technologies from supporting the trend towards larger systems that further exploit the benefits of chiplet-based systems remain:

[1] Based on maximum reported pin data rates [27, 103, 111]

[2] Based on 40Gbps PAM-4 transceivers and 4-16 wavelengths per link [6, 9]

[3] Based on minimum reported μBump pitches [95, 110, 112]

[4] Based on recently reported and utilized interconnects [30, 101, 103, 113]

Interconnect Restrictions. Current integration technologies can enable sufficient IO bandwidth density but are distance-limited due to the energy consumption of electrical signaling and large chiplet dimensions which prevents direct connectivity to chiplets other than direct neighbors on the substrate. Therefore, future systems are restricted to topologies with relatively high average hop counts, with each hop incurring tens of nanoseconds in latency– significantly degrading system performance.

NUMA Challenge. The implications of the interconnect challenge on the average hop count and latency will cause substantially larger variances in memory access latency compared to current systems, which are already NUMA [26]. Extracting performance from NUMA systems is notoriously difficult for programmers, leaving much of the potential of the compute resources untapped. Moving closer to UMA (rather than in the opposite direction) is therefore critical, but is hard to attain in chiplet-based systems. Recent efforts making all chiplets communicate through an IO chip go in the right direction [92], but the limited scalability of a central interconnection fabric makes this only a temporary solution.

Scalability and Disintegration Limits. The limitations of electrical signaling over longer distances is one of the main reasons for the interconnect challenge and limits the scalability of future chiplet-based systems in which the large communication overheads may not be acceptable. The only option will then be to use fewer large chiplets to reduce off-chiplet communication and network diameter (and, in turn, latency); however, this removes one of the main motivators for chiplet-based systems, namely the lower cost of the higher manufacturing yield of smaller chiplets and the freedom of choosing the most efficient process node. A scalable, low-latency interconnection fabric is therefore key to future cost reductions.

Packaging Challenges. The size of high-end chiplet-based systems is already large and will further increase in the future, making a inexpensive substrate material increasingly important and organic substrates the preferred solution. Those, however, can only provide sufficient IO bandwidth when combined with technologies like silicon bridges which are based on electrical signaling and thus distance-limited. Ideally, one would desire an interconnect technology on an organic substrate that overcomes the distance-related energy overheads of electrical interconnects, while providing high IO pin bandwidth.

5.1.3 Opportunities with Silicon Photonics

Integrated optical interconnects enabled by SiPhs offer properties that can be exploited to overcome all of the previous challenges of electrical interconnects. Optics offer virtually distance-independent energy consumption, near speed-of-light signal propagation latency, and high bandwidth density through wavelength-division multiplexing (WDM) which enables to transmits on multiple wavelengths in parallel inside the same optical link. Moreover, SiPh devices can perform wavelength-selective routing, i.e., data is routed based on the wavelength channel, which allows to connect a chiplet to multiple other chiplets through a single waveguide. In addition, SiPhs can be integrated on organic package substrates, which allows to overcome the interconnection challenges while enabling the use of a relatively cheap packaging substrate (compared to Si interposer). In particular, SiPhs can be used to solve the challenges of chiplet-based systems as follows:

Solving the Interconnect Challenge. WDM not only enables high IO pin bandwidth but also allows to communicate with several chiplets through the same pin (on different wavelengths). Moreover, energy consumption in now distance-independent and signaling fast enough to reach each chiplet on a package in less than a nanosecond (depending on the material, a few hundred ps/cm [7, 10, 79]). This solves any IO limitations in terms of communication bandwidth and enables low-latency low-diameter/high-radix (and even point-to-point depending on scale) networks.

Solving the NUMA Challenge. A scalable low-latency low-diameter network could allow large-scale chiplet-based systems to become UMA, thereby significantly facilitating programmability and, in turn, paving the way for easier and more efficient extraction of performance from the available computing resources.

Enabling Scalability and Further Disintegration. In addition to exploiting the latency, bandwidth density, and energy properties of SiPh interconnects and their integration on inexpensive organic substrates, their superior scalability (compared to electrical interconnects) can be used to further support disintegration of chiplets into even smaller chiplets. This not only improves manufacturing yield of chiplets, but also opens-up new opportunities for dis- and re-integration. For instance, large LLCs could be moved on separate chiplets and either be manufactured on a more efficient process for SRAM for lower leakage power or facilitate the integration of alternative memory technologies like non-volatile STT-RAM [114].

Enabling Low-cost Package Substrates. SiPh interconnects can be integrated on organic package substrate, thereby offering high pin IO bandwidth density without requiring expensive Si interposers or being restricted to short-distance communication.

It should be noted, however, that SiPh come with their own challenges, one being manufacturing yield which is currently lower than for CMOS devices. Besides, the majority of SiPh chips are currently processed on older technology nodes (mostly 45nm/65nm [115–117], 28nm has recently emerged [118]) leading to lower-volume fabrication compared to current 7nm or 14nm nodes for CMOS. This not only increases cost, but also makes it unreasonable to monolithically integrate chiplets with SiPh transceivers as the SiPh devices could render a correctly functioning chiplet defective. Moreover, thermal control of components has to be considered carefully to avoid malfunctioning of SiPh devices. In the following, this chapter discusses these issues in more detail and introduce practical solutions to largely overcome these challenges. In this context, a system architecture is proposed that utilizes the properties of SiPhs to design a high-performance, energy-efficient, scalable chiplet-based UMA system on a cheap organic substrate.

5.2 Scalable Chiplet-based Uniform Memory Architecture

Figure 5.3 depicts a high-level view of the proposed system architecture, which will be discussed in detail in the following, consisting of many-core processor chiplets, LLChiplets containing last level caches (LLCs), memory and directory controllers, SiPh transceiver chiplets (TRX), and a SiPh all-to-all interconnection die in the same package.

This design incorporates several techniques to address the interconnect, NUMA, disintegration, and packaging challenges outlined in the previous sections and has the following **goals**:

1. Utilize a scalable, low-latency, high-bandwidth, low-energy point-to-point SiPh interconnection fabric for inter-chiplet communication to enable a scalable chiplet-based uniform memory and cache architecture not attainable with electrical interconnects.

2. Enable dis- and re-integration of LLC structures from processor into separate chiplets to

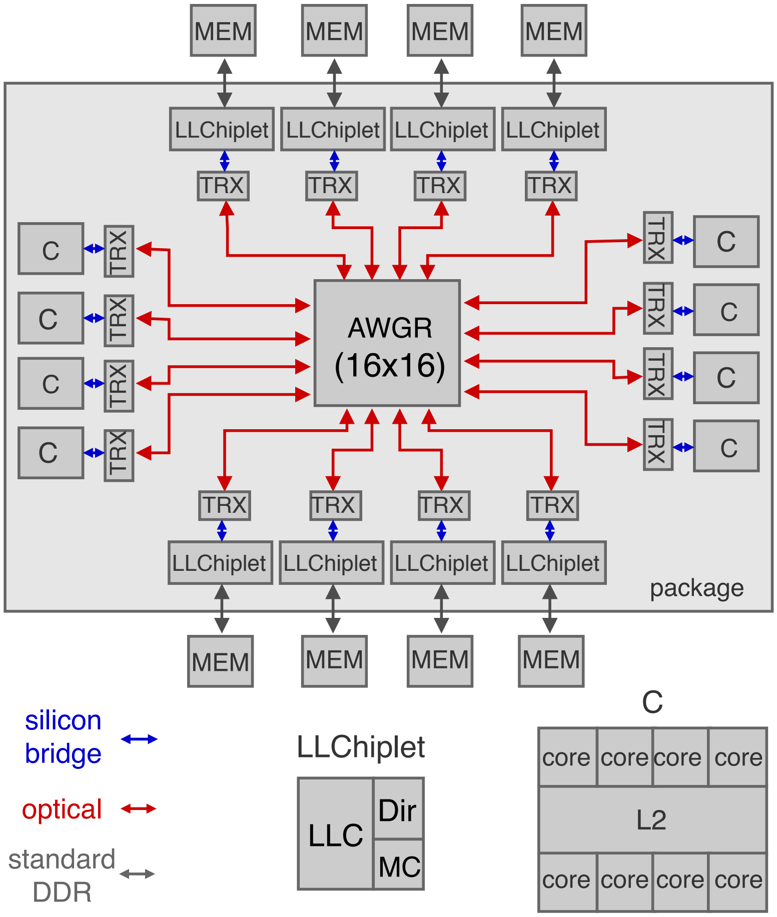


Figure 5.3: Target System (not to scale) with SiPh interconnection die (which is an AWGR point-to-point fabric), processor chiplets (C), LLChiplets (containing LLC, directory coherence controller (Dir) and memory controller (MC)), and SiPh transceiver chiplets (TRX).

reduce manufacturing cost and to provide more flexibility in choosing the most power-efficient process for the large LLCs.

3. Offer a practical, cost-efficient chiplet-based system architecture with advanced SiPh packaging to support the trend of higher numbers of chiplets inside the same package.

This section introduces the proposed chiplet-based system, in particular the new techniques proposed to target the challenges of scalable chiplet-based systems.

5.2.1 Addressing the Interconnect Challenge

Electrical interconnects suffer from high distance-dependent energy and latency overheads on relatively long inter-chiplet links and offer limited IO pin bandwidth. Therefore, connecting chiplets to not physically-adjacent chiplets is prohibitive in terms of pin availability and energy, leading to multi-hop networks with larger latency overheads. Integrated optical interconnects enabled by SiPh overcome these challenges by offering almost distance-independent energy, high pin bandwidth density with WDM, and wavelength-selective routing which enables compact point-to-point switching fabrics with optical links growing linearly with the number of nodes (rather than quadratically as in the electrical domain).

5.2.2 Addressing the NUMA Challenge

The overall performance in large-scale computing systems running workloads with ever-growing data sets extremely depends on the performance of the memory subsystem. In fact, to scale aggregate processing power, current system designs (e.g., Centaurs on IBM powers [119] or buffer chips in Oracle M series [120]) increasingly rely on NUMAs implementing ever-deepening memory hierarchies to achieve lower Average Memory Access Time (AMAT) compared to UMA systems. NUMAs offer lower AMATs as combining memory requests satisfied by ”local” memory (close to the processor with relatively low latency/energy) with ”remote” accesses (with significantly higher latency/energy) results in lower AMAT compared to UMA systems which have constant latencies across different locations in memory. The ratio of local and remote memory accesses highly depends on the problem size and as data sets keep growing, NUMAs should increasingly provide larger local memory capacity to achieve high performance.

Caches already occupy up to 40% of the chip area [92] which not only contributes significantly to power consumption but also adds complexities and inefficiencies to the fabrication process due to the different technological requirements of SRAM cells compared to the compute logic. Moreover, NUMAs also impact the programmability of systems by further emphasizing the importance of data locality to achieve higher performance [121].

UMA designs, on the other hand, lead to easier programmability but suffer from poor scalability as their AMAT is proportional to their size. A UMA with low average access latency could match the performance benefits of locality while providing the programming flexibility to fully exploit the compute power of the system independent from the problem size. Unfortunately, NUMAs emerged because UMAs have become prohibitively expensive for current system scales due to the energy and latency overheads of electrical interconnects.

S-UMA solves the NUMA challenge by utilizing a point-to-point SiPh fabric that enables uniform memory access without increasing AMAT compared to NUMA by enabling low-latency access to both local and remote memories. Moreover, SiPhs enable UMA with large bisection bandwidth without excessive energy overheads even for large physical distances, offering significantly better scalability than electrical interconnects and could thereby enable scalable chiplet-based UMA systems.

5.2.3 Addressing Disintegration Limits

In addition to exploiting SiPhs to implement a UMA system with a scalable low-latency high-bandwidth interconnection network, this chapter hypothesizes that the scalability of SiPhs can also be leveraged to fuel further dis- and re-integration of processor and memory chiplets. In particular, the proposed architecture aims to disintegrate the L3 (LLC) cache, directory coherence, and memory controllers from the processor chiplet and re-integrate them into a separate chiplet called *LLChiplet*. In this scenario, as depicted in Figure 5.3, S-UMA consists of the following compute and memory chiplets in the same package:

Processor chiplets accommodate multiple cores with their corresponding L1I/L1D and L2 caches along with IO circuitry for inter-chiplet communication.

LLChiplets contain the LLC, the directory coherence controller, and the memory controller (MC). Considering an MC with a DDR4 DRAM interface in this chapter, which could, however, be replaced by any other memory interface (e.g. serial HMC, HBM, etc.).

Co-locating the directory and memory controllers with the LLC on the same chiplet adds no overheads to the off-chip memory access time compared to designs monolithically integrating the controllers and LLC on the processor chiplet. Moreover, re-integration of L3s on the LLC chiplets eliminates the need for memory buffers (i.e., L4) by providing the same functionality through a flattened hierarchy–by bringing the LLCs and memory ”closer” to each other.

With LLC sizes growing (currently up to 16MiB per chiplet [92]), enabling separate manufacturing processes for processor/memory chiplets allows the utilization of the most cost efficient process nodes and/or most power efficient for SRAM memory (e.g. by reducing leakage power of memory cells). Alternatively, non-volatile memory technologies like STT-RAM, which represent a promising solution to replace SRAM based LLCs [114], would also bene-

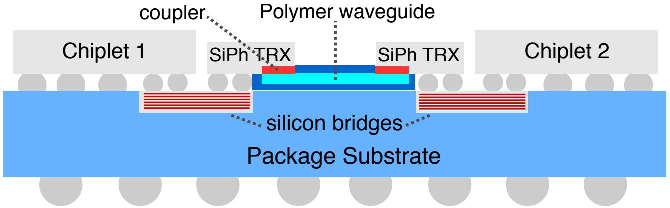


Figure 5.4: Optically-interconnected SiPh transceivers with transceiver-chiplet Si bridges on an organic substrate

fit from separate manufacturing as it reduces complexity of monolithic integration of different memory and processing technologies.

Aside from these benefits, such disintegration of LLCs would increase access latency and energy as the LLC is no longer on the same chiplet and possibly far away on the package substrate; however, the low-latency SiPh point-to-point fabric can mitigate these overheads to an extent where they do not cause noticeable performance degradation or energy overheads.

5.2.4 Addressing the Packaging Challenge

Figure 5.4 illustrates the cross-sectional view of the proposed packaging approach which loosely adopts a previous technique for inter-package communication [122] and applies it to inter-chiplet communication. Rather than monolithically integrating SiPh transceivers (TRXs) into processor chiplets, this chapter propose to implement dedicated SiPh transceiver chiplets connected to their respective processor chiplet on one side through Si bridges and to each other through SiPh transceivers and polymer waveguides on the other side.

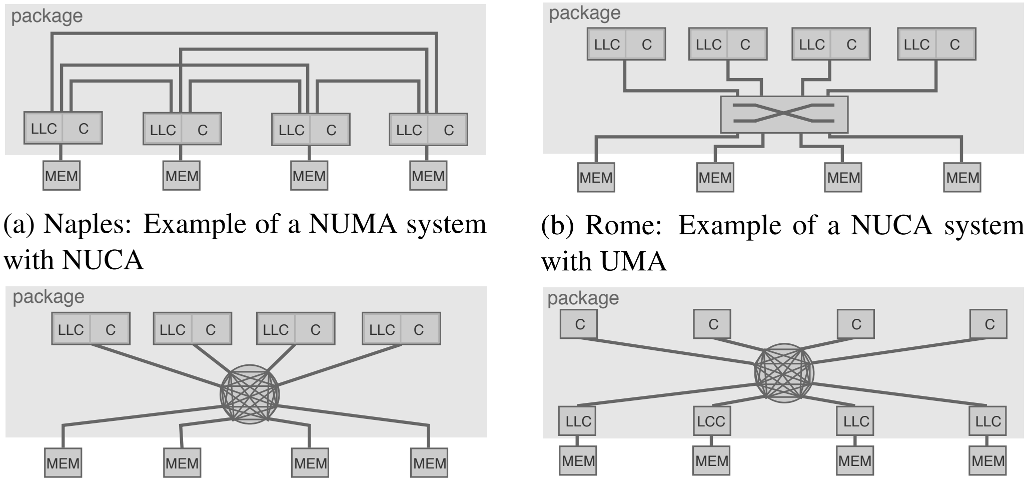
Polymer waveguides (PWGs) are integrated on top of the organic package substrate to provide optical connectivity between the chiplets. Scalable integration of PWGs (186 optical IOs [123]) has successfuly been demonstrated enabling high flexibility and connectivity in the interconnection network. Light is guided in and out of the chiplets using a vertical adiabatic coupler [123], which is a tapered waveguide (i.e., the waveguide width incrementally decreases) inside the chiplet placed on top of the on-package PWG. The tapering region of the waveguide confines the light and guides it into (and out of) the chiplets into/out of the PWG at low loss and low susceptibility to misalignment and mismatches. Reader can refer to the work of Dangel et al. [123, 124] for more details on the integration process.

The proposed approach combines SiPhs and Si bridges and utilizes each interconnection technology where it is the most efficient: SiPhs for long-distance interconnect between chiplets and Si bridges for short-distance interconnect between the TRXs and the chiplets. This approach does not reduce IO bandwidth density compared to direct optical communication between chiplets with monolithically integrated transceivers as the very fine-pitch electrical interconnects of Si bridges provide IO density similar to on-chip wires. Moreover, energy and latency overheads are negligible due to the small size of Si bridges and their low-loss vertical contacts [95].

Note that, while an additional TRX chip increases the distance between the actual chiplets compared to alternative techniques, the physical size of the optical TRX is very small as it only integrates transceiver circuitry and Si bridges enable tight packaging of just 100μm between the TRX and chiplets. The largest element on the optical TRX chip is the vertical adiabatic optical coupler, which, although providing very compact coupling width (5-20μm [123, 124]) should be at least 200μm long to offer low-loss optical coupling (coupling loss decreases with longer coupling structures). Modeling with DSENT [13] on a 45nm technology node shows less than $1mm^2$ for a transceiver matching the bandwidth of AMD's IF [93]. Nevertheless, these distance and area overheads will largely be outweighed by the benefits of this approach, which are

- significantly higher manufacturing yield compared to monolithic integration of SiPh TRX within the chiplets.
- more flexibility in choosing the most appropriate and efficient technology node and process to manufacture SiPh TRX and processor chiplets.
- elimination of any area concerns of μm-scale SiPh devices compared to nm-scale CMOS.

While materials that are not silicon (e.g. Germanium) required in SiPh device manufacturing necessitate modifications to standard CMOS processes and therefore cannot exploit their infrastructure, SiPh manufacturing and photonic integrated circuit design (along with advanced tooling to increase productivity) have seen significant growth and investment in the last ten years, now allowing low-cost SiPh integration [5]. For a more detailed cost roadmap, the reader can refer to [125].



(a) Naples: Example of a NUMA system with NUCA

(b) Rome: Example of a NUCA system with UMA

(c) S-UMA: System with UMA and NUCA

(d) S-UMA-Dis: Disintegrated S-UMA with UMA and UCA

Figure 5.5: Logical topologies of chiplet-based systems with different memory architectures

5.3 Methodology

System Comparisons. This chapter compares S-UMA to representative state-of-the-art chiplet-based systems, the logical topologies of which are shown in Figure 5.5 (examples are with 4 processor chiplets for illustration purposes, all networks will be evaluated with 8 processor chiplets like in Figure 5.3). In particular, this study compares a NUMA/NUCA system (similar to AMD's Naples [26]) shown in Figure 5.5a and a UMA/NUCA system (similar to AMD's Rome with a central electrical crossbar switch (”IO die”) [92]) shown in Figure 5.5b to S-UMA both with (S-UMA-Dis) and without (S-UMA) LLC disintegration (Figure 5.5d and 5.5c, respectively) to independently study both the impact of a point-to-point optical fabric on current chiplet-based systems in general as well as the performance impact of LLC disintegration. Finally, an electrical version S-UMA-Dis (S-UMA-E) is added to this study providing point-to-point interconnection with electrical links (which are unrealistic, and infeasible due to IO pin limitations) to analyze the benefits of a technology shift towards optical interconnects.

Performance modeling. The performance modeling was done using the cycle-level simulator gem5 [48] in *full-system mode* running the Linux operating system. For each network, the modeled system matches the bandwidth of AMD's Infinity Fabric (IF) on the network links

	Naples	Rome	S-UMA(-dis/-E)
Topology	Hyper Cube	Star with Xbar	Point-to-point
Config	64 x86 cores (8 chiplets) @ 3GHz L1_I:64KB/core, assoc:4, private L1_D: 32KB/core, assoc:8, private L2: 512KB/core, assoc:8, private L3: 8MB/(8 cores), assoc:16, shared Memory: 8 Channels, DDR4, 2GB		

Table 5.2: System Configurations

which provides 20GB/s (160Gb/s) unidirectional link bandwidth[1] and assumes an electrical link traversal latency of 10ns [26], optical link traversal of 1ns [7, 79], 1 core clock cycle router traversal latency on the chiplets [26], and 25ns latency through the IO die in Rome-like UMA-NUCA design. Table 5.2 lists the configurations of the systems under investigation.

The networks are evaluated with both synthetic and application traffic to study how the networks perform under a set of modern application workloads and how they would perform based on the network injection rate (which allows infering how other applications or multi-programmed workloads with much higher traffic would perform). In synthetic workloads, packets consist of 4 flits with a flit width of 32 bits (according to AMD's IF which has a 32-bit wide interface) and apply uniform random and bit complement traffic to stress different corner cases of the topologies. For application workloads, this study evaluates the chiplet-based systems for a variety of high-performance computing workloads from the NAS Parallel Benchmarks (NPB) with "C" server-class input sets [126], Rodinia [127], PARSEC3.0 and Splash-2x with large input sets [46], collecting statistics during the parallel region of the workloads. Different applications with different data sharing and memory access patterns as well as working sets were chosen to identify the impact of S-UMA for a variety of application workloads.

Power modeling. The power consumption of the electrical interconnection fabrics is based

[1] IF operates at MEMCLK of DRAM (2666MHz in this case) and has SERDES circuitry to transmit $4\times 32-bits$ per MEMCLK [26]

Paramer	Value	Parameter	Value
Optical Fiber	5e-6 dB/cm	Photodetector loss	0.1 dB
Modulator Insertion loss	1 dB	Power Margin	3 dB
Waveguide loss	0.5 dB/cm	Filter through loss	0.1 dB
Filter drop loss	1.5 dB	AWGR loss	1.8 dB
Coupler: Fiber-to-Package	3 dB	AWGR crosstalk	-20 dB
Coupler: Package-to-Chiplet	0.5 dB	Laser efficiency	14%
Receiver Sensitivity	-17 dBm		

Table 5.3: SiPh Device Parameters

on the energy-per-bit values reported for AMD's IF in their chiplet-based systems, i.e., 2pJ/bit per [26][2]. This study uses DSENT [13] for energy modeling of the switch and silicon bridge traversals based on a 14nm technology node (modeling links on Si bridges as BEOL links).

Modeling the power consumption of the SiPh links is based on a demonstrated 25Gb/s transceiver from Li et al. [128, 129] in 65nm CMOS (including static external laser power, serializer/deserializer, clock generation/recovery, drive rcircuitry, and microring tuning) combined with loss values corresponding to the SiPh packaging (i.e., demonstrated polymer waveguides and adiabatic couplers for both fiber-to-package and package-to-chiplet coupling [123, 124]) introduced in the previous section. In addition, SPICE models are used to scale down Li et al.'s transceiver to 28nm and 14nm to analyze power for both demonstrated technologies (65nm/28nm [118, 128–130]) and future projections (14nm). AWGR loss and crosstalk is based on a fabricated 16 × 16 SiN AWGR [12]. These assumptions based on manufactured and measured devices allows reporting realistic power consumption of the SiPh components. Tables 5.3 summarizes parameters and values.

[2] This is an optimistic assumption as 2pJ/bit in AMD's IF is consumed on intra-socket links. In Naples, some of the hypercube links are inter-socket links with higher pJ/bit.

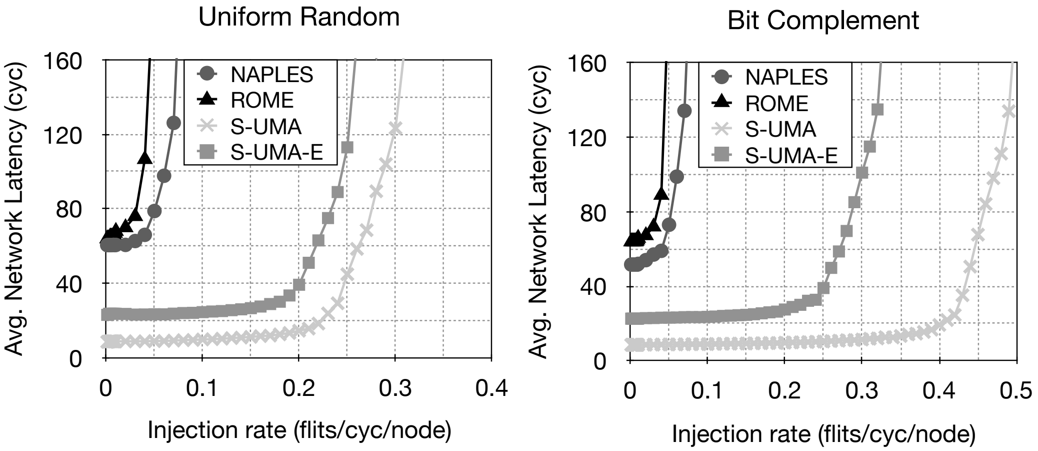


Figure 5.6: Average network latency (cycles) vs. injection rate (flits/cyc/node) under synthetic workloads

5.4 Simulation Results

5.4.1 Synthetic Workloads

5.4.1.1 Performance

Figure 5.6 depicts the average packet latency results (note that S-UMA and S-UMA-Dis are based on the same network–only the type of nodes would change–which is why the latter is left out in these charts for clarity). The general trends are similar in both traffic patterns: as expected, S-UMA and S-UMA-E–both based on all-to-all interconnection–offer both higher throughput due to higher bisection bandwidth and lower network latency due to lower average hop count compared to Naples (hypercube topology) and Rome (start/crossbar IO chip). In particular, network latency is significantly reduced by S-UMA as it not only has a lower diameter, but also lower link traversal latency due to optical communication (which is clearly more significant for inter-chiplet links than on-chip in NoCs where electrical communication over short distances can compete with optical links in terms of latency). Both latency critical applications and memory-bound workloads with high traffic between LLCs and/or off-chip memory would thus tremendously benefit from an optical all-to-all interconnect.

5.4.1.2 Power Consumption

Figure 5.7 shows the network power consumption for different injection rates. S-UMA-E consumes the highest power due to and the higher energy-per-bit compared to the SiPh links (2pJ/bit vs. 0.89pJ/bit (14nm) vs. 1.124pJ/bit (28nm) vs. 1.92pJ/bit (65nm)) and the larger number of

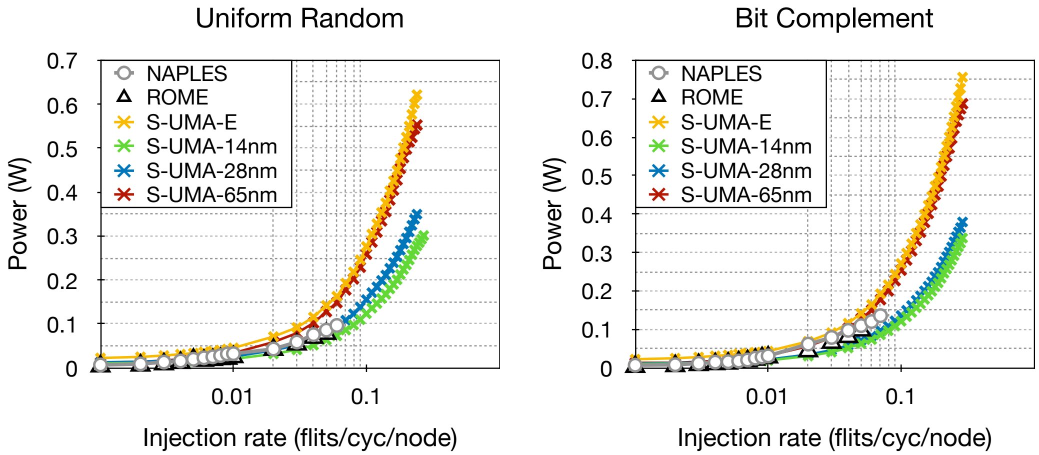


Figure 5.7: Power consumption (W) vs. injection rate (flits/cyc/node) under synthetic workloads

links compared to Naples and Rome (despite having a lower average number of hops). Although the electrical networks provide power consumption similar to S-UMA-E for low injection rates (mostly due to the static laser power overheads of the SiPh transceivers), S-UMA power efficiency becomes increasingly superior with rising injection rates due to its lower average hop count and lower-energy transceivers. S-UMA, therefore, not only provides higher bandwidth and lower latency, but also consumes less power, allowing system designers to dedicate more of the power budget towards the compute and memory–particularly for transceiver technologies below 65nm.

5.4.2 Application Workloads

5.4.2.1 Performance

Figure 5.8 illustrates the speed-up normalized to the Naples baseline. It can be observed that S-UMA provides an average speed-up of 23% compared to Naples and 12% compared to Rome, showing that the execution time of the considered workloads significantly benefits from the SiPh interconnection fabric. However, these speed-up benefits vary significantly. For instance, running Breadth First Search (*bfs*) with S-UMA results in 2.3× speed-up compared to Naples as graph traversals significantly benefit from all-to-all connectivity. Workloads with smaller data sets and less frequent communications such as *fft* do not stress the interconnect as much, thus exhibiting only marginal speed-ups (˜10%).

After analyzing the workloads, this study discovered that the speed-up benefits mostly de-

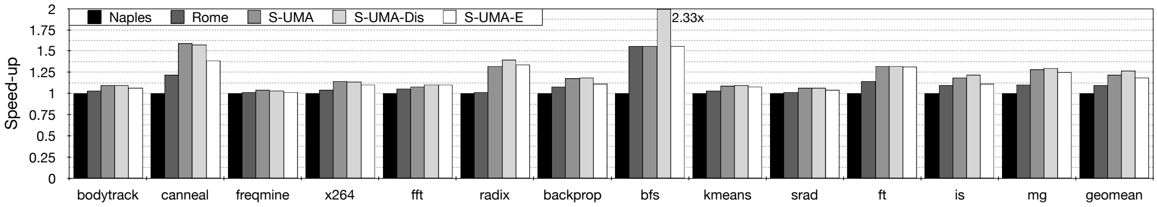


Figure 5.8: Speed-up normalized to Naples

pend on 1) the amount of data sharing within an application, 2) (ir)regularities in memory access patterns, and 3) the working set size. Particularly 3) is an important attribute as the simulation set-up, like state-of-the-art chiplet-based systems, assumes very large on-chiplet cache (1×8MiB L3 + 8×512KiB), meaning that large parts of the working set fit into a chiplets cache, even when choosing the largest input sets available for such systems. In these scenarios, most of the traffic throughout workload execution is due to data sharing and coherence traffic (rather than cache capacity), leading to very low network utilization (and, in turn, over-provisioning of the networking resources). Nevertheless, as Figure 5.8 shows, we still see a fair speed-up of S-UMA compared to the electrical networks due to its much lower network latency.

In the case of S-UMA-Dis, moving L3 caches off-chiplet increases L3 access latency compared to the other designs as on-chiplet accesses to a monolithically-integrated L3 are faster than inter-chiplet accesses; however, the speed and bandwidth provided by the all-to-all optical interconnect minimizes these latency overheads for most applications. In fact, co-locating memory controllers and directories with L3 slices on a dedicated chiplet reduces inter-chiplet traffic in case of L3 misses due to cache conflicts since write-backs go straight to the memory and are not sent off-chiplet. For applications with relatively high L3 miss rates like bfs or radix (which exhibit almost random memory access patterns), this actually leads to a performance speed-up compared to S-UMA.

These results suggest that the proposed SiPh point-to-point network allows to further exploit the benefits of chiplet disintegration while providing the benefits of a UMA design and still offering higher performance than state-of-the-art designs with less disintegration.

5.4.2.2 Power Consumption

Figure 5.9 shows the total power of each network normalized to Naples. On average, both S-UMA and S-UMA-Dis reduce power consumption by 30% compared to Naples and 48%

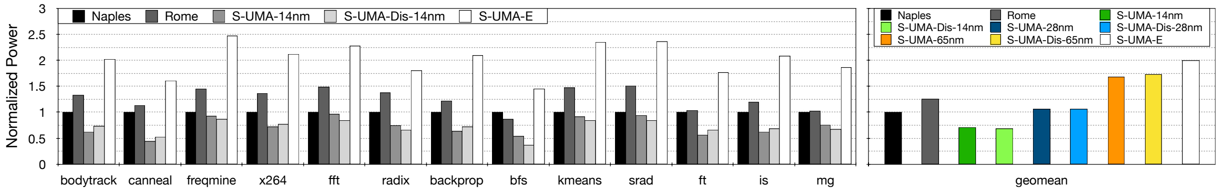


Figure 5.9: Power consumption normalized to Naples

compared to Rome for 14nm, which are both in line with the synthetic traffic results indicating that the lower average hop count and energy-per-bit of the SiPh TRXs offer significant power saving benefits. While the all-to-all topology with 28nm TRXs matches power consumption of Naples on average, older 65nm TRXs amount to significant power overheads. Generally, the higher the network utilization in the simulated workloads, the higher the power benefits of S-UMA(-Dis). This is mostly due to the fact that static power in SiPh (laser, MRR tuning) is a significant contributor to the total power and is amortized by high utilization.

5.4.3 Discussions

The simulation results indicate that a point-to-point SiPh interconnection fabric can offer substantial improvements in terms of average network latency, application execution time, and energy consumption compared to state-of-the-art electrical networks in chiplet-based systems . Not only does this proposal improve all figures of merit, it does so while enabling further disaggregation of LLC caches without performance hits and enables a scalable UMA architecture which will be significant to programmability and performance extraction of future systems. However, even with SiPh, point-to-point networks will sooner or later reach its scalability limits. Although offering much better scalability than electrical networks and enabling architectures that would otherwise not be practical in the electrical domain,

The scalability of S-UMA will ultimately be limited by the crosstalk inside the AWGR, which limits its port count. The analytical models of state-of-the-art SiN AWGRs [12, 131] show that they can scale up to 32 × 32 with acceptable power penalty, which would limit S-UMA to 32 chiplets for current AWGR technologies. However, AWGRs made from other materials like Silica scale to much higher port counts (up to 512 × 512), albeit with larger area footprint ($16mm \times 11mm = 176mm^2$) [60]. Nevertheless, this area footprint might be negligible for S-UMA with +32 chiplets, which are $213mm^2$ each in current designs [132].

Aside from scalability limits due to AWGR crosstalk, optical link bandwidth has a significant impact on laser power and MRR tuning [6, 9, 10]), especially relative to network size. This might become a limiting factor to system scalability as the number of laser sources and their power impact the overall power efficiency of the system. Presented results indicate, however, that for the workloads considered in this study, latency rather than bandwidth was the limiting factor. Therefore, for various HPC workloads considered in this study, link bandwidth could likely be reduced to save power and enable further scalability without significant performance hits.

Nevertheless, systems like multi-GPU systems have high link bandwidth demands, possibly making a straight-forward scaling of point-to-point networks power inefficient; however, as static laser power in SiPh is a well-known issue for such systems, several solutions such as adaptive laser sources which can be turned on and off based on the current bandwidth demands have been proposed to solve this issue [61, 133–135]. Chiplet-based systems in the HPC domain might particularly benefits from adaptive lasers as HPC systems have to process a variety of different workloads. Prior to launching applications, lasers could be turned on and off to provide the bandwidth necessary for the system. While this study assumes external off-chip lasers in this study, the time to control such lasers could be amortized by executing laser control in parallel to the set-up phase of an application. Alternatively, on-chip lasers, although currently still suffering from low manufacturing yield and thermal issues, could be an attractive solution in the future as they allow nanosecond-scale control time, potentially even allowing dynamic bandwidth reconfiguration during an application execution.

Despite the scalability limits (which are still well-above what is possible in the electrical domain), the advancements and amount of research dedicated to SiPh devices is highly encouraging, with improvements in losses, laser efficiencies, and receiver sensitivities constantly published in the literature [136]. These have a significant impact on the total laser power could allow low laser power requirements even for high-bandwidth all-to-all fabrics in chiplet-based systems of medium or large scale, further fueling the scaling-up of shared memory chiplet-based HPC compute nodes and thereby reducing the data movement problem of scaling-out systems.

5.5 Conclusion

Given the increasing cost of monolithic chip design, multi-chip packages using smaller chiplets are becoming more common. However, these modular designs put significant strain on electrical interconnection networks leading to non-uniform memory architectures with large remote memory access latencies. The study presented in this chapter is published in the international symposium on memory systems (MEMSYS) [137]. The study presented in this chapter showed that SiPhs can enable uniform memory architectures in chiplet-based systems with low memory access latency and energy. SUMA exploits the high bandwidth density in optics to overcome issues arising due to IO pin bandwidth limitations and their wavelength-selective routing properties to enable point-to-point networks with low pin IO demands. Simulation results suggest that SiPhs could enable scalable chiplet-based uniform memory architectures and thus be of high importance to scale-up performance and, in turn, reduce the data movement overheads of scaling-out in HPC systems.

Chapter 6

HTA: A Scalable High-Throughput Accelerator for Irregular HPC Workloads

The advent of exponentially-growing data-intensive applications across several domains has created a category of throughput-oriented workloads. This class of *irregular* applications impose new challenges for computer architects as their data sets are increasingly sparse and they exhibit poor locality in memory accesses. Unlike traditional compute-intensive applications, computing solutions designed for irregular applications should focus on reducing the latency and energy overheads of inevitable data movements.

The computing community has been utilizing GPUs as data-parallel accelerators given their massive throughput offerings. Though GPUs have proved to be effective as high throughput accelerators for many regular applications, we explore *specializing* data-parallel accelerators for efficient execution of *irregular* data-parallel workloads. These applications exhibit random memory access patterns, essentially making any shared component an architectural bottleneck limiting the obtainable throughput. Our main insight in designing HTA is to reduce the *contention* within the memory system and reduce the energy and performance cost of data movement.

On the scalability front, as we reach the end of transistor scaling, we cannot simply rely on increasing the number of compute units on a single die to scale. An alternative approach is to design processors utilizing multiple "chiplets" [91]. Chiplets assembled using advanced packaging technologies, such as multi-chip-modules (MCMs), can offer a scalable design compared to one large monolithic chip. However, the inter-chiplet communication and its energy efficiency are known as the dominant factors towards performance and scalability due to significant power penalties brought by MCM designs [138]. We propose to address this challenge by taking advantage of recent advances in 2.5D/3D packaging with Silicon Photonics, which offers advantages of significantly lower energy per bit and scalability to much larger interposers than what today's reticle size limits allow. For example, recently TSMC and Broadcom announced 1700 mm^2 interposer [139] which is twice the size of the maximum reticle size by proposing to stitch together multiple interposers together.

This chapter present the design, evaluation, and 2.5D/3D packaging solution of the high-throughput scalable accelerator architecture called **HTA**. HTA's memory architecture exploits a partitioned memory controller (PMC) and all-to-all SiPh interconnects replacing conventional cross-bar based systems to support nearly-contention-free, high-throughput, and scalable data movement between the compute cores and the main memory. The partitioned memory controller reduces the queuing latency by 10% to 30% which translate to 5% to 26% reduction on overall memory access latency. In addition, addressing the contention in the memory controller reduces the variations in access latency by 10% to 60% in terms of 95^{th} percentile latency. Furthermore, HTA improves the performance of the memory system and reduces L1 misses penalty by 2.3× to 5×. Evaluating our design at scale shows 1.5× speedup on average for HTA compared to a multi-GPU system for the same number of compute units.

The rest of this chapter is organized as follows. Section6.1 presents challenges towards scaling the memory system in the state-of-the-art data-parallel accelerators. Section 6.1.1 describes the architecture of partitioned memory controller, utilizing an interconnect fabric described in Section 6.1.2. Section 6.1.4 presents HTA architecture which builds on top of the proposed memory system. Through simulations with the methodology described in Section 6.2, the performance of partitioned controller and the proposed HTA architecture are evaluated in Section 6.3, followed by the conclusions in Section 6.4.

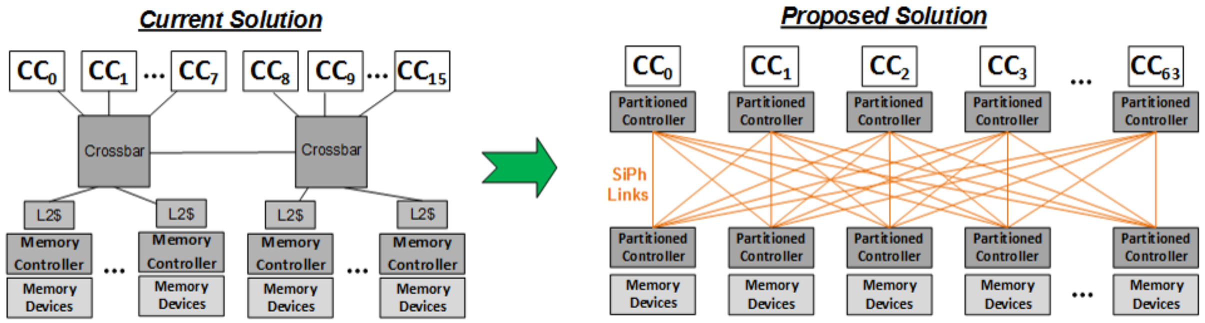


Figure 6.1: (left) Overview of baseline memory system where different core clusters (CCs) share a crossbar, a single read/write queue per channel, and a last level cache. (right) Proposed memory system addresses the contention by providing dedicated queues for each core cluster to send memory request to every channel through an all-to-all interconnect.

6.1 HTA - Background, Rationale, and Design

GPUs are the de facto choice for high throughput accelerators in the HPC domain. The left side of Figure 6.1 shows an overview of state-of-the-art GPUs. We identify four key challenges to the architecture shown in Figure 6.1 when it comes to scaling irregular applications.

1) Crossbar Radix: Increasing the number of core clusters requires increasing the radix of the electrical crossbar between the cores and the L2 caches as current systems implement a mostly uniform L2 architecture. In addition to the power and area overheads of the crossbar, it imposes a trade-off between latency and bandwidth: to increase the bisection bandwidth there must be more layers in the crossbar increasing both latency and area.

2) Overheads of Data Movement: Moving the data through multiple levels of memory hierarchy adds to memory access latency and results in increased energy consumption. This challenge becomes more important as physical distance between different levels increases in multi-chip module systems. In fact, the performance and energy overheads of data movements are known to be the main limiting factor towards scalability of multichip modules systems [138].

3) Bandwidth to Memory: Scaling the number of compute units in the system increases the demand for bandwidth to memory. Already limited by the latency-bandwidth trade-off due to the crossbar design, the number of available pins (between the compute dies and memory) add another constraint on bandwidth, especially in chiplet-based designs.

4) Variability in Memory Latency: Memory requests from different processing units share many deep queues including the crossbar, an L2 bank, the memory controller queues, DRAM bus, and DRAM banks. The contention from different compute units at these components increases the queuing delay which leads to variations in access latency and adds to the complexity of the scheduling for the memory controller and GPU cores.

Recent design trends from NVIDIA and AMD have taken steps to address these challenges. These solutions are inspired by similar techniques used in CPUs, and as a result, they do not address the underlying problem (i.e., contention) especially as we go towards scaling these systems. For instance, on a single GPU, NVIDIA's Ampere architecture [140] increases the number of compute units by 50% (from 84 in Volta to 128 in Ampere). To maintain a reasonable radix for the crossbar, the crossbar in is partitioned into two pieces. However, this approach introduces non-uniform latency and bandwidth to the memory, increasing the programming complexity on these systems. AMD's RDNA architecture [141] reduces the radix of the crossbar by adding a L1 cache which filters requests from all Compute Units (CUs) within a core cluster. While this approach simplifies the crossbar design, and reduces the pressure on the globally shared L2 cache, it adds to variability in memory access latency and only helps workloads which have regular memory access patterns or temporal reuse. AMD's CDNA architecture [142] eliminates the L1 cache along with the fixed-functions logic dedicated for graphics application to free up area and power for adding more CUs. However, the crossbar (and subsequently the L2 cache) is divided into two slices to achieve a reasonable radix for the state-of-the-art electrical interconnect technologies. Similar to NVIDIA's design, this approach increases the programming complexity by introducing non-uniformity in both latency and bandwidth, and further increases the variability in memory access latency.

The *main idea* underlying our proposal for HTA is to eliminate the contention in the memory subsystem as much as possible. We focus on three sources of contention: the on-chip crossbar, the globally shared L2 cache, and the memory controller queues. Our proposal makes the following **contributions** towards addressing the sources of contention in data-parallel accelerators.

(a) To reduce the contention at the request queues, we partition the memory controller into two parts: core-side controller with dedicated queues per core cluster (CC), and memory-side controller in charge of scheduling and issuing DRAM commands. This reduces the contention on read/write queues by offering dedicated queues for each core cluster and reduces queuing latency by avoiding the head-of-line blocking in scheduling. We will discuss the architecture and scheduling of proposed memory controller in Section 6.1.1

(b) The contention at the crossconnect is reduced by providing direct point-to-point links. However, implementing such a topology using electrical links would be extremely challenging due to bandwidth, energy, and routing limitations. To that end, and to reduce the overhead of data movements, we leverage an efficient all-to-all passive optical fabric (called Arrayed Waveguide Grating Router or AWGR) enabled by silicon photonics by taking advantage of 2.5D packaging. Describing the key enabling technology for our architecture, the details of proposed interconnect and packaging solutions are presented in Section 6.1.2 and Section 6.1.3 respectively.

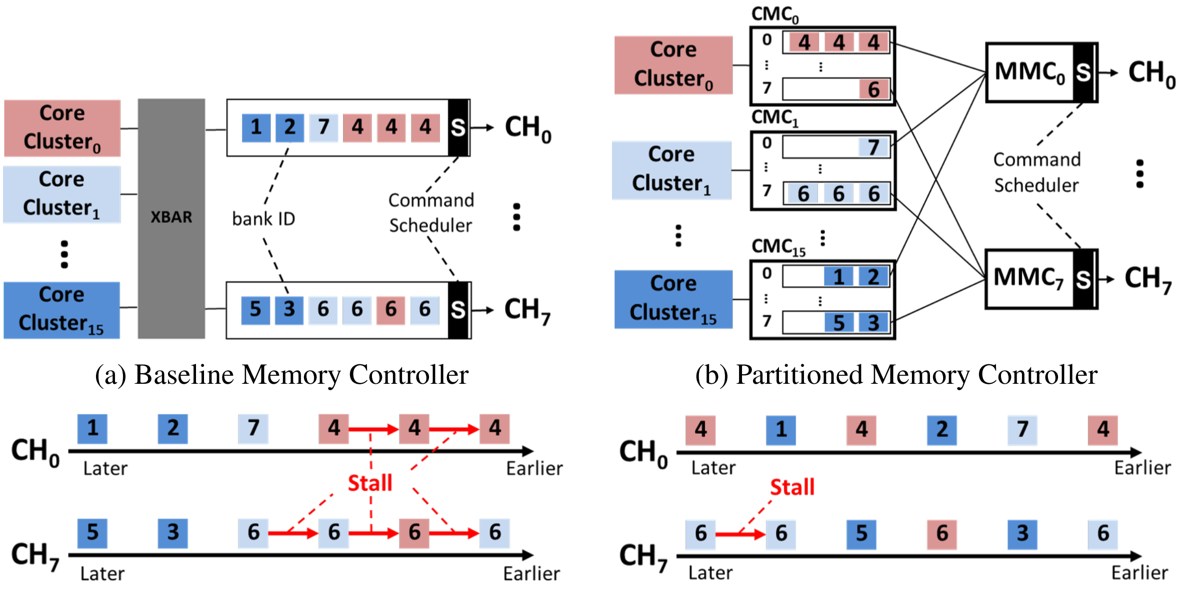
(c) We utilize the partitioned memory controller design, and propose HTA in Section 6.1.4, which benefits from a scalable unified memory architecture and avoid NUMA challenges.

6.1.1 Partitioned Memory Controller

In this section, we present the details of our proposed Partitioned Memory Controller (PMC) which consists of two parts: the compute-side memory controller (CMC) and the memory-side memory controller (MMC). For the discussions and evaluations presented in this chapter, we target HBM as the DRAM device, but the core idea of our proposal is agnostic to DRAM micro-architecture and can be applied to other DRAM technologies (e.g., GDDR, DDR, etc.) in a similar fashion as we focus only on the memory controller design and require no changes to DRAM core architecture (see Section 6.1.3 for details).

Figure 6.2b presents an overall view of the components within PMC. The key idea is to eliminate the contention on request queues and improve bank utilization by avoiding stalls due to bank conflicts between requests from different core clusters. With dedicated set of queues per channel for each core cluster, the variation in the memory access latency will be limited to unavoidable conflicting patterns from a single core cluster.

While dedicated queues eliminate the contention, the memory controller still needs to have a single scheduler per bank as point of reference for DRAM timings. Thus, we *partition* the memory controller into two parts. We keep the *front-end* (containing dedicated read/write queues) on



(a) Baseline Memory Controller

(b) Partitioned Memory Controller

(c) Scheduling timeline for the baseline

(d) Scheduling timeline for PMC

Figure 6.2: Working example of PMC, showing how the scheduling of memory requests is improved compared to the baseline. The stalls are avoided by scheduling requests from different core clusters, and is limited only to the conflicting requests within a core cluster. This is mainly achieved by expanding the scheduling window for each channel beyond the size of a single read or write queue. The probability of finding non-conflicting requests is increased, and a better quality of service is provided to each core cluster.

the accelerator side, and move the *back-end* (including scheduling logic, and command queues) to the memory side.

Our design requires an all-to-all interconnect between the front-end and the back-end. Section 6.1.2 describes how a multi-wavelength routing device called AWGR can be used to replace the long-latency electrical crossbar while offering high-throughput contention-free communication.

Compute-side Memory Controller (CMC) As Figure 6.2b illustrates, we keep read and write queues on the processor side, with dedicated read and write queues for each channel. The idea is to limit the contention only to requests from the CUs within a single core cluster, and not all core clusters within the system. These queues are the result of breaking down the single shared read/write queue in the baseline memory controller shown in Figure 6.2a into per core cluster queues.

Requests from L1 caches in each core cluster are routed to proper queues according to the address mapping scheme, similar to how corresponding L2 banks are selected for each request in the baseline architecture. Each Compute-side Memory Controller (CMC) has dedicated links to communicate with the Memory-side Memory Controller (MMC) for a given channel.

Memory-side Memory Controller (MMC) Figure 6.2b shows two channels of our proposed memory controller, and connectivity between MMC and read/write queues from different CMCs. The scheduler looks at requests from all core clusters regardless of their queue occupancy. Therefore, the scheduler can continue servicing memory requests even when one requester has several conflicting requests issued within a short period of time—a common case in high-throughput accelerators illustrated in Figure 6.2.

At each cycle, all CMCs send a copy of the request at the head of their queues. Then, an MMC selects a request to serve, it broadcasts back the requester ID (i.e., the winner) and the bank number to all CMCs. Thus, other requesters with requests for the same bank at the head of their queues can wait until the response is provided. Requests from different requesters (i.e., core clusters) are serviced in a round-robin fashion with an FR-FCFS scheduling policy similar to the baseline.

Figure 6.2 illustrates how the partitioned memory controller can address the head-of-line blocking problem. One core cluster (CC_0) is sending several conflicting requests (going to the same bank) to the first channel (CH_0). This results in several stalls during the scheduling. However, these stalls can be avoided by addressing non-conflicting requests from other CCs between the conflicting requests. PMC achieves this by allowing the MMC to select from dedicated queues for each channel within each CMC. Current systems use deep associative queues to avoid these stalls by finding requests to different banks within the queue. However, one CC can fill the queue with conflicting requests in a short period of time. This leaves the scheduler with no other options to choose from, even using the most sophisticated logic-intensive associative queues, and results in unnecessary back pressure applied to the whole system.

We should note that the processor's total queue size remains unchanged for each core cluster. We are essentially breaking down a large shared queue into *n* (*i.e.*, number of channels) smaller dedicated queues. The overhead of this is approach is limited to a small fraction to replicate the logic needed for maintaining those queues. On the memory side, there will be a small overhead for the added queues and we envision this logic to be implemented on the logic layer in 3D stacked memories.

6.1.2 Interconnect

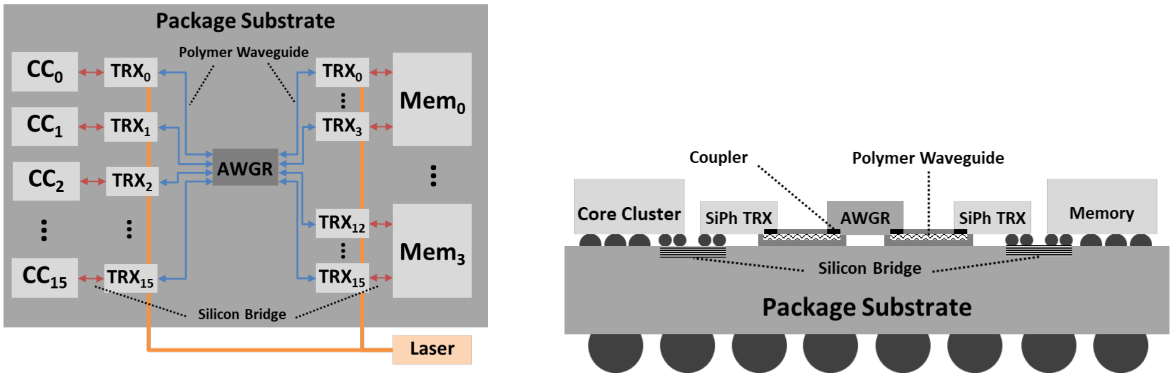
To address the contention at the crossconnect, our design utilizes a point-to-point connectivity between the core clusters and memory controllers. Besides addressing the contention, our proposed architecture requires an all-to-all connectivity between CMCs and MMCs. This connectivity allows for our scheduling policy to make local decisions at the MMC and updating CMCs through broadcasting.

As discuss earlier, designing a scalable high-throughput accelerator requires addressing the cost of data movements. Disaggregating the monolithic chip into multiple smaller chiplets allows for more input/output interfaces for each core cluster. However, chiplet electrical interconnection suffers from high distance-dependent signal loss and limited I/O bandwidth [6]. Therefore, interconnecting many non-adjacent chiplets require multi-hop networks with repeaters, incurring large latency and energy overheads. These challenges can be overcome by silicon photonic technology: reducing latency with almost distance-independent communication energy and providing high pin bandwidth density through wavelength-division multiplexing (WDM) [143]. In the following sections, we present a summary on the principle of operation for optical links used in our design, and discuss the details about our proposed interconnect fabric and packaging solution.

6.1.2.1 Silicon Photonics

As discussed earlier in Chapter 2, integrated optical interconnects, enabled by silicon photonics, offer properties that can be exploited to address the performance and energy overheads of data movements in high-throughput accelerators.

An external WDM laser (in form of an optical frequency comb source or individual lasers) generates the optical signal at the required wavelengths, which are then coupled from a fiber into on-chip waveguides. On-chip modulators encode bits onto wavelengths (one modulator for each wavelength). Then, the modulated wavelengths traverse the waveguides and are filtered out and converted back into the electrical domain by on-chip photodetectors. In terms of latency, electrical-to-optical (EO) and optical-to-electrical (OE) signal conversions are done at one cycle



(a) Top-view.

(b) Side view.

Figure 6.3: Example of proposed packaging solution, where Compute and Memory dies are optically-interconnected through an AWGR using SiPh transceivers with transceiver-chiplets and Si bridges on an organic substrate.

and incur no additional latency to the transmission line.

6.1.2.2 Arrayed Waveguide Grating Router

One interesting property of WDM technology (aside from its bandwidth benefits), is that it allows connecting a single node to multiple receiver nodes by leveraging wavelength-selective routing devices. This method allows implementing an all-to-all network without a large number of point-to-point ports.

Among different SiPh wavelength routing devices that have been demonstrated [6], we utilize the Arrayed Waveguide Grating Router (AWGR) with a footprint of ~1mm^2 [12] to provide contention-less point-to-point connectivity between all chiplets. AWGR is a *passive* SiPh fabric which provides all-to-all connectivity between any input and any output port. Several studies explored AWGRs as a uniquely compact solution for all-to-all interconnection with lower loss and crosstalk compared with other SiPh devices providing similar connectivity [56, 79, 137]. The reader can refer to the following articles for what concerns the physics, design principle, and scalability of AWGRs [37, 38, 64].

6.1.3 Packaging

Figure 6.3 presents an overview of the packaging approach we use in our design. We adopt a previously proposed technique for intra-package communication [122, 137, 144] which can be applied to our memory controller design. This approach considers developing dedicated SiPh transceiver chiplets connected to their respective (compute or memory) dies.

The advantage of this design decision is that it can be leveraged to provide support for off-the-shelf memory devices (e.g., HBM, GDDR, etc.) by choosing the proper command scheduler in MMCs. By integrating the MMC and SiPh TRx (on the memory side) on the same die, no extra logic is required on memory dies, and MMCs can be designed to work with existing PHY interfaces - with minimal distance for data movements on electrical wires.

The dedicated SiPh transceiver chiplets connected to their respective dies on one side through Si bridges and to AWGR (the fabric providing all-to-all connectivity) through polymer waveguides (PWGs). These polymer waveguides are integrated on top of the organic package substrate and provide inter-chiplet optical connectivity. The reader can refer to the work of Dangel et al. [123, 124] for the details on the overall integration process for polymer waveguides.

Combining SiPh and Si bridges, our proposal utilizes each interconnection technology where it is the most efficient: SiPh for long-distance cross-package interconnect between chiplets and Si bridges for short-distance electrical interconnect between the TRXs and the memory controller.

SiPh manufacturing processes exploits well established CMOS processes, and photonic integrated circuit design kits (PDKs) have seen significant growth in the past ten years, resulting in cost-effective SiPh integration [5]. The reader can refer to [125, 137] for more detailed cost analyses and roadmap.

6.1.4 HTA Architecture

We discussed the challenges in scaling the memory architecture for today's high-throughput accelerator and how our proposed memory architecture addresses them. In this section we build on top of the proposed memory system, and introduce a high-throughput accelerator (HTA) architecture which takes the advantage of low-latency all-to-all optical fabric and allows elimination of the shared last level cache.

Elimination of last-level caches provide significant advantages in terms of dedicating more area for compute, reducing access latency, and improving predictability in memory access time. The photonic interconnect used in our proposal provides us higher bandwidth at a lower energy per bit cost to make the underlying design tradeoffs such as eliminating the last level caches feasible, especially for irregular workloads with poor locality.

6.1.4.1 Scalability of HTA

One of the main benefits of SiPh interconnects is their distance-independent energy consumption and performance. Combining this with the benefits of packaging solution discussed in Section 6.1.3 allows HTA to scale.

Considering the area saving from eliminating L2 cache (occupying ∼50% of chip area), a single package instance of HTA can support 4× more compute units. Moreover, multiple packages can be utilized to scale further, and realize a scalable high-throughput accelerator with a unified address space without considerable energy and performance overheads.

The major component in HTA that needs to scale with the system is the AWGR. In this chapter, we study HTA with 64 and 256 CUs which can be realized using 16×16 and 64×64 AWGR respectively. Scaling above 256 CUs requires AWGR with more than 64 ports. While 512×512 AWGR has been demonstrated [145], the main challenge for implementing AWGRs with high port counts (i.e., >64) is the optical crosstalk. However, the new Thin-CLOS architecture successfully demonstrated by Proietti et al. [38] can utilize multiples of smaller AWGRs (lower port count) in parallel to provide the same functionality of a larger AWGR at lower crosstalk. While these solutions have larger footprints, the area overhead might be negligible in large accelerators with more than 256 CUs.

The bandwidth between any input-output pair in AWGR is limited to the information that can be carried out by a single modulated wavelength. If the bandwidth requirements exceed what a single wavelength offers, there are two alternative options. The first one is to leverage multiple free spectral ranges (FSRs) of an AWGR [18, 79], and virtually create a parallel channel of communication. The second one is to use spatial-division multiplexing (SDM), i.e., integrating and transmitting data through parallel AWGRs (either planar or 3D-stacked [17]). Multi-FSR strategy requires a broader laser spectrum and higher laser power to compensate for higher crosstalk inside the AWGR and to guarantee the required minimum optical power at the receiver. The SDM approach has similar laser power requirements but does not need a broader laser spectrum. However, it needs a larger die area or more SiPh layers, as well as more optical IO pins.

Table 6.1: Simulation Parameters

Compute Cluster			
Number of CUs	**64**	**CUs per CC**	**4**
Memory Hierarchy			
L0 V$	**16kB (per CU)**	**L0 I$**	**32kB (per CC)**
L0 K$	**16kB (per CC)**	**L1 $**	**64kB (per CC)**
L2 $	**2MB (8 banks)**	**DRAM**	**4GB HBM2 [148]**

6.2 Methodology

6.2.1 System Comparisons

To evaluate our proposed HTA architecture, we compare it against a system similar to AMD's RDNA architecture with details of the memory hierarchy shown in Figure 6.1. CUs within a core cluster have private caches ("L0") and share the L1 cache, which centralizes all caching functions within each cluster [141]. L1 caches are connected to a globally shared L2 cache through a long-latency crossbar interconnect, resulting in ~100 cycles hit latency for L2 [146]. Therefore, for our simulations, we modelled the electrical crossbar with a latency of 50 cycles in each direction.

Within the memory controller of a given channel, all requests from different CUs share a read and a write queue. In each cycle, the scheduler performs an associative search and issues commands for requests in a First-Ready First-Come-First-Served (FR-FCFS [147]) fashion. For our evaluations, we refer to this design as the baseline memory controller. While we use AMD's RDNA memory hierarchy as our baseline, the challenges in scaling the memory hierarchy of GPUs are common in NVIDIA's systems and our proposal can be applied there similarly.

One example of HTA can host 64 CUs by utilizing a 16 × 16 AWGR to interconnect eight compute chiplets (each with four CUs) to four stacks of HBM2 memory. SiPh links use WDM with 16 wavelengths and perform modulation/demodulation at 32Gbps. On the compute side, each compute chiplet uses one SiPh WDM TRX with 64GB/s bandwidth in each direction, making a total of 16 SiPh TRXs for CMCs. On the memory side, four SiPh WDM TRXs can match the 256GB/s bandwidth of a single stack of HBM2 which results in a total of 16 SiPh TRXs for MMCs.

6.2.2 Simulations

6.2.2.1 Performance

To model our target systems we use MGPUSim [149] which models the Graphics Core Next 3 (GCN3) ISA. We extended the simulator to model a three level cache hierarchy. We integrated the timing model from DRAMSim3 [148] after extending it to model our proposed partitioned memory controller design discussed in Section 6.1. We utilize MGPUSim for collecting the traces on the memory system, and piped those traces on detailed timing model on DRAMSim.

For the performance of the interconnect technologies used in this chapter, we used latency reporting in the previous work [137, 146]. The details of the modeled system in the simulator for different components are listed in Table 6.1. It should be noted that the trace-based evaluation approach limits our reporting to the performance of the memory system, and does not allow us to obtain execution times for the two systems under comparison. However, since a significant portion of the pipeline stalls are due to memory accesses, the performance of the memory system would be a reasonable candidate for our evaluation. To this end, we will look at the penalty of L1 misses when comparing the baseline with PMC in Section 6.3.

For evaluating our proposal we used benchmarks from AMD's Accelerated Parallel Processing (APP) Software Development Kit (SDK), Hetero-Mark suite [150], and Scalable Heterogeneous Computing (SHOC) suite [151].

Among those supported by MGPUSim, we chose different benchmarks with different memory behaviours to evaluate our proposal under different scenarios. Breadth-first Search *bfs* and Page Rank *pr* represent applications with irregular memory access patterns (i.e., poor locality). AES-256 Encryption (*aes*), Fast Fourier Transform (*fft*), and FIR Filter (*fir*) represent typical compute intense HPC applications with considerable amount of data reuse (i.e., medium locality). Simple Convolution (*conv*) implementation used for this work divides the image into sub-images to maximize data reuse (i.e., high locality).

6.3 Evaluation

In this section, we present the evaluation results on three aspects of our proposal.

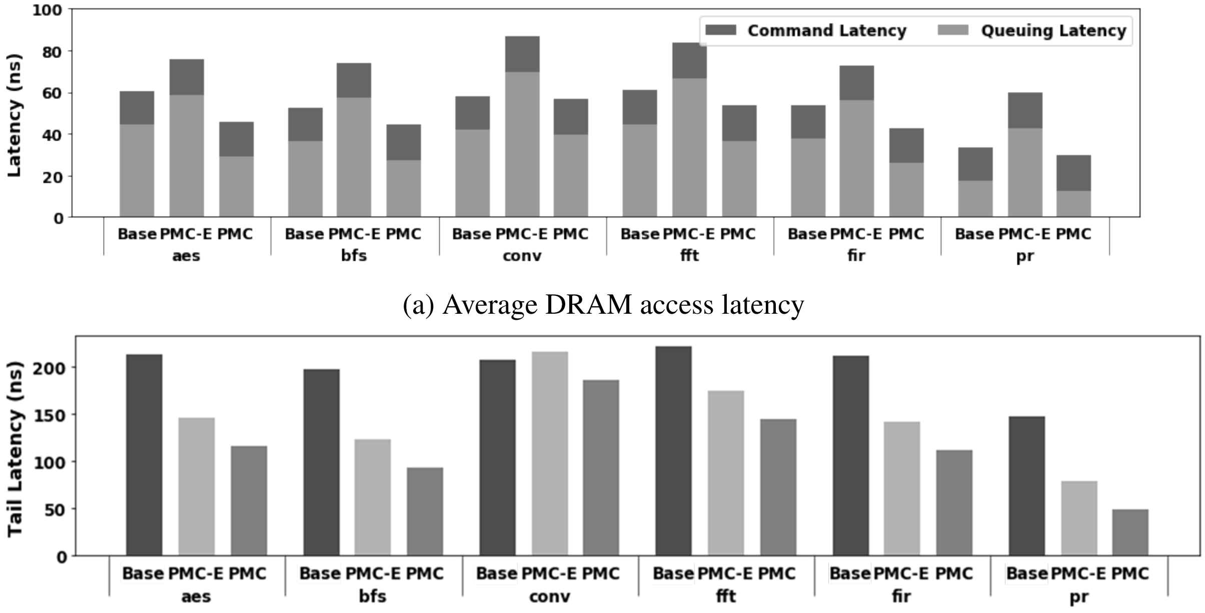
First, we look at the performance of the proposed memory controller design compared to the baseline memory controller discussed in Section 6.1. This analysis is done under the same cache hierarchy. In these experiments, we look at the average DRAM access latency in both designs, as well as 95^{th} percentile latency as a measure of divergence in the access latency. Second, we evaluate HTA design against the baseline GPU architecture. In this set of analyses, we evaluate our memory controller design combined with a new cache hierarchy, and model a system like the one shown on right in Figure 6.1. We report the average miss penalty for L1 caches in the form of Average Memory Access Time (AMAT) for L1 misses. Third, we evaluate our proposal at scale by comparing the performance of HTA with 256 CUs against a multi-GPU system with 4× 64CU GPUs.

6.3.1 Evaluation of Partitioned Memory Controller

As our first step in evaluating our proposed architecture, we compare the performance of the partitioned memory controller against the baseline memory controller, both using the same cache organization. To emphasize on the importance of the enabling technology used in our design, an implementation of PMC using electrical links (PMC-E) is evaluated.

PMC design reduces access latency divergence by avoiding head-of-line blocking in scheduling. In the baseline design where all requesters share a single queue, if one requester sends a stream of requests over a short window (a common case in data-parallel accelerators), requests from other requesters are blocked until DRAM manages to return pending requests. PMC avoids this by having dedicated queues for each requester and directly applies the back-pressure to the original requester and not the whole system. Figure 6.4b shows the 95^{th} percentile in access latency, indicating a significant reduction in memory latency variation for PMC over the baseline memory controller. Depending on the access pattern in each workload, the 95^{th} percentile in access latency is improved by 10% to 60%. The benefits gained through scheduling are strong enough to result in improved tail latency even for the electrical implementation of the PMC which suffers from high-latency links.

Besides improving the predictability in access latency, PMC improves the access latency by



(a) Average DRAM access latency

(b) 95th percentile latency for DRAM access

Figure 6.4: DRAM performance for the baseline memory controller (Base) compared to a system utilizing Partitioned Memory Controller with implemented with electrical and SiPh links (PMC-E and PMC, respectively). (a) In terms of access latency, PMC improves the queuing latency by 10% to 30% resulting in 5% to 26% reduction on overall access latency compared to the baseline memory controller. (b) The 95th percentile latency for DRAM access is improved by 10% to 60% by reducing contention at read and write queues within the memory controller.

increasing parallelism in bank accesses within the DRAM. Figure 6.4a depicts the average memory latency for the baseline memory controller and the proposed PMC. PMC achieves a lower average access latency by avoiding a portion of bank conflicts in the memory requests. If one requester sends several conflicting requests, those would limit bank activations in the baseline design, while in PMC, the scheduler can schedule requests from other requesters. Therefore, the queuing portion of memory access is reduced by 10% to 30% depending on the access pattern exhibited by each workload.

Both PMC-E and PMC take advantage of the scheduling scheme offered by PMC and avoid head-of-line blocking which translates to improvements in tail latency. This is purely due to the scheduling scheme in PMC, and it is independent of the technology used to implement the point-to-point fabric. However, as described in Section 6.1.1, the PMC design makes the crossbar latency part of the memory access. Therefore, the latency overhead imposed by the interconnect

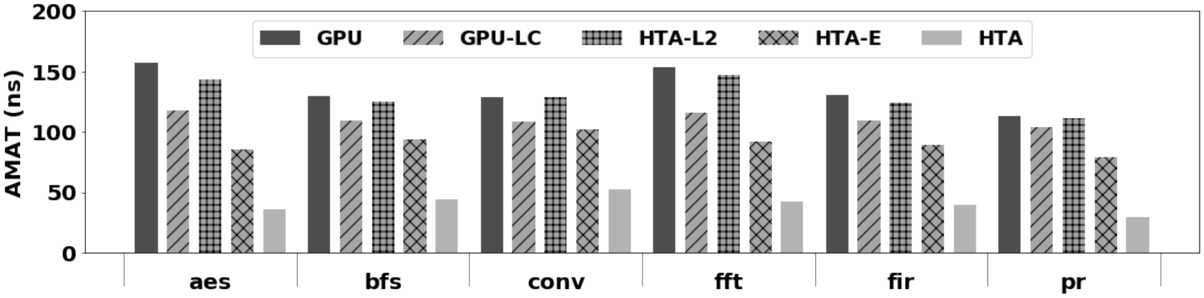


Figure 6.5: Average Memory Access Time (AMAT) for L1 misses. The baseline (GPU) is compared to a GPU with 40MB of last level cache (GPU-LC), a similar system using PMC (HTA-L2), an implementation of HTA using electrocal links (HTA-E), and ultimately the proposed HTA. HTA improves the average L1 miss penalty by 2.3× to 5× compared to the baseline GPU architecture by avoiding data transfers over a high-latency crossbar.

used in PMC is a critical part of this design. While the PMC design improves the average access latency by 10%-30% (*i.e.*, 5-20ns), these improvements can be masked when using a long-latency crossbar (*e.g.*, 50ns). As illustrated in Figure 6.4b, the implementation of PMC using electrical links (PMC-E) improves the tail latency. However, as shown in Figure 6.4a, the average access latency is significantly increased as the result of long-latency electrical links used in this design. This analysis shows the importance of interconnect technology used for our proposal, making SiPh and AWGR the key enablers for this design.

6.3.2 Evaluation of HTA

As the next step, we investigate the performance of proposed HTA system which allows for elimination of the last level cache against the baseline GPU described in Section 6.2, along with a GPU with 40MB of last level cache. In order to analyse different architectural differences between HTA and the baseline, we present evaluate two middle point between the two systems. First, we modeled a system similar to the baseline which utilizes the PMC under the same cache organization (labeled HTA-L2). Moreover, we modeled HTA implemented using electrical interconnects to separate the architectural changes from the benefits gained purely from SiPh technology (labeled HTA-E).

As we discussed earlier in Section 6.2, our trace-based evaluation does not allow us to report runtime numbers. Thus, we choose to report the overall performance of the memory system. Figure 6.5 presents the L1 miss penalty, as a measure of performance of the memory system for both architectures under investigation. Average miss penalty for L1 caches is calculated in the form of AMAT for L1 misses.

As the third bar (HTA-L2) in Figure 6.5 shows, DRAM access latency improvements gained from PMC result in 10-15% reduction in L1 miss penalty. However, the latency-intensive (50 cycles) consult with the last level cache is hiding most of the benefits achieved. With L2 caches eliminated in HTA, all L1 misses are directly added to the CMCs, where requests are transferred over the all-to-all fabric to the MMCs.

Even the HTA system using electrical links (with 50 cycles of latency between CMCs and MMCs) significantly reduces the L1 miss penalty. Taking advantage of low-latency (3 cycles) interconnect fabric enabled by SiPh, HTA reduces the latency cost of L1 misses by 2.3x to 5x.

Reductions on the average miss penalty for L1 caches are mostly obtained through improvements on the 95^{th} percentile in access latency, emphasizing the importance of variations in memory access latency in the overall performance of the memory system for high-throughput accelerators.

The second bar (GPU-LC) represents a GPU with a large (*i.e.*, 40MB) last-level cache, similar to the architectural approach taken by NVIDIA [140], lowering the AMAT by reducing the traffic to DRAM. This approach benefits workloads with high locality. However, as can be seen in Figure 6.5, it will only achieve a small fraction of improvements offered by HTA for irregular HPC workloads with sparse data accesses.

6.3.3 Comparison with Multi-GPU systems

A key motivation for our HTA design is to achieve scalability. Utilizing a 64×64 AWGR, HTA can deliver an accelerator with 256 CUs. The state-of-the-art GPU systems can achieve this scale only by combining multiple GPUs.

For the last part in evaluating HTA, we compared its performance against a multi-GPU system with the same number of compute units (256 CUs). It should be noted that not all the benchmarks provided support for multi-GPU execution, and we only had a few options to run this experiment. Also, we should note that the speedups reported in Figure 6.6 are mainly a lower-bound for what the HTA can achieve. As of today, MGPUSim lacks a memory controller with timing details, and DRAM responses are satisfied at a flat latency. That is the main limiting

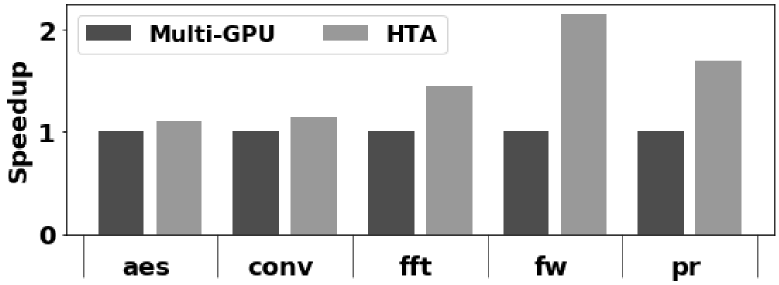


Figure 6.6: The speedup of HTA with 256 CUs compared to a multi-GPU system with 4 GPUs each with 64 CUs. The overhead of data movements in multi-GPU setup result in a speedup of up to 2× for HTA.

factor for us to evaluate PMC in terms of execution time. However, to show the potential benefits of a scalable system enabled by HTA, we modeled a system with the average DRAM access latency measured in DRAMSim for the baseline controller and PMC. This approach does not take into account the benefits of lower variations in memory access achieved by PMC, and does not reveal the full performance potential of HTA.

According to the evaluation results shown in Figure 6.6, HTA can achieve 1.5× speedup on average compared to a multi-GPU system. This improvement is mainly achieved in HTA by avoiding the cross-GPU communication and scheduling overheads in a multi-GPU system.

One interesting observation here is the overhead of a multi-GPU system for different workloads. As can be seen in Figure 6.6, applications like *aes* or *conv* with smaller data sharing between their kernels experience less overhead (~10%) in the multi-GPU system. On the other hand, applications with more inter-kernel data dependencies such as Page Rank (*pr*), *fft*, and Floyd Warshal (*fw*) require more data movements between kernels (running on different GPUs), and result in larger slowdowns (up to 2×) in a multi-GPU setup. These variations depend on both architecture and workload, and impose several barriers in utilizing multi-GPU systems. HTA allows the programmers to migrate their applications to a scalable platform, and avoids considerable performance overheads especially for applications with significant data sharing across different compute units.

6.4 Conclusion

The study presented in this chapter is accepted for publication in the International Supercomputing Conference (ISC-HPC). In this study, we proposed a novel partitioned memory controller (PMC) to reduce the contention in memory system of high-throughput accelerators. Utilizing the PMC design along with a scalable all-to-all optical fabric, we proposed a new high-throughput accelerator. Our simulation results show improvements for PMC on DRAM access latency and memory access divergence, and reduced miss penalty in L1 caches. Our chiplet-based design combines our novel PMC design and SiPh technology to support 4× more compute units.

Given the lack of publicly available area/power models of state-of-the-art GPUs, it is difficult to do a fair and accurate comparison of HTA with GPUs in terms of power and area. However, we can present a qualitative analysis. In terms of power consumption, SiPh links used in this work require 1.65-0.66 pJ/bit depending on the technology node used ranging from 65nm to 14nm. In terms of area overheads, PMC design does not add any logic for queuing as dedicated queues are result of breaking down the single shared queue in the baseline controller. Moreover, the SiPh components used in our design (the AWGR, and SiPh TRXs) have small footprints compared to size of the processor dies (less than 0.01% for typical compute dies [64]).

In this work we have assumed that the compute units in the HTA are similar to that of a GPU. However the proposed HTA architecture can apply to many different types of processors and accelerators. The combination of the significantly lower memory latency and more deterministic memory access time enables unexplored areas for micro-architecture design of advanced computing units and accelerators. Chapter 7 discusses the design space enabled by this work, and directions for exploring this space in the future.

Chapter 7

Conclusions and Future Work

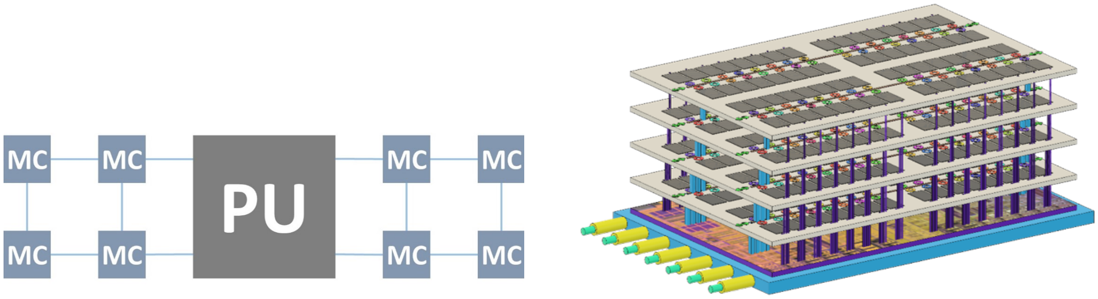
The architectures studied in this dissertation open up a variety of system level designs to be explored. This chapter summarizes the key findings of this work and presents directions for future research.

The architecture presented in Chapter 4 can be the structure of future studies on memory networks with a variety of topologies. The memory architecture shown in Figure 7.1a paves the way towards memory disaggregation by taking advantage of distant-independent energy and latency offered by SiPh links. Investigating this architecture over a variety of applications along with implementation in a full-system cycle-accurate simulator like gem5 [152] can reveal the potential of using optics for processor to memory communication.

Moreover, this idea can be extended to other DRAM architectures (*e.g.* HBM) and increase the parallelism in the memory accesses. In terms of DRAM micro-architecture, as shown in Figure 7.1b WDM-based photonic interconnects can be used to provide direct optical IO to each bank. This allows for much higher concurrency in access and significantly less queuing delays at the memory controller.

Memory accesses in GPUs takes hundreds of cycles to be serviced, and this latency can drastically change during the application execution as different compute units compete for receiving their data through shared memory channels [146]. GPU architects have addressed this issue by increasing the number of contexts executed simultaneously on GPUs [153]. However, this design choice comes with several challenges:

Context Scheduler: Allowing execution of multiple contexts at the same time requires



(a) Memory Network

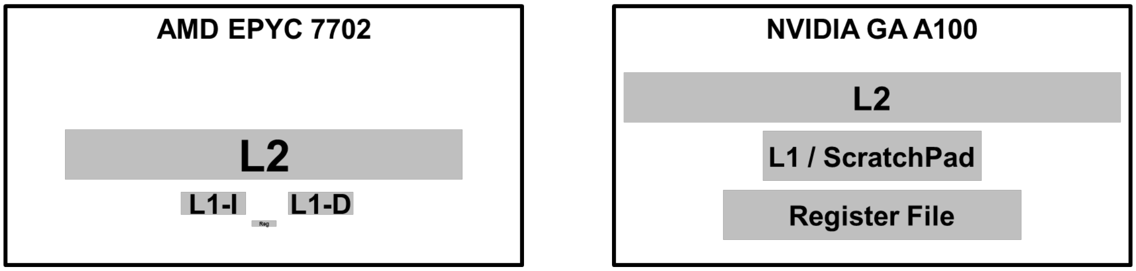
(b) SiPh DRAM

Figure 7.1: (a) Memory disaggregation through an optically-interconnected memory network (b) Fine-grain DRAM with optical IO increasing the parallelism in memory access

dedicated logic to maintain and track the state for each of them. Moreover, based on the state of contexts, additional logic is required to perform scheduling with proper arbitration and decoding units involved [154][155].

Physical Register Files: GPUs rely on large register files to store data required for computation. Providing support for tens of contexts to be executed simultaneously translates in larger register files, scaling almost linearly with the number of contexts supported.

Both area and power dedicated to the operations discussed above are obstacles towards achieving scalability for GPUs [156][157]. In fact, this architectural decision have a significant impact on the overall shape of memory hierarchy in GPUs. Figure 7.2 compares the memory hierarchy in modern GPUs against state-of-the-art CPUs. As shown in Figure 7.2, supporting higher number of contexts in GPUs for hiding the memory latency resulted in significantly larger register files in GPUs compared to CPUs. This architectural impact for compute-focused GPUs (e.g., AMD's Radeon Instinct and NVIDIA's P100/V100/A100 lines) is more relevant for two reasons. Firstly, these systems target a higher compute capacity, and mainly achieve this goal by providing more compute units. Increasing the number of compute units linearly increases the size of physical register file, limiting the scalability in these systems due to overall die size. In addition, increasing the compute capacity expands the bandwidth gap between the two end of the memory system (*i.e.*, the bandwidth offered at the DRAM and the bandwidth at the L0s). In order to fill the gap between the memory bandwidth and the rate SIMD units ask for data, GPU architects are simply increasing the size of the L2s (*e.g.*, by 7× in NVIDIA's Ampere [140]). Given that the cache sizes are in the range of 40 MB, already occupying significant area on GPU



(a) CPU Memory Hierarchy

(b) GPU Memory Hierarchy

Figure 7.2: Memory organization in state-of-the-art CPUs (AMD EPYC 7702) compared to the state-of-the-art GPUs (NVIDIA GA A100). AMD EPYC 7702 has a total of 32 MB L2 cache, 2 MB L1-Data cache, 2 MB L1-Instruction cache, and 180 kB physical register file. NVIDIA GA A100 offers a total of 40 MB L2 cache, 17.3 MB of memory to be allocated as L1 cache and scratchpad, and 27 MB of register file. Increasing number of contexts in GPUs while each requires physical register files results in significantly larger register file size compared to CPUs.

dies, this approach does not seem to be a scalable solution going forward.

The memory architecture proposed in Chapter 6 mitigates the aforementioned overheads by eliminating the L2 cache, lowering the access latency, and improving the predictability in memory access. These benefits can be leveraged towards redesigning the GPU core architecture, especially for workloads with poor locality in their memory access. In particular, this dissertation suggests two architecture to be explored:

A Lightweight GPU Core for Scalability: Reducing the variation in memory access latency mitigates the need for execution of several contexts in parallel. As discussed earlier, GPUs utilize these concurrent contexts to keep their pipelines busy when the memory behavior is unpredictable. Lowering the access latency and improving the predictability in memory access alleviates this architectural requirement and allows GPU cores to operate using fewer contexts since there is less latency (both in absolute terms and in its variations) to hide.

In terms of the core micro-architecture, supporting fewer contexts reduces the total size of physical register files per core, effectively freeing up area and power. Moreover, the logic required for the context scheduler would be less complicated. These optimizations can be combined with the power and area savings from the architecture discussed in Chapter 6 to design a lightweight GPU core with reduced logic footprint and power consumption. Moreover, as discussed in Chapter 6, the proposed SiPh interconnect paves the way towards processor disintegration in GPUs by offering lower energy and performance overheads for data movements compared to state-of-the-art electrical interconnects. In addition, the all-to-all connectivity offered by AWGR can enable sharing of data among different compute units which can be utilized to perform synchronization among different contexts. Such GPU architecture can be utilized to scale the number of core in the system, and achieve a significantly higher compute capacity for the system.

Sparse Compute on GPUs with Large Scratchpads: Unlike the approach discussed above, the GPU core design can take a different approach in utilizing the reduction in memory access variations offered by the HTA architecture presented in Chapter 6. By reducing the total number of context needed for keeping the pipelines full and maintaining the total size of physical register files, we are effectively increasing the amount of local memory per context. In addition, the logic savings from eliminating the L2 caches can also be utilized to further expand the size of local memory available to each core.

This design choice is suitable for applications with sparse memory accesses, where hardware managed caches loose their efficiency in the absence of locality in data accesses. Therefore, a GPU core with large local memory managed through software (*i.e.*, a combination of user-level instructions and compiler optimizations) can handle data placements more efficiently and achieve higher throughput. This architecture can utilize previously proposed techniques [156] to dynamically change the memory allocation among registers, cache, and scratchpad on a per-application basis, and enable more aggressive prefetching of data into register files [157].

Data-Driven Graph Accelerator: Though GPUs, as a successful example employing single-instruction multiple-data (SIMD) model of execution, have proved effective in terms of harnessing data parallelism in many workloads, they may not be the best candidates for graph processing applications. In fact, SIMD execution model naturally requires high data locality for the underlying hardware to be exploited. For graph processing application with poor data locality, this model of execution struggles to deliver high throughput. Today's GPUs implement a wide range of techniques (*e.g.*, aggressive prefetching, coalescing, etc) to fetch the data required for their wide SIMD pipelines. However, these solutions comes with their set of challenges and do not scale well with the rate that data sets are growing. The oracle model of execution for graph applications would be a dataflow architecture where the instructions are

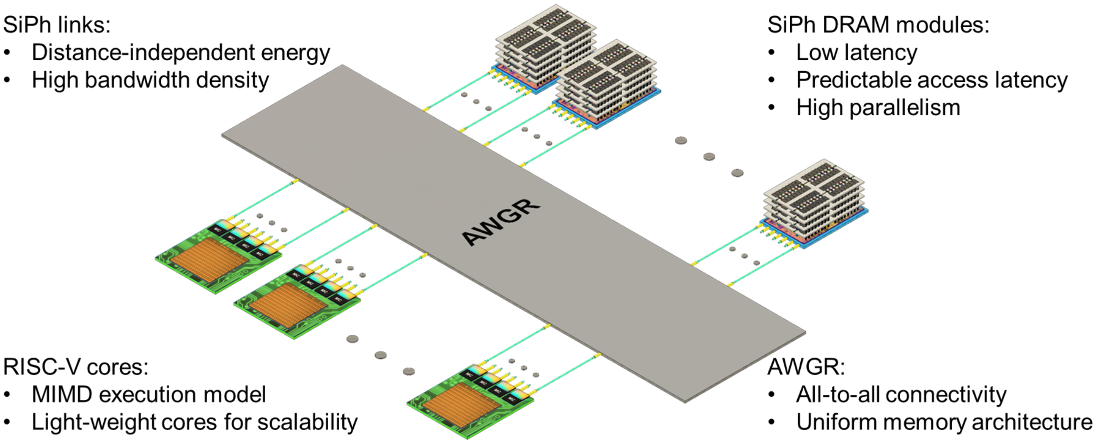

SiPh links:
• Distance-independent energy
• High bandwidth density
SiPh DRAM modules:
• Low latency
• Predictable access latency
• High parallelism
AWGR
RISC-V cores:
• MIMD execution model
• Light-weight cores for scalability
AWGR:
• All-to-all connectivity
• Uniform memory architecture

Figure 7.3: A data-driven accelerator for graph processing workloads utilizing RISC-V processors, SiPh interconnect with a flat topology using AWGR, and SiPh DRAM dies with dedicated IO to each bank.

executed as soon as their data dependencies are met. However, realizing a high-throughput dataflow machine in the presence of memory wall combined with significant overheads of data movement is challenging.

The architecture presented in this dissertation can be combined to design a data-driven accelerator for graph applications. Lowering the access latency, improving the predictability in memory access, and increasing the parallelism in the memory accesses can all come together to enable the shift towards a multiple-instruction multiple-data (MIMD) processor with a flat memory hierarchy. This dissertation envisions the system architecture as shown in Figure 7.3. On each of the processor dies, a cluster of RISC-V processors with large software-managed scratch pad memories are used to perform the arithmetic and logic operations. While each of these RISC-V cores are not high performance cores, their offerings in terms of energy consumption and footprint allow for scalability. On the memory, SiPh DRAM modules with dedicated optical IO to each bank provide data at lower latency with smaller variations compared to traditional DRAM architectures. In terms of communication between the processors and memory dies, SiPh links mitigate the energy and latency overheads of data movements and offer high bandwidth to memory through use if WDM. Utilizing the AWGR as the interconnect fabric provide all-to-all communication between the processors and memory dies and effectively implements a flat memory hierarchy. Last, but by no means the least, the distance independent energy consumption of SiPh links eliminates any area constraints (*e.g.*, imposed by realizable interposer size) or heat dissipation limitations (*e.g.*, imposed by package thermal design point). Therefore, the components of this system (*i.e.*, processor dies, interconnect fabric, and memory dies) can be individually fabricated, packaged, and tested while the physical distance between them has virtually no impact on the overall performance of the system.

7.1 Conclusions

This dissertation investigated the design space of AWGR-based interconnects for both processor and memory systems in HPC domain, and compared them to the state-of-the-art SiPh fabrics and aggressive electrical baselines.

Chapter 3 evaluated the use of AWGR-based SihP Network-on-Chip to enable energy-efficient all-to-all connectivity for large-scale interposer-based HPC systems. Based on the presented results in this chapter, many future studies could be conducted around bit parallelism in AWGRs, dynamic bandwidth allocation, and adaptive laser sources, which can use AWGRs as the physical interconnection fabric and could further improve the total energy efficiency of photonic NoCs.

Chapter 4 studied off-chip memory networks capable of providing tera-bytes of memory capacity by optically interconnecting several 3D stacked DRAM modules. By exploiting SiPhs for integrated optical links, a compelling case for the suitability of AWGRs as a layout-efficient, mature, and highly energy-efficient interconnection fabric between the processor and the vaults inside the MC was made.

Chapter 5 investigated the scaling limitations in chiplet-based systems, in particular, large inter-chiplet NUMA latencies, distance-related energy overheads, and limited IO bandwidth. By exploiting the properties of optical interconnects, it proposed a scalable uniform memory architecture (S-UMA) that overcomes all NUMA-related performance challenges. Simulation results suggest that SiPhs could enable scalable chiplet-based uniform memory architectures and thus be of high importance to scale-up performance and, in turn, reduce the data movement overheads of scaling-out in HPC systems.

Chapter 6 researched the architecture of state-of-the-art GPUs, the impact of memory access latency variations on overall performance and system design, and key challenges in scaling memory and compute capacity in these systems, and proposes a novel GPU architecture called Scale-Up GPU which aims to reduce the contention within the memory system with the help of a partitioned memory controller and an all-to-all passive optical interconnect.

The state-of-the-art computing systems employ increasingly complex hardware and software stacks to meet their performance goals. A large portion of these added complexities is due to limitations at the technology level (*e.g.*, memory wall, pin wall, reticle size, energy cost of data movements, etc). These limitations shift the architects away from their ideal design choices, and result in many compromises at the design stage. The unique properties of SiPh links in terms of their energy-consumption and bandwidth-density can be utilized to change the way we think about computing systems today. This dissertation explored a subset of the design space for computing systems enabled by SiPh, and provided pointers for several studies in the future.

References

[1] John L Hennessy and David A Patterson. A new golden age for computer architecture: Domain-specific hardware/software co-design, enhanced security, open instruction sets, and agile chip development. *Turing Lecture*, 2018.

[2] Jason Lowe-Power. *On Heterogeneous Compute and Memory Systems*. PhD thesis, University of Wisconsin–Madison, 2017.

[3] Milo MK Martin, Mark D Hill, and Daniel J Sorin. Why on-chip cache coherence is here to stay. *Communications of the ACM*, 55(7), 2012.

[4] Christopher J Nitta, Matthew K Farrens, and Venkatesh Akella. On-chip photonic interconnects: A computer architect's perspective. *Synthesis Lectures on Computer Architecture*, 8(5), 2013.

[5] Jian Wang and Yun Long. On-chip silicon photonic signaling and processing: a review. *Science Bulletin*, 2018.

[6] Keren Bergman, Luca P Carloni, Aleksandr Biberman, Johnnie Chan, and Gilbert Hendry. *Photonic network-on-chip design*. Springer, 2014.

[7] Sebastian Werner, Javier Navaridas, and Mikel Luján. Amon: An advanced mesh-like optical NoC. In *IEEE 23rd Annual Symposium on High-Performance Interconnects (HOTI)*. IEEE, 2015.

[8] Christopher Nitta. *Design and analysis of large scale nanophotonic on-chip networks*. University of California, Davis, 2011.

[9] Sebastian Werner, Javier Navaridas, and Mikel Luján. A survey on optical network-on-chip architectures. *ACM Computing Surveys (CSUR)*, 50(6), 2017.

[10] Parisa Khadem Hamedani, Natalie Enright Jerger, and Shaahin Hessabi. Qut: A low-power optical network-on-chip. In *Eighth IEEE/ACM International Symposium on Networks-on-Chip (NoCS)*. IEEE, 2014.

[11] S Kamei, M Ishii, M Itoh, T Shibata, Y Inoue, and T Kitagawa. 64× 64-channel uniform-loss and cyclic-frequency arrayed-waveguide grating router module. *Electronics Letters*, 39(1), 2003.

[12] Kuanping Shang, Shibnath Pathak, Chuan Qin, and SJ Ben Yoo. Low-loss compact silicon nitride arrayed waveguide gratings for photonic integrated circuits. *IEEE Photonics Journal*, 9(5), 2017.

[13] Chen Sun, Chia-Hsin Owen Chen, George Kurian, Lan Wei, Jason Miller, Anant Agarwal, Li-Shiuan Peh, and Vladimir Stojanovic. DSENT-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling. In *Sixth IEEE/ACM International Symposium on Networks on Chip (NoCS)*. IEEE, 2012.

[14] Ishan G Thakkar, Sai Vineel Reddy Chittamuru, and Sudeep Pasricha. Improving the reliability and energy-efficiency of high-bandwidth photonic NoC architectures with multilevel signaling. In *Proceedings of the Eleventh IEEE/ACM International Symposium on Networks-on-Chip*. ACM, 2017.

[15] S. Moazeni, S. Lin, M. Wade, L. Alloatti, R. J. Ram, M. Popovi, and V. Stojanovi. A 40-Gb/s PAM-4 transmitter based on a ring-resonator optical DAC in 45-nm SOI CMOS. *IEEE Journal of Solid-State Circuits*, 52(12), Dec 2017.

[16] Yang Zhang, Amir Hosseini, Xiaochuan Xu, David Kwong, and Ray T Chen. Ultralow-loss silicon waveguide crossing using Bloch modes in index-engineered cascaded multimode-interference couplers. *Optics letters*, 38(18), 2013.

[17] Tiehui Su, Guangyao Liu, Katherine E Badham, Samuel T Thurman, Richard L Kendrick, Alan Duncan, Danielle Wuchenich, Chad Ogden, Guy Chriqui, Shaoqi Feng, et al. Interferometric imaging using $Si_3\ N_4$ photonic integrated circuits for a SPIDER imager. *Optics Express*, 26(10), 2018.

[18] Paolo Grani, Gengchen Liu, Roberto Proietti, and SJ Ben Yoo. Bit-parallel all-to-all and flexible AWGR-based optical interconnects. In *Optical Fiber Communications Conference and Exhibition (OFC)*. IEEE, 2017.

[19] Yangdong Deng and Wojciech P Maly. Interconnect characteristics of 2.5-D system integration scheme. In *Proceedings of the 2001 international symposium on Physical design*, 2001.

[20] Natalie Enright Jerger, Ajaykumar Kannan, Zimo Li, and Gabriel H Loh. Noc architectures for silicon interposer systems: Why pay for more wires when you can get them (from your interposer) for free? In *47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2014.

[21] Ajaykumar Kannan, Natalie Enright Jerger, and Gabriel H Loh. Enabling interposer-based disintegration of multi-core processors. In *48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2015.

[22] Ajaykumar Kannan, Natalie Enright Jerger, and Gabriel H Loh. Exploiting interposer technologies to disintegrate and reintegrate multicore processors. *IEEE Micro*, 36(3), 2016.

[23] NVIDIA. NVIDIA tesla V100 GPU architecture. `http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf`, 2017. [Online; accessed 06-05-2021].

[24] Joe Macri. AMD's next generation GPU and high bandwidth memory architecture: FURY. In *IEEE Hot Chips 27 Symposium (HCS)*. IEEE, 2015.

[25] Gabriel H Loh, Natalie Enright Jerger, Ajaykumar Kannan, and Yasuko Eckert. Interconnect-memory challenges for multi-chip, silicon interposer systems. In *Proceedings of the 2015 International Symposium on Memory Systems*. ACM, 2015.

[26] WikiChip. Zen - microarchitectures - AMD. `https://en.wikichip.org/wiki/amd/microarchitectures/zen#Die-die_memory_latencies`, . [Online; accessed 06-05-2021].

[27] Akhil Arunkumar, Evgeny Bolotin, Benjamin Cho, Ugljesa Milic, Eiman Ebrahimi, Oreste Villa, Aamer Jaleel, Carole-Jean Wu, and David Nellans. MCM-GPU: Multi-chip-module GPUs for continued performance scalability. *ACM SIGARCH Computer Architecture News*, 45(2), 2017.

[28] Furkan Eris, Ajay Joshi, Andrew B Kahng, Yenai Ma, Saiful Mojumder, and Tiansheng Zhang. Leveraging thermally-aware chiplet organization in 2.5 D systems to reclaim dark silicon. In *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2018.

[29] Luca Ramini, Paolo Grani, Sandro Bartolini, and Davide Bertozzi. Contrasting wavelength-routed optical NoC topologies for power-efficient 3D-stacked multicore processors using physical-layer analysis. In *Proceedings of the Conference on Design, Automation and Test in Europe*. EDA Consortium, 2013.

[30] S. Moazeni, S. Lin, M. Wade, L. Alloatti, R. J. Ram, M. Popović, and V. Stojanović. A 40-Gb/s PAM-4 smitter based on a ring-resonator optical DAC in 45-nm SOI CMOS. *IEEE Journal of Solid-State Circuits*, 52(12), 2017.

[31] Xian Xiao, Yu Zhang, Roberto Proietti, and SJB Yoo. Scalable AWGR-based all-to-all optical interconnects with 2.5 D/3D integrated optical interposers. In *2018 IEEE Photonics Society Summer Topical Meeting Series (SUM)*. IEEE, 2018.

[32] Thiruvengadam Vijayaraghavany, Yasuko Eckert, Gabriel H Loh, Michael J Schulte, Mike Ignatowski, Bradford M Beckmann, William C Brantley, Joseph L Greathouse, Wei Huang, Arun Karunanithi, et al. Design and analysis of an APU for exascale computing. In *IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2017.

[33] Kevin Tran and June Ahn. HBM: Memory solution for high performance processors. *Proceedings of MemCon*, 2014.

[34] Yvain Thonnart and Mounir Zid. Technology assessment of silicon interposers for manycore SoCs: Active, passive, or optical? In *Eighth IEEE/ACM International Symposium on Networks-on-Chip (NoCS)*. IEEE, 2014.

[35] Sebastian Werner, Javier Navaridas, and Mikel Luján. Designing low-power, low-latency networks-on-chip by optimally combining electrical and optical links. In *IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2017.

[36] Yigit Demir, Yan Pan, Seukwoo Song, Nikos Hardavellas, John Kim, and Gokhan Memik. Galaxy: A high-performance energy-efficient multi-chip architecture using photonic interconnects. In *Proceedings of the 28th ACM international conference on Supercomputing*. ACM, 2014.

[37] Yu Zhang, Xian Xiao, Kaiqi Zhang, Siwei Li, Anirban Samanta, Yichi Zhang, Kuanping Shang, Roberto Proietti, Katsunari Okamoto, and SJ Ben Yoo. Foundry-enabled scalable all-to-all optical interconnects using silicon nitride arrayed waveguide router interposers and silicon photonic transceivers. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(5), 2019.

[38] R Proietti, X Xiao, K Zhang, G Liu, H Lu, P Fotouhi, J Messig, and SJB Yoo. Experimental demonstration of a 64-port wavelength routing thin-CLOS system for data center switching architectures. *Journal of Optical Communications and Networking*, 10 (7), 2018.

[39] Ajay Joshi, Christopher Batten, Yong-Jin Kwon, Scott Beamer, Imran Shamim, Krste Asanovic, and Vladimir Stojanovic. Silicon-photonic clos networks for global on-chip communication. In *2009 3rd ACM/IEEE International Symposium on Networks-on-Chip*. IEEE, 2009.

[40] Cheng Li, Mark Browning, Paul V Gratz, and Samuel Palermo. LumiNOC: A power-efficient, high-performance, photonic network-on-chip. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 33(6), 2014.

[41] Ian O'Connor, Matthieu Briere, Emmanuel Drouard, Art Kazmierczak, Faress Tissafi-Drissi, David Navarro, Fabien Mieyeville, Joni Dambre, Dirk Stroobandt, Jean-Marc Fedeli, et al. Towards reconfigurable optical networks on chip. volume 5, 2005.

[42] M. Ortín-Obón, M. Tala, L. Ramini, V. Viñals-Yufera, and D. Bertozzi. Contrasting laser power requirements of wavelength-routed optical NoC topologies subject to the floorplanning, placement, and routing constraints of a 3-D-stacked system. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 25(7), July 2017.

[43] Taizhong Huang, Xiaoyuan Liu, Hai Zhang, Senming Gong, and Anmin Zhang. Athermal arrayed waveguide grating wavelength division multiplexer, 2016. US Patent 9,519,103.

[44] Li Li, Pierre Chia, Paul Ton, Mohan Nagar, Sada Patil, Jie Xue, Javier Delacruz, Marius Voicu, Jack Hellings, Bill Isaacson, et al. 3D SiP with organic interposer for ASIC and memory integration. In *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*. IEEE, 2016.

[45] Wim Heirman, Trevor Carlson, and Lieven Eeckhout. Sniper: Scalable and accurate parallel multi-core simulation. In *8th International Summer School on Advanced Computer Architecture and Compilation for High-Performance and Embedded Systems (ACACES-2012)*. High-Performance and Embedded Architecture and Compilation Network of Excellence (HiPEAC), 2012.

[46] Xusheng Zhan, Yungang Bao, Christian Bienia, and Kai Li. PARSEC3.0: A multicore benchmark suite with network stacks and SPLASH-2x. *ACM SIGARCH Computer Architecture News*, 44(5), 2017.

[47] Niket Agarwal, Tushar Krishna, Li-Shiuan Peh, and Niraj K Jha. GARNET: A detailed on-chip network model inside a full-system simulator. In *IEEE International Symposium on Performance Analysis of Systems and Software*. IEEE, 2009.

[48] Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R Hower, Tushar Krishna, Somayeh Sardashti, et al. The gem5 simulator. *ACM SIGARCH Computer Architecture News*, 39(2), 2011.

[49] Meisam Bahadori, Sébastien Rumley, Dessislava Nikolova, and Keren Bergman. Comprehensive design space exploration of silicon photonic interconnects. *Journal of Lightwave Technology*, 34(12), 2016.

[50] Masahiro Nada, Shigeru Kanazawa, Hiroshi Yamazaki, Yasuhiko Nakanishi, Wataru Kobayashi, Yoshiyuki Doi, Takaharu Ohyama, Tetsuichiro Ohno, Kiyoto Takahata, Toshikazu Hashimoto, et al. High-linearity avalanche photodiode for 40-km transmission with 28-Gbaud PAM4. In *Optical Fiber Communication Conference*. Optical Society of America, 2015.

[51] Yu Zhang, Kuanping Shang, Yi-Chun Ling, and SJ Ben Yoo. 3D integrated silicon photonic unit cell with vertical U-turn for scalable optical phase array. In *2018 Conference on Lasers and Electro-Optics (CLEO)*. IEEE, 2018.

[52] George Kurian, Jason E Miller, James Psota, Jonathan Eastep, Jifeng Liu, Jurgen Michel, Lionel C Kimerling, and Anant Agarwal. Atac: a 1000-core cache-coherent processor with on-chip optical network. In *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*. ACM, 2010.

[53] Yan Pan, Prabhat Kumar, John Kim, Gokhan Memik, Yu Zhang, and Alok Choudhary. Firefly: Illuminating future network-on-chip with nanophotonics. In *ACM SIGARCH Computer Architecture News*, volume 37. ACM, 2009.

[54] Shirish Bahirat and Sudeep Pasricha. Meteor: Hybrid photonic ring-mesh network-on-chip for multicore architectures. *ACM Transactions on Embedded Computing Systems (TECS)*, 13(3s), 2014.

[55] Dana Vantrease, Robert Schreiber, Matteo Monchiero, Moray McLaren, Norman P Jouppi, Marco Fiorentino, Al Davis, Nathan Binkert, Raymond G Beausoleil, and Jung Ho Ahn. Corona: System implications of emerging nanophotonic technology. In *ACM SIGARCH Computer Architecture News*, volume 36. IEEE Computer Society, 2008.

[56] Sebastian Werner, Pouya Fotouhi, Roberto Proietti, Xian Xiao, and SJ Ben Yoo. Towards energy-efficient high-throughput photonic NoCs for 2.5D integrated systems: A case for

AWGRs. In *2018 Twelfth IEEE/ACM International Symposium on Networks-on-Chip (NOCS)*. IEEE, 2018.

[57] William James Dally and Brian Patrick Towles. *Principles and practices of interconnection networks*. Elsevier, 2004.

[58] Ritesh Parikh, Reetuparna Das, and Valeria Bertacco. Power-aware NoCs through routing and topology reconfiguration. In *Proceedings of the 51st Annual Design Automation Conference*. ACM, 2014.

[59] Reetuparna Das, Satish Narayanasamy, Sudhir Satpathy, and Ronald G Dreslinski. Catnap: energy proportional multiple network-on-chip. In *ISCA*, 2013.

[60] Stanley Cheung, Tiehui Su, Katsunari Okamoto, and SJB Yoo. Ultra-compact silicon photonic 512× 512 25 ghz arrayed waveguide grating router. *IEEE Journal of Selected Topics in Quantum Electronics*, 20(4), 2014.

[61] Yigit Demir and Nikos Hardavellas. SLaC: Stage laser control for a flattened butterfly network. In *HPCA*. IEEE, 2016.

[62] Géza Kurczveil, Di Liang, Marco Fiorentino, and Raymond G Beausoleil. Robust hybrid quantum dot laser for integrated silicon photonics. *Optics express*, 24(14), 2016.

[63] Geza Kurczveil, Chong Zhang, Antoine Descos, Di Liang, Marco Fiorentino, and Raymond Beausoleil. On-chip hybrid silicon quantum dot comb laser with 14 error-free channels. In *2018 IEEE International Semiconductor Laser Conference (ISLC)*. IEEE, 2018.

[64] P. Fotouhi, S. Werner, R. Proietti, X. Xiao, and S. J. Ben Yoo. Enabling scalable disintegrated computing systems with AWGR-based 2.5D interconnection networks. *IEEE/OSA Journal of Optical Communications and Networking*, 11(7):333–346, 2019.

[65] M. Poremba, I. Akgun, J. Yin, O. Kayiran, Y. Xie, and G. H. Loh. There and back again: Optimizing the interconnect in networks of memory cubes. In *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, June 2017.

[66] P. Stanley-Marbell, V. C. Cabezas, and R. P. Luijten. Pinned to the walls impact of packaging and application properties on the memory and power walls. In *IEEE/ACM International Symposium on Low Power Electronics and Design*, Aug 2011.

[67] Patrick Siegl, Rainer Buchty, and Mladen Berekovic. Data-centric computing frontiers: A survey on processing-in-memory. In *MEMSYS*. ACM, 2016.

[68] JEDEC. High bandwidth memory (HBM) DRAM. `https://www.jedec.org/standards-documents/docs/jesd235a`, 2015. [Online; accessed 06-05-2021].

[69] JEDEC. High bandwidth memory (HBM2) DRAM. `https://www.jedec.org/news/pressreleases/jedec-updates-groundbreaking-high-bandwidth-memory-hbm-standard`, 2016. [Online; accessed 06-05-2021].

[70] Micron. Hybrid memory cube. https://web.archive.org/web/20130513053443/http://www.hybridmemorycube.org/files/SiteDownloads/HMC_Specification%201_0.pdf, 2013. [Online; accessed 06-05-2021].

[71] Avinash Sodani. Knights landing (knl): 2nd generation intel® Xeon phi processor. In *IEEE Hot Chips 27 Symposium (HCS)*, pages 1–24. IEEE, 2015.

[72] Business Wire. Hybrid memory cube (HMC) and high-bandwidth memory (HBM global market report (2018-2023)). https://www.businesswire.com/news/home/20180312005484/en/Hybrid-Memory-Cube-HMC-High-bandwidth-Memory-HBM, 2015. [Online; accessed 06-05-2021].

[73] Jia Zhan, Itir Akgun, Jishen Zhao, Al Davis, Paolo Faraboschi, Yuangang Wang, and Yuan Xie. A unified memory network architecture for in-memory computing in commodity servers. In *MICRO*, pages 1–14. IEEE, 2016.

[74] Zhehui Wang, Zhengbin Pang, Peng Yang, Jiang Xu, Xuanqi Chen, Rafael KV Maeda, Zhifei Wang, Luan HK Duong, Haoran Li, and Zhe Wang. MOCA: An inter/intra-chip optical network for memory. In *DAC*. IEEE, 2017.

[75] Gwangsun Kim, John Kim, Jung Ho Ahn, and Jaeha Kim. Memory-centric system interconnect design with hybrid memory cubes. In *PACT*, pages 145–156. IEEE Press, 2013.

[76] Micron. Hybrid memory cube specification 2.1. https://web.archive.org/web/20170202004433/http://hybridmemorycube.org/files/SiteDownloads/HMC-30G-VSR_HMCC_Specification_Rev2.1_20151105.pdf, 2014. [Online; accessed 06-05-2021].

[77] Ke Wen, Hang Guan, David M Calhoun, Sebastien Rumley, Keren Bergman, David Donofrio, and John Shalf. Silicon photonic memory interconnect for many-core architectures. In *HPEC*. IEEE, 2016.

[78] Scott Beamer et al. Re-architecting DRAM memory systems with monolithically integrated silicon photonics. In *ACM SIGARCH Computer Architecture News*, volume 38. ACM, 2010.

[79] Paolo Grani, Roberto Proietti, Venkatesh Akella, and SJ Ben Yoo. Design and evaluation of AWGR-based photonic NoC architectures for 2.5 D integrated high performance computing systems. In *IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2017.

[80] Yasuhiko Arakawa et al. Silicon photonics for next generation system integration platform. *IEEE Communications Magazine*, 51(3), 2013.

[81] Scott Beamer, Krste Asanović, Christopher Batten, Ajay Joshi, and Vladimir Stojanović. Designing multi-socket systems using silicon photonics. In *ICS*. ACM, 2009.

[82] Ashok V Krishnamoorthy, Ron Ho, Xuezhe Zheng, Herb Schwetman, Jon Lexau, Pranay Koka, GuoLiang Li, Ivan Shubin, and John E Cunningham. Computer systems based on silicon photonic interconnects. *Proceedings of the IEEE*, 97(7), 2009.

[83] Dong J Shin, Kwan S Cho, H-C Ji, Beom S Lee, Sung G Kim, Jin K Bok, Sang H Choi, Yong H Shin, Jung H Kim, Shin Y Lee, et al. Integration of silicon photonics into DRAM process. In *OFC/NFOEC*. IEEE, 2013.

[84] Chen Sun, Mark T Wade, Yunsup Lee, Jason S Orcutt, Luca Alloatti, Michael S Georgas, Andrew S Waterman, Jeffrey M Shainline, Rimas R Avizienis, Sen Lin, et al. Single-chip microprocessor that communicates directly using light. *Nature*, 528(7583), 2015.

[85] Luca Ramini and Davide Bertozzi. Power efficiency of wavelength-routed optical NoC topologies for global connectivity of 3D multi-core processors. In *NoCs*. ACM, 2012.

[86] J Thomas Pawlowski. Hybrid memory cube: breakthrough DRAM performance with a fundamentally re-architected DRAM subsystem. In *Hot Chips*, volume 23, 2011.

[87] Christian Bienia, Sanjeev Kumar, Jaswinder Pal Singh, and Kai Li. The PARSEC benchmark suite: Characterization and architectural implications. In *Proceedings of the 17th international conference on Parallel architectures and compilation techniques*. ACM, 2008.

[88] Steven Cameron Woo, Moriyoshi Ohara, Evan Torrie, Jaswinder Pal Singh, and Anoop Gupta. The SPLASH-2 programs: Characterization and methodological considerations. In *ACM SIGARCH computer architecture news*, volume 23. ACM, 1995.

[89] Wim Bogaerts and Shankar Kumar Selvaraja. Compact single-mode silicon hybrid rib/strip waveguide with adiabatic bends. *IEEE Photonics Journal*, 3(3), 2011.

[90] Sebastian Werner, Pouya Fotouhi, Roberto Proietti, and S. J. Ben Yoo. AWGR-based optical processor-to-memory communication for low-latency, low-energy vault accesses. In *Proceedings of the International Symposium on Memory Systems*, MEMSYS '18, New York, NY, USA, 2018. Association for Computing Machinery.

[91] Sujal Das. It's time for disaggregated silicon! `https://www.netronome.com/blog/its-time-disaggregated-silicon/`, 2018. [Online; accessed 06-05-2021].

[92] Anton Shilov. AMD previews epyc 'rome' processor: Up to 64 zen 2 cores. `https://www.anandtech.com/show/13561/amd-previews-epyc-rome-processor-up-to-64-zen-2-cores`, 2018. [Online; accessed 06-05-2021].

[93] Noah Beck, Sean White, Milam Paraschou, and Samuel Naffziger. Zeppelin: An SoC for multichip architectures. In *ISSCC*. IEEE, 2018.

[94] Deborah Patterson, Isabel De Sousa, and Lousi-Marie Achard. The future of packaging with silicon photonics. *Chip Scale Review*, 21(1), 2017.

[95] Intel. Embedded multi-die interconnect bridge (EMIB). https://www.intel.com/content/www/us/en/foundry/emib-an-interview-with-babak-sabi.html. [Online; accessed 06-05-2021].

[96] Pranay Koka, Michael O McCracken, Herb Schwetman, Xuezhe Zheng, Ron Ho, and Ashok V Krishnamoorthy. Silicon-photonic network architectures for scalable, power-efficient multi-chip systems. In *ACM SIGARCH Computer Architecture News*, volume 38. ACM, 2010.

[97] ibm. IBM zenterprise 196 technical guide. (2011). https://www.redbooks.ibm.com/redbooks/pdfs/sg247833.pdf, 2011. [Online; accessed 06-05-2021].

[98] M. Doggett. Xenos: XBOX360 GPU (2005). http://fileadmin.cs.lth.se/cs/Personal/Michael_Doggett/talks/eg05-xenos-doggett.pdf. [Online; accessed 06-05-2021].

[99] Nintendo. Wii U announcement. http://iwataasks.nintendo.com/interviews/#/wiiu/console/0/0. [Online; accessed 06-05-2021].

[100] Jieming Yin, Zhifeng Lin, Onur Kayiran, Matthew Poremba, Muhammad Shoaib Bin Altaf, Natalie Enright Jerger, and Gabriel H Loh. Modular routing design for chiplet-based systems. In *ISCA*. IEEE, 2018.

[101] Sergey Shumarayev. Stratix 10: Intel's 14nm heterogeneous FPGA system-in-package (SiP) platform. In *HC29*. IEEE, 2017.

[102] Ian Cutress. Naples, rome, milan, zen 4: An interview with AMD CTO, mark papermaster. https://www.anandtech.com/show/13578/naples-rome-milan-zen-4-an-interview-with-amd-cto-mark-papermaster, . [Online; accessed 06-05-2021].

[103] John W Poulton, William J Dally, Xi Chen, John G Eyles, Thomas H Greer, Stephen G Tell, John M Wilson, and C Thomas Gray. A 0.54 pj/b 20 Gb/s ground-referenced single-ended short-reach serial link in 28 nm CMOS for advanced packaging applications. *J. Solid-State Circuits*, 48(12), 2013.

[104] Dylan Stow, Yuan Xie, Taniya Siddiqua, and Gabriel H Loh. Cost-effective design of scalable high-performance systems using active and passive interposers. In *ICCAD*. IEEE Press, 2017.

[105] Ahmad Usman, Etizaz Shah, Nithanth B Satishprasad, Jialou Chen, Steven A Bohlemann, Sajjad H Shami, Ali A Eftekhar, and Ali Adibi. Interposer technologies for high-performance applications. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 7(6), 2017.

[106] Suresh Ramalingam. 3D-ICs: Advances in the industry. In *Proc. IEEE 64th ECTC*, 2014.

[107] Stéphane Bellenger et al. Silicon interposers with integrated passive devices: Ultra-miniaturized solution using 2.5 d packaging platform. *Caen, France, IPDiA White Paper Silicon Interposers_260214*, 2014.

[108] Alexandre Ayres de Sousa. *3D Monolithic Integration: performance, Power and Area Evaluation for 14nm and beyond*. PhD thesis, Université Grenoble Alpes, 2017.

[109] Suresh Ramalingam. HBM package integration: Technology trends, challenges and applications. In *Hot Chips 28 Symposium (HCS), 2016 IEEE*. IEEE, 2016.

[110] Eric Beyne. The 3-d interconnect technology landscape. *IEEE Design & Test*, 33(3), 2016.

[111] David Greenhill et al. A 14nm 1Ghz FPGA with 2.5 D transceiver integration. In *ISSCC*. IEEE, 2017.

[112] Subramanian S Iyer. Heterogeneous integration using the silicon interconnect fabric. In *2018 IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM)*. IEEE, 2018.

[113] Behzad Dehlaghi and Anthony Chan Carusone. A 0.3 pj/bit 20 Gb/s/wire parallel interface for die-to-die communication. *IEEE Journal of Solid-State Circuits*, 51(11), 2016.

[114] Kunal Korgaonkar et al. Density tradeoffs of non-volatile memory as a replacement for sram based last level cache. In *ISCA*. IEEE, 2018.

[115] Chen Sun, Mark Wade, Michael Georgas, Sen Lin, Luca Alloatti, Benjamin Moss, Rajesh Kumar, Amir Atabaki, Fabio Pavanello, Rajeev Ram, et al. A 45nm soi monolithic photonics chip-to-chip link with bit-statistics-based resonant microring thermal tuning. In *2015 Symposium on VLSI Circuits (VLSI Circuits)*. IEEE, 2015.

[116] Michael Georgas, BR Moss, Chen Sun, Jeffrey Shainline, JS Orcutt, M Wade, Y-H Chen, Kareem Nammari, JC Leu, Aravind Srinivasan, et al. A monolithically-integrated optical transmitter and receiver in a zero-change 45nm SOI process. In *VLSI Circuits Digest of Technical Papers, 2014 Symposium on*. IEEE, 2014.

[117] Cheng Li, Rui Bai, Ayman Shafik, Ehsan Zhian Tabasy, Geng Tang, Chao Ma, Chin-Hui Chen, Zhen Peng, Marco Fiorentino, Patrick Chiang, et al. A ring-resonator-based silicon photonics transceiver with bias-based wavelength stabilization and adaptive-power-sensitivity receiver. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International*. IEEE, 2013.

[118] Rajeev J Ram. Photonic-electronic integration with polysilicon photonics in bulk CMOS. In *Silicon Photonics X*, volume 9367. International Society for Optics and Photonics, 2015.

[119] Balaram Sinharoy, JA Van Norstrand, Richard J Eickemeyer, Hung Q Le, Jens Leenstra, Dung Q Nguyen, B Konigsburg, K Ward, MD Brown, José E Moreira, et al. IBM POWER8 processor core microarchitecture. *IBM Journal of Research and Development*, 59(1), 2015.

[120] Stephen Phillips. M7: Next generation sparc. *IEEE Hot Chips*, 2014.

[121] Fabien Gaud, Baptiste Lepers, Jeremie Decouchant, Justin Funston, Alexandra Fedorova, and Vivien Quema. Large pages may be harmful on NUMA systems. In *2014 USENIX Annual Technical Conference (USENIX ATC 14)*, pages 231–242, Philadelphia, PA, 2014. USENIX Association.

[122] Ian Cutress. Intel launches stratix-10-tx leveraging EMIB with 58G transceivers. `https://www.anandtech.com/show/12477/intel-launches-stratix-10-tx-leveraging-emib-with-58g-transceivers-`, . [Online; accessed 06-05-2021].

[123] Roger Dangel, Antonio La Porta, Daniel Jubin, Folkert Horst, Norbert Meier, Marc Seifried, and Bert J Offrein. Polymer waveguides enabling scalable low-loss adiabatic optical coupling for silicon photonics. *IEEE Journal of Selected Topics in Quantum Electronics*, 24(4), 2018.

[124] Roger Dangel, Jens Hofrichter, Folkert Horst, Daniel Jubin, Antonio La Porta, Norbert Meier, Ibrahim Murat Soganci, Jonas Weiss, and Bert Jan Offrein. Polymer waveguides for electro-optical integration in data centers and high-performance computers. *Optics express*, 23(4), 2015.

[125] Jeppix. Cost roadmap. `http://www.pics4all.jeppix.eu/public/downloads/Roadmaps/Roadmap_2018_OFC_Preview.pdf`. [Online; accessed 06-05-2021].

[126] David H Bailey. Nas parallel benchmarks. In *Encyclopedia of Parallel Computing*. Springer, 2011.

[127] Shuai Che et al. Rodinia: A benchmark suite for heterogeneous computing. In *IISWC*. Ieee, 2009.

[128] Hao Li et al. A 25 Gb/s, 4.4 V-swing, AC-coupled ring modulator-based WDM transmitter with wavelength stabilization in 65 nm CMOS. *IEEE Journal of Solid-State Circuits*, 50(12), 2015.

[129] Kunzhi Yu et al. A 25 Gb/s hybrid-integrated silicon photonic source-synchronous receiver with microring wavelength stabilization. *IEEE Journal of Solid-State Circuits*, 51 (9), 2016.

[130] Yanfei Chen, Masaya Kibune, Asako Toda, Akinori Hayakawa, Tomoyuki Akiyama, Shigeaki Sekiguchi, Hiroji Ebe, Nobuhiro Imaizumi, Tomoyuki Akahoshi, Suguru Akiyama, et al. 22.2 a 25Gb/s hybrid integrated silicon photonic transceiver in 28nm

CMOS and SOI. In *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*. IEEE, 2015.

[131] Molly Piels, Jared F Bauters, Michael L Davenport, Martijn JR Heck, and John E Bowers. Low-loss silicon nitride awg demultiplexer heterogeneously integrated with hybrid iii–v/silicon photodetectors. *Journal of Lightwave Technology*, 32(4):817–823, 2014.

[132] WikiChip. EPYC 7601 - AMD. `https://en.wikichip.org/wiki/amd/epyc/7601`, . [Online; accessed 06-05-2021].

[133] Yigit Demir and Nikos Hardavellas. Energy-proportional photonic interconnects. *ACM Transactions on Architecture and Code Optimization (TACO)*, 13(4), 2016.

[134] Yigit Demir and Nikos Hardavellas. Ecolaser: an adaptive laser control for energy-efficient on-chip photonic interconnects. In *2014 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE, 2014.

[135] Soumyajit Poddar, Prasun Ghosal, Priyajit Mukherjee, Suman Samui, and Hafizur Rahaman. Design of an noc with on-chip photonic interconnects using adaptive cdma links. In *S2012 IEEE International OC Conference (SOCC)*. IEEE, 2012.

[136] David Thomson et al. Roadmap on silicon photonics. *Journal of Optics*, 18(7), 2016.

[137] Pouya Fotouhi, Sebastian Werner, Jason Lowe-Power, and SJ Ben Yoo. Enabling scalable chiplet-based uniform memory architectures with silicon photonics. In *Proceedings of the International Symposium on Memory Systems*, pages 222–334, 2019.

[138] Akhil Arunkumar, Evgeny Bolotin, David Nellans, and Carole-Jean Wu. Understanding the future of energy efficiency in multi-module gpus. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2019.

[139] TSMC. Enhancing the CoWoS platform. `https://pr.tsmc.com/english/news/2026`, 2020. [Online; accessed 06-05-2021].

[140] NVIDIA. A100 tensor core GPU architecture. `https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf`. [Online; accessed 06-05-2021].

[141] AMD. Introducing RDNA architecture. `https://www.amd.com/system/files/documents/rdna-whitepaper.pdf`, 2019. [Online; accessed 06-05-2021].

[142] AMD. Introducing AMD CDNA architecture. `https://www.amd.com/system/files/documents/amd-cdna-whitepaper.pdf`, 2020. [Online; accessed 06-05-2021].

[143] David AB Miller. Device requirements for optical interconnects to silicon chips. *Proceedings of the IEEE*, 97(7), 2009.

[144] Mark Wade, Erik Anderson, Shahab Ardalan, Pavan Bhargava, Sidney Buchbinder, Michael L Davenport, John Fini, Haiwei Lu, Chen Li, Roy Meade, et al. TeraPHY: A chiplet technology for low-power, high-bandwidth in-package optical I/O. *IEEE Micro*, 40(2), 2020.

[145] Stanley Cheung, Tiehui Su, Katsunari Okamoto, and SJB Yoo. Ultra-compact silicon photonic 512× 512 25 ghz arrayed waveguide grating router. *IEEE Journal of Selected Topics in Quantum Electronics*, 20(4):310–316, 2013.

[146] Zhe Jia, Marco Maggioni, Benjamin Staiger, and Daniele P Scarpazza. Dissecting the NVIDIA Volta GPU architecture via microbenchmarking. *arXiv preprint arXiv:1804.06826*, 2018.

[147] Scott Rixner, William J Dally, Ujval J Kapasi, Peter Mattson, and John D Owens. Memory access scheduling. *ACM SIGARCH Computer Architecture News*, 28(2), 2000.

[148] Shang Li, Zhiyuan Yang, Dhriaj Reddy, Ankur Srivastava, and Bruce Jacob. DRAMsim3: a cycle-accurate, thermal-capable DRAM simulator. *IEEE Computer Architecture Letters*, 2020.

[149] Yifan Sun, Trinayan Baruah, Saiful A Mojumder, Shi Dong, Xiang Gong, Shane Treadway, Yuhui Bao, Spencer Hance, Carter McCardwell, Vincent Zhao, et al. MGPUSim: enabling multi-GPU performance modeling and optimization. In *Proceedings of the 46th International Symposium on Computer Architecture*, 2019.

[150] Yifan Sun, Xiang Gong, Amir Kavyan Ziabari, Leiming Yu, Xiangyu Li, Saoni Mukherjee, Carter McCardwell, Alejandro Villegas, and David Kaeli. Hetero-mark, a benchmark suite for CPU-GPU collaborative computing. In *2016 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 2016.

[151] Anthony Danalis, Gabriel Marin, Collin McCurdy, Jeremy S Meredith, Philip C Roth, Kyle Spafford, Vinod Tipparaju, and Jeffrey S Vetter. The scalable heterogeneous computing (SHOC) benchmark suite. In *Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units*, 2010.

[152] Jason Lowe-Power, Abdul Mutaal Ahmad, Ayaz Akram, Mohammad Alian, Rico Amslinger, Matteo Andreozzi, Adrià Armejach, Nils Asmussen, Srikant Bharadwaj, Gabe Black, et al. The gem5 simulator: Version 20.0+. *arXiv preprint arXiv:2007.03152*, 2020.

[153] Jianmin Chen, Xi Tao, Zhen Yang, Jih-Kwon Peir, Xiaoyuan Li, and Shih-Lien Lu. Guided region-based GPU scheduling: Utilizing multi-thread parallelism to hide memory latency. In *2013 IEEE 27th International Symposium on Parallel and Distributed Processing*, 2013.

[154] Timothy G. Rogers, Mike O'Connor, and Tor M. Aamodt. Cache-conscious wavefront scheduling. In *2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*, 2012.

[155] Timothy G. Rogers, Mike O'Connor, and Tor M. Aamodt. Divergence-aware warp scheduling. New York, NY, USA, 2013. Association for Computing Machinery.

[156] Mark Gebhart, Stephen W. Keckler, Brucek Khailany, Ronny Krashinsky, and William J. Dally. Unifying primary cache, scratch, and register file memories in a throughput processor. In *2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*, 2012.

[157] Mohammad Sadrosadati, Amirhossein Mirhosseini, Seyed Borna Ehsani, Hamid Sarbazi-Azad, Mario Drumond, Babak Falsafi, Rachata Ausavarungnirun, and Onur Mutlu. LTRF: Enabling high-capacity register files for gpus via hardware/software cooperative register prefetching. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '18, New York, NY, USA, 2018. ACM.