2017 IEEE International Symposium on High Performance Computer Architecture

Design and Evaluation of AWGR-based Photonic NoC Architectures for 2.5D Integrated High Performance Computing Systems

Paolo Grani, Roberto Proietti, Venkatesh Akella, S. J. Ben Yoo
Department of Electrical and Computer Engineering, University of California, Davis, CA, USA
Email: pgrani,rproietti,akella,sbyoo@ucdavis.edu

***Abstract*—In future performance improvement of the basic building block of supercomputers has to come through increased integration enabled by 3D (vertical) and 2.5D (horizontal) die-stacking. But to take advantage of this integration we need an interconnection network between the memory and compute die that not only can provide an order of magnitude higher bandwidth but also consume an order of magnitude less power than today's state of the art electronic interconnects. We show how Arrayed Waveguide Grating Router-based photonic interconnects implemented on the silicon interposer can be used to realize a 16 $\times$ 16 photonic Network-on-Chip (NoC) with a bisection bandwidth of 16 Tb/s. We propose a baseline network, which consumes 2.57 pJ/bit assuming 100% utilization. We show that the power is dominated by the electro-optical interface of the transmitter, which can be reduced by a more aggressive design that improves the energy per bit to 0.454 pJ/bit at 100% utilization. Compared to recently proposed interposer-based electrical NoC's we show an average performance improvement of 25% on the PARSEC benchmark suite on a 64-core system using the Gem5 simulation framework.**

I. Introduction

It is becoming increasingly clear that performance improvement through technology scaling (the so-called Moore's law) is slowing down if not coming to an end. What this means is that performance has to improve by increasing the level of integration. Much like we moved from discrete transistors to Large Scale Integration (LSI) and later on Very Large Scale Integration (VLSI) chips in the late 1970s, we are seeing the emergence of a new type of chip that consists of a set of vertically stacked die (3D integration) connected to each other using a silicon interposer (2.5D integration). 3D/2.5D integrated systems have many crucial advantages such as the ability to provide significantly higher memory bandwidth and memory capacity without going off-chip, the ability to integrate heterogeneous accelerators/co-processors and non-volatile memory, and the ability to realize a chip-scale multiprocessor with many smaller die in different technologies to improve yield and lower cost [1].

However, there are several challenges with 3D/2.5D integration that need to be addressed. Loh et al. [2] describe the requirements of an interposer-based integrated system that consists of four stacks of High Bandwidth Memory (HBM), [3]) connected to a 64-core processor organized either as a single die or *defragmented* into four smaller dies housing 16 cores each. Each HBM stack is a 4-layer vertically integrated 2 Gb DRAM die with a 1024 bit interface operating at 1 Gb/s for a total I/O bandwidth of 128 GB/s (in second generation HBM systems, such as from Hynix, the bandwidth can be as high as 256 GB/s [4]). Since the stacked DRAM capacity is limited (nowadays just 1 GB stack), Loh et al. [2] assume the chip has four additional channels of conventional DDR4 operating at 2666 MHz for a total bandwidth of 83 GB/s. The key requirement of the connection between the processor and memory is, in this case, a bisection bandwidth of roughly 298 GB/s [2] for adversarial traffic pattern and about 148 GB/s for uniform traffic pattern. A more futuristic compute node suitable for Exascale computing is described by AMD researchers in [5], where each node is a 2.5D/3D integrated system with 32 processor cores, a large GPU, connected to about eight stacks of 8-layer DRAM, with each stack providing about 16 GB capacity and 1 Tb/s HBM interface as before. Each node is expected to provide a computational capacity of 10 TFlops. This is along the lines of the target set by DoE except that the DoE requirement [1] calls for 1 TB of memory and a memory-access bandwidth of about 32 Tb/s.

The first challenge is how to develop the interconnection network for 2.5D integrated systems that can provide a bandwidth that is more than an order of magnitude higher than what is possible today. The related challenge is how to provide such a high bandwidth at a power consumption that is practical. This is important because not only are the 3D die-stacks more susceptible to thermal issues, but also, power budget of the chip has to be *shared* by the processors, memory, accelerators, and the inter-die interconnection network. If more power is spent in the interconnection between the die, then there is less power available to do the actual computation. Indeed, the overall power budget for the chip is still set by the Thermal Design Power (TDP) and is unlikely to be much different from today. So, the energy per bit of the interconnection network has to improve significantly (by an order of magnitude or more) for 2.5D integration to actually deliver the improvement in performance. Even with 8 GB or 16 GB per memory stack in the future, scaling a system to 1 TB or more to meet the Exascale computing requirement,

[1]FastForward 2 R&D Draft Statement of Work, report LLNL-PROP-652542, Lawrence Livermore Nat'l Lab., 2014; https://asc.llnl.gov/fastforward/rfp/04 DraftSOW 04-03-2014.pdf

2378-203X/17 $31.00 © 2017 IEEE
DOI 10.1109/HPCA.2017.17

means that the interconnection network has to span multiple chips, so we need an interconnection network that can *scale* with the same energy efficiency to tens of chips.

We argue that photonics is a compelling solution to address the challenges outlined above for the following reasons. In conventional electrical signaling, data is transported by charging and discharging capacitance of an electrical wire, which is not only extremely wasteful, but also scales poorly with distance. Optical communication completely avoid this by using the photo-electric effect (to transport data), which allows the generation of a large voltage in a detector with very little energy (the so-called *quantum impedance conversion* [6]). This results in exceptional energy efficiency, especially when the distance is long as is the case with 2.5D integrated systems. With an optical link it is possible to modulate the signals at very high rates, and transmit data in parallel without interference using different wavelengths (Wavelength Division Multiplexing, WDM), which can be harnessed to create interconnection networks with low latency and high bandwidth density (bandwidth per unit area) to meet the challenges listed above.

Moreover, over the past decade, there has been significant progress in silicon photonics with the development and experimental demonstration of efficient modulators, switches, receivers, waveguides, and lasers (see [7], [8], [9], [10]), which has resulted in better understanding of the design space tradeoffs of photonic networks (see [11], [12], [13]) which can help us develop *cost-effective* photonics NoCs to meet the requirements of interposer-based systems.

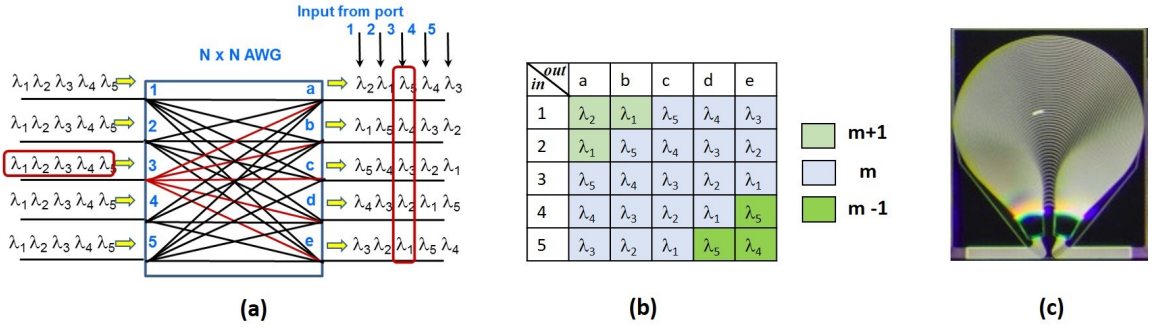
There has been a significant amount of prior work in intra-die photonic interconnection networks (see [11], [13], [14]) but the design tradeoffs for intra-die networks is different from an interposer-based inter-die one. First, the interposer is a large separate die almost 900 mm^2 in area that can be exclusively used for the interconnection. Therefore, fully connected topologies and switch fabrics that use a large number of waveguides can be used, which is not the case with an intra-die network where the optical interconnects share the die area with the rest of the processor electronics. Second, as opposed to an ad-hoc network that is closely coupled to the processors, for interposer-based systems we need a flexible, free-standing, general purpose switch fabric that can be used to interface CPUs, GPUs, HBM memory stacks, and other on-chip accelerators (heterogeneous architectures). Third, the network should be cost-effective and practical in the near term (as the 7 nm technology node is just a couple of generations away and there are no easy solutions to scale CMOS beyond that) which means it should be easy to fabricate and not rely on technologies that are speculative, like architectures exploiting WDM values higher than 64, and with hundreds of thousands of microrings like some of the intra-die networks proposed in literature(see [9], [15]).

Keeping these issues in mind we propose to use an Arrayed Waveguide Grating Router (AWGR) as the optical switch fabric to construct a photonic NoC that is suitable for interposer-based implementation.

The AWGR (see [16], [17]) is a passive optical cross-connect that is realized using two star couples connected by waveguides of unequal length. The functionality of a diffraction grating is obtained by letting the length of the waveguides increase linearly. In this router, when every input port carries the same set of optical wavelengths, each output port will receive the set of wavelengths with each wavelength coming from a different input. This provides, in the strict sense, a nonblocking cross connect. Though an AWGR is a mature technology with widespread use in telecommunication applications, it was not considered for intra-die photonic interconnects because its area was still quite large and it was hard to scale it to a very large number of ports. However, recent developments in fabrication (described in Section II) have made it possible to implement extremely compact AWGRs (which takes a few mm^2); moreover, for interposer-based NoCs, we do not need AWGRs with a very large number of ports, because the number of die on an interposer is not going to be arbitrarily large (in the order of tens). Therefore, it is time to evaluate the potential of AWGR-based photonic NoCs in computing applications. Specifically, we design and evaluate a 16 $\times$ 16 photonic NoC that can connect eight compute dies and eight vertically stacked DRAM dies, based on the High Bandwidth Memory (HBM) standard. Though AWGR is capable of implementing all-to-all, fully connected network, we deploy it as a crossbar with an all-optical control plane for arbitration. The network provides a total bisection bandwidth of 16 Tb/s.

The main contributions of the paper as follows: (1) We show how Free Spectral Range (FSR) and parallelism can exploited to realize a multi-bit AWGR-based interposer photonic network that can be used to interconnect 16 dies on a chip with a 16 Tb/s crossbar topology. (2) We identify the electrical/optical interface (SERializer-DESerializer, SERDES) as the main contributor to the latency/power consumption, and provide an optimized a network that reduces the SERDES requirements by a factor of 6 in terms of energy consumption and by a factor of 3 in terms of latency. (3) We evaluate the performance of the proposed networks on PARSEC benchmarks by modeling it in the Gem5 simulation environment and demonstrate an average performance improvement of $\sim$25% compared to the state-of-the-art electrical interposer-based NoCs proposed in recent research literature [1]. (4) We show that the high bandwidth and energy efficiency of photonic NoCs could be exploited to reduce the cache sizes of the processors which not only frees up die area to add more cores or accelerators, but also, helps in reducing the overall power consumption/energy efficiency of 2.5D/3D integrated compute nodes.

The rest of the paper is organized as follows. We start with the detailed discussion of the enabling technology, the AWGR and the design space of multi-bit AWGRs,



in \ out	a	b	c	d	e
1	λ_2	λ_1	λ_5	λ_4	λ_3
2	λ_1	λ_5	λ_4	λ_3	λ_2
3	λ_5	λ_4	λ_3	λ_2	λ_1
4	λ_4	λ_3	λ_2	λ_1	λ_5
5	λ_3	λ_2	λ_1	λ_5	λ_4

Figure 1. (a) shows the Wavelength Routing Property in a 5 × 5 AWGR - input 1 uses λ_2, input 2 uses λ_1, input 3 uses λ_5, input 4 uses λ_4, and input 5 uses λ_3 to go to output a as described in the wavelength assignment table (b). *m*, *m+1* and *m-1* denote the current, the next, and the previous Free Spectral Range (FSR). (c) shows the realized layout of an 8 × 8 AWGR.

including the implementation issues in Section II. Then we will show how to use the AWGR to construct the photonic NoC architecture in Section III. This is followed by the performance analysis and power analysis in Section IV and the discussion of the results in Section V. We conclude the paper with an overview of related work and future work.

II. Enabling Technology

In this section, we will describe the key enabling technology underlying the proposed photonic interconnect architecture. Given that the number of dies on a chip in a 2.5D/3D is going to be relatively small (in the tens, not hundreds), we believe that a topology that connects the computational units and memory directly without intermediate routers, is the key. We propose to use Arrayed Waveguide Grating Router (AWGR) to implement a crossbar topology. AWGR is a mature technology that has been used in the telecom industry for many years [16] and it is possible to make compact chip-scale AWGR with as many as 512 ports, with an 8 × 8 AWGR consuming only about 1 mm^2, as experimentally demonstrated in our laboratories [18], [19]. To the best of our knowledge, this is the first time a chip-scale AWGR is being proposed for an intra-die interconnect.

A. AWGR Principle of Operation

Figure 1 illustrates that the routing property of an AWGR allows any input port to communicate with any output port simultaneously using different wavelengths. As the wavelength assignment table in Figure 1(b) shows, input *1* uses wavelength λ_2 to communicate with output *a*, wavelength λ_1 to communicate with output *b*, wavelength λ_5 to communicate with output *c*, wavelength λ_4 to communicate with output *d*, and wavelength λ_3 to communicate with output *e*. Thus, a passive $N \times N$ AWGR intrinsically provides simultaneous, all-to-all communication capability (a fully connected topology) as long as each input is equipped with *N* transmitters and *N* receivers, which is prohibitively expensive and unnecessary most of the time. Hence, we restrict the topology to a crossbar by allowing the connection between one input-output at any given time. However, since multiple inputs might seek to connect to the same output, some mechanisms for contention resolution are needed at the receiver side, as in any crossbar topology. One interesting feature of AWGR operation is that the wavelength routing is *cyclic* with the period, being called the *Free Spectral Range (FSR)*, that we will be denoted in the following by the symbol Δ. This means that a given output *j* can be reached from input *i* using wavelength $\lambda_{ij \pm k\Delta}$, where *k* is an integer.

B. Multibit AWGR

AWGR, as described so far, is a single bit device, which means that at a given time a single bit can be transmitted from an input port *i* to an output port *j* by modulating λ_{ij} using a modulation scheme like simple ON/OFF keying. In computing application, we need to be able to transmit multiple bits in parallel which introduces additional complexity and design tradeoffs especially in terms of power consumption and area. There are several possibilities. The simplest option is to use *Time-Division Multiplexing* (TDM) by using *m* times higher data rate to send m bits at a time. However, this requires the electronic interface including the SERDES to operate at a much higher rate, which increases the power consumption. The next option is to take advantage of the FSR to send *k* bits simultaneous to an output. In the example above, if *k=4*, input *1* could use wavelengths λ_2, $\lambda_{2+\Delta}$, $\lambda_{2+2\Delta}$, $\lambda_{2+3\Delta}$, to send four bits to output *a*. However, this means that the tunable range of the laser is *k* times larger, or we need *k* times more lasers, which could be a challenge. Another option is to use more advanced modulation schemes such as Quadrature Amplitude Modulation (QAM) where multiple bits could be encoded by the quadrature and the phase of the optical signals using multiple levels. However, this requires very complex receivers and transmitters that are not appropriate for on-chip photonic networks. The final option is to use *Space Division Multiplexing* (SDM) or multiple AWGRs in parallel with each AWGR transmitting

Technology	Size	Loss	Crosstalk
Silica PLC	Large (e.g., 40 mm × 40 mm for 8 × 8 100 GHz spacing)	Low (e.g., 4 dB for 8 × 8)	Low (e.g., -27 dB for 8 × 8)
SiN_x/SiON or SiN_x/SiO_2 PICs	Medium (e.g., 4 mm × 4 mm for 8 × 8 100 GHz spacing)	Medium (e.g., 5 dB for 8 × 8)	Medium (e.g., -20 dB for 8 × 8)
SiP	Small (e.g., 1 mm × 1 mm for 8 × 8 100 GHz spacing)	High (e.g., 7 dB for 8 × 8)	High (e.g., -12 dB for 8 × 8)

Table I
AWGR IMPLEMENTATION OPTIONS AND TRADEOFFS.

one or more bits using the first three approaches. This is the *bit-slice* approach, which has been used since the early days of hardware. Each AWGR is responsible for transmitting a particular subset of contiguous bits of the flit or memory word. However, wiring the multiple AWGRs together, especially when the number of ports of the AWGR (N) is large, is challenging because of the extremely large number of waveguide crossing that increases the optical loss and crosstalk. Furthermore, if there are p AWGRs in parallel the number of lasers per node increases by a factor of p.

We propose to use only a single set of lasers for all the p parallel AWGRs with splitters, if necessary. We investigate using fewer AWGR slices in parallel but each operating at a much higher data rate (32 Gb/s or so). As far as dealing with laying out multiple AWGR slices (a 2D planar layout on the interposer), given the relative large area of the interposer, we think that it is possible to layout the AWGRs on the same plane. Indeed, the AWGR chips are very small (e.g., in the order of 1 mm^2, as shown in Table I). Using a planar layout, the 16 chiplets must be integrated in the same plane. From [1] a 16-core CPU requires ∼75 mm^2. From [3] a HBM stack module occupies ∼35 mm^2. Therefore the total footprint required to place the 8 CPU and 8 HBM modules, the required AWGR chips plus space for the waveguide *place and route*, is ∼1000 mm^2. This according to [1] is a well established value for silicon interposer size. According to TSMC [2] projects, even bigger silicon interposer with an area of 1200 mm^2 can be realized.

C. Implementation Issues

Tunable lasers and AWGRs are practical technologies that can be implemented today in many commercial foundries such as NeoPhotonics, Enablence, NEL, LioNIX, AIMPhotonics, and IME. Table I summarizes the different implementation options for an AWGR and the design tradeoffs.

Silicon Photonics (SiP) offers the smallest footprint but optical losses and crosstalk are relatively high but it has the additional benefit that low-power modulators and detectors can be integrated with the AWGR. The design and fabrication of 512 × 512, 25-GHz AWGR with a channel spacing of 25 GHz and a footprint of 11 mm × 16 mm is demonstrated in [18] and a 16 port AWGR with 50 GHz channel spacing on a Silicon Nitride (SiN) platform exhibiting significantly lower loss was demonstrated in [20]. The 16 port AWGR is quite compact (just 3.7 mm × 0.7 mm) and is closer to what we need in terms of a building block for the proposed interconnection network in this paper.

[2]http://www.eetimes.com/document.asp?doc_id=1329217 &print=yes

III. INTERPOSER-BASED PHOTONIC ARCHITECTURE

In this section we describe a 16 × 16 network that uses a crossbar topology realized using a multi-bit AWGR described before. The main advantage of using an AWGR-based interconnection is that to have the possibility to *flat* the topology, achieving a lower number of hops between the dies. The network has four key design parameters. N is the number of nodes (that correspond to the number of dies on the chip where a die could be a vertically integrated die stack corresponding to a processor, GPU, hardware accelerator or HBM memory), k is the number of FSRs used, p is the number of AWGR slices used to realize a multi-bit datapath, and L is the line rate of each port.

A. Implementation

Figure 2(a) shows the high-level implementation of the optical NoC that we call *Baseline*. Here we assume *N=16*, *k=4*, *p=8*, and *L=2* Tb/s corresponding to HBM2 memory standard specification [4]. The crossbar network has a total bandwidth of 16 Tb/s as every node can be communicating with another (disjoint) node concurrently. Figure 2(b) shows the details of the electro/optical interface for each port. Note that there are eight transmitters (dark circles) since *p=8*, one for each slide of the multi-bit AWGR and similarly eight receivers (gray circles). We assume that processing nodes such as GPUs, CPUs, etc, have 256 I/Os operating at 4 GHz to realize the 1 Tb/s bandwidth in each direction, for a total bidirectional bandwidth of 2 Tb/s (denoted as the parameter *L*). We have four tunable lasers per transceiver, one for each FSR (note that *k=4*) and four, 8:1 serializers for time domain multiplexing, eight bits per time slot. Therefore, the resultant optical data rate is 32 Gb/s to achieve 4 × 8 × 32 = 1 Tb/s link bandwidth in each direction. The HBM I/O interface is slightly different. It has 1024 I/Os operating at 1 Gb/s (in each direction, see [4]), and, therefore, we use a 32:1 serializer for the HBM modules. A receiver as a similar architecture, as shown in Figure 2(b), bottom side. As we will show in Section IV, the transmitter side of the electro-optical interface dominates the energy per bit of the optical Baseline. We propose to increase the parallelism, i.e.,

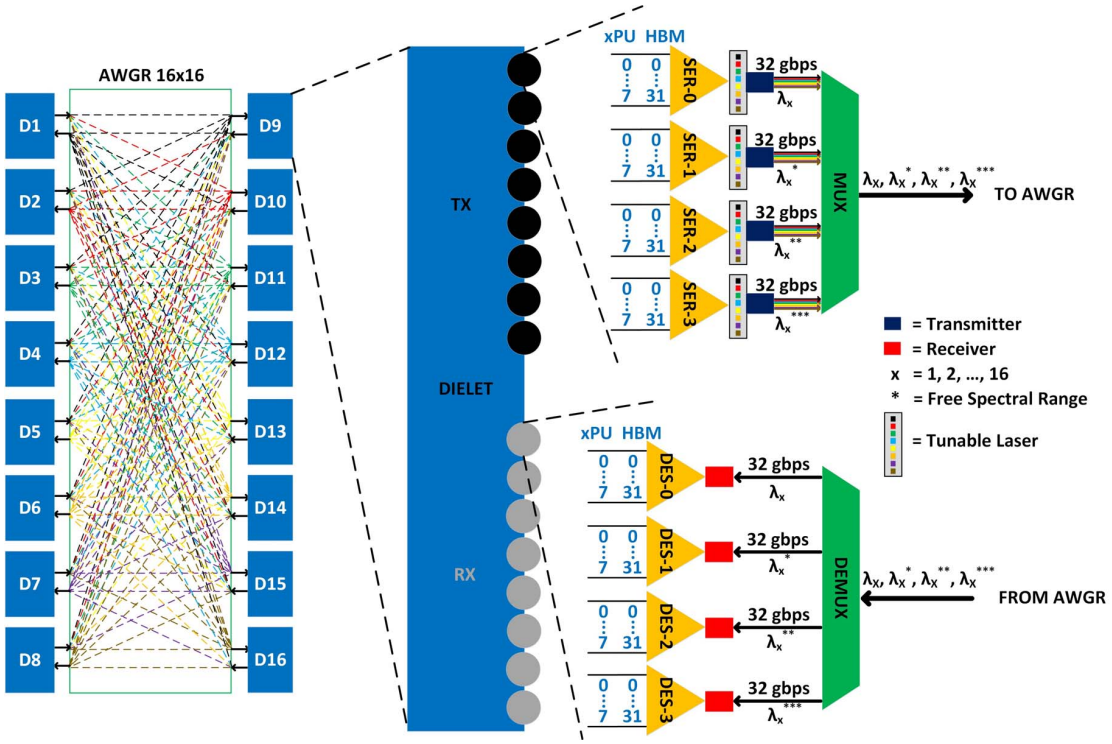


Figure 2. (a) High level view of the interposer-based photonic NoC called Baseline connecting 16 dielets. (b) Details of I/O interface of the NoC. The notation *, **, *** is to denotes the four wavelengths corresponding with different FSR, since *k=4*. In general, when realizing the inter-chip network, some of the ports of the NoC will be connected to in other chips.

to increase *p*, which means to use more parallel AWGRs, to reduce the clock frequency of the electro-optical interface, maintaining the same overall bandwidth. Indeed, one of the key insight/ideas in the paper, is how to find the right balance between parallelism, SERDES date rate (which affects power), number of FSRs, to achieve the target bandwidth and reduce energy per bit significantly. This optical network is called *Optimized*, and it has the following design parameters. *N* is still 16, *k=12*, *p=17*, and the data rate on the optical link is about 5 Gb/s instead of 32 Gb/s of the Baseline network. The product of *k* and *p* gives us the total number of wavelengths which happens to be 204. When multiplied by the frequency of the optical link (5 Gb/s), gives us 1 Tb/s link bandwidth per direction. We would like to note that these parameters were chosen so that the optical link rate and the I/O clock rate of the compute nodes are the same (5 Gb/s). In this case, we can eliminate the SERDES in the transmitter and receiver sides for the compute nodes which can save significant power. This design is aggressive since *k=12* and *p=17* are challenging, but we believe they are feasible in the future with the 3D writing technology.

B. Optical Contention Resolution

The gain saturation effect in a Reflective Semiconductor Optical Amplifier (R-SOA) [21] can be used to realize optical contention resolution. *N* different nodes can make requests R_1, R_2, ..., R_n to the R-SOA associated with a given AWGR output port using different wavelengths λ_1, λ_2, ..., λ_n (see Figure 3). The first request, say R_i that arrives at the R-SOA saturates it, which results in some fraction of the power (P_{tot}) power reflected back to the sender node *i*. The R-SOA stays saturated as long as the request on λ_i is held. A detector that is set to trigger at P_{tot} produces the grant signal. If another request R_j (on λ_j) from node *j* arrives while R_i is still active, the power reflected at λ_j will be $\sim P_{tot}/2$ (because of the saturation effect in the R-SOA), which is not enough to set the trigger condition; hence, the second request will be excluded [see Figure 3(c)]. If two requests arrive at approximately the same time at the R-SOA (with a time interval comparable or lower than the R-SOA gain dynamics, i.e., few hundreds of picoseconds), both the requestors receive approximately $P_{tot}/2$ reflected power and hence the detectors at neither node triggers, which corresponds to a situation that neither requestor has been granted. Note that this is different from a classic electronic mutex element where eventually one of the requestor gets a grant. In the R-SOA-based mutex element, it is possible for none of the requestors to get a grant, that is okay because the requirement of mutual exclusion element is that at most one of the requestors be granted the resource, zero grants is okay

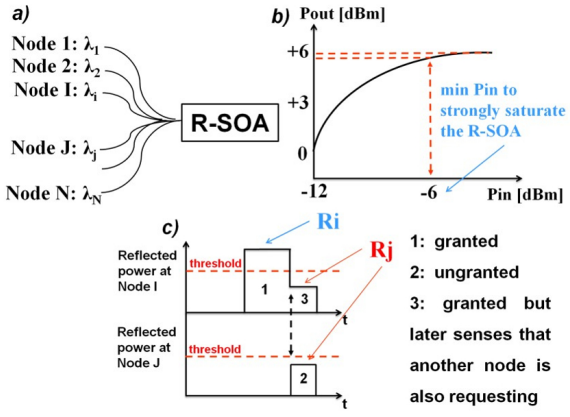


Figure 3. (a) N:1 Mutual Exclusion scheme with R-SOA. (b) Typical R-SOA P_{out}/P_{in} characteristic and minimum input power to operate the R-SOA in strong saturation regime. (c) Three different possible cases in the R-SOA-based mutual exclusion scheme. Figure from [22].

from the correctness of the protocol perspective. R-SOA-based mutual exclusion has an interesting property. Suppose, a request R_j arrives at the R-SOA after request R_i has been granted, i.e., while R_i is still asserted, then the reflected power to node i drops from P_{tot} to $P_{tot/2}$, which could serve as a signal to node i that some other node has made a request to the same resource. This information is used to ensure fairness while removing the overhead for arbitration. This all-optical contention resolution scheme, inspired by Carrier Sense Multiple Access with Collision Detection technique, performs well for short distance communications. More details can be found in [22], [23].

IV. Results

In this section we will develop analytical models for latency and power consumption of the proposed photonic NoC and evaluate its performance on benchmarks. We will also compare our performance results with state-of-the-art interposer-based electrical NoCs from recent literature [1].

A. Modeling the Latency of the Photonic NoC

In the proposed networks, the main contributors to the latency are: the SERDES, the waveguides in AWGR-based switch fabric, the control plane for arbitration resolution, and the wavelength tuning time for the tunable lasers. For the control plane we assume the all-optical contention resolution mechanism proposed in [22] where the control plane latency for a 16-node network is shown to be around 10 ns.

We use the equation from [24] ($8.3 \times 10^{-3} \times$ L + 70.9 ps, where L is the length of the waveguide) to estimate the propagation delay through the waveguides and the optical interface circuitry (modulators, detectors, and drivers). In our proposed architecture the length of the waveguides is of the order of 10 mm so the propagation latency is around 80 ps. The SERDES is the main contributor to the latency.

Parameter	Description
Cores	64, 64 bits, out-of-order, 2 GHz
L1 cache	32/16 kB(I)+32/16 kB(D), 2-way, 1 cycle hit
L2 cache	512/256 kB banks, 8-way, 3/12 cycles tag/tag+data
Directory	MOESI, 64 slices, 3 cycles
E-Interposer	Concentrated Mesh (8:1), 2 GHz, 64 bits, 1 cycle/hop
O-Interposer	All-to-All (AWGR), 32/5 GHz, 8/17 bits, 1 cycle/hop
Memory	64 GB, 16/8 channels, 128 bits, 1.6 GHz, ∼200 cycles

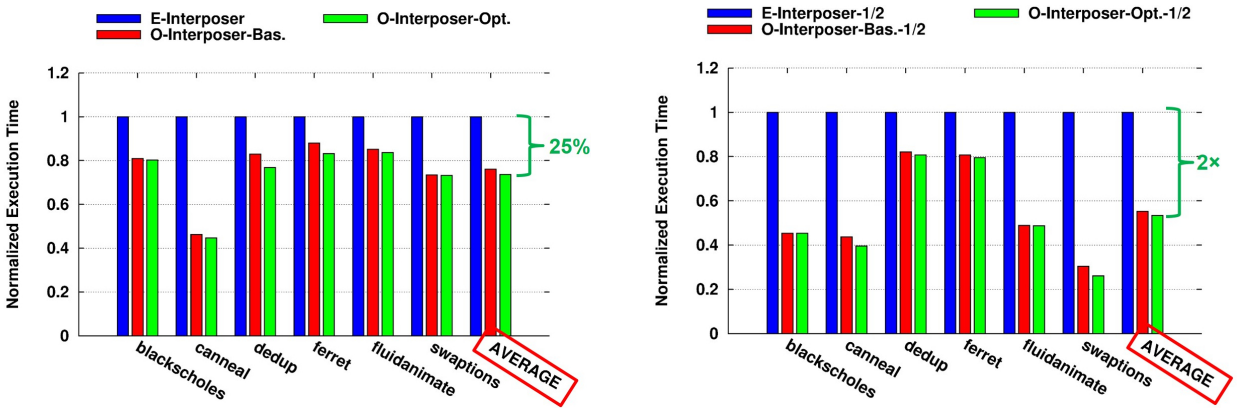
Table II
Parameters of the simulated architecture.

For the Baseline network, we assume four, 8:1 and 1:8 serializers/deserializers for the processing dies operating at 4 Gb/s and for the memory dies we need 32, 32:1 and 1:32 serializers/deserializers operating at 1 Gb/s to achieve the 1 Tb/s bandwidth. Using data from [25] and [26], we estimate that the total SERDES latency to be around 36 ns. Finally, according to [27], the wavelength tuning time is about 8 ns. Therefore the total end-to-end latency in the worst-case Baseline network is about 54 ns. As pointed out in Section III-A in the Optimized case, we completely avoid SERDES in the processor dies by reducing the data rate from 32 Gb/s to 5 Gb/s. Therefore, the estimated overall latency for the Optimized network is around 18 ns.

B. Experimental Setup

Our goal is to model the system shown in Figure 2 which consists of 16 dies, eight of which are compute nodes and the remaining eight are memory nodes. Each compute node is a 8-core processor for a total of 64 cores. We assume the intra-die network is a 2D mesh-based electrical network and the inter-die network is the proposed photonic NoC. We use the Gem5 simulator [28] in Full-System (FS) mode to evaluate the performance of the proposed architecture. The simulator booted a complete Linux 2.6.27 Operating System (OS) for multi-threaded application scheduling and support. We modeled Chip Multi Processors (CMP) architectures with 64 cores, based on the Alpha Instruction Set Architecture (ISA) in the simulator. Each core has private L1 caches for instruction and data and a slice of a shared distributed L2 cache (Last-Level Cache, LLC). The directory information is distributed, and a directory-based, coherence protocol (i.e., MOESI) manages the caches. Table II summarizes the main architectural features of the overall architecture.

We use the PARSEC benchmarks suite [29], a collection of heterogeneous parallel applications spanning different application domains (e.g., media processing, search and filtering, 3D, and physics simulations) and representative of a diverse workload. Benchmarks were modified to enforce that each spawned thread is pinned to a fixed core of the processor (i.e., core affinity). This approach prevents some non-determinism in the parallel benchmark execution. We compared the performance results obtained with our proposed architectures against a recent electronic interposer-based architecture discussed in [1], which we will call



(a) Full-Cache **(b) Half-Cache**

Figure 4. Execution time results of the proposed solutions in the (a) Full-cache case, and (b) Half-cache one, normalized to the E-Interposer setup. The results are presented for both the O-Interposer-Baseline and the O-Interposer-Optimized cases.

E-Interposer in the rest of this paper. We think that the adopted methodology is a rigorous method to evaluate the potential benefits of the proposed architectures. Table II shows the parameters of the different configuration simulated. We modeled the DRAM-stacks as described in [1] using the same amount of DRAM in both the electrical and photonic configurations except that it is distributed in four stacks in the E-interposer case, and in eight stacks for our proposed optical architectures. *O-Interposer-Baseline* and *O-Interposer-Optimized* are the configuration with the baseline photonic NoC and the optimized NoC as described in Section III. In the E-Interposer case, the memory stacks are deployed on the electronic interposer substrate so they can be accessed in the same way that one core can reach another core, it is just another hop to the memory. We are assuming that Gem5 models the hop delay and the latency of the electrical network including the arbitration/control overhead for the given topology appropriately. In the proposed design, memory is deployed on the optical interposer. We modeled the latency of having the undergo electric to optical conversion, SERDES, overhead of the control plane, and AWGR-based switch fabric, and the receiver, and the final optical to electrical conversion as described in Section IV-A.

C. Performance Analysis

We experimented two configurations to estimate the benefits of a high bandwidth inter-die interconnection network.

- *Full-Cache* case: 32 kB(I)+32 kB(D), 2-way, L1 cache, and 512 kB per core, 8-way, L2 cache
- *Half-Cache* case: 16 kB(I)+16 kB(D), 2-way, L1 cache, and 256 kB per core, 8-way, L2 cache

Figure 4a shows the performance of the *Full-Cache* case. Note that the photonic network outperforms the E-Interposer on all the benchmarks. On average, the O-Interposer-based setups outperform the E-Interposer setup (blue columns in the figure) of $\sim$25%. Especially *canneal* benchmark does particularly well due to its more random memory traffic pattern which requires all-to-all communication as described in [29]. This fits well with the photonic NoC topology described in this paper and is especially encouraging because emerging applications such as large scale data analytics and large scale graph mining are expected to be more irregular and with large working sets and little data reuse. Another interesting thing to note is that there is not much difference in terms of performance between the O-Interposer-Baseline and the O-Interposer-Optimized configurations. As we will show in the Section IV-D, the main benefit of the optimized network is that improves the power consumption significantly by reducing the SERDES power. Figure 4b shows the performance of *Half-cache* configuration. Clearly, reducing the cache has a more significant impact on the E-Interposer as opposed to the O-Interposer. In fact, the O-Interposer-based system is almost 2$\times$ better in terms of the performance. This is because the E-interposer-based system has a NoC based on multi-hop mesh topology and has lower bisection bandwidth than the photonic NoCs.

Figure 5 is a better demonstration of the impact of the reduction in the cache size. Note that the simulated systems only have a relatively small L1 caches (32 kB), so halving the L1 cache size (16 kB) and the L2 cache size (256 kB per core), has a profound effect on the overall miss rate. Given that the DRAM access latency even with HBMs is still quite high, the overall performance suffers. However, it is still encouraging to note that even with a 50% reduction in the cache sizes, the performance of our proposed systems with photonic NoC is only about 30% lower than a system with full cache sizes and an electrical interposer-based NoC. This might work better with systems with a large L3 as Last-Level Cache, where reducing the size of the L3 will have less impact on performance because the L2 and the savings in area and power consumption will also be more substantial.

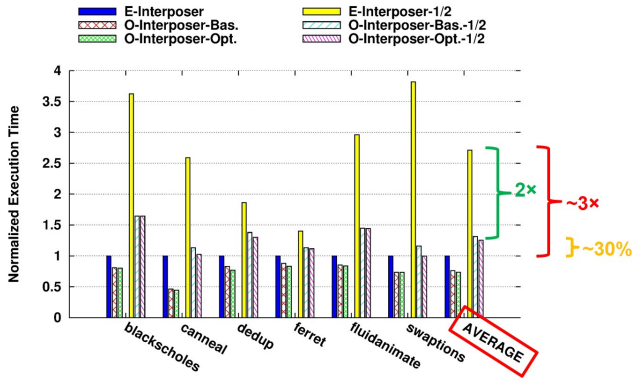


Figure 5. Execution time comparison between the Full-cache and Half-cache scenarios, normalized to the E-Interposer, Full-Cache case (blue solid columns). The yellow solid columns represent the E-Interposer, Half-Cache case. The dashed columns represent the O-Interposer-Baseline and the O-Interposer-Optimized setups for Full-cache (red-dashed and green-dashed columns) and the O-Interposer-Baseline and the O-Interposer-Optimized setups Half-cache (lightblue-dashed and purple-dashed columns).

D. Power Analysis

We used the parameters listed in Table III to construct a power model for the proposed photonic NoCs. The tunable laser parameters for the O-Interposer-Baseline network were obtained from [30], [31]. The transmitter parameters were obtained from [32] by scaling the frequency from 48 GHz to 32 GHz and the receiver values were obtained from [33]. The SERDES values were obtained from [32]. Similarly, for the O-Interposer-Optimized network, the parameter values were obtained from [34] and scaled to meet the design parameters of our network. The values for the control plane were obtained from [23]. The data for insertion and waveguide losses were obtained from [35], [36], [37], [38], [7], [39] and are summarized in Table IV. Based on these parameters, we estimate the total power consumption for the O-Interposer-Baseline network to support the bisection bandwidth of 16 Tb/s is ∼45 Watts assuming 100% link utilization, and the relative energy consumption is ∼2.57 pJ/bit. For the O-Interposer-Optimized case, we estimate the power is ∼7 Watts or 0.454 pJ/bit to support the 16 Tb/s bandwidth assuming 100% utilization.

Figure 6(a) and Figure 6(b) show the breakdown of the power in the two networks. Note how, by taking advantage of higher parallelism (both in terms of additional FSRs and number of parallel AWGRs, and scaling down the data rates appropriately, the transmitter energy was reduced from 1.52 pJ/bit to only 16 fJ/bit for an overall total energy reduction of a ∼6 factor. How does this compare with high bandwidth state-of-the-art electronic and optical switches from recent research literature such as the work from HP reported in [8]? The paper from HP is based on switches operating at 320 Gb/s (32 bits @ 10 Gb/s) per input port. Hence, a very high radix switch (we estimate 144 port switch) is required to provide the same bandwidth as the

Parameter Name	Baseline	Optimized
Frequency [GHz]	32	5
Laser Efficiency	4.5%	4.5%
Laser-Output-Power [mW/bit]	0.5	2.5
Laser Tuning [mW]	8.76	28
Photodetector Sensitivity [dBm]	-10	-15
Transmitter (Driver+Mod+PLL) [fJ/bit]	1520.8	16.6
Receiver (Diode+TIA+Clock) [fJ/bit]	708.3	135.7
Serializer xPU [fJ/bit]	39	-
Serializer HBM [fJ/bit]	156.25	41.6
Deserializer xPU [fJ/bit]	39	-
Deserializer HBM [fJ/bit]	156.25	67
Reflective SOA [mW]	100	100
Total TX+RX per xPU [W]	2.64	0.3
Total TX+RX per HBM [W]	2.88	0.3
Total-Network+I/O [W]	∼45	∼7

Table III
POWER / ENERGY PARAMETERS.

NoCs considered in this paper. According to the HP paper (Table 4 in [8]) the power consumption of a 144 port electronic switch (operating at 320 Gb/s per port in a 22 nm implementation) is ∼154 Watts and a microring resonator-based optical switch is about ∼76 Watts. Though these switches might have other advantages such as the ability to allocate bandwidth more flexibly and resilience, we believe that the high power consumption makes them unsuitable for interposer-based applications where there is a stringent chip-level power budget which has to be shared by the compute units, memory, and the interconnection network.

V. DISCUSSION

In this section we will interpret the results presented in the last section and revisit the underlying assumption and opportunities for extending this work. First, note that the related work in interposer-based electrical NoCs [1], does not present absolute power numbers in terms of energy per bit for their networks and the utilization (they only present relative power numbers comparing different topologies of their own). Therefore, we cannot exactly compare our results

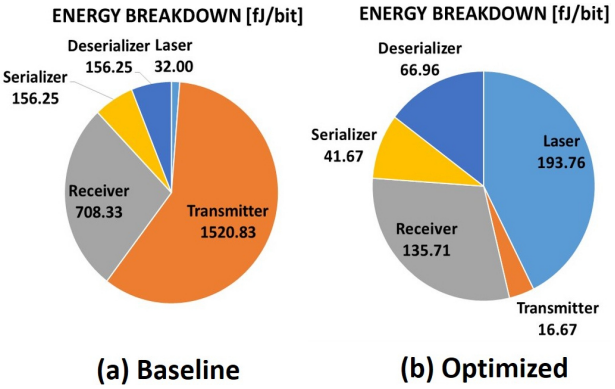


Figure 6. Energy breakdown for the O-Interposer-Baseline (a) and O-Interposer-Optimized (b) NoCs, HBM case. The total energy consumption for case (a) is 2.57 pJ/bit, while for case (b) is 0.454 pJ/bit, ∼6× reduction.

Parameter Name	Baseline	Optimized
Fiber [dB/cm]	-0.0001	-0.0001
Grating Coupler [dB]	-1	-1
Waveguide [dB/cm]	-0.1	-0.1
Coupler (x2) [dB]	-1	-1
AWGR [dB]	-3	-3
Splitters [dB]	-9	-12
Photodetector	-0.1	-0.1
Mux/Demux	-1	-1
Total-Power-Penalty	$\sim$-15	$\sim$-19

Table IV
INSERTION LOSS PARAMETERS.

with theirs. Second, one could argue that the electrical interposer-based NoC has lower bisection bandwidth than the proposed network so it might be unfair to compare them directly. However, it should be noted that the overall memory bandwidth for the PARSEC benchmarks [29] is well below the bandwidth of both the electrical and photonic networks compared here. The benefit of the photonic network is due to lower latency of the crossbar topology of the photonic NoC, as opposed to a 8×8 mesh in the case of the electrical interposer-based NoC and the fact that we have more parallelism (eight DRAM stacks versus four).

A. Architectural Implications

Interposer-based photonic NoCs with extremely high energy-efficient bandwidth might require new thinking in terms of the architecture of 2.5D integrated systems. For example, recent work in graph exploration [40] and in the evaluation of the benefits of stacked high bandwidth DRAM [41], has shown that memory-level parallelism is quite low, which means that it is questionable whether a very high bandwidth network to memory can be fully utilized. In Section IV we proposed the idea of using smaller LLC which not only has the benefit of reducing the die size but also reducing the power consumption of the compute die (especially leakage) since a significant fraction of the die area is occupied by the cache itself. Memory bandwidth and power consumption due to off-chip communications, constrained the design space of computer architecture for decades but with very high bandwidth memory interconnect and with very low energy consumption per bit, the 2.5D systems could employ different architectures. These solutions can trade area/power and simpler programming models at the expense of more communications such as simper cache coherence protocols, off-die LLCs. Another possibility is to implement a mechanism so that the bandwidth of the network can be scaled *dynamically*. This is somewhat challenging in a photonic network because the external laser is always ON. Therefore, power is always consumed even if the network is not fully utilized. In the proposed NoC we address this issue in two ways. First, we use a *tunable* laser instead of an Optical Frequency Comb (OFC) source to generate the different wavelengths at each port, to reach the different destinations. We assume that it is very likely that a compute node is communicating with some node at any given time. Therefore, the laser can be constantly used or, in other words, using a tunable laser improves the *utilization* of the light source itself. Second, we designed the NoCs to support *k* bandwidth states [analogous to power management states in Dynamic Voltage and Frequency Scaling (DVFS) solutions by taking advantage of the FSR-based parallelism in the AWGR. Indeed, in many computing applications, the traffic is bursty [42] hence we may need a very high bandwidth for a short period of time and a much lower bandwidth during other periods. For example, as shown in Figure 2(b), by turning OFF one transmitter section (say SER-3) we can provide $3/4$ of the link bandwidth; by turning OFF two sections (say SER-3 and SER-2) we can provide $1/2$ the peak link bandwidth, and by turning OFF three sections we can provide $1/4$ the peak bandwidth. Though turning the sections back ON (including recalibration of the lasers) could take several microseconds, we think it is still advantageous to use bandwidth scaling. It is possible for the communication scheduler in the memory/network controller to initiate the turning ON/OFF basing on the appropriate sections of the transmitter in advance.

B. Scaling to Larger Systems

We expect future high performance compute nodes that are geared towards meeting the requirements of Exascale computing to have a 3-level interconnect strategy - (a) *intra-die* network that is most likely going to be based on the electrically signaling in the foreseeable future, (b) *inter-die* network that is based on the interposer, which was the main focus of this paper, and (c) *inter-chip* network that connects multiple chips together on a board or multiple boards. In today's technology the interposer is around $\sim$900 mm^2 [1], and assuming 8 GB HBM memory stacks (that are possible in the near future), we can support 64 GB of DRAM on a chip. Therefore, to build a compute node with 1 TB of memory, to meet the DoE requirements outlined in the introduction of this paper, we need 16 chips. One way to interconnect these 16 chips is to use a second level network implemented on a separate chip that is identical to the one proposed here as first level network. The ability to scale to larger system with modest degradation in latency and power is the advantage of photonic interconnects and the proposed network is designed to be modular so that it can be integrated with an identical second level network on a board.

VI. RELATED WORK

There has been significant interest in 3D/2.5D integration in recent years. In [2] researchers describes silicon interposer technology for on-chip architectures, and tradeoffs about memory integration, thermal managements and cost analysis. The authors in [43] envision the next generation 3D-stacked chip integrating different and novel technologies like non-volatile memories, efficient heat removal, and *in-memory*

computation model. As noted before, [1] is perhaps the first work in research literature to thoroughly evaluate the design space of interposer-based electrical NoCs and propose two enhacements to existing network topologies called folded torus and enhanced butterfly. We use this as a baseline to compare our performance results. However, this work does not mention absolute power numbers, so we are not sure exactly what their energy per bit numbers are. But, given the distances involved in interposer-based implementation, they are likely to scale poorly compared with the photonic NoCs described here in terms of bandwidth and energy efficiency.

The benefits of photonics, design tradeoffs of photonic and electronic links, and the design of key building blocks such as modulators, waveguides, and lasers are described in [6], [14], [7]. Researchers at MIT/Berkeley [10] demonstrated a chip with dual-core RISC-V processors with 1 MB of SRAM on a commercial 45 nm CMOS SOI process without any changes to the foundry processes. This *zero change* proves that photonic circuits can be fabricated using standard design flows and tool.

The related work in the area of photonic interconnects in computing applications can be classified into four categories - intra-die networks such as HP's Corona [9] and its derivatives [44], [45], CPU-DRAM interface [40], [46], [10], inter-chip networks such as MIT's work [47], Galaxy [39], and high-radix switches [8], [48]. A majority of the existing work uses microring-based resonators and often a bus-based crossbar (though work such as [49] uses a CLOS topology as the switching fabric). The core of any interconnection network is the switch that implements the cross-connect. Instead of an AWGR, it is possible to realize an optical switch using MRRs-based. In this case, a large number of microrings and waveguides are needed to realize an all-to-all network topology similar to what is possible with an AWGR. Exploiting AWGR th switch fabric is *microrings-free* and we need rings just in the transmitter and receiver sides. Therefore, a much lower number of microrings (and a lower energy consumption due to thermal tuning) is achievable with an AWGR-based interconnection. Also, as noted in Section I, the design space tradeoffs of intra-die photonic networks are different from interposer-based implementation. For example, the intra-die networks assume very Dense WDM (DWDM, with 64 wavelenghts) so they do not have to worry about SERDES. The novelty of our approach is the use of Arrayed Waveguide Grating Router (AWGR) as the passive optical switch fabric. To the best of our knowledge this has not be considered for computing applications because, as described in Section II, in the past, the area requirements of AWGR was quite large, and it was not scalable to a very large number of ports (a key requirement in many intra-die networks) due to crosstalk between the waveguides. Furthermore, AWGR has always been used as a serial device (one bit at a time). Only very recently, with the advances in fabrication technology, low loss millimeter scale AWGRs have become practical. These advances made AWGR-based architectures attractive to implement new interposer photonic networks. Furthermore, given that there is a physical overall area constraint (the size of the interposer cannot be too large), these interposer-based architectures do not require AWGRs with a very large number of ports. In this case, losses due to crosstalk and propagation can be quite low. In this paper we showed how to realize multi-bit AWGRs by taking advantage of multiple FSRs and transmission parallelism in computing applications. Another key difference with related work is that, in this work, we use tunable lasers as opposed to Optical Frequency Combs to realize WDM-based communication. As already explained in Section V-A, this is more cost-effective (in terms of the number of components required), given that we do not need 64 wavelengths, and the fact that the lasers will be implemented on the interposer itself.

VII. Conclusion and Future Work

In this paper, we argued that 3D/2.5D die-stacked architectures are inevitable considering the fact that we are approaching the end of the era of performance improvement through cost-effective CMOS scaling. However, connecting the different die-stacks with a high-bandwidth interconnect, especially one that can scale to Terabytes of memory and tens of Terabits per second bisection bandwidth, is a huge challenge. We proposed an interposer-based photonic NoC that can provide a bandwidth of 16 Tb/s at an extremely high energy efficiency of 454 fJ/bit. The novelty of our solution is to use Arrayed Waveguide Grating Router (AWGR) as the optical switching fabric to support computing applications. With photonic networks we can scale the interconnect to multiple chips to provide tens of Tflops of computing and tens of Terabytes of memory. This is really exciting because from the dawn of computing, architects have always been shackled by the so-called *memory wall* - both memory capacity and bandwidth (given that it was off-chip) have always been a bottleneck. But with the advent of 3D die-stacked memories that can provide very high capacity and high bandwidth interface, and of extremely energy efficient, scalable inter-die *photonic* networks, it is perhaps time to think again about the role and size of caches and cache coherence protocols. Instead of plunking existing compute die on an interposer, perhaps the architecture of 2.5D computing systems has to be rethought without being clouded by the memory wall but, more generally, from how to maximize the performance per Watt, keeping in mind that memory latency is still a huge issue as none of these technologies directly mitigate it. This will form our future work.

Acknowledgment

This work was supported in part under DoD Agreement Number: W911NF-13-1-0090.

REFERENCES

[1] A. Kannan, N. E. Jerger, and G. H. Loh, "Enabling interposer-based disintegration of multi-core processors," in *Proceedings of the 48th International Symposium on Microarchitecture*, ser. MICRO-48. New York, USA: ACM, 2015.

[2] G. H. Loh, N. E. Jerger, A. Kannan, and Y. Eckert, "Interconnect-memory challenges for multi-chip, silicon interposer systems," in *Proceedings of the 2015 International Symposium on Memory Systems*, ser. MEMSYS '15. New York, USA: ACM, 2015, pp. 3–10.

[3] D. U. Lee, K. W. Kim, K. W. Kim, H. Kim, J. Y. Kim, Y. J. Park, J. H. Kim, D. S. Kim, H. B. Park, J. W. Shin, J. H. Cho, K. H. Kwon, M. J. Kim, J. Lee, K. W. Park, B. Chung, and S. Hong, "25.2 a 1.2v 8gb 8-channel 128gb/s high-bandwidth memory (hbm) stacked dram with effective microbump i/o test methods using 29nm process and tsv," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, Feb 2014, pp. 432–433.

[4] D. U. Lee, K. S. Lee, Y. Lee, K. W. Kim, J. H. Kang, J. Lee, and J. H. Chun, "Design considerations of hbm stacked dram and the memory architecture extension," in *2015 IEEE Custom Integrated Circuits Conference*, 2015, pp. 1–8.

[5] M. J. Schulte, M. Ignatowski, G. H. Loh, B. M. Beckmann, W. C. Brantley, S. Gurumurthi, N. Jayasena, I. Paul, S. K. Reinhardt, and G. Rodgers, "Achieving exascale capabilities through heterogeneous computing," *IEEE Micro*, vol. 35, no. 4, pp. 26–36, July 2015.

[6] D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proceedings of the IEEE*, vol. 88, no. 6, pp. 728–749, June 2000.

[7] J. Cardenas, C. B. Poitras, J. T. Robinson, K. Preston, L. Chen, and M. Lipson, "Low loss etchless silicon photonic waveguides," *Opt. Express*, vol. 17, no. 6, Mar 2009.

[8] N. Binkert, A. Davis, N. P. Jouppi, M. McLaren, N. Muralimanohar, R. Schreiber, and J. H. Ahn, "The role of optics in future high radix switch design," in *2011 38th Annual International Symposium on Computer Architecture (ISCA)*, 2011.

[9] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn, "Corona: System implications of emerging nanophotonic technology," in *2008 International Symposium on Computer Architecture*, June 2008, pp. 153–164.

[10] C. Sun, M. T. Wade, Y. Lee, J. S. Orcutt, L. Alloatti, M. S. Georgas, A. S. Waterman, J. M. Shainline, R. R. Avizienis, S. Lin *et al.*, "Single-chip microprocessor that communicates directly using light," *Nature*, vol. 528, no. 7583, 2015.

[11] C. J. Nitta, M. K. Farrens, and V. Akella, "On-chip photonic interconnects: A computer architect's perspective," *Synthesis Lectures on Computer Architecture*, vol. 8, no. 5, 2013.

[12] P. Grani and S. Bartolini, "Design options for optical ring interconnect in future client devices," *J. Emerg. Technol. Comput. Syst.*, vol. 10, no. 4, pp. 30:1–30:25, Jun. 2014.

[13] K. Bergman, L. P. Carloni, A. Biberman, J. Chan, and G. Hendry, *Photonic network-on-chip design*. Springer, 2014.

[14] R. G. Beausoleil, "Large-scale integrated photonics for high-performance interconnects," *J. Emerg. Technol. Comput. Syst.*, vol. 7, no. 2, pp. 6:1–6:54, Jul. 2011.

[15] C. Nitta, M. Farrens, and V. Akella, "Dcaf - a directly connected arbitration-free photonic crossbar for energy-efficient high performance computing," in *Proceedings of the 2012 IEEE 26th International Parallel and Distributed Processing Symposium*, ser. IPDPS '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 1144–1155.

[16] S. Kamei, M. Ishii, M. Itoh, I. Shibata, Y. Inoue, and T. Kitagawa, "64× 64-channel uniform-loss and cyclic-frequency arrayed-waveguide grating router module," *Electronics Letters*, vol. 39, no. 1, pp. 83–84, 2003.

[17] K. Takada, M. Abe, M. Shibata, M. Ishii, and K. Okamoto, "Low-crosstalk 10-ghz-spaced 512-channel arrayed-waveguide grating multi/demultiplexer fabricated on a 4-in wafer," *IEEE Photonics Technology Letters*, vol. 13, no. 11, pp. 1182–1184, Nov 2001.

[18] S. Cheung, T. Su, K. Okamoto, and S. Yoo, "Ultra-compact silicon photonic 512× 512 25 ghz arrayed waveguide grating router," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 4, pp. 310–316, 2014.

[19] R. Yu, S. Cheung, Y. Li, K. Okamoto, R. Proietti, Y. Yin, and S. J. B. Yoo, "A scalable silicon photonic chip-scale optical switch for high performance computing systems," *Opt. Express*, vol. 21, no. 26, pp. 32 655–32 667, Dec 2013.

[20] S. Pathak, K. Shang, and S. J. B. Yoo, "Experimental demonstration of compact 16 channels-50 ghz si3n4 arrayed waveguide grating," in *2015 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2015, pp. 1–3.

[21] M. J. Connelly, *Semiconductor optical amplifiers*. Springer Science & Business Media, 2007.

[22] R. Proietti, C. J. Nitta, Y. Yin, R. Yu, S. J. B. Yoo, and V. Akella, "Scalable and distributed contention resolution in awgr-based data center switches using rsoa-based optical mutual exclusion," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, no. 2, March 2013.

[23] R. Proietti, Y. Yin, R. Yu, C. Nitta, V. Akella, and S. J. B. Yoo, "An All-Optical Token Technique Enabling a Fully-Distributed Control Plane in AWGR-Based Optical Interconnects," *Journal of Lightwave Technology*, vol. 31, 2013.

[24] S. Sun, A. H. A. Badawy, V. Narayana, T. El-Ghazawi, and V. J. Sorger, "The case for hybrid photonic plasmonic interconnects: Low-latency energy-and-area-efficient on-chip interconnects," *IEEE Photonics Journal*, vol. 7, 2015.

[25] J. Wang, L. Guan, Z. Sang, J. Chapman, T. Dai, B. Zhou, and J. Zhu, "Characterization of a serializer asic chip for the upgrade of the atlas muon detector," *IEEE Transactions on Nuclear Science*, vol. 62, no. 6, pp. 3242–3248, 2015.

[26] B. Deng, M. He, J. Chen, D. Gong, D. Guo, S. Hou, X. Li, F. Liang, C. Liu, G. Liu, P. K. Teng, A. C. Xiang, T. Xu, Y. Yang, J. Ye, X. Zhao, and T. Liu, "Component prototypes towards a low-latency, small-form-factor optical link for the atlas liquid argon calorimeter phase-i trigger upgrade," *IEEE Transactions on Nuclear Science*, vol. 62, no. 1, 2015.

[27] S. Matsuo and T. Segawa, "Microring-resonator-based widely tunable lasers," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 15, no. 3, pp. 545–554, May 2009.

[28] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, "The gem5 simulator," *SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, Aug. 2011.

[29] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The parsec benchmark suite: Characterization and architectural implications," in *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques*, ser. PACT '08. New York, USA: ACM, 2008, pp. 72–81.

[30] X. Zheng, S. Lin, Y. Luo, J. Yao, G. Li, S. S. Djordjevic, J. H. Lee, H. D. Thacker, I. Shubin, K. Raj, J. E. Cunningham, and A. V. Krishnamoorthy, "Efficient wdm laser sources towards terabyte/s silicon photonic interconnects," *Journal of Lightwave Technology*, vol. 31, no. 24, 2013.

[31] T. Chu, N. Fujioka, and M. Ishizaka, "Compact, lower-power-consumption wavelength tunable laser fabricated with silicon photonic wire waveguide micro-ring resonators," *Opt. Express*, vol. 17, no. 16, pp. 14 063–14 068, Aug 2009.

[32] A. A. Hafez, M. S. Chen, and C. K. K. Yang, "A 32-to-48gb/s serializing transmitter using multiphase sampling in 65nm cmos," in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb 2013, pp. 38–39.

[33] K. Yu, H. Li, C. Li, A. Titriku, A. Shafik, B. Wang, Z. Wang, R. Bai, C. H. Chen, M. Fiorentino, P. Y. Chiang, and S. Palermo, "22.4 a 24gb/s 0.71pj/b si-photonic source-synchronous receiver with adaptive equalization and microring wavelength stabilization," in *IEEE International Solid-State Circuits Conference - Digest of Technical Papers*, 2015.

[34] K. T. Settaluri, S. Lin, S. Moazeni, E. Timurdogan, C. Sun, M. Moresco, Z. Su, Y. H. Chen, G. Leake, D. LaTulipe, C. McDonough, J. Hebding, D. Coolbaugh, M. Watts, and V. Stojanovi, "Demonstration of an optical chip-to-chip link in a 3d integrated electronic-photonic platform," in *ESSCIRC Conference 2015 - 41st European Solid-State Circuits Conference (ESSCIRC)*, Sept 2015, pp. 156–159.

[35] R. Morris, A. K. Kodi, and A. Louri, "Dynamic reconfiguration of 3d photonic networks-on-chip for maximizing performance and improving fault tolerance," in *2012 45th Annual International Symposium on Microarchitecture*, 2012.

[36] P. Koka, M. O. McCracken, H. Schwetman, C. H. O. Chen, X. Zheng, R. Ho, K. Raj, and A. V. Krishnamoorthy, "A micro-architectural analysis of switched photonic multi-chip interconnects," in *2012 39th Annual International Symposium on Computer Architecture (ISCA)*, June 2012, pp. 153–164.

[37] S. Koohi, Y. Yin, S. Hessabi, and S. J. B. Yoo, "Towards a scalable, low-power all-optical architecture for networks-on-chip," *ACM Trans. Embed. Comput. Syst.*, vol. 13, 2014.

[38] A. V. Krishnamoorthy, R. Ho, X. Zheng, H. Schwetman, J. Lexau, P. Koka, G. Li, I. Shubin, and J. E. Cunningham, "Computer systems based on silicon photonic interconnects," *Proceedings of the IEEE*, vol. 97, no. 7, pp. 1337–1361, 2009.

[39] Y. Demir, Y. Pan, S. Song, N. Hardavellas, J. Kim, and G. Memik, "Galaxy: A high-performance energy-efficient multi-chip architecture using photonic interconnects," in *Proceedings of the 28th ACM International Conference on Supercomputing*. New York, USA: ACM, 2014.

[40] S. Beamer, K. Asanovic, and D. Patterson, "Locality exists in graph processing: Workload characterization on an ivy bridge server," in *Proceedings of the 2015 IEEE International Symposium on Workload Characterization*, ser. IISWC '15. Washington, DC, USA: IEEE Computer Society, 2015.

[41] M. Radulovic, D. Zivanovic, D. Ruiz, B. R. de Supinski, S. A. McKee, P. Radojković, and E. Ayguadé, "Another trip to the wall: How much will stacked dram benefit hpc?" in *Proceedings of the 2015 International Symposium on Memory Systems*. New York, USA: ACM, 2015.

[42] M. Badr and N. E. Jerger, "Synfull: Synthetic traffic models capturing cache coherent behaviour," *SIGARCH Comput. Archit. News*, vol. 42, no. 3, pp. 109–120, Jun. 2014.

[43] M. M. S. Aly, M. Gao, G. Hills, C. S. Lee, G. Pitner, M. M. Shulaker, T. F. Wu, M. Asheghi, J. Bokor, F. Franchetti, K. E. Goodson, C. Kozyrakis, I. Markov, K. Olukotun, L. Pileggi, E. Pop, J. Rabaey, C. R, H. S. P. Wong, and S. Mitra, "Energy-efficient abundant-data computing: The n3xt 1,000x," *Computer*, vol. 48, no. 12, pp. 24–33, Dec 2015.

[44] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary, "Firefly: Illuminating future network-on-chip with nanophotonics," *SIGARCH Comput. Archit. News*, vol. 37, no. 3, pp. 429–440, Jun. 2009.

[45] C. Nitta, M. Farrens, and V. Akella, "Dcof-an arbitration free directly connected optical fabric," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2012.

[46] S. Beamer, C. Sun, Y.-J. Kwon, A. Joshi, C. Batten, V. Stojanović, and K. Asanović, "Re-architecting dram memory systems with monolithically integrated silicon photonics," *SIGARCH Comput. Archit. News*, vol. 38, no. 3, 2010.

[47] C. Batten, A. Joshi, V. Stojanovic, and K. Asanovic, "Designing chip-level nanophotonic interconnection networks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 2, pp. 137–153, June 2012.

[48] X. Ye, Y. Yin, S. J. B. Yoo, P. Mejia, R. Proietti, and V. Akella, "Dos: A scalable optical switch for datacenters," in *Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, ser. ANCS '10. New York, USA: ACM, 2010, pp. 24:1–24:12.

[49] A. Joshi, C. Batten, Y. J. Kwon, S. Beamer, I. Shamim, K. Asanovic, and V. Stojanovic, "Silicon-photonic clos networks for global on-chip communication," in *3rd ACM/IEEE International Symposium on Networks-on-Chip*, 2009.