This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2015.2510656, Journal of Lightwave Technology

Flat-Topology High-Throughput Compute Node with AWGR-based Optical-Interconnects

Paolo Grani, Roberto Proietti, Stanley Cheung and S. J. Ben Yoo, *Fellow, IEEE, Fellow, OSA*

***Abstract*—This work presents simulation studies on the execution time and energy consumption of optical Multi-Socket Boards (MSBs) with on-chip, all-to-all, and contention-less Arrayed Waveguide Grating Routers (AWGR)-based interconnection. This study considers throughput and energy-efficiency optimizations based on Dynamic Voltage and Frequency Scaling (DVFS) under realistic, shared memory, and cache-coherent PARSEC benchmarking traffic. The benchmark results show how a low-latency, optical inter-socket interconnection can provide significant execution time reduction, and up to 3× energy savings when using dynamic variable bandwidth communication techniques when compared to an electronic baseline. The proposed architecture can be used as "building block" for future energy-aware large-scale systems.**

***Index Terms*— Arrayed Waveguide Grating Routers, Chip Multi-Processor Systems, Optical Interconnects, Tiled CMP Architectures.**

I. Introduction

Next generation exascale computing and data systems must support significantly increased data traffic at all scales due to extreme-scale data-sets and working-sets. These requirements can lead to high power consumption and low energy efficiency due to: (a) poor scalability of their interconnection architectures (typically thousands of boards interconnected through multi-stage switching networks as Fat-Tree, Flattened Butterfly, or Torus [1-4]); and (b) inefficient utilization of transmission modules that continue to consume power to transmit synchronization and framing bits [5] (in order to keep the receivers locked) even when no actual data information bits need to be transmitted [6]. Currently, the energy cost of moving data is becoming the dominant factor for energy consumption and has overshadowed the energy cost of data processing and storage. A recent analysis [7], shows that data movement approximately represents about half of the power budget of a regular desktop and almost one-fourth of a server. It is evident that future exascale computing systems, utilizing classical electronic infrastructure, would not be able to sustain power consumption below 30 MW by 2020 [8, 9].

Recently, these large-scale architectures adopt Multi-Socket Boards (MSBs) to increase both the computation density as well as the number of resources (e.g., RAM and cores) that applications can access with limited overhead due to physical proximity. These MSBs, as further discussed in Section III.A, typically have a standard, non-tiled electronic architecture and utilize a protocol (i.e., Quick Path Interconnect (QPI) [10] or HyperTransport (HT) [11]) for inter-socket communications. As shown in Fig. 1 (left), a non-tiled electronic architecture has more than one core (C in Fig. 1) connected to the same Network Interface (NI) for the inter-socket transmissions through a shared bus, also exploited to access the Last-Level Cache (LLC). To achieve inter-socket communications, an electronic MSB requires the crossing of logics (bus and crossbar) and the interfacing towards the I/O pins to cross the package, resulting in additional communication latency on the order of tens of nanoseconds [11, 12].

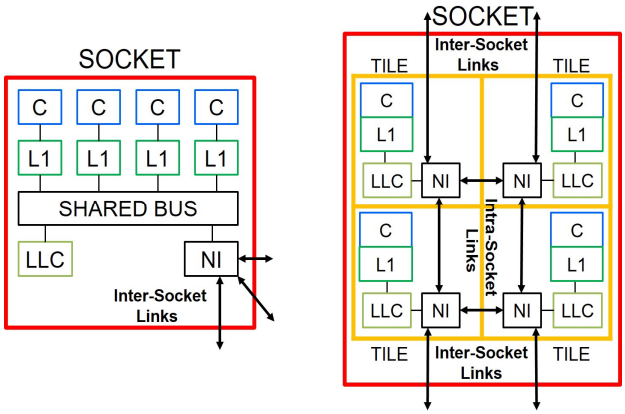

Fig. 1. Non-tiled (left), and tiled (right) electronic architecture for inter-socket communication in a Multi-Socket Board (MSB) [12]. C: Core; L1: Level-1 Cache; LLC: Last-Level Cache; NI: Network Interface.

Last generation electronic Chip Multi-Processor (CMPs) architectures adopt *tiled* topologies [13, 14]. A tiled CMP, as shown in Fig. 1 (right), is a multi-core system with a shared LLC and Non-Uniform Cache Access (NUCA) architecture [15] in which all its cores share the physically distributed cache banks. Each *tile* is composed of one switch (NI in Fig. 1) for the communication with the other tiles, one core, and some cache resources (typically a private L1 cache and a slice of a shared and distributed LLC). These systems demand high memory bandwidth because of the increased number of cores per chip [16, 17] and because of the requirement of low-latency memory access [18]. A high-radix switch supporting contention-less, low-latency, and power-efficient Network-on-Chip (NoC) is necessary to exploit the full potential of a multi-threaded application running on a tiled CMP.

This work was supported in part under DoD Agreement Number: W911NF-13-1-0090. P. Grani, R. Proietti, S. Cheung, and S. J. Ben Yoo are with the Department of Electrical and Computer Engineering, University of California, Davis, CA, 95616, USA (e-mail: pgrani@ucdavis.edu, rproietti@ucdavis.edu, stacheung@ucdavis.edu, and sbyoo@ucdavis.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2015.2510656, Journal of Lightwave Technology

Silicon Photonics (SiP) [19] is a promising technology for on-chip and on-board interconnections due to its intrinsic low-latency, low-power, and compact features. This has the potential to increase both performance and energy efficiency of future CMPs [20]. The impact of SiP is expected to be significant particularly in tiled CMPs because they facilitate placement of chip-scale optical switches for NoC. In this paper we analyze the implications of tight integration of optical and electrical components inside the processor package, considering to have the inter-chip optical links directly connected to the same switches that serve the on-chip links of the tiled CMP. With the capabilities offered by dense optical integration, it is easier to deploy communication resources closer to the computational ones and, therefore, it is possible to shorten the path for the inter-socket transmissions, avoiding the aforementioned latencies introduced by a classical electronic architecture. This intimate integration is enabled by advanced 3D integration techniques involving Through-Silicon-Vias (TSVs) [21, 22].

Silicon photonic interconnects utilizing Arrayed Waveguide Grating Routers (AWGR) [23, 24] offer contention-free, all-to-all, low-latency, and high-bandwidth interconnections for future CMPs. While optical interconnects offer communication bandwidth several orders of magnitude greater than electronic counterparts, power consumption is a critical issue.

As mentioned above, a significant part of the power consumption lies in the transmission of synchronization bits, even when no actual data transmission is required. For instance, standard Clock and Data Recovery (CDR) techniques assume that the transmitters (TXs) continue to use line-coding (e.g., 64b/66b) to limit the maximum run length of the CDR circuitry [5]. Therefore, TXs continue to send modulated signals (synchronization bits and framing bit sequences) at their specified maximum line-rate even when they have no data to transmit (their buffers are empty). In comparison, Dynamic Voltage and Frequency Scaling (DVFS) with source-synchronous transmission [25-28] allows one to effectively utilize the transmission modules to prevent the waste of communication resources and power, as discussed in Section III.B. The evaluation and benchmarking of energy efficient computing systems is a non-trivial problem. The power efficiency of various architectures has been simply estimated in terms of energy-per-bit (Joules/bit) and fails to take into account the context of the system in the network such as the actual utilization rate of the system, and the application it is running during the evaluation. We can achieve a more accurate estimation of the energy efficiency by considering the throughput instead of the raw bit rates. However, even this method overestimates the energy efficiency since the actual useful bits are the ones seen by the application. Hence, the *goodput* (i.e., the number of useful information bits delivered by the network to a certain destination per unit of time), is considered a better metric than the system throughput [29].

This paper reports a benchmark study of optical Multi-Socket Boards exploiting both AWGR-based, all-to-all interconnection capabilities, and dynamic bandwidth reconfiguration techniques based on Source Synchronous communications with DVFS. In particular, we focus on the detailed analysis of next generation CMP performance (execution time and energy consumption). Our benchmarking studies compare the proposed architectures with a state-of-the-art electronic topology based on current QPI or HT inter-socket transmission protocols.

The main contributions of this paper are:

- Investigated the design and modeling of a fully connected Multi-Socket Board (MSB) architecture exploiting AWGR-based interconnection. We modeled a detailed state-of-the-art electronic configuration to have a fair comparison with our proposed optical architecture;
- Evaluated performance studies of the proposed architecture using the PARSEC-2.1 [30] benchmark suite. In particular, we show the achievable results in terms of execution time and Energy Delay Product (EDP), and provide comparison against a state-of-the-art electronic switch counterpart;
- We evaluated different optimization techniques (DVFS) to explore tradeoffs under the load of real benchmarking traffic and to achieve better application level throughput (goodput) avoiding the transmission of synchronization bits.

The remainder of the paper is organized as follows. In Section II we introduce the MSB architecture, and we analyze the optical technologies enabling this work. In Section III, we present the methodology. Section IV discusses the achieved result, and finally, Section V summarizes key findings.

II. Multi-socket board with SI-LIONS-based all-to-all interconnection

A. Multi-Socket Board

State-of-the-art commercial boards have the ability to host four sockets with each one directly connected to separate, but shared, DRAM banks. Each socket can host 16 cores chips (e.g., AMD 6300-class) or 12 hyperthreaded core chips (e.g., Intel E5-4657L v2). Programmers can use a shared-memory paradigm (something that is extremely desirable for transparency and flexibility [31]) in which all the memory space is accessible from all the processors on a flat topology. Multi-Socket Boards (MSBs) allow processors to be closer than the board-to-board communication configuration, thus allowing faster and more energy-efficient communication between the sockets [12]. Optical interconnects utilizing AWGR can provide contention-free communications, eliminating the need for repeating or regenerating the data between the sockets. Furthermore, optical communications enable the direct integration of transmission modules on-chip. In our proposed architecture, data transmission can directly arrive at the integrated Hubs instead of passing by the package transceivers and move ahead from that point towards the cores.

Fig. 2 shows a logic scheme of the proposed Multi-Socket Board architecture. The board contains N sockets optically interconnected through the AWGR for all-to-all, contention-less communications. We considered using off-chip Optical Frequency Comb (OFC) generators and splitters in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2015.2510656, Journal of Lightwave Technology

proposed architecture, although on-chip lasers and splitters are conceivable in future systems. Also, we exploit an embedded electronic switch (named Hub in Fig. 2) acting as the interface between the electronic (intra-socket) domain and the optical (inter-socket) AWGR-based, all-to-all network. All the modules required for the optical communications link (Electro-Optical (EO) and Optical-Electro (OE) converters, drivers, Dual Clock FIFO buffers and microring resonators) are monolithically integrated.

The Network Interface (NI) integrated into the Hub has been modeled according to [32] neglecting the implementation of the credit mechanism (contention-less communication, no backpressure) and taking into consideration the required transmission parallelism. Section III describes this in detail. For the CDR case, we are not transmitting the wavelength associated with the clock. To achieve an inter-socket communication, packets first propagate through an electronic Mesh to reach the Hub. The packets then cross the chip boundary in the optical domain. The Hub is an electronic embedded switch connecting the intra-chip tiles with the tiles in other sockets through optical links. Embedding the switch directly in the chip reduces the latency of the initial/final hop. Each socket has one embedded electronic switch with Wavelength Division Multiplexing (WDM) optical I/Os, as shown in Fig. 2. The embedded switch performs the routing function within the socket to forward the packets to the proper transceiver (TRX) port based on the destination address of the packets. Each electronic intra-chip port has its own buffering structures, and the optical interface comprises all the required modules (microring resonators, transmitter, receiver, serializer, deserializer, multiplexer, and demultiplexer) integrated into the Hub. Each MSB requires p and μ wavelengths for intra-board and inter-board communication. Note that, the simulation studies in this paper focus only on the intra-board domain. Ref. [33] shows a solution for inter-board interconnect architecture.

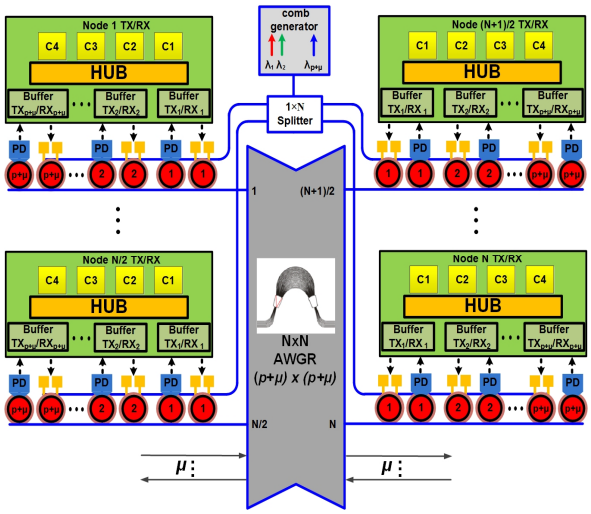

Fig. 2. Optically interconnected Multi-Socket Board, N sockets. Each socket has four cores and p+μ transmitters and receivers for intra- and inter-board communications through AWGR.

An AWGR [23, 34] is a wavelength routing device which allows any input port to communicate with any output port simultaneously and without contention. Thus, a $N\times N$ AWGR provides all-to-all communication among N compute nodes in a flat topology using N wavelengths. The optical MSB in Fig. 2 can be implemented as a Si-LIONS architecture [33]. Fig. 3 shows a recently fabricated 8×8 Si-LIONS with a total device area of approximatively 1.2 mm by 3.6 mm. The scalability of the Si-LIONS depends on the scalability of AWGR. In theory, fabrication of large-port-count AWGRs is possible, but limiting factors such as difficulties in accurate wavelength registration [35, 36], high crosstalk due to dense channel spacing, and Silicon On Insulator (SOI) technology tolerances, prevent such system from being deployed on a large scale.

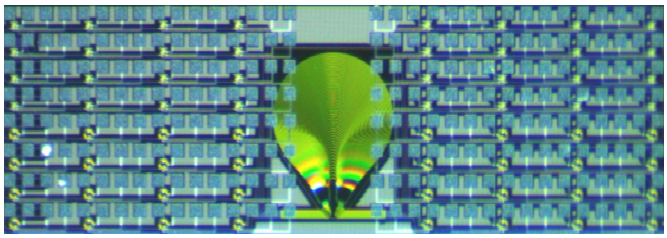

Fig. 3. Silicon-Photonic LIONS chip with a 1.2 mm × 3.6 mm footprint.

However, it is feasible to use small AWGRs with a fewer number of wavelengths, while supporting the same connectivity as large-port-count AWGR [37]. In the case of an N-port AWGR interconnecting N nodes in an all-to-all fashion, each node requires N TRXs, with a total number of N^2 TRXs and N wavelengths per node. We can achieve the same all-to-all connectivity by using M^2 W-port AWGRs (M groups of M W-port AWGRs), with $N=M\times W$. In this case, each node requires N TRXs grouped as M groups of W wavelengths.

Interconnecting N nodes is possible by using W ($W < N$) wavelengths and $W\times W$ AWGRs. Meanwhile, the Free Spectral Range (FSR) of the ring resonators should be larger than the channel spacing of the $N\times N$ AWGR, so that the ring resonance wavelength aligns with only one of the AWGR passbands. Assuming a ring resonator with a 5 μm radius [38] (approximatively 2.4-THz FSR near 1550 nm wavelength), the Si-LIONS can easily accommodate 32 (48 maximum) wavelength channels with a 50 GHz channel spacing. Based on the above analysis, a larger scale switch can be constructed from 32×32 AWGRs [38] and ring resonators with 5 μm radius.

III. Simulation Methodology

We used the GEM5 simulator [39] 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 CMP architectures with up to 64 cores distributed on 4 or 16 sockets. Specifically, we relied on the Alpha Instruction Set Architecture (ISA) already integrated inside the GEM5 simulator. It should be noted that, in FS mode, an actual Alpha platform only supports up to four processors. To overcome this limitation, we relied on a variant of the Alpha platform that can take up to 64 processors [40].

In the considered tiled architecture, each core has private L1 caches (Instruction+Data), a slice of a shared and distributed L2 cache (LLC), and an electronic router/switch used for network communications. The directory information is also distributed, and a coherence protocol (i.e., MOESI) manages the directory-based, cache-coherent communication.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2015.2510656, Journal of Lightwave Technology

Table 1 summarizes the main architectural features of the overall tiled CMP architecture. A 2D Mesh via a Network-on-Chip (NoC) interconnects the intra-socket cores (tiles).

The simulation results were evaluated for the PARSEC-2.1 benchmarks suite [30], a collection of heterogeneous parallel applications spanning different application domains (e.g., media processing, search and filtering, 3D, and physics simulations) and representative of diverse workloads that can be run on nowadays CMP devices. The characterization of this suite in terms of memory access patterns for the different benchmarks has been extensively analyzed in [30, 41] and it is beyond the scope of this paper. This suite splits the application load on all the available processors and cores, and relies on shared memory and cache-coherency for the communication and synchronization between the working threads mediated by the simulated Linux OS. 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 an electronic baseline discussed in details in the following section. For the proposed optical architecture, we considered three setups: a 4S4C system with four sockets, each one composed of four cores (for a total number of 16 cores); a 16S4C system with 16 sockets, each one also composed by four cores (for a total number of 64 cores); a 4S16C setup with four sockets, each one composed of 16 cores, for a total of 64 cores. The latter setup is also the reference for the electronic baseline in the 64-core configuration, as explained in the next section.

We limit the total number of wavelengths managed by the AWGR to a maximum value of 64 (WDM). By exploiting multiple AWGR Free Spectral Ranges (FSRs) [23], the AWGR-based architecture (in the 4S4C and 4S16C systems) allows four inter-socket optical links with 16-bit parallelism for a total of 48 wavelengths (p=48 intra-board wavelengths). As a result, each Hub should be able to communicate with other Hubs in the considered architectures. In the four socket case (4S4C and 4S16C) this means that each Hub has to reach three other Hubs. Therefore, using 16-bit parallelism, a total of 16×3=48 wavelengths are required. We could push this configuration to exploit up to a 21-bit parallelism (21×3 = 63 lambdas), but we decided to use a power of two *flit* (FLow control digITs) value, as typically considered in an architectural analysis. Furthermore, we aimed to leave some lambdas (μ wavelengths) for inter-board communications. This topic is beyond the scope of this paper but it has been discussed in some previous works [33, 42] for synthetic traffic analysis.

TABLE 1. PARAMETERS OF THE SIMULATED ARCHITECTURE.

Cores	16/64 64-bit processors, 5 GHz
L1 caches	16kB (I)+16kB (D), 2/4-way (D/I), 1 cycle
L2 cache	16 MB, 8-way, 16×1 MB or 64×256 kB banks, 4/12 cycles tag/tag+data
Directory	MOESI coherence protocol, 16/64 slices, 1 cycle
ENoC intra	Mesh, 5 GHz, 4 cycles, 32-bit, 2.5 mm
ENoC inter	p2p links, 5 GHz, 30 ns latency, 32 bit/flit, 12.5 cm
ONoC inter	4/16 bit/flit p2p optical links and AGWR, 10GHz, 12.5 cm, 13 cycles

The 16S4C setup makes use of only 4-bit parallelism for a total of 60 lambdas (4×15=60). Note that, for all the considered setups, one of these lambdas is dedicated to the clock distribution when utilizing Source Synchronous technique instead of Clock and Data Recovery (CDR) [43] (see Section III.B). The maximum line rate for each lambda is set to 10 Gb/s. We modeled the electronic baseline for the inter-socket interconnection according to the current performance of HyperTransport-3.1 [11], as discussed in the next section.

A. Electronic Baseline

As discussed in Section II.A, Multi-Socket Boards are largely diffused as a "building block" for large-scale architectures. However, MSBs performance heavily depends on the effectiveness of the peer-to-peer interconnection capability of processors and the boards themselves. Intel QPI [10] and AMD HyperTransport (HT) [11] are state-of-the-art examples of board-level interconnections capable of implementing various topologies. Over the years, the number and bandwidth of the links has increased. For instance, AMD Opteron-6378 has four 12.8 GBps (102 Gbps) HT links. For these solutions, inter-chip transmission latencies of around 40 ns are reported [12]. Sources of latency are due to the distance between sockets, the need for the signals to cross the processors packages, the inter-socket voltage domain changes, and the non-negligible bus interface/adaptation logic (crossbars and electronic circuitry) for supporting the protocol [44]. Specifically, the implementation of split transactions, i.e., the capabilities of sending out different requests without waiting for the related replies from the memory, comports a further latency increase. Indeed, the aforementioned board-level interconnections need buffers to maintain the unsolved requests, and also a protocol to manage the status of the outstanding requests as soon as the replies arrive. HyperTransport-3.1 is rated to work up to 3.2 GHz Double Data Rate (DDR) and with up to 32-bit parallelism. Our reference baseline has an aggressive 32-bit parallelism at 5GHz, 160 Gbps bandwidth, and 30 ns inter-socket head-flit latency (i.e., the time for the first *flit-bits* to reach the destination).

The main difference between electronic non-tiled MSB architectures and the proposed optical tiled solutions is that, in the optical domain, we can directly cross over the chip boundary to provide a seamless interconnection model by bringing the communications closer to the cores, within the integrated Hub. As already discussed, we are considering two setups in our analysis: 4S4C (four sockets each with four cores) and 16S4C (16 sockets each with four cores). The 16S4C setup for the electronic baseline is a challenge. In recent multi-socket architectures [44], pin constraints imposed a limit of four HT-3.1 ports on the package. To reach a full connectivity between the considered four chips, the architecture presented in [44] makes use of an electronic switch acting as an interface for the transmission [45] between the processors, introducing further latency in the inter-chip communications. Therefore, we withdraw the consideration of applying the 16S4C setup to the 64-core case for the electronic baseline. In fact, to achieve a full connectivity between 16 endpoints (the 16 sockets), up to

15×16=240 inter-socket HT-links would be necessary, requiring a complicated place-and-route layout which results in unacceptable CPU pin burning.

To achieve a 64-core design for the electronic baseline, we considered the 4S16C (four sockets each with 16 cores) setup that keeps the number of inter-socket HT-links limited to six for full connectivity. Note that the increase of inter-socket connections is true also in a classical optical p2p solution, but the AWGR wavelength multiplexing capability significantly reduces the number of waveguides needed.

B. Optimization Techniques

As mentioned in Section I, a significant part of power inefficiency comes from poorly adapting the communication bandwidth to the traffic patterns. To reach better results in term of goodput, we analyzed the application of different optimizations and adaptive algorithms for the considered setups (4S4C and 16S4C).

In a shared-memory system, the traffic characteristics are very challenging and bursty [46]. Conventional communication systems consume energy even when no actual data transmission is required (this is typically needed to keep the Clock and Data Recovery (CDR) circuit alive). On the other hand, the Source Synchronous model [6] does not require the transmission of synchronization bits. Furthermore, by exploiting Dynamic Voltage and Frequency Scaling (DVFS) [25-28], the system can dynamically set the transmitter frequency and voltage supply to different values depending on the traffic load. The DVFS technique can reduce, as a consequence to the load, the dynamic power of CMOS transistors in the transceivers proportionally to a $(f * Vdd^2)$ factor. Vdd is the driving voltage and f is the clock speed. Since Vdd is lowered for circuits with low f, there can be a significant improvement in energy efficiency when the clock frequency can be lowered in combination with the DVFS technique. We have implemented two different optimizations for DVFS: 1) using 2-level thresholds, maximum (10 GHz, 1.2 V) and minimum (1 GHz, 0.53 V), according to the optical inter-socket links utilization, and 2) using 3-level thresholds in which the system dynamically sets the transmitter frequency and voltage supply to a maximum (10 GHz, 1.2 V), medium (5 GHz, 0.63 V) and minimum (1 GHz, 0.53 V) value, depending on the amount of messages in each Hub's buffer. These voltage and frequency values have been derived from [47]. In both optimizations, we considered the scaling of both voltage and frequency [48]. Specifically, we applied this technique only to the transmission part where the transmitter leads this scaling and the receiver simply follows it to minimize 1) the complexity of the circuitry and the protocol for managing the receiver voltage and frequency values, and 2) the latency introduced by additional circuits and control loops.

C. Performance and Energy Metrics and Model

In our simulation results, we consider two main metrics: execution time and Energy Delay Product (EDP). PARSEC-2.1 benchmarks are composed of a) a well-defined initialization portion, in which the required threads are spawned, b) the parallel region (called "Region of Interest", ROI) and c) the final part in which benchmark resources are released. As for performance analysis, in line with similar works, we considered the execution time of the entire ROI of each benchmark. The execution time is an important metric because it directly reflects the final execution time to accomplish a specified task (application runtime). EDP provides a way to combine both energy efficiency and throughput into the same metric.

TABLE 2. OPTICAL POWER PENALTIES.

Component	Loss (4S4C, 16-bit parallelism)	Loss (16S4C, 4-bit parallelism)
Grating Coupler	-1 [dB]	-1 [dB]
Fiber	-0.0001 [dB/cm]	-0.0001 [dB/cm]
Coupler	-1 [dB]	-1 [dB]
Splitter	-0.2 [dB]	-0.2 [dB]
Modulator + Mux	-5.4 [dB]	-5.4 [dB]
Modulator Array	-0.3 [dB]	-0.069 [dB]
Waveguide (12.5 cm)	-0.1 [dB/cm]	-0.1 [dB/cm]
AWGR	-5 [dB]	-5 [dB]
Photodetector	-0.1 [dB]	-0.1 [dB]
Demultiplexing	-1.184 [dB]	-1.046 [dB]
Total-Power-Penalty	-15.4 [dB]	-15 [dB]

Table 2 shows the optical component parameters, which have been derived from Ref. [33, 49-51]. Table 3 shows the optical energy consumption of all the required modules (i.e., laser, transmitter, and receiver) considered for all optimization cases (with CDR or with DVFS 2- or 3-level).

TABLE 3. ENERGY OPTICAL INTER-SOCKET PARAMETERS.

Photodetector-sensitivity (*cons* / *aggr*)	-17 / -22 [dBm]
Laser-efficiency (*cons* / *aggr*)	4.5% / 10%
P-Lambda (4S4C / 16S4C) *aggr*	0.218 / 0.199 [mW]
P-Lambda (4S4C / 16S4C) *cons*	0.691 / 0.631 [mW]
P-Laser (4S4C / 16S4C) *aggr*	418 / 1910 [mW]
P-Laser (4S4C / 16S4C) *cons*	2948 / 13461 [mW]
P-Transmitter	1.35 [mW]
P-Receiver	3.95 [mW]
P-SER+DES-Lambda-Static	0.076 [mW]
P-SER+DES-Lambda-Dynamic	0.0171 [pJ/bit]
P-CDR-Bit	5 [mW]
Microring-Tuning	0.113 [mW]
Total-E-Dynamic	0.53 [pJ/bit]
Total-P-Static DVFS (4S4C / 16S4C) *aggr*	476 / 2200 [mW]
Total-P-Static DVFS (4S4C / 16S4C) *cons*	3006 / 13751 [mW]
Total-P-Static CDR (4S4C / 16S4C) *aggr*	1436 / 7000 [mW]
Total-P-Static CDR (4S4C / 16S4C) *cons*	3966 / 18551 [mW]

In our analysis we considered two scenarios: an *aggressive* one (*aggr* in the tables), in which we assumed a laser efficiency of 10% [52] and a photodetector sensitivity of -22 dBm [53], and a *conservative* setup (*cons* in the tables), in which we considered a laser efficiency of 4.5%, and a photodetector sensitivity of -17 dBm [37]. The required power per lambda (*P-Lambda* in Table 3) can be calculated considering the difference between the photodetector sensitivity (for both the aggressive and conservative cases) and the *Total-Power-Penalty* reported in Table 2 for each setup (4S4C and 16S4C). Table 3 reports the worst-case power required for each scenario and each architecture. The laser power consumption (*P-Laser* in Table 3, 16-bit parallelism for 4S4C setup, and 4-bit parallelism for 16S4C one) is calculated considering all the optical losses along the end-to-end path in the worst-case scenario. Specifically, starting from the P-Lambda parameter in each configuration, we applied the considered laser efficiency (4.5% for the conservative case and 10% for the aggressive one) and finally

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2015.2510656, Journal of Lightwave Technology

multiplied all considered lambdas (48 or 60, depending on the case) and for the number of considered chips (and, therefore, lasers) depending on the architecture (4 in the 4S4C and 16 in the 16S4C). Transmitter and receiver values are based on [54]. We derived the CDR circuitry power consumption from [55], and applied this power consumption both to the electronic baseline and to the proposed CDR-based optical solutions. The microring resonators are thermally tuned to operate at evenly spaced wavelengths around 1550 nm, thus forming high bandwidth WDM channels [56]. To calculate the total dynamic consumption per bit (*Total-E-Dynamic* in Table 3), we considered the power consumption of the transmitters, receivers (which perform the electro/optical and the optical/electro conversions, respectively), and the dynamic contribution of the Serializer and Deserializer circuitry (SERDES, [57]). To calculate the total static power consumption (*Total-P-Static* entries in Table 3, for each setup), we considered the static consumption of the laser, the tuning currents, the SERDES static consumption, and the CDR circuitry (where utilized).

TABLE 4. ELECTRONIC PARAMETERS OF THE SIMULATED ARCHITECTURE.

E-link-intra	0.25 pJ/bit (2.5 mm)
E-link-inter	15.5 pJ/bit (12.5 cm)
E-switch-dyn (3-port)	5.29 pJ/flit (×16)
P-switch-static (3-port)	8.76 mW (×16)
E-switch-dyn (4-port)	7.24 pJ/flit (×16 / ×32 / ×64)
P-switch-static (4-port)	11.84 mW (×16 / ×32 / ×64)
E-switch-dyn (6-port)	30.38 pJ/flit (×16)
P-switch-static (6-port)	30.69 mW (×16)
E-hub-dyn (5-port)	9.38 pJ/flit (×4 / ×16)
P-hub-static (5-port)	15.05 mW 4 (×4 / ×16)

Table 4 shows the setup for the electrical parameters (intra-socket, and inter-socket HT-based, p2p links for the electronic baseline). Electronic energy parameters shown in this table are derived from the DSENT [58] and from [59]. The Hubs (introduced in Section II.A) are switches which interface the intra-socket network with the inter-socket optical one. In all electronic switches, we assumed to have one additional port for local communications towards caches and core. The number of ports of each switch is based on the considered topology.

IV. BENCHMARKS RESULTS DISCUSSION

This section discusses the achieved results for all the considered setups discussed above, from both execution time and energy efficiency perspectives.

A. Performance Results

Fig. 4 shows the achieved results in terms of performance for the 4S4C case (Fig. 4 (a)) and the 16S4C one (Fig. 4 (b)). The optical architecture with CDR always achieves the best execution time (the lower the better). Note that, the system with CDR is not affected by any latency related to the frequency and voltage scaling. Instead, with the DVFS setup, we assume to have 10 ns latency for voltage and frequency scaling operation [48]. With the considered real cache-coherent traffic, short and latency-critical 64-bit control messages represent 75% of the network load. Moreover, the traffic is very bursty. Therefore, DVFS has to often pay a 10-ns latency before each transmission.

Fig. 4 (b) shows that, for the bigger 64-core setup, the DVFS 3-level (DVFS-3 bars in the figure) performs significantly worse than the DVFS 2-level (DVFS-2 bars in the figure) even if it is still able to outperform, on average, the electronic baseline. In fact, when a higher number of cores request the inter-socket links, the probability of optical communication increases. Looking at the buffer utilization by using three different voltage and frequency levels, and assuming to have 10 ns latency to change the state for each transition, the DVFS 3-level optimization is most likely to pay this latency and results in the worst performance results.

B. Energy Delay Product Results

Fig. 5 shows the achieved results in terms of Energy Delay Product (EDP) for the 4S4C (four sockets each with four cores) setup with 16 cores in total. Fig. 5 (b) shows that all the optical solutions are, on average, performing better than the electronic baseline, achieving up to around 70% EDP reduction when exploiting the DVFS techniques. Among the considered optical solutions, Fig. 5 also shows that the CDR has the worst results in terms of EDP. In fact, with CDR, all the required optical

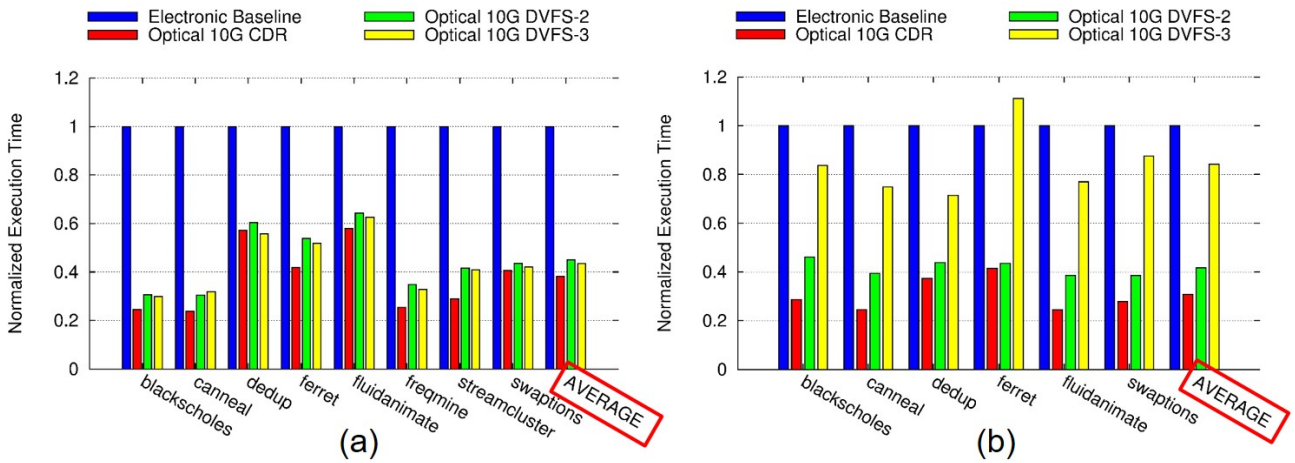

Fig. 4. Execution time results normalized to the electronic baseline for the 4S4C setup (a) and for the 16S4C one (b). The first bars in the figures represent the electronic baseline; the second ones the optical configuration with the CDR; the third ones represent the optical configuration with source synchronous and 2-level voltage and frequency scaling based on link utilization and, finally, the forth ones the optical configuration with source synchronous and 3-level voltage and frequency scaling based on buffer thresholds.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2015.2510656, Journal of Lightwave Technology

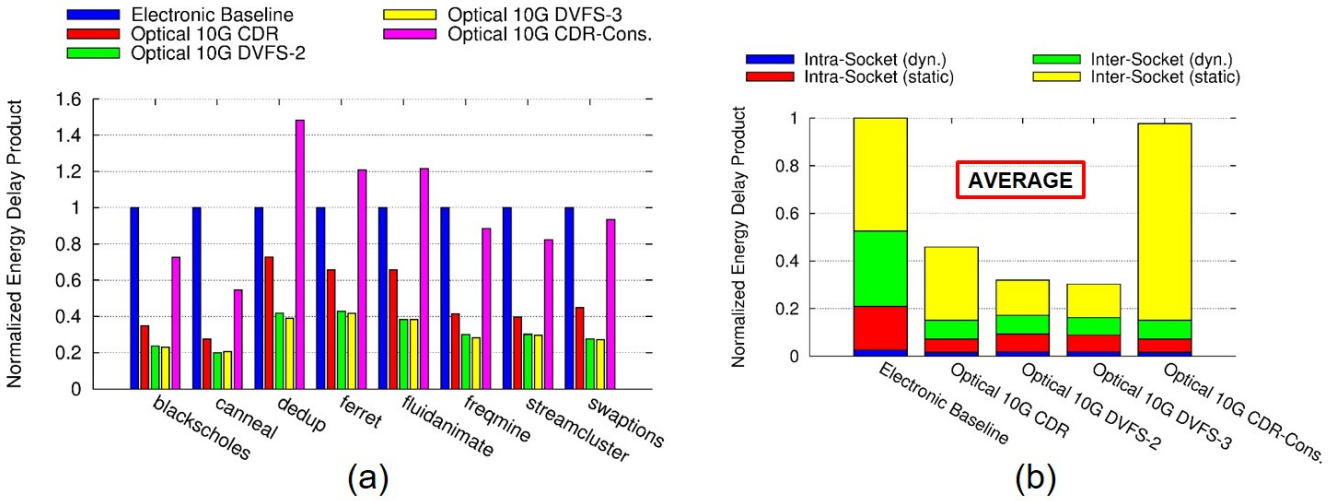

Fig. 5. Energy Delay Product results normalized to the electronic baseline for the 4S4C setup. The first bars in (a) represent the electronic baseline; the second ones the optical configuration with the CDR in the aggressive case; the third ones represent the optical configuration with source synchronous and 2-level voltage and frequency scaling based on link utilization; the forth ones the optical configuration with source synchronous and 3-level voltage and frequency scaling based on buffer thresholds, and, finally, the last bars represent the optical configuration with the CDR in the conservative case. (b) represents the EDP average value for all the considered benchmarks in the different setups. Each columns in (b) is split in stacks representing intra- and inter-Socket dynamic (dyn.) and static energy consumption.

modules are always working at the maximum frequency and voltage (no energy optimizations are performed and the goodput is getting worse). To stress the optical solution, we also reported the results achieved with a conservative setup for the CDR case (last bars in Fig. 5). As described in Section III.C, the laser efficiency and the receiver sensitivity are in line with mature current technology. The achieved results show that using these consolidated values, the EDP increases considerably (the higher the worse), mainly because the static consumption of the laser is increasing. However, for the smaller 4S4C case with the optical CDR solution, it is still able to slightly outperform the electronic baseline while maintaining good execution time performance as shown in Fig. 4 (a). Note that, the DVFS 2-level performs similarly to the DVFS 3-level even if the applied methodology is different, as explained in Section III.B. Fig. 6 shows the achieved results in terms of EDP for the 16S4C (16 sockets each with four cores) setup with 64 cores in total. In this case, there are significant differences between the two DVFS solutions. When the optical inter-socket links are more likely to be utilized, the DVFS 3-level scenario experiences higher execution time, as shown in Fig. 4 (b) and, therefore, the EDP increases proportionally. Specifically, the *Inter-Socket (static)* part (yellow stacks in Fig. 6 (b)), which represents the static energy consumption, increases significantly. In the 16S4C case, for the conservative CDR case, even if the execution time achieved by the proposed optical CDR solution is much better than the electronic baseline (as shown in Fig. 6 (b)), the laser power consumption becomes a

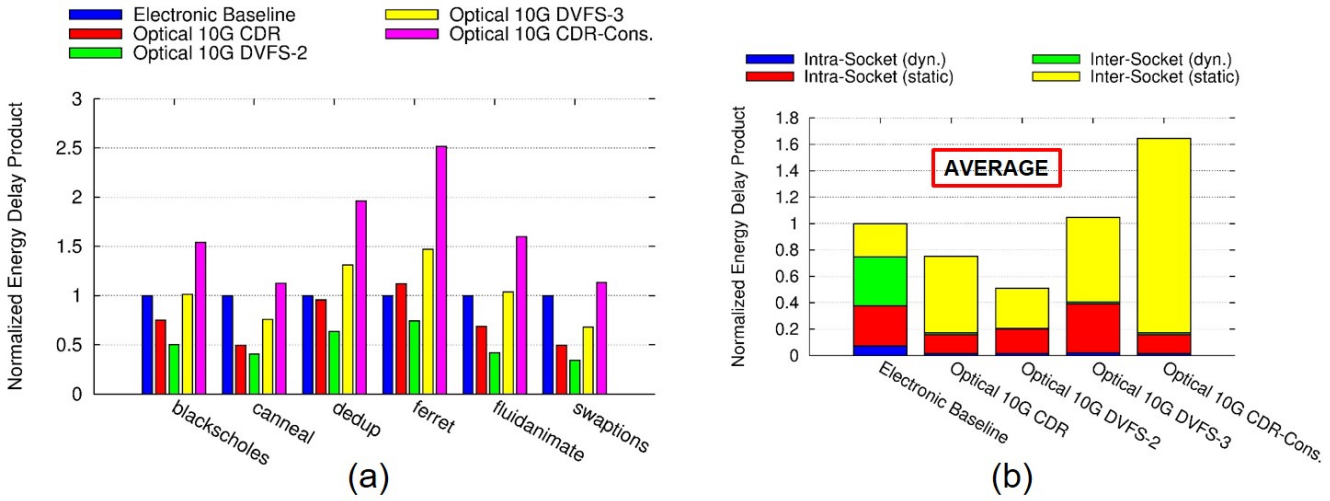

Fig. 6. Energy Delay Product results normalized to the electronic baseline for the 16S4C setup. The first bars in (a) represent the electronic baseline; the second ones the optical configuration with the CDR in the aggressive case; the third ones represent the optical configuration with source synchronous and 2-level voltage and frequency scaling based on link utilization; the forth ones the optical configuration with source synchronous and 3-level voltage and frequency scaling based on buffer thresholds, and, finally, the last bars represent the optical configuration with the CDR in the conservative case. (b) represents the EDP average value for all the considered benchmarks in the different setups. Each columns in (b) is split in stacks representing intra- and inter-Socket dynamic (dyn.) and static energy consumption.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2015.2510656, Journal of Lightwave Technology

significant factor on the overall EDP results. In fact, the static energy consumption (see Fig. 6 (b)), on average, increases a lot, making the analyzed optical solution not competitive with the considered electronic baseline (see last bars in Fig. 6 (b)). This suggests that technology improvements of the current technology (especially in terms of laser efficiency) are crucial to achieve good energy efficiency in the full-system architecture. At the same time, the high power required by the laser is a point in favor of placing the lasers and splitters off-chip. Furthermore, the possibility of having integrated on-chip lasers with WDM capabilities is still not considered a fully available and stable solution.

Finally, comparing the 4S4C (Fig. 5) and the 16S4C (Fig. 6) results, for all the considered optical solutions, the *Inter-Socket (dynamic)* part is getting smaller (the smaller the better) in the 16S4C case. Indeed, in the 16S4C setup, we assumed to have a 4-bit transmission parallelism. Therefore, the lower number of optical modules brings to a lower dynamic consumption for each single optical transmission. Finally, we want to remark that, also in the larger 16S4C setup, the CDR technique experiences higher EDP compared to the DVFS 2-level setup.

C. Architectural Designs Comparison

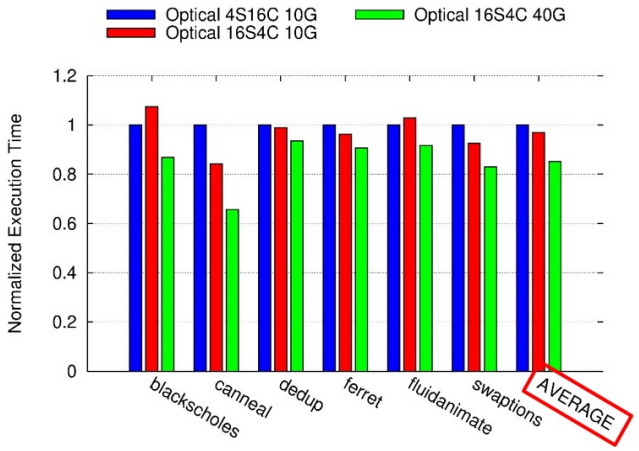

Fig. 7. Normalized execution time presented as a comparison between the 4S16C optical setup (first bars) and the 16S4C one, with 10GHz (second bars) and 40GHz (third bars) with CDR technique.

This section analyzes two different architectural choices. Specifically, we compare the 4S16C (four sockets each with 16 cores) and 16S4C (16 sockets each with four cores) optical setups (both with a total of 64 cores) to evaluate different tradeoffs between the achieved execution time results. Fig. 7 shows the results of this analysis for the CDR case. Similar performance can be expected for DVFS optimizations.

In the optical 4S16C setup (first bars in Fig. 7), the high number of electronic intra-socket hops among the 16 cores negatively affect the performance (the lower, the better in Fig. 7). In fact, in our topology, we supposed to place the Hub in the central zone of the socket and only the four cores surrounding it have a direct connection towards it. If, for instance, the initial request starts from one of the cores in the corner of the considered intra-socket, 2D electronic Mesh, at least two electronic hops must be performed to reach the Hub. Therefore, additional latency is introduced in the architecture. Furthermore, due to our assumptions on the optical frequency (10 GHz) and maximum number of lambdas (64 wavelengths), the 16S4C setup is limited to a 4-bit parallelism (4-bit @ 10 Gbps = 40 Gbps), as explained in Section III. Therefore, the optical 16S4C setup is penalized, in terms of bandwidth, in comparison to the 4S16C case (16-bit @ 10 Gbps = 160 Gbps). Nevertheless, the 16S4C @ 10 GHz is capable to slightly outperform, on average, the 4S16C case. To have a fair bandwidth comparison, we increased the optical frequency to 40 GHz for the 16S4C setup (4-bit @ 40 Gbps = 160 Gbps). Fig. 7, with its third bars, shows a performance improvement over the 4S16C setup (more than 15% on average), confirming the fact that the intra-socket electronic hops can degrade the achieved performance.

V. Conclusions

We proposed an optical interconnected Multi-Socket Board (MSB) architecture for future Chip Multi-Processor systems (CMPs) enabled by an AWGR-based, all-to-all, and contention-less inter-socket communication model. Through realistic benchmarking studies, we have examined the final application execution time and Energy Delay Product (EDP), and demonstrated that the proposed compute node (i.e., optical interconnected MSB) is significantly beneficial as a building block for large-scale, high-throughput, and energy-efficient HPC Data Center architectures. Indeed, our architectures can achieve more than 2× execution time improvement and up to 3× energy consumption reduction compared with a Multi-Socket, HT-based electronic baseline. We have implemented and simulated different optimization techniques and architectural designs to systematically examine different tradeoffs between execution time and energy efficiency. We have shown that the CDR technique allows for the best execution time, achieving on average, more than 2× improvement. Furthermore, we demonstrated that, through a proper architecture design (avoiding the intra-socket electronic hops as much as possible), it is possible to achieve an additional 15% performance improvement. From an energy consumption standpoint, the use of the voltage and frequency scaling optimization and the Source Synchronous transmission model makes possible to obtain up to a ~ 3× reduction in energy consumption. We based this comparison on state-of-the-art optoelectronics technologies involving high laser efficiency and photodetector sensitivity. On the other hand, if conservative and conventional optoelectronic solutions are considered, the results achieved by an on-chip optical communication model can be significantly different, and the challenges of exploiting an integrated optical communication compared to a state-of-the-art electronic solution arise. This suggests that (a) further technology improvements are crucially important in achieving the energy efficiency and goodput in the full-system architecture for future HPC systems, and that (b) the adoption of adaptive algorithms, intelligent transmission models, and the utilization of the proper architectural designs, are key considerations for exploiting the benefits offered by on-chip optical communications.

VI. References

[1] G. Huaxi, X. Jiang, and W. Zheng, "A novel optical mesh network-on-chip for gigascale systems-on-chip," in *Circuits and Systems, 2008. APCCAS 2008. IEEE Asia Pacific Conference on*, 2008.

[2] Y. Ye, J. Xu, X. Wu, W. Zhang, W. Liu, and M. Nikdast, "A Torus-Based Hierarchical Optical-Electronic Network-on-Chip for Multiprocessor System-on-Chip," *J. Emerg. Technol. Comput. Syst.,* 2012.

[3] J. Kim, J. Balfour, and W. J. Dally, "Flattened Butterfly Topology for On-Chip Networks," *Computer Architecture Letters,* 2007.

[4] G. Huaxi, X. Jiang, and Z. Wei, "A low-power fat tree-based optical Network-On-Chip for multiprocessor system-on-chip," in *Design, Automation & Test in Europe Conference & Exhibition.*, 2009.

[5] R. Walker and R. Dugan, "64b/66b low-overhead coding proposal for serial links," *IEEE,* vol. 802, pp. 11-13, 2000.

[6] C. Gray, D. Keezer, O. Liboiron-Ladouceur, and K. Bergman, "Multi-Gigahertz Source Synchronous Testing of an Optical Packet Switching Network," in *Mixed-Signals Test Workshop*, 2006.

[7] S. Borkar, "Exascale computing - A fact or a fiction?," in *Parallel & Distributed Processing (IPDPS)*, 2013.

[8] D. McMorrow and M. Corporation, *Technical Challenges of Exascale Computing*: MITRE Corporation, 2013.

[9] P. M. Kogge and T. J. Dysart, "Using the TOP500 to trace and project technology and architecture trends," in *International Conference for High Performance Computing, Networking, Storage and Analysis*, Seattle, Washington, 2011, pp. 1-11.

[10] Intel, "An introduction to the quickpath interconnect," *Online,* vol. http://www.intel.com/technology/quickpath/introduction.pdf, 2009.

[11] B. Holden, "Latency comparison between HyperTransport and PCI-Express in communications systems," *HyperTransport* 2006.

[12] A. Ros, B. Cuesta, M. E. Gomez, A. Robles, and J. Duato, "Cache miss characterization in hierarchical large-scale cache-coherent systems," in *Parallel and Distributed Processing with Applications (ISPA), 2012 IEEE 10th International Symposium on*, 2012.

[13] S. R. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, A. Singh, T. Jacob, S. Jain, V. Erraguntla, C. Roberts, Y. Hoskote, N. Borkar, and S. Borkar, "An 80-Tile Sub-100-W TeraFLOPS Processor in 65-nm CMOS," *Solid-State Circuits, IEEE Journal of,* vol. 43, 2008.

[14] A. Ros, M. E. Acacio, and J. M. Garcia, "A scalable organization for distributed directories," *J. Syst. Archit.,* vol. 56, pp. 77-87, 2010.

[15] C. Kim, D. Burger, and S. W. Keckler, "An adaptive, non-uniform cache structure for wire-delay dominated on-chip caches," *SIGARCH Comput. Archit. News,* vol. 30, pp. 211-222, 2002.

[16] M. Pavlovic, Y. Etsion, and A. Ramirez, "On the memory system requirements of future scientific applications: Four case-studies," in *Workload Characterization, International Symposium on*, 2011.

[17] B. M. Rogers, A. Krishna, G. B. Bell, K. Vu, X. Jiang, and Y. Solihin, "Scaling the bandwidth wall: challenges in and avenues for CMP scaling," *SIGARCH Comput. Archit. News,* vol. 37, 2009.

[18] P. Conway and B. Hughes, "The AMD Opteron Northbridge Architecture," *Micro, IEEE,* vol. 27, pp. 10-21, 2007.

[19] J. Chan, G. Hendry, K. Bergman, and L. P. Carloni, "Physical-Layer Modeling and System-Level Design of Chip-Scale Photonic Interconnection Networks," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 30, 2011.

[20] M. Petracca, B. G. Lee, K. Bergman, and L. P. Carloni, "Design Exploration of Optical Interconnection Networks for Chip Multiprocessors," in *High Performance Interconnects*, 2008.

[21] Z. Yunhui, M. Shenglin, S. Xin, F. Runiu, Z. Xiao, B. Yuan, C. Meng, C. Jing, M. Min, L. Wengao, and J. Yufeng, "Development and characterization of a through-multilayer TSV integrated SRAM module," in *Electronic Components and Technology Conference (ECTC), 2013 IEEE 63rd*, 2013.

[22] C. Ming-Hung, H. Wei-Chih, W. Pei-Chen, C. Ching-Te, C. Kuan-Neng, W. Chen-Chao, T. Chun-Yen, C. Kua-Hua, C. Chi-Tsung, T. Ho-Ming, and H. Wei, "Multi-layer adaptive power management architecture for TSV 3DIC applications," in *Electronic Components and Technology Conference (ECTC), IEEE*, 2013.

[23] S. Kamei, M. Ishii, M. Itoh, T. Shibata, Y. Inoue, and T. Kitagawa, "64 x 64-channel uniform-loss and cyclic-frequency arrayed-waveguide grating router module," *Electronics Letters,* 2003.

[24] 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," *Optics Express,* 2013.

[25] A. K. Kodi and A. Louri, "Power-Aware Bandwidth-Reconfigurable Optical Interconnects for High-Performance Computing (HPC) Systems," in *Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International*, 2007, pp. 1-10.

[26] P. P. Dash, G. Cowan, and O. Liboiron-Ladouceur, "A variable-bandwidth, power-scalable optical receiver front-end in 65 nm," in *Circuits and Systems (MWSCAS), 2013 IEEE 56th International Midwest Symposium on*, 2013, pp. 717-720.

[27] J. E. Proesel, B. G. Lee, A. V. Rylyakov, C. W. Baks, and C. L. Schow, "Ultra-low-power 10 to 28.5 Gb/s CMOS-driven VCSEL-based optical links [Invited]," *Optical Communications and Networking, IEEE/OSA Journal of,* vol. 4, pp. B114-B123, 2012.

[28] X. Chen, L.-S. Peh, G.-Y. Wei, Y.-K. Huang, and P. Prucnal, "Exploring the design space of power-aware opto-electronic networked systems," in *High-Performance Computer Architecture, 2005. HPCA-11. 11th International Symposium on*, 2005.

[29] S. J. B. Yoo, "Energy Efficiency in the Future Internet: The Role of Optical Packet Switching and Optical-Label Switching," *Selected Topics in Quantum Electronics, IEEE Journal of,* vol. 17, pp. 406-418, 2011.

[30] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The PARSEC benchmark suite: characterization and architectural implications," presented at the Proceedings of the 17th international conference on Parallel architectures and compilation techniques, Toronto, 2008.

[31] M. M. K. Martin, M. D. Hill, and D. J. Sorin, "Why on-chip cache coherence is here to stay," *Commun. ACM,* vol. 55, 2012.

[32] L. Ramini, P. Grani, H. T. Fankem, A. Ghiribaldi, S. Bartolini, and D. Bertozzi, "Assessing the energy break-even point between an optical NoC architecture and an aggressive electronic baseline," in *Design, Automation and Test in Europe Conference and Exhibition (DATE)*, 2014.

[33] R. Proietti, C. Zheng, C. J. Nitta, L. Yuliang, and S. J. B. Yoo, "A Scalable, Low-Latency, High-Throughput, Optical Interconnect Architecture Based on Arrayed Waveguide Grating Routers," *Lightwave Technology, Journal of,* vol. 33, pp. 911-920, 2015.

[34] B. Glance, I. P. Kaminow, and R. W. Wilson, "Applications of the integrated waveguide grating router," *Lightwave Technology, Journal of,* vol. 12, pp. 957-962, 1994.

[35] S. Kamei, M. Ishii, A. Kaneko, T. Shibata, and M. Itoh, "N x N Cyclic-Frequency Router With Improved Performance Based on Arrayed-Waveguide Grating," *Lightwave Technology, Journal of,* vol. 27, pp. 4097-4104, 2009.

[36] K. Sato, H. Hasegawa, T. Niwa, and T. Watanabe, "A large-scale wavelength routing optical switch for data center networks," *Communications Magazine, IEEE,* vol. 51, pp. 46-52, 2013.

[37] Z. Xuezhe, L. Shiyun, L. Ying, Y. Jin, L. Guoliang, S. S. Djordjevic, L. Jin-Hyoung, H. D. Thacker, I. Shubin, K. Raj, J. E. Cunningham, and A. V. Krishnamoorthy, "Efficient WDM Laser Sources Towards Terabyte/s Silicon Photonic Interconnects," *Lightwave Technology, Journal of,* 2013.

[38] P. Cheben, J. H. Schmid, A. Delâge, A. Densmore, S. Janz, B. Lamontagne, J. Lapointe, E. Post, P. Waldron, and D. X. Xu, "A high-resolution silicon-on-insulator arrayed waveguide grating microspectrometer with sub-micrometer aperture waveguides," *Optics express,* vol. 15, 2007/03/05 2007.

[39] 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, 2011.

[40] N. L. Binkert, R. G. Dreslinski, L. R. Hsu, K. T. Lim, A. G. Saidi, and S. K. Reinhardt, "The M5 Simulator: Modeling Networked Systems," *Micro, IEEE,* vol. 26, pp. 52-60, 2006.

[41] N. Barrow-Williams, C. Fensch, and S. Moore, "A communication characterisation of Splash-2 and Parsec," in *Workload Characterization, IISWC. IEEE International Symposium on*, 2009.

[42] C. Zheng, R. Proietti, and S. J. B. Yoo, "Scalable and high performance HPC architecture with optical interconnects," in *Photonics Conference (IPC), 2014 IEEE*, 2014, pp. 180-181.

[43] R. Jinsoo, C. Kwang-Chun, and C. Woo-Young, "A 10-Gb/s power and area efficient clock and data recovery circuit in 65-nm CMOS technology," in *SoC Design Conference (ISOCC)*, 2012.

[44] P. Conway, N. Kalyanasundharam, G. Donley, K. Lepak, and B. Hughes, "Cache Hierarchy and Memory Subsystem of the AMD Opteron Processor," *Micro, IEEE,* vol. 30, pp. 16-29, 2010.

[45] N. Sambo, A. D'Errico, C. Porzi, V. Vercesi, M. Imran, F. Cugini, A. Bogoni, Poti, x, L., and P. Castoldi, "Sliceable transponder architecture including multiwavelength source," *Optical Communications and Networking, IEEE/OSA Journal of,* 2014.

[46] M. Badr and N. E. Jerger, "SynFull: synthetic traffic models capturing cache coherent behaviour," *SIGARCH Comput. Archit. News,* vol. 42, pp. 109-120, 2014.

[47] R. Proietti, C. J. Nitta, Z. Cao, M. Clements, G. Tzimpragos, and S. J. B. Yoo, "Flexible-Bandwidth Power-Aware Optical Interconnects with Source Synchronous Technique," presented at the IEEE Optical Interconnects, 2015.

[48] D. N. Truong, W. H. Cheng, T. Mohsenin, Y. Zhiyi, A. T. Jacobson, G. Landge, M. J. Meeuwsen, C. Watnik, A. T. Tran, X. Zhibin, E. W. Work, J. W. Webb, P. V. Mejia, and B. M. Baas, "A 167-Processor Computational Platform in 65 nm CMOS," *Solid-State Circuits, IEEE Journal of,* vol. 44, 2009.

[49] R. Morris, A. K. Kodi, and A. Louri, "Dynamic Reconfiguration of 3D Photonic Networks-on-Chip for Maximizing Performance and Improving Fault Tolerance," in *Microarchitecture (MICRO), 2012 45th Annual IEEE/ACM International Symposium on*, 2012.

[50] P. Koka, M. O. McCracken, H. Schwetman, C. H. O. Chen, Z. Xuezhe, R. Ho, K. Raj, and A. V. Krishnamoorthy, "A micro-architectural analysis of switched photonic multi-chip interconnects," in *Computer Architecture (ISCA), International Symposium on*, 2012.

[51] 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, pp. 1-30, 2014.

[52] A. J. Zilkie, B. J. Bijlani, P. Seddighian, D. C. Lee, W. Qian, J. Fong, R. Shafiiha, D. Feng, B. J. Luff, X. Zheng, J. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari, "High-efficiency hybrid III-V/Si external cavity DBR laser for 3-μm SOI waveguides," in *Group IV Photonics (GFP), IEEE 9th International Conference on*, 2012.

[53] G. Masini, A. Narasimha, A. Mekis, B. Welch, C. Ogden, C. Bradbury, C. Sohn, D. Song, D. Martinez, D. Foltz, D. Guckenberger, J. Eicher, J. Dong, J. Schramm, J. White, J. Redman, K. Yokoyama, M. Harrison, M. Peterson, M. Saberi, M. Mack, M. Sharp, P. de Dobbelaere, R. LeBlanc, S. Leap, S. Abdalla, S. Gloeckner, S. Hovey, S. Jackson, S. Sahni, S. Yu, T. Pinguet, W. Xu, and Y. Liang, "CMOS photonics for optical engines and interconnects," in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC), 2012 and the National Fiber Optic Engineers Conference*, 2012, pp. 1-3.

[54] X. Zheng, D. Patil, J. Lexau, F. Liu, G. Li, H. Thacker, Y. Luo, I. Shubin, J. Li, J. Yao, P. Dong, D. Feng, M. Asghari, T. Pinguet, A. Mekis, P. Amberg, M. Dayringer, J. Gainsley, H. F. Moghadam, E. Alon, K. Raj, R. Ho, J. E. Cunningham, and A. V. Krishnamoorthy, "Ultra-efficient 10Gb/s hybrid integrated silicon photonic transmitter and receiver," *Optics Express,* vol. 19, 2011.

[55] R. Jinsoo, C. Kwang-Chun, and C. Woo-Young, "A 10-Gb/s power and area efficient clock and data recovery circuit in 65-nm CMOS technology," in *SoC Design Conference (ISOCC), 2012 International*, 2012, pp. 104-107.

[56] P. Dong, W. Qian, H. Liang, R. Shafiiha, D. Feng, G. Li, J. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari, "Thermally tunable silicon racetrack resonators with ultralow tuning power," *Optics Express,* vol. 18, 2010/09/13 2010.

[57] M. Ortín-Obón, L. Ramini, V. Viñals, and D. Bertozzi, "Capturing the sensitivity of optical network quality metrics to its network interface parameters," *Concurrency and Computation: Practice and Experience,* vol. 26, pp. 2504-2517, 2014.

[58] S. Chen, C. H. O. Chen, G. Kurian, W. Lan, J. Miller, A. Agarwal, P. Li-Shiuan, and V. Stojanovic, "DSENT - A Tool Connecting Emerging Photonics with Electronics for Opto-Electronic Networks-on-Chip Modeling," in *Networks on Chip, IEEE/ACM International Symposium on*, 2012.

[59] S. Beamer, K. Asanovic, C. Batten, A. Joshi, and V. Stojanovic, "Designing multi-socket systems using silicon photonics," presented at the Proceedings of the 23rd international conference on Supercomputing, Yorktown Heights, NY, USA, 2009.

Paolo Grani received the M.S. degree in Computer Engineering from the University of Pisa, Pisa, Italy, in 2009, and the Ph.D. degree in Information Engineering and Science from University of Siena, Siena, Italy, in 2015. He is currently a Postdoctoral Researcher with the University of California, Davis, CA, USA. His research interests include low-latency and scalable optical interconnects for data centers and high performance computing, and on-chip optical integration from the architectural point of view.

Roberto Proietti received the M.S. degree in Telecommunications Engineering from the University of Pisa, Pisa, Italy, in 2004, and the Ph.D. degree in optical communication systems and networking from Scuola Superiore Sant'Anna, Pisa, in 2009. He is currently a Project Scientist with the University of California, Davis, CA, USA. His research interests include optical switching technologies and architectures for supercomputing and data center, high-spectrum efficiency coherent transmission systems, elastic optical networking, and radio over fiber systems.

Stanley Cheung received the B.S. degree in Electrical Engineering from the University of Southern California, Los Angeles, CA, USA, the M.S. degree from Columbia University, New York, NY, USA and the Ph.D. degree in Electrical Engineering at the University of California, Davis, CA, USA. He is currently a research scientist at Hewlett Packard Labs in Palo Alto, CA. His research interests include InP/InGaAsP semiconductor mode-locked lasers, silicon photonic integrated circuits, and hybrid Si/III-V active devices.

S. J. Ben Yoo (S'82–M'84–SM'97–F'07) received the B.S. degree in Electrical Engineering with distinction, the M.S. degree in Electrical Engineering, and the Ph.D. degree in Electrical Engineering with a minor in Physics, all from Stanford University, Stanford, CA, USA, in 1984, 1986, and 1991, respectively. He currently serves as a Professor of electrical engineering at the University of California, Davis, CA, USA. His research interests include high performance all-optical devices, systems, and networking technologies for future computing and communications. Prior to joining UC Davis in 1999, he was a Senior Research Scientist at Bellcore, leading technical efforts in optical networking research and systems integration. He participated in ATD/MONET testbed integration and some standardization activities including GR-2918-CORE, GR-2918-ILR, GR-1377-CORE, and GR-1377-ILR on dense WDM and OC-192 systems. He is a Fellow of the Optical Society of America, and is a recipient of the DARPA Award for Sustained Excellence (1997), the Bellcore CEO Award (1998), the Outstanding Mid-Career Research Award (UC Davis, 2004), and the Outstanding Senior Research Award (UC Davis, 2011).