Received: 9 March 2020 | Revised: 26 June 2020 | Accepted: 27 August 2020

IET Optoelectronics

DOI: 10.1049/ote2.12018

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Silicon circuits for chip-to-chip communications in multi-socket server board interconnects

Miltiadis Moralis-Pegios[1] | Stelios Pitris[1] | Charoula Mitsolidou[1] | Konstantinos Fotiadis[1] | Hannes Ramon[2] | Joris Lambrecht[2] | Johan Bauwelinck[2] | Xin Yin[2] | Yoojin Ban[3] | Peter De Heyn[3] | Joris Van Campenhout[3] | Tobias Lamprecht[4] | Andreas Lehnman[5] | Nikos Pleros[1] | Theoni Alexoudi[1]

[1]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

[2]IDLab, Department of Information Technology, Ghent University IMEC, Ghent, Belgium

[3]IMEC, Leuven, Belgium

[4]Vario-optics AG, Heiden, Switzerland

[5]Amphenol FCI, Berlin, Germany

Correspondence

Miltiadis Moralis-Pegios, Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece.
Email: mmoralis@csd.auth.gr

Funding information

Horizon 2020 Framework Programme (ICT), Grant/Award Number: STREAMS 688172

Abstract

Multi-socket server boards (MSBs) exploit the interconnection of multiple processor chips towards forming powerful cache coherent systems, with the interconnect technology comprising a key element in boosting processing performance. Here, we present an overview of the current electrical interconnects for MSBs, outlining the main challenges currently faced. We propose the use of silicon photonics (SiPho) towards advancing interconnect throughput, socket connectivity and energy efficiency in MSB layouts, enabling a flat-topology wavelength division multiplexing (WDM)-based point-to-point (p2p) optical MSB interconnect scheme. We demonstrate WDM SiPho transceivers (TxRxs) co-assembled with their electronic circuits for up to 50 Gb/s line rate and 400 Gb/s aggregate data transmission and SiPho arrayed waveguide grating routers that can offer collision-less time of flight connectivity for up to 16 nodes. The capacity can scale to 2.8 Gb/s for an eight-socket MSB, when line rate scales to 50 Gb/s, yielding up to 69% energy reduction compared with the QuickPath Interconnect and highlighting the feasibility of single-hop p2p interconnects in MSB systems with >4 sockets.

1 | INTRODUCTION

The explosive growth of intra-data centre (DC) traffic, along with the ongoing transition to new architectural paradigms such as resource disaggregation [1,2] has created the need for higher interconnect capacity and computation density within a reasonable energy and cost envelope [3]. With the number of high-performance cores integrated within the same processor die already facing practical real estate limitations [4], efforts are now focused on multi-socket server board (MSB) schemes that exploit the physical proximity of interconnected sockets to increase computational power, while maintaining low latency and energy consumption. Current MSBs are classified in two categories: 'glueless' architectures formed by point-to-point (p2p) interconnected sockets, where socket connectivity is limited to four- or eight-socket topologies [5], and 'glued' architectures, where scaling beyond eight sockets is achieved via the use of active switches [6,7] at the expense of increased energy consumption, latency and complexity.

Increasing MSB computational performance should combine a high number of interconnected sockets in a single-hop glueless architecture to retain low latency while avoiding the energy burden of switches. To achieve that, single-hop links have been suggested via the replacement of electrical with optical interconnects, utilizing arrayed waveguide grating router (AWGR)-based schemes [8,9]. This solution has already been investigated at 10 Gb/s line rates via cycle-accurate simulations [9]. Its experimental evidence has been demonstrated in the C-band, performing, however, at a rather low rate of 0.3 Gb/s and incorporating both the transceiver (TxRx) and the AWGR on the same chip [8].

Increasing the number of the socket interfaces however necessitates a disintegrated implementation approach, where TxRxs and AWGR will reside in separate chips, optically communicating over an electro-optic printed circuit board (EOPCB) technology, relaxing in this way the real estate limitations. With EOPCBs typically offering a low waveguide loss figure at the O-band [10] and rather high propagation losses at the C-band, the experimental AWGR-based demonstrations reported so far, almost exclusively in the C-band regime, are not compatible with MSB topologies with >4 sockets. Under these circumstances, we have recently demonstrated the main subsystems that allow for the realization of an O-band MSB optical interconnect [11–17], including an 8×50 Gb/s O-band silicon photonics (SiPho) TxRx [14], a 16×16 O-band SiPho AWGR [16] and an automated thermal drift compensation system [17].

Here, we demonstrate advances in O-band WDM SiPho TxRx and AWGR interconnect circuitry, configured in a novel crosstalk (Xtalk)-aware flat topology for scaling to >4 socket optical glueless MSBs. Following an overview of the current electrical MSB interconnects and their main challenges, we present an optical MSB interconnect scheme that can yield all-to-all single-hop interconnections. We demonstrate a full-scale eight-socket MSB architecture along with a brief review of the progress in each constituent building block. An energy efficiency (EE) analysis reveals significant savings compared with the current most popular electrical interconnect, which can reach up to 38% and 69% when performing at 25 Gb/s and 50 Gb/s line rates, respectively.

Section 2 presents an overview of the current state-of-the-art MSB electrical interconnects. In Section 3, we introduce the benefits of optical MSB interconnects. Section 4 provides a brief overview of the progress made in the constituent SiPho building blocks for optical MSB architectures. Finally, Section 5 presents an energy and optical power budgets analysis of the proposed scheme and provides a performance versus the QuickPath Interconnect (QPI)-based electrical baseline.

2 | MULTI-SOCKET ELECTRICAL INTERCONNECTS

Current MSB electrical interconnects follow two deployment directions: glueless topologies, where processor sockets communicate over p2p interconnects, and glued topologies, where the number of interconnected sockets can increase by employing an active switch to connect different glueless socket islands. Figure 1a shows a four-socket glueless interconnect topology, where each node can exchange traffic with all other nodes using direct links. The glueless topology can scale out to an eight-socket interconnection, as shown in Figure 1b, where each socket connects directly via a p2p direct link to its three neighboring sockets but employs an indirect dual-hop link to connect with the other four sockets. Going beyond the eight-socket interconnection can be only realized via glued set-ups that employ active switches between clusters of glueless sockets [6], as shown in Figure 1c for the case of a 16-socket glued configuration. The main interconnect-related performance parameters in both glueless and glued MSB configurations, together with the main challenges, are summarized next.

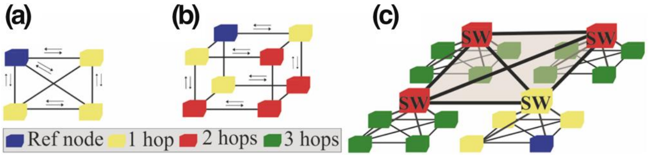

FIGURE 1 (a) Four-socket glueless architecture, (b) 8-socket glueless architecture and (c) 16-socket glued architecture with switches

2.1 | Connectivity

Computational power increases with the number of interconnected processors in MSBs, turning connectivity into a critical performance parameter. Glueless interconnects allow for a limited connectivity up to eight sockets, with four-socket set-ups being the typical case and relying exclusively on p2p links, while its scaled-out eight-socket version incorporates also dual-hop paths. Scaling beyond eight sockets can be accomplished only via glued schemes, as current state-of-the-art processors do not support a higher number of interconnected central processing units (CPUs) without an active switch circuitry [5], at the expense however of high energy consumption, latency and cost, with commercial prototypes reporting up to 16 connectivity [7].

2.2 | Bandwidth

The turning of modern server boards into multi-socket environments transformed interconnect bandwidth into a key factor for increasing processing power in DCs. Intel's QPI bidirectional link offers a peak bandwidth of 19.2 GB/s and 25.6 GB/s when utilizing a 4.8-giga transfer (GT)/s and a 6.4-GT/s bit rate, respectively, over its 16-bit data bus [3]. It should be noted, that GT refers to the number of data transfer operations per second per lane and that the total QPI bandwidth calculation takes into account the bidirectionality of the QPI. Recently, QPI was replaced by the UltraPath Interconnect (UPI) to allow for higher speeds of up to 10.4 GT/s and a peak bandwidth of up to 41.6 GB/s per bidirectional link. Considering the glued interconnects, NVIDIA's NVLink technology allows for the interconnection of up to 16 sockets in a single server with a line rate of 25 Gb/s and a six-lane link bidirectional peak bandwidth of up to 300 Gb/s.

2.3 | Cache coherency update bandwidth

The need for cache coherency in MSBs yields communication patterns with strong multi- and broadcast characteristics that tend to consume a large portion of the available interconnect bandwidth [6]. In glueless MSBs, cache coherency messages often account for >30% of the total bandwidth, while eight-socket implementations may even require up to 65% of the bandwidth [6], wasting in this way a significant portion of the available resources.

2.4 | Energy efficiency

Current MSBs rely to a large degree on data movement, with electrical interconnects forming the main energy consuming factor for chip-to-chip communication. EE of the glueless QPI interconnection between two sockets is equal to 16.2 pJ/bit [18], only slightly improved in Intel's recent UPI version. Glued configurations like NVIDIA's NVlink v2 consume up to 60 pJ/bit for chip-to-chip interconnect purposes, clearly outlining the overhead associated with the use of active switches [7].

2.5 | Latency

With computational speed relying on the communication between the sockets, interconnect latency turns into a critical operational parameter for the ultimate execution speed. The link latency associated with CPU-to-memory communication of a four-socket QPI/UPI implementation can reach up to 140 ns [5], becoming even higher when scaling to eight-socket systems where dual-hop links are present. In >8-socket glued set-ups, the use of the active switch for allowing a higher number of interconnected sockets comes at the expense of a higher latency value that can even reach 1 μs, when using PCIe lanes for socket interconnection [19,20].

3 | A FLAT TOPOLOGY OPTICAL INTERCONNECT CONCEPT

Optical AWGR-based architectures have recently been shown as a promising route towards high-bandwidth and low-latency direct p2p interconnection in MSB schemes [8,9]. These architectures take advantage of the cyclic wavelength-routed characteristics of the AWGR, requiring the employment of optical WDM TxRx circuits for associating every possible socket communication link with a distinct wavelength. Although this architecture has already revealed system-scale benefits compared with the electrical interconnected MSBs [9], its experimental deployment is still confronting the following challenges:

- It has been demonstrated on chip photonic TxRx circuits [8] with rather low line rates, not exceeding 0.3 Gb/s.
- It has been oriented towards the C-band [8,9], restricting its perspectives to a higher number of sockets since EOPCBs favour the O-band for lower loss connectivity.
- The high in-band Xtalk of integrated AWGRs [8,9,15] hinders the experimental demonstration of a fully loaded all-to-all interconnect.

Figure 2 presents the optical interconnect topology for p2p communication between a number of N sockets. The proposed topology consists of the following building blocks: a) the N WDM optical Txs, with each Tx containing $N-1$ channels at different wavelengths that are subsequently multiplexed through a ($N-1$):1 WDM multiplexer (MUX) to exit via a single optical interface, b) the N WDM Rxs, with every Rx comprising a 1:($N-1$) WDM demultiplexer (DEMUX) with its $N-1$ ports connecting to respective photoreceivers and c) the mid-board passive routing platform implemented by a cyclic $N\times N$ AWGR. The Tx and Rx engines connect each socket with the server board to enable connectivity with the remaining N-1 sockets, while the AWGR allows for all-to-all connectivity among sockets in a collision-less, buffer-less and low-latency way. The following two inter-socket communications are supported via this optical interconnect.

Simultaneous any-to-any operation, where each socket sends different data to any of the other sockets. As shown in Figure 2, every wavelength channel within the WDM Tx of a socket corresponds to a unique communication path, establishing communication with a different destination socket whenever this channel gets activated. For example, if Socket #1 wants to communicate with Socket #N, then the electrical data emerging at Socket #1 electrical interface modulate the optical channel at λ_1, which is then multiplexed with the other possible WDM channels and launched at the first input port of the AWGR for routing to the destination Socket #N. The received optical signal is then demultiplexed (DEMUX) and sent to the photoreceiver of Socket #N, which comprises a photodiode (PD) and a transimpedance amplifier (TIA). The remaining channels λ_2–λ_{N-1} within the Socket#1 Tx will be modulated with electrical data when a different destination socket is targeted. At the same time, if Socket #3 wants to send data to Socket #N, it simply has to encode the respective electrical data onto a different wavelength channel that gets routed to the AWGR output port connected with Socket #N. This means that data can be transmitted simultaneously from multiple sockets to the same destination, since the respective channels reaching the same AWGR output will be encoded onto different wavelengths. Thus, the introduction of wavelength-selective routing into on-board MSBs can effectively reduce the energy consumption and cabling requirements, while at the same time increase the bandwidth through WDM parallelism. Connecting N sockets necessitates the use of an $N\times N$ AWGR and $N-1$ wavelengths, yielding a total number of $N\times(N-1)$ p2p links. The p2p collision-less and buffer-less interconnect scheme allows for single-hop communication between all sockets, significantly reducing the interconnect latency towards approaching the limit of light's time-of-flight value.

Broadcast/multicast operation, where each socket can send the same data to all other sockets (broadcasting) or a subset of sockets (multicasting). For example, if Socket #N wants to broadcast data to the rest of $N-1$ sockets as depicted in Figure 2, the same electrical data are modulating all $N-1$ wavelengths within the Socket #N Tx, which are then directed to the first input port of the AWGR, so that every wavelength gets routed to a different AWGR output port and as such a different destination socket. In case that Socket #N wants to multicast its data to Sockets #1,2,3, electrical data carrying common information will modulate the proper three different wavelengths that subsequently enter the AWGR input port and get routed at the proper

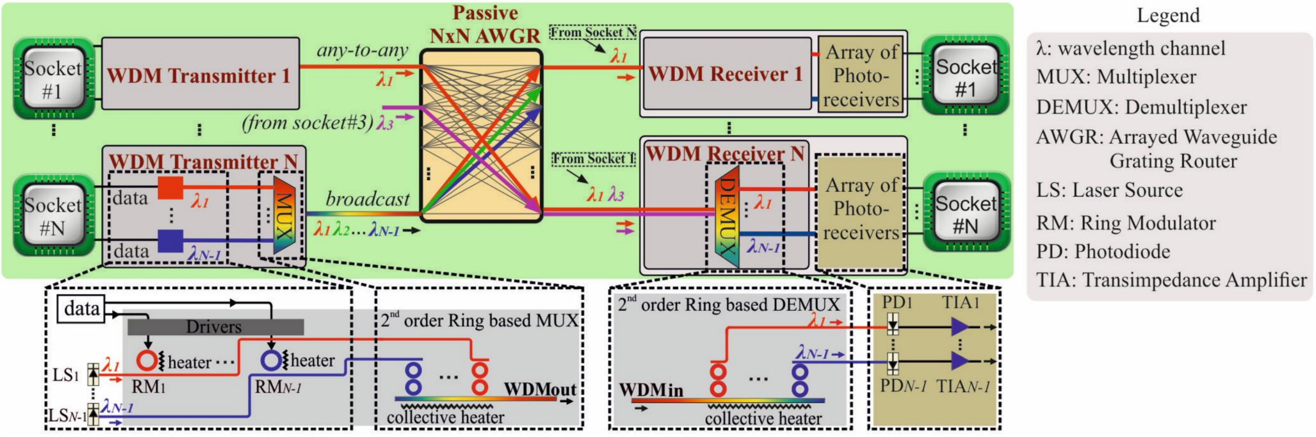

FIGURE 2 AWGR-based optical interconnection architecture employing ring-based WDM transceivers

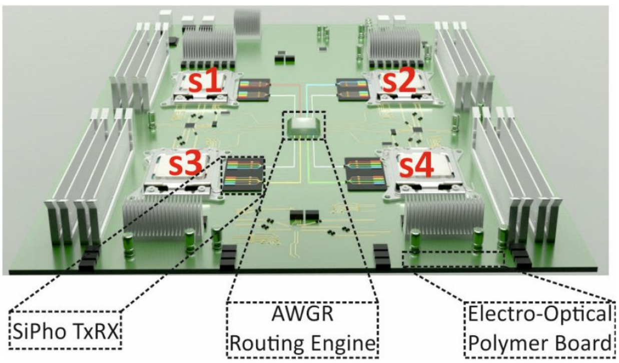

FIGURE 3 Artistic perspective of the envisaged MSB optical interconnect platform

output ports. The broadcast/multicast properties suggested by this scheme can be highly useful in MSBs as they can handle the corresponding communication requirements arising during cache coherency updates among the sockets, where the cache memory of a socket has to broadcast its state to the last-level caches of all other sockets to retain synchronized cache content along the entire setting [6].

4 | SIPHO CIRCUITS FOR CHIP-TO-CHIP INTERCONNECTS

SiPho can offer a rich portfolio of advantages in optical MSB interconnects, allowing for high-line rate, small-footprint, low-energy consumption and O-band operation. Their CMOS-compatible and high-volume manufacturing credentials, together with their increased level of maturity, has already led to commercially available 4×25 Gb/s TxRxs for optical interconnects, offering a solid ground for turning SiPho circuits into a flat-topology interconnect technology even for on-board chip-to-chip communication. Figure 3 illustrates an artistic perspective of the optical MSB architecture when deployed onto an EOPCB for the case of four interconnected sockets. Four different sockets, denoted as s1, s2, s3 and s4, are attached to respective mid-board WDM SiPho TxRx interfaces that subsequently convert the electrical data emerging from the CPU socket into optical data and launch them into optical polymer waveguides embedded into the EOPCB board. All-to-all communication is ensured via a mid-board SiPho AWGR passive router chip plugged almost at the centre of the EOPCB and connecting to all four different sockets, following the architectural principles analysed in Figure 2. In this section, we present a brief review on the evolution and recent advances in the field of O-band WDM SiPho TxR and AWGRs for optical multi-socket board interconnects, together with the progress on the respective EOPCBs, [10] bound to serve as the hosting platform for the interconnect system.

4.1 | SiPho WDM mid-board transceivers

Each socket is equipped with a SiPho WDM TxRx, responsible for optically interfacing the CPU with the server board, by employing multiple RM modulators resonating at different wavelengths while powered from WMD continuous wave (CW) lasers. Figures 4 and 5 depict a time and performance evolution of SiPho Tx and TxRx demonstrations for MSBs, which were developed within the European research project H2020-ICT-STREAMS, progressing from single-lane 50 Gb/s towards WDM layouts that can reach an aggregate bandwidth of 400 Gb/s [11–14]. More specifically,

- Figure 4a illustrates an O-band SiPho RM transmitter co--packaged with a low power consumption (PC) CMOS 1V driver [11]. This first demonstration validated the low-power and high-speed credentials of RM-based Tx, achieving 50 Gb/s line rates and a low PC of 0.8 pJ/bit, excluding the heater and laser power requirements, as this first prototype lacked an integrated thermal heater. Indicative results of the performance of the O-band Tx are depicted in Figure 4b, including an optical eye diagram at the Tx egress at 50 Gb/s, along with bit error rate (BER) measurements performed after 52 km of SSMF fibre, validating the capabilities of the Tx even for inter-DC interconnections.

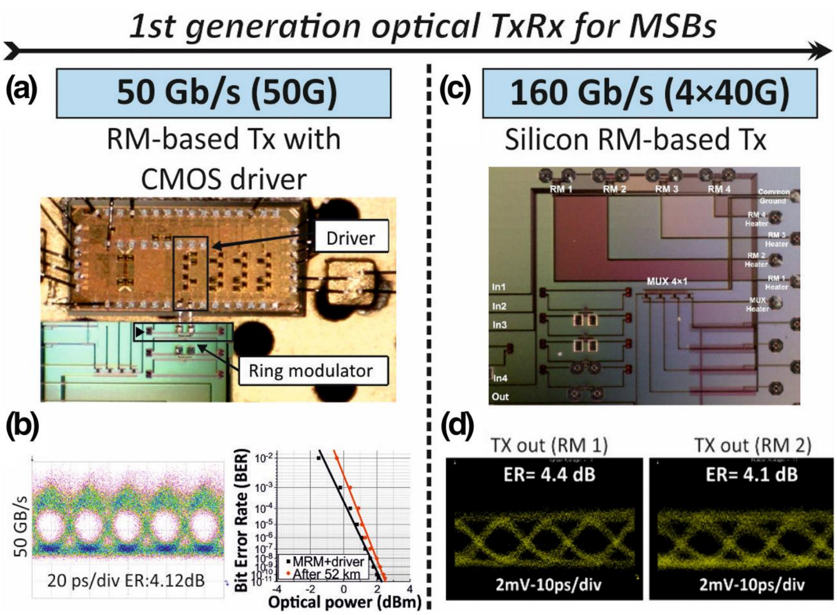

FIGURE 4 Evolution and indicative results of SiPho RM-based Tx/TxRx for optical multi-socket server boards: (a, b) 50 Gb/s Tx assembly (c), (d) 160 (4×40) Gb/s Tx

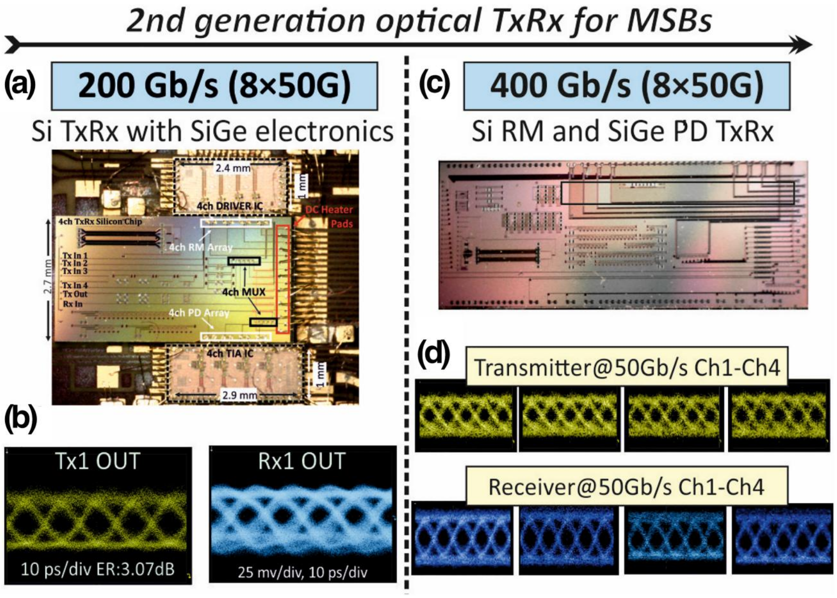

FIGURE 5 Evolution and indicative results of SiPho RM-based Tx/TxRx for optical multi-socket server boards: (a,b) 200 (4×50) Gb/s TxRx paired with SiGe electronics (c, d) 400 (8×50) Gb/s TxRx

- Figure 4c illustrates a four-channel SiPho Tx [12] that comprises a four-channel RM modulator array interconnected by a cascaded double-ring resonator-based multiplexer (MUX), with a channel spacing of 2.25 nm. An aggregate Tx bandwidth of 160 Gb/s was achieved with each RM operating at 40 Gb/s, with the demonstrator featuring a high power efficiency of ~1 pJ/bit, excluding the laser. Figure 4d depicts the performance of two of the channels of the WDM Tx, revealing clearly open-eye diagrams and an ER of 4.1 dB for a driving voltage of approximately 2 Vpp.
- Figure 5a depicts a 200G-capable SiPho WDM O-band optical TxRx [13] comprising four-element arrays of RMs and Ge PDs co-packaged with a SiGe BiCMOS integrated driver (DR) and a SiGe TIA chip. By exploiting the synergy between the recent advantages in SiGe electronic technology with state-of-the-art SiPho components, this demonstrator achieved a record breaking, among RM-based TxRx, aggregate bandwidth of 200 Gb/s while featuring a total PC of 4.2 pJ/bit, including the heater and excluding the laser power requirements. Figure 5b illustrates some indicative results of the 200G TxRx, including the Tx performance of one Tx and one Rx channel. It should be noted that the Rx achieved an interpolated BER value of 10E-12 referenced to a low-input optical modulation amplitude of −9.5 dBm at 50 Gb/s.
- Finally, Figure 5c illustrates a scaled-up version of the previous WDM TxRx [14], which encompasses an eight-

element array of both RM modulators and SiGe PDs, both interconnected by double-ring based MUX, to achieve an impressive aggregate bandwidth of 400 Gb/s, when operating at 50 Gb/s line rates. The 8×1 and 1×8 MUX and DEMUX, respectively, featured a channel spacing of 1.17 nm and an insertion loss of 0.68-2.47 dB, while the TxRx achieved a high EE of ~2 pJ/bit, including the heater and excluding the laser power requirements. Figure 5d depicts some indicative results of the TxRx performance for channels 1 to 4, with the Tx achieving am average ER of 4.5 dB for a driving voltage of 2.1 Vpp and the Rx featuring a Q factor ~5, which is theoretically mapped to a BER value of 10E−7, referenced to an average input optical power of −3 dBm.

4.2 | SiPho AWGR mid-board passive router

The core of the optical MSB architecture comprises an AWGR passive router that provides communication between multiple sockets in a buffer-less and collision-less way. In this context, by harnessing the ultra-dense and low-cost credentials of the SiPho platform, two Si AWGR O-band prototypes were developed: an 8×8 [15] designed for coarse wavelength division multiplexing (CWDM) and a scaled-up version in 16×16 I/O port configuration [16] for dense wavelength division multiplexing (DWDM) operation.

- Figure 6a illustrates a microscope photo of the fabricated 8×8 CWDM AWGR [15] that features a 10 nm channel spacing, a 5.7-nm 3-dB -bandwidth, insertion losses of 4.2 dB and a worst-case crosstalk of 15.4 dB, with the fabricated device occupying a footprint of 0.27×0.7 mm^2. The optical spectra for all eight inputs and respective outputs are illustrated in Figure 6b, revealing the successful cyclic routing properties of the device.
- Figure 6c depicts a microscope photo of the fabricated 16×16 DWMD AWGR [16] that features a channel spacing of 1.063 nm, a free spectral range of 17.8 nm, and a 3 dB bandwidth of 0.655 nm. The fabricated device occupied a small footprint of 0.27×0.71 mm^2, while the insertion losses ranged from 3.9 dB to 8.37 dB and the optical crosstalk had a mean value of 21.65 dB.

Although the AWGRs had successfully confirmed their cyclic wavelength routing properties [15], a fully loaded routing scheme allowing for the simultaneous all-to-all communication has not been demonstrated due to the constraints arising from the in-band Xtalk effects of the AWGR structures [22]. The in-band Xtalk can significantly impair system operation when the same wavelength λref is used simultaneously at several AWGR input ports, since interference noise originating from the other routing paths causes severe performance degradation on the data signal carried by λref and exiting through the desired output. According to [22], an in-band Xtalk value of -34 dB is needed for an 8×8 AWGR structure; however, the reported AWGRs exhibit significantly higher XTalk. To overcome this limitation, we have developed a XTalk-aware routing scheme towards enabling fully loaded AWGR-based interconnects even with AWGRs that do not meet the necessary Xtalk value [22], as has been typically the case for the integrated AWGRs reported so far [8,9,15,21]. The proposed XTalk-aware scheme [23] exploits a number of slightly detuned wavelengths around the nominal AWGR channel central wavelength for each spectral band, considering, however, the maximum possible wavelength utilization factor (i.e. the number of different AWGR inputs that can use the same wavelength and still obtain error-free operation at the output).

4.3 | Single-mode EOPCB

A single-mode EOPCB acts as the hosting platform for the sockets, the TxRx interface circuitry and the AWGR offering optical polymer waveguides for interconnecting the TxRx to the AWGR ports. Figure 7a depicts a microscope photo of the fabricated EOPCB for single-mode O-band operation that comprises both optical and electrical waveguides. The EOPCB allows the SiPho TxRx chip to be coupled to the polymer board via Si-to-polymer adiabatic couplers. The EOPCB comprises an array of polymer waveguides with a rectangular cross section slightly smaller than 6×6 μm^2 and propagation losses between 0.35 and 0.5 dB/cm, while the adiabatic coupling scheme between the polymer and the SiPho waveguides has been shown to exhibit an average value of optical losses equal to 0.5 dB [24]. This configuration provides a wide optical bandwidth while relaxing the lateral alignment tolerances, forming in this way a promising route towards low-loss and wavelength-insensitive interfacing SiPho TxRx and AWGR-based routing chips. Moreover, the EOPCB is equipped with copper traces with nearly vertical sidewalls and well-controlled 40 μm gap width, yielding low impedance variations and allowing features size down to 40 μm, with a cutback section of a four-layer EOPCB being depicted in Figure 7b. The high-frequency operation of the RF traces has been already confirmed for electrical transmission of 112 Gb/s PAM-4 signals through 4-cm long lines [10].

5 | MAIN OPERATION CHARACTERISTIC OF THE AWGR-BASED INTERCONNECT

Table 1 presents the optical loss break down and the PC of the active components deployed in the transmission link of the envisioned optical 8×8 AWGR-based interconnect, considering a case of an 8×8 interconnection Table 1 provides also a comparison of the proposed 8×8 AWGR-based interconnection its performance versus QPI [5] for line-rate operation at 25 Gb/s and 50 Gb/s, respectively. The EE of the optical interconnect has been calculated by summing the PC of all link components presented in Table 1 and dividing by the line rate. First, to estimate the PC of the laser source (LS), we have calculated the average optical power level, required at their output to achieve error-free operation for the all-to-all

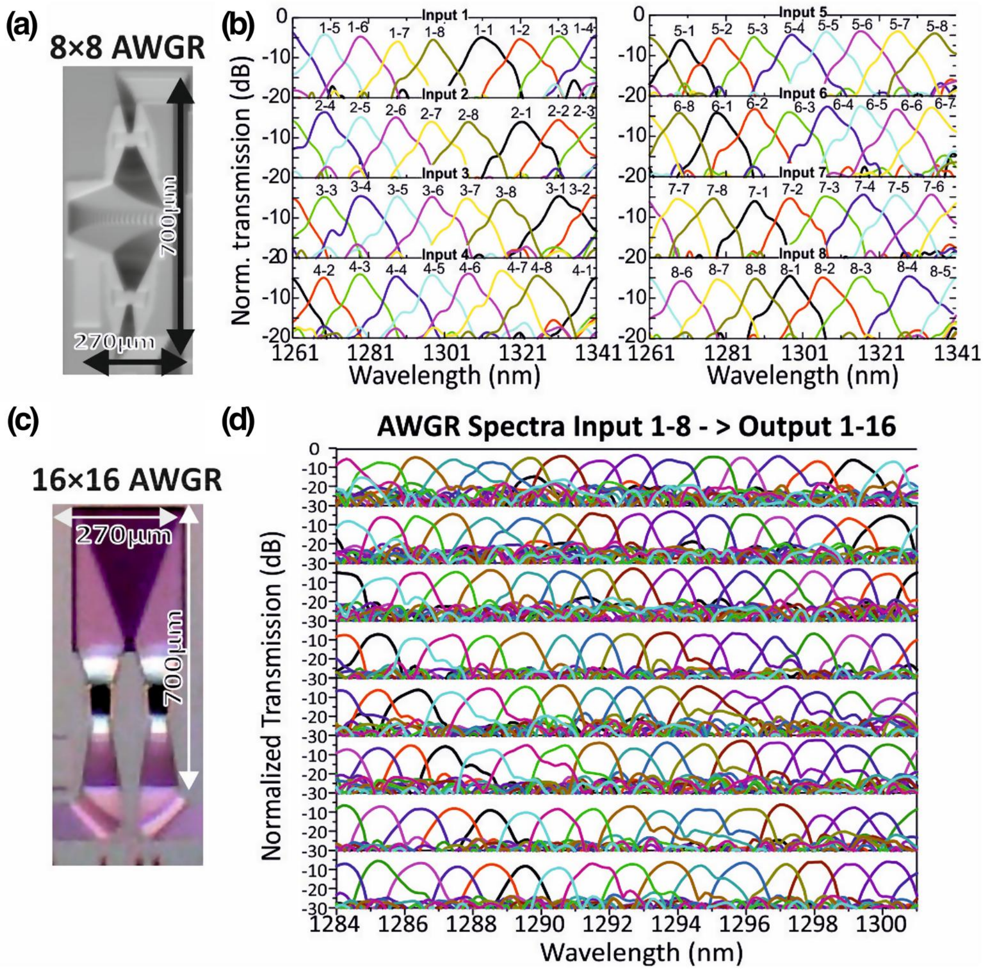

FIGURE 6 SiPho AWGR mid-board passive router fabricated chips microscope photos and indicative results (a, b) 8×8 CWDM implementation (c, d) 16×16 DWDM implementation

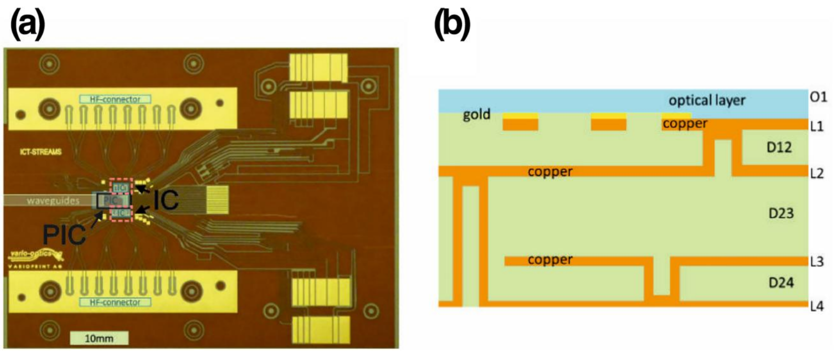

FIGURE 7 (a) Photograph of an EOCB prototype board (68 mm × 52 mm) with indication of PIC, IC, HF-connector, and polymer waveguides (b) illustration of layer build-up for four-layer boards (L1–L4) and two-layer boards (L1–L2). High frequency (HF); integrated circuit (IC); photonic intergrated circuit (PIC)

communication. As shown in Table 1, this value has been found to be equal to 4.5 dBm by taking into account the total link losses and the average receiver sensitivity needed in the input of the PDs to achieve a BER value of 10E−9. The receiver offered a sensitivity of −10 dBm at up to 50 Gb/s, referenced to a BER of 10E−12 [25], while the total link losses were calculated to be equal to 14.5 dB by adding the loss values for all constituent circuitry shown in Table 1. Assuming a 10% wall-plug efficiency, the 4.5 dBm of average optical power required at the output of the laser sources translates to an optical PC of 28.2 mW. The PC of each TxRx's electronics was found to be equal to 223 mW by adding the PC of the RM's Heater and DR and the PC of the TIA [18]. The respective PC values are presented in Table 1. Adding the PC values of the integrated electronics and the LS, yields a total PC of 251.2 mW, which is translated into an EE of 10.04 pJ/bit for the 25 Gb/s operation. The EE decreases to 5.02 pJ/bit when scaling the operational speed to 50 Gb/s, as has been the line rate

TABLE 1 Main characteristics of the proposed AWGR-based interconnect

Optical power budget			
1:8 MUX/ DEMUX insertion loss (x2)		3 dB	
8×8 AWGR insertion loss		4 dB	
Polymer waveguides propagation loss (x2)		1 dB	
Silicon-to-polymer couplers insertion loss (x4)		2 dB	
Grating coupler insertion loss		1.5 dB	
Ring modulator insertion loss		3 dB	
Total link losses		**14.5 dB**	
Power consumption analysis			
Receiver sensitivity		-10 dBm	
Laser source output power		4.5 dBm	
Laser source power consumption (10% wall plug-in efficiency)		28.2 mW	
Heater power consumption		50 mW	
Driver power consumption		61 mW	
TIA power consumption		112 mW	
Total single-channel power consumption		**251.2 mW**	
System performance			
	This work 25 Gb/s	This work, 50Gb/s	QPI
Energy efficiency (pJ/bit)	10.04	5.02	16.2
Number of hops	1	1	1
Number of sockets	8	8	4
Socket line rate (Gb/s)	25	50	9.6
Socket capacity (Gb/s)	175	350	307.2
Interconnect capacity (Tb/s)	1.4	2.8	2.45

demonstrated experimentally for the four-channel SiPho TxRx [13]. These EE values suggest a significant improvement that goes up to 38% and 69% for the 25 Gb/s and 50 Gb/s operation, respectively, compared with the 16.2 pJ/bit EE of Intel QPI.

Moreover, the optical interconnect enables doubling, or even quadrupling, of the number of the interconnected sockets when comparing the four-socket single-hop implementation of the QPI. It also outperforms QPI's line rate of 9.6 Gb/s, since it holds the credentials for 25 Gb/s and 50 Gb/s operation. Although QPI exhibits higher socket capacity with respect to the optical interconnect at the line rate of 25 Gb/s, scaling to a 50 Gb/s line rate allows the optical interconnect to achieve 350 Gb/s/socket capacity or even 750 Gb/s, when the 16×16 AWGR is employed, which is higher than the respective value of QPI (307.2 Gb/s). Similarly, the interconnect capacity of the proposed architecture at the line rate of 25 Gb/s is lower than the respective value of the QPI. However, when scaling to a line rate of 50 Gb/s, the MSB interconnect yields a 2.8 Tb/s or 5.6 Tb/s interconnect capacity, at 8- and 16-socket implementations, respectively, which is at least 14.5% higher than QPI's capacity.

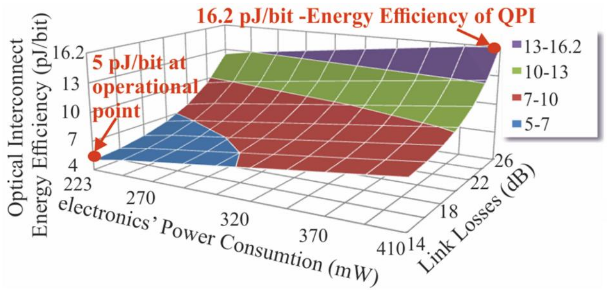

FIGURE 8 Energy efficiency (EE) of the optical interconnect versus the link losses and the integrated electronics' power consumption at 50 Gb/s

Finally, Figure 8 presents the EE of the photonic interconnect versus the PC of the integrated electronics and the link losses for the data rate of 50 Gb/s. As shown, the photonic interconnect exhibits a tolerance of 158.8 mW to the electronics' PC on top of the 251.2 mW until the total EE reaches the EE of the QPI standard, when the laser power is allowed to reach up to 16 dBm. This means that the electronics' PC value is allowed to range for up to 410 mW. Figure 8 also reveals that the link losses are allowed to reach the value of 26 dB, meaning that the tolerance of the photonic interconnection to the link losses is equal to 11.5 dB.

6 | CONCLUSION

We presented an overview of the electrical MSB interconnects and their challenges. We have discussed how optics can facilitate direct p2p interconnects communication between >4 sockets. We presented the recent progress towards demonstrating an O-band SiPho interconnect for MSBs that is based on WDM SiPho WDM TxRxs, a high-speed hosting EOPCB and a SiPho AWGR. A comparison between the operational characteristics of the QPI interconnect and the presented AWGR-based optical interconnect was performed, accompanied with an EE analysis revealing savings that can reach up to 69% for 50 Gb/s line rates. Finally, it should be noted that the validation of the thermal drift compensation system performed in [17], which allowed automatic alignment of the operating points of the main photonic building blocks and provided a safety net against on-chip temperature variations, by employing contact less integrated probes [26], an ASIC controller and the integrated heaters of the main building blocks of the proposed architecture, paves the way towards more complex system demonstrations that could encompass alignment and stabilization of the operating points of both the lasing and photonic circuitry.

ACKNOWLEDGEMENT

This work was supported by the European Commission through the H2020-ICT-STREAMS project (No. 688172).

ORCID

Miltiadis Moralis-Pegios https://orcid.org/0000-0002-9401-730X
Stelios Pitris https://orcid.org/0000-0001-5010-8843
Hannes Ramon https://orcid.org/0000-0002-2986-7017
Xin Yin https://orcid.org/0000-0002-9672-6652

REFERENCES

1. Moralis-Pegios, M., et al.: A 1024-port optical uni- and multicast packet switch fabric J. Lightw. Technol. 15(4), 1415–1423 (2019)
2. Terzenidis, N., et al.: High-port low-latency optical switch architecture with optical feed-forward buffering for 256-node disaggregated data centers. Opt. Express. 26, 8756–8766 (2018)
3. McMorrow, D., Corporation, M.: Technical Challenges of Exascale Computing. MITRE Corporation (2013)
4. Alexoudi, T., et al.: Optics in computing: from photonic network-on-chip to chip-to-chip interconnects and disintegrated architectures. J. Lightw. Tech. 37(2), 363–379 (2019)
5. Mulnix, D.: Intel Xeon Processor Family Technical Overview (2017) https://software.intel.com/content/www/us/en/develop/articles/intel-xeon-processor-scalable-family-technical-overview.html. Accessed 29 May 2019
6. Bull, S.A.S.: An Efficient Server Architecture for the Virtualization of Business-critical Applications. White Paper (2012)
7. Beck, N., et. al.: 'Zeppelin': An SoC for multichip architectures. IEEE International Solid-State Circuits Conference, San Francisco, CA (2018)
8. Yu, R., et al.: A scalable silicon photonic chip-scale optical switch for high performance computing systems. OSA Optic. Express. 21(26), 32655–32667 (2013)
9. Grani, P., et al.: Flat-topology high-throughput compute node with AWGR-based optical-Interconnects. IEEE/OSA J.Lightw. Tech. 34(12), 2959–2968 (2015)
10. Lamprecht, T., et al.: EOCB-platform for integrated photonic chips direct-on-board assembly within Tb/s applications. IEEE Electronic Components and Technology Conference, San Diego, CA (2018)
11. Moralis-Pegios, M., et al.: 52 km-long transmission link using a 50 Gb/s O-band silicon microring modulator co-packaged with a 1V-CMOS Driver. Photon. J. 11(4), 1–7 (2019)
12. Pitris, S., et al.: O-band silicon photonic transmitters for Datacom and Computercom interconnects. J. Light. Technol. 37, 5140–5148 (2019)
13. Moralis-Pegios, M., et al.: A 4-channel 200 Gb/s WDM O-band silicon photonic transceiver sub-assembly. Opt. Express. 28, 5706–5714 (2020).
14. Pitris, S., et al.: A 400 Gb/s O-band WDM (8×50 Gb/s) silicon photonic ring modulator-based Tranceiver. Optic Fiber Commununications Conference Exhibition (OFC), Paper M4H.3 (2020)
15. Pitris, S., et al.: Silicon photonic 8 × 8 cyclic arrayed waveguide grating router for O-band on-chip communication. Opt. Ex. 26(5), 6276–6284 (2018)
16. Fotiadis, K., et al.: 16×16 Silicon Photonic AWGR for Dense Wavelength Division Multiplexing (DWDM) O-band Interconnects. SPIE Photonics West (2020) Paper 11285-13
17. Moralis-Pegios, M., et al.: Automated thermal drift compensation in WDM-based silicon photonic multi-socket interconnect systems. Opt. Fiber Commun. Conf. Exhibit. (OFC) (2020) paper W3G.2
18. Gough, C., Steiner, I., Saunders, W.A.: Energy Efficient Servers - Blueprints for Data Center Optimization. Apress (2015)
19. Neugebauer, R., et al.: Understanding PCIe performance for end host networking. In: Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '18) New York, NY, USA, pp. 327–341 (2018)
20. Derradji, S., et al.: The BXI interconnect architecture. In: 2015 IEEE 23rd Annual Symposium on High-Performance Interconnects, Santa Clara, CA, pp. 18–25 (2015)
21. Idris, N.A., Tsuda, H.: 6.4-THz-spacing, 10-channel cyclic arrayed waveguide grating for T- and O-band Coarse WDM. IEICE Electron. Express. 13(7) (2010)
22. Takahashi, H., Oda, K., Toba, H.: Impact of crosstalk in an arrayed-waveguide multiplexer on N× N optical interconnection. IEEE/OSA J. Lightw. Techn. 14(6), 1097–1105 (1996)
23. Pitris, S., et al.: Crosstalk-aware wavelength-switched all-to-all optical interconnect using sub-optimal AWGRs. Phot. Techn. Lett. 31(18), 1507–1510 (2019)
24. Dangel, R., et al.: Polymer Waveguides Enabling Scalable Low-Loss Adiabatic Optical Coupling for Silicon Photonics. IEEE J. Sel. Top. Quant. Electr. 24(4), 1–11 (2018)
25. Lambrecht, J., et al.: 90-Gb/s NRZ optical receiver in silicon using a fully differential transimpedance amplifier. J. Lightw. Technol. 37(9), 1964–1973 (2019)
26. Annoni, A., et al.: Automated routing and control of silicon photonic switch fabrics. IEEE J. Sel. Top. Quant. Electr. 22(6), 169–176 (2016)

How to cite this article: Moralis-Pegios M, Pitris S, Mitsolidou C, et al. Silicon circuits for chip-to-chip communications in multi-socket server board interconnects. *IET Optoelectron*. 2021;15:102–110. https://doi.org/10.1049/ote2.12018