Application-Oriented On-Board Optical Technologies for HPCs

Pavlos Maniotis, Nikolaos Terzenidis, Apostolos Siokis, Konstantinos Christodoulopoulos, Emmanuel Varvarigos, Marika Immonen, Hui Juan Yan, Long Xiu Zhu, Kobi Hasharoni, Richard Pitwon, Kai Wang, and Nikos Pleros

***Abstract*—The increased communication bandwidth demands of high performance computing (HPC) systems calling at the same time for reduced latency and increased power efficiency have designated optical interconnects as the key technology in order to achieve the target of exascale performance. In this realm, technology advances have to be accompanied by the development of corresponding design and simulation tools that support end-to-end system modeling in order to evaluate the performance benefits offered by optical components at system scale. In this paper, we present recent advances on electro-optical printed circuit boards (EOPCB) technology development pursued within the European FP7 PhoxTroT research program and directed toward system-scale performance benefits in real HPC workload applications. We report on high-density and multilayered EOPCBs together with all necessary building blocks for enabling true optical blade technology, including multimode polymer-based single- and dual-layer EOPCBs, a board-compatible optically interfaced router chip, and passive board-level connectors. We also demonstrate a complete optical blade design and evaluation software simulation framework called OptoHPC that tailors optical blade technology development toward optimized performance at HPC system scale, allowing for its validation with synthetic workload benchmark traffic profiles and for reliable comparison with existing HPC platforms. The OptoHPC simulator is finally utilized for evaluating and comparing a 384-node HPC system relying on optically enabled blades with the state-of-the-art Cray XK7 HPC network when performing with a range of synthetic workload traffic profiles, revealing the significant throughput and delay improvements that can be released through application-oriented optical blade technology.**

***Index Terms*—Electro-optical PCBs, flexplane technology, HPC network simulation, Omnet++, optical interconnects, optoelectronic router chip.**

I. Introduction

The predictions and expectations for exaflop High Performance Computing Systems (HPCs) by 2020 [1] rely mainly on the aggregation of vast numbers of Chip Multiprocessors (CMPs) within the HPC platforms, constantly pushing the performance envelope at all three critical factors: bandwidth, latency and energy efficiency. With the currently on-board electrical and off-board optical employed interconnect system comprising still a major bottleneck, optical interconnect and photonic integration technologies are being promoted as highly promising interconnect solutions with the aim to translate their proven high-speed, low-latency and energy-efficient data transfer advantages into respective benefits at system-level. Optics are rapidly replacing electrical interconnects with Active Optical Cables (AOCs) forming already a well-established technology in rack-to-rack communications. At the same time, mid-board optical subassemblies and compact board-level flexible modules, like FlexPlane [1], have recently entered the market targeting the replacement of conventional on-board interconnects for chip-to-chip communication purposes.

Emerging optical technologies are continuously penetrating at deeper hierarchy levels. In [2] a mid-board optical (MBO) transceiver with 12 Coolbit Optical Engines [3] is presented, performing at ~25.8 Gbps and offering a total bandwidth of 300 Gbps per square inch. The Coolbit Engine development process begins with the semiconductor fabrication of the VCSEL and Photodiode ICs, moves to the automated wafer assembly of the VCSEL, photodiode and other ICs and ends with the operational testing of the wafer. Moreover, INTEL has introduced the hybrid silicon 100G PSM4 QSFP28 Optical Transceiver [4], targeted for use in optical interconnects for data communications applications. Luxtera from the other side has successfully entered the Silicon Photonics market [5] by presenting the LUX62608 OptoPHY and LUX42604 QSFP optical transceiver modules [6] that offer a small form-factor, high speed, and low power consumption solution for Cloud Data Interconnect, Local

Manuscript received September 15, 2016; revised December 24, 2016 and February 25, 2017; accepted March 5, 2017. Date of publication March 13, 2017; date of current version June 24, 2017. This work was supported by the European Commission through FP7-ICT project PhoxTroT under Contract 318240.

P. Maniotis, N. Terzenidis, and N. Pleros are with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54621, Greece (e-mail: ppmaniot@csd.auth.gr; nterzeni@csd.auth.gr; npleros@csd.auth.gr).

A. Siokis is with the Computer Technology Institute and Press "Diophantus," Patras 26504, Greece, and also with the Computer Engineering and Informatics Department, University of Patras, Patras 26504, Greece (e-mail: siokis@ceid.upatras.gr).

K. Christodoulopoulos and E. Varvarigos are with the Computer Technology Institute and Press "Diophantus," Patras 26504, Greece, and also with the National Technical University of Athens, Athens 15780, Greece (e-mail: kchristodou@ceid.upatras.gr; manos@ceid.upatras.gr).

M. Immonen is with TTM Technologies, Salo 24100, Finland (e-mail: marika.immonen@ttmtech.com.hk).

H. J. Yan and L. X. Zhu are with TTM Technologies, Shanghai 201600, China (e-mail: huijuan.yan@smst.ttmtech.com.cn; longxiu.zhu@smst.ttmtech.com.cn).

K. Hasharoni is with Compass Networks, Netanya 42505, Israel (e-mail: kobi.hasharoni@dustphotonics.comkobihash@gmail.com).

R. Pitwon and K. Wang are with the Department of Photonics Research and Development, Seagate, Havant PO9 1SA, U.K. (e-mail: richard.pitwon@seagate.com; kai.wang@seagate.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JLT.2017.2681972

Area Network (LAN) and High Performance Computing (HPC) applications.

Going a step further, Optical Printed Circuit Board (OPCB) layouts can offer high-density, energy efficient and low-loss Tb/s on-board data transmission forming a promising solution for completely replacing the copper printed wires and their associated low bandwidth and distance- and speed-dependent energy dissipation problems. OPCBs have successfully revealed various optical wiring solutions like (a) optical fiber to the board [7], (b) optical polymer waveguides to the board [8]–[12], (b) embedded optical polymer waveguides [13], [14] and (c) embedded glass optical waveguides [7], [15], [16], while at the same time very high density parallel interfaces have been presented [17], [18]. Single-layered and single-mode arrays of optical waveguides in OPCBs have been recently presented to offer as low as 0.6 dB/cm propagation losses at 1310 nm and a total density of 50 wires/cm [19]. Bringing multiple optical layers hybridly integrated in Electro-Optical PCB (EOPCB) layouts with several electrical interconnect layers comprises the next big goal towards increasing the number of wiring and routing paths, with recent works reporting already on successful implementations of multi-layer embedded optical [16] and polymer [20], [21] waveguides. Although extensive hardware reliability studies have yet to be done [22] and low-cost mass-manufacturing processes have to deployed for enabling their market adoption, the EOPCB technology holds much promise for eliminating the low bandwidth and distance- and speed-dependent energy dissipation problems originating from copper printed wires.

This roadmap, combined with the rapid progress on mid-board optical transceiver chips [9], [23]–[25] has also triggered expectations for on-board optoelectronic routing schemes either via optically interfaced electronic router ASICs [26], or via silicon photonic switching platforms [27]. After the successful examples of circuit-switched optical solutions in Data Center environments [28], [29], the approach of on-board optically enabled routing seems to gain momentum as the line-rates of ASIC I/O ports reached already 25 Gb/s [30], [31]. Bringing optics as close as possible to the ASIC I/Os can yield significant power benefits at board-level signal routing, mimicking the case of the board-to-board connectivity where the recent release of fiber-coupled router ASIC from Compass EOS allows for just 10 pJ/bit consuming optical I/O ports [26].

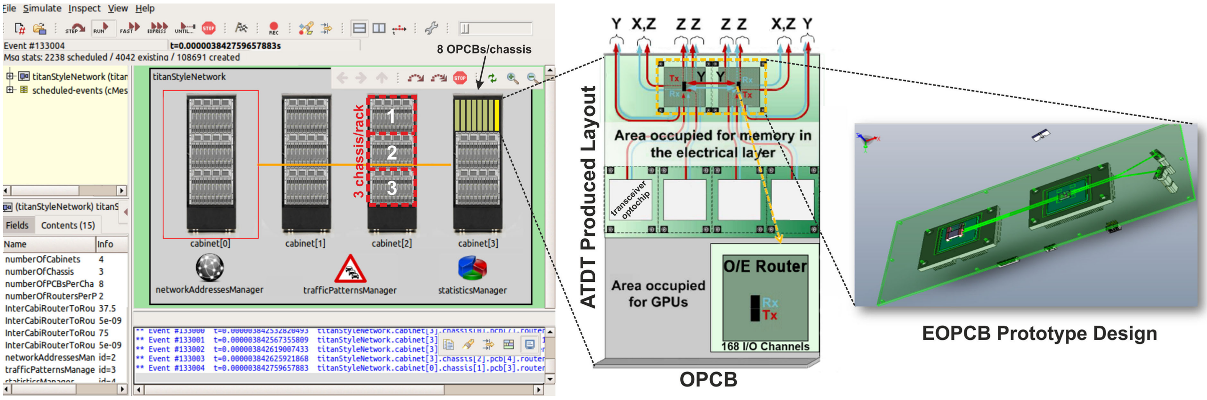
However, the rapid progress witnessed in the fields of board-level optical interconnects and optoelectronic routing technologies has still not been provenly neither tailored nor reflected in system-scale benefits in HPC environments. Although advantages at link-level are being thoroughly addressed, the EOPCB layout and the performance of a complete HPC engine that exploits EOPCBs and performs with workload applications is usually still an unknown parameter. One main reason for the disassociation between hardware technology development and HPC-scale performance lies also in the lack of a corresponding system-scale simulation engine that would allow for optimally exploiting the new technology toolkit through performance evaluation at HPC level. Although photonics have already emerged in chip-scale simulation platforms like PhoeniXSim [32] suggesting optimal technology and network architecture design rules through system-scale performance [33], state-of-the-art sophisticated HPC simulators still cannot efficiently support the use of advanced electro-optic router and interconnect solutions at board-level. Extreme-scale Simulator (xSim) [34] and SST + gem5 [35] are some of the few open-source simulators that are free of charge and available to the research community but none of them is focused on or can even efficiently explore the adoption of optical technology advancements in the HPC field.

In this paper, we present the recent technology highlights accomplished within the European project PhoxTrot towards implementing and demonstrating a fully functional Optical Blade along with a complete optically enabled HPC hardware/architecture ecosystem that tailors EOPCB design around application-oriented optimized HPC performance. We report on the development of the most basic building blocks on the way to board-level optoelectronic router blades, spanning from single- and multi-layered multi-mode polymer-based EOPCBs with a high electronic layer count, through board-level coupling interfaces and up to optically enabled board-adaptable router chips. Technology development goes hand-by-hand with application-oriented design through the combined employment of the Automatic Topology Design Tool (ATDT) [36] and the OptoHPC-Sim [37] toolkits that allow for system-scale-optimized on-board optical interconnect layouts. ATDT is a software design suite that is capable of providing the optimum OPCB interconnect layout for a given layout strategy, while the OptoHPC-Sim engine is a complete HPC network simulator supporting the employment of optical technologies and focusing on analyzing the performance of the entire HPC network under a wide range of synthetic and realistic application traffic profiles. Finally, we exploit our hardware/architecture design ecosystem and present a comparative performance analysis between world's no. 3 Supercomputer Titan CRAY XK7 (as of June 2016) [38], and a respective HPC architecture where PhoxTrot optical blades have replaced the electronic CRAY blades. The results presented in this work reveal that the employment of board-level optics in appropriate layouts can lead to optically enabled HPC systems that can significantly outperform top-class HPC machines, on average offering throughput improvements higher than 190% for a number of 8 workload benchmarks.

The rest of this paper is organized as follows: Section II describes the optical blade design layout as pursued within the PhoxTroT project and all the technological advancements achieved towards electro-optical boards employing optical interconnects and optoelectronic router chips for use in future HPC systems. Section III presents the ATDT and optoHPC-Sim, while Section IV proceeds with a performance evaluation analysis by comparing an HPC network system employing state-of-the-art optoelectronic routers and optical interconnects with a system employing a purely electrical board layout as is being used in Titan CRAY XK7. Section V concludes the paper.

II. On-Board Optical Technology Platform

The application-oriented technology development roadmap is illustrated in Fig. 1. It presents an example HPC network of 4 racks, as it appears at the GUI interface of the OptoHPC-Sim simulator. The internal rack architecture hierarchy follows the architecture of the Titan CRAY XK7 supercomputer [38], where



Event #133004 t=0.000003842759657883s
Msg stats: 2238 scheduled / 4042 existing / 108691 created

Name	Info
numberOfCabinets	4
numberOfChassis	3
numberOfPCBsPerCha	8
numberOfRoutersPerP	2
InterCabiRouterToRou	37.5
InterCabiRouterToRou	5e-09
InterCabiRouterToRou	75
InterCabiRouterToRou	5e-09
networkAddressesMan	id=2
trafficPatternsManage	id=3
statisticsManager	id=4

```
** Event #133000  t=0.000003842532820493  titanStyleNetwork.cabinet[3].chassis[0].pcb[7].router
** Event #133001  t=0.000003842567355809  titanStyleNetwork.cabinet[3]
** Event #133002  t=0.000003842619007433  titanStyleNetwork.cabinet[3]
** Event #133003  t=0.000003842625921868  titanStyleNetwork.cabinet[3].chassis[2].pcb[4].router
** Event #133004  t=0.000003842759657883  titanStyleNetwork.cabinet[0].chassis[1].pcb[3].router
```

Fig. 1. *OptoHPC-Sim's* main GUI frame demonstrating an example HPC model incorporating four racks. Each rack consists of 24 OPCBs being grouped in 3 chassis of 8 OPCBs each. OPCB layout has been designed with the ATDT tool. At the right the EOPCB design with the 2 router interfaces is also demonstrated.

8 CRAY XK7 Blades are grouped together forming a chassis and three chassis are grouped together forming an HPC rack. At the top of the 4th HPC rack, a cluster of 8 electro-optical PCBs forming a chassis is highlighted and illustrated as inset in more detail. It shows a single OPCB with the optical links having been generated by the ATDT tool [36], whose role is to provide the optimum OPCB interconnect layout for a given layout strategy. The OPCB includes proper sockets for hosting 4 transceiver optochips and 2 optoelectronic router chips along with the proper pin connections between them. Transceiver optochips serve as the interface between the CPU chips and the board-level optical waveguides, while the optoelectronic router chips connect the CPU chips together as well as with the outer world off-board devices. The inset at the right side of Fig. 1 presents the EOPCB prototype design that is currently being fabricated within the PhoxTroT project in order to validate the basic blade functionality required by the 4-rack HPC network. This EOPCB prototype is capable of hosting two Compass EOS optoelectronic router chip modules [26] that allow both for chip-to-chip as well as for off-board communication by optical means. The critical technology blocks required for enabling this EOPCB prototype include a) the EOPCB, b) the board-adaptable electro-optic router ASIC together with the appropriate chip-to-board interfaces, and c) the board-level connectors and coupling interfaces. The following subsections describe in more detail the progress along all these individual technology blocks towards realizing an optical blade capable to serve the needs of the 4-rack HPC network shown in Fig. 1.

A. *High End Routing Platform Using Optical Interconnects*

This section briefly reviews the optoelectronic router chip developed by Compass Networks [26], which will be utilized in its board-adaptable version for serving as the on-board routing machine. The router ASIC developed is a traffic manger operating as a queuing machine, in which the conventional electrical serial Input/Output (I/O) have been replaced with a pure optical I/O interface. In CMOS design, high frequency chip I/O ports can be located only on the perimeter of the package since an areal array of I/Os is more form factor efficient than a linear array at the same linear pitch. More details on optical and electrical escape bandwidth can be found in [39]. The I/O ports number is thus limited by the package size and the ability to connect these serial I/Os to the metal traces on the PCB (trace breakout). An optical chip interface allows overcoming this radix limitation via the assembly of dense, two-dimensional arrays of lasers and photodiodes directly on the ASIC. A matching 2 D fiber array is assembled above the optoelectronic matrices allowing for direct coupling of light to and from the fibers. This technology allows for very high data densities and to increase the BW-distance product of the device while minimizing the power required for chip I/O.

Using this approach, several linecards can be connected directly using fibers thereby eliminating the need for a backplane. In a typical router application, traffic from the packet processing unit is routed to the traffic manager ASIC with its on-chip parallel optical interconnect connected via parallel fiber arrays to several traffic managers on different linecards and racks. The high BW available from the large optical matrices allows building large, non-blocking Clos networks since the router nodes have enough optical BW to connect directly without the need for an intermediate switch fabric. This approach is followed in [26], where various linecards are connected with fiber optics thereby eliminating the package constraints and greatly simplifying the linecard architecture: the traffic from the packet processing unit is routed to a traffic manager/queuing machine ASIC with an on-chip parallel optical interconnect which is linked via parallel fiber arrays to several traffic managers on different linecards with minimal queuing constraints.

A cross-sectional view of the parallel optical interconnect assembled on the traffic manager chip is shown in Fig. 2. This is a mixed signal chip with both digital and analog functionalities. Two dimensional matrices of InGaAs/GaAs Vertical-Cavity Surface-Emitting Lasers (VCSELs) and InGaAs/InP photodiodes (PDs) are directly attached to their analog circuits in the chip. Each VCSEL is located directly above a Tx cell containing the laser driver and serializer. Similarly, each PD is located above an Rx cell containing the TIA, limiting amplifier, equalizer, de-serializer and clock data recovery circuit (CDR). This is a localized design with each optoelectronic pixel electrically

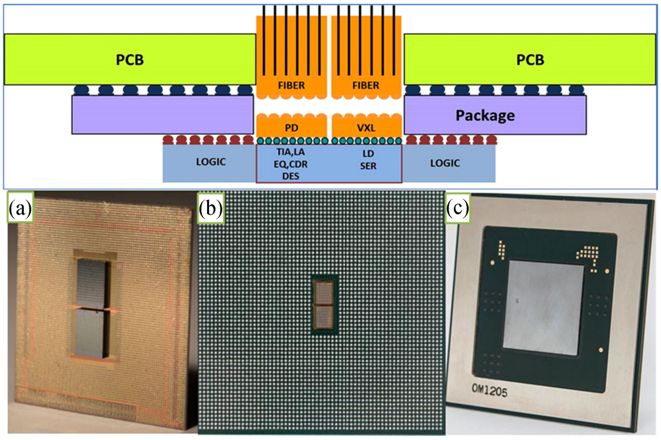


Fig. 2. Schematic outline of the hybrid optical interconnect showing chip and optical packaging (top) and the actual chip with the CMOS die and assembled VCSEL and PD matrices (bottom, A) and the packaged chip with a cutout hole in the package for optical coupling (bottom, B & C) [41].

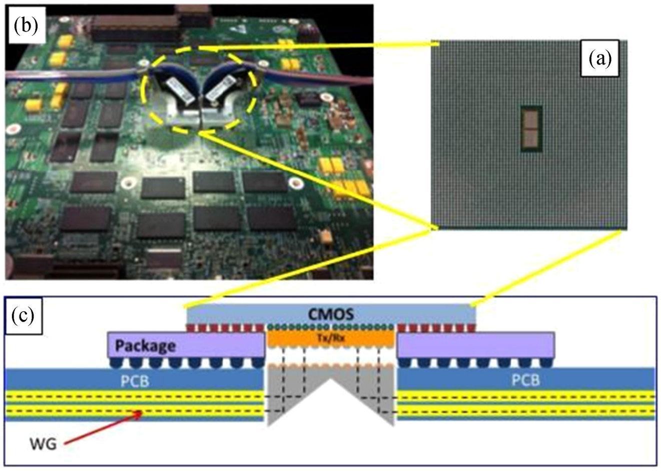


Fig. 3. Optical coupling of the on-chip optical interconnect (A) with a 2 D fiber bundle assembled on the PCB (B); coupling to a double-layered embedded waveguide array in a PCB using microlens and prisms (C) [41].

isolated from all other pixels, with pixel denoting either an individual VCSEL or a PD module. The transmission length from the analog circuit to the pixel is in the 100 μm range thereby minimizing the effect of parasitics on the link.

This design consumes significantly lower power with respect to conventional chip I/O. In standard design, most of the power dissipation is in the serializer/deserializer (SerDes) arrays. However, unlike the typical chip package case where the SerDes drives a lossy ($\sim$25 dB) copper trace; in this case, it drives a low loss ($<$0.2 dB) fiber optic link. This allows packaging all of the analog components densely in a 250 × 250 μm cell. With an 8 Gb/s clock, the power for an optical link is 10 pJ/bit, including the SerDes arrays. For a 168 element device (1.34 Tb/s BW, full duplex), the total power consumption of the optical interconnect and its analog circuits is about 12 W– significantly lower than any comparable copper I/O technology. Future generation of this technology based on 16 nm CMOS will have power efficiency in the 3–4 pJ/bit (including the SerDes arrays) [40].

Thermal management of the optoelectronics is a main concern as they are mounted on top of a high power ($\sim$100 W) ASIC. The requirement is thus twofold: i) remove the heat from the laser active region; and ii) avoid heat flow from the ASIC to the lasers and PDs. This is carried out by adding electrically isolated pillars that conduct heat from the optoelectronics active region to the chip below. The excess heat is removed by forced air incident on the ASIC heatsink.

The 2 D optoelectronic chips cannot be assembled on the ASIC using standard 850 nm VCSELs as they would be illuminating into the CMOS die since the power bumps are in the same plane as the laser/PD aperture. Therefore the optoelectronics are made back illuminating with light going through the III-V substrate. The operating wavelength has to be red-shifted to 1060 nm where the substrate is transparent; this is carried out by increasing the Indium percentage in the heterojunction.

The reliability of the VCSEL is not affected by these changes and FIT values of $\sim$30 have been measured at 80 C. To further enhance reliability, a 20% laser redundancy has been included in the router chip. However, no failures have been noted to date during field operation of about 135 000 VCSELs for more than 35 000 h.

The mixed signal ASIC die has an area of $\sim$450 mm^2 and the Tx/Rx analog circuits occupy about 10% of the die area, with the rest being digital logic. The chip was fabricated using TSMC 65 nm CMOS process and the wafers are post-processed for Cu under bump metallization (UBM) and eutectic SnPb bump deposition using standard processes. Flip-chip technology is used to position the VCSEL and PD dies on the ASIC. Differences in the thermal expansion coefficient between Silicon and the III-V substrates are compensated using high aspect ratio Gold pillars.

The high bump count of the final packaged chip requires the use of a high density organic substrate for connecting the ASIC die with the PCB by rewiring of the CMOS bumps to a BGA matrix with $\sim$4000 balls. Flipchip is used also here to assemble the die on the organic substrate. Since light needs to be coupled to and from the VCSEL and PD matrices, a cutout hole is made in the package allowing direct access to them. More details about the fabrication processes and procedure can be found in [26].

Optical coupling of the on-chip interconnect can be carried either to a 2D fiber optic array mounted on the PCB [see Fig. 3(b)] or to an embedded waveguide array [see Fig. 3(c)]. In either case, a 2-lens relay is used for light coupling. A cutout hole is required thus also in the PCB. In the fiber coupling case, the Tx and Rx fiber arrays are mounted vertically above the VCSEL and PD matrices. The fiber bundles are glued to the PCB surface and this solution has passed extensive reliability tests, already being applied in real field applications of the optically-interconnected router chip. In the case of waveguide coupling solution. which is still in the development phase, a right angle prism has to be used with the microlens array to facilitate in-plane light coupilng. This approach can also be used for fiber coupling in future generations.

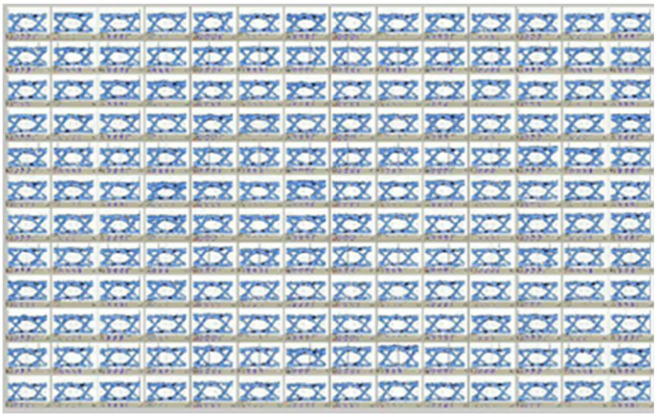


Fig. 4. PRBS31 eye diagrams from a 12 × 14 VCSEL matrix at 8 Gb/s line-rate [26].

The assembled PCB with 2 fiber bundles (Tx and Rx) is shown in Fig. 3 [41]. The two fiber bundles are connected to the system optical backplane, which is also a fiber bundle (similar approaches have been also developed in the past [42]). Using this arrangement, the link between any two ASICs in the system is carried out with a passive fiber link.

The high bandwidth (BW) of the optical interconnect is obtained by using large matrices in the transceiver. The device has 168 optical channels in a 12 × 14 layout and uses 8 Gb/s optoelectronic chips, leading to an aggregate BW of 1.34 Tb/s with a data density of 64 Gb/s/mm^2. This chip is currently in the process of serving as the board-adaptable router chip in the Optical Blade Design presented in the next sections; however the recent progress towards 336-element optical I/O matrix size [41] raises expectation for future on-board router chips with record high aggregate capacity values. Fig. 4 shows the eye diagrams from a 168 element VCSEL matrix performing at 8 Gb/s line-rates and producing a 2^{31}–1 PRBS test pattern [26]. All 168 eyes exhibit BER $< 10^{-12}$ at the center of the eye and are clearly open with an extinction ratio of about 5 dB and high Signal-to-Noise Ratio (SNR) values. The optical crosstalk between neighboring cells was in the −32 dB range indicating good optical and electrical isolation in the matrix. Sensitivity measurements with a 2 m, 200 m and 300 m multimode OM3 fiber reported a sensitivity level of about −10 dBm at a BER of 10^{-12} for all fiber length. This is indicative that the receiver CDR design is sensitive enough to overcome the distortions resulting from modal dispersion. With an average VCSEL power of ~2 dBm, this result indicates a dynamic range of about 10 dB [26].

B. Multi-Mode Electro-Optical PCB Technology

Fig. 5 depicts the mask layout for the EOPCB prototype that can host two optoelectronic router chips and follows the EOPCB design illustrated in the inset of Fig. 1. This prototype layout aims at all-optical chip-to-chip connectivity using multimode polymeric waveguide arrays embedded in conventional multilayer PCB card with up to 16 electrical layers. The two optoelectronic chips are located at a distance of 15 cm and have their optical I/O matrix facing the PCB, so that the VCSEL transmitter matrix of the first chip can connect to the PD receiver matrix of the second chip via a 14-element multimode polymer waveguide array. The waveguide loop has been implemented in the EOPCB prototype in order to allow for a step-wise experimental testing approach for the chip-to-chip connectivity. This loop allows the transmission of data from a router chip back to the same chip, so that the experimental testing can be carried out even when assembling the first chip prior having both router chips deployed on-board.

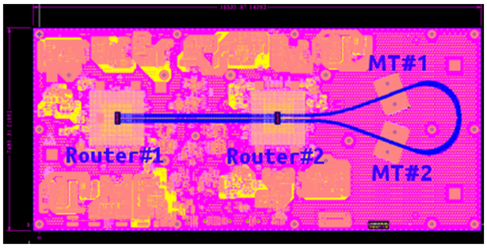


Fig. 5. Optical/Electrical PCB demonstrator with 16 copper and 1 optical embedded layer. Black rectangles at Router areas represent the optical I/Os that couple to the underlying embedded waveguide layer following the concept of Fig. 3 (c). MT#1 and MT#2 areas are targeted for vertical out-of-board connections via MT Termination Push-On sites for fiber-to-waveguide connections. Optical waveguide tracing layout is shown in blue.

Waveguides were designed to have rectangular core cross-section with nominal size of 50 × 50 μm, and channel-to-channel spacing of 250 μm. Dow Corning liquid silicone materials were used to fabricate the waveguides. Core material (WG-1010) with refractive index n = 1.519 were surrounded by lower index cladding material (OE-4141) with n = 1.501 to result NA = 0.24 channels. Clad layer thicknesses were 50 μm and 25 μm for bottom and top cladding, respectively.

Fig. 6(a) presents a schematic of chip-to-waveguide coupling concept. The beam folding element to couple light from/to TX/RX to waveguide comprises of beam splitter with microlens array (MLA) [see Fig. 6(b)] glued on one of the flat surfaces. After EOPCB fabrication, the folding element is assembled into cut out hole in the PCB (size 5.13 × 11.5 mm) [see Fig. 6(c)].

Previous OPCB link demonstrators show optical waveguides as separate entity from PCB [13], as layer built on board surface [9]–[12] or embedded part of PCBs with varying complexity and a limited number of up to 4 electrical layers [8], [14], [43]–[46]. In this paper, the optical waveguide array has been embedded for the first time inside a high layer count PCB product with 16 to 20 copper layers including one or two optical layers in the stack for compliance with product form factors, constructions and board materials [14]. The board contains all required electrical layers and via structures (Plated-Through Holes(PTHs), n-PTH, stacked and buried microvias) built around optical cores, following certain process and design strategies during the development for: (a) rerouting of all signals to avoid areas with optical waveguides, (b) managing processing of sub-cores with different copper thickness (17 μm for signal (S), 35 μm for power (P) and 70 μm for ground (G) layers), (c) providing three microvia layers as part of the EOPCB, (d) controlling registration and material movement during lamination of dissimilar

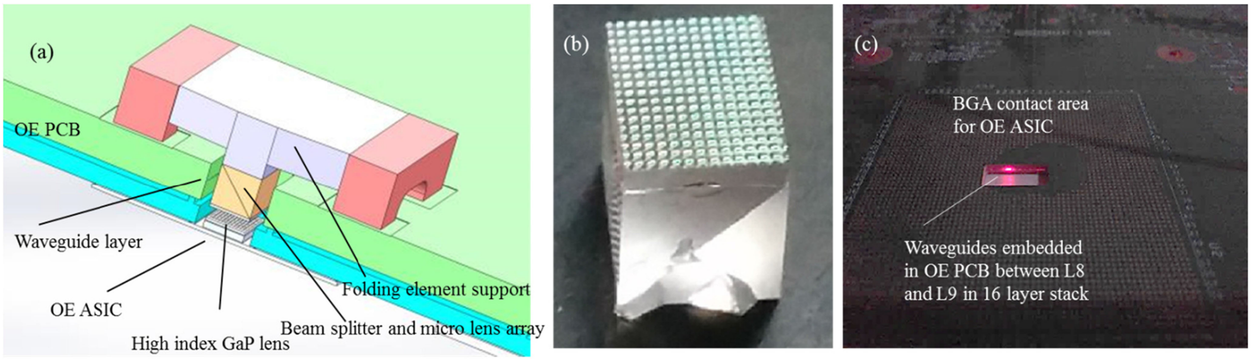

(a)
OE PCB
Waveguide layer
OE ASIC
High index GaP lens
Beam splitter and micro lens array
Folding element support
(b)
(c)
BGA contact area for OE ASIC
Waveguides embedded in OE PCB between L8 and L9 in 16 layer stack

Fig. 6. (a) Schematic of chip-to-waveguide coupling concept (b) folding element comprising of beam splitter and microlens array (MLA), (c) cut out section in OEPCB with waveguide illuminated.

materials and (e) providing a process flow with minimal thermal load to waveguides.

The final fabricated EOPCB board has an outline of 190 mm × 420 mm and comprises 16 electrical layers for signal and power line interconnects and one optical waveguide layer stacked between copper layers 8 (L8) and 9 (L9). The construction of optical/electrical build is 8 electrical + 1 Optical + 8 electrical. However, this design uses only a small percentage of the actual optoelectronic router chip interconnect size, which has a 12 × 14 layout. For assembling large high I/O count O/E ASIC packages on board, high flatness in the BGA areas as well as very low bow/twist must be achieved. For that, balanced board construction imposing minimal thermo-mechanical stress to optical layer and providing high rigidity e.g. bow/twist compliant with d-factor specification <7% . . . <5% was objected. Low dielectric constant Dk (3.6–3.8 @ 10 GHz) and dissipation factor Df (0.0070∼0.0090 @ 10 GHz) resin system (Hitachi HE679 G(S) with low CTE ($\alpha 1$) Z-axis 30–40 ppm/°C was selected as dielectric material due to its high dimensional stability required to achieve low movement and predictable fabrication in a complex hybrid O/E construction. Hitachi HE679 GS is halogen free and high heat resistance material used in high frequency applications. Board stack was equalized on copper content and number of copper layers top/bottom adjoining the optical layer. Further impacts with non-functional dielectric layers and parameters were optimized to maximize stack stability and minimize laminate movement and stress during fabrication and assembly, which are critical to control in PCBs with embedded optical elements.

Besides chip-to-chip connectivity via embedded polymer waveguides, the EOPCB prototype hosts two mid-board Multi-fiber Termination Push-On (MTP) sites for fiber-to-waveguide connections. These MTP sites provide out-of-plane waveguide connection with embedded micro-mirrors, which were embedded directly into the waveguide substrate as part of the PCB fabrication progress and connected to lensed MT ferrules assembled in a slot perpendicular to the mirrors. Except from the two chip-to-board interfaces presented in Fig. 5, two mid-board MTP fiber-to-WG test connectors can be seen at the right side of the board. In addition, Fig. 7 shows an overview of the fabricated board as well as a cross-section across the stack detailing the electrical layers and the embedded waveguides.

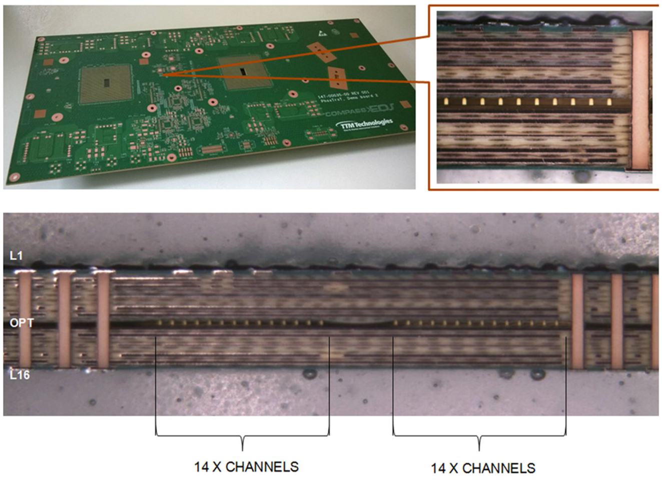


Fig. 7. (Top) Fabricated EOPCB with embedded MM polymer waveguide layer, (bottom) Cross-section of the EOPCB showing 14 + 14 waveguides.

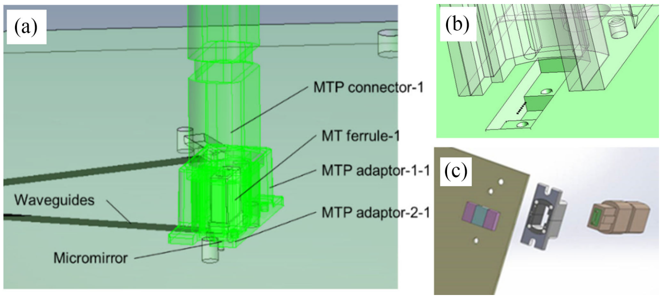


Fig. 8. (a) Schematic of fiber-to-waveguide coupling concept (b) recess cavities formed for micromirror and lensed MT ferrule, (c) breakdown of the final assembly with second MT ferrule in the bottom side to serve as high accuracy mating part.

Coupling of light from the MT fiber array to the WG array is carried out using a commercial lensed MT ferrule (see Fig. 8). Here, Enplas (PN 037 type 2) and TE (PN 2123368-1 TELLMI) were used. The folding micro mirror is a Au coated glass prism. Since the WG array is about 1 mm below the PCB surface, a double cavity is cut in the board on the site of the MT connectors as shown in Fig. 8(b). The lower cavity (3.5 × 0.55 × 0.25 mm) contains the mirror element facing towards the WG array and is located at the level of the WG layer. The upper cavity is larger

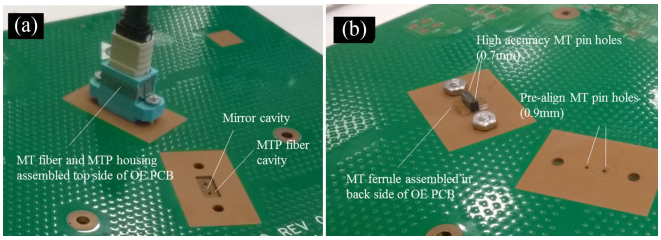
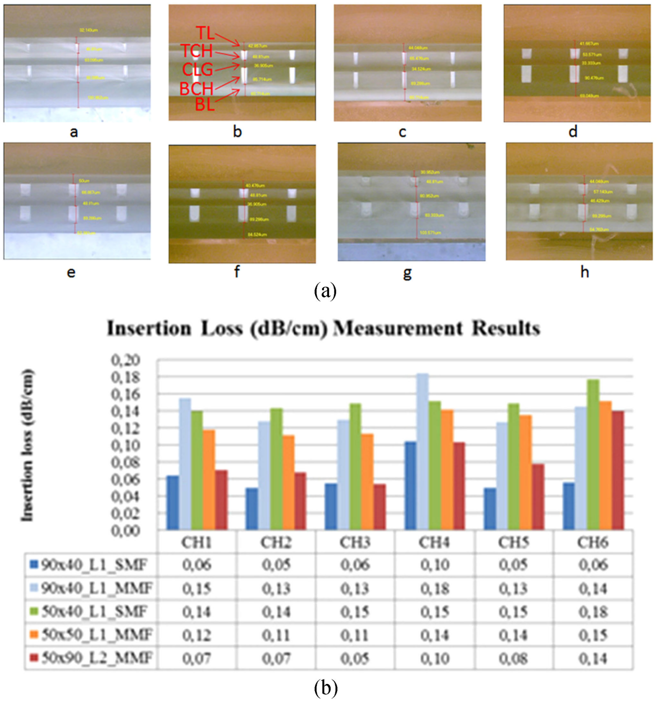

(a)
MT fiber and MTP housing assembled top side of OE PCB
Mirror cavity
MTP fiber cavity
(b)
High accuracy MT pin holes (0.7mm)
Pre-align MT pin holes (0.9mm)
MT ferrule assembled in back side of OE PCB

Fig. 9. (a) Assembled MT fiber and MTP housing on recess cavity formed in OE PCB top side of the board (b) MT ferrule assembled in the back side of the OE PCB providing high accuracy mating for fiber-to-WG coupling.

(3.5 × 4.8 × 0.8764 mm) and houses the MT ferrule and the lens array. The glass prisms were coated by 1000/3000A Ti/Au thin film for high reflectance. The mirror loss measured at 850 nm was 0.6 ± 0.1 dB. Fig. 8(c) presents the breakdown of the final assembly with the second MT ferrule in the bottom side to serve as high accuracy mating part.

In order to facilitate the pluggable operation of the MT mid-board connectors, the MT guide pins must be accurately inserted into their corresponding mating holes. Since PCB drill technology is not capable enough to provide sufficient accuracy, the following procedure was used to align the MT guide pins with looser tolerance. Two large diameter guide holes were drilled on both sides of the embedded mirror throughout the board. The guide-hole diameter is 0.9 mm while the MT pin diameter is 0.7 mm. A second MT ferrule is positioned on the bottom surface of the board and during assembly, the guide pins are inserted into it while the align ferrule is moved until optimum coupling is achieved. The ferrule is glued to the board and will serve to align the lensed MT ferrule upon each insertion. Fig. 9(a) presents the assembled MT fiber and MTP housing on recess cavity formed in OE PCB top side of the board and Fig. 9(b) presents the MT ferrule assembled in the back side of the OE PCB providing high accuracy mating for fiber-to-WG coupling.

Focusing now to 16" × 20" standard production panels and taking advantage of the established fabrication processes [47], we report for the first time that the developed process has been scaled up to support EOPCBs with two optical layers. Fig. 3(c) presents the targeted coupling scheme that is based on microlenses and prisms, which is in principle transferrable also to the fiber-connector scheme and could allow for reducing the form factor of the 2 D fiber bundle approach that was presented in Fig. 3(b). Using this coupling scheme in the dual-layer embedded waveguide board layout, as shown in Fig. 3(c), would allow for the utilization of the two outer rows of the router's 12 × 14 I/O optical matrix. In such an arrangement, the first outer-row 48 peripheral IO pins connect to the first PCB waveguide layer and the second-periphery row 40 pins connect to the second waveguide layer, so that finally only 88 out of the 168 optical pins can be accommodated. In order to fully exploit the whole 12 × 14 optical I/O matrix of the router at on-board setups, the electro-optical PCB should be replaced by flexplane technology (presented in Section II-D), since multi-layer EOPCB deployments are still facing severe difficulties towards accommodating more than 2 waveguide layers.



	CH1	CH2	CH3	CH4	CH5	CH6
■ 90x40_L1_SMF	0,06	0,05	0,06	0,10	0,05	0,06
■ 90x40_L1_MMF	0,15	0,13	0,13	0,18	0,13	0,14
■ 50x40_L1_SMF	0,14	0,14	0,15	0,15	0,15	0,18
■ 50x50_L1_MMF	0,12	0,11	0,11	0,14	0,14	0,15
■ 50x90_L2_MMF	0,07	0,07	0,05	0,10	0,08	0,14

Configuration	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
TOP LAYER (TL) (um)	32.143	42.857	44.048	41.667	50	40.476	30.952	44.048
TOP CORE HEIGHT (TCH) (um)	48.81	48.81	65.476	53.571	66.667	48.81	48.81	57.143
CORE LAYER GAP (CLG) (um)	63.095	36.905	34.524	33.333	48.81	36.905	80.952	46.429
BOTTOM CORE HEIGHT (BCH) (um)	88.095	85.714	89.286	90.476	89.286	89.286	83.333	89.286
BOTTOM LAYER (BL) (um)	130.952	60.714	60.714	69.048	52.381	84.524	103.571	54.762

(c)

Fig. 10. (a) Dual Layer embedded optical waveguides with different geometrical specifications. (b) Insertion loss (IL) measurement results in dual layer construction (c) table with geometrical specifications of (a).

The first fabricated units were realized with varying core size with a width equal to 20 μm, 35 μm, 50 μm and 60 μm and a height varying from 45 μm to 90 μm. On top of that they show an excellent waveguide layer-to-layer registration of less than +/–5 µm between two optical layers [see Fig. 10(a)]. Insertion loss (IL) measurement results of selected waveguide core sizes in dual layer construction are given in Fig. 10(b). Cross-talk and insertion loss for complex waveguide geometries including bends and over crossings have been reported earlier in [14]. The IL results normalized by sample length (dB/cm) are shown for 90 μm × 40 μm, 50 μm × 40 μm, 50 μm × 50 μm and 50 μm × 90 μm (width × height) waveguides. 90 μm × 40 μm channels with length 28.1 cm were characterized with both low mode fill (single mode fiber input, SMF) and with high mode fill (multimode fiber input, MMF) conditions to extract coupling loss with standard OM4 MMF 50 μm fiber. Difference in the high mode fill and low mode fil measurement result gave average coupling loss of 2.29 dB. Measurement results show that core size optimization to a specific channel termination (fiber type, diameter, NA, and coupling optics) can lead low loss system link loss with polymer waveguides in dual layer construction.

In all cases, the measurements were conducted at $\lambda = 850$ nm, output power captured by area photodetector and index fluid ($n = 1.47$) used at the input waveguide facet. To extract loss at the TX wavelength more samples were further characterized at 1000 nm. Average loss measured of 12 channels was 0.060 dB/cm and 0.075 dB/cm for 850 nm and 1000 nm respectively, indicating 25% higher loss at the wavelength of operation compared to datacom wavelength. These results are well in line with the spectral loss reported earlier for the same waveguide material [48].

C. Passive Optical Connector and Polymer Coupling Interfaces

In order to fully utilize the number of available channels and exploit the off-board interconnect capabilities of integrated O/E routing chips with high numbers of optical I/Os, appropriate passive coupling interfaces and pluggable connectors need to be developed. Passive parallel optical interfaces based on the MT standard can accommodate up to 6 rows of 12 optical channels per connector ferrule, whereby adjacent channels will have a center-to-center separation of 0.25 mm. MT ferrules are designed to house arrays of multimode or single mode optical fibers. In order to ensure that each connecting fiber pair in the connecting ferrules can make full physical contact with each other even when the connecting MT facets are not completely parallel, the fibers are arranged to protrude slightly out of the MT ferrule facet. MT ferrules are by far the most common parallel optical connector interface available. The MT ferrule has been manufactured by USConec [49] and the MLAs which can be connected onto the MT ferrule interface have been manufactured by the Japanese company Enplas [50]. The MT ferrules have been manufactured using a durable thermoplastic composite, Poly-phenylene Sulfide (PPS). The termination process involves multiple parallel fibers being aligned and secured to the ferrules with an optical connector grade thermal cure epoxy and polished with a variety of commercially available batch connector polishing machines in order to achieve the interface requirements set out in the IEC standard 61754-5. The fibers are arranged and polished such that the tip of each fiber should protrude by between 0.5 μm and 2.5 μm from the surface of the MT ferrule facet, depending on the quality required. A new generation of parallel optical connector was developed by USConec in 2013 in collaboration with Intel and Facebook as part of the Open Compute project [51] to address the problem of scaling such connectors into future mega Data Centers. The expanded beam PrizmMT ferrules incorporate microlens arrays into the fiber holding structure to ensure that, at the exposed connecting interfaces, the optical beam width is actually increased to about 3.5 times the size of the multimode fiber aperture, thus making it far less susceptible to contamination. The MXC connector, which formed a key part of the publicity drive surrounding the OpenCompute project houses a PrizmMT ferrule in a plastic shell and clip and is designed for host side access.

Moving to polymer coupling interfaces, a suite of receptacles to allow coupling of MT fiber interfaces to PCB embedded multimode polymer waveguides has been developed. Fig. 11(a) shows two waveguide coupling interfaces on an electro-optical PCB with embedded multimode polymer waveguides. One type of receptacle, allows in-plane fiber-to-waveguide coupling, whereby the optical axis of the connecting fiber will be co-linear with the axis of the embedded waveguide. The other receptacle types allow out-of-plane fiber-to-waveguide coupling, whereby the axis of the connecting fiber will be orthogonal to the waveguide axis. The receptacle of Fig. 11(b) includes a discrete micro-mirror system. This will allow MT ferrule-based connectors to plug to the top of the PCB and launch or receive light to and from the embedded waveguides. The receptacles are passively aligned and attached to the polymer waveguide interface using a proprietary assembly method [see Fig. 11(c)]. Fig. 11(d) shows a test board with generic waveguide coupling interfaces, designed to accommodate either in-plane or out-of-plane receptacles. An MTP fiber optic cable is attached to an out-of-plane receptacle and illuminates an embedded multimode polymer waveguide with visible 650 nm light.

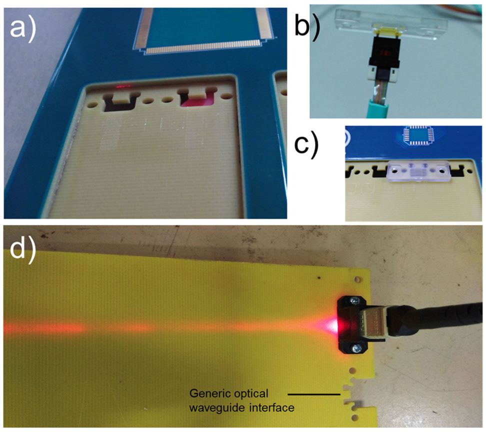


Fig. 11. (a) Electro-optical backplane with embedded waveguides, (b) out-of-plane receptacle connected to an MT ferrule, (c) Out-of-plane receptacle passively aligned onto optical waveguide interface, d) polymer waveguide test board with generic waveguide interfaces and waveguide illuminated with 650 nm light from out-of-plane test cable.

D. Fiber and Polymer Waveguide Flexplane Technologies

Following a realistic scenario that combines a dual-layer embedded polymer waveguide PCB with the Compass EOS router chip, we can only use the two outer rows of the router's 12 × 14 I/O optical matrix. In this arrangement, the first outer-row 48 peripheral IO pins connect to the first PCB waveguide layer and the second-periphery row 40 pins connect to the second waveguide layer. In order to fully exploit the whole 12 × 14 optical I/O matrix of the router without migrating to still immature deployments of multi-layer OPCBs with more than 2 waveguide layers, the electro-optical PCB should be replaced by flexplane technology. Fiber flexplanes are laminated fiber-optic circuits, in which optical fibers are pressed and glued into place on a substrate. These structures benefit from the reliability of conventional optical fiber technology and are currently promoted as the mature and low-cost alternative to EOPCBs, which have

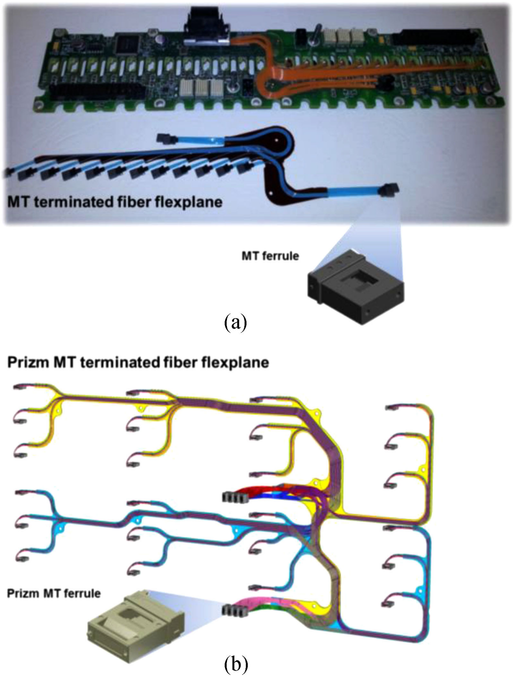

MT terminated fiber flexplane
MT ferrule
(a)
Prizm MT terminated fiber flexplane
Prizm MT ferrule
(b)

Fig. 12. Optical fiber flexplanes deployed within the PhoxTroT project: (a) Photo of electro-optical midplane with MT terminated flexplane, (b) Schematic view of Prizm MT terminated flexplane.

still to reach certain reliability and cost-efficiency metrics prior entering the market. However, unlike embedded optical waveguides, these circuits cannot accommodate waveguide crossings in the same layer i.e. fibers must cross over each other and cannot cross through each other. Moreover, each additional fiber layer necessitates typically the addition of backing substrates in order to hold the fibers in place, thus significantly increasing the thickness of the circuit. As such, flexplanes can be attached at best as separate entities onto the surface of a conventional PCB.

Fig. 12(a) shows a 196 fiber flexplane with MT ferrule terminations in an optically enabled data storage and switch test platform [52] for data centers. The average insertion of the flexplane alone is ∼0.32 dB and has been measured using an 850 nm VCSEL source from an Intel XFP transmitter. Fig. 12(b) depicts the design of a more complex 196 fiber flexplane with Prizm MT terminations, which will be more suitable for forced air environments in Data Centers.

III. Application-Oriented Interconnect Layout and Performance Analysis Design Tools

The deployment of on-board technology even with brilliant physical layer performance characteristics cannot ensure on its own an excellent performance at HPC-scale environments. With the network topology and bandwidth allocation between the nodes in a HPC comprising significant performance factors on top of the underlying technology, we demonstrate here for the first time a software engine that can incorporate optical device technology in a HPC network and produce the optimal network layout and its expected performance for a range of application workloads. The software tool comprises two main building blocks: a) the Automatic Topology Design Tool (ATDT), which is responsible for generating the optimal EOPCB topology, and b) the OptoHPC-Sim simulation platform, which adopts the PCB design provided by ATDT and evaluates throughput and latency over a wide range of application benchmarks. This synergy between ATDT and OptoHPC-Sim can yield valuable feedback on the technology development towards conforming to application-driven performance requirement, facilitating critical decisions such as the number of optical links finally required and the number of optoelectronic chips that need to be hosted on a EOPCB.

A. *Interconnect Layout: The Automatic Topology Design Tool*

The Automatic Topology Design Tool (*ATDT*) has been deployed as a software suite that aims to aid topology design for EOPCBs, making also sure that physical layer constraints related to power budget and available board area are satisfied [36]. The building blocks it takes into account are transceiver optochips, router chips and various polymer waveguide structures (straight waveguides, waveguide bends & waveguide crossings). Transceiver optochips are considered to be the active Tx/Rx interface modules connecting the electronic chips like processors to the EOPCB embedded optical waveguides. Following the example of optical I/O technology of the optoelectronic router chip, transceiver chips may rely on identical matrices as used in the router chips so as to ensure compatibility at all physical layer parameters between the processor-router communication link. The *ATDT* routing elements can be in general router chips with integrated optical I/Os that will be connected to transceiver chips. In this work, the *ATDT* router chip modules have been considered to rely on the board-adaptable version of the optoelectronic router chip described in Section II-A [26].

ATDT generates the optimal on-board topology within a specific set of topology families, which (a) satisfies given physical- and packaging-related parameters as well as performance requirements, (b) while taking into account that the EOPCBs are parts of a larger system. The traffic pattern assumed for evaluating the performance and concluding the optimal layout has been the Uniform Random Traffic (URT) profile, so as to produce a more general purpose network that doesn't match only to a specific workload problem-set. Performance in *ATDT* is estimated using analytical formulas to calculate throughput and average distance [36]. The set of topology families currently supported are meshes, tori and fully connected networks.

The main performance metrics used as optimal topology criteria in ATDT are *"network speedup"* and *"average distance"*. *Network speedup* is defined as "the ratio of the total input bandwidth of the network to the network's ideal *capacity"* [53]. *Network speedup* is unitless, and it is closely related to the ideal throughput. A network with *network speedup* equal to 1 allows 100% ideal throughput under URT. A network with *network speedup* greater than 1 allows non idealities in the implementation and ensures better performance under non Uniform traffic patterns. *Average distance* relates to the latency of the network, being an indicator for the expected packet latency when assuming light network load and uniform distribution of the traffic destinations.

TABLE I
ATDT INPUT PARAMETERS

router footprint: *52 x 52 (mm x mm)*	***90° bend. radius:*** *10 mm*
optochip footprint: *52x52 (mm x mm)*	***10mm bending radius loss:*** *1.2 dB*
VCSEL power: *3 dBm*	**Propagation Loss:** *0.005 dB/m*
PD sensitivity: *-10 dBm*	**90° crossing loss:** *0.023 dB*
Coupling loss (chip to waveguide and opposite): *3 dB*	

Network speedup is given as input by the user, while *average distance* can be used as an optimization criterion to solve ties. The user can set the desired *network speedup* value greater than 1 in order to relax the non-ideal assumptions and to derive topologies performing better under adversarial traffic patterns.

The physical implementation of a logical topology on optical boards employs various waveguide structures such as waveguide crossings with different crossing angles, waveguide bends, splitters and combiners. The feasibility of a physical implementation for a given topology for the on-board level of the packaging hierarchy is largely determined by its layout. The layout determines a) the worst case losses, i.e. the highest loss value among the losses experienced by all available optical paths, b) the layout area (height, width) as well as c) the volume (number of waveguide layers). A topology is considered as feasible only if its layout satisfies the given optical power budget as well as the board-area constraints. The ATDT input includes (i) footprint parameters: router and optochip footprints, (ii) chip-waveguide interface parameters: VCSEL transmission power, PD sensitivity, chip-to-waveguide coupling loss (and the opposite), and (iii) waveguide propagation parameters: 90° bends loss, propagation loss, 90° crossing loss. *ATDT* follows structured "circural manhattan" waveguide routing strategies, where all waveguide structures appear in a specific deterministic order, with half of the waveguide bends being followed by all the waveguide crossings which are followed by the remaining bends [36]. Due to the deterministic nature of the layout strategies, both layout area requirements and worst case losses can be estimated. Up to 2 waveguide layers have been assumed in the current version of the tool, so as to comply with the EOPCB technology developments described in Section II. However, the layout strategies can be easily extended for more than 2 optical layers. The input parameters of the ATDT used in our study are presented in Table I.

The *ATDT* operates in 2 phases, with its process flow being depicted in Fig. 13. During the first phase, all feasible designs for given physical layer (board size, module footprints and losses) and performance inputs (required *network speedup*, injected bandwidth from hosts, total system size)are generated. More specifically, the total number of on-board hosts/transceiver chips starts to increase gradually assuming also increasing number of on-board router chips. For every combination of hosts/transceiver and router chips, all feasible networks within the available topology families are generated. A design is said to be feasible when (i) the performance constraints in terms of *network speedup* are satisfied and (ii) there is at least one layout of the logical topology satisfying the on-board worst case losses and board area constraints. This strategy considers that off-board connectivity is not limited by optical power losses, since usually the signal has to undergo an opto-electro-optical conversion at the board-edge in order to leave the board through conventional active optical cable transceivers.

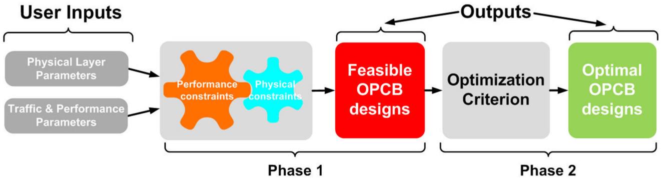


Fig. 13. ATDT process flow.

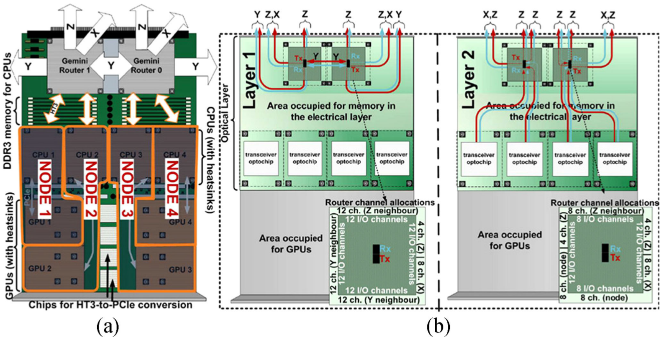


Fig. 14. 1 × 4 × 1 torus layout for a) a CRAY XK7 blade and b) a Dual Layer EOPCB.

The second phase considers then all feasible designs generated by the first phase and selects the optimal one, with the optimality criterion being the maximization of the number of the on-board transceiver optochips (hosts) while requiring the minimal number of router chips. Ties are solved by minimizing the *average distance*. Note that other optimization criteria can be also applied without having to re-execute phase 1.

In order to allow for a direct comparison between an HPC network architecture relying on application-driven optical blade technology with the CRAY XK7 systems employed in world's no. 3 supercomputer Titan, the topology type and size in *ATDT* for both the whole network and for the individual boards were kept constant and equivalent to CRAY XK7 systems. Fig. 14 shows the detailed layout of a single EOPCB, as this has been obtained by the *ATDT*. The EOPCB includes 4 sockets for hosting the transceiver optochips and 2 optoelectronic router chips along with the proper pin connections between them. Transceiver optochips serve as the interface between the CPU traffic generating modules, called *computing nodes*, and the board-level optical waveguides, while the optoelectronic router chip version is here shown to support 168 multi-mode optical I/Os, following the relevant layout of the commercially available chip offered by Compass EOS [26] and described in more detail in Section II.

B. From OptoBoard to HPC Systems: The OptoHPC Simulation Engine

The layout design through the ATDT tool ensures that throughput and latency values are optimized when using uniform random traffic profiles, however it doesn't provide any information about the network performance when different traffic profiles are employed, as is usually the case during workload execution in HPC environments. This would require the use of HPC network simulation engines, however state-of-the-art sophisticated HPC simulators still don't support the use of advanced electro-optic router and interconnect solutions at board-level. Among the few HPC open-source simulators that are free of charge and available to the research community, none of them is focused on or can even efficiently explore the adoption of optical technology advancements in the HPC field. The Extreme-scale Simulator (xSim) [34] implements a parallel discrete event HPC simulator but is mainly targeting the investigation of parallel applications' performance at extreme-scale Message Passing Interface (MPI) environments. SST + gem5 [35] is a scalable simulation infrastructure for HPCs and comes as the result of the integration of the highly detailed gem5 performance simulator into the parallel Structural Simulation Toolkit (SST). SST is a system of disparate hardware simulation component entities integrated via a simulator core, which provides essential services for interfacing, executing, synchronizing and monitoring the various components with gem5 [54] being integrated as one of them. However, gem5 gives emphasis in simulating detailed CPU-cores and computer memory hierarchies, yielding high simulation times due to its highly-detailed CMP hardware models.

This section describes a new simulation engine called OptoHPC-Sim, which exploits the ATDT outcome as input towards evaluating throughput and latency of the complete HPC network based on EOPCBs for a range of traffic profiles typically used for benchmarking in HPCs. The OptoHPC-Sim simulation platform comes as an extension of the OptoBoard Performance Analysis Simulator (OBPAS) simulator [55] towards supporting the use of electro-optical boards and routing technologies in complete and fully operational HPC network architectures. OptoHPC-Sim forms a powerful, modular and light-weight solution being implemented on top of the Omnet++ discrete event simulation framework [56]. It relies on a careful balance between the model detail and the simulation execution time, employing a queue-based HPC model and including only the absolutely necessary details for reliably evaluating an optically enabled HPC system. OptoHPC-Sim offers a user-friendly Graphical User Interface (GUI) that allows the detailed exploration of complete HPC topologies and can successfully be used for both demonstration and education purposes.

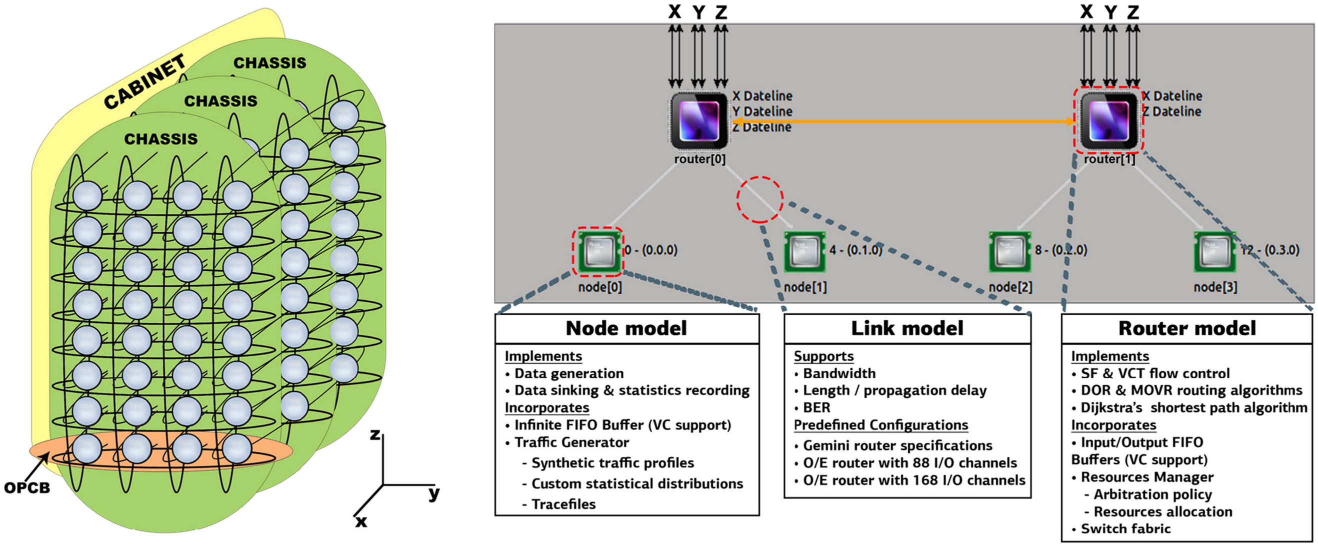
OptoHPC-Sim's GUI is presented in Fig. 1, where an example HPC network of 4 racks along with the internal rack architecture hierarchy is demonstrated. The same rack architecture is also employed in Titan CRAY XK7 supercomputer [57], where 8 CRAY XK7 Blades are grouped together forming a chassis and three chassis are grouped together forming an HPC rack. Depending on the size of network determined as the number of computing nodes, the number of racks may vary between 1–3 racks up to 49–320 racks. Building for example a class0 network of 96–288 computing nodes would require 1–3 racks organized in a single rack-row, while a class2 network of 1632–4608 nodes would require 17–48 racks organized in two rack-rows [58].

At the top of OptoHPC-Sim's GUI in Fig. 1, the main menu's toolbar allows the management of the simulation process providing the options for a step-by-step, fast and express simulation mode. Along with the main menu's toolbar simulation, kernel-statistics are reported including the simulation clock-time and the number of scheduled/executed events. At the left side of OptoHPC-Sim's GUI, the parameters explorer allows the exploration of the configurations regarding the current simulation setup. At the bottom of GUI, the event-list section informs the user for the executed events. Last but not least, the network explorer appears in the middle of GUI allowing the top-down exploration of the simulation model hierarchy by double-clicking to the individual modules.

OptoHPC-Sim currently supports both Mesh and Torus network topologies in up to 3-dimensional arrangements, as being widely used in many of the industry's HPC systems [57]. Fig. 15(a) presents an example topology of a single rack 3 D torus architecture where a total number of 24 PCBs are organized in groups of 8 PCBs, where each of the groups forms a chassis. Using the OptoHPC-Sim's GUI and moving down through the HPC hierarchy, we reach the PCB-layer view demonstrated in Fig 15(b). In this example, two router modules are connected together using a specifically configured instance of the link module, with each router being directly connected to two node modules by using again a specifically configured instance of the same link module. This specific OPCB model represents the ATDT-produced EOPCB layout depicted in Fig. 14.

Router model represents the router chips used in the HPC network and is responsible for all the routing decisions which are taken on a hop-by-hop basis. Router model comes with support for Dimension Order Routing (DOR) and Minimal Oblivious Valiant Routing (MOVR) algorithms that ensure deadlock free operation by eliminating any cyclic dependencies that could arise through the individual routing decisions [53]. During the OptoHPC-Sim's initialization stage, the router model is responsible for generating the routing-table structures that are necessary for taking the routing decisions. Routing tables are organized in rows where the number of rows is equal to the total number of routers in the network minus one since traffic should never be routed to the source router again. Each routing table row is organized in two columns, where the first column contains a unique router address and the second column contains a set of one or more possible output gates that should be followed in order to route any data destined to the router of the first column. The routing table generation is based on the Dijkstra's shortest paths algorithm ensuring minimal routing operation for both DOR and MOVR routing algorithms [53].

Router model comes with a set of three predefined configurations, where all the router network-level characteristics have been taken into account, such as the input and output port organization as well as their specific bandwidth specifications. The first configuration has been derived by considering the Gemini



(a)

(b)

Fig. 15. (a) A single rack 3 D-Torus topology example where a total number of 24 PCBs are organized in groups of 8 PCBs where each of the group forms a chassis. Each PCB incorporates 4 computing nodes, (b) *OptoHPC-Sim's* PCB-layer view where two *router* modules are connected together using the *link* module where each of them is directly connected to two *node* modules by using again another instance of the *link* module.

router's specifications, which is currently used in Titan Cray's XK7 blades. The other two configurations have been derived by considering the specifications of the first Optoelectronic (OE) Router that has recently entered the market [26]. Regarding the first OE Router configuration, named OE-Router-88ch, we consider a total number of 88 bi-directional Input/Output (IO) links where every link operates at 8Gbps. In this case, we follow a realistic scenario of using only the two outer rows of the router's 12 × 14 IO optical matrix over a dual-layer embedded polymer waveguide PCB. In this arrangement, the first outer-row 48 peripheral IO pins connect to the first PCB waveguide layer and the second-periphery row 40 pins connect to the second waveguide layer. In order to fully exploit the whole 12 × 14 optical I/O matrix of the router without migrating to still immature deployments of multi-layer OPCBs with more than 2 waveguide layers, we also consider the case where all 168 optical IOs are utilized by using a fiber-optic Flexplane technology (see Section II-D) for realizing the on-board interconnections. This OE Router configuration is named OE Router-168 ch.

Router model incorporates also the buffer, resourcesManager and switchFabric models that are necessary for the internal router organization but are not depicted in Fig. 15(b). Buffer model implements a basic First-In-First-Out (FIFO) policy and supports Virtual Channel (VC) organization, which ensures deadlock-free operation with regard to the wrap-around links existing in Torus networks. VC organization is also essential for MOVR routing algorithm in order to eliminate any cyclic dependences arising by the individual routing decisions [53]. The Buffer model can be used for modeling either an input- or an input-output-buffer router organization. ResourcesManager implements a FIFO arbitration policy with respect to the router's input buffers, while at the same time orchestrates the output ports resource allocation. ResourcesManager module is also responsible for driving the switchFabric module that forwards the input buffers transmitted data to the proper output ports.

Link model incorporates all the physical-layer relevant parameters, such as the link bandwidth, link length/propagation delay and Bit-Error-Rate (BER). The link module is utilized in all HPC network connections and not only at on-board level, as shown in the example of Fig. 15(b), using the corresponding parameters for every hierarchy level.

Node model simulates the HPC's computing nodes and is responsible for the per node traffic generation according to the applications running on the HPC and described later along with trafficPatternsManager. Node also sinks any incoming data updating at the same time the per node simulation statistics (global statistics management described later along with statisticsManager). Node model incorporates both the buffer and trafficGenerator models that are necessary for the internal node organization.

Buffer model is the same with the one incorporated in the router model, where in the case of node it is capable of simulating an infinite depth queue which separates the packet source (trafficGenerator) from the simulated network. It is important to note here that the traffic injection process is operated in lockstep with the rest of the network simulation, achieving in this way a bounded memory footprint even for network saturation conditions [53].

TrafficGenerator manages the actual traffic generation by generating and forwarding proper messages to the node's infinite buffer. Due to the fact that messages may be arbitrarily long, they are further divided into one or more packets that have a predefined maximum length. Each packet carries a segment of the message's payload and a packet header is always preceding. Considering the SF flow control mechanism, both the header and payload data are packed together into a single group of bits and are transmitted to node's buffer. When the Virtual Cut-Through (VCT) flow control mechanism is followed, the packet payload is further divided into zero or more body flits that are followed by a tail flit. In this case all the header, body flits and tail flit are individually transmitted to the node's buffer and subsequently to the entire network.

Three more auxiliary modules, namely networkAddressesManager, trafficPatternsManager and statisticsManager, have been incorporated to support the successful network initialization setup and the correct simulation operation process. All these three modules can be seen in the OptoHPC-Sim's GUI network explorer of Fig. 1 and are accessible directly below the four racks of the HPC network example.

NetworkAddressesManager is responsible for the network's addresses allocation along both the computing nodes and the routers. Two automatic address allocation schemes are supported with the first one following a sequential address allocation policy like in the case of Titan CRAY XK7 [57] and the second one following a random-uniform address allocation policy. If desired, custom address-allocation schemes can be fed to the simulator in the form of input text files. For all the cases each node is assigned both a decimal address and a location identifier in the form of X.Y.Z coordinates with regard to its absolute position in the Torus/Mesh grid. Taking as an example the second node of Fig. 15(b), its decimal address equals to 4 where its location identifier equals to 0.1.0. All addresses are unique and start counting from zero up to the number of nodes minus one. The same address allocation scheme is also applied to the router nodes. Finally, networkAddressesManager is responsible for defining the dateline routers, which are essential for ensuring deadlock free operation in the Torus topologies [53]. Considering the example of Fig. 15(b), the first router serves as dateline in all three X, Y and Z dimensions, while the second router serves as dateline only in X and Z dimensions.

TrafficPatternsManager's main responsibility is to define and manage the applications executing in the simulated system by means of traffic pattern distributions. OptoHPC-Sim currently supports 8 most well-known synthetic traffic patterns in the literature [53]: 1) Random Uniform, 2) Bit Complement, 3) Bit Reverse, 4) Bit Rotation, 5) Shuffle, 6) Transpose, 7) Tornado, and 8) Nearest Neighbor. Two more configuration options are additionally offered, where the simulator can be fed with either real-world packet traces or files describing the traffic pattern distribution among the computing nodes. On top of that, the user can choose between constant and exponential message inter-arrival times as well as constant and variable message size distributions.

StatisticsManager's role is to handle the global result collection during the simulation process and to record the results into proper output files when the simulation comes to an end. One of its most significant features is that it can detect whether a steady state has been reached through continuously monitoring the global network's performance metrics, informing the simulation kernel via a special termination signal that denotes that a steady state has been reached.

OptoHPC-Sim's configuration procedure can be easily handled by only a single configuration file, which specifies the network configuration parameters that must be taken into account.

TABLE II
ROUTER CONFIGURATIONS' IO CAPACITIES

Router Port Type	ConventionalRouter	OE-Router-88ch	OE-Router-168ch
Node-Router (Gbps)	83.2	64	120
X dimension* (Gbps)	75	64	120
Y dimension* (Gbps)	75 (Mezzanine) 37.5 (Cable)	96	192
Z dimension* (Gbps)	120 (Backplane) 75 (Cable)	128	240
Max Capacity (Tbps)	0.706	0.704	1.344

**per direction*

IV. EOPCB-BASED HPC NETWORK PERFORMANCE ANALYSIS AND COMPARISON WITH CRAY XK7 HPC

In this section we use the OptoHPC-Sim in order to evaluate and compare the performance of an HPC network that employs three different types of on-board routing: a) the OE-Router-88 ch, b) the OE-Router-168 ch and c) a Conventional Router model that complies with the Gemini router's specifications along with a purely electrical board layout, as is being used in the world's 3rd fastest supercomputer [38]. For the OE-Router models, router channel allocation has been realized in both OE-Router-88 ch and OE-Router-168 ch cases with the ATDT tool in order to offer optimum saturation throughput for the case of Uniform Random traffic pattern when considering optimal routing conditions. Table II summarizes the IO link capacities per dimension for the 2 OE-Router and the Conventional Router configurations, as well as the maximum router capacity for all the three cases. For the optimum channel allocation *network speedup* equal to 1 was assumed, leading to maximum injection bandwidth of 64 Gbps and 120 Gbps for the *OE-Router-88 ch* and *OE-Router-168 ch* cases.

In our analysis, we assume a 4 × 12 × 8 3 D Torus HPC network which can be classified as a class1 network incorporating a total number of 384 computing nodes [58]. The computing nodes are organized in a single rack-row of 4 racks, where each rack incorporates 3 chassis of 8 PCB Blades. Each PCB Blade incorporates 2 directly connected router modules, where each router module is directly connected to 2 computing nodes. A sequential address allocation policy is followed and we use all the eight synthetic traffic patterns presented in Section III-B. DOR has been employed in all cases as the routing algorithm, as it has been shown to outperform the MOVR algorithm in the Conventional Router–based network topology in terms of saturation throughput and for both the Uniform Random and Nearest Neighbor synthetic traffic patterns [37]. Regarding the Conventional Router configuration, the VCT flow control mechanism has been utilized complying with the respective mechanism of the Gemini router that is used in the Titan CRAY XK7 supercomputer [57]. In both cases of OE-Router-configurations, both

TABLE III
SIMULATION PARAMETERS

Parameter Name	Value
Network Size	4 x 12 x 8
Traffic patterns	Uniform Random & Nearest Neighbor
Message generation distribution	Exponential
Header Size (Bytes)	64
Packet Size* (Bytes)	1536
Router Buffer Size (KBytes)	250
Flow Control Mechanism	Store and Forward (SF), Virtual Cut Through (VCT)

Store-and-Forward (SF) and VCT flow control methods have been evaluated. The rest of the simulation parameters employed is being summarized in Table III.

Figs. 16 and 17 illustrate the simulation comparison results among all the three OE-Router-88 ch, OE-Router-168 ch and Conventional Router (termed as CRAY in the figure) cases and for all the 8 synthetic profiles supported by OptoHPC-sim. Fig. 16 presents the mean node throughput versus mean node offered load while Fig. 17 present the respective mean message delay versus mean node offered load considering all the messages exchanged among the computing nodes. As expected, for all throughput measurements and for both OE-Router cases, no variations between the SF and VCT flow control methods are observed irrespective of the traffic pattern applied.

Fig. 16 reveals that the use of Uniform Random pattern leads to the highest saturation throughput among all 8 traffic patterns for both OE-Router cases. This comes in agreement with the channel allocation and design strategy that were followed by the ATDT tool towards ensuring maximum throughput for Uniform Random patterns. However, given that ATDT considers optimal routing conditions that are certainly not met by realistic routing algorithm implementations like DOR and that the router channel allocation was obtained assuming *network speedup* equal to 1 (leaving no room for non-idealities), both OE-Router-based cases saturate below the 100% offered load that should be theoretically expected. On the other hand, although the maximum capacity of the Conventional Router is slightly higher compared to the OE-Router-88 ch, the Conventional Router CRAY system throughput saturates much earlier at ∼14.5 Gbps, being ∼3.3 times lower compared to the 48 Gbps saturation point of the OE-Router-88 ch. This particular observation reveals the important role of total router's bandwidth channel allocation strategy, highlighting the benefit of supporting the ATDT tool-enabled channel allocation strategy in the case of OE-Router-88 ch. The throughput performance is significantly improved in the case of the OE-Router-168 ch compared to the OE-Router-88 ch due to the 1.9× higher router capacity offered in this case.

Beyond the corresponding saturation points, a slight throughput drop for all the three router configurations is observed. This behavior stems from the channel arbitration unfairness introduced by the network routers with respect to the individual packet flows of Uniform Random pattern. In our scenarios, we employ a per router First-In-First-Out (FIFO) arbitration policy with respect to the desired output router port. Packets are grouped together according to the desired output port and are prioritized according to the absolute arrival time at the input ports of each individual router. This would eventually allow packets that require fewer hops and therefore fewer resource arbitrations to get a higher proportion of the available bandwidth, since no global routing criteria are taken into consideration. Hence, some flows may become starved and their throughput can drop dramatically as the load increases beyond saturation. Solutions like (a) the adoption of age-based arbitration criteria (e.g. # of hops) or (b) the implementation of non-interfering networks with one virtual channel per destination (unrealistic for big networks) are well-known in the literature for offering network stabilization beyond saturation point [53]. However, the implementation and analysis of such advanced solutions falls out of the scope of this analysis.

Proceeding to the remaining traffic patterns shown in Fig. 16, mean node throughput increases proportionally to the offered load until reaching the corresponding saturation points for all three router configurations, similarly to the case of Uniform Random. In the cases of Tornado's CRAY and OE-Router-88 ch and of Bit Complement's CRAY, saturation throughput is reached even from the first measurement at an offered load of 10 Gb/s. The OE-Router-88 ch configuration outperforms the CRAY system for all traffic patterns, with the only exception offered in the case of the Nearest Neighbor traffic profile. In the case of the Nearest Neighbor pattern, the CRAY-based network saturates at ∼36 Gbps, offering ∼14.6% better performance compared to the OE-Router-88 ch and confirming in this way that the Titan CRAY XK7 design favors the use of this specific traffic pattern.

For the rest of the patterns, the comparative analysis yields almost similar behavior as for the Uniform Random; although the total maximum capacity of the CRAY Conventional Router is slightly higher compared to the OE-Router-88 ch, the system throughput saturates much earlier resulting in significantly worse performance. In the case of the OE-Router-168 ch-based layout, the network throughput outperforms both the OE-Router-88 ch and the CRAY cases for all traffic patterns including the Nearest Neighbor, taking advantage of the highest router capacity employed in this network.

For the Nearest Neighbor and Bit Rotation cases, the network continues to deliver the peak throughput even after reaching the saturation point, designating the behavior of a stable network. For the Tornado and Bit Complement traffic patterns, the throughput drops beyond the corresponding saturation points, following a similar behavior as in the case of the Uniform Random pattern. Again, this stems from the channel arbitration unfairness introduced by the network routers with respect to the individual packet flows of each pattern. The significantly sharper drops experienced in these two patterns compared to the Uniform Random pattern indicate that the unfairness related to these patterns is much more severe than for the Uniform Random case.

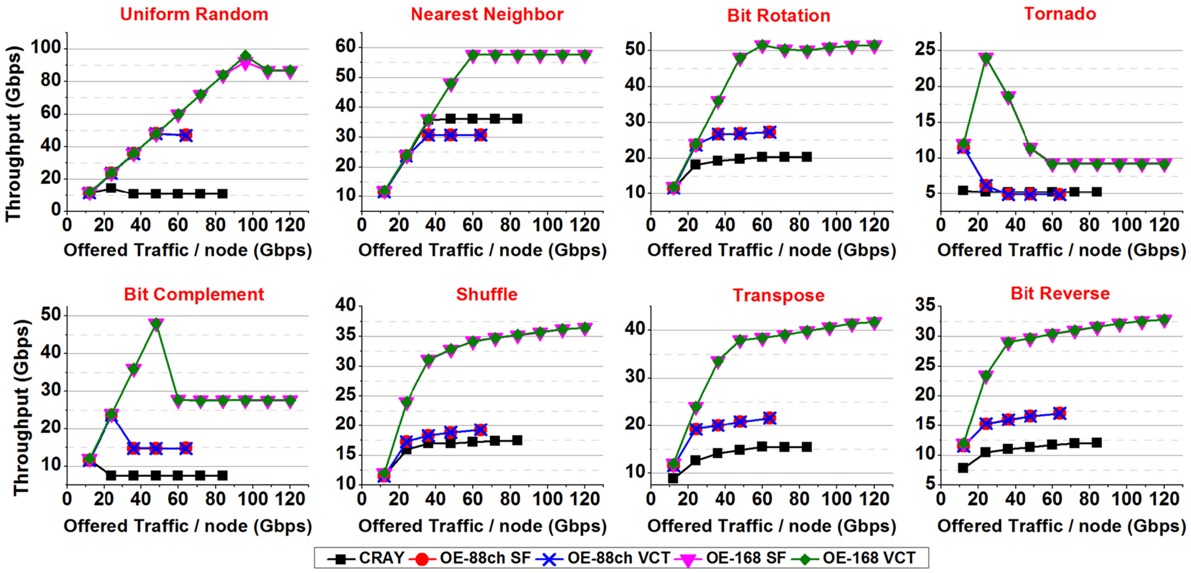


Fig. 16. Throughput simulation results for 8 synthetic traffic profiles.

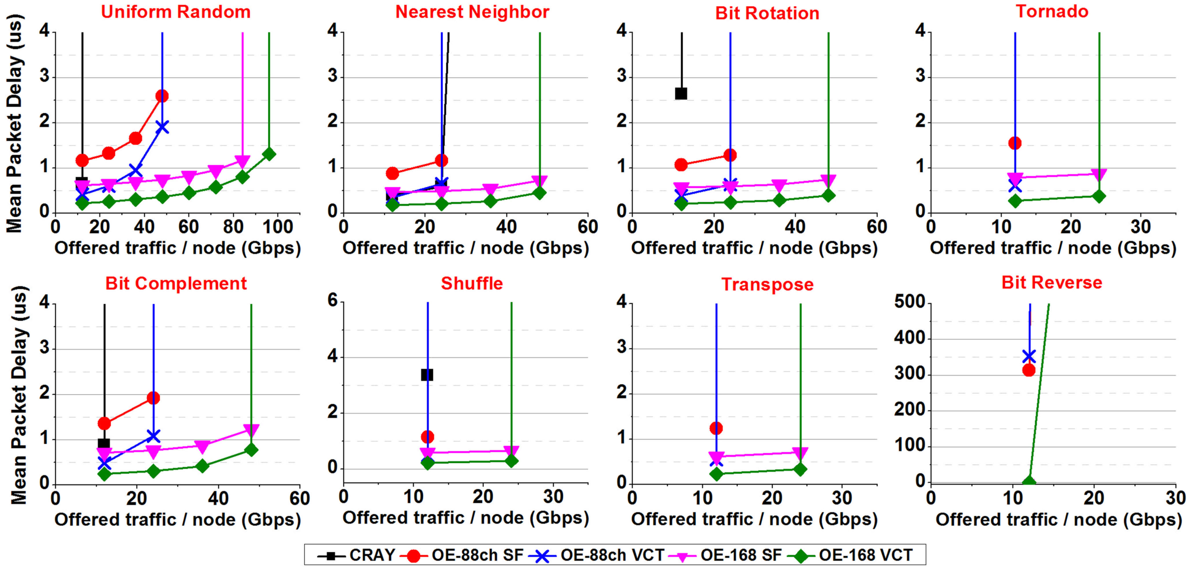


Fig. 17. Mean packet delay simulation results for 8 synthetic traffic profiles.

For the Shuffle, Transpose and Bit Reverse traffic profiles, the mean node throughput continues to increase even beyond the respective saturation points but at a significantly lower rate. This can be explained by the use of different link capacities in the different dimensions of the network. In multi-dimensional networks with different link capacities per dimension (see Table II), there may be some dimensions that get saturated earlier depending on the applied traffic pattern. As such, a portion of the traffic gets favored as this has only to travel through unsaturated areas of the network, resulting in a lower-rate throughput increase even beyond the saturation point.

Proceeding to the mean message delay measurements shown in Fig. 17 for all router configurations, the constant mean message delay for the Uniform Random case increases until becoming unbounded at the saturation point. However, before reaching the saturation point, the VCT flow control method performs better offering lower mean message delay values compared to the SF for every individual OE-Router configuration, being fully in agreement with respective theoretical expectations [53]. In both the VCT and SF flow control methods, the OE-Router-168 ch system outperforms the respective OE-Router-88 ch case taking advantage of its 1.9 × higher capacity value. Finally, all OE-Router cases outperform the respective CRAY system, which leads to unbounded delay values even from the second measurement at a 20 Gb/s offered load.

Similar behavior is witnessed for the mean message delay performance of the network for all traffic patterns shown in Fig. 17, with the Nearest Neighbor forming again the sole exception as

TABLE IV
SIMULATION RESULTS

	CRAY		OE-Router-88ch-VCT		OE-Router-168ch-VCT	
Traffic profile	**Throughput (Gb/s)**	**Delay (us)**	**Throughput (Gb/s)**	**Delay (us)**	**Throughput (Gb/s)**	**Delay (us)**
Uniform Random	14.28	0.66	48 *(+236.13%)*	0.62 *(-6.06%)*	92 *(+544.25%)*	0.22 *(-66.6%)*
Nearest Neighbor	20.2	0.58	27.2 *(+34.65%)*	0.49 *(-15.51%)*	51.46 *(+169.6%)*	0.21 *(-63.79%)*
Bit Rotation	11.7	2.64	23.67 *(+102.3%)*	0.57 *(-78.40%)*	48 *(+310%)*	0.20 *(-92.42%)*
Tornado	12	30050 ***	17 *(+41.6%)*	0.78 (-∞)	32.8 *(+173.3%)*	0.27 (-∞)
Bit Complement	17.4	0.88	19.25 *(+10.63%)*	0.70 *(-20.45%)*	36.43 *(+109.36%)*	0.24 *(-72.72%)*
Shuffle	5.23	3.37	11.51 *(+120%)*	0.58 *(-82.78%)*	24 *(+358.9%)*	0.22 *(-93.47%)*
Transpose	15.45	18942 ***	21.63 *(+40%)*	0.61 (-∞)	41.76 *(+170.29%)*	0.23 (-∞)
Bit Reverse	36	17703 ***	30.7 *(-14.7%)*	0.57 (-∞)	57.6 *(+60%)*	0.23 (-∞)
MEAN:	~16.5	~1.35	~24.9 *(+50.9%)*	~0.6 *(-54.8%)*	~48 *(+190.9%)*	~0.2 *(-83.7%)*

****not taken into account for the MEAN calculation*

the CRAY system offers in this case again lower delay values compared to the OE-Router-88 ch system. Table IV provides a summary of the results for both the throughput and delay values and for all available traffic patterns and router configurations. Performance of the CRAY system is illustrated in 2 columns; one presenting the mean node throughput in Gb/s, and the other presenting the mean packet delay in us. The corresponding columns for the OE-Router-88 ch and OE-Router-168 ch systems include, apart from their individual throughput and delay metrics, the difference as percentage compared with the respective CRAY performance. For all three configurations the reference for throughput metrics is considered the saturation point of the CRAY system. Regarding the delay metrics, performance of the CRAY system just before the saturation point is considered as the reference except for Tornado, Transpose and Bit Reverse patterns where the CRAY system becomes saturated before the measurement of 10 Gb/s injection bandwidth and consequently this first point is considered as reference. Important to note is that the OE-Router-88 ch system provides on average a 50% higher throughput value and a 54% lower delay value compared to CRAY despite the router module has a slightly lower capacity than the Gemini router employed in the CRAY XK7 configuration. The OE-Router-168 ch system, when compared to the CRAY system provides even more significant performance improvements, yielding almost 190% higher throughput and 83% lower delay which comes from the combination of the larger bandwidth offered by the optical technology and the optimization of the router channel allocation realized with the ATDT tool.

V. CONCLUSION

We have presented for the first time, to the best of our knowledge, an application-driven electro-optical on-board technology design and development framework for yielding optimized HPC throughput and delay values at system-scale level. We have demonstrated the recent technological advances achieved within the FP7 research project PhoxTroT towards implementing high-density and multi-layered Electro-optical Printed Circuit Boards (EOPCBs) with on-board optoelectronic routing along with a complete optically enabled ecosystem featuring HPC hardware, architectures and software tools that tailor EOPCB design to optimized HPC performance. The software tools allow the design and utilization of optical interconnect and electro-optical routing technologies at system-scale, offering at the same time complete end-to-end simulation of HPC-systems and allowing for reliable comparison with existing HPC platforms. The comparison analysis between an HPC network system employing state-of-the-art optoelectronic routers and optical interconnects with a system following the Cray XK7 system platform specifications reveals the benefits that can be gained by incorporating these technology advancements to future HPC networks in terms of both throughput and mean message delay. The proposed OptoHPC-Sim simulation engine has all the credentials for being enriched with energy consumption performance analysis and with real HPC application workloads, which comprise the goals of our future work.

REFERENCES

[1] M. Taubenblatt, "Optical interconnects for high-performance computing," *J. Lightw. Technol.*, vol. 30, no. 4, pp. 448–457, Feb. 2012.

[2] [Online]. Available: http://www.te.com/content/dam/te-com/custom/styles/everyconnectioncounts/assets/docs/data-communications-coolbit-te-25g-active-optics-white-paper.pdf.

[3] [Online]. Available: https://www.pdf-archive.com/2015/02/08/05-tec-product-description-english/05-tecproduct-description-english.pdf.

[4] [Online]. Available: http://www.intel.com/content/www/us/en/architecture-and-technology/silicon-photonics/optical-transceiver-100g-psm4-qsf-p28-brief.html.

[5] [Online]. Available: http://www.luxtera.com/luxtera/Luxtera_1M_final_4.pdf.

[6] [Online]. Available: http://www.luxtera.com/luxtera/products.

[7] H. Schröder *et al.*, "Electro-optical backplane demonstrator with integrated multimode gradient-index thin glass waveguide panel," *Proc. SPIE*, vol. 9368, 2015, Art. no. 93680U.

[8] Y. Matsuoka *et al.*, "20-Gb/s/ch high-speed low-power 1-Tb/s multilayer optical printed circuit board with lens-integrated optical devices and CMOS IC," *IEEE Photon. Technol. Lett.*, vol. 23, no. 18, pp. 1352–1354, Sep. 2011.

[9] F. E. Doany *et al.*, "Terabit/s-class optical PCB links incorporating 360-Gb/s bidirectional 850 nm parallel optical transceivers," *J. Lightw. Technol.*, vol. 30, no. 4, pp. 560–571, Feb. 2012.

[10] J. Beals *et al.*, "Terabit capacity passive polymer optical backplane," in *Proc. 2008 Conf. Lasers Electro-Opt. 2008 Conf. Quantum Electron. Laser Sci.*, 2008, pp. 1–2.

[11] N. Bamiedakis *et al.*, "40 Gb/s Data transmission over a 1 m long multimode polymer spiral waveguide for board-level optical interconnects," *J. Lightw. Technol.*, vol. 33, no. 4, pp. 882–888, Feb. 2015.

[12] T. Ishigure *et al.*, "Low-loss design and fabrication of multimode polymer optical waveguide circuit with crossings for high-density optical PCB," in *Proc. 63rd Electron. Compon. Technol. Conf.*, 2013, pp. 297–304.

[13] K. Schmidke *et al.*, "960Gb/s optical backplane ecosystem using embedded polymer waveguides and demonstration in a 12G SAS storage array," *J. Lightw. Technol.*, vol. 31, no. 24, pp. 3970–3974, Dec. 2013.

[14] M. Immonen, J. Wu, H. J. Yan, L. X. Zhu, P. Chen, and T. Rapala-Virtanen, "Electro-optical backplane demonstrator with multimode polymer waveguides for board-to-board interconnects," in *Proc. Electron. Syst.-Integr. Technol. Conf.*, 2014, pp. 1–6.

[15] H. Schroder *et al.*, "Thin glass based electro-optical circuit board (EOCB)," in *Proc. IEEE CPMT Symp. Jpn.*, 2015, pp. 109–117.

[16] L. Brusberg, S. Whalley, C. Herbst, and H. Schröder, "Display glass Forlow-loss and high-density optical interconnects in electro-optical circuit boards with eight optical layers," *Opt. Express*, vol. 23, pp. 32528–32540, 2015.

[17] I. Papakonstantinou, D. R. Selviah, K. Wang, R. A. Pitwon, K. Hopkins, and D. Milward, "Optical 8-channel, 10 Gb/s MT pluggable connector alignment technology for precision coupling of laser and photodiode arrays to polymer waveguide arrays for optical board-to-board interconnects," in *Proc. 58th Electron. Compon. Technol. Conf.*, 2008, pp. 1769–1775

[18] R. Pitwon *et al.*, "FirstLight: Pluggable optical interconnect technologies for polymeric electro-optical printed circuit boards in data centers," *J. Lightw. Technol.*, vol. 30, no. 21, pp. 3316–3329, Nov. 2012

[19] A. Sugama, K. Kawaguchi, M. Nishizawa, H. Muranaka, and Y. Arakawa, "Development of high-density single-mode polymer waveguides with low crosstalk for chip-to-chip optical interconnection," *Opt. Express*, vol. 21, pp. 24231–24239, 2013.

[20] F. R. Libsch *et al.*, "MCM LGA package with optical I/O passively aligned to dual layer polymer waveguide in PCB," in *Proc. 56th Electron. Compon. Technol. Conf.*, 2006, pp. 1693–1699.

[21] M. Shishikura, Y. Matsuoka, T. Shibata, and A. Takahashi, "A high coupling efficiency multilayer optical printed wiring board with a cube core structure for high-density optical interconnections," in *Proc. Electron. Compon. Technol. Conf.*, 2007, pp. 1275–1280.

[22] C. D. Martino *et al.*, "Lessons learned from the analysis of system failures at petascale: The case of blue waters," in *Proc. 44th Annu. IEEE/IEIP Int. Conf. Dependable Syst. Netw.*, 2014, pp. 610–621.

[23] L. Schares *et al.*, "Terabus: Terabit/second-class card-leveloptical interconnect technologies," *IEEE J. Sel. Topics Quantum Electron.*, vol. 12, no. 5, pp. 1032–1044, Sep 2006.

[24] T. Takemoto *et al.*, "A 25-Gb/s 2.2-W 65-nm CMOS optical transceiver using a power-supply-variation-tolerant analog front end and data-format conversion," *IEEE J. Solid-State Circuits*, vol. 49, no. 2, pp. 471–485, Feb. 2014.

[25] R. Dangel *et al.*, "Polymer-waveguide-based board-level optical interconnect technology for datacom applications," *IEEE Trans. Adv. Packag.*, vol. 31, no. 4, pp. 759–767, Nov. 2008.

[26] K. Hasharoni *et al.*, "A high end routing platform for core and edge applicationsbased on chip to chip optical interconnect," in *Proc. Opt. Fiber Commun. Conf. Expo./Nat. Fiber Opt. Eng. Conf.*, 2013, pp. 1–3.

[27] B. G. Lee *et al.*, "Monolithic silicon integration of scaled photonic switch fabrics, CMOS logic, and device driver circuits," *J. Lightw. Technol.*, vol. 32, no. 4, pp. 743–751, Feb. 2014.

[28] Polatis.com, "SERIES 7000 - 384x384 port software-defined optical circuit switch," 2016. [Online]. Available: http://www.polatis.com/series-7000-384x384-port-software-controlled-optical-circuit-switch-sdn-enabled.asp. Accessed on: Sep. 10, 2016.

[29] Y. Yan *et al.*, "All-optical programmable disaggregated data centre network realized by FPGA-based switch and interface card," *J. Lightw. Technol.*, vol. 34, no. 8, pp. 1925–1932, Apr. 2016.

[30] Broadcom.com, "High-density 25/100 Gigabit Ethernet StrataXGS Tomahawk ethernet switch series," 2014. [Online]. Available: https://www.broadcom.com/products/Switching/Data-Center/BCM56960-Series. Accessed on: Sep. 10, 2016.

[31] Mellanox.com, "Mellanox Spectrum Ethernet Switch," 2016. [Online]. Available: http://www.mellanox.com/page/products_dyn?product_family=218&mtag=spectrum_ic. Accessed on: Sep. 10, 2016.

[32] J. Chan *et al.*, "Phoenixsim: A simulator for physical-layer analysis of chip-scale photonic interconnection networks," in *Proc. Conf. Des., Autom. Test Eur.*, 2010, pp. 691–696.

[33] D. Vantrease *et al.*, "Corona: System implications of emerging nanophotonic technology," in *Proc. Int. Symp. Comput. Archit.*, Jun. 2008, pp. 153–164.

[34] S. Bohm and C. Engelmann, "xSim: The extreme-scale simulator," in *Proc. Int. Conf. High Perform. Comput. Simul.*, 2011, pp. 280–286.

[35] M. Hsieh, K. Pedretti, J. Meng, A. Coskun, M. Levenhagen, and A. Rodrigues, "Sst + gem5 = A scalable simulation infrastructure for high performance computing," in *Proc. 5th Int. ICST Conf. Simul. Tools Tech.*, 2012, pp. 196–201.

[36] A. Siokis, K. Christodoulopoulos, and E. Varvarigos, "Laying out interconnects on optical printed circuit boards," in *Proc. ACM/IEEE Symp. Archit. Netw. Commun. Syst.*, 2014, pp. 101–112.

[37] N. Terzenidis, P. Maniotis, and N. Pleros, "Bringing OptoBoards to HPC-scale environments: An OptoHPC simulation engine," in *Proc. 1st Int. Workshop Adv. Interconnect Solutions Technol. Emerging Comput. Syst.*, 2016, Art. no. 6.

[38] Top500.org, "Top 500 supercomputers' list of June 2016," 2016. [Online]. Available: http://www.top500.org. Accessed on: Sep. 10, 2016.

[39] M. Rittera, Y. Vlasova, J. Kasha, and A. Benner, "Optical technologies for data communication in large parallel systems," *J. Instrum.*, vol. 6, Jan. 2011, Art. no. C01012.

[40] A. Mamman, S. Stepanov, and K. Hasharoni, "On-chip optics for data center interconnects. Prospects and limitations," in *Proc. 2016 IEEE CPMT Symp. Jpn.*, 2016, pp. 83–86.

[41] S. Benjamin *et al.*, "336-channel electro-optical interconnect: Underfill process improvement, fiber bundle and reliability results," in *Proc. IEEE 64th Electron. Compon. Technol. Conf.*, May 2014, pp. 1021–1027.

[42] T. Mikawa, "Low-cost, high density optical parallel link modules and optical backplane for the last 1 meter regime applications," *Proc. SPIE*, vol. 6899, pp. 689902-1–689902-10, 2008.

[43] R. T. Chen *et al.*, "Fully embedded board-level guided-wave optoelectronic interconnects," *Proc. IEEE*, vol. 88, no. 6, pp. 780–793, Jun. 2000.

[44] C.C. Chang *et al.*, "Fabrication of fully embedded board-level optical interconnects and optoelectronic printed circuit boards," in *Proc. 11th Electron. Packag. Technol. Conf.*, 2009, pp. 973–976.

[45] R. C. A. Pitwon *et al.*, "Pluggable electro-optical circuit board interconnect based on embedded graded-index planar glass waveguides," *J. Lightw. Technol.*, vol. 33, no. 4, pp. 741–754, Feb. 2015.

[46] C. L. Schow *et al.*, "160-Gb/s, 16-channel full-duplex, single-chip CMOS optical transceiver," in *Proc. 2007 Conf. Opt. Fiber Commun. Nat. Fiber Opt. Eng. Conf.*, 2007, Paper OThG4.

[47] M. Immonen, J. Wu, H. J. Yan, L. X. Zhu, P. Chen, T. Rapala-Virtanen, "Development of electro-optical PCBs with embedded waveguides for data center and high performance computing applications," *Proc. SPIE*, vol. 8991, 2014, Art. no. 899113.

[48] N. Bamiedakis, J. Beals, R. V. Penty, I. H. White, J. V. DeGroot, and T. V. Clapp, "Cost-effective multimode polymer waveguides for high-speed on-board optical interconnects," *IEEE J. Quantum. Electron.*, vol. 45, no. 4, pp. 415–424, Apr. 2009.

[49] [Online]. Available: http://www.usconec.com/.

[50] [Online]. Available: http://www.enplas.co.jp.

[51] Intel Corporation, "Design guide for photonic architecture," 2013. [Online]. Available: http://www.opencompute.org/wp/wp-content/uploads/2013/01/Open_Compute_Project_Open_Rack_Optical_Interconnect_Design_Guide_v0.5.pdf. Accessed on: Sep. 10, 2016.

[52] R. Pitwon, A. Worrall, P. Stevens, A. Miller, K. Wang, and K. Schmidtke, "Demonstration of fully-enabled data centre subsystem with embedded optical interconnect," *Proc. SPIE*, vol. 8991, 2014, Art. no. 899110.

[53] W. J. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. San Fransisco, CA, USA: Morgan Kaufmann, 2004, pp. 233–247.

[54] N. Binkert *et al.*, "The gem5 simulator," *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, May 2011.

[55] S. Markou *et al.*, "Performance analysis and layout design of optical blades for HPCs using the OptoBoard-Sim simulator," in *Proc. Opt. Interconnects Conf.*, 2015, pp. 4–5.

[56] A. Varga, "The OMNeT++ discrete event simulation system," in *Proc. Eur. Simul. Multiconf.*, 2001, Art. no. 60.

[57] M. Ezell, "Understanding the impact of interconnect failures on system operation," in *Proc. Cray User Group Conf.*, May 2013. [Online]. Available: https://www.osti.gov/scitech/biblio/1086655

[58] Cray, "Cray XT and Cray XE system overview," [Online]. Available: https://www.nersc.gov/assets/NUG-Meetings/NERSCSystem-Overview.pdf. Accessed on: Sep. 10, 2016.

Authors' biographies not available at the time of publication.