Performance Analysis and Layout Design of Optical Blades for HPCs using the *OptoBoard-Sim* simulator

S. Markou[1], A. Siokis[2], P. Maniotis[1], K. Christodoulopoulos[2], E. Varvarigos[2] and N. Pleros[1]
1: Department of Informatics, Aristotle University of Thessaloniki, Greece and Information Technologies Institute, Center for Research and Technology Hellas, Thessaloniki, Greece
2: Computer Engineering and Informatics Dept., University of Patras, and Computer Technology Institute and Press-Diophantus, Patra, Greece
e-mail address:{samarkou, ppmaniot, npleros}@csd.auth.gr, {siokis, kchristodou, manos}@ceid.upatras.gr

Abstract: We demonstrate the *Optical Board Simulator* platform for optical PCB layout design and performance evaluation. Performance of two optical Blades is compared to CRAY-XK7 Blade for the FFTW benchmark, revealing significant throughput and latency improvements.
OCIS codes: (200.4650)Optical interconnects;(060.4258)Networks, network topology;(060.4256)Networks, network optimization

1. Introduction

To sustain performance advances towards Exascale, modern HPC systems rely on increasing the density of compute nodes and integrating multicore chips [1], putting more pressure on the system interconnect within HPCs. To overcome this drawback electrical interconnects are already being replaced by Active Optical Cables (AOCs) in rack-to-rack communication, while mid-board optical subassemblies and compact board-level flexible modules (FlexPlane) [2] have already entered the market. Optical Printed Circuit Boards (OPCBs) are researched towards replacing electrical PCB interconnects and low-loss embedded polymer waveguide interconnects have been demonstrated to yield Tb/s on-board transmission capabilities [3]. In all this deployment, communication is realized with optical I/O's to an ASIC electronic router chip [2] that can even reach 168 bi-directional links [4]. However, all these impressive technology advances can be translated into system-scale benefits only when evaluated through a design and performance toolkit, similar with the simulation platform for optical networks on chip presented in [5]. Thus we demonstrate for the first time to our knowledge a holistic design and simulation engine, named *Optical Board-Simulator (OptoBoard-Sim)*, that comes to bridge the gap between physical-layer PCB technologies and their system-level enabling capabilities. The *OptoBoard-Sim* takes into account all necessary physical layer and OPCB geometrical specifications and automatically produces all possible OPCB and physically feasible interconnect layouts (within certain topology families), selecting finally the optimally performing on-OPCB topology, which is subsequently evaluated with respect to throughput and latency values with a variety of real HPC traffic profiles.

2. Producing Optical PCB layouts with *OptoBoard-Sim*

OptoBoard-Sim is composed of two subsystems: (a) the *Automatic Topology Design Tool* (*ATDT*) (described in [6]),, which is responsible for providing the optimum OPCB interconnect layout for the followed layout strategy (see [6]), and (b) the *OptoBoard Performance Analysis Simulator* (*OBPAS*), being responsible for evaluating *ATDT's* selected topologies with real application profiles. *OBPAS* is implemented on top of Omnet++ to model the on-board processors/router chips and optical links. Processors are emulated by means of traffic generation modules that can produce either synthetic or realistic traffic profiles. The latter are trace files extracted after executing a number of HPC applications on HellasGrid cluster HG-06-EKT and profiled using IPM tool. The router model follows a queue-based configuration and the optical link model incorporates all network-level relevant parameter originating from *ATDT*, like link bandwidth and link latency. By modifying the router and link model parameters, *OBPAS* allows also for the performance analysis of electrical PCB layouts, rendering a simulation engine credible for comparison between electrical versus OPCB-based interconnects.

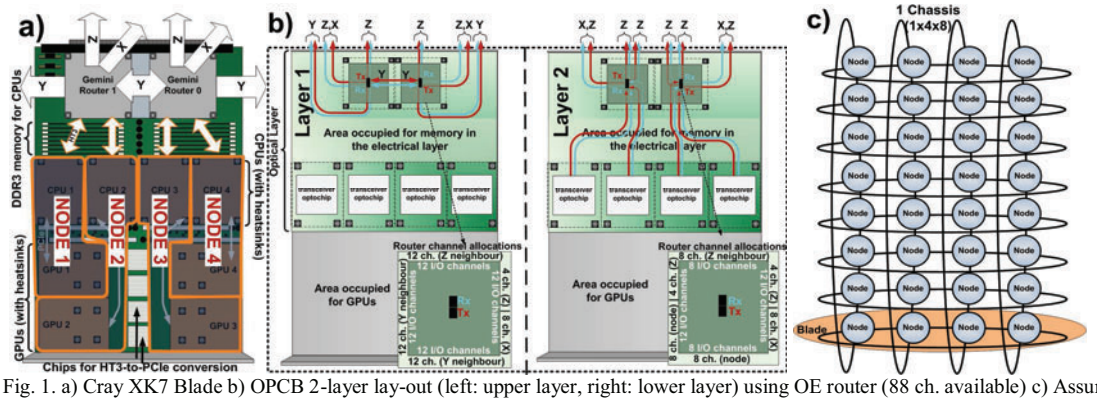

Fig. 1. a) Cray XK7 Blade b) OPCB 2-layer lay-out (left: upper layer, right: lower layer) using OE router (88 ch. available) c) Assumed placement of examined board in the 1x4x8 chassis (for simulations).

3. Optical Blade versus CRAY XK7 card performance analysis

To validate the *OptoBoard-Sim* platform we compared the performance of an OPCB interconnect with state-of-the-art specifications with that of the CRAY XK7 Blade (Fig. 1(a)) employed in world's no. 2 Supercomputer [7]. A compute node in

Cray XK7 is composed of a 16-Core AMD Opteron CPU communicating with an NVIDIA Tesla "K20x" GPU over PCIe, while CPU communicates directly with a Gemini router chip, using the HT3 protocol. Nodes in Cray XK7 are interconnected in a 3D torus topology and a single PCB accommodates a 1x4x1 sub-network. Fig. 1(b) shows the respective OPCB layout produced by the *OptoBoard-Sim* when considering an optoelectronic (O/E) router with a total number of 88 bi-directional I/O links with every link operating at 8Gbps, following a realistic scenario of using only the two outer I/O rows of an O/E router with 12x14 optical I/O's matrix [4] over a dual-layer embedded polymer waveguide PCB [2]. In this arrangement, the first outer-row 48 peripheral I/O pins of the O/E router's I/O matrix connect to the first PCB waveguide layer and the second-periphery row 40 I/O pins connect to the second waveguide layer. We also evaluate the performance of the optical Blade in case all available 12x14 optical I/O's of the O/E router [4] are utilized, assuming a fiber-optic Flexplane technology for the off-board interconnection. In both cases, the compute nodes were assumed to be optically interfaced and connected to the O/E router optical I/O ports over the OPCB. In our experiments we assumed different compute node-to-router capacities, indicating the use of computing nodes with stronger processing capabilities compared to the CRAY XK7 nodes, if these could be supported by the OPCB infrastructure.

In all examined cases the board is connected through the off-board links to the 3D torus system, the size of which is 1x4x8 in case of a single rackmount chassis with 8 PCBs, shown in Fig. 1(c), and 4x12x8 in case of 4 interconnected racks with a total number of 96 PCBs. The various physical layer parameters used for the case of the optical Blade analysis are depicted in Table I [6]. In all cases, performance analysis was focused on a single PCB, with the different X, Y, Z channel rates and the location of the particular blade as part of the larger system configurations taken also into account [8]. Table II summarizes the I/O link capacities per direction for the 2 OPCB scenarios (resulted by using the topology optimization logic of the *ATDT*) which are also the simulation parameters used for the performance evaluation of the topology using *OBPAS*.

Table I [6]

router footprint: *52 x 52 (mm x mm)*	***90° bend. radius:*** *10 mm*
optochip fooprint: *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*

Table II

	88 chan.	**168 channels**
CPU-router	*8 ch.*	*15 ch.*
X direction	*8 ch.*	*15 ch.*
Y direction	*12 ch.*	*24 ch.*
Z direction	*16 ch.*	*30 ch.*

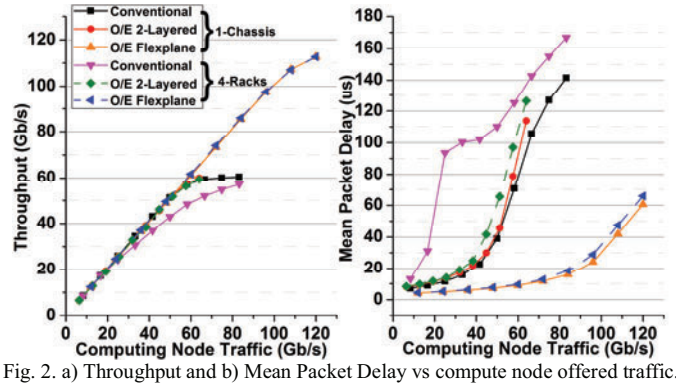

Fig. 2. a) Throughput and b) Mean Packet Delay vs compute node offered traffic.

Fig. 2 presents the performance comparison of the two optical blades with CRAY-XK7 blade for the case of the FFTW traffic profile [9]. Fig. 2(a) illustrates the *received traffic* (Throughput) vs *offered traffic* for a single on-PCB compute node (note that throughput was similar for all 4 on-PCB nodes), while Fig. 2(b) shows the mean packet delay. *Mean node throughput* was defined as the number of packets that were generated by a node and successfully reached their destinations divided by the number of generated packets , while *Mean packet delay* was calculated over all packets generated in the network. As shown in Fig. 2(a), the maximum offered traffic (100%) from a single compute node to the respective routing element is 83.5 Gbps, 64 Gbps and 120 Gbps for the cases of the CRAY XK7 Blade with the Gemini interconnect and for the optical Blade with a 88-channel (2-layered OPCB) O/E router and a 168-channel (Flexplane) O/E router, respectively. The maximum throughput of the original CRAY XK7 system is 60.5 Gbps for the 1 chassis system configuration and slightly lower for the 4-rack system. The maximum throughput of the OPCB equipped with the 88-channel O/E router is marginally higher than that of the original CRAY XK7 system, since the 88-channel O/E router capacity is 704 Gbps which is slightly higher than the Gemini router chip capacity. By increasing the number of optical links to 168 the optical Blade leads to significantly higher throughput of 112.5 Gbps when operating under 100% offered load, i.e. The mean packet delay was slightly worse to CRAY for the 88-channel O/E router while it is much lower for the 168 optical links case. The comparison of two optical Blades with CRAY-XK7 Blade showed a throughput improvement up to 27% and a mean packet delay reduction up to 60.5%.

Acknowledgement

This work was supported by the EC FP7-ICT project PhoxTrot (contract number 318240)

References

[1] J. Feehrer et al., "The Oracle Sparc T5 16-Core Processor Scales to Eight Sockets," IEEE Micro, Vol. 33, no. 2, pp. 48-57, 2013
[2] M.A. Taubenblatt, "Optical Interconnects for High-Performance Computing," Journal of Lightwave Technology, 30(4), pp.448,457, 2012.
[3] Y.Matsuoka, et. al., "20-Gb/s/ch High-Speed Low-Power 1-Tb/s Multilayer Optical Printed Circuit Board With Lens-Intergrated Optical Devices and CMOS IC," IEEE Photon. Technol. Lett., 23(18), pp. 1352 - 1354, 2011
[4] K. Hasharoni et al, "A High End Routing Platform for Core and Edge Applications Based on Chip To Chip Optical Interconnect", OFC 2013.
[5] J. Chan, et. al., "PhoenixSim: A simulator for physical-layer analysis of chip-scale photonic interconnection networks," DATE, 2010
[6] A. Siokis, et. al. "Laying out Interconnects on Optical Printed Circuit Boards", ANCS 2014, 2014.
[7] Top 500 Supercomputers' list of November 2013 (http://www.top500.org)
[8] Whitepaper: "The Gemini Network", Rev. 1.1, 2010 (https://wiki.ci.uchicago.edu/pub/Beagle/SystemSpecs/Gemini_whitepaper.pdf)
[9] M. Frigo and S. G. Johnson, "The design and implementation of FFTW3," *Proceedings of IEEE*, vol. 93, no. 2, pp. 216–231, 2005.