On-OPCB Interconnection Networks

Apostolos Siokis, Konstantinos Christodoulopoulos, and Emmanouel Varvarigos
Computer Engineering and Informatics Department, University of Patras, Greece
and Computer Technology Institute and Press – Diophantus, Patra, Greece
e-mail: {siokis,kchristodou,manos}@ceid.upatras.gr

ABSTRACT
Photonic solutions are penetrating the lowest levels of the packaging hierarchy of Data Centers (DC) and High-Performance Computing (HPC) systems (board-to-board, chip-to-chip, intra-chip), as a solution for the increased bandwidth demands and to avoid an explosion in energy consumption. To fully exploit their offered benefits, HPC and DCs architectures need to be reconsidered. In this paper we focus on the on-OPCB (optical printed circuit board) packaging level and describe a lay-out strategy for direct interconnection networks. We also outline a methodology for on-OPCB interconnection network design that takes as input a set of packaging and required performance parameters and incorporates our proposed lay-out strategy. We apply our OPCB designing methodology, using realistic parameters, to highlight potential bottlenecks and to explore the benefits of photonic technological advancements (namely, smaller bending radiuses, crossing angles and chip footprints).
Keywords: optical printed circuit board (OPCB), optical interconnects, topology layout, direct networks, interconnection networks design methodology

1. INTRODUCTION

Photonics is the most promising technology for the next generation Data Centers (DC) and High-Performance Computing (HPC) systems, in order to allow them to cope with the ever-increasing computation density and the bandwidth demanding applications. Optics have successfully replaced electronics in many networking domains [1]: fiber optics have already replaced copper in telecom systems in the range of 10's to 1000's of km's, and have also penetrated shorter distances in campus and enterprise LANs. Active optical cables are currently used for rack-to-rack communication in DC and HPC systems while optical technologies, under active research, target to be deployed in even shorter distances in the near future: board-to-board, on-board and even on-chip.

To take advantage of the new photonic short-distance technologies, we need to reconsider the architectures for HPC systems and DCs at all the different hierarchical levels. In this paper we focus on the packaging of optical modules on boards. In particular, we propose lay-out strategies for optical printed circuit boards (OPCB) and we also present a general methodology for designing interconnects on OPCB, using a set of packaging and required performance parameters as inputs. Our methodology incorporates the lay-out strategies we propose but can also be enriched with other strategies. We apply our designing OPCB methodology to highlight potential bottlenecks and to explore the benefits of technological advancements in photonic integration.

In Section 2 we present our lay-out strategy for interconnects on OPCBs. In Section 3 we describe our methodology for on-OPCB interconnects design, which we later apply to obtain the results presented in Section 4.

2. INTERCONNECTION NETWORK LAY-OUTS FOR OPCBs

In this section we outline a lay-out strategy for interconnection networks on OPCBs. Taking into account IP Phox-Trot [2], a EU funded project on photonics for HPC and DC, we consider the building blocks to be (optoelectronic or all-optical) routers and transceiver optochips/hosts (active Tx/Rx interface modules, on top of which the processors or memory modules are located) that communicate via waveguides. We focus on direct interconnection networks in which every host is directly connected to a routing element. More specifically, we focus on mesh and torus topology families as well as fully connected networks.

Our lay-out strategy translates lay-outs proposed for copper interconnects [3] to a form suitable for OPCBs. The model used in [3] assumes at least 2 layers of wiring, where odd layers include horizontal wires, while even layers the vertical ones, to avoid crossings. All connections between nodes are realized on a 2-D grid, and all bends are 90^0, implemented using "vias" connecting the two layers. The main differences between the optical waveguided communication and the aforementioned model for copper interconnects is that a bending radius is required and that crossings are allowed at the same layer (90^o are preferable due to losses and crosstalk) [4], [5]. Taking those into account we construct the on-OPCB interconnection network with network nodes that consist of one or more hosts and a single router (Section 2.1), and then we connect the routers of the nodes to form a direct network (Section 2.2). In this study we do not assume the use of WDM technology, which would enable multiple wavelengths to be transferred within a single waveguide.

2.1 Node Construction

We first describe how we organize and lay-out network *nodes* with which we build the direct interconnection network. A network node consists of a router chip and a number of optochip hosts, connected in a star topology.

We construct nodes with 2-pinout sides (North and West) – as used in network creation (see next section), assuming routers with peripheral pinout (4-sides) and optochips with a single side pinout (Fig. 1a). In both cases, we arrange the router chip and host chips in a 2-D array, the router chip is in the top-left position and appropriate space is left between rows and columns of the 2-D array and pins are allocated in a specific order so as create the star topology and avoid any waveguides crossing. Different inter-node and intra-node bending radiuses (r_o and r_o' respectively, where $r_o \geq r_o'$, since losses for intra-node network are lower) can be used in order to save area.

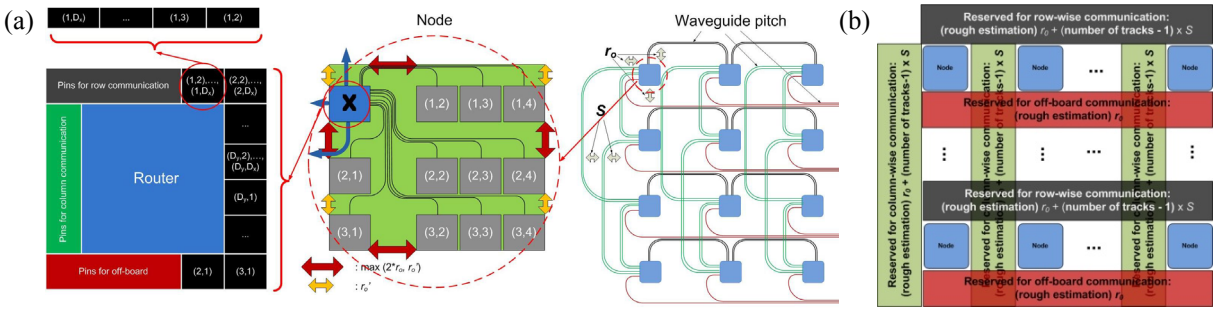

Figure 1: (a) Node layout and pin allocations of the router for router chips with peripheral pinout and (right) a 2-D grid array (3×4) lay-out of 3×2×2 mesh on-OPCB; (b) Lay-out design rules on 2D grid for OPCBs. Space reserved for row-wise, column-wise and off-board communication.

2.2 Direct interconnection network lay-out on OPCB

We examine two types of direct network topology lay-outs: collinear and 2-D. In the former all nodes are placed along a line, while in the latter nodes are placed along rows and columns. The 2-D lay-outs are constructed using collinear lay-outs along their rows and columns, so in the following we focus on collinear lay-outs. We apply the topology lay-outs for electrical interconnects of [3], taking into account that waveguide bends require a bending radius, and crossings are allowed at the same layer.

Connections are done with "Waveguide tracks" (or bundles) which are multi-waveguide links routed together. Waveguides distance within a track is standard pitch (250 μm) – or the waveguide pitch preferred, but since bending radius and chip sizes are at least two orders of magnitude larger, we neglect the tracks' width in our calculations. The first track parallel to the lay-out direction is placed at r_o space from the node, while the space S left between following tracks is related to the desired waveguide crossing angle θ and the bending radius r_o: $S = (1 - cos\theta) \cdot r_o$. Thus, if 90º crossings are used, the tracks spacing equals the bending radius. Smaller bending radius and smaller crossing angles lead to less area required, but also to higher losses. Note that in the adopted layout strategy no crossings occur among waveguides initiating at a node and node sizes are large enough to neglect the spacing between the tracks vertical to the lay-out direction. Also note that the bends and crossings appear in a specific and deterministic order: for every waveguide, an initial bend (or bends) take place, followed by all the crossings, followed by a final bend (or bends).

To lay-out a topology on an OPCB we reserve area for row-wise, column-wise and off-board communication. Our generalized approach for 2-D grid lay-outs is depicted in Fig. 1b (this also shows how we calculate the total required lay-out area). It assumes that nodes have pinouts from the North and West sides for inter-node interconnection (following the node design – Fig. 1a). For the communications of the nodes in the same row, we reserve the space above the nodes. The required area depends on the number of waveguide tracks, which is determined by the topology. For the communications of the nodes in the same column, we reserve the space left to the nodes, while for off-board communication we reserve the space beneath the nodes. Figure 1a (right) depicts an example of a 2-D (3×4) lay-out of a 3×2×2 mesh for OPCBs. A single row of the 2-D lay-out is a collinear lay-out of 3 nodes, requiring 1 waveguide track. A single column is a collinear lay-out of 4 nodes (2×2), requiring 3 waveguide tracks. Two waveguides form a bundle and are used within column and row tracks, while one waveguide/node is used for off-board communication.

3. OPCB INTERCONNECTS DESIGN METHODOLOGY

In this section, we describe briefly our methodology for designing OPCBs, which incorporates the lay-out strategy we presented in Section 2. Our methodology has been implemented in an Automatic Topology Design Tool (ATDT), to aid topology design. The ATDT takes into account two performance metrics, namely speedup and average distance (closely related to throughput and latency). We design networks assuming Uniform Traffic, that is, each source is equally likely to send data to each destination. The speedup of a network is defined as the ratio of the available bandwidth of the bottleneck channel to the amount of traffic crossing it, and is unitless. Speedup equal to one means that the injected traffic equals the available bandwidth of the bottleneck channels (bisection channels for Uniform Traffic). So, under *ideal conditions* (perfect routing and flow control) the network can accommodate the injected traffic with no congestion. Average distance (number of routers traversed on average) depends on the topology of the interconnection network and the traffic pattern.

The OPCB design methodology in ATDT follows 2 stages. In the first stage, given physical and performance inputs (such as module footprints and pinouts, channel rates, losses, power budget), the injected bandwidth from hosts and the probability for off-board communication per host, all the feasible designs are generated. More specifically, we examine different number of optochips on board. For every such case, we examine different number of routers to form nodes. For every such case all feasible mesh, torus and fully connected networks are generated. A design is feasible if the performance constraints are satisfied (the resulting design offers enough bisection bandwidth to achieve on-board speedup at least equal to 1 and the board pinout is large enough to achieve off-board speedup at least equal to 1) and if there is at least one layout of the network that satisfies the board area and worst case losses (power budget related) constraints. In the second stage the optimal design is chosen. The optimality criterion is the maximization of the number of the transceiver optochips (hosts) on-OPCB with the minimal number of utilized router chips. Ties are solved by minimizing the on-OPCB average distance.

4. RESULTS

In this section we first examine the potential benefits of photonic technological advancements, such as smaller module footprints, smaller bending radiuses and smaller crossing angles, on the required board area using our lay-out approach. Then we apply our proposed methodology for OPCB design, using the ATDT, for specific and realistic device and module attributes. We focus on multi-mode optical modules, since at this point they are more mature than single-mode modules, although we can apply our solutions to both.

We assume 50 μm × 50 μm polymer waveguides with a minimum parallel separation (waveguide pitch) of 250 μm. The propagation loss is 0.05 dB/cm for 850 nm wavelength and based on [4], [5] we assume r_o = 20 mm with 1 dB loss per bend, and r_o' = 10 mm. Loss per crossing is given by $L_C = 1.0779 \cdot \theta^{-0.8727}$ [5], and for the baseline scenario we assumed θ = 90º crossings. We assume two symmetrical optical layers of waveguides, one layer for each communication. For the router optochip we followed the PHOXTROT [2] specifications of the router chips for multi-mode communication. The former provides 168 Tx (VCSELs) and Rx (PDs) elements at 8Gbps. The router chip footprint is 52 mm × 52 mm. We assume that all channel pins are available, using all four sides of the router. For our purposes, we assume that the host optochips will accommodate only processors, the channel rate to be 8 Gbps and the on-OPCB optochip footprint to be 52 mm × 52 mm (equal to the router). The number of channels required for host-to-router connection is 12 (assuming processor chips of 1 TFLOPS – Intel Xeon Phi 3100 – and communication-to-computation ratio equal to 0.1 bps/FLOPs). Assuming VCSELs operating at 850 nm with $P_{VCSEL} = 4.7$ dBm power, photodiodes with sensitivity of $PD_{sens} = -13$ dBm and 3 dB loss for chip-to-waveguide or waveguide-to-chip coupling, the power budget is: $B = P_{VCSEL} - P_{couplings} - PD_{sens}$ =11.7 dBm.

4.1 Impact of photonic integration technological advancements on required lay-out area

In this subsection we apply our lay-out approach for a single topology and we examine the benefits on the required lay-out area, varying a single technological parameter at a time. Specifically, we examine the impact of very small bending radiuses (1 mm for both intra- and inter-node connections), smaller crossing angles (45º) and smaller chip footprints (26 mm × 26 mm, and 10 mm × 10 mm) on the required area. The topology we chose is a 3 × 3 torus, laid out in a 2D (3 × 3) fashion, where every router accommodates 4 optochips. 5 router channels are used for off-board and 14 channels for router-to-router connection. Module footprints and sizes for the baseline scenario were described above. The estimated node and network lay-out areas are presented in Table 1. Node areas are rectangles since a node contains an odd number of chips (4 hosts and 1 router). The 50 mm × 50 mm square area in 10 mm × 10 mm chip size case, is due to host-to-router bending radius (also 10 mm). The total area in that case it is a 272 mm × 332 mm rectangle due to the extra waveguide tracks for off-board communication. Different crossing angles do not reduce node area since no crossings occur within nodes. As it can be seen in table 1, all aforementioned improvements in OPCB technologies lead to reduced required area. However, it is clear that the greatest benefit regarding required board area can be obtained by reducing the chip footprints. The impact of the utilization of half size chips (26 mm × 26 mm) is similar to the impact of the (extremely aggressive) assumption of 1 mm bending radius. 10 mm × 10 mm chips (the footprint of the single-mode all-optical router developed in PHOXTROT) leads to less required area than the 1mm bending radiuses.

Table 1. Impact of technological advancements on required area.

	Node area (mm × mm)	Total Lay-out area (mm × mm)
Baseline	176 × 134	524 × 710
Smaller bending radius (= 1 mm)	158 × 107	329 × 485
Smaller crossing angle (= 45º)	176 × 134	481 × 667
Smaller chips (half size)	98 × 82	368 × 476
Smaller chips (= 10 mm × 10 mm)	50 × 50	272 × 332

4.2 Impact of photonic integration technological advancements on on-OPCB interconnection networks

We now apply our proposed methodology for OPCB design, using the ATDT, for specific device and module attributes, to evaluate how these parameters interplay and examine their impact on on-board interconnects design.

We assume board area equal to A4 paper size (297 mm × 210 mm) and board pinout equal to 96 (PHOXTROT's target for multi-mode OPCBs). The rest baseline parameters were described above. The results are presented as graphs. Points in the graphs are denoted by (N_{node}, T, W_b), where N_{node} is the number of hosts (optochips)/node, W_b is the waveguides within a waveguide bundle for router-to-router communication and T represents the topology which is "t" for torus, "m" for mesh, "f" for fully connected, followed by the dimensions of the specific router-router networks. Networks with a single node are not classified to belong to any family.

In Fig. 2a we present the resulting designs, varying the percentage of off-board destined traffic per host. We compare the baseline scenario with scenarios utilizing: (i, ii) smaller chips (26 mm × 26 mm and 10 mm × 10 mm), (ii) smaller bending radiuses (1 mm for both intra- and inter-node connections) assuming 1 dB loss (equal to 20 mm radius loss) and (iii) vertical cabling. In vertical cabling scheme, off-board communication takes place through fiber cables connected to the routers, not through waveguides, leading to less crossings and thus less losses, while board pinout is neglected. We also examine the case where 45º crossings are used (loss L_c as described above), assuming crosstalk is not an issue.

As depicted in Fig. 2a, the highest integration of clients (hosts) on-OPCB can be achieved using smaller chips. Smaller bending radius and vertical cabling also allow more hosts on board. For off-board traffic equal and higher to 0.5, board pinout becomes the bottleneck, reducing the number of hosts that can be accommodated. For the vertical cabling case the main bottleneck is the board area (or the router chip pinout): more routers are added to accommodate the hosts' requirements for off-board traffic, which after a point is constrained by space (A4 board area). For board size = A4, 45º crossings do not improve the baseline scenario.

In Fig. 2b we examine the same scenarios, but keep constant the required off-board traffic (equal to 0.9) and vary the board pinout. 48 pinout is the state-of-the-art for OPCBs, while 96 is targeted in PHOXTROT for multi-mode boards. As explained, the board pinout does not affect vertical cabling scheme designs. Also remember that in all designs a requirement is to ensure that off-board speedup is at least equal to 1. Results indicate that state-of-the-art 48 board pinout only allow very few hosts integrated on-OPCB, while a large portion of the board area remains unused: 144×154 is the total layout area for the (2, 1, 0) baseline. PHOXTROT 96-pinout boards slightly improves that. A 200-pin OPCB would allow more hosts on board, allowing at the same time the area benefits obtained from smaller chips and smaller bending radiuses. A far larger board pinout (400) and the use of 10 mm × 10 mm chips would allow very dense integration (151×230) and more efficient usage of board area.

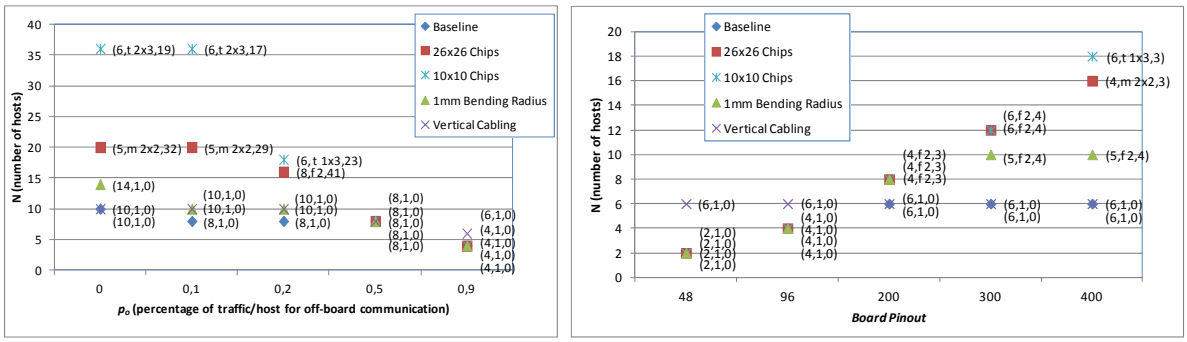

Figure 2: (a) Number of hosts on-OPCB with board pinout = 96, varying the percentage of off-board traffic; (b) Number of hosts on-OPCB for 90% off-board traffic, varying the board pinout.

5. CONCLUSION

We presented lay-out strategies for interconnects on-OPCBs, and a methodology for designing on-OPCB interconnects, using a set of packaging and performance parameters as inputs. We applied our methodology for on-OPCB interconnects design, with realistic parameters, to examine potential benefits of photonics technological advancements and to highlight potential bottlenecks. Our results indicate that reducing the footprints of the chips and also increasing the board pinout, can allow more hosts to be accommodated on OPCBs.

ACKNOWLEDGEMENTS

This work was supported by the EC through PHOXTROT (ICT 318240).

REFERENCES

[1] M. Taubenblatt, "Optical interconnects for high performance computing", in *Proc. OFC,* 2011, paper OThH3.
[2] http://www.phoxtrot.eu/
[3] C.-H. Yeh, E.A. Varvarigos, and B. Parhami, "Multilayer VLSI lay-out for interconnection networks", in *Proc. Int'l Conf. Parallel Processing*, 2000, pp. 33-40.
[4] K. Wang *et al*., "Optical waveguide modelling, measurement and design for optical printed circuit board", in *Proc. Workshop on Interconnections within High Speed Digital Systems (HSDS)*, 2008.
[5] K. Wang *et al*., "Photolithographically manufactured acrylate polymer multimode optical waveguide loss design rules", in *Proc. Electronics System-Integration Technology Conference*, 2008.