Laying out Interconnects on Optical Printed Circuit Boards

Apostolos Siokis, Konstantinos Christodoulopoulos, Emmanouel (Manos) Varvarigos
Computer Engineering and Informatics Department, University of Patras, Greece, and
Computer Technology Institute and Press – Diophantus, Patras, Greece
{siokis,kchristodou,manos}@ceid.upatras.gr

ABSTRACT

Short distance optical interconnections, on-printed circuit boards, on-backplanes, and even on-chip, are a promising solution for replacing copper interconnections in future Data Center and HPC systems. Since photonic technology introduces new network building blocks, topology design for all the packaging levels should be reconsidered. This paper focuses on the on-board level of the packaging hierarchy, and proposes lay-out strategies for optical interconnection networks on optical printed circuit boards (OPCBs), based on direct topology families (tori, meshes and fully connected networks). We also describe a methodology for designing OPCBs given a set of input parameters, including building blocks specifications as well as traffic demands. The on-board topology design methodology generates all the feasible designs within the topology families examined, following our proposed OPCB lay-out approach, and selects the optimal designs based on specific optimization criteria.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design - *Network topology*; C.5.4 [**Computer System Implementation**]: VLSI Systems

Keywords

Optical interconnects; Optical printed circuit boards; Waveguides; On-board topology lay-out; Direct networks

1. INTRODUCTION

The ever-increasing network load in Data centers (DC) and High-Performance Computing (HPC) systems pushes the electrical-copper interconnection technologies to their limits. As the need for bandwidth grows, electrical interconnects cannot keep pace due to wiring density [1,2], high power dissipation, increased signal degradation and crosstalk between neighboring channels. On the other hand, photonic technologies offer superior bandwidth-distance product at much lower energy consumption. These reasons lead to the replacement of copper based communication in a from-outside-to-the-inside manner [3]: fiber optics have already replaced copper in long-haul (MAN and WAN) telecom systems in the range of 10's to 10000's of km's, and are penetrating shorter distances (<= 100's of meters) in campus and enterprise LANs. At this point active optical cables are the norm for rack-to-rack communication in DC and HPC systems.

ANCS'14, October 20–21, 2014, Los Angeles, CA, USA.

http://dx.doi.org/10.1145/2658260.2658277

Even so, power consumption of data communication is still daunting. Prediction studies for performance, bandwidth requirements and power consumption, back in 2010, projected that a 10PF HPC machine in 2012 would require 5MW [24]. One of the top500 HPC systems, 2011 K-supercomputer with 10PF performance, requires more than double the predicted amount of power (~12.7 MW) [25]. The global demand for electricity from data centers was around 330bn kWh in 2007 and it is projected to rise to more than 1000bn kWh by 2020 [26]. So, to cope with both the energy and bandwidth limitations, optical technologies target to be deployed in even shorter (in-the-box) distances in the near future: board-to-board, on-board (module-to-module), and even on-chip (distances < 20 mm).

This new era brings an entirely new technology portfolio of network modules for short-distance communication. These include: Optical Printed Circuit Boards (OPCBs), printed with multi-mode (usually polymer) or single-mode (polymer or glass) waveguides, optical transceiver chips (usually VCSELs – Vertical Cavity Surface-Emitting Laser for Tx, and PDs – PhotoDiodes for Rx), optoelectronic and photonic routers, Arrayed Waveguide Gratings (AWGs), backplanes for passive board/daughtercard optical interconnection, chip-to-board coupling technologies, optical RAMs, among others. Recently completed and ongoing research efforts include IBM's "Terabus" for tranceiver optochips-on-optoboard and "C2OI" for intra-chip and off-chip communication [4], IBMs-Columbia University research on photonic networks-on-chip [5], Intel-UCSB joint initiative research on silicon laser, modulator and amplifier configurations [6]. Several European research initiatives have focused on specific optical interconnection technologies (like FP7 POLYSIS [7]).

FP7 PHOXTROT [8] investigates the development of low cost and energy efficient optical interconnects at chip-to-chip, board-to-board and rack-to-rack levels of the packaging hierarchy. Within PHOXTROT, various photonic technologies are being developed, but the research activities also examine how these can be deployed at the different packaging levels. Thus, to take advantage of the new photonic technologies we need to reconsider the architectures for HPC systems and DCs at the different hierarchical levels. Architectural issues such as on-board, on-backplane and system level topologies, number of waveguides for chip-to-board communication, number of routers on board, number of channels/waveguides for router-to-router communication, lay-out of topologies in waveguide levels, board pinout, switching paradigms (packet vs. circuit) are issues that need to be re-visited, re-addressed, and re-evaluated.

A key difference between electrical and optical interconnects is the physical layer, which constrains the interconnects that can be designed. While related constraints in electrical interconnects are well understood and subsystems and whole systems building methodologies are well established, there is no related work in optical interconnects, especially for the lower hierarchical levels. A recent survey [9] discusses the benefits of photonic technologies for next generation HPC systems and DCs, but

mainly for inter-rack communication. A number of on-board optical architectures that use such "in-the-box" photonic technologies have been proposed, including a shared optical bus [18], a high-speed clock distribution tree [19], a meshed waveguide architecture for optical backplanes [20] and an optical bus for optical backplane interconnections [21].

Since the underlying technologies and the packaging constraints of the various levels of the packaging hierarchy determine the feasible system level topologies/architectures, we chose a bottom-up approach and focused on the packaging of optical modules on boards. In particular, we propose lay-out strategies for on-optical printed circuit board (OPCB) topologies, "translating" existing lay-out strategies for electrical PCBs into a form suitable for OPCBs. We mainly target HPC system designs and thus we focus on direct networks such as meshes, tori and fully connected networks. 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 it can also be enriched with other strategies. To the best of our knowledge, this is the first work that presents a structured lay-out strategy for OPCBs. There are software suites for OPCB design, but, however, they focus on physical layer and propagation modeling of the waveguides and do not provide design guidelines/methodologies for topology lay-outs on OPCBs. Although we focus on OPCBs, we plan to re-apply our approach (somewhat modified) for higher packaging levels (rack, set of racks, up to the whole system). Note that methodologies for topology design have been presented in the past [10,11], but aimed for electrical/copper interconnects. Compared to that, we put more emphasis on and have a more detailed model for the lay-out strategy of the (optical) topologies.

The contributions of this paper are:

- We outline the similarities and differences between lay-out models for electrical interconnects and optical waveguided communications in order to capture the peculiarities of the latter in a lay-out model suitable for optical interconnects (Subsection 2.3).
- We propose a way to organize and lay-out an optical router chip and hosts chips in network nodes, suitable for direct network topologies (Subsection 2.4).
- We propose a lay-out strategy for optically interconnected direct topologies of nodes, suitable for OPCBs (Subsection 2.3).
- We propose an articulate methodology for on-OPCB topology design that incorporates our lay-out strategies and takes into account network performance metrics while keeping in mind that the OPCB will be part of a bigger system (Section 4). This methodology maximizes the number of hosts on-OPCB given the available board area while using the minimum number of (active) router chips, but other optimization objectives can be easily employed.
- We apply our designing OPCB methodology to highlight potential bottlenecks and to explore the benefits of technological advancements in photonic integration (Section 5).

This paper is structured as follows. In Section 2 we present lay-out strategies for interconnects on OPCBs. In Section 3 we introduce the performance metrics that we consider. In Section 4 we present our methodology for designing on-board interconnects, and we apply it to obtain the results presented in Section 5, using PHOXTROT subsystem specifications as input. Conclusions follow in Section 6.

2. LAY-OUT STRATEGY FOR INTERCONNECTS ON OPCBs

In this section we present lay-out strategies for interconnection networks on OPCBs. These lay-out strategies are incorporated in our topology design methodology described in section 4. To determine the number of modules and the topologies (within the topology families considered) that are *feasible* in the on-board packaging level we need to calculate the required area and worst case losses of each design. So we start in Subsection 2.1 by describing the topology families (tori, meshes, fully connected) we consider, followed in Subsection 2.2 by general electronic lay-out strategies for them. Then, in Subsection 2.3 we reveal the differences between copper and waveguided interconnections and describe our strategy for laying out on OPCBs. In Subsection 2.4 we describe the way we organize optical interconnection building blocks (routers and transceiver optochips) in network nodes. Finally, in Subsection 2.5 we briefly discuss waveguide length matching.

2.1 Considered network families

Interconnects can be distinguished in two classes: (a) direct networks in which every host is directly connected to a routing element, and (b) indirect networks in which there are routing elements with no hosts connected to them. The majority of direct topologies are either configurations of or isomorphic to meshes, tori, k-ary n cubes, while popular indirect topologies are trees (including fat-trees), clos, and butterfly networks.

In this work we target mainly HPC environments and thus we focus on direct networks. More specifically, we focus on meshes, tori and fully-connected networks (FCN). Note that several HPC systems in the Top500 are built with tori/meshed networks [12,13,14].

Formally [15], a n-dimensional mesh has k_1 x k_2 x ... x k_n nodes, k_i nodes along dimension i, where $k_i \geq 2$ and $1 \leq i \leq n$. Note that since this is a direct network, the nodes correspond to a routing element and one or more hosts that are directly connected to it. A node x is logically identified by n coordinates ($x_1, x_2, \ldots x_n$), where $1 \leq x_i \leq k_i$ for $1 \leq i \leq n$. Two nodes x and y are neighbours and are connected through a link if and only if $y_i = x_i$ for all i, $1 \leq i \leq n$, except for one coordinate j, where $y_j = x_j \pm 1$.

A n-dimensional torus is the equivalent mesh with added wrap-around links. Formally, in a k_1 x k_2 x ... x k_n torus, two nodes x and y are neighbours if and only if $y_i = x_i$ for all i, $1 \leq i \leq n$, except one, j, where $y_j = (x_j \pm 1) \bmod k_j$.

Finally, in a fully connected network (FCN) of N_r nodes, every node is connected with the other N_r - 1 nodes.

2.2 Lay-outs for electrical interconnection networks

The authors in [27] present lay-outs for a variety of interconnection network topologies, and provide formulas for the required lay-out area and required tracks, assuming copper wiring.

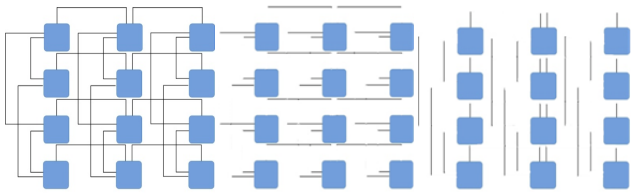

Figure 1. (a) 2-D grid array (3x4) lay-out of 3x2x2 mesh, (b) and (c) layer one and two implementing the horizontal and vertical wires.

The model used, following the well known Thomson model, 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 (when viewing both layers), implemented using "vias" connecting the two layers. Note that as discussed above, for the indirect networks under study, nodes consist of a routing element and one or several host attached to it, while the network is build by connecting the routing elements. In the next section we will discuss how to build nodes, but here we consider them as a single block.

In this paper we examine two types of network topology lay-outs: collinear and 2-D. In the former all network nodes are placed along a line, while in the latter nodes are placed along rows and columns, forming a 2-D grid array. Figure 1(a) depicts an example of a 3x2x2 mesh, laid out in a 2-D grid of 3x4 nodes, with wires also laid out in a 2-D grid. Note that the wiring, although depicted in one layer, is done in 2 (or more) layers, and Figure 1 (b) and (c) show the related 2 level implementation. 2-D lay-outs are constructed using collinear lay-outs along the rows and columns. A single row of the 2-D lay-out in Figure 1(a) is a collinear lay-out of 3 nodes, requiring 1 wiring track. A single column of the 2-D lay-out is a collinear lay-out of 4 nodes (2x2), requiring 3 wiring tracks. In what follows we will only consider collinear lay-outs, having in mind that 2-D lay-outs are constructed by using them.

We have calculated the number of tracks for collinear lay-outs of meshes and tori of arbitrary dimensions using the strategies in [27]. For a collinear lay-out of a k_1 x k_2 x ... x k_n torus, the number *Y* of tracks parallel to the lay-out direction is:

$$Y = \sum_{i=0}^{n-1}(a_i \cdot \prod_{j=0}^{i} k_j),\ a_i = \begin{Bmatrix} 1, if\ k_{i+1} = 2 \\ 2, if\ k_{i+1} > 2 \end{Bmatrix},\ k_0 = 1 \qquad (1)$$

For the equivalent mesh, the number of tracks would be the same as Eq. (1) but with $a_i = 1$ in all cases. The number of tracks for a strictly optimal collinear lay-out of a fully connected network (FCN) is $\lfloor N^2/4 \rfloor$ [27]. We consider only collinear FCNs, because this topology family is difficult to be layed-out in a 2-D grid.

We have also calculated the worst case crossings for tori and meshes of arbitrary dimensions as well as FCNs, for the above discussed lay-out strategies. Both the worst-case crossings number and the number of tracks will be used for the estimation of the worst-case losses.

In the FCN, a *type-i* link connects two nodes whose addresses differ by *i*. The $N(N-1)/2$ links of the FCN can be classified into type 1, 2, 3, …, $N-1$, and there are $N-i$ type-*i* links. In the FCN lay-outs of [27], all nodes meet the same number of crossings in the worst case. The largest number of crossings appears in the $\lceil \frac{N}{2} \rceil$ and $\lfloor \frac{N}{2} \rfloor$ links of every node. We calculate the number of crossings for link $(1, 1+N/2)$. This link will meet $\left(\lfloor \frac{N}{2} \rfloor - 1\right)$ links from nodes 2, 3, …$\lfloor \frac{N}{2} \rfloor$. Thus, the total number of worst case crossings is $\left(\lfloor \frac{N}{2} \rfloor - 1\right) \cdot \left(\lfloor \frac{N}{2} \rfloor - 1\right)$.

For k_1 x k_2 x ... x k_n meshes and tori we do not use a closed-type formula, but we calculate the number of crossings in a recursive manner. The mesh and torus collinear lay-outs of [27] are also built recursively using a bottom-up approach, starting with a single dimensional ring (or chain array for mesh) and inductively moving to higher dimensions. The worst case crossings appear in (but not necessarily only in) the highest dimension (dimension *n*) links: the links that connect k_n segments of the k_1 x k_2 x ... x k_{n-1} subnetworks (or a wraparound link in a torus). In brief, for every dimension *i*, $1 \le i \le n$, we create a vector of size $k_1 \cdot k_2 \cdot \ldots \cdot k_{n-1}$, using subvector patterns of size $k_1 \cdot k_2 \cdot \ldots \cdot k_{i-1}$ and repeat them $k_i \cdot \ldots \cdot k_{n-1}$ times. A vector in dimension *i* will contain the number of crossings that the highest dimension (dimension *n*) links will exhibit due to links of dimension *i* at hand. Adding elementwise the resulting *n* vectors we get a final $k_1 \cdot k_2 \cdot \ldots \cdot k_{n-1}$ vector that contains the total number of crossings for the highest dimension links. The number of the worst case crossings is the max element of this vector.

2.3 Lay-outs of interconnection networks on OPCBs

The main differences between optical waveguided communication and the copper interconnects described above, from the lay-out point of view, are:

(a) waveguide bends require a bending radius, and

(b) crossings are allowed in the same layer (a crossing angle of 90^o is preferable due to losses and crosstalk) [22, 23].

The lay-outs of direct topologies described in Subsection 2.2 can be applied on OPCBs with the following modifications. As before, network nodes form the building blocks and are constructed by one router chip and hosts following the strategy described in the next Subsection. We assume two symmetrical layers, each for one direction of communication between nodes. Given the collinear lay-out of nodes (remember that 2-D lay-outs are constructed from row- and column-wise collinear layouts), at each layer the links are laid out in a 2-D grid, bends have a given radius, and crossings are allowed to occur.

In case where more than one link is needed between two nodes and since bends are (space and loss) expensive, we route multi-waveguide links together, as bundles, in a single "waveguide track". Waveguides distance within a track is standard pitch (250μm) – or the waveguide pitch preferred, but since bending radius and chips sizes are at least two orders of magnitude larger, we neglect tracks width in our calculations. The first track parallel to the collinear lay-out direction of nodes 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 as follows:

$$S = (1 - cos\theta) \cdot r_o \qquad (2)$$

Thus, according to Eq. (2), if 90^o crossings are used, the tracks spacing equals the bending radius (S=r_o). Smaller bending radius and smaller crossing angles lead to less required area, but to higher losses. Since crossings are allowed in the same layer, even only one layer would suffice if the worst case losses (due to bends, crossing, and distance) allow that.

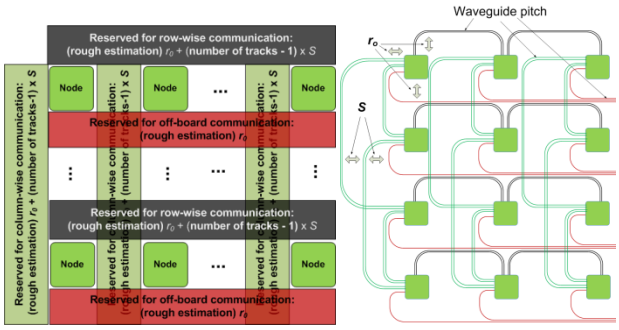

Figure 2. (a) Lay-out design rules on 2D grid for OPCBs. Space reserved for row-wise, column-wise and off-board communication. (b) Lay-out of the 3x2x2 mesh of Figure 1 on OPCB, following the strategy shown in (a).

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-, column-wise and off-board communication. Our generalized approach for 2-D grid lay-outs is depicted in Figure 2(a). It assumes that network nodes have pinouts from two of their sides for inter-node interconnection (North and West sides – see next Subsection for a way to construct such nodes). 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 row-wise collinear topology (Subsection 2.2). For the communications of the nodes in the same column, we reserve the space left to the nodes, again depending on the required tracks. Finally, for off-board communication we reserve the space beneath the nodes that has width equal to r_o, since we assume that all off-board waveguides from all nodes at the same row are routed in parallel with standard pitch (or the pitch preferred) between them, at r_o distance from the nodes. If nodes use a single side for pinout, then the required area for waveguides is the same, but more bends are required. For simple collinear lay-outs, the proposed strategy is that of a single row of a 2D, as depicted in Figure 2(a), but the required distance between nodes is r_o, because no column-wise communication takes place. Figure 2(a) also gives an estimation of the total required area. In Figure 2(b) a 2-D (3x4) lay-out of a 3x2x2 mesh is depicted (equivalent to network of Figure 1). Two waveguides form a bundle and are used within column and row tracks, while one waveguide/node is used for off-board communication.

In principle, the reduced link-to-link separation (waveguide pitch) and the allowance of crossings in the same layer (compared to electrical interconnects) allow denser integration and reduction of PCB thickness (layer count). The usage of WDM, will further increase the data density in Gbps/mm. However, a potential issue is crosstalk with respect to crossing angle, for angles less than 90°. To the best of our knowledge there is not yet a design rule/formula for crosstalk as a function of the crossing angle. Measurements for crosstalk can be found in [21], but only for the examined bus architecture. Another manufacturing issue for OPCBs is that the performance of multimode waveguide components depends on the launch conditions at the component input (see discussion in [23]). Note that we have not assumed WDM, which would enable multiple wavelengths to be transferred within a single waveguide. So links are point-to-point, as in electrical interconnects, and we plan to explore the benefits of WDM in future work.

2.4 Node construction and lay-out

We now describe how we organize and lay-out network nodes suitable for direct network architectures that are laid out according to the previous Subsection. In our approach a network node consists of a router chip and one or several optochip (hosts) connected in a star topology. The transceiver optochips provide optical inputs/outputs to processors or memory modules. We construct nodes with 2-pinout sides (North and West), Figure 3(a), as used in network creation (see Figure 2(a)), assuming router chips with peripheral pinout (4-sides) and optochips with a single side pinout. We have also developed lay-outs with single side pinout routers, Figure 3(b), since this is considered easier to manufacture and mount on OPCBs. The predominant Tx modules for optoelectronic or photonic router chips are Vertical-Cavity Surface Emitting Lasers (VCSELs) while the Rx modules are PhotoDiodes (PD). We assume that the VCSELs and PD arrays are laid out at the peripheral of the chip, forming as many rows as the layers supported in the OPCB platform (two in our case). Else, if the chip pins are laid out in a matrix, we assume that the Tx and Rx pins can be mapped in a way that enables to view the chip as a bulding block with peripheral pinout.

In both cases, to construct the node we arrange the router chip and host chips in 2-D arrays. We assume that we have $M = N_{node} + 1$ chips of the same size, where N_{node} is the number of optochips – adding one for the router chip. We arrange these chips in a 2-D array with $D_x = \lceil \sqrt{M} \rceil$ columns and $D_y = \lceil \sqrt{M} - 0.5 \rceil$ rows, while placing the router at the top-left position. Note that placing the router in the middle and leaving space between transceiver optochips in order to save waveguide bends, would ultimately lead to more required area. Such alternative lay-outs are not ruled out and are left for future studies. Also, note that depending on M, some of the array positions maybe left blank. In both cases (2 and 1 pinout side) nodes can be constructed without waveguide crossings, by appropriate spacing and allocation of router pins. Finally, note that, to save space, we use for intra-node connections (hosts-to-router) a smaller bending radius r_i, as compared to r_o used for inter-node communication, since the traveled distances within a node are smaller and no crossings occur.

The node lay-out strategy with waveguides exiting 2 (North and West) sides using router chips with peripheral pinout (4-sides) is depicted in Figure 3(a). We show the required space between the modules to allow the waveguides to take the required turns (of r_i radius) and we also depict the allocation of the Tx router pins (for 2-D lay-outs) to make the star topology and exit the node without crossings. They are arranged in the following manner (clockwise): {pins for row communication, pins for intra-node communication for hosts: $[(1,2),(1,3),\ldots,(1,D_x)],\ldots,[(D_y,2),(D_y,3),\ldots,(D_y,D_x)]$, $(D_y,1)$, $(D_{y-1},1)$,..., (3,1), (2,1), pins for off-board communication, pins for column communication}. If more pins are needed for a specific type of connection, then more pins can be reserved from the "neighboring pin areas", maintaining however the ordering. For collinear node topologies, both pinout router areas for row and column communication will be used just for the row-wise communication.

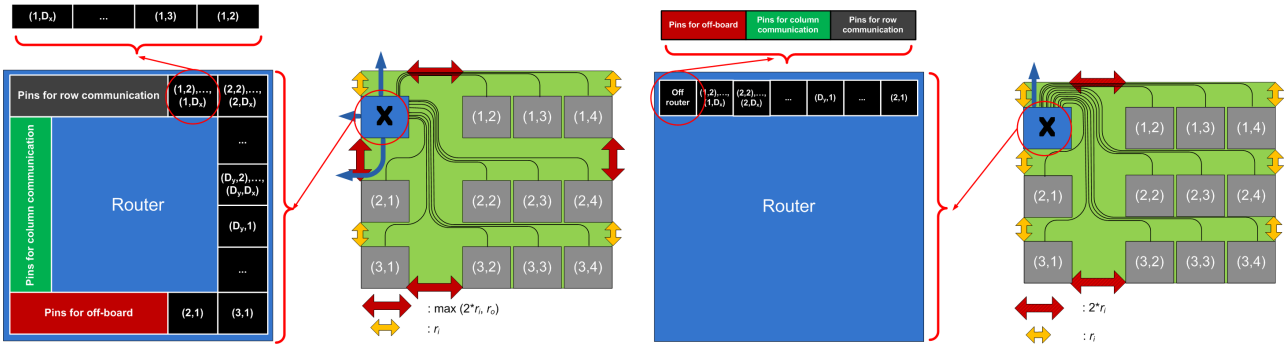

Figure 3. Node lay-out and Tx pin allocations of the router (similar for Rx) for (a) router chips with peripheral pinout, (b) router chips with North-side pinout.

A similar pin allocation pattern is followed for the Rx router pins. The node lay-out strategy with waveguides exiting a single (North) side using router chips with single-side pinout is depicted in Figure 3(b).

2.5 Waveguide length matching

An important issue for electrical PCB designers is trace/link length matching. The trace length mismatch tolerance is determined by (i) the protocols tolerance in timing skew and (ii) the propagation speed of the signals in the medium. The majority of the high-performance, high-speed protocols are serial (eg InfiniBand**,** Serial RapidIO, PCI-Express). In these protocols a serial lane is composed of two differential signaling pairs per direction. A single link between two devices can consist of multiple serial lanes where the data is striped across these lanes. The length matching requirements between differential pairs are usually very tight, while length matching requirements between lanes are looser. Differential signaling is used in electrical interconnects since differential signals are less susceptible to crosstalk and also tend to produce less electromagnetic interference (EMI) than single-ended signals. Optical signals do not suffer from the aforementioned problems, they have low crosstalk, allowing denser waveguide spacing [1, 3]. This is particularly important, since the much smaller pitch between parallel lanes makes the length mismatch very small. Moreover the propagation speed is higher in polymer optical waveguides than the related electrical lanes [29], relaxing the trace length mismatch tolerance even further. To provide a concrete example, we will take the lane-to-lane skew matching values for of PCI-express 3.0. In PCI-express lane-to-lane skew should be less than 2*UI*+500ps which corresponds to 750ps for PCI-e 3.0 (*UI*=125ps, since the channel rate of a differential pair is 8Gb/s). Assuming polymer waveguides with refractive index $n \approx 1.5$, thus signal propagation speed $c/n \approx 2 \cdot 10^8$ m/s, the tolerated differences in waveguide lengths would be $(2 \cdot 10^8$ m/s) / 750ps = 150 mm for lane-to-lane waveguides. We can calculate the difference in the waveguide lengths (taking into account the tracks-width) between the two links: Assuming bending radius r_o, then the length of a 90° bend for the inmost waveguide is $S_i = 2 \cdot \pi \cdot r_o \cdot 90^o/360^o = \pi \cdot r_o / 2$ and the length of the equivalent 90° bend for the outmost waveguide is $S_o = \pi \cdot (r_o + pitch) / 2$, where *pitch* is the waveguide pitch. The length difference for 2 bends is $\Delta S = 2 \cdot (S_o - S_i) = \pi \cdot pitch$. For standard pitch of 250μm, ΔS = 785 μm << 150 mm. If we assume 32 waveguides in a single layer, for a single link (PCI-e x32), 50μm wide waveguides and standard pitch (250 μm). In this case, length difference between the inmost and the outmost waveguide would be (for 2 bends): $\Delta S = \pi \cdot [31 \cdot (250+50)] = 29.2$ mm < 150 mm. Taking all the above into account, in the following paragraphs we will assume a single-ended signaling, serial, multi-lane protocol, tolerant to lane-to-lane mismatches.

However, for the sake of completeness, we will discuss some potential solutions for de-skewing. A design option could be to route differential signals in different waveguide layers over same paths. Another option could be to route the differential signals in the same layer, in successive waveguides, and use S-shaped bends in the shorter lane to increase its length. S-shaped waveguide bends (see [28] and references cited there) are smoother and in principle far less expensive in losses than 90° bends. S-bends can be generated by two circular arcs of constant radius *R*, sine, cosine or raised cosine functions. A waveguide S-bend structure made of two circular arcs with a constant radius of curvature *R* is specified as: $R = \pm \frac{L^2}{4d}\left(1 + \frac{d^2}{L^2}\right)$, where *L* is the transition length and *d* is the lateral offset. The path length of such an S-bend is $S = 2 \cdot R \cdot \theta \cdot \pi/180$, where θ is in degrees. If S-bends are used for length matching, then the designer should pay attention where these will be placed, in order to maintain the crossing angles between (row-wise, column-wise, off-board) waveguides. To the best of our knowledge, there isn't yet a design rule/formula for S-bend losses as a function of *R*, *L*, *d*, θ. However, measurements for *d* =10 mm can be found in [28]. Finally, more aggressive waveguide pitch (such as 62.5 μm) would further reduce length mismatches.

3. PERFORMANCE METRICS FOR NETWORK DESIGN

We discuss now the performance metrics that we take into account in our OPCB design methodology described in Section 4. The layout strategies discussed in the previous section specify if a topology is feasible in terms of area and losses, while the performance metrics discussed here characterize the topology, irrespective of the actual layout. Both of these features are taken into account in the Automatic Topology Design Tool we present in the next section.

We assume that our system of *N* hosts (optochips) in total consists of N_r nodes, where each node consists of a single router interconnected with N_{node} hosts, and thus N_r is also the number of routers on the board. The two most representative and general metrics of performance for interconnection networks are throughput and latency [15]. Both throughput and latency are functions of topology, routing policy, flow control, interconnect characteristics, switch architecture, as well as traffic characteristics. We use two metrics that are closely related to throughput and latency, namely Speedup and Average Distance that will be described below. We design networks assuming Uniform Traffic (commonly used for topology design), that is each source is equally likely to send to each destination.

While the quality of an interconnection network should be measured by how well it satisfies the communication requirements of targeted applications, on the other hand problem-specific networks are inflexible and thus good "general purpose" networks should be opted for. This is why Uniform Traffic which is quite generic is typically chosen for the topology design phase. It is also useful for emulating global exchange traffic with no underlying data locality (such as HPC applications based on Fast Fourier Transformation - FFT). In the future we plan to evaluate the resulting designs under realistic traffic patterns using simulations.

Ideal Throughput and Speedup. Throughput is the number of bits per second the network can transport from input to output. It is a function of topology, routing policy, flow control, interconnect characteristics, switch architecture, as well as traffic characteristics. The throughput that a topology can carry can be calculated assuming ideal routing (perfect load balancing over alternative paths) and flow control (no idle cycles on the bottleneck channels), what is defined in the literature as *Ideal Throughput* λ_{ideal} [15]. It equals the input bandwidth that saturates the bottleneck channel(s) for a specific traffic pattern, assuming that the hosts have infinite injection bandwidth so as to reach the saturation point. Considering a real system where hosts have a specific maximum injection bandwidth λ_{max} limited by hosts' pinout and channel rates, then the Ideal Throughput under traffic injection constraints $\lambda_{ideal-ic}$ is defined as follows:

$$\lambda_{ideal-ic}(\lambda_{max}) = \begin{cases} \lambda_{ideal}, & if\ network\ is\ saturated\ before\ \lambda_{max} \\ \lambda_{max}, & otherwise \end{cases}$$

The *Speedup* of the network is defined as the ratio of the available bandwidth of the bottleneck channel to the amount of traffic crossing it (under ideal conditions as discussed above). As opposed to Ideal Throughput, Speedup is unitless.

$$Speedup(\lambda_{max}) = \frac{Bottleneck\_channel_{bandwidth}}{Bottleneck_{Traffic}(\lambda_{max})} \quad (3)$$

Speedup is very useful when designing networks. Speedup equals to 1 means that, under ideal conditions, the network can accommodate the injected traffic with no congestion. That is, hosts can inject their maximum bandwidth λ_{max} without reaching network saturation point. Designing a network with Speedup greater than 1, allows non-idealities in the implementation. Speedup is related to ideal throughput under traffic injection constraints as follows:

$$\lambda_{ideal-ic}(\lambda_{max}) = \begin{cases} Speedup(\lambda_{max}) \cdot \lambda_{max}, & if\ Speedup(\lambda_{max}) \leq 1 \\ \lambda_{max}, & otherwise \end{cases}$$

$$\rightarrow \lambda_{ideal-ic}(\lambda_{max}) = \begin{cases} \lambda_{ideal}, & if\ Speedup(\lambda_{max}) \leq 1 \\ \lambda_{max}, & otherwise \end{cases} \quad (4)$$

where we used Speedup to identify whether the network is saturated with the maximum injection bandwidth or not. In our model, we assume that data that has to be transmitted from a router to a router is distributed evenly over their connecting waveguides, and we assume infinite flow granularity. To calculate the Speedup, we must calculate the values of $Bottleneck\_channel_{bandwidth}(\lambda_{max})$ and $Bottleneck\_Traffic$.

For Uniform Traffic, the bottleneck channels are the bisection channels and the traffic that crosses the bisection width is distributed uniformly [15]. Thus, we must calculate the bisection bandwidth and the traffic that crosses the bisection channels. The bisection bandwidth B_b is calculated as follows

$$B_b = 2 \cdot B_w \cdot C \quad (5)$$

where B_w is the bisection width of the examined topology, and C is the channel rate. In this, we take into account traffic in both directions, since we assume uni-directional waveguides. Closed type formulas for B_W of tori and meshes can be found in [16], while for a FCN of N_r nodes we have $B_w = \left\lfloor \left(N_r/2\right)^2 \right\rfloor$. The amount of traffic that crosses bisection channels is found as follows. For N hosts, there are $N-1$ candidate destinations on board (not considering self-traffic) and N_{node}-1 candidate destinations that are connected to the same router. Let's assume that p_{on} of the host's injected traffic is destined for on-board communication. Since every host injects λ traffic, the total inter-router traffic is $N \cdot \lambda_{max} \cdot p_{on} \cdot (1 - \frac{N_{node}-1}{N-1})$. In Uniform traffic, half of that will cross bisection channels. So taking into account self-traffic as well, we have

$$Bottleneck_{Traffic}(\lambda_{max}) = \frac{N \cdot \lambda_{max} \cdot p_{on} \cdot \left(1 - \frac{N_{node}-1}{N-1}\right) \cdot \left(1 + \frac{1}{N_r - 1}\right)}{2} \quad (6)$$

Note that parameter p_{off}=1-p_{on}, which corresponds to the percentage of off-board traffic, is one of the key parameters considered in the methodology (Section 4). As such we examine how this parameter affects the performance, in the related Section 5.

Average distance and zero load latency. The average distance (number of hops traversed on average) for meshes, under Uniform Traffic, is calculated by adding the average distance for each dimension [17]. Following the same rationale, we can calculate the average distance for Torus. In Tori, average distance in dimension i equals to $\frac{k_i}{4} - \frac{1}{4k_i}$ if k_i is odd and $\frac{k_i}{4}$ if k_i is even [15]. Average distance in an FCN equals $\frac{(N_r-1)}{N_r}$. All the above take into account self-traffic. Zero load latency is the latency experienced by packets on average, at a load where no contention occurs. Assuming store and forward switching, the zero load latency T_0 is:

$$T_0 = h_{av} \cdot \left(T_r + T_{trans} + T_{prop}\right) \quad (7)$$

where h_{av}, T_r, T_{trans}, T_{prop} are average distance, average router delay, transmission delay, and propagation delay, respectively.

4. METHODOLOGY FOR DESIGNING INTERCONNECTS ON OPCBs

In this section, we present our methodology for designing OPCBs, which incorporates the lay-out strategies we presented in Section 2 to identify whether a design is feasible and uses the performance metrics described in Section 3 to judge its efficiency. Our methodology has been implemented in an Automatic Topology Design Tool (ATDT), to aid topology design. We assume two different schemes for off-board communication: off-board communication through waveguides, or alternatively through vertical cabling. In the second case no waveguides for off-board communication is needed and no board pinout constraint is imposed.

The OPCB design methodology in ATDT follows 2 stages. In the first stage, given physical (such as module footprints and pinouts, channel rates, losses, power budget, board pinout) and performance (required Speedup) inputs, 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 on board. For every

such case all feasible mesh, torus and fully connected networks are generated. A design is feasible if

(i) The performance constraints are satisfied (the resulting design offers enough bisection bandwidth and the board pinout is large enough to achieve on- and off-board Speedup at least equal to the required), and

(ii) There is at least one lay-out of the network that satisfies the board area and worst case losses (power budget related) constraints.

The second stage takes all the feasible designs and chooses the optimal one. The optimality criterion used 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 zero load latency. Note that other optimization criteria can be applied, without having to re-execute phase 1, and this is one of the main reasons we followed such a two-phase approach.

Below, we present our methodology using pseudocode, and then we elaborate on several details.

Algorithm: OPCB design

/***Goal:** Maximize hosts on-board, while ensuring on- and off-board Speedup ≥ S*/

Inputs:
s_r: side in mm of the (square shaped) router chip
s_h: side in mm of the (square shaped) host chip
A: OPCB area (as *board height* x *board width*)
r_i: bending radius for host-router waveguides
r_o: bending radius for router-router waveguides
U_R: router chip pinout (number of pairs of Tx/Rx)
U_B: board pinout – set to inf. for vertical cabling
W_h: waveguides (pairs) for host-router connection[1]
θ: waveguide crossing angle
B: power budget
L_p: propagation loss per mm
L_b: power loss due to a waveguide bend
L_c: power loss due to a crossing
N_{max}: max number of hosts to be attempted to fit on board
r: channel rates
p_{off}: percent of off-board traffic/host ($p_{on} = 1 - p_{off}$)

Outputs:
N: number of host optochips on OPCB
N_r: number of nodes (routers utilized) on board
W_b: waveguides (pairs) within a waveguide bundle
U_{off}: number of router channels for off-board communication
topology: (mesh or torus or FCN) router-router topology on OPCB

```
/*phase 1: find all feasible OPCB designs*/
1.  for (increase N by 2, until N_max)
2.   for (increase N_r by 1, until current value of N)
3.    N_node ← N/N_r /*number of hosts connected to a router*/
4.    if (U_R – N_node· W_h < 0)||(N_node is not an integer)
5.     continue /*infeasible – not enough router channels*/
6.    endif
7.    Node_construct(s_r,s_h,r_i,r_o,N_node) /*node lay-out*/
8.    for (2 iterations:at 2nd swap node height with width)
9.     for (FCN, all meshes, all tori topologies of size N_r)
10.  W_b ← Rq_waveguides(S,topology,N,N_r,N_node,W_h,r,p_off)
11.    U_off ← U_R – N_node· W_h - W_b· topology_degree
12.    if N_r·U_off > U_B
13.     U_off ← ⌊U_B/N_r⌋ /*nodes share the available U_B*/
14.    endif
15.    Speedup_off ← (U_off · r)/(p_off · W_h · r · N_node)
16.    if (Speedup_off < S)
17.     continue /*infeasible: not enough offboard pinout*/
18.    endif
19.    for (all possible lay-outs: collinear and 2D)[2]
20.     if (A suffices) && (B ≥ worst case loss)
21.      (Keep OPCB design as feasible)
22.     endif
23.    endfor
24.   endfor
25.  endfor
26. endfor
27.endfor

/*phase 2: choose optimal OPCB design*/
28.  for (all feasible OPCB designs of phase 1)
29.   (Choose as optimal the design maximizing N and solve ties
      by maximizing N/N_r and then minimizing latency)
30.  endfor
```

A design that exhibits high Speedup would allow non idealities in the implementation (not perfect routing and flow control). Since the on-OPCB networks are small, we chose $S = 1$ as the minimum acceptable on- and off-board Speedup value. p_{on} is indirectly related with the size of the system. For uniform traffic $p_{on} = (N-1)/(N_{total}-1)$, where N_{total} is the total number of transceiver optochips in the system. The function in line 7 implements the node construction strategy described in Subsection 2.4. The function in line 10 calculates the number of waveguides required in a router-to-router bundle (the "fatness" of the links) for a given topology, in order to achieve on-board Speedup (equal to or) higher than S, using Eq. (3)-(6). In line 12, if the total required board pinout (from all routers) is greater than the available board pinout, then the board pins are equally distributed in all on-board routers. In the case of vertical cabling (U_B=inf), and this is not imposed. If the router pins for off-board do not suffice for its off-board communication (off-board Speedup < S: line 16), then the OPCB design is not feasible. Lay-outs generated in line 19 are based on the lay-out strategies presented in Subsection 2.3, taking into account whether vertical cabling is used. In line 20, worst case loss is identified as the maximum of the total loss of the worst row-wise and column-wise router-to-router waveguide. Loss calculations include the longest path length, the bends, and the crossings (to create the topology on the direction under study, plus meeting the vertical direction networks, plus the off-board cabling for column-wise networks), following Subsections 2.2 and 2.3.

5. APPLYING THE METHODOLOGY: PERFORMANCE RESULTS

In this section we apply our proposed methodology for OPCB design (presented in the previous Section) for specific and realistic device and module attributes. We focus on multi-mode optical

[1] Depends on processor computational power and communication -to-computation ratio

[2] Function of node-height-and-width, N_r, topology, W_b, θ, r_i, s_w, w_p

interconnection modules, since at this point they are more mature than single-mode modules. However, our topology lay-out strategy and methodology can be used for both multi- and single-mode OPCBs.

First, we list the specifications of the multi-mode modules used as a baseline scenario, most of them driven by Phoxtrot [8] so as to have realistic device and module attributes. Then, we examine the potential benefits of photonic technological advancements, such as smaller module footprints, smaller bending radiuses and smaller crossing angles on the required area using our lay-out approach (Subsection 5.1). In Subsection 5.2 we apply our proposed methodology for OPCB design, using the ATDT, for the specific and realistic device and module attributes described below. Finally, in Subsections 5.3, 5.4, 5.5 we examine the impact of board pinout, off-board communication schemes and router pinout respectively on-OPCB network design.

Polymer Multi-mode Waveguides. The size of the polymer waveguides assumed as baseline is 50μm x 50μm, with a minimum parallel separation (waveguide pitch) of 250μm. The propagation loss is L_p=0.05dB/cm for 850nm wavelength. We assume two optical layers of waveguides, one layer for each communication direction. Bending radius values for polymer waveguides at (around) 850nm and corresponding losses can be found in [22]. Based on that, we assume r_o=20 mm for inter-node communication with L_b=1 dB loss per bend, and r_i=10 mm for intra-node. Optical waveguides allow crossings with various crossing angles [22, 23]. As baseline we assumed θ=90^o crossing with L_c=0.023dB loss, but we will however examine the impact of smaller crossing angles, on lay-out area (Subsection 5.1) – assuming that crosstalk is not an issue.

Router chip. We assume an optoelectronic packet switching router chip that provides U_R=168 Tx (VCSELs) and 168 Rx (PDs) elements at r=8Gbps channel rate. It is actually an electronic chip with embedded parallel optical interfaces. The Tx and Rx modules are arranged in two 12 x 14 matrices. The optoelectronic router chip footprint (on-board) is s_r x s_r = 52mm x 52 mm. These specifications correspond to a commercially available router chip, already used in actual network products. The state-of-the-art commercial application of the chip was realized by directly connecting fiber cables to the optical Tx and Rx interfaces (vertical cabling). Within PHOXTROT the goal is to integrate the chip on-board to realize multi-mode polymer waveguide OPCBs. The chip has electrical SerDes in 25 Gbps for the processor-to-router connections. However, in this work we will assume that the processor-to-router connections are realized optically using the Tx/Rx interfaces (see host optochip below). The router chip-to-board integration is still under active research. We will first assume that all channel pins are available, using all four sides of the router: 42 Rx and 42 Tx channels per router side (baseline scenario). We will also make more conservative assumptions (current status): 12 Tx and 12 Rx available from each side, (48 bidirectional channels in total).

Host optochip. A host optochip is an active Tx/Rx interface module on top of which the processors or memory modules will be located. For these studies, we assume that the host optochips will accommodate only processors, making the simplifying assumption that processor-to-memory connections are realized electrically in a separate layer, not affecting the optical layer in terms of area. The channel rate is r=8Gbps. Regarding the optochip's footprint on-board, we will assume optochip footprint to be s_h x s_h=52mm x 52 mm (equal to the router). The number of channels we will assume is W_h=12 (assuming processor chips of 1 TFLOPS – as Intel Xeon Phi 3100 – and communication-to-computation ratio equal to 0.1 bps/FLOPs).

Power Budget. The transmitters assumed are VCSELs operating at 850 nm, using $P_{VCSEL} = 4.7\ dBm$ power. The receivers are PDs with sensitivity $PD_{sens} = -13dBm$. We assume that chip-to-board and board-to-chip couplings are realized with a microlens system of mirrors (MLA). Egress and input lens losses (VCSEL-to-mirror and mirror-to-PD) are 1 – 2dB, and waveguide input and egress facet losses (mirror-to-waveguideand waveguide-to-mirror) are 1 – 3 dB. Thus, assuming a total 3dB loss for chip-to-waveguide and waveguide-to-chip couplings, we have a power budget of: $B=P_{VCSEL}-P_{couplings}-PD_{sens}$=11.7dBm. This is the power budget for a single optical waveguide on-board, connecting either two router chips or a router chip with a host optochip. This budget can be *spent* on lay-out-related losses, that is, the waveguide length, the bends and crossings.

5.1 Topology lay-out strategy and required 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 smaller chip footprints (s_h = s_r = 26 mm, and s_h = s_r = 10 mm), smaller bending radiuses (r_o = r_i = 10 mm and 1mm - for both intra- and inter-node connections), and smaller crossing angles (θ = 60^o, 45^o) on the required area. Note that some of these values are unrealistic and are used as the reference/extreme scenario, in an attempt to understand the effect of them on the layout area. Such small bending radius and chip sizes would make the waveguide pitch relevant if there were many waveguides within a waveguide bundle/track and many waveguide tracks – large topologies (not the case for the simulation results of sections 5.2, 5.3), but we neglect such issues here. The topology we chose is a 4 x 4 torus, laid out in a 2-D (4 x 4) fashion, where every router accommodates N_{node}=4 optochips. U_{off}=2 router channels are used for off-board and W_h=2 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. The first column contains the total area required (in mm x mm) for a single node (a single router and 4 optochips). The second column contains the total area required (mm x mm) for the 2-D lay-out of the 4 x 4 torus. The third column shows the improvement in the total-layout area as a result of the modification of the related single lay-out parameter (chip footprints, bending radiuses, crossing angles). The fourth column contains the *lay-out area efficiency* values. We define lay-out efficiency as the ratio of the total area of the chips (routers and hosts) to the total lay-out area:

$$a = \frac{chips\ area}{total\ layout\ area} \qquad (8)$$

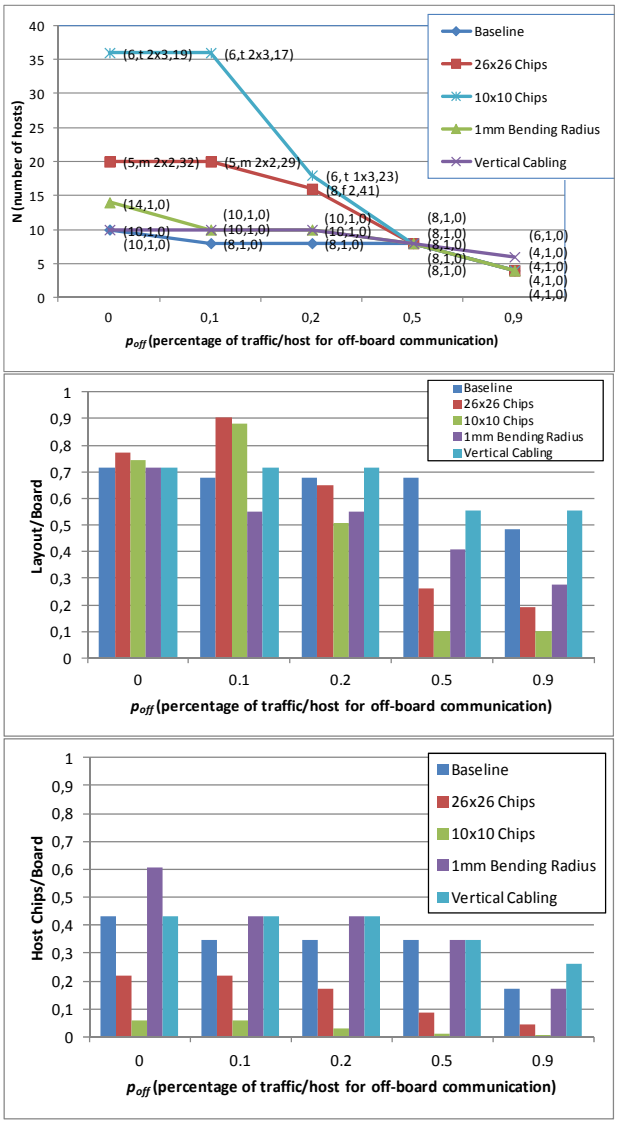

Figure 4. (a) Number of hosts on-OPCB, (b) α_b, and (c) α_h, for board pinout U_B = 96, varying the percentage of off-board traffic (p_{off}).

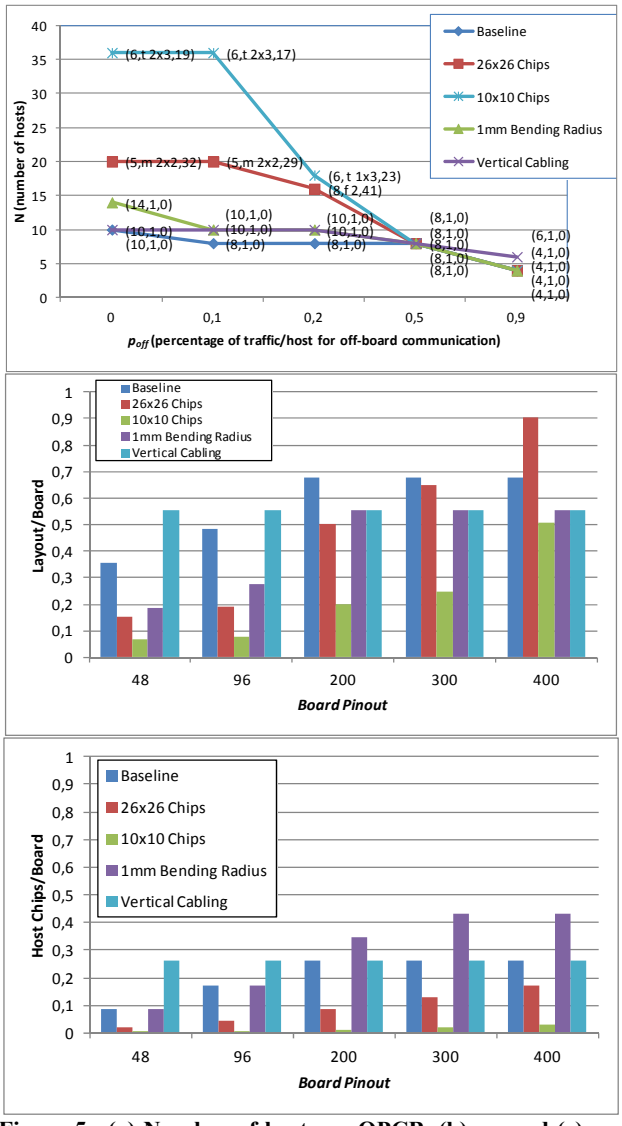

Figure 5. (a) Number of hosts on-OPCB, (b) α_b, and (c) α_h, for p_{off} = 0.90 off-board traffic, varying the board pinout.

Different lay-out strategies for a given topology would result in different α values. Eq. (8) would be equal to 1 in the ideal lay-out scenario where the lay-out of a topology would require the same area as the total area of the chips.

Node areas are rectangles since a node contains an odd number of chips (4 hosts and 1 router). The 50mm x 50mm square area in the 10mm x 10mm chip size case, is due to host-to-router bending radius (also 10mm). The total area in that case it is a 362mm x 442mm 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. Using 10mm bending radiuses also does not reduce node area, since, in the baseline scenario r_i was also set to 10mm. As it can be seen in Table I, all aforementioned improvements in OPCB technologies lead to reduced required area. However, it is clear that the greatest benefit regarding the required area can be obtained by reducing the chip footprints. The impact of the utilization of half size chips (26mm x 26mm) is somewhat similar to the impact of the (extremely aggressive) assumption of 1mm bending radius. Having 10mm x 10mm chips (the footprint of the single-mode all-optical router developed in PHOXTROT) leads to less required area than the 1mm bending radius. The lay-out area efficiency (metric a, see Eq. (8)) increases by reducing the "layout-related" overheads, namely bending radiuses and crossing angles. On the other hand, reducing chip sizes leads to reduction of the lay-out efficiency, since although the total area is reduced, a larger portion of the layout area is the "overhead", that is, it is empty. The greatest benefits for lay-out efficiency come from reducing the bending radiuses.

5.2 On-OPCB methodology and off-board traffic

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.

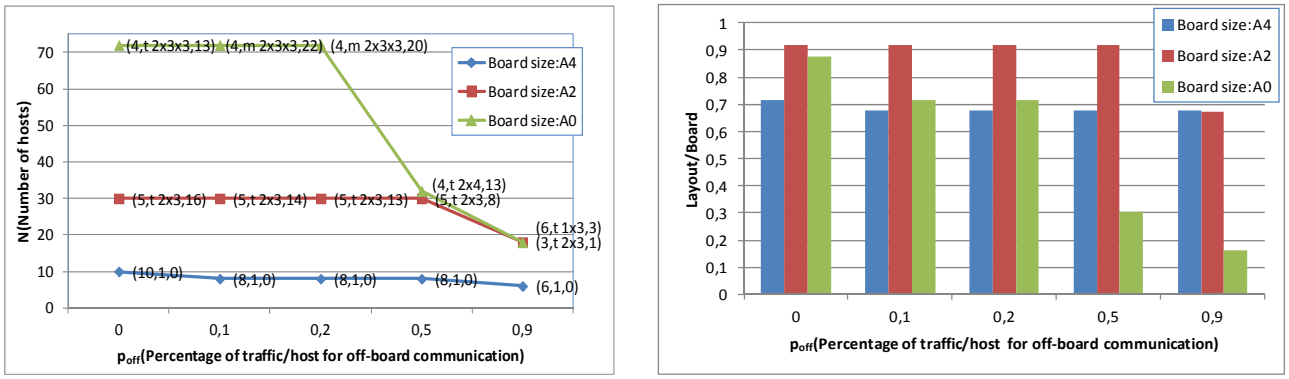

Figure 6. (a) Impact of board size on-OPCB network design using waveguides for off-board communication. (b) Related a_b metric.

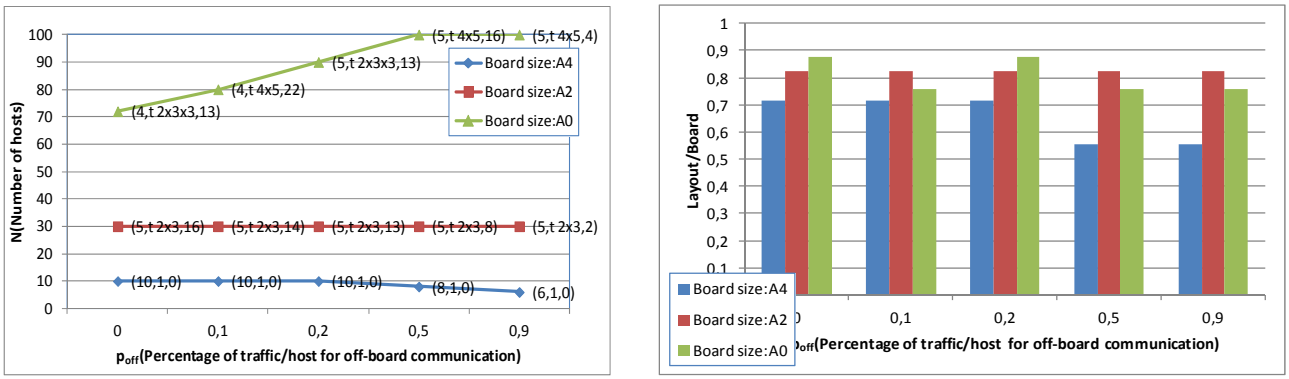

Figure 7. (a) Impact of board size on-OPCB network design using vertical cabling for off-board communication. (b) Related a_b metric.

TABLE I: Impact of smaller chip footprints, waveguide bending radiuses and crossings angles in layout area.

	Node area	Lay-out	%	α
Baseline	176x134	698x946	-	0.328
Chips with s_h=s_r=26	98x182	490x634	53	0.174
Chips *with* s_h=s_r=10	50x50	362x442	75.8	0.050
r_0 =r_i=10mm	176x134	618x826	22.7	0.424
r_0 =r_i =1mm	158x107	438x646	57.1	0.765
θ=60°	176x134	658x906	9.7	0.363
θ=45°	176x134	641x889	13.7	0.380

We assume board area equal to A4 paper size (297mm x 210mm) and board pinout U_B=96 (PHOXTROT's target for multi-mode OPCBs). In what follows, by board size we actually refer to the board area available for the optical layer. The rest baseline parameters were described in the beginning of this Section. 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. For example, in Figure 4(a), the first point (6, *t* 2x3, 19) for the 10mm x 10mm Chips scenario denotes a 2x3 torus network. A single node in this network is comprised of a router and 6 optochips. 19 bidirectional channels (19 Tx and 19 Rx waveguides) are used for router-to-router connections. The (10, 1, 0) point for the baseline scenario denotes a single node network: 1 router with 10 hosts connected (for single node networks we skip using any of the "*t*", "*m*" or "*f*" symbols). We also define two new metrics, the *board utilization* a_b and the *hosts' board utilization* a_h as follows:

$$a_b = \frac{layout\ area}{board\ area} \quad (9), \quad a_h = \frac{host\ chips\ area}{board\ area} \quad (10)$$

In Figure 4(a) we present the resulting designs by varying the percentage of off-board destined traffic per host p_{off}. This can be viewed as examining boards destined for systems with different total sizes. We compare the baseline scenario with scenarios utilizing: (i) and (ii) smaller chips (26mm x 26mm and 10mm x 10mm, respectively), (iii) smaller bending radiuses (1 mm for both intra- and inter-node connections) assuming 1 dB loss (equal to 20mm radius loss – an extreme assumption, but made in order to examine the impact of ideally small bending radiuses) and (iv) vertical cabling. In vertical cabling scheme, off-board communication takes place through fiber cables connected to the routers, not through waveguides, leading to fewer crossings and thus smaller losses, while board pinout U_B is neglected. Figures 4(b) and (c) depict the related a_b and a_h values, respectively.

As depicted in Figure 4(a), 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 percentage p_{off} equal and higher to 0.5, board pinout becomes the bottleneck, reducing the number of hosts that can be accommodated. The usage of smaller chips or radiuses would allow more modules on board. This is also apparent from Figure 4(b), where at p_{off}=0.5 the utilization drops

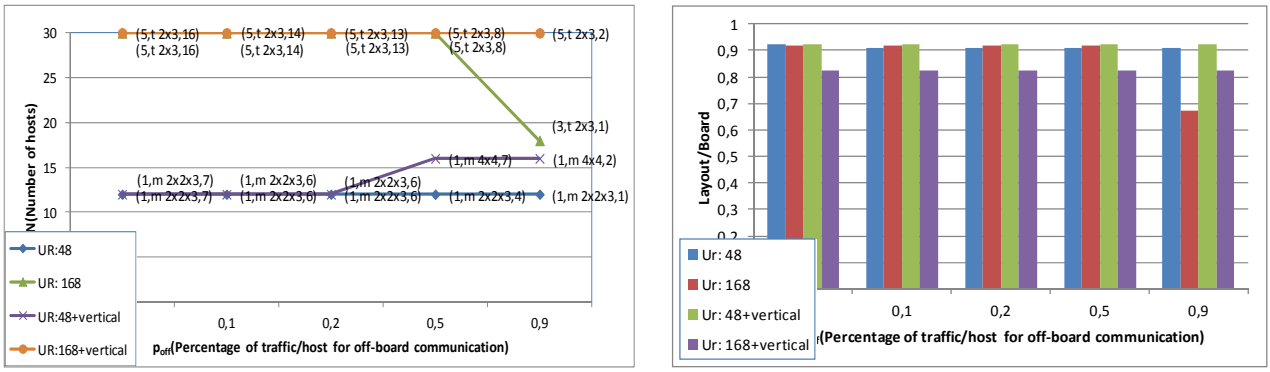

Figure 8. (a) Impact of less available router pinout on-OPCB network design (board size: A2). (b) Related a_b metric.

dramatically for scenarios (i), (ii) and (iii). This also affects a_h (Figure 4(c)) which also drops as p_{off} increases. The "spike" in Figure 4(b), at p_{off}=0.1 off-board traffic for 26mm x 26mm and 10mm x 10mm chips is due to the extra waveguide tracks required for off-board communication (not needed for at p_{off}=0 off-board traffic). For the vertical cabling case (scenario (iv)) 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). Regarding the resulting topologies, mesh and torus networks are more suitable when a large number of routers is needed (provided that there is enough available area and board pinout to be accommodated on board), since FCNs have greater connectivity degree (more demanding on router-to-router channels) and are laid-out only in collinear fashion.

5.3 Impact of board pinout

In Figure 5(a) we examine the same scenarios, but keep constant the required off-board traffic p_{off} (and in particular p_{off}=0.9) and vary the board pinout U_B. Figures 5(b) and (c) depict the related a_b and a_h values, respectively. U_B=48 pinout is the state-of-the-art for OPCBs, while U_B=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: 144x154 is the total lay-out area for the (2, 1, 0) baseline. PHOXTROT's targeted 96-pinout board slightly improves that. A 200-pin OPCB would allow more hosts on board, allowing at the same time to harvest the area benefits that can be obtained from smaller chips and smaller bending radiuses. A far larger board pinout (400) and the use of 10mm x 10mm chips would allow denser integration (151x230) and more efficient usage of board area. This is also apparent from Figures 5(b) and (c), where a_b and a_h tend to increase as board pinout increases.

5.4 Impact of board area and off-board communication schemes

Figures 6(a), 7(a) illustrate the impact of off-board communication schemes (waveguided or vertical cabling) and board sizes on OPCB network design, while Figures 6(b), 7(b) depict the relative a_b values. The board pinout value used is U_B=400. For the waveguided off-board communication, as p_{off} grows, less hosts can be accommodated due to pinout constraints (for $p_{off} \geq 0.5$). On the contrary, in off-board communication via cables, the board pinout constraint is not imposed. Thus, as available board area increases, the number of on-OPCB hosts tend to increase. For the baseline scenario in Figure 7(a), the constraint is the available board area: as p_{off} increases, more routers are required in order to accommodate the increasing off-board destined traffic. However, they can not be accommodated due to OPCB area constraints.

Thus, since a single router can not satisfy the increased off-board traffic requirements, fewer hosts must be connected to the single router. Board utilization (Figure 7(b)) does not necessarily increase as p_{off} increases in vertical cabling scheme: the lay-out of a topology using cabling for off-board communication is more lay-out area efficient (8) than the lay-out of the same topology using waveguided communication for off-board cabling, since the former does not introduce on-OPCB "lay-out overhead" due to off-board waveguides. Furthermore, the same router-to-router topology could accommodate a larger number of hosts with the same board utilization: eg topologies (4, t 4x5, 22) and (5, t 4x5, 16) use about the same portion of the board. This is due to blank positions in the intra-node 2D array (Subsection 2.4) that in the latter case are filled with hosts.

5.5 Impact of router pinout

Figure 8(a) illustrates the impact of available router pinout U_R on-OPCB network design, for both vertical cabling and waveguided off-board communication, while Figure 8(b) depicts the related board utilization. The board pinout value used is U_B=400 and board area is equal to A2. Apparently, larger router pinout leads to more hosts on-OPCB in all examined cases.

6. CONCLUSIONS

We proposed lay-out strategies for on-optical printed circuit board (OPCB), and we also presented a general methodology for designing optical interconnection networks/architectures, using a set of packaging and required performance parameters as inputs. Our methodology incorporates the lay-out strategies we proposed but it can also be enriched with more lay-out strategies. The topology design methodology consists of two phases. In the first phase we generate all the feasible designs (in terms of area, losses and performance) within the topology families examined, using our proposed OPCBs lay-outs. In the second phase, we select the optimal designs based on specific optimization criteria. We applied our methodology for the on-board level of packaging hierarchy using PHOXTROT subsystem specifications as input. 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. Our future work includes the

expansion of our methodology for higher packaging layers and the incorporation of WDM, and enriching the topology families with bus like topologies. We also plan to evaluate our designs performance under realistic traffic patterns using simulations.

7. ACKNOWLEDGMENTS

This work was supported by the European Commission through the FP7 ICT-PHOXTROT (ICT 318240) project.

8. REFERENCES

[1] J. H. Collet, F. Caignet, F. Sellaye, and D. Litaize, "Performance constraints for onchip optical interconnects," *IEEE J. Sel. Topics in Quantum Electron.,* vol. 9, no. 2, pp. 425−432, Mar./Apr. 2003.

[2] G. Astfalk, "Why optical data communications and why now?," *Appl. Phys., vol. A 95*, no. 4, pp. 933−940, 2009.

[3] M. Taubenblatt, "Optical interconnects for high performance computing," *in Proc. Optical Fiber Communications (OFC/NFOEC)*, Los Angeles, CA, 2011, paper OThH3.

[4] C. L. Schow et. al., "A 24-Channel, 300 Gb/s, 8.2 pJ/bit, Full-Duplex Fiber-Coupled Optical Transceiver Module Based on a Single "Holey" CMOS IC", *IEEE/OSA J. Lightwav. Technol.*, vol. 29, no. 4, Feb. 2011

[5] http://lightwave.ee.columbia.edu/?s=partners&p=funding_agencies

[6] http://www.intel.com/content/www/us/en/research/intel-labs-hybrid-silicon-laser.html

[7] http://www.ict-polysys.eu/

[8] http://www.phoxtrot.eu/

[9] Kachris, Christoforos, and Ioannis Tomkos. "A survey on optical interconnects for data centers." *Communications Surveys & Tutorials, IEEE* 14.4 (2012): 1021-1036.

[10] Basak, Debashis, and Dhabaleswar K. Panda. "Designing clustered multiprocessor systems under packaging and technological advancements." *Parallel and Distributed Systems, IEEE Transactions on 7.9 (1996):* 962-978.

[11] Gupta, Amit K., and William J. Dally. "Topology optimization of interconnection networks." *Computer Architecture Letters* 5.1 (2006): 10-13.

[12] Chen, Dong, et al. "Looking under the hood of the ibm blue gene/q network." *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. IEEE Computer Society Press, 2012.

[13] Abts, Dennis. "Cray XT4 and Seastar 3-D Torus Interconnect." *Encyclopedia of Parallel Computing* (2011): 470-477.

[14] Ajima, Y., Takagi, Y., Inoue, T., Hiramoto, S., Shimizu, T. "The tofu interconnect." High Performance Interconnects (HOTI), 2011 *IEEE 19th Annual Symposium on*. IEEE, 2011.

[15] W. J. Dally and B. Towles. "Principles and Practices of Interconnection Networks", *Morgan Kaufmann*, 2004.

[16] Aroca, Jordi Arjona, and Antonio Fernández Anta. "Bisection (band) width of product networks with application to data centers." *Theory and Applications of Models of Computation. Springer Berlin Heidelberg*, 2012. 461-472.

[17] Grange, Matt, et al. "Optimal network architectures for minimizing average distance in k-ary n-dimensional mesh networks." *Networks on Chip (NoCS), 2011 Fifth IEEE/ACM International Symposium on*. IEEE, 2011.

[18] X. Dou *et al.*, "Optical bus waveguide metallic hard mold fabrication with opposite 45◦ micro-mirrors," *in Proc. Optoelectron. Interconnects Compon. Integr. IX*, 2010, pp. 76070P-1–76070P-6.

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

[20] J. Beals *et al.*, "A terabit capacity passive polymer optical backplane based on a novel meshed waveguide architecture," *Appl. Phys. A: Mater. Sci. Process.*, vol. 95, no. 4, pp. 983–988, 2009.

[21] Bamiedakis, N., Hashim, A. , Penty, R.V. , White, I.H. "A 40 Gb/s Optical Bus for Optical Backplane Interconnections", *Lightwave Technology*, vol.32 , issue: 8, Apr. 2014.

[22] Wang, Kai, et al. "Photolithographically manufactured acrylate polymer multimode optical waveguide loss design rules." *Electronics System-Integration Technology Conference, 2008. ESTC 2008. 2nd*. IEEE, 2008.

[23] A Hashim, N Bamiedakis, RV Penty, "Multimode Polymer Waveguide Components for Complex On-Board Optical Topologies." *Journal of Lightwave Technology* 31.24 (2013): 3962-3969.

[24] Pepeljugoski, Petar K., et al. "Low power and high density optical interconnects for future supercomputers." *Optical Fiber Communication Conference*. Optical Society of America, 2010.

[25] http://www.top500.org/system/177232#.U3mxKnbm5Gk

[26] Make IT Green: Cloud Computing and its Contribution to Climate Change. Greenpeace International, 2010.

[27] Yeh, C.-H., E.A. Varvarigos, and B. Parhami, "Multilayer VLSI lay-out for interconnection networks," *Proc. Int'l Conf. Parallel Processing*, 2000, pp. 33-40.

[28] Bamiedakis, Nikolaos, et al. "Cost-effective multimode polymer waveguides for high-speed on-board optical interconnects." *Quantum Electronics, IEEE Journal of* 45.4 (2009): 415-424.

[29] Haurylau, Mikhail, et al. "On-chip optical interconnect roadmap: challenges and critical directions." *Selected Topics in Quantum Electronics, IEEE Journal of* 12.6 (2006): 1699-1705.