The role of optical interconnects in the design of data center architectures

15

A. Siokis[1,2], K. Christodoulopoulos[1,2] and E. Varvarigos[1,2]
[1]University of Patras, Patras, Greece, [2]Computer Technology Institute and Press – Diophantus, Patras, Greece

15.1 Introduction

Data centers (DCs) experience exponential increases in traffic volumes both in their connection to the end-user and also in the internal communication among servers. This trend is due to both evolution in processor technology and the expansion of the Internet, amplified by the ever-increasing use of wireless and cellular networks and the related information-centric services and applications for these platforms. A particularly interesting observation is that the majority of the DC traffic (76%) stays within the DC [1]. The traditional fat-tree topology built out of electronic switches [2] presents several problems, since it scales superlinearly to the number of servers and servers' rate, leading to increased requirements for switching equipment and power consumption.

Optical technology is a promising, energy-efficient solution for satisfying the increased bandwidth requirements for both telecoms and datacoms. In telecoms, optical fibers that were widely used in long-haul networks have now replaced most of the copper technology in WAN and MAN, and are gradually finding their way to datacom networks *inside* the DCs [3]. Several optical interconnection solutions have been proposed for specific parts of the DC network, while more ambitious holistic all-optical architectures are being researched. The first applications of optics are for rack-to-rack or server-to-switches communications. Since optics can once more be the solution to the bandwidth and energy problems for next generation DCs, their adaptation at the lower layers of the packaging hierarchy, including board-to-board, on-board, and on-chip, is impending. A promising technology are the Optical Printed Circuit Boards (OPCBs): boards with integrated optical waveguides can be used at the on-board and board-to-board (backplane-boards) levels to interconnect optically-enabled modules, such as opto-electronic chips packed with vertical-cavity surface-emitting laser (VCSEL), and photodiodes (PD). Laying out topologies via optical waveguides on-OPCBs presents a number of issues that have to be addressed when designing architectures for DCs.

In this chapter we outline the similarities and differences between layout models for electrical interconnects and optical waveguided communications, in order to

Optical Interconnects for Data Centers. DOI: http://dx.doi.org/10.1016/B978-0-08-100512-5.00015-2

understand the peculiarities of the latter. Taking these into account, we then present a layout model suitable for optical interconnects. We also describe strategies that can be used for laying out logical topologies on-OPCBs, assuming the aforementioned model.

15.2 Overview of optical interconnects technologies

Optical networks have been widely used in the long-haul and metropolitan telecom networks (MAN) providing low power and latency, and increased throughput. Initially, the copper channels were replaced by fiber for point-to-point communication to form fat circuit pipes. In this case, opto-electro-optical regeneration took place at every network node. Long-haul and MAN networks have now evolved to all-optical (but still circuit switched) approaches to avoid power hungry conversions from the electrical to the optical domain, and vice versa. The use of optics has been extended to cover smaller distances in LAN as well as DC networks. Currently, optics have replaced electrical links between Top-of-Rack switches, to achieve higher bandwidth, reducing the power consumption and latency somewhat. Even so, power consumption of data communication is still daunting. In order to cope with both the energy and bandwidth limitations of the electrical interconnects, optical technologies have to be deployed at even shorter distances in the near future: optics are gradually becoming more cost-effective for board-to-board, on-board, and even on-chip communications.

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, chip-to-board coupling technologies, optical transceiver chips (equipped e.g., with VCSELs—Vertical Cavity Surface-Emitting Laser for Tx, and PDs—PhotoDiodes for Rx), photonic switching and routing elements, (de)multiplexing elements, Wavelength Selective Switches (WSS), Arrayed Waveguide Gratings (AWGs), and optical RAMs, among others.

15.3 Applications of optical interconnects in the individual layers of the packaging hierarchy

In this section we review optical interconnection network architectures presented in the literature for all the layers of the packaging hierarchy: rack-to-rack, on-board, board-to-board, and on-chip.

15.3.1 Rack-to-rack interconnections

The proposed rack-to-rack optical interconnects architectures for DCs fall into two categories: (1) hybrid approaches that enhance the legacy DC architecture with

optical interconnects, and (2) optical switch architectures targeting higher radices in order to lead to flatter DC architectures (such as fat trees with fewer tiers).

Regarding the first approach, two well-known hybrid architectures have been proposed that rely on both electrical (commodity) packet switches and optical circuit switches, as proposed by Farrington et al. [4] and Wang et al. [5]. In the approach of Wang et al. [5], the Top-of-Rack switches are connected both to an electrical packet-based network (based on commodity switches), and to an optical circuit-based network. The optical switch must be configured so as to connect pairs of racks with high bandwidth demands through this optical switch. In Farrington et al. [4], a similar approach is followed, but wavelength division multiplexed (WDM) links are used for the optical circuits. The electrical packet switches are used for all-to-all communication of the pod switches, while the optical circuit switches are used for high bandwidth, slowly changing (and usually long lived) communication between the pod switches.

A number of optical switch architectures have been proposed for all-optical (non-hybrid) communication between racks in the DCs. Singla et al. [6] proposed an architecture based on Wavelength Selective Switches (WSS) and Micro-Electro-Mechanical Systems Switches (MEMS). Each port has several optical transceivers operating at different wavelengths. The optical wavelengths are combined using a multiplexer, the output port of which is routed to a WSS. The outputs of the WSS are connected in the MEMS optical switch. At the output stage (after the MEMS), all of the wavelengths are demultiplexed and routed to the optical transceiver in the output port. The architecture proposed by Luijten et al. [7] is based on wavelength- and space-division multiplexing taking place in two different stages. In the first stage, multiple wavelengths are multiplexed in a common WDM line and are broadcast to all the modules of the second stage through a coupler. The second stage uses SOAs (Semiconductor Optical Amplifiers) as fiber-selector gates to select the wavelength that will be forwarded to the output. Another approach based on the combination of wavelength- and space-division multiplexing is proposed by Castoldi et al. [8], in which a Space-Time (ST) switched architecture is also proposed. Shacham and Bergman [9] proposed an architecture based on 2×2 SOA-based switches that can be scaled efficiently in tree-based topologies. Zhu et al. [10] proposed an optical switching platform, the key idea of which is the combination of Wavelength Division Multiplexing (WDM) and Space Division Multiplexing (SDM) utilizing an $N \times 1$ Wavelength Selective Switch (WSS). A number of high radix optical switch architectures have been proposed based on Arrayed Waveguide Grating Routing (AWGR) elements. Xia et al. [11] proposed a switch architecture based on a three-stage AWGR-based Clos network, using tunable wavelength converters. The architecture proposed by Ye et al. [12] consists of an array of tunable wavelength converters (one TWC for each node), an AWGR, and a loopback shared buffer. Each port can connect with any other port through the AWGR by configuring the transmitting wavelength of the tunable wavelength converter. Gripp et al. [13] proposed a three-stage architecture using AWGR. The first stage is a space switch based on AWGRs that distributes the packets uniformly across the ports of the second stage. The second stage, a time switch,

holds the packets until the third stage; another round-robin space switch based on AWGRs provides a path to the output port. Proietti et al. [14] investigated the scalability issues in the AWGR-based interconnect architectures and pursued an active AWGR switch architecture assuming a distributed control plane.

15.3.2 On-board and board-to-board interconnections

Optical interconnects for on-OPCB and OPCB-to-OPCB levels of the packaging hierarchy is an active research field, which includes research on a wide range of technologies such as optical waveguides (single-mode and/or multi-mode) of various materials (polymer, glass, ...), optical transmitters (e.g., VCSELs), various coupling techniques, etc. Architecture-wise a number of (mainly passive) interconnection architectures have been proposed for use in optical backplanes such as large parallel waveguide arrays [15], a waveguide-based optical bus structure [16], meshed waveguide architectures [17] and [18], a shared optical bus [19], and a regenerative bus structure of 40 Gbps [20]. On-OPCB architecture is a research area that will probably attract more interest in the future in order to define next generation optical DC architectures spanning from the higher/rack packaging levels to the lower ones. OPCB is the packaging level on which we will focus in the remaining of this chapter, beginning with Section 15.4.

15.3.3 Networks-on-Chip

Networks-on-Chip (NoC) constitute the lowest layer of the network system, which can benefit greatly from the introduction of photonics [21]. A great deal of research effort has been put into this area. A number of traditional topologies have been implemented for NoC environments, such as buses [22−24], crossbars [25,26], Butterfly and variants [27−29], torus [21], mesh [30], Clos [31,32], and fat-tree [33] topologies.

15.4 On-optical printed circuit boards (OPCB) layout strategies

In this section we focus on the on-OPCB layer of packaging hierarchy. First, in Section 15.4.1, we briefly discuss models presented in the literature for laying out logical topologies on electrical PCBs. In Section 15.4.2 we outline the differences between these models and the characteristics of optical waveguided communication, in order to adjust the former to a model suitable for laying-out topologies on OPCBs. Based on these modifications, we present simple layout strategies that can be used for both point-to-point as well as multi-point networks. In Section 15.4.3 we give some examples using the presented layout techniques for a number of topologies.

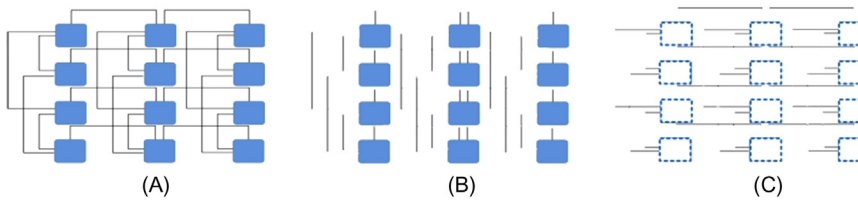

Figure 15.1 2D (3 × 4) Lay-out of a 3 × 2 × 2 mesh using the Thompson model. (A) Both layers. (B) Layer 1 (chips and vertical wiring). (C) Layer 2 (horizontal wiring).

15.4.1 Layout strategies for electrical interconnections on boards

Classic layout models for electrical interconnection networks rely on the Thompson model and variants [34]. In the Thompson model the interconnection network is modeled as a graph whose nodes represent processing elements and whose edges represent wires/links. The graph is mapped on a 2D grid. The wires run either horizontally or vertically along the grid lines, called tracks. The Thompson model assumes two wiring layers: one layer is used to lay out the horizontal segments of the wires and the other one the vertical segments. When a wire makes a turn (a 90 degree bend), the horizontal and vertical segments in the two layers are interconnected using inter-layer connectors (vias). The required area is the area of the smallest rectangle in the 2D grid containing all nodes and wires. An example of a layout using the Thompson model is depicted in Fig. 15.1. The Thompson model has been generalized from two wiring layers to the multilayer (layer > 2) 2D grid model. In the multilayer 2D grid model all the nodes are located in a single layer (active layer), while this first layer and the remaining layers contain only wiring. The multilayer 2D grid model has been further extended to the multilayer 3D grid model where the nodes of the network are embedded in more than one layer.

Yeh et al. [34] and [35] present a variety of layouts for various topologies based on the aforementioned models. In the following section we will examine (and adjust for OPCBs) two such network topology layouts: collinear and 2D. In the former, all network nodes are placed along a line, while in the latter nodes are placed along rows and columns, forming a 2D grid array. Note that in both cases the wires run on 2D grid lines. Fig. 15.2B depicts an example of a 3 × 2 × 2 mesh, laid out in a 2D grid of 3 × 4 nodes, with wires also laid out in a 2D grid. Note that the wiring, although depicted in one layer, is done in two (or more) layers. 2D layouts are constructed using collinear layouts along the rows and columns. A single row of the 2D layout in Fig. 15.2A is a collinear layout of three nodes, requiring one wiring track. A single column of the 2D layout is a collinear layout of four nodes (2 × 2), requiring three wiring tracks.

15.4.2 Layout strategies for OPCB

In this section we present layout strategies for point-to-point and multi-point interconnection networks on OPCBs. A node of the network topology could be either

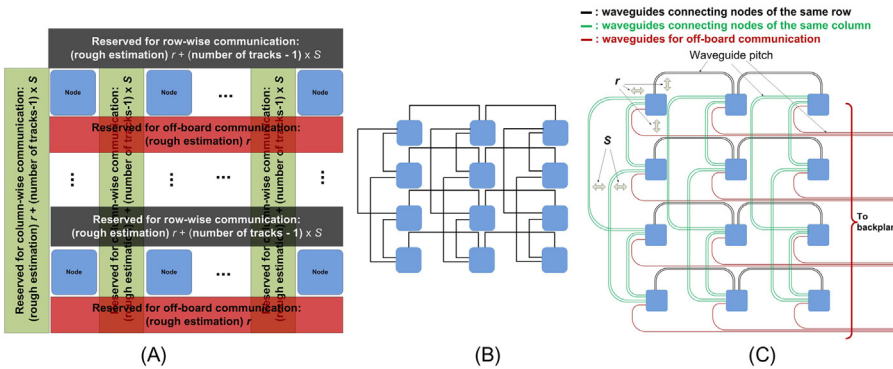

Figure 15.2 (A) Layout design rules on 2D grid for OPCBs. Space reserved for row-wise, column-wise and off-board communication. (B) 2D (3×4) Lay-out of a $3 \times 2 \times 2$ mesh. (C) The same layout on an OPCB, following the strategy shown in (A).

a single chip or a group of chips, e.g., a number of optical host chips connected in an optical/opto-electronic router forming a star network (see Siokis et al. [36]). Assuming only collinear layouts, the nodes could also be assumed to be the coupling points of linecards and the 2D grid surface, the optical backplane. In the text that follows we will view a node as a square of size d (the layout strategies can be easily generalized for nodes in the shape of rectangles). We will examine layouts both for point-to-point topologies and multi-point topologies, and we will also discuss how WDM can be used in order to implement point-to-point topologies using multi-point layouts. All the layouts will be presented without length matching. If length matching is required (to meet the requirements of the chip-to-chip protocol in timing skew) when more than one waveguide connects two nodes, additional small S-bends can be used in the shorter waveguide to even up the length of the waveguides (see also discussion in Siokis et al. [36]).

15.4.2.1 Layouts for point-to-point topologies

The main differences between optical waveguided communication and the models described for copper interconnects in Section 15.4.1, from the layout point of view, are:

1. Waveguide bends require a (non-sharp) bending radius r in order to allow the propagation of light. Smaller r means more losses. In electrical interconnects a bend with $r \approx 0$ is possible (in the Thompson model it is implemented as an inter-layer connection though a via).
2. Crossings are allowed in the same layer (a crossing angle of 90 degree is preferable due to losses and crosstalk). Crossings in the same layer are not possible in the electrical interconnects, since this would lead to a closed circuit.

The layout strategies described in Section 15.4.1 can be applied on OPCBs with the following modifications. We assume two layers for waveguide routing, each for

one direction of communication between nodes: so for each communicating nodes the first layer implements the Tx→Rx connection, and the second (almost identical) layer the Rx→Tx connection. This two-layer approach does not impose important restrictions on the placement of the Tx and Rx elements on the node. For example, assuming separate arrays of Tx and Rx elements on-chip, the Tx and Rx arrays could be placed either side-by-side, or the Tx array right behind or in front of the Rx array. Given the collinear layout of nodes (remember that 2D layouts are constructed from row- and column-wise collinear layouts), at each layer the links are laid out in a 2D grid, bends have a given radius, and crossings are allowed to occur. Alternatively, a single layer can be used to accommodate both directions of communication (Tx→Rx links and Rx→Tx links) side-by-side in a single "waveguide track" or bundle (see below). This approach lends itself to an alternating TxRx|TxRx|... pinout placement of the Tx and Rx elements on the chip.

In a case where more than one link is needed between two nodes, and since bends are (space and loss) expensive, to save on area we route multi-waveguide links together, as bundles, in a single "waveguide track". The distance of waveguides within a track can be as low as 250 μm, considered as the standard pitch in our study, or higher as preferred. Since the bending radius r and the chips' sizes are at least two orders of magnitude larger than the standard pitch, we neglect track width in our calculations. The first track parallel to the collinear layout direction of nodes is placed at space r from the node, while the space S left between the following tracks is related to the desired waveguide crossing angle θ and the bending radius r as follows:

$$S = (1 - \cos\theta)r \tag{15.1}$$

Thus, according to Eq. (15.1), if 90 degree crossings are used, the track spacing equals the bending radius ($S = r$). Smaller bending radii 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 (assuming also an alternating TxRxTxRx... pin placement on the nodes).

Also note that in the adopted strategy the bends and crossings appear in a specific and deterministic order: for every waveguide, an initial bend (or bends) takes place, followed by all the crossings, followed by a final bend (or bends).

To layout a topology on an OPCB we reserve an area for row-, column-wise, and off-board communication. Our generalized approach for 2D grid layouts is depicted in Fig. 15.2A. It assumes that network nodes have pinouts from two of their sides for inter-node interconnection. For the communication of the nodes in the same row, we reserve the area above the nodes (black area in Fig. 15.2A). The required area depends on the number of waveguide tracks, which is determined by the row-wise collinear topology. For the communication of the nodes in the same column, we reserve the space left to the nodes (green space in Fig. 15.2A), again depending on the required tracks. Finally, for off-board communication we reserve the space beneath the nodes (red space in in Fig. 15.2A) that has a width

equal to r, 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 distance r from the nodes. If nodes use a single side for pinout, instead of the two sides assumed above, then the required area for waveguides will be the same, but more bends will be required. For simple collinear layouts, the proposed strategy is that of a single row of 2D, as depicted in Fig. 15.2A, but because no column-wise communication takes place, the required distance between nodes is $2r-d$ if $r \geq d/2$ (assuming that the waveguides originate from about the center of the chip) or 0 otherwise (nodes are positioned as close as possible next to each other). Fig. 15.2A also gives an estimation of the total required area. In Fig. 15.2C a 2D (3×4) layout of a $3 \times 2 \times 2$ mesh is depicted (equivalent to the network of Fig. 15.2B). Two waveguides form a bundle and are used within column and row tracks, while one waveguide/node is used for off-board communication. The off-board waveguide tracks can be omitted completely if the off-board communication takes place via vertical cabling. In the latter case, off-board routing is implemented using fiber optics (such as in Hasharoni et al. [37]). However, in racks containing a large number of boards with a large number of on-board modules, optical fibers across boards could lead to a cabling mess. Furthermore, the incorporation of on-OPCB waveguides for off-board communication would result in pluggable boards offering ease of installation.

The layout strategies outlined above based on the Thompson model follow, by definition, a X−Y routing approach (or Manhattan routing). A more general routing approach can be used by applying λ-geometry, where λ represents the number of possible routing directions and π/λ the routing angles allowed [38]. $\lambda = 2$, 3, and 4 correspond to the Manhattan architecture, Y-architecture, X-architecture, respectively. In the Manhattan architecture there are only vertical or horizontal routing options as described above (0, 90, 180, and 270 degrees). In Y-architecture (or hexagonal routing) and X-architecture (or octagonal routing) the routing options vary by 60 and 45 degrees respectively. These approaches are depicted in Fig. 15.3A−C. λ-geometry routing approaches with $\lambda > 2$ lead to alternative mesh architectures with higher connectivity degrees. Fig. 15.3D, E−F depict meshes based on these approaches, adjusted for OPCBs, assuming nodes with pinout from all four sides and non-unit side size. The four-side pinout allows the nodes of the regular 4×4 mesh to be placed as close as possible to each other (Fig. 15.3D). In Fig. 15.3E we present a generalization of the Y-mesh for arbitrary routing angles (normally $\theta = 60$ degree in 3-geometry) and in Fig. 15.3F a generalization of the X-mesh for arbitrary crossing angles (normally 90 degree crossings are present in 4-geometry). The required layout area can be estimated approximately using basic geometric shapes: isosceles triangles for the generalized Y-mesh and rectangles for the generalized X-mesh.

Redefining the routing grid in order to allow crossings of various crossing angles could lead to various such “extended” mesh architectures with even higher connectivity degrees. These extended architectures (using $\lambda > 2$) would require more area if implemented using Manhattan routing. For example, Fig. 15.3G depicts a X-mesh implemented using X-Y routing and 90 degree crossing angles.

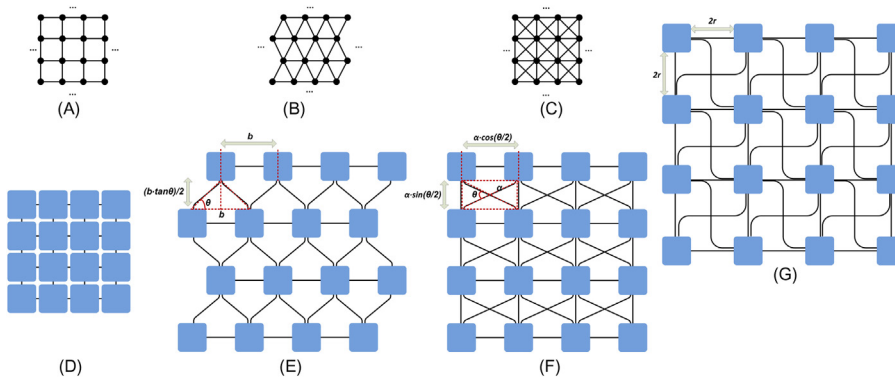

Figure 15.3 λ-geometry with: (A) $\lambda = 2$ (Manhattan routing), (B) $\lambda = 3$ (Y-routing), (C) $\lambda = 4$ (X-routing). Adjustment of λ-geometry approaches for OPCBs for 16 nodes with pinout from all four sides and non-unit side size: (D) 4×4 Mesh ($\lambda = 2$). (E) 4×4 generalized Y-Mesh (normally $\theta = 60$ degree in the Y-routing approach with $\lambda = 3$). (F) 4×4 generalized X-Mesh (normally $\theta = 90$ degree in the X-routing approach with $\lambda = 4$). (G) a X-Mesh implemented using Manhattan routing and crossing angles 90 degree.

If we set for simplicity $2r = d$, then this topology would require a $7d \times 7d$ area. Setting, for a fair comparison, crossing angle $\theta = 90$ degree (as in the $\lambda = 4$ routing approach) and $\alpha \cdot \sin(45 \text{ degree}) = d$ in the layout approach of Fig. 15.3F, the latter would require an area equal to $4d \times 7d$ (in the vertical dimension the nodes would be placed as close as possible next to each other). In principle, the allowance of crossings in the same layer and the reduced link-to-link separation (waveguide pitch), compared to electrical interconnects, allow denser integration and reduction of PCB thickness (layer count). However, a potential issue is crosstalk with respect to the crossing angle, for angles less than 90 degree . To the best of our knowledge there is not yet a design rule/analytical formula for crosstalk as a function of the crossing angle. Measurements for crosstalk can be found in the work of Bamiedakis et al. [20], but only for the examined bus architecture. Another manufacturing issue for OPCBs is that the performance of waveguide components depends on the launch conditions at the component input, e.g., whether light enters the waveguide using multi-mode MMF, or SMF, or mirrors for chip-to-board coupling. Furthermore, in multi-point topologies where splitters/combiners are used, it is possible for the light entering the first splitter along a multi-mode waveguide path to resemble light input from a well-aligned MMF, while the light entering the following splitters on the waveguide path resembles light input from a displaced MMF toward the bent output of the splitter [20]. In principle, splitters and combiners are the most expensive components, followed next by bends and finally by crossings. The layout strategies presented for both point-to-point and multi-point topologies in this chapter are detailed, requiring specific numbers of waveguide components, appearing in a specific order on the waveguide paths, while at the same time they are general enough, abstracting implementation details, thus offering flexibility to the designers.

15.4.2.2 Layouts for multi-point topologies

In this section we present layout strategies for multi-point interconnection networks on OPCBs. The most popular multi-point architecture is the bus, a legacy topology for interconnection networks, offering simplicity and reduced hardware requirements. We distinguish between two types of layouts for a single bus: collinear (or 1D) and 2D. In Fig. 15.4 we present several options for a single 1D bus that can be laid out using a single waveguide layer. Each 1D bus layout requires specific placement of the Tx/Rx modules on the chips. These bus architectures have been presented in the literature (discussed below). We have adjusted them for on-OPCB application using bending radius r and crossing angles of 90 degree. The feasibility of the layouts depends on the available area, the power budget, and the optical modules losses (with splitters and combiners being the most expensive). Regeneration units placed in strategic points can be used in order to render a layout that is infeasible, due to losses, feasible. Note that in Fig. 15.4 the bus layouts are presented without using any regeneration. In the following we briefly discuss the depicted architectures.

The architecture depicted in Fig. 15.4A is based on the bus architecture presented by Dou et al. [19]. It is a bidirectional bus consisting of two waveguides with splitting/combining occurring at the Tx and Rx points of the nodes (with the exception of the Tx of the first node and the Rx of the last node). It assumes that the transmitters and receivers are located on opposite sides of the node. The bus architecture in Fig. 15.4B consists of two separate multi-point channels, one for every communication direction [39,40], which therefore is called a dual bus. It assumes that the transmitters and the receivers for the first link are located at the same side of the node, with the transmitters and receivers of the second link at the opposite side of the node in reverse order. It also assumes that the separation distance between a Tx element and an Rx element in a single side of the node is r. An alternative layout of the same architecture where the distance between the Tx and Rx elements is the used waveguide pitch is depicted in Fig. 15.4C. In this case more bends and layout area are required. The architecture in Fig. 15.4D is

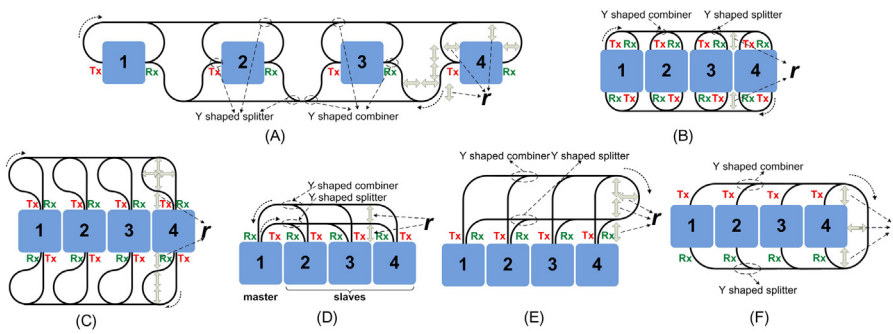

Figure 15.4 (A) Bi-dir bus, (B) dual bus, (C) dual bus (alternative), (D) master-slave bus, (E) folded bus 1, (F) folded bus 2.

Table 15.1 Comparison of the 1D bus layouts assuming N nodes

	Width	Height	Split.	Comb.	Bends	Cross.
Bi-dir bus	$N \cdot (d + 2r) + (N - 1) \cdot 2r$	$4r$	$N - 1$	$N - 1$	4	–
Dual bus	$N \cdot d$	$d + 2r$	$N - 1$	$N - 1$	2	–
Dualb (alt.)	$\begin{cases} N \cdot d, d \geq 4r \\ N \cdot d + 2r - d/2, d < 4r \end{cases}$	$d + 6r$	$N - 1$	$N - 1$	4	–
Master-slave bus	$\begin{cases} N \cdot d, d \geq 2r \\ d + 2r \cdot (N - 1), d < 2r \end{cases}$	$d + 2r$	$N - 2$	$N - 2$	2	$N - 2$
Folded bus 1	$N \quad d + r$	$d + 3r$	$N - 1$	$N - 1$	4	$N - 1$
Folded bus 2	$N \quad d + r$	$d + 2r$	$N - 1$	$N - 1$	4	–

a master-slave parallel optical bus [41] consisting of two parallel buses. The master node broadcasts signals on the bus using the first waveguide, where any slave node can receive them and send data back to the master using the second waveguide. The bus layouts in Fig. 15.4E and F are folded buses using a single waveguide [39,42]. The first folded bus layout assumes that the Tx and Rx elements are located at the same side of the node, separated by a distance equal to waveguide pitch. The second folded bus layout assumes that the Tx and Rx elements are located at the opposite side of the nodes.

Table 15.1 summarizes the characteristics of the bus layouts presented above in terms of area (width, height) as well as number of splitters, combiners, crossings, and bends in the worst case (for a single waveguide channel). We count each S-bend as two waveguide bends. The dual bus options need twice the number of Tx and Rx modules than the other ones. Splitters and combiners are both present in all the layout approaches (thus there are $2(N - 1)$ splitting/combining elements in the "worst-case waveguide"), with the exception of the master-slave bus where only splitters or combiners are present in a single waveguide.

All the aforementioned bus layouts can be extended using multiple waveguide layers and identical waveguide routing in every layer to increase aggregate bandwidth. Alternatively, more waveguides can be added using the same waveguide layer (or a combination of both approaches). Fig. 15.5 depicts how the bus layouts can be extended in the same layer using more waveguides.

Adding bus waveguides in the same layer for the bi-directional bus presents problems due to the presence of splitters/combiners at the Tx and Rx points of the nodes (using only 90 degree crossing angles and bending radiuses equal to r). The addition of a single extra bus waveguide increases the layout height by $2r$ (for all bus layouts). It also increases the required width by r in the folded bus approaches. The worst case for the number of splitters, combiners, and bends remains the same. The worst case for the number of crossings assuming W bus waveguides occurs for the waveguides that are located closest to the nodes. Table 15.2 summarizes the total (worst case) crossings assuming W bus waveguides in the same layer.

The 1D bus layouts considered above may be restricting in terms of area since they require a lot of area for their width. In Fig. 15.6 we provide two serpentine

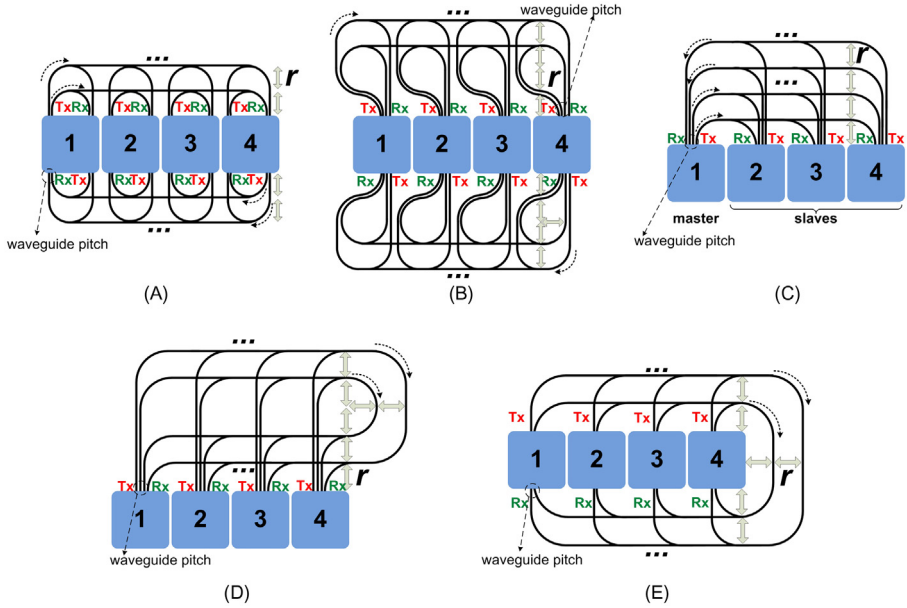

Figure 15.5 Additional bus waveguides in the same layer for increased aggregate bandwidth. (A) Dual bus, (B) Dual bus (alternative), (C) Master-slave bus, (D) Folded bus 1, (E) Folded bus 2.

Table 15.2 Number of crossings for the 1D bus lay-outs assuming W bus channels

Dual bus	Dual bus (alt.)	Master-slave bus	Folded bus 1	Folded bus 2
$(2(N-2)+2)\cdot W$	$(2(N-2)+2)\cdot W$	$(N-2)\cdot (2W-1)$	$(N-1)\cdot (2W-1)$	$2(N-1)\cdot (W-1)$

2D layout approaches (requiring a single layer) for a dual bus and a folded bus, allowing better balancing between the required height and the required width.

Finally, there could be combinations between point-to-point and multi-point architectures such as mesh of buses [43]. A 2-layer layout of a 4×4 mesh of buses is depicted in Fig. 15.7.

15.4.2.3 WDM for point-to-point topologies using multi-point layouts

WDM (Wavelength Division Multiplexing) is an important advantage of optical technology, giving the ability for a single waveguide to support multiple optical channels simultaneously using different wavelengths. This allows the implementation of

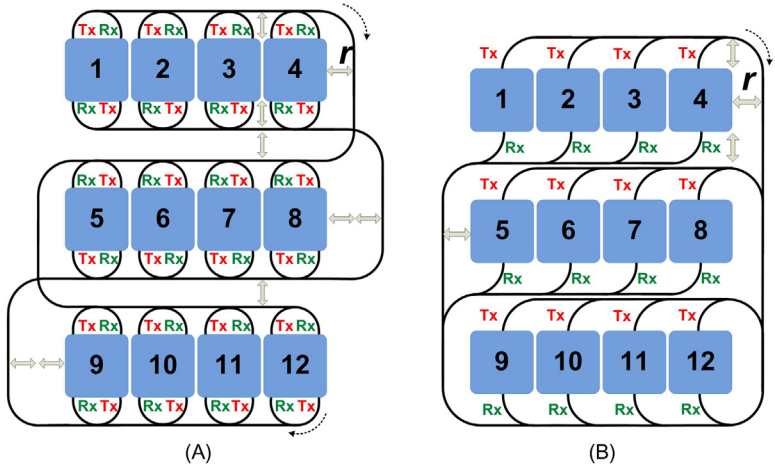

Figure 15.6 Serpentine 2D layouts for: (A) a dual bus, (B) a folded bus.

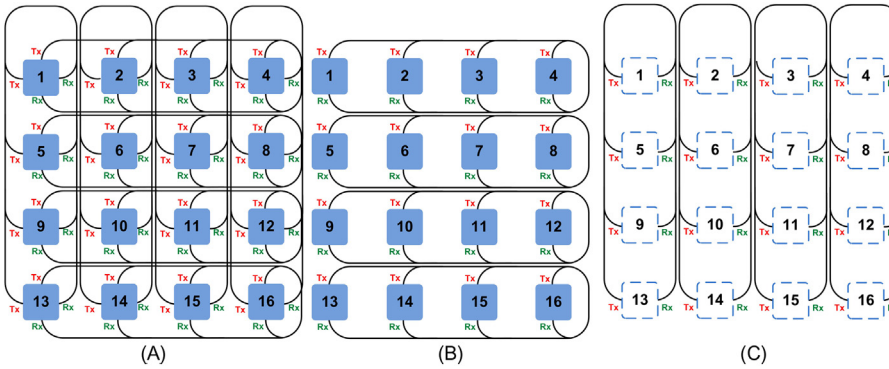

Figure 15.7 (A) 2D 4×4 mesh of buses topology (2 layers): (B) Layer 1, (C) Layer 2.

many point-to-point connections over physical waveguides laid out as busses. For example, a fully-connected network could be implemented using a single bus layout, and a mesh of fully-connected networks could be implemented as a mesh of buses. The multiplexers/demultiplexers required to create the WDM signal on a waveguide can take place either on the chip, or on the OPCB. In the latter case it can be realized by using $1 \times N$ splitters/combiners on the OPCB to combine different signals (of different wavelengths) in a single waveguide.

A simple approach to implement a point-to-point topology using an optical bus would be to use as many wavelengths as the number of uni-directional links of the topology. For example, a uni-directional ring of N nodes has N links, while an equivalent bi-directional ring has $2N$ links. Thus, for their implementation using a bus architecture, N and $2N$ wavelengths would be needed respectively. The

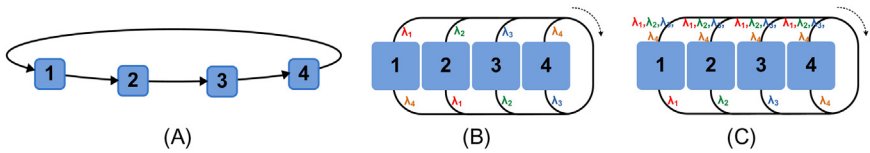

Figure 15.8 (A) Logical topology of a four uni-directional ring, and (B) its implementation using a (folded) bus. (C) An *N*-Tx, 1-Rx implementation for point-to-point connections using a (folded) bus layout.

number of Tx/Rx pairs for a single node is equal to the degree of the node (or twice that number for the dual buses). Fig. 15.8A and B depict the logical topology of a four uni-directional ring and its implementation, respectively.

Another approach is to use N wavelengths (equal to the number of nodes), smaller than the number of links of the topology, and configure the connectivity dynamically using a wavelength assignment algorithm. This would require every node to have:

- A Tunable transmitter and a burst mode receiver, or
- N separate Tx elements, N separate Rx elements, or
- N Rx elements, 1 Rx element—see Fig. 15.8C (each node transmits in a single wavelength determined by the wavelength assignment algorithm in order to ensure that no other node transmits in the same wavelength), or
- One Tx element, N Rx elements (each node transmits in a single wavelength and receives all wavelengths).

Note that in a topology composed of multiple buses such as a mesh of buses the same wavelengths can be re-used (in both the horizontal and vertical buses in a mesh of buses), since there is a different set of Tx/Rx for the second dimension.

15.4.3 Applying the proposed on-OPCB layout strategies: illustrative examples

In this section we apply the layout approaches described in the previous section for four logical topologies: a $3 \times 2 \times 2$ mesh, a 4×4 torus, a 9-fully-connected network, and a 9×9 mesh of fully-connected networks (a topology resembling a 9×9 mesh where every row and column is a fully-connected network instead of a linear array as in mesh networks). We omit the waveguide tracks for off-board communication in all cases for simplicity. We assume that at most two sides of the node can be used pinout. For the point-to-point networks we assume that two waveguide layers are used. Fig. 15.9 depicts: (A) a collinear layout for a $3 \times 2 \times 2$ mesh network, (B) a 2D (4×4) lay-out for a 4×4 torus network, and (C) a collinear layout for a 9-fully-connected network using the strategies described by Yeh et al. [34] and modified as described in Section 15.4.2 using bending radius r and crossing angles equal to 90 degree.

Table 15.3 presents the required area for all four topologies, as well as the number of crossings, using various layout approaches, assuming $d = 50$ mm and

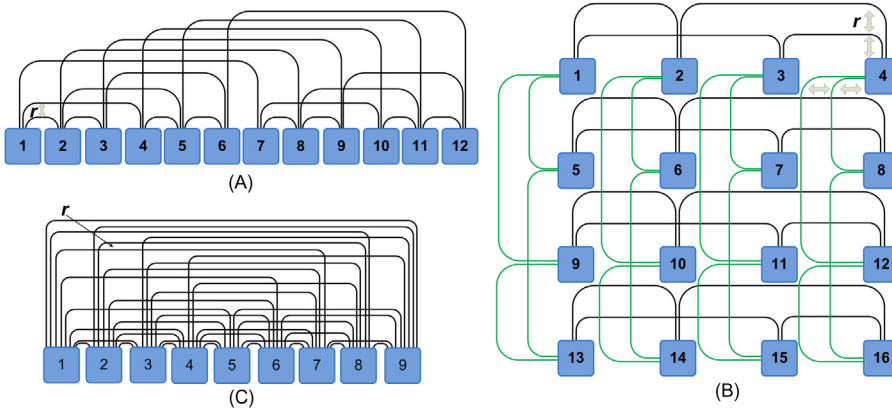

Figure 15.9 (A) Collinear layout for a $3 \times 2 \times 2$ mesh network. (B) 2D (4×4) layout for 4×4 torus networks. (C) Collinear layout for a 9-fully-connected network.

Table 15.3 Area requirements, number of crossings and bends using the point-to-point and multi-point layout techniques for various topologies ($d = 50$ mm, $r = 10$ mm)

		Width (in mm)	Height (in mm)	Crossings
$3 \times 2 \times 2$ mesh	collinear layout	820	150	9
	collinear layout ($\theta = 45$ degree)	820	86	9
	3×4 2D layout	240	240	3
	3×4 2D layout ($\theta = 45$ degree)	189	240	3
	2×6 2D layout	410	180	5
	dual bus layout	600	70	–
	folded bus layout	610	70	–
	2D 3×4 folded bus layout	220	240	–
4×4 torus	collinear layout	1100	150	8
	collinear layout ($\theta = 45$ degree)	1100	86	8
	4×4 2D layout	280	280	6
	4×4 2D layout ($\theta = 45$ degree)	252	252	6
	4×4 mesh of buses layout	280	280	–
	dual bus layout	800	70	–
	folded bus layout	810	70	–
	2D 4×4 folded bus layout	220	280	–
9 fully-connected	collinear layout	610	250	12
	collinear layout ($\theta = 45$ degree)	610	116	12
	dual bus layout	450	70	–
	folded bus layout	460	70	–
	2D 5×4 folded bus layout	270	140	–
9×9 mesh of fully-connected networks	2D layout	2250	2250	160
	2D mesh of buses layout	630	630	–

$r = 10$ mm. The bus layouts need only one layer (excluding mesh of buses layouts). The crossings column gives the number of crossings in a single layer for point-to-point layouts. For some layouts, the required height and width was calculated assuming crossing angles equal to 45 degree, using Eq. (15.1) for all tracks (except the first one as described in Section 15.4.2). Even though the 45 degree crossing angle could present practical problems due to crosstalk, it was used in an attempt to understand its effect on the layout area. As expected, collinear layout areas are rectangles with greater width than height. The 45 degree crossing angle result in area savings whenever a large number of tracks is required along a single dimension. For example, in the required height for the collinear layouts of the examined mesh, torus, and fully-connected topologies. Similarly, the WDM multi-point layouts for point-to-point networks implementation are more efficient for point-to-point topologies that require many waveguide tracks. For example, the 9×9 mesh of buses implementation for a 9×9 mesh of fully-connected networks would result in impressive savings in area (92% reduced area requirements by using the former). Similarly, the bus implementations for the 9 fully-connected network lead to significantly smaller lay-out areas (e.g, 79% less area is needed if the folded bus lay-out is used).

15.5 Conclusion

Optics have already found their way inside the DC for rack-to-rack and server-to-rack connections. In order to cope with both the energy and bandwidth requirements, and to overcome the limitations of the electrical interconnects, DCs will have to deploy optical technologies, if possible, in all packaging levels of their architecture. As short distance optical interconnects and nano photonics mature, new architectures for DCs using these building blocks will be proposed in order to maximally exploit the benefits of optics.

A large number of architectures for optical switches (for rack-to-rack communication) as well as Networks-on-Chip have already been presented, and new architectures will continue to be proposed. Implementing and laying out complex topologies via optical waveguides on the on-OPCB level presents a number of issues that have to be considered when designing architectures for DCs. To this end we have outlined layout strategies for both point-to-point and multi-point topologies for OPCBs, general enough to be easily applied by designers.

Acknowledgments

This work was supported by the European Union (European Social Fund—ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) − Research Funding Program: Thales

Investing in knowledge society through the European Social Fund and by the European Commission through the FP7 ICT-PHOXTROT (ICT 318240) project.

References

[1] Cisco Networks White Paper. Cisco global cloud index: forecast and methodology, 2013−2018, Cisco Systems. [Online] Available from: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html; 2013 [accessed 01.09.15].
[2] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture. ACM SIGCOMM Comput Commun Rev 2008;38(4):63−74.
[3] Taubenblatt MA. Optical interconnects for high-performance computing. J Lightwave Technol 2012;30(4):448−57.
[4] Farrington N, et al. Helios: a hybrid electrical/optical switch architecture for modular data centers. ACM SIGCOMM Comput Commun Rev 2011;41(4):339−50.
[5] Wang G, et al. c-Through: Part-time optics in data centers. ACM SIGCOMM Comput Commun Rev 2011;41(4):327−38.
[6] Singla A, et al. Proteus: a topology malleable data center network. In: Proc. 9th ACM SIGCOMM workshop on hot topics in networks; 2010, p. 8.
[7] Luijten R, et al. Optical interconnection networks: the OSMOSIS project. In: The 17th annual meeting of the IEEE lasers and electro-optics society; 2004.
[8] Castoldi P, et al. Energy efficiency and scalability of multi-plane optical interconnection networks for computing platforms and data centers. In: Optical fiber communication conference, Optical Society of America; 2012, p. OW3J-4.
[9] Shacham A, Bergman K. An experimental validation of a wavelength-striped, packet switched, optical interconnection network. J Lightwave Technol 2009;27(7):841−50.
[10] Zhu Z, et al. Fully programmable and scalable optical switching fabric for petabyte data center. Opt Express 2015;23(3):3563−80.
[11] Xia K, et al. Petabit optical switch for data center networks. Polytechnic Institute of New York University, New York, Tech. Rep; 2010.
[12] Ye X, et al. DOS: a scalable optical switch for datacenters. In: Proc. 6th ACM/IEEE symposium on architectures for networking and communications systems; 2010, p. 24.
[13] Gripp J, et al. Photonic terabit routers: the IRIS project. In: Optical fiber communication conference, Optical Society of America; 2010, p. OThP3.
[14] Proietti R, et al. Scalable optical interconnect architecture using AWGR-based TONAK LION switch with limited number of wavelengths. J Lightwave Technol 2013;31(24):4087−97.
[15] Schmidtke K, et al. 960 Gb/s optical backplane ecosystem using embedded polymer waveguides and demonstration in a 12G SAS storage array. J Lightwave Technol 2013;31(24):3970−5.
[16] Chen RT, et al. Fully embedded board-level guided-wave optoelectronic interconnects. Proc IEEE 2000;88(6):780−93.
[17] Pitwon RC, et al. FirstLight: pluggable optical interconnect technologies for polymeric electro-optical printed circuit boards in data centers. J Lightwave Technol 2012;30(21):3316−29.
[18] Beals IV J, et al. A terabit capacity passive polymer optical backplane based on a novel meshed waveguide architecture. Appl PhysA 2009;95(4):983−8.

[19] Dou X, et al. Optical bus waveguide metallic hard mold fabrication with opposite 45 micro-mirrors. In: OPTO, International Society for Optics and Photonics; 2010, p. 76070P-76070P.
[20] Bamiedakis N, et al. A 40 Gb/s optical bus for optical backplane interconnections. J Lightwave Technol 2014;32(8):1526–37.
[21] Shacham A, Bergman K, Carloni LP. Photonic networks-on-chip for future generations of chip multiprocessors. IEEE Trans Comput 2008;57(9):1246–60.
[22] Vantrease D, et al. Corona: system implications of emerging nanophotonic technology. ACM SIGARCH Comput Arch News 2008;36(3):153–64.
[23] Beamer S, et al. Re-architecting DRAM memory systems with monolithically integrated silicon photonics. ACM SIGARCH Comput Arch News 2010;38(3):129–40.
[24] Pan Y, Kim J, Memik G. Flexishare: channel sharing for an energy-efficient nanophotonic crossbar. In 2010 IEEE 16th international symposium on high performance computer architecture (HPCA); 2010, p. 1–12.
[25] Kirman N, et al. Leveraging optical technology in future bus-based chip multiprocessors. In Proc. 39th annual IEEE/ACM international symposium on microarchitecture; 2006, p. 492–503.
[26] Kurian, et al. ATAC: a 1000-core cache-coherent processor with on-chip optical network. In: Proc. 19th international conference on parallel architectures and compilation techniques; 2010, p. 477-488.
[27] Batten C, et al. Building manycore processor-to-dram networks with monolithic silicon photonics. In: 16th IEEE symposium on high performance interconnects, HOTI'08; 2008, p. 21–30.
[28] Morris R, Kodi AK. Exploring the design of 64-and 256-core power efficient nanophotonic interconnect. IEEE J Sel Top Quantum Electron 2010;16(5):1386–93.
[29] Koka P, et al. Silicon-photonic network architectures for scalable, power-efficient multi-chip systems. ACM SIGARCH ComputArch News 2010;38(3):117–28.
[30] Cianchetti MJ, Kerekes JC, Albonesi DH. Phastlane: a rapid transit optical routing network. ACM SIGARCH Comput Arch News 2009;37(3):441–50.
[31] Joshi A, et al. Silicon-photonic clos networks for global on-chip communication. In: Proc. 2009 3rd ACM/IEEE international symposium on Networks-on-Chip; 2009, p. 124–33.
[32] Pan Y, et al. Firefly: illuminating future network-on-chip with nanophotonics. ACM SIGARCH Comput Arch News 2009;37(3):429–40.
[33] Gu H, Xu J, Zhang W. A low-power fat tree-based optical network-on-chip for multiprocessor system-on-chip. In: Proc. conference on design, automation and test in Europe; 2009, p. 3–8.
[34] Yeh CH, Varvarigos E, Parhami B. Multilayer VLSI layout for interconnection networks. In: Proc. 2000 international conference on parallel processing; 2000, p. 33–40.
[35] Yeh CH, et al. VLSI layout and packaging of butterfly networks. In: Proc. twelfth annual ACM symposium on parallel algorithms and architectures; 2000, p. 196–205.
[36] Siokis A, Christodoulopoulos K, Varvarigos E. Laying out interconnects on optical printed circuit boards. In Proc. tenth ACM/IEEE symposium on architectures for networking and communications systems; 2014, p. 101–12.
[37] Hasharoni K, et al. A high end routing platform for core and edge applications based on chip to chip optical interconnect. In: Optical fiber communication conference. Optical Society of America; 2013, p. OTu3H-2.

[38] Chen H, et al. The Y-architecture: yet another on-chip interconnect solution. In: Proc. ASP-DAC 2003. Asia and South Pacific design automation conference. IEEE; 2003, p. 840–6.
[39] Li K, Pan Y, Zheng SQ. Parallel computing using optical interconnections. Boston, MA: Springer Science & Business Media; 1998.
[40] Guo Z, et al. Pipelined communications in optically interconnected arrays. J Parallel Distributed Comput 1991;12(3):269–82.
[41] Tan M, et al. A high-speed optical multi-drop bus for computer interconnections. Appl Phys A 2009;95(4):945–53.
[42] Melham RG, Chiarulli DM, Levitan SP. Space multiplexing of waveguides in optically interconnected multiprocessor systems. Comput J 1989;32(4):362–9.
[43] Iwama K, Miyano E, Kambayashi Y. Routing problems on the mesh of buses. In: Algorithms and computation. Berlin: Springer; 1992, p. 155–64.