Multipoint Architectures for On-Board Optical Interconnects

Apostolos Siokis, Konstantinos Christodoulopoulos, and Emmanouel (Manos) Varvarigos

***Abstract*—Optical technology offers a high-bandwidth, energy-efficient solution for the increased communication requirements of data center and high-performance computing environments and is expected to be gradually deployed at all levels of the packaging hierarchy: from board-to-board, to on-board, and even on-chip communication. In this work we focus on the on-board architecture level, outlining layout strategies for multipoint networks, such as single buses as well as a mesh of buses (MB) on optical printed circuit boards (OPCBs). Driven by that, we discuss how related point-to-point topologies, such as a mesh of fully connected networks (MFCN), can be realized using wavelength division multiplexing (WDM) and multipoint layouts. We also provide closed-form formulas for the network capacity and the average internodal distance of these two mesh-like topology families, MB and MFCN, and demonstrate how the proposed techniques and formulas can be used for designing reconfigurable mesh-like architectures on OPCBs.**

***Index Terms*—Bus; Mesh of buses; Mesh of fully connected networks; Multipoint networks; Optical interconnects; Optical printed circuit boards; Topology layout; WDM.**

I. Introduction

Increased bandwidth requirements, energy consumption, and cost are the main challenges in today's data centers (DCs) and high-performance computing (HPC) systems. The continuous growth in big data analytics and cloud-based applications stresses the interconnection networks of DCs. Annual global DC traffic is expected to reach 10.4 ZB in 2019 (from 3.5 in 2014) [1], while power consumption is projected to rise to 1012 billion kWh in 2020 [2] (to put that into perspective, the *total* energy consumption of the European Union in 2013 was 2798 billion kWh). A similar trend is observed in the supercomputer community, where cost, rather than power, poses the main challenge for exascale HPC interconnects [3]. The HPC industry expects to provide the first exaflop system around 2020, yielding 30 times higher performance than today's No. 1 supercomputer (measured at 33.9 Pflops/s in Nov. 2015) [4].

Manuscript received January 22, 2016; revised September 8, 2016; accepted September 11, 2016; published October 19, 2016 (Doc. ID 258107).

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

K. Christodoulopoulos and E. Varvarigos are with the National Technical University of Athens, Greece, and the Computer Technology Institute and Press "Diophantus," Patras, Rio, Greece.

http://dx.doi.org/10.1364/JOCN.8.000863

Optical technology is a promising energy-efficient solution for satisfying the ever-increasing bandwidth requirements of DCs and HPCs. In telecommunication networks, optical fibers have replaced most of the copper technology in wide area networks (WANs) and metropolitan area networks (MANs) and have already found their way to datacom networks—inside DCs and HPC systems [5]. Today, optics has already replaced electrical links in the network interconnecting the top-of-rack switches, achieving the required bandwidth with low power consumption. Even so, bandwidth requirements and power consumption of data communication still pose daunting issues.

To cope with the energy and bandwidth limitations of electrical interconnects, optical technologies will 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 communication. Optochips with integrated Tx (transmitter) and Rx (receiver) elements have been demonstrated, for example, in the Terabus program [6]. Electro-optical router chips with integrated vertical-cavity surface-emitting laser (VCSEL) and photodiode (PD) arrays are already available [7]. Optical printed circuit boards (OPCBs), which are extensions of the established electrical boards also equipped with integrated optical waveguides (multimode or single-mode), are currently being developed with various manufacturing methods. Multimode waveguides, usually made of polymer [8,9] and used for 850 μm applications, are more mature. Single-mode waveguides (for wavelengths of 1310 and 1550 nm) made of polymer [10] or glass [11] are expected to achieve high spectral efficiency and render wavelength division multiplexing (WDM) feasible on OPCBs. Another important step toward the goal of photonics integration in DCs and HPC systems is the emergence of silicon photonics. Silicon-based transceiver chips with hybrid bonded lasers, integrated modulators, multiplexers, and photodetectors are being integrated in multicore processor, memory, or through-silicon-via (TSV) processing units [12], while the development of silicon photonic switch fabrics is an active research field. Silicon switches using microring switching elements [13] or Mach–Zehnder interferometers [14] have recently been demonstrated and continue to attract research interest.

In general, the ability of optics to provide high bandwidth is expected to constitute an important step in reducing the cost/performance ratio [3] and power requirements, overcoming the limitations of electrical interconnects in these fields. This is further enhanced by the application

of WDM and the use of colored optical transceivers mounted as close as possible to the compute units to implement multiple logical links over a single physical link. To fully exploit the features of these new building blocks, the architectures for next-generation DCs and HPC systems that will rely upon them should be carefully designed and optimized at all levels of the packaging hierarchy. Architecture-wise, several approaches for rack-to-rack communication based on optical interconnects have already been proposed [15]. Chip-level network architectures have also received a lot of attention. Many traditional topologies have been implemented for networks-on-chip environments [16]. For the intermediate on-OPCB and OPCB-to-OPCB packaging levels, most of the proposed solutions are limited to (mainly passive) architectures targeting backplane deployment, such as parallel waveguide arrays [17], a waveguide-based optical bus structure [18], meshed waveguide architectures [19], a shared bus [20], and a regenerative bus structure [21]. In [22] we proposed a layout strategy for optically interconnected point-to-point topologies of optochips suitable for OPCBs, taking into account the differences of electrical to optical on-board communication. The main differences that were identified were that i) waveguide bends require a (nonsharp) bending radius to allow the propagation of light and ii) crossings are allowed in the same layer (with crossing angles of 90° being preferable due to losses and crosstalk). These strategies were incorporated in a methodology and a tool for on-OPCB architecture design, taking into account topology performance characteristics and also the off-board communication requirements, as the OPCBs actually form part of a larger system.

In the current work we focus on topologies based on multipoint links, as opposed to the point-to-point links considered in [22]. Multipoint links offer certain advantages over point-to-point links for optical interconnects over short distances, especially when combined with WDM, due to their simple broadcast/multicast and select nature. In Section II we describe metrics for both communication and physical layer performance estimation to be used in subsequent sections. In Section III we give formal definitions for multipoint networks and the particular mesh-like topologies that we examine. In Section IV we outline layout strategies for single bus networks and for mesh of buses (MB) networks, which constitute two important multipoint topologies for optical interconnects in DC and HPC. We also show the way point-to-point connections can be realized using WDM and multipoint layouts, obtaining in this way layouts for a point-to-point variation of MBs, the mesh of fully connected network (MFCN) topology. In addition to their area (height and width), the layouts presented are evaluated with respect to several parameters of interest that determine the physical layer quality of the signal, such as the number of waveguide structures (bends, splitters, etc.) they require in the worst case. In Section V we discuss some structural properties of the MB and MFCN topologies, which are important for addressing routing problems in them, and establish the related notation. In Section VI we turn our attention to communication aspects and provide closed-form formulas for the network capacity and the average internodal distance of the multipoint MB and the point-to-point MFCN. In Section VII we demonstrate how the techniques of Section IV and the closed-form formulas of Section VI can be used for designing dynamically configured mesh-like architectures based on optical interconnects for the on-board level (OPCB) of the packaging hierarchy. We conclude the paper in Section VIII.

II. Physical Layer and Communications Performance Characteristics

The efficiency of interconnection networks is measured both in terms of i) communications performance ("quantity") and ii) physical layer performance ("quality"). An optical interconnects topology/architecture should efficiently meet the requirements of the given applications, while its physical (on-board) implementation should satisfy all the physical layer constraints. In this section we briefly describe some metrics and parameters that are important for communication performance estimation as well as the physical layer constraints (targeting the on-board level of packaging hierarchy) that must be met. Both physical layer and communication performance will be examined using these parameters in subsequent sections.

The most representative communication performance metrics for the logical topologies of interconnection networks are *throughput* and *latency*. Throughput is closely related to channel loads (the demanded bandwidth from the channels) given the network architecture, the traffic pattern, and the routing strategy. The *ideal throughput* is the throughput of the network assuming ideal/optimal flow control and routing [23]. The ideal throughput under uniform random traffic (URT), where every node sends an equal amount of traffic to every node of the network, is often referred to as the *capacity of the network*, which is, in turn, related to the *bisection width* and *bisection bandwidth* of the network topology. The former is the smallest number of links that have to be removed to split the network in two equal parts, while the latter bounds the amount of data that can be moved between them. Another metric closely related to ideal throughput is *speedup*, which is defined as the ratio of the total input bandwidth of the network to the network's capacity, or, equivalently, as the ratio of the available bandwidth of the bottleneck channel(s) to the amount of traffic crossing it (assuming URT), and it is unitless. A network with speedup equal to 1 provides contentionless transmission under ideal conditions. Finally, a metric closely related to latency is the *average internodal distance*, which is an indicator for the expected packet latency assuming light network load and uniform distribution of the traffic destinations. Latency is the time that elapses from the moment a packet is generated at a source node until the time it is delivered to the destination, which for light load (low queueing delays) is proportional to the hop count (distance).

The feasibility of a topology's physical implementation for the on-board level of the packaging hierarchy is largely determined by its *layout*. So, a topology is feasible only if the layout satisfies both power budget [quality-of-transmission (QoT)

constraint] and (board) area constraints. The layout determines the worst-case losses (losses experienced by an optical path in the worst case) and the layout area (height, width) and volume (number of waveguide layers). The power budget is the difference between the optical transmission power and the photodetector's sensitivity. The optical signal quality deteriorates as the light travels from the light source through on-chip photonic modules [such as (de)multiplexing and switching elements], chip-to-board and board-to-chip coupling elements, and on-board waveguide structures until it reaches the photodetector in another chip. The total loss of the optical path can be estimated by adding the insertion losses of the respective elements along the path. Considering the worst path losses and comparing that to the power budget determines if the layout is QoT feasible.

Given the layout of a particular topology and the module footprints (chip sizes, waveguide bending radius, etc.), the exact layout area can be estimated. Basic on-board waveguide structures/building blocks, found in almost all OPCB platforms (single mode or multimode), regardless of the materials used and the fabrication methods, are depicted in Fig. 1. Namely, a crossing with crossing angle θ, a 90° bend with bending radius ρ, and a 1×2 Y-shaped splitter/combiner are depicted in Figs. 1(a), 1(b), and 1(c), respectively. Crossing angles of 90° for the optical waveguides are preferable due to lower losses and negligible crosstalk. S-shaped bends are also possible as well as waveguide bends with bending angles other than 90°. The splitters/combiners can be implemented so that they offer a non-equal splitting/combining ratio between the two output (input) ports (see [21] for such an example using multimode waveguides). By combining multiple 1×2 splitters (combiners), splitters (combiners) with more than two outputs (inputs) can be obtained. The exact footprints and insertion losses of the optical components are determined by the material and the fabrication method. In general, the greatest contribution to the area is due to the waveguide bends. For example, in [21] a bending radius ρ of 9 mm is chosen for the 90° bends, achieving bending losses below 1 dB, while in [11] 180° bends having radii smaller than 30 mm result in high losses. For comparison, the waveguide width is usually 50 or 100 μm and the standard waveguide pitch (waveguide spacing between straight waveguides) is 250 μm, which are 2 orders of magnitude lower than the bending radius ρ and are thus safely neglected in our area calculations. The most "expensive," in terms of losses, building blocks are the chip-to-board and board-to-chip coupling elements (implemented, for example, as micromirrors) followed by the splitters and combiners, which are followed by the bends and finally by the crossings.

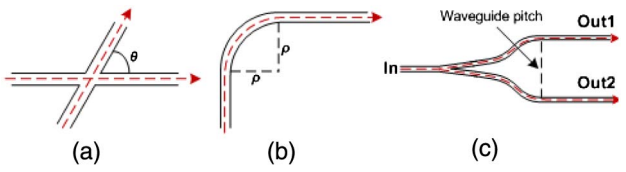

Fig. 1. (a) Crossing with crossing angle θ, (b) 90° bend with bending radius ρ, and (c) a 1×2 Y-shaped splitter.

III. Multipoint and Point-to-Point Topologies

We call *multipoint* a link, such as a bus, where more than two nodes are connected. A data packet transmitted by a node is received by all nodes attached to the multipoint link (almost) simultaneously. For the rest of the paper, we will consider a node to be a single optochip hosting the processing elements and optical components (transmitters, receivers, and switching elements if required) or a group of such chips. A multipoint topology is a network that uses multipoint links, each of which connects a subgroup of its nodes. The simplest and most popular multipoint architecture is the bus (where all nodes are connected to a single link), a legacy topology for interconnection networks, offering simplicity and reduced hardware requirements.

The MB is a multidimensional family of multipoint topologies, with each dimension being a bus. Assume that the network is represented by an undirected graph $G = (V, E)$, where V is the set of vertices or nodes and E is the set of (unordered) pairs of vertices, called edges. Then, formally, we define a d-dimensional MB as a network with $N = |V| = \prod_{i=1}^{d} k_i$ nodes, with k_i nodes along dimension i connected in a bus, where $k_i \geq 2$. A node x is logically identified by d coordinates $(x_1, x_2, \ldots, x_d)$, where $1 \leq x_i \leq k_i$ for $1 \leq i \leq d$. Two nodes x and y are neighbors and are connected on the same (j-dimensional) bus if and only if $y_i = x_i$ for all i, $1 \leq i \leq d$, except for one coordinate j, where $y_j \neq x_j$.

By using an MB optical layout and WDM, several point-to-point mesh-like logical topologies can be implemented. The most popular such topologies are the regular meshes and the meshes with wraparound links (tori). In this work we also examine the performance attributes of another topology, namely MFCN. MFCN (also called generalized hypercube) is a point-to-point mesh-like network of any dimension and radix, where the nodes along a single dimension are interconnected by a fully connected network (FCN). The formal definition of MFCN is similar to that of MB, but in the case of MFCN, two neighboring nodes are connected with a point-to-point link instead of a bus. So, the set of neighboring nodes of a node x is the same, but x is connected to them by point-to-point links instead of a single multipoint bus. MFCN first appeared in [24], where a bus-based topology similar to MB is also examined.

Mesh-like networks, whether based on multipoint or point-to-point links, have efficient layouts and simple self-routing properties (e.g., zig-zag routing, or crossing dimensions in a particular order). At the same time, MB, MFCN, regular meshes, and tori exhibit different topology characteristics that make them suitable for different kinds of HPC and DC workloads. On the one extreme there are the MB networks with the minimum number of links in every dimension (a single link/bus) but with big aggregate channel bandwidths. On the other extreme there are the MFCN with the maximum number of links in every dimension (full connectivity) but with skinnier point-to-point links, while mesh and torus topologies fall somewhere in between. Combining MB layouts (Subsection IV.B) with WDM (Subsection IV.C), which optical technology offers, we can implement various point-to-point topologies

(Subsection IV.C) and even create reconfigurable mesh-like on-board architectures to meet diverging application scenarios (Section VII).

IV. Layouts for Multipoint and Point-to-Point Mesh Topologies

In this section we present layout strategies for multipoint interconnection networks on OPCBs. In Subsection IV.A we describe different implementations and layouts of a single bus network for on-board optical interconnects. We discuss how the proposed optical bus architectures (not destined for OPCBs) can be adjusted for OPCB application. We also propose variations of these layouts, offering additional bandwidth and more efficient use of board area. Note that a key distinction compared to electrical buses is that optical transmission is directed, a fact that is taken into account in the proposed layouts. Then, in Subsection IV.B we present layouts for the MB topology family, and in Subsection IV.C we describe the way point-to-point connections (thus also point-to-point topologies, such as meshes or tori) can be implemented over multipoint layouts using WDM.

A. Single Bus Optical Layouts

We distinguish between two types of layouts for a single bus: collinear (or 1D) and 2D. In Fig. 2 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 and are discussed in more detail in the following paragraph. We have adjusted them for on-OPCB application using bending radius ρ and crossing angles of 90°. The details regarding the splitter and combiner implementation are abstracted in the layout strategies we outline. We note only that the required 1×2 combining/splitting elements can be implemented so that the branch connected to the bus artery (where more couplers/splitters follow) experiences lower losses than the other branch (as in [21]). Regeneration units placed at appropriate points can also be used to render an otherwise infeasible, due to losses, connection feasible. Note that in Fig. 2 the bus layouts are presented without using any regeneration. In the following we briefly discuss the depicted architectures.

The architecture depicted in Fig. 2(a) and presented (and fabricated using polymer waveguides for OPCBs) in [20] 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 Txs and Rxs are located at opposite sides of the node. The dual-bus architecture presented in [25] consists of two separate multipoint channels, one for every communication direction. It assumes that the Tx and Rx elements for the first link are located at the same side of the node, and the Tx and Rx elements of the second link are at the opposite side of the node in reverse order. This architecture achieves lower channel loads than single-bus approaches [as the approaches of Figs. 2(a), 2(e), and 2(f)] since it uses two separate buses, requiring, however, more Tx and Rx elements. In Fig. 2(b) we adjusted the dual bus architecture for OPCB application. In this layout, the separation distance between a Tx element and an Rx element in a single side of the node is ρ. In Fig. 2(c) we present an alternative layout of the same architecture where the distance between the Tx and Rx elements is the used waveguide pitch. In this case, more bends and layout area are required. The architecture presented in [26] is a master–slave parallel optical bus 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. This architecture does not allow direct slave-to-slave communication (only through the master node can two slaves communicate). It is suitable for I/O and memory systems. In Fig. 2(d) we present a simple layout for the master–slave architecture, where the distance between the Tx and Rx elements is the preferred waveguide pitch. The bus layouts in Figs. 2(e) and 2(f) are straightforward adaptations of the folded bus architectures presented in [25,27] for OPCB application using a single waveguide. The first folded bus layout assumes that the Tx and Rx elements are located at the same side of the node, separated by distance equal to waveguide pitch. The second folded bus layout assumes that the Tx and Rx are at opposite sides of the nodes.

Table I summarizes the characteristics of the bus layouts discussed 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), assuming nodes with height = width = h and footprint equal to $h \times h$. Similar results for nonsquare nodes can be easily obtained. For the layout width and height calculation, we neglect the waveguide width and waveguide pitch, since these are at least 2 orders of magnitude smaller than the bending radiuses. We count each S-bend as two waveguide bends.

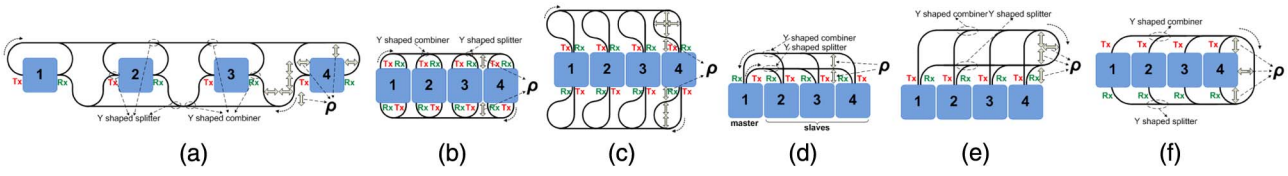

Fig. 2. (a) Bidirectional bus, (b) dual bus 1, (c) dual bus 2, (d) master–slave bus, (e) folded bus 1, and (f) folded bus 2.

TABLE I
COMPARISON OF THE 1D BUS LAYOUTS (N NODES)

Layout	Width	Height	Splitters/Cominbers	Bends	Crossings
BD	$N \cdot (h + 2\rho) + (N - 1) \cdot 2\rho$	4ρ	$N - 1$	4	–
DB1	$N \cdot h$	$h + 2\rho$	$N - 1$	2	–
DB2	$\begin{cases} N \cdot h, & h \geq 4\rho \\ N \cdot h + 2\rho - h/2, & h < 4\rho \end{cases}$	$h + 6\rho$	$N - 1$	4	–
MS	$\begin{cases} N \cdot h, & d \geq 2\rho \\ h + 2\rho \cdot (N - 1), & d < 2\rho \end{cases}$	$h + 2\rho$	$N - 2$	2	$N - 2$
F1	$N \cdot h + \rho$	$h + 3\rho$	$N - 1$	4	$N - 1$
F2	$N \cdot h + \rho$	$h + 2\rho$	$N - 1$	4	–

The dual bus options need twice the number of Tx and Rx modules as the other options. The number of splitters equals that of combiners. 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 (a) multiple waveguide layers and identical waveguide routing in every layer, or (b) additional waveguides routed in parallel on the same waveguide layer, or a combination of both approaches in order to increase aggregate bandwidth. The aggregate bandwidth can also be increased, without modifying the layout, using WDM. In this case, if arrays of Tx and Rx elements with different wavelengths are added, maybe the chip footprint will increase (this relates to the chip design/layout and the footprints of the Tx/Rx elements). A combination of WDM and the preceding approach (b) is presented in Subsection VII. In Fig. 3 we present the way the bus layouts can be extended at the same layer using waveguides routed in parallel [approach (b)]. These are depicted without length matching. Length matching may be needed due to the protocol's intolerance to timing skew. To equalize the lengths, additional S-bends can be used in the shorter waveguide. Adding bus waveguides in the same layer for the bidirectional bus [Fig. 1(a)] presents several complications due to the presence of splitters/combiners at the Tx and Rx points of the nodes (using only 90° crossing angles and bending radiuses equal to ρ). The addition of a single extra bus waveguide increases the layout height by 2ρ (for all bus layouts). It also increases the required width by ρ in the folded bus approaches [Figs. 1(e) and 1(f)]. The maximum number of splitters, combiners, and bends remains the same. The maximum number of crossings occurs for the waveguides located closest to the nodes. Table II summarizes the total (worst case) crossings assuming W waveguides (in the same layer) for the implementation of a single bus. For all aforementioned cases, if angles $\theta < 90°$ are used for the crossings of the additional waveguides, the layouts can be somewhat "squeezed," leading to area reductions (see [22] for point-to-point links). In this case, the space between a node and its closest waveguide is ρ (as for $\theta = 90°$), while the space left between following waveguides equals $(1 - \cos\theta) \cdot \rho$ ($= \rho$ for $\theta = 90°$).

The scalability with respect to losses of the layouts presented above depends on the power budget, the insertion losses of the waveguide structures, and the use (or not) of signal amplification or regeneration. To give an example, assume the folded bus layout of Fig. 2(f) implemented with single-mode waveguides, using splitters/combiners with a 50:50 ratio. In this case, the splitting/combining losses are ideally 3 dB for both cases (this is not true for multimode waveguides, in which combining can be achieved without the 3 dB penalty/join [21]). Assuming 3 dB loss for both chip-to-board and board-to-chip couplings and 0.5 dB loss per bend, the total losses in the worst case [experienced by the signal arriving at node 1 and generated by node 1, meeting 4 bends and $2(N - 1)$ splitters/combiners] are $6N - 1$ ignoring the propagation loss (for the single-mode waveguides in [11] it is 0.05 dB/cm). For $N = 4$ or 5, the resulting losses are 23 and 29 dB, respectively (ignoring the combining losses, assuming ideal conditions and multimode waveguides, the respective losses are 14 and 17 dB). If the power budget is 15 dB

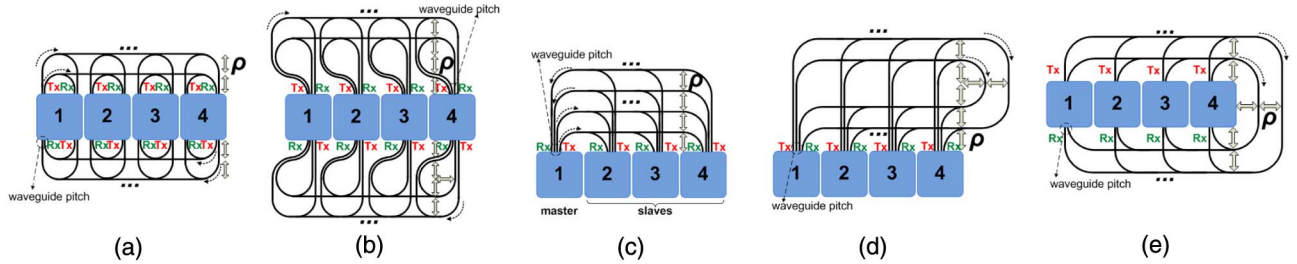

Fig. 3. Additional bus waveguides in the same layer for increased aggregate bandwidth. (a) Dual bus 1, (b) dual bus 2, (c) master–slave bus, (d) folded bus 1, and (e) folded bus 2.

TABLE II
CROSSINGS FOR 1D BUSES (W WAVEGUIDES)

Layout	Number of Crossings
Dual bus 1	$(2(N-2)+2)\cdot W$
Dual bus 2	$(2(N-2)+2)\cdot W$
Master–slave Bus	$(N-2)\cdot(2W-1)$
Folded bus 1	$(N-1)\cdot(2W-1)$
Folded bus 2	$2(N-1)\cdot(W-1)$

(commercially available VCSEL and PD components), a single regenerator can halve the total insertion losses to 11.5 and 14.5 dB, respectively, for the single-mode case. Using more regenerators, more nodes can be added in a single bus (a similar approach based on 3R regeneration is followed in the passive bus architecture of [21]). Careful choice of varying splitting/combining ratios over the optical path could render an infeasible layout feasible without requiring regeneration units [21].

The 1D bus layouts considered above may be restrictive in terms of area, since they require a lot of area only in one direction. In Fig. 4 we provide two serpentine 2D layout approaches (again requiring a single layer) for a dual bus and a folded bus, allowing better balancing between the required height and the required width, but requiring additional bends. As in the 1D layouts, regenerators need to be used, according to the specific power budgeting. Some extra regenerators are expected to be required to account for the additional losses introduced by the serpentine bends.

B. Mesh of Buses Layouts

In this section we present a layout strategy for the MB topology. A d-dimensional MB requires $d \times d$ switching elements in every node for the appropriate bus selection. The switching element could be on the same chip with the processing elements. Alternatively, there could be optochips hosting the processing elements connected on-board with a separate optochip hosting the switching element. In the latter case a single node in Fig. 5 would actually be a group of optochips interconnected in a small star topology.

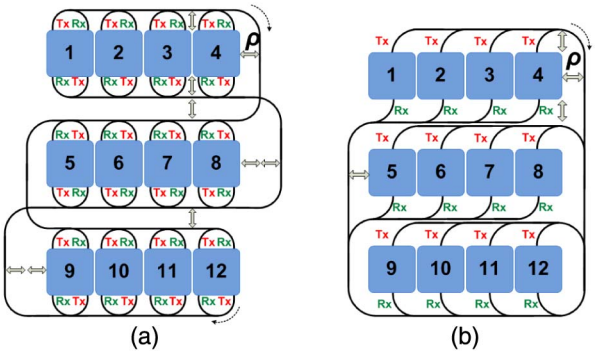

Fig. 4. Serpentine 2D layouts for (a) a dual bus and (b) a folded bus.

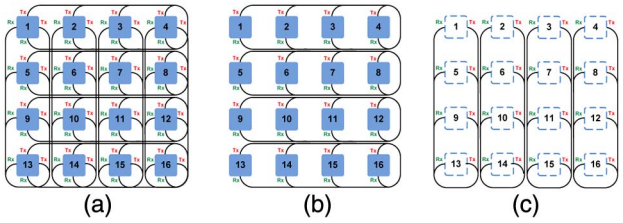

Fig. 5. (a) 2D 4 × 4 MB topology layout (two layers): (b) layer 1 and (c) layer 2.

A two-layer layout of a 4×4 MB is depicted in Fig. 5 using folded bus layouts and a single waveguide for each of them. The required layout width for a $k_1 \times k_2$ MB using W_1 and W_2 waveguides in the buses along dimensions 1 (k_1-*buses*) and 2 (k_2-*buses*) using a folded layout [the one depicted in Fig. 1(f)] is equal to $k_1 \cdot (h + 2 \cdot \rho \cdot W_2) + \max(0, W_1 - W_2) \cdot \rho$, while the respective height is equal to $k_2 \cdot (h + 2 \cdot \rho \cdot W_1) + \max(0, W_2 - W_1) \cdot \rho$. The distance between two neighboring nodes in dimension 1 (in a row) is $2 \cdot \rho \cdot W_2$ and $2 \cdot \rho \cdot W_1$ in dimension 2 (a column). The crossings and the bends for the MB in the worst case are equal to the crossings and bends of the 1D layout used. The respective numbers for 2D MB layouts using the other bus implementations can also be easily calculated.

A d-dimensional MB can be laid out in a 2D manner requiring d optical layers (one for every dimension). It can be constructed recursively, starting with the 2D subnetworks laid out as described above. Then the 2D subnetworks are treated as building blocks that are placed next to each other (along either rows or columns) in order to layout the 3D subnetwork and so on. An example is depicted in Fig. 6 for a $3 \times 3 \times 3$ MB where the 3×3 blocks are put next to each other (row-wise). For the node-to-node connection in the third dimension, three buses are needed in which only the respective nodes participate (only a single row of the third OPCB layer is shown for simplicity). Inevitably this leads to increased distance between nodes in columns. For a $k_1 \times k_2 \times k_3$ MB with W_1, W_2, and W_3 waveguides in the buses of the respective dimensions, a layout similar to that in Fig. 6 requires the nodes to be spaced apart at a distance equal to $k_1 \cdot 2 \cdot \rho \cdot W_3$ column-wise and $2 \cdot \rho \cdot W_2$ row-wise. The worst case for the number of crossings in the third layer is $2 \cdot (k_1 - 1) \cdot (k_3 - 1)$ and occurs for the waveguide that is closest to the nodes.

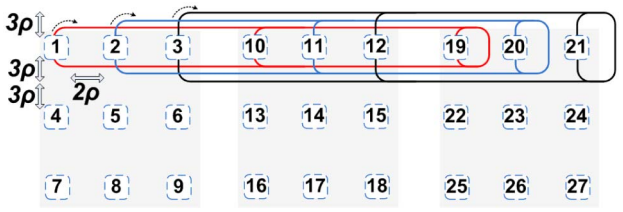

Fig. 6. Third layer of a 3 × 3 × 3 MB layout (only a single row is shown).

C. Implementation of Point-to-Point Connections Using Multipoint Layouts and WDM

WDM is an important advantage of optical technology, giving the ability for a single waveguide to simultaneously support multiple optical channels using different wavelengths. This allows the implementation of many point-to-point connections over multipoint bus waveguides. For example, a fully connected network could be implemented using a single bus layout, and an MFCN could be implemented as an MB. The typical silicon-based (de)multiplexing structures that are used to create the WDM signal and are targeted to be deployed on-chip are arrayed waveguide gratings (AWGs) and Echelle diffraction gratings (EDGs). Note that layouts for point-to-point topologies, such as tori, meshes and FCNs without the use of WDM were provided in [22]. MFCN point-to-point layouts can be easily obtained by applying the fully connected layouts strategies in every line of the MFCN.

A simple approach to implement a point-to-point topology over an optical bus would be to use as many wavelengths as the number of links of the topology. For example, a unidirectional ring of N nodes has N links, while an equivalent bidirectional ring has $2N$ links. Thus, for their implementation using a bus architecture, N and $2N$ wavelengths would be needed, respectively. The 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). Figures 7(a) and 7(b) depict the logical topology of a 4 unidirectional ring and its implementation, respectively. With this approach, for an N-FCN implementation over a single bus of N-nodes, $N\cdot(N-1)$ wavelengths are required. For a bus with W waveguides and Z wavelengths (the number of wavelengths a single node can use), and a point-to-point topology with $2|E|$ unidirectional links (with $|E|$ being the number of bidirectional links), $(W\cdot Z)/(2\cdot|E|)$ wavelengths can be used for a single unidirectional channel.

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

- a tunable Tx and a burst mode Rx, or
- N separate Tx elements, N separate Rx elements, or
- N Tx elements, 1 Rx element—see Fig. 2(c) (each node transmits in a single wavelength determined by the wavelength assignment algorithm to ensure that no other node transmits in the same wavelength), or
- 1 Tx element, N Rx elements (each node transmits in a single wavelength and receives all wavelengths).

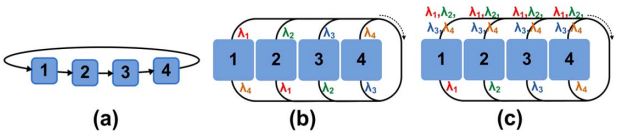

Fig. 7. (a) Logical topology of a 4 unidirectional ring and (b) its implementation using a (folded) bus approach. (c) N-Tx, 1-Rx implementation for point-to-point connections using a similar layout.

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

V. Structural Properties of MB and MFCN

In this section we present some structural properties of the (logical) MB and MFCN topologies and introduce notation that will be useful in examining the communication performance properties of these topologies in Section VI.

Following the formal definition of an MB, given in Section III, a node x in a d-dimensional MB with $k_i \geq 2$ nodes along dimension i, $i=1,2,\ldots,d$, is identified by a d-dimensional coordinate vector $(x_1,x_2,\ldots,x_d)$, where $1\leq x_i\leq k_i$ for $1\leq i\leq d$. Two nodes x and y are neighbors and connected on the same (j-dimensional) bus if and only if $y_i=x_i$ for all i, $1\leq i\leq d$, except for one coordinate j, where $y_j\neq x_j$.

We will denote a k_i-*bus* in the MB as $(x_1,x_2,\ldots,x_{i-1},*,x_{i+1},\ldots,x_d)$; this is the bus that contains node $x^{(i)}=(x_1,x_2,\ldots,x_{i-1},1,x_{i+1},\ldots,x_d)$ and all its k_i-1 neighbors along dimension i. Vector $x^{(i)}$ whose ith component is equal to 1 will be called the *representative* of that bus (or the "lower end" of the bus). The number of k_i-*buses* is $|E_{k_i}|=\prod_{j=1,j\neq i}^{d}k_j$ and the total number of buses in an MB network is $|E|=\sum_{i=1}^{d}|E_{k_i}|$.

Let $D=\{1,2,\ldots,d\}$ be the set of the network dimension, S be a subset of $D(S\subseteq D)$, of cardinality $|S|=r\leq d$, and let $S'=D-S$ be the complementary set of S. The representative $x^{(S)}$ of a node $x=(x_1,x_2,\ldots,x_d)$ along the set of r dimensions in S is defined as the node that has all coordinates equal to those of x, except for the coordinates of the dimensions in S, which are all set equal to 1.

An r-dimensional subnetwork $\mathrm{MB}(\boldsymbol{x},S)$ of MB that includes node $x=(x_1,x_2,\ldots,x_d)$ is denoted by the set of d-tuple vectors $(a_1,a_2,\ldots,a_d)$, whose ith coordinates a_i, for all $i\in S$, take all the values in $\{1,2,\ldots,k_i\}$, while their jth coordinates for all $j\in S'$ are fixed to $a_j=x_j$. For example, $\mathrm{MB}(\boldsymbol{x},\{i\})$ denotes the k_i-*bus* that passes from node $x=(x_1,x_2,\ldots,x_i,\ldots,x_d)$, that is, bus $(x_1,x_2,\ldots,x_{i-1},*,x_{i+1},\ldots,x_d)$, which is a 1D MB network. Note that bus $\mathrm{MB}(x,\{i\})=\mathrm{MB}(\boldsymbol{x}^{(i)},\{i\})$, as it consists of the k_i neighbors of x along dimension i, and $x^{(i)}$ is their representative. Generally, for any x, we have $\mathrm{MB}(x,S)=\mathrm{MB}(x^{(S)},S)$, since $x^{(S)}$ is the representative of $\boldsymbol{x}$ with respect to the dimensions contained in S ["lower end" node of the r-dimensional $\mathrm{MB}(\boldsymbol{x},S)$]. If we denote by $1=(1,1,\ldots,1)$, the all ones vector, then we can see that $\mathrm{MB}(x,S)=\mathrm{MB}(x^{(S)},S)$ is strongly isomorphic to $\mathrm{MB}(1,S)\overset{\mathrm{def}}{=}\mathrm{MB}(S)$. By strongly isomorphic we mean that there is a mapping of one graph to the other, obtained by just renumbering the nodes. (Strongly) isomorphic graphs are structurally identical. Thus, when we do not care about the exact nodes that the r-dimensional

subnetwork MB(x, S) contains or about its exact position in the MB, but we care only about its topological properties, we can omit the dependence on the nodes x that it connects, or on their representative $x^{(S)}$. Note that, if $|S| = r$ and $i \in S$, then MB(x, S) denotes an r-dimensional subnetwork of MB that contains the k_i-*bus* $(x_1, x_2, \ldots, x_{i-1}, *, x_{i+1}, \ldots, x_d)$ as a subnetwork. For example, in a $4 \times 5 \times 7$ MB network, MB$((1,5,1),\{1\}) = (*,5,1)$ is a k_1-*bus* with representative (1,5,1), and is contained in the MB$((1,1,1),\{1,2\}) = (*,*,1)$, which is a 2D (4×5) subnetwork of the entire MB. Also, MB$((1,5,1),\{1,3\}) = (*,5,*)$ is another 2D (4×7) network that contains the k_1-*bus* MB$((1,5,1),\{1\}) = (*,5,1)$. Also, MB$((1,1,1),\{1,2,3\}) = (*,*,*)$ is the entire $4 \times 5 \times 7$ MB network.

We consider now the k_i-*bus* MB$(x^{(i)},\{i\}) = (x_1, x_2, \ldots, x_{i-1}, *, x_{i+1}, \ldots, x_d)$ with representative $x^{(i)} = (x_1, x_2, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_d)$, and we define the set of all r-dimensional subnetworks of MB graphs that contain it:

$$C(x^{(i)},\{i\},r) = \{\text{MB}(x^{(S)},S) | \text{ all } S \subseteq D \text{ with } |S| = r \text{ and } i \in S\}.$$

Thus, $C(x^{(i)},\{i\},r)$ represents the set containing all r-dimensional subnetworks of MB, of which the k_i-*bus* $(x_1, x_2, \ldots, x_{i-1}, *, x_{i+1}, \ldots, x_d)$ is a part. The cardinality of this set equals

$$|C(\boldsymbol{x}^{(i)},\{i\},r)| = C_r = \binom{d-1}{r-1}.$$

Clearly, $|C(x^{(i)},\{i\},r)|$ does not depend on $x^{\{i\}}$ and i but only on the total number of dimensions d and the dimensionality r of the subnetworks. For example, for a $4 \times 5 \times 7$ MB network and a k_1-*bus* $(*, x_2, \ldots, x_d)$, with representative $(1, x_2, \ldots, x_d)$, the set of 2D subgraphs that contain it is $C(\boldsymbol{x}^{(1)},\{1\},2) = \{\text{MB}((1,5,1),\{1,2\}), \text{MB}((1,5,1),\{1,3\})\}$ and has cardinality $C_2 = 2$, while the set of 3D subgraphs that contain it is $C(x^{(1)},\{1\},3) = \{\text{MB}((1,5,1),\{1,2,3\})\}$, with cardinality $C_3 = 1$.

In an MFCN, a k_i-*link* is a point-to-point link that connects two neighboring nodes along dimension i with coordinates $(x_1, x_2, \ldots, x_i, \ldots, x_d)$ and $(x_1, x_2, \ldots, x_i', \ldots, x_d)$, where $x_i' \neq x_i$ and $1 \leq x_i, x_i' \leq k_i$. A specific k_i-*link* can be referred to by using the coordinates of the two neighboring nodes as $[x_1, x_2, \ldots, (x_i, x_i'), \ldots, x_d]$. Considering that a FCN of N nodes requires $N \cdot (N-1)/2$ point-to-point links, an MFCN requires $|E_{k_i}| = \frac{k_i(k_i-1)}{2} \cdot \prod_{j=1, j\neq i}^{d} k_j k_i$-*links*. The total number of links in an MFCN network is $|E| = \sum_{i=1}^{d} |E_{k_i}|$.

Node $\boldsymbol{x}^{(i)} = (x_1, x_2, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_d)$ in an MFCN is the representative for the 1D FCN along dimension i. Representative $\boldsymbol{x}^{(S)}$ of a node $\boldsymbol{x}$ is defined in a similar way as for MB networks. Also, MFCN$(\boldsymbol{x},S) =$ MFCN$(\boldsymbol{x}^{(S)},S)$, $|S| = r$, denotes an r-dimensional subnetwork of MFCN. For example, MFCN$(\boldsymbol{x}^{(i)},\{i\})$ is the FCN along dimension i with representative $\boldsymbol{x}^{(i)} = (x_1, x_2, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_d)$, containing k_i-*link* $s(x_1, x_2, \ldots, (x_i, x_i'), \ldots, x_d)$, $\forall x_i'$, x_i with $1 \leq x_i, x_i' \leq k_i$ and $x_i' \neq x_i$. In a similar way, as in MB, we also define sets of graphs containing all the r-dimensional subnetworks of MFCN.

In the following, given a set S with $S \subseteq D$, we will denote

$$K_S = \prod_{i \in S} (k_i - 1). \tag{1}$$

We will also use the notion of elementary symmetric polynomials [28], defined as

$$e_r(X_1, X_2, \ldots, X_n) = \sum_{1 \leq i_1 < i_2 < \cdots < i_r \leq n} X_{i_1} X_{i_2} \cdots X_{i_r}. \tag{2}$$

From this definition follows that $e_0(X_1, X_2, \ldots, X_n) = 1$ and $e_n(X_1, X_2, \ldots, X_n) = X_1 X_2 \cdots X_n$.

VI. Communication Performance Under Uniform Traffic

In this section we give closed-form formulas for the channel loads, ideal throughput, and average internodal distance for both MB and MFCN assuming URT. Under URT, every node sends an equal amount of traffic to every node of the network (including itself).

When designing architectures, one would like to use topologies optimal for the specific application(s). On the other hand, it is almost always preferable to use a good general purpose network than to design a network with a topology matched exactly to a specific problem. This is why we focus on URT, which is quite generic and does not assume any locality for the communications. Using the capacity of the network, the designer has the option to design the topology with more channel bandwidths required to ideally sustain URT (thus designing with speedup greater than 1) to allow nonidealities in the implementation (such as nonideal routing) and to allow better performance under adversarial, nonuniform traffic patterns. We used this approach also in [22] and [29].

In what follows we will assume that each source node generates 1 unit of traffic in total (thus, every node will be sending $1/N$ units of traffic to every destination node, with $N = k_1 \cdot k_2 \cdot \cdots \cdot k_d$; note that $1/N$ units of traffic are self-traffic) and that the routing of the traffic is perfectly load-balanced among all the shortest paths (SP-LB) connecting each source–destination pair. Typically, for point-to-point networks, no assumption is needed for the routing strategy to calculate the network capacity. In this case, the bisection width of the network can be used. However, this is not possible for multipoint networks such as MB. The reason we chose SP-LB is because we know that it is an optimal routing strategy for edge-symmetric mesh-like topologies such as tori networks [23]. In this section we prove that SP-LB is also optimal for MFCN (which is a point-to-point network). Based on this and the relationship between MFCN and MB networks, we conclude that by estimating throughput under URT and SP-LB we also estimate the network capacity of MB networks.

For both MB and MFCN, we will refer to a shortest path for a source–destination pair that belongs on an r-dimensional ($r \leq d$) subnetwork of the original network

as an *r-path*. For both topologies, the total number of r-paths for a given source–destination pair is $r!$ (the number of permutations of the r dimensions). Thus, the amount of traffic that such a path carries is (for both MB and MFCN) equal to $\frac{1}{r!N}$.

A. Channel Loads and Network Capacity

Theorem 1. *In an MB with SP-LB and URT, every k_i-bus is loaded with $k_i - 1$ units of traffic.*

Proof. Without loss of generality, we focus on dimension $i = 1$ and examine the load of a single (any) k_1-*bus* $\mathrm{MB}(\boldsymbol{x}^{(1)}, \{1\}) = (*, x_2, \ldots, x_d)$. We use the simplified notation $\mathrm{MB}(S)$ (with $1 \in S$) instead of $\mathrm{MB}(\boldsymbol{x}^{(S)}, S)$, for the subset that contains the examined k_1-*bus* $(*, x_2, \ldots, x_d)$, and the notation $C(r)$ instead of $C(\boldsymbol{x}^{(1)}, \{1\}, r)$ for the set of r-dimensional MB subgraphs that contain it. The traffic load on the k_1-*bus* can be classified in four categories based on the location of the generating source and destination nodes in the MB network:

- l_{SD}: source and destination nodes on the k_1-*bus*;
- l_S: source node on the k_1-*bus* and destination node not on the k_1-*bus*;
- l_D: source node not on the k_1-*bus* and destination node on the k_1-*bus*; or
- l_I: source and destination nodes not on the k_1-*bus*.

We will denote the traffic load on the k_1-*bus* due to the four aforementioned categories as L_{SD}, L_S, L_D, and L_I, respectively. The total traffic load for the k_1-*bus* is $L = L_{SD} + L_S + L_D + L_I$. Figure 8 illustrates paths belonging in the aforementioned categories for a $4 \times 5 \times 4$ MB. We will examine each one of the four categories separately.

i) L_{SD} calculation.

For the first category of traffic load, we have

$$L_{SD} = k_1 \frac{k_1 - 1}{N} = \frac{k_1 K_{\{1\}}}{N}. \tag{3}$$

This is because, for a given source on the k_1-*bus*, there are $k_1 - 1$ destinations on the k_1-*bus*, and a single shortest path carrying $1/N$ of traffic. Thus every node injects $\frac{k_1 - 1}{N}$ traffic on the k_1-*bus*, and there are k_1 such source nodes. The definition of $K_{\{1\}}$ is given in Eq. (1). In the special case of an 1D MB network (a single bus), there exists only the l_{SD} type of traffic, with total traffic load equal to $L_{SD} = k_1 \frac{k_1 - 1}{N} = k_1 \frac{k_1 - 1}{k_1} = k_1 - 1$ (while $L_S = L_D = L_I = 0$).

ii) L_S calculation.

For the second category of traffic load, we have

$$L_S = k_1(k_1 - 1)\left\{ \underbrace{\frac{(k_2 - 1)}{2N} + \cdots + \frac{(k_d - 1)}{2N}}_{C_2 = \binom{d-1}{1} = d-1} + \cdots + \underbrace{\cdots + \frac{\prod(k_i - 1)}{rN} + \cdots}_{C_r = \binom{d-1}{r-1},\ \forall i \in S - \{1\}} + \cdots + \underbrace{\frac{(k_2 - 1)\cdots(k_d - 1)}{dN}}_{C_d = \binom{d-1}{d-1} = 1} \right\}.$$

The detailed calculations for L_S are given in Appendix A. Using Eqs. (1) and (2), the previous expression is rewritten as

$$L_S = \frac{k_1 K_{\{1\}}}{N} \sum_{r=2}^{d} \frac{e_{(r-1)}(K_{\{2\}}, K_{\{3\}}, \ldots, K_{\{d\}})}{r}. \tag{4}$$

In Eq. (4) the variables X_i of the elementary symmetric polynomials are $X_i = K_{\{i+1\}}$, with $i = 1, 2, \ldots, d-1$.

iii) L_D calculation.

For the third category of traffic load, we have

$$L_D = L_S. \tag{5}$$

Under URT, all source–destination pairs have equal traffic, and the paths used for communication between a particular S-D pair are the opposite of those used for the corresponding D-S pair. So the load contributed to the k_1-*bus* from the destination nodes located on the k_1-*bus* from a set of sources would be equal to the load contributed for the communication in the opposite direction.

iv) L_I calculation.

For the fourth category of traffic load, we have

$$L_I = \underbrace{\frac{k_1 \cdot K_{\{1,2,3\}}}{3N} + \frac{k_1 \cdot K_{\{1,2,4\}}}{3N}}_{I_3 = \binom{d-1}{2} = \frac{(d-2)(d-1)}{2}} + \cdots + \cdots + \underbrace{(r-2)\frac{k_1 \cdot K_S}{rN}}_{I_r = \binom{d-1}{r-1}} + \cdots + \cdots$$
$$+ \underbrace{(d-2)\frac{k_1 \cdot K_D}{dN}}_{I_d = \binom{d-1}{d-1} = 1}$$
$$\Rightarrow L_I = \frac{k_1 K_{\{1\}}}{N} \sum_{r=3}^{d} \frac{e_{(r-1)}(K_{\{2\}}, K_{\{3\}}, \ldots, K_{\{d\}})}{r}(r-2). \tag{6}$$

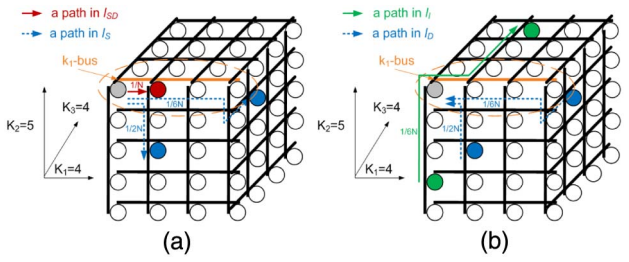

Fig. 8. Depiction of (a) paths in l_S and l_{SD} and (b) paths in l_D and l_I in a $4 \times 5 \times 4$ (logical) MB.

The detailed calculations for L_I, as well as the definition of I_r shown above, can be found in Appendix B.

We can now calculate the total traffic load $L = L_{SD} + L_S + L_D + L_I$ for the k_1-*bus*, using Eqs. (3)–(6), as

$$L = \frac{k_1 K_{\{1\}}}{N}\left(1 + \sum_{r=2}^{d} 2\frac{e_{(r-1)}(K_{\{2\}}, K_{\{3\}}, \ldots, K_{\{d\}})}{r} + \sum_{r=3}^{d} \frac{e_{(r-1)}(K_{\{2\}}, K_{\{3\}}, \ldots, K_{\{d\}})}{r}(r-2)\right)$$

$$\Rightarrow L = \frac{k_1 K_{\{1\}}}{N}\sum_{r=1}^{d} e_{(r-1)}(K_{\{2\}}, K_{\{3\}}, \ldots, K_{\{d\}}). \quad (7)$$

By substituting $K_{\{i\}} = k_i - 1 = a_i$, we have

$$A_d = \sum_{r=1}^{d} e_{(r-1)}(a_2, \ldots, a_d) = 1 + \underbrace{a_2 + \cdots + a_d}_{d-1 \text{ elements}} + \underbrace{+a_2a_3 + \cdots + a_{d-1}a_d}_{2\text{-combinations of } d-1 \text{ elements}} + \cdots + \underbrace{a_2a_3\cdots a_{d-1}a_d}_{(d-1)\text{-combinations of } d-1 \text{ elements}}$$
$$= A_{d-1} + a_d A_{d-1} = (1 + a_d)A_{d-1} = (1 + a_d)(1 + a_{d-1})A_{d-2}$$
$$= (1 + a_d)(1 + a_{d-1})\cdots(1 + a_2).$$

Reverting back to the initial notation, $a_i + 1 = k_i$, we obtain

$$A_d = k_2 \cdots k_d. \quad (8)$$

Thus, the total load on the k_1-*bus* using Eqs. (7) and (8) is

$$L = \frac{k_1(k_1 - 1)}{N}(k_2 k_3 \cdots k_d) = k_1 - 1.$$

By following similar reasoning we can derive similar formulas for all k_i-*buses* in the MB network. ∎

Corollary 1.1. *In an MB with SP-LB and URT the bottleneck links are the k_j-buses, with $j = \text{argmax}_{i \in D} k_i$, loaded with $k_j - 1$ units of traffic.*

Theorem 2. *In an MFCN with SP-LB and URT every k_i-link is loaded with $2/k_i$ units of traffic.*

Proof. We will examine the load of a single (any) k_1-*link*. Again, we use the simplified notation MFCN(S) and $C(r)$ as for the MB network. As before, we classify the traffic load on the k_1-*link* in four categories: l_{SD}, l_S, l_D, and l_I.

i) L_{SD} calculation.

For the first category of traffic load, we have

$$L_{SD} = \frac{2}{N}. \quad (9)$$

In a k_1-*link* only two nodes participate. Thus the k_1-*link* is loaded with $1/N$ units of traffic for each communication direction between the two nodes. In a 1D MFCN (a FCN), $S_D = \frac{2}{N} = \frac{2}{k_1}$, and $L_O = L_D = L_I = 0$.

ii) L_S calculation.

For the second category of traffic load, we have

$$L_S = \frac{2}{N}\sum_{r=2}^{d} \frac{e_{(r-1)}(K_{\{2\}}, K_{\{3\}}, \ldots, K_{\{d\}})}{r}. \quad (10)$$

The analysis is similar to that for MB networks, with the only difference being the number of source and destination nodes. In the general case, there are $C_r = \binom{d-1}{r-1}$ terms for the destination nodes belonging to all r-dimensional subnetworks MFCN(S), $|S| = r$. A single node on the k_1-*link* has $K_{S-\{1\}}$ destinations and there are two source nodes on the k_1-*link*, thus there are $2 \cdot K_{S-\{1\}}$ pairs of nodes in total (in an MB we had $k_1 \cdot K_S$ source and destination nodes).

iii) L_D calculation.

For the third category of traffic load, we have

$$L_D = L_S. \quad (11)$$

Under URT, the $l_S type$ load on the k_1-*link* equals the load l_D contributed by the communication in the opposite direction.

iv) L_I calculation.

For the fourth category of traffic load, we have

$$L_I = \frac{2}{N}\sum_{r=3}^{d} \frac{e_{(r-1)}(K_{\{2\}}, K_{\{3\}}, \ldots, K_{\{d\}})}{r}(r-2). \quad (12)$$

The analysis is similar to that for MB networks, with the difference that, while in an MB we had $k_1 \cdot K_S$ source and destination nodes, in an MFCN we have $2 \cdot K_{S-\{1\}}$.

So, the total load $L = L_{SD} + L_S + L_D + L_I$ is calculated using Eqs. (9)–(12):

$$L = \frac{2}{N}\sum_{r=1}^{d} e_{(r-1)}(K_{\{2\}}, K_{\{3\}}, \ldots, K_{\{d\}}). \quad (13)$$

Using Eqs. (13) and (8), the total load of the k_1-*link* is equal to

$$L = \frac{2}{N}(k_2 k_3 \cdots k_d) = \frac{2}{k_1}.$$

We can derive similar results for all k_i-*links*. ∎

The channel load for a unidirectional k_i-*link* in an MFCN under URT is $1/k_i$ units of traffic (half the traffic of the respective bidirectional link).

Corollary 2.1. *In an MFCN with SP-LB and URT, the bottleneck links are the k_j-links, with $j = \text{argmin}_{i \in D} k_i$, loaded with $2/k_j$ units of traffic.*

Theorem 3. *The throughput under URT, assuming perfect load balancing over all shortest paths (SP-LB)*

and channels with bandwidth b for MB networks and MFCN is $b/(\max(k_i)-1)$ and $b\cdot\min(k_i)/2$, respectively, $i\in D=\{1,2,\ldots,d\}$.

Proof. Given the communication traffic pattern and the routing strategy, the throughput for a network with channels of bandwidth b can be calculated as

$$\Theta=\frac{b}{\gamma_{\max}}, \tag{14}$$

where $\gamma_{\max}$ is a dimensionless number, equal to the ratio of the bandwidth demanded from the bottleneck (unidirectional) channel to the injected bandwidth of the input ports, or equivalently equal to the amount of traffic that will cross the bottleneck channel (channel load) if each input generates 1 unit of traffic in total. Thus, the calculation of throughput for MFCN and MB networks under URT and perfect load balancing over all shortest paths is straightforward using corollaries 1.1, 2.1, and Eq. (14) (keeping in mind that for MFCN the load on unidirectional channels is required for the calculation):

$$\Theta_{\mathrm{MB}}(\mathrm{URT})=\frac{b}{\max_{i\in D}k_i-1},\qquad \Theta_{\mathrm{MFCN}}(\mathrm{URT})=b\cdot\min_{i\in D}k_i.$$

■

Theorem 4. *Perfect load balancing over shortest paths (SP-LB) is an optimal routing strategy for an MFCN when the minimum dimension of the MFCN* ($\min_{i\in D}k_i$) *is even, assuming URT and channels with bandwidth b.*

Proof. Ideal throughput is the throughput of the network assuming ideal/optimal flow control and routing. Ideal routing is the routing strategy that achieves the minimum maximum load on the bottleneck channel(s). An upper bound for ideal throughput for point-to-point networks under URT is

$$\Theta_{\mathrm{ideal}}(\mathrm{URT})\le\frac{2B_B}{N}=\frac{2bB_C}{N}, \tag{15}$$

where B_C is the bisection width, and B_B is the bisection bandwidth of the network. Since the bisection width for a 1D FCN with k_1 nodes is $\lfloor(\frac{k_1}{2})^2\rfloor$, the bisection width for an MFCN when the minimum dimension of the MFCN is even is

$$B_C=\min_{i\in D}\left\lfloor\left(\frac{k_i}{2}\right)^2\right\rfloor\cdot\prod_{j\in D,\,j\neq i}k_j=\min_{i\in D}\left\lfloor\left(\frac{k_i}{2}\right)^2\right\rfloor\cdot\prod_{j\in D,\,j\neq i}k_j. \tag{16}$$

Thus, from Eqs. (15) and (16) we get

$$\Theta_{\mathrm{ideal}}(\mathrm{URT})\le\frac{2b\left(\min_{i\in D}\left(\frac{k_i}{2}\right)^2\cdot\prod_{j\in D,j\neq i}k_j\right)}{\prod_{\forall j\in D}k_j}=\frac{b\cdot\min_{i\in D}k_i}{2}.$$

Since the upper bound for the ideal throughput of an MFCN is equal to the throughput achieved when using perfect load balancing over shortest paths (Theorem 3), we conclude that the routing strategy where the traffic load is perfectly balanced among all shortest paths is optimal. ■

When the minimum dimension of the MFCN is odd, then the throughput is lower than the upper bound of Eq. (15). This is also true for tori (with SP-LB still being optimal nevertheless). We conclude that SP-LB is also ideal/optimal for the corresponding MB networks under URT, thus Theorem 3 gives the capacity of MB networks.

B. Average Internodal Distance

In this section we give a closed-form formula for the average internodal distance of MB and MFCN under URT and perfect load balancing. We follow the reasoning found in [30], where mesh networks were examined.

Theorem 5. *For both MFCN and MB networks, the average internodal distance under URT is equal to* $\sum_{i=1}^{d}\frac{k_i-1}{k_i}$.

Proof. Average distance under URT is equal to

$$\bar{D}(\mathrm{URT})=\frac{\sum_{A\in V}\sum_{B\in V}d_{A,B}}{\sum_{A\in V}\sum_{B\in V}1},$$

where $d_{A,B}$ is the shortest distance between A and B. The average distance for a 1D MFCN or MB network with $N=k_1$ nodes is equal to

$$\overline{D_{1D}}(\mathrm{URT})=\frac{\sum_{i=1}^{k_1}\sum_{j=1}^{k_1}d_{i,j}}{\sum_{i=1}^{k_1}\sum_{j=1}^{k_1}1},\qquad d_{i,j}=\begin{cases}1, & \text{if } i\neq j\\ 0, & \text{if } i=j\end{cases}$$
$$\Rightarrow\overline{D_{1D}}(\mathrm{URT})=\frac{k_1(k_1-1)}{k_1^2}=\frac{k_1-1}{k_1}. \tag{17}$$

The shortest path length in a d-dimensional MFCN or MB can be found by adding the distances of the destination node from the source node along each of the d dimensions (as in regular mesh networks). Thus, the average internodal distance can be calculated by adding the average internodal distance across each dimension, yielding

$$\overline{D_{1D}}(\mathrm{URT})=\bar{D}(\mathrm{URT})=\sum_{i=1}^{d}\frac{k_i-1}{k_i}.$$

■

VII. Application to Example Topologies and Technologies

In this section we apply the results obtained in previous sections to some example topologies, assuming specific technologies developed in the PhoxTrot project [31].

The channel loads and average distances under URT for 4×4, 3×6, and $3\times4\times7$ MFCN and MB networks are presented in Table III. The loads for the MFCNs are for the respective unidirectional links (= half the load of the bidirectional links as given by Theorem 2). The results

TABLE III
CHANNEL LOADS AND AVERAGE DISTANCE FOR VARIOUS MFCN AND MB NETWORKS

	Channel Loads MFCN			Channel Loads MB			
Topology	k_1-links	k_2-links	k_3-links	k_1-buses	k_2-buses	k_3-buses	Av. Dist. (both)
4×4	0.25	0.25	–	3	3	–	1.5
3×6	0.33	0.17	–	2	5	–	1.5
$3 \times 4 \times 7$	0.33	0.25	0.14	2	3	6	2.274

of Table III were also verified using simulations. In particular, we modeled the topologies as flow networks in MATLAB, implemented SP-LB routing, used URT traffic matrices with 1 unit injected traffic, and verified that the obtained results match exactly the ones provided by the formulas of Section VI. The results in Table III give the minimum required channel bandwidths needed to sustain the injected traffic. For example, to design a 4×4 MFCN with speedup equal to 1, the available (unidirectional) channel bandwidths should be equal to 0.25 units of traffic. Designing with speedup greater than 1 will allow nonideal routing and flow control and better performance for arbitrary traffic patterns. The closed-form formulas for the communication performance of MFCN and MB could be used in conjunction with the layout strategies of Section IV (knowing the chip footprints, board area constraints, and the losses for crossings, splitters, combiners, and bending radius ρ) to determine feasibility of on-board topologies in terms of both network performance and area and loss constraints in the optical layer. This is the approach that we followed in [22] for point-to-point topologies where both performance and layout strategies were incorporated in a tool performing exhaustive topology research, leading to optimal on-OPCB point-to-point topologies.

We now give an example of an on-board layout for 16 nodes using the layout strategy of Section IV, assuming building blocks developed in the PhoxTrot project [31]. Part of the PhoxTrot technology portfolio for the single-mode board platform is a 4×4 space switch (a similar switching element was presented also in [14]). Based on this switch, a 48×48 switching element of 1920 Gbps will be achieved (with 40 Gbps/channel), using 12 wavelengths. The 48 input signals are partitioned in 4 groups of 12 channels. The 12 signals within a group are multiplexed in a single WDM signal which then enters a single input port of the 4×4 space switch. The WDM signal exiting from a single output port of the 4×4 switch is then demultiplexed in 12 output signals. We assume that the switching elements are hosted on the same chip with the processing elements (a similar approach was followed in the IBM Blue Gene/Q using solely electrical interconnections) [32]. The processing of the data takes place in the electrical domain, and then the output data are buffered and are finally converted to the optical domain for transmission. Vice versa, the received optical data are converted to the electrical domain for processing. For simplicity, in this example we will assume that all 48 channels are used for interconnecting the 16 nodes on-board. To design a board—part of a larger system, a number of routing channels should be reserved for off-board node-to-node connections (see [22]) or for connecting to another routing element at the edge of the board that connects the on-board cluster with similar boards/clusters.

We lay out the 16 nodes in a 4×4 MB fashion, as described in Subsection IV.B, using two waveguide layers. Since a 2D MB requires 2×2 switching elements in every node, the two inputs/outputs of the 4×4 space switching element are treated as one link. Thus, for both row and column buses, we have two waveguides ($W_1 = W_2 = 2$). The required layout area is $4 \cdot (h + 4 \cdot \rho) \times 4 \cdot (h + 4 \cdot \rho)$ (see Subsection IV.B). The total number of splitters, combiners, crossings, and bends is equal to the splitters, combiners, crossings, and bends of the 1D bus layout used. Since in this case a folded layout is used [as in Fig. 2(f)], three combiners (Table I), three splitters (Table I), six crossings (Table II), and two bends (Table I) are required in the worst case. We assume that the CPU-to-memories communication takes place electrically in a separate board layer (an OPCB with a single optical layer stacked between multiple electrical PCB layers has been demonstrated in the PhoxTrot project [31], in which two electro-optical router chips, as described in [7], communicate using multimode polymer waveguides). All the aforementioned are depicted in Fig. 9. The layout of the on-chip optical modules needed to achieve the appropriate pin placement for the 4×4 MB (which is Tx exiting the north and east chip sides and Rx entering the south and west chip sides) is also depicted. An alternative architecture is to move all switching in the electrical domain. In this case, the electrical interface connecting the processors and the electronic processing element located between the Tx and Rx arrays in Fig. 9 should connect to all four such processing elements, and the 4×4 optical switch would not be required (this is the approach followed in [7] using multimode optical links and no WDM).

We also assume that a topology configuration element, located on the electrical layer of the board, is connected to the on-board nodes, allowing reconfiguration of the logical topology. For the implementation of a logical 4×4 MB topology, all ($2 \cdot 12 =$)24 available channels in a single bus should be used for the implementation of a single broadcast link with an aggregate bandwidth of 960 Gbps. For the implementation of point-to-point topologies, a subset of the available channels should be used for the implementation of a single point-to-point unidirectional link. We consider four different topology configurations, namely 4×4 MB, MESH, TORUS, and MFCN topologies. Since the number of unidirectional links for a 4-chain array (1D-MESH), a 4-ring (1D-TORUS), and a 4-FCN (1D-MFCN) is 6, 8,

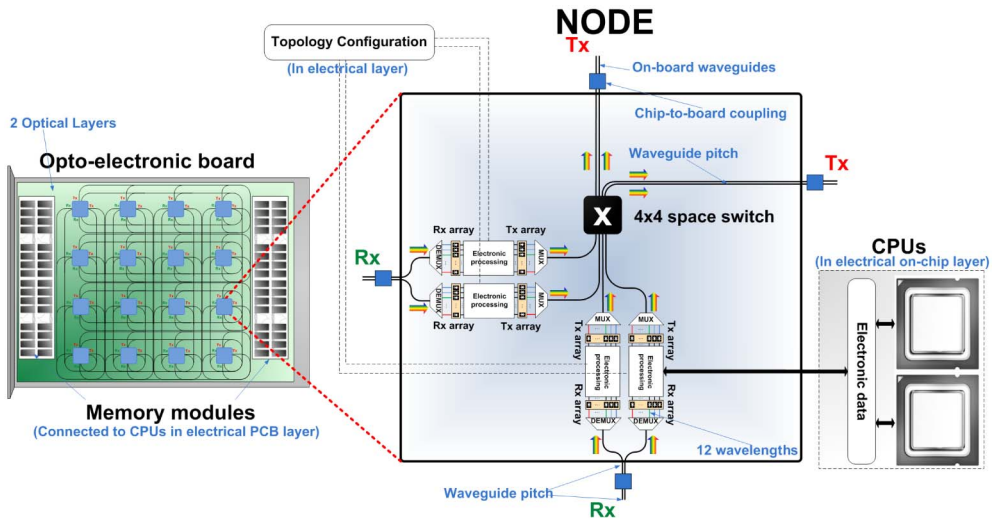

Fig. 9. Layout of a 4 × 4 MB, potential placement of the memory modules on-OPCB, and layout of the optical modules on a single chip.

and 12, respectively, 4, 3, and 2 wavelengths are required for the implementation of a single unidirectional link for these topologies (Subsection IV.C) with link bandwidths of 160, 120, and 80 Gbps.

Finally, we assume two processing elements per node, generating 320 Gbps of traffic under URT (the equivalent to the 1 unit of traffic in the analysis of Section VI). This corresponds to $1/N = 20$ Gbps of self-traffic, while the remaining 300 Gbps is the injection bandwidth of data from the processors to the routing elements. These numbers correspond to processor chips of 1 Tflops (as Intel Xeon Phi 3100) and communication-to-computation ratio 0.15 bit/flop per processor. For the configuration described above, the routing bandwidth to injection bandwidth ratio for a single node is 6.4 (the equivalent ratio for, e.g., Cray XE6 nodes is 5 [33]).

The respective topology results are presented in Fig. 10. For the speedup and average distance estimation of MB and MFCN, we used the formulas of Section VI. For the speedup/ideal-throughput of mesh and torus we used the bisection-based upper bounds [Eq. (15)]. For the respective average distances we used the formulas in [30].

For the specific scenario, the 4 × 4 MB and MFCN achieve speedup equal to 1 and thus ideal throughput of 320 Gbps. The former has a far lower connectivity degree but greater aggregate link bandwidth, while the latter has a greater connectivity degree but skinnier channels. The MESH and TORUS configurations achieve speedup equal to 0.5 and 0.75 and thus ideal throughput of 160 and 240 Gbps, respectively. The average distance for the four topology configurations is 1.5 for both MB and MFCN, and 2.5 and 2 for the mesh and torus configurations, respectively. Finally, note that the MB configuration will be preferable in traffic scenarios where extensive multicasting communication is required.

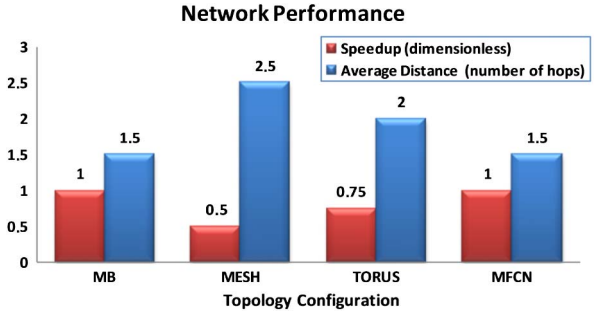

Fig. 10. Performance (speedup, average distance) for 4 × 4 MB, MESH, TORUS, and MFCN OPCB network configurations.

VIII. Conclusion

We outlined layout strategies for multipoint architectures such as buses and meshes of buses. We also provided closed-form formulas for the network capacity and average distance for both the mesh of buses (MB) and mesh of fully connected networks (MFCN) topology families. Finally, we demonstrated how the aforementioned results can be used for on-board topology design of reconfigurable mesh-like architectures. Bus-based optical interconnection architectures are good candidates for small network architectures and thus suitable for the on-board level of the packaging hierarchy. The techniques presented in this work can be a useful tool for architecture designers, paving the way toward the adoption of optically interconnected data communications in all levels of the HPC and DC systems.

Appendix A: Calculation of L_s in MB Networks

In the general case, there are $C_r = \binom{d-1}{r-1}$ terms for the destination nodes belonging to all r-dimensional subnetworks MB(S), $|S| = r$. A single node on the k_1-*bus* has K_S destinations in l_S. For a specific source and destination, there are $r!r$-paths. From these paths, only those that use the k_1-*bus* as a first dimension will be counted in the l_S category examined here. The number of these paths is the

number of permutations of r objects $(k_1, k_2, \ldots, k_r)$ in which k_1 is put first, which are $(r-1)!$ cases. Thus, for a single source–destination pair, the k_1-*bus* will be burdened with $(r-1)!/(r!N) = 1/rN$ units of traffic, and there are k_1 source nodes, thus $k_1 \cdot K_S$ pairs of nodes in total.

For $r = 2$, the (first) $C_2 = d-1$ terms of the summation relate to the traffic load from source nodes belonging in all 2D subnetworks of the MB, that is, subnetworks in $C(2) = \{\mathrm{MB}(\{1,2\}), \mathrm{MB}(\{1,3\}), \ldots, \mathrm{MB}(\{1,d\})\}$. A single node on the k_1-*bus* under examination has $K_{\{1,j\}} = (k_1-1)\cdot(k_j-1)$ destinations, $j \in D-\{1\}$. Since, for a specific source node in the k_1-*bus* and a specific 2D subnetwork and destination, there are two shortest paths (2-paths) and there are k_1 source nodes on the k_1-*bus*, every path is credited with $\frac{1}{2N}$ units of traffic. Thus, the traffic load on the k_1-*bus* for every subnetwork $\mathrm{MB}(\{1,j\})$, $j \in D-\{1\}$ of $C(2)$ is $k_1 \frac{K_{\{1,j\}}}{2N}$.

For $\boldsymbol{r = d}$, there is only one term. It is the case where the coordinates of the source and destination nodes differ in all their d subcoordinates. There are K_D destinations for a single source and k_1 sources and thus $k_1 \cdot K_D$ pairs of nodes in total. For a single source–destination pair, the k_1-*bus* will be loaded with $1/dN$ units of traffic.

Appendix B: Calculation of L_I in MB Networks

For the estimation of the total amount of l_I type traffic on the k_1-*bus*, we will have to enumerate all possible source–destination pairs that use the k_1-*bus* under examination as an intermediate link on their shortest paths. That is, the l_I type traffic consists of the total traffic carried by all r-paths that connect source and destination nodes belonging to two subnetworks that include k_1-*bus*, but none of the source and destination nodes belong to k_1-*bus*, thus constraining r to be $3 \le r \le d$. To carry out this calculation, we will use the sets $\mathrm{I}(r)$ that correspond to the l_I type traffic of the *r-paths* that burden the k_1-*bus*. A formal definition of $\mathrm{I}(r)$ will be given below, but first we describe the notation that we use for the enumeration of the source–destination pairs: we denote as $\mathrm{MB}(x^{(S_1)}, S_1) \to \mathrm{MB}(x^{(S_2)}, S_2)$, $S_1 \cap S_2 = \{1\}$ a pair of subnetworks of MB both containing the k_1-*bus* $(*, x_2, \ldots, x_d)$ as a subnetwork, where the left part symbolizes a "source subnetwork of MB" and the right part a "destination subnetwork of MB." In what follows, for simplicity we will denote $\mathrm{MB}(x^{(S_1)}, S_1) \to \mathrm{MB}(x^{(S_2)}, S_2)$ as $\mathrm{MB}(S_1) \to \mathrm{MB}(S_2)$. A single node in the source subnetwork $\mathrm{MB}(S_1)$ has K_{S_2} destinations at the destination subnetwork $\mathrm{MB}(S_2)$ (we exclude the on-the-k_1-*bus* nodes and the nodes not burdening the k_1-*bus* with traffic). There are $k_1 \cdot K_{S_1-\{1\}}$ such source nodes. All paths connecting the source nodes to the destination nodes will be r-paths (will use r dimensions), where $r = |S|$ and $S = S_1 \cup S_2$. Note that if $S_1 \cap S_2 \supset \{1\}$, then $\mathrm{MB}(S_1) \to \mathrm{MB}(S_2)$ degenerates into the source–destination pair of subnetworks $\mathrm{MB}(S_1') \to \mathrm{MB}(S_2')$, $S_1' \cap S_2' = \{1\}$, since only the source nodes belonging to $\mathrm{MB}(S_1')$ and destination nodes belonging to $\mathrm{MB}(S_2')$ will burden the k_1-*bus* with traffic; for the remaining nodes in $\mathrm{MB}(S_1)$ and $\mathrm{MB}(S_2)$, the k_1-*bus* is not part of their shortest paths (thus not loaded with traffic). Note that the l_I type traffic for $\mathrm{MB}(S_1) \to \mathrm{MB}(S_2)$ subnetworks (2-paths) is 0, since the related traffic is included in the calculations of L_{SD} or L_S or L_D.

We will now define the set $I(r)$ that includes all source–destination pairs that use r-dimensional shortest paths (*r-paths*). For this purpose, we first define two more sets:

$$\mathrm{I}(S,r,s) = \{\mathrm{MB}(S_1) \to \mathrm{MB}(S_2) | \text{ all } S_1, S_2 \text{ with } S = S_1 \cup S_2, \; |S| = r, \quad S_1 \cap S_2 = \{1\}, \quad |S_1| = s\}.$$

From the definition above, it follows that $|\mathrm{I}(S,r,s)| = I_{r,s}^S = \binom{r-1}{s-1}$. We also define

$$\mathrm{I}(S,r) = \{\mathrm{I}(S,r,s) | \quad \text{for all } s = 2, \ldots, (r-1)\}.$$

From the definition above, it follows that $|\mathrm{I}(S,r)| = I_r^S = r-2$. Finally, we define

$$\mathrm{I}(r) = \{\mathrm{I}(S,r) | \text{ all } S \subseteq D \text{ with } |S| = r\}.$$

Remember that $D = \{1, 2, \ldots, d\}$ is the set of indices of all MB dimensions. From the definition above it follows that $|\mathrm{I}(r)| = I_r = \binom{d-1}{r-1} = C_r$, the same number of terms as for l_S type traffic. Now we can calculate the l_I type traffic for all r-paths.

Let us examine the last term in Eq. (6) wherein $r = d$ and we examine $\mathrm{I}(d)$ (d-paths). The calculations for $\mathrm{I}(d)$ will help us estimate the total traffic for the general case, $\mathrm{I}(r)$. For $r = d$, we have $I_d = \binom{d-1}{d-1} = 1$ set in $\mathrm{I}(d)$ and, in particular, the set $\mathrm{I}(\{1,2,\ldots,d\}, d)$. $\mathrm{I}(\{1,2,\ldots,d\}, d)$ contains $d-2$ sets: $\mathrm{I}(\{1,2,\ldots,d\}, d, s)$, $\forall s = 2, \ldots, (d-1)$. Every such set contains $\binom{d-1}{s-1}$ source–destination subnetwork pairs. For every such source–destination pair, there are $k_1 \cdot K_D = k_1(k_1-1)(k_2-1)\cdots(k_d-1)$ source and destination nodes in total. Every shortest path will carry $\frac{1}{d!N}$ load of traffic. Only the shortest paths in which the $s-1$ dimensions from the "source subnetwork" are chosen first, followed by dimension k_1, will burden the k_1-*bus*. So, in the general case, only routing with the following order: $\underbrace{-\cdots-}_{s-1}\underline{k_1}\underbrace{-\cdots-}_{d-s}$, that is, where dimension 1 is crossed sth by the path, will load the k_1-*bus*. Thus, we have $(s-1)!$ permutations that appear $(d-s)!$ times, and so we have in total $(s-1)!(d-s)!$ paths that load the k_1-*bus* with traffic. To sum up, for d-dimensional shortest paths, we have $\binom{d-1}{s-1}$ combinations of subnetwork pairs, $k_1 \cdot K_D$ nodes in all cases, $(s-1)!(d-s)!$ paths and $s = 2, 3, \ldots, d-1$. Thus, for d-paths we have traffic load equal to

$$\sum_{s-1=1}^{d-2} \frac{k_1 \cdot K_D}{d!N} \binom{d-1}{s-1} (s-1)!(d-s)! = \sum_{s-1=1}^{d-2} \frac{k_1 \cdot K_D}{dN} = (d-2)\frac{k_1 \cdot K_D}{dN}.$$

For the general case of r-dimensional shortest paths, we substitute d with r and K_D with K_S ($S \subseteq D$, $1 \in S$, $|S| = r$).

So the terms in $I(r)$ have the following form: $(r-2)\frac{k_1 \cdot K_S}{rN}$ and we have $I_r = \binom{d-1}{r-1}$ $(= C_r)$ such terms. Note that $L_I = 0$ for $d \leq 2$ (1D or 2D MB networks).

Acknowledgment

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

References

[1] "Cisco Global Cloud Index: Forecast and Methodology, 2014–2019," Cisco White Paper, 2016.

[2] "Make IT Green: Cloud Computing and Its Contribution to Climate Change," Greenpeace International, 2010.

[3] C. Minkenberg, "HPC networks: Challenges and the role of optics," in *Optical Fiber Communication Conf. (OFC)*, 2015, paper W3D-3.

[4] "TOP500 supercomputer list of June 2014" [Online]. Available: http://s.top500.org/static/lists/2014/06/TOP500_201406_Poster.png.

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

[6] F. E. Doany, C. L. Schow, C. W. Baks, D. M. Kuchta, P. Pepeljugoski, L. Schares, R. Budd, F. Libsch, R. Dangel, F. Horst, B. J. Offrein, and J. A. Kash, "160 Gb/s bidirectional polymer-waveguide board-level optical interconnects using CMOS-based transceivers," *IEEE Trans. Adv. Packag.*, vol. 32, no. 2, pp. 345–359, 2009.

[7] K. Hasharoni, S. Benjamin, A. Geron, G. Katz, S. Stepanov, N. Margalit, and M. Mesh, "A high end routing platform for core and edge applications based on chip to chip optical interconnect," in *Optical Fiber Communication Conf. (OFC)*, 2013, paper OTu3H-2.

[8] A. Hashim, N. Bamiedakis, R. V. Penty, and I. H. White, "Multimode polymer waveguide components for complex on-board optical topologies," *J. Lightwave Technol.*, vol. 31, no. 24, pp. 3962–3969, 2013.

[9] T. Ishigure, K. Shitanda, T. Kudo, S. Takayama, T. Mori, K. Moriya, and K. Choki, "Low-loss design and fabrication of multimode polymer optical waveguide circuit with crossings for high-density optical PCB," in *IEEE Electronic Components and Technology Conf. (ECTC)*, 2013, pp. 297–304.

[10] R. Dangel, J. Hofrichter, F. Horst, D. Jubin, A. Porta, N. Meier, I. Soganci, J. Weiss, and B. J. Offrein, "Polymer waveguides for electro-optical integration in data centers and high-performance computers," *Opt. Express*, vol. 23, no. 4, pp. 4736–4750, 2015.

[11] L. Brusberg, H. Schroder, M. Queisser, and K. Lang, "Single-mode glass waveguide platform for DWDM chip-to-chip interconnects," in *IEEE Electronic Components and Technology Conf. (ECTC)*, 2012, pp. 1532–1539.

[12] Y. Vlasov, "Silicon photonics for next generation computing systems," presented at the 34th European Conf. on Optical Communications (ECOC), Brussel, Belgium, 2008.

[13] D. Nikolova, S. Rumley, D. Calhoun, Q. Li, R. Hendry, P. Samadi, and K. Bergman, "Scaling silicon photonic switch fabrics for data center interconnection networks," *Opt. Express*, vol. 23, no. 2, pp. 1159–1175, 2012.

[14] N. Dupuis, "Modeling and characterization of a nonblocking 4 × 4 Mach–Zehnder silicon photonic switch fabric," *J. Lightwave Technol.*, vol. 33, no. 20, pp. 4329–4337, 2015.

[15] C. Kachris and I. Tomkos, "A survey on optical interconnects for data centers," *IEEE Commun. Surv. Tutorials*, vol. 14, no. 4, pp. 1021–1036, 2012.

[16] C. Batten, A. Joshi, V. Stojanovć, and K. Asanović, "Designing chip-level nanophotonic interconnection networks," in *Integrated Optical Interconnect Architectures for Embedded Systems*. New York: Springer, 2013, pp. 81–135.

[17] K. Schmidtke, F. Flens, A. Worrall, R. Pitwon, F. Betschon, T. Lamprecht, and R. Krähenbühl, "960 Gb/s optical backplane ecosystem using embedded polymer waveguides and demonstration in a 12 G SAS storage array," *J. Lightwave Technol.*, vol. 31, no. 24, pp. 3970–3975, 2013.

[18] R. T. Chen, L. Lin, C. Choi, Y. J. Liu, B. Bihari, L. Wu, S. Tang, R. Wickman, B. Picor, M. K. Hibbs-Brenner, J. Bristow, and Y. S. Liu, "Fully embedded board-level guided-wave optoelectronic interconnects," *Proc. IEEE*, vol. 88, no. 6, pp. 780–793, 2000.

[19] J. Beals, N. Bamiedakis, A. Wonfor, R. V. Penty, I. H. White, J. V. DeGroot, Jr., K. Hueston, T. V. Clapp, and M. Glick, "A terabit capacity passive polymer optical backplane based on a novel meshed waveguide architecture," *Appl. Phys. A*, vol. 95, no. 4, pp. 983–988, 2009.

[20] X. Dou, A. X. Wang, X. Lin, H. Huang, and R. T. Chen, "Optical bus waveguide metallic hard mold fabrication with opposite 45 micro-mirrors," *Proc. SPIE*, vol. 7607, 76070P, 2010.

[21] N. Bamiedakis, A. Hashim, R. V. Penty, and I. H. White, "A 40 Gb/s optical bus for optical backplane interconnections," *J. Lightwave Technol.*, vol. 32, no. 8, pp. 1526–1537, 2014.

[22] A. Siokis, K. Christodoulopoulos, and E. Varvarigos, "Laying out interconnects on optical printed circuit boards," in *Proc. of the 10th ACM/IEEE Symp. on Architectures for Networking and Communications Systems*, 2014, pp. 101–112.

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

[24] L. N. Bhuyan and D. P. Agrawal, "Generalized hypercube and hyperbus structures for a computer network," *IEEE Trans. Comput.*, vol. C-33, no. 4, pp. 323–333, 1984.

[25] K. Li, Y. Pan, and S. Q. Zheng, *Parallel Computing Using Optical Interconnections*. New York: Springer, 1998.

[26] M. Tan, P. Rosenberg, J. S. Yeo, M. McLaren, S. Mathai, T. Morris, H. P. Kuo, J. Straznicky, N. P. Jouppi, and S. Wang, "A high-speed optical multi-drop bus for computer interconnections," *Appl. Phys. A*, vol. 95, no. 4, pp. 945–953, 2009.

[27] R. G. Melham, D. M. Chiarulli, and S. P. Levitan, "Space multiplexing of waveguides in optically interconnected multiprocessor systems," *Comput. J.*, vol. 32, no. 4, pp. 362–369, 1989.

[28] I. G. MacDonald, *Symmetric Functions and Hall Polynomials*. Oxford University, 1998.

[29] S. Markou, A. Siokis, P. Maniotis, K. Christodoulopoulos, E. Varvarigos, and N. Pleros, "Performance analysis and layout design of optical blades for HPCs using the OptoBoard-Sim simulator," in *IEEE Optical Interconnects Conf. (OIC)*, San Diego, Apr. 2015.

[30] M. Grange, R. Weerasekera, D. Pamunuwa, A. Jantsch, and A. Y. Weldezion, "Optimal network architectures for minimizing average distance in k-ary n-dimensional mesh networks," in *Proc. Fifth ACM/IEEE Int. Symp. on Networks-on-Chip (NoCS)*, Pittsburgh, 2011.

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

[32] R. Haring and the IBM BlueGene Team, "The BlueGene/Q compute chip," in *23rd Symp. on High Performance Chips (Hot Chips)*, vol. 4, Palo Alto, CA, 2011.

[33] B. Alverson, E. Froese, L. Kaplan, and D. Roweth, "Cray XC Series Network," White Paper WP-Aries01-1112, 2012.