Silicon Photonic Switch Topologies and Routing Strategies for Disaggregated Data Centers

Article *in* IEEE Journal of Selected Topics in Quantum Electronics · December 2019

DOI: 10.1109/JSTQE.2019.2960950

10 authors, including:

Qixiang Cheng
Columbia University
78 PUBLICATIONS **655** CITATIONS
SEE PROFILE

Keren Bergman
Columbia University
584 PUBLICATIONS **9,883** CITATIONS
SEE PROFILE

Meisam Bahadori
University of Illinois, Urbana-Champaign
56 PUBLICATIONS **484** CITATIONS
SEE PROFILE

Madeleine Glick
The University of Arizona
155 PUBLICATIONS **1,437** CITATIONS
SEE PROFILE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2960950, IEEE Journal of Selected Topics in Quantum Electronics

Silicon Photonic Switch Topologies and Routing Strategies for Disaggregated Data Centers

Qixiang Cheng, *Member, IEEE*, Yishen Huang, Hao Yang, Meisam Bahadori, Nathan Abrams, Xiang Meng, *Member, IEEE*, Madeleine Glick, *Senior Member, IEEE*, Yang Liu, Michael Hochberg, and Keren Bergman, *Fellow, IEEE*

(Invited paper)

***Abstract*— Disaggregation enabled by silicon photonic switch fabrics is a path to low-cost and energy-efficient data centers. The routing strategy, which can be seamlessly incorporated into the switch control plane, potentially provides an additional dimension for the physical-layer performance optimization, at no extra cost. In this paper, we analyze the role of optical routing strategies for silicon photonic switch fabrics. We define and quantify the number of global switching states in various switching topologies and discuss their relationship to the number of switch permutations. We propose a topology-agnostic approach that is shown to optimize fabric-wide switch path power penalties and consequently reduce the dynamic-range requirement on receivers. Additionally, it potentially compensates for device fabrication variations by taking advantage of the redundancy in switching states over switch permutations; thus, increasing fabrication tolerance. Significant power penalty improvements are demonstrated via both our simulation and test platforms, even for moderate-scale silicon switches.**

***Index Terms*—Optical switches, optical routing strategies, photonic integrated circuits.**

I. Introduction

With the explosive growth in data analytics that rely on machine and deep learning, there are increasing stresses on the computation within the nodes and the communication between the nodes. Current resources in datacenters are largely organized according to legacy architectures with static node configurations [1]. These static configurations of resources (compute, memory, storage) often result in the inefficient use of resources, with some being left idle while others are overtaxed. This is particularly the case for the different stages of machine learning algorithms of training and inference which use different mixes of resources [2]. Disaggregation of the traditional server has been proposed as a solution to improve efficiency [3], in which similar resources are pooled, with the possibility of the resources being adaptively configured for optimized performance and also independently upgraded. A disaggregated data center requires an interconnection fabric that must carry the additional traffic engendered by the disaggregation, and be high bandwidth and low latency in order to maintain high and improved performance. In addition, this interconnection network requires photonic switching fabrics to adaptively provision the computing resources [4], [5], as, for example, in the concept shown by Fig. 1.

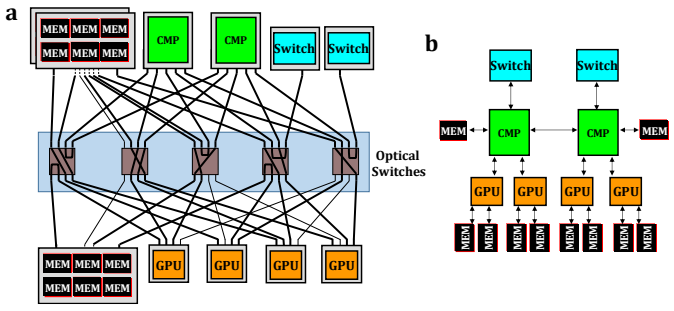

Fig. 1 (a) Concept of a Photonic Integrated Networked Energy efficient data center (PINE) network where optical switches are used in a disaggregated network. The nodes can be assembled as (b) by configuring the optical switches. Adopted from [3].

A number of optical switching technologies have been studied [6]. Manipulating light with free-space optics, such as three-dimensional micro-electro-mechanical systems (MEMS) [7] and beam-steering technology [8], provides the most mature optical switching solution that offers high port count scaling compared with other optical switching technologies. However, the requirement for rigorous calibration and installation of discrete components is costly. In addition, their electrostatic driver typically has a high yielding voltage of over 100 V [7]. On the other hand, lithography-based integration technologies hold great promise for large-scale optical integrated switch fabrics by reducing the device footprint and also the overhead in terms of assembly, calibration, and synchronization. Planar integrated optical switches have been developed so far on

Manuscript received xx xx, 2019. This work is supported by the ARPA-E ENLITENED program (DEAR00000843). (Corresponding author: Qixiang Cheng)

Q. Cheng, Y. Huang, H. Yang, N. Abrams, X. Meng, M. Glick, and K. Bergman are with the Department of Electrical Engineering, Columbia University, New York, NY 10027, USA. Email: qc2228@columbia.edu; yh2785@columbia.edu; hy2408@columbia.edu; nca2123@columbia.edu; xm2137@columbia.edu; msg144@columbia.edu; bergman@ee.columbia.edu.

M. Bahadori, Y. Liu and M. Hochberg are with Elenion Technologies, 171 Madison Ave. STE 1100, New York, NY 10016, USA. E-mail: meisam.bahadori@elenion.com; yhvhliuyang@gmail.com; michael.hochberg@elenion.com.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2960950, IEEE Journal of Selected Topics in Quantum Electronics

TABLE I
COMMONLY APPLIED OPTICAL SWITCH ARCHITECTURES

Blocking Characteristic	Architecture	Order of Crosstalk	Total Number of Switch Elements	Number of Switch Stages	Global Switching States
Blocking	Banyan-type Network	First	$\frac{N}{2}log_2N$	log_2N	$2^{\frac{N}{2}log_2N}$
RNB	Beneš	First	$\frac{N}{2}(2log_2N-1)$	$2log_2N$-1	$2^{\frac{N}{2}(2log_2N-1)}$
RNB	Dilated Beneš	Second	$2Nlog_2N$	$2log_2N$	$2^{\frac{N}{2}(2log_2N-1)}$
RNB	N-stage planar	First	$\frac{N}{2}(N-1)$	N	$2^{\frac{N}{2}(N-1)}$
WSNB	PILOSS	First	N^2	N	N!
WSNB	Crossbar	First	N^2	$2N-1$	N!
SNB	Switch-and-Select	Second	$2N(N-1)$	$2log_2N$	N!
SNB	Dilated Banyan	Second	$2N(N-1)$	$2log_2N$	N!

RBB: Rearrangeably non-blocking; WSNB: Wide-sense non-blocking; SNB: Strictly non-blocking

various material platforms, such as indium phosphide, lithium niobate, silica, and silicon [9]–[15].

Given the requirement for high bandwidth density at low cost and low power consumption for data centers [4], it is not surprising that silicon photonics, fabricated in high volume CMOS compatible foundries, draws the most attention [16]. The most widely applied switching mechanisms in silicon photonics include phase manipulation in interferometric structures, i.e. Mach-Zehnder interferometer (MZI) [15] and microring resonator (MRR) [13], both thermally or electrically actuated, and MEMS-actuated coupling between adjunct waveguide layers [12]. N×N optical switch fabrics can be built up by interconnecting multiple stages of elementary switch cells in a defined switch topology via passive waveguide shuffles. With the scale-up of switch port count, the requirement of efficient calibration as well as control/routing methods for such complex integrated circuits are severe and urgent.

While a number of calibration techniques have been studied to facilitate fast and accurate characterization of optical switching circuits [17]–[25], research on photonic switch routing control has been sparsely reported. The routing algorithms in electronic switches are primarily developed to resolve contentions, such as the classic looping algorithm [26], where all paths are seen equal. However, the optical integrated switches, as repeaterless fabrics where signal's amplitude and timing margins are not restored, generally have path-dependent performance. Such path-dependent variation normally lies in imperfect fabrication and design limitations, which could cause a discrepancy in the different switching states of the elementary cells, bound the switch performance, and lead to excess loss in waveguide shuffling. A photonic routing strategy should take these factors into consideration. To date, the scanty reported routing algorithms in optical switch fabrics tend to be on a case by case basis, such as loss and input power dynamic range (IPDR) improvement in Clos network [27], and crosstalk reduction in dilated Banyan topology [28]. These topology-specific routing strategies do not represent a generic solution and may not often provide a global performance optimization considering the fabrication variations.

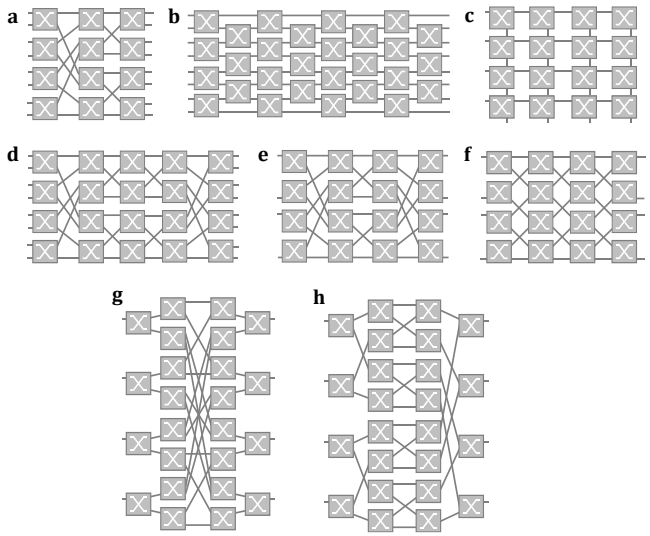

Fig. 2 Schematic of switch architectures: (a) Banyan-type, (b) N-stage planar, (c) crossbar, (d) Beneš, (e) dilated Beneš, (f) PILOSS, (g) switch-and-select, and (h) dilated Banyan. Note that (a), (b) and (d) have a scale of 8×8, while the rest are in 4×4 due to the space limitation.

We recently proposed a generic optical routing strategy that optimizes fabric-wide power penalties [29] and this paper extends the work by providing an in-depth analysis of the role of optical routing strategies for photonic integrated switch fabrics with different blocking characteristics. We review the commonly applied optical switching architectures and quantify the number of global switching states, revealing their relationship to the number of switch permutations. The generic penalty-optimized routing approach that opts for the most favorable switch configuration is presented and verified via both our simulation and test platforms. This method is shown to significantly improve the worst-case path power penalty even for a moderate-scale switch fabric (8×8), which is critical for disaggregated data centers [4].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2960950, IEEE Journal of Selected Topics in Quantum Electronics

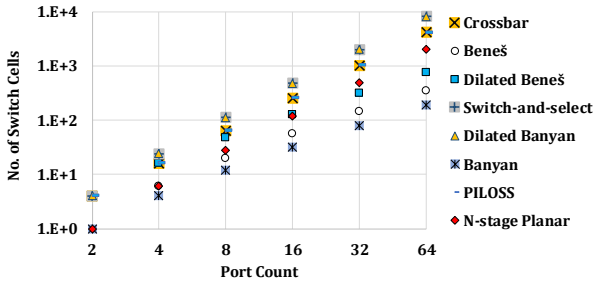

Fig. 3 Comparison of the total number of switch cells among various switch topologies as a function of port number N in an N×N network.

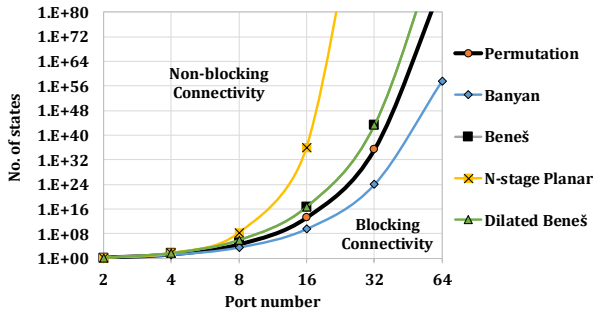

Fig. 4 Global switching states for Banyan-type, Beneš, N-stage planar and dilated Beneš networks, with the number of permutations.

The remainder of this paper is organized as follows. Section II reviews the commonly applied optical switching architectures, illustrates their global switching states and describes the role of optical routing strategy. Section III presents the proposed generic penalty-optimized routing approach and also shows the verification of the proposed method via both simulation and experiments. Finally, conclusions are drawn in Section V.

II. Switch Topologies and Global Switching States

It has been recognized that switching device performance depends critically on the selection of topology, including switch blocking characteristics, crosstalk suppression, total number of switch cells and number of cascading stages [4]. Some of the classical switch architectures, such as crossbar, Banyan-type, Clos and Beneš networks, are adopted from electronic switch network designs, while the others are made by pioneers in optical switch fabrics, especially those with innovations to offset the limitations of photonic integration technologies. For instance, N-stage planar architecture was proposed to eliminate waveguide crossings, path-independent loss (PILOSS) network was designed to achieve a loss uniformity across all paths, and dilated networks were exploited to completely cancel the first order crosstalk. We list the commonly applied optical switch architectures, namely Banyan-type, N-stage planar, crossbar, Beneš, dilated Beneš, PILOSS, switch-and-select, and dilated Banyan (shown in Fig. 2a-h), in Table I based on their blocking characteristics, together with other key figure of merits, such as the order of crosstalk, total number of switch elements and number of switch stages. We focus on multi-stage architectures based on 2×2 switching elements and define the *global switching states* as the total number of switch configurations

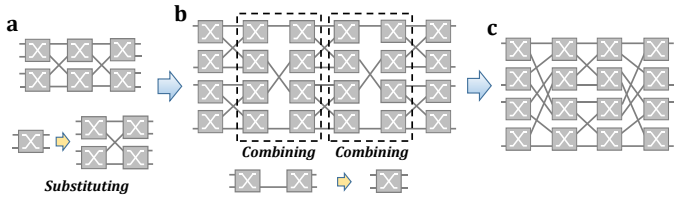

Fig. 5 Construction of a dilated Beneš network: (a) Substituting 2×2 switch cell by dilated four-switch-element in a Beneš network, and (b) combining redundant switching cells. (c) Schematic of a dilated Beneš network.

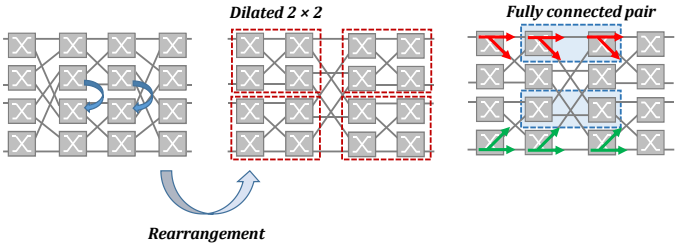

Fig. 6 Rearrangement of a 4×4 dilated Beneš network to identify its global switching states. The two pairs of MZI cells outlined in dashed blue boxes logically form a dilated 2×2 block, named as a fully connected pair.

that lead to valid input/output permutations. The number of global switching states is topology dependent, relating to the total number of switch cells and the way of their logical connections. In this section, we investigate the global switching states for various switch topologies with relevance to their blocking features, and reveal their relationship to the number of switch permutations. We also show that while the idle cells in a switch fabric do not contribute to the global switching states, their settings affect the switch fabric-wide performance.

A. *Banyan-type blocking networks*

Banyan-type switch fabric, which was proposed for use in computer networks, is also attractive for high-speed electronic and optical switching applications. A Banyan-type network is defined as a class of multistage networks that have no path diversity, and there exists some variants including banyan, omega, baseline and n-cube [30], differing in the inter-stage connecting pattern.

The Banyan-type switch fabric provides the minimum diameter for a fully connected but blocking network, which means that any input port has a full connectivity to any output port provided that no contentions with existing connections in the network [31]. It thus has the minimum number of switch cells as illustrated by Fig. 3, which compares the number of switch cells among the switch topologies listed in Table I, leading to the minimum footprint. Banyan-type network suffers first-order crosstalk as any 2×2 elementary switch cell is traversed by two signals at once. Each switch cell configures two switching states, i.e. cross and bar, leading to $2^{\frac{N}{2}log_2 N}$ global switching states. This is smaller than the number of permutations (N!, i.e. factorial N) in an N×N switch fabric, as shown in Fig. 4, evidencing its blocking connectivity. This type of network exhibits no redundancy in path settings and therefore, leaves no room for routing strategies.

B. *Rearrangeably non-blocking networks*

Rearrangeably non-blocking switch fabrics have so far

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2960950, IEEE Journal of Selected Topics in Quantum Electronics

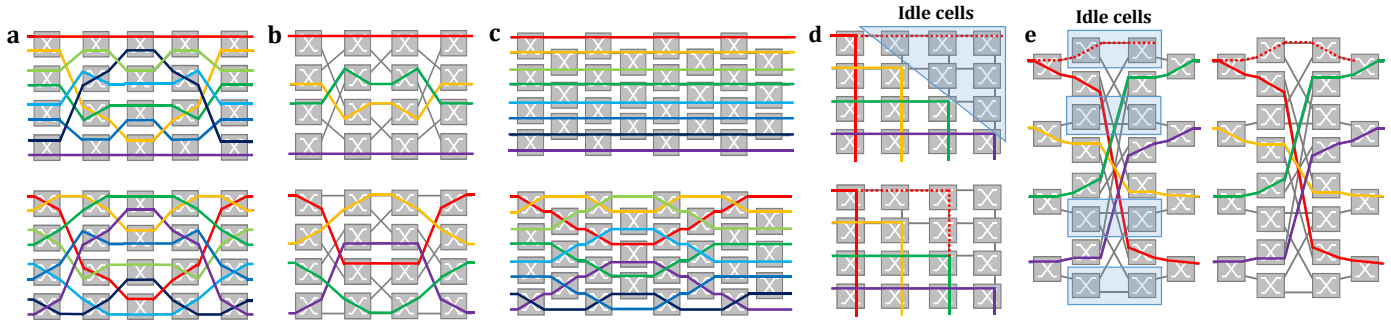

Fig. 7 Routing variations for (a) Beneš, (b) dilated Beneš, (c) N-stage planar, (d) crossbar, and (e) switch-and-select networks. Dotted lines represent paths of first-order crosstalk. (d) and (e) show the impact of settings of idle cells.

drawn the most attention, since this type of switch networks typically require smaller number of switch cells, as revealed in Fig. 3. This is a critical consideration in photonic integrations that translates into lower design and fabrication complexity. However, one should note that this type of networks can only establish connections of all permutations of input ports to output ports if rerouting existing connections are allowed.

Beneš topology is undoubtedly the most popular selection, because it has the minimum number of switch cells to construct a non-blocking N×N network. Therefore, it can have small footprint and power consumption. It was derived from the Clos architecture, with a bank of 2×2 switch cells forming each of the input and output stages and recursively replacing the center stage by smaller size Clos networks until it reaches 2×2. This produces $2log_2N$-1 stages with in a total of $\frac{N}{2}(2log_2N-1)$ cells. Beneš topology also encounters first-order crosstalk. Comparing to Banyan-type blocking networks, the additional switch stages not only enhance the connectivity but also provide the path diversity, resulting in $2^{\frac{N}{2}(log_2N-1)}$ global switching states, larger than the switch permutations.

N-stage planar network is also widely picked for building modest-scale switch fabrics [32], [33]. As mentioned above, this design was motivated to avoid optical crossovers [34], as a derivative from the crossbar architecture. The N-stage planar architecture has $\frac{N}{2}(N-1)$ primitive 2×2 cells, and a path-dependant performance with paths traversing N/2 to N switch cells. The global switching states can be obtained at $2^{\frac{N}{2}(N-1)}$, as plotted in Fig. 4. The N-stage arrangement provides a strong path redistribution capability that largely increases path diversity and thus, leading to a huge discrepancy between the global switching states and switch permutations. This, however, also leads to a larger footprint than the Beneš network.

The first-order crosstalk imposes stringent requirements on the design of elementary switch cells. Dilation can be used to modify the switch fabric so that each individual switch cell only carries one signal at once, dramatically reducing crosstalk (to the square of the crosstalk ratio of an elementary switch cell). The construction of a dilated Beneš network is indicated in Fig. 5, which is at the expense of using a greater number of switching elements ($2Nlog_2N$), in $2log_2N$ stages. Deciding the global switching states is, however, not straight-forward. Each switch cell can still be configured in two states, but only carries one signal at a time, and thus, not all switch configurations are valid. We can, however, rearrange the switch cells as shown in Fig. 6, and, while keeping the same logical connections, implement a fabric that comprises 2×2 dilated blocks [35]. The dashed blue boxes outline two pairs of MZI elements from adjacent dilated 2×2 blocks connected via a shuffle. The two pairs of MZI cells logically form a dilated 2×2 block, named as a fully connected pair. Considering each MZI cell only carries one signal at once, the fully connected cell pair can only be configured in 2 states. Therefore, we can see that each 2×2 dilated block has 4 switching states except the last stage who can only be set in 2 states. We can then have its global switching states as $2^{\frac{N}{2}(2log_2N-1)}$, which is the same as a Beneš network.

For the topologies that have global switching states larger than the switch permutations, there exists an element of repetition so that different switch configurations can lead to the same switch permutation (two routing variations are shown in Fig. 7a-c for Beneš, dilated Beneš, and N-stage planar, respectively). Therefore, a smart routing strategy would take advantage of such repetition and opt for the most favorable configuration. Particularly, in the rearrangeably non-blocking networks, such a routing strategy serves for both path mapping that cancels out the path diversity and fabric-wide performance optimization. We show details in Section III.

C. Wide-sense and strictly non-blocking networks

Wide-sense and strictly non-blocking networks can set up paths between any idle inputs to any idle outputs without interference with existing connections. The difference lies in that routing in the former type of networks needs to follow certain rules while the latter has no restrictions. This type of networks is generally favored due to the simplified switching control system.

Crossbar and PILOSS are two typical wide-sense non-blocking networks. Both topologies require N^2 switching cells and are not fully immune to first-order crosstalk, but PILOSS wins with respect to loss uniformity since any path traverses exactly N switch cells [36]. This number, however, varies from 1 to 2N-1 for crossbar architecture. The two-fold increase in the number of switch elements also doubles their footprint comparing to the N-stage planar network. They both have regulations that optical routing paths contain strictly one switch

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2960950, IEEE Journal of Selected Topics in Quantum Electronics

cell in bar state with the rest setting to cross, which rules out path diversity. This leads to global switching states of N! that equals to the switch permutations.

Both switch-and-select and dilated Banyan architectures belong to the strictly non-blocking network category. Switch-and-select topology follows the binary-tree logic that comprises two linear switching arrays, respectively for signal fan-out and fan-in, interconnected by a perfect central shuffle. Dilated Banyan has the same logical connections but relocates switch cells. They both have $2N(N-1)$ switch cells arranged in $2log_2N$ stages and completely block first-order crosstalk, which gives rise to the largest footprint among the networks listed in Table I. Note that the rearrangement in the dilated Banyan topology reduces the total number of waveguide crossings in the waveguide shuffle networks comparing to the switch-and-select network, making the dilated Banyan slightly stand out in terms of footprint. Such connect logic provides a dedicated path for any input to output pair with no path diversity. This regulates their global switching states both at N!.

Nevertheless, one should note that, even if their global switching states equal to switch permutations, there often exist idle cells in this category of switch networks, given the large number of switch cells used. In a switch permutation, while the idle cells do not contribute to the global switching states, i.e. they do not get traversed by optical signals, their settings still impact the overall switch performance. Two examples are shown in Fig. 7d and 7e, respectively for crossbar and switch-and-select networks, revealing that the idle cells decide the crosstalk paths. Whereas the problem can be simply solved in a crossbar switch by regulating all the idle cells in cross state, it requires a more intelligent method to configure the switch-and-select networks in order to achieve a fabric-wide performance optimization. This can be realized using the proposed routing strategy described in Section III.

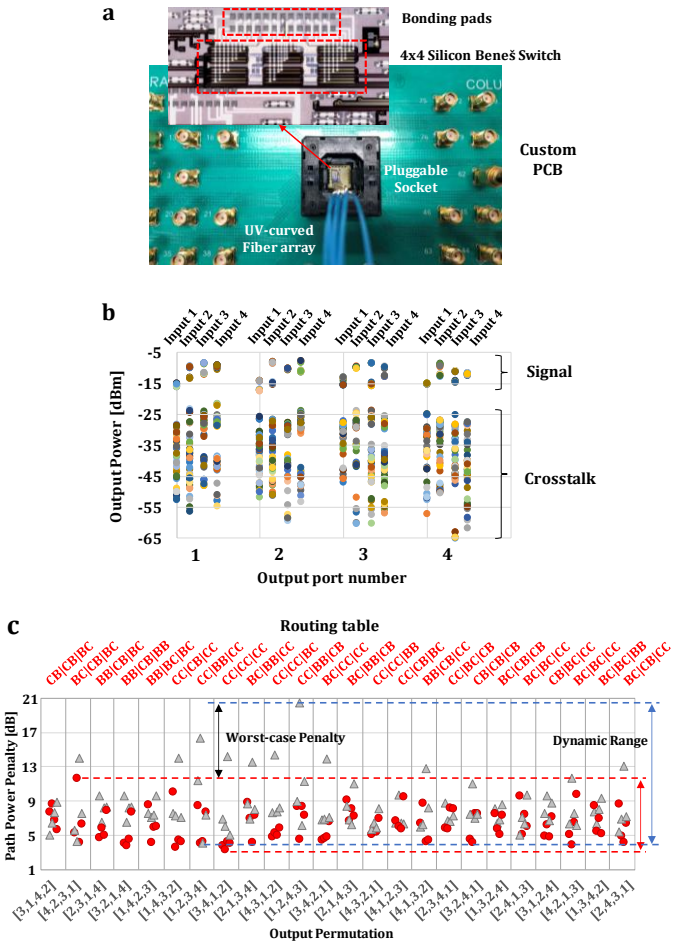

Fig. 8 (a) Switch test-bed with enlarged inset showing the OPSIS 4×4 silicon MZI-based Beneš switch photo. (b) Measured full optical power map. (c) Detailed port-to-port power penalty of both selected and worst-case switch settings for all 24 permutations. "C" and "B" stand for cross state and bar state, respectively. The routing table is column-wise.

III. Proposed Routing Strategy

In this section, we present the proposed generic routing strategy that is capable of optimizing the fabric-wide switch performance, quantified in optical power penalties. The routing strategy is topology-agnostic and can be incorporated into the switch control plane, by re-defining the look-up table. Quantifying path power penalties requires a full characterization of the switch device, which can leverage the developed fast and automated calibration methods [17]–[25]. For instance, calibrating a 32×32 PILOSS silicon switch containing 1024 MZI cells can be done within 90 seconds [22]. The full device characterization consequently takes fabrication variation and cross-wafer non-uniformity into consideration, creating routing tables that are aware of physical-layer parameters. We also demonstrate the proposed approach via both simulation and experiment platforms.

A. Methodology

Optical power transfer functions are needed to be first verified in the N×N switch matrix for each switching state, i.e. sweeping all switch cells in both cross and bar states, acquiring a full optical power transfer map. With one input injected with optical signal (ϱ_i) each time, the power transfer map comprises the output signal power ($\rho_{i,j}$, for optical path from input i to output j) together with N-1 leakages ($\sigma_{i,j,k}$, where k denoting the leakage channel and $k \neq j$), revealing the path insertion loss ($\varepsilon_{i,j}$) and crosstalk ratio ($\kappa_{i,j,k}$) to the rest ports, where:

$$\varepsilon_{i,j} = \rho_{i,j}/\varrho_i \quad (1)$$

$$\kappa_{i,j,k} = \sigma_{i,j,k}/\rho_{i,j} \quad (2)$$

For one switching state, the relation between input port i and output port j have to be a bijection, i.e. $i = F(j)$, $j = F^{-1}(i)$, which means each output port has its one-to-one correspondent input port. The aggregated crosstalk power at full switch load to output k can be found as:

$$\mu_k = \sum_{j=1,j\neq k}^{N} \sigma_{i=F(j),j,k}. \quad (3)$$

and the extinction ratio for output k can be written as:

$$\epsilon_k = \mu_k/\rho_{i,j}\,. \quad (4)$$

The aggregated crosstalk-induced power penalty can thus be estimated as [37]:

$$\delta_k = -10\log\left(1 - 2\sqrt{\epsilon_k}\right). \quad (5)$$

We can then have the total path power penalty as:

$$PP_{i,j} = -10\log\left(\varepsilon_{i,j}\right) + \delta_k \quad (6)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2960950, IEEE Journal of Selected Topics in Quantum Electronics

TABLE II
EXTRACTED PHYSICAL-LAYER PARAMETERS

Parameters	Value
Waveguide Propagation Loss	2 dB/cm
Waveguide Crossover Loss	0.05 dB
Waveguide Crossover Induced Crosstalk	-37 dB
Directional Coupler Loss	0.1 dB
Directional Coupler Coupling Coefficient	0.67

in dB. A path power penalty map can be simply captured using the optical power transfer functions.

The subsequent procedure would be the classification of optical switching states based on N! switch permutations. One can quickly convert any optical switching state, i.e. switch fabric configuration, to a switch permutation using the transfer matrix method [38]. In this stage, the defined global switching states in the rearrangeably non-blocking architectures can be categorized into each switch permutation, while for the wide-sense and strictly non-blocking architectures, the redundant switching states due to the idle cells can be identified, referring back to Section II. We can then process all the repetitive switching states for each switch permutation and opt for the most favourable option.

Deciding the selection metric for the wide-sense and strictly non-blocking switch fabrics is straight-forward, and because the routing paths have no diversity, the switching state settings for idle cells should optimize the crosstalk-induced penalty, thus the overall device performance. For simplify, we only need to pay attention to the term δ_k in Eq. (6), and select the switching state that minimizes δ_k. However, for rearrangeably non-blocking networks, the path diversity should be taken into consideration also. A few weighting metrics can be used. For instance, select the switching state that has the best worst-case path power penalty, or choose the one with the minimum power penalty deviations. The former guarantees the maximum power penalty reduction; however, we specifically examine the latter case in this paper considering that the dynamic range of receivers also stands out as a key limiting factor in the switching system. For one switching state that comprises of N routing paths, the root-mean-square error (RMSE), η, of path power penalties can be calculated, where:

$$\overline{PP} = (\sum_{i=1}^{N} PP_{i,j})/N, \tag{7}$$

$$\eta = \sqrt{(\sum_{i=1}^{N}(PP_{i,j} - \overline{PP}))^2/N}. \tag{8}$$

The results can be automatically sorted for all repetitive switching states that map to one switch permutation and opted for the one with η_{min}. This process repeats for all permutations generating a full look-up table. Note this method is agnostic to switch scales (N) and can be seamlessly incorporated into the switch control plane.

The pre-defined look-up table is part of a typical control architecture for photonic switch circuits [39]. Therefore, the routing algorithm is transparent to the switch control plane, introducing no extra complexity or latency.

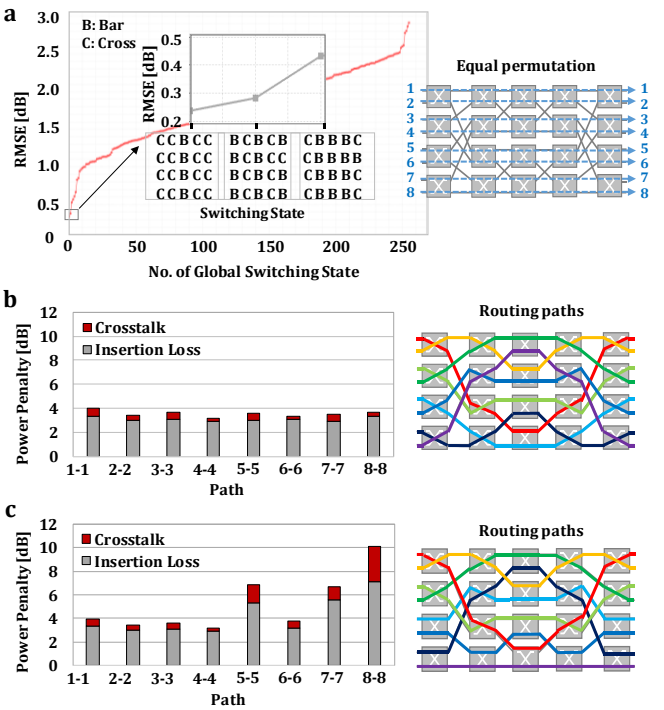

Fig. 9 (a) Sorted RMSE of path power penalties for 256 switching states that map to the equal permutation. (b) Path power penalties for the selected routing paths for the equal permutation. (c) Path power penalties for the worst-RMSE routing paths.

B. *Routing Table for an Electro-Optic 4×4 Beneš Switch Device*

We use Beneš, the most popular rearrangeably non-blocking network, to demonstrate the proposed routing strategy that provides optimal routing paths to simultaneously cancel path diversity and minimize fabric-wide path power penalties. The 4×4 electro-optic Beneš switch device, as shown by the microscope photo in Fig. 8a, was fabricated at the Institute of Microelectronics via an OpSIS multi-project-wafer (MPW) run. It consists of six 2×2 MZI switch cells and each arm of the cell is equipped with a thermal tuner for device calibration and a fast PN modulator for switching. The 4-port Beneš switch can be configured in 64 (2^6) global switching states that map to 24 (4!) permutations. The device was die-bonded onto a chip carrier clamped on a custom PCB fan-out board, as shown by Fig. 8a. A fiber-array was UV-cured on the top surface of the silicon chip coupling to an array of grating couplers. The switch cells were initialized in Cross state by calibrating the two arms with a thermal tuner. Switching to Bar state can be done by applying a voltage to the PN phase modulator with a V_π of ~1.3 V.

The switch device was subsequently characterized for all 64 switching states by measuring the output power of all four ports with input power injected from port 1 to 4 in sequence. An automated process can be leveraged as shown in [20]. Aside from one output carrying switched signal, the leakage power to the other three ports denotes crosstalk. Figure 8b summarizes the output power of all 1024 cases (64 switching states × 4 input × 4 outputs), grouping into switched signals (64×1×4 = 256

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2019.2960950, IEEE Journal of Selected Topics in Quantum Electronics

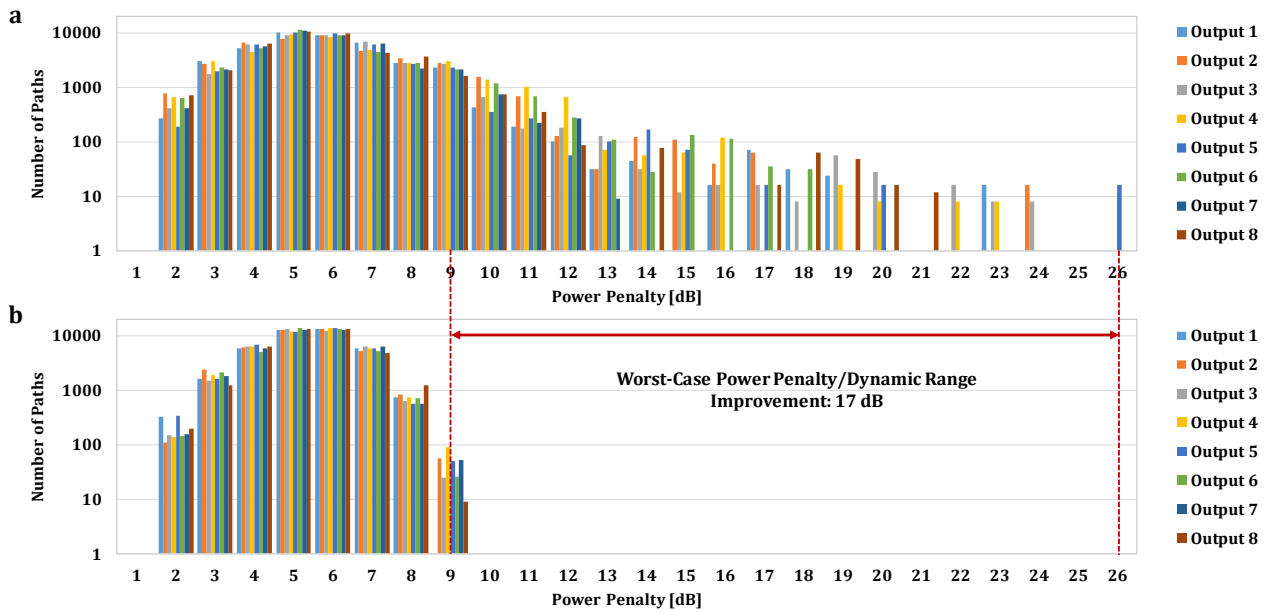

Fig. 10 Histogram of path power penalties for (a) worst-case routing and (b) optimized routing for all 40320 (8!) switch permutations, indicating an improvement of ~17 dB for both the worst-case path power penalty and dynamic range.

paths), and crosstalk leakages (in total 64×3×4). The 4-port device exhibits on-chip insertion losses ranging between 1 dB and 7.5 dB, i.e. a path-dependent loss of 6.5 dB, and crosstalk ratio below -15 dB. The loss variations mainly lie in optical shuffling that comprises different propagation distances and numbers of crossovers, and the loss discrepancy in Bar and Cross states due to free-carrier absorption. The electro-absorption loss can be improved through the use of optical phase bias [14] that decreases the phase shift need to half pi. This reduces optical loss and the bounded crosstalk ratio. Further discussions on component loss improvement are provided in Section IV.

As shown by Eqs. (3)-(6), by aggregating leakage crosstalk and translating into crosstalk-induced penalty, a power penalty map can be obtained for all 256 light paths corresponding to 64 global switching states. For each switching state, the RMSE of power penalties for the four light paths is programmed to be calculated at the full switch load case. The results are then automatically sorted with the least value selected, and stored as a routing look-up table that equalizes path-dependent power penalties and avoids the worst-case routing paths. Figure 8c presents the detailed path power penalties of both selected (in red circles) and worst-case (in grey triangles) switching states for all 24 permutations, in the form of output permutation ([1,2,3,4,]→[a,b,c,d]). One can see that with the defined routing table, the worst-case path power penalty can be improved by 8.8 dB. This can be attributed to the reduced number of Bar states that reside in the selected switching states, since electro-absorption loss induced by the phase shift in the Bar state contributes to both insertion loss and crosstalk. The use of routing strategy also leads to a dramatic reduction in the dispersion of path power penalties, from 16.1 dB to 8.3 dB. This results in a large relaxation on the dynamic range required for the receivers. A complete routing table for the integrated 4×4 silicon photonic switch is also shown in Fig. 8c, where C and B stands for “Cross state” and “Bar state”, respectively.

C. Simulated 8×8 Beneš Switch Fabric

To further study the benefit of the proposed routing strategy in a scaled switch fabric, we set up an 8×8 Beneš switch model using our generic cross-layer simulation platform, *PhoenixSim* [40]. This simulation tool enables integrated and interactive design space exploration over the physical, networking and application layers. It builds up physical-layer compact models of silicon photonic components, including waveguides, bends, crossovers, directional couplers, PIN phase shifters, and thermal tuners, and from which system-level metrics such as loss and crosstalk can be extracted.

We use the physical-layer parameters that fitted from the measurements, shown in Fig. 8b. The passive loss of each path is a linear combination of five loss sources: MZI in Cross state, MZI in Bar state, passive shuffle with crossover, passive shuffle with straight waveguide, and grating coupler. Thus, parameters can be extracted and averaged from a combination of measurements, as listed in Tab II. An 8×8 Beneš switch fabric is subsequently defined, where multiple stages of MZI cells are installed with a comprehensive model of passive shuffles that reflects the discrepancy in waveguide propagation lengths and number of crossovers for different paths. Eight input ports are simultaneously injected with optical powers to enable fast assessment of eight paths, and this process is consecutively executed for all 2^{20} switching states. The results are then processed by individually calculating RMSE and sorting out under 40320 (=8!) permutations to generate a full routing table, similar to the procedures described in section III.B.

We first take a close look at the case of equal permutation, i.e. [1,2,3,4,5,6,7,8]→[1,2,3,4,5,6,7,8], containing the highest switching diversity (256 switching states). Figure 9a outlines

RMSE of power penalties for all switching states sorted by their values, with the optimum selected as "CCCC | CCCC | BBBB | CCCC | CCCC" (column-wise). The detailed power penalty breakdown as well as routing paths is illustrated in Fig. 9b. For comparison, the switching state that has the worst RMSE is also presented in Fig. 9c, together with its routing paths, indicating that the worst-case power penalty can be reduced by up to 6 dB for the equal permutation. It is clearly shown that both path loss and crosstalk-induced penalty can be effectively reduced benefiting from the routing strategy.

An overview of path power penalties for all 40320 (8!) switch permutations is shown in Fig. 10 by two histograms that describe the power penalty distribution respectively for the worst-case and optimized routing. It can be seen that the distribution for the worst-case routing (Fig. 10a) is significantly more diverse than the one with the proposed routing strategy (Fig. 10b). The worst-case power penalty is improved by ~17 dB and so does the dynamic range. The worst-case power penalty has a large contribution from the crosstalk, which is believed to be caused by worst-case aggregation of crosstalk power at the output port. This gives rise to a high crosstalk penalty that significantly deteriorates the link performance. We conclude that such a case can be avoided by applying the proposed routing strategy.

D. Discussion

It is not a surprise that the worst-case power penalty improvement in the 8×8 Beneš switch fabric is evidently higher than the 4×4 one, since larger N×N switch matrices, assembled by a larger number of stages, feature a higher number of global switching states (as shown in Fig. 4). The larger number of switching stages provides stronger redistribution capability and thus exhibit stronger diversity. Such diversity generates not only path diversity but also differences in loss and crosstalk in the paths due to differences in the characteristics of the bar and cross states of each elementary cell, as well as the variations in passive shuffle networks as the example shown in [41]. Therefore, a larger scale silicon switch fabric, such as 16×16 and 32×32, is expected to benefit more from utilizing the proposed routing strategy in terms of worst-case power penalty reduction for disaggregated data centers.

To further reduce switch power penalties for data center adoption [6], it is vital to improve the intrinsic loss of silicon components. For example, losses should be reduced as in the following: waveguide propagation loss of <0.5 dB/cm [42], waveguide crossing loss of <0.01 dB [43], Y-junction coupler loss of <0.1 dB [44], thermo-optic MZI element loss of <0.2 dB [15], and coupling loss of <1 dB [45]. Note that the switch control scheme will also have an impact on the device performance, such as the push-pull approach that reduces drive voltage for lowering electro-absorption loss and crosstalk ratio [23], [24] in contrast to the single-arm drive used in this paper, which can be combined with the routing strategy.

IV. Conclusions

Silicon photonic switch fabrics hold great promise for disaggregated data center networks that provide agile bandwidth reconfigurability to dramatically improve compute performance and energy efficiency. In addition to their low cost and high energy efficiency, the inserted power penalty stands out as a key challenge. The routing strategy, which can be seamlessly incorporated into the switch control plane, potentially provides an additional dimension for the physical layer performance optimization, at no extra cost.

In this paper, we analyze the role of optical routing strategies for silicon photonic switch fabrics. By defining and quantifying the number of global switching states in various switching topologies, we reveal their relationship to the number of switch permutations and show how to leverage such a redundancy to optimize fabric-wide switch path power penalties. Significant power penalty improvements with low diversity from path to path are demonstrated via both our simulation and test platforms, even for moderate-scale silicon switches. This is also shown as a great potential to compensate for device fabrication variations and thus, increasing fabrication tolerance.

References

[1] A. Singh *et al.*, "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network," in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication - SIGCOMM '15*, London, United Kingdom, 2015, pp. 183–197.

[2] W. J. Dally, C. T. Gray, J. Poulton, B. Khailany, J. Wilson, and L. Dennison, "Hardware-Enabled Artificial Intelligence," in *2018 IEEE Symposium on VLSI Circuits*, 2018, pp. 3–6.

[3] P. X. Gao *et al.*, "Network Requirements for Resource Disaggregation," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, Berkeley, CA, USA, 2016, pp. 249–264.

[4] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "Recent advances in optical technologies for data centers: a review," *Optica*, vol. 5, no. 11, pp. 1354–1370, Nov. 2018.

[5] K. Wen *et al.*, "Flexfly: Enabling a Reconfigurable Dragonfly through Silicon Photonics," in *SC '16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2016, pp. 166–177.

[6] Q. Cheng, S. Rumley, M. Bahadori, and K. Bergman, "Photonic switching in high performance datacenters [Invited]," *Opt. Express*, vol. 26, no. 12, pp. 16022–16043, Jun. 2018.

[7] J. Kim *et al.*, "1100 x 1100 port MEMS-based optical crossconnect with 4-dB maximum loss," *IEEE Photonics Technol. Lett.*, vol. 15, no. 11, pp. 1537–1539, Nov. 2003.

[8] A. Dames, "Beam steering optical switch," US20070091484A1, 26-Apr-2007.

[9] M. Ding, A. Wonfor, Q. Cheng, R. V. Penty, and I. H. White, "Hybrid MZI-SOA InGaAs/InP Photonic Integrated Switches," *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 1, pp. 1–8, Jan. 2018.

[10] P. J. Duthie and M. J. Wale, "16*16 single chip optical switch array in lithium niobate," *Electron. Lett.*, vol. 27, no. 14, pp. 1265–1266, Jul. 1991.

[11] S. Sohma, T. Watanabe, N. Ooba, M. Itoh, T. Shibata, and H. Takahashi, "Silica-based PLC Type 32 x 32 Optical Matrix Switch," in *2006 European Conference on Optical Communications*, 2006, pp. 1–2.

[12] T. J. Seok, T. J. Seok, K. Kwon, J. Henriksson, J. Luo, and M. C. Wu, "240×240 Wafer-Scale Silicon Photonic Switches," presented at the Optical Fiber Communication Conference, 2019, p. Th1E.5.

[13] Q. Cheng *et al.*, "Ultralow-crosstalk, strictly non-blocking microring-based optical switch," *Photonics Res.*, vol. 7, no. 2, pp. 155–161, Feb. 2019.
[14] L. Qiao, W. Tang, and T. Chu, "32 × 32 silicon electro-optic switch with built-in monitors and balanced-status units," *Sci. Rep.*, vol. 7, p. 42306, Feb. 2017.
[15] K. Suzuki *et al.*, "Low-Insertion-Loss and Power-Efficient 32 × 32 Silicon Photonics Switch With Extremely High-Δ Silica PLC Connector," *J. Light. Technol.*, vol. 37, no. 1, pp. 116–122, Jan. 2019.
[16] M. Hochberg *et al.*, "Silicon Photonics: The Next Fabless Semiconductor Industry," *IEEE Solid-State Circuits Mag.*, vol. 5, no. 1, pp. 48–58, winter 2013.
[17] P. Dumais *et al.*, "Silicon Photonic Switch Subsystem With 900 Monolithically Integrated Calibration Photodiodes and 64-Fiber Package," *J. Light. Technol.*, vol. 36, no. 2, pp. 233–238, Jan. 2018.
[18] A. Annoni *et al.*, "Automated Routing and Control of Silicon Photonic Switch Fabrics," *IEEE J. Sel. Top. Quantum Electron.*, vol. 22, no. 6, pp. 169–176, Nov. 2016.
[19] M. S. Hai *et al.*, "Automated characterization of SiP MZI-based switches," in *2015 IEEE Optical Interconnects Conference (OI)*, 2015, pp. 94–95.
[20] Y. Huang, Q. Cheng, N. C. Abrams, J. Zhou, S. Rumley, and K. Bergman, "Automated Calibration and Characterization for Scalable Integrated Optical Switch Fabrics without Built-in Power Monitors," in *2017 European Conference on Optical Communication (ECOC)*, 2017, pp. 1–3.
[21] D. A. B. Miller, "Setting up meshes of interferometers - reversed local light interference method," *Opt. Express*, vol. 25, no. 23, pp. 29233–29248, Nov. 2017.
[22] S. Suda *et al.*, "Fast and Accurate Automatic Calibration of a 32 × 32 Silicon Photonic Strictly-Non-Blocking Switch," presented at the Photonics in Switching, 2017, p. PTu3C.5.
[23] Y. Huang, Q. Cheng, and K. Bergman, "Automated Calibration of Balanced Control to Optimize Performance of Silicon Photonic Switch Fabrics," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, 2018, pp. 1–3.
[24] Y. Huang, Q. Cheng, and K. Bergman, "Crosstalk-aware Calibration for Fast and Automated Functionalization of Photonic Integrated Switch Fabrics," in *Conference on Lasers and Electro-Optics (2018), paper STh3B.6*, 2018, p. STh3B.6.
[25] A. Gazman *et al.*, "Tapless and topology agnostic calibration solution for silicon photonic switches," *Opt. Express*, vol. 26, no. 25, pp. 32662–32674, Dec. 2018.
[26] D. C. Opferman and N. T. Tsao-wu, "On a class of rearrangeable switching networks part I: Control algorithm," *Bell Syst. Tech. J.*, vol. 50, no. 5, pp. 1579–1600, May 1971.
[27] M. Ding, Q. Cheng, A. Wonfor, R. V. Penty, and I. H. White, "Routing algorithm to optimize loss and IPDR for rearrangeably non-blocking integrated optical switches," in *2015 Conference on Lasers and Electro-Optics (CLEO)*, 2015, pp. 1–2.
[28] Y. Qian *et al.*, "Crosstalk optimization in low extinction-ratio switch fabrics," in *OFC 2014*, 2014, pp. 1–3.
[29] Q. Cheng, Y. Huang, M. Bahadori, J. Zhou, M. Glick, and K. Bergman, "Fabric-Wide, Penalty-Optimized Path Routing Algorithms for Integrated Optical Switches," presented at the Optical Fiber Communication Conference, 2019, p. Th3A.4.
[30] W. Kabacinski, *Nonblocking Electronic and Photonic Switching Fabrics*. Springer US, 2005.
[31] Y. Huang *et al.*, "Multi-Stage 8×8 Silicon Photonic Switch based on Dual-Microring Switching Elements," *J. Light. Technol.*, pp. 1–1, 2019, DOI: 10.1109/JLT.2019.2945941.
[32] D. Zheng, J. D. Doménech, W. Pan, X. Zou, L. Yan, and D. Pérez, "Low-loss broadband 5 × 5 non-blocking Si3N4 optical switch matrix," *Opt. Lett.*, vol. 44, no. 11, pp. 2629–2632, Jun. 2019.
[33] J. Xing, Z. Li, P. Zhou, X. Xiao, J. Yu, and Y. Yu, "Nonblocking 4x4 silicon electro-optic switch matrix with push-pull drive," *Opt. Lett.*, vol. 38, no. 19, pp. 3926–3929, Oct. 2013.
[34] R. A. Spanke and V. E. Benes, "N-stage planar optical permutation network," *Appl. Opt.*, vol. 26, no. 7, pp. 1226–1229, Apr. 1987.
[35] Q. Cheng, A. Wonfor, R. V. Penty, and I. H. White, "Scalable, Low-Energy Hybrid Photonic Space Switch," *J. Light. Technol.*, vol. 31, no. 18, pp. 3077–3084, Sep. 2013.
[36] T. Shimoe, K. Hajikano, and K. Murakami, "Path-independent insertion loss optical space switch," presented at the Optical Fiber Communication Conference, 1987, p. WB2.
[37] N. Dupuis and B. G. Lee, "Impact of Topology on the Scalability of Mach–Zehnder-Based Multistage Silicon Photonic Switch Networks," *J. Light. Technol.*, vol. 36, no. 3, pp. 763–772, Feb. 2018.
[38] Q. Chen, F. Zhang, R. Ji, L. Zhang, and L. Yang, "Universal method for constructing N-port non-blocking optical router based on 2 × 2 optical switch for photonic networks-on-chip," *Opt. Express*, vol. 22, no. 10, pp. 12614–12627, May 2014.
[39] I. H. White *et al.*, "Control architecture for high capacity multistage photonic switch circuits," *J. Opt. Netw.*, vol. 6, no. 2, pp. 180–188, Feb. 2007.
[40] S. Rumley, M. Bahadori, K. Wen, D. Nikolova, and K. Bergman, "PhoenixSim: Crosslayer Design and Modeling of Silicon Photonic Interconnects," in *Proceedings of the 1st International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems*, New York, NY, USA, 2016, pp. 7:1–7:6.
[41] Q. Cheng, M. Bahadori, Y. Hung, Y. Huang, N. Abrams, and K. Bergman, "Scalable Microring-Based Silicon Clos Switch Fabric With Switch-and-Select Stages," *IEEE J. Sel. Top. Quantum Electron.*, vol. 25, no. 5, pp. 1–11, Sep. 2019.
[42] T. Horikawa, D. Shimura, and T. Mogami, "Low-loss silicon wire waveguides for optical integrated circuits," *MRS Commun.*, vol. 6, no. 1, pp. 9–15, Mar. 2016.
[43] Y. Zhang, A. Hosseini, X. Xu, D. Kwong, and R. T. Chen, "Ultralow-loss silicon waveguide crossing using Bloch modes in index-engineered cascaded multimode-interference couplers," *Opt. Lett.*, vol. 38, no. 18, p. 3608, Sep. 2013.
[44] Zhen Sheng *et al.*, "A Compact and Low-Loss MMI Coupler Fabricated With CMOS Technology," *IEEE Photonics J.*, vol. 4, no. 6, pp. 2272–2277, Dec. 2012.
[45] X. Wang, X. Quan, M. Liu, and X. Cheng, "Silicon-Nitride-Assisted Edge Coupler Interfacing With High Numerical Aperture Fiber," *IEEE Photonics Technol. Lett.*, vol. 31, no. 5, pp. 349–352, Mar. 2019.

Qixiang Cheng (M'17) received his B.S from Huazhong University of Sci. & Tech., China in 2010 and Ph.D. from the University of Cambridge, UK in 2014. He then joined Shannon Lab., Huawei researching future optical computing systems.

Since September 2016, he has been a Post-doc and then a Research Scientist at the Lightwave Research Lab, Columbia University, New York, USA. His current research interests focus on system-wide photonic integrated circuits for optical communication and optical computing applications, including

a range of optical functional circuits such as packet-, circuit-, and wavelength-level optical switch fabrics, massively parallel transceivers, optical neural networks, and optical network-on-chip.

Yishen Huang received his B.Sc. in Engineering Physics from Queen's University and M.S. in Electrical Engineering from Columbia University. He is now a Ph.D. candidate at Columbia University, Lightwave Research Laboratory. His current research interests include large scale silicon photonic switch fabrics and high performance silicon photonic modulators, with specific focus on novel structures and control schemes to enable energy efficient optical link designs.

Hao Yang received the B.Sc. degree in physics from the Peking University, Beijing, China in 2013, and the MS and the PhD in electrical engineering from Columbia University, New York, USA, in 2019. He is currently a post-doc at at the Lightwave Research Lab, Columbia University, New York, USA. His current research interests include silicon photonic devices, and optical switch.

Meisam Bahadori received his B.Sc. degree in electrical engineering, majoring in Communication Systems, with honors from Sharif University of Technology in 2011. After that, he worked toward M.Sc. degree in electrical engineering, majoring in Microwaves and Optics, at the same school and graduated with the highest honors in June 2013.

From fall 2011 to spring 2014, he worked as a research assistant at the Integrated Photonics Laboratory at Sharif University of Technology. He joined the Lightwave Research Laboratory at Columbia University in fall 2014 where he obtained the PhD degree in Electrical Engineering in 2018 with a focus on Silicon Photonics. His current research interests include silicon photonic devices, thin-film Lithium Niobate photonics, and nano-photonics.

Nathan Abrams received both his B.S. in Electrical Engineering from Columbia University in 2014 and his B.A. in Natural Mathematics and Sciences from Whitman College in Walla Walla, WA in 2014. He is currently working towards his M.S. and Ph.D at Columbia University, with research interests relating to photonic devices.

Xiang Meng received his M.S. and Ph.D in Electrical Engineering from Columbia University, New York, NY in 2012 and 2017. He received his B.Sc. in Computer Science and B.Eng. in Electrical Engineering from University of Saskatchewan, Canada in 2011. His research interests include scientific parallel computing and numerical analysis on emerging nanophotonic devices, ranging from nano-lasers, nano-sensors, to high-speed optical transceivers and energy efficient photonic interconnects mainly for applications in high performance computing and data center platform.

Madeleine Glick (M'99–SM'16) received the Ph.D. degree in physics from Columbia University, New York, NY, USA, for research on electro-optic effects of GaAs/AlGaAs quantum wells. After receiving the degree, she joined the Department of Physics, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, where she continued her research in electro-optic effects in GaAs and InP-based materials. From 1992 to 1996, she was a Research Associate with CERN, Geneva, Switzerland, as part of the Lightwave Links for Analogue Signal Transfer Project for the Large Hadron Collider. From 2002 to 2011, she was a Principal Engineer at Intel (Intel Research Cambridge UK, Intel Research Pittsburgh) leading research on optical interconnects for computer systems. Her research interests are in applying photonic devices and interconnects to computing systems.

Dr. Strom Glick is a Fellow of the Institute of Physics and a Senior Member of OSA.

Yang Liu received his B.S. in Physics from Peking University, China, and M.S and Ph.D. degrees in Electrical Engineering from the University of Washington in 2014. He was a senior photonics engineer at Elenion Technologies. His current research interests include the modeling and design of high speed integrated optical communication systems in the SOI platform.

Michael Hochberg obtained a BS in Physics in 2002 and a PhD in Applied Physics in 2006, both from the California Institute of Technology. He has held faculty positions and run research groups at the University of Washington, University of Delaware, and the National University of Singapore, and has held appointments in various departments including Electrical Engineering, Chemical and Biomedical Engineering, and Materials Science. He was a founder at Luxtera, which was recently acquired by Cisco, and at a number of other silicon photonics startups. He was the Director of the OpSIS foundry-access service, which built a community of hundreds of silicon photonic designers around the world; OpSIS was the first organization to offer silicon photonic multi-project wafer runs including a library of passive devices, high-speed modulators and detectors, and an integrated PDK. He is now the CTO of Elenion Technologies in New York City.

Keren Bergman (S'87–M'93–SM'07–F'09) received the B.S. degree from Bucknell University, Lewisburg, PA, in 1988, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1991 and 1994, respectively, all in electrical engineering. Dr. Bergman is currently a Charles Batchelor Professor at Columbia University, New York, NY, where she also directs the Lightwave Research Laboratory. She leads multiple research programs on optical interconnection networks for advanced computing systems, data centers, optical

packet switched routers, and chip multiprocessor nanophotonic networks-on-chip. Dr. Bergman is a Fellow of the IEEE and OSA.