A Novel Simulation Methodology for Silicon Photonic Switching Fabrics

A Novel Simulation Methodology for Silicon Photonic Switching Fabrics

Markos Kynigos*, Javier Navaridas†, Jose Pascual† and Mikel Luján*
*Department of Computer Science, *University of Manchester*, Manchester, M13 9PL, United Kingdom
†Faculty of Computer Engineering, *University of the Basque Country*, San Sebastian, Spain
Corresponding Emails: markos.kynigos@manchester.ac.uk, javier.navaridas@ehu.eus

***Abstract*—Optical communication based on silicon photonics is a promising candidate for future networks. However, a key component that still presents challenges is a practical, silicon photonics-based, high performance switch with a high port count. The impracticality of buffering traffic in the optical domain mandates the use of circuit switching at the transmission level. This renders the photonic power penalty dependent on many factors, including architectural aspects and, most importantly, the switch load. Since the latter changes dynamically with network traffic we argue that simulating silicon photonics-based switches requires considering the photonic power penalty under dynamic workloads, which is not supported by state-of-the-art techniques. In this paper, we show how to simultaneously simulate both the overall switch as well as the photonic power penalty, by proposing a novel combination of the bufferless nature of photonic fabrics, flow-level simulation and optical beam propagation modelling. This approach enables a simulator to consider different kinds of switching fabrics and photonic components. We focus on how to model Beneš photonic switching fabrics formed with Mach-Zehnder Interferometers and consider their deployment as switching cores for top-of-rack switches. We compare our simulation with the published data from two fabricated chips and found accuracy is within $0.5dB$ with respect to insertion loss, and within $3dB$ with respect to crosstalk. As a use-case, we evaluate the impact of routing algorithms on the photonic power penalty and found this can reduce the worst-case photonic power penalty by up to $4dB$.**

***Index Terms*—Simulation, Photonic Switching Fabrics, Performance Analysis**

I. Introduction

Optical communication technology is considered to be a viable candidate for supplanting conventional electronic interconnects, due to its ability to increase transmission speed, provide massive data density per link and maintain signal quality relative to distance. Furthermore, the use of Silicon Photonics has enabled co-integration of microelectronics and photonics due to its amenability to CMOS fabrication processes. These favourable characteristics have led the community to consider deploying optical communication systems closer to the source of computation in high-performance systems [1]–[3].

However, in spite of advances in photonic device design, one key component that has yet to materialise is a practical photonic switch that features low link-level loss, low wiring complexity, high bandwidth density and a large port count. One promising avenue for photonic switches is to use multiple 2×2 switching devices, which are tiled and interconnected to form a Photonic Switching Fabric (PSF). These PSFs can simultaneously achieve energy efficiency and high bandwidth density through Dense-Wavelength-Division Multiplexing (DWDM), in which communication traffic is encoded onto multiple wavelengths, or λs. However, employing switches with DWDM mandates using broadband switch devices, which actuate uniformly across a contiguous spectral segment. Mach-Zehnder Interferometers (MZIs) are often considered since they are inherently broadband, have low wiring complexity and can switch at GHz rates by using the Free-Carrier Dispersion effect (FCD). However FCD generates photonic crosstalk, which cascades through the PSF. At the PSF output ports, photonic crosstalk degrades the signal quality in the form of interference. To some extent, this can be compensated with a higher laser power. This power increase is called the crosstalk power penalty. However, there is a limit to the amount of crosstalk a signal can support beyond which data corruption is not recoverable.

In PSFs, crosstalk is affected by three factors: device design, PSF topology and PSF use (see Section II). By increasing device Extinction Ratio (ER), the crosstalk of the PSF can be decreased. The topology of the PSF, also impacts crosstalk; non-dilated rearrangeably-non-blocking topologies such as the Beneš network are susceptible to first-order crosstalk, while others such as the Dilated Beneš are not, but require double the switching devices for the same scale of PSF I/O. Finally, the use of the PSF, defined by the PSF routing algorithm and the serviced traffic workload, also determines the crosstalk profile. This aspect affects topologies which offer path diversity, where different paths expose the photonic carrier to a variable number of devices and at different states. Routing algorithms can be designed for these to optimise PSF use with respect to crosstalk. However, designing these requires analysing the routing algorithm behaviour under realistic traffic, since crosstalk is dependent on the switch saturation and the path characteristics imposed on the network traffic. This analysis is impractical when using traditional PSF modelling techniques, since they focus either on the photonic layer (i.e. device level), or on the switch control plane architecture [4], [5].

Contrary to electronic switching fabrics, PSFs are bufferless internally. They must therefore enforce circuit switching at the transmission level. As a consequence, the performance impact of both traffic dynamics and design choices in the switch

architecture and control plane are amplified for PSFs.

We observe, however, that the timing variability inherent to electronic switching due to buffering, does not exist in bufferless PSFs. Based on this aspect, we propose that flow-level network simulation, normally used for generalised interconnection networks, can be appropriate for simulating photonic switching fabrics. Extending a flow-level simulator with an optical beam propagation model allows to co-simulate the physical and control layers of a PSF under dynamic traffic and to make informed optimisation decisions for PSFs.

This paper therefore addresses the above by presenting a novel PSF modelling technique for capturing both the device and control planes (Section III). We target PSFs formed with MZIs in the Beneš topology, a popular PSF configuration that requires the fewest MZIs to form a rearrangeably-nonblocking network. Although we focus our evaluation on MZIs and the Beneš topology, our proposal is applicable to other PSF topologies and architectures. Our technique follows a traffic-driven simulation of photonic PSFs, which is based on flow-level interconnection network simulation, and it is augmented with a photonic beam propagation model. We establish the photonic model accuracy by comparing it to two state-of-the-art chips (Section IV). To demonstrate our technique, we evaluate the impact of routing algorithm selection on the photonic power penalty under a range of realistic workloads in Section V. Finally, in Section VI, we discuss the key differentiating factors between our technique and the state-of-the-art.

II. Background

A. Motivation for Photonics-based ToR Switches

Modern DC and HPC deployments currently rely on electronic packet switches (Infiniband or Ethernet), with optical communication being relegated to inter-switch transmission. There exists a large variety of commercial DC switches, featuring various radices, switching capacities and form factors; but they tend to be extremely power hungry. To illustrate this and to estimate the impact on energy consumption, Table I compares a number of commercially available ToR switches. We include the radix, maximum per-port data rate, maximum capacity at that data rate and the estimated peak power dissipation. Based on the peak power dissipation and switching capacity, we estimate the switching energy per bit. In this way, we can illustrate the impact of the switching technology on power consumption, isolated from the link transmission technology. We consider peak power dissipation without optics; where this is not reported, we subtract $radix * optics\_wattage$ from the reported peak power, assuming 20W optics for 400Gb/s, 4.5W for 100Gb/s and 2.5W for 40Gb/s links.

Based on these estimates, switching energy efficiency in commodity electronic switches ranges between 42 and 330 pJ/bit, depending on the device. The most energy efficient and highest bandwidth switch is the MQM9700 by NVIDIA-Mellanox, with 42.4 pJ/bit with 400Gb/s links which, however, comes with a power envelope of approx. 1KW. With hundreds of switches being employed in modern large-scale DCs, the total power footprint of the network increases dramatically.

TABLE I
Power and energy consumption of switches (excl. optics).

Device Model	Switch Radix	Data Rate (Gb/s)	Switching Cap. (Tb/s)	Power Diss. (W)	Energy (pJ/bit)
CISCO Nexus 3636C-R	36	100	3.6	1,179	327.5
Aruba CX 8320	32	40	1.3	230	179.7
Aruba CX 8325	32	100	3.2	406	126.9
CISCO Nexus 3464C	64	100	6.4	712	111.3
Huawei CloudEngine 9860	128	100	12.8	1,051	82.1
Arista 7368X4 Series	32	400	12.8	966	75.5
NVIDIA MQM9700	64	400	25.6	1,084	42.3
NVIDIA SN2700	32	100	3.2	135	42.2
MZI PSF Switch	16	512	8.2	21.2	2.6
	32	512	16.4	23	1.4
	64	512	32.8	27.3	0.8
	64	400	25.6	27.3	1.1

In contrast, we estimate the switching energy for ToR switches that employ MZI PSFs by extrapolating from the MZI-based 16× 16 switching fabric characterised in [6]. Such PSF would exhibit a very small switching power envelope (between 1.2W and 7.3W for 16 to 64 endpoints). To this we would need to add a network controller, which can be implemented in a Virtex-7 FPGA. Considering the power budgets reported for such devices in [7] we take a pessimistic power envelope of 20W. We assume a deployment scenario with different switch radices (512Gb/s links with 32 wavelengths), as well as a comparative scenario assuming 64 ports and 400Gb/s links similar to the MQM9700. Switches with these characteristics will feature energy per bit figures of between 0.8 and 2.6 pJ/bit. Clearly, the peak switching power and switching energy per bit can be potentially reduced by 1-2 orders of magnitude moving from electronics to photonics. This can be highly compelling for photonic ICNs, as their adoption can potentially reduce the total cost of ownership or increase the power budget for other components such as CPUs, I/O, etc.

B. Photonic Switching Fabric Technology

PSF operation differs substantially to that of electronic switches. We therefore detail the fundamental governing principles of PSF devices here, focusing on our target technology.

PSFs are constructed using multiple 2×2 switching devices. These are interconnected using waveguides and, where necessary, waveguide crossings (*wgx*s). We focus on switching devices formed using interference-based Mach-Zehnder Interferometers (MZIs). MZIs consist of two waveguide arms, connected on either side by 2×2 3dB couplers or Multi-Mode Interferometers (MMIs), serving as input and output couplers. MZI arms are equipped with thermal or electrical tuners, which enforce a phase difference on light traversing one arm with respect to the other. This tuning sets the MZI state to either ”cross” or ”bar”; in the ”bar” state, light entering the MZI egresses at the mirrored output port, whereas in the “cross” state light egresses at the complement port. Light entering the MZI from either input port is split into both MZI arms by the input coupler; based on the tuning-induced phase difference, the light either constructively or destructively interferes in the output coupler, to egress the MZI at the desired port. Tuning principle affects the switching speed, insertion loss (ILoss), ER,

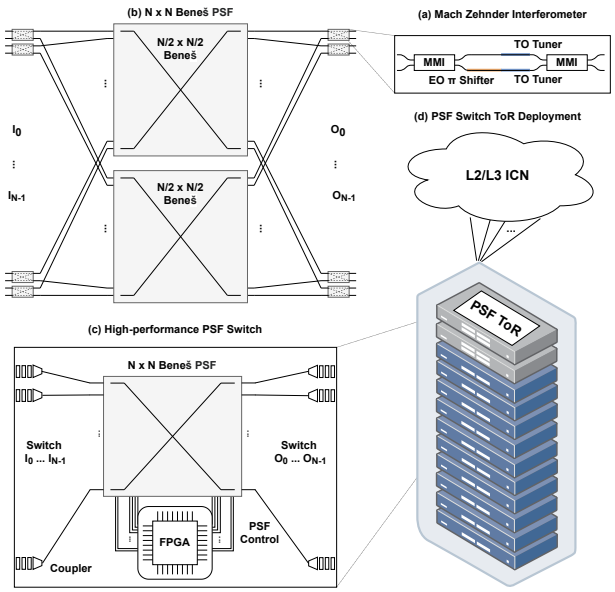

Fig. 1. *(a)* Schematic of a 2×2 EO/TO MZI switching element. *(b)* An $N \times N$ MZI Beneš PSF. *(c)* A high-performance switch containing the FPGA-controlled PSF. *(d)* Deployed ToR switch within a DC or HPC rack.

broadband nature and footprint of the device. These, together with the topology, affect the capabilities and achievable size of the PSF. An example TO/EO-tuned MZI is depicted in Fig. 1a.

MZI structure — Most MZIs entail an equal arm length and rely on tuning for calibration. Nested MZI switches have also been proposed to reduce the crosstalk in PSFs [8]. These have an increased footprint, higher complexity and a smaller tuning spectral region, but a higher ER leading to less crosstalk. Using these, DWDM can be achieved by adopting a smaller channel spacing; e.g., 50GHz instead of the ITU standard 100GHz. Tri-state MZIs have also been proposed [9]. These include a third state which decreases the overall crosstalk power penalty of PSFs, and can be used for practical larger PSF sizes.

Tuning principle — The thermo-optic (TO) or the electro-optic (EO) effect are used for MZI tuning. In the former, a heating element changes the refractive index of the material to induce phase change; this provides low ILoss and a high ER to the MZI, reducing the attenuation and leakages which lead to photonic crosstalk. TO tuning happens at the μs scale, which is too slow for many applications in high-performance switching (e.g., TDM). EO tuning takes a few ns and is therefore suitable for TDM, but leads to free-carrier absorption (FCA) [10]. This increases ILoss and reduces the ER, generating leakage power at the output ports in the form of crosstalk. As stated by Lee *et al*. [11], ILoss can be mitigated through amplification, but crosstalk can not. Crosstalk is most detrimental when two interfering light-beams are coherent. The power penalty from coherent crosstalk can limit PSF size.

Tuning application — Tuning can be induced on either one or both MZI arms; the former is referred to as single-ended tuning, the latter as push-pull [12]. In single-ended tuning, one tuner must provide the entire π phase shift relative to the light traversing the other MZI arm. In TO tuning this increases the heating element size and, in EO-tuning, this increases FCA, decreasing ER. In push-pull tuning, both arms provide a $\pi/2$ phase shift, which mitigates EO-tuning crosstalk penalties. If only TO-tuning is used, it both calibrates the MZI to either state and switches to the complement state. If only EO-tuning is used, the device is initially in the quadrature state, and tuning induces either MZI state. If both tuning options are used, TO-tuning is used for calibration, and EO-tuning is used to switch.

PSF Topology — The connection pattern within in a PSF is defined by the topology, which governs how many devices are required to connect N inputs to N outputs. This, in turn, defines the photonic loss that each carrier beam incurs when traversing the PSF, and the photonic crosstalk that it is exposed to from other carrier beams. It also defines the PSF path diversity, i.e. how many potential paths exist from a source to a destination. The topology also dictates the blocking characteristics of the fabric, whether it is blocking (BNG), rearrangeably non-blocking (RNB) or strictly non-blocking (SNB). Various topologies have been adopted from the electronics domain, or proposed specifically for their application to PSFs, as reviewed in [13]. Of these, the Beneš network has been frequently used, as it requires the fewest 2×2 devices to connect $N \times N$ endpoints in a RNB fashion. This lowers the electronic backend complexity and ILoss compared to other topologies. Both aspects are favourable for practical PSF-enabled network switches, which is why we focus our analysis on the Beneš network. However, the Beneš network is prone to first-order crosstalk; each victim carrier suffers leakage from aggressors at every stage. We depict an FPGA-controlled PSF formed with MZIs deployed as a ToR switching core in Fig. 1.

The lack of buffering is a challenge for controlling Beneš - based PSFs. Traffic must be blocked at the PSF input port while the switch state is rearranged or be mis-routed or lost. An interesting property of PSFs is that the total ILoss and photonic power penalty varies with the employed routing algorithm within the switching fabric. This occurs due to *wgx*s and if there is an imbalance in ILoss or ER between the MZI states. Thus the literature has captured different routing strategies which aim to allocate paths that incur the least amount of ILoss from *wgx*s, and/or MRR/MZI traversal [14]–[17].

III. Modelling Photonic Switching Fabrics

A. Flow-level Simulation for Photonic Switching Fabrics

To co-evaluate photonic metrics and switch architecture, we extend INRFlow, an open source, flow-level network simulator [18]. This relies on a simple but potent observation. Electronic network switch cores include internal ingress and egress buffers (virtual output queues) between input and output ports, as well as intermediate buffers. Packets or flits entering the switch are buffered before and after arbitration. This buffering takes time which varies with external factors, such as switch load and port contention, and thus, it impacts the switch latency.

Conversely, *photonic* switches are *bufferless* internally. Once traffic has entered the PSF through an input port, it must stream uninterrupted through the photonic hardware and reach the destination output port. If no path is available, traffic

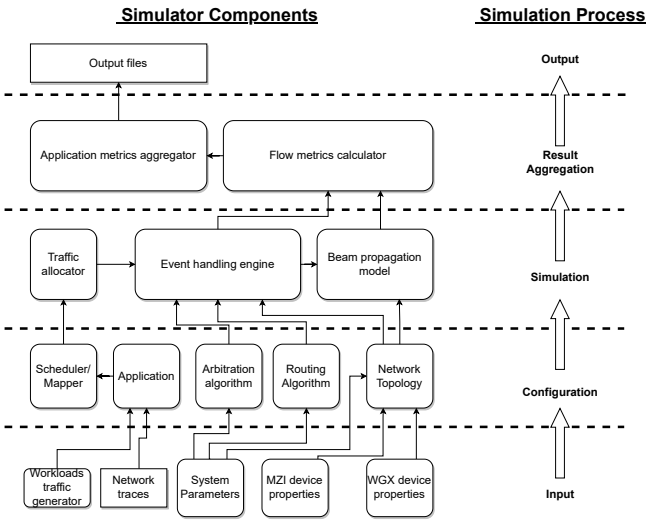

Fig. 2. Augmented simulator components & simulation process.

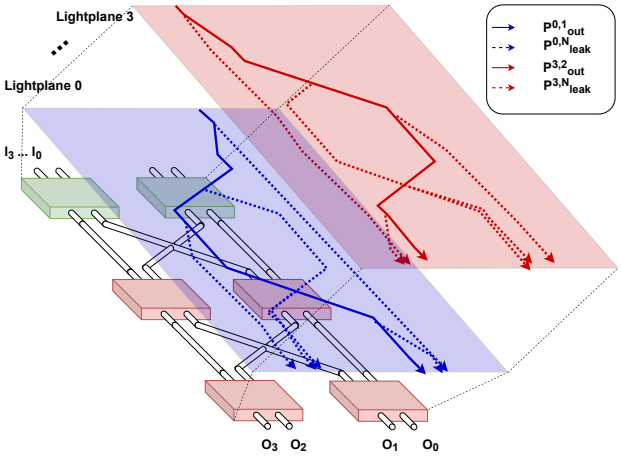

Fig. 3. The signal and leakage powers generated by a lightpath are partitioned into logical lightplanes (colour online). Lightpaths (contiguous arrows) are established from $I_0 \rightarrow O_1$ (blue) and $I_2 \rightarrow O_3$ (red). 1^{st}-order leakages shown as dotted arrows. "Bar"-state MZIs shown in red, "cross"-state in green. Higher order leakages omitted for illustration.

must queue in the electronic backend, until a path becomes available; otherwise, signals will interfere with each other and data will be corrupted. The timing and latency variability caused by internal buffering is therefore negated. We note that capturing timing variability due to the physical characteristics of the medium (lightpath length, geometry, *etc.*) [19] requires modelling the PSF geometry; this is usually conducted with photonic design automation tools (e.g. Lumerical suite [20]). This paper assumes these are accounted for by separating transmission phases with guard time slots [21], [22]. Under this assumption, the flow-level abstraction, which does not account for buffering time, is suitable for *switch-level* modelling of data streams encoded in light traversing photonic hardware.

By using the simple node design, unidirectional traffic modelling and event-driven transmission time computation afforded by INRFlow, arbitrary-sized PSFs formed with 2×2 switches can be investigated using a wide variety of communication workloads. Note that while we focus here on the Beneš topology, other topologies based on 2×2 switches can be modelled by adding a new topology file to the simulator.

B. Modelling Beam Propagation

We extend INRFlow by augmenting its pre-existing components and supplying some new ones. Fig. 2 shows the components of the simulator and the simulation process. We constrain the discussion to the beam propagation model and how it interacts with the central data structure, i.e. the *node*. Nodes model both traffic producer/consumer endpoints and MZIs. For MZIs, the *node* is extended to include the ILoss and crosstalk ratio profiles in dB for each device state, as well as the required tuning power in mW. These are inputted using property files at runtime and consist of the ILoss and ER for each discrete λ. *Optical ports* are added to the *node*; these contain an array of *lightplane* instances used for the beam propagation model, and an array of *wgx* instances. The *wgx* data structure contains four ports, each of which also contains an array of *lightplanes* and the connected neighbour identifier of *wgx*. This can be either another *wgx*, or a switching device.

The *lightplane* is a logical plane used to model the photonic signal power and leakages of a *lightpath*. A *lightpath* is established when a source node communicates with a destination node. With this information, leakage propagation from a *lightpath* can be modelled across the entire PSF. This is depicted in Fig. 3, which shows a partial permutation, with lightpaths established from $I_0 \rightarrow O_1$ and $I_2 \rightarrow O_3$. Here, each signal is depicted as a contiguous arrow and the 1^{st}-order leakages as dotted arrows. Higher order leakages, although captured by our model, are not shown for simplicity.

Lightplanes contain an array of *optical channel* instances, which in turn contains the photonic power and leakages of carrier beams at a particular λ. *Optical channels*, modelled within *lightplanes*, which are included in the *optical ports* of *nodes* are swept by the beam propagation model, to model the effects of *lightpaths* traversing the PSF.

Photonic carrier beam propagation is modelled after the routing phase and before the next event handling phase. Each flow is carried by an individual photonic beam carrier. The carrier beams enter the PSF input ports i.e. the switching devices at the first PSF stage, with an input power of $0dBm$. Depending on the input port and switching device state, they are traced to the corresponding switching device output port. There, the state ILoss penalty is enforced upon the power level of each beam. State-dependent leakage power is added to the complement output port on the carrier *lightplane* of each beam.

The model then propagates the power and leakages from the current stage switch devices to the adjacent *wgx*. In each *wgx*, the photonic beam is propagated from the input port of *wgx* input *lightplane* to the *lightplane* of the output port, based on the propagation direction of the beam. Leakages from the beam are added to the corresponding *lightplane* of the ports perpendicular to the beam's direction of propagation (excl. reflection). The beams and leakages are propagated to the

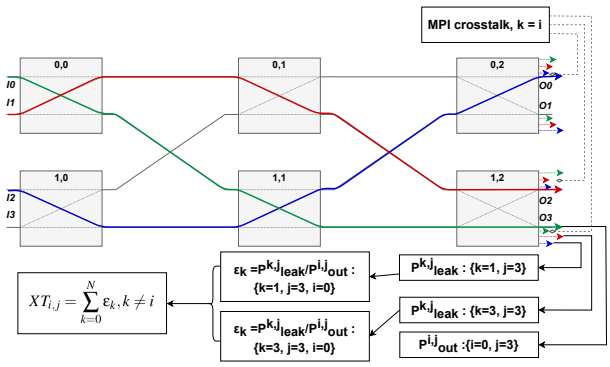

Fig. 4. Process for calculating the crosstalk ratio of aggregate leakages, depicted in a 4×4 Beneš (colour online).

adjacent *wgx* in the propagation direction. This process iterates, until the next stage is encountered. *wgx* between two PSF stages are processed *column-wise*; for each switch in the stage, the first *wgx* of each port is processed, then the second *wgx* etc. This preserves leakage calculation sequence and maintains multi-order leakages (crosstalk arising from other crosstalk).

Once all *wgx* are processed, the beams and leakages from each *lightplane* of the *wgx* output ports are propagated to the switching device input port of the next PSF stage. The next PSF stage becomes the current, and the process iterates until the output ports of the final PSF stage are reached. At the end of the photonic beam propagation process, all flows that are traversing the PSF on a lightpath have populated a *lightplane* trace of power and leakages at each PSF output port.

Once a victim flow reaches the destination node, the simulator uses the populated *lightplanes* to derive the ILoss suffered from traversing the switch, as well as the crosstalk suffered from other flows concurrently traversing the PSF, called *aggressors*. Based on these, the power penalty is derived. The following equations are used, adopted from the work of Ramaswami *et al*. [23] and Cheng *et al*. [24], and adapted to our simulator. In these equations, i, j refer to the input and output endpoint of the victim signal, k refers to the input port of the aggressor signal, while N is the number of PSF endpoints.

The crosstalk ratio of a leakage signal from aggressor k at port j, with respect to the victim signal power level $\{i, j\}$, is

$$\epsilon_k = P^{k,j}_{leak}/P^{i,j}_{out}. \tag{1}$$

Through Eq. 1, we can express the crosstalk in dB from an individual leakage arising from aggressor k as

$$L_{i,j,k} = 10log_{10}(P^{k,j}_{leak}/P^{i,j}_{out}). \tag{2}$$

In the Beneš topology, when multiple photonic flow carriers occupy the PSF, each flow causes leakages which cascade and cause higher-order leakages, which aggregate at every output port. Therefore, *every* flow acts as an aggressor to every other flow simultaneously present in the PSF. The crosstalk ratio of the aggregate leakages to the victim flow $\{i, j\}$ is

$$XT_{i,j} = \sum_{k=0}^{N} \epsilon_k, k \neq i. \tag{3}$$

The process expressed in Eq. 3 is visualised in Fig. 4, which shows photonic carrier beams transmitting (coloured arrow lines) during a partial permutation in a 4×4 PSF. The dashed coloured arrows at the PSF outputs symbolise accrued leakages, with multi-path interference (MPI) depicted as well. At each PSF output, non-MPI leakages are summed to form the aggregate crosstalk. Note that, unlike other simulation techniques (e.g. [11], [25]–[27]), this enables our simulator to capture multi-order crosstalk levels for *partial permutations*.

In the DWDM scenario, we assume every flow is being carried by the same λ group. As mentioned by Lee and Dupuis [11], the severity of crosstalk varies with relative phase, polarisation and λ between aggressors and victims. In the worst case, crosstalk is termed to be *coherent*, meaning co-polarised, exactly out of phase and approximately at the same λ. In this case and when $XT_{i,j} \ll 1$, the power penalty from crosstalk imposed on the victim signal is expressed as

$$PP^{i,j}_{XT} = 10\log_{10}(1 - 2\sqrt{XT_{i,j}}). \tag{4}$$

Note that Eq. 4 forms a critical threshold for $XT_{i,j}$. Once it is surpassed, no amount of input laser power can compensate for the effect of aggregate crosstalk . In this case, $PP^{i,j}_{XT} \to \infty$.

We also include inter-channel crosstalk, which occurs when the aggressor and victim signal are carried at different λs, and the λ difference is sufficiently large compared to the receiver bandwidth [23]. When DWDM λ is filtered by an MRR resonator at the output before the receiver, the number of aggressor signals for inter-channel crosstalk for every victim λ_l depends on the ER of the MRR filter and the channel spacing of the DWDM λs. With 100GHz channel spacing, λ_l receives inter-channel crosstalk only from λ_{l-1} and λ_{l+1}. Thus, the power penalty from inter-channel crosstalk is

$$PP^{i,j}_{XT-inter} = 10\log_{10}((1 - \sqrt{XT_{i,j,\lambda_{l-1}}})(1 - \sqrt{XT_{i,j,\lambda_{l+1}}})). \tag{5}$$

Ramaswami *et al*. showed that inter-channel crosstalk is much less detrimental than coherent crosstalk, and is frequently discounted in most studies. In contrast, we are able to include it in our model, since the per-λ power penalty is isolated within *channels* and *lightplanes*. Thus, the total power penalty on the victim signal in the simulations is

$$PP_{i,j} = IL_{i,j} + PP^{i,j}_{XT} + PP^{i,j}_{XT-inter}, \tag{6}$$

TABLE II
ILOSS AND CROSSTALK OF PSF COMPONENTS.

	EOMZI (1560 ± 15 **nm**)	**TOMZI** (1560 ± 5 **nm**)	**TOMZI** (1560 **nm**)
IL_{prop} (per stage)	$\sim$ 0.44 dB	$\sim$ 0.35 dB	$\sim$ 0.35 dB
IL_{wgx}	0.05 dB	0.05 dB	—
XT_{wgx}	-30 dB	-30 dB	—
$IL_{MZI,cross}$	0.4 dB	0.32 dB	0.32 dB
$IL_{MZI,bar}$	1.4 dB	0.32 dB	0.32 dB
$XT_{MZI,cross}$	-30 dB	-30 dB	-35 dB
$XT_{MZI,bar}$	-18 dB	-30 dB	-35 dB

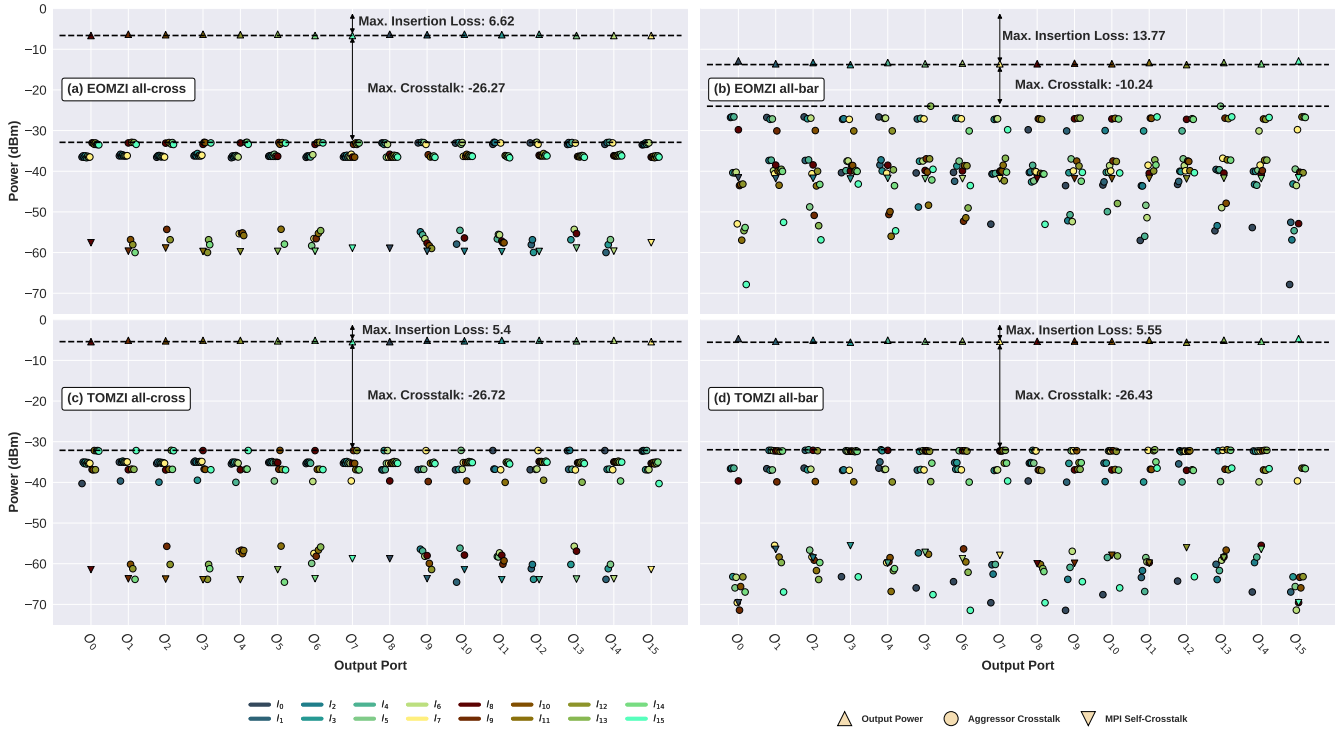

Fig. 5. Power levels a the PSF outputs of the EOMZI and TOMZI PSFs.

where the ILoss is $IL_{i,j} = 10\log_{10}(P_{out}^{i,j}/P_{in}^{i,j})$.

IV. Simulation Accuracy

A. PSF Chip Validation Targets

We now establish the simulation accuracy of the beam propagation model. We assume the MZI, waveguide and waveguide crossing specifications of two 16×16 switching fabrics from the literature, namely [6] and [28]. We call these the EOMZI and TOMZI PSF respectively, as the former switches electrically, while the latter switches thermally. In both works, which operate in the C-band centred around $1560nm$, the authors specify the performance in ILoss and crosstalk of MZIs and waveguide crossings, as well as the waveguide propagation loss. Lu*et al.* report the ILoss and crosstalk for their MZIs and waveguide crossings as a worst-case over a $30nm$ wavelength region centred on the central wavelength. Zhao*et al.*, on the other hand, detail the performance of their MZIs and waveguide crossings both on the central wavelength and for a $10nm$ region centred around $1560nm$. The reported values across the wavelength regions are collected in Table II.

They also both depict the transmission spectra of the 16×16 PSFs in the "all-cross" and "all-bar" states, and report on the maximum ILoss and crosstalk for each PSF at each state.

We model both PSFs in the simulator in the two states and excite all input ports. As the beam propagation is separated into lightplanes based on the input port, we collect all signal and leakage values at the output ports for each state, and depict them in fig. 5. It is noted here that for the EOMZI PSF, the authors found that one of the MZIs (sixth MZI in stage 3) shows a decreased extinction ratio. We therefore correct for this in our simulation by explicitly raising the crosstalk ratio of that MZI for the bar state in the simulation setup.

B. Validating the Simulator against 16×16 PSF Chips

We first use the EOMZI specifications [6] in the "all-cross" state, the ILoss is at most $6.7\pm1dB$, with the crosstalk being at most $-30dB$ for the $30nm$ λ segment. In comparison, when modelling the "all-cross" state with all output ports excited, we find the maximum ILoss to be $6.62dB$, and the max. crosstalk to be approx. $-27dB$, which is slightly higher than the measured data. In the "all-bar" state, they report an ILoss of at most $\sim 14dB$, whereas the crosstalk is at most $-10dB$. In our model the ILoss is at most $13.77dB$, while the crosstalk is at most $-10.4dB$, $0.4dB$ lower than the measured data. Remember that Lu *et al.* reported the worst-case performance of their devices for the whole λ region, rather than for a specific λ as calculated by our simulation. As with the TOMZI PSF, ILoss and crosstalk ratio varies with λ, with the lowest value being at the central λ. To emulate this, we conduct a parameter sweep on the crosstalk of the MZIs and the *wgx*s, over a $10dB$ region at $0.25dB$ increments for both states, depicted in Fig. 6.

In the parameter sweep (especially in Fig. 6–b), we observe the tightly coupled relationship between PSF maximum crosstalk, MZI crosstalk and wgx crosstalk. The highest device crosstalk value limits the reductions in PSF crosstalk achievable by optimising the other device type; reducing MZI-induced crosstalk only reduces PSF crosstalk if the wgx crosstalk is also reduced, and vice versa. Also, the measured PSF crosstalk value for the cross state (i.e. $-30dB$) is achieved when "cross-state" MZIs and wgx have a crosstalk ratio of approx. $-33.5dB$. As

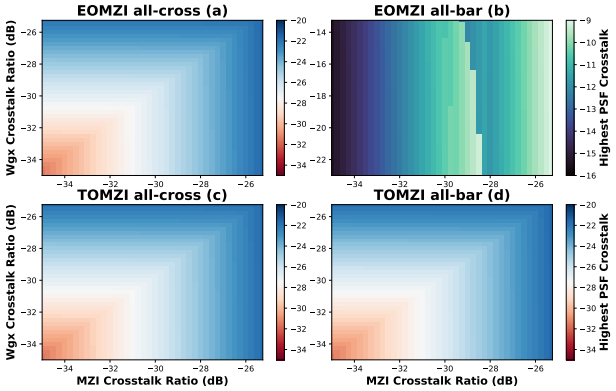

Fig. 6. Impact of MZI and wgx crosstalk ratios on worst-case PSF crosstalk.

crosstalk ratios of the devices are lower at their central λ, we infer that the crosstalk ratio of "cross-state" MZIs and wgx from Lu *et al*. is approximately $-33.5dB$.

For the TOMZI PSF, Zhao *et al*. report that their PSF ILoss and crosstalk is at most $\sim 5.2 \pm 1dB$ and $-30dB$ respectively for both states. We model the TOMZI PSF, report the powers and leakages in Fig. 5, and also perform a parameter sweep. Our model shows the ILoss for the "all-cross" state to be $5.4dB$, which is within $0.5dB$ of the measured data. The discrepancy is attributed to path-length variation. The modelled crosstalk is $\sim -27dB$, which is slightly increased compared to the measured data. As before, the measured wgx crosstalk values at the central λ may be slightly lower; this is reflected in the parameter sweep, which indicates that the measured crosstalk value for the PSF is achieved at the central λ with a wgx crosstalk ratio of $-33.5dB$. In the "all-bar" state, our model exhibits $5.5dB$ ILoss, which is very close to the measured data. The modelled PSF crosstalk is again slightly higher ($\sim 26dB$) with the nominal crosstalk ratio for the wgx, but reaches the measured value when the wgx crosstalk ratio is reduced.

V. Traffic-driven Photonic ToR Simulation

Aside from establishing the accuracy of the beam propagation model, the previous section examined the performance of two state-of-the-art PSFs in two PSF states, namely "all-cross" and "all-bar". These two states however, represent corner cases of the PSF functionality. In fact, a $N \times N$ Beneš PSF can be configured in $2^{Nlog_2(N)-\frac{N}{2}}$ states, with multiple potential candidate states being able to service partial or full permutations; for full permutations, $N!$ states are required [29]. This, along with the fact that the power penalty imposed on a connection is determined by the PSF state, means that the power penalty of connections can be optimised by selecting one PSF state over the other. Additionally, a PSF state that services a full permutation might not be optimal for a corresponding partial permutation. Partial permutations occur when the workload traffic driving the PSF endpoints causes output contention, includes causal relationships or when switch fabric contention is present. These factors must be accounted for when investigating power penalty optimisations in Beneš PSFs.

We present in this section an analysis of the performance of 6 routing algorithms with respect to the power penalty. We model PSFs ranging from 4×4 to 32×32 endpoints to examine how the variation of crosstalk and power penalty scales with the network size. We consider PSFs based on MZIs from the EOMZI PSF and assume circuit switching; we specifically disregard the faulty MZI and assume an ideal scenario where all MZIs operate identically. We detail our modelling setup below.

A. Routing, Switching & Arbitration

We first describe the configuration of the control plane which we consider for the ToR switch, namely the employed routing algorithms, switching context and arbitration.

The standard routing algorithm for controlling Beneš networks is the "Looping Algorithm" (***LA***) [29]. The *LA* configures the network to serve full permutations by exploiting the topological symmetry, and can be adapted for partial permutations. As it was designed for electronics-based Beneš networks, we will see that due to the unique constraints of photonics (lack of intermediate buffering, variable path power penalty) it can be out-performed in terms of exhibited photonic power penalty.

Another set of algorithms which has been recently proposed for Beneš PSFs is the "hardware-inspired routing" set (HIRs) [16]. The HIRs operate using bufferless switching and compute potential paths from a PSF input to a PSF output regardless of the contending paths. They then rank the potential paths based on traversed MZI states and traversed number of waveguide crossings, in order to assign paths with reduced ILoss to flows. We compare the following HIRs against *LA*:

- ***m_b***: ranking is based on the number of MZIs in the "bar" state per path.
- ***m_x***: ranking is based on the number of waveguide crossings per path.
- ***m_xb***: ranking is based on the number of waveguide crossings and ties are broken by the number of MZIs in "bar" state.
- ***m_bx***: ranking is based on the number of MZIs in "bar" state and ties are broken by the number of waveguide crossings.
- ***rnd***: selects a path randomly, without taking underlying hardware asymmetries into account.

Two popular switching variants have been examined for Beneš PSFs, namely circuit switching (CS) and time-division multiplexing (TDM). While our simulator can evaluate both scenarios, we assume CS here, as TDM places timing constraints on PSF state computation. As *LA* solves permutations in $NlogN$ time, this would be unrealistic for TDM.

B. Employed Workloads

We investigate the behaviour of the PSF under a wide variety of synthetic and pseudo-realistic workloads:

RandomApp (RA) — Selects the source and destination uniformly at random. This is a typical networking benchmark used to stress the IC. According to [30], the traffic mix run on a typical DC is unstructured and essentially random in nature.

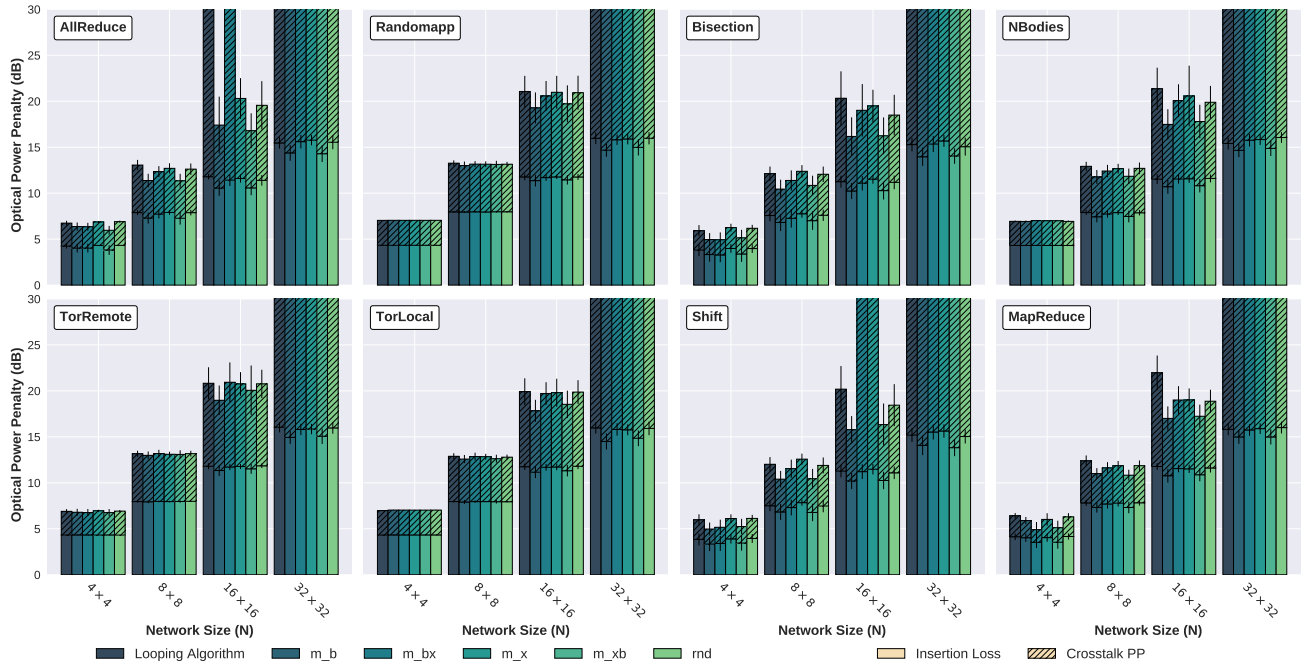

Fig. 7. The evolution of the total photonic power penalty with PSF size under 8 workloads and 6 PSF routing algorithms. ILoss is shown in the bottom of each bar, crosstalk power penalty on the top (hatched).

Bisection (BI) — Tasks perform pair-wise communications swapping pairs randomly every round. This benchmark was introduced in [31] as a way to estimate bisection bandwidth.

Shift (SH) — In this workload, tasks send messages to destinations at a given *stride*, t. The destination, D, is calculated as a function of the source, S: $D = (S + t) \mod N$. This is akin to the adversarial traffic proposed in [32].

AllReduce (AR) — An optimised, binary implementation of the AllReduce collective [33], widely used in parallel applications from a range of domains. This workload sends a total of $N \cdot logN$ flows.

NBodies (NB) — A typical scientific pattern, where a collection of bodies (e.g., planets, subatomic particles, etc.) interact with each other to model the evolution of physical phenomena. Tasks are arranged in a virtual ring and each task starts a chain of messages that travel clockwise across half of the ring [34]. This results in a total of $N^2/2$ flows.

MapReduce (MR) — This is a representative application from the data center domain. First the master server scatters data to the slave tasks, these communicate among themselves using an all-to-all traffic pattern and finish with a gather phase to send the results back to the master server.

TorLocal (TL) — This workload models the traffic handled by a ToR switch within a DC. It is based on the analysis of the actual traffic captured in 10 DCs from different domains [35]. TL considers that most traffic is local, while 20% of the traffic is extra-rack, as reported for the CLD5 system.

TorRemote (TR) — This workload is similar to TorLocal, but uses the configuration with the highest proportion of remote traffic. In TR, 90% of the traffic is extra-rack, as observed in the EDU1 system of [35].

Each workload is executed 100 times for each PSF size using a different random seed per run. Unless otherwise specified, the workloads send 1000 flows from the sources to the destinations. It is noted that the workloads include causality between the flows and therefore undergo phases of high and low network pressure. Additionally, as is standard practice for DCs and clusters, we assume that the system scheduler models the system as a flat network with no locality information. his results in tasks being distributed randomly across the network [36], [37]. Finally, to conform with common practice in DCs [38], we assume an oversubscription of 3:1 at the ToR level. As an example, a 16×16 switch will have 12 ports connected to servers and 4 uplinks connected to higher levels of the ICN.

C. Evolution of Photonic Power Penalty with PSF Size

We examine the worst-case performance of the PSF under different traffic scenarios in Fig. 7. We show that irrespective of the workload, the crosstalk power penalty increases dramatically with PSF size. For 32×32 PSFs, the aggregate crosstalk surpasses the critical threshold; the signal quality is degraded to a point where no input laser power increase can compensate for the presence of crosstalk. On one hand, enlarging the PSF adds more MZI and wgx devices to a path, which increases the ILoss and the occurrences of crosstalk leakage. On the other hand, due to the Beneš topology, first-order crosstalk is generated at every stage, and propagates through the network causing higher-order crosstalk and accumulating at the output ports. The larger the network, the more crosstalk generation instances. As the Beneš topology is not designed to mitigate crosstalk, the crosstalk power penalty limits the achievable size. Our simulations confirm the results of Lee *et al.* [11].

In terms of the routing algorithms, it is interesting to note that, in terms of photonic penalties, *LA* is always out-performed by the most effective HIRs, i.e. *m_b* and *m_xb*, especially as the PSF size increases. Increased PSF size entails increased path diversity, allowing the HIRs to provide more options for route provision through the PSF. A decreased power penalty from crosstalk also correlates with decreased ILoss in these cases; this is expected, as the design of the assumed MZIs entails a higher ILoss and crosstalk ratio in the "bar" state due to free carrier absorption. It is also interesting to note that HIRs that prioritise paths with the fewest wgx devices, i.e. *m_x* and *m_bx* exhibit a higher photonic power penalty than *m_b* and *m_xb*; compared to *LA* the worst-case power penalty for these is within one standard deviation, and they are therefore ineffective at reducing the metric. The *rnd* HIR performs better than the *LA* in some cases (e.g. Allreduce, Mapreduce and Shift for 16 endpoints) but has comparable performance in terms of worst-case photonic power penalty for all other cases, with the metric being within one standard deviation of *LA*.

Finally, with respect to workloads, we observe two distinct groups of workload behaviour. Randomapp, TorRemote and TorLocal show highly similar power penalty profiles as the PSF size increases, with only slight differences (1-2 dB) within one standard deviation of each other. As discussed in [39], these workloads do not include high levels of causality between the flows and suffer from switch fabric contention and output contention, leading to less-pronounced impacts from the routing algorithm's path selection on the photonic power penalty imposed on the flows. Even so, the *m_b* and *m_xb* HIRs reduce the power penalty by $\sim$2 dB.

The remaining workloads show a more distinct impact of routing algorithm selection on the photonic power penalty. Bi-section, which exhibits full saturation with *LA*, is most severely impacted; the total power penalty can be reduced by up to ~ 4 dB by the best-performing HIR. AllReduce, MapReduce, Shift and NBodies all include a high level of causality between the flows; i.e, flows must wait for previous flows to complete before being transmitted. For this reason, switch saturation is lower than workloads with less causality. The case of AllReduce with a 16×16 PSF is particularly interesting: in some runs under the *LA* and the *m_bx* strategy, there exists at least one PSF state requested by the controller which induces too much crosstalk for the photonic power penalty to be realistic. This can be negated by selecting *m_b* or *m_xb*. In summary, the above effects are leveraged by the HIRs to reduce the total power penalty by $\sim 3-4$ dB for a 16×16 PSF.

VI. Discussion

Photonic ICN simulation is an active domain, with various frameworks developed by industry and academia. We discuss these, and summarise the key innovations that set our technique apart from the state-of-the-art, as captured in Table III.

Industry-developed tools target the device layer. They focus on multi-physics solvers for analysing the propagation of light through the physical medium, and encapsulate the results into "compact models", which are used as building blocks to form complex networks. Representative products have been developed by Ansys-Lumerical (Lumerical Suite), Luceda Photonics (IPKISS), and Cadence (EPDA environment).

TABLE III
Qualitative comparison with state-of-the-art simulators.

	Permutation Setup	**Power Penalty for Partial Permutation Simulation**	**Dynamic Traffic Integration**	**Multi-order Crosstalk Simulation**	**Simulation Target**
Industry tools (Lumerical suite, Luceda IPKISS, Cadence EPDA)	Manual	Yes	No	Yes	Physical Layer
PhoenixSim	Manual	No	No	Yes	Physical Layer
Dupuis & Lee	Automatic	No	No	Yes	Physical Layer
DSENT	Manual	No	No	No	Physical Layer
This Work	Automatic & Manual	Yes	Yes	Yes	Physical Layer & Control Plane

Various simulation environments have also been proposed by the academic community. DSENT [4], which focuses on the co-integration of MRR-based photonic interconnects with the electronic backend, has been used to propose various photonic interconnects in the computer architecture domain (including MZI-based PSFs). However, it only supports random traffic and is therefore unable to capture the dynamic interactions of traffic flows we describe here. PhoenixSim [25], which targets the photonic layer, has also been used to propose novel routing strategies for MZI-based PSFs. As with DSENT, PhoenixSim does not include traffic modelling and requires manual connection pattern setup. Other frameworks (e.g. [11] [26] [27]), while able to capture the behaviours they target, are limited in their ability to co-simulate multi-order crosstalk, traffic configuration and waveguide crossing models with MZI-based PSFs.

In contrast to the above, the simulation framework that encompasses the techniques described in this work bridges the gap between the switch control plane and the photonics design plane. By capturing beam propagation at the lightplane level we isolate the propagation behaviour of each lightpath as it traverses the PSF (signal power and leakages), before forming the photonic power penalty at the outputs. By doing so, we are able to capture the photonic behaviour of partial PSF saturation for arbitrary partial permutations. We can also investigate the impact of faults in individual devices within the PSF, as exemplified in Section IV. Further, by incorporating the beam propagation model into a flow-level simulator driven by dynamic traffic, we can model the effects of switch control plane optimisations on the photonic behaviour of the PSF as we have shown in Section V. Our technique therefore unlocks new avenues for research into PSF control and optimisation.

VII. Conclusions

This paper has presented a novel, traffic-driven simulation methodology for co-evaluating the switch control plane and device design for PSF ToR switches. Our approach combines flow-level interconnection network simulation, which is ideal for the bufferless nature of PSFs, and optical beam propagation. This yields a simulation environment where both the physical layer of PSFs and their control algorithms can be

simultaneously evaluated and optimised, thereby bridging the gap between device and control plane simulation.

We have established the accuracy of the photonic beam propagation model by modelling two state-of-the-art PSF chips. The simulations are accurate within $\sim$ 0.5 dB in terms of ILoss compared to the published chip data. However, it tends to slightly overestimate crosstalk; it exhibits ~ -27 dB crosstalk where the chips report ~ -30 dB. In addition, we have presented how the crosstalk ratios of MZI and wgx devices affect the worst-case PSF crosstalk; due to the Beneš connection pattern, the device with the highest crosstalk level determines the PSF crosstalk level, and thus the photonic power penalty.

We have shown that with the DWDM scenario, the photonic power penalty for 32×32 PSFs grows exponentially, rendering them unrealistic. Depending on the PSF size, PSF routing algorithm selection can reduce the worst-case photonic power penalty by up to 4 dB for 16×16 PSFs. This can reduce laser power and increase communication energy efficiency. In summary, our PSF modelling technique enables future research directions for silicon photonic switch control plane design by providing a wider evaluation spectrum.

ACKNOWLEDGEMENTS

We thank Liang Yuan Dai and Keren Bergman of the Lightwave Research Laboratory, University of Columbia, for their conversations on photonics theory. Mikel Luján is supported by an Arm/RAEng Research Chair Award and a Royal Society Wolfson Fellowship. Javier Navaridas is supported by a Rámon y Cajal Fellowship RYC2018-024829-I from the Spanish Ministry of Science, Innovation and Universities. This work is partially supported by the Basque Government (projects KK-2021/00095 and IT1504-22).

REFERENCES

[1] C. Sun, D. Jeong, M. Zhang, W. Bae, C. Zhang, P. Bhargava, D. Van Orden, S. Ardalan, C. Ramamurthy, E. Anderson *et al.*, "Teraphy: An o-band wdm electro-optic platform for low power, terabit/s optical i/o," in *2020 IEEE Symposium on VLSI Technology*. IEEE, 2020, pp. 1–2.

[2] A. Agrawal and C. Kim, "Intel tofino2–a 12.9 tbps p4-programmable ethernet switch," in *2020 IEEE Hot Chips 32 Symposium (HCS)*. IEEE Computer Society, 2020, pp. 1–32.

[3] H. Li, G. Balamurugan, T. Kim, M. N. Sakib, R. Kumar, H. Rong, J. Jaussi, and B. Casper, "A 3-d-integrated silicon photonic microring-based 112-gb/s pam-4 transmitter with nonlinear equalization and thermal control," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 19–29, 2020.

[4] C. Sun, C.-H. O. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L.-S. Peh, and V. Stojanovic, "Dsent-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip*. IEEE, 2012, pp. 201–210.

[5] N. Binkert *et al.*, "The gem5 simulator," *ACM SIGARCH computer architecture news*, vol. 39, no. 2, pp. 1–7, 2011.

[6] L. Lu *et al.*, "16×16 optical switch based on electro-optic mach-zehnder interferometers," *Optics express*, vol. 24, no. 9, pp. 9295–9307, 2016.

[7] *Xilinx 7 Series FPGA Power Benchmark Design Summary*, Xilinx, 2015. [Online]. Available: https://www.xilinx.com/publications/technology/power-advantage/7-series-power-benchmark-summary.pdf

[8] N. Dupuis *et al.*, "Ultralow crosstalk nanosecond-scale nested 2× 2 mach–zehnder silicon photonic switch," *Optics letters*, vol. 41, no. 13, pp. 3002–3005, 2016.

[9] Z. Lu *et al.*, "High-performance silicon photonic tri-state switch based on balanced nested mach-zehnder interferometer," *Scientific reports*, vol. 7, no. 1, pp. 1–7, 2017.

[10] F. Testa and L. Pavesi, *Optical switching in next generation data centers*. Springer, 2017.

[11] N. Dupuis and B. Lee, "Impact of topology on the scalability of mach–zehnder-based multistage silicon photonic switch networks," *Journal of Lightwave Technology*, vol. 36, no. 3, pp. 763–772, 2017.

[12] B. G. Lee, "Photonic switch fabrics in computer communications systems," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*. IEEE, 2018, pp. 1–22.

[13] Q. Cheng *et al.*, "Recent advances in optical technologies for data centers: a review," *Optica*, vol. 5, no. 11, pp. 1354–1370, 2018.

[14] Q. Cheng, M. Bahadori, and K. Bergman, "Advanced path mapping for silicon photonic switch fabrics," in *CLEO*, 2017, pp. 1–2.

[15] P. Yuen and L. Chen, "Optimization of microring-based interconnection by leveraging the asymmetric behaviors of switching elements," *Journal of lightwave technology*, vol. 31, no. 10, pp. 1585–1592, 2013.

[16] M. Kynigos *et al.*, "On the routing and scalability of mzi-based optical beneš interconnects," *Nano Communication Networks*, vol. 31, 2020.

[17] R. Yao and Y. Ye, "Toward a high-performance and low-loss clos–benes-based optical network-on-chip architecture," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 12, pp. 4695–4706, 2020.

[18] J. Navaridas *et al.*, "Inrflow: An interconnection networks research flow-level simulation framework," *Journal of Parallel and Distributed Computing*, vol. 130, pp. 140 – 152, 2019.

[19] G. T. Reed and A. P. Knights, *Silicon photonics: an introduction*. John Wiley & Sons, 2004.

[20] G. Michelogiannakis *et al.*, "Challenges and opportunities in system-level evaluation of photonics," *Metro and Data Center Optical Networks and Short-Reach Links II*, 2019.

[21] H. Ballani, P. Costa, R. Behrendt, D. Cletheroe, I. Haller, K. Jozwik, F. Karinou, S. Lange, K. Shi, B. Thomsen *et al.*, "Sirius: A flat datacenter network with nanosecond optical switching," in *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, 2020, pp. 782–797.

[22] X. Xue and N. Calabretta, "Nanosecond optical switching and control system for data center networks," *Nature communications*, vol. 13, no. 1, p. 2257, 2022.

[23] R. Ramaswami, K. Sivarajan, and G. Sasaki, *Optical networks: a practical perspective*. Morgan Kaufmann, 2009.

[24] Q. Cheng *et al.*, "Silicon photonic switch topologies and routing strategies for disaggregated data centers," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 2, pp. 1–10, 2019.

[25] J. Chan, G. Hendry *et al.*, "Phoenixsim: A simulator for physical-layer analysis of chip-scale photonic interconnection networks," in *DATE 2010*. IEEE, 2010, pp. 691–696.

[26] J. E. Miller, H. Kasture *et al.*, "Graphite: A distributed parallel simulator for multicores," in *HPCA-16*. IEEE, 2010, pp. 1–12.

[27] H. Zhang *et al.*, "Comparative analysis of simulators for optical network-on-chip (onoc)," in *PAAP-12*. IEEE, 2021, pp. 19–23.

[28] S. Zhao *et al.*, "16x16 silicon mach-zehnder interferometer switch actuated with waveguide microheaters," *Photon. Res.*, vol. 4, no. 5, pp. 202–207, Oct 2016.

[29] D. Opferman and N. Tsao-Wu, "On a class of rearrangeable switching networks part i: Control algorithm," *The Bell System Technical Journal*, vol. 50, no. 5, pp. 1579–1600, 1971.

[30] S. Kandula *et al.*, "The nature of data center traffic: Measurements & analysis," in *9th ACM SIGCOMM Conf. on Internet Measurement*, 2009, p. 202–208.

[31] X. Yuan *et al.*, "A new routing scheme for jellyfish and its performance with hpc workloads," in *Procs. of the Intl. Conf. on High Performance Computing, Networking, Storage and Analysis*, ser. SC '13. New York, NY, USA: Association for Computing Machinery, 2013. [Online]. Available: https://doi.org/10.1145/2503210.2503229

[32] J. Kim *et al.*, "Technology-driven, highly-scalable dragonfly topology," in *2008 Intl. Symposium on Computer Architecture*, 2008, pp. 77–88.

[33] R. Thakur and W. Gropp, "Improving the performance of collective operations in mpich," in *European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting*, 2003, pp. 257–267.

[34] C. Seitz, "The cosmic cube," *Commun. ACM*, vol. 28, pp. 22–33, 01 1985.

[35] T. Benson, A. Akella, and D. Maltz, "Network traffic characteristics of data centers in the wild," in *Procs. of the 10th ACM SIGCOMM Conf. on Internet measurement*, 2010, pp. 267–280.
[36] S. Zaheer *et al.*, "Locality-aware process placement for parallel and distributed simulation in cloud data centers," *Journal of Supercomputing*, 08 2019.
[37] R. Fujimoto, "Research challenges in parallel and distributed simulation," *ACM Trans. Model. Comput. Simul.*, vol. 26, no. 4, May 2016.
[38] A. Greenberg *et al.*, "Vl2: A scalable and flexible data center network," in *ACM SIGCOMM 2009 Conference on Data Communication*, ser. SIGCOMM '09, 2009, p. 51–62.
[39] M. Kynigos *et al.*, "Power and energy efficient routing for mach-zehnder interferometer based photonic switches," in *ACM ICS*, 2021, pp. 177–189.