







Understanding the Impact of Arbitration in MZI-Based Beneš Switching Fabrics

Javier Navaridas , Markos Kynigos , Jose A. Pascual , Mikel Luján , Jose Miguel-Alonso ,
and John Goodacre 

Abstract—Top-of-rack switches based on photonic switching fabrics (PSF) could provide higher bandwidth and energy efficiency for datacenters (DC) and high-performance computers (HPC) than these with traditional electronic crossbars. However, because of their bufferless nature, PFS are affected by contention much more drastically than traditional packet-switched electronic networks where traffic can advance towards its destination, getting buffered upon encountering contention and resuming transmission once resources are freed. In contrast, PSFs stop the injection of all traffic that generate contention. Consequently, it is important to understand how the order in which flows are serviced affects performance metrics. Our contribution is to quantify this impact through a comprehensive simulation-based evaluation focusing on a recently fabricated PSF prototype. Our experiments include configurations with three routing algorithms, two switching methods, three ToR switch sizes and 9 representative workloads from the DC and HPC domains. We found that the effect of arbitration on raw throughput is negligible but, when considering more realistic loads, selecting an appropriate arbitration policy can improve communication time and energy efficiency. Indeed, the communication time can be reduced by between 10% and 30% by employing appropriate arbitration. Switching energy efficiency can also be improved between 4% and 13%. Finally, insertion loss is barely affected, with differences below 2%. LFU and ARR were found to obtain the best results. LFU is very good with regular workloads but one of the worse with irregular workloads. ARR obtains good results regardless of the type of workload.

Index Terms—Top-of-rack photonic switches, arbitration, mach-zehnder interferometers, performance evaluation, simulation.

I. INTRODUCTION

SCALING datacenter (DC) and high-performance computing (HPC) interconnection networks (ICNs) is a continuous

Manuscript received 3 June 2022; revised 4 April 2023; accepted 21 November 2023. Date of publication 28 November 2023; date of current version 5 January 2024. This work was supported in part by the EU Horizon2020 under Grant 754337, in part by the EPSRC under Grant EP/T026995/1, and in part by Basque Country Government under Projects KK-2023/00012, KK-2023/00090, and IT1504-22. The work of Javier Navaridas was supported by the Spanish Ministry of Science, Innovation and Universities under Grant RYC2018-024829-I. The work of Mikel Luján was supported by an Arm/RAEng Research Chair Award and a Royal Society Wolfson Fellowship. Recommended for acceptance by A. Bhatele. (Corresponding author: Javier Navaridas.)

Javier Navaridas, Jose A. Pascual, and Jose Miguel-Alonso are with the Department of Computer Architecture and Technology, University of the Basque Country, 20018, Donostia-San Sebastián, Spain (e-mail: javier.navaridas@ehu.es; joseantonio.pascual@ehu.es; j.miguel@ehu.es).

Markos Kynigos, Mikel Luján, and John Goodacre are with the Department of Computer Science, University of Manchester, M13 9PL Manchester, U.K. (e-mail: markos.kynigos@manchester.ac.uk; mikel.lujan@manchester.ac.uk; john.goodacre@manchester.ac.uk).

Digital Object Identifier 10.1109/TPDS.2023.3336703

challenge as their communication demands continue to grow. All-optical networks incorporating silicon photonics, hereafter referred to as *Photonic ICNs*, are a promising approach for such large scale systems. Deploying photonic switching fabrics (PSFs) within HPC and DC network switches provides significant advantages compared with standard electronic crossbars. Photonics technology offers greater data density (approx one order of magnitude) due to wavelength division multiplexing (WDM), and can accommodate more bandwidth per link. Photonic ICNs also exhibit very low propagation latency and relatively distance-independent energy consumption [1]. These benefits, together with the rapid advancement on the technology side suggest that photonic ICNs are getting closer to become a commonplace technology. A significant step has been the introduction of CMOS-compatible photonic devices [2].

However, there still remain many challenges to develop and deploy efficient photonic ICNs. Although some attempts have been made, it is currently not possible to buffer light in optical form for practical amounts of time [3]. This precludes the deployment of photonic packet-switching at the transmission level in high-performance photonic network switches. Relying on electronic buffering requires extra opto-electric and electro-optic conversions, which detracts from the benefits of optical transmission. Also, the physical characteristics of PSFs, especially insertion loss (hereafter *ILoss*) and photonic crosstalk, can affect the required laser power to a point where it negates the benefits of photonics. To side step these effects and avoid excessive energy consumption while maintaining low wiring complexity, bufferless PSFs based on Beneš networks with Mach-Zehnder Interferometers (MZIs) [4], [5] are a promising technology we use as a workbench in this paper. Such networks are normally controlled either using circuit switching (CS) or time-division multiplexing (TDM) [6], [7], [8].

A Beneš network is a rearrangeably non-blocking (RNB) recursive topology, as seen in Fig. 1. It is also known to feature the lowest *ILoss* among RNB topologies, when using photonic 2-port switching cells such as Mach-Zehnder Interferometers (MZIs). In contrast to electronic Beneš networks, standard switch control algorithms in PSFs, such as the *Looping Algorithm* [9], are unable to route network traffic in an energy-efficient manner. On the other hand, PSF-focused switch control algorithms [10], [11], can not completely eliminate contention. For instance, Fig. 2 presents a contention scenario for MZI-based PSFs. If the $I_0 \rightarrow O_1$ and $I_3 \rightarrow O_2$ transmissions are scheduled first, all feasible paths for $I_2 \rightarrow O_0$ (dotted lines) are blocked

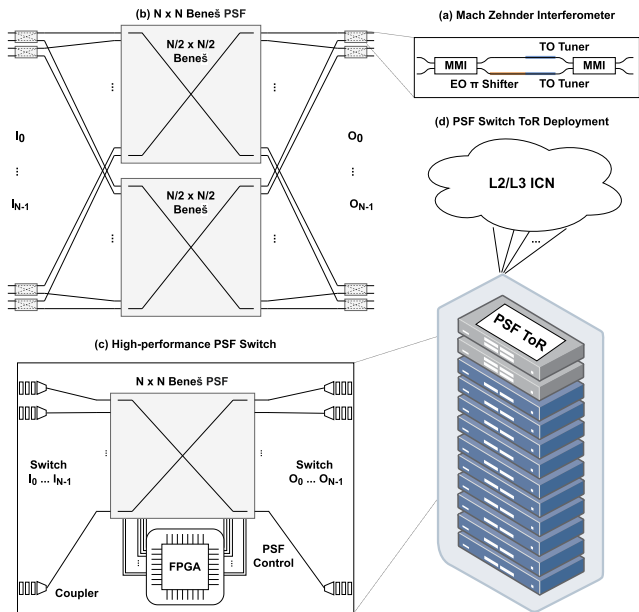


Fig. 1. (a) Schematic of a 2×2 EO/TO MZI switching element. (b) An $N \times N$ MZI Beneš PSF. (c) A high-performance switch containing the FPGA-controlled PSF. (d) Deployed ToR switch within a DC or HPC rack.

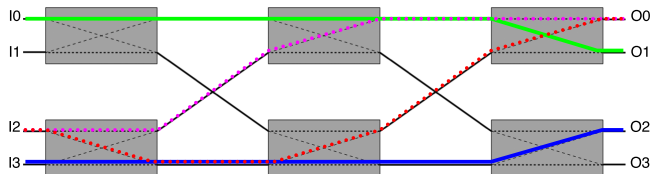


Fig. 2. Fabric contention in a 4×4 Beneš network.

because they require resources that are already-allocated (solid lines). Thus, the incoming transmission must wait until resources are freed, incurring time and throughput penalties. We refer to this phenomenon as *switch fabric contention*. Similarly, in *output contention*, several input ports want to send to the same output port but only one of them can transmit at a time, so other ports are consequently blocked.

The existence of either form of contention means that the order in which input ports are serviced, i.e., *PSF arbitration*, can have a significant impact on the overall performance as exemplified by Fig. 3. The figure shows the arbitration of a 4-port Beneš PSF. Time (μs) flows from left to right and input ports can be transmitting (in gray), blocked (in red) or idle (empty). With CS, if a short flow is blocked by long ones, it would incur a significant latency penalty which is generally detrimental to performance. With TDM, a fair interleaving of slots tends to be advantageous because it ensures a balanced sharing of network bandwidth among ports and, in turn, among traffic flows [11]. Based on the above, while many works investigated practical aspects of photonic ICNs such as routing and switching, arbitration has not been investigated, to the best of our knowledge.

Our objective is, therefore, to address this gap by carrying out a comprehensive study of the impact of arbitration in MZI-based PSFs. In particular, we consider their use within Top of Rack (ToR) switches (as per Fig. 1) and carry out a simulation-based

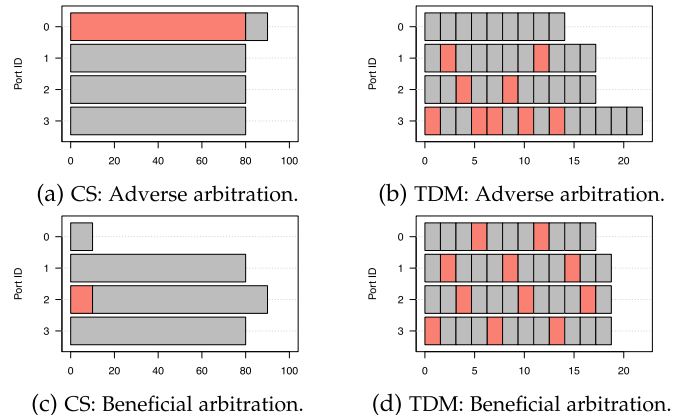


Fig. 3. Examples of beneficial and adverse flow arbitration in CS and TDM 4-port PSFs.

study using experimental data from a recently manufactured 16×16 PSF chip. The study employs 3 state-of-the-art routing algorithms, 2 switching techniques, 9 workloads from the DC and HPC domains, and 3 ToR switch sizes. We investigate 5 classical arbitration policies and propose 2 new round robin variants: accelerated round-robin and multi-level round-robin. With this setup, we can analyze the interactions between routing, switching and arbitration in Beneš PSF-enabled ToR switches. In particular we consider four metrics: switch throughput, communication time, insertion loss and switching energy per bit. We find that switch throughput is barely affected by arbitration when dealing with uniform traffic from independent sources. In contrast, with more realistic workloads that include causality among tasks, appropriate arbitration policies can yield communication time savings *without sacrificing energy efficiency*. In addition, we see that switching and arbitration exhibit interrelated effects depending on the workload. Routing and arbitration, on the other hand, can be designed independently, as the impact of arbitration is consistent across the examined routing algorithms. Finally, we investigate switch size scalability and find that the effect of arbitration is consistent across sizes; just increasing slightly with the switch radix.

With respect to the policies, we found that classic policies such as round robin and random arbitration suffer of starvation related issues that have a substantial impact on performance. The least frequently used policy is outstanding with regular workloads but can be detrimental with irregular ones. Accelerated round robin achieves comparable performance but is effective for all types of traffic.

II. BACKGROUND

A. Opportunities for Photonic Switching

Silicon photonics switches are very appealing for their potential to increase the energy efficiency of communication within DCs and HPCs. PSFs are composed of multiple active photonic devices acting as 2×2 switch cells (e.g., MRRs or MZIs), which are tiled into switch matrices and connected using passive photonic devices (waveguides and, if necessary, waveguide

TABLE I
POWER AND ENERGY CONSUMPTION OF SWITCHES

Device Model	Switch Radix	Data Rate	Switching Capacity	Power Dissipation	Switching Energy
CISCO Nexus 3636C-R	36 ports	100 Gb/s	3.6 Tb/s	1,179 Watt	327.5 pJ/bit
Aruba CX 8320	32 ports	40 Gb/s	1.3 Tb/s	230 Watt	179.7 pJ/bit
Aruba CX 8325	32 ports	100 Gb/s	3.2 Tb/s	406 Watt	126.9 pJ/bit
CISCO Nexus 3464C	64 ports	100 Gb/s	6.4 Tb/s	712 Watt	111.3 pJ/bit
Huawei CloudEngine 9860	128 ports	100 Gb/s	12.8 Tb/s	1,051 Watt	82.1 pJ/bit
Arista 7368X4 Series	32 ports	400 Gb/s	12.8 Tb/s	966 Watt	75.5 pJ/bit
NVIDIA MQM9700	64 ports	400 Gb/s	25.6 Tb/s	1,084 Watt	42.3 pJ/bit
NVIDIA SN2700	32 ports	100 Gb/s	3.2 Tb/s	135 Watt	42.2 pJ/bit
MZI PSF Switch	16 ports	512 Gb/s	8.2 Tb/s	21.2 Watt	2.6 pJ/bit
	32 ports	512 Gb/s	16.4 Tb/s	23.0 Watt	1.4 pJ/bit
	64 ports	512 Gb/s	32.8 Tb/s	27.3 Watt	0.8 pJ/bit
	64 ports	400 Gb/s	25.6 Tb/s	27.3 Watt	1.1 pJ/bit

crossings). PSFs can be deployed as the switching core of high performance network switches.

The photonics switch we examine is presented in Fig. 1 which shows (a) a thermally-electrically tuned MZI switch cell conforming the fabric, (b) the recursive Beneš topology in which the fabric is organized, (c) the architecture of a photonics switch based on the Beneš fabric and a controller FPGA and (d) a DC or HPC rack employing a photonics ToR switch connecting to higher ICN tiers. In particular, this architecture is formed using broadband MZIs which, due to their operating principles, are able to switch multiple wavelengths simultaneously at ns time and without being affected by the data rate carried by individual wavelengths. This latter characteristic is called *bandwidth transparency* (BWT). The BWT of MZI-based PSFs can be leveraged to adopt dense-wavelength-division multiplexed (DWDM) links. This reduces the individual data rate per wavelength but increases signal quality and energy efficiency while maintaining high aggregate data rates. Note that DWDM comes at the expense of more complex photonic transceivers. Another advantage is that electronic switches must either be upgraded at every data rate generation to support new transceivers, or transceivers must remain constrained by legacy capabilities. In contrast, BWT allows photonic switches to accommodate future data rates, as their performance is less dependent on per-wavelength data rates or number of wavelengths, so infrastructure investment can be amortized over a longer-term.

B. Comparison With Electronic 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 popular ToR switches from Aruba, NVIDIA-Mellanox, Arista, Huawei and CISCO. 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 20 W optics for 400 Gb/s, 4.5 W for 100 Gb/s and 2.5 W for 40 Gb/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 400 Gb/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. In contrast, we estimate the switching energy for PSF switches extrapolating from the MZI-based 16×16 switching fabric characterized in [4]. Such PSF would exhibit a very small switching power envelope (between 1.2 W and 7.3 W 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 devices¹ we take a pessimistic power envelope of 20 W. We assume a deployment scenario with different switch radices, 512 Gb/s links with 32 wavelengths, as well as a comparative scenario assuming 64 ports and 400 Gb/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 reduced by 1-2 orders of magnitude. This can be highly compelling for photonic ICNs, as their adoption can reduce the total cost of ownership or increase the power budget for other components such as CPUs, I/O, etc.

C. Trends in Photonic Architectures

Recently, the topic of exploiting hardware asymmetries in photonic architectures through routing has gained traction in the community. Recent works proposed an exhaustive analysis of paths to select the most effective permutation by evaluating all potential fabric states in both MZI [10] and microring resonator designs [12]. Although it is an interesting approach, their brute-force design quickly becomes intractable as the number

¹See Xilinx 7 Series FPGA Power Benchmark Design Summary. Available (March 2023) at: <https://www.xilinx.com/publications/technology/power-advantage/7-series-power-benchmark-summary.pdf>

of ports and possible fabric states scales up. For this reason, other approaches rely on greedy heuristics [11] or other forms of automated learning [13], providing lightweight high-quality solutions. The general objective of all these algorithms is to allocate paths that incur the least amount of ILoss from waveguide crossings, and/or MRR/MZI traversal.

In terms of switching methods, optical technologies tend to use a combination of space-division multiplexing (SDM), TDM and/or WDM in order to maximize throughput and to use bandwidth more effectively. A survey of different approaches can be found in [14]. The *Data Vortex* optical ICN uses TDM/WDM switching [15]. Other authors study different mixes of SDM/TDM/WDM [8], [16], [17]. None of these works, however, takes into consideration the effects that arbitration may have in the use of resources neither in the temporal, spatial or wavelength domains.

In fact, we should remark that the research on arbitration for photonics is scarce and limited to the optical network-on-chip (ONoC) domain [18], [19]. This justifies the timeliness and novelty of our research which focuses on a different photonic technology (MZIs) applied at a different system level (ToR switch) and with rather different topological constraints (a Beneš network). Note that while we focus our analysis on MZI-based Beneš PSFs which have been thoroughly investigated by the community, the impact of arbitration would still be present for many flavors of PSFs, including other topologies and device types.

III. ARBITRATION

This section discusses the technology and the switch architecture we focused on. It also describes the arbitration policies that we consider in our study. As explained above, arbitration in Beneš MZI PSFs can impact communication time, see again Fig. 3.

A. Switch Design

As a test bench for the effects of arbitration, we consider in our study a ToR switch based on the 16×16 photonic switch demonstrated in [4]. From it, we extrapolate to 32×32 and 64×64 switches to investigate the scalability of the arbitration policies and their applicability to future designs. Note that, with current devices and processes, 32×32 switches are already at the limit of the technology due to exceedingly high crosstalk [5], [20]. We assume a deployment scenario where these devices operate as ToR switches connected to both servers and the higher tier of the IC.² We consider WDM transmission with 32λ working at 16 Gb/s data rate using an On-Off Keying (OOK) scheme [21]. This yields a 512 Gb/s aggregate bandwidth per port, with endpoints modulating on all λ simultaneously.

The modeled MZIs require TO tuning to reach the *cross* state and additional EO tuning to reach the *bar* state. With

²Note that the tier above the ToR needs to translate the signal from the optical domain back to the electronic domain because the crosstalk induced within the ToR switch onto the optical signal would already be near its limit. This incurs substantial latency but, in exchange, allows for buffering traffic and, thus, for packet-switching, at this level.

EO tuning, which takes a few *ns*, and all MZIs being switched simultaneously, the switch fabric reconfiguration time becomes relatively short and the bandwidth and latency overheads are adequate for both CS and TDM switching methods.

We consider a centralized controller for the switch fabric, e.g., an FPGA or an ASIC. During boot-up the controller generates and stores pre-computed paths for the source-destination pairs. At run time, the fabric state is built incrementally, serving communication requests sequentially in the order specified by the arbitration policy. For each request it will allocate one of the pre-computed paths as directed by the routing algorithm. If no path is available, the controller blocks the input port. Given that the Beneš network is *rearrangeably* non-blocking and offers a relatively high path diversity of $N/2$ (for N ports), most routing algorithms should maintain a sufficiently low level of switch contention. However, when servicing full permutations or if output contention arises, blocking can still occur. Thus, the order in which ports are serviced has a substantial effect on both the availability of paths and the characteristics of the allocated path. For instance, the first request to be serviced is able to select among all possible paths, whereas the last ones are very likely to be blocked, and even if they are not, the number of paths to select from is reduced.

B. Arbitration Policies

Our experiments consider the following classic arbitration policies, and propose some extensions that suit themselves nicely for our target architecture:

First-in, First-out (FIFO) — The ports are serviced in the order in which requests are received.

Least recently used (LRU) — The priority of the ports increases over time, so lower priority is given to the inputs that have been serviced more recently.

Least frequently used (LFU) — The inputs that have transmitted the least traffic have the highest priority so that a balanced use of all ports is maintained.

Random (RND) — Ports are serviced in random order, without following any priority scheme, which is expected to ensure a fair utilization of resources [22], [23]. As we will see below, this is not the case for the PSFs under study.

Round-robin (RR) — Ports are serviced sequentially starting from an index value. At each round of arbitration the index is incremented by one. As will be shown later, this indexing mechanism is not very effective in the context of PSFs, so two new RR variants are proposed below.

Accelerated round-robin (ARR) — A modification to RR. Instead of increasing the index by one, ARR updates it to the first port in the round that requested a path but was blocked.

Multi-level round-robin (MRR) — Another modification to RR. In this case, we split the switch into 4 consecutive sets of ports, each of them with their own index, plus an extra index for selecting a set. Each round increments the set index, plus the index of the selected set.

IV. EXPERIMENTAL METHODOLOGY

This section discusses our experimental methodology. It describes the simulated models, including the network architecture and workloads, and explains how results are presented.

A. Simulation Model and Workloads

We use INRFlow [24], an open source, light-footprint, highly scalable, flow-level network simulator which we have extended to support PSFs.³ The results of the simulator have been compared with empirical measurements of two different fabricated chips, observing relatively good accuracy [20]. In particular, we evaluate two switching methods: **CS**, where flows reserve a path and use it to send all the required data, and **TDM**, where flows are segmented into slices of a predefined size [14], [15], [16], [17]. In general a shorter slot provides better flow interleaving and lower internal fragmentation, but requires a more frequent reconfiguration of the switch fabric which imposes some delay and throughput penalties. For simplicity, we consider a 100 KB slot size ($\approx 1.5\mu s$), which was found to be a reasonable compromise for the available bandwidth and the tuning delay [8]. As it is common practice in DCs [25], 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.

Endpoints are modeled as traffic producer/consumer nodes. For the throughput analysis, we use uniformly distributed synthetic traffic from independent sources. The rest of experiments use more realistic traffic, following the dynamics defined by a range of application-inspired workloads based on representative HPC and DC applications and well-known benchmarks. These workloads include causality among tasks, which need to wait for traffic from other tasks. This means that most workloads go through phases of high and low network pressure. Unless otherwise stated, workloads send 5,000 flows and all flows are 1 MB long. In the descriptions, N represents the number of tasks of a given workload, which in our experiments is the same as the switch radix. We consider the following workloads:

All2All (AA) — This is a typical collective operation in HPC applications and also representative of DC traffic, as it constitutes the core of MapReduce. Tasks communicate among themselves sending flows to all other tasks. Thus, the total number of flows is $N \cdot (N - 1)$.

AllReduce (AR) — An optimized, binary implementation of the AllReduce collective [26], widely used in parallel applications from a range of domains. This workload sends a total of $N \cdot \log N$ flows.

Bisection (BI) — Tasks perform pair-wise communications swapping pairs randomly every round. This benchmark is commonly used [27], [28] as an estimator of the bisection bandwidth, a topological characteristic of ICNs closely related to their performance and resiliency [29].

HotRegion (HR) — A classic networking benchmark where traffic is generated at random, but non-uniformly: 25% of the traffic goes to the *hot region*, which comprises 12.5% of the

output ports; the rest of the traffic is sent uniformly at random. This creates an unbalanced use of network resources which intensifies output contention.

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 [30]. This results in a total of $N^2/2$ flows.

RandomApp (RA) — Selects the source and destination uniformly at random but introducing causality among messages to better resemble real applications. This is a typical networking benchmark which, depending on the context can stress the ICN because of the low-locality or be benign as it produces a balanced use of network resources. In the case of all-optical interconnects it induces a substantial amount of output contention due to the lack of buffering. According to [31], the traffic mix run on a typical DC is unstructured and essentially random in nature.

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 : $D = (S + t) \bmod N$. This is akin to the adversarial traffic proposed in [32].

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 [33]. 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 [33].

In the discussions below, we classify these workloads into two distinct categories: in *Regular* workloads, all tasks progress at the same pace, with homogeneous communication phases of fixed size. Thus, the critical path of all tasks is similar. AA, AR, BI, NB and SH belong to this category. In contrast, in *Irregular* workloads, each task progresses at a different pace, dictated by traffic causality. In this case, communication phases are different for each task and their critical paths differ substantially. HR, RA, TL and TR belong to the irregular category. Following the standard practice for DCs and clusters, we assume that the system scheduler models the system as a flat network with no locality information. This results in tasks being distributed randomly across the network [34], [35].

B. Routing Algorithms

To investigate the impact of arbitration on routing, we consider three routing schemes. In particular we use random path as our baseline and also two routing algorithms which exploit underlying hardware asymmetries to minimize ILoss (from [11]).

Minimize Bar States (mb) — Prioritizes the paths with the least MZIs in Bar state, since this is the state with higher ILoss and power consumption.

Minimize Crossings (mx) — Since waveguide crossings is another substantial contributor to ILoss, this routing selects the path with the minimum number of them.

³Available at: <https://gitlab.com/ExaNeSt/phinflow>

TABLE II
SIMULATION PARAMETERS

Component	ILoss	Tuning Type	Power Cons.
Bar MZI	1.4 dB	Thermal	0-26 mW
Cross MZI	0.4 dB	Mean, STD	15.725, 6.608
Wg. Crossing	0.05 dB	Electrical	3.28-5.88 mW
Wg. Propag.	1.18 dB/cm	Mean, STD	5.166, 0.428

(a) Insertion Loss.

(b) Power consumption.

Random Path (rnd)—Selects a path randomly, without taking into account any characteristic of the path.

Note that these routing algorithms are of quite different nature. The former two have different objectives and consider different aspects of the underlying architecture, while the later one is completely agnostic of the architecture.

C. Methodology

We simulate different system configurations consisting of workload, arbitration policy, routing, switching method and network size. Table II presents the simulation parameters for the photonic components. Each configuration is simulated 100 times with different random seeds and the mean and standard deviations of the following performance metrics are gathered.

Aggregated switch bandwidth with random uniform traffic at full load to measure the effect that arbitration policies have on the raw throughput of the switch.

Communication time to assess the impact of arbitration policies on the execution speed of the workloads.

Switching energy per bit – we measure the total energy consumed for MZI tuning and divide it by the total amount of traffic traversing the switch. This metric is used to show the impact of arbitration policies on energy efficiency.

Maximum ILoss used to estimate the impact of arbitration policies on laser power. We found that the differences in terms of ILoss are very small and, indeed, that confidence intervals overlap in most cases and, thus, are not statistically significant. Hence we do not report them in the paper due to space constraints.

Given the large number of experiments, our analysis only brings up a subset of representative results showing the general findings of our study. The results excluded are similar to the ones that are discussed below but are not presented for the sake of brevity. The complete set of results is available through an OSF repository.⁴ To make comparisons easier, and to isolate the effects of the arbitration policies from other aspects of the architecture, we normalize all results to the arbitration policy producing the best result. This way, the best policy has a 1, and it is easy to see the degradation suffered with other policies. For example, if a policy obtains a result of 1.1, it means it requires 10% more time or energy than the best result.

The plots include 95% confidence intervals to capture the variability exhibited by the different configurations. For clarity,

⁴Available at: https://osf.io/285d4/?view_only=60d0d30da13e4948a90350b215ac4490

the arbitration policies that are based on priorities are colored in different shades of red, the ones based on round-robin are colored in shades of blue, and the random policy that follows none of these approaches is colored gray.

V. ANALYSIS OF EXPERIMENTS

Focusing on the effects of arbitration and their interrelation with other aspects of photonic switch architectures, we discuss the results of our experimental work. We start by analyzing the impact that switch arbitration has on the raw throughput of a switch. Then, we move to experimenting with more realistic workloads to provide a deeper understanding of their relation with various aspects of the switch architecture: routing algorithms, switch radix and switching methods.

A. Aggregated Switch Bandwidth

We begin by investigating the impact that arbitration policies may have on the throughput of photonic switches. Fig. 4 shows the aggregated switch bandwidth under uniform traffic from independent sources injecting at maximum load. As expected, throughput grows linearly with switch radix. Furthermore, for a given radix, we observe very small differences with respect to the routing or arbitration employed. Routing-wise, the differences are negligible, within a 1%. Regarding arbitration, the differences are slightly higher, but still insubstantial. In particular, switches using policies based on round-robin saturate at a scarcely higher load due to a small reduction of switch fabric contention. This suggests that serving ports sequentially might have some unexpected benefits.

However, these raw throughput results do not have any consideration about the dynamics of real applications or the way they generate traffic. For this reason, it is essential to carry out a deeper analysis where tasks dependencies are considered. The remaining of our evaluation will use the application-inspired workloads discussed above.

B. Routing

We begin by investigating the interactions between routing algorithms and arbitration policies. Fig. 5 shows the communication time with a 16-port switch using CS and the three routing algorithms. The first thing we notice is that the differences between arbitration policies can be substantial, up to around 30%. This is in stark contrast to the minute differences in terms of aggregated switch bandwidth studied above and illustrates the need for deeper analysis using realistic application-inspired workloads.

In general, the potential benefits of arbitration vary according to the workload. They are the highest for All2All and AllReduce, where all endpoints transmit exactly the same volume of data and, therefore, a fair way of sharing the network bandwidth is beneficial. In addition, AllReduce is one of the workloads with the highest level of causality between flows. This means that any delay suffered by one flow is transmitted to the other flows within the causality chain. In contrast, the

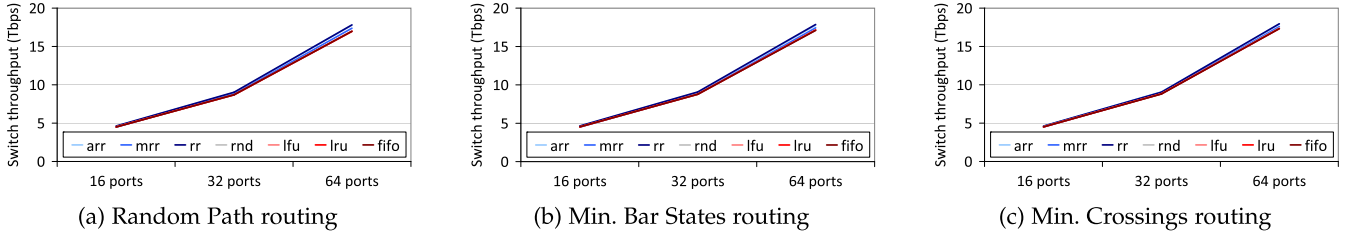


Fig. 4. Aggregated switch bandwidth with uniform traffic.

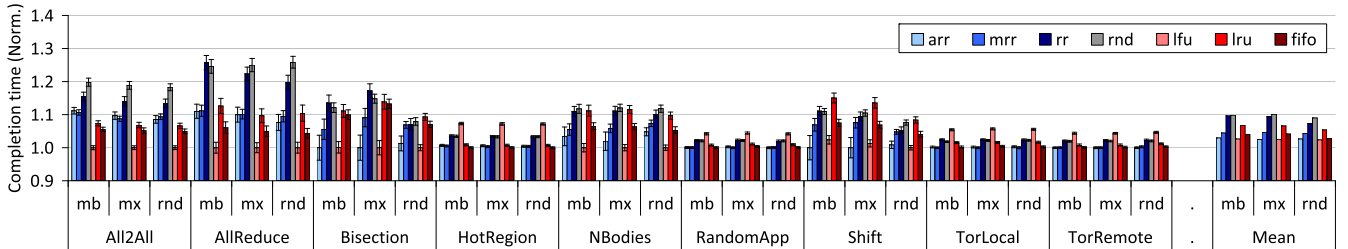


Fig. 5. Normalized communication time for different configurations – Routing algorithms.

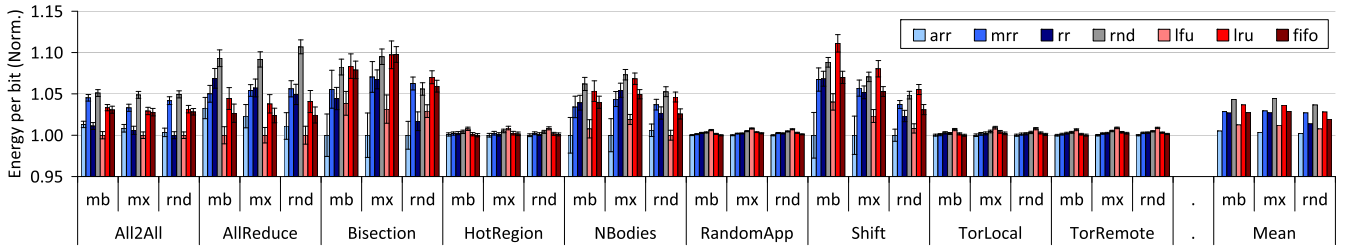


Fig. 6. Normalized energy-per-bit for different configurations – Routing algorithms.

irregular workloads (HotRegion, RandomApp, TorLocal, TorRemote) are the ones where lower differences can be observed. This is because the critical path of all tasks is different and the algorithms we are investigating are not capable of detecting and optimizing this. HotRegion and TorRemote are of particular interest as their resource usage is highly unbalanced, sending a disproportionate amount of traffic to the hot region and the uplink ports, respectively. This means that fair arbitration could be counterproductive, as traffic addressed to these bottlenecks may be blocked by traffic directed to other areas. Although some of the arbitration policies investigated here are capable of achieving some small benefits for unbalanced scenarios, there seems to be room for further improvement through specific arbitration techniques based, for instance, on learning the critical paths of applications or giving priority to traffic going to the most heavily loaded areas.

With respect to the relative performance of arbitration policies, LFU generally supports the fastest execution, sometimes with wide margins of over 25%. This is reasonable since it provides the fairest sharing of resources which can be highly

beneficial for regular workloads. As an example, Fig. 7(a) shows the timeline of execution of LRU, where transmission slots are distributed fairly among tasks. With irregular workloads, however, providing fair use of resources is far from the best strategy and LFU generally produces the worst results.

Regarding the round-robin based policies, we can see that the standard RR does not perform very well. The reason is that, by increasing the index one at a time, short periods of starvation occur. For example, on a 16-port switch, if the index is 0, port 15 will be the last to be serviced, so it is very likely to be blocked. The next round, the index will be 1 and port 15 will be the penultimate port to be serviced, and still likely to be blocked. The probability of being allocated a path increases in every arbitration round, but it remains small for a few more rounds of arbitration (notice the diagonal red stripe in Fig. 7(b), where most tasks are blocked for many consecutive rounds). ARR and MRR, reduce this effect by providing faster shuffling of the index, which in turn leads to a more efficient interleaving of flows. ARR provides the best results among the RR variants and is, indeed, the best policy for irregular workloads.

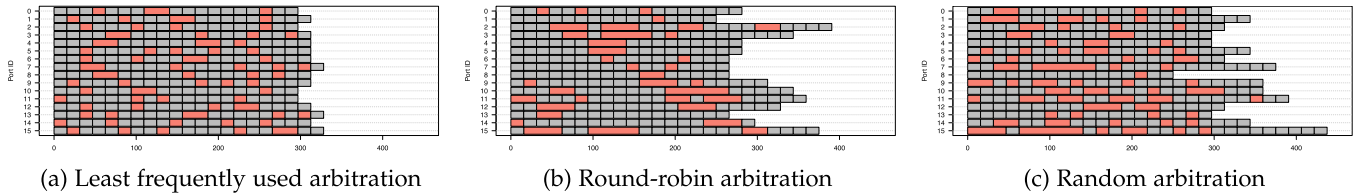


Fig. 7. Timeline of All2All using rnd routing. Time (μs) flows from left to right. Gray: transmitting. Red: blocked.

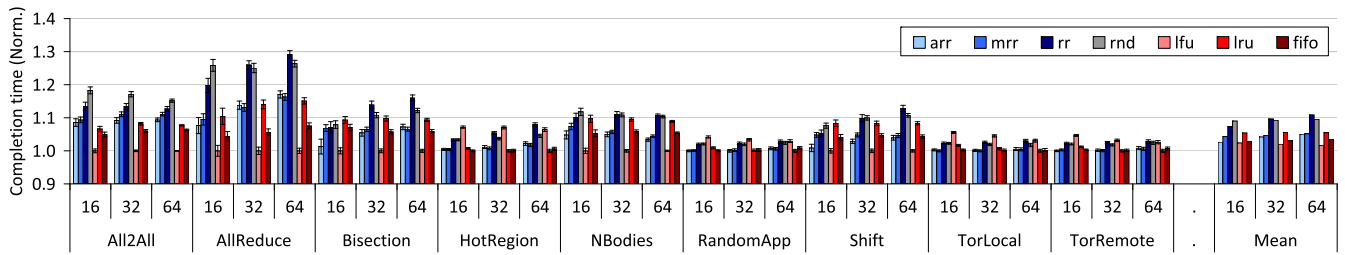


Fig. 8. Normalized communication time for different configurations – Switch radix.

Finally, while RND is expected to provide fair bandwidth sharing, it is actually one of the worst performing policies. This occurs because RND provides fairness in the long-run, but not in the short term. This is similar to what occurs with RR and has a negative impact on the overall performance. As port order is chosen at random, it is likely that there are one or multiple ports that, by chance, get serviced among the last ones in several consecutive rounds of arbitration. Therefore, it is very unlikely that they are allocated a path, which effectively means they are suffering from starvation. See Fig. 7(c), where ports 15, 11 and 7 have very high blocking ratios (46.4%, 40% and 37.5%, respectively). In contrast, port 8 is the luckiest and is only blocked 6% of the time.

Fig. 6 shows the energy-per-bit results where we can see that the impact of arbitration on energy efficiency is less substantial but still significant with the largest differences being around 10%. It is also noticeable that there is a general correlation between communication time and energy-efficiency results. There are some exceptions to that correspondence: for Bisection and Shift and, to a lesser extent, NBodies, the relation between energy and communication time is magnified when compared with other workloads. Also with All2All, RR has a long execution time but, comparatively, a remarkably low energy consumption. These anomalies would require further investigation which is outside the scope of this paper but, for our purposes, it suffices to note that they happen in all routing schemes. If we focus on the arbitration policies we can see that ARR provides the lowest energy consumption.

Finally, it is also worth mentioning that the results for all routing algorithms are very similar; no matter what routing was used, the workloads that benefit the most from arbitration are the same, and the benefits obtained are analogous. The similarity of these results is unexpected, as these routing algorithms are rather different in nature. However, this is a beneficial feature for the design and implementation of Beneš photonic switches, as it

suggests that flow routing and port arbitration can be engineered independently. For the sake of brevity, the remaining discussion will concentrate on random path routing since it features the smallest variability, i.e., it has the tightest confidence intervals.

C. Switch Radix

We now discuss how the performance of the arbitration policies scales with switch radix. Figs. 8 and 9 show the results for communication time and energy-efficiency, respectively, for the investigated switch radices (16, 32 and 64 ports).

The first thing to notice is that the results remain relatively consistent across different switch fabric scales. We found that as the number of ports increases, the observed differences in performance across arbitration policies increase slightly but, overall, the relative merit of each policy is similar for all tested radices. This was expected because the general structure of the workloads is maintained regardless of the number of communicating tasks. We can find some differences in performance when comparing ARR versus MRR arbitration. With 16 ports, ARR was able to significantly outperform MRR in terms of communication time. However, as we increase the number of ports, differences in communication times become insignificant. Energy-wise, ARR keeps being the most efficient policy.

Performance metrics also vary with radix in some configurations with the NBodies workload; communication times with MRR get significant improvements from 16 ports to 32, which remain when scaling further to 64 ports. The reason for this is that the causality inherent to the workload increases with the number of communicating nodes. This translates into longer dependency chains among flows which, as discussed above, means that the delays incurred due to flows getting blocked add up and, subsequently, communication time increases. Longer chains translate into larger differences and fair arbitration policies such as the ones above can extract larger benefits. An analogous effect can

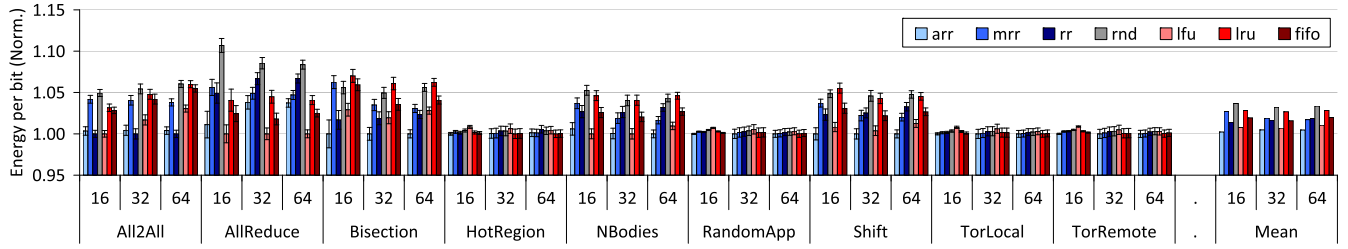


Fig. 9. Normalized energy-per-bit for different configurations – Switch radix.

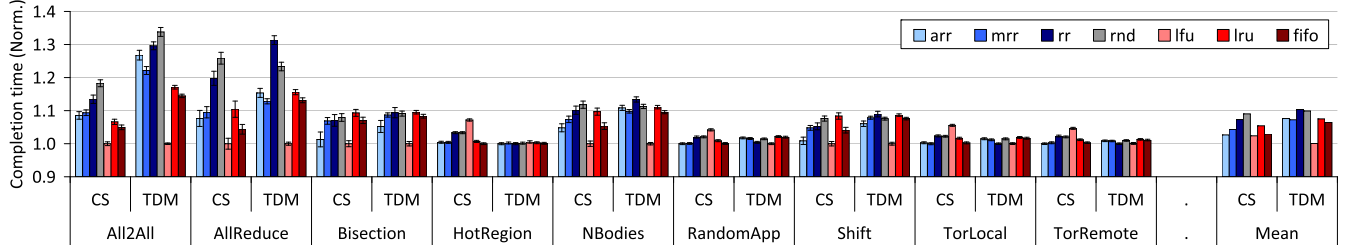


Fig. 10. Normalized communication time for different configurations – Switching methods.

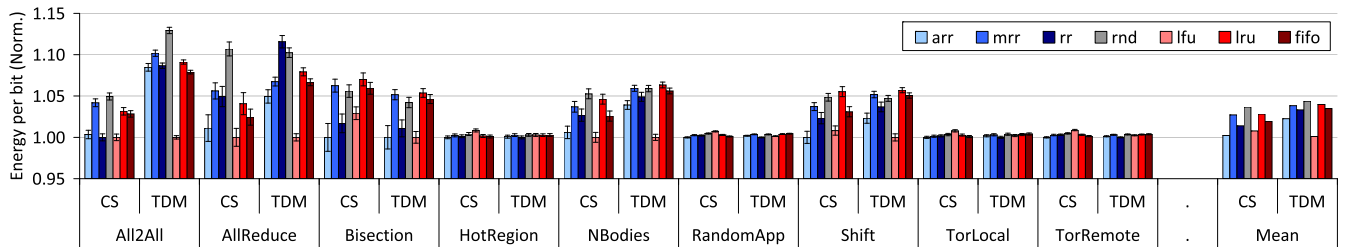


Fig. 11. Normalized energy-per-bit for different configurations – Switching methods.

be seen for AllReduce, albeit to a lesser extent. At any rate, it is clear that, in terms of execution speed, LFU seems to scale better than the other policies as its lead increases with switch scale. In contrast, RR shows the worst performance scalability. This is due to the length of the starvation periods which increases with the number of ports which suggests that unfairness grows with switch size.

D. Switching Method

Finally, we assess how the performance of arbitration policies changes with the switching method (CS versus TDM). Figs. 10 and 11 present communication time and energy efficiency, respectively. While most of the trends explained in the previous subsections still remain, we observe a tighter relation between arbitration and switching method.

For example, for the All2All and AllReduce workloads, the relative merits (reductions in communication times and also in energy per bit) of the different arbitration policies vary substantially between the switching methods. Differences between arbitration policies are much larger for TDM than for CS when using workloads that have dependency chains. The effects of causality are exacerbated by segmenting the flows into TDM slots as all segments of a flow need to be received in order

to trigger dependent flows. Thus causality delay is added to the extra delay derived from flow interleaving, which renders fair arbitration particularly important for TDM scenarios. In contrast, with TDM the choice of arbitration has a smaller impact for the irregular workloads (HotRegion, RandomApp, TorLocal, TorRemote). Indeed, LFU, which offered the worst results with CS for these workloads, is very competitive and ARR, which obtained the best results, performs worse with TDM.

In general, we conclude that the choice of switching method has a more noticeable effect on the performance of arbitration policies than other aspects of the architecture. This is because TDM essentially changes the granularity at which arbitration is conducted. Even so, the leading arbitration policy is still LFU both in terms of faster execution and higher energy efficiency. These results pave the way to future research on specific arbitration policies for TDM switches.

VI. CONCLUSION

We have presented a comprehensive simulation-based evaluation of the impact of arbitration policies for photonic ToR switches. Our experimental methodology harnesses the characterization data of a recently manufactured 16×16 Beneš prototype based on MZIs, but the impact of arbitration would

also be a factor for other architectures and devices. In particular we have evaluated different figures of merit: switch throughput, communication time, insertion loss and energy per bit. The results have revealed that the effect of the arbitration policies is consistent across routing algorithms and switch radices, with performance variations slightly increasing with size. Conversely, we found a closer relation between arbitration and switching method, as the behavior of the tested arbitration policies clearly differed between TDM and CS configurations. The reason for this is that TDM implies a finer-grain arbitration. With TDM, the effects of arbitration policies have been generally more noticeable for regular workloads, at the expense of being barely appreciable for irregular workloads, when compared with CS. With regards to the impact on different metrics, we have found that communication time is the most sensitive to arbitration, with differences among policies around 10% and a few cases where they can be over 30%. The impact on energy efficiency was less significant, with typical differences of 4–5% and a few cases maxing at around 12–13%. Finally, the impact on ILoss was found to be insignificant, which suggests laser power is barely affected by arbitration.

Policy-wise, we have found that LFU is the policy obtaining the lowest average runtime, and also features low energy-per-bit in all tested configurations. LFU particularly excels with regular workloads and can greatly outperform other policies as it achieves the highest level of fairness. However, we found that with irregular workloads it fails to distribute traffic appropriately and is one of the worst performing. Another interesting finding is that RR, one of the most common arbitration policies, produces very poor performance metrics. We identified the reason for this to be the standard port selection, which tends to cause short periods of starvation, so two related policies with improved port selection mechanisms were proposed. In particular, when compared with RR our first proposal, MRR, can reduce execution time up to a 12% and, on average, this reduction is a 4%. Our second proposal, ARR performs even better with reductions of up to a 15% and an average of 6%. Indeed, ARR achieves comparable performance to LFU, but has the best performance with irregular workloads, and features the lowest energy in most cases and on average. Therefore, there exists a trade-off between LFU and ARR based on the application mix that is to be executed in the system. If it is expected that a majority of the workloads are irregular ARR might be preferable, whereas with a majority of regular applications LFU should be the policy of choice.

As future work we plan to explore the impact of arbitration in other PSF designs based on different photonics devices and topologies. In addition, we aim to analyze in more detail the effects that arbitration policies may have on highly unbalanced workloads such as HotRegion and TorRemote. This has the potential to lead to specific arbitration algorithms for such workloads. Two algorithm favors seem of particular relevance: First, we will investigate priority-based algorithms that prioritize traffic going towards the most heavily loaded ports. A second approach is to apply learning algorithms capable of identifying and prioritizing flows that are part of the critical path. This second approach has the benefit of being more general and, in principle, amenable to all possible kinds of workloads.

REFERENCES

- [1] S. Werner, J. Navaridas, and M. Luján, “A survey on optical network-on-chip architectures,” *ACM Comput. Surv.*, vol. 50, no. 6, pp. 89:1–89:37, 2017.
- [2] D. Thomson et al., “Roadmap on silicon photonics,” *J. Opt.*, vol. 18, no. 7, 2016, Art. no. 073003.
- [3] E. Bernier et al., “Switches and routing for on-chip photonic networks,” in *Proc. 24th OptoElectronics Commun. Conf. Int. Conf. Photon. Switching Comput.*, 2019, pp. 1–3.
- [4] L. Lu et al., “16×16 optical switch based on electro-optic Mach-Zehnder interferometers,” *Opt. Exp.*, vol. 24, no. 9, pp. 9295–9307, 2016.
- [5] N. Dupuis and B. Lee, “Impact of topology on the scalability of Mach-Zehnder-based multistage silicon photonic switch networks,” *J. Lightw. Technol.*, vol. 36, no. 3, pp. 763–772, Feb. 2018.
- [6] H. Gu, Z. Wang, B. Zhang, Y. Yang, and K. Wang, “Time-division-multiplexing-wavelength-division-multiplexing-based architecture for ONoC,” *J. Opt. Commun. Netw.*, vol. 9, no. 5, pp. 351–363, May 2017.
- [7] E. Harstead, D. van Veen, V. Houtsma, and P. Dom, “Technology roadmap for time-division multiplexed passive optical networks (TDM PONs),” *J. Lightw. Technol.*, vol. 37, no. 2, pp. 657–664, Jan. 2019.
- [8] M. Kynigos et al., “Power and energy efficient routing for Mach-Zehnder interferometer based photonic switches,” in *Proc. ACM Int. Conf. Supercomputing*, 2021, pp. 177–189.
- [9] D. Opferman and N. Tsao-Wu, “On a class of rearrangeable switching networks Part I: Control algorithm,” *Bell Syst. Tech. J.*, vol. 50, no. 5, pp. 1579–1600, 1971.
- [10] Q. Cheng, M. Bahadori, and K. Bergman, “Advanced path mapping for silicon photonic switch fabrics,” in *Proc. Conf. Lasers Electro-Opt.*, 2017, pp. 1–2.
- [11] M. Kynigos et al., “On the routing and scalability of MZI-based optical beneš interconnects,” *Nano Commun. Netw.*, vol. 27, 2021, Art. no. 100337.
- [12] P. Yuen and L. Chen, “Optimization of microring-based interconnection by leveraging the asymmetric behaviors of switching elements,” *J. Lightw. Technol.*, vol. 31, no. 10, pp. 1585–1592, May 2013.
- [13] R. Yao and Y. Ye, “Toward a high-performance and low-loss clos-beneš-based optical network-on-chip architecture,” *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 39, no. 12, pp. 4695–4706, Dec. 2020.
- [14] C. Kachris and I. Tomkos, “A survey on optical interconnects for data centers,” *IEEE Commun. Surv. Tut.*, vol. 14, no. 4, pp. 1021–1036, Fourth Quarter 2012.
- [15] Q. Yang et al., “WDM/TDM optical-packet-switched network for supercomputing,” in *Proc. Opt. Comput.*, 2000, pp. 555–561.
- [16] S. Yan et al., “Archon: A function programmable optical interconnect architecture for transparent intra and inter data center SDM/TDM/WDM networking,” *J. Lightw. Technol.*, vol. 33, no. 8, pp. 1586–1595, Apr. 2015.
- [17] A. A. M. Saleh, A. S. P. Khope, J. E. Bowers, and R. C. Alferness, “Elastic WDM switching for scalable data center and HPC interconnect networks,” in *Proc. OptoElectronics Commun. Conf. Int. Conf. Photon. Switching*, 2016, pp. 1–3.
- [18] S. Werner, J. Navaridas, and M. Luján, “Subchannel scheduling for shared optical on-chip buses,” in *Proc. IEEE 25th Annu. Symp. High-Perform. Interconnects*, 2017, pp. 49–56.
- [19] G. Hendry et al., “Silicon nanophotonic network-on-chip using TDM arbitration,” in *Proc. IEEE Symp. High Perform. Interconnects*, 2010, pp. 88–95.
- [20] M. Kynigos, J. Navaridas, J. Pascual, and M. Lujan, “A novel simulation methodology for silicon photonic switching fabrics,” in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw.*, Raleigh, NC, 2023, pp. 114–123.
- [21] A. Jain, R. Bahl, and A. Banik, “Demonstration of RZ-OOK modulation scheme for high speed optical data transmission,” in *Proc. IFIP Int. Conf. Wireless Opt. Commun. Netw.*, 2014, pp. 1–5.
- [22] K. Ogawa, Y. Sangenya, M. Morikura, K. Yamamoto, and T. Sugihara, “IEEE 802.11ah based M2M networks employing virtual grouping and power saving methods,” in *Proc. IEEE 78th Veh. Technol. Conf.*, 2013, pp. 1–5.
- [23] L. Chen and N. Chrysos, “Throughput of random arbitration for approximate matchings,” in *Proc. IEEE/ACM 6th Symp. Architectures Netw. Commun. Syst.*, 2010, pp. 1–2.
- [24] J. Navaridas et al., “INRFLOW: An interconnection networks research flow-level simulation framework,” *J. Parallel Distrib. Comput.*, vol. 130, pp. 140–152, 2019.

- [25] A. Greenberg et al., "VL2: A scalable and flexible data center network," in *Proc. ACM SIGCOMM Conf. Data Commun.*, 2009, pp. 51–62.
- [26] R. Thakur and W. Gropp, "Improving the performance of collective operations in MPICH," in *Proc. Eur. Parallel Virtual Mach./Message Passing Interface Users' Group Meeting*, 2003, pp. 257–267.
- [27] T. Hoefler, T. Schneider, and A. Lumsdaine, "Multistage switches are not crossbars: Effects of static routing in high-performance networks," in *Proc. IEEE Int. Conf. Cluster Comput.*, 2008, pp. 116–125.
- [28] X. Yuan, S. Mahapatra, W. Nienaber, S. Pakin, and M. Lang, "A new routing scheme for jellyfish and its performance with HPC workloads," in *Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal.*, 2013, pp. 1–11.
- [29] W. J. Dally and B. P. Towles, *Principles and Practices of Interconnection Networks*. Amsterdam, The Netherlands: Elsevier, 2004.
- [30] C. Seitz, "The cosmic cube," *Commun. ACM*, vol. 28, pp. 22–33, 1985.
- [31] S. Kandula et al., "The nature of data center traffic: Measurements & analysis," in *Proc. 9th ACM SIGCOMM Conf. Internet Meas.*, 2009, pp. 202–208.
- [32] J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," in *Proc. Int. Symp. Comput. Archit.*, 2008, pp. 77–88.
- [33] T. Benson, A. Akella, and D. Maltz, "Network traffic characteristics of data centers in the wild," in *Proc. 10th ACM SIGCOMM Conf. Internet Meas.*, 2010, pp. 267–280.
- [34] S. Zaheer et al., "Locality-aware process placement for parallel and distributed simulation in cloud data centers," *J. Supercomputing*, vol. 75, pp. 7723–7745, 2019.
- [35] R. Fujimoto, "Research challenges in parallel and distributed simulation," *ACM Trans. Model. Comput. Simul.*, vol. 26, no. 4, 2016, Art. no. 22.



Javier Navaridas received the MEng and PhD degree from the University of the Basque Country, where his PhD thesis received an Extraordinary Doctorate award (Top 16 out of 306 theses). He is a Ramón y Cajal research fellow with the University of the Basque Country. The Ramón y Cajal fellowships are highly competitive and he ranked 3rd out of 107 candidates in the ICT area. Previously he was a Royal Society Newton Research fellow, a lecturer and a senior research fellow with The University of Manchester. During his PhD he was a visiting researcher with

the Interconnect Fabrics group with IBM Zürich Research Labs. He has a broad experience within the areas of networking, computer architecture and system modeling as demonstrated by his extensive list of publications (more than 75 papers).



Markos Kynigos received the MSc and PhD degrees in computer science from the University of Manchester. After a period as a post-doctoral researcher within the Advanced Processor Technologies Group of the University of Manchester, he is currently employed by Huawei U.K. R&D. His research interests include high-performance computing, optical interconnects, optical switching through silicon photonics, and energy efficiency within the scope of computer architecture.



Jose A. Pascual received the MEng and PhD degrees in computer science from the Department of Computer Architecture and Technology of the University of the Basque Country UPV/EHU, Spain. After a period as a postdoctoral researcher with The University of Manchester. He is currently a lecturer in operating systems and networks with The University of the Basque Country. His research interests include high performance computing, scheduling for parallel supercomputers, and performance evaluation of parallel systems.



Mikel Luján received the PhD degree in computer science from The University of Manchester, U.K., in 2002. He is currently a professor with the Department of Computer Science, The University of Manchester, where he holds the ARM/Royal Academy of Engineering Research Chair on Computer Systems. His research interests include runtime environments, low-power architectures, and application-specific systems and optimizations.



Jose Miguel-Alonso received the graduated degree in computer science, and the PhD degree both from the University of the Basque Country UPV/EHU in Spain, in 1989 and 1996, respectively. He is a full professor with the Department of Computer Architecture and Technology of the UPV/EHU, and a member of the Intelligent Systems Group of this university. He carries out research related to networks and parallel-distributed systems, in areas such as cybersecurity, performance modeling (with focus on the interconnection network), resource management

in supercomputers, cloud infrastructures and high-performance scientific and technical applications. He has published two books, 33 journal articles and 30 papers in international conferences. He is a member of the IEEE Computer Society and the HiPEAC Network of Excellence on High Performance and Embedded Architecture and Compilation.



John Goodacre received the bachelor's degree in computer science from the University of New York, New York, in 1987. He is a professor of computer architectures with the Department of Computer Science, The University of Manchester, Manchester, U.K., having previously spent 17 years with Arm Ltd. as the director of technology and systems where he defined and introduced the first multicore processors and other core technologies. He is also appointed by the U.K. government's Research and Innovation agency as the director of the Digital Security by

Design Challenge Fund, a £200 million programme to enable industry and researchers to create a step change in approach to cybersecurity, blocking vulnerabilities by design, and protecting the operation and data by default. His research interests include web-scale servers, exascale efficient systems, and secure and ubiquitous computing.