



Power and energy efficient routing for Mach-Zehnder interferometer based photonic switches

DOI:
[10.1145/3447818.3460363](https://doi.org/10.1145/3447818.3460363)

Document Version
Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

Kynigos, M., Pascual, J., Navaridas, J., Goodacre, J., & Luján, M. (2021). Power and energy efficient routing for Mach-Zehnder interferometer based photonic switches. In *ICS 2021 - Proceedings of the 2021 ACM International Conference on Supercomputing* (pp. 177-189). (Proceedings of the International Conference on Supercomputing). Association for Computing Machinery. <https://doi.org/10.1145/3447818.3460363>

Published in:
ICS 2021 - Proceedings of the 2021 ACM International Conference on Supercomputing

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact openresearch@manchester.ac.uk providing relevant details, so we can investigate your claim.



Power and Energy Efficient Routing for Mach-Zehnder Interferometer based Photonic Switches

Markos Kynigos
Department of Computer Science
The University of Manchester
Manchester, United Kingdom
markos.kynigos@manchester.ac.uk

Jose A. Pascual
The University of the Basque Country
San Sebastian, Spain

Javier Navaridas
The University of the Basque Country
San Sebastian, Spain

John Goodacre
Department of Computer Science
The University of Manchester
Manchester, United Kingdom

Mikel Luján
Department of Computer Science
The University of Manchester
Manchester, United Kingdom

ABSTRACT

Silicon Photonic top-of-rack (ToR) switches are highly desirable for the datacenter (DC) and high-performance computing (HPC) domains for their potential high-bandwidth and energy efficiency. Recently, photonic Beneš switching fabrics based on Mach-Zehnder Interferometers (MZIs) have been proposed as a promising candidate for the internals of high-performance switches. However, state-of-the-art routing algorithms that control these switching fabrics are either computationally complex or unable to provide non-blocking, energy efficient routing permutations.

To address this, we propose for the first time a combination of energy efficient routing algorithms and time-division multiplexing (TDM). We evaluate this approach by conducting a simulation-based performance evaluation of a 16x16 Beneš fabric, deployed as a ToR switch, when handling a set of 8 representative workloads from the DC and HPC domains.

Our results show that state-of-the-art approaches (circuit switched energy efficient routing algorithms) introduce up to 23% contention in the switching fabric for some workloads, thereby increasing communication time. We show that augmenting the algorithms with TDM can ameliorate switch fabric contention by segmenting communication data and gracefully interleaving the segments, thus reducing communication time by up to 20% in the best case. We also discuss the impact of the TDM segment size, finding that although a 10KB segment size is the most beneficial in reducing communication time, a 100KB segment size offers similar performance while requiring a less stringent path-computation time window. Finally, we assess the impact of TDM on path-dependent insertion loss and switching energy consumption, finding it to be minimal in all cases.

CCS CONCEPTS

- **Hardware** → Emerging optical and photonic technologies;
- **Networks** → Bridges and switches; Data center networks.

KEYWORDS

Top-of-Rack, Photonic Switches, Mach-Zehnder Interferometers, TDM, Routing

ACM Reference Format:

Markos Kynigos, Jose A. Pascual, Javier Navaridas, John Goodacre, and Mikel Luján. 2021. Power and Energy Efficient Routing for Mach-Zehnder Interferometer based Photonic Switches. In *2021 International Conference on Supercomputing (ICS '21)*, June 14–17, 2021, Virtual Event, USA. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3447818.3460363>

1 INTRODUCTION

All-optical interconnects (OINs) based on silicon photonics are a promising emerging technology for scaling datacenter (DC) and high-performance computing (HPC) interconnects. Many research demonstrations have been produced frequently for all levels of interconnection network, or IC (hereafter, IC refers to on-chip, inter-chip, board level, top-of-rack and L2/L3 network tiers). The fabrication platform's CMOS compatibility, combined with the large data density in links due to wavelength division multiplexing (WDM), the low propagation latency inherent to photonics as well as low energy consumption relative to link distance [30] make silicon photonics a viable candidate for augmenting conventional ICs.

Although optical networks have been present since the late 1980s, there are still many challenges to developing and deploying efficient, all-optical (i.e. photonic) IC systems. For instance, while some attempts have been made, it is currently not possible to efficiently store light in optical form for practical amounts of time [3], making photonic buffering a non-option. As such, ICs that employ optical technology must rely on circuit switching (CS) techniques at the transmission level to remain photonic, or suffer conversion to the electric domain at every hop, which increases energy consumption substantially and detracts for the benefits of optical links.

This presents interesting research challenges for the whole OIN and especially for photonic switches, as CS may lead to contention in the fabric which reduces overall performance. Furthermore, physical level characteristics of the photonic components which form switches, e.g. insertion loss (hereafter ILoss) and crosstalk, increase required laser power prohibitively, resulting in scalability challenges. Therefore, switch design must aim towards reducing these metrics to avoid excessive energy consumption, which justifies the use of multi-stage fabrics such as the Beneš network.

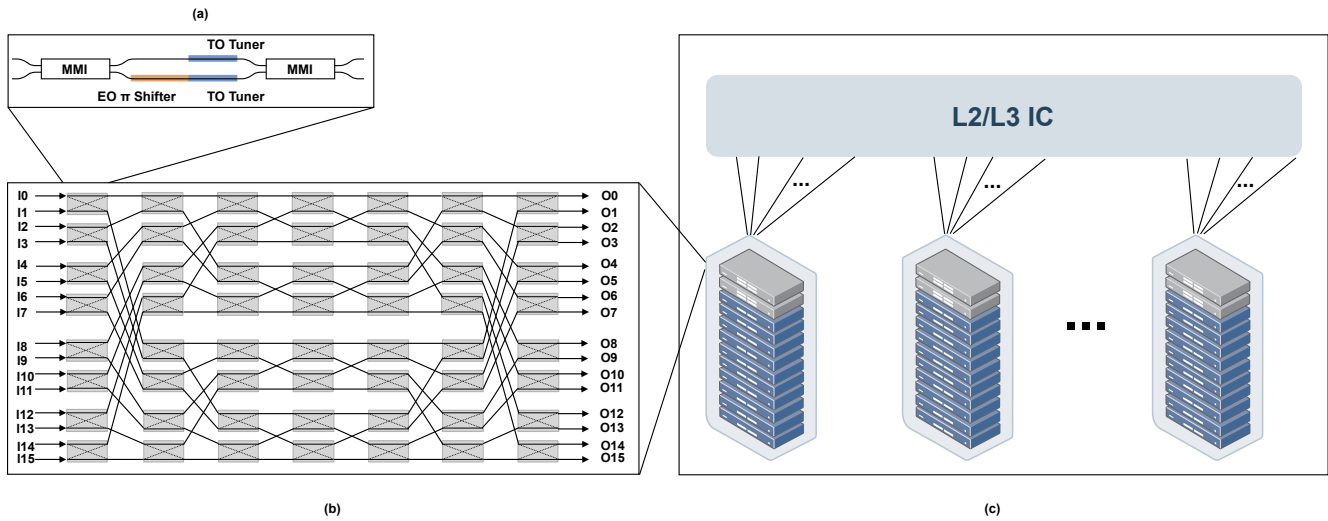


Figure 1: (a) Schematic of a 2×2 EO/TO MZI Switch. (b) A Beneš-based ToR switch formed with MZI switches (c) A hypothetical photonic interconnect with photonic ToR switches.

The Beneš network is a rearrangeably non-blocking topology composed of the least amount of 2×2 switches necessary to connect $N \times N$ endpoints, leading to the least optical loss when using photonic 2×2 switches such as Mach-Zehnder Interferometers (MZIs). However, standard network control algorithms such as the “Looping Algorithm” [23] are unable to provide energy and power efficient configurations for photonic Beneš fabrics, while algorithms that do so, such as “hardware-inspired routing”, introduce contention in the switch fabric [16].

Switch fabric contention in rearrangeably non-blocking networks is when a connection from a source to a destination cannot be established due to other connections being serviced. Hereafter, we refer to connections as flows. Switch fabric contention is different to output contention, where multiple flows attempt to access the same output. As section 3.2 explains, although these networks can serve any permutation, switch fabric contention may occur when flows are serviced incrementally.

This work addresses this problem by presenting for the first time a combination of time division multiplexing and energy-efficient hardware-inspired routing, which we propose as the control mechanism for a recently fabricated and characterized 16×16 photonic Beneš switch fabric formed with thermally-electrically tuned MZIs [20], deployed as a top-of-rack (ToR) switch. This approach partitions the flows into segments and provides energy efficient configurations to service flow segments, while at the same time alleviating the effects of switch fabric contention in the Beneš network. We evaluate our approach through simulation, employing 8 realistic and synthetic workloads from the DC and HPC domains.

Our contributions are as follows:

- We investigate the prevalence of switch fabric contention when using circuit switching (CS) and previously proposed hardware-inspired routing algorithms, finding that it can be as high as 23% for the heaviest workloads.

- We present and evaluate a combination of TDM and hardware-inspired routing algorithms, showing communication time reductions up to 20% in the best case.
- We assess the impact of flow segment size and observe that around 100KB is most beneficial, as decreasing the size further offers diminishing improvements (at most 3%).
- We assess the impact of TDM on critical-path ILoss and switching energy, finding it to be minimal.

To our knowledge, this is the first simulation-driven evaluation of TDM for photonic EO/TO MZI-based Beneš ToR switches grounded on a fabricated device.

2 BACKGROUND & RELATED WORK

2.1 Opportunities for Photonic Switching

Modern DC and HPC deployments currently adopt electrical packet switches based on Infiniband or Ethernet at all layers of the DC interconnect, including the ToR level, with optical transmission being relegated to optical links. There exists a large variety of commercial DC switches, featuring various radices, switching capacities and form factors; however these switches can be extremely power-hungry [24]. For example, the Arista 7368X4 Series switch offers up to 128 ports of 100GbE (32 port 400GbE); however, the average power consumption reported is $\sim 961\text{W}$ excluding optics or cables and the peak consumption rises to $\sim 1998\text{W}$ assuming 4.5W CWDM optics [19]. Conversely, the deployment we investigate considers using dense-wavelength division multiplexing with 32 wavelengths modulated at low data rates, which can lead to $\leq 0.1\text{W}$ per port in required laser power for traversing the switch (not including coupling losses) [17] combined with an MZI tuning power requirement of $\sim 1\text{W}$, this can lead to a substantial reduction of power requirements for a ToR switch, motivating for the evolution of photonic ToRs such as the one we examine here.

Additionally, electrical switches must either be upgraded at every data rate generation to support new transceivers, or transceivers must remain constrained by legacy capabilities, thereby increasing costs. Photonic switches, on the other hand, have the potential of ameliorating these costs and accommodating future data rates more easily, as their performance is less dependent on per-wavelength data rates and number of wavelengths. This is especially the case with MZI-based switching fabrics, as MZIs can provide broadband switching at ns switching time. However, using photonic switching fabrics still entails many challenges which we discuss below.

2.2 MZI-based Optical Switching Fabrics

Over the past decade, various proposals for MZI-based switching fabrics have been produced targeting different levels of IC. Li et al. [18] and more recently Cheng et al. [4] provide comprehensive reviews of silicon photonic technology for DC interconnects. A large number of the cited proposals concern Beneš-based switch fabrics with MZIs in both works. MZI-based approaches have also been formulated for the on-chip domain, e.g. [36] or [10]. MZI-based switch fabrics organized in either the Beneš or dilated-Beneš topologies (e.g. [25], [26], [6] or [7]) have been recently demonstrated. These demonstrations with sizes of $16\times$ and $32\times$ fabrics surmount many of the technological and fabrication challenges associated with increasing the switch radix, showing promise for adoption in the medium term. However, many challenges, such as decreasing optical losses (ILoss and crosstalk) or optimal communication arbitration and routing strategies, must be addressed before deployment can occur.

To address these challenges, Cheng et al. propose a path mapping strategy for $8\times$ Beneš fabrics which evaluates all potential states for a permutation; however this quickly becomes intractable as a Beneš network scales up [5]. Yuen and Chen [35] also propose a methodology for exploiting hardware asymmetries in MRR-based photonic ICs. In [16], we proposed routing algorithms which leverage the underlying hardware constraints of EO/TO MZI-based Beneš ICs, showing reductions in optical losses and switching energy. These are the algorithms we are leveraging with TDM to improve overall performance. ILoss, one of the main optical losses, is a defining factor for the required power of the laser beams which carry information through the switch; minimising this as well as switching energy can improve the total energy efficiency of a switch fabric and therefore of the whole interconnect. The algorithms aim to allocate paths that incur the least amount of ILoss from waveguide crossings and MZIs in the “bar” state; However, they may introduce switch fabric contention in the network as they do not guarantee non-blocking operation. For this reason, we analyse the effect of using a TDM scheme in order to allow a better sharing of network resources, while maintaining low ILoss and high energy efficiency.

2.3 Enhancing Optical Interconnects with TDM

This section focuses on describing the most relevant related work for Optical Interconnects with TDM. While optical IC systems have recently been the subject of thorough research (e.g. [29] [1] [15] [21]), the research on the deployment and the practical application of

MZI-based switching fabrics is quite novel, even when the technology is highly promising. For this reason, we were unable to find much research on routing or arbitration for this technology.

However, it is clear from the related work that other optical technologies tend to employ a combination of space-division multiplexing (SDM), TDM or WDM in order to maximize throughput and to use bandwidth fairly. A survey of different approaches can be found here [13]. In [32], an optical IC using SDM/TDM for intra-datacenter and WDM for inter-datacenter traffic is reported. They employ FPGA-based ToR switches that send traffic either through slotted-TDM/Ethernet or optical bandwidth variable transmitters (BVTs). TDM-based optical ICs have also been researched for supercomputing. In [33], the “Data Vortex” optical interconnect is used with a TDM/WDM routing function, while in [27] the authors motivate for a microring-based elastic crossbar switch which, when augmented with TDM, can be considered for both HPC and data centre use cases.

TDM arbitration has also been proposed within the optical network-on-chip (ONoC) domain. Werner et al. propose a mixed WDM-TDM approach for bus-based ONoCS [31] based on micro-ring resonators (MRRs). Hendry et al. employ MRR-based broadband nanophotonic switches organized in a mesh topology which, when coupled with a TDM arbitration scheme, show substantial efficiency gains with respect to both circuit-switched ONoCs and electronic equivalents [11]. In contrast to these works we examine for the first time a photonic Beneš ToR switch formed with EO/TO MZIs.

3 ADDING TDM TO AN MZI-BASED TOR SWITCH

3.1 Network Topology

The ToR switch we investigate is based on the demonstrated 16×16 photonic switch found in [20]. Fig. 1 visualises the structure of our envisioned system. In the top left, we show how the MZIs are composed from their constituent parts: Multi-Mode Interferometers, waveguides, thermal and electrical tuners. In the bottom left we represent the fabric organization based on 2×2 MZIs, including the waveguide crossings. Finally, in the right hand side there is an sketch of the deployment scenario, where the photonic ToR switch is connected to in-rack servers and to the higher tiers of the interconnect. The switch fabric is formed using thermally-electrically tuned MZIs organized in a Beneš network. As explained, this topology requires the fewest MZIs for a rearrangeably non-blocking switch fabric. Although a higher radix switch would be desirable and, indeed, should benefit even more from TDM, we select this size because a larger size may prohibitively increase first-order crosstalk as indicated in [8]. We consider a WDM scenario with 32λ , each modulated at 16Gb/s with an On-off Keyring (OOK) scheme [12], yielding a 512Gb/s aggregate bandwidth per port, with endpoints modulating on all λ simultaneously to reduce flow transmission time in the switching fabric.

3.2 Switch Fabric Contention

The Beneš network is a rearrangeably non-blocking network which means that, in principle, it is capable of servicing any connection permutation. However, when operating in CS, traffic is serviced incrementally which means that switch fabric contention for the use

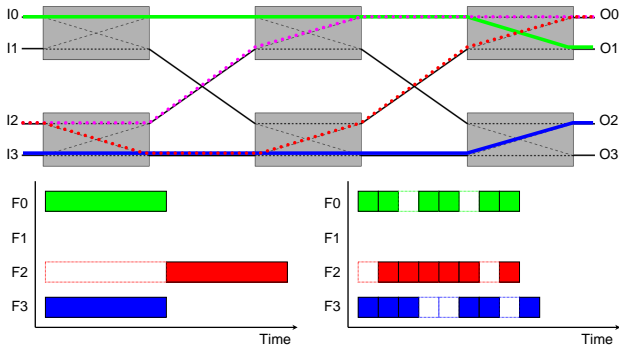


Figure 2: Top: Example of switch fabric contention in a small 4×4 Beneš network. The green, F0, and blue, F3, flows have a path allocated. A third flow, F2, from I2 to O0 arrives but can not be served because resources are busy. Bottom: Timeline of execution using CS (left) and TDM (right). With CS, F2 has to wait for the others, with TDM the transmission of all flows is interleaved.

of resources can not always be avoided. *Switch fabric contention* is the event in which a flow does not have any available path because resources are busy serving other flows. An example of switch fabric contention is shown in Fig. 2. In the figure, two flows (blue and green) have an allocated path when a third flow arrives to the switch. The new flow has two possible paths to its destination (red and magenta), but both paths require resources that are already allocated to the other flows. In a pure CS scheme, this means the new flow has to wait until any of the other flows completes transmission. With a TDM scheme, however, the 3 flows would be interleaved along time, resulting in a fairer utilization of network resources. This in turn results in a faster transmission of all flows and, arguably, in lower average latency and, more importantly for some applications, lower jitter.

One solution to switch fabric contention, known as the “Looping algorithm” [23], leverages the network’s symmetry in order to solve a permutation in $O(n \log n)$ time. However, a key component of the “Looping Algorithm” is its ability to rearrange connections in the presence of configurations that would cause switch fabric contention. In such cases, the algorithm reconfigures the switch states, thereby eliminating switch fabric contention and servicing the whole permutation. While this can be favourable in electrical networks which are buffered, the switch we examine is inherently *bufferless*; consequently, reconfiguration of the switch state requires either early termination of the flows or, ultimately, loss of data. To mitigate this, the “Looping Algorithm” could be augmented with TDM. In this approach, the fabric is reconfigured if necessary in each timeslot. However on the one hand, as detailed in section 3.4, timeslots are extremely short which would lead to excessive computation demands on the network controller. On the other hand, due to the nature of the algorithm, it is unable to take into account underlying hardware constraints such as ILoss. We also note that, while the “Looping algorithm” aims at solving full permutations with relatively low time complexity, most DC traffic does not fit a

perfect permutation explicitly. Our approach surmounts these challenges by using pre-computed routing tables and energy efficient routing algorithms.

3.3 Power Efficient Routing in Photonic Beneš Networks

Recently, the topic of exploiting hardware asymmetries to reduce laser power has gained traction in the photonic architecture community. This work builds upon this idea and shows that enhancing such functions with TDM switching can offer significant benefits in both execution time and energy efficiency. Hence we consider a subset of the hardware-aware routing strategies we proposed in [16]. Our objective is to demonstrate that the approach can be generalized and is independent of other switching aspects.

The routing algorithms operate under the following principle. For each source/destination pair, a routing table of size $N/2$ is constructed. N is the number of endpoints, a power of 2; for example, in Fig.3(a), $N = 8$. The routing table contains the entries for all potential paths. Each entry comprises of a routing signature (expressed as $\log \frac{N^2}{2}$ bits) plus scoring ranks which determine the priority of the path for each routing strategy. In each bit of a routing signature, a 0 denotes egress from the top port of a 2×2 MZI and a 1 from the bottom port. The first $\log \frac{N}{2}$ bits of the signature are a bit permutation, while the latter $\log N$ bits are used for the destination tag. For each path, the number of waveguide crossings and MZI states are calculated. The paths are subsequently sorted and ranked based on the minimisation criterion required by each routing strategy (e.g. fewest crossings, fewest MZIs in the “bar” state etc.). The rank of each path is then stored in the scoring rank fields. Fig.3 depicts all possible paths from I2 to O6 (a) and how these are encoded in

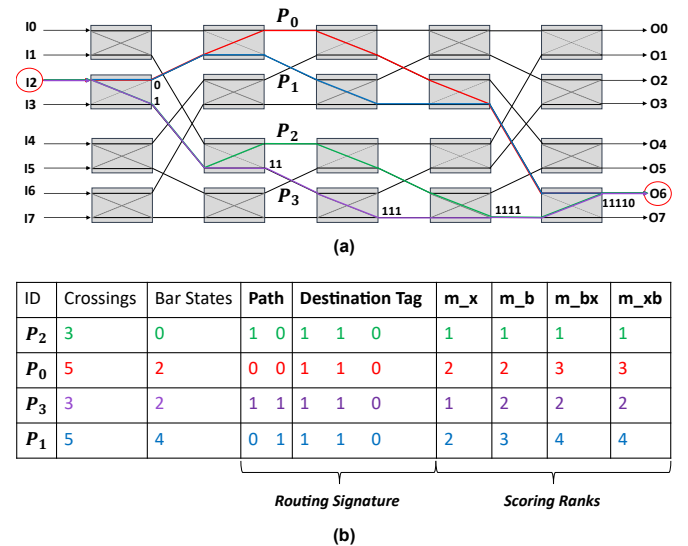


Figure 3: (a) Path diversity in an $8 \times$ Beneš network (in colour online), including routing signatures for P_4 . (b) Routing table for I2→O6. Paths are ranked by the “m_b” rank. The ID, Crossings and Bar States columns are added for convenience, but do not need to be stored in the routing table.

the routing table (b). The routing table depiction also includes the number of crossings and MZIs in the “bar” state for each path. For simplicity the figure assumes an $8 \times$ Beneš fabric.

For our analysis here, we consider the following strategies:

- **Minimise Crossings (m_x)** the ranking is based on the number of waveguide crossings.
- **Minimise “Bar” States (m_b)** the ranking is based on the number of MZIs in “bar” state.
- **m_{xb}** the ranking is based on the number of waveguide crossings and ties are broken by the number of MZIs in “bar” state.
- **m_{bx}** the ranking is based on the number of MZIs in “bar” state and ties are broken by the number of waveguide crossings.
- **Random Path (rnd)** selects a path randomly, without taking underlying hardware asymmetries into account.

The state of the switch fabric is constructed incrementally servicing requesting input ports one by one. For each port, the network controller checks the availability of paths based on the fabric state. If there is no available path, the controller stops the injection. Otherwise, it selects the available path with the lowest rank in the routing table, depending on the applied routing strategy. For instance, if the m_{bx} strategy is used, the order will be determined by the m_{bx} rank.

In our experimental work, we investigate how often flows suffer from switch fabric contention with CS, as a motivation for using TDM switching. In addition we show that the algorithms can be combined with TDM to mitigate the communication time penalty, while at the same time offering network configurations with reduced ILoss and therefore energy consumption.

3.4 TDM & Routing Implementation

Controlling a photonic MZI-based Beneš switch fabric is a non-trivial process. The MZIs we model require thermal tuning to reach a “cross” state and additional electrical tuning to reach a “bar” state, each of which takes time. As explained previously, thermal tuning requires time in the order of microseconds, while electrical tuning is substantially faster (ns scale). This is where electro-optical tuning becomes more advantageous than thermo-optical; if switching of MZI states from cross to bar state happens at the ns scale and all MZIs are switched simultaneously, the switch reconfiguration time overhead becomes small enough to be realistic for TDM. While this is barely relevant when using CS, it becomes essential with TDM. When using TDM, the required state of the switch at the next timeslot must be calculated within the time boundary of the current timeslot, which must complete before the network controller can issue the required power to the thermal or electrical MZI contacts. The tuning must then occur so that the switch acquires the state required to progress, and then the next timeslot’s communication may proceed. These constraints mean that the routing algorithm required to calculate the MZI states must run within a very strict time window.

To illustrate this, Table 1 shows the timeslot duration of each corresponding TDM segment size for various segment sizes, based on the aggregate data rate we target. In principle, shorter timeslots would allow for a fairer distribution of network resources to flows,

Table 1: TDM Segment Size & Slot Duration.

Segment size	Slot duration
10 KB	156ns
20 KB	312ns
40 KB	624ns
50 KB	780ns
100 KB	1.56 μ s
200 KB	3.12 μ s
500 KB	7.80 μ s

whereas larger timeslots are more prone to internal fragmentation. However, for shorter timeslots where the fabric must reconfigure more frequently, total tuning time would increase. As tuning time cannot be used for communication, a balance must be found between decreasing communication time and increasing tuning time penalty. In section 5.3, we conduct a parameter sweep over these segment sizes to evaluate the impact of this effect.

In our approach, we consider a centralised controller such as an FPGA or an ASIC for the switch fabric. The controller would generate and store pre-computed paths for the pairs requesting communication as detailed in section 3.3. As the Beneš network offers a path diversity of $N/2$ for each input-output pair, the state-space of the Beneš network scales exponentially. However, for the $16 \times$ variant we assess here, the topology’s symmetry can be exploited and combined with the routing strategies to reduce the memory footprint of the stored routing tables to the order of KB.

Since there are $2 \log N - 1$ MZIs per path for N input-output pairs simultaneously requesting access, the controller would have to accommodate $O(N^2(2 \log N - 1))$ comparisons of required versus current MZI state. This can be parallelised and also further optimised by eliminating paths that cannot be accessed, due to the state of the previous MZI; in Fig.3 for instance, if the top MZI in the second stage is already serving a flow and therefore in the “bar” state, P_1 need not be considered. This compute time, together with the memory access overhead, must be less than the TDM timeslot. For the scale of $16 \times$ endpoints, this worst-case computation overhead can be accommodated by current FPGA systems and even more easily by an ASIC. However, this is a research question in itself and out of scope for this work.

4 EXPERIMENTAL METHODOLOGY

4.1 Simulator & Model

We use PhINRFlow (Photonic Interconnection Network for Research Flow-level Simulation Framework), an in-house developed flow-level simulator dedicated to photonic interconnects. This simulator affords a light footprint, is highly scalable and includes the main technological aspects necessary for modelling photonic interconnects based on MZI switches. Additionally, the simulator includes a variety of workloads which emulate the behaviour of real applications. These capabilities enable us to evaluate the system under realistic loads, giving us insight to its viability as a ToR switch. The simulator inherits functionality from INRFlow [22], wherein a detailed description of the simulator’s methodology, organisation and workloads may be found. We model the ToR switch

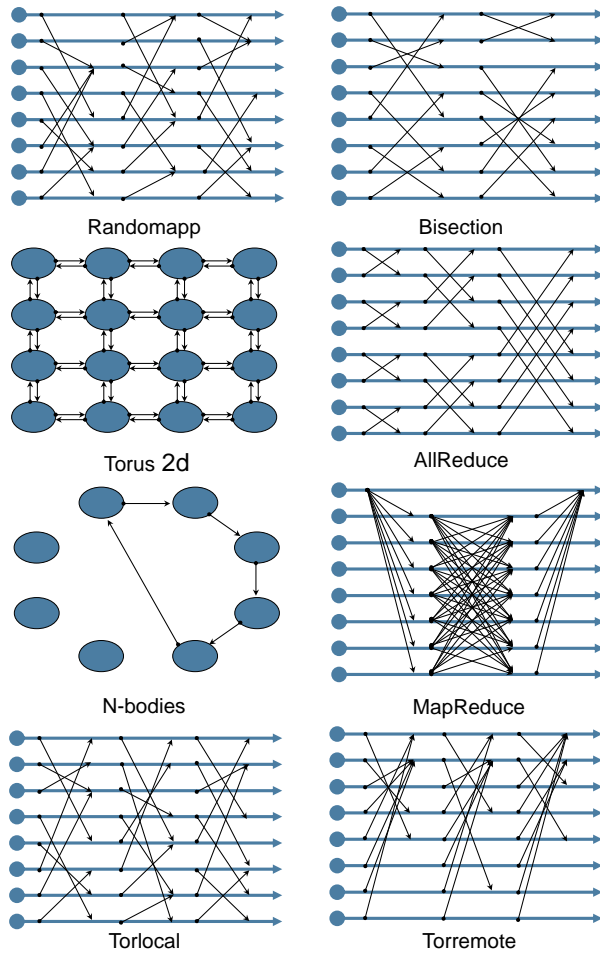


Figure 4: Schematic representation of the eight workloads used, shown with eight endpoints.

as a new topology, with unidirectional links and traffic flowing from “left” to “right”. Each endpoint connected to the ToR switch is modelled as a simple traffic producer/consumer node.

We evaluate two use-cases for this switch fabric: a circuit-switched variant and a TDM-enhanced system. The latter works by partitioning the flows into segments of a defined size, over which we conduct a parameter sweep to explore tradeoffs. In general a shorter segment provides finer grain flow interleaving and lower internal fragmentation, but also requires a more frequent re-configuration of the switch fabric, which imposes some delay and throughput penalties as data can not be transmitted while the switch fabric is getting reconfigured. Currently switch arbitration is done randomly, but research on more advanced techniques will be essential to ensure the technology uses resources in an efficient way.

4.2 Workloads

In our experiments below, we use a range of workloads which model some representative applications and well-known benchmarks. Note that these workloads include causality among the

messages, so most applications go through phases of high and low network pressure:

- *Randomapp* – Selects the source and destination uniformly at random. This is a typical networking benchmark which is used to stress the IC and, according to [14], the traffic mix run on a typical DC is unstructured and has some resemblance to random traffic.
- *Bisection* – Nodes are split into pairs at random and nodes in a pair communicate with each other. This was proposed by [34] as a means to estimate the bisection bandwidth of interconnection networks.
- *Torus 3d* – A common communication pattern in scientific applications where large matrices are split into tasks such that each task only communicates with neighbouring tasks having contiguous chunks of the matrices.
- *Nbodies* – Another typical scientific application where a collection of particles (bodies) interact with each other to model the evolution of physical phenomena (e.g. planets, atoms, etc). Tasks are arranged in a virtual ring in which each task starts a chain of messages that travel clockwise across half of the ring.
- *AllReduce* – An optimised, binary implementation of the AllReduce collective [28], widely used in parallel applications from a range of domains.
- *Mapreduce* – 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* – Models the traffic handled by a ToR switch within a DC based on the analysis of the traffic captured in 10 DCs from different domains [2]. It considers the most local traffic configuration, where 20% of the traffic is extra-rack, as reported for CLD5 in [2]. We assume a 3:1 oversubscription ratio so that 12 ports are connected to servers and the other 4 are uplinks connected to higher level switches.
- *Torremote* – Similar to ToR Local, but considering the most remote traffic configuration shown in [2], with 90% of extra-rack traffic, as observed in EDU1 of [2].

Table 2 summarises the number of flows and size for each workload. We include a visual representation of the workloads in Fig. 4. The black arrows represent messages where a reception before a send represents causality among messages. We note that instead of

Table 2: Number of flows per workload.

Workload	# Flows	Flow size (KB)
Bisection	16	1000
Randomapp	1000	1000
Torus 3d	1000	1000
AllReduce	64	1000
Nbodies	128	1000
Mapreduce	270	1000
Torlocal	1000	1000
Torremote	1000	1000

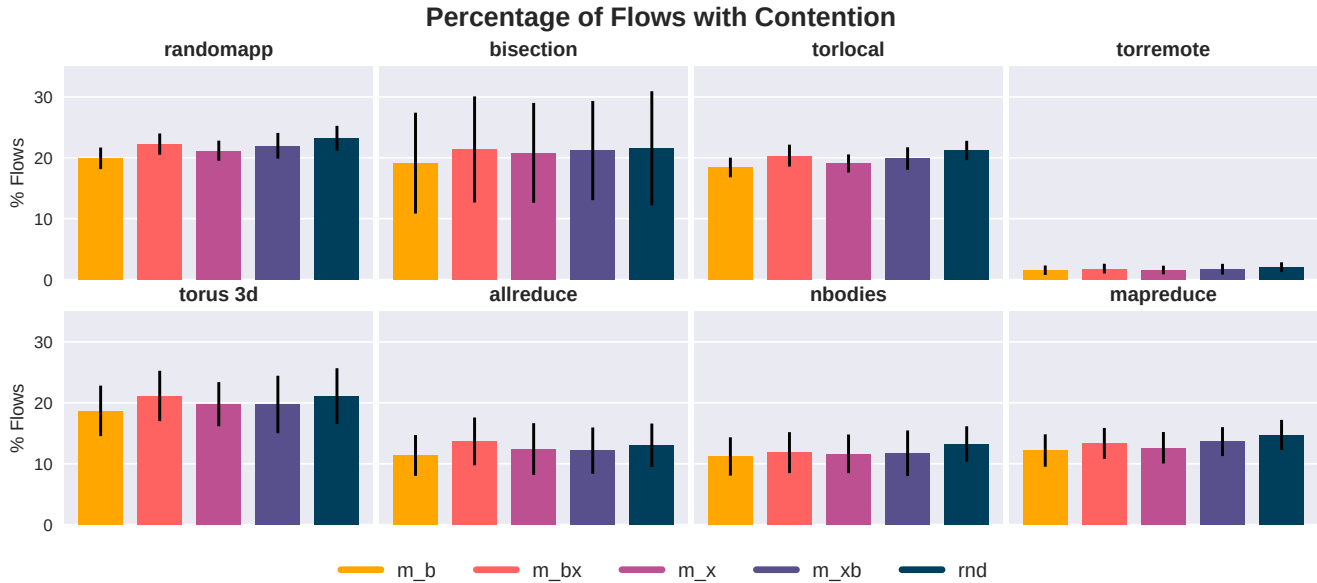


Figure 5: Percentage of flows that suffer switch fabric contention.

a 3d Torus workload, we depict a 2d Torus which is similar with one less dimension. In Nbodies, we depict the chain of messages started by a single task. All the other tasks produce a similar chain of messages which are not shown for the sake of clarity.

4.3 Experiment Process & Figures of Merit

In our experiments, we do 100 repetitions for each configuration, each of them with a different random seed. Following the standard practice for DCs and clusters, we assume the system scheduler models the system as a flat network and incorporates no locality information, so tasks are distributed randomly across the network [9, 37]. We then gather the mean and standard deviations of the following metrics:

- Percentage of flows suffering from switch fabric contention as an indicator of how much can a workload benefit from TDM.
- Normalised workload communication time to assess the impact of TDM on applications performance.
- Maximum path-dependent ILoss to measure the impact of TDM on the maximum laser power needed at the endpoints.
- Switching energy per bit, dissipated from MZI usage, to show the impact of TDM on the efficiency of the switch.

For our energy calculations, we consider an optimistic MZI tuning policy which minimizes the static power consumption of inactive MZIs. Our model only takes into account the MZIs that are used for flow communication during each timeslot, assuming the MZIs that are not used are off – i.e. that they draw no power. While in reality some extra tuning power might be needed by unused MZIs, the model is adequate for our purposes since it benefits CS: a higher static power consumption translates to higher energy when communication time is increased and, as we will see in section 5.2,

CS requires more time than TDM to perform the same communication. However, our evaluation in section 5.4 shows that even when employing this methodology, TDM can maintain energy efficiency.

5 RESULTS & DISCUSSION

In our experiments, we first investigate the prevalence of switch fabric contention when using blocking routing strategies with CS. This serves as a motivation for the use of TDM because, as explained before, a fine grain interleaving of flows is beneficial against this pathological phenomenon. Secondly, we assess using the routing strategies with TDM and compare the communication time against the routing strategies with CS, to highlight the potential savings. Thirdly, we examine the impact of flow segment size on communication time and discuss the consequences of using smaller sizes on path computation constraints. Lastly, we evaluate whether using TDM with the routing strategies increases critical-path Insertion Loss and switching energy consumption. As the routing strategies aim to provide energy efficiency by reducing these metrics, it is essential that their benefits are not negated by augmenting the strategies with TDM.

5.1 Switch Fabric Contention Occurrence

Here, we investigate what percentage of flows suffer from switch fabric contention for the eight workloads using the routing algorithms and CS, with the results depicted in Fig. 5.

Firstly, the most switch fabric contention is exhibited in the synthetic *randomapp* workload (average of 19-23%). This is expected, considering that there is no causality between the messages, which in other workloads makes the workload’s flows more amenable to the amount of path diversity offered by the switch (e.g. *nbodies*, *bisection*). Interestingly, in *randomapp*, switch fabric contention is

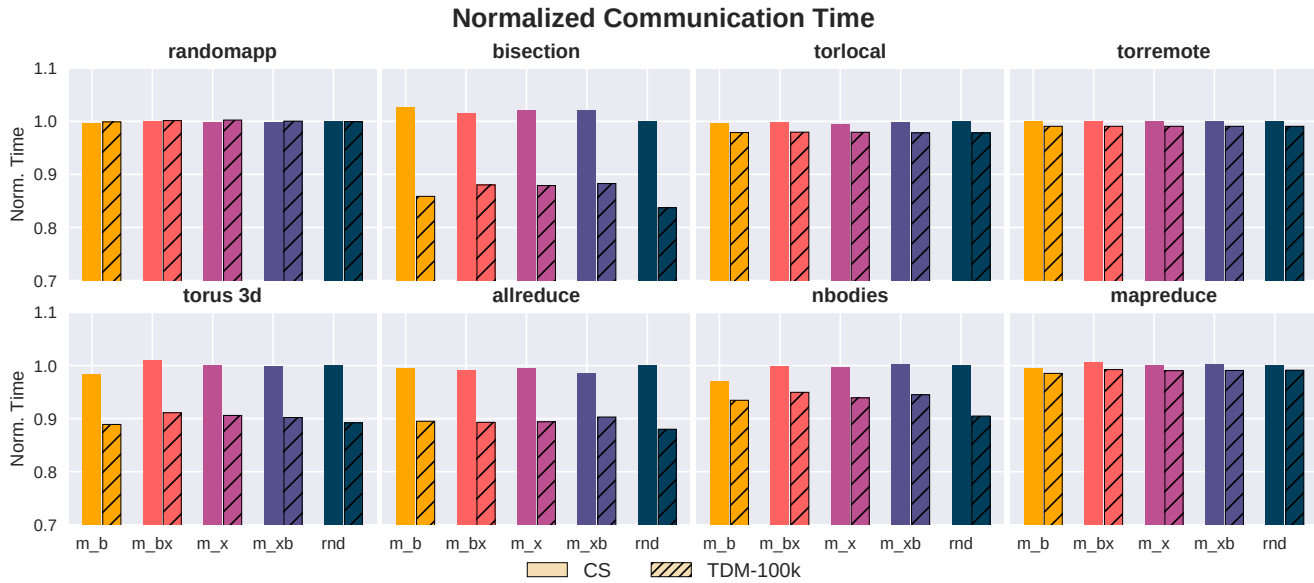


Figure 6: Normalised communication time for CS and TDM.

exhibited uniformly among the routing strategies, within one standard deviation; the random nature of the flow source/destinations cannot be taken advantage of in terms of reducing switch fabric contention by routing algorithms that are aimed at leveraging hardware asymmetries. The same behaviour occurs with the *torlocal* workload, since 80% of the traffic is similar to *randomapp* with 20% assigned to the uplink.

The *bisection* workload also presents an interesting behaviour. The percentage of flows suffering switch fabric contention, between 19-21% on average, is slightly lower than that of *randomapp*, but with highly divergent behaviour for all routing algorithms. This is attributed to two factors. Firstly, the rearrangeably-non-blocking nature of CS in the Beneš network means that depending on the ordering or the allocation of source/destination pairs, determined by randomness, flows may or may not get blocked as they are allocated paths sequentially. This aspect, coupled with the fact that the number of flows in *bisection* is small (see table 2), means a blocked flow has a pronounced effect on the total metric.

The *torus 3d* workload also suffers from significant switch fabric contention in all examined cases (18-21% average). This is because all nodes are communicating with their neighbours, which produces a heavy load and increases the chances of switch fabric contention appearing. Interestingly, it is the workload with the second highest variability after *bisection*.

The *allreduce*, *nbodies* and *mapreduce* workloads all suffer from medium amounts of switch fabric contention compared to the other workloads, between 11-15% on average, with low divergence across the routing algorithms. Interestingly, the switch fabric contention profile of each routing algorithm is similar across the three workloads (within 1%), with the “m_b” and “m_x” algorithms exhibiting the lowest switch fabric contention levels. However, these levels are all within one standard deviation of each other, indicating

that switch fabric contention cannot be reduced using hardware-inspired routing alone.

Lastly, the *torremote* workload exhibits very low levels of switch fabric contention. This is expected, considering that 90% of flows compete for access to the uplinks, thereby being blocked at the receiver and consequently leading to low network saturation in the switch. The remaining 10% can be accommodated for easily.

In summary, CS with hardware-inspired routing in a Beneš-based photonic ToR switch can indeed exhibit high levels of switch fabric contention. As this can lead to unwanted delays in communication time and an unfair use of resources, a TDM methodology is justified as there is margin for improvement by reducing the effects of switch fabric contention. In the next section, we examine the impact of TDM through flow segmentation on workload communication time.

5.2 Workload Communication Time

We continue by examining workload communication time for both CS and TDM approaches, with a flow segment size of 100KB, portrayed in Fig. 6. The depicted communication time results are normalized per workload against each workload’s communication time using the *rnd* routing algorithm and CS, in order to highlight the differences in runtime of the workloads under the two approaches.

The *randomapp* workload has average communication times between 1.248-1.254 ms for CS and 1.252-1.256 ms for TDM. Despite the large amount of switch fabric contention when using CS, TDM is unable to provide substantial changes in communication time (<1%), and always within one standard deviation. As in this workload flow destinations are assigned randomly, there is significant output contention, forcing flows to be blocked. Simply segmenting the flows is unable to alleviate the effects of output contention, leading to negligible impact on communication time from TDM.

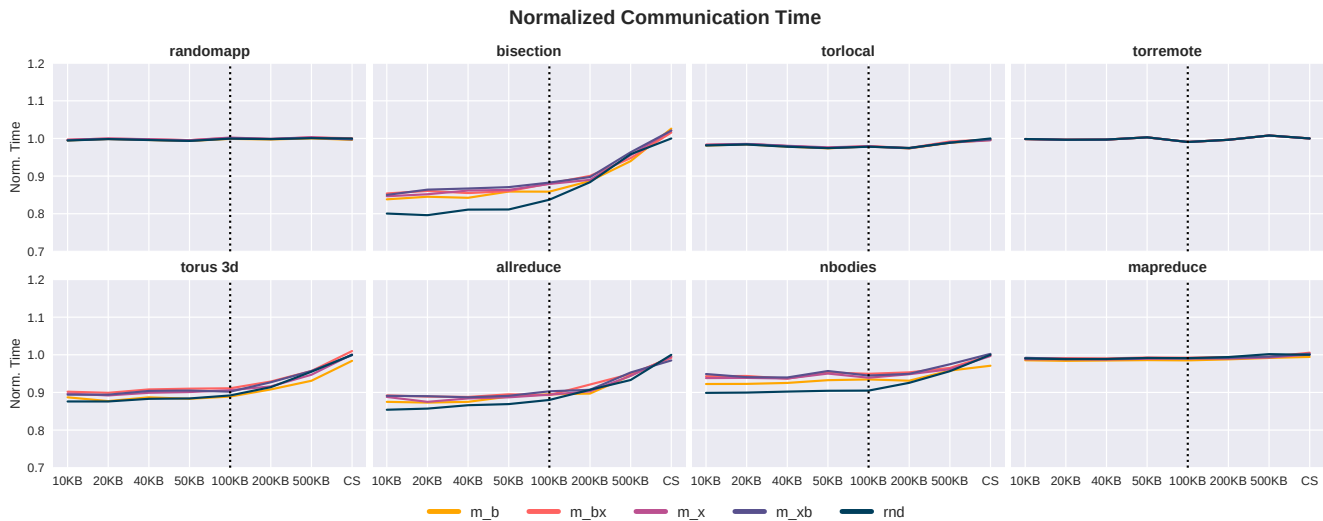


Figure 7: Normalized communication time for various flow segment sizes.

Considering the high prevalence of switch fabric contention in this workload, a more nuanced selection of flow segments in TDM could help reducing communication time. As this is out of the scope of this paper, we leave the switch arbitration as future work.

The communication time of the *bisection* workload shows substantial decreases with TDM (24.8-26.4 μ s) compared to CS (29.6-30.4 μ s). As this is a permutation workload, the flow segments are interleaved gracefully, thereby avoiding the case of one large flow being blocked, which causes the communication time to increase in CS. This effect leads to substantial time savings between 12-17%.

The *torus 3d* workload also significantly benefits from TDM (129.6-131.5 μ s) compared to CS (143.4-147.2 μ s). As this workload experienced high levels of switch fabric contention using CS, the more graceful interleaving of flows with TDM decreases flow waiting time, leading to time decreases between 10-12% compared to CS.

Allreduce also shows marked savings in communication time with TDM (108.8-109.9 μ s) over CS (119.9-121.7 μ s) with savings ranging between 10-11%. Again, this is due to the fact that TDM spreads the waiting time among the flow segments, thereby decreasing the total waiting time.

Interestingly, the *nbodies* workload exhibits slightly smaller savings with TDM (218.4-229.6 μ s) against CS (234.4-242.4 μ s) than the previous two workloads, despite the similar levels of switch fabric contention shown previously. However, as explained in section 4.2, the *nbodies* workload has a high level of causality between the tasks; this lowers the benefits introduced by TDM by 1-4% compared to the previous, high-contention workloads. Nevertheless, the savings against the baseline range between 7-10% less communication time.

Surprisingly, TDM does not benefit *mapreduce* significantly, with the communication times being reduced compared to CS by only 1-2%. In spite of the substantial level of switch fabric contention seen with CS, this workload cannot benefit from flow interleaving as much as other workloads, as flow segments exhibit output

contention, thereby being forced to wait and increasing the communication time compared to other cases. Nevertheless, as is with *randomapp*, this use-case shows that even under unfavourable workload conditions, TDM does not detriment communication time.

Expectedly, the *torlocal* workload does not benefit substantially from TDM either, with reductions between 1-2% across the routing algorithms. Like *randomapp*, this workload suffers from output contention, leading to decreased benefits from TDM. Lastly, *torremote* does not benefit either; this is a corner-case workload where, as 90% of flows are sent to the uplinks, they compete for the same resources, something that flow interleaving cannot mitigate. Again, this is expected behaviour.

In summary, inducing TDM by splitting the flows into smaller segments can lead to communication time savings in the ToR switch we examine here, which can be significant for workloads that can take advantage of path diversity. However, for some cases with relatively high switch fabric contention (e.g. *randomapp* or *mapreduce*), a more complex methodology is needed to yield communication time savings. Several alternatives are possible for this. One is to use a more intelligent arbitration mechanism which shares resources in a fairer way. Another is to attempt to select specific flow segments for transmission that fill a permutation for a given TDM timeslot. In any case, further research is needed to enhance the benefits of TDM in the context of MZI-based switching fabrics.

5.3 Flow Segment Size

We continue by conducting a parameter sweep over a range of flow segment sizes between 10KB and 500KB. Our aim is to discover the size which benefits communication time the most. As discussed, a smaller flow segment helps to decrease communication time by allowing a finer-grain interleaving of flows; however, having smaller segments means that the TDM timeslot becomes smaller, leading to tighter constraints on reconfiguration time and route solving. It is therefore important to discern how substantial the communication time reductions are and whether they justify

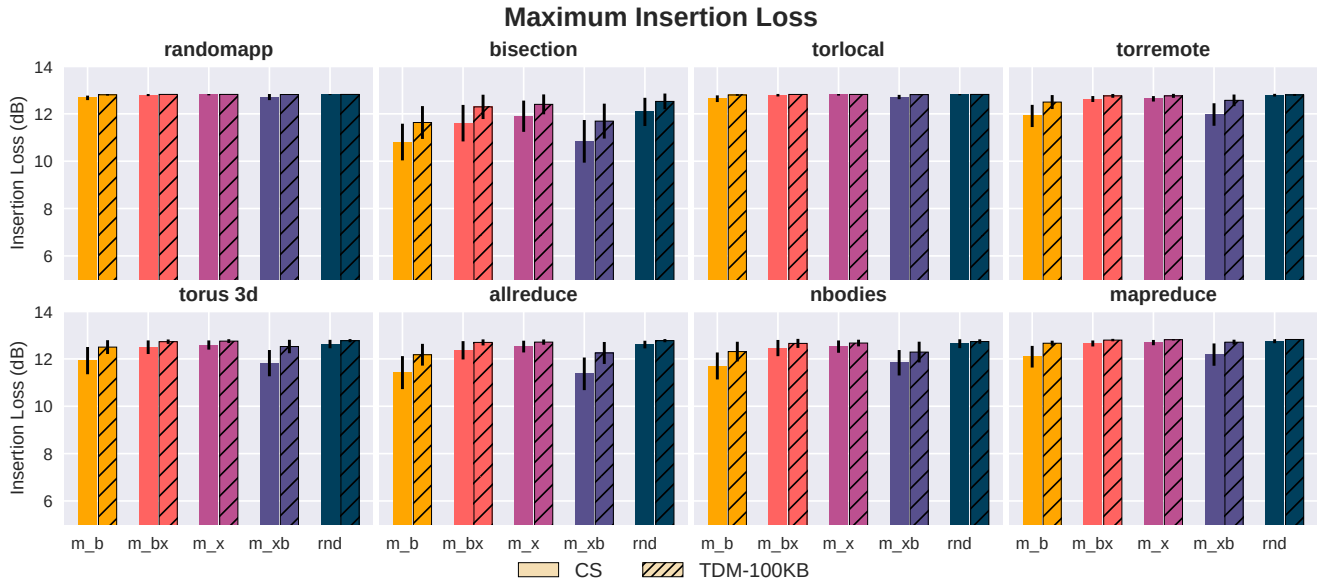


Figure 8: Worst-case exhibited ILoss for circuit-switching and TDM.

the additional constraints. We normalize the communication time results per workload against each workload’s communication time using the *rnd* routing algorithm and CS to highlight the benefits and detriments of a TDM approach compared to CS. The results are depicted in Fig. 7.

Firstly, it is interesting to note that as in section 5.2, TDM does not benefit all workloads. The *randomapp*, *torlocal*, *torremote* and *mapreduce* workloads all show negligible benefits from TDM with very little variation as segment size is increased. Considering our previous findings, this is expected behaviour; in these workloads, flow segments attempt to access the same receiver simultaneously, forcing them to be blocked. However, for the other four workloads that do not exhibit this effect, there are significant variations in communication time.

Under the *bisection* workload, communication time increases very gradually as flow segment size increases up to 200KB (50KB, for *rnd*), with the difference in time compared to 10KB segments being at most 5%. However, for larger segment sizes, communication time increases more drastically, ultimately reducing the benefit of TDM on communication time compared to CS. As *bisection* is a permutation workload, ever larger flow segments spend ever longer time intervals being blocked due to less graceful flow interleaving, ultimately leading to the behaviour discussed with CS.

Allreduce shows similar behaviour, albeit slightly less pronounced. The use of TDM reduces communication time by 11-13% up to a segment size around 100KB. For larger segment sizes, communication time increases gradually, eventually leading to the behaviour of CS.

The *torus 3d* workload also presents interesting behaviour. Communication time shows a gradual increase with segment size up to around 100KB, with the communication time for 100KB segments being within 1-2% of that for 10KB segments. Above this

size, communication time increases similarly to *allreduce*, with the *m_b* strategy maintaining a ~1% decrease in time relative to the other algorithms.

The *nbodies* workload is also affected by increasing flow segment size. Communication time remains relatively unaffected across the routing algorithms until a size of 100KB, with variations being within 1% of each other. The only exception is with segments of 50KB where for the “*m_x*” and “*m_xb*” routing strategies, communication time increases by 3%. However, for the two larger segment sizes, the TDM approach is unable to provide much benefit, ultimately leading to the behaviour seen with CS.

In summary, where TDM is impactful, increasing the flow segment size slightly increases communication time up to the inflection point at around 50-200KB segment size. Sizes above that can lead to unacceptable communication time increases. Also, choosing a very large segment size can exacerbate unfair flow segmentation, severely impacting the metric. Based on the above, a 10KB flow segment size is indeed the most impactful for reducing communication time. However, as previously discussed, flow segment size determines the TDM timeslot, which in turn enforces constraints on the routing algorithm. For example, a 10KB segment means a timeslot of 0.156 ns at link aggregate speeds of 512 Gbps. Using segments of between 50KB and 200KB size would increase the timeslot by 5-20× while only increasing communication time by 1-3%, allowing for more complex routing algorithms. Therefore, TDM is most impactful with flow segment sizes of around 100KB.

5.4 Insertion Loss & Switching Energy

We conclude the study by examining the maximum ILoss exhibited by the flows traversing the network and the switching energy consumption, presented in Figs. 8 and 9, respectively. As explained previously, the objective of the routing strategies is to increase

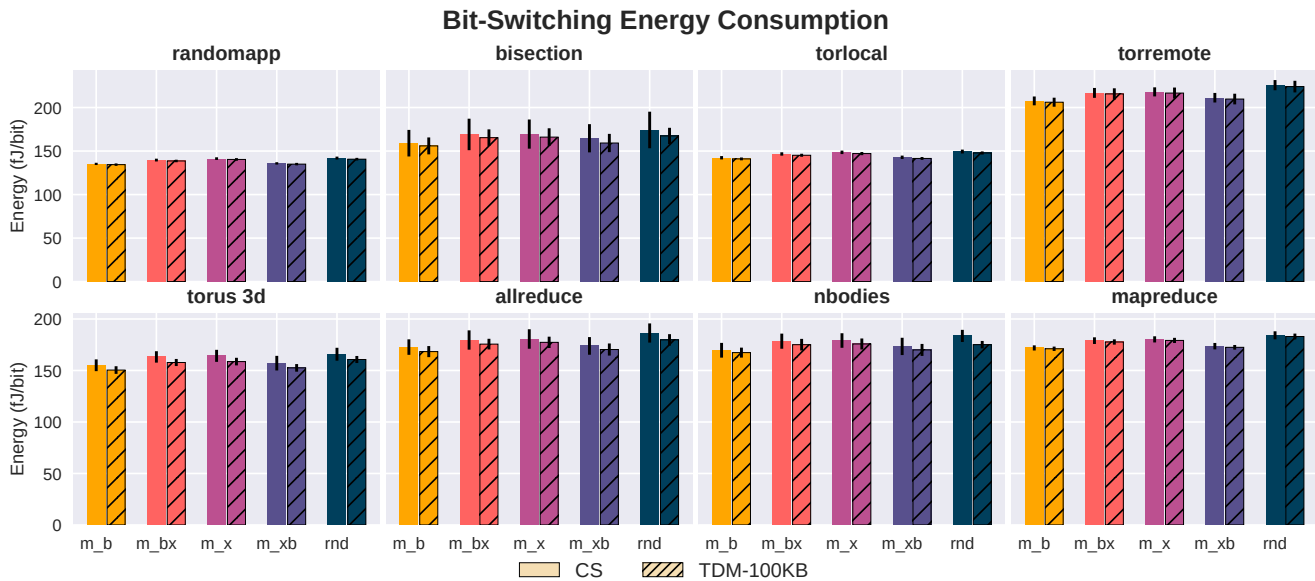


Figure 9: Bit-switching energy consumption for circuit-switching and TDM.

the energy efficiency of the switch. As such, it is important to assess whether introducing TDM reduces energy efficiency, thereby detracting from the benefits of the routing strategies. We therefore investigate whether segmenting the flows to induce TDM adversely affects either of the two metrics. We show measurements for CS and TDM, using the 100KB segments previously shown to reduce communication time. We note that in our experimental setup we have also measured average ILoss. However, the average ILoss for TDM and CS are almost identical, with discrepancies of at most 0.1 dB and within one standard deviation of each other. Therefore, we refrain to include these results for the sake of brevity.

In terms of worst-case ILoss, depicted in Fig. 8, it is interesting to note that where TDM is effective in reducing communication time, it increases worst-case ILoss by a small margin (0.5-1 dB). This is most prevalent in *bisection* under all routing conditions and in *allreduce*, *torus 3d*, *nbodies* and *mapreduce* when using the “m_b” or “m_xb” strategies. As TDM allows for flow segment interleaving, segments are allocated a less ILoss-optimal path, i.e. a path with one extra “bar” state or more waveguide crossings. This leads to higher switching fabric saturation which is reflected in the communication time reductions explained previously. Increases in ILoss can increase the required laser power, therefore increasing the energy cost. However, as seen previously, communication time is decreased substantially relative to using CS. This presents an interesting trade-off for laser power, where slightly more power is required for less time. Additionally, the worst-case ILoss shown here is the maximum incurred by flow segments. Compared to long flows with CS, short flow segments incurring ILoss for a slightly less ILoss-optimal path but for less time would arguably reduce the overall energy footprint from the lasers. This is also reflected by our results on average ILoss (not shown here) which, as explained, are negligibly affected by the introduction of TDM.

Conversely, average bit-switching energy consumption is slightly reduced for the workloads in which TDM is most impactful, between 1-4%. As TDM reduces communication times, MZIs are used for less time under those workloads, thereby reducing the energy-per-bit. This is despite the fact that MZIs in the “bar” state consume more energy. However, as the bit-switching energy consumption reductions with TDM are within one standard deviation of the energy in CS, we do not consider this effect to be significant.

In summary, where TDM is impactful, worst-case ILoss exhibited by the flow segments is slightly increased by 0.5-1 dB whereas energy consumption from switching remains virtually unaffected.

6 CONCLUSIONS

In this work, we have proposed for the first time a combination of energy efficient routing with TDM as a control mechanism for a recently fabricated 16×16 photonic Beneš switch fabric formed with thermally-electrically tuned MZIs, deployed as a ToR switch. We have evaluated our approach through simulation, employing eight realistic and synthetic workloads from the DC and HPC domains.

We have investigated switch fabric contention between communication flows when using a state-of-the-art approach (CS and hardware-inspired routing), finding that switch fabric contention occurs frequently for *randomapp* (19-23%), *bisection* (19-21%), *torus 3d* (18-21%) and *torlocal* (19-23%), with medium levels under *allreduce*, *mapreduce* and *nbodies* (11-15%).

We have evaluated the impact of our approach on communication time, finding that in some cases, it can reduce communication time substantially, e.g. up to 17% for *bisection* and 10-15% for *torus 3d* and *allreduce* when using 100KB segments. We have conducted a parameter sweep on flow segment size, finding that although communication time is least with a 10KB size, the savings compared

to sizes around 100KB are at most 3% and, therefore, do not justify the stricter time constraints imposed on path computation.

Lastly, we have assessed the impact of TDM on worst-case path-dependent insertion loss and bit-switching energy consumption and found it to be small (0.5-1 dB increase and 1-4% decrease respectively), if not slightly beneficial in the case of switching energy. To our knowledge, this is the first simulation-driven evaluation of TDM in photonic Beneš ToR switches based on a fabricated device.

This research work opens several new avenues for improving the architecture of photonic switches that we leave as future work. As discussed, we observe in our results that the way segments are interleaved may have an impact on the performance of TDM and, hence, we plan to investigate how different arbitration policies may affect the behaviour of TDM. We also plan to implement the switch controller in an FPGA, to determine further optimizations to the switching mechanism. Lastly, as higher-radix photonic ToR switches are highly desirable, we plan to assess the combination of a TDM mechanism with a wavelength-dilation scheme as a means to reduce crosstalk and therefore enable switch scalability.

ACKNOWLEDGEMENTS

This work was partly funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 754337, EuroEXA project. Prof. Mikel Luján is funded by an Arm/RAEng Research Chair Award and a Royal Society Wolfson Fellowship. Dr. Javier Navaridas is funded by a Ramón y Cajal RYC2018-024829-I from the Spanish Ministry of Science, Innovation and Universities.

REFERENCES

- [1] Theonitsa Alexoudi, Nikolaos Terzenidis, et al. 2019. Optics in computing: from photonic network-on-chip to chip-to-chip interconnects and disintegrated architectures. *Journal of Lightwave Technology* 37, 2 (2019), 363–379.
- [2] Theophilus Benson, Aditya Akella, and David A Maltz. 2010. Network traffic characteristics of data centers in the wild. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 267–280.
- [3] Eric Bernier, Dominic J Goodwill, et al. 2019. Switches and routing for on-chip photonic networks. In *2019 24th OptoElectronics and Communications Conference (OECC) and 2019 International Conference on Photonics in Switching and Computing (PSC)*. IEEE, 1–3.
- [4] Qixiang Cheng, Meisam Bahadori, Madeleine Glick, Sébastien Rumley, and Keren Bergman. 2018. Recent advances in optical technologies for data centers: a review. *Optica* 5, 11 (2018), 1354–1370.
- [5] Qixiang Cheng, Keren Bergman, Yishen Huang, Hao Yang, Meisam Bahadori, Nathan Abrams, Xiang Meng, Madeleine Glick, Yang Liu, and Michael Hochberg. 2019. Silicon Photonic Switch Topologies and Routing Strategies for Disaggregated Data Centers. *IEEE Journal of Selected Topics in Quantum Electronics* PP (12 2019), 1–1. <https://doi.org/10.1109/JSTQE.2019.2960950>
- [6] Tao Chu, Lei Qiao, Weijie Tang, Defeng Guo, and Weike Wu. 2018. Fast, high-radix silicon photonic switches. In *2018 Optical Fiber Communications Conference and Exposition (OFC)*. IEEE, 1–3.
- [7] Patrick Dumais, Dominic J Goodwill, et al. 2017. Silicon photonic switch subsystem with 900 monolithically integrated calibration photodiodes and 64-fiber package. *Journal of Lightwave Technology* 36, 2 (2017), 233–238.
- [8] Nicolas Dupuis and Benjamin G Lee. 2017. Impact of topology on the scalability of Mach-Zehnder-based multistage silicon photonic switch networks. *Journal of Lightwave Technology* 36, 3 (2017), 763–772.
- [9] Richard M. Fujimoto. 2016. Research Challenges in Parallel and Distributed Simulation. *ACM Trans. Model. Comput. Simul.* 26, 4, Article 22 (May 2016), 29 pages. <https://doi.org/10.1145/2866577>
- [10] Minming Geng, Zhenhua Tang, Kan Chang, Xufang Huang, and Jiali Zheng. 2017. N-port strictly non-blocking optical router based on Mach-Zehnder optical switch for photonic networks-on-chip. *Optics Communications* 383 (2017), 472–477.
- [11] Gilbert Hendry, Johnnie Chan, et al. 2010. Silicon nanophotonic network-on-chip using TDM arbitration. In *2010 18th IEEE Symposium on High Performance Interconnects*. IEEE, 88–95.
- [12] Adarsh Jain, R Bahl, and Alak Banik. 2014. Demonstration of RZ-OOK modulation scheme for high speed optical data transmission. *IFIP International Conference on Wireless and Optical Communications Networks, WOCN*, 1–5. <https://doi.org/10.1109/WOCN.2014.6923082>
- [13] Christoforos Kachris and Ioannis Tomkos. 2012. A survey on optical interconnects for data centers. *IEEE Communications Surveys & Tutorials* 14, 4 (2012), 1021–1036.
- [14] Srikanth Kandula, Sudipta Sengupta, Albert Greenberg, Parveen Patel, and Ronnie Chaiken. 2009. The Nature of Data Center Traffic: Measurements & Analysis. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement (Chicago, Illinois, USA) (IMC '09)*. Association for Computing Machinery, New York, NY, USA, 202–208. <https://doi.org/10.1145/1644893.1644918>
- [15] Kostas Katrinis, Dimitris Syryvelis, et al. 2016. Rack-scale disaggregated cloud data centers: The dReDBox project vision. In *Proceedings of the 2016 Conference on Design, Automation & Test in Europe*. EDA Consortium, 690–695.
- [16] Markos Kynigos, Jose A Pascual, Javier Navaridas, Mikel Luján, and John Goodacre. 2019. Scalability analysis of optical Beneš networks based on thermally/electrically tuned Mach-Zehnder interferometers. In *Proceedings of the 12th International Workshop on Network on Chip Architectures*. 1–6.
- [17] Markos Kynigos, Jose A Pascual, Javier Navaridas, Mikel Luján, and John Goodacre. 2020. On the Routing and Scalability of MZI-based Optical Beneš Interconnects. *Nano Communication Networks* 31, 10 (2020).
- [18] Yu Li, Yu Zhang, Lei Zhang, and Andrew W Poon. 2015. Silicon and hybrid silicon photonic devices for intra-datacenter applications: state of the art and perspectives. *Photonics Research* 3, 5 (2015), B10–B27.
- [19] Linear Technology 2020. *Arista 7368X4 Series 100/200/400G Data Center Switches Data Sheet*. Linear Technology. <https://www.arista.com/assets/data/pdf/Datasheets/7368X4-Datasheet.pdf>
- [20] Liangjun Lu, Shuoyi Zhao, Linjie Zhou, Dong Li, Zuxiang Li, Minjuan Wang, Xinwan Li, and Jianping Chen. 2016. 16×16 non-blocking silicon optical switch based on electro-optic Mach-Zehnder interferometers. *Optics express* 24, 9 (2016), 9295–9307.
- [21] Cyriel Minkenbergh et al. 2018. Reimagining Datacenter Topologies With Integrated Silicon Photonics. *J. Opt. Commun. Netw.* 10, 7 (Jul 2018), B126–B139. <https://doi.org/10.1364/JOCN.10.00B126>
- [22] Javier Navaridas, Jose A. Pascual, Alejandro Erickson, Iain A. Stewart, and Mikel Luján. 2019. INRFLOW: An interconnection networks research flow-level simulation framework. *J. Parallel and Distrib. Comput.* 130 (2019), 140 – 152. <https://doi.org/10.1016/j.jpdc.2019.03.013>
- [23] DC Opferman and NT Tsao-Wu. 1971. On a class of rearrangeable switching networks part I: Control algorithm. *The Bell System Technical Journal* 50, 5 (1971), 1579–1600.
- [24] Roberto Proietti, Pouya Fotouhi, Sebastian Werner, and S.J. Ben Yoo. 2020. *Intra-Datacenter Network Architectures*. Springer International Publishing, Cham, 757–778. https://doi.org/10.1007/978-3-030-16250-4_23
- [25] Lei Qiao, Weijie Tang, and Tao Chu. 2016. 16×16 non-blocking silicon electro-optic switch based on Mach-Zehnder interferometers. In *Optical Fiber Communication Conference*. Optical Society of America, Th1C–2.
- [26] Lei Qiao, Weijie Tang, and Tao Chu. 2017. 32×32 silicon electro-optic switch with built-in monitors and balanced-status units. *Scientific Reports* 7, 1 (2017), 1–7.
- [27] A. A. M. Saleh et al. 2016. Elastic WDM switching for scalable data center and HPC interconnect networks. In *2016 21st OptoElectronics and Communications Conference (OECC) held jointly with 2016 International Conference on Photonics in Switching (PS)*. 1–3.
- [28] Rajeev Thakur and William D Gropp. 2003. Improving the performance of collective operations in MPICH. In *European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting*. Springer, 257–267.
- [29] Ke Wen et al. 2016. Flexfly: Enabling a Reconfigurable Dragonfly through Silicon Photonics. 166–177. <https://doi.org/10.1109/SC.2016.14>
- [30] Sebastian Werner, Javier Navaridas, and Mikel Luján. 2017. A Survey on Optical Network-on-Chip Architectures. *ACM Comput. Surv.* 50, 6, Article 89 (Dec. 2017), 37 pages. <https://doi.org/10.1145/3131346>
- [31] S. Werner, J. Navaridas, and M. Luján. 2017. Subchannel Scheduling for Shared Optical On-chip Buses. In *2017 IEEE 25th Annual Symposium on High-Performance Interconnects (HOTI)*. 49–56.
- [32] Shuangyi Yan, Emilio Hugues-Salas, et al. 2015. Archon: A function programmable optical interconnect architecture for transparent intra and inter data center SDM/TDM/WDM networking. *Journal of Lightwave Technology* 33, 8 (2015), 1586–1595.
- [33] Qimin Yang, Mark F Arendt, et al. 2000. WDM/TDM optical-packet-switched network for supercomputing. In *Optics in Computing 2000*, Vol. 4089. International Society for Optics and Photonics, 555–561.
- [34] Xin Yuan, Santosh Mahapatra, Wickus Nienaber, Scott Pakin, and Michael Lang. 2013. A New Routing Scheme for Jellyfish and Its Performance with HPC Workloads. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (Denver, Colorado) (SC '13)*. Association for Computing Machinery, New York, NY, USA, Article 36, 11 pages.

- <https://doi.org/10.1145/2503210.2503229>
- [35] Piu-Hung Yuen and Lian-Kuan Chen. 2013. Optimization of microring-based interconnection by leveraging the asymmetric behaviors of switching elements. *Journal of lightwave technology* 31, 10 (2013), 1585–1592.
- [36] Zhao Yunchou, Jia Hao, Ding Jianfeng, Zhang Lei, Fu Xin, and Yang Lin. 2016. Five-port silicon optical router based on Mach–Zehnder optical switches for photonic networks-on-chip. *Journal of Semiconductors* 37, 11 (2016), 114008.
- [37] Saad Zaheer, Asad Malik, Anis Rahman, and Safdar Khan. 2019. Locality-aware process placement for parallel and distributed simulation in cloud data centers. *The Journal of Supercomputing* (08 2019). <https://doi.org/10.1007/s11227-019-02973-9>