EXPLORING PHOTONIC BENEŠ SWITCHING FABRICS FOR FUTURE HPC AND DATACENTRES

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE AND ENGINEERING

2022

Markos Kynigos

Department of Computer Science

Contents

Word Count: 57049

List of Tables

List of Figures

Abstract

EXPLORING PHOTONIC BENEŠ SWITCHING FABRICS FOR FUTURE HPC AND DATACENTRES
Markos Kynigos
A thesis submitted to The University of Manchester
for the degree of Doctor of Philosophy, 2022

Scalable photonic interconnection networks are highly desirable for both the High-Performance Computing (HPC) and the datacentre domains. Their potential energy efficiency and increased bandwidth capacity compared to networks based on electronics are the appeal. One of the main challenges in realising large-scale photonic interconnection networks is the adoption of network switches that can internally employ (1) high-port-count, and (2) fast, broadband photonic switching fabrics (PSFs). These fabrics are created by composing multiple stages of individual photonic devices which, when controlled with thermal or electrical tuning, can act as network switches.

Including PSFs at any level of the network still faces many obstacles, related both to photonic device design, and to control functionality for the switching fabric. This thesis contributes to the latter. It presents a simulation-based, network-traffic driven evaluation of PSFs, that are constructed using electrically/thermally tuned Mach-Zehnder Interferometers (MZIs), and formed using the Beneš network topology. These MZIs are broadband and fast switching, as they exhibit switching behaviour in *ns*-time across a continuous 30*nm* spectral segment. The Beneš topology requires the fewest MZIs, thereby reducing the PSF control complexity and increasing the photonic performance. Furthermore, the thesis enables simulating the deployment of such switching fabrics in the context of future HPC systems and datacentres.

First, the thesis discusses the main concepts enabling photonic communication, as well as the state-of-the-art in PSFs, and outlines the design challenges related to photonic switching.

It then describes a simulation-driven methodology for evaluating the relationship among communication traffic configuration, PSF-internal routing algorithm and photonic performance for a given PSF. The methodology is evaluated by simulating two

state-of-the-art PSFs selected from the literature, and comparing with their reported performance. The simulation accuracy is established against the published data (insertion loss within 0.05 dB, photonic crosstalk within 3 dB).

The thesis then proposes the concept of “Hardware-Inspired Routing strategies” (HIRs), which are a collection of routing algorithms for the studied PSFs. They leverage both the state-based asymmetry in device photonic performance and the path-based asymmetry offered by the switch fabric topology, to reduce photonic losses and switching energy-per-bit when using Circuit Switching (CS). Depending on the communication traffic configuration, the two best HIRs can be effective at reducing the photonic losses which compose the combined photonic power penalty. The power penalty determines the required signal power for the PSF and therefore the energy efficiency. Compared to the state-of-the-art “Looping Algorithm,” the HIRs can reduce the photonic power penalty by $\sim 15-20\%$ on average and by $\sim 19-15\%$ in the worst case as the PSF size increases. When considering an on-chip deployment scenario, this can lead to laser power savings between $\sim 20-77\%$ on average and $\sim 24-42\%$ in the worst case.

It then proposes augmenting the HIRs with Time-Division Multiplexing (TDM), and investigates deploying a 16×16 PSF, which is selected from the literature, within a top-of-rack switch. When using TDM, flows are partitioned into equal-sized segments, which are then interleaved by the PSF controller to reduce the timing penalty of switch fabric contention incurred by CS. The simulations show that when employing TDM, communication time within the PSF can be reduced by up to $\sim 20\%$ compared to CS, depending on the employed workload, while not affecting insertion loss or switching energy per bit.

The thesis concludes by investigating the joint impact of traffic arbitration policy, PSF-internal routing algorithm and workload on the switch performance (insertion loss, communication time within the PSF, switching energy per bit). The results indicate that communication time is affected the most by the arbitration policy with differences generally at $\sim 10\%$ and, in some extreme cases, over 30%. Switching energy per bit is affected less significantly, with differences around $\sim 4-5\%$ (at most 15%), while insertion loss is negligibly affected. These indicate that arbitration in these PSFs could be designed independently from routing. The least-frequently used policy was found to be the best overall and particularly with regular workloads, in which tasks progress at the same pace, with clear communication phases of fixed size. In these, the communication time is reduced by the arbitration policy by $\sim 30\%$, while in irregular workloads the communication time is increased due to the policy by $\sim 6\%$. On the other hand, one of the novel policies proposed, accelerated round-robin, excels with irregular workloads; in these, tasks progress at a pace dictated by traffic causality.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

First and foremost, I thank my mother and father, who offered me every opportunity under the sun. Your nurture and ongoing support thereafter have helped me become the person I am today. Thank you both for teaching me how to think critically, to behave with integrity, to interact with empathy and, most importantly, to persevere. Special thanks also to my brother, David, whose supportive words helped me throughout this journey. We may be far apart in distance, but never in spirit.

I also thank my supervisors. Dr. Javier Navaridas, even during the adversity of the pandemic, offered me continuous support and direction. I thank you for your dedication, for uncounted hours of discussion, for teaching me how to become a researcher and for constantly encouraging me to push myself and grow. Prof. Mikel Luján, who stepped in as main supervisor in the middle of my PhD, supported me throughout. I thank you for your reviews of my work, your assistance in focusing it and your support with with the submission.

Many thanks also to Jose Pascual for his technical assistance and long discussions, which significantly helped with the direction of my thesis.

I thank Prof. Keren Bergman from the Lightwave Research Lab at Columbia University, for enabling me to collaborate with her team, as well as Richard Dai and Kristoff Yan, whose discussions greatly improved my understanding on the physical layer of silicon photonics.

I thank everyone from the APT group for making it a great place to pursue research and for their support and friendship, both in and outside of the university. Special thanks to Nikos Kyparissas, Konstantinos Iordanou and Kyriakos Paraskevas for their help in reviewing chapters of the thesis and, of course, Luca Perez, for the coffee breaks which kept us sane during the final semester.

Last but in no way least, I thank my partner, Katerina, for her never-ending support. I thank you for sharing my successes and failures, for being there for me every single day and through every adversity. You are extraordinary and I hope to be as steadfast when you submit your thesis.

Acronyms

Arrayed Waveguide Grating Coupler (AWGR)

A passive photonic device used to interconnect multiple endpoints.

Bit-error Rate (BER)

The number of bit errors per unit time, where bit errors are received bits which have been altered due to noise, interference, distortion or synchronisation errors.

Circuit Switching (CS)

A switching technique which allocates the entire set of required resources to the transmitted traffic for the required time of transmission, and blocks requests for allocated resources.

Coarse Wavelength-division Multiplexing (CWDM)

A form of wavelength division multiplexing which employs few wavelengths (less than 8), differing significantly in frequency.

Complementary metal-oxide-semiconductor (CMOS)

A family of processes used to fabricate integrated circuits.

Datacentre (DC)

A facility which is composed of interconnected compute, storage and network infrastructure, with the purpose of delivering shared applications and data.

Datacentre Network (DCN)

An interconnection network used to connect compute and storage machines within a datacentre.

Dense wavelength-division Multiplexing (DWDM)

A form of wavelength-division multiplexing, in which the employed wavelength channels align to a standardized frequency grid.

Electro-optic (EO)

Used to refer to electro-optic tuning, which employs the free-carrier dispersion effect to induce phase change or resonance.

Electro-Optically tuned Mach-Zehnder Interferometer (EOMZI)

A Mach-Zehnder Interferometer that employs electro-optical tuning to change state.

Hardrare-Inspired Routing Strategies (HIRs)

A set of path-wise routing strategies which select paths through switches based on the properties or state of the underlying traversed photonic medium.

High-Performance Computer (HPC)

A special-purpose computer composed of multiple servers, connected using an interconnection network.

Insertion Loss (ILoss)

Power attenuation incurred by the photonic carrier as it traverses a photonic medium.

Interconnect (IC)

Used to describe a communication link (e.g. copper or fibre cable, waveguide) that connects compute or memory resources to each other.

Mach-Zehnder Interferometer (MZI)

A photonic device consisting of two inputs and two outputs, which can be used as a switching cell.

Microring Resonator (MRR)

A photonic device formed using a circular silicon waveguide structure, used to change the propagation direction of traversing light, based on resonance mechanics.

Multi-mode Interferometer (MMI)

A photonic device with two inputs and two outputs, which can be used as a coupler within switching devices or as a waveguide crossing.

Multipath Interference (MPI)

Refers to a type of crosstalk, which is generated by delayed versions of a photonic signal leaked through multiple physical paths through a photonic medium.

Network-on-Chip (NoC)

A network which interconnects multiple chip components.

Non-Return-to-Zero (NRZ)

A binary line code in which ones are represented by one significant condition and zeroes by some other significant condition, with no other neutral or rest condition.

On-Off Keying (OOK)

A modulation scheme which represents digital data as the presence or absence of a carrier wave.

Optical Interconnection Network (OIN)

An interconnection network which employs optical links and electronic switches.

Optical Network-on-Chip (ONoC)

An optical network which interconnects multiple chip components.

Photonic Integrated Circuit (PIC)

An integrated circuit containing at least two optical devices.

Photonic Interconnection Network (PIN)

An interconnection network in which traffic does not suffer electro-optic or opto-electric conversion in intermediate network nodes.

Photonic Switching Fabric (PSF)

A collection of active photonic devices tiled and interconnected by passive photonic devices, with the purpose of performing switching functionality.

Rearrangeably non-blocking (RNB)

A network topology which can route arbitrary input-output connection permutations, but incremental route allocation may cause some pre-allocated connections to require rerouting.

Return-to-Zero On-Off Keying (RZ-OOK)

On-off keying modulation using a return-to-zero binary line code.

Signal-to-Noise Ratio (OSNR)

The measure of the ratio of signal power to noise power in an optical channel.

Silicon Photonics (SiPh)

A material platform which uses silicon-on-insulator, from which photonic integrated circuits can be made.

Space Division Multiplexing (SDM)

A multiplexing technique for optical data transmission where multiple spatial channels are utilized.

Strictly non-blocking (SNB)

A network topology in which the connections of any input-output permutation can be allocated incrementally, without rerouting pre-allocated connections.

Switching Element (SE)

A device that performs switching.

Thermo-optic (TO)

Used to refer to thermo-optic tuning, with which the refractive index of a waveguide is modified to induce a phase change or resonance.

Time-Division Multiplexing (TDM)

A communication channel sharing technique, in which the channel capacity is partitioned into time-slots, with competing traffic being transmitted using interleaving.

Top-of-Rack (ToR)

Used to refer to a switch, a Top-of-Rack switch connects in-rack servers to the interconnection network.

Vertical Cavity Surface-emitting Laser (VCSEL)

A type of laser.

Wavelength-division Multiplexing (WDM)

A multiplexing technique through which a data stream is split and encoded on multiple wavelengths traversing the same photonic medium.

Chapter 1

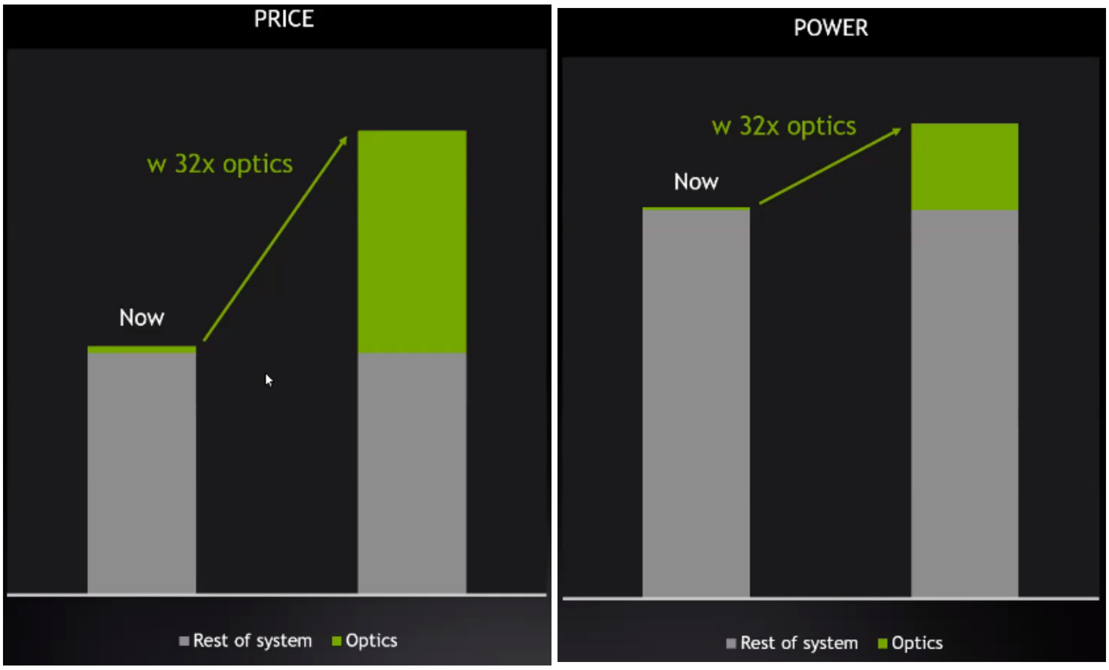
Introduction

1.1 Motivation

The field of High-Performance Computing (HPC) is transitioning into the "exa-scale" era, in which large-scale, special-purpose computers have the ability to reach peak performances of ≥ 1 *EFLOPS*. The Fugaku supercomputer [Don20], deployed in 2020, Japan, is currently the most powerful HPC system and can achieve ~ 415 *PFLOPS* peak performance, while the United States of America aims to deploy two exa-scale HPC systems, Frontier and Aurora, by 2023 [(OL21, SRM+19].

HPC systems rely on massive parallelism, enabled by high-performance interconnection networks, to distribute compute tasks to the system nodes and to communicate data among them. Performance tends to be reported using benchmark applications, such as LINPACK [DLP03], which is a computation-heavy workload, or HPCG [DHL16], which stresses the communication to memory and, to an extent, the interconnection network. The performance of Fugaku under the LINPACK benchmark is 80.9% of peak performance; however, its performance under the HPCG benchmark is 2.8% of the peak. While this is still the best performance for HPCG globally, the mismatch in performance between the two benchmarks highlights the growing gap in resource provision between computation and communication capability in current HPC systems [Ber21]. This occurs due to the increasing cost and power consumption of the interconnection network. As HPC performance becomes increasingly dependent on data communication efficiency, it is expected that future high-performance interconnects will require a paradigm shift in order to meet communication demands [RNH+15].

At the same time, the data centre world has transitioned into the "Zettabyte era",



(a) Cost increase.

(b) Power consumption increase.

Figure 1.1: A qualitative comparison of the price and cost increases associated with improving pluggable optics. Image copied from [SZC+21].

in which large-scale Data Centres (DCs) are tasked with serving ever-growing Internet Protocol (IP) traffic, which has already surpassed 2 $Zettabytes/year$ [Net17]. This demand, driven by both increasing video traffic and use of AI, is substantially increasing the power consumption of data centres. In spite of significant advances in DC energy efficiency, they are projected to consume between $3-13\%$ of global power production by the year 2030 [AE15, GKL+21].

In spite of their differences, both HPCs and DCs rely on interconnection networks, which are composed of copper or optical links, and electronic packet routers and switches. In both cases, the number of employed switches grows with the scale of the system, and therefore the power consumption for switches increases. In addition, optical cables are being used more frequently, as the demand for more bandwidth within HPCs and DCs increases. Pluggable optical cables, however, are costly and further increase power consumption [SZC+21]; this is depicted in Fig. 1.1, which qualitatively shows the price and power consumption increase when improving the capabilities of current pluggable optical cables. Communicating data must be transformed from the electrical domain, where the data is switched at every hop within electronic packet switches, to the optical domain for transmission and vice versa (O/E/O conversion), a process which consumes power and prohibits scaling. This in turn enforces a dependency between electronic switches and pluggable optics; switches must either be upgraded at every optical cable data rate generation to support new transceivers, which perform O/E/O conversion, or transceivers must be constrained by legacy capabilities.

Augmenting interconnection networks by employing photonic switches is considered a highly promising approach for surmounting many of the challenges examined above [NRC$^+$15]. On one hand, this could reduce the dependency on pluggable optics and therefore O/E/O conversions. On the other hand, recently demonstrated photonic switches employing Silicon Photonics technology (SiPh) can be highly power efficient, thereby reducing the power consumption of the interconnection network. SiPh devices, most notably Microring Resonators (MRRs) and Mach-Zehnder Interferometers (MZIs) are capable of switching multiple wavelengths simultaneously in the $\mu s - ns$ scale, thereby allowing for data rate scaling using Dense-wavelength-Division Multiplexing (DWDM), and can be cascaded to form photonic switching fabrics, thereby increasing the port count. These switching fabrics can then be used to either augment current interconnection networks by providing efficiency through new capabilities (e.g. bandwidth steering [MST$^+$19]), or to potentially replace some electronic packet switches, thereby forming hybrid electronic/photonic interconnection networks [FPR$^+$10].

However, many challenges still exist in using photonic switching fabrics within high-performance network switches, such that they can compete with standard electronic packet switches. The lack of practical data buffering in the optical domain makes packet switching challenging, necessitating the use of circuit switching at the transmission level. This in turn affects the complexity of the switching fabric controller and routing algorithm internal to the switch, which is responsible for providing uninterrupted lightpaths from a source to a destination port. At the same time, optical losses which are inherent to the device technology, namely insertion loss and photonic crosstalk, limit the scalability of switching fabrics [DL17].

This thesis examines the challenges outlined above, in the context of rearrangeably-non-blocking photonic switching fabrics formed with MZIs. In addition, it improves on the state-of-the-art by proposing a set of routing techniques which can be used to optimise the photonic power penalty, while not prohibitively increasing switch controller complexity.

1.2 Thesis Contributions

This thesis examines the potential of photonic switching fabrics formed with thermally/electrically tuned MZIs using the Beneš network topology, for their deployment in future HPCs and datacenters. The **research question** that is addressed by this thesis comprises of the following components:

1. What are the limitations of these photonic switching fabrics in terms of scalability, photonic metrics (insertion loss, photonic crosstalk, power budget and required signal power) and performance, that is communication time?

2. Can the properties of the adopted topology and underlying technology be leveraged by intra-switch routing algorithms to reduce the photonic power penalty, switching energy, signal power and therefore required laser power? As required laser power is the main limitation in scalability, can the use of these routing algorithms reduce required signal power enough for larger switching fabrics to become realistic?

3. What is the impact of network traffic configuration on photonic metrics (insertion loss, optical crosstalk, photonic power penalty, signal power, switching energy) and performance metrics (blocking in the form of contention, communication time)? How is this impact affected by routing algorithm selection?

4. How does the performance of the proposed routing algorithms evolve with the network size, with respect to the metrics under consideration?

5. The routing algorithms proposed to optimise for photonic metrics introduce blocking in the form of contention in the switch, thereby decreasing the performance of the fabric. Can time-division multiplexing ameliorate the performance degradation? To what degree and under which traffic conditions?

To address the research question, this thesis makes the following **contributions**:

1. It proposes a methodology and tool for evaluating the interactions between network traffic configuration, internal routing algorithm selection and photonic metrics (exhibited insertion loss, photonic crosstalk, signal power). To the best of the author's knowledge, the use of network traffic-driven simulation consisting of causality-enabled traffic models as a tool for evaluating the performance of photonic switching fabrics, and assessing the above interactions has not been

previously investigated. The methodology, as well as the simulation framework that has been developed to support it, are described in Chapter 3. The methodology's use is involved in addressing all the components of the research question.

2. It contributes to the state-of-the-art in routing algorithms for the internals of photonic Beneš switching fabrics based on thermally/electrically tuned MZIs by proposing a set of circuit-switched Hardware-Inspired Routing strategies (HIRs). These routing strategies leverage the inherent asymmetry in the state-based insertion loss, crosstalk and tuning energy of this type of MZI to provide loss-optimised paths inside the switching fabric. This contribution comprises of the following parts:

   (a) The first part consists of the routing strategies discussed in Chapter 4. Each routing strategy proposed here optimises against a single criterion, namely minimizing the number of waveguide crossings, minimizing the number of required state changes, and minimizing the number of MZIs in the "bar" state. The *rnd* strategy, which randomly selects a potential path, also belongs here.

   (b) The second part consists of the hybridised routing strategies proposed in Chapter 5. These routing strategies combine the optimisation criteria used in the previous to elicit further improvements in the considered metrics.

   Compared to the state-of-the-art Looping algorithm [Ben64], most HIRs show significant reductions on signal power, aggregate crosstalk, insertion loss and switching energy exhibited per flow, both on average and in the worst case. This contribution, discussed in Chapters 3-5, addresses the second and third components of the research question. It also address the fourth component, by focusing on insertion loss, switching energy and communication time for Chapters 4 & 5. It is noted that the latter two chapters do not focus on photonic crosstalk, as the crosstalk model was developed after the finalisation of the publications they contain.

3. The proposed HIRs introduce a trade-off, as they select loss-optimised paths by sacrificing the non-blocking characteristics of the Beneš network, thereby introducing contention in the fabric. To address this, the third contribution of this thesis is the augmentation of HIRs with time-division multiplexing (TDM), rather than circuit switching. Circuit switching and TDM are defined in Sections

2.4.2 and 2.4.3, while their application to the PSF is detailed in Chapter 6. The contributed technique ameliorates the effects of contention while maintaining the insertion loss savings of the HIRs. This approach is evaluated using the proposed methodology and considering a 16×16 Beneš switching fabric deployed at the top-of-rack level. The findings show that employing TDM can reduce communication time by up to 20% in the best case compared to CS, thereby decreasing the impact of contention, while having minimal impact on the exhibited critical path insertion loss ($\sim 0.5dB$ increase) and average required switching energy ($\sim 5\%$ decrease). A trade-off analysis is also performed with respect to TDM segment size, finding that it can be increased tenfold with negligible performance impact, thereby reducing constraints in path computation time. This contribution addresses the fifth component of the research question.

In addition to the above, the thesis includes an investigation on the role of arbitration algorithms in the control of the photonic switching fabrics under investigation. This investigation has been submitted for publication in the Journal of Optical Communication Networks (JOCN). The thesis author has contributed to this research article by co-authoring the paper, in particular sections 1-4 of the paper, as well as providing assistance for the simulation experiments contained in the work and their analysis. The domain investigated in the article is an evolution of the thematics of this thesis and is based on the methodology proposed as a thesis contribution. However, as the thesis author did not develop the arbitration algorithms examined in the article, it is not considered to be within the thesis authors' contributions and is therefore not included above.

1.3 Publications

As this thesis is organized in the journal format, the majority of the contained work has either been published or submitted as the following research articles:

1. M. Kynigos, J. Pascual, J. Navaridas, M. Luján, J. Goodacre, *Scalability analysis of optical Beneš networks based on thermally/electrically tuned Mach-Zehnder interferometers*, Published in the ACM Proceedings of the 12th International Workshop on Network on Chip Architectures (NoCArc), 2019 (Chapter 4).

2. M. Kynigos, J. Pascual, J. Navaridas, M. Luján, J. Goodacre, *On the routing and scalability of MZI-based optical Beneš interconnects*, Published in the Elsevier

Journal of Nano Communication Networks, 2020 (Chapter 5).

3. M. Kynigos, J. Pascual, J. Navaridas, J. Goodacre, M. Luján, *Power and energy efficient routing for Mach-Zehnder interferometer based photonic switches*, Published in the Proceedings of the ACM International Conference on Supercomputing (ICS), 2021 (Chapter 6).

4. J. Navaridas, M. Kynigos, J. Pascual, M. Luján, J. Miguel-Alonso, J. Goodacre, *Understanding the Impact of Arbitration in MZI-based Beneš Switching Fabrics*, Submitted for publication to the IEEE Transactions on Parallel and Distributed Systems (TPDS), 2022 (Chapter 7).

In addition to the above, the following papers which are based on this thesis are in progress to be submitted:

- The first paper will present the photonic interconnection network simulator that has been augmented for this thesis, PhINRFLow. Based on the study included in Chapter 3, this paper shall discuss the capability of simulating bufferless photonic switching fabrics using flow-level network simulation, thereby enabling network traffic-driven analyses of photonic switching fabrics.

- The second paper will be based on the comparative study of intra-switch routing algorithms presented in Chapter 3. It will propose a technique for partitioning the employed wavelengths into groups, forming a set of *wavelength arbitration* policies to be used in conjunction with the HIRs. The goal of these policies will be to reduce photonic crosstalk in MZI-Beneš photonic switching fabrics and allow for their greater scalability.

1.4 Outline

Chapter 2 discusses silicon photonic devices and their composition into photonic pathways, and describes the photonic losses and power penalties that arise from the their use. It also compares photonic interconnect topologies for multi-stage switching fabrics and discusses the different types of switching methodology that have been previously employed. Lastly, it introduces the photonic switching fabric assumed as a basis for the simulation-driven studies in later chapters and describes the interdependent design challenges which arise from the formation of photonic networks.

Chapter 3 proposes a methodology for evaluating the effect of network traffic configuration and of intra-switch routing algorithms on the performance of Beneš-based photonic switching fabrics, which is the first contribution of this thesis. This methodology focuses on the application of network traffic-driven simulation to photonic switching fabrics, allowing for investigations into the interactions between network traffic, routing algorithm and their impact on the photonic performance of the fabric. It describes the simulation framework which has been augmented to pursue the research aims of the thesis, compares the results of the simulation models against state-of-the-art switching fabrics and investigates the impact of routing algorithm selection and traffic configuration on photonic metrics.

Chapter 4 proposes the concept of hardware-inspired routing strategies (HIRs) for Beneš networks formed with thermally-electrically tuned Mach-Zehnder Interferometers (MZIs). These routing strategies, aim at reducing the path-dependent insertion loss and switching energy incurred by flows as they traverse the photonic switching fabric.

Chapter 5 improves on the concepts proposed in Chapter 4 and proposes the hybridised HIRs. These, together with those described in Chapter 4, are the second contribution of this thesis. These routing strategies aim to further reduce the insertion loss and switching energy exhibited. The chapter also includes an investigation of their impact on communication time.

Chapter 6 introduces the third contribution of this thesis, as it investigates the application of time-division multiplexing as a switching technique for the photonic switching fabric. It demonstrates that using time-division multiplexing can mitigate contention-induced performance degradation.

Chapter 7 investigates the impact of arbitration algorithm selection on switch performance, by examining various arbitration algorithms and their combination with routing algorithms and switching techniques. It is shown that routing and arbitration can be designed independently from each other, as the combination of routing algorithm and arbitration algorithm does not significantly affect the metrics under investigation. However, arbitration and switching technique are more closely interrelated, as communication time was affected significantly for different arbitration algorithms when using TDM compared to CS.

Chapter 8 summarises the contributions of this thesis and details opportunities for future research.

Chapter 2

Silicon Photonics-based Interconnects: Technology Review

2.1 Introduction

Although optical interconnection networks (OINs) have been researched since the 1980s, many challenges remain in their deployment. OINs are distinct from photonic interconnection networks (PINs) in that the latter require the communicated information to remain in optical form end-to-end [NNLBX17]; information is modulated optically at the transmitter, carried optically by the interconnect and demodulated into the electrical domain at the receiver. Although the two terms are often used interchangeably in the literature, this distinction shall be adopted for this work.

OINs, i.e. interconnection networks that employ optical links and electronically buffered intermediaries, are attractive in the short to medium term for large-scale computing contexts such as within datacentres or high-performance computers. Their physical characteristics reduce the dependency between interconnect energy consumption and communication distance [Bor13] compared to electronic variants. However, these networks require network traffic to suffer conversion into the electrical domain for storage at every network hop, and re-conversion into the optical domain for transmission to the next hop. As has been indicated in many studies (e.g. [PC20] and [KT12]), this requirement leads to poor scalability of these networks from a power and energy perspective; as interconnect demands increase, this solution can be significantly improved upon in terms of power and energy efficiency.

Conversely, PINs operate on the premise of no electro-optic conversion during network path traversal, with conversions occurring only at the transmitting and receiving node. This is achieved by deploying photonic switches that implement circuit switching at the lightpath or wavelength level. These networks can offer a comparatively improved energy profile, making them attractive candidates for high performance interconnection networks in datacentres or HPC. However, many challenges remain to realising high-performance photonic interconnects at the system level, which will be discussed in Section 2.9. Adopting PINs requires photonic switches and these switches are still challenging to materialize, leading to datacentre and HPC interconnect architectures requiring significant changes for adoption to occur. Therefore, the cost of replacing electrical links for optical fibres while maintaining electronic switching (i.e. OINs) is currently dwarfed by the cost of adopting the PIN paradigm [Bah18]. In summary, significant challenges exist at the photonic device, switching fabric, switching methodology, and network control levels.

This chapter discusses the fundamentals of photonic communication in PINs and is structured as follows. Section 2.2 disambiguates the terminology used in this thesis. Section 2.3 provides an overview of photonic communication. Section 2.4 discusses switching and multiplexing methodologies that are used in photonic communication. Section 2.5 discusses the photonic devices , while section 2.6 describes the photonic losses and power penalties that arise from the individual devices, as they are traversed by a carrier beam travelling from a sender to a receiver. Section 2.7 discusses photonic interconnect topologies for multi-stage switching fabrics and their merits. Section 2.8 details the photonic switching fabric chip that has been assumed as a baseline for the simulation-driven studies in this thesis. The chapter concludes with Section 2.9, where the interdependent design challenges that arise from combining the above to form a photonic network are discussed.

2.2 Nomenclature

In photonics communication systems, much of the terminology has been adopted from electronic communication, leading to ambiguity. This section clarifies the meaning of the terminology used in this thesis.

A **Link** is a point-to-point connection between two nodes. A **photonic link** is a link where the communicated traffic is encoded from the electrical to the optical domain at the sender node and decoded from the optical to the electronic domain at the receiver, after traversing a passive photonic medium.

A **wavelength**, or λ, corresponds to a frequency within a specific spectrum which is used for communication.

A **carrier beam** refers to a beam of light, usually generated by lasers, comprising of one or more wavelengths. Communication traffic is modulated on to the carrier beam or a subset of it.

A **channel** in photonics refers to a wavelength. **Intra-channel**, as in intra-channel crosstalk, describes an effect that occurs between two carrier beams at the same wavelength, whereas **Inter-channel** describes an effect occurring between different wavelengths.

A **switch**, or switch device, refers to a device with more than one inputs and more than one outputs, which is capable of controlling the desired output port of a photonic signal which ingresses through an input port.

A **connection** between two nodes refers to an uninterrupted transmission of traffic between those nodes, such as a circuit (Circuit-Switching) or a TDM timeslot.

A **lightpath** is a path between two nodes between which light passes through unmodified, meaning on the same wavelength(s) and without electronic conversion. A lightpath includes all devices through which a carrier beam travels from a source to a destination. This is differentiated from a **path**, which refers to the set of connected devices which *may* be used by a carrier beam to travel from a source to a destination. In this thesis, and particularly in chapter 3, lightpath refers to the tracing of the carrier beam through a switching fabric, as it travels along a path.

A **photonic pathway** is used as an umbrella term that encompasses all the stages of communication between two nodes in a PIN.

A **victim** is a photonic signal, encoded onto a carrier beam, which suffers interference effects from other photonic signals.

An **aggressor** is a photonic signal, encoded onto a carrier beam, which interferes with other photonic signals.

Insertion loss is the attenuation in power of the carrier beam of a photonic signal, as it traverses one or more photonic devices.

crosstalk is the power level of an aggressor photonic signal or set of signals, as measured at the output port the victim signal uses to egress from the photonic switching fabric.

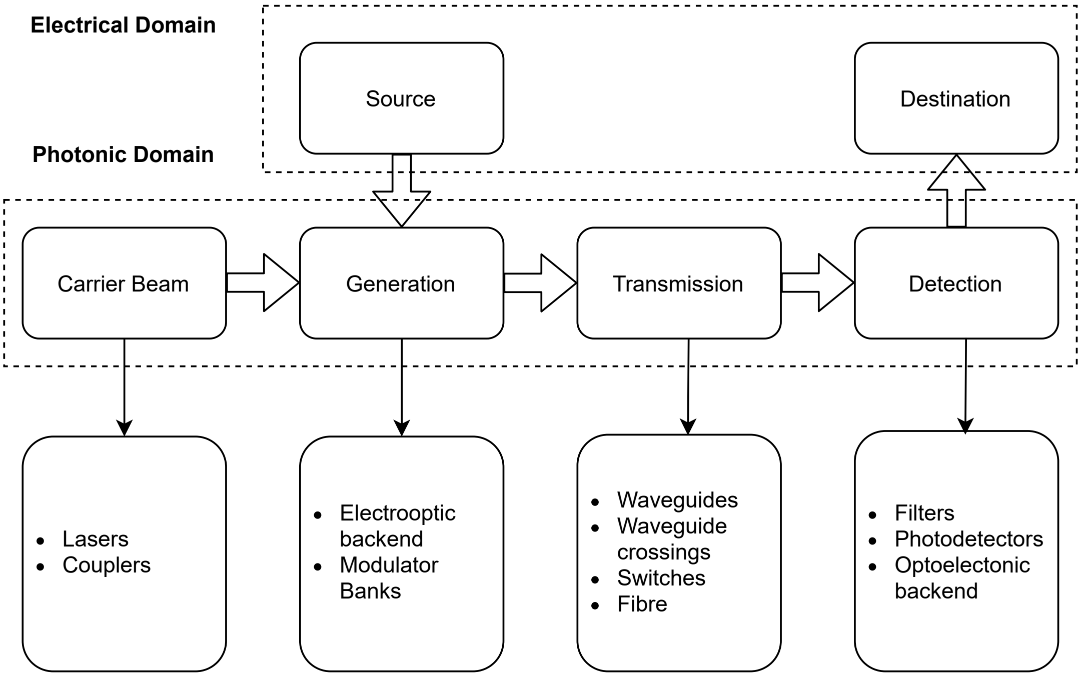


Figure 2.1: High-level depiction of photonic communication.

2.3 Anatomy of a Silicon Photonic Pathway

Fig. 2.1 outlines the main stages of photonic communication, as well as their constituent parts. The source and destination nodes are in the electrical domain. A carrier beam, which is commonly emitted by a laser source, is coupled into a modulator array during photonic signal generation; there, communication traffic from the electronic domain is modulated onto the beam. The beam is then transmitted over the medium in the transmission phase. During this phase, the photonic signal attenuates as it traverses devices, and can suffer interference from other photonic signals if switches are traversed. These two effects, known as insertion loss and crosstalk, are discussed in Section 2.6. The transmission phase ends when the photonic signal reaches the detection array at the destination. During the detection phase, the communication traffic is de-modulated from the carrier beam into the electronic domain so that it may be used by the destination node.

A connection between a source and a destination using a PIN is called a photonic pathway. A typical pathway is composed by a laser source, couplers, modulators, waveguides and fibre, photonic switching fabrics, and detectors. Photonic switching fabrics, in turn, comprise of waveguides, switching devices and, depending on the fabric topology, waveguide crossings. In the case of on-chip communication, the modulators and detectors are coupled to the switching fabric through waveguides. The carrier beam source can either be co-located using on-chip lasers, or can be coupled into the chip from an off-chip laser source.

In the case of rack-level communication, carrier beam generation, photonic signal generation and detection occur inside pluggable *transceivers*. The transceiver endpoints are then coupled to switches and/or servers represented here as abstracted sender and receiver nodes.

2.4 Switching & Multiplexing Techniques

Although some switching methodologies (e.g. circuit switching) and some multiplexing techniques (e.g. time division multiplexing) are similar between electronic and photonic communication, physical differences in the latter introduce limitations but also new opportunities in communication. In this section the switching and multiplexing methodologies most commonly considered for photonic switches will be discussed. The current fundamental limitation of photonics will also be explained, along with its impact on switching techniques.

It is noted that multiplexing techniques relying on polarization and mode manipulation (e.g. [WHD14] [LOC$^+$14]), i.e. polarization division multiplexing and mode division multiplexing, are also interesting avenues for increasing the available bandwidth beyond wavelength division multiplexing. However, as the focus of this work is on single-mode MZI switching fabrics using WDM, the above methodologies are out of scope for this work and will not be expanded upon.

2.4.1 Buffered vs. Bufferless

Although many works have endeavoured to do so (e.g. [AZS$^+$08, KNS$^+$14, TAM$^+$19]), storing information in photonic form is currently challenging for practical amounts of time and in practical data volume [AKP20]. This means that the concept of *photonic buffering* is currently infeasible. This limitation has substantial consequences for the PIN paradigm, as PINs must either employ *bufferless* communication, or rely on electrical buffers and therefore become OINs. The latter option, as discussed, would entail including photonic transceivers every time information must be buffered, leading to the power and performance penalties discussed previously.

Although *bufferless* communication enables photonic communication, it also introduces a fundamental constraint in available switching paradigms. *Bufferless* communication systems preclude the use of packet switching, which ultimately relies on packets or packet segments being stored in buffers in intermediate nodes before they are forwarded to their destination. Based on this concept, flow control techniques (e.g. cut-through, wormhole, etc.), which have been developed to more efficiently share the network resources and buffering capacity, become impossible without buffers. Photonic communication must therefore rely on circuit-switching at the transmission level in order to remain *bufferless* and therefore energy efficient.

2.4.2 Circuit Switching

Historically, the Circuit Switching (CS) methodology evolved for telephony networks and found use in early telecommunication and interconnection networks. With CS, the source node issues a header packet which travels along the selected path (usually determined by the network controller), reserving the required network resources until reaching the destination. These reserved network resources are called a *circuit*. Data transmission then commences, with other requests which require the already-reserved resources being blocked while the circuit is being serviced. Once data transmission is complete, a tear-down packet is issued, allowing intermediate nodes to de-allocate the resources devoted to the circuit and prepare them for the next request. The shortcomings of CS in electronic networks such as unfair allocation of communication resources have led to the methodology being mostly superseded by packet switching.

In bufferless photonic communication using switching fabrics, intermediate nodes are photonic devices. These devices do not provision for packet reads as this would require detectors and backend circuitry, but are controlled by a centralised network controller, commonly using an FPGA device. Source nodes communicate with the network controller using a separate control network. With CS, a source node sends a path request to the network controller. The controller assesses the state of the photonic switching fabric and, if a path is available to the requested destination, sets the state of the switches within the fabric that correspond to the required path. It then sends a response to the source node to begin transmitting. Once the source node has completed transmission, it informs the network controller so that the reserved resources may be released.

2.4.3 Time Division Multiplexing

Time Division Multiplexing (TDM) is a communication sharing technique which was developed to overcome the inefficiencies of CS networks. Here, the channel capacity is partitioned into *time slots*, usually of a fixed length [DT04]. Information to be transmitted by nodes nodes across a shared medium is partitioned at each node into segments conforming to the time slot, and segments are transmitted using the process known as interleaving. The process of selecting which node may transmit is called arbitration, with round-robin, least frequently used or random selection being the most common arbitration strategies for TDM. These are investigated in detail in Chapter 7.

As is explained in chapter 6, in photonic switching fabrics that employ TDM, the time slot length is defined by the photonic device switching time, and the routing algorithm solving time. The photonic device switching time is defined by the device type and tuning strategy, as different devices employ different tuning strategies. For example, MRR-based switching devices commonly employ thermal tuning to induce switching through the thermo-optic effect, which is done in the μs range. MZI-based devices that employ electrical tuning to induce switching through the electro-optic effect switch in the *ns* range, allowing for shorter time slots. However, the TDM time slot must also be long enough for the employed routing algorithm to compute the state of the switching fabric, as the state computation for the next time slot must be completed during the current time slot at any given time. It is noted that this is not the general case in switching fabrics that employ TDM, where switch states that align to input-output permutations can be pre-computed and stored ahead of time in the switch controller memory; in these cases, the switch controller must retrieve the appropriate switch state from memory based on the input-output permutation to be serviced. However, when employing routing algorithms which route paths through the switching fabric *based on the current network state*, as is the case with some of the routing algorithms considered in this work, routing computation must be done on-line and therefore solving time must be incorporated within the TDM timeslot. The routing algorithm is also dependent on the switch fabric topology, with rearrangeably-non-blocking topologies generally requiring more computationally complex (and therefore time-consuming) routing algorithms.

Cheng *et al.* reason that, in the case of photonic switching fabrics within datacenters, reducing the tuning time of active photonic devices below the μs range provides less benefit than is widely considered by the community [CRBB18]. They base their reasoning on the following arguments. Firstly, they consider a datarate per wavelength of $\sim 25Gbps$, which is needed for a $\sim 100Gbps$ CWDM4 optical link. At that datarate, the bit error rate (BER) is highly sensitive to phase variations within the link. Therefore, if the link contains active photonic elements which induce phase variations (such as when including a photonic switching fabric between the modulator and the receiver bank), these must be calibrated to achieve a steady state before transmission; the required calibration time is referred to as "link training time". Secondly, they estimate that link training time is a process that takes at least $10-100ns$ [CRBB18]. Based on these aspects, they argue that device switching time below $100ns$ would be less beneficial than expected, as this would be dominated by the link training time. However, on the one hand their reasoning disregards demonstrations of sub-nanosecond ($< 600ps$) clock and data recovery locking [CBB$^+$18]. This reduces the timing penalty of link training substantially enough for *ns*-scale switching to become realistic, as shown by Benjamin *et al.* [BGL$^+$20]. On the other hand, reducing the per-wavelength data rate would reduce the sensitivity to phase variations; high aggregate data rates can then be obtained by increasing the number of wavelengths per link, with the added benefit of higher energy efficiency albeit at a greater hardware cost at the transceiver. Switching fabrics based on broadband photonic switches such as silicon photonic MZIs are instrumental to adopting this paradigm, as they can switch multiple wavelengths concurrently without incurring additional hardware costs, by using WDM.

2.4.4 Wavelength Division Multiplexing

Wavelength Division Multiplexing (WDM) is enabled by the fact that a single photonic medium (waveguide or fibre-optic cable) can carry multiple wavelengths of light concurrently, with little to no interference between them. These wavelengths can be transmitted, modulated upon and detected independently from each other using wavelength-selective filters and can carry data at different bit-rates or encodings. This aspect of optical and photonic communication is one of the main benefits, as it can lead to simplified network design, reduced hardware cost and increased energy efficiency while at the same time offering increased bandwidth capacity through higher spectral efficiency in photonic links.

Coarse Wavelength Division Multiplexing (CWDM) refers to the sparse use of a few wavelength channels (commonly 4 or 8) in the C or O communication bands [ITU12]. Information is multiplexed on the channels and then propagated across a common link, similar to a network bus. These channels differ substantially in wavelength and therefore show very little crosstalk interference between them, leading to increased data rate per channel [LVVR+19b]. This has led to the adoption of CWDM in commercial optical links based on transceivers; 100Gbps links that utilise four NRZ-modulated wavelengths at 25Gbps per wavelength (known as CWDM4) find widespread use in DCs and HPCs (circa 2008) [BP20]. 200Gbps CWDM4 links that employ PAM4 encoding are starting to be adopted, while 400Gbps links employing high Baud rate PAM4 are in development (circa 2020) [BP20]. These advances in link data rate are enabled primarily due to the use of SiPh technology, which allows for higher integration density in the hardware while allowing the transceiver form factor to remain compact.

However, scaling the CWDM channel data rate also has its limits. Increasing data rates lead to increased energy and thermal dissipation as well as increased complexity in the optoelectronic backend [LVVR+19b]. While higher order signal coding such as PAM4 or quadrature-based coding has also been considered to further increase bandwidth without scaling the channel data rate, it is more susceptible to interference effects which limits their use in switching fabrics.

Dense Wavelength Division Multiplexing (DWDM), on the other hand, employs a comparatively increased number of wavelength channels (10s - 100s) to increase spectral efficiency. The employed wavelengths are aligned to frequency grids, with the 100GHz and 50GHz grids being considered the standard. A tighter grid allows for more DWDM channels but can increase inter-channel interference in the form of crosstalk.

In the context of photonic networks, DWDM is encountered at the link level in bus based wavelength routed systems, especially Optical Networks-on-Chip (ONoCs) such as [VSM+08] or [PKK+09]. In these, source and destination pairs communicate using distinct wavelengths via a waveguide bus. DWDM can also be employed at the connection level when bandwidth-transparent photonic hardware is used. Here, the connection (circuit) carries data modulated on multiple wavelengths concurrently. By doing so, data rates per wavelength can be reduced while maintaining high aggregate data rates in order to decrease connection length and increase energy efficiency. Adopting connection-level DWDM also means that simple coding schemes such as OOK employing Return-to-Zero or more commonly Non-Return-to-Zero (NRZ) can be used while maintaining high aggregate data rates. This is advantageous as it decreases the complexity of the optoelectronic backend. This is, in fact, the multiplexing scenario considered in the experimental component of this thesis, where 32 wavelengths modulated using OOK coding are assumed per connection.

2.5 Silicon Photonic Devices

Photonic devices are the building blocks of silicon photonic pathways found in PINs. As photonics is being considered for deployment ever closer to the source of computation and memory storage, Silicon Photonics (SiPh) is particularly attractive due to the compatibility of fabrication processes with CMOS fabrication [NFA13], facilitating adoption cost for chip-level, interposer-level or board-level photonics. As a material, silicon has a very high refractive index (approx. 3.48 at 1550 nm [Lee16]) at optical communication wavelengths, leading to light modes being closely confined within the carrier medium which reduces optical losses. It also exhibits a thermo-optic and electro-optic response [PL$^+$16]. These properties mean that silicon is a favourable material for both waveguide and switching device construction. However, it is noted that as silicon is an indirect band-gap material, challenges exist with producing high-efficiency silicon lasers.

This section details the principal components forming silicon photonic pathways. It is noted that each subsection here is a research domain in itself and therefore a systematic review of each is out of scope for this thesis. However, as end-to-end energy efficiency is a primary aim for scaling PINs and is directly affected by the device level, a background in silicon photonic devices is discussed to provide context and construct a holistic view of the optical interconnection system in section 2.9.

2.5.1 Lasers

Lasers are the most commonly considered light source for photonic interconnects, although research is being conducted on the use of other light sources, e.g. LEDs. For example, Bie *et al.* report on a promising $MoTe_2$ based LED/photodetector devices [BGH$^+$17], which could allow for direct on-chip WDM. They also claim that the technology can be used to fabricate narrowband lasers with very high coupling efficiencies.

Current laser research generally falls within two categories; either the use of on-chip, or off-chip lasers. Off-chip lasers are generally more efficient in terms of wall-plug efficiency, but must be powered on for the full operation of communication irrespective of the sending endpoint. For an in-depth discussion on laser source efficiency for Optical Networks-on-Chip (ONoCs), the reader is referred to [WNL17], while prominent examples of off-chip lasers can be found in [WVM$^+$17].

On-chip lasers are in principle very desirable for on-chip optical communication, as they could be powered on-demand and used for direct modulation onto the optical carrier. Silicon, however, is an indirect band-gap material, meaning that its structural properties (misalignment of free electrons in the conduction band to free holes in the valence band) do not make it a good gain medium for light emission. The use of silicon as a gain medium for lasers is a highly active research domain which is, however, out of scope for this work; the reader is referred to the detailed discussion by Liang and Bowers [LB10]. As such, the complementary use of more exotic materials is required [WAD$^+$17] [BHG$^+$18] [DOJ$^+$16], which in turn creates a range of problems, most importantly with regards to laser efficiency and wafer yield [CLW$^+$16] [ZYM15]. It is noted that on-chip lasers are a very active research topic; prominent novel example technologies which show promise are VCSEL lasers and Transistor Lasers; discussions on these can be found in [TLN$^+$18] and [TFH13] respectively. Nevertheless, most proposed systems to date, including the interconnection network examined in this thesis, consider off-chip laser sources.

As this work focuses on the control and routing of photonic switch fabrics and not on the light source, the use of an off-chip comb laser such as [LLS$^+$18] is assumed. These laser sources produce high quality multi-wavelength beams with a high degree of uniformity in per-wavelength optical power.

2.5.2 Modulators

In order for information to be transmitted in the photonic domain, it must first be modulated onto a carrier beam and with a predefined encoding; this is done using *modulators*. Modulation can either be performed on the whole carrier beam [SB07] or on a per-wavelength basis [DOG$^+$12]. The latter technique enables DWDM, as discussed in section 2.4.4.

Modulation can be performed by the laser itself in the case of an on-chip laser source(referred to as direct modulation); however this entails limitations in achievable data rate. Further, this technique does not lend itself to per-wavelength modulation which is beneficial for high bandwidth density [SXH$^+$15]. Direct-modulated on-chip laser sources do not allow for per-wavelength modulation in the case of comb sources; in the case of single-wavelength sources such as VCSELs, the required number of lasers scales with the number of channels, which increases packaging complexity. The most common approach for achieving high bandwidth density is therefore to use external modulation, in which modulator devices modulate the communication traffic onto the carrier beam after it has egressed from the laser source. Example modulator

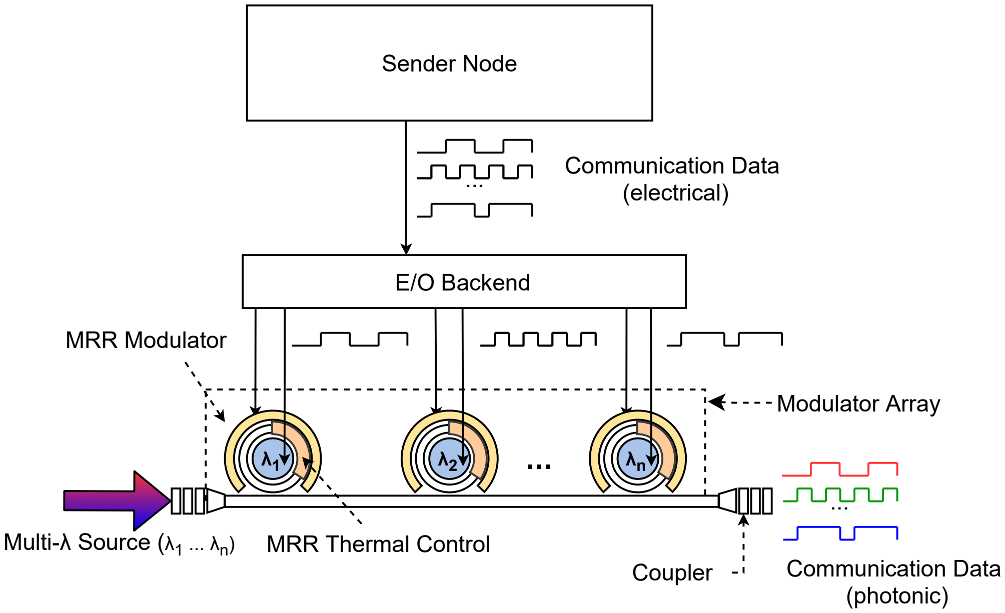


Figure 2.2: Structure of photonic signal generation based on MRR modulators.

devices are MRRs (discussed in sec. 2.5.4) which employ carrier injection or depletion [LVVR+19a], or MZIs which use interference effects.

MRR-based modulators are comparatively smaller in footprint and require less driving voltage than MZI-based ones, leading to their wider consideration for energy efficient DWDM when combined with multi-wavelength sources such as comb lasers. As such, in DWDM many cascaded wavelength-selective MRR modulators are employed, each modulating on a single wavelength to achieve high aggregate data rates at low energy consumption [XLDD+18]. This is depicted in Fig. 2.2. Here, a multi-wavelength source is coupled into the modulator array; each MRR modulator is fed information from the E/O Backend and modulates it on a specific wavelength. The multi-wavelength beam is then coupled into the transmission medium.

In terms of encoding schemes, many works consider simple On-Off Keying schemes (OOK) [GLM+11]. While higher order modulation formats such as PAM4 have been considered (e.g. [NFF+18, Jon19]), they require hardware provisioning for additional digital signal processing and forward error correction, leading to significant energy consumption and latency overheads [LVVR+19a].

As this thesis focuses on the routing and control of photonic switch fabrics, a systematic review of modulator technology is out of scope. For the simulation work in this thesis, OOK-encoded DWDM through the use of cascaded MRR modulators and a multi-wavelength laser source (e.g. a comb laser [LLS$^+$18]) are assumed. These assumptions correspond to those made in the DSENT PIN simulator [SCK$^+$12], whose laser model is incorporated in to the simulation work in later chapters.

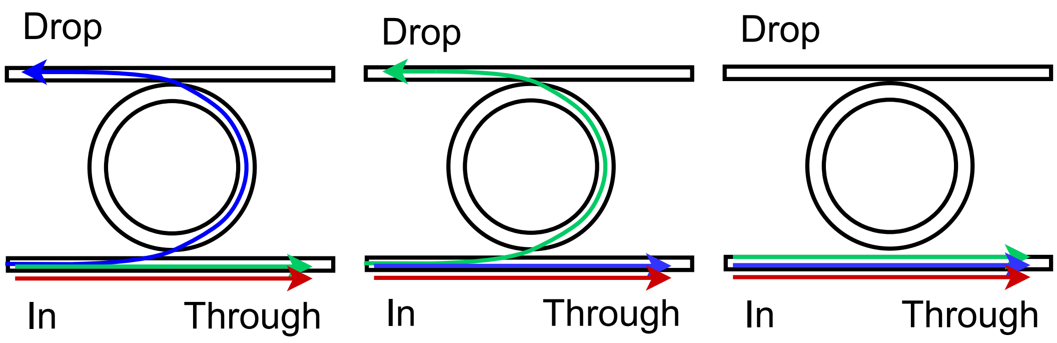
2.5.3 Waveguides

Waveguides are the optical equivalent of the electrical wire and enable the most prominent paradigm shift in interconnects, namely the ability to send multiple information streams in parallel over the same physical medium in an energy efficient and relatively distance-independent fashion. Waveguide technology for PINs is widely studied and a wide variety of demonstrations have been put forth recently. For instance, in this thesis a conservative waveguide propagation insertion loss penalty of 1.18 dB/cm is considered [LZZ$^+$16]; Thraskias *et al.* on the other hand mention waveguide-incurred insertion loss of as low as 0.2 dB/cm [TLN$^+$18]. It is noted that propagation loss due to waveguides is highly dependent on device technology; nevertheless, the aforementioned survey illustrates the rate of progress on the technology front. As is shown in chapter 4, other factors that contribute to the insertion loss penalty (i.e. number of waveguide crossings, MZI states) have a much greater impact on the scalability of switch fabrics than the propagation loss due to waveguides.

2.5.4 Switching Devices

Switches are another fundamental building block of PINs, with various types being explored throughout the literature. A comprehensive review of SiPh switches can be found in [CBG$^+$18].

Most commonly, optical switching elements are based on microring resonators, whereby organisations of MRRs are coupled to waveguides to form N×N switching elements. Most MRR implementations are wavelength selective [LZZP15], although multi-wavelength MRRs have been reported (e.g. [LBSD$^+$09] [BDL$^+$07]). Due to this aspect, they are usually used for wavelength-routed photonic networks, with cascaded arrays of MRRs being used for multi-wavelength switching. However, this approach entails a number of drawbacks, such as increased area and MRR tuning-induced power consumption, both of which are problematic with respect to scalability in constrained interconnects such as ONoCs [WNL17]. Additionally, MRR-based switching elements have limited bandwidth, since only a small subset of wavelengths is used between each pair of endpoints. Other drawbacks that constrain scalability include increasing wiring



(a) MRR Switch switching the "blue" wavelength.

(b) MRR Switch switching the "green" wavelength.

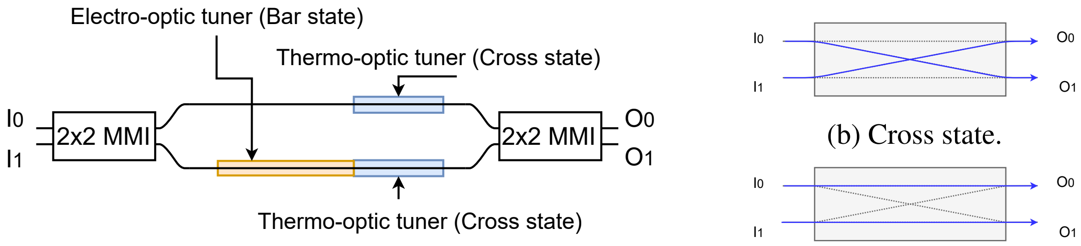
(c) MRR in the through state, switching no wavelengths.

Figure 2.3: A 2×2 wavelength-selective switch formed with an MRR. Image adapted from [BCB+14].

complexity in the device electronic control circuitry as well as increased packaging costs. A 2 × 2 wavelength-selective switch formed with an MRR in its operation states is depicted in Fig. 2.3.

More recently, Mach-Zehnder Interferometers (MZIs) have been considered for optical switching. MZI switches trade off wavelength selectivity for reduced tuning power and wiring complexity compared to MRR switches [SXH+15] by switching all incoming wavelengths simultaneously. They are composed of two 2 × 2 couplers, either directional couplers or multi-mode interferometers (MMIs), and two waveguide arms, either of equal or different length. Stand-alone MZIs can be used as 2×2 switching elements which are then organised in a multi-stage fabric such as the one examined in this thesis. MZI switches normally have two states; "cross" and "bar". More novel MZI designs include a third, blocking state, which is aimed at reducing optical losses between the stages [LCM+17]. Optical losses will be discussed in detail in Section 2.6. Another approach that has been put forth is to use nested MZI organisations which can compose higher-radix base switching elements [DRS+16].

The operating principles of MZIs are the following. Light is coupled into one or both of the input ports of the MZI, and is split by the coupler (MMI or DC) into the two arms. Heat or electricity is applied to one or both of the arms; this induces a phase shift in the light present on the arm due to the thermo-optic or electro-optic effect. This is known as thermal or electrical tuning. With proper tuning, the light beam components on either arm acquire a phase shift of π relative to each other. As they then traverse the outgoing coupler, the beam components interfere constructively or destructively,


Electro-optic tuner (Bar state)
Thermo-optic tuner (Cross state)
I0
I1
2x2 MMI
2x2 MMI
O0
O1
Thermo-optic tuner (Cross state)
I0
I1
O0
O1
I0
I1
O0
O1

(a) Schematic adapted from [LZZ$^+$16]

(b) Cross state.

(c) Bar state.

Figure 2.4: A 2×2 EO/TO MZI switch with associated states.

thereby egressing the device on one of the two outgoing ports.

MZIs are commonly switched by means of either thermal or electrical tuning. Thermo-optical (TO) tuning has a relatively slow response (μs scale [DCY20]) but incurs less optical loss. Electro-optical (EO) tuning is much faster; for example, [GLZ$^+$17] report an EO switching time in the order of a few ns. However, due to effects such as free carrier absorption, EO tuning entails a reduced tuning range compared to TO [DL17]. As discussed in later chapters, deploying MZI-based switches on-chip or in high-performance scenarios such as in TDM multi-stage switching fabrics for DC top-of rack requires sub-μs switching time, which cannot be accommodated with TO tuning. For this work, a combination of EO and TO tuning as proposed in [LZZ$^+$16] is assumed. In this model, TO tuning is used to compensate for fabrication defects and reach MZIs to the "cross" state, while EO tuning is applied to switch the state to "bar". This is depicted in Fig. 2.4.

There are two commonly used ways to induce switching in EO/TO-tuned MZIs; *single-ended* tuning, or *push-pull*. In *single-ended* tuning, a phase difference of π between the two MZI arms is reached using the tuners of one arm. Here, the bias point of the MZI, which is controlled by the thermal tuners, is set to one of the two MZI states. Electrical tuning is then used to switch state. On the other hand, *push-pull* drive is used to reduce the imbalance in optical losses between states. Here, thermal tuning is used to set the bias point of MZIs to their quadrature point. Then, electrical tuning is used to provide a $\frac{\pi}{2}$ phase shift on both arms, thus reaching the desired relative phase difference of π between the two arms. Examples of push-pull driven MZI switches have shown reduced optical losses compared to single-ended tuning [DLR$^+$15b, DLR$^+$15a, HCB18]. The optical loss reductions afforded by push-pull drive may prove to be instrumental to increasing the scale (i.e. port count) of multi-stage switching fabrics, as exemplified in [QTC17]. However, push-pull drive entails an increased wiring and control complexity due to the presence of more tuning pads; due to this and to comparatively simpler control methodology, MZIs operated with single-ended drive are selected for the simulation-based studies presented later.

Another device type which has shown to be a promising candidate for implementinc photonic interconnection networks is the Arrayed Waveguide Grated Coupler (AWGR). AWGRs are passive photonic devices which provide $N \times N$ connectivity through wavelength permutation, that is each input permutes N wavelengths to N outputs, making them an interesting candidate for wavelength routing ($\lambda - routing$) designs, as exemplified by Werner *et al.* [WFP+18]. DWDM can be employed in AWGRs using the following feature: the permutation of wavelengths from an input to N outputs is *cyclical with the free spectral range*, meaning that multiple wavelengths following a period defined by the free spectral range can be routed to a single output. This is discussed by Fotouhi *et al.* [FWP+19] and experimentally demonstrated by Grani *et al.* [GLPY17]. In contrast to AWGRs, MZIs such as the ones considered in this work implement space switching to entire spectral segments (rather than individual wavelengths that follow a periodicity defined by the physical properties of the device). This aspect of MZIs can, in principle, provide a much greater degree of flexibility for the network designer in selecting the number of employed wavelengths and effective data rate per wavelength, thereby potentially leading to high bandwidth density per input port.

2.5.5 Detection

Detection of photonic signals occurs using photodetectors, which are commonly fabricated using germanium or semiconductor compound (III-V) materials [CVF+19, LRARK17], although more recently materials such as graphene have been considered [WG17]. In WDM contexts, multiple photodetectors are fed by individual, wavelength selective filters (usually MRR-based). As photodetectors generate current far below the threshold required by electronic receivers [KB12], trans-impedance amplifiers are usually employed between the photodetector and the receiver. Electronic signals are then subjected to de-serialisation and clock data recovery, after which the detection process is complete. A diagrammatic description of the process is shown in Fig. 2.5.

In terms of photonic power budget, the most important metrics are the following. Responsivity of the photodetector determines the lower limit of detection and, together with the extinction ratio of the modulator, the electrical properties of the electronic

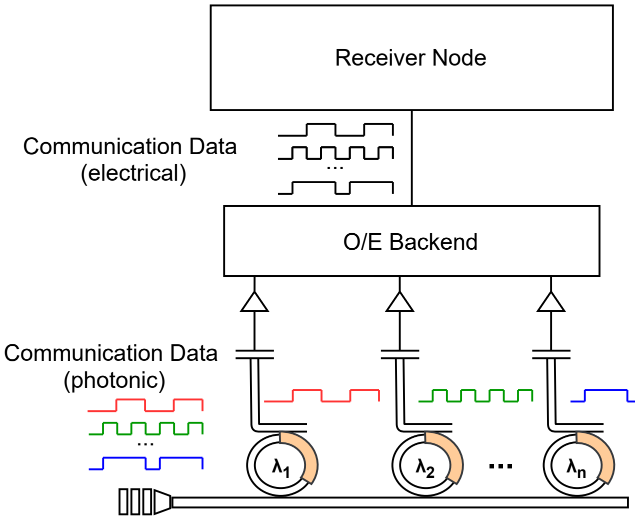


Figure 2.5: Structure of a photonic receiver based on MRR filters.

backend and the data rate, define the receiver sensitivity [SCK+12]. The receiver sensitivity in turn defines the required signal power at the input, when combined with the optical losses of the path. Inter-channel crosstalk, generated between successive MRR filters, is also an important metric in WDM. Inter-channel crosstalk here is dependent on the free spectral range of the rings as well as the channel spacing. For channel spacing which conforms to the 100GHz spectral grid standardised under the ITU-T G.694.1 standard [ITU12] as is considered in this work, inter-channel crosstalk in MRR filters is considered negligible, with much more significant intra-channel crosstalk being produced in the switching fabric.

2.6 Photonic Losses & Power Penalty

The detractive effect that accumulative photonic losses have on the power of a laser carrier beam after it has traversed a photonic pathway is called the photonic power penalty. This penalty reduces the power of the laser beam at the receiver, and must be compensated for either through higher input power or through amplification in order for the information carried by the beam to be detected at a low bit error rate (BER). The photonic losses that cause the power penalty are accrued when the beam traverses passive and active devices as depicted in Fig. 2.6, and also through interference effects with other laser carrier beams such as photonic crosstalk. Here, the carrier beam is inputted at P_{in}. It then accrues losses and therefore power penalty as it traverses the components of the optical pathway; the optical power at the receiver is $P_{out} < P_{in}$, as the

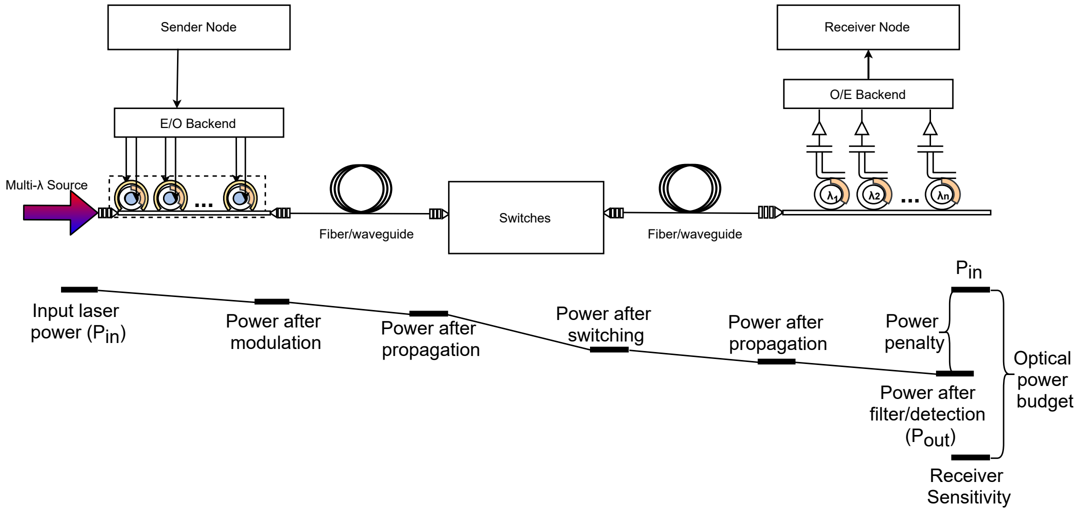


Figure 2.6: Evolution of the power penalty in a photonic pathway.

carrier beam suffers photonic losses. P_{out} must be greater than the receiver sensitivity, which forms the lower bound of the optical power budget. The two main contributors to the photonic power penalty are called insertion loss and crosstalk. Crosstalk is differentiated into inter-channel and intra-channel crosstalk below.

The equations in this section are adopted from the work of Ramaswami *et al.* [RSS09], and use the following notation. It is considered that a device or network has at least one input port i and at least one output port j. The laser carrier beam enters the device or network from i and exits from j, forming a lightpath. P^i_{in} is the carrier beam power (in W) at the device or network input i, P^j_{out} is the beam power (in W) at the output j.

2.6.1 Insertion Loss

Insertion loss is the degradation of beam power as a device or network is traversed and is expressed in decibels (dB) as a ratio of output to input power:

$$IL_{i,j} = -10log(P^j_{out}/P^i_{in}) \tag{2.1}$$

Although insertion loss represents a reduction in power and therefore should be of negative sign when expressed in dB, it is conventionally expressed as a positive number and then subtracted from input power to calculate the power penalty.

2.6.2 Crosstalk

At the photonic signal level, crosstalk in a photonic device is defined as “the general term given to the effect of other signals to the desired signal” [RSS09]. Here, a desired photonic signal is defined as a collection of information that is transmitted over a carrier laser beam that uses a lightpath. In crosstalk analyses, this signal is commonly termed the “victim”, while the output it egresses from is termed the “victim port”. The signal is modulated on at least one channel (or wavelength); interference effects of other signals, or “aggressors”, on the victim signal are therefore split into two categories, “intra-channel” and “inter-channel” crosstalk, with “intra-channel” crosstalk having a more severe impact [ZCC$^+$96]. The severity of crosstalk effects is also dependent on the relative phase difference between the victim and aggressor, as well as their relative polarisation difference. Phase and polarization differences are discussed in depth in [ZCC$^+$96], where they determine that crosstalk effects are most severe when the victim and aggressor signals are co-polarised and out of phase. In practice, these effects vary depending on thermal conditions, laser output conditions (e.g. laser frequency drift [DL17]) and over time. Most analyses, however, assume and optimise against worst-case conditions, as is done in this work.

In photonic devices, crosstalk arises due to the imperfect isolation present between the output ports, due to effects such as free-carrier absorption or non-linear effects such as four-wave mixing [ZXS$^+$17]. In networks of photonic devices, a crosstalk signal *cascades* from one device to another along the direction of propagation and generates more crosstalk, thereby interfering with other signals, other crosstalk signals or delayed and attenuated versions of itself. Crosstalk is split into *orders* of crosstalk. A crosstalk signal originating from a carrier beam is considered first-order crosstalk. A crosstalk signal originating from first-order crosstalk is considered second-order crosstalk and so on. It is for this reason that interconnected photonic devices must be designed to exhibit the smallest crosstalk ratio possible to their ports; the larger the crosstalk ratio, the greater the impact of higher crosstalk orders on the aggregate crosstalk at the network endpoint and therefore the smaller the Optical Signal-to-Noise Ratio (OSNR). The smaller the OSNR, the worse the achievable BER [ZXS$^+$17].

2.6.3 Inter-channel Crosstalk

In WDM networks, the operation wavelength channels are selected with respect to a frequency grid. This will be discussed further in section 2.4.4 but, for the purposes

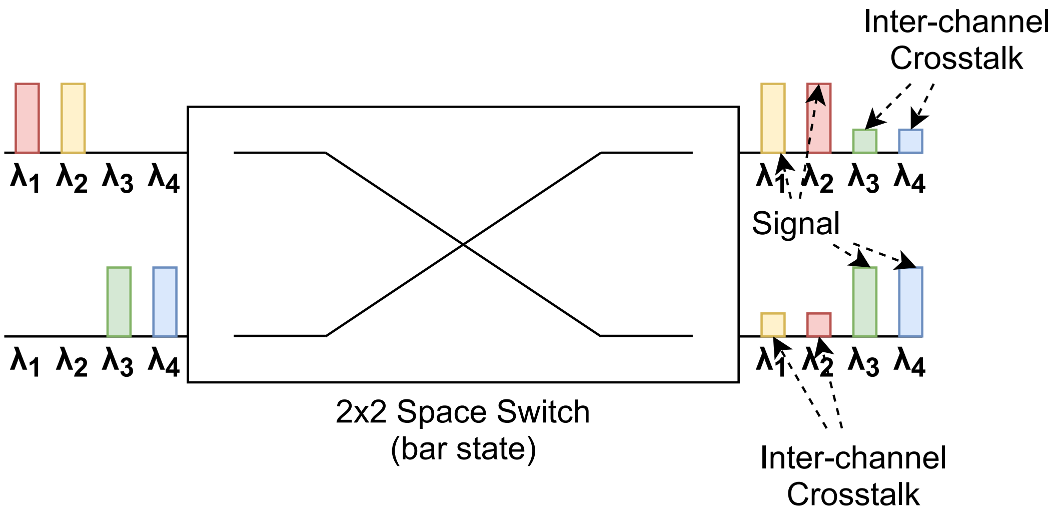


Figure 2.7: Inter-channel crosstalk in a 2×2 Space switch in the bar state.

of this section, it is sufficient to state that the more channels employed, the smaller the difference is in wavelength ($\delta\lambda$) between two adjacent channels. A smaller ($\delta\lambda$) correlates with increased inter-channel crosstalk.

Inter-channel crosstalk in WDM systems is defined as the effect of an aggressor signal on the victim signal when these signals are at a sufficiently different wavelength that the difference is larger than the receiver's electrical bandwidth [RSS09]. For non-amplified networks such as the ones examined in this work, the power penalty in dB from inter-channel crosstalk can be expressed as follows:

$$PP_{XT\_inter\ i,j} = -10log(1-\varepsilon) \tag{2.2}$$

Here,the crosstalk coefficient at output j is $\varepsilon = P^{j}_{leak}/P^{i}_{in}$, that is the ratio of the leakage power at output j to the input power at input i. For this equation to be representative, it is identified that $\varepsilon \ll 1$, with typical values found in [RSS09]. In the case of inter-channel crosstalk, if there are L leakages at an output, ε is given by the following:

$$\varepsilon = \sum_{i=1}^{L} \varepsilon_i \tag{2.3}$$

and $\varepsilon_i = P^{j}_{leak}/P^{i}_{in}$.

In practice, this type of crosstalk is much less detrimental than intra-channel and, in wavelength-routed networks, can be filtered out at the receiver thereby completely negating its effect. In WDM networks where the desired information is carried by

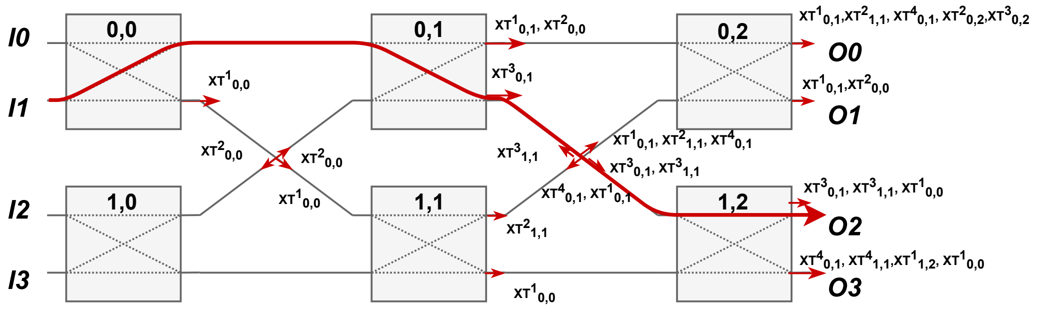

I0
I1
I2
I3
0,0
1,0
0,1
1,1
0,2
1,2
XT$^1_{0,0}$
XT$^2_{0,0}$
XT$^2_{0,0}$
XT$^1_{0,0}$
XT$^1_{0,1}$, XT$^2_{0,0}$
XT$^3_{0,1}$
XT$^3_{1,1}$
XT$^2_{1,1}$
XT$^1_{0,0}$
XT$^1_{0,1}$, XT$^2_{1,1}$, XT$^4_{0,1}$
XT$^3_{0,1}$, XT$^3_{1,1}$
XT$^4_{0,1}$, XT$^1_{0,1}$
XT$^1_{0,1}$,XT$^2_{1,1}$, XT$^4_{0,1}$, XT$^2_{0,2}$,XT$^3_{0,2}$
O0
XT$^1_{0,1}$,XT$^2_{0,0}$
O1
XT$^3_{0,1}$, XT$^3_{1,1}$, XT$^1_{0,0}$
O2
XT$^4_{0,1}$, XT$^4_{1,1}$,XT$^1_{1,2}$, XT$^1_{0,0}$
O3

Figure 2.8: Intra-channel crosstalk in a 4 × 4 switching fabric.

multiple wavelengths simultaneously, inter-channel crosstalk cannot be filtered out. It is, however, very low compared to intra-channel crosstalk [ZCC+96]. An example of inter-channel crosstalk in a 2 × 2 space switch (e.g. an MZI) in the bar state is depicted in Fig. 2.7. Here, each input port serves a different set of wavelengths; λ_1 and λ_2 ingress at I_0 and egress at O_0, while λ_3 and λ_4 ingress at I_1 and egress at I_1. Leakages of the wavelengths accumulate at the victim ports.

2.6.4 Intra-channel Crosstalk

Intra-channel, or coherent crosstalk, is the effect of an aggressor signal on the victim where they are at the same wavelength or when their wavelength difference is within the receiver's electrical bandwidth. The power penalty in dB from intra-channel crosstalk is expressed as such for non-amplified systems:

$$PP_{XT\_intra\ i,j} = -10log(1-2\sqrt{\varepsilon}) \tag{2.4}$$

In intra-channel crosstalk, $\sqrt{\varepsilon}$ is given by:

$$\sqrt{(\varepsilon)} = \sum_{i=1}^{L} \sqrt{(\varepsilon_i)} \tag{2.5}$$

and $\varepsilon_i = P^j_{leak}/P^i_{in}$.

Intra-channel crosstalk has a much more severe effect on the photonic power penalty than inter-channel. Cascading intra-channel crosstalk is particularly problematic for networks such as photonic switching fabrics as has been frequently identified by the community [GEE94] [TOT96] [HH90]. This type of crosstalk is depicted in Fig. 2.8, which exhaustively shows the crosstalk terms accrued at each device output in a simple example. Here, six 2 × 2 MZIs and two waveguide crossings are connected in the

Beneš topology (detailed in the next section), forming a 4×4 switching fabric.

Here, a single-wavelength beam traverses from input 1 (I_1) to output 2 (O_2) in a 4×4 switching fabric composed of 2×2 MZIs and MMI-based waveguide crossings. The MZIs are numbered with respect to the row/column they belong to, and MZIs that are not traversed by the lightpath are set to the cross state.

Crosstalk terms are expressed as $XT^k_{l,m}$; l,m denotes the MZI in which the crosstalk term was generated through leakage, while k denotes the crosstalk order. For example, $XT^1_{0,0}$ is first order crosstalk that is leaked from MZI $0,0$, while $XT^2_{0,0}$ is second order crosstalk, generated by $XT^1_{0,0}$ as it traverses the waveguide crossing. The total power of crosstalk terms present at the network outputs (O_j) is then the leakage power, P^j_{leak}.

As can be seen, intra-channel crosstalk cascades from one photonic device to the other. It therefore accumulates across the direction of light propagation and generates more crosstalk at the victim ports of each traversed device. This can be clearly seen at output O_3, where two first-order crosstalk terms accumulate; one generated from the lightpath traversing MZI $1,2$ in the bar state ($XT^1_{1,2}$) , and one from it traversing MZI $0,0$ in the cross state ($XT^1_{0,0}$). Note that crosstalk terms degenerate in power as they traverse the network similarly to the lightpath; $XT^1_{0,0}$ at O_3 will be reduced due to insertion loss, relative to $XT^1_{0,0}$ at MZI $0,0$.

In practice, higher orders of crosstalk have a very small power level. As a hypothetical example, first assume an input signal power of $1mW$ ($0dBm$) and a crosstalk ratio of $-30dB$ for waveguide crossings and each MZI state. In this case, $XT^1_{0,0} \approx 1\mu W$ at MZI $0,0$, while $XT^2_{0,0} < 1nW$ at the first waveguide crossing. Higher orders of crosstalk fall well below the detection limit of photodetectors. Assuming a crosstalk ratio of $-20dB$ however, $XT^1_{0,0} \approx 10\mu W$ while $XT^2_{0,0} \approx 100nW$. This shows the importance of optimising device design to reduce crosstalk. Nevertheless, unless the devices traversed have a very high crosstalk ratio, high orders of crosstalk (i.e. third-order and more) have a negligible impact on the final term and can be disregarded, at least for small networks.

It is also clear that the number of 1st and 2nd order crosstalk terms reaching the outputs scales with the number of traversed devices; scaling the network size therefore increases the leakage power level accumulated at the outputs. Additionally, a switching fabric is used to serve multiple lightpaths concurrently; if these lightpaths are on the same wavelength and assuming they are co-polarized and out of phase, intra-channel crosstalk from one lightpath will interfere maximally with all other lightpaths, thereby also increasing the aggregate leakage power.

2.6.5 Photonic Power Penalty

Based on the insertion loss and intra-channel crosstalk, the total photonic power penalty in dB affecting a carrier beam traversing a photonic switch from input i to output j can be estimated using the following formula:

$$PP_{i,j} = IL_{i,j} + PP_{XT\_intra\ i,j} \tag{2.6}$$

As explained, the photonic power penalty present in a photonic pathway must be compensated for by either increasing input signal power or amplification in order to be above the receiver threshold, thereby forming the power budget.

When considering networks such as photonic switching fabrics formed with MZIs in the Beneš topology, it will be shown in chapter 3 that both terms of the photonic power penalty are dependent on the path and state of the switch fabric as well as its saturation, that is how many photonic signals are present in the switch fabric at each time. The more signals present, the greater the probability of higher insertion loss paths being allocated to signals due to switch saturation and the more 1st and 2nd order crosstalk terms at the outputs.

2.7 Multi-Stage Switch Fabric Network Topologies

Switch fabrics formed with silicon photonic switching devices are a promising technology candidate for surmounting many technological challenges in electronic interconnection networks regarding concurrently increasing bandwidth and power/energy efficiency. As explained in section 2.5.4, the selection of switching device determines transmission parameters such as photonic power penalty due to device design, and bandwidth due to wavelength selectivity. As this work targets DWDM broadband photonic switching fabrics formed with MZIs, this section first describes the fabricated chip that has been chosen as a target for simulation and discusses the design trade-offs present. It then details other commonly selected network topologies and analyses the trade-offs involved with topology choice.

MZIs are typically formed as 2×2 switching devices. Although nested MZI structures providing three ports per side have been investigated, their additional ports and the corresponding additional states have been investigated for their crosstalk reduction potential, not for forming connections [LCM+17]. Therefore, switching fabrics composed from cascaded MZI stages are organised using topologies based on 2×2

Table 2.1: Most commonly adopted topologies for multi-stage switching fabrics based on 2 × 2 switches. RNB: Rearrangeably non-blocking, SNB: Strictly non-blocking.

Topology	**# Stages**	**# Switches**	**Max. # Crossings per path**	**Order of Crosstalk**	**Max. # Stages w. 1^{st}-order Crosstalk**	**Blocking Behaviour**	**# Stages in Path**[a]	**Path Diversity**
Banyan	log_2N	$\frac{N}{2}log_2N$	$N-log_2N-1$	First	log_2N	Blocking	log_2N	None
Beneš	$2log_2N-1$	$\frac{N}{2}(2log_2N-1)$	$2N-2log_2N-2$	First	$2log_2N-1$	RNB	$2log_2N-1$	$\frac{N}{2}$
Dilated Beneš	$2log_2N$	$2Nlog_2N$	$2\sum_{i=1}^{log_2N}\frac{N}{2^i}-1$	Second	0	RNB	$2log_2N$	$\frac{N}{2}$
N-Stage Planar	N	$\frac{N}{2}log_2N$	0	First	N	RNB	$\frac{N}{2}\rightarrow N$	$1\rightarrow N$
Crossbar	$1\rightarrow 2N-1$	N^2	0	First	$N-1$	SNB	$1\rightarrow 2N-1$	None
PILOSS	N	N^2	$N-1$	First	$N-2$	SNB	N	None
Switch and Select	$2log_2N$	$2N(N-1)$	$(N-1)^2$	Second	0	SNB	$2log_2N$	None
Double Layer Network	$2log_2N-1$	$(N\frac{5N}{4}-2)$	$3N-2log_2N-4$	First[b]	1	SNB	$2log_2N-1$	None

[a] The number of stages in a path is the number of traversed switches, which signifies the insertion loss.

[b] The DLN exhibits second-order crosstalk for all stages except for the middle stage, where it exhibits first-order.

switching devices. These topologies are either adapted from the electronics domain, or proposed specifically for photonics by trading off architectural aspects such as hardware complexity to reduce optical losses.

Depending on the number of cascaded switch stages and the logical connection pattern between stages, these topologies exhibit different levels of blocking behaviour. A non-blocking switch can simultaneously service all connections between inputs and outputs such that they form a permutation of inputs to outputs [DT04]. There are three classes of blocking behaviour for switches; *blocking*, where a switch cannot ensure accommodating a path from an input to an output without conflicts, *rearrangeably-non-blocking* (RNB), where a switch can route permutations but may require re-arranging previous connections to do so when the permutation is set up incrementally, and *strictly non-blocking* (SNB), where the permutation can be serviced irrespective of the set-up order. Blocking behaviour is a critical consideration for bufferless photonic networks such as multi-stage switches, as will be discussed in Section 2.4; the topologies examined here will therefore be ordered based on their blocking behaviour.

The most commonly used topologies can be found in Table 2.1. The Banyan topology shown in Fig. 2.9, proven by Wu & Feng to be isomorphic to the baseline, reverse baseline, omega, and indirect binary $n-$cube among others [WF80], represents a class of blocking multistage topologies. This topology offers full connectivity for minimum network diameter, meaning that any input can be connected to any output provided that no contention exists in the network and this can be done using the fewest 2 × 2 switches. However this is a *blocking* topology which, as discussed in Section 2.4 has important implications for routing and scheduling. Additionally, as this class of networks does not offer path diversity, path-dependent optimisations of the optical power

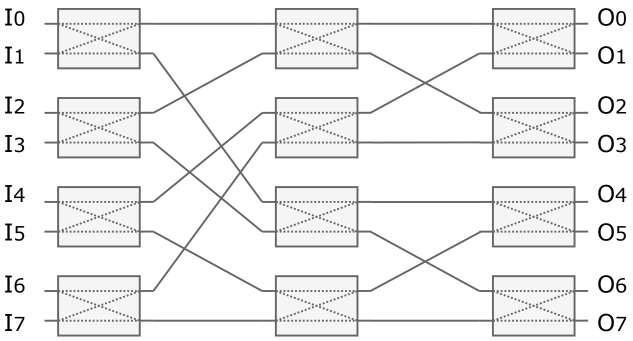


Figure 2.9: An 8 × 8 Banyan network.

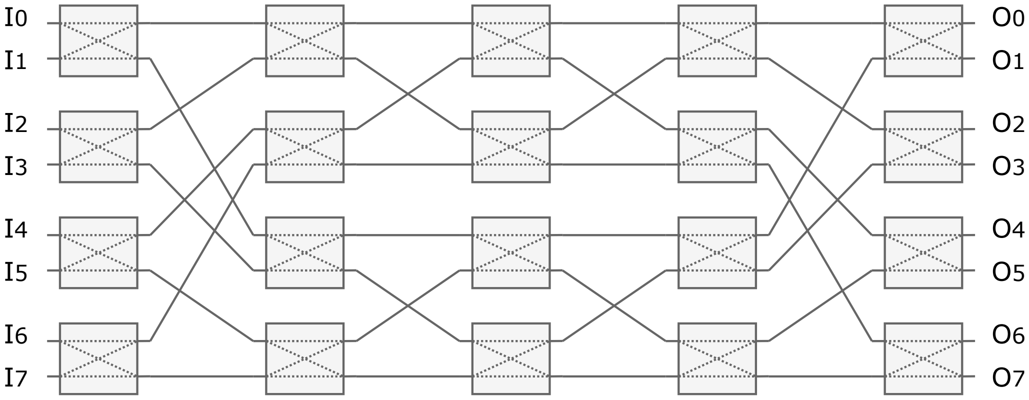


Figure 2.10: An 8 × 8 Beneš network.

budget such as those proposed in this thesis cannot be applied.

Rearrangeably non-blocking (RNB) topologies offer a compromise between a larger network diameter (i.e. more stages, more MZIs and therefore higher hardware complexity) and blocking characteristics, as they can route full permutations with the appropriate routing algorithm and where reconfiguration during transmission is allowed. The Beneš [Ben64] network (Fig. 2.10) is a specialized type of Clos network [Clo53]. It is the most widely adopted topology originating from the electronic domain, as it requires the fewest 2 × 2 switches in order to fully connect N inputs to N outputs in an RNB Fashion. This also means that it is the most scalable topology when considering hardware complexity. However, photonic losses such as ILoss and coherent crosstalk adversely impact their scalability, as the Beneš network suffers from first-order crosstalk. As this work focuses on MZI-based photonic switch fabrics adopting the Beneš topology, these effects shall be discussed in depth in the following chapters.

The dilated Beneš topology (Fig. 2.11) was proposed to surmount the challenges that first-order crosstalk poses in photonic Beneš networks [PN87]. This topology

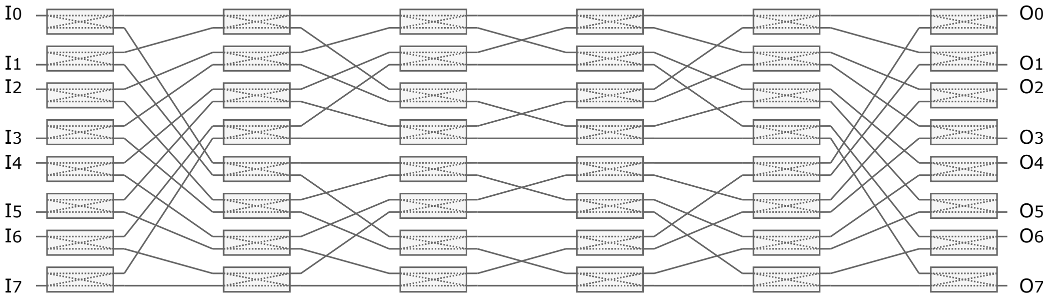


Figure 2.11: An 8 × 8 Dilated Beneš network.

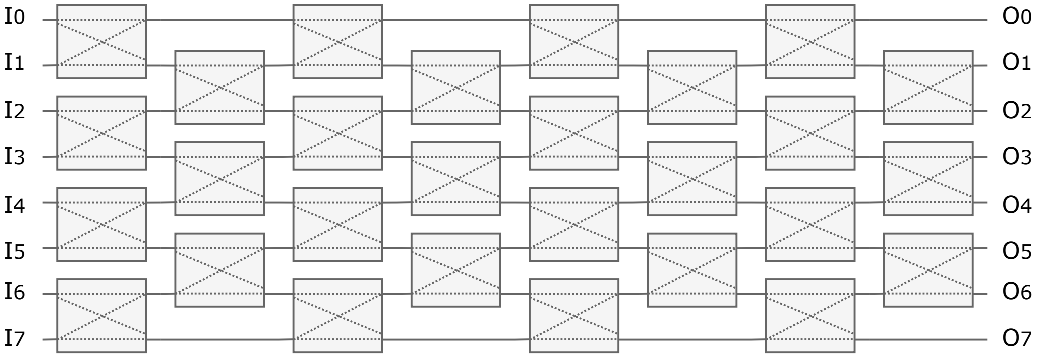


Figure 2.12: An 8 × 8 N-stage-planar network.

leverages *space dilation*, in which a base topology is augmented with more base switches per stage and half the inputs and outputs of the switch fabric are disconnected.

The N-stage-planar or Spanke-Beneš topology (Fig. 2.12) has also been frequently considered for its use in smaller scale switch fabrics [SB87]. The topology was considered as it avoids waveguide crossings, thereby reducing crosstalk and overall design complexity. However, it scales more poorly than the Beneš network in terms of switching elements, leading to increased ILoss. The photonic power penalty is also highly non-uniform in this topology, due to the variation in stages per path.

Strictly non-blocking topologies are attractive for their simple routing and non blocking characteristics relative to RNB ones. The most common topology inherited from electronics is the crossbar (Fig. 2.13). Although this topology does not require waveguide crossings in a planar layout, the number of switches and therefore hardware complexity scales quadratically with the number of endpoints. Also, the crossbar topology is susceptible to first-order crosstalk and shows high insertion loss, as well as non-uniform photonic power penalty due to the variable number of stages per path.

The PILOSS topology (Fig. 2.14) was proposed to reduce the dynamic range in path-dependent insertion loss [SHM87], as it offers a relatively uniform insertion loss

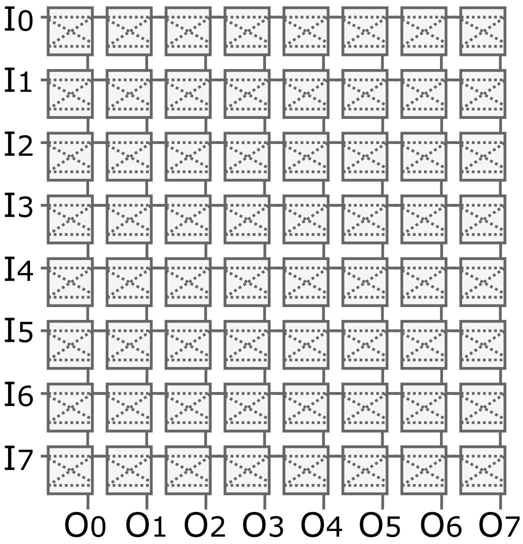


Figure 2.13: An 8 × 8 crossbar network.

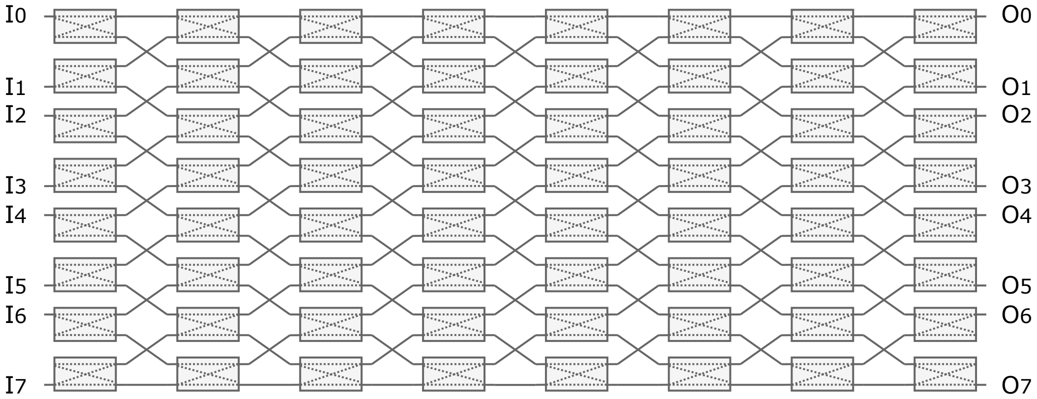


Figure 2.14: An 8 × 8 PILOSS network.

per path (same number of switches traversed). However, is also not immune to first-order crosstalk and scales poorly compared to RNB topologies. Although only half the input and output ports are connected in the PILOSS, it has not been formally classified as a *space-dilated* topology to the author's knowledge.

Space dilation has also been employed in topologies other than the dilated Beneš. Dilated topologies such as the Switch-and-Select [Spa86] and Double Layer Network (DLN) [LT94] in Figs. 2.15 and 2.16 have also been proposed for their crosstalk reduction properties, as they are immune to first-order crosstalk in the switches. These topologies exhibit tree-like characteristics, with a $1 \times N$ tree demultiplexer switch at

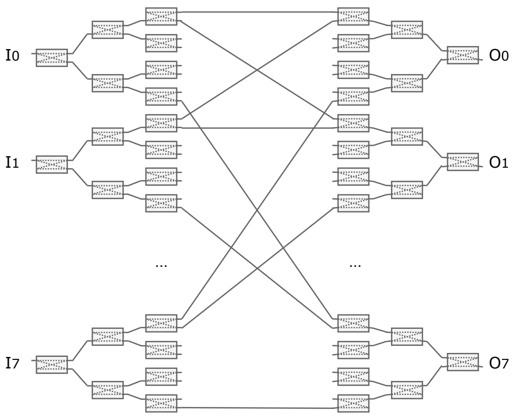


Figure 2.15: An 8 × 8 Switch-and-Select network.

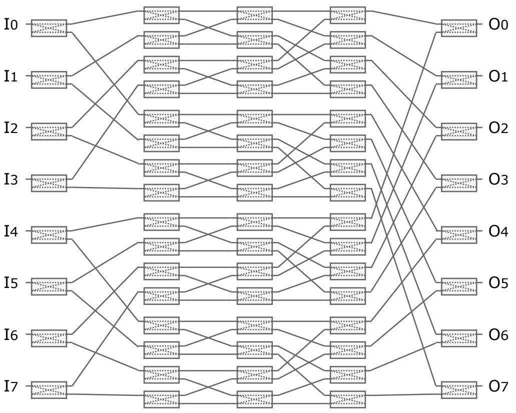


Figure 2.16: An 8 × 8 DLN network.

each input, a central shuffle stage and a $N \times 1$ multiplexer stage at the output, as mentioned by Lee and Dupuis [LD18]. The Switch-and-Select consists of two mirrored linear switching arrays with a perfect shuffle stage connecting them. The Double Layer Network is a recursive topology, with the input and output stages of a radix-N network being 2×2 switches with one used input or output respectively. The centre stage consists of four radix-$\frac{N}{2}$ DLNs. The DLN scales better than PILOSS in terms of stages having the same amount as the Beneš and is immune to first-order crosstalk except in the central stage. However it requires a larger amount of waveguide crossings than either the Beneš topology or PILOSS and the switch count scales quadratically, leading to scalability challenges for a larger switch radix.

In terms of insertion loss and crosstalk, the above topologies compare as follows based on the metrics shown in Table 2.1. The DLN arguably exhibits the best metrics in both categories, as it has the fewest stages in a path and only exhibits first-order crosstalk in the middle stage; however it does so with an increased footprint due to the number of switches. Note that the DLN also requires a comparatively large number of waveguide crossings for a planar layout, which increases both insertion loss and crosstalk slightly. The Switch-and-Select follows, with one extra stage compared to the DLN, which increases insertion loss; crosstalk however is reduced, as it completely isolates first-order crosstalk. Again, this topology comes at the expense of an increased footprint due to the number of switches and number of waveguide crossings, which also increase insertion loss and crosstalk slightly.

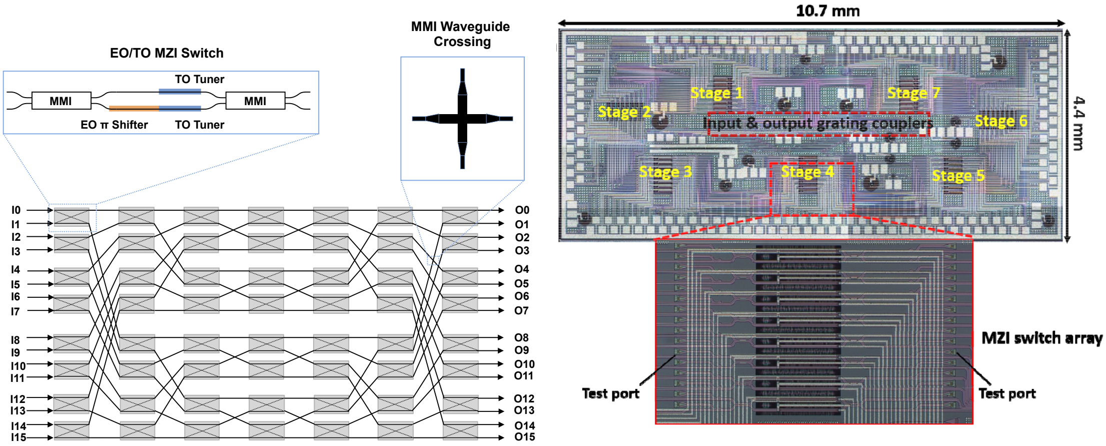
The dilated Beneš topology performs similarly to the Switch-and-Select in insertion loss and crosstalk due to switches, and does so while requiring fewer switches for an $N \times N$ fabric, leading to a simpler photonic switching fabric. However it is not strictly non-blocking and also requires an increased amount of crossings compared to other RNB topologies. The non-dilated Beneš topology requires the fewest switches among RNB topologies, and exhibits insertion loss as low as the DLN topology. However first-order crosstalk is incurred at every stage of the fabric, leading to increased crosstalk at the output ports. The Banyan topology, on the other hand, exhibits much less insertion loss than the DLN and employs fewer switches; but, like the Beneš, it suffers from first-order crosstalk in every stage. It is usually not preferred due to its blocking characteristics.

Finally, the crossbar, N-Stage Planar and PILOSS topologies all perform poorly in insertion loss and crosstalk, compared to the other topologies examined here. The crossbar and N-Stage Planar topologies are particularly problematic, as they exhibit a variable number of stages in the path, leading to a great variability in both the insertion loss and crosstalk at the outputs. However, they do not require waveguide crossings for a planar layout, which reduces the fabric complexity. Interestingly, the PILOSS topology is also a poor performer as regards insertion loss and crosstalk, with the caveat that it exhibits high uniformity in these metrics at the output ports.

2.8 Target Switch Fabric

The simulation studies performed in this thesis are based on a fabricated 16×16 switch fabric formed with thermally and electrically tuned MZIs [LZZ$^+$16]. The 56 MZIs used are interconnected using the Beneš topology, a rearrangeably-non-blocking topology. Fig. 2.17a shows an MZI composed of 2×2 MMI couplers, thermal and electrical tuners, a waveguide crossing formed from an orthogonal 2×2 MMI coupler and the full switch fabric, while Fig. 2.17b shows the produced chip.

The photonic power penalties are measured for the individual devices across a $30nm$ bandwidth centred around 1560 nm, which is the centre operation wavelength of the fabric. Insertion loss and crosstalk ratios are then reported for all tested wavelengths, with the maximum reported. The waveguide crossings have an optimised design with respect to photonic penalties, and are shown to have an insertion loss of $0.05dB$ with a crosstalk ratio lower than $-35dB$. The MZIs show an insertion loss of $0.5dB$ and a crosstalk ratio lower than $-30dB$ in the *cross* state, which is reached using the thermal tuners. MZIs in the bar state, reached using electrical tuning, show an increased insertion loss of $1.5dB$ and crosstalk lower than $-18dB$. The switch fabric



(a) Schematic adapted from [LZZ$^+$16].

(b) Chip microscope image copied from [LZZ$^+$16].

Figure 2.17: 16×16 EO/TO Beneš switch fabric.

demonstrates on-chip insertion loss (that is excluding couplers) of $6.7dB$ and crosstalk lower than $-20dB$ in the all-cross state, that is all MZIs tuned to cross. This insertion loss and crosstalk is the minimum across all switch states. In the all-bar state, insertion loss and crosstalk increase to $14dB$ and $-10dB$ respectively, the maximum across all sates. The significant insertion loss and crosstalk difference between the all-cross and all-bar states indicate a large dynamic range in path-dependent photonic penalty. The switch fabric also exhibits a maximum operation power of $1.17W$ for switching the MZIs.

The authors of [LZZ$^+$16] use a Beneš topology, as it is known to provide rearrangeably-non-blocking $N \times N$ connectivity using the smallest number of radix-2 switches [Ben64]. This characteristic is advantageous in reducing the insertion loss that carrier beams are exposed to when traversing the fabric. It also reduces the control and hardware complexity, as fewer radix-2 switches require less wiring complexity to achieve switching behaviour. However, it comes at the expense of exposure to first-order crosstalk, which can degrade signal quality. In photonic Beneš networks at full switching load, first-order crosstalk is applied to carrier beams in all the fabric stages, as well as all the waveguide crossings they encounter while traversing a path. The crosstalk then cascades along the vector of light propagation, creating higher orders of crosstalk and interfering with more victim signals.

2.8.1 Optimising the Power Penalty in Photonic Switching Fabrics

Photonic power penalties can be optimised in switch fabrics in the following ways. Firstly, the devices themselves can be optimised in this direction, as the authors of [LZZ^{+}16] have shown with optimising the waveguides and orthogonal MMI waveguide crossings. To further reduce the insertion loss and crosstalk of the MZIs in the bar state, which is greatly increased relative to the cross state, a "push-pull" drive can be employed for the electrical tuning [DLR^{+}15a]. As explained in Sec. 2.5.4, "push-pull" electro-optic MZIs include two p-i-n junctions, one on each MZI arm, each operating at a halved phase-shift capacity. They are then used in tandem to elicit switching behaviour while decreasing the effects of FCA and therefore insertion loss and crosstalk.

To reduce the effects of first-order crosstalk, topologies different to the Beneš network have been proposed. These are discussed in depth in Sec. 2.7 but, in brief, they trade off increased hardware complexity (i.e. number of radix-2 switches) and increased path-dependent insertion loss for decreases in crosstalk. However, if the objective of switch fabric design is to increase radix, a low hardware complexity in terms of photonic devices and electronic backend is paramount. Increasing switch fabric radix is essential for photonic switch fabrics to compete with commercial switches and interconnection networks based on electronics. This fact, combined with the relatively low photonic power penalties showcased by [LZZ^{+}16], motivated for this fabric's adoption as a baseline for the study comported in this work.

In chapter 4, a set of path-wise heuristic routing strategies that optimise for photonic losses are proposed and evaluated, while extrapolating the switch size to assess their potential at reducing insertion loss, required laser power (as evaluated through DSENT) and switching energy. Chapter 5 expands on the routing strategies by assessing routing strategy combinations for their further potential at reducing these metrics, as well as their impact on communication time.

2.9 Design Challenges

The previous sections of this chapter have described the inter-related challenges of designing photonic switching fabrics, which encompass design choices from the device level in terms of photonic performance (wavelength selective/broadband, device radix, photonic losses etc.), to the switch fabric control layer (topology, switching, multiplexing). In summary, the following properties are desirable in a photonic switching fabric:

A **high port count** is desirable irrespective of which level of the IN hierarchy the photonic switch is deployed in. Increasing the port count ultimately increases the ability to scale out the HPC or DC.

Low hardware complexity in terms of the number of component devices is also desirable as, on the one hand it enables higher port counts while decreasing photonic losses, while on the other hand it simplifies the backend electronics which are required to control switching fabrics.

Low control complexity is also beneficial for photonic switching fabrics. Due to their bufferless nature, these fabrics will have to reconfigure their state very frequently to serve incoming traffic, especially in the case of TDM. A highly complex routing scheme would increase the timing penalty of switch state computation, thereby decreasing the overall performance of the fabric. To surmount this, simple routing schemes for photonic switching fabrics are mandated.

Keeping a **low photonic loss level** is essential to providing scalability to the photonic switch fabric. Insertion loss, which is determined by the number of switch stages and the presence of waveguide crossings, can enforce unrealistic demands on lasers. Crosstalk, determined by switch device and waveguide crossing design, determines the quality of the received signal and the increase in signal power required to compensate. Device design, topology choice but also routing can affect the photonic loss level.

Fast switching capability, determined by device design, is also important to the overall performance of the fabric; μs-scale switching time prevents TDM for the same reason as control complexity.

Finally, **broadband** devices, i.e. non-wavelength selective, allow for more flexibility in spectrum allocation within the switch. Denser channel spacings or number of wavelengths can be employed, allowing the designer to reduce the data rate per wavelength without sacrificing aggregate bandwidth per path.

The Beneš switching fabric that has been selected as a target of study for this thesis fulfils a number of these criteria. The EO/TO-tuned MZIs are broadband and relatively wavelength-transparent, providing flexibility in spectrum allocation for DWDM. They are also fast-switching, leading to ns-scale reconfiguration time of the switching fabric. As the switching fabric employs the Beneš topology, the fewest MZIs are present both in total and per path for an RNB topology with good path diversity; this entails both a lower hardware complexity and a lower insertion loss per path, compared to other switch fabric topologies.

However this switching fabric has some disadvantages. While the employed single-ended EO-tuning provides fast switching and simple control per switching device, it also increases the insertion loss and crosstalk level in the "bar" state. Given this fact, whether this technology can be used for higher port count switching fabrics must be investigated. The Beneš topology also imposes first-order crosstalk to lightpaths at every stage; it must therefore also be investigated whether the cascading effect of photonic crosstalk prohibits scaling the switching fabric.

Lastly, as will be discussed in Chapter 3, the standard routing algorithm for Beneš networks is both of high complexity, and unable to account for photonic losses. Therefore, computationally simpler routing algorithms which can lead to lower photonic losses, both in terms of insertion loss and crosstalk, must be investigated to improve the photonic performance of the switch.

Motivated from the above design challenges, the next chapters of this thesis investigate these issues by proposing a methodology of evaluating the inter-dependent effects of routing algorithm, traffic configuration and photonic switch fabric performance, as well as a set of routing algorithms which reduce the photonic losses of lightpaths while remaining computationally simple. The effects of combining these with TDM as well as of arbitration algorithms are also studied.

Chapter 3

Evaluation Methodology

The previous Chapter introduced the main concepts behind PINs based on SiPh technology. The fundamental devices used to compose PINs were examined, along with the most prominent network topologies, and the trade-offs involved in SiPh PIN design were discussed. The switch fabric chip assumed as a baseline for simulation, that is the Beneš EO/TO MZI switch fabric shown in [LZZ$^+$16], was also described.

This chapter discusses the limitations of current methodologies for evaluating the performance of photonic switching fabrics. It then contributes to the state-of-the-art by proposing a methodology for evaluating the effects of network traffic configuration and intra-switch routing algorithm selection on the optical performance of photonic switching fabrics. It introduces the simulation framework designed to support the methodology, and describes the key insights that enable the abstraction of event-based network simulation to be applied at the level of a photonic switching fabric, which is the novel aspect of the methodology. It then continues by describing the implementation of the photonic loss model, which allows for estimations of photonic power penalty and therefore required signal power and laser power. It then includes a comparative analysis of the implemented model, where it is compared with two fabricated photonic switch fabrics as presented in the literature. The comparison demonstrates the level of accuracy of the model with regards to photonic losses. Finally, the chapter discusses the implemented routing algorithms, which are introduced as a contribution in Chapters 4 and 5, and compares them against the standard routing algorithm for the Beneš network.

3.1 The Simulation Landscape for Photonic Interconnection Networks

The advent of SiPh technology has led researchers from both academia and industry to create many different simulation methodologies for photonics technology, either for academic or commercial use. These methodologies span from photonic device design at Photonic Integrated Circuit (PIC) level, to that of network simulation at the PIN level [BC18]. As argued in Section 2.9, photonic device characteristics have a significant impact on the performance of photonic switch fabrics at the network level; however the *usage* of photonic switch fabrics (i.e. the traffic communication pattern and routing scheme) affects these metrics as well, as will be shown in this chapter as well as Chapters 4 – 7.

In fact, as argued by Michelogianakis *et al.* [MWT$^+$19], the photonics simulation landscape is currently suffering from a lack of standardisation in methodology and design flow, which embiggens the adoption barrier of the technology. This is especially true when considering photonic communication network simulation, where network designers commonly rely on in-house, purpose-built codes and frameworks.

Optical physics level simulation is used to capture the behaviour of light within materials and is commonly complemented by multi-physics device-level simulators; the objective here is to enable photonic device design. Commercially available examples are Lumerical-FDTD Solutions, Lumerical MODE, Lumerical Device or Luceda-IPKISS.

PIC level simulation targets designing photonic circuits based on photonic device models. A higher level of abstraction is adopted compared to device-level simulation and device behaviour is simplified into “compact models”, which operate using scatter matrices; these compact models are usually based on process design kits provided by photonic chip foundries. The objective here is to enable larger integration, thereby yielding larger designs. Circuit level simulators also incorporate chip floor planning design flow, allowing for their output to be used by foundries for chip fabrication. Examples of proprietary simulation software include Lumerical INTERCONNECT, Cadence-EPDA and Luceda-IPKISS which falls in both categories.

PIN-level simulation further raises the level of abstraction, with the goal of simulating the interactions of elements of complex PICs when composed into networks.

This methodology is commonly used when investigating photonic links for datacenters and HPCs, or for the ONoC domain. Many simulators that provide capability for research in these domains have evolved from academia. DSENT [SCK+12][1] is arguably the state-of-the-art open-source simulator for evaluating ONoCs, while PhoenixSim [CHB+10] [RBW+16], which is not available in the public domain, has been extensively used to evaluate photonic links and system-level PINs. DSENT has also been integrated in the Graphite simulator [MKK+10][2] for simulating multi- and many-core systems that include ONoCs. Various NoC or board-level investigations have used ocin_tsim [PGGH10][3]. Others have endeavoured to extend NoC simulators for the electronic domain such as Sniper [HCE12][4] or INSEE [NMAPR11][5].

However, to the best of the author's knowledge, the only PIN-level simulator that has native support for MZIs is PhoenixSim, which is not open source. DSENT does not include MZI models or network traffic models, meaning that the effect of network traffic configuration and routing algorithm on photonic metrics cannot be established there. Sniper extends Graphite (which uses DSENT for the photonics) and is targeted at the NoC level, while ocin_tsim or INSEE would also require re-working to include models for the photonic devices and for abstraction to the HPC or DC switch level. To this end, the decision was made to augment a flow-level interconnection network simulator called INRFlow [NPE+19], as will be described below. The augmented version of the simulator framework is named PhINRFLow.

In summary, PhINRFlow provides capabilities for performing *network traffic-driven* analyses of arbitrarily-sized photonic switching fabrics formed with multiple stages of 2×2 photonic switching devices. PhINRFlow can support multiple planar topologies for switching fabrics, so long as they are formed with 2×2 switches, as the topology model is decoupled from the photonic loss calculation, network traffic and laser power models. Photonic device parameters such as state-based insertion loss, crosstalk and switching power, as well as the photonic properties of passive devices (e.g. propagation loss, waveguide crossing crosstalk and insertion loss), can be inputted at runtime; this allows designers to extrapolate the photonic performance of switch fabrics formed with these devices, allowing for rapid exploration of the switch fabric domain. It also integrates the laser power model of DSENT, allowing for preliminary evaluations of

[1] https://www.rle.mit.edu/isg/technology.htm

[2] https://github.com/mit-carbon/Graphite

[3] http://www.ece.tamu.edu/ocintsim/

[4] https://snipersim.org//w/The_Sniper_Multi-Core_Simulator

[5] https://sourceforge.net/projects/insee/

laser power requirements and the co-examination of transceiver properties and switch properties, such as trading off data-rate per wavelength and number of wavelengths to reduce crosstalk. Through the integration of mature network traffic models, PhINR-FLow allows for investigating the interaction between the photonic device level and the network traffic and routing levels. To the best of the author's knowledge, this is the first simulator capable of modelling the interactions of the routing, network traffic and photonic device levels in PSFs, when considering the metrics of insertion loss, photonic crosstalk, switching energy and communication time.

3.2 Introduction to PhINRFlow

PhINRFlow is a light-weight, modular, flow-level interconnection network simulator which is designed for evaluating large-scale photonic interconnection networks at the DC or HPC level. It is extended from INRFLow, from where it derives most of its functionality, and includes abstract models of photonic links and routers/switches which can be leveraged for modelling photonic switching fabrics. It includes a dynamic engine which is able to capture temporal and causal relationships between communication flows and includes a large variety of communication models. Due to these factors and with proper modification, PhINRFlow is able to capture the behaviour of RNB photonic switch fabrics and constitutes a useful framework for investigating photonic switch fabric optimization.

3.2.1 Leveraging Bufferless Communication for Photonic Switch Simulation

Using a flow level PIN simulator to evaluate photonic switch fabrics relies on a simple but potent observation. In the electronic domain, a high performance network switch would include internal ingress and egress buffers between input and output ports in the form of virtual output queues. Packets or flits entering the switch are stored in these buffers before and after arbitration, with this process occurring between packet ingress to the switch and packet egress out of the switch. As this buffering takes time, it significantly affects the latency afforded by the switch. This effect is compounded when one considers interconnection networks comprising of many switches, as in these packets may have more than one route to reach their destination especially if considering adaptive routing. There, buffer capacity and latency cause an even greater variability in the performance of a switch, which cannot be captured by simulation at the flow level. Based on this reasoning, flow level simulation is unsuitable for the internals of *electronic* switches.

Photonic switches, on the other hand, are *bufferless* internally. Once network traffic has entered the photonic switch through an input port, it must stream uninterrupted through the photonic hardware and reach the destination output port. If a path through the photonic hardware is not available, packets or circuits must queue in the electronic backend, that is before streaming across the network, until a path becomes available. The timing and latency variability caused by internal buffering is therefore negated. This means that the flow-level abstraction, which does not account for buffering time, is suitable for *switch-level* modelling of data streams encoded in light traversing photonic hardware.

Other abstractions offered by PhINRFlow also play a key role in enabling photonic switch fabric simulation. Switch devices are modelled as abstract nodes with ports, and can therefore easily be expanded to model MZIs operating as 2×2 switches. Waveguides connecting cascaded MZIs can be modelled indirectly, through the connection of one node's ports to the next. Communication can be modelled unidirectionally through the topology and routing algorithm. Photonic losses of individual devices are inputted to the simulator at runtime using property files, while path-based losses are calculated using the dynamic engine functionality afforded by the simulator. Communication time is derived in an event-based fashion, whereby the timestamp of the next event relies on the highest aggregate data rate employed to transmit a flow.

Using these characteristics, arbitrary-sized photonic switch fabrics formed with 2×2 switches organised in the Beneš topology can be investigated using a wide variety of communication patterns or workloads. It is noted that by modifying the PhINRFlow simulator to include photonic losses, other topologies based on 2×2 switches can also be investigated simply by adding a new topology file and adhering to the simulator formalisms. However, as this thesis focuses on Beneš networks, implementing models for other network topologies within the simulator is considered out of scope.

3.2.2 Simulator Structure

As the simulator extends from INRFlow, the core components of the simulator are similar. Servers are modelled as abstract nodes that produce and consume traffic using event queues, where send, receive or computation events are stored. These events are generated using network traces, traffic patterns or pseudo-applications, as explained

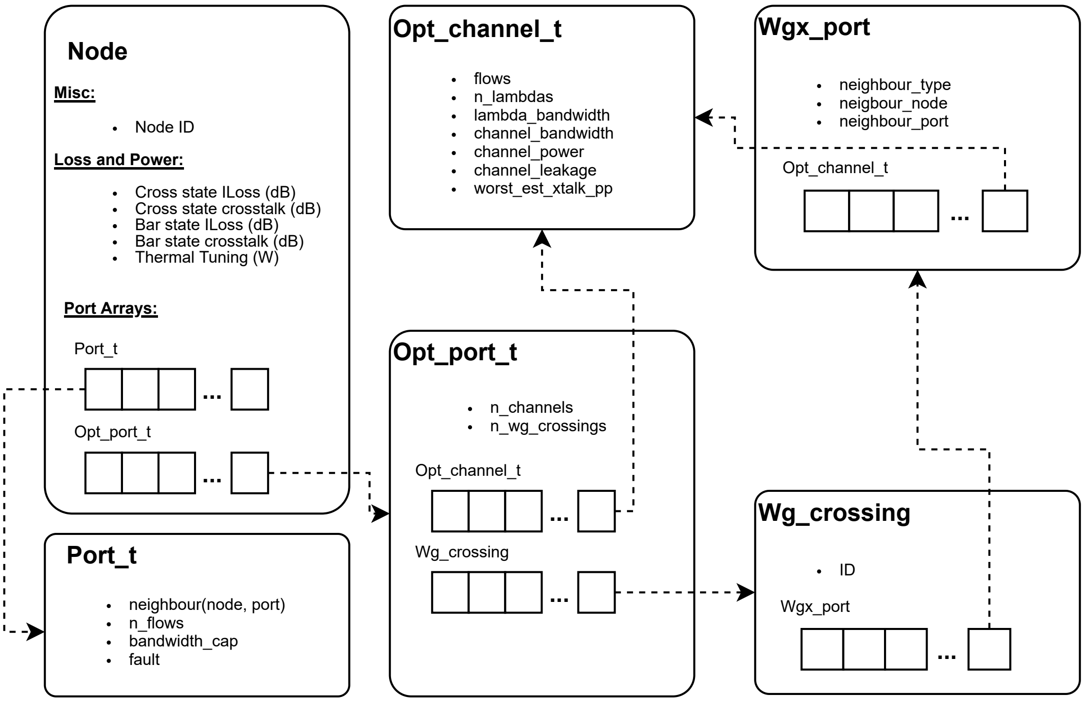


Figure 3.1: Structures used for modelling photonic nodes.

below. Event queues are encapsulated in the application model, which holds the parameters of the application, the information of the network flows and the collection of metrics gathered during simulation, which are reported after the simulation completes. Switches, on the other hand, are modelled as abstract nodes with port arrays.

Nodes and Modifications for Photonics Simulation

In PhINRFlow, a *node* is an abstract data structure containing an identifier and two sets of port arrays, one for connectivity information and one for information relating to photonics. This is depicted in Fig. 3.1. The separation of port arrays is done to isolate the photonics information and therefore, in principle, allow for the simulator to more efficiently model photonic links and electronic switches for full-system interconnection networks. Connectivity ports, or *port_t* data structure instances, hold the port's neighbour identifier (node/port tuple), the number of flows being concurrently served by the port, port bandwidth capacity (a function of the channels and the capacity per channel) and a fault parameter, indicating whether the port is operational or not. In the setups used for this thesis, one flow can use a port at each time, and the optical power corresponding to each flow is modelled on an individual channel. Additionally, network faults have not been used, as fault tolerance is not the focus of this thesis (although the structure to model them exists). The *node* also includes state-based insertion loss and crosstalk parameters, as well as the tuning wattage required to reach the modelled states. These values are parametrised and can be loaded into the simulator at runtime using property files.

Photonic ports, or *opt_port_t* data structure instances, hold an array of channel data structure instances as well as an array of waveguide crossing objects. Channels hold the channel bandwidth as well as the photonic power and leakage values for that channel, i.e. flow. As such, photonic power and leakage (discussed below) can be modelled for each individual port which contains a channel instance, facilitating power and leakage propagation to connecting ports.

Waveguide crossing data structures model the orthogonal MMI waveguide crossings assumed in this thesis. These are very similar to nodes, holding an identifier and an array of ports, called *wgx_ports*. Each port instance holds a pointer to a neighbour instance (either a node or another crossing) and an array of optical channels, similar to the *opt_port_t* structure. This allows for modelling photonic power and leakage propagation through the MMI crossings, thereby increasing the accuracy of the crosstalk model and allowing for higher orders of crosstalk.

Applications and Network Traffic Model

The data streams are modelled as flows inside the simulation engine, where a flow is a collection of data to be transmitted from a source to a destination. Flows are generated based on communication patterns and injected into event queues. The simulation engine assesses the network resources for each flow in the queue (i.e. the paths available from a flow source to its destination) and, if available, reserves them for the communication of the prospective flow.

As the simulator extends from INRFLow, it provides the same workload functionality. Network endpoints are modelled as simple traffic producers and consumers. However, a large variety of workloads are modelled, ranging from synthetic communication traffic to pseudo-realistic traffic generators created from analysing network traces. Some examples are included below.

Synthetic traffic patterns are commonly used to assess the performance of a network, while considering only the distribution of traffic to the nodes (i.e. no temporal characteristics). PhINRFlow supports a wide range of these:

- **Random-based**: Flows at a source node are assigned a destination based on a given probability distribution. When using the uniform distribution, all nodes have the same probability of being assigned as a destination. Non-uniform distributions are modelled also: in hot-spot and hot-region, nodes or node groups have a higher probability of being selected as a destination. Modelling the network performance while some regions of the network suffer congestion is possible with the latter two.

- **Bisection**: With one task per network node, the tasks are split into pairs. Tasks whose nodes are in the same pair communicate with each other.

- **All-to-one, all-to-all**: In the former, one target node is chosen uniformly at random and every other node sends a flow to the target node. In all-to-all, every node sends a flow to every other node in the network.

The simulator also includes pseudo-realistic communication workloads inspired from applications in the scientific computing and datacenter domains. These workloads include causality among messages, so most applications go through phases of high and low network pressure:

- **Scientific applications**: The workloads in this subset emulate scientific codes traditionally used by the HPC community. 2D and 3D stencil and sweep codes are incorporated, as well as an *nbodies* application.

- **Datacenter applications**: The workloads in this group emulate applications that are commonly used in the datacenter domain. This includes Mapreduce, which is a popular datacenter application which comprises of a scatter phase (one-to-all), an all-to-all phase and a gather (all-to-one) phase of communication. Unstructured applications are also included; *dcntraffic* adheres to the “elephant and mouse” traffic model reported in [KSG$^+$09], while *Torlocal* and *Torremote*, explained in Chap. 6, model traffic based on the analysis reported in [BAM10].

In addition, the simulator is capable of performing simulation based on network traces from real applications. Descriptions for the workloads employed to evaluate the contributions in this thesis can be found in Chapters 4 – 7. In particular, Fig. 6.4 in Chapter 6 depicts the distribution of messages for different workloads. For further verification of the expected behaviour of the workloads, the simulator includes a network traffic visualisation tool; an example output of the tool can be found in Chapter 7.

Workloads and their flows are modelled using the “application” structure. This holds all the relevant information about the workload (e.g. app. size, number of tasks, traffic pattern and related parameters, allocation and mapping strategies for the tasks). It also contains lists for event tracking (e.g. which events have occurred, which are pending etc.) and a structure used for reporting the application metrics.

Simulation Engine

Once the topology is defined, the application(s) initiated and the communication traffic generated, the simulation commences.

The simulation engine accesses the events from the application queues and injects them into the network. For flow send events, arbitration is performed on the access order of the injected flow sends. Arbitrated flows are accessed for their source and destination, and paths within the network are explored using routing functions. The effect of arbitration techniques on the switch performance is discussed in Chapter 7. Path availability is determined as such: if the sending and receiving nodes are available, and if the switches belonging to a prospective path can assume the state required by the routing function, and if the ports required by that path are not serving other flows, then a path is available. If a path is available within the network, the flow is assigned the path; if not, it is blocked.

Once a flow is assigned a path, the nodes, ports and channels belonging to that path are traced through the interconnect model and marked as busy. Based on the state of the network and the encountered photonic devices, a path is then assigned accumulated insertion loss, crosstalk and power penalty, and tuning wattage. Flows which are assigned paths are moved to the sending queues.

The next time-step is then derived, by dividing the size of send events (in orders of bytes) by the data rate they are sent on. The minimum of these and of the length of computation (if applicable) is assumed as the next time-step for the simulation.

3.2.3 Extensions conducted for PSF Modelling

As PhINRFLow was initially designed to model high-performance OINs (rather than PIN-based switching fabrics), adaptation was required to enable PSF modelling. A diagram depicting the functional components of PhINRFLow is shown in Fig. 3.2, which also contains a depiction of the simulation process. The functional components which were extended or added to the simulator to enable PSF modelling are framed in red.

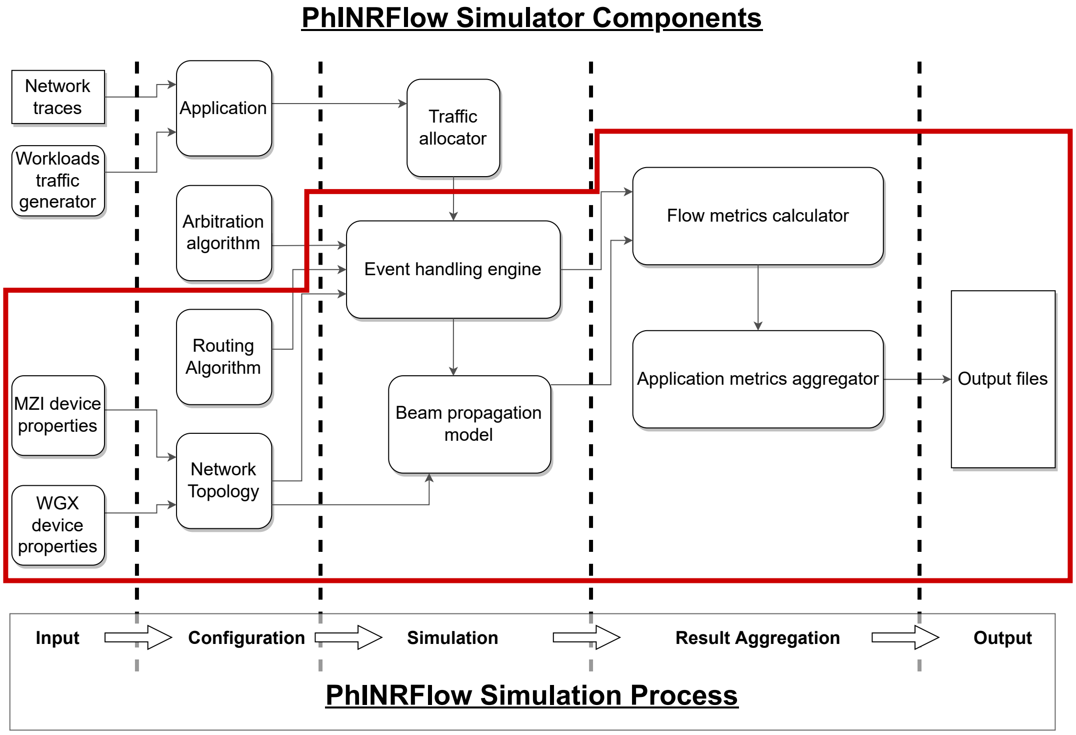


Figure 3.2: Structures used for modelling photonic nodes.

The input and configuration phases were extended to include photonic device models, such as MZIs and waveguide crosses. These are used by the "network topology" component, which constructs the network which will be fed input from the traffic allocator through the event handling engine during the simulation phase. The routing algorithm component (which is encapsulated within the topology files in the simulator) was extended to include the HIRs and "looping algorithm".

In the simulation phase, the "event handling engine" component was adapted such that it can call the topology-based routing algorithms. Contrary to the OIN paradigm, where routing is performed node-by-node, routing in PSFs is conducted in advance, with paths being pre-computed in their entirety and stored into routing tables during the configuration phase. This component was also extended to perform flow partitioning for TDM, as well as to enable the selection of port order through arbitration. To that end, this component also includes the arbitration policies discussed in Chapter 7.

The simulation phase also includes the “beam propagation model”, which was developed using the principles discussed above in this thesis. The way that these principles are used to model light propagation within the simulator is discussed in Section 3.4.

After the simulation phase is completed, the result aggregation phase commences. The purpose of this step is to aggregate the output metrics from the simulation phase, and organise the data for storing into output files during the output phase. The “flow metrics calculator” and “application metrics aggregator” components were extended to include the photonic and contention metrics.

Extending PhINRFlow

The simulator is open source[6], and is designed with extendability in mind. Contributors can extend it to cover additional PSF topologies by modelling them and their respective routing algorithms within the “network topology” component. Various 2×2 switching devices can be modelled by providing their photonic profiles during the input phase, through property files.

Contributors can also extend the simulator with new workloads either by augmenting the workloads “traffic generator component” to include them, or by running the simulator in “trace mode”, and inputting network traces instead. For a description of required network trace format, the reader is referred to the INRFlow paper [NPE^{+}19]. Novel arbitration algorithms can be included by extending the “arbitration algorithm” component.

Modelling switching devices with different port structures would require modifying the “beam propagation model” component as well as the repective “network topology” component and device models. The “beam propagation model” component can also be extended by contributors to model additional photonic effects (e.g. interchannel crosstalk).

3.3 Modelling the Beneš Switch Fabric

The Beneš EO/TO MZI switch fabric under investigation is modelled in the simulator as a new network topology. Here, the endpoints which produce and consume network traffic are modelled as nodes with two ports, one output port which connects to the

[6] Available at: `https://gitlab.com/ExaNeSt/phinrflow`

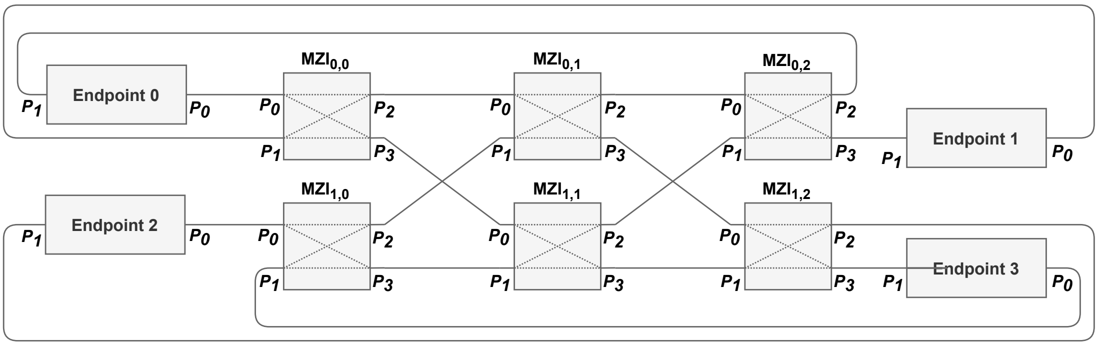


Figure 3.3: A 4-endpoint Beneš network with connected endpoints.

network input and one input port, connected to the network output. This is depicted in Fig. 3.3, which shows a small 4 × 4 model for illustrative purpose with enumerated ports, MZIs and endpoint numbers. The MZI switches that form the switch fabric are modelled as nodes with four ports, with ports P_0, P_1 acting as inputs and ports P_2, P_3 acting as outputs. They are denoted as $MZI_{i,j}$, where i, j are the corresponding row and stage of the MZI, respectively. Here, $i \leq 2logN - 1$ and $j \leq \frac{N}{2}$, where N is the number of endpoints in the Beneš network.

MZI switches are organised into stages and are either internal or edge switches. Internal switches have their input ports connected to MZI switch output ports from the previous stage. Their output ports connect to the input ports of an MZI switch in the next stage. Edge MZI switches connect in two ways; either their input ports are connected to endpoint outputs, or their output ports are connected to endpoint inputs. The remainder is connected to internal MZI switches.

The Beneš network is a recursively constructed network, with the 2 × 2 switch, or *radix* 2 Beneš, being the smallest network. An $N - endpoint$ Beneš is constructed using two $\frac{N}{2} - endpoint$ Beneš subnetworks and two additional stages of 2 × 2 switches. The left additional stages connect to the subnetworks using a connection shuffle, while the subnetworks connect to the right stage using a reverse shuffle. It is also observed that the Beneš network is symmetrical to the vertical axis.

In the simulated model, these properties of the topology are exploited to interconnect the network stages recursively. The network is split vertically at the centre stage and the two halves are processed separately before being connected to the centre stage. In the left half, switches in the current stage connect to the next stage in a column-wise order, with P_2 connecting to the next available input port in the upper subnetwork in the next stage, and P_3 to next available input port in the lower subnetwork. The connection algorithm recursively performs the same steps for the switches corresponding to the inner subnetworks and terminates when the centre stage is reached. The right half is connected using the same method but in reverse order, that is right-to-left. This method of interconnecting the switches is used to integrate the calculation of waveguide crossings per connection.

The calculation of waveguide crossings per connection relies on the assumptions that the switch fabric is implemented in a planar layout using orthogonal MMIs as crossings, and that no topological transformation is performed to reduce the total number of crossings. Such optimization is detailed for delta networks by Wang *et al.* [WWX$^+$16] and could be extended to Beneš networks; however this is out of scope here. The number of waveguide crossings in a connection beginning from port P in switch $MZI_{i,j}$ is calculated using the following for the left half:

$$wgx = \begin{cases} \bmod(i, var), & \text{if } P = P_2 \\ var - 1 - \bmod(i, var), & \text{otherwise} \end{cases} \tag{3.1}$$

and for the right half:

$$wgx = \begin{cases} \bmod(i, var), & \text{if } P = P_0 \\ var - 1 - \bmod(i, var), & \text{otherwise} \end{cases} \tag{3.2}$$

where $var = \frac{L}{2}$ and L is the number of endpoints of the subnetwork $MZI_{i,j}$ belongs to.

Once these are calculated, they are instantiated using a structure and added to P. The waveguide crossings are used to model the propagation and leakages of photonic and crosstalk signals.

3.4 Modelling Photonic Losses and Switching Energy in PhINRFlow

As explained, modelled switches, connections and crossings represent MZIs, waveguides and orthogonal MMIs respectively. Each of these applies insertion loss and leakage to a carrier beam traversing it, relative to the wavelengths of the carrier beam. Further, stateful elements (i.e. MZIs) apply a different level of insertion loss and crosstalk leakage to the carrier beam depending on the state. The maximum insertion loss and crosstalk values over the examined wavelength region are inputted into the simulator using property files at runtime. These are stored in corresponding fields within the node structure and accessed during simulation to calculate the leakage and output signal power during simulation. These are then used to calculate the insertion loss and crosstalk imposed on a flow as it is allocated a lightpath through the switching fabric by the routing algorithm.

3.4.1 Insertion Loss & Crosstalk

Once all the permitted flows in a particular time-step are allocated paths through the switch, the leakage calculation process commences. Input power in mW is assigned to flows at the input ports of the switch fabric. The devices in the switching fabric are swept in the direction of light propagation, stage by stage; based on the state of the switches in a stage, insertion loss is applied to the flow power as it is propagated to the switch output ports the flows egress from. Insertion loss is also applied to the leakages at the input ports as they are propagated to the output ports. Leakage power based on the state-dependent crosstalk ratio is applied to the rest ports of the switches, relative to each traversing flow. Leakage power is also applied based on the leakages present at the input ports.

Then, for each switch output port, the waveguide crossings are accessed. For each waveguide crossing, insertion loss is applied to the flow power which is then propagated to the crossing output port. Leakage power based on the waveguide crossing crosstalk ratio is applied to the relative rest ports of the crossing. The leakage power of the waveguide crossing ports is propagated along the vector of light propagation (after insertion loss is applied); leakage from P_0 of a waveguide crossing propagates to P_3 and leakage from P_1 to P_2 respectively. After all crossings in a path are processed, insertion loss for waveguide propagation per stage is applied to the power level of the leakages and flow powers of the final crossing. These are then propagated to the input ports of the next switch.

This process continues until the output stage of the switching fabric is processed. The power and leakages are then be read from the output ports of the switch. Insertion loss, crosstalk and power penalty are then added to each flow to be used after the simulation has completed for reporting the metrics.

3.4.2 Required Laser Power

Once the power penalty imposed on a flow has been calculated, it can be used to estimate the required signal power, and therefore laser power, for that flow. The DSENT [SCK$^+$12] laser model has been incorporated into PhINRFlow to facilitate this. The number of employed wavelengths, the data rate per wavelength and the combined power penalty of the flow are fed into the laser power model to calculate the required laser power. This is then added to the flow to be used after the simulation completes, for reporting the metrics.

3.4.3 Switching Energy

Switching energy is estimated in the following fashion. Firstly, the simulator is inputted property files including the thermal and electrical tuning wattage to reach the cross and bar states, respectively. For simulating networks with a different size than that of Lu *et al.*, the tuning wattages are fitted to a normal distribution using the Box-Muller method. Tuning wattages per state are then stored in the nodes representing the MZIs. For every time increment within a simulation, the tuning wattages corresponding to the network state are aggregated; the sum is multiplied by the time increment to get the total energy for the switch state. The energy from all the time increments is summed up and divided by the number of flows and the size of the flows in bits, to estimate the energy per bit from switching.

3.5 Comparison of Photonic Loss Simulation Model Against Demonstrated Switch Fabrics

In this section, the photonic signal propagation model is evaluated by comparing against two demonstrated switching fabrics from the literature. The goal is to ascertain the models' level of accuracy with respect to insertion loss and crosstalk modelling of switching fabrics formed with MZIs. To this effect, two photonic switch fabrics are selected: one based on thermally/electrically tuned MZIs [LZZ$^+$16], and one composed of thermally tuned MZIs [ZLZ$^+$16].

These two demonstrations are selected as they disclose the loss and crosstalk characteristics of their devices, as well as the switching fabrics. This can enable simulations where the switching fabric performance is extrapolated from the performance of

Table 3.1: I/O Permutations served by the “all-cross” and “all-bar” States.

All-cross State															
I_0	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}	I_{13}	I_{14}	I_{15}
O_8	O_9	O_{10}	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}	O_0	O_1	O_2	O_3	O_4	O_5	O_6	O_7
All-bar State															
I_0	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}	I_{13}	I_{14}	I_{15}
O_0	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}

the component devices. Zhao *et al.* include insertion loss and crosstalk ratio for the MZIs at the centre wavelength (1560 nm), while Lu *et al.* [LZZ$^+$16] report the worst-case insertion loss and crosstalk for the whole spectral range they investigate. They adopt the same MMI waveguide crossings, for which the insertion loss and crosstalk value is reported. They also report measured insertion loss and crosstalk values at the output ports for the entire switching fabric at two switching states, namely “all-cross” and “all-bar”, which indicate the best and worst-case states in terms of insertion loss and crosstalk, where there exists a difference between loss profiles of the two MZI states. The experiment is conducted by iteratively coupling a laser beam into each of the input ports and sweeping the wavelength region, then measuring the power at the output port as well as the leakage power on the rest ports. They report the results by grouping the output and leakage powers at each output port.

With light being coupled into all inputs, the input/output permutations which correspond to the two states are shown in table 3.1.

The experiments conducted by Lu *et al.* and Zhao *et al.* are replicated within PhINRFlow for each of the reported switch states. As the propagation model in PhINRFlow calculates leakages on a per-channel basis, simulating simultaneous communication of all the flows on the same channel would lead to all leakages being accumulated at the endpoints, therefore skewing the results and misrepresenting the experiment conditions. The flows are therefore routed on separate channels, one per flow, at 1560 nm. This way, the leakage of each ingressing beam to the output ports is isolated.

3.5.1 Simulating Light Propagation in a 16×16 EOMZI Switch Fabric

In their work, Lu *et al.* include a characterisation of the 16×16 Beneš switching fabric, where they evaluate the fabrics’ performance across a 30 nm spectral region centred

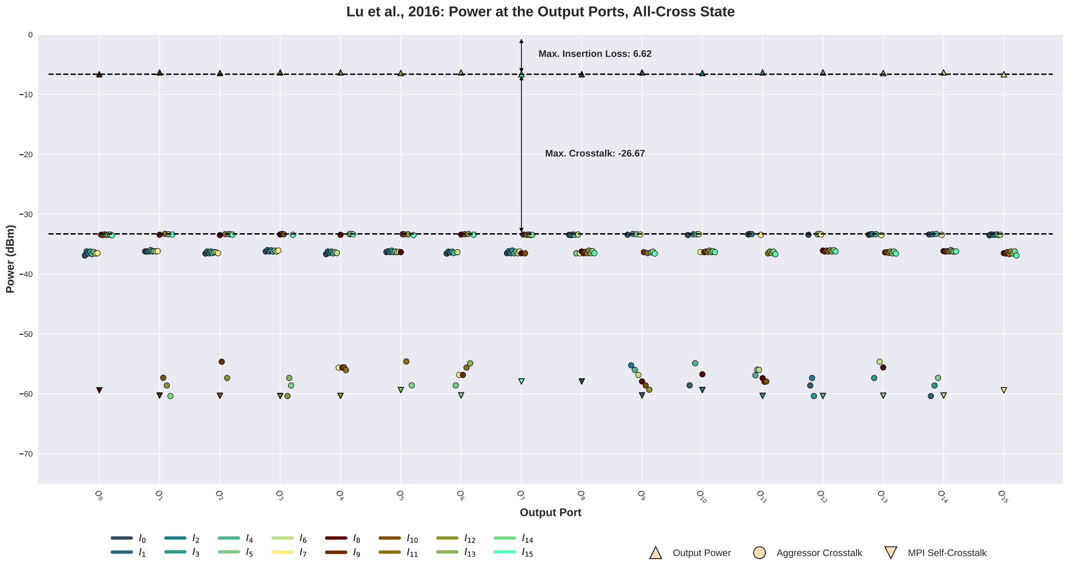


Figure 3.4: Simulated power and leakage values for the "all-cross" state at 1560 nm.

at 1560 nm, which they consider to be the central operational wavelength. They also report the performance of the individual 2×2 MZI switches across the wavelength region. In the "cross" and "bar" states respectively, the max. insertion loss is $0.4dB$ and $1.4dB$, while the crosstalk is $\leq -30dB$ and $\leq -10dB$. They then tune the switching fabric to the "all-cross" state. They find that the chip offers a worst-case insertion loss of approx. $6.7 \pm 1dB$ depending on the path corresponding to the permutation mapping. They also find that the worst-case crosstalk value at 1560 nm is $-30dB$. They then follow the same process for the "all-bar" state, finding that the insertion loss increases to approx. $14dB$ and the worst-case crosstalk across the wavelength region is $-10dB$.

In PhINRFlow, the fabric is first simulated in the "all-cross" state, using the permutation depicted in Table 3.1. The output power and leakage at each output from each ingressing beam is depicted in Fig. 3.4. Leakages to the victim ports are denoted as "aggressor crosstalk", while leakages of ingressing beams cascading to their designated output ports through multipath interference are denoted as "MPI Self-crosstalk". The worst-case insertion loss and crosstalk are then derived. In terms of insertion loss, the simulator reports a worst-case of $6.6dB \pm 0.5dB$. This value is in line with the reported measurements by Lu et. al, considering that waveguide propagation loss is simulated as identical in every path. The worst-case crosstalk obtained by simulation is approx. $-27dB$, which is increased within $3dB$ of the reported crosstalk. This is attributed to the fact that the simulation assumes an identical worst-case crosstalk ratio

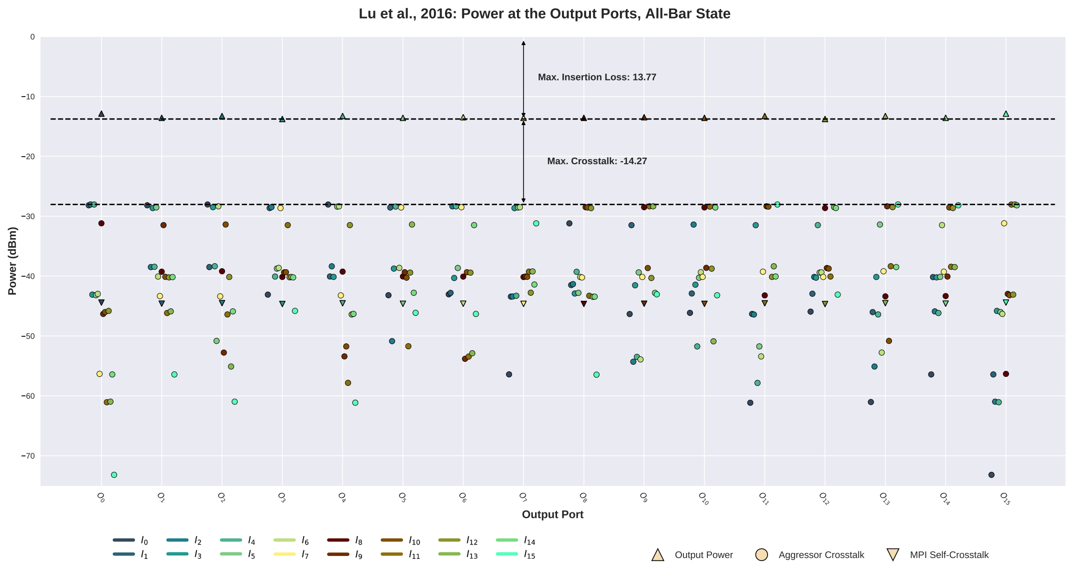


Figure 3.5: Simulated power and leakage values for the “all-bar” state at 1560 nm.

for all MZIs and waveguide crossings, which is not necessarily the case in the fabricated chip.

The fabric is then simulated in the “all-bar” state, using the permutation in Table 3.1 and the insertion loss and crosstalk values for MZIs in the “bar” state. Ingressing beams are routed again through individual channels to isolate the leakages, with the results depicted in Fig. 3.5. The simulated worst-case insertion loss of the bar state is approx. $13.8 \pm 0.5dB$, which is in line with the measured insertion loss when the propagation loss assumption is taken into account. The worst-case crosstalk obtained is $\sim -14dB$, which is $4dB$ less than the figure reported by Lu *et al.*; however, they report $\sim -10dB$ for the whole spectral region and not for the specific wavelength. They also report that one MZI has a lower than expected extinction ratio, which leads to higher levels of interference (and therefore crosstalk). As they do not report the crosstalk value for that specific MZI, the simulation does not include this anomaly.

3.5.2 Simulating Light Propagation in a 16×16 TOMZI Switch Fabric

The same process as above is used to simulate the “all-cross” and “all-bar” states for the switch fabric presented in [ZLZ+16]. They report $0.32dB$ insertion loss and $\leq -35dB$ of crosstalk at 1560 nm for an individual 2×2 thermally tuned MZI at both the cross and the bar state. They also derive the waveguide propagation loss for the switching

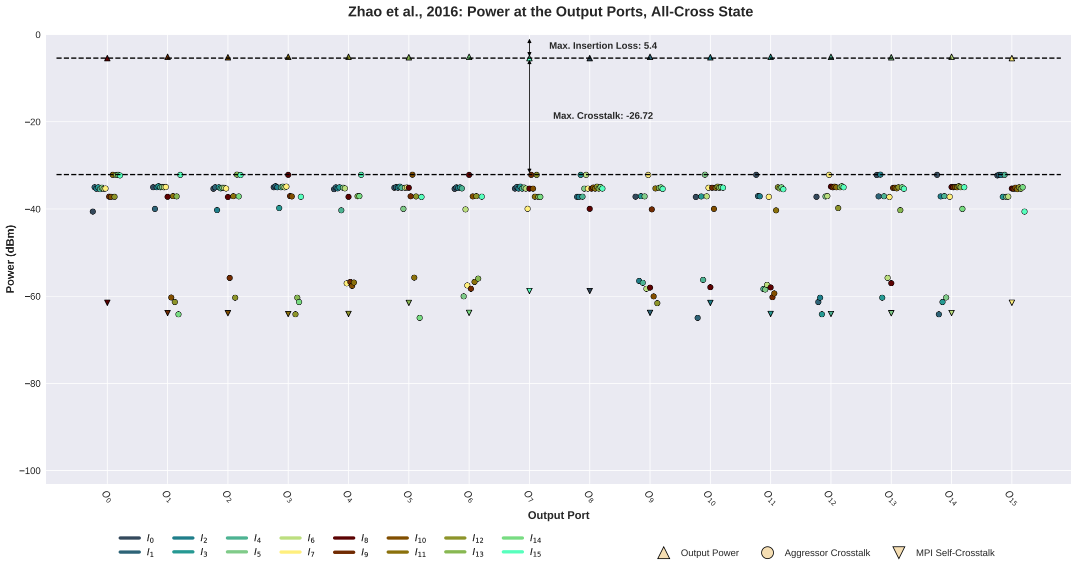


Figure 3.6: Simulated power and leakage values for the “all-cross” state at 1560 nm.

fabric to be 1.18 dB/cm but include the same waveguide crossing design as [LZZ+16]. They then tune the switch fabric, first to the “all-cross” and then to the “all-bar” state. The measured insertion loss is $5.2 \pm 1dB$ and the crosstalk is $\leq -30dB$ for both states, at a 10 nm bandwidth around 1560 nm.

The two states of the switching fabric are simulated using the reported waveguide propagation loss, as well as the insertion loss and crosstalk measurements from the MZI and waveguide crossing. The output power and leakages at all ports are depicted in Figs. 3.6 and 3.7 for the “all-cross” and “all-bar” states, respectively. The simulated insertion loss is $5.4 \pm 0.2dB$ for the “all-cross” state and $5.5 \pm 0.55dB$ for the “all-bar” state, which is in line with the measurements in [ZLZ+16]. The simulated crosstalk is $\leq -26.72dB$ for the “all-cross” state and $\leq -26.43\ dB$ for the “all-bar” state. As with [LZZ+16], the simulated crosstalk for [ZLZ+16] is higher than the measured values. This is, again, expected, considering that the crosstalk value employed for the MZIs and waveguide crossings is the worst-case crosstalk from a single MZI.

Based on the above results, the simulation is considered to be accurate with respect to insertion loss, and accurate within $3dB$ with respect to crosstalk. It is also noted here that with both switching fabrics, the results show good symmetry against the vertical axis in Figs. 3.4-3.7. The effects of aggressors on O_0 resemble those of O_{15}, those on O_1 resemble those of O_{14} and so on. This is encouraging, since it indicates that the Beneš network’s symmetry along the horizontal axis is closely matched by the results. Expanding the simulation to include more wavelengths is feasible by including

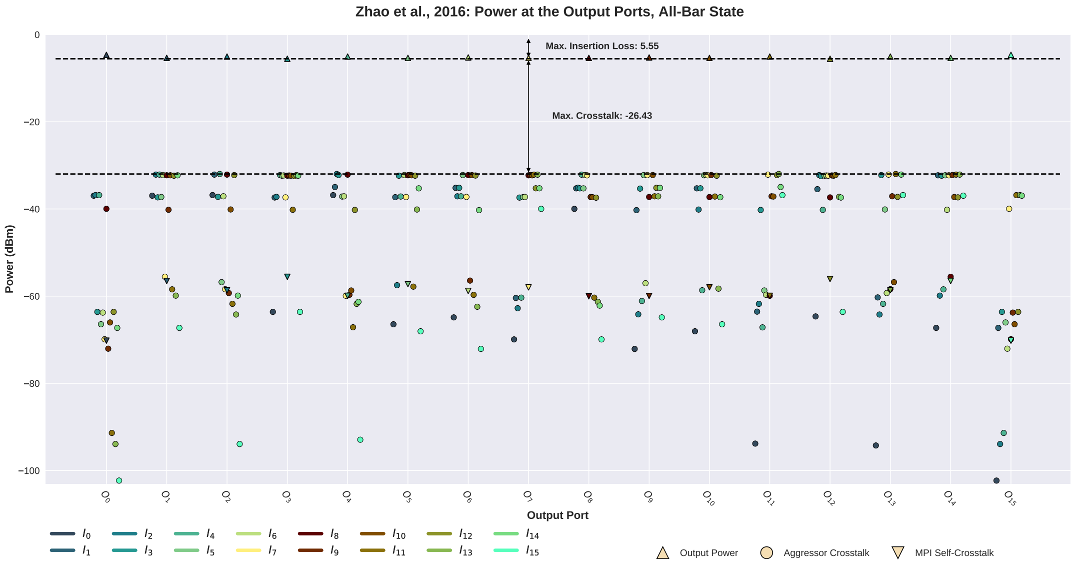


Figure 3.7: Simulated power and leakage values for the "all-bar" state at 1560 nm.

an insertion loss and crosstalk ratio profile for the devices for each wavelength to be simulated and each state of the active devices. There, each wavelength would be simulated discreetly within one channel, with inter-channel crosstalk being calculated at the output of each device using the equations derived in by Ramaswami *et al.* [RSS09], to account for cascading inter-channel crosstalk.

3.6 PhINRFlow Routing Models

In order to serve as switches within an interconnection network, photonic switching fabrics must be able to route traffic from multiple inputs to multiple outputs concurrently, either in a non-blocking or a blocking fashion. The routing algorithm internal to network switches is responsible for reserving network resources (virtual queue space, circuits etc.) for the routed information between source and destination ports of the switch, as well as resolving potential conflicts. This is distinct from the *network routing algorithm*, which routes packets or circuits across a network. In the case of photonic switching fabrics, the network resource to be reserved is the lightpath, an uninterrupted pathway which traverses a set of waveguides and stateful elements within the switching fabric. The choice of internal routing algorithm has a substantial impact on both the performance of the network switch as well as the complexity of the network controller, which in turn affects the scalability of the switch.

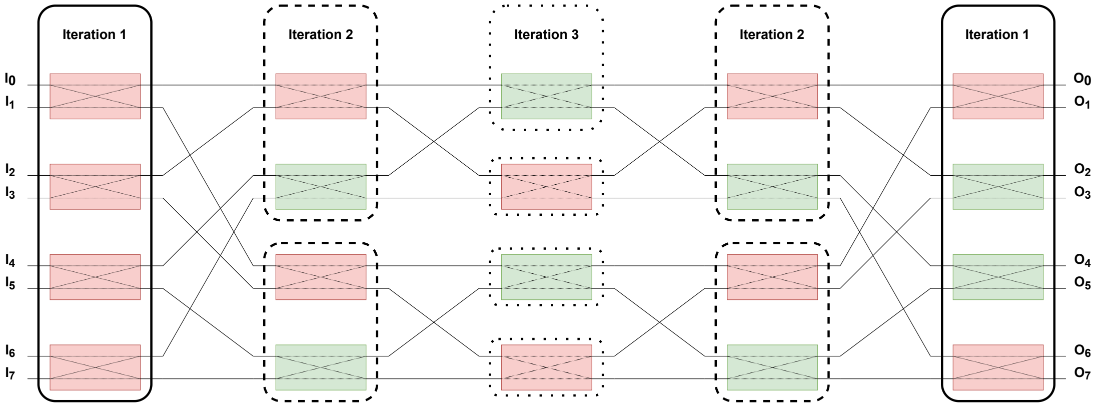


Figure 3.8: Recursive switch state setup in the Looping Algorithm. The red switches are set to the “bar” state, while the green switches are set to the “cross” state.

Table 3.2: A permutation example for Looping Algorithm setup.

Permutation							
I_0	I_1	I_2	I_3	I_4	I_5	I_6	I_7
O_6	O_7	O_3	O_2	4_5	O_4	O_0	O_1

3.6.1 The Looping Algorithm

The standard routing algorithm for controlling Beneš switching fabrics is called the “Looping Algorithm”[OTW71]. In this algorithm, the source and destination pairs, or connections, in a permutation to be routed form a bipartite graph [DT04]. Paths for connections are computed by setting the switch states of the outer stages of the Beneš network so that they can connect to the inner subnetworks, and then following the same process for the inner subnetworks using recursion. The recursion terminates when the inner subnetworks are the 2×2 switches in the centre stage. The algorithm begins by randomly selecting the first connection to route. It then assigns a path to that connection exclusively using either the upper or the lower subnetworks at each recursive step.

Once the first connection has been routed (i.e. assigned a path), the neighbour node of the connection’s destination is selected; the connection with this node as its destination is selected for routing. In this context, neighbouring nodes share the MZI they connect to at the outer stages of the Beneš network. The selected connection is then assigned a path using the subnetwork which is complement to that of the previous connection; if the previous was assigned the upper subnetworks, the current connection will be assigned the lower subnetworks and vice versa. Once it is routed, the neighbour node of the connection's source is selected; the connection with this node as its source is selected for routing, using the complement subnetwork. This process continues until the permutation is routed.

Depending on the permutation, a situation may arise during the looping algorithm where the neighbour source or destination is already routing a connection. In this case, the algorithm selects a random un-routed connection via its source and attempts to route it through the upper subnetwork. It then continues in the neighbour-wise fashion until either all connections have been routed or until another loop is detected.

The algorithm's operation is depicted in Fig. 3.8, showing an 8 × 8 Beneš serving a permutation. The switches in the left-most stage and right-most stage are assigned states based on the order of configuration the looping algorithm has enforced on the connection permutation (iteration 1). The upper subnetwork is then selected, with the left-most and right-most stages of the subnetwork being configured based on the looping order for the connections accessing the subnetwork (iteration 2). The upper switch in the centre stage is then selected and configured (iteration 3). The algorithm then configures the lower subnets. By computing paths recursively and using the looping order, all connections can be routed through the Beneš network without conflicts.

3.6.2 Challenges of the Looping Algorithm in Photonic Beneš Switching Fabrics

As it was developed for buffered circuit-switched electronic networks, the looping algorithm follows a set of assumptions that render it inefficient for photonic Beneš networks. Firstly, the original looping algorithm requires a full permutation to operate correctly. However, in communication contexts such as intra-datacenter interconnects, HPC interconnects or photonic NoCs, the communication data that a switching fabric must handle at a given time does not necessarily conform to a permutation. Although modifications to the algorithm have been proposed to circumvent this (e.g. [KAM$^+$14] [CC14]), these require additional computation and therefore increase the complexity of the algorithm.

Secondly, the looping algorithm operates *given* the permutation that it is meant to route through the Beneš network at operation time; that is, there is a dependency between the algorithm's operation and the permutation. This aspect requires packets (or information carried in circuits) to be buffered at the input until the looping algorithm operates. In bufferless photonic networks, this is intractable.

The looping algorithm also requires the ability to rearrange connections that have already been routed, when these block a new connection. Although this can be done computationally before any connection has been routed through the Beneš network, in photonic networks that are already routing photonic carrier beams this becomes intractable. The photonic source must cease transmitting, and the remaining light inside the network must propagate to the respective destinations before the network can be reconfigured. This can cause unacceptable time overheads and even packet corruption or loss. Therefore to maintain the integrity of pre-existing connections which block a new one from accessing the network, the new connection must be blocked until the required resources are available. This aspect defeats the purpose of a rearrangeably non-blocking network. This is discussed at length in Chapter 6.

Finally, as it was developed for electronics, the looping algorithm is unable to accommodate the concept of path priority. Routing in the looping algorithm is done using only turn-based logic and rearrangement, with all paths being considered equal. However, in photonic Beneš networks formed with MZIs, each path exposes the carrier beam to a different level of insertion loss and crosstalk leakage, depending on the state of the MZI switches, the number of waveguide crossings and the propagation loss due to waveguide length. Energy dissipation can also vary with respect to MZI state; paths can exhibit a variable footprint in energy dissipation. Therefore, available paths for flows can be assigned priorities based on the insertion loss they will exhibit, on the crosstalk, on the energy dissipation or combinations of these features. Although this aspect of Beneš network routing can be exploited to reduce the optical penalties and energy dissipation in photonic Beneš switch fabrics, this cannot be done by the looping algorithm.

3.6.3 Hardware-Inspired Routing Strategies

To address the problems outlined above, the concept of “Hardware-inspired Routing Strategies” (HIRs), which is one of the contributions of this thesis, is introduced. HIRs operate by providing greedy guaranteed-delivery routing for an incoming connection, using pre-computed paths which are ranked based on the characteristics of the switch hardware and compared with the network state at the time the routing algorithm is called. The routing capability is on the one hand greedy, in that the strategies do not ensure a non-blocking use of the fabric; on the other hand, as blocked traffic retransmits once resources have been freed, delivery is guaranteed. The routing algorithm operates in three decoupled phases: pre-compute, path ranking, and routing.

Pre-Compute Phase:

Here, the paths available from a source to a destination through the Beneš network are computed, along with the required MZI states and the number of crossings. The route computation relies on the fact that the Beneš network is equivalent to two mirrored butterfly networks with their centre stage being fused [DT04].

Computing the route signatures: Assume an $N \times N$ Beneš network, with sources connecting to the left-most stage and destinations to the right-most. Connections from sources to destinations are formed left-to-right and for each switch in the fabric, the input ports are located on the left and output ports on the right of the switch. The upper output port of each switch is marked using a 0, and the lower port with a 1 sign.

Starting from the left-most stage of the network, a path to a destination can be described by the sequence of output ports that must be traversed at each stage, until the switching fabric is traversed and the destination is reached. A routing signature can then be formed in binary with length $2logN-1$, using the signs of the traversed ports from each stage in order to reach the destination. Thus, each bit b_i in the routing signature corresponds to output from the upper or lower port of a switch in stage i in the fabric.

As the Beneš network is equivalent to two mirrored butterfly networks with the centre stage being fused, the $\frac{(2logN-1)}{2}$ right-most bits in the routing signatures for every path from a source to a destination, are identical and correspond to the destination tag of the destination node. This follows from the self-routing principle of the butterfly topology; there, destinations are assigned a binary destination tag, and each bit b_i of the tag denotes egress from the upper or lower port at stage i. This property has been identified by Raghavendra *et al.* [RB91] and is exploited here to form the right half of the proposed routing signature. The remaining bits form a bit permutation, and correspond to the path number. Using the bit permutation and the destination tag, all available routes from a source to a destination can be expressed in the Beneš network.

Computing the switch states: With paths and their routing signatures computed, the input ports of every switch in the paths are now marked using a 0 for the upper or 1 for the lower input port. For each path, a secondary routing signature is formed using the input port signs at each stage. By using the XOR operation on the two routing signatures, the state of the traversed MZI switches is computed.

Computing the waveguide crossings: For every computed path, the total number of waveguide crossings within the path is computed. This is done stage-by-stage, using equations 3.1 and 3.2.

Path Ranking Phase:

The second phase ranks the pre-computed paths for a source-destination pair based on the required switch states, the number of crossings, the number of state changes required from the current network state to route a path, or combinations of these. These criteria are selected for the following reasons.

Number of waveguide crossings: It has been observed that the number of crossings varies between paths for a source-destination pair. By ranking the paths depending on the number of crossings they include, paths with fewer crossings can be preferred by the routing algorithm. As crossings increase insertion loss and crosstalk, preferring paths with fewer waveguide crossings can reduce the level of these two metrics.

Number of switches in the "cross" or "bar" state: When the switches are formed using MZIs, the employed MZI tuning mechanism can cause an imbalance between the states in insertion loss, crosstalk and required switching energy. This is especially the case for single-ended EO tuning, where free-carrier absorption deteriorates MZI performance in one of the two states. Therefore, by ranking paths by the number of MZIs in one of the states and sorting them in ascending order, the three metrics can be reduced.

Number of switch state changes: Depending on the MZI tuning mechanism, there may be an imbalance in the required tuning energy between the switch states. Ranking paths in ascending order depending on the number of switch state changes from the current network state can therefore potentially reduce the switching energy for the network through MZI reuse.

The above ranking criteria can be combined such that ties in one routing rank are resolved using another rank. For instance, if paths are ranked first by number of crossings and then by number of "bar" state MZIs, a new rank may be formed where paths are sorted by state with the ties solved by the number of waveguide crossings. The routing criteria, their combinations as well as their behaviour at different switch scales are examined in Chapter 5.

Routing Phase:

In this final phase, the pre-computed paths for a source-destination pair are used to route the connection in the network. Based on the employed routing strategy, the paths are assessed one at a time using the corresponding rank as the assessment order until an available path is found. For each assessed path, the required MZI states for the path are compared against the current network state. If any of the requested MZIs cannot be switched to the requested state because of a pre-existing connection, the path assessment is stopped and the next path is assessed. If all MZIs can assume the required state, the path is selected, marked as occupied in the controller and the routing algorithm terminates. If no path is available, the source is notified and the connection is blocked.

3.7 Impact of Routing Algorithm Choice on Photonic Metrics

As has been argued in this chapter, the choice of routing algorithm for controlling EO/TO-tuned photonic Beneš switching fabrics can have a significant impact on the performance of the fabric with respect to insertion loss, crosstalk, optical power penalty and, ultimately, required signal power. In this section, this impact is investigated for five of the proposed HIRs as well as the looping algorithm, across five sizes of switching fabric. The goal is to investigate the potential benefits of HIRs over the looping algorithm, as well as identify the limits of the switch fabric in terms of scalability.

3.7.1 Experiment Design

The bisection workload is selected to evaluate the impact of the routing algorithms. Bisection is a permutation workload, meaning that the served outputs are a permutation of the inputs; therefore, the looping algorithm can function without modification when used to configure the fabric. It also causes the switching fabric to be fully loaded; as all inputs are active simultaneously, crosstalk has the highest impact on the optical power penalty, allowing for investigation of the worst-case in terms of traffic saturation.

In terms of load offered to the switch, the looping algorithm excels at servicing the bisection workload. This is because the looping algorithm is designed to service complete permutations with no switch fabric contention, while the bisection workload comprises of flows whose inputs and outputs are a permutation. Therefore, all flows transmit, leading to the switch being at full load. The HIRs on the other hand, are designed to optimise for insertion loss, power penalty and switching energy consumption; to do so, they induce switch fabric contention, as detailed in Chapter 6. There, it is shown that with the HIRs $\sim 20\%$ of the flows suffer contention; therefore, the switch load with the HIRs decreases accordingly.

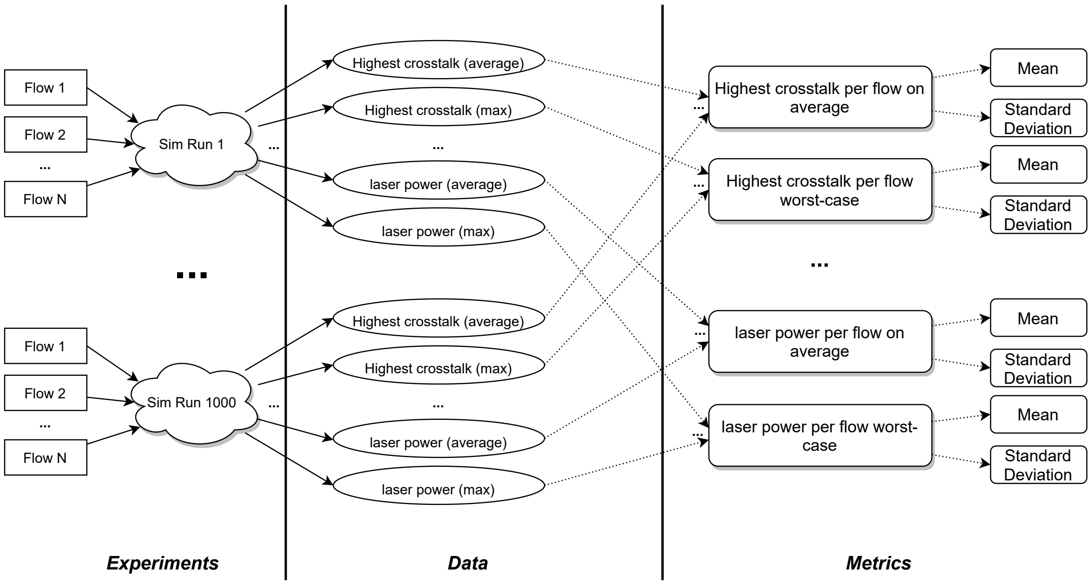


Figure 3.9: The process of composing simulation data into metrics.

The following routing algorithms are selected for evaluation. As explained above, the Looping Algorithm is the standard routing algorithm for Beneš networks, while the latter algorithms are a subset of the HIRs proposed in this thesis:

- **Looping Algorithm**: The standard unmodified looping algorithm discussed in Sec. 3.6.1.

- **m_b**: Selects the first available path based on the least number of switches in the "bar" state.

- **m_bx**: Selects the first available path based on the least number of waveguide crossings in the path. Ties are solved by sorting against the number of switches in the "bar" state.

- **m_x**: Selects the first available path based on the least number of waveguide crossings in the path.

- **m_xb**: Selects the first available path based on the least number of switches in the "bar" state. Ties are solved by sorting against the number waveguide crossings in the path.

- **rnd**: The paths from a source to a destination are shuffled. The first available path of the shuffled set of paths is selected.

Using each of the routing algorithms, 1000 simulations with a different random seed are executed for each network size, ranging between a 4×4 and a 64×64 network. For each run, the following data are obtained:

- **Highest crosstalk level per flow**: Each flow that traverses the fabric during a simulation is exposed to leakages from aggressor flows. The leakage with the highest power level is obtained for each flow. Examining the highest crosstalk level per flow reveals the highest level of interference from an aggressor flow, in isolation from leakages caused by other aggressors.

- **Aggregate crosstalk leakages per flow**: For each flow during a simulation, the sum of the crosstalk leakages from aggressor flows is obtained. As in these experiments all flows are considered to be modulated on the same wavelength set, every crosstalk leakage from an aggressor will contribute to intra-channel crosstalk. The aggregate crosstalk determines the crosstalk power penalty.

- **Insertion loss per flow**: The insertion loss that is imposed on each flow traversing a path through the switch fabric is obtained.

- **Crosstalk power penalty per flow**: The crosstalk power penalty, based on the aggregate crosstalk leakages, is reported per flow.

- **Laser Power per flow**: The laser power required to service each flow, in the presence of insertion loss and crosstalk, is reported.

The above data are then composed into *metrics*, as depicted in Fig. 3.9. Each of the 1000 simulations comprises of a number of flows. For each simulation, the average of each datum over the flows is collected. The examined metric is composed of the mean and standard deviation of the data over 1000 simulations. Therefore, the standard deviation corresponds to the metric, not the data that compose it; it therefore reflects the behaviour of the data over the set of simulations. This is done to examine the effects of the routing algorithms on the metrics in the average case, related to the exposed traffic configuration; these metrics are suffixed with “on average”.

Further, the maximum of each datum over the flows is collected for each simulation. The examined metric is again composed of the mean and standard deviation of the data over the runs. Again, the standard deviation corresponds to the metric and not its components. These metrics, suffixed with “in the worst case”, are obtained to analyse the worst-case effects of the routing algorithms related to the traffic configuration.

Table 3.3: Photonic parameters for simulation.

Parameter	Value	Corr. Ph. Pathway Segment
MZI Insertion Loss “bar”	1.4 dB	Switch path
MZI Insertion Loss “cross”	0.4 dB	Switch path
MZI Crosstalk “bar”	-18 dB	Switch path
MZI Crosstalk “Cross”	-30 dB	Switch path
Wg. Crossing Insertion Loss	0.05 dB	Switch path
Wg. Crossing Crosstalk	-30 dB	Switch path
Insertion Loss (Propagation)	1.18 dB/cm	Switch path
Coupling Loss	2 dB	Couplers
Laser Insertion Loss	1 dB	Laser
Laser Efficiency	25%	Laser
Ring Filter Loss “thru”	0.01 dB	Modulator/Filter
Ring Filter Loss “drop”	1 dB	Modulator/Filter
Mod. Insertion Loss	2.5 dB	Modulator
Extinction Ratio	15 dB	Modulator
Detector Loss	1 dB	Detector
Detector Responsivity	1.1 A/W	Detector
Link Data Rate	$16Gb/s \times 32\lambda$	Link Level
Target Bit Error Rate	10^{-15}	Link Level

The optical parameters assumed for the simulations are shown in Table 3.3, organised by the corresponding segment of the photonic pathway which they refer to. The parameters related to the intra-switch lightpath are adopted from [LZZ+16]. It is assumed that the crosstalk value per MZI state is identical for each wavelength. Obviously in practice this varies with wavelength but, since Lu *et al.* only report the worst-case value, this is adopted here. Parameters related to the laser, modulator/filter, detector and couplers are aligned to the default values used in DSENT [SCK+12], from which the laser model is adopted. The modulator characteristics are obtained from [WGK+16]. Finally, the driver electronics for the modulator and receiver use the 45 nm CMOS model detailed in DSENT.

3.7.2 Impact on Highest Crosstalk

The impact of routing algorithm selection on the highest crosstalk level is first examined. In the context of this thesis, the power in dB of an individual crosstalk signal at the network output arising from carrier beam k is defined by the following ratio:

$$XT_{i,j,k} = 10log_{10}(P_{leak}^{k,j}/P_{out}^{i,j}) \tag{3.3}$$

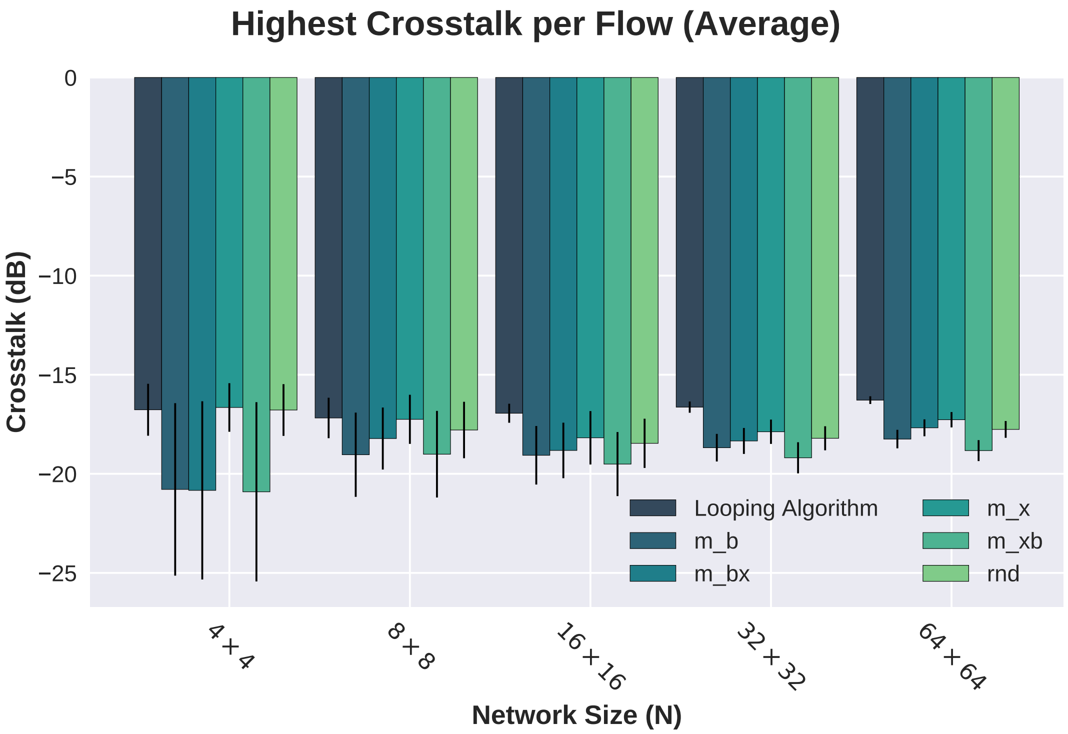


Figure 3.10: Highest crosstalk per flow on average. Lower is Better.

where i, j are the input and output ports of the victim signal. For each flow within a simulation, the crosstalk signal with the highest power level that this flow is affected by is obtained; the highest power levels are then averaged from all the flows to report the highest crosstalk power level on average, shown in Fig. 3.10. The maximum of the highest power levels is also obtained and reported in Fig. 3.11, to investigate how the dynamic range of crosstalk evolves for each different routing algorithm as the network scales in size.

For the looping algorithm, the highest crosstalk power levels on average range between $-16.8 \pm 1.3dB$ and $-16.3dB \pm 0.2dB$ as the network size increases. This algorithm shows the poorest performance which is expected, since the looping algorithm guarantees full saturation of the switch fabric at any given time when using a bisection workload; all input ports are active concurrently and the algorithm offers no control over path selection, leading to paths being assigned with more MZIs in the “bar” state and more waveguide crossings compared to other routing algorithms. This also means that path selection with the looping algorithm causes the insertion loss to increase proportionately to the highest level of leakage power as the network scales, leading to a similar highest crosstalk level on average. It is noted that as the network size increases,

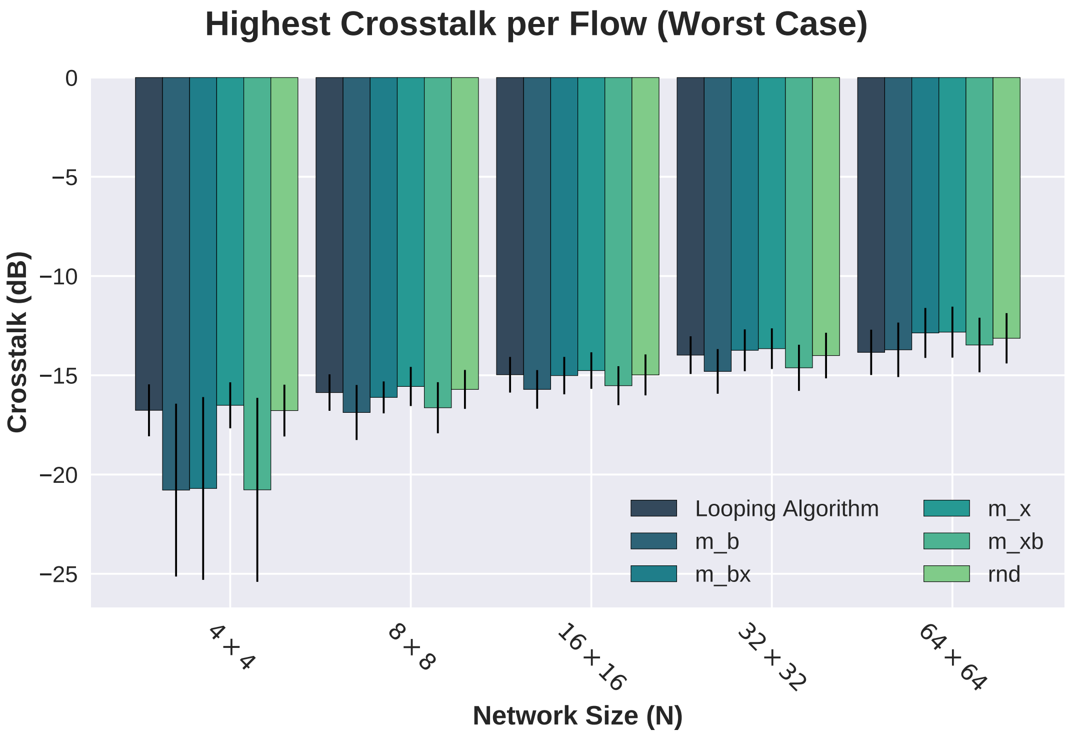


Figure 3.11: Highest crosstalk per flow in the worst case. Lower is Better.

the standard deviation across the simulation batch decreases; scaling the network size decreases the variability of the highest crosstalk power level on average. Conversely, the maximum highest crosstalk power level per flow increases with network size, with power levels ranging from $-16.8 \pm 1.3dB$ to $-13.8dB \pm 1.1dB$. Again, this is expected as the maximum highest crosstalk power level reflects the size of the network in terms of number of devices per path.

In terms of the HIRs, the first notable observation is that for networks with 4 inputs, high variability and very low crosstalk levels can be observed for routing strategies that prefer paths with the fewest "bar"-state MZIs, i.e. *m_b*, *m_xb* and *m_bx*. This is explained by the fact that in these networks there are only 3 MZIs per path and only two paths between each source and destination. Therefore, on the one hand the highest crosstalk per flow on average and in the worst case are similar; the worst-case highest crosstalk power level has a greater impact because fewer flows are present in the network compared to larger sizes, thereby skewing the average. On the other hand, changing the permutation of source/destination pairs, which occurs by running the simulation with different random seeds, forces the routing algorithm to select the sub-optimal path for a flow more frequently in a higher number of cases, leading to the increased variability and therefore larger standard deviation. Conversely, the behaviour when using *m_x* is very similar to both *rnd* and the looping algorithm. This is because for a network with 4 inputs, very few waveguide crossings (0-2) may exist in any path from a source to a destination. This leads to the *m_x* strategy being ineffective at reducing the highest level of exhibited crosstalk when compared to the looping algorithm.

For larger network sizes, the highest level of crosstalk power on average and in the worst case increases irrespective of the routing strategy used. The best strategies at reducing the highest crosstalk level are *m_b* and *m_xb* and, to a smaller degree, *m_bx*. The levels for *m_b* range between $-20.8 \pm 4.4dB \rightarrow -18.5 \pm 0.5dB$ on average and $-20.8 \pm 4.4dB \rightarrow -13.8 \pm 1.4dB$ in the worst case. The levels for *m_xb* are similar to *m_b* for all network sizes; however, on average *m_xb* outperforms *m_b* by approx. $0.5dB$, while in the worst case *m_b* shows approx. $0.2dB$ lower levels. Nevertheless, both routing strategies significantly outperform the looping algorithm on average and in the worst case. In the average case *m_b* presents savings ranging from $4 \rightarrow 2dB$ compared to the looping algorithm as the network size increases. In the worst case, the savings shown by *m_xb* over the looping algorithm range from $3.5 \rightarrow 0.2dB$ with *m_b* increasing the savings further.

The *m_bx* routing strategy also outperforms the looping algorithm in the average case, although to a lesser degree than either *m_bx* or *m_b*. The savings in the average case range from approx. $3.7dB \rightarrow 1.4dB$. However, in the worst case the performance of *m_bx* worsens as the network size increases and is ultimately outperformed by the looping algorithm for the largest network size by approx. $1dB$. As this routing strategy prefers paths that include less waveguide crossings and solves ties in paths through the least amount of “bar”-state MZIs, the performance indicates that selecting paths based on the number of waveguide crossings is not a viable criterion for reducing the highest level of crosstalk power across the ports, especially in the worst case.

This indication is reinforced by the performance of the *m_x* strategy, which is worse than any of the previous. The performance exhibited by this strategy improves for the average case as the network scales in size, which is different to the previously analysed strategies; this is because the average number of waveguide crossings per path increases proportionately to the network size, enabling *m_x* to have an increasing impact. In the average case, *m_x* outperforms the looping algorithm by up to $1.5dB$ for 64 input ports. However, in the worst case, *m_x* performs poorly, showing increases of 0.5 to $1dB$ compared to the looping algorithm. In fact, *m_x* performs little better than random as shown by the *rnd* strategy.

Finally, it is interesting to note that the *rnd* strategy behaves similarly to *m_x* in both the average and the worst case. The looping algorithm is slightly outperformed by *rnd* as the network size increases; however, in both cases the differences are less than $0.5dB$. For smaller network sizes, *rnd* performs worse than the looping algorithm in spite of a lower network saturation, indicating that it is unsuitable for routing in this network.

In summary, HIRs that optimise for fewer MZIs in the "bar" state per path consistently outperform all other routing algorithms, both in terms of highest crosstalk power level on average and in the worst case. HIRs that minimise the number of crossings in the path show improved performance as the network scales up but are frequently outperformed by the looping algorithm, and in smaller network sizes perform no better than random. Finally, it is noted here that when considering the "highest crosstalk power level" metric, the crosstalk reductions of the HIRs compared to the Looping Algorithm are modest. However, in DWDM systems where all inputs are modulated on the same group of wavelengths, the *aggregate crosstalk* is the metric which is used to compute the power penalty. As will be shown in the next section, the HIRs show significant reductions in aggregate crosstalk compared to the Looping Algorithm.

3.7.3 Impact on Aggregate Crosstalk per Port

Next, the impact of routing algorithm choice on the aggregate crosstalk is examined. For every flow in a simulation, the aggregate crosstalk encountered by the flow at output j is defined as:

$$XT_{i,j} = \sum_{k=0}^{N} \varepsilon_k, k \neq i \tag{3.4}$$

where $\varepsilon_k = P_{leak}^{k,j}/P_{out}^{i,j}$ and i, j are the input and output ports of the victim flow. The aggregate crosstalk per output port (and therefore per flow) defines the power penalty from crosstalk that the flow is exposed to as such:

$$PP_{XT}^{i,j} = 10\log_{10}(1 - 2\sqrt{XT_{i,j}}) \tag{3.5}$$

This power penalty is then aggregated with the insertion loss as a parameter of signal power. It is noted here that if $XT_{i,j} \geq 0.25$, $PP_{i,j} \rightarrow \infty$, meaning that the signal power required for the data carried by the flow to be read at $BER = 10^{-15}$ tends to

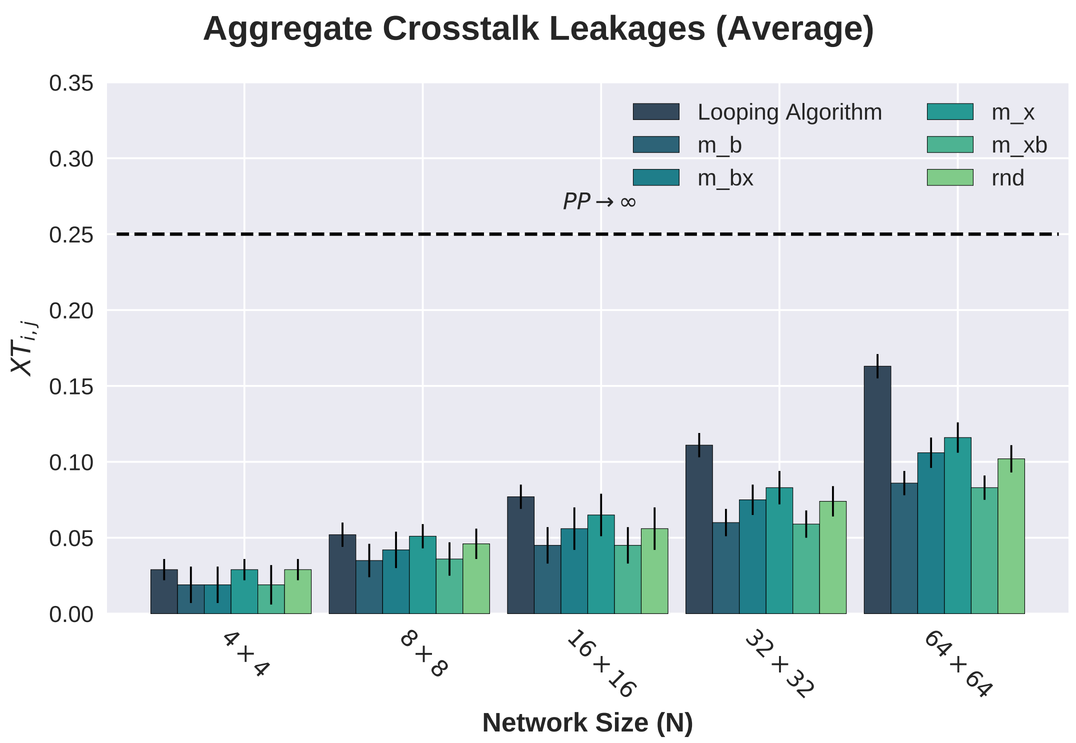


Figure 3.12: Aggregation of crosstalk on average (a). Lower is Better.

infinity. It is therefore important to assess in which cases the upper limit for $XT_{i,j}$ is reached.

To this end, the average and maximum aggregate crosstalk of the flows for each simulation is obtained over the simulation batch. The results are depicted in Figs. 3.12 and 3.13.

As with highest crosstalk level per flow, the looping algorithm generally exhibits the highest aggregate crosstalk per flow on average, compared to the HIRs. The exception to this is for 4×4 networks, where *m_x* and *rnd* exhibit similar aggregate crosstalk. As before, there are only two waveguide crossings in 4×4 networks, leading to decreased impact of HIRs that do not optimise for switch state.

On average, *m_b* and *m_xb* and, to a smaller degree, *m_bx* outperform the looping algorithm for all network sizes. The savings of *m_b* compared to the looping algorithm in aggregate crosstalk on average range between 35% for a 4×4 to 47% for a 64×64 network; the aggregate crosstalk scales more gracefully for *m_b* as the network size increases. For *m_xb* the savings range between 31% and 49%, while *m_bx* shows comparatively reduced savings against the looping algorithm, ranging from 35% for 4×4 and 64×64 networks and 19% for 8×8. The *rnd* strategy shows moderate savings

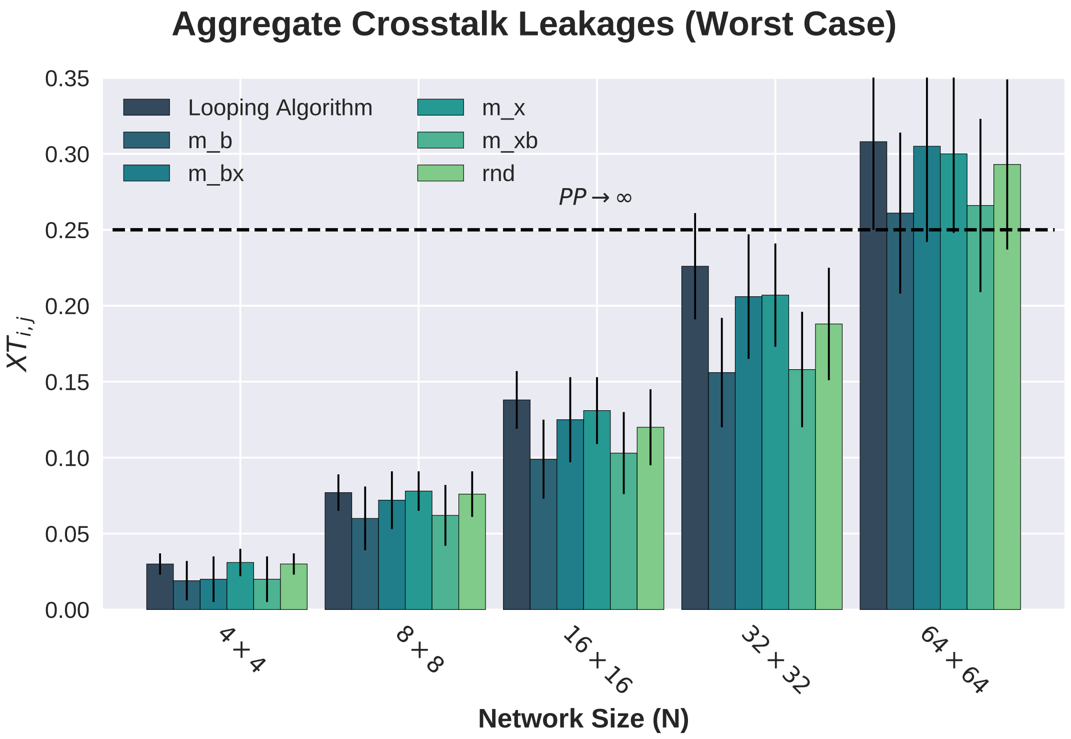


Figure 3.13: Aggregation of crosstalk in the worst case (b). Lower is Better.

for aggregate crosstalk on average in smaller networks; however the metric increases more gracefully than with the looping algorithm as the network size scales, leading to relatively good performance for 64×64. Lastly, *m_x* has the worst performance of the HIRs, but still shows less aggregate crosstalk per flow in the average case than the looping algorithm.

However, when examining the worst-case aggregate crosstalk per flow, the performance of the HIRs in terms of comparative savings against the looping algorithm changes. The *m_b* strategy still performs best overall, however the comparative savings steadily decrease as the network size increases, ranging from 37% savings for 4×4 to 15% for the largest size. *m_xb* performs similarly, albeit $2-3\%$ worse than *m_b* in all cases. The *m_bx* strategy performs similarly to *m_xb* for the smallest network size; however the performance quickly degrades, with the aggregate crosstalk increasing significantly with the network size, leading to the strategy exhibiting savings of 1% for 64×64 against the looping algorithm. *rnd* shows modest improvements compared to the looping algorithm for medium-sized networks, with the greatest comparative savings at approx. 17% for a 32×32 network. Lastly, *m_x* has the worst performance of the HIRs; in contrast to the case with average aggregate crosstalk, in the worst case this

Table 3.4: Percentage of simulations in which $XT_{i,j} \geq 0.25$ at least once.

	32×32	64×64
Looping Algorithm	23.8%	100%
m_x	11.2%	97.6%
m_b	1.6%	54.9%
m_bx	13.9%	95.3%
m_xb	2.3%	58.8%
rnd	6%	85.4%

strategy is slightly outperformed by the looping algorithm in most cases by around 1% for smaller networks. However as the network size increases, *m_x* slightly outperforms the looping algorithm ($< 5\%$)

In fact, for network sizes of 32×32 in some cases, and 64×64 and in most cases, the worst-case aggregate crosstalk surpasses the condition $XT_{i,j} \geq 0.25$. This means that at least one flow in at least one simulation was exposed to prohibitively high crosstalk. In this case the power penalty cannot be estimated using the formula assumed here, as it tends to infinity; as the aggregate crosstalk power becomes too great, it cannot be surmounted by simply increasing the signal power to counteract it and keep the SNR at an acceptable range. The portion of runs in which this situation occurs is shown in Table 3.4.

The number of simulation runs in which this effect occurs at least once differs depending on the employed routing algorithm, reflecting the results of Fig. 3.13. The looping algorithm shows the most occurrences for both network sizes; in fact for the largest size, there is always a flow which suffers this condition in the examined data. HIRs *m_x* and *m_bx* perform slightly better; however they perform worse than *rnd*, especially for the largest size. *m_xb* and *m_b* show the least occurrences, with *m_b* performing better in both network sizes; even so, with a network size of 64×64 the number of occurrences are approx. 55% with *m_b*. This indicates that scaling this network past 32 inputs is untenable, when using EO-tuned MZIs with the assumed crosstalk ratios, with all flows transmitting on the same wavelengths and assuming full path diversity. Lowering the path diversity by disallowing the least favourable available path would reduce the number of occurrences; however, this would come with the cost of reducing performance, as a lower path diversity would block more flows and therefore increase communication time.

In summary, the *m_b* and *m_xb* HIRs outperform the looping algorithm both on average and in the worst case in terms of aggregate crosstalk as well. They are therefore

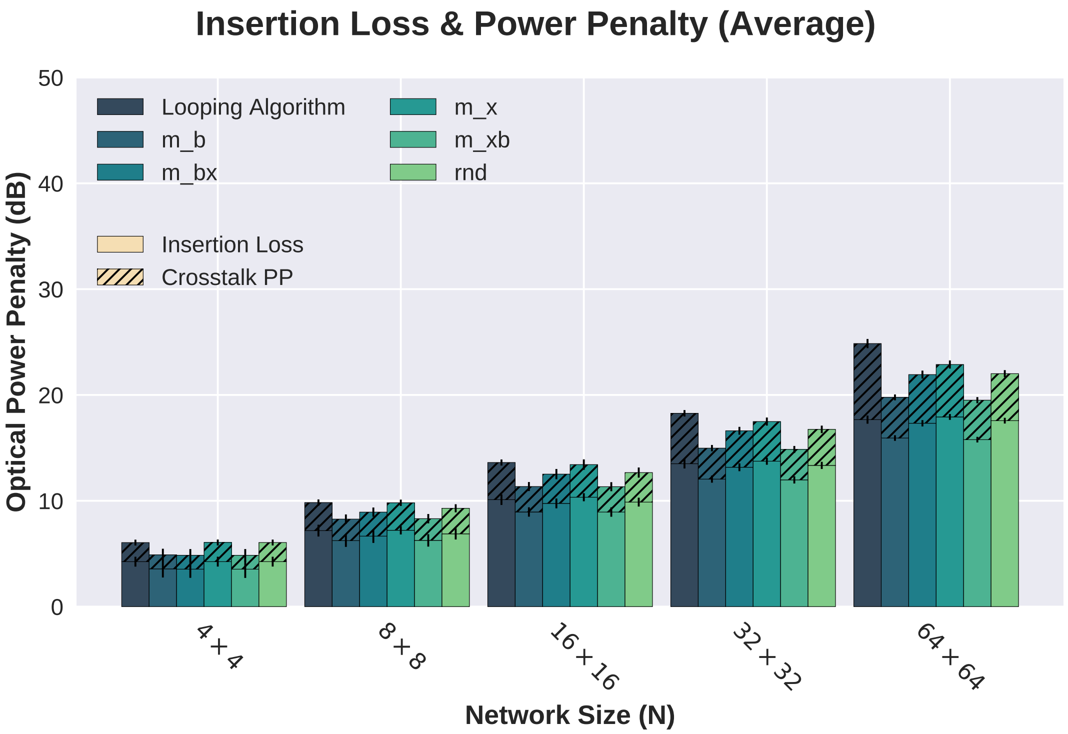


Figure 3.14: Insertion Loss and Power Penalty on average. Lower is Better.

an effective solution for reducing crosstalk in photonic Beneš switching fabrics formed with EO-tuned MZIs. However, as the number of devices in the network increases, the HIRs are progressively less able to mitigate crosstalk in the worst case, leading to the crosstalk surpassing the threshold for sizes above 16×16.

3.7.4 Impact on Optical Power Penalty

The analysis continues by examining the impact of routing algorithm choice on the optical power penalty enforced by the switching fabric on the traversing laser beam carriers. The combined optical power penalty enforced on a flow traversing from input i to output j of the fabric is defined as:

$$PP_{i,j} = IL_{i,j} + PP_{XT}^{i,j} \tag{3.6}$$

where $IL_{i,j}$ is the insertion loss, defined as:

$$IL_{i,j} = 10\log_{10}(P_{out}^{i,j}/P_{in}^{i,j}) \tag{3.7}$$

The aim here is to examine the combined effect that HIRs have on reducing the power penalty compared to the looping algorithm, by simultaneously reducing insertion loss and crosstalk power penalty. The average insertion loss for a run is combined with the power penalty ensuing from the average aggregate crosstalk for a run and depicted over the simulation runs in Fig. 3.14. The maximum insertion loss per run is combined with the maximum aggregate crosstalk per run and depicted in Fig. 3.15. For 32×32 and 64×64 networks, if the percentage of runs in which $XT_{i,j} \geq 0.25$ for a routing algorithm is below 3% (see Table 3.4), those runs are not included in Fig. 3.15; this is to examine whether the reductions in power penalty increase or decrease with network size. It is nevertheless noted that for size 32×32 networks there exists at least one occurrence for all routing algorithms.

The optical power penalty is first examined in the average case, starting from the looping algorithm. The power penalty ranges from $\sim 6dB$ to $\sim 24.9dB$ as the network size increases. The crosstalk accounts for between 30 and 25% of the power penalty, slightly decreasing with network size. As with the looping algorithm the power penalty from crosstalk is the greatest among the examined routing algorithms, this indicates that as the network size increases, insertion loss has an increasing effect on power penalty in terms of the combined penalty on average.

As expected and corresponding to the previous results, the HIRs are effective at reducing the combined optical power penalty compared to the looping algorithm. The most effective HIRs are *m_b* and *m_xb*, showing a power penalty reduction of $15-20\%$. This reduction comprises of a $10-15\%$ reduction in average insertion loss and a $25-45\%$ reduction in crosstalk power penalty. The least effective strategy is again *m_x*, showing at most 6% improvement compared to the looping algorithm. The *m_bx* strategy shows significant improvement ($\sim 20\%$) for the smallest network, but the performance degrades as the network size increases, ultimately showing savings of $\sim 8\%$ for the largest network. Conversely, *rnd* exhibits no improvement against the looping algorithm for the smallest network but improves as the network size increases, ultimately showing $\sim 10\%$ savings for the largest network.

Next, the optical power penalty in the worst case is examined. Here, from size 4×4 to 16×16, the power penalty of the looping algorithm is between $6-18dB$ compared to $6-13.6dB$ for the average case. For larger sizes where the aggregate crosstalk increases beyond the prohibitive limit, the power penalty from crosstalk tends to infinity. This occurs with *m_x* and *m_bx* for 32×32 and with all routing algorithms for 64×64.

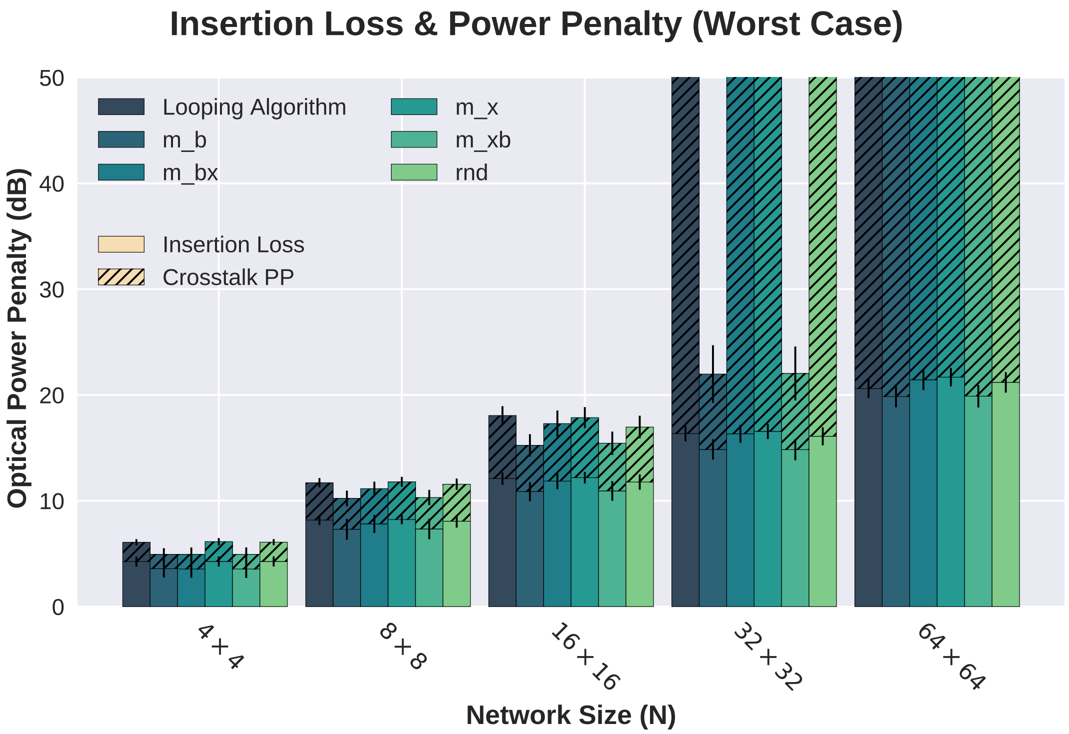


Figure 3.15: Insertion Loss and Power Penalty in the worst case. Lower is Better.

For the smaller sized networks, *m_b* shows savings between $\sim 19-15\%$ with the benefit decreasing with network size. *m_xb* shows similar savings but performs marginally worse (2%) for the smallest network. The performance of *m_bx* is comparable to *m_b* but deteriorates with the network size increase, while *rnd* shows no improvement for the smallest case but 5% improvement for 16 endpoints. Lastly, *m_x* is outperformed by all other routing algorithms.

As mentioned, for size 32 × 32, there is at least one simulation run where the prohibited crosstalk condition occurs for all routing algorithms. However, with *m_x*, *m_bx* and *rnd* very few of these cases occur; the combined power penalty can therefore be examined for the majority of cases. Here, the *m_b* strategy exhibits a maximum combined power penalty of approx. $21.9dB$, comprising of $14.8dB$ of insertion loss and $7.1dB$ crosstalk power penalty. *m_xb* shows a combined power penalty of $22dB$ ($14.8dB$ insertion loss, $7.2dB$ crosstalk power penalty). This indicates that with larger network sizes, the *m_b* routing strategy is the most effective at reducing the combined optical power penalty.

In summary, the HIRs can be effective at reducing the combined power penalty both in the average and the worst case. However they are not immune to the effects of crosstalk, which become prohibitive as the network size increases past 16 × 16. However, on average, the aggregate crosstalk never reaches the prohibitive condition for all the examined network sizes. This fact, combined with the low number of occurrences of the prohibitive condition for 32 × 32 and some HIRs, indicates that there is room for improvement and that the worst-case optical power penalty can be improved to a degree where larger network sizes become realistic.

3.7.5 Impact on Laser Power

The analysis concludes by examining the impact of routing algorithm choice on required laser power per flow in the average and the worst case. This is given by the following equation:

$$P_{laser} = N_{\lambda} E_{laser} P_{sense} 10^{PP_{i,j}/10} \tag{3.8}$$

where P_{sense} is the signal power required at the destination's photodetector, N_{λ} is the number of wavelengths, E_{laser} is the laser efficiency and $PP_{i,j}$ is composed by all the losses and power penalties encountered by the carrier beam from the source to the destination of the flow. This includes insertion losses from the laser, coupling in and out of the chip, modulator, filters and photodetector, as well as the combined power penalty examined previously. These are inputted as parameters to the simulator and used for each simulation; the properties can be found in Table 3.3. It is noted that an off-chip laser is assumed per input port for simplicity. As mentioned, the laser power model is integrated into PhINRFlow from the DSENT photonic network simulator. The laser power per port in the average case, obtained from the average combined power penalty as derived in Sec. 3.7.4, is depicted in Fig. 3.16 for all examined network sizes.

As expected, laser power per flow on average scales exponentially with the network size in all cases. When using the looping algorithm, the average required laser power per flow ranges between 0.02 ± 0.003 to $1.6 \pm 0.3W$. The variability in the results also increases with network size, as indicated by the error bars; as the network size increases, more carrier beams cross paths when the network is in full saturation. This leads to the permutation offered by the bisection workload to have a greater impact in the variability of average laser power per flow.

The exponential scaling of laser power with network size leads to loss reductions offered by the HIRs having an increasing impact as the network size increases. The *m_b* routing strategy exhibits an average per-flow laser power between 0.016 ± 0.005

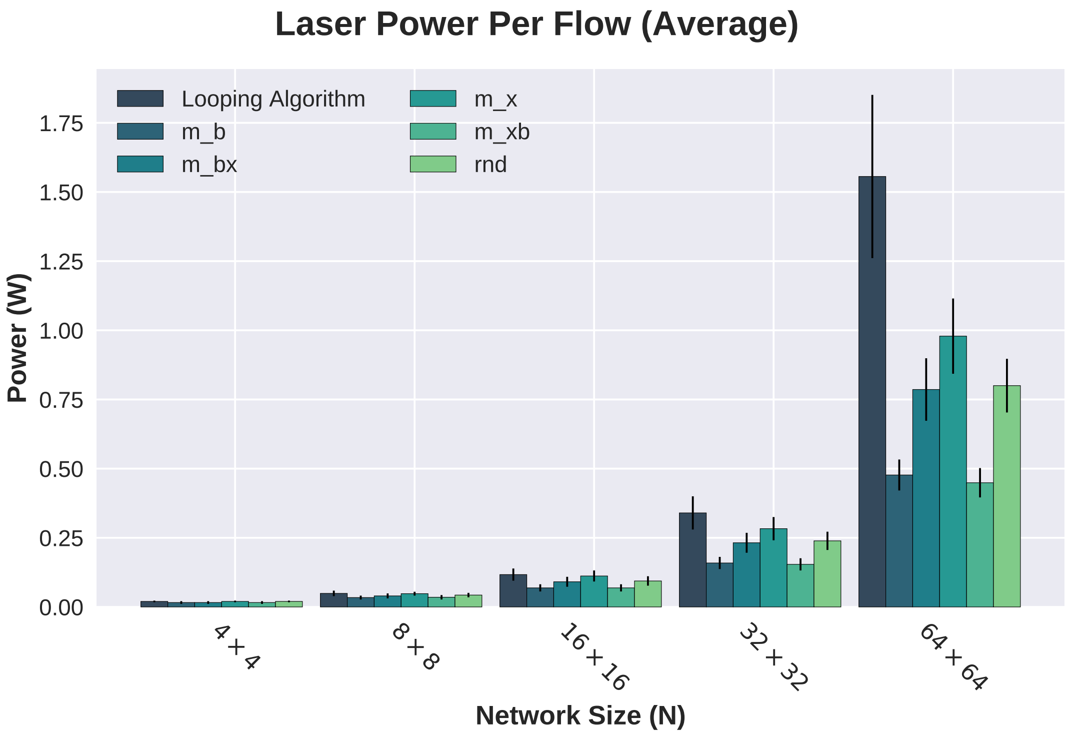


Figure 3.16: Required laser power per port in the average case. Lower is Better.

and $0.47\pm0.06W$. This shows a reduction ranging between $\sim 20-70\%$ compared to the looping algorithm. However, in the smallest network the reduction is not significant, as the means of the looping algorithm and *m_b* are within one standard deviation of each other. The *m_xb* routing strategy performs marginally better due to the concurrent reduction in insertion loss and power penalty; the average laser power per flow ranges between 0.016 ± 0.005 and $0.45\pm0.05W$, with the comparative savings ranging between $\sim 20-71\%$.

In fact, due to the exponential relationship of power penalty and signal power, even the HIRs which showed worse performance still exhibit substantial laser power savings. *m_bx* exhibits savings between $\sim 20-49\%$, *rnd* between $0-48\%$ and *m_x* between $0-37\%$. It is noted that for the latter two strategies there are no savings in the smallest network and the moderate savings shown in medium-sized networks come at the expense of higher contention and therefore decreased performance, as discussed in Chapter 5. However, these cases exemplify how comparatively small improvements in combined power penalty can lead to meaningful savings in signal power and therefore laser power, at least in the average case.

As laser power per flow depends on the combined power penalty, in the worst case

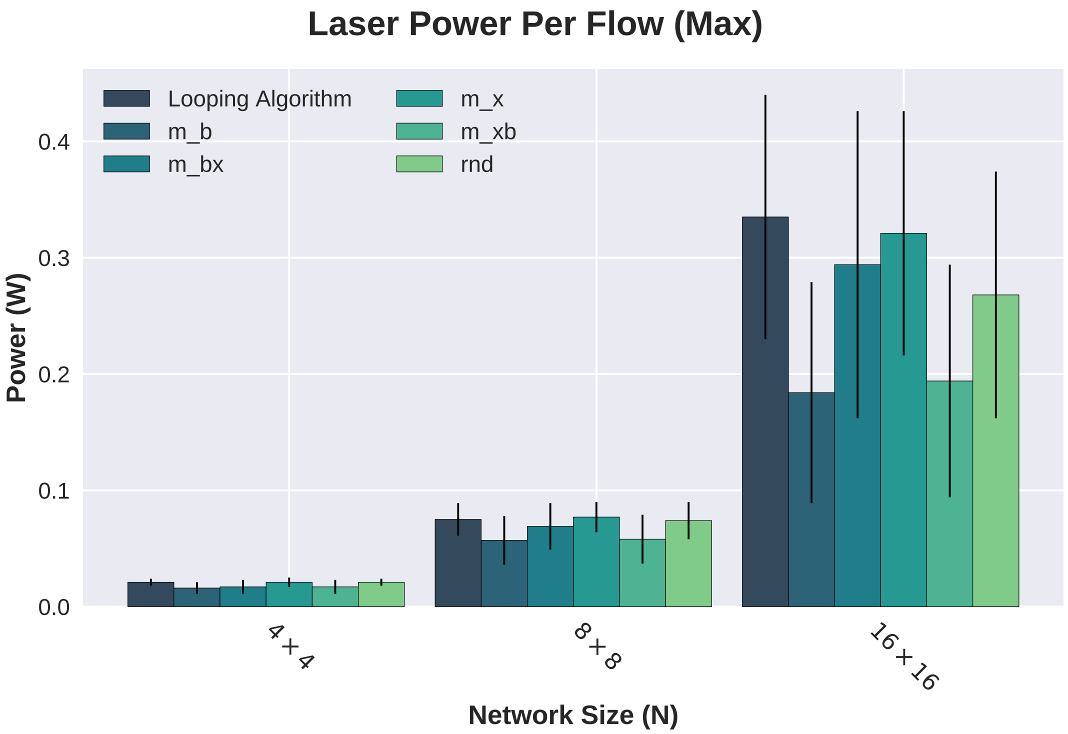


Figure 3.17: Required laser power per port in the worst case. Lower is Better.

it cannot be estimated for large networks using the model assumed for this thesis. Therefore, the worst-case laser power per flow, depicted in Fig. 3.17, is shown only up to a network size of 16 × 16.

When using the looping algorithm, the maximum laser power ranges between $0.021 \pm 0.003 - 0.34 \pm 0.1W$. The variability increases with network size, indicating that the permutation of the flows has an increasing effect as the network size scales up; this reflects the findings of Sections 3.7.2 - 3.7.4. The performance of the HIRs also remains consistent with the previous findings; *m_b* and *m_xb* show similar performance, with *m_b* showing comparative savings of $\sim 24-45\%$, and *m_xb* showing $\sim 19-42\%$. For size 4×4 the savings are slightly better than the average case due to the marginally worse performance of the looping algorithm; the benefits are lessened compared to the average case as the network size increases. This is shown clearly for *m_bx*, which in the average case showed 21% savings for 16×16; in the worst-case laser power, this drops to $\sim 8\%$. The *m_x* strategy performs similarly to the looping algorithm in the worst case; the 5% savings exhibited over the simulation runs is well within one standard deviation of the looping algorithm and therefore not significant. Lastly, *rnd* shows modest savings compared to the looping algorithm for the largest size ($\sim 15\%$), but no significant savings for smaller sizes.

In summary, HIRs can be effective at reducing signal power and therefore laser power in both the average and the worst case.

3.8 Known Modelling Limitations

PhINRFlow is extended from an electronic interconnection network simulator operating at the flow level. Coupled with the fact that communication time is derived based on events which are scheduled during the simulation (send, receive and computation events), this means that time and therefore energy per bit are modelled at a coarse granularity.

In addition, photonic elements in PhINRFLow are not modelled physically, i.e. the effect of their dimension is not studied. Insertion loss, crosstalk and tuning power values are assumed from constructed chips demonstrated in the literature. The propagation loss due to waveguide length which is considered is the average reported in [LZZ$^+$16]; therefore insertion loss deviation due to waveguide length is not considered.

Based on these two factors, propagation time of photonic signals in switching fabrics is not modelled in PhINRFlow. Although this could be estimated and factored into the model, a more appropriate methodology would be to simulate a chip in one of the circuit-level simulators discussed in Sec. 3.1. As the integration of photonic lightpath modelling in PhINRFlow has the objective of revealing the dynamic relationship of photonic effects, routing algorithms and employed communication workloads, this was considered to be out of scope. Nevertheless, it is to be noted that due to the lack of propagation time, modelling energy consumption from lasers would be challenging in this tool.

In addition to the above, this work assumes that all photonic beams traversing the network are co-polarised and out of phase, leading to a worst case crosstalk estimation. While this is a justified assumption as discussed in [RSS09], it entails that the impact of polarization and phase relationships between photonic signals on the crosstalk is not captured by the model. Such modelling has been conducted by Dupuis and Lee [DL17].

Finally, non-linear effects of the interaction between photonic beams and devices (e.g. scattering effects, FCD/FCA, four-wave mixing) are not modelled.

3.9 Summary

This chapter has contributed to the state-of-the-art by proposing a methodology for evaluating the effect of network traffic and intra-switch routing algorithms on the performance of photonic switching fabrics, focusing on fabrics composed by EO/TO-tuned MZIs in the Beneš topology. To the best of the author's knowledge, this is the first proposed methodology for performing this type of network traffic-driven analyses on PSFs, where the photonic device level, network control level and application traffic level are combined. The network simulator that has been augmented to investigate this methodology, PhINRFlow, has been described. The model for the switching fabric in focus has been discussed. The accuracy of the model has been evaluated and compared against measured data from two constructed photonic switching fabrics from the literature.

Further, the chapter has discussed the standard routing algorithm used to control Beneš networks, as well as the concept of HIRs, which are proposed as contributions of the thesis in the next chapters. The known modelling limitations have also been examined. Finally, the novel methodology is demonstrated by evaluating the effect of routing algorithm choice on exhibited insertion loss, crosstalk, combined power penalty and required signal power and laser power. This is conducted considering different sizes of switching fabric and using a bisection workload.

The findings of the evaluation indicate that routing algorithm choice can have a significant impact on the photonic metrics examined. The most impactful routing algorithms show comparative savings to the state-of-the-art looping algorithm for all metrics, with the savings in laser power increasing proportionately to the network size. The most impactful routing strategy proposed in Chapter 4 is effective at reducing the highest level and aggregate photonic crosstalk, as well as the combined power penalty. The combined power penalty reduction compared to the looping algorithm can lead to laser power savings for the largest network size, ranging from $\sim 71\%$ in the average case to $\sim 45\%$ in the worst case. The most impactful strategy proposed in Chapter 5 exhibits slightly improved performance in the average case, while other routing strategies exhibit lower comparative savings.

Comparing the findings of Sections 3.7.2 and 3.7.3 indicates that there is scope for improvement by assigning different sets of wavelengths to different flows which use the network concurrently. The results in Section 3.7.2 depict the highest crosstalk power level encountered by the flows in the average and worst case, while those in 3.7.3 depict the difference in aggregate crosstalk leakages in the average and worst case.

This aggregate crosstalk considers that all flows use the same group of wavelengths and therefore the leakages aggregate as intra-channel crosstalk. However, as it is known that inter-channel crosstalk is much less impactful to the power penalty, a wavelength group arbitration scheme may be considered; in this scheme, the available wavelengths would be separated into groups, or λ-groups, with each flow being assigned a λ-group based on a pre-defined arbitration scheme. This way, intra-channel crosstalk would be reduced, therefore reducing the impact of the crosstalk power penalty and potentially enable the use of larger network sizes. Of course, reducing the number of λs per flow would reduce the performance of the network; however if the crosstalk power penalty is reduced enough by this scheme, the data-rate per λ could be increased to mitigate the performance loss. A lower channel spacing that what has been considered in this work could also be assumed, thereby increasing the total number of available λs. However, this endeavour would require re-working the simulator to include insertion loss and crosstalk profiles for wavelength regions, as well as gathering this data from real devices. It is therefore considered to be future work in the context of this thesis.

Chapter 4

Paper 1: Scalability Analysis of optical Beneš networks based on Thermally/Electrically Tuned Mach-Zehnder Interferometers

Scalability Analysis of optical Beneš networks based on Thermally/Electrically Tuned Mach-Zehnder Interferometers

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

Jose A. Pascual
University of the Basque Country
San Sebastian, Spain
joseantonio.pascual@ehu.es

Javier Navaridas
School of Computer Science, The University of Manchester
Manchester, United Kingdom
javier.navaridas@manchester.ac.uk

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

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

ABSTRACT

Silicon Photonic interconnects are a promising technology for scaling computing systems into the exa-scale domain. However, significant challenges exist in terms of optical losses and complexity. In this work, we examine the applicability of thermally/electrically tuned Beneš network based on Mach-Zehnder Interferometers for on-chip interconnects as regards its scalability and how optical loss and laser power scale with the number of endpoints. In addition, we propose three hardware-inspired routing strategies that leverage the inherent asymmetry present in the switching components. We evaluate a range of NoC sizes, from 16 up to 1024 endpoints, using 4 realistic workloads and found very promising results. Our routing strategies offer an optical loss reduction of up to 32% as well as a laser power reduction by 33% for 32 endpoints.

CCS CONCEPTS

• **Hardware** → **Emerging optical and photonic technologies**; • **Networks** → *Network on chip*; Network experimentation.

KEYWORDS

Silicon Photonics, Optical Beneš Networks, Scalability Analysis

ACM Reference Format:
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 *12th International Workshop on Network on Chip Architectures (NoCArc '19), October 13, 2019, Columbus, OH, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3356045.3360715

1 INTRODUCTION

As high-performance computing (HPC) advances into the exa-scale domain, numerous system scalability challenges present themselves. HPC commonly supports massively parallel workloads, which require a substantial level of communication between compute elements. It is widely acknowledged that interconnection networks constitute a scalability bottleneck for future HPC systems [19]. Furthermore, recent evidence suggests that conventional electrical interconnects will not be able to keep up with system scalability trends in terms of performance, while satisfying the ever-more stringent power and area constraints [23].

Optical interconnects based on Silicon Photonics (SiPh) have emerged as a promising technology to augment, if not substitute, traditional interconnects at the NoC and inter-chip level. The technology's CMOS compatibility, its capacity for high-bandwidth through dense-wavelength-division multiplexing (DWDM) as well as distance-independent energy consumption show substantial promise [15]. However, state-of-the-art devices suffer from limitations that can lead to increased optical losses, complexity and package cost [4].

In this paper, we investigate the scalability potential of an optical Beneš Network formed with thermally/electrically tuned Mach-Zender Interferometers [8]. Our aims are:

- To evaluate the performance of a network based on this technology under realistic workloads.
- To propose three hardware-inspired routing strategies which leverage the network's underlying asymmetrical behaviours.
- To demonstrate the benefits of the strategies at different system scales and ascertain the network's applicability to the on-chip domain.
- To evaluate the optical losses of the system at scale and identify the chief contributors.
- To assess the laser power required for driving the network.

In addition, we aim to identify the best use cases for this network; that is, whether it is beneficial as the main switching fabric on a chip serving many endpoints, or if it is more realistic, in terms of losses and energy consumption, to constrain the size for on-chip territory and consider a nested topology paradigm for inter-chip fabric. The way that optical loss scales due to the chief contributors in conjunction with laser power yields valuable insight for the applicability of this network.

NoCArc '19, October 13, 2019, Columbus, OH, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6949-7/19/10…$15.00
https://doi.org/10.1145/3356045.3360715

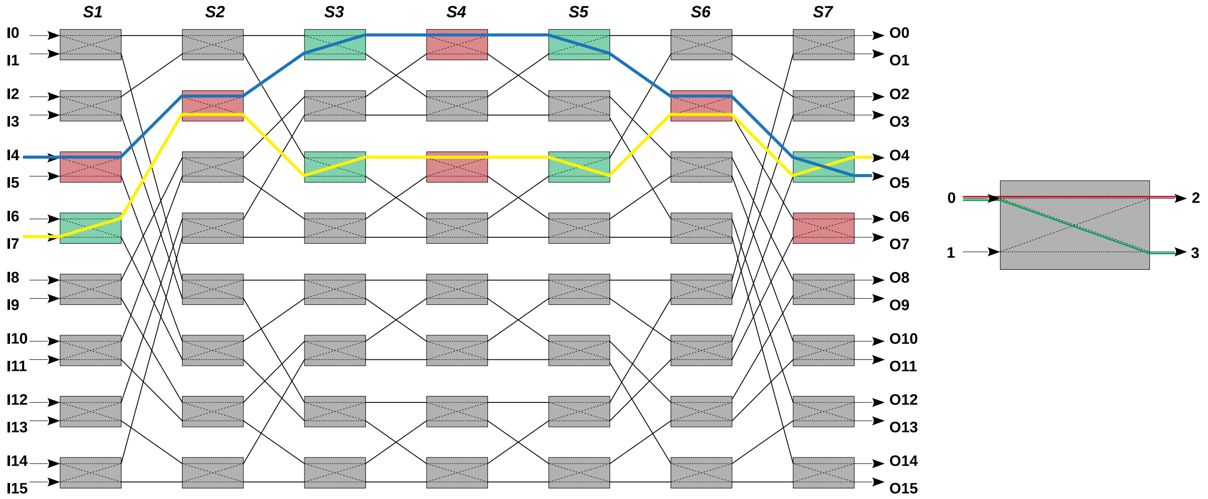


Figure 1: Left: Topology diagram of the 16x16 Beneš network. The blue flow traverses from I4 to O5 and the yellow from I7 to O6, green shows MZIs in "cross" state and red in "bar" state. Right: An MZI with port numbers.

2 BACKGROUND

In this work, our interest is to examine the scalability of Beneš networks [5] for their use with electro-optic Mach-Zender Interferometers (MZIs) [11]. Our analysis is based on [8], where a 16×16 Beneš network is constructed out of seven stages of 2×2 MZIs. The authors contribute an experimental demonstrator and extract a full characterisation of the underlying components. They describe the design and fabrication process, as well as the optimisations undertaken to reduce the optical loss exhibited by the components (e.g. optimised tapers for waveguide crossings, optimised MZI design etc.). In addition, the thermal and electrical tuning power is reported for each MZI which we used as a basis for our analysis. Fig. 1 depicts a 16×16 Beneš topology; the rectangular components represent 2×2 MZI switches. Note that we consider binary MZIs which are either at a"cross" or a "bar" state. Not to be confused with tri-state MZIs [9], with a third, "blocking" state where the phase tuning forces the MZI to be completely blocked. Tri-state MZIs could afford interesting possibilities; idle elements within a $N \times N$ fabric can be tuned to the "blocking" state to dramatically reduce crosstalk, one of the main limitations to scalability.

The Beneš network is a Clos-network variant constructed from 2×2 switches. It requires the minimum number of 2×2 switches to connect 2^i ports in a rearrangeably non-blocking fashion [5]. As such, this paradigm lends itself well to the case of using MZIs as base switching components. Additionally, due to the inherently buffer-less nature of optical communications, packet-switching in optical networks requires electro-optic conversions, which generate huge energy and latency overheads which are undesirable in terms of scalability [24]. However, these can be ameliorated by using circuit switching techniques [2]. The Beneš network's properties can therefore be taken advantage of in terms of path diversity, as we explore in this work.

2.1 SiPh Interconnects and Technology

SiPh is acknowledged by the interconnects community as a key enabler for scaling interconnect systems [14, 18]. With recent advances on photonic integrated circuits (PICs) using CMOS-compatible processes, interest has been generated for Optical Networks-on-Chip (ONoCs). A comprehensive study of these can be found here [24] with Corona [21], Amon [23]and Venus [17] being notable examples. SiPh-enabled architectures at the system level have also been proposed (e.g. [7] [10] [22]). SiPh-enabled systems have also emerged for use in data-centre networks (e.g. [12] or [3]).

The underlying components that make SiPh interconnects possible (e.g. waveguides, microring resonators, MZIs, multimode interferometers, transceivers, lasers etc.) are the subject of wide research with novel components being proposed very frequently [20].

For instance, in this work we consider a waveguide (hereafter WG) propagation optical loss penalty of 1.18 dB/cm. Thraskias et al. on the other hand mention WG-incurred optical losses of as low as 0.2 dB/cm [20]. We note that propagation loss due to WGs is highly dependent on device technology; nevertheless, the aforementioned survey illustrates the rate of progress on the technology front. One other key set of components necessary for interconnects is switches; a comprehensive review of the state of the art of SiPh switches can be found in [4], and on MEMS switches for more general optical communications here [25]. As with the model we investigate in this work, the SiPh switches examined in [4] are commonly based on the Beneš topology as well as MZIs with thermal/electrical tuning.

3 SIPH BENEŠ NETWORK AT SCALE

3.1 Scalability Challenges and Experiment Motivation

As discussed, the MZIs we consider in this paper are *thermally* tuned to reach a "cross" state and *electrically* tuned to reach a "bar" state from that "cross" state. As such, more power is required for an MZI to hold the "bar" state. Furthermore, an MZI in the "bar" state exhibits substantially more Insertion Loss (hereafter, ILoss) than an MZI in a "cross" state. Thus, using MZIs in the "bar" state generates significant undesirable overheads. In addition, note that an ILoss penalty is incurred for each WG crossing and that each connection between MZIs entails a different number of crossings. Aggregating the ILoss penalty encountered by a flow can lead to excessive demands on the lasers as the system scales up.

In fact, depending on the photodetector specification, ILoss is a defining factor for the power of the laser beam which carries the flow [16]. Excessive ILoss means a more powerful laser is required, which affects the applicability context of the technology in terms of chip power budget. As such, it is important to reduce these metrics to achieve scale both for on-chip and inter-chip domains.

These effects, combined with the need to evaluate the system under realistic workloads, outline our experimental motivation. We aim to understand the scalability implications relating to ILoss and energy consumption, as well as to evaluate a set of hardware-inspired routing strategies which we outline below.

3.2 Routing in a SiPh Beneš network

In order to facilitate route allocation and choice, each modelled MZI must be electrically/thermally tuned to the required state. At each flow injection, the control process calculates the possible routes the flow may take through the network. For N endpoints, each flow can use a maximum of $N/2$ different paths. The route calculation generates potential paths by varying the interconnected MZIs to be traversed per stage in the left half of the potential path to produce path diversity. The right half of the path is kept stable, ensuring correct destination addressing.

Once all options have been calculated, the best path is selected based on the hardware-inspired routing strategies outlined in the next subsection. After selection, the control process iterates through the potential flow paths and, for each encountered MZI, assesses its ability to preserve or switch to the required state. Note that for an MZI in a "bar" state where a previous flow has reserved ports 0 and 2 for example, ports 1 and 3 may be used by another flow. The corresponding scenario applies for the "cross" state as well. If the path assessment completes successfully, the path is reserved by tuning the corresponding MZIs if needed. Otherwise, the process continues for the remainder of the potential paths.

Path selection and flow scheduling in this network is a non-trivial problem. MZI states substantially affect the network's ability to fully take advantage of path diversity and allocate the best path. To illustrate this, consider the scenario depicted in Fig. 1, where a flow is scheduled from I7 to O6, shown in yellow. However, the network is *already* serving a flow from I4 to O5, shown in blue. The control process has already assigned MZI states to serve the blue flow; in this scenario, the two senders can share the second MZI they encounter (which has a state assigned), constraining the available paths the controller can assign for the yellow flow.

There are two main approaches for designing controllers for this network. The first option is to design a centralised controller, which controls MZI states and receives path requests from the endpoints. This controller has full knowledge of the network state and can therefore allocate paths to flows based on state-aware decisions; we examine some of these in the next section. The second option is distributed control; *every* MZI is controlled by a separate controller, which is connected to its neighbouring MZIs through an underlying control network. This option enables cascaded path selection, whereby senders request a path from their neighbouring MZI, the request gets forwarded to the MZI's neighbour and so on. In this case, an MZI which cannot change state and serve a flow would send a failure message to the previous stage in the network.

However, various challenges arise with controller design. Controller complexity, flow scheduling and latency are aspects which affect both the centralised and distributed design types. For the distributed design especially, how the back-propagation of failure and success messages affect latency is a substantial question, as is whether the design's greedy nature would be able to reach near-optimal paths. As such, controller design is a research question in itself, which is out of scope for this work.

3.3 Hardware-inspired Routing Strategies

Yuen and Chen [26] present an interesting approach to reducing ILoss and power consumption for switches based on micro-ring resonators (MRRs). They propose a heuristic for leveraging asymmetric behaviours in the underlying switching components. They demonstrate improvements with respect to the baseline case, with more than 30% reduction in ILoss and power savings which increase with the degree of asymmetry. Inspired by this methodology, we propose the following three hardware-inspired routing strategies for networks based on e/o tuned MZIs:

- **min_crossings** prioritises the paths with the least amount of WG crossings to reduce ILoss.
- **min_state_changes** prioritises the paths with the fewer MZI state changes to overlap flows and reduce energy consumption through MZI reuse.
- **max_state_cross** prioritises the path with the most MZIs in "cross" state aiming to reduce both ILoss and energy consumption.

4 EXPERIMENTAL SETUP

The primary focus of our experimental work is to evaluate the proposed routing strategies as well as the scalability of the Beneš network and its applicability to the on-chip domain. To make a more realistic evaluation, we consider various realistic traffic models.

4.1 Simulator and Model

We use phINRFlow (photonic Interconnection Network for Research Flow-level Simulation Framework), an in-house flow-level simulator for photonic interconnects. It affords a light footprint, is highly scalable and includes the main aspects necessary for photonic interconnects. It also includes various 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 candidate for exa-scale systems. The simulator inherits functionality from INRFlow [13] wherein a detailed description of the simulator's methodology and workloads may be found.

The system is modelled within the simulator as a new topology. All the links are uni-directional with traffic flowing from "left"

Table 1: Simulation Parameters

(a) Optical Loss and Power Consumption, as reported in [8].

Component	Insertion Loss	Tuning Type	Power Cons.
WG	1.18 dB/cm	Thermal	0-26 mW
Beneš Stage	0.4386 dB	Mean, STD	15.725, 6.608
WG Crossing	0.05 dB		
"Cross" MZI	0.4 dB	Electrical	3.28-5.88 mW
"Bar" MZI	1.4 dB	Mean, STD	5.166, 0.428

(b) DSENT Simulation Parameters.

Parameter	Values
Core Rate	2 GHz
#λ per laser	32
Laser efficiency	0.25
Detector Loss	1.0 dB
Extinction Ratio	1.0 dB

(c) Number of flows per workload.

Workload	# Flows
Randomapp	10000
Bisection	N
Mesh	$\leq 4N$
Hotregion	10000

to "right". For the purposes of this analysis, we assume that each endpoint is supplied with a laser source and can communicate with all other nodes independently. Lu et al. [8] report the power consumption of both thermal and electric tuners located within the MZIs. In our experimental process, we use the reported tuner power consumption metrics by fitting them to a normal distribution, from which we then assign values to each MZI for the larger network sizes. In order to evaluate laser power and energy consumption we use DSENT [16]. We explore the impact of varying the data rate per wavelength (denoted as λ), as well as that of the routing strategies. We then use the best data rate per λ in our switching energy consumption evaluation.

Lastly, we consider a centralised control process which allocates paths to flows in a first-come-first-served fashion by controlling the MZIs. Based on the network state at each request, the controller uses the enabled the routing strategy to recommend a path.

4.2 Workloads & Metrics of Interest

We use the following workloads supported by phINRFlow. Note that they 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.
- **Bisection** Nodes are split into pairs at random and nodes in a pair communicate with each other.
- **Mesh** A 2D stencil commonly exhibited by scientific codes.
- **Hotregion** Generates a non-uniform load, with 25% of the traffic being directed to the upper 12.5% of the network. The rest is allocated a destination randomly.

Under each workload, every endpoint injects a number of flows into the network per configuration, depending on network size and workload properties. For clarity, Table 1c summarises the configuration setup parametrised to the number of endpoints (N) where required. For a more detailed description of the workloads, see [13].

Lu et al. [8] also report the ILoss per component for the proposed wavelength region. Based on this, the ILoss per component and power consumption we consider for our model are found in Table 1a. Our study assesses the following metrics using both simulators:

- **Average ILoss** ILoss is measured on a per-flow basis and is defined by the state of each MZI traversed, the number of crossings and the length of the path taken by the flow. Note that for ILoss induced by WG length, we assume 0.4386 dB per Beneš stage as described in [8].
- **Max ILoss**: The worst-case ILoss experienced by a flow.
- **Power per Laser**: We measure the power-per-laser requirements for different network sizes.

5 RESULTS AND DISCUSSION

5.1 Insertion Loss

As mentioned, ILoss exacerbates the power consumption problem of optical interconnects. The three main ILoss contributors in our model are WG crossings, WG length and MZI-incurred loss for each state. Fig. 2 shows the average and max. ILoss per strategy under each workload. For completeness, we include the absolute max. ILoss, calculated from the original device parameters (see Table 1a) in the worst-case configuration, i.e. max. WG length, max. number of crossings and max. number of MZIs in the "bar" state for each size. We include the average ILoss to illustrate the variability among different paths and to motivate for organisations with more balanced ILoss. To help understanding how the factors contribute to the overall ILoss, Fig. 3 presents a broken-down view.

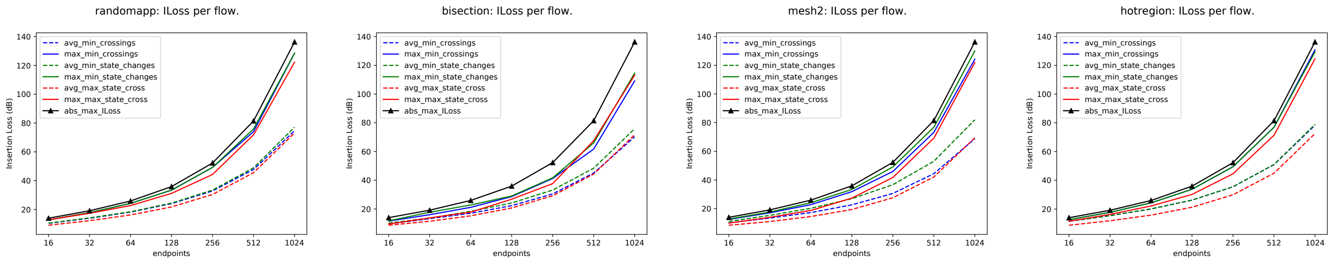


Figure 2: Insertion Loss (dB)

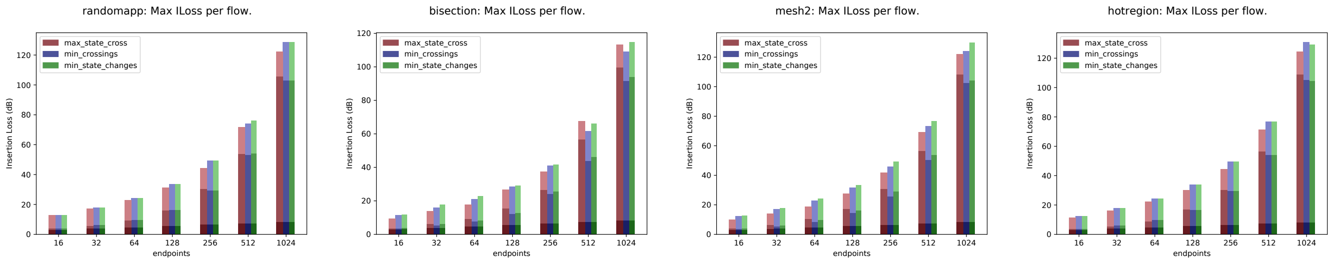


Figure 3: Maximum Insertion Loss Breakdown (dB). Darker: WG ILoss. Mid: Crossings ILoss. Lighter: MZI ILoss.

Firstly, average and maximum ILoss scale proportionately to network size in all cases. However, the exhibited max. ILoss is always less than the absolute max. Therefore, the original's report of 14 dB worst-case ILoss was conservative, based on these results. This demonstrates the benefits of routing-based solutions for underlying hardware constraints. Interestingly, as the network size scales up, all the routing strategies exhibit an increasing reduction in max. ILoss. The largest reduction is exhibited under the bisection workload, and the least under hotregion. In both cases, *max_state_cross* exhibits the most reduction.

The *min_crossings* strategy has little impact on ILoss for most workloads, up to 128 endpoints. In most cases, the strategy's behaviour is almost identical to the others for both average and maximum ILoss. The largest benefit is exhibited under the mesh2 workload with respect to *min_state_changes*, but again is very small. Nevertheless, it never outperforms the *max_state_cross* strategy up until that size.

For sizes ranging from 256 to 1024 endpoints, the max. ILoss incurred by crossings increases substantially. This is because the number of crossings per path scales proportionately to the number of endpoints rather than the number of stages, as is the case with MZI and WG-incurred ILoss. For more than 128 pairs, crossings-incurred ILoss dominates over the other factors; indicatively, for mesh2 under *min_state_changes*, max. ILoss from crossings is approximately 46%, 61% and 74% of the total for 256, 512 and 1024 endpoints respectively. Clearly, crossings-incurred ILoss is the primary scalability concern for optical Beneš networks. This may be ameliorated through chip floor-planing optimisation by minimising the number of crossings; this is a direction we plan to investigate in the future.

The *min_state_changes* strategy yields very little benefit overall. The only instances where it outperforms any other strategy are for 16 endpoints, and then only by 1-2 dB. The only advantage is for larger sizes, where it reduces crossings-induced ILoss with respect to *max_state_cross*; however, it never manages to reduce MZI-incurred max. ILoss as much as the latter. One exception is under bisection for a network with 512 endpoints; however the reduction is approx. 1 dB.

Max_state_cross is the best strategy overall in decreasing average ILoss per flow for sizes up to 256 endpoints. The most reduction is encountered under the bisection workload (32%, 64 endpoints). This permutation workload keeps the network near saturation and exploits path diversity, thereby allowing the routing strategy to have a pronounced impact as discussed previously. From that size onward, max. ILoss due to crossings reduces this strategy's impact. Nevertheless, this strategy reduces total ILoss substantially enough to outperform its contenders in most cases, with large sizes in bisection being the only exceptions. As per the insights above, *min_crossings* and *max_state_cross* are the two most useful strategies to adopt for routing. Combining these two strategies is an interesting future work possibility.

5.2 Laser Power

Here, we discuss the laser power required for various sizes of the network, as well as the impact of the data rate per λ. We conduct a parameter sweep using DSENT [16] and the max. ILoss derived from our phINRFlow experiments under randomapp. The DSENT configuration parameters we use are shown in Table 1b. We chose 32λ as this would allow for more degrees of freedom for DWDM while sticking to the 100 GHz channel spacing (ITU-T G.694.1 standard). We derive laser energy consumption from laser power (DSENT), execution time (phINRFlow) and payload for various data rates. We conduct a similar parameter sweep in phINRFlow for the corresponding switching energy per data rate, shown in Fig. 4.

Firstly, switching energy scales more gracefully than laser energy, which can be up to 3 orders of magnitude larger for a 256-endpoint NoC. A larger data rate per λ increases the laser energy consumption for each network size (Fig. 4 left) but reduces that from switching (Fig. 4 right). However, increasing the rate from 4 to 8Gbps/λ does not increase laser energy consumption substantially (<1%), whereas increasing from 8 to 16 Gbps/λ increases consumption by approx. 33%. Consequently, the least energy consumption from lasers and switching is afforded at 8Gbps/λ, which for 32λ adds up to a total of 256Gbps per endpoint.

Based on this data rate, we show the impact of our routing strategies on required power per laser for different network sizes (Fig. 5). These results conclusively show that laser power, affected by max. ILoss, is the main scaling inhibitor. Indicatively, with *max_state_cross*, a network of 256 endpoints requires 206 W per laser. Reducing this without changing technology parameters would entail sacrificing throughput, by reducing the number of λ. However, with *max_state_cross*, a 32-endpoint NoC requires 13.4 W for lasers. Considering that the NoC may take up to 24% of a SoC's power budget [1], a typical 100W budget as exhibited by regular server-grade processors could very well accommodate for this. Larger many-cores such as Intel Xeon Phi have a much higher power budget (around 300W), which would allow for 75W to be

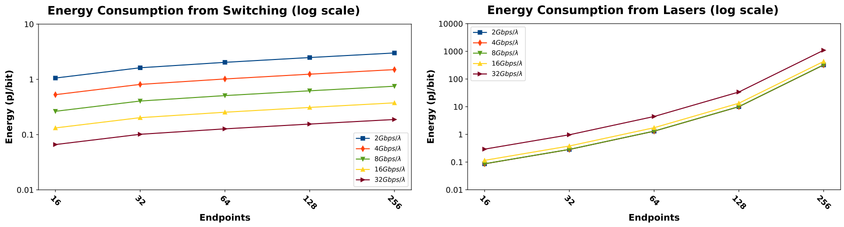


Figure 4: Energy consumption from lasers and switching.

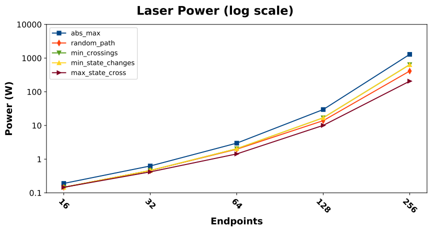


Figure 5: Laser Power.

dedicated to the interconnect. Another advantage is that this technology offers a rearrangeably non-blocking network as opposed to a 2D electrical mesh, which is prone to contention.

Based on the above a 32-endpoint optical MZI-based network can be considered for the on-chip domain. Using more efficient lasers could reduce the power requirements of the larger network sizes into this domain as well. The use of on-chip lasers and adaptive laser control techniques [6] can also provide substantial laser power reductions. Lastly, considering the large bandwith per link (256Gbps), core aggregation at the endpoints can also be viable.

Fig. 5 also shows the routing strategies' impact with respect to laser power; all strategies exhibit substantial savings at every size with respect to the baseline, which is exhibited by the max. ILoss calculated from the original device. *Min_crossings*, *min_state_changes* and *random_path* exhibit similar savings, ranging from 23% for 16 endpoints to 50% for 256 endpoints. *Max_state_cross* performs even better, with savings ranging from 33% to 85% across the network sizes.

6 CONCLUSIONS & FUTURE WORK

In this work, we have evaluated the limitations of scaling out a thermally/electrically tuned MZI-based optical Beneš network. We have presented three hardware-inspired routing strategies which aim to leverage the asymmetric behaviours of internal switching elements. We show that these strategies always reduce the max. ILoss. Furthermore, we show that maximising the number of MZIs in "cross" state can reduce max. ILoss by 32% in the best case (Bisection, 64 endpoints). Through our laser power analysis, we show that a network of 32 endpoints is suitable for the on-chip domain, and show substantial laser power reduction with the best routing strategy, ranging from 33% to 85% depending on the number of endpoints. In the future, we plan to investigate combining the routing strategies to further reduce max. ILoss and energy consumption, as well as to explore nested network topologies using variable sizes of this model.

ACKNOWLEDGMENTS

This work was funded by the European Union's Horizon 2020 research and innovation programme, grant agreement No 754337.

REFERENCES

[1] Arghavan Asad, Aniseh Dorostkar, and Farah Mohammadi. 2018. A novel power model for future heterogeneous 3d chip-multiprocessors in the dark silicon age. *EURASIP Journal on Embedded Systems* 2018, 1 (2018).

[2] Janibul Bashir, Eldhose Peter, and Smruti R Sarangi. 2019. A survey of on-chip optical interconnects. *ACM CSUR* 51, 6 (2019).

[3] Nicola Calabretta et al. 2013. On the performance of a large-scale optical packet switch under realistic data center traffic. *IEEE/OSA Journal of Optical Communications and Networking* 5, 6 (2013).

[4] Qixiang Cheng et al. 2018. Recent advances in optical technologies for data centers: a review. *Optica* 5, 11 (2018).

[5] William James Dally and Brian Patrick Towles. 2004. *Principles and practices of interconnection networks*. Elsevier.

[6] Yigit Demir and Nikos Hardavellas. 2014. Ecolaser: an adaptive laser control for energy-efficient on-chip photonic interconnects. In *2014 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE.

[7] Odile Liboiron-Ladouceur et al. 2008. The data vortex optical packet switched interconnection network. *J. Lightw. Technol.* 26, 13 (2008).

[8] Liangjun Lu et al. 2016. 16× 16 non-blocking silicon optical switch based on electro-optic Mach-Zehnder interferometers. *Optics express* 24, 9 (2016).

[9] Zeqin Lu, Dritan Celo, Hamid Mehrvar, Eric Bernier, and Lukas Chrostowski. 2017. High-performance silicon photonic tri-state switch based on balanced nested Mach-Zehnder interferometer. *Scientific reports* 7, 1 (2017).

[10] Ronald Luijten and Richard Grzybowski. 2009. The OSMOSIS optical packet switch for supercomputers. In *Optical Fiber Communication Conference*. Optical Society of America.

[11] Rekha Mehra and Jitender Tripathi. 2010. Machzehnder Interferometer and it's Applications. *International Journal of Computer Applications* 1, 9 (2010).

[12] Cyriel Minkenberg et al. 2018. Reimagining datacenter topologies with integrated silicon photonics. *Journal of Optical Communications and Networking* 10, 7 (2018).

[13] 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).

[14] Sébastien Rumley et al. 2017. Optical interconnects for extreme scale computing systems. *Parallel Comput.* 64 (2017).

[15] Richard Soref. 2006. The past, present, and future of silicon photonics. *IEEE Journal of selected topics in quantum electronics* 12, 6 (2006).

[16] Chen Sun et al. 2012. DSENT-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling. In *2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip*. IEEE.

[17] Wei Tan et al. 2017. Venus: A low-latency, low-loss 3-D hybrid network-on-chip for kilocore systems. *Journal of Lightwave Technology* 35, 24 (2017).

[18] Marc A Taubenblatt. 2011. Optical interconnects for high-performance computing. *Journal of Lightwave Technology* 30, 4 (2011).

[19] David Thomson et al. 2016. Roadmap on silicon photonics. *Journal of Optics* 18, 7 (2016).

[20] Christos A Thraskias et al. 2018. Survey of photonic and plasmonic interconnect technologies for intra-datacenter and high-performance computing communications. *IEEE Communications Surveys & Tutorials* 20, 4 (2018).

[21] Dana Vantrease et al. 2008. Corona: System implications of emerging nanophotonic technology. In *ACM SIGARCH*, Vol. 36. IEEE Computer Society.

[22] Ke Wen et al. 2016. Flexfly: Enabling a reconfigurable dragonfly through silicon photonics. In *SC'16*. IEEE.

[23] Sebastian Werner, Javier Navaridas, and Mikel Luján. 2017. Efficient sharing of optical resources in low-power optical networks-on-chip. *IEEE/OSA J. Opt. Commun. Netw.* 9, 5 (2017).

[24] Sebastian Werner, Javier Navaridas, and Mikel Luján. 2018. A survey on optical network-on-chip architectures. *ACM CSUR* 50, 6 (2018).

[25] Ming C Wu and Tae Joon Seok. 2018. Large-scale silicon photonic switches. In *2018 ACP*. Ieee.

[26] Piu-Hung Yuen and Lian-Kuan Chen. 2013. Optimization of microring-based interconnection by leveraging the asymmetric behaviors of switching elements. *J. Lightw. Technol.* 31, 10 (2013).

Chapter 5

Paper 2: On the Routing and Scalability of MZI-based Optical Beneš Interconnects

On the Routing and Scalability of MZI-based Optical Beneš Interconnects

Markos Kynigos[a,], Jose A. Pascual[b], Javier Navaridas[a], Mikel Luján[a], John Goodacre[a]

[a]*School of Computer Science, The University of Manchester, Oxford Road, M13 9PL Manchester, United Kingdom*
[b]*University of the Basque Country, San Sebastian, Spain*

Abstract

Silicon Photonic interconnects are a promising technology for scaling computing systems into the exa-scale domain. However, there exist significant challenges in terms of optical losses and complexity. In this work, we evaluate the applicability of thermally/electrically tuned Beneš network based on Mach-Zehnder Interferometers for on-chip and inter-chip interconnects as regards its scalability. We examine how insertion loss, laser power and switching energy consumption scale with the number of endpoints. In addition, we propose a set of hardware-inspired routing strategies that leverage the inherent asymmetry present in the switching components. We evaluate a range of network sizes, from 16 up to 256 endpoints, using 8 realistic and synthetic workloads and found very promising results. Our routing strategies offer a reduction in path-dependent insertion loss of up to 35% in the best case, as well as a laser power reduction of 31% for 32 endpoints. In addition, bit-switching energy is reduced by between 8% and 15% using the most efficient routing strategy, depending on the communication workload. We also show that workload execution time can be reduced with the best strategies by 5-25% in some workloads, while the worst-case increases are at most 3%. Using our routing strategies, we show that under the examined technology parameters, a 32-endpoint interconnect can be considered for the NoC domain in terms of insertion loss and laser power, even when using conservative parameters for the modulator.

1. Introduction

As high-performance computing (HPC) advances into the exa-scale domain, numerous challenges present themselves in terms of system scalability. HPC commonly supports massively parallel workloads, which in turn require very strict communication constraints between compute elements. It is widely acknowledged that interconnection networks constitute a scalability bottleneck for future HPC systems [1]. Furthermore, recent evidence suggests that conventional electrical interconnects will not be able to keep up with system scalability trends in terms of performance, while satisfying the ever-more stringent constraints in power consumption and area [2] [3].

Optical interconnects (OINs) based on Silicon Photonics (SiPh) have emerged as a promising candidate technology to augment, if not substitute, traditional interconnects at the NoC, interposer, inter-chip and top-of-rack level. The technology's compatibility with CMOS processes, its capacity for high-bandwidth data transmission through dense-wavelength-division multiplexing (DWDM) as well as relatively distance-independent energy consumption make it a promising solution for future systems [4]. However, the current state-of-the-art devices as specified in [5] suffer from intrinsic limitations that can lead to increases in optical losses (namely insertion loss and crosstalk), complexity and package cost.

In this paper, we investigate the scalability potential of an optical network based on a Beneš topology formed with thermally/electrically tuned Mach-Zehnder Interferometers [6]. Our aims are:

- To evaluate the network's performance under realistic workloads.
- To evaluate the insertion loss of the system at scale, identify the chief contributors and quantify their impact.
- To assess the network's energy consumption in terms of switching and laser power.
- To propose a set of hardware-inspired routing strategies which leverage the network's underlying asymmetrical behaviours to reduce insertion loss and energy consumption.
- To demonstrate the benefits of the strategies at different system scales and ascertain the network's applicability to the on-chip domain.

In addition to the above, we aim to identify the use cases in which this type of network would be applicable; that is, whether it is beneficial as the main switching fabric on a chip serving a large number of endpoints, or if it is more realistic, in terms of losses and energy consumption, to constrain the size for on-chip territory and consider a nested topology paradigm for inter-chip fabric. The way that insertion loss scales due to the chief contributors in

Email addresses: markos.kynigos@manchester.ac.uk (Markos Kynigos), joseantonio.pascual@ehu.es (Jose A. Pascual), javier.navaridas@manchester.ac.uk (Javier Navaridas)

conjunction with laser power yields valuable insight for the applicability context of this network.

In our evaluation, we quantify the amount of maximum insertion loss that is incurred from the networks' different components at different system scales. We also show that using our enhanced, variability-aware routing strategies can reduce this metric substantially, depending on the evaluated workload. The best-case reduction is exhibited under sweep2 (strategy *m_xb*, 64 endpoints, 35.5% savings). We also show that the most efficient data rate as regards to energy per bit is 8Gbps per wavelength and that increasing it prohibitively increases energy consumption. Furthermore, we demonstrate that our routing strategies can also decrease energy consumption from switching when compared to random path selection, by between 3-18%. Additionally, we show that with the best strategy execution time can be reduced by 5-25% in some workloads (e.g. bisection & nbodies) while remaining virtually unaffected in others (torus2, randomapp, hotregion) and that the worst-case increase is never over 3% relative to the baseline. Lastly, we show that by using the most impactful routing strategy, a 32-endpoint system in this technology may be suitable for the NoC domain in terms of laser power budget, as laser power can be decreased by 31% for 32 endpoints under random traffic.

2. Background

SiPh has been acknowledged by the interconnects community as a key enabler for scaling interconnect systems [7, 8]. Here, we discuss the most prominent advances in the domain starting from the device substrate. We then continue by outlining examples of SiPh interconnects at the NoC, inter-chip and datacenter level. Afterwards, we discuss other scalability analyses of OINs and how they relate to this work. The section concludes by introducing the optical network we focus on in this work, where we also discuss the use of Mach-Zehnder Interferometers (MZIs) as a building block for multi-stage networks.

2.1. SiPh Technology

The underlying components that make SiPh interconnects possible (e.g. waveguides, microring resonators, MZIs, multi-mode interferometers, transceivers, lasers etc.) are the subject of wide research with novel components being proposed very frequently [9]. This section discusses advances in devices which affect the metrics we examine in this work. It is to be noted that although modulator and detector arrays are indispensable to OINs, their study is out of scope for this work. For our experimental analysis we assume the modulator/detector components detailed in [10].

Lasers are the most commonly considered light source for OINs, although research is being conducted on the use of other light sources, e.g. LEDs. For example, Bie et al. report on a promising $\mathrm{MoTe_2}$ based LED/photodetector devices [11], which could allow for direct on-chip waveguide-division multiplexing. They also claim that the technology can be used to fabricate narrowband lasers with very high coupling efficiencies.

Current laser research generally falls within two categories; either the use of on-chip, or off-chip lasers. Off-chip lasers are generally more efficient in terms of wall-plug efficiency, but must be powered on for the full operation of communication irrespective of the sending endpoint. For an in-depth discussion on laser source efficiency for ONoCs, we refer the reader to [12]. Prominent examples of off-chip lasers can be found here [13]. On-chip lasers are in principle very desirable for on-chip optical communication, as they could be powered on-demand and used for direct modulation onto the optical carrier. However, due to silicon's indirect band gap, the material doesn't lend itself well to lasing. As such, the complementary use of more exotic materials is required [14] [15] [16], which in turn creates a range of problems, most importantly with regards to laser efficiency and wafer yield [17] [18]. We note that on-chip lasers are a very active research topic; prominent novel example technologies which show promise are VCSEL lasers and Transistor Lasers; discussions on these can be found in [9] and [19] respectively. Nevertheless, most proposed systems to date, including the interconnect we examine here, consider off-chip laser sources. We later discuss how laser power is affected by the size of the network and identify it as the primary scalability constraint.

Waveguides are the optical equivalent of the electrical wire and enable the most prominent paradigm shift in interconnects, namely the ability to send multiple information streams in parallel over the same physical link in an energy efficient and relatively distance-independent fashion. Waveguide technology for OINs is widely studied and a wide variety of demonstrations have been put forth recently. For instance, in this work we consider a conservative waveguide propagation insertion loss penalty of 1.18 dB/cm; Thraskias et al. on the other hand mention waveguide-incurred insertion loss of as low as 0.2 dB/cm [9]. We note that propagation loss due to waveguides is highly dependent on device technology; nevertheless, the aforementioned survey illustrates the rate of progress on the technology front. As we show in our experimental work, other factors that contribute to the insertion loss penalty (i.e. number of waveguide crossings, MZI states) have a much greater impact on scalability than the propagation loss due to waveguide length.

Switches are another fundamental building block of OINs, with various types being explored throughout the literature. A comprehensive review of SiPh switches can be found in [5], and on MEMS switches for more general optical communications here [20].

Most commonly, optical switching elements are based on microring resonators (MRR), whereby organisations of MRRs are coupled to waveguides to form N×N switching elements. Most MRR implementations are wavelength selective [21], although multi-wavelength MRRs have been reported (e.g. [22] [23]). Due to this aspect, they are usually used for wavelength-routed OINs, with cascaded arrays of MRRs being used for multi-wavelength switching. However, this approach entails a number of drawbacks, such as increased area and MRR tuning-induced power consumption, both of which are problematic with respect to scalability in constrained interconnects such as ONoCs [12]. Additionally, MRR-based switching elements have limited bandwidth, since only a small subset of wavelengths is used between each pair of endpoints. Other drawbacks that constrain scalability include increasing wiring complexity in the device electrical control circuitry as well as increased packaging costs.

More recently, Mach-Zehnder Interferometers (MZIs) have been considered for optical switching. MZI switches trade off wavelength selectivity for reduced tuning power and wiring complexity compared to MRR switches [24] by switching all incoming wavelengths simultaneously. Standalone MZIs can be used as 2×2 switching elements which are then organised in a multi-stage fabric such as the one we examine here. MZI switches normally have two states; "cross" and "bar". More novel MZI designs include a third, blocking state, which is aimed at reducing optical crosstalk between the stages [25]. Another approach that has been put forth is to use nested MZI organisations which can compose higher-radix base switching elements [26].

MZIs are commonly switched by means of either thermal or electrical tuning. Thermo-optical (TO) tuning has a relatively slow response (μs scale [27]) and is generally superseded by electro-optical (EO) tuning which is much faster; for example, [28] report an EO switching time in the order of a few ns. However, due to effects such as free carrier absorption, EO tuning entails a reduced tuning range compared to TO [29]. As the on-chip domain requires ns switching time which TO tuning cannot accommodate, we adopt a combination of EO and TO tuning as proposed in the model we base our analysis on [6]. In this model, TO tuning is used to compensate for fabrication defects and reach MZIs to the "cross" state, while EO tuning is applied to switch the state to "bar".

2.2. SiPh Interconnects

Many proposals for Interconnects enabled by SiPh have emerged in the past decade, targeting application domains from the NoC level to the intra-datacenter level. We will discuss some the most prominent examples of these here. The discussed proposals are summarised in Table 1, which depicts the deployment scenario, number of photonic endpoints of deployed prototypes or proposals, switching-interconnect technology, interconnect type, photonic topology and routing scheme in each work. We note that rather than perform an exhaustive survey of the state of the art, we include a selection of works that illustrates the large variability of technology and deployment scenarios proposed, thereby underpinning the need for the design-space exploration for the MZI-based technology we examine here. For exhaustive surveys, we refer the reader to those quoted in the discussion below.

With recent advances allowing photonic integrated circuits (PICs) using CMOS-compatible processes, a lot of interest has been generated for Optical Networks-on-Chip (ONoCs). A comprehensive study of these can be found here [12]. Notable ONoC examples are ATAC [33], Corona [32], METEOR [30], QuT [31], Amon [2] and more recently Venus [34]. These proposals commonly employ MRR-based interconnects using routing techniques such as DWDM and wavelength routing.

With the advent of interposer technology, various examples with optical interconnects targeting the chiplet level have been put forth [49] [36] [37]. In [37] specifically, a 2.5D design with processor disintegration, HBM memories and an AWGR-based ONoC fabric is investigated. Resource disintegration enabled by SiPh interconnects at the PCB level is another interesting approach [50], whereby spatially far resources are brought logically close enough to form a virtual chip, e.g. in Oracle Macrochip [39] or Galaxy [40].

The community has also proposed many SiPh-enabled architectures at the system level such as Data Vortex [43], Osmosis [51], Flexfly [45] or the Hipoλaos-enabled system described in [41]. As with the model we examine here, Flexfly employs 2x2 MZIs composed in a Beneš multistage fabric. SiPh-enabled systems have also emerged for use in data-centre networks [47][48][42].

2.3. Beneš-based SiPh Interconnects

In this work, our interest is to examine the scalability of Beneš networks [52] for their use with EO/TO Mach-Zender Interferometers (MZIs) [53] and to identify feasible deployment use cases. Our analysis is based on [6], where a 16×16 Beneš network is constructed out of seven stages of 2×2 MZIs. They contribute an experimental demonstrator and extract a full characterisation of the underlying components. They describe the underlying process used to design and fabricate the basic components, as well as several optimisation processes undertaken to reduce the insertion loss exhibited by the components (e.g. optimised tapers for waveguide crossings, optimised MZI design etc.). They report the insertion loss of the waveguide crossings and MZIs in each state, which we adopt for our analysis. In addition, the thermal and electrical tuning power is reported for each individual MZI which we used as a basis for our analysis.

A diagram of the 16×16 Beneš topology can be seen in Fig. 1; the rectangular components represent 2×2 MZI switches. Note that we consider binary MZIs where an MZI is either at a"cross" or a "bar" state, rather than the tri-state MZIs discussed previously [25]. Composing an optical network using tri-state MZIs could afford interesting possibilities for future designs; idle elements within a $N \times N$ fabric can be tuned to the "blocking" state to dramatically reduce optical crosstalk.

Table 1: State-of-the-art photonic network proposals.

Name	Deployment	# Photonic Endpoints	Opt. Tech.	Type	Phot. Topology	Phot. Routing
[2] Amon	On-chip	$\leq$64	MRRs	All-Optical	mesh-like	λ routing
[30] METEOR	On-chip	4	MRRs	All-Optical	ring	Selective XY routing
[31] QuT	On-chip	$\leq$ 128	MRRs	All-Optical	ring-like	λ routing
[32] Corona	On-chip	$\leq$64	Phot. Crossbar	All-Optical	Crossbar + ring	token-based
[33] ATAC	On-chip	$\leq$64	WDM bus	Hybrid	ring	WDM Broadcast
[34] Venus	On-chip	$\leq$256	MRRs & SWMR crossbar	Hybrid	hierachical	hierarchical
[35] Lego	On-chip	64	WDM busses	Hybrid	generalised hypercube	XY-over elec.
[36] N/A	Interposer	$\leq$32	AWGRs	All-Optical	all-to-all	λ routing
[37] N/A	Interposer	16	AWGRs	All-Optical	all-to-all	λ routing
[38] Baldur	Interposer-PCB	256	Trasistor Lasers	All-Optical	multi-butterfly	optical packet routing
[39] Macrochip	Dissagg. PCB	64	Various	Hybrid	crossbar	static point-to-point
[40] Galaxy	Dissagg. PCB	N/A	Phot. Crossbar	Hybrid	modified Firefly	token-based
[41] Hipoλaos	Dissagg. System	16	AWGRs	Hybrid	all-to-all	λ routing
[42] dReDBox	Dissagg. System	N/A	VCSEL Array	Hybrid	full mesh	N/A
[43] Data Vortex	System	N/A	Broad. SOA	Hybrid	banyan-based	deflection-routing
[44] Osmosis	Switch	64	SOA	All-Optical	broadcast-and-select	cell-switching
[45] Flexfly	DCN System	4	MZI	All-Optical	Beneš	link stealin
[46] DOS	DCN Switch	32	AWGR	All-Optical	all-to-all	λ routing
[47] Topanga	DCN Switch	12	Opto-ASIC	Hybrid	generalised hypercube	FRP
[48] N/A	DCN Switch	$\leq$64	SOA-AWG	Hybrid	Spanke	distributed
This Work	TBC	16-256	MZI	All-Optical	Beneš	Routing Strats.

The Beneš network is a Clos network variant constructed from 2×2 switching elements (SE). It requires the minimum number of SEs to connect 2^i ports in a rearrangeably non-blocking fashion and enables all routing paths to traverse the network through an equal number of SEs [52]. As such, this paradigm lends itself well to the case of using MZIs as base switching components.

The fact that this topology requires the minimum number of 2×2 SEs amongst other, state-of-the-art topologies is significant for two reasons. The PILOSS [54] and DLN [55] topologies incur a switch count which scales quadratically with the number of endpoints. Considering that MZIs are relatively large in terms of area, these two topologies would become prohibitively large as the number of endpoints is scaled out. The second reason is that these two topologies require a much larger amount of waveguide crossings than the Beneš network [29]. As we will discuss in the evaluation section, waveguide crossings are critical to scalability in terms of insertion loss and can severely restrict scaling.

Additionally, photonic buffering is impractical for meaningful amounts of time in the context of switching and buffering [56]. This means that in proposals that employ packet switching, which inherently requires buffering, packets must be converted from the optical to the electrical domain in intermediate switches for storage, and back to the optical domain for transmission in the next hop (unless they are dropped, which causes high retransmission penalties). These conversions generate large energy and latency overheads which are undesirable in terms of scalability [12]. However, these overheads can be ameliorated by using circuit switching techniques in a buffer-less network [3], as we explore in this work.

2.4. Related Scalability Analyses

As OINs have undergone significant research in the interconnect community, various studies have been conducted on their scalability. In [57], a comparative study is conducted between the Beneš network and a full mesh under different scales, for their use with cascaded MRR switches. Bianco et al. also study the scalability of OINs based on MRRs, whereby they propose a set of 2×2 switches to reduce crosstalk and insertion loss [58]. The latter study also mentions the scalability limitations as regards to area due to the large number of required MRRs. Neither of these studies, however, evaluates the proposed interconnects under different traffic scenarios ([57] uses random traffic and [58] has no mention of this).

In [29], the impact of topology choice on the scalability of MZI-based OINs with respect to crosstalk is discussed. The study concludes that the DLN topology is most favourable; however, they do not evaluate using different traffic scenarios either. In contrast, this work focuses on the routing aspects and how, by making informed pathing choices, insertion loss and therefore required laser power can be consistently reduced and, indeed, we consider a wide range of traffic scenarios to show the flexibility of our approach.

3. SiPh Beneš Switch at Scale

3.1. Scalability Challenges and Experiment Motivation

As discussed, the MZIs we consider in this paper are *thermally* tuned to reach a “cross” state and *electrically* tuned to reach a “bar” state from that “cross” state. As such, more power is required for an MZI to hold the “bar” state. Furthermore, an MZI in the “bar” state exhibits substantially more Insertion Loss (hereafter, ILoss) than an MZI in a “cross” state. Thus, using MZIs in the “bar”

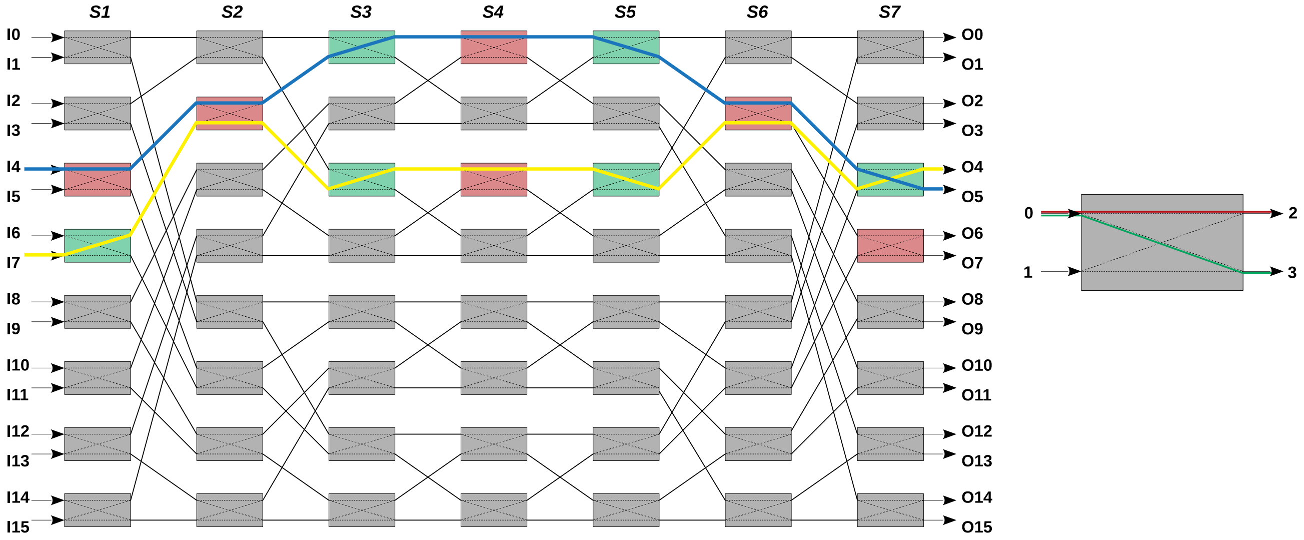


Figure 1: Left: Topology diagram of the 16x16 Beneš network. The blue flow traverses from I4 to O5 and the yellow from I7 to O6, green shows MZIs in "cross" state and red in "bar" state. Right: An MZI with port numbers.

state generates significant overheads and is therefore considered unfavourable, albeit necessary in order to produce the full path diversity of the Beneš network. In addition, an ILoss penalty is incurred for each waveguide crossing and each connection between MZIs entails a different number of crossings. Considering these factors, ILoss is path dependent in the Beneš network. As we will see in the next section, allocating a path for one flow constrains the available paths for other flows; allocating an ILoss-optimal path for one flow can lead to other flows being allocated less ILoss-optimal paths.

In fact, depending on the type of photodetector used in the receiver, ILoss is a defining factor for the power of the laser beam which carries the flow [10]. Excessive ILoss means a more powerful laser is required, which affects the applicability context of the technology in terms of chip power budget. As such, it is important to consider ways of reducing these metrics to achieve scale both for on-chip and inter-chip domains.

Another key constraint to scalability is interconnect area, irrespective of deployment scenario. This is especially the case with MZIs which, compared to MRR switching elements, occupy a substantially larger area. Indicatively, Lee and Dupuis report MZIs occupying between 0.04-0.1 mm^2 [59]. In Table 2, we report how the number of MZIs and waveguide crossings scale w.r.t. the number of endpoints in a Beneš network. The total number of crossings in a logical Beneš scheme scales according to the following function,

$$\sum_{b=1}^{log_2 \frac{r}{2}} \frac{r}{2^{b+1}} \cdot 2^b \cdot (2^b - 1) \tag{1}$$

while as in [59] the number of MZIs scales according to

$$(r)log_2(r) - r/2 \tag{2}$$

where r is the number of endpoints in the network.

As Lu et al. [6] do not report the area footprint of individual photonic components in their demonstrator, it is not possible to provide accurate area estimates for larger interconnects based on their technology. Furthermore, a comprehensive analysis on photonic layer area requires a physical level design analysis which is a research question of its own and out of scope for this paper.

Table 2: Component scalability and area estimation

# Endpoints	# Crossings	# MZIs
16×	88	56
32×	416	144
64×	1824	352
128×	7680	832
256×	31616	1920
512×	128512	4352
1024×	518656	9728

When implementing planar Beneš-based interconnects on a real layout, ILoss may also be affected adversely due to the increase of waveguide length and the addition of waveguide crossings mandated by the place-and-route process. Topology optimization aiming to mitigate place-and-route induced ILoss is an interesting problem; in [60] Wang et al. propose a floor-plan optimization process for delta networks which can dramatically reduce ILoss even at high radices. However, on the one hand they consider MRR-based switching elements rather than MZIs and on the other, they focus on delta networks and optimization at the floor-plan level. In contrast, this work focuses on how to further decrease path-dependent Iloss and switching energy consumption through routing once the network has been implemented. Another implementation consideration is the type of laser source considered, as well as the optical in/out power coupling to the interconnect. However, as in this work we perform a design space exploration which focuses on reducing path-dependent Iloss and energy through routing, which has no effect on these considerations, we consider them to be out of scope for our evaluation, and assume the laser source properties detailed in [10].

These effects, combined with the need to evaluate the system under realistic workloads, outline our experimental motivation. We aim to understand the scalability implications relating to path-dependent ILoss and energy consumption, as well as to evaluate a set of hardware-inspired routing strategies that minimise these key metrics, which we outline below.

3.2. Routing in a SiPh Beneš network

To correctly utilise the model, each element within the MZI array must be electrically/thermally tuned in order to facilitate route allocation and choice.

Each time a flow is to be injected from a source endpoint, the control process calculates the possible routes the flow may take through the network. For N endpoints, each flow can use a maximum of $N/2$ different paths. The convention we use is that flows traverse the switch from "left" to "right"; flows are always inputted in ports 0 or 1 and outputted at ports 2 or 3. The route calculation process generates potential paths by varying the interconnected MZIs to be traversed per stage in the left half of the potential path to produce path diversity. The right half of the path is kept stable to ensure the destination endpoint is correctly addressed.

Once all options have been calculated, the best path is selected based on the hardware-inspired routing strategies outlined in the next subsection. After selection, the control process iterates through the potential flow paths and, for each encountered MZI, assesses its ability to preserve or switch to the required state. Note that for an MZI in a "bar" state where a previous flow has reserved ports 0 and 2 for example, ports 1 and 3 may be used by another flow. The corresponding scenario applies for the "cross" state as well. If the path assessment completes successfully, the path is reserved by tuning the corresponding MZIs if needed. Otherwise, the process continues for the remainder of the potential paths.

Path selection and flow scheduling in this type of network is a non-trivial problem. MZI states, which compose the network state, directly affect the network's ability to fully take advantage of the topology's rearrangeably non-blocking nature and allocate the best path. To illustrate this point, consider the scenario depicted in Fig. 1, where a flow is scheduled from I7 to O6, shown in yellow. However, the network is *already* serving a flow from I4 to O5, shown in blue. The control process has already assigned MZI states to serve the blue flow; in this scenario, the two sender endpoints can share the MZI in column S2 but, since it already has a state assigned, it constrains the available paths for the yellow flow.

There are two main approaches for designing controllers for this network. The first option is to design a centralised controller, which controls MZI states and receives path requests from the endpoints. This controller has full knowledge of the network state and can therefore allocate paths to flows based on state-aware decisions; we examine some of these in the next section.

The second option is to design a distributed control mechanism; *every* MZI is controlled by a separate controller, which is connected to its neighbouring MZIs. This option enables cascaded path selection, whereby senders request a path from their neighbouring MZI, the request gets forwarded to the MZI's neighbour and so on. In this case, an MZI which cannot change state and serve a flow would send a failure message to the previous stage in the network, enabling it to "fail back". However, various challenges arise with controller design. Controller complexity, flow scheduling and latency are aspects which affect both the centralised and distributed design types. For the distributed design especially, how the back-propagation of failure and success messages affect latency is a substantial question, as is whether the design's greedy nature would be able to reach near-optimal paths. As such, controller design is a research question in itself, which is out of scope for this work.

4. Reducing Insertion Loss & Energy Consumption with Routing

As discussed, max. ILoss varies significantly with respect to the path that is allocated to a communication flow. These variations depend on both the state of the MZIs that compose the network and the amount of waveguide crossings that a flow is subjected to. Also, the MZI-based SEs we examine use both thermal and electrical tuning to reach the cross and bar states respectively. Each of these tuning mechanisms incurs a different amount of power consumption with electrical tuning requiring much less power, making the total power and therefore energy consumption inherently state dependent. Minimising the amount of traversed waveguide crossings or the amount of MZIs in the "bar" state can therefore obtain significant ILoss savings, thereby reducing required laser power. In addition, promoting MZI reuse can, in theory, reduce energy consumption; prioritising this can be an appropriate strategy for allocating paths to flows.

Yuen and Chen [61] present an interesting approach to reducing ILoss and power consumption for MRR-based switches, whereby they propose a heuristic for leveraging

asymmetric behaviours in the underlying switching components. They demonstrate significant improvements with respect to the baseline case, with more than 30% reduction in ILoss and power savings. This approach, however, does not take into account non-uniformity in the paths with respect to the number of waveguide crossings and, indeed, is limited to be used with MRR-based SEs.
Another approach is outlined by Cheng et al. [62], who propose a path mapping strategy for 8× MZI-based Beneš fabrics by exhaustively evaluating all potential switch fabric states for a permutation; however this quickly becomes intractable as the switch scales out. In contrast, we propose and evaluate routing strategies at the path level rather than the permutation level.

Having these factors in mind and building upon our previous work [63], we propose the following single-criterion hardware-inspired routing strategies:

- **Minimise Crossings** (**m_x**) prioritises the paths with the least amount of waveguide crossings to reduce ILoss.

- **Minimise State Changes** (**m_c**) prioritises the paths with the fewer MZI state changes to overlap flows and reduce energy consumption through MZI reuse.

- **Minimise "Bar" States** (**m_b**) prioritises the path with the least MZIs in "bar" state aiming to reduce both ILoss and energy consumption.

In addition to these, we also propose the following hybrid routing strategies, whose aim is to combine the benefits of the routing choice criteria outlined above. This is done by consecutively applying them to path selection:

- **m_xb** applies the *m_x* strategy followed by *m_b*.

- **m_bx** applies the *m_b* strategy followed by *m_x*.

- **m_cb** first applies the *m_c* strategy followed by *m_b*.

- **m_bc** first applies the *m_b* strategy followed by *m_c*.

- **m_xc** first applies the *m_x* strategy followed by *m_c*.

- **m_cx** first applies the *m_c* strategy followed by *m_x*.

For completeness we also consider the following baseline for power and energy evaluation as a representative routing algorithm.

- **Random Path** (**rnd**) selects a path randomly, without taking underlying hardware asymmetries into account.

To illustrate the operation of the routing strategies, Table 3 shows the number of MZIs encountered by flows in the "cross" and "bar" state, as well as the number of encountered crossings, for each of the routing strategies in a 32× Beneš example, when using a bisection workload. We include metrics for both the average of all the flows generated by the workload and for the ILoss-critical path encountered. We also include the path with the worst ILoss, i.e. the most "bar" states and most crossings, for comparison.

Table 3: Number of "cross"/"bar" MZIs and waveguide crossings encountered by flows with each routing strategy in a 32× Beneš, when exposed to a Bisection workload. Includes metrics exhibited on average and by the critical path for each strategy. Abs. Max. denotes the worst-case path in the whole configuration.

32× Beneš	Average			Critical Path (*IlMax*)		
Strategy	"Cross" MZIs	"Bar" MZIs	WGX	"Cross" MZIs	"Bar" MZIs	WGX
m_x	4.1875	4.8125	23.875	1	8	31
m_c	4.1875	4.8125	26.375	2	7	34
m_b	5.5625	3.4375	25.625	4	5	43
m_xb	5.6875	3.3125	24.75	3	6	22
m_bx	4.3125	4.6875	23.813	1	8	31
m_cb	5.875	3.125	26.125	4	5	41
m_bc	5.8125	3.1875	25.875	4	5	39
m_xc	4.125	4.875	26.125	1	8	34
m_cx	4.1875	4.8125	23.875	1	8	31
rnd	3.9375	5.0625	26.875	1	8	34
Abs. Max	-	-	-	0	9	52

For instance, the single-criterion **m_x** and hybrid **m_bx** and **m_cx** variants, show a consistent reduction in encountered waveguide crossings both on average and in the critical path case, when compared to both the "Abs. Max" path and "rnd". This, however, comes at the expense of more "bar" states, especially in the case of critical path. As we will see in our evaluation section, this effect increases ILoss but, as the size of the network increases, the impact of waveguide crossings also increases substantially.

On the other hand, **m_b**, **m_xb** and **m_cb** accomplish their aim to consistently reduce the number of "bar" states both on average and in the critical path, at the expense of using paths with more waveguide crossings. Indeed, the **m_cb** slightly out-performs the other two for this configuration; as we will see, this is not the case for all the workloads we examine.

In our evaluation section, we will examine whether the order of application of each criterium has an impact on the figures of merit that we assess under various workloads. For instance, it is expected that the *m_xb* and *m_bx* hybrids will present a different behaviour to each other under the different workloads we examine, due to the fact that choosing the least amount of crossings and choosing the least MZIs in the "bar" state are independent criteria. Conversely, we expect the *m_cb* and *m_bc* strategies to present very similar if not identical behaviour to each other, as both the component criteria leverage the state of the MZIs. As such, we expect the order of hybridisation to have little to no impact for the latter two strategies. Nevertheless, it is worth experimentally examining whether this supposition holds true and we therefore elaborate on this in the evaluation section.

5. Experimental Setup

As discussed, the primary focus of our experimental work is to evaluate the proposed routing strategies as well as the scalability of the Beneš network and its applicability to the on-chip domain. To make a more realistic evaluation, we consider various realistic traffic models.

5.1. Simulator and Model

We use phINRFlow (photonic Interconnection Network for Research Flow-level Simulation Framework), an inhouse 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 photonic interconnects. 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 candidate for exa-scale systems. The simulator inherits functionality from INRFlow [64] wherein a detailed description of the simulator's methodology, organisation and workloads may be found.

The system is modelled within the simulator as a new topology. All the links are uni-directional with traffic flowing from "left" to "right". For the purposes of this analysis, we assume that each endpoint is supplied with a laser source and can communicate with all other nodes independently.

In their description of the experimental demonstrator, Lu et al. [6] report the power consumption of both thermal and electric tuners located within the MZIs. In our experimental process, we use the reported tuner power consumption metrics by fitting them to a normal distribution, from which we then assign values to each MZI for the larger network sizes. We also adopt their usage of 32 wavelengths for all our experiments, which we explain further in section 6.2.

In order to evaluate laser power and energy consumption we use DSENT [10]. We explore the impact of varying the data rate per wavelength (denoted as λ), as well as that of the routing strategies. We then use the best data rate per λ in our switching energy consumption evaluation.

Lastly, in terms of MZI state control, we consider a centralised control process which allocates paths to flows in a first-come-first-served fashion. Based on the network state at each path request, the controller makes an informed decision with respect to the routing strategy that is being used.

5.2. Workloads

We use the following workloads supported by phINRFlow. Note that all 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.
- **Bisection** Nodes are split into pairs at random and nodes in a pair communicate with each other.
- **Mesh** A 2D stencil pattern commonly exhibited by scientific codes.
- **Hotregion** Generates a non-uniform load, with 25% of the traffic being directed to the upper 12.5% of the network. The rest is allocated a destination randomly.
- **AllReduce** An optimised, binary implementation of the AllReduce collective [65].
- **Nbodies** Tasks are arranged in a virtual ring in which each task starts a chain of messages that travel clockwise across half of the ring.
- **Torus2** A toroidal 2D stencil pattern commonly exhibited by scientific codes.
- **Sweep2** Performs a wavefront communication in a grid of tasks, which is traversed diagonally starting from the upper left corner.

Under each workload, every endpoint injects a number of flows into the network per configuration, depending on the size of the network and the workload properties. For the sake of clarity, Table 4 summarises the configuration setup parametrised to the number of endpoints (N) where required. 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. For a more detailed description of the workloads, see [64].

Table 4: Number of flows per workload.

Workload	**# Flows**
Randomapp	10000
Bisection	N
Mesh2	$\leq 4N$
Hotregion	10000
AllReduce	$Nlog_2N$
Nbodies	$N^2/2$
Torus2	$4N$
Sweep2	$\leq 2N$

5.3. Metrics of Interest

As previously discussed, the experimental demonstrator we base this work on gives a thorough report on ILoss for the proposed wavelength region. Based on this, the ILoss per component and power consumption we consider for our experiments are found in Table 5. Our study assesses the following metrics using both simulators:

- **Max ILoss** is the worst-case, path-dependent ILoss exhibited by any flow during the simulation of a workload. It is composed by the state of each MZI traversed, the number of crossings and the length of the path taken by the flow. Note that for ILoss induced by waveguide length, we assume 0.4386 dB per Beneš stage as described in [6].

- **Power per Laser** We measure the power-per-laser requirements for different network sizes.
- **Energy per Bit** As the simulator is aware of elapsed time for each event, the total Energy per Bit is derived based on elapsed time and power metrics, for both lasers and switching.

In previous work [63], we have also shown the average ILoss encountered by flows in the network. However, in terms of assessing ILoss to derive laser power, max. ILoss is the more pertinent figure of merit, as this is the ILoss that the laser will have to compensate for in the worst case. Consequently, this will be the figure that will determine the laser requirements.

Table 5: Insertion Loss and Power Consumption, as reported in [6].

Component	**Insertion Loss**
Waveguide	1.18 dB/cm
Beneš Stage	0.4386 dB
Waveguide Crossing	0.05 dB
"Cross" MZI	0.4 dB
"Bar" MZI	1.4 dB
Tuning Type	**Power Cons.**
Thermal	0-26 mW
Mean, STD	15.725, 6.608
Electrical	3.28-5.88 mW
Mean, STD	5.166, 0.428

5.4. Experimental Methodology

The mesh2, sweep2 and torus2 workloads we use to evaluate the model are deterministic or use pre-obtained flow traces. The rest use randomness to determine sources and destinations, as described in section 5.2.

Consequently, in order to ensure a robust evaluation for the workloads with inherent randomness, we run the workloads for each size of the network and for each routing strategy 100 times with varying random seeds. We then derive the mean and standard deviation for each batch of 100 runs in order to report our findings. The workloads not affected by randomness are run once per size per strategy with the same, randomly selected random seed for comparison purposes.

6. Results and Discussion

6.1. Insertion Loss

As has been frequently reported in the literature, ILoss can exacerbate the power consumption problem of OINs, as it leads to needing more powerful laser sources. Therefore, minimising this metric is key to achieving larger scale systems in the presence of stringent power constraints, as is the case with NoCs. The three main factors contributing to ILoss in our model are waveguide crossings, waveguide length and MZI-incurred loss for each state. The ILoss results we obtained have been split into three figures for readability and can be seen in Figs. 2, 3 and 4.

Each of the figures shows the max ILoss presented by each depicted routing strategy under the 8 workloads we described previously. The ILoss figures have been broken down into the contributing factors which compose the total figure of merit. For each workload, we have included the absolute maximum ILoss, which is exhibited by the worst case path that can be encountered. It comprises of the maximum waveguide length, the maximum number of crossings and the maximum number of MZIs in the "bar" state for each size point (see Table 5). Our experiments show that using the routing strategies always yields significant savings over this upper bound. Although the *rnd* routing strategy, which we use as a baseline, sometimes out-performs the other routing strategies in terms of Max. ILoss, it does so negligibly and is usually out-performed by at least one other strategy. For the three workloads whose flows are affected by randomness, we show the mean max. ILoss with error bars, as explained in section 5.4.

Firstly, it is clear that maximum ILoss scales proportionately to network size, irrespective of the routing strategy. However, in all cases, the maximum exhibited ILoss is less than the absolute maximum using any routing strategy. Therefore, the original report of 14 dB ILoss in the worst case was conservative, based on these results. This is encouraging, as it demonstrates the benefits of routing-based solutions for underlying hardware constraints. Interestingly, as the network size scales up, all the routing strategies exhibit a decreasing reduction in maximum ILoss; this is because, as we will see, the way that the number of crossings increases with network scale inhibits the strategies from impacting on the metric. However, there are some cases where this trend does not apply as we will discuss. The largest reductions over the absolute max. are exhibited under sweep2 (*m_x*, 256 endpoints, 33% and *m_xb*, 64 endpoints, 35.5%) and the least under the randomapp workload (strategy *rnd*, 32 endpoints, 7% savings).

6.1.1. Single Criterion Routing Strategies

Here, we analyse the performance of the single criterion routing strategies, namely *m_x*, *m_c* and *m_b*. We contextualise our numerical findings in the form of percentage savings of the max. ILoss exhibited by the routing strategies over the absolute max. ILoss. The three strategies are also compared against the baseline, to ascertain whether their benefits are enough to justify the additional path selection complexity.

The *m_x* strategy contributes substantially towards minimising the max. ILoss for most workloads. The ILoss savings that this strategy achieves over the abs. max

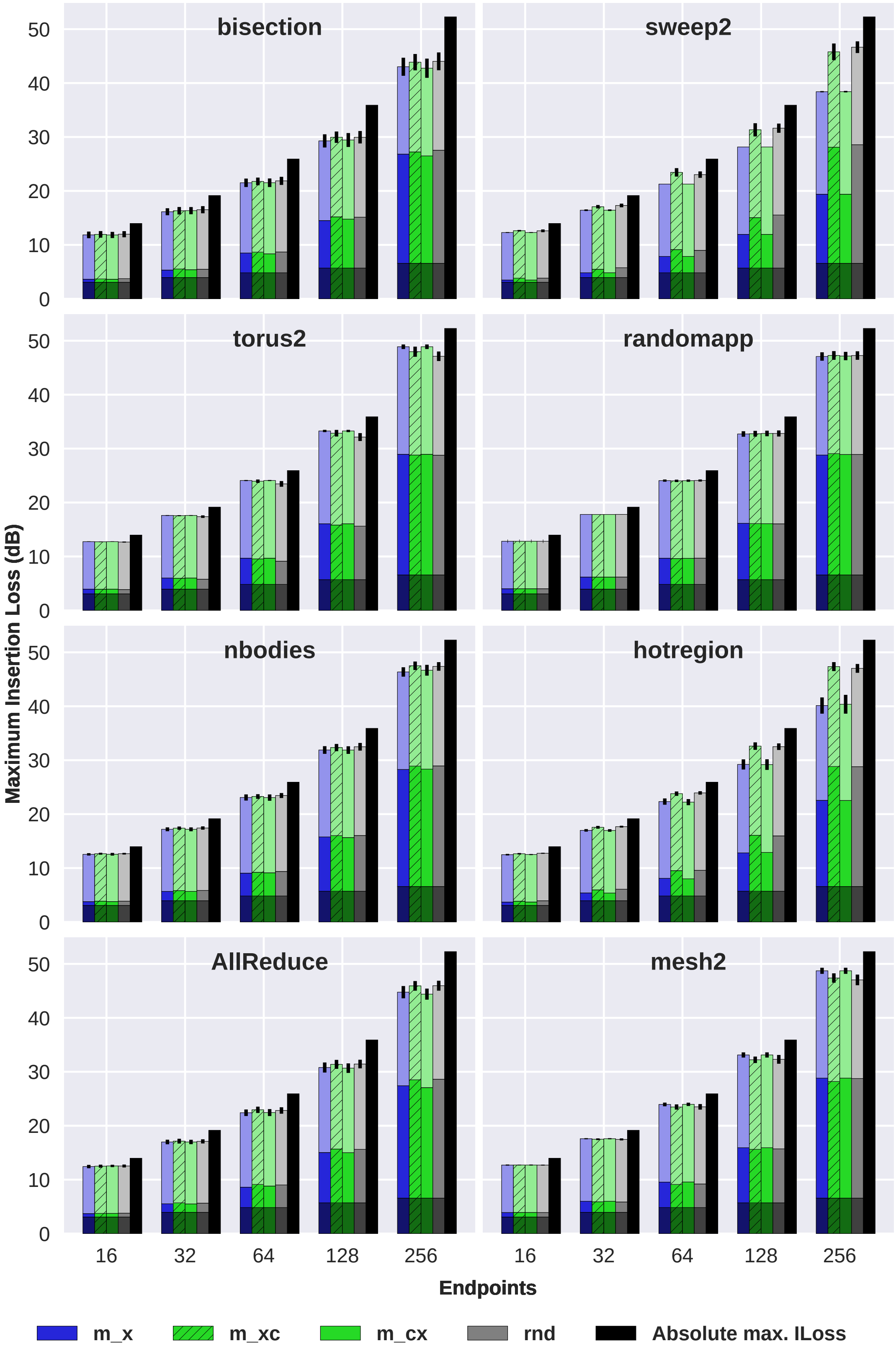


Figure 2: Maximum Insertion Loss Breakdown (dB) using the m_x, m_xc, m_cx and rnd routing strategies. The max. ILoss derived from the specs is depicted in black. All subplots share axes. Darker: Waveguide ILoss. Mid: Crossings ILoss. Lighter: MZI ILoss.

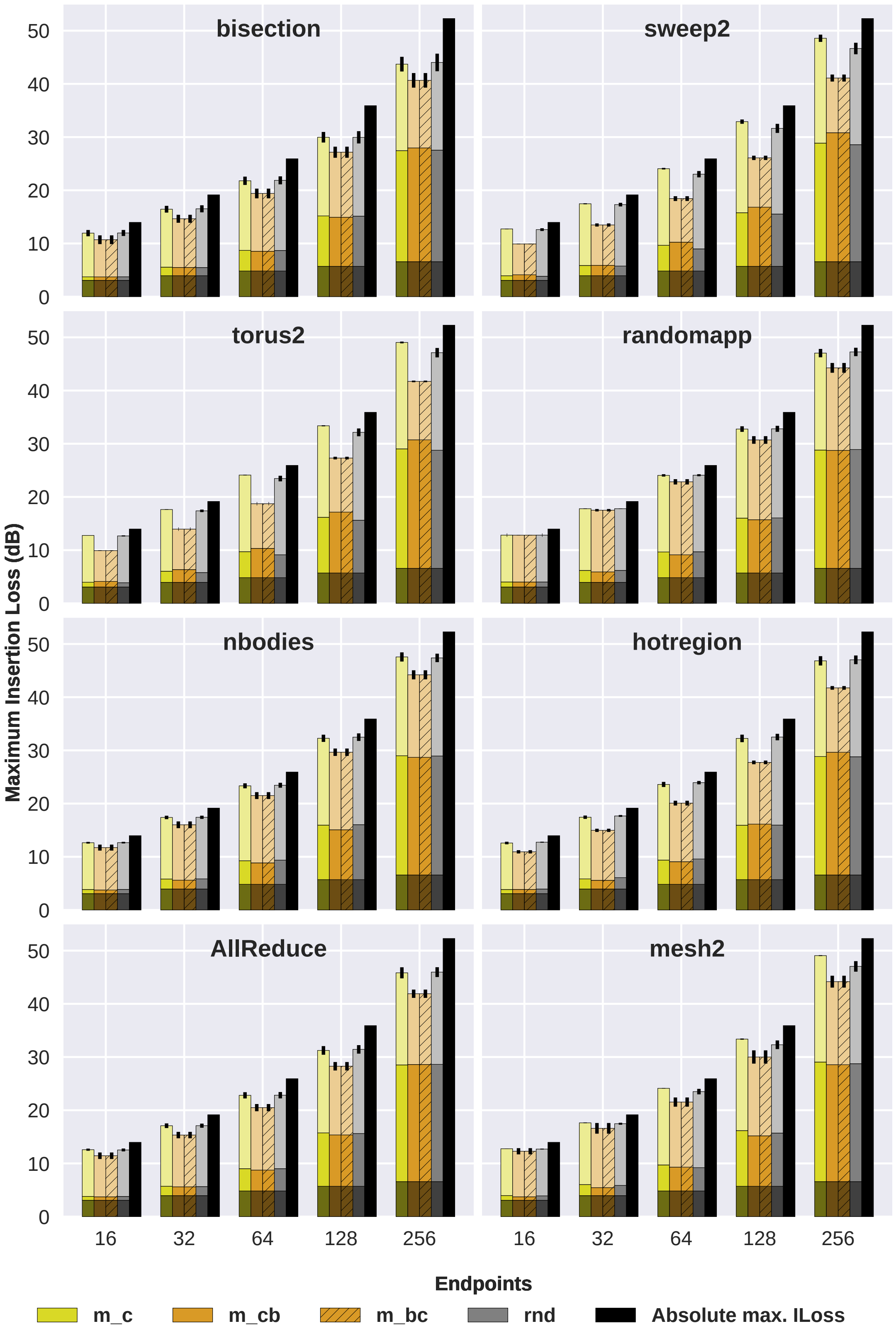


Figure 3: Maximum Insertion Loss Breakdown (dB) using the *m_c*, *m_cb*, *m_bc* and *rnd* routing strategies. The max. ILoss derived from the specs is depicted in black. All subplots share axes. Darker: Waveguide ILoss. Mid: Crossings ILoss. Lighter: MZI ILoss.

range from 7.1-18% for most workloads. However, there are two outlier cases; one is exhibited under the hotregion workload. Here, *m_x* achieves ILoss savings ranging from 10% to 23%, with the savings increasing with the network scale. The same trend is observed with the sweep2 workload as well, where ILoss savings range from 14% to 33%. This interesting effect is ascribed to the organisation of the source and destination pairs in these workloads. As discussed in the workloads description, hotregion forces a large amount of traffic towards a small number of endpoints. As the flows contend for access to these endpoints, they become blocked leaving the network in a less saturated state. Therefore, the remainder of the flows that are allocated random destinations can be allocated more ILoss-favourable paths, thus reducing the worst-case ILoss experienced by the flows. The behaviour of the sweep2 workload is due to the causality in the workload. Increased causality leads to less flows being present in the network at the same time, leading to the strategy being able to select better paths for the flows. This effect occurs with the hybrid strategies that include minimising crossings as a path selection criterion as well. However, aside from these two outlier workloads, *m_x* performs similarly to the baseline strategy (within 5% max. ILoss). Nevertheless, these results show that, under the right routing conditions, prioritising paths with the least amount of crossings can incur savings in Max. ILoss.

The *m_c* strategy presents ILoss savings between 9-16.6% compared to the absolute maximum. However, it yields very little benefit w.r.t. maximum ILoss per flow overall when compared with the baseline strategy. In fact, *m_c* is sometimes outperformed by the baseline with the most notable example being under mesh2 for 16 endpoints and sweep2 for 64, where *rnd* presents approximately 6% and 7% more savings than *m_c*. For the other cases, these two strategies exhibit savings within a 4% margin of each other. Therefore, this strategy is not as beneficial for reducing max. ILoss.

m_b is the best single-criterion strategy overall in decreasing average ILoss per flow. The max. ILoss savings compared to the absolute maximum exhibited by this strategy range from 8.3% (randomapp, 16 endpoints) to 30.6% (sweep2, 64 endpoints). This strategy outperforms the baseline in almost all cases; the only situation where the benefits aren't very significant is for small network sizes under the randomapp and torus2 workloads for 16 endpoints, where *m_b* is within 2% of *rnd*. For the other workloads, *m_b* yields substantially larger savings; indicatively, under the hotregion workload for 64 endpoints, *m_b* saves 17% more max. ILoss than *rnd*. In the other cases, the additional savings range from approx. 3% to 19%. Consequently, this strategy is generally the best single-criterion contender.

However, it is interesting to note that with *m_b*, the max. ILoss savings increase up to a network size of 64 endpoints; from that size onwards, the savings start decreasing substantially. The most representative case of this phenomenon is under the sweep2 workload, where the savings over abs. max reach a peak of 30.6% for 64 endpoints, and then gradually reduce to 19.7% for 256 endpoints. This occurs because of the number of crossings present in the larger network sizes; as the number of waveguide crossings increases, it reduces the capacity of this routing strategy to save on max. ILoss. Especially for the sweep2 workload, as we observed the *m_x* is able to take more advantage of the path selection criterion it uses, allowing it to significantly outperform *m_b*.

As per the insights above, *m_x* and *m_b* are the two most useful single-criterion strategies to adopt for routing.

6.1.2. Hybrid Routing Strategies

We continue our analysis by evaluating the hybrid routing strategies, namely *m_cb*, *m_bc*, *m_xc*, *m_cx*, *m_xb* and *m_bx*. As mentioned in the strategy description previously, these hybrids attempt to reduce max. ILoss by enforcing two path selection criteria in order to leverage the benefits of both. We contextualise our analysis by comparing the savings presented by the hybrids with those presented by the single criterion strategies in order to ascertain the benefit of hybridisation.

Firstly, it is interesting to note that the *m_cb* and *m_bc* hybrids have identical behaviour in all cases, for all network sizes. This indicates that the two selection criteria used in this strategy are too tightly coupled for there to be a complementary effect from their combination. In fact, when compared to *m_b*, the benefits or detriments of these two hybrids fall within a margin of 0.5% difference for the bisection, sweep2 and randomapp workloads, and are identical for the rest. Therefore, hybridising *m_b* with *m_c* is ineffectual as regards max. ILoss, and should therefore be avoided.

The *m_xc* and *m_cx* also have similar behaviour, although the *m_cx* strategy is almost always slightly more beneficial than the inverse. The difference in savings that *m_cx* entails over *m_xc* is between 0-6% in most cases. However, there is the exception of hotregion and sweep2, where *m_cx* consistently outperforms *m_xc*. Interestingly, *m_cx* has a very similar behaviour to *m_x*; as with the previous two hybrids the only differences in max. ILoss savings between *m_cx* and *m_x* are within a margin of 1% or less for the bisection, sweep2 and randomapp workloads. As such, although the *m_cx* hybrid avoids the detrimental behaviour of *m_c*, it does not manage to yield significant savings compared to *m_x*. Therefore, both these hybrid strategies are also ineffectual with reducing max. ILoss.

The last pair of hybrids we examine, namely *m_xb* and *m_bx*, combine the two most useful single-criterion strategies. Here, the *m_xb* strategy is significantly more beneficial than *m_bx*, yielding additional max. ILoss savings between 2-10% for most cases. Exceptions to this are with hotregion, sweep2 and torus2; as with their single criteria counterparts, *m_xb* exhibits higher savings in smaller network sizes which reduce as size increases, while *m_bx*

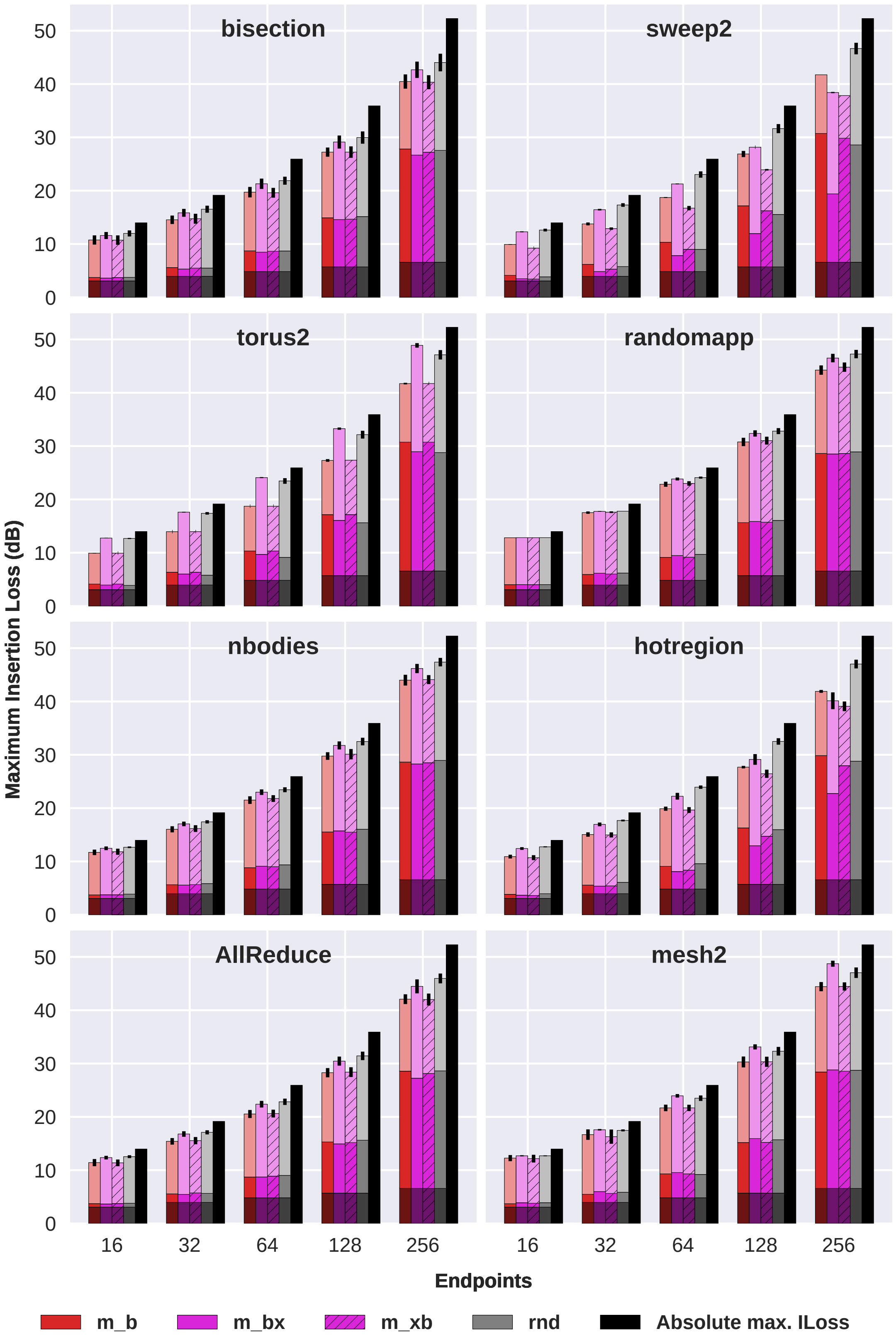


Figure 4: Maximum Insertion Loss Breakdown (dB) using the *m_b*, *m_bx*, *m_xb* and *rnd* routing strategies. The max. ILoss derived from the specs is depicted in black. All subplots share axes. Darker: Waveguide ILoss. Mid: Crossings ILoss. Lighter: MZI ILoss.

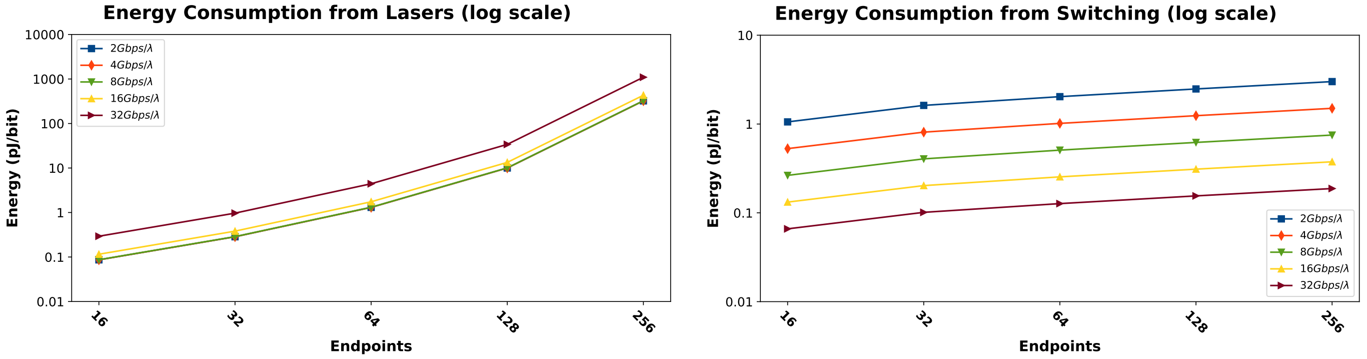


Figure 5: Energy consumption (pJ/bit) from lasers and switching.

starts with negligible ILoss savings and increases to almost on-par with *m_xb*. As explained, as the number of waveguide crossings increases with network size, minimizing them becomes more impactful over the ILoss exhibited by the flows in the workloads. Lastly, it is worth noting that *m_xb* exhibits the highest max. ILoss reduction out of all the routing strategies (35.5%, 64 endpoints, sweep2).

Based on the above, for networks with a size ranging from 16 to 256 endpoints, the *m_xb* routing strategy is the most beneficial overall in reducing max. ILoss. Additionally, the minimal impact exhibited by the other four strategies indicates that further hybridisation, including all three criteria in different orders, would not have a substantial positive effect. This means that the additional complexity of the path selection process, and therefore of the network controller, would be unjustified.

Table 6: DSENT Simulation Parameters.

Parameter	Conservative	Aggresive
Core Rate	2 GHz	
#λ per laser	32	
Laser efficiency	0.25	
Modulator Loss	6 dB [66]	2.5 dB [67]
Extinction Ratio	10 dB [66]	15 dB [67]

6.2. Optimal Data Rate per Wavelength

We continue by discussing the impact of the data rate per λ on the energy consumed by lasers as well as that consumed by switching. We conduct a parameter sweep using DSENT [10] and the max. ILoss derived from our phINRFlow experiments with the randomapp workload described previously and the *rnd* strategy. We constrain this evaluation to randomapp as DSENT also uses random traffic. The DSENT configuration parameters we use are shown in Table 6. We chose 32λ because in principle, this allows for more degrees of freedom for DWDM while sticking to the 100 GHz channel spacing as defined by the ITU-T G.694.1 standard [68]. We derive laser energy consumption from laser power (DSENT), execution time (phINRFlow) and payload for various data rates. For completeness, we conduct a similar parameter sweep in phINRFlow for the corresponding switching energy consumption per data rate, showing the energy dissipated from MZI tuning. The results are depicted in Fig. 5.

Our results show that switching energy scales much more nicely than laser energy, which can be up to 3 orders of magnitude larger for a 256-endpoint NoC. Thus, this is clearly the main limiting factor to scalability in terms of energy per bit. With regards to the data rate per λ, it increases the laser energy consumption for each network size (Fig. 5 left) but reduces that from switching (Fig. 5 right). As we can see, increasing the data rate up to 8Gbps/λ does not increase laser energy consumption substantially (<1%), whereas increasing from 8 to 16 Gbps/λ increases the energy per bit by 33% and further increasing to 32 Gbps/λ entails a 150% energy increase. Consequently, the sweet spot for minimising energy consumption from lasers and switching is 8Gbps/λ, which for 32λ adds up to an aggregate data rate of 256Gbps per endpoint. We will use this data rate for the rest of our experiments.

6.3. Switching Energy Consumption

Moving on to switching energy consumption, we collect the average energy per bit as specified in Section 5.3 for all workloads. As seen previously, the most effective routing strategies can decrease max. ILoss substantially; this, in turn, decreases laser power and therefore energy. In this section, we focus on the energy dissipated from MZI tuning, in order to evaluate whether further energy savings can be achieved from this contributor when using the routing strategies. We compare the routing strategies against *rnd*, which we use as a baseline. The results can be seen in Figs. 6 and 7, split for readability.

6.3.1. Single Criterion Routing Strategies

As before, we first evaluate the routing strategies that use a single criterion. Firstly, it is clear that energy consumption from switching increases logarithmically with the size of the network, which is due to the way in which the number of MZIs scales.

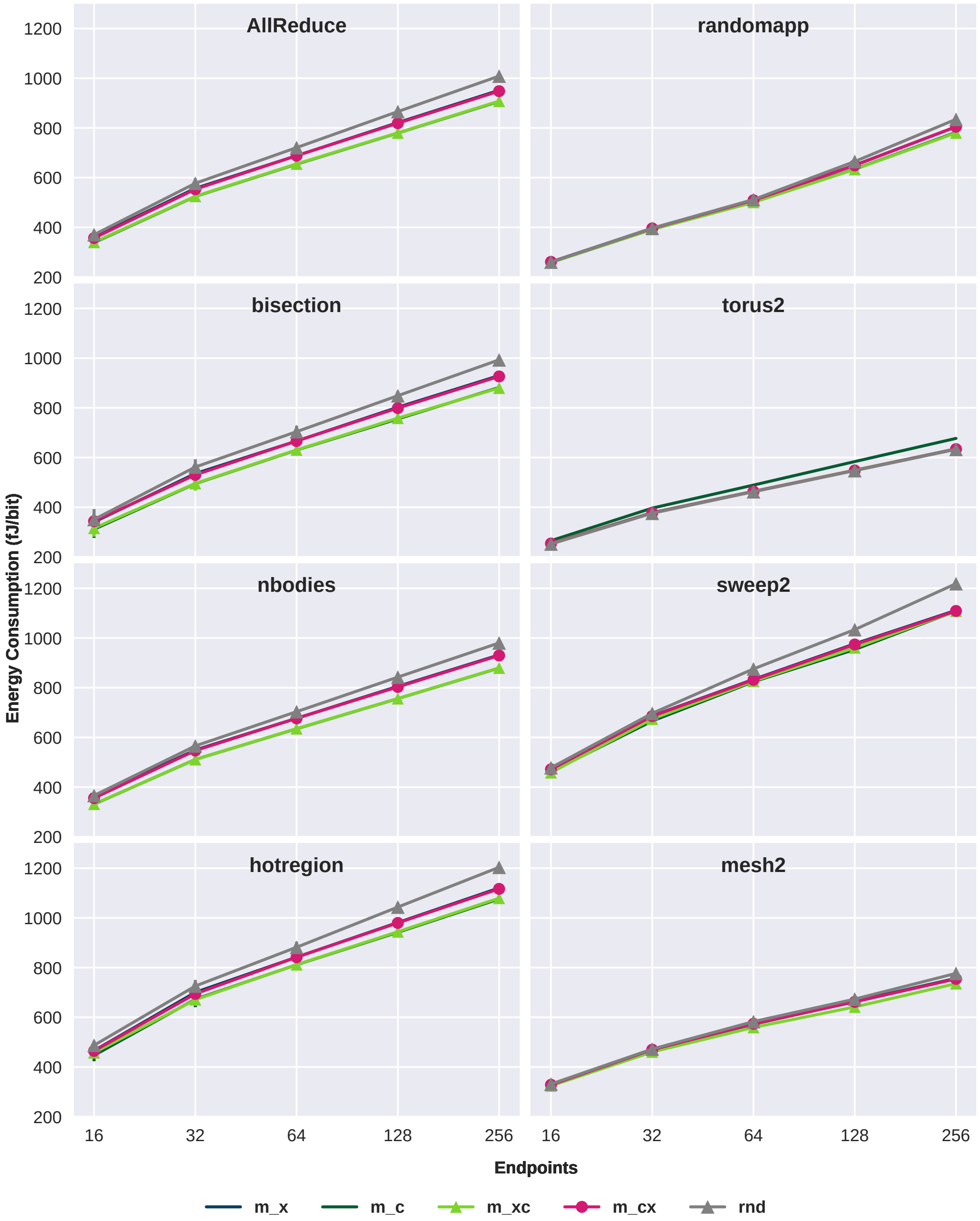


Figure 6: Switching energy consumption (fJ/bit) for the m_x, m_c, m_xc, m_cx and rnd routing strategies.

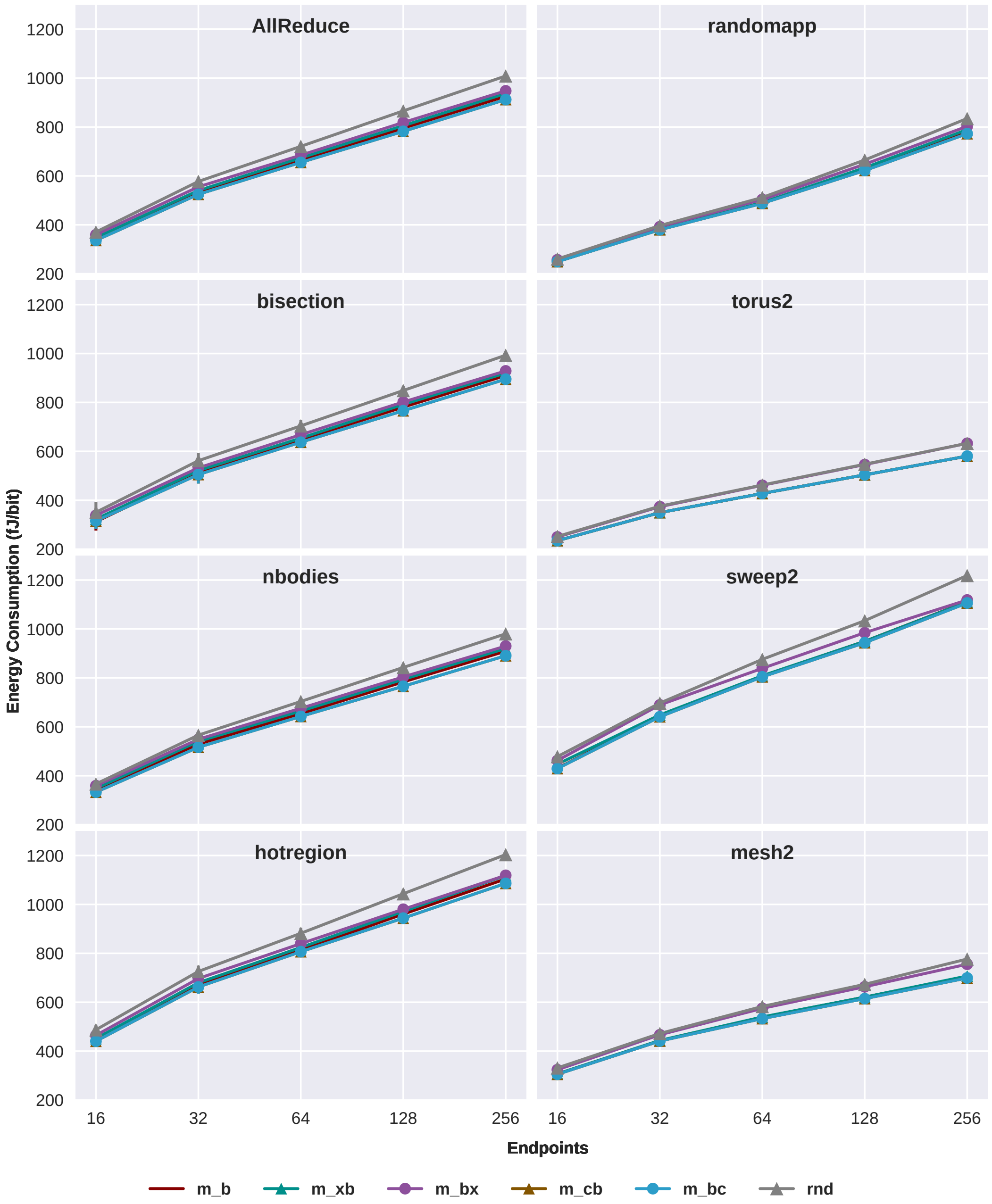


Figure 7: Switching energy consumption (fJ/bit) for the m_b, m_xb, m_bx, m_cb, m_bc and rnd routing strategies.

The *m_x* strategy is generally the worst performer of the three strategies with respect to the baseline yielding the most modest energy savings. It's performance ranges from a 1% *increase* in consumption (torus2, 16 endpoints) to a 7% decrease against *rnd* (hotregion, 256 endpoints). However, as this strategy is more aimed at decreasing max. ILoss, it is expected that it will perform worse than the other strategies w.r.t. switching energy. Nevertheless, it still provides benefits.

The *m_c* strategy performs better, exhibiting the highest energy savings of the single-criterion strategies for bisection, butterfly and nbodies, ranging between 10-11%, 8-10% and 9-10% respectively. However, the strategy exhibits a 5-6% increase in switching energy under torus2 and is less effective than *m_c* in the rest of the workloads by approximately 1-2%. Therefore, this strategy's performance w.r.t. switching energy is not beneficial enough to offset its' poor performance w.r.t. max. ILoss.

M_b consistently outperforms our baseline and yields more savings than the other two strategies in most cases, with the exceptions where it is outperformed by *m_c*. The most savings are presented for the bisection, where savings against *rnd* are between 8.3-11%. In the rest of the workloads the strategy yields savings between 6-9% with the exception of randomapp, where savings are more modest (4-6%). Consequently, this is the best strategy to use either on its own, or through hybridisation.

6.3.2. Hybrid Routing Strategies

Here, we evaluate the hybrid routing strategies in a pair-wise fashion as before, comparing their performance against our baseline.

As with max. ILoss, the *m_cb* and *m_bc* hybrids have identical behaviour in all cases, for all network sizes. Again, this indicates that these two path selection criteria do not produce a benefit when combined. They do, however, consistently outperform both the baseline and *m_b* by 1-3%. However, these additional savings are very small (20-30 fJ/bit at best), meaning that they are ineffectual w.r.t. energy consumption. Nevertheless, this case shows how routing criteria hybridisation can have a positive impact on switching energy. Even so, considering the fact that adding a second path selection criterion increases the complexity of the path selection algorithm, it is not worthwhile in this case.

In contrast to their behaviour as regards max. ILoss, the *m_xc* and *m_cx* present different savings to each other w.r.t. energy consumption, with *m_xc* consistently providing 3-5% more savings in energy. The only exception is under torus2, where the benefits are within 1%; however, in this case the baseline outperforms these strategies. The *m_xc* also shows some of the largest energy savings of all strategies; however, considering its relatively poor performance in max. ILoss it is not justified. Again, these two hybrids do not provide enough of an impact to justify their usage when considering the added complexity in the path selection algorithm.

The *m_xb* and *m_bx* strategies present significantly different behaviour w.r.t. each other. *M_bx* consistently outperforms *m_x* by 1-2% and *m_xb* consistently outperforms *m_bx* as well. However, the *m_xb* hybrid is slightly outperformed by *m_b*, with the latter yielding an additional 1-5% energy savings across the workloads and network sizes. The *m_xb* strategy's performance, coupled with the hybrid's consistent max. ILoss savings over the baseline, make it the most useful hybrid strategy for reducing energy consumption across both metrics.

6.4. Execution Time

We now evaluate the impact of our routing strategies on workload execution time, depicted in Fig. 8. We scale the execution time of each strategy against that of our baseline, *rnd*, for each workload and network size, in order to determine whether there are adverse impacts that may offset the ILoss and energy benefits we discussed above.

Firstly, the results show that no routing strategy increases execution time by more than 3% over *rnd*. In fact, in most cases they perform as well as or significantly better than the baseline. We note that with a rearrangeably non-blocking network using circuit switching, execution time should not be greatly affected either way. Nevertheless, in some cases the routing strategies show execution time savings between 5-10% and, in the case of *m_c* and *m_xc*, up to 26%.

The routing strategies have virtually no effect on torus2 and hotregion. In the case of hotregion, this is because of the flow distribution. As 25% of flows are directed to 12.5% of the network, the flows will contend for the destinations (as seen with ILoss) and be blocked at the receiver. As we are employing circuit switching, blocked flows must wait for the whole transmission of the flow being currently serviced before they may access the endpoint. Therefore, attempting to assign specific paths to flows will not be able to reduce execution time. Torus2, on the other hand, is a high-causality workload; flow destinations are assigned according to a strict rule, which yields path selection algorithms ineffective at reducing total execution time. Nevertheless, the metric is not inhibited by the strategies, meaning that with the lower insertion loss exhibited previously, the best strategies could arguably reduce total energy consumption for the communication.

Bisection presents an interesting case. Here, the behaviour of the routing strategies is particularly erratic, with execution time increasing slightly (e.g. ≈2% for *m_bx* at 16 endpoints) in some cases and decreasing substantially (26% for *m_c* and *m_xc* at 256 endpoints) in others. However, this is expected considering that on the one hand the workload has very few flows and on the other it is a permutation workload, aiming to saturate the network with flows. This behaviour causes intermediate contention in the switching fabric where, because flows are assigned paths incrementally and the network is rearrangeably non-blocking, leading to there being no available paths for a small amount of flows. The most common solution to this

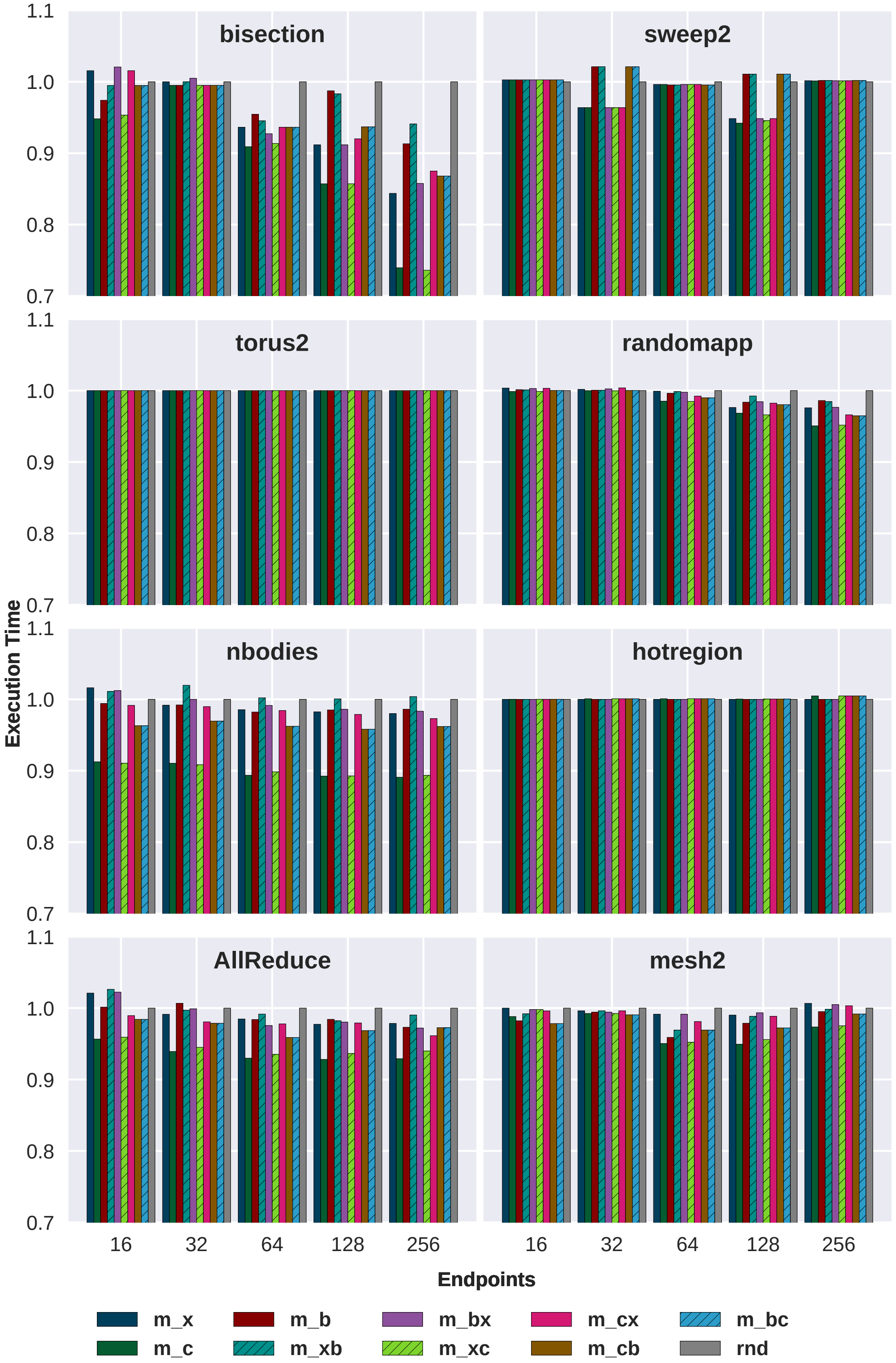


Figure 8: Execution time relative to *rnd* for all strategies.

problem is known as the "Looping algorithm" which can successfully solve full or near-full permutations with high compute efficiency. Based on our work, we plan to adapt the algorithm in the future to take into account underlying hardware constraints for photonic networks.

The strategies present similar behaviour in the nbodies, allreduce and mesh2 workloads. Most routing strategies are within 5% execution time relative to the baseline except for *m_c* and *m_xc*; the latter show substantial execution time reductions, especially in nbodies. Interestingly, in contrast to *m_xc*, *m_bc* is not able to reduce execution time substantially; this is another indication that the component criteria are too tightly coupled to afford pronounced benefits when hybridising.

In sweep2, another workload with high causality and few flows, the routing strategies also show little effect on execution time. The exception is for 32 and 128 endpoints, where execution time is increased slightly (2-4%) by *m_b*, *m_xb*, *m_cb* and *m_bc* but decreased slightly with the rest (3-5%). Lastly, the routing strategies show little effect on randomapp, with a slight decrease in execution time for larger networks across all routing strategies.

In conclusion, the routing strategies do not prohibitively increase execution time and, in some cases, can yield substantial savings with respect to random allocation. In addition, in spite of their poor performance w.r.t. max. ILoss, the *m_c* and *m_xc* provide substantial benefits for some of the workloads. This reduction in execution time could arguably offset the increase in laser power incurred by the max. ILoss increases; we plan to investigate this and develop additional routing algorithms using this technique in the future.

6.5. Laser Power

In our previous work, we showed the power per laser for different network sizes assuming a modulator with a very low extinction ratio, in order to outline a worst-case scenario for laser power. Here, we analyse the required power per laser using two carrier depletion MRR-based modulators (MRMs), representing a conservative [66] and aggressive technology case [67] (see table 6). We note that MRMs based on the carrier injection principle can afford better properties [69]; however, as DSENT models carrier depletion-based MRMs, we selected examples in the literature which conform to this model. The results are shown in fig. 9, where we analyse how laser power is impacted by our most efficient routing strategies. These results conclusively show that laser power, affected by max. ILoss, is a main scaling inhibitor. Indicatively, using the *m_b* strategy, a network of 256 endpoints requires 118 and 49 W per laser in the conservative and aggressive cases respectively. Considering that the NoC may take up to 24% of a SoC's power budget [70], a typical 100W power budget as exhibited by regular server-grade processors could not accommodate for this. Larger many-cores such as Intel Xeon Phi have a much higher power budget (around 300W), which would allow for 75W to be dedicated to the interconnect; however even with the larger power envelope, NoCs of more than 64 endpoints are unrealistic. Reducing laser power without changing technology parameters would entail sacrificing throughput, by reducing the number of λ. However, 32 endpoints become reasonable for different applications for both cases. For instance, using the *m_xb* strategy, lasers require a total of 8.3 W to drive the network in the conservative case and 3.5 W with the more modern modulator. Using a simple random path allocator cannot accommodate for further scaling even in the aggressive case, as a 64-endpoint NoC requires 30 W for lasers as per our baseline. However, using the *m_xb* strategy, laser power falls 23.6 W for all lasers, potentially bringing a 64-endpoint NoC under the boundary of feasibility.

Another advantage is that for the specified power budget, this technology offers a rearrangeably non-blocking network as opposed to a 2D electrical mesh, which is prone to contention; the flat latency exhibited is also beneficial when compared to the non-uniform latency characteristics of a mesh interconnect.

Based on the above, as laser power is the primary contributor to power consumption in this technology, a 32-endpoint EO/TO MZI-based network can be considered for the on-chip domain. Using novel, more efficient lasers could reduce the power requirements of the larger network sizes into this domain, as would using carrier injection modulators. Another way to drastically reduce laser power and therefore achieve better scaling has been proposed by Demir et al. in [71]; they report laser power reductions of up to 92% through the use of on-chip lasers and adaptive laser control techniques. As mentioned in 2.1, however, on-chip lasers are still a research challenge for Silicon Photonics.

The power per laser results also show the substantial impact afforded by the routing strategies w.r.t. laser power; all strategies exhibit substantial laser power savings at every size w.r.t. the absolute maximum, which is exhibited by the max. ILoss calculated from the original device parameters. These savings are exhibited irrespective of the choice of modulator. *M_x*, *m_c* and *rnd* exhibit similar power savings, ranging from 23% for 16 endpoints to ≈70% for 256 endpoints. *M_b* performs even better, with savings ranging from 23% to 85% across the network sizes. It is also interesting to note that with random traffic, most hybrid strategies perform worse than *m_b*. The only hybrid that performs marginally better is *m_cb* and there it manages to reduce laser power by ≤1% more than the simpler strategy. However, as explained in section 6.1, the choice of routing strategy has a profound impact on max. ILoss depending on the workload's communication characteristics. Considering that laser power increases exponentially with max. ILoss and that the best hybrid strategies consistently yielded additional savings w.r.t. *m_b*, the additional complexity of hybrid routing strategies can be justified.

Based on the discussion above, an interesting use case scenario for this technology is to employ core aggregation

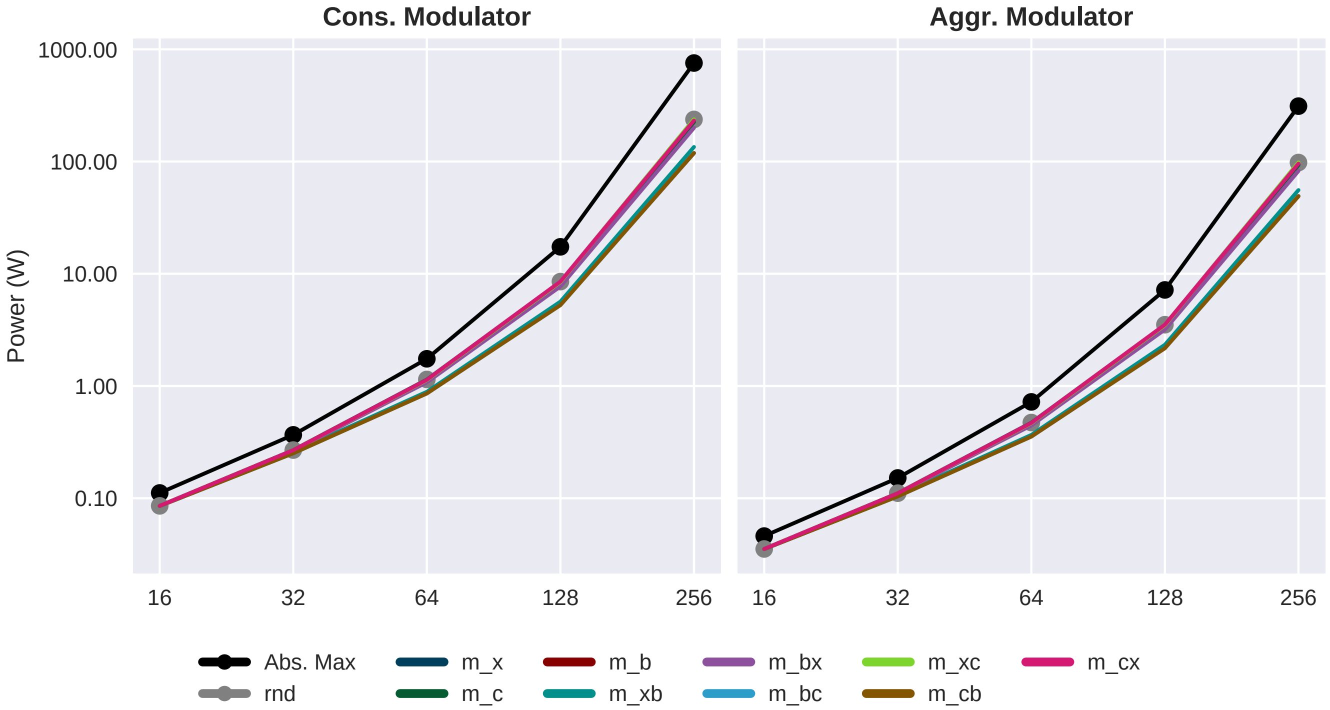


Figure 9: Required power per laser. Left: conservative modulator [66]. Right: aggressive modulator [67].

at the endpoints, enabled by the large bandwidth per endpoint (256Gbps) demonstrated in our model. The cores within each core group could be interconnected using a conventional electrical crossbar to save on laser power and each group would contain a photonic network interface, which would convert the electrical signalling into the optical domain. The interface would then use an appropriate arbitration mechanism, such as time-division multiplexing (TDM), to assign carrier capacity to each communication request, thereby composing a communication flow. More complex wavelength partitioning techniques, such as λ-routing or DWDM may also be used to compose the flow in an efficient manner. This flow would then be offloaded into the Beneš interconnect, after the interface has been allocated a path by the central control process, in order for it to reach the destination group. The destination group's interface would then demodulate the optical stream, assessing which segment of information is destined for which core and finalise the communication.

This use-case, however, entails various challenges. Firstly, the optimal number of cores per group would need to be determined, such that on the one hand the amount of communication requests is substantial enough to benefit from the optical network, while on the other hand fitting within the constrained NoC power and area envelope. Secondly, the most efficient stream partitioning strategy must be determined; if each group contains a different number of cores to the number of λ, simple λ-routing will be inefficient and more complex techniques, such as DWDM, must be explored. Thirdly, due to both the circuit switching and non-wavelength selective nature of the Beneš interconnect we examine, this use case would be able to realise communication requests such that only two groups can communicate with each other at any one time. Appropriately mapping communication workloads to the groups in order to exhibit performance gains is an interesting research direction we plan on investigating in the future.

7. Conclusions & Future Work

In this work, we have evaluated the benefits and challenges of scaling out a thermally and electrically tuned MZI-based optical Beneš network. We have shown that up to 128 endpoints, laser light incurs the most maximum ILoss from MZIs, whereas for larger sizes the aggregation of waveguide crossings contributes most to maximum ILoss.

In order to minimise ILoss and energy consumption in the network, we have presented a set of hardware-inspired routing strategies which leverage the asymmetric behaviours of internal switching elements. We then evaluated their usage while exposing our model to 8 communication workloads. We have demonstrated that using these strategies always reduces the maximum ILoss exhibited under an information flow. Furthermore, we showed that minimising the number of MZIs in "bar" state can reduce maximum ILoss by 30.6% in the best case (sweep2, 64 endpoints), while minimising the number of crossings shows a 33% best-case reduction over the absolute maximum (sweep2, 256 endpoints). Through our experiments, we showed that minimising the amount of state changes within the network does not reduce max. ILoss as much as other strategies, but reduces switching energy consumption. Furthermore, we demonstrated that routing strategy hybridisation is beneficial, with the *m_xb* hybrid consistently reducing max. ILoss by more than the single-criteria counterparts and offering the highest max. ILoss savings (35.5%, 64 endpoints, sweep2).

We then determined the most efficient data rate per λ to be 8 Gbps/λ, as this offers the least energy consumption from both lasers and switching. We also showed that the best routing strategy can consistently offer between 8% and 15% reduction in bit-switching energy consumption with respect to random path selection depending on the communication workload.

We continued by evaluating the impact of the routing strategies on workload execution time finding that in the worst case, execution time increases by at most 3% compared to the baseline and, in some cases, is decreased very significantly (5-25%).

Finally, we showcased that a network of 32 endpoints is suitable for the on-chip domain using a conservative modulator, and demonstrated substantial laser power reduction with the best routing strategy (23-85% across the network sizes). A 64-endpoint NoC is potentially achievable using the best routing strategy and a more modern modulator as well. We also discussed the merits of core aggregation at the endpoints of a 32-endpoint NoC and outlined potential routing challenges.

In the future, we aim to augment our simulation framework, such that a comprehensive examination of optical crosstalk can be conducted. Optical crosstalk is an issue commonly encountered with multi-stage optical interconnects [29] [72]. However, as expanded on in [29], analytical simulation cannot accurately describe phenomena related to crosstalk, such as inter-aggressor skew. Consequently, evaluating crosstalk requires a statistical approach, which we plan to integrate into our simulator. We expect that this will allow for a fruitful insight on how crosstalk is incurred depending on the workload specification and will reveal novel ways of reducing the penalty through routing. We also plan to investigate nested network topologies using variable sizes of this model as well as the impact of the routing strategies in such a use-case.

8. Acknowledgement

This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 671553.

9. References

References

[1] D. Thomson, et al., Roadmap on silicon photonics, Journal of Optics 18 (7) (2016) 073003.

[2] S. Werner, J. Navaridas, M. Luján, Efficient sharing of optical resources in low-power optical networks-on-chip, J. Opt. Commun. Netw. 9 (5) (2017) 364–374. doi:10.1364/JOCN.9.000364.

[3] J. Bashir, E. Peter, S. R. Sarangi, A survey of on-chip optical interconnects, ACM Comput. Surv. 51 (6) (2019) 115:1–115:34. doi:10.1145/3267934.

[4] R. Soref, The past, present, and future of silicon photonics, IEEE Journal of selected topics in quantum electronics 12 (6) (2006) 1678–1687.

[5] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, K. Bergman, Recent advances in optical technologies for data centers: a review, Optica 5 (11) (2018) 1354–1370. doi:10.1364/OPTICA.5.001354.

[6] L. Lu, et al., 16x16 non-blocking silicon optical switch based on electro-optic mach-zehnder interferometers, Opt. Express 24 (9) (2016) 9295–9307. doi:10.1364/OE.24.009295.

[7] M. A. Taubenblatt, Optical interconnects for high-performance computing, J. Lightwave Technol. 30 (4) (2012) 448–457.

[8] S. Rumley, et al., Optical interconnects for extreme scale computing systems, Parallel Computing 64 (2017) 65–80.

[9] T. C. A, et al., Survey of photonic and plasmonic interconnect technologies for intra-datacenter and high-performance computing communications, IEEE Communications Surveys & Tutorials (2018).

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

[11] Y.-Q. Bie, G. Grosso, M. Heuck, M. M. Furchi, Y. Cao, J. Zheng, D. Bunandar, E. Navarro-Moratalla, L. Zhou, D. K. Efetov, et al., A mote 2-based light-emitting diode and photodetector for silicon photonic integrated circuits, Nature nanotechnology 12 (12) (2017) 1124.

[12] S. Werner, J. Navaridas, M. Luján, A survey on optical network-on-chip architectures, ACM Comput. Surv. 50 (6) (2017) 89:1–89:37. doi:10.1145/3131346.

[13] Z. Wang, K. Van Gasse, V. Moskalenko, S. Latkowski, E. Bente, B. Kuyken, G. Roelkens, A III-V-on-Si ultra-dense comb laser, Light: Science & Applications 6 (5) (2017) e16260–e16260. doi:10.1038/lsa.2016.260.

[14] Z. Wang, A. Abbasi, U. Dave, A. De Groote, S. Kumari, B. Kunert, C. Merckling, M. Pantouvaki, Y. Shi, B. Tian, et al., Novel light source integration approaches for silicon photonics, Laser & Photonics Reviews 11 (4) (2017) 1700063.

[15] D. J. Blumenthal, R. Heideman, D. Geuzebroek, A. Leinse, C. Roeloffzen, Silicon nitride in silicon photonics, Proceedings of the IEEE 106 (12) (2018) 2209–2231.

[16] G.-H. Duan, S. Olivier, C. Jany, S. Malhouitre, A. Le Liepvre, A. Shen, X. Pommarede, G. Levaufre, N. Girard, D. Make, et al., Hybrid iii-v silicon photonic integrated circuits for optical communication applications, IEEE Journal of Selected Topics in Quantum Electronics 22 (6) (2016) 379–389.

[17] S. Chen, W. Li, J. Wu, Q. Jiang, M. Tang, S. Shutts, S. N. Elliott, A. Sobiesierski, A. J. Seeds, I. Ross, et al., Electrically pumped continuous-wave iii–v quantum dot lasers on silicon, Nature Photonics 10 (5) (2016) 307.

[18] Z. Zhou, B. Yin, J. Michel, On-chip light sources for silicon photonics, Light: Science & Applications 4 (11) (2015) e358.

[19] H. W. Then, M. Feng, N. Holonyak, The transistor laser: Theory and experiment, Proceedings of the IEEE 101 (10) (2013) 2271–2298.

[20] M. C. Wu, T. J. Seok, Large-scale silicon photonic switches, in: 2018 Asia Communications and Photonics Conference (ACP), Ieee, 2018, pp. 1–3.

[21] Y. Li, Y. Zhang, L. Zhang, A. W. Poon, Silicon and hybrid silicon photonic devices for intra-datacenter applications: state of the art and perspectives [Invited], Photonics Research 3 (5) (2015) B10. doi:10.1364/prj.3.000b10.

[22] B. G. Lee, A. Biberman, N. Sherwood-Droz, C. B. Poitras, M. Lipson, K. Bergman, High-speed 2x2 switch for multiwave-

length silicon-photonic networks–on-chip, Journal of Lightwave Technology 27 (14) (2009) 2900–2907.

[23] A. Biberman, P. Dong, B. G. Lee, J. D. Foster, M. Lipson, K. Bergman, Silicon microring resonator-based broadband comb switch for wavelength-parallel message routing, in: LEOS 2007-IEEE Lasers and Electro-Optics Society Annual Meeting Conference Proceedings, IEEE, 2007, pp. 474–475.

[24] H. Subbaraman, X. Xu, A. Hosseini, X. Zhang, Y. Zhang, D. Kwong, R. T. Chen, Recent advances in silicon-based passive and active optical interconnects, Optics Express 23 (3) (2015) 2487. doi:10.1364/oe.23.002487.

[25] Z. Lu, D. Celo, H. Mehrvar, E. Bernier, L. Chrostowski, High-performance silicon photonic tri-state switch based on balanced nested mach-zehnder interferometer, Scientific reports 7 (1) (2017) 12244.

[26] N. Dupuis, A. V. Rylyakov, C. L. Schow, D. M. Kuchta, C. W. Baks, J. S. Orcutt, D. M. Gill, W. M. Green, B. G. Lee, Ultralow crosstalk nanosecond-scale nested 2× 2 mach–zehnder silicon photonic switch, Optics letters 41 (13) (2016) 3002–3005.

[27] F. Duan, K. Chen, Y. Yu, High-speed and low-power thermally tunable devices with suspended silicon waveguide, Optical and Quantum Electronics 52 (1) (2019) 5. doi:10.1007/s11082-019-2124-1.
URL https://doi.org/10.1007/s11082-019-2124-1

[28] Z. Guo, L. Lu, L. Zhou, L. Shen, J. Chen, 16x16 silicon optical switch based on dual-ring-assisted mach–zehnder interferometers, Journal of Lightwave Technology 36 (2) (2017) 225–232.

[29] N. Dupuis, B. G. Lee, Impact of topology on the scalability of mach–zehnder-based multistage silicon photonic switch networks, Journal of Lightwave Technology 36 (3) (2017) 763–772.

[30] S. Bahirat, S. Pasricha, Meteor: Hybrid photonic ring-mesh network-on-chip for multicore architectures, ACM Trans. Embed. Comput. Syst. 13 (3s) (Mar. 2014). doi:10.1145/2567940.
URL https://doi.org/10.1145/2567940

[31] P. K. Hamedani, N. E. Jerger, S. Hessabi, Qut: A low-power optical network-on-chip, in: 2014 Eighth IEEE/ACM International Symposium on Networks-on-Chip (NoCS), IEEE, 2014, pp. 80–87.

[32] D. Vantrease, , et al., Corona: System implications of emerging nanophotonic technology, in: ACM SIGARCH Computer Architecture News, Vol. 36, IEEE Computer Society, 2008, pp. 153–164.

[33] G. Kurian, J. E. Miller, J. Psota, J. Eastep, J. Liu, J. Michel, L. C. Kimerling, A. Agarwal, Atac: A 1000-core cache-coherent processor with on-chip optical network, in: Proceedings of the 19th International Conference on Parallel Architectures and Compilation Techniques, PACT '10, ACM, New York, NY, USA, 2010, pp. 477–488. doi:10.1145/1854273.1854332.

[34] W. Tan, H. Gu, Y. Yang, K. Wang, X. Wang, Venus: A low-latency, low-loss 3-d hybrid network-on-chip for kilocore systems, J. Lightwave Technol. 35 (24) (2017) 5448–5455.

[35] S. Werner, J. Navaridas, M. Luján, Designing low-power, low-latency networks-on-chip by optimally combining electrical and optical links, in: 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA), IEEE, 2017, pp. 265–276.

[36] P. Fotouhi, S. Werner, R. Proietti, X. Xiao, S. B. Yoo, Enabling scalable disintegrated computing systems with awgr-based 2.5 d interconnection networks, IEEE/OSA Journal of Optical Communications and Networking 11 (7) (2019) 333–346.

[37] S. Werner, P. Fotouhi, R. Proietti, X. Xiao, S. B. Yoo, Towards energy-efficient high-throughput photonic nocs for 2.5 d integrated systems: A case for awgrs, in: 2018 Twelfth IEEE/ACM International Symposium on Networks-on-Chip (NOCS), IEEE, 2018, pp. 1–8.

[38] M. R. Jokar, J. Qiu, F. T. Chong, L. L. Goddard, J. M. Dallesasse, M. Feng, Y. Li, Baldur: A power-efficient and scalable network using all-optical switches, in: 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA), IEEE, 2020, pp. 153–166.

[39] P. Koka, M. O. McCracken, H. Schwetman, X. Zheng, R. Ho, A. V. Krishnamoorthy, Silicon-photonic network architectures for scalable, power-efficient multi-chip systems, ACM SIGARCH Computer Architecture News 38 (3) (2010) 117–128.

[40] Y. Demir, Y. Pan, S. Song, N. Hardavellas, J. Kim, G. Memik, Galaxy: A high-performance energy-efficient multi-chip architecture using photonic interconnects, in: Proceedings of the 28th ACM international conference on Supercomputing, ACM, 2014, pp. 303–312.

[41] T. Alexoudi, N. Terzenidis, S. Pitris, M. Moralis-Pegios, P. Maniotis, C. Vagionas, C. Mitsolidou, G. Mourgias-Alexandris, G. T. Kanellos, A. Miliou, et al., 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.

[42] K. Katrinis, D. Syrivelis, D. Pnevmatikatos, G. Zervas, D. Theodoropoulos, I. Koutsopoulos, K. Hasharoni, D. Raho, C. Pinto, F. Espina, et al., Rack-scale disaggregated cloud data centers: The dredbox project vision, in: Proceedings of the 2016 Conference on Design, Automation & Test in Europe, EDA Consortium, 2016, pp. 690–695.

[43] O. Liboiron-Ladouceur, et al., The data vortex optical packet switched interconnection network, J. Lightwave Technol. 26 (13) (2008) 1777–1789.

[44] C. Minkenberg, F. Abel, P. Muller, R. Krishnamurthy, M. Gusat, B. R. Hemenway, Control path implementation for a low-latency optical hpc switch, in: 13th Symposium on High Performance Interconnects (HOTI'05), IEEE, 2005, pp. 29–35.

[45] K. Wen, et al., Flexfly: Enabling a reconfigurable dragonfly through silicon photonics, 2016, pp. 166–177. doi:10.1109/SC.2016.14.

[46] X. Ye, Y. Yin, S. J. B. Yoo, P. Mejia, R. Proietti, V. Akella, Dos: A scalable optical switch for datacenters, in: Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems, ANCS '10, Association for Computing Machinery, New York, NY, USA, 2010. doi:10.1145/1872007.1872037.
URL https://doi.org/10.1145/1872007.1872037

[47] C. Minkenberg, et al., Reimagining datacenter topologies with integrated silicon photonics, J. Opt. Commun. Netw. 10 (7) (2018) B126–B139. doi:10.1364/JOCN.10.00B126.

[48] N. Calabretta, R. P. Centelles, S. D. Lucente, H. J. S. Dorren, On the performance of a large-scale optical packet switch under realistic data center traffic, J. Opt. Commun. Netw. 5 (6) (2013) 565–573. doi:10.1364/JOCN.5.000565.

[49] Y. Thonnart, M. Zid, Technology assessment of silicon interposers for manycore socs: Active, passive, or optical?, in: 2014 Eighth IEEE/ACM International Symposium on Networks-on-Chip (NoCS), IEEE, 2014, pp. 168–169.

[50] N. Terzenidis, M. Moralis-Pegios, S. Pitris, C. Mitsolidou, G. Mourgias-Alexandris, A. Tsakyridis, C. Vagionas, T. Alexoudi, N. Pleros, K. Vyrsokinos, Optics for disaggregating data centers and disintegrating computing, Tech. rep., Aristotle University of Thessaloniki (2019).

[51] R. Luijten, R. Grzybowski, The osmosis optical packet switch for supercomputers, in: Optical Fiber Communication Conference and National Fiber Optic Engineers Conference, Optical Society of America, 2009, p. OTuF3. doi:10.1364/OFC.2009.OTuF3.

[52] W. J. Dally, B. P. Towles, Principles and practices of interconnection networks, Elsevier, 2004.

[53] R. Mehra, J. Tripathi, Machzehnder interferometer and it's applications, International Journal of Computer Applications 1 (9) (2010) 110–118.

[54] T. Shimoe, K. Hajikano, K. Murakami, Path-independent insertion loss optical space switch, in: Optical Fiber Communication Conference, Optical Society of America, 1987, p. WB2.

[55] C.-C. Lu, R. A. Thompson, The double-layer network architecture for photonic switching, Journal of lightwave technology 12 (8) (1994) 1482–1489.

[56] E. Bernier, D. J. Goodwill, et al., 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, 2019, pp. 1–3.

[57] R. Hendry, D. Nikolova, S. Rumley, K. Bergman, Modeling and evaluation of chip-to-chip scale silicon photonic networks, in: 2014 IEEE 22nd Annual Symposium on High-Performance Interconnects, IEEE, 2014, pp. 1–8.

[58] A. Bianco, D. Cuda, R. Gaudino, G. Gavilanes, F. Neri, M. Petracca, Scalability of optical interconnects based on microring resonators, IEEE Photonics Technology Letters 22 (15) (2010) 1081–1083.

[59] B. G. Lee, N. Dupuis, Silicon photonic switch fabrics: Technology and architecture, Journal of Lightwave Technology 37 (1) (2019) 6–20.

[60] Z. Wang, Z. Wang, J. Xu, P. Yang, L. H. K. Duong, Z. Wang, H. Li, R. K. V. Maeda, Low-loss high-radix integrated optical switch networks for software-defined servers, Journal of Lightwave Technology 34 (18) (2016) 4364–4375.

[61] P.-H. Yuen, L.-K. Chen, Optimization of microring-based interconnection by leveraging the asymmetric behaviors of switching elements, J. Lightwave Technol. 31 (10) (2013) 1585–1592.

[62] Q. Cheng, M. Bahadori, K. Bergman, Advanced path mapping for silicon photonic switch fabrics, in: 2017 Conference on Lasers and Electro-Optics (CLEO), 2017, pp. 1–2.

[63] M. Kynigos, J. A. Pascual, J. Navaridas, M. Luján, J. Goodacre, 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, ACM, 2019, p. 9.

[64] J. Navaridas, J. A. Pascual, A. Erickson, I. A. Stewart, M. Luján, Inrflow: An interconnection networks research flow-level simulation framework, Journal of Parallel and Distributed Computing 130 (2019) 140 – 152. doi:https://doi.org/10.1016/j.jpdc.2019.03.013.

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

[66] P. Dong, S. Liao, D. Feng, H. Liang, D. Zheng, R. Shafiiha, X. Zheng, G. Li, K. Raj, A. V. Krishnamoorthy, et al., High speed silicon microring modulator based on carrier depletion, in: Optical Fiber Communication Conference, Optical Society of America, 2010, p. JWA31.

[67] Z. Wang, Y. Gao, A. S. Kashi, J. C. Cartledge, A. P. Knights, Silicon microring modulator for dispersion uncompensated transmission applications, Journal of Lightwave Technology 34 (16) (2016) 3675–3681.

[68] Spectral grids for wdm applications: Dwdm frequency grid, Standard, Telecommunication Standardisation Sector of the International Telecommunication Union, Geneva, CH (Feb. 2012).

[69] Y. London, T. Van Vaerenbergh, L. Ramini, A. J. Rizzo, P. Sun, G. Kurczveil, A. Seyedi, J. Rhim, M. Fiorentino, K. Bergman, Performance requirements for terabit-class silicon photonic links based on cascaded microring resonators, Journal of Lightwave Technology (2019).

[70] A. Asad, A. Dorostkar, F. Mohammadi, A novel power model for future heterogeneous 3d chip-multiprocessors in the dark silicon age, EURASIP Journal on Embedded Systems 2018 (1) (2018) 3.

[71] Y. Demir, N. Hardavellas, Ecolaser: an adaptive laser control for energy-efficient on-chip photonic interconnects, in: 2014 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), IEEE, 2014, pp. 3–8.

[72] P. Dumais, D. J. Goodwill, D. Celo, J. Jiang, C. Zhang, F. Zhao, X. Tu, C. Zhang, S. Yan, J. He, et al., Silicon photonic switch subsystem with 900 monolithically integrated calibration photodiodes and 64-fiber package, Journal of Lightwave Technology 36 (2) (2017) 233–238.

Chapter 6

Paper 3: Power and Energy Efficient Routing forMach-Zehnder Interferometer based Photonic Switches

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

ICS '21, June 14–17, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8335-6/21/06…$15.00
https://doi.org/10.1145/3447818.3460363

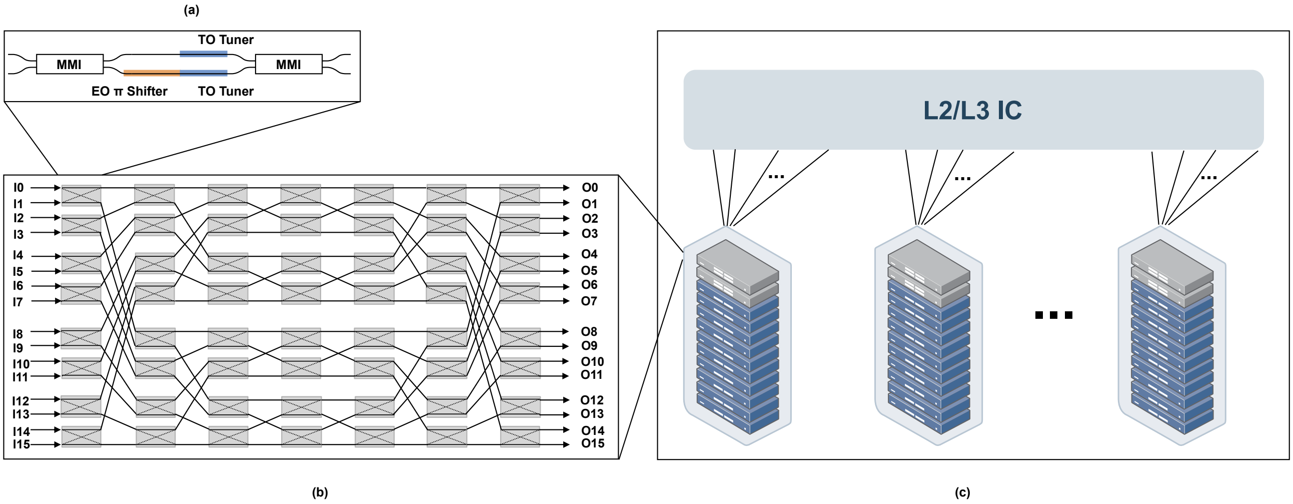


Figure 1: (a) Schematic of a 2×2 EO/TO MZI Switch. (b) A Beneš-based ToR switch formed with MZI switches (c) A hypothetical photonic interconnect with photonic ToR switches.

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.

The Beneš network is a rearrangeably non-blocking topology composed of the least amount of 2×2 switches necessary to connect N×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 ~961W excluding optics or cables and the peak consumption rises to ~1998W 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 ≤0.1W per port in required laser power for traversing the switch (not including coupling losses) [17] combined with an MZI tuning power requirement of ~1W, 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× and 32× 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× 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

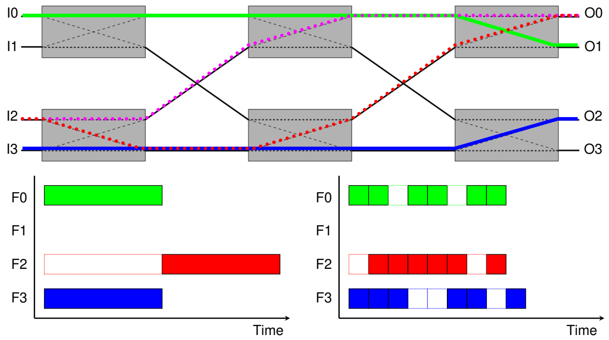


Figure 2: Top: Example of switch fabric contention in a small 4×4 Beneš network. The green, F0, and blue, F3, flows have a path allocated. A third flow, F2, from I2 to O0 arrives but can not be served because resources are busy. Bottom: Timeline of execution using CS (left) and TDM (right). With CS,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

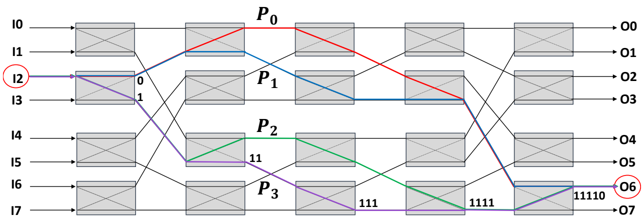
incrementally which means that switch fabric contention for the use of resources can not always be avoided. *Switch fabric contention* is the event in which a flow does not have any available path because resources are busy serving other flows. An example of switch fabric contention is shown in Fig. 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 $O(nlogn)$ 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 $logN$ 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



(a)

ID	Crossings	Bar States	Path	Destination Tag	m_x	m_b	m_bx	m_xb
P_2	3	0	1 0	1 1 0	1	1	1	1
P_0	5	2	0 0	1 1 0	2	2	3	3
P_3	3	2	1 1	1 1 0	1	2	2	2
P_1	5	4	0 1	1 1 0	2	3	4	4

Routing Signature *Scoring Ranks*

(b)

Figure 3: (a) Path diversity in an 8× 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× 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, 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.

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	$1.56\mu s$
200 KB	$3.12\mu s$
500 KB	$7.80\mu s$

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× 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 $2logN - 1$ MZIs per path for N input-output pairs simultaneously requesting access, the controller would have to accommodate $O(N^2(2logN - 1))$ 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× 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

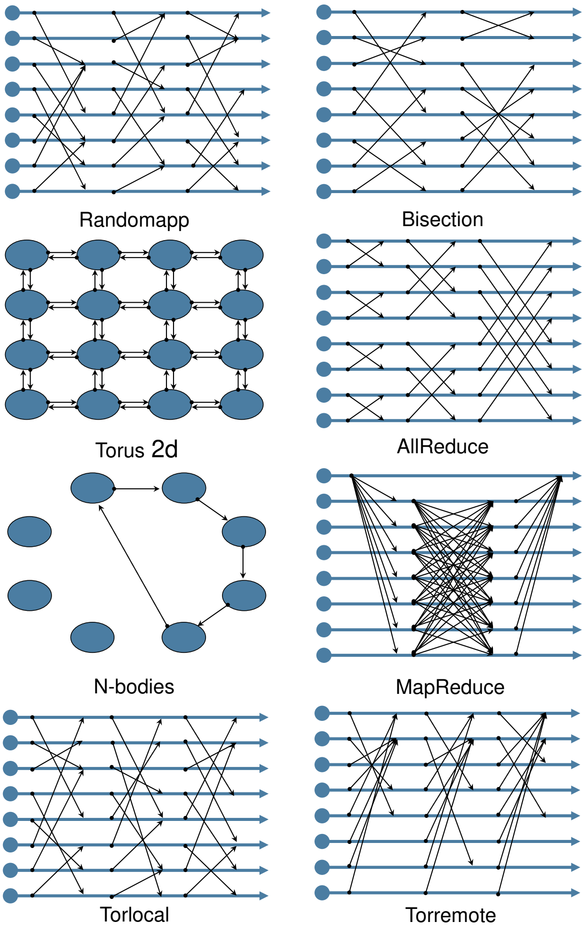


Figure 4: Schematic representation of the eight workloads used, shown with eight endpoints.

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

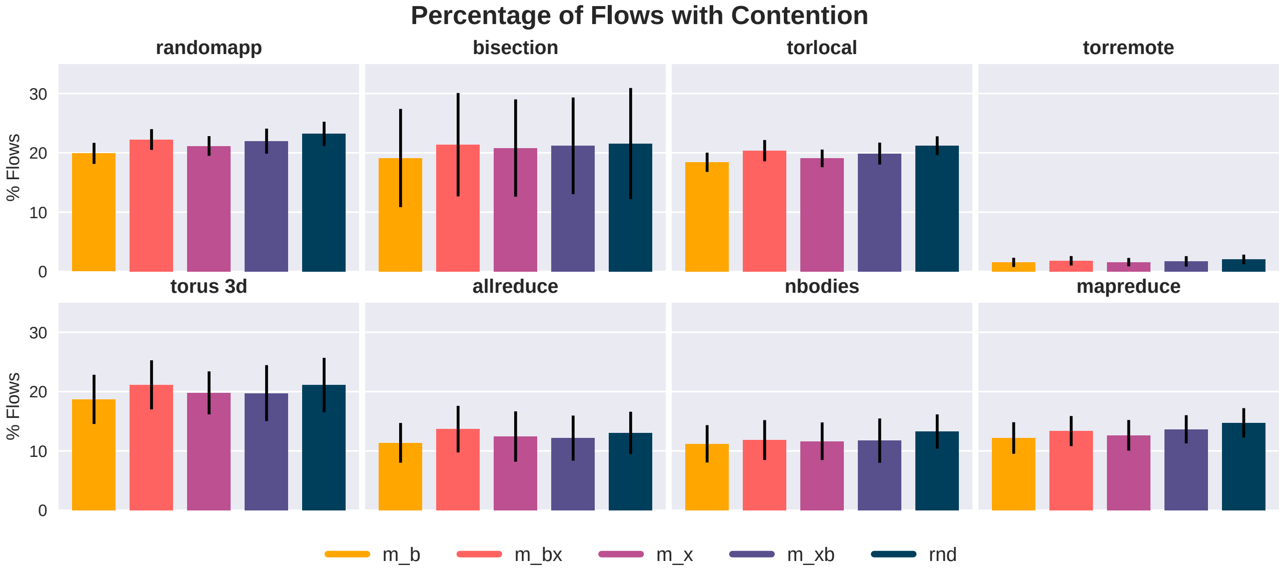


Figure 5: Percentage of flows that suffer switch fabric contention.

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

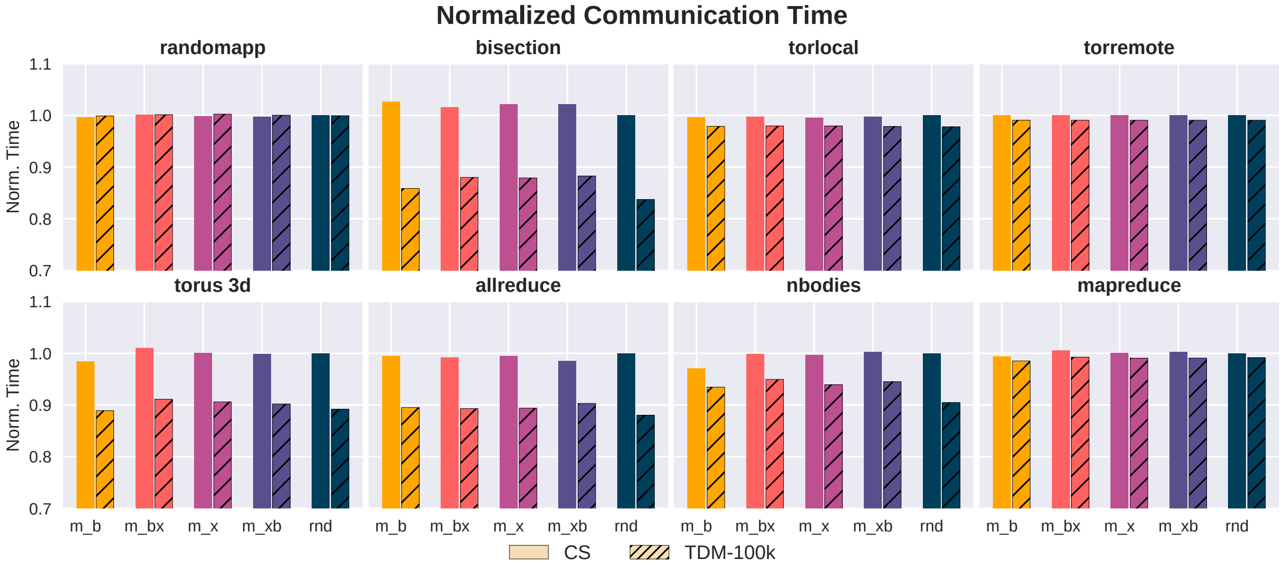


Figure 6: Normalised communication time for CS and TDM.

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.

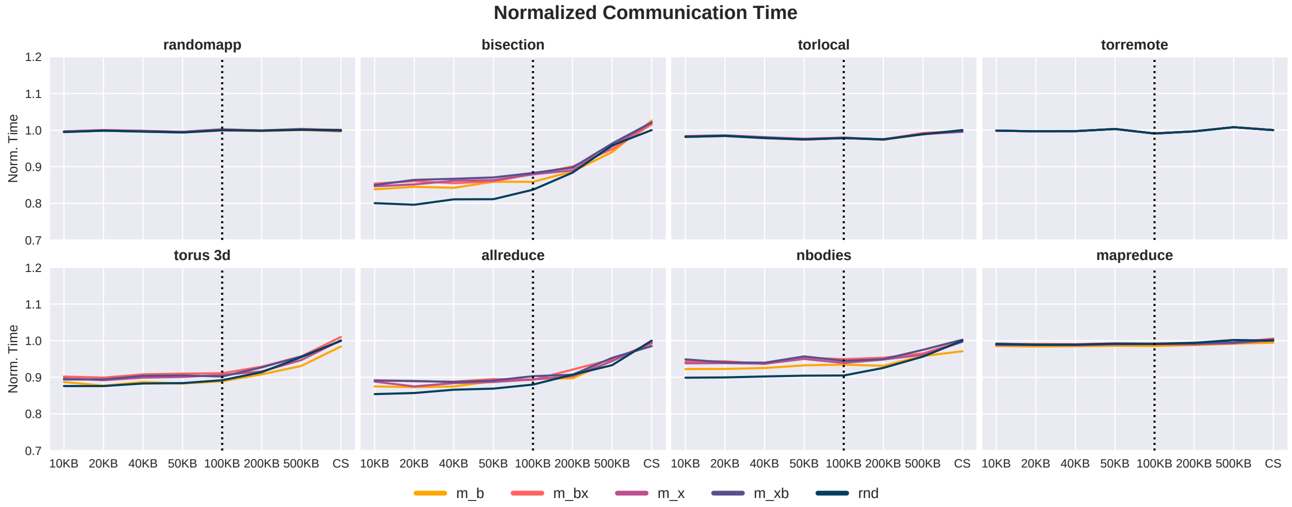


Figure 7: Normalized communication time for various flow segment sizes.

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 μs) compared to CS (29.6-30.4 μs). 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 μs) compared to CS (143.4-147.2 μs). 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 μs) over CS (119.9-121.7 μs) 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 μs) against CS (234.4-242.4 μs) 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

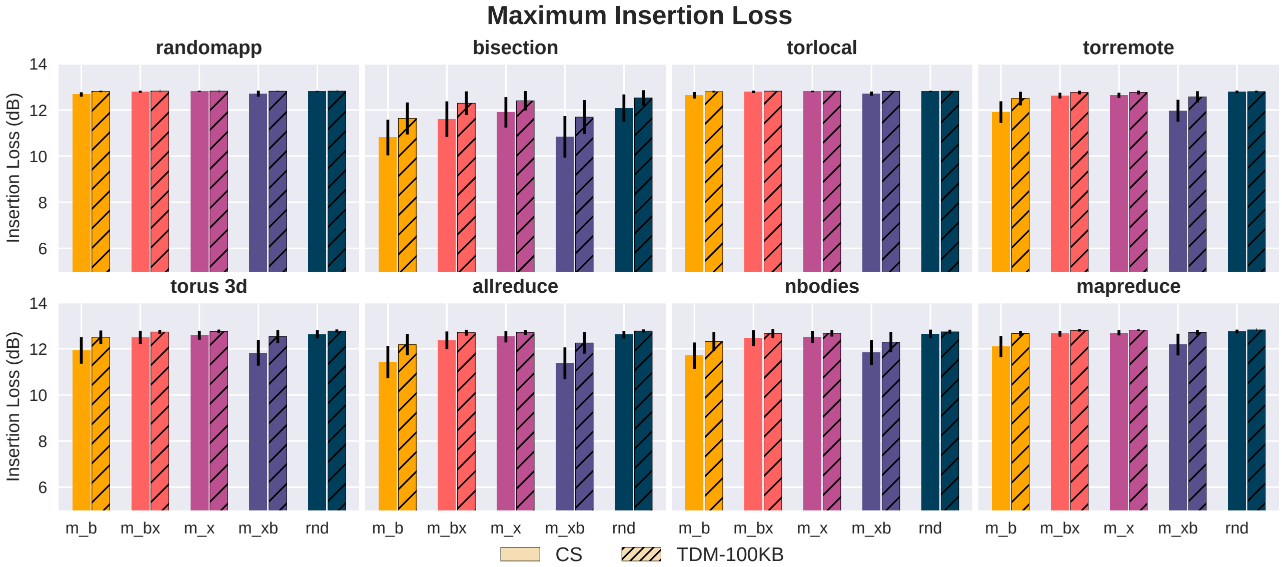


Figure 8: Worst-case exhibited ILoss for circuit-switching and TDM.

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

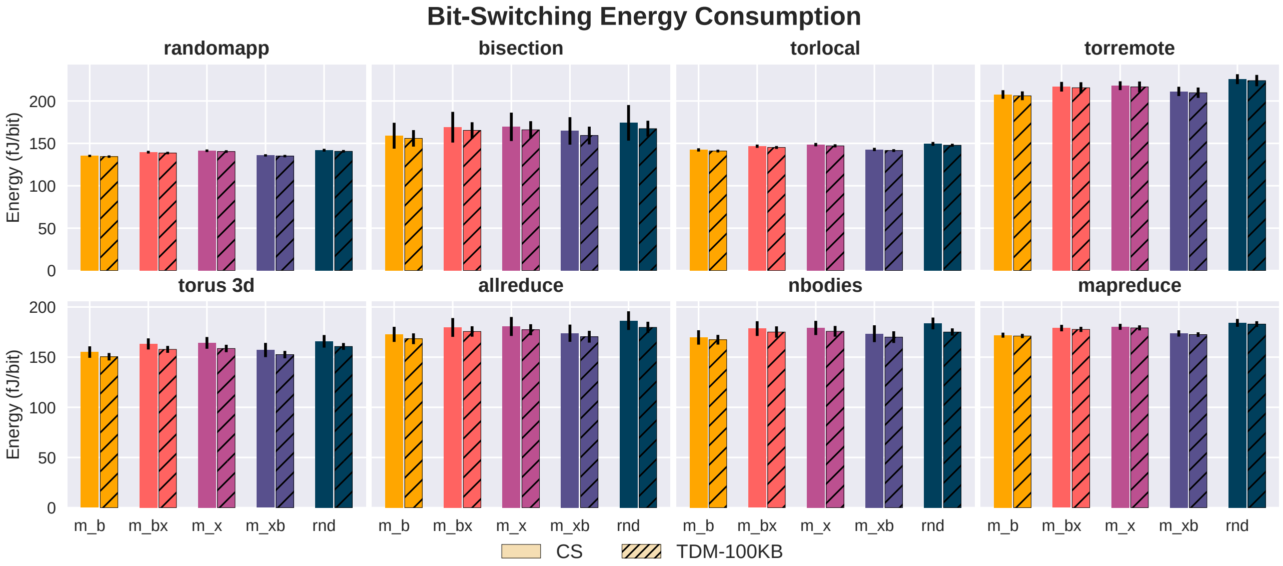


Figure 9: Bit-switching energy consumption for circuit-switching and TDM.

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-ber-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 Syrivelis, 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 Minkenberg 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 Arend, 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

Chapter 7

Paper 4: Understanding the Impact of Arbitration in MZI-based Beneš Switching Fabrics

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 will provide higher bandwidth and energy efficiency for datacenters and high-performance systems than those employing traditional electronic crossbars. While recent works focus on photonic devices, routing and switching capabilities, not enough attention has been dedicated to other practical deployment aspects. As photonic fabrics are intrinsically bufferless, traffic is likely to be subject to contention and, thus, the order in which flows are served has an impact on performance metrics. This is especially the case for rearrangeably non-blocking topologies, such as the Beneš network. We analyse the impact of arbitration on the performance of such fabrics. Our study uses the experimental data from a recently manufactured 16×16 Beneš prototype using Mach-Zehnder Interferometers, but a similar impact can be expected from other architectures. Our evaluation establishes the impact that arbitration policies have on the performance of photonics switches including configurations with three routing algorithms, two switching methods, three ToR switch sizes and 9 representative workloads from the DC and HPC domains. We evaluate five classic arbitration policies and, upon revealing the weaknesses of the round-robin policy, propose two variants. We found that the effect of arbitration on raw throughput is negligible but, when considering realistic loads, selecting an appropriate arbitration policy can improve communication time without sacrificing 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%.

Index Terms—Top-of-Rack Photonic Switches, Arbitration, Mach-Zehnder Interferometers, Performance Evaluation, Simulation.

1 Introduction

Scaling datacenter (DC) and high-performance computing (HPC) networks is a continuous challenge as their communication demands continue to grow. All-optical interconnection networks (ICNs) 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 could provide significant advantages compared with standard electronic crossbars. Photonics technology offers greater data density (approx one order of magnitude) due to coarse- and dense-wavelength division multiplexing (CWDM & DWDM), and can accommodate more bandwidth per link. Photonic ICNs can 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 approaching adoption. A significant step has been the recent introduction of CMOS-compatible photonic devices [2]. However, developing and deploying efficient photonic ICNs is still challenging. 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 widely used because, among all RNB topologies, it requires the fewest 2-port switches and switching stages necessary to connect 2^k endpoints. These characteristics offer both reduced ILoss and wiring complexity compared to other alternatives. When using broadband photonic 2-port switching cells such as Mach-Zehnder Interferometers (MZIs), Beneš PSFs become particularly appealing as a low-loss, low hardware complexity DWDM-enabled solution.

In contrast to electronic based Beneš networks, standard switch control algorithms, such as the *Looping Algorithm* [9], are unable to route network traffic in an energy efficient manner in PSFs. Alternative PSF-focused switch control algorithms [10], [11] can not completely eliminate *fabric contention*. This phenomenon is illustrated in Fig. 2. Here, the $I_0 \rightarrow O_1$ and $I_3 \rightarrow O_2$ transmissions have already been assigned a path. However, if a $I_2 \rightarrow O_0$ transmis-

- *J. Navaridas, JA. Pascual, and J. Miguel-Alonso are with the Department of Computer Architecture and Technology, University of the Basque Country. Corresponding author: javier.navaridas@ehu.eus*
- *M. Kynigos, M. Luján and J. Goodacre are with the Department of Computer Science, University of Manchester.*

This is a manuscript draft that has been submitted to TPDS and is awaiting peer review.

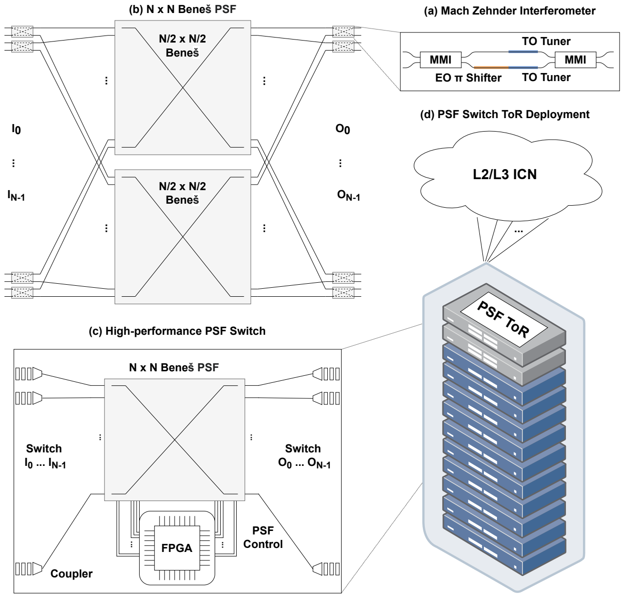


Fig. 1: *(a)* Schematic of a 2 × 2 EO/TO MZI switching element. *(b)* An $N \times N$ MZI Beneš PSF. *(c)* A high-performance switch containing the FPGA-controlled PSF. *(d)* Deployed ToR switch within a DC or HPC rack.

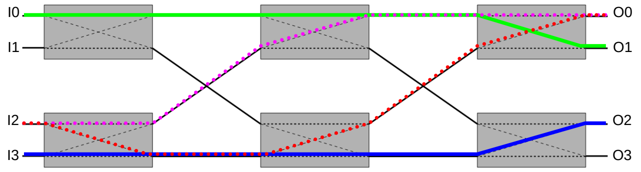
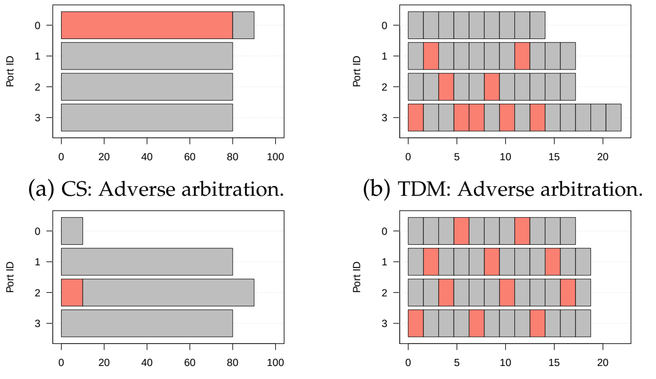


Fig. 2: Fabric contention in a 4×4 Beneš network.



(a) CS: Adverse arbitration.

(b) TDM: Adverse arbitration.

(c) CS: Beneficial arbitration.

(d) TDM: Beneficial arbitration.

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

sion was requested next, it could not be served because all possible paths (dotted lines) require resources that are already-allocated (solid lines). Thus, the incoming transmission must wait until resources are freed or all established connections need to be torn down and the switch fabric reconfigured. Either way, incurring time and throughput penalties. A similar phenomenon is *output contention*, where 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. It shows the arbitration of a 4-port Beneš PSF. Time (μs) flows from left to right and input ports can be transmitting (in grey), blocked (in red) or idle (blank). With CS, if a short flow is blocked by long ones, it incurs 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. Based on the above and to the best of our knowledge, while many works have investigated routing and switching in Beneš PSF-enabled high-performance switches, arbitration has not been investigated.

To address this gap, this paper contributes the first comprehensive study of the impact of arbitration in MZI Beneš-based PSFs (Section 5). In particular, we consider their use within Top of Rack (ToR) switches (as per Fig. 1) and carry out a simulation-based 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 (Section 4). We investigate the use of 5 classical arbitration policies. We identify the weaknesses for round robin, one of the most common arbitration policies, and propose 2 new round robin variants: accelerated round-robin (ARR) and multi-level round-robin (MRR) (Section 3). With this setup, we can analyse the interactions between routing, switching and arbitration in Beneš PSF-enabled ToR switches. In particular we consider four metrics: switch throughput, communication time, ILoss and switching energy per bit. We find that switch throughput under uniform traffic is barely affected by arbitration. In contrast, with application traffic, 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.

2 Background

2.1 Opportunities for photonic switching

Silicon photonic 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 employed 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 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, (b) the recursive Beneš topology in which the fabric is organised, (c) the architecture of a photonic switch based on the Beneš fabric and a controller FPGA and (d) a DC or HPC rack employing a photonic 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 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. BWT in photonic switches affords another advantage; photonic switches can more easily accommodate future data rates, as their performance is less dependent on per-wavelength data rates or number of wavelengths. In contrast, electronic switches must either be upgraded at every data rate generation to support new transceivers, or transceivers must remain constrained by legacy capabilities. Therefore, employing photonic switches based on PSFs can allow infrastructure investment to be amortised over a longer term.

2.2 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 1 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 for fair comparison; where this is not reported, we subtract $radix*optics\_wattage$ from the reported peak power, assuming 20W optics for 400Gb/s, 4.5W for 100Gb/s and 2.5W for 40Gb/s links.

Based on these estimates, switching energy efficiency in commodity electronic switches ranges between 42 and 330 pJ/bit. The most energy efficient and highest bandwidth switch is the MQM9700 by NVIDIA-Mellanox, with 42.4 pJ/bit and 400Gb/s links. However, it 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 characterised in [4]. Such PSF switches would exhibit a very small switching power envelope (between 1.2W and 7.3W for 16 to 64 endpoints). To this we would need to add a network controller, which can be implemented in a Virtex-7 FPGA. Considering the power budgets reported for such devices in [20] we take a pessimistic power envelope of 20W. We assume a deployment scenario with different switch radices (512Gb/s links with 32 wavelengths), as well as a comparative scenario assuming 64 ports and 400Gb/s links similar to the MQM9700. Switches with these characteristics will feature energy per bit figures of between 0.8 and 2.6 pJ/bit. Clearly, the peak switching power and switching energy per bit can be potentially reduced by 1-2 orders of magnitude by adopting PSF switches. This can be highly compelling for photonic ICNs, as their adoption can potentially reduce the total cost of ownership or increase the power budget for other components such as CPUs, I/O, etc.

2.3 MZI-based photonic switching fabrics

MZIs are commonly selected for their ability to exhibit switching behaviour across contiguous segments of the transmission spectrum, which enables BWT. MZIs are interferometric structures composed of two waveguide arms, connected on either side by 2 × 2 3dB couplers or MMI couplers. Light entering the MZI from either input port is split into the two arms. A controllable phase difference is induced by thermal or electrical tuning to define the MZI state, which can be either "cross" or "bar". Based on the phase difference, the light from the two arms constructively, or destructively, interferes inside the second coupler. Thus, the light outputs via the top or bottom port. Tuning principle, tuning application and MZI structure affect the characteristics of a device: switching speed, ILoss, *extinction ratio* (ER), broadband nature and footprint. These in turn, affect the capabilities and achievable size of the PSF.

Tuning principle — The thermo-optic or the electro-optic effect are used for MZI tuning. In the former, a heating element changes the refractive index of the material to induce phase change; this provides low ILoss and a high extinction ratio to the MZI, reducing the attenuation and leakages which lead to photonic crosstalk. Thermo-optic tuning, however, happens at the μs scale, which is too slow for many applications in high-performance switching (e.g., TDM). Electro-optic tuning takes a few ns , but as described by the Kramers-Kronig equations [21], this leads to free-carrier absorption (FCA). Such absorption increases ILoss and reduces the ER. As stated by Lee *et al.* [5], ILoss can be mitigated through amplification, but crosstalk can not. Crosstalk is most detrimental when two interfering light-beams are coherent. The power penalty from coherent crosstalk can limit Beneš PSF scalability.

Tuning application — Tuning can be induced on either one or both MZI arms; the former is referred to as single-ended tuning, the latter as push-pull [22]. In single-ended tuning, the tuning mechanism on one arm must provide the entire π phase shift relative to the light traversing the other MZI arm. In TO tuning this increases the size of the heating element and, in EO tuning, this increases FCA, decreasing ER. Push-pull tuning mitigates this, as both MZI arms provide a $\pi/2$ phase shift, leading to decreased crosstalk penalties with EO

TABLE 1: Power and energy consumption of commercial electronic switches *vs* prototypical photonic switches.

Device Model	Switch Radix	Data Rate	Switching Capacity	Power Dissipation	Switching Energy
CISCO Nexus 3636C-R [12]	36 ports	100 Gb/s	3.6 Tb/s	1,179 Watt	327.5 pJ/bit
Aruba CX 8320 [13]	32 ports	40 Gb/s	1.3 Tb/s	230 Watt	179.7 pJ/bit
Aruba CX 8325 [14]	32 ports	100 Gb/s	3.2 Tb/s	406 Watt	126.9 pJ/bit
CISCO Nexus 3464C [15]	64 ports	100 Gb/s	6.4 Tb/s	712 Watt	111.3 pJ/bit
Huawei CloudEngine 9860 [16]	128 ports	100 Gb/s	12.8 Tb/s	1,051 Watt	82.1 pJ/bit
Arista 7368X4 Series [17]	32 ports	400 Gb/s	12.8 Tb/s	966 Watt	75.5 pJ/bit
NVIDIA MQM9700 [18]	64 ports	400 Gb/s	25.6 Tb/s	1,084 Watt	42.3 pJ/bit
NVIDIA SN2700 [19]	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

tuning. If only TO-tuning is used, it both calibrates the MZI to either state and switches to the complement state. If only EO-tuning is used, the device is initially in the quadrature state, and tuning induces either MZI state. If both tuning options are used, TO-tuning is responsible for calibrating the MZI to either state, and EO-tuning is used to switch.

MZI structure — Most MZI proposals entail an equal arm length and the previously described structure, relying on tuning to calibrate the MZI. Nested MZI switches have also been proposed to reduce the crosstalk in PSFs [23]. These have an increased footprint, higher complexity and a smaller tuning spectral region, but a higher ER leading to less crosstalk. Using these, BWT and DWDM can be achieved by adopting a smaller channel spacing; e.g., 50GHz instead of the 100GHz spacing we assume in this work [24]. Tri-state MZIs have also been proposed [25]. These include a third state to decrease the overall crosstalk power penalty of PSFs, and can lead to practical larger PSF sizes.

MZI-based PSFs — Various proposals for MZI-based PSFs have been produced recently, targeting different levels of the ICN. Li *et al.* [26] and more recently Cheng *et al.* [27] provide comprehensive reviews of silicon photonics for DC interconnects. A large number of the reviewed papers considers Beneš-based switch fabrics with MZIs. MZI-based approaches have also been formulated for the on-chip domain, e.g. [28], [29]. More specifically, MZI-based PSFs organised in either the Beneš or dilated-Beneš topologies have been demonstrated recently [30], [31], [32], [33]. The PSF sizes observed in the literature have progressively increased up to 32 ports, by addressing many technological and manufacturing challenges associated with increasing switch radix.

Based on this trend, BWT-capable, 64×64 Beneš PSFs with $ns-$switching and adequately low crosstalk could appear in a not too far future. Thus, we extend our simulations to 64×64 PSFs. Nonetheless, many challenges, such as providing optimal arbitration and routing algorithms, must be resolved before production deployments can occur. Indeed, while photonic ICN systems have recently been the subject of intensive investigation (e.g., [34], [35], [36], [37], [38], [39]), research on the deployment and practical application of MZI-based switching fabrics is scarce, especially for architectural aspects, such as routing, switching or arbitration.

2.4 Trends in photonic architectures

Recently, the topic of exploiting hardware asymmetries in photonic architectures through routing has gained traction in the community. Cheng *et al.* [10] propose a path mapping strategy for 8-port Beneš fabrics which evaluates all potential states for a permutation and selects the most effective one. Although it is an interesting approach, their brute-force design quickly becomes intractable as the number of ports scales up. Yuen and Chen [40] propose a homologous methodology for exploiting hardware asymmetries. However, they focus on micro ring resonators (MRR) based photonic ICNs instead of MZI-based ICNs, and their proposal does not account for waveguide crossings. Similarly, Kynigos *et al.* [11] propose a collection of routing algorithms which leverage the underlying hardware constraints of MZI-based Beneš PSFs. Yao and Ye [41] propose several routing algorithms, including loss-aware adaptive routing, a priority-based algorithm and a Q-learning-based heuristic routing. 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 maximise throughput and to use bandwidth more effectively. A survey of different approaches can be found in [42]. We highlight several works whose characteristics render them interesting candidates for both HPC and DC use cases. Yang *et al.* describe the *Data Vortex* optical ICN which uses TDM/WDM switching [43]. Yan *et al.* [44] propose an optical ICN using SDM/TDM for intra-DC and WDM for inter-DC traffic. They employ FPGA-based ToR switches that send traffic either through slotted-TDM/Ethernet or optical bandwidth variable transmitters (BVTs). Saleh *et al.* [45] introduce a MRR-based elastic crossbar switch augmented with TDM. Kynigos *et al.* [8] propose leveraging TDM switching, which provides execution time reductions while maintaining low ILoss. None of these works, however, takes into consideration the effects that arbitration may have in the use of resources.

In fact, we should remark that the research on arbitration for photonics is scarce and limited to the optical network-on-chip (ONoC) domain. Werner *et al.* [46] propose a mixed WDM-TDM approach for bus-based ONoCS based on MRRs, incorporating an ad-hoc arbitration scheme to maximise bandwidth utilisation. Hendry *et al.* [47] use mesh switches based on broadband nanophotonic MRRs that, when coupled with an arbitration scheme for TDM, show efficiency gains over both circuit-switched ONoCs and electronic equivalents. This highlights the novel contributions 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 base 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 topologies and device types, and could be applicable to other forms of PSFs.

3 Arbitration

We now describe the technology and the switch architecture our research is focused on, and also the arbitration policies that we consider in our study. As explained above, when the PSF state is set up incrementally, the order in which connections are allocated may lead to some of them being blocked, causing contention. As highlighted in Fig. 3, blocking increases the total communication time of a workload, thus arbitration can impact communication time significantly.

3.1 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. We assume a deployment scenario where these devices operate as ToR switches connected to both servers and the higher tier of the IC. We consider WDM transmission with 32λ working at 16Gb/s data rate using an On-Off Keying (OOK) scheme [48]. This yields a 512Gb/s aggregate bandwidth per port, with endpoints modulating on all λ simultaneously.

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

We consider a centralised 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.

3.2 Arbitration policies

Our experiments consider the following arbitration policies:

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 [49], [50]. 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 blocked port in the round that requested a path. This way, the next round it will have the highest priority and, thus, will be able to transmit. Ideally, this policy should minimize the number of consecutive rounds a port is 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. Hence, we ensure port priority is interleaved across the PSF, obtaining a better spread of input ports than in the baseline RR.

4 Experimental Methodology

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

4.1 Simulation model and workloads

We use INRFlow [51], an open source, light-footprint, highly scalable, flow-level network simulator which we have extended to support PSFs[1]. 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 corresponding to timeslots of a predefined size [42], [43], [44], [45]. In general a shorter timeslot provides better flow interleaving and lower internal

[1] Available at:https://gitlab.com/ExaNeSt/phinrflow

fragmentation, but requires a more frequent reconfiguration of the switch fabric which imposes some delay and throughput penalties. For simplicity, we consider a 100KB slice size ($\approx 1.5\mu s$ timeslot length), which was found to be a reasonable compromise for the available bandwidth and the tuning delay [8]. As it is common practice in DCs [52], 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 modelled as traffic producer/consumer nodes following the dynamics defined by a range of workloads based on representative HPC and DC applications and well-known benchmarks. These workloads include causality among flows, so most applications go through phases of high and low network pressure. Unless otherwise stated, workloads send 5,000 flows and all flows are 1MB 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 optimised, binary implementation of the AllReduce collective [53], widely used in parallel applications from a range of domains. This workload sends a total of $N \cdot logN$ flows.

Bisection (BI) — Tasks perform pair-wise communications swapping pairs randomly every round. This benchmark was introduced in [54] to estimate bisection bandwidth.

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

RandomApp (RA) — Selects the source and destination uniformly at random. This is a typical networking benchmark which is used to stress the IC. According to [56], 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: $D = (S + t) \mod N$. This is akin to the adversarial traffic proposed in [57].

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 [58]. TL considers most traffic as 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 [58].

In the discussions below, we classify these workloads into two distinct categories: in *Regular* workloads, all tasks progress at a similar 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. Tasks are therefore distributed randomly across the network [59], [60].

4.2 Routing algorithms

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

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

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

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 PSF architecture.

4.3 Methodology

We simulate different system configurations consisting of workload, arbitration policy, routing, switching method and network size. Table 2 shows the photonic component simulation parameters. We simulate each configuration 100 times with different random seeds, and gather the mean and standard deviations of the following performance metrics.

Aggregated bandwidth with uniform traffic at full load to measure the effect of arbitration on switch throughput.

Maximum ILoss used to estimate the impact of arbitration policies on laser power.

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.

Given the large number of experiments, our analysis only shows a subset of representative results. The complete set of results is available through an OSF repository[2]. To make comparisons easier, and to isolate the effects of the arbitration policies from other aspects of the architecture, we normalise 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, the arbitration policies that are based on priorities are coloured in different shades of red, the ones based on round-robin are coloured in shades of blue, and the random policy that follows none of these approaches is coloured grey.

5 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 analysing the impact that switch arbitration has on the raw throughput of a switch. Then, we move to experimenting with realistic workloads to provide a deeper understanding of the relation between applications and the various aspects of the switch architecture: routing algorithms, switch radix and switching methods.

5.1 Insertion Loss

We begin by examining the effects of arbitration on max. ILoss. Max. ILoss, i.e. the critical path ILoss in the worst exhibited configuration of the PSF, indicates an upper bound in required laser power. The impact of arbitration on max. ILoss is therefore critical to assess, since any increase induced by arbitration policies would detract from the energy efficiency of the PSF. Fig. 4 depicts the maximum ILoss results for all combinations of routing strategy, arbitration policy, and radix. We only show the results under the bisection workload using CS, which exposes the network to full-saturation conditions. Under these conditions, the impact on max. ILoss is most pronounced, thereby serving as an upper-bound for the workloads considered in this paper. The results show that the impact of arbitration on max. ILoss is negligible, with variations in the metric between arbitration policies well below 1% for all routing strategies and for all radices. In comparison, in previous analyses we found the impact of routing on this metric is usually in the range 10-30% [11]. Furthermore, we can see that the performance of the arbitration policies regarding max. ILoss is highly consistent, as indicated by the very small confidence intervals. The lack of significant variability in the metric indicates that arbitration alone does not lead to the PSF state being set up in less ILoss-optimal configurations.

All these results show that the benefits of arbitration in other examined metrics do not come at the expense of higher max. ILoss or laser power. Furthermore, they suggest that there is no dependency on 2 × 2 switching cell type, meaning that they can be applicable when employing any of the switching cell types discussed in Section 2.3. Lastly, we note that the impact of arbitration on max. ILoss is similarly negligible for all other considered workloads, whose results are not shown for the sake of brevity.

TABLE 2: Simulation parameters.

Component	ILoss
Bar MZI	1.4 dB
Cross MZI	0.4 dB
Wg. Crossing	0.05 dB
Wg. Propag.	1.18 dB/cm

(a) Insertion Loss.

Tuning Type	Power Cons.
Thermal	0-26 mW
Mean, STD	15.725, 6.608
Electrical	3.28-5.88 mW
Mean, STD	5.166, 0.428

(b) Power consumption.

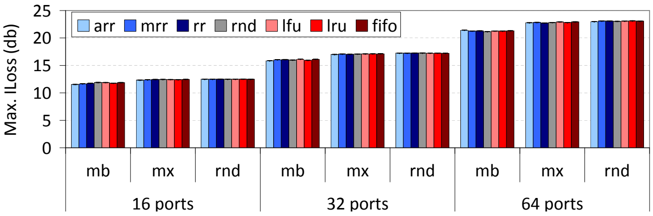


Fig. 4: Critical path ILoss for different configurations – Routing Algorithms

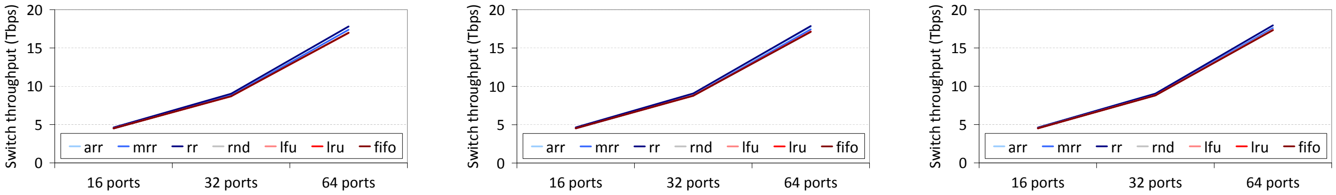
5.2 Aggregated Bandwidth

We move now to investigate the impact that arbitration policies may have on the throughput of photonic switches. In these experiments traffic is generated at maximum load and fabric contention is the only limiting factor. Fig. 5 shows the aggregated bandwidth under uniform traffic supported by the different configurations under study, using CS. As expected, the maximum throughput grows linearly with switch radix: duplicating the number of ports increases the aggregated bandwidth approximately by a factor of two.

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 more significant, but still insubstantial. In particular, switches using policies based on round-robin saturate at a scarcely higher load. The reason for this improvement is a small reduction of switch fabric contention, which suggests that serving ports sequentially might be beneficial.

However, these throughput results only consider uniform traffic, without any concern about the way traffic is generated by real applications. As explained in [8], contention, which is affected by both routing algorithm and arbitration policy, varies depending on the application traffic distribution. The performance of applications with different traffic distributions and causal chains can therefore be affected by arbitration. For this reason, it is essential to carry out a deeper analysis where dependencies between tasks of applications are considered.

[2] Available at: https://osf.io/285d4/?view_only=60d0d30da13e4948a90350b215ac4490



(a) Random Path routing (b) Min. Bar States routing (c) Min. Crossings routing

Fig. 5: Aggregated switch bandwidth with uniform traffic.

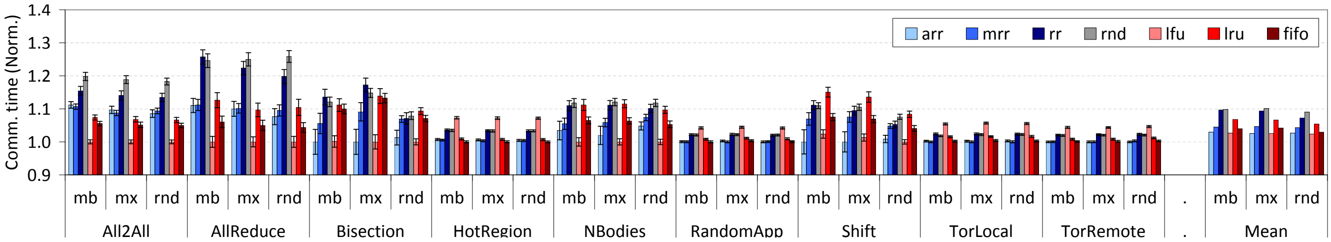


Fig. 6: Normalised communication time for different configurations — Routing algorithms.

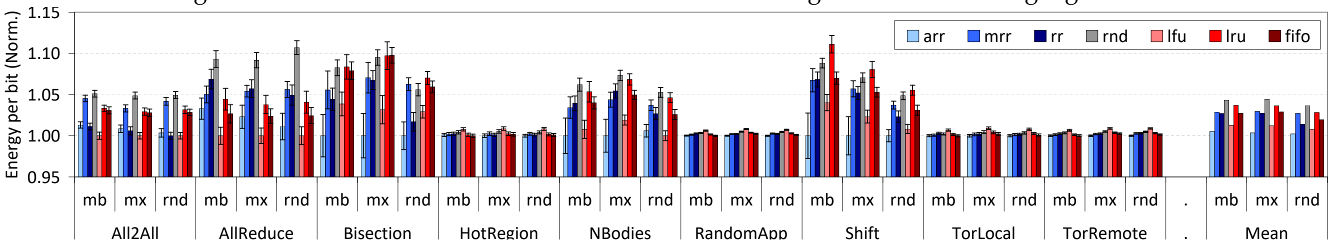


Fig. 7: Normalised energy-per-bit for different configurations — Routing algorithms.

5.3 Interaction with Routing

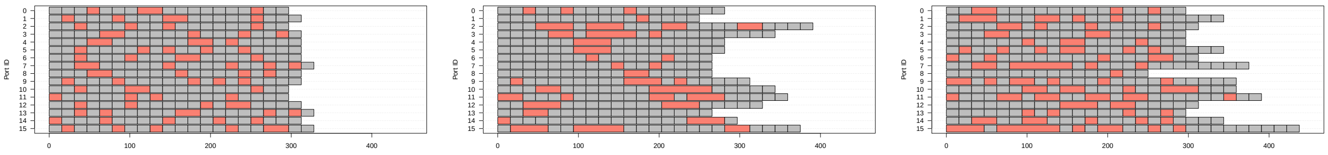
We continue by investigating the interactions between routing algorithms and arbitration policies. Fig. 6 shows the communication time with a 16-port switch using CS and the three routing algorithms. The first thing we notice is that the differences between arbitration policies can be substantial, up to around 30%. This is in stark contrast to the minute differences in terms of aggregated bandwidth studied above and illustrates the need for deeper analysis using application traffic.

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 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 optimising this. For instance, if a task needs to send twice as much traffic as the others, having twice as much bandwidth allocated would be the ideal to ensure it advances at the same pace.

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. 8a 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 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



(a) Least frequently used arbitration (b) Round-robin arbitration (c) Random arbitration

Fig. 8: Timeline of All2All using rnd routing. Time (μs) flows from left to right. Grey: transmitting. Red: blocked.

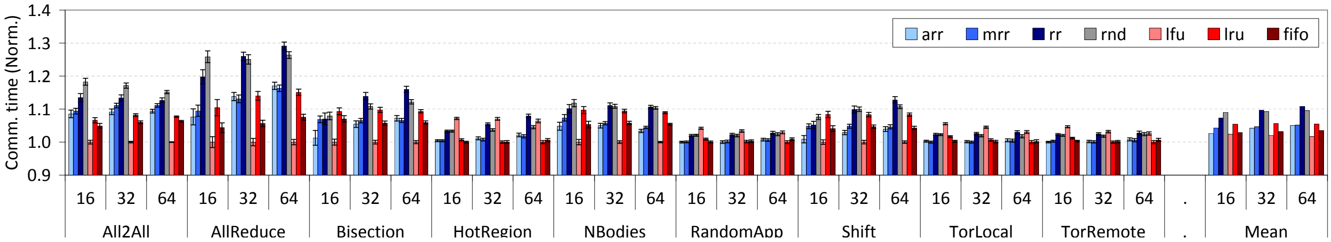


Fig. 9: Normalized communication time for different configurations — Switch radix.

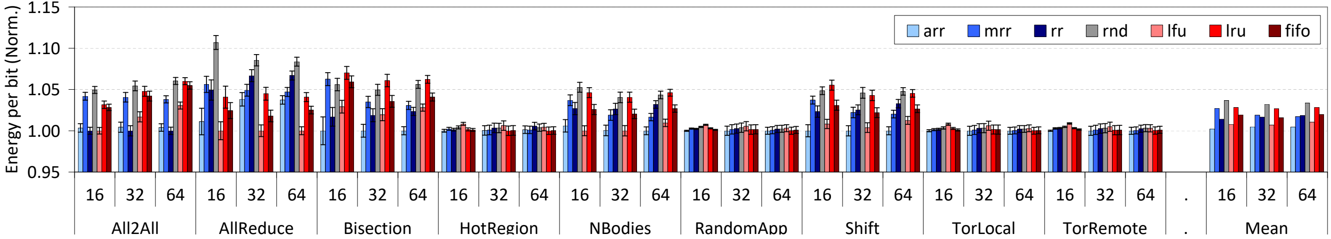


Fig. 10: Normalized energy-per-bit for different configurations — Switch radix.

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. 8b, 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.

Finally, while RND is known 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. 8c, 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.

Moving on to energy efficiency, we have summarised the energy-per-bit results in Fig. 7. There, we can see that the impact of arbitration on energy efficiency is less substantial than it was on communication time, 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. This is reasonable because time is one of the components of energy. 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. These anomalies would require further investigation but, for this paper, 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, suggesting that it is a good candidate when reducing energy consumption is critical.

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. This similarity is somewhat unexpected, as all the tested 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 subsections will concentrate on random path since it features the smallest variability, i.e., it has the tightest confidence intervals.

5.4 Scalability

We now discuss how the performance of the arbitration policies scales with switch radix. Fig. 9 and 10 show the results for communication time and energy-efficiency, respectively, for the investigated switch radices (16, 32 and 64 ports). As explained above, we only present the results for random path but other routing algorithms demonstrate similar behaviour.

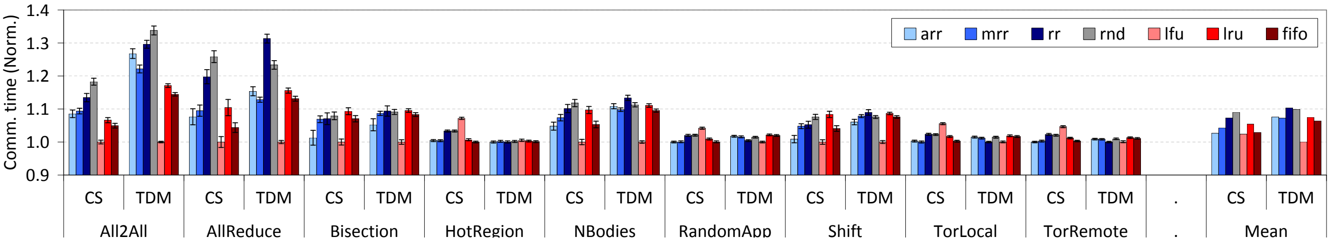


Fig. 11: Normalised communication time for different configurations — Switching methods.

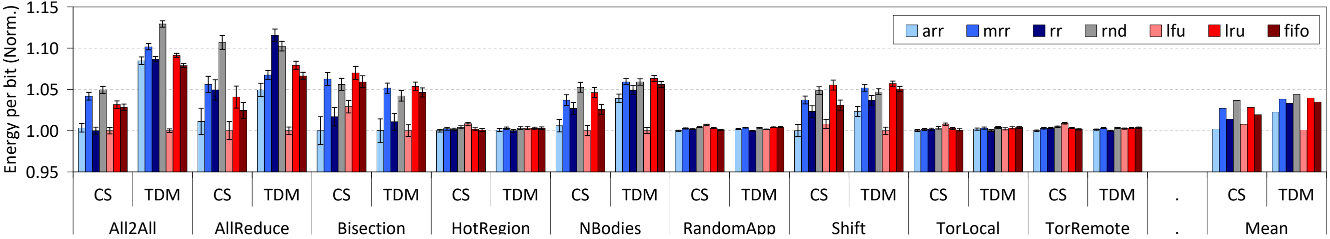


Fig. 12: Normalised energy-per-bit for different configurations — Switching methods.

Firstly, 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 vs. 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 overall 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 be seen for AllReduce, albeit to a lesser extent.

At any rate, it is clear that, in terms of communication time, LFU seems to scale better than the other policies as its lead increases with switch scale. In contrast, RR shows the worst performance scalability, with communication time growing significantly with the switch scale, especially for regular workloads. This is due to the length of the starvation periods which increases with the number of ports; unfairness grows with switch size.

Finally, our two proposed RR variants, i.e. ARR and MRR, scale better than RR in terms of communication time. ARR mitigates the performance penalties of RR for regular workloads, as the communication time grows less with the radix than in the case of RR. With MRR, this scaling penalty is virtually eliminated (except for the AR workload), since the communication time grows negligibly with switch radix, or even reduces slightly in the case of NBodies and Shift.

5.5 Effects of Switching method

Finally, we assess how the performance of arbitration policies changes with the switching method (CS vs. TDM). Fig. 11 and Fig. 12 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 architecture aspects. 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 communication time and energy efficiency. These results pave the way to future research on specific arbitration policies for TDM switches.

6 Conclusions

In this work, we have presented the first comprehensive simulation-based evaluation of the impact of arbitration policies for photonic ToR switches based on MZI Beneš PSFs. Our experimental methodology harnesses the characterisation data of a recently manufactured 16×16 Beneš prototype using MZIs. However, the impact of arbitration would also be a factor for other architectures and devices.

In particular we have evaluated four figures of merit: switch throughput, communication time, insertion loss and energy per bit. We have evaluated five well-established arbitration policies and, upon revealing the weaknesses of the popular RR policy, we have proposed two variants: ARR and MRR. Our 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 behaviour 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 in all cases, which suggests laser power is barely affected by arbitration. This indicates that, arbitration can elicit performance gains without sacrificing energy efficiency. Moreover, arbitration is unaffected by the properties of photonic devices, indicating that our methodology is applicable to Beneš PSFs with a wide range of switch cell design.

Policy-wise, we have found that LFU is the best policy overall, as it is the most effective in terms of performance 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. We also found that 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. ARR, one of our proposals, achieves comparable performance to LFU, but has the best performance with irregular workloads, and consumes the lowest energy in most cases.

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 analyse 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 flavours seem of particular relevance: First, we will investigate priority-based algorithms that prioritise traffic going towards the most heavily loaded ports. A second approach is to apply learning algorithms capable of identifying and prioritising 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.

Acknowledgements

This work has been funded by the EU Horizon2020 (grant 754337); by the Spanish Ministry of Science and Innovation (grant PID2019-104966GB-I00); and by the Basque Country Government (grants KK-2021/00095, IT-1244-19) Dr. Javier Navaridas is a Ramón y Cajal fellow from the Spanish Ministry of Science, Innovation and Universities (RYC2018-024829-I). Prof. Mikel Luján is funded by an Arm/RAEng Research Chair Award and a Royal Society Wolfson Fellowship.

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, Dec. 2017.

[2] D. Thomson *et al.*, "Roadmap on silicon photonics," *Journal of Optics*, vol. 18, no. 7, p. 073003, 2016.

[3] E. Bernier *et al.*, "Switches and routing for on-chip photonic networks," in *24th OptoElectronics and Communications Conf. (OECC) and Intl. Conf. on Photonics in Switching and Computing (PSC)*, 2019.

[4] L. Lu *et al.*, "16×16 optical switch based on electro-optic machzehnder interferometers," *Optics express*, 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," *Journal of Lightwave Technology*, pp. 763–772, 2017.

[6] H. Gu *et al.*, "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 *et al.*, "Technology roadmap for time-division multiplexed passive optical networks (tdm pons)," *Journal of Lightwave Technology*, vol. 37, no. 2, pp. 657–664, 2019.

[8] M. Kynigos *et al.*, "Power and energy efficient routing for machzehnder interferometer based photonic switches," in *Procs. of the ACM Intl. Conf. on Supercomputing*, 2021, pp. 177–189.

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

[10] Q. Cheng, M. Bahadori, and K. Bergman, "Advanced path mapping for silicon photonic switch fabrics," in *Conf. on Lasers and Electro-Optics (CLEO)*, 2017, pp. 1–2.

[11] M. Kynigos *et al.*, "On the routing and scalability of mzi-based optical beneš interconnects," *Nano Communication Networks*, vol. 31, no. 10, 2020.

[12] *Cisco Nexus 3636C-R Switch Data Sheet*, CISCO, 2020. [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/switches/nexus-3000-series-switches/datasheet-c78-740345.html

[13] *Aruba CX 8320 Switch Series*, Aruba Networks, 2020. [Online]. Available: https://www.arubanetworks.com/assets/ds/DS_8320Series.pdf

[14] *Aruba CX 8325 Switch Series*, Aruba Networks, 2020. [Online]. Available: https://www.arubanetworks.com/assets/ds/DS_8325Series.pdf

[15] *Cisco Nexus 3464C Switch Data Sheet*, CISCO, 2020. [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/switches/nexus-3000-series-switches/datasheet-c78-740836.html

[16] *Huawei CloudEngine 9860 Switch Datasheet*, Huawei, 2021. [Online]. Available: https://e.huawei.com/en/material/networking/data-center-network/726347c1089a4591b4993c66f007a571

[17] *Arista 7368X4 Series 100/200/400G Data Center Switches Data Sheet*, Linear Technology, 2020. [Online]. Available: https://www.arista.com/assets/data/pdf/Datasheets/7368X4-Datasheet.pdf

[18] *NVIDIA QM9700 NDR InfiniBand Switch Data Sheet*, NVIDIA, 2021. [Online]. Available: https://nvdam.widen.net/s/tf6tkwsmmn/infiniband-quantum2-datasheet-web

[19] *Mellanox SN2700 Switch Data Sheet*, NVIDIA-Mellanox, 2018. [Online]. Available: https://www.mellanox.com/related-docs/prod_eth_switches/PB_SN2700.pdf

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

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

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

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

[24] "Spectral grids for wdm applications: Dwdm frequency grid," Telecommunication Standardisation Sector of the International Telecommunication Union, Geneva, CH, Standard, Feb. 2012.

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

[26] Y. Li *et al.*, "Silicon and hybrid silicon photonic devices for intra-datacenter applications: state of the art and perspectives," *Photonics Research*, vol. 3, no. 5, pp. B10–B27, 2015.

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

[28] Z. Yunchou *et al.*, "Five-port silicon optical router based on mach—zehnder optical switches for photonic networks-on-chip," *Journal of Semiconductors*, vol. 37, no. 11, p. 114008, 2016.

[29] M. Geng *et al.*, "N-port strictly non-blocking optical router based on mach-zehnder optical switch for photonic networks-on-chip," *Optics Communications*, vol. 383, pp. 472–477, 2017.

[30] L. Qiao, W. Tang, and T. Chu, "16× 16 non-blocking silicon electro-optic switch based on mach-zehnder interferometers," in *Optical Fiber Communication Conf.* Optical Society of America, 2016.

[31] ——, "32× 32 silicon electro-optic switch with built-in monitors and balanced-status units," *Scientific Reports*, vol. 7, pp. 1–7, 2017.

[32] T. Chu *et al.*, "Fast, high-radix silicon photonic switches," in *2018 Optical Fiber Communications (OFC)*. IEEE, 2018, pp. 1–3.

[33] P. Dumais *et al.*, "Silicon photonic switch subsystem with 900 monolithically integrated calibration photodiodes and 64-fiber package," *Journal of Lightwave Technology*, vol. 36, no. 2, 2017.

[34] K. Wen *et al.*, "Flexfly: Enabling a reconfigurable dragonfly through silicon photonics," 11 2016, pp. 166–177.

[35] K. Katrinis *et al.*, "Rack-scale disaggregated cloud data centers: The dredbox project vision," in *Procs. of the 2016 Conf. on Design, Automation & Test in Europe*. EDA Consortium, 2016, pp. 690–695.

[36] C. Minkenberg *et al.*, "Reimagining datacenter topologies with integrated silicon photonics," *J. Opt. Commun. Netw.*, vol. 10, no. 7, pp. 126–139, Jul 2018.

[37] T. Alexoudi *et al.*, "Optics in computing: from photonic network-on-chip to chip-to-chip interconnects and disintegrated architectures," *Journal of Lightwave Technology*, vol. 37, no. 2, 2019.

[38] T. Hirokawa *et al.*, "A wavelength-selective multiwavelength ring-assisted mach–zehnder interferometer switch," *J. Lightwave Technol.*, vol. 38, no. 22, pp. 6292–6298, Nov 2020.

[39] A. S. P. Khope *et al.*, "Scalable multicast hybrid broadband-crossbar wavelength selective switch: proposal and analysis," *Opt. Lett.*, vol. 46, no. 2, pp. 448–451, Jan 2021.

[40] P. Yuen and L. Chen, "Optimization of microring-based interconnection by leveraging the asymmetric behaviors of switching elements," *Journal of lightwave technology*, pp. 1585–1592, 2013.

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

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

[43] Q. Yang *et al.*, "Wdm/tdm optical-packet-switched network for supercomputing," in *Optics in Computing*, 2000, pp. 555–561.

[44] S. Yan *et al.*, "Archon: A function programmable optical interconnect architecture for transparent intra and inter data center sdm/tdm/wdm networking," *Journal of Lightwave Technology*, vol. 33, no. 8, pp. 1586–1595, 2015.

[45] A. Saleh *et al.*, "Elastic wdm switching for scalable data center and hpc interconnect networks," in *OptoElectronics and Communications Conf. (OECC) and Photonics in Switching (PS)*, 2016, pp. 1–3.

[46] S. Werner, J. Navaridas, and M. Luján, "Subchannel scheduling for shared optical on-chip buses," in *2017 IEEE 25th Annual Symposium on High-Performance Interconnects (HOTI)*, 2017, pp. 49–56.

[47] G. Hendry *et al.*, "Silicon nanophotonic network-on-chip using tdm arbitration," in *IEEE Symposium on High Performance Interconnects*, 2010, pp. 88–95.

[48] A. Jain, R. Bahl, and A. Banik, "Demonstration of rz-ook modulation scheme for high speed optical data transmission," 09 2014.

[49] K. Ogawa *et al.*, "Ieee 802.11ah based m2m networks employing virtual grouping and power saving methods," in *2013 IEEE 78th Vehicular Technology Conf. (VTC Fall)*, 2013, pp. 1–5.

[50] L. Chen and N. Chrysos, "Throughput of random arbitration for approximate matchings," in *Procs. of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, ser. ANCS '10, New York, NY, USA, 2010.

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

[52] A. Greenberg *et al.*, "Vl2: A scalable and flexible data center network," in *ACM SIGCOMM 2009 Conference on Data Communication*, ser. SIGCOMM '09, 2009, p. 51–62.

[53] R. Thakur and W. D. Gropp, "Improving the performance of collective operations in mpich," in *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, 2003, pp. 257–267.

[54] X. Yuan *et al.*, "A new routing scheme for jellyfish and its performance with hpc workloads," in *Intl. Conf. on High Performance Computing, Networking, Storage and Analysis*, ser. SC '13, 2013.

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

[56] S. Kandula *et al.*, "The nature of data center traffic: Measurements & analysis," in *9th ACM SIGCOMM Conf. on Internet Measurement*, ser. IMC '09, 2009, p. 202–208.

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

[58] T. Benson, A. Akella, and D. Maltz, "Network traffic characteristics of data centers in the wild," in *Procs. of the 10th ACM SIGCOMM Conf. on Internet measurement*, 2010, pp. 267–280.

[59] S. Zaheer *et al.*, “Locality-aware process placement for parallel and distributed simulation in cloud data centers,” *Journal of Supercomputing*, 08 2019.

[60] R. Fujimoto, “Research challenges in parallel and distributed simulation,” *ACM Trans. Model. Comput. Simul.*, vol. 26, no. 4, 2016.

Chapter 8

Conclusions

8.1 Epilogue

Recent advances in Silicon Photonics have led to the demonstration of medium-scale photonic switching fabrics from the research community [CBG$^+$18]. This paves the way for research into their adoption in future data centres and high-performance computers. However, research into both device-level improvements of photonic switching fabrics and control-level improvements, by means of switch fabric control and routing algorithm, are required before this can be achieved. This thesis contributes to the latter by proposing improvements to the routing and control internal to photonic switching fabrics formed with EO/TO-tuned MZIs organised in the Beneš topology. It also contributes to the state-of-the-art by offering a methodology and tool for simulating the interactions between network traffic configuration, routing algorithm selection and photonic performance in photonic switching fabrics.

This chapter reviews the findings presented in this thesis (Sec. 8.2) and discusses future avenues for study in Sec. 8.3, which are made possible by the work presented here. It then concludes by outlining the significance of the thesis findings within the field of photonic switching fabrics, and offers the author's reflection (Sec. 8.4).

8.2 Summary of Contributions

Before describing the thesis contributions, **Chapter 2** offered a background in the photonic links, devices, and metrics investigated in this thesis. It discussed the most prominent photonic switching fabric topologies that have been adopted, as well as the fabricated chip used as a hardware baseline for the simulation-driven studies of the thesis. The chapter also discussed the differences between photonic and electronic interconnection networks, as well as the most important switching and multiplexing techniques used in photonics, and concluded with the major design challenges entailed in considering large scale photonic switching fabrics.

Chapter 3 proposed a methodology and tool for evaluating the effect of network traffic and routing algorithm selection on the performance of photonic switching fabrics formed with EO/TO-tuned MZIs in the Beneš topology. It discussed the network simulator which has been developed to investigate the methodology, and compared the photonic loss model against two photonic switch fabric chips from the literature. It then demonstrated the methodology by evaluating the effect of routing algorithm choice on insertion loss, crosstalk, combined power penalty , signal power and laser power. This was done by assuming a full-saturation traffic pattern (*bisection*) under 1000 different traffic configurations, and comparing the metrics as exhibited when using the state-of-the-art “Looping Algorithm” [Ben64] with those exhibited when using the “hardware-inspired routing strategies” proposed in the later thesis chapters. It showed that the routing algorithm choice can have a significant impact on the considered photonic metrics, with laser power being reduced by $\sim 75\%$ on average and $\sim 42\%$ in the worst case, for a switch fabric with 16×16 endpoints when using the best routing strategy from Chapter 4. It then demonstrated that using the best routing strategy from Chapter 5 shows slight comparative savings only in laser power on average, with other strategies exhibiting lower savings against the “Looping Algorithm”. The chapter concluded by discussing the future scope for improvement in photonic crosstalk, by partitioning the available spectrum into “λ-groups” and employing them in conjunction with the routing strategies.

Chapter 4 evaluated the limitations of scaling out a thermally/electrically tuned MZI-based optical Beneš network. Three hardware-inspired routing strategies were presented; these strategies aim to leverage the asymmetric behaviours of internal switching elements, as well as the path-based asymmetry offered by the network topology, to reduce insertion loss and switching energy. It was shown that these strategies always reduce both the average and the maximum insertion loss exhibited by the workload flows. Maximising the number of MZIs in “cross” state reduced max. ILoss by 32% in the best case (Bisection, 64 endpoints). The laser power analysis showed a substantial laser power reduction with the best routing strategy, ranging from 33% to 85% depending on the number of endpoints. Also, from the point of insertion loss, increasing the fabric size above 256 endpoints was shown to be untenable.

Chapter 5 expanded the analysis of Chapter 4 and proposed a collection of hybrid routing strategies. The impact of the component devices on insertion loss relative to switch size was investigated; for more than 128 endpoints, impact of waveguide crossings surpasses that of MZIs. The routing strategies from Chapter 4 underwent an extended evaluation, with respect to maximum insertion loss, switching energy consumption and communication time. The experiments used 8 synthetic workloads and the routing strategies were compared against random path selection, which was considered to be the baseline. The analysis showed that *m_b* can reduce maximum ILoss by 30.6% in the best case (sweep2, 64 endpoints), while *m_c* shows a 33% best-case reduction over the absolute maximum (sweep2, 256 endpoints). Minimising the amount of state changes within the network is less effective at reducing max. ILoss, but more impactful on switching energy consumption. Furthermore, routing strategy hybridisation via the combination of minimization criteria, was shown to be beneficial, with the *m_xb* hybrid consistently reducing max. insertion loss by more than the single-criteria counterparts and offering the highest savings (35.5%, 64 endpoints, sweep2).

The trade-off between energy consumption from switches and lasers was also investigated. The routing strategies also demonstrated switching energy savings, with the best one offering between 8% and 15% savings compared to the baseline, depending on the communication workload. The investigation of routing strategy impact on communication time found that in the worst case, execution time increases by at most 3% compared to the baseline and, in some cases, is decreased very significantly (5-25%). Finally, the applicability of the switching fabric to the on-chip domain was investigated in terms of the impact of insertion loss on the power budget. It was shown that when considering only insertion loss, a network of 32 endpoints can be suitable for the on-chip domain using a conservative modulator; a substantial laser power reduction was demonstrated with the best routing strategy (23-85% across the network sizes).

Chapter 6 proposed combining the energy efficient routing strategies with TDM to form a control mechanism for the 16×16 photonic Beneš switch fabric formed with thermally-electrically tuned MZIs, when deployed as a ToR switch. The approach was evaluated through simulation, employing eight realistic and synthetic workloads from the DC and HPC domains. Switch fabric contention between communication flows was investigated when using circuit-switching 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%). The impact of the proposed approach on communication time was then evaluated, 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. A parameter sweep was conducted 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, the impact of TDM on worst-case path-dependent insertion loss and bit-switching energy consumption was assessed and found to be small (0.5-1 dB increase and 1-4% decrease respectively), if not slightly beneficial in the case of switching energy.

Finally, **Chapter 7** investigated a collection of classical and novel arbitration policies for Beneš-based photonic switches implementing a subset of the routing schemes proposed in Chapters 4-5. The arbitration policies were assessed for their effect on communication time, insertion loss and switching energy per bit. The results revealed that the effect of arbitration policies is consistent across routing algorithms and switch radices, with performance differences among arbitration policies slightly increasing with size. Conversely, a closer relation between arbitration and switching was found, as the behaviour of the tested arbitration policies differed between TDM and CS configurations. With TDM, arbitration policies were generally more effective for balanced workloads, at the expense of being less effective for unstructured workloads, when compared with CS. It was also shown that communication time varied most depending on the chosen policy, with differences among policies around 10% and, in some outliers, over 30%. The impact on energy efficiency was less significant, with most differences being below 5% and a few cases maxing at around 15%. Finally, the impact on ILoss was found to be insignificant, suggesting that signal power is barely affected by arbitration. Policy-wise, LFU was found to be the best-performing policy overall. It particularly excelled with structured traffic but performed poorly with unstructured workloads. RR, one of the most common arbitration policies used in networking, resulted in most cases in poor performance metrics. This was identified to be due to the standard indexing, which creates short periods of starvation in all of the nodes. To circumvent this issue, two new policies were proposed. One of them, ARR, achieved slightly lower performance figures than LFU, but had the best performance with unstructured workloads and featured the lowest switching energy consumption in most cases.

8.3 Future Work

- Broadening the scope of simulation to include inter-channel crosstalk and device loss profiles for multiple device types.

Although the simulation framework that was augmented for this thesis has enabled the preliminary analysis of photonic metrics and the effect of network traffic and routing algorithms on these for MZI-Beneš networks, it is far from complete. Currently, device losses and crosstalk ratios are included as absolute maxima around the considered wavelength region. Although this design decision abstracts the details of the device layer offers simplicity to simulation and decouples it from requiring spectrum measurements from a physical chip, it currently limits both the simulation capacity and the accuracy compared to real devices. Including device loss and crosstalk profiles for the wavelength region based on measured data from devices, as well as including their calculation within the light propagation model of PhINRFlow, would both increase the accuracy of the simulation and allow for the modelling of spectrum allocation. Assuming those, the simulator could potentially be used for the rapid analysis of the behaviour of planar Beneš-based MZI switching fabrics by device designers at arbitrary switching states. This could help designers identify the impact of device faults as well as trade-offs between components. Designers would also then be able to gain a first-order estimation of the switch fabric behaviour under different traffic scenarios, as has been conducted in preliminary fashion in Chapter 3. Another similar fruitful direction would be the inclusion of different device types, such as nested MZI structures or MRR-based switches, thereby broadening the scope of research into photonic Beneš switching fabrics.

- Spectrum partitioning for crosstalk reduction and integration with routing strategies.

Following up on the previous, Chapter 3 showed that there is scope for improvement in reducing the impact of coherent crosstalk on the switching fabric. One option for this is to consider partitioning the available spectrum into λ-groups and assigning specific groups to paths based on assignment policies, and mitigate the reduction in aggregate bandwidth by increasing the data rate. These assignment policies can then be examined in tandem with the routing strategies. Due to the integration of the DSENT laser model in PhINRFLow, this could then allow for an end-to-end evaluation of signal power and therefore laser power requirements.

- Modelling additional switch topologies.

As detailed in Chapter 2, Beneš is one of various topologies that have been adopted for the creation of planar switching fabrics. Including additional topologies that are based on 2×2 switch elements would enable research into the behaviour and comparison of different switching fabrics, when exposed to various workloads.

- Implementing the routing algorithms and control structure in an FPGA.

One of the main contributions in this work, which is the combination of TDM and HIRs described in Chapter 6, has discussed the trade-offs involved in designing the control structures for these switching fabrics. Implementing these control structures in an FPGA or ASIC would significantly deepen the analysis of these trade-offs, allowing for research into optimizations related to their implementation and the deployment of prototypes, assuming access to a photonic switching fabric chip.

8.4 Concluding Remarks

8.4.1 Reflection on Photonic Switching Fabrics

This thesis has investigated the application of photonic switching fabrics based on EO/TO-tuned MZIs in future DCs and HPCs. The merits of MZIs (broadband, data-rate transparent, fast-switching) over a subset of photonic switches (MRRs, AWGRs) have been discussed. Coherent crosstalk, insertion loss and switching energy has been investigated in Beneš -based EO/TO-tuned MZI switching fabrics, as well as various routing algorithms which can be used for these networks. Based on these, this section reflects on the viability of these specific switching fabrics for different levels of the interconnection network of a future high-performance system, and discusses potential areas of improvement on the technology.

MZI-based switching fabrics are unlikely to be viable for the on-chip domain, regardless of topology. The advantage that MZIs can offer over MRRs is their flexibility with respect to spectral bandwidth, that is that they can provide a relatively uniform behaviour over a contiguous segment of the spectrum; MRRs on the other hand resonate with one or multiple individual wavelengths in the spectrum, based on their radius and group index. This advantage means that MZI-based ONoCs could be particularly applicable to DWDM, thereby achieving much higher aggregate data rates compared to MRR-based ONoCs. However, MZIs are very large compared to MRRs (1 order of magnitude difference in length), meaning that large ONoCs formed with MZIs would be challenging in terms of chip footprint. Also, for them to be fast-switching, MZIs need to employ electro-optic tuning; contrary to MRR-based switches, EO-tuned MZIs suffer very adverse impacts from FCA, which increases their insertion loss and decreases their extinction ratio, leading to crosstalk. As has been demonstrated in Chapter 3, this can lead to prohibitive power penalties even for relatively few endpoints (32×32) fabrics. As detailed in Chapter 5 however, MRR-based switching fabrics have been demonstrated with much higher endpoint counts. In fact, while Chapter 5 found that 32×32 fabrics employing $32 - \lambda$ DWDM could be applicable on-chip in terms of insertion loss, Chapter 3 found that when considering crosstalk this ceases to be realistic. However, the findings of Chapter 3 indicated that this may become so if a bandwidth partitioning scheme were to be employed.

MZI-based switching fabrics are more likely to be viable when deployed as network switches within DCs and HPCs, either at the top-of-rack or at higher network levels. In these scenarios, the physical footprint of the devices would be less prohibitive, allowing for larger endpoint counts. However, photonic losses and crosstalk in particular would continue to pose a challenge which, ultimately, must be met at the level of photonic device design. However, the analysis in this thesis has shown that it is possible to mitigate the effects of trade-offs made at device design (e.g. higher insertion loss and crosstalk traded off for low switching time) through the use of these devices, i.e. through design choices in routing, switching or arbitration.

In terms of routing and control, Chapters 3- 6 have discussed the major contributions of the thesis, which are the HIRs and their combination with TDM. Chapter 3 compared these to the “Looping Algorithm” with respect to the photonic metrics of interest. While the “Looping Algorithm” exploits the symmetry of the Beneš network to provide rearrangeably-non-blocking connectivity, the HIRs do not. This can be considered to be a limitation of the HIRs, although as discussed in Chapter 6, the “Looping Algorithm” would also present challenges for bufferless photonic networks.

In summary, while Beneš-based photonic switching fabrics formed with EO/TO-tuned MZIs have their drawbacks, they pose an interesting new technology choice for the domain of high-performance switches in DCs and HPCs.

8.4.2 Author’s Reflection

On a personal note, I will reflect here on the way that I have undertaken the doctoral programme which has culminated in this thesis.

I have found the study of PINs to be a profound and humbling challenge. Throughout my research, I strived to bring together concepts from the domains of Silicon Photonic devices, computer architecture and network design. Although I would not formally characterise my thesis as interdisciplinary, researching the nascent field of PINs requires, in my opinion, a more holistic approach that draws concepts from the aforementioned fields. I found that balancing these concepts and studying their interrelated effects was a challenging endeavour.

With hindsight, I believe I would have benefitted with respect to this thesis from a more formal study of optical physics and the device layer in the initial phase of the doctoral programme. This could have allowed me to conduct more rigorous modelling in the first years of the programme leading to an increased research quality. Following from this, it would have been more beneficial to focus more on the beam propagation model in the first years, rather than on publishing my first paper. Lastly, it would have benefitted me to conduct a research collaboration earlier in the programme and, obviously, under different conditions than the ones mandated by the pandemic-related restrictions. Even so, the collaboration significantly deepened my understanding of silicon photonics and was, ultimately, a growing experience for which I am grateful.

Bibliography

[AE15] Anders SG Andrae and Tomas Edler. On global electricity usage of communication technology: trends to 2030. *Challenges*, 6(1):117–157, 2015.

[AKP20] Theoni Alexoudi, George Theodore Kanellos, and Nikos Pleros. Optical RAM and integrated optical memories: a survey. *Light: Science & Applications*, 9(1):1–16, 2020.

[AZS$^+$08] Dimitris Apostolopoulos, Panagiotis Zakynthinos, Leontios Stampoulidis, Efstratios Kehayas, Rob McDougall, Robert A. Harmon, Alistair J. Poustie, Graeme D. Maxwell, Ruth Van Caenegem, Didier Colle, Mario Pickavet, Eduward Tangdiongga, Harmen J. S. Dorren, and Hercules Avramopoulos. Contention resolution for burst-mode traffic using integrated SOA-MZI gate arrays and self-resetting optical flip-flops. *IEEE Photonics Technology Letters*, 20(24):2024–2026, 2008.

[Bah18] Meisam Bahadori. *Physical Layer Modeling and Optimization of Silicon Photonic Interconnection Networks.* Columbia University, 2018.

[BAM10] Theophilus Benson, Aditya Akella, and David A. Maltz. Network traffic characteristics of data centers in the wild. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 267–280, 2010.

[BC18] Wim Bogaerts and Lukas Chrostowski. Silicon photonics circuit design: Methods, tools and challenges. *Laser & Photonics Reviews*, 12(4):1700237, 2018.

[BCB$^+$14] Keren Bergman, Luca P. Carloni, Aleksandr Biberman, Johnnie Chan, and Gilbert Hendry. *Photonic Network-on-Chip Design.* Springer, 2014.

[BDL$^+$07] Aleksandr Biberman, Po Dong, Benjamin G Lee, Justin D Foster, Michal Lipson, and Keren Bergman. Silicon microring resonator-based broadband comb switch for wavelength-parallel message routing. In *LEOS 2007-IEEE Lasers and Electro-Optics Society Annual Meeting Conference Proceedings*, pages 474–475. IEEE, 2007.

[Ben64] Václad E. Beneš. Optimal rearrangeable multistage connecting networks. *Bell system technical journal*, 43(4):1641–1656, 1964.

[Ber21] Keren Bergman. Energy efficient multi-terabit photonic connectivity for disaggregated computing. Presented in the Photonics in Switching and Computing (PSC), 2021.

[BGH$^+$17] Ya-Qing Bie, Gabriele Grosso, Mikkel Heuck, Marco M Furchi, Yuan Cao, Jiabao Zheng, Darius Bunandar, Efren Navarro-Moratalla, Lin Zhou, Dmitri K Efetov, et al. A MoTe$_2$-based light-emitting diode and photodetector for silicon photonic integrated circuits. *Nature nanotechnology*, 12(12):1124–1129, 2017.

[BGL$^+$20] Joshua L. Benjamin, Thomas Gerard, Domaniç Lavery, Polina Bayvel, and Georgios Zervas. Pulse: Optical circuit switched data center architecture operating at nanosecond timescales. *Journal of Lightwave Technology*, 38(18):4906–4921, 2020.

[BHG$^+$18] Daniel J. Blumenthal, Rene Heideman, Douwe Geuzebroek, Arne Leinse, and Chris Roeloffzen. Silicon nitride in silicon photonics. *Proceedings of the IEEE*, 106(12):2209–2231, 2018.

[Bor13] Shekhar Borkar. Exascale computing-a fact or affliction? *Keynote presentation at IPDPS*, 10, 2013.

[BP20] Brad Booth and David Piehler. System aspects for optical interconnect transceivers. In *Springer Handbook of Optical Networks*, pages 779–793. Springer, 2020.

[CBB$^+$18] Kari Clark, Hitesh Ballani, Polina Bayvel, Daniel Cletheroe, Thomas Gerard, Istvan Haller, Krzysztof Jozwik, Kai Shi, Benn Thomsen, Philip Watts, Hugh Williams, Georgios Zervas, Paolo Costa, and Zhixin Liu. Sub-nanosecond clock and data recovery in an optically-switched data centre network. In *2018 European Conference on Optical Communication (ECOC)*, pages 1–3, 2018.

[CBG$^+$18] Qixiang Cheng, Meisam Bahadori, Madeleine Glick, Sébastien Rumley, and Keren Bergman. Recent advances in optical technologies for data centers: a review. *Optica*, 5(11):1354–1370, 2018.

[CC14] Amitabha Chakrabarty and Martin Collier. $O(log\bar{m}.logN)$ routing algorithm for $(2logN-1)$-stage switching networks and beyond. *Journal of Parallel and Distributed Computing*, 74(10):3045–3055, 2014.

[CHB$^+$10] Johnnie Chan, Gilbert Hendry, Aleksandr Biberman, Keren Bergman, and Luca P Carloni. PhoenixSim: A simulator for physical-layer analysis of chip-scale photonic interconnection networks. In *2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010)*, pages 691–696. IEEE, 2010.

[Clo53] Charles Clos. A study of non-blocking switching networks. *Bell System Technical Journal*, 32(2):406–424, 1953.

[CLW$^+$16] Siming Chen, Wei Li, Jiang Wu, Qi Jiang, Mingchu Tang, Samuel Shutts, Stella N. Elliott, Angela Sobiesierski, Alwyn J. Seeds, Ian Ross, et al. Electrically pumped continuous-wave III-V quantum dot lasers on silicon. *Nature Photonics*, 10(5):307–311, 2016.

[CRBB18] Qixiang Cheng, Sébastien Rumley, Meisam Bahadori, and Keren Bergman. Photonic switching in high performance datacenters. *Optics express*, 26(12):16022–16043, 2018.

[CVF$^+$19] Papichaya Chaisakul, Vladyslav Vakarin, Jacopo Frigerio, Daniel Chrastina, Giovanni Isella, Laurent Vivien, and Delphine Marris-Morini. Recent progress on Ge/SiGe quantum well optical modulators, detectors, and emitters for optical interconnects. In *Photonics*, volume 6, page 24. Multidisciplinary Digital Publishing Institute, 2019.

[DCY20] Fei Duan, Kai Chen, and Yonglin Yu. High-speed and low-power thermally tunable devices with suspended silicon waveguide. *Optical and Quantum Electronics*, 52(1):1–10, 2020.

[DHL16] Jack Dongarra, Michael A Heroux, and Piotr Luszczek. High-performance conjugate-gradient benchmark: A new metric for ranking high-performance computing systems. *The International Journal of High Performance Computing Applications*, 30(1):3–10, 2016.

[DL17] Nicolas Dupuis and Benjamin G Lee. Impact of topology on the scalability of mach–zehnder-based multistage silicon photonic switch networks. *Journal of Lightwave Technology*, 36(3):763–772, 2017.

[DLP03] Jack J Dongarra, Piotr Luszczek, and Antoine Petitet. The LINPACK benchmark: past, present and future. *Concurrency and Computation: practice and experience*, 15(9):803–820, 2003.

[DLR$^+$15a] Nicolas Dupuis, Benjamin G Lee, Alexander V Rylyakov, Daniel M Kuchta, Christian W Baks, Jason S Orcutt, Douglas M Gill, William MJ Green, and Clint L Schow. Design and fabrication of low-insertion-loss and low-crosstalk broadband 2×2 mach–zehnder silicon photonic switches. *Journal of Lightwave Technology*, 33(17):3597–3606, 2015.

[DLR$^+$15b] Nicolas Dupuis, Benjamin G Lee, Alexander V Rylyakov, Daniel M Kuchta, Christian W Baks, Jason S Orcutt, Douglas M Gill, William MJ Green, and Clint L Schow. Modeling and characterization of a non-blocking 4×4 mach–zehnder silicon photonic switch fabric. *Journal of Lightwave Technology*, 33(20):4329–4337, 2015.

[DOG$^+$12] Kapil Debnath, Liam O'Faolain, Frederic Y Gardes, Andreas G Steffan, Graham T Reed, and Thomas F Krauss. Cascaded modulator architecture for WDM applications. *Optics Express*, 20(25):27420–27428, 2012.

[DOJ$^+$16] Guang-Hua Duan, Segolene Olivier, Christophe Jany, Stéphane Malhouitre, Alban Le Liepvre, Alexandre Shen, Xavier Pommarede, Guillaume Levaufre, Nils Girard, Dalila Make, et al. Hybrid III-V silicon photonic integrated circuits for optical communication applications. *IEEE Journal of Selected Topics in Quantum Electronics*, 22(6):379–389, 2016.

[Don20] Jack Dongarra. Report on the fujitsu fugaku system. *University* *of Tennessee-Knoxville Innovative Computing Laboratory, Tech. Rep. ICLUT-20-06*, 2020.

[DRS+16] Nicolas Dupuis, Alexander V Rylyakov, Clint L Schow, Daniel M Kuchta, Christian W Baks, Jason S Orcutt, Douglas M Gill, William MJ Green, and Benjamin G Lee. Ultralow crosstalk nanosecond-scale nested 2× 2 mach–zehnder silicon photonic switch. *Optics letters*, 41(13):3002–3005, 2016.

[DT04] William James Dally and Brian Patrick Towles. *Principles and Practices of Interconnection Networks*. Elsevier, 2004.

[FPR+10] Nathan Farrington, George Porter, Sivasankar Radhakrishnan, Hamid Hajabdolali Bazzaz, Vikram Subramanya, Yeshaiahu Fainman, George Papen, and Amin Vahdat. Helios: A hybrid electrical/optical switch architecture for modular data centers. In *Proceedings of the ACM SIGCOMM 2010 Conference*, pages 339–350, 2010.

[FWP+19] Pouya Fotouhi, Sebastian Werner, Roberto Proietti, Xian Xiao, and S. J. Ben Yoo. Enabling scalable disintegrated computing systems with awgr-based 2.5d interconnection networks. *J. Opt. Commun. Netw.*, 11(7):333–346, Jul 2019.

[GEE94] EL Goldstein, L Eskildsen, and AF Elrefaie. Performance implications of component crosstalk in transparent lightwave networks. *IEEE Photonics Technology Letters*, 6(5):657–660, 1994.

[GKL+21] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Chasing carbon: The elusive environmental footprint of computing. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 854–867. IEEE, 2021.

[GLM+11] Michael Georgas, Jonathan Leu, Benjamin Moss, Chen Sun, and Vladimir Stojanović. Addressing link-level design tradeoffs for integrated photonic interconnects. In *2011 IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–8. IEEE, 2011.

[GLPY17] Paolo Grani, Gengchen Liu, Roberto Proietti, and S. J. Ben Yoo. Bit-parallel all-to-all and flexible awgr-based optical interconnects. In *Optical Fiber Communication Conference*, page M3K.4. Optica Publishing Group, 2017.

[GLZ+17] Zhanzhi Guo, Liangjun Lu, Linjie Zhou, Lin Shen, and Jianping Chen. 16× 16 silicon optical switch based on dual-ring-assisted mach–zehnder interferometers. *Journal of Lightwave Technology*, 36(2):225–232, 2017.

[HCB18] Yishen Huang, Qixiang Cheng, and Keren Bergman. Advanced control for crosstalk minimization in MZI-based silicon photonic switches. In *2018 IEEE Optical Interconnects Conference (OI)*, pages 17–18. IEEE, 2018.

[HCE12] Wim Heirman, Trevor Carlson, and Lieven Eeckhout. Sniper: Scalable and accurate parallel multi-core simulation. In *8th International Summer School on Advanced Computer Architecture and Compilation for High-Performance and Embedded Systems (ACACES-2012)*, pages 91–94. High-Performance and Embedded Architecture and Compilation Network of . . . , 2012.

[HH90] Pierre A Humblet and Walid M Hamdy. Crosstalk analysis and filter optimization of single-and double-cavity fabry-perot filters. *IEEE Journal on selected areas in communications*, 8(6):1095–1107, 1990.

[ITU12] Spectral grids for WDM applications: DWDM frequency grid. Standard, Telecommunication Standardisation Sector of the International Telecommunication Union, Geneva, CH, February 2012.

[Jon19] Kevan Jones. Enabling technologies for in-router DWDM interfaces for intra-data center networks. In *Optical Fiber Communication Conference*, pages M1F–1. Optical Society of America, 2019.

[KAM+14] Abbas Karimi, Kiarash Aghakhani, Seyed Ehsan Manavi, Faraneh Zarafshan, and SAR Al-Haddad. Introduction and analysis of optimal routing algorithm in beneš networks. *Procedia Computer Science*, 42:313–319, 2014.

[KB12] Joohwa Kim and James F Buckwalter. A 40-Gb/s optical transceiver front-end in 45 nm SOI CMOS. *IEEE Journal of Solid-State Circuits*, 47(3):615–626, 2012.

[KNS$^+$14] Eiichi Kuramochi, Kengo Nozaki, Akihiko Shinya, Koji Takeda, Tomonari Sato, Shinji Matsuo, Hideaki Taniyama, Hisashi Sumikura, and Masaya Notomi. Large-scale integration of wavelength-addressable all-optical memories on a photonic crystal chip. *Nature Photonics*, 8(6):474–481, 2014.

[KSG$^+$09] Srikanth Kandula, Sudipta Sengupta, Albert Greenberg, Parveen Patel, and Ronnie Chaiken. The nature of data center traffic: Measurements & analysis. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pages 202–208, 2009.

[KT12] Christoforos Kachris and Ioannis Tomkos. A survey on optical interconnects for data centers. *IEEE Communications Surveys & Tutorials*, 14(4):1021–1036, 2012.

[LB10] Di Liang and John E Bowers. Recent progress in lasers on silicon. *Nature photonics*, 4(8):511–517, 2010.

[LBSD$^+$09] Benjamin G Lee, Aleksandr Biberman, Nicolás Sherwood-Droz, Carl B Poitras, Michal Lipson, and Keren Bergman. High-speed 2 × 2 switch for multiwavelength silicon-photonic networks–on-chip. *Journal of Lightwave Technology*, 27(14):2900–2907, 2009.

[LCM$^+$17] Zeqin Lu, Dritan Celo, Hamid Mehrvar, Eric Bernier, and Lukas Chrostowski. High-performance silicon photonic tri-state switch based on balanced nested mach-zehnder interferometer. *Scientific reports*, 7(1):1–7, 2017.

[LD18] Benjamin G Lee and Nicolas Dupuis. Silicon photonic switch fabrics: Technology and architecture. *Journal of Lightwave Technology*, 37(1):6–20, 2018.

[Lee16] Jong-Moo Lee. Athermal silicon photonics. *Silicon Photonics III*, pages 83–98, 2016.

[LLS^{+}18] ZG Lu, JR Liu, CY Song, J Weber, Y Mao, SD Chang, HP Ding, PJ Poole, PJ Barrios, D Poitras, S. Janz, and M. O'Sullivan. High performance InAs/InP quantum dot 34.462-GHz c-band coherent comb laser module. *Optics express*, 26(2):2160–2167, 2018.

[LOC^{+}14] Lian-Wee Luo, Noam Ophir, Christine P Chen, Lucas H Gabrielli, Carl B Poitras, Keren Bergmen, and Michal Lipson. WDM-compatible mode-division multiplexing on a silicon chip. *Nature communications*, 5(1):1–7, 2014.

[LRARK17] R. R. LaPierre, M. Robson, K. M. Azizur-Rahman, and P. Kuyanov. A review of III–V nanowire infrared photodetectors and sensors. *Journal of Physics D: Applied Physics*, 50(12):123001, 2017.

[LT94] Chien-Chun Lu and Richard A Thompson. The double-layer network architecture for photonic switching. *Journal of Lightwave Technology*, 12(8):1482–1489, 1994.

[LVVR^{+}19a] Yanir London, Thomas Van Vaerenbergh, Luca Ramini, Anthony J Rizzo, Peng Sun, Geza Kurczveil, Ashkan Seyedi, Jinsoo Rhim, Marco Fiorentino, and Keren Bergman. Performance requirements for terabit-class silicon photonic links based on cascaded microring resonators. *Journal of Lightwave Technology*, 38(13):3469–3477, 2019.

[LVVR^{+}19b] Yanir London, Thomas Van Vaerenbergh, Anthony J Rizzo, Peng Sun, Jared Hulme, Geza Kurczveil, Ashkan Seyedi, Binhao Wang, Xiaoge Zeng, Zhihong Huang, et al. Energy efficiency analysis of comb source carrier-injection ring-based silicon photonic link. *IEEE Journal of Selected Topics in Quantum Electronics*, 26(2):1–13, 2019.

[LZZ^{+}16] Liangjun Lu, Shuoyi Zhao, Linjie Zhou, Dong Li, Zuxiang Li, Minjuan Wang, Xinwan Li, and Jianping Chen. 16× 16 non-blocking silicon optical switch based on electro-optic mach-zehnder interferometers. *Optics express*, 24(9):9295–9307, 2016.

[LZZP15] Yu Li, Yu Zhang, Lei Zhang, and Andrew W Poon. Silicon and hybrid silicon photonic devices for intra-datacenter applications: state of the art and perspectives. *Photonics Research*, 3(5):B10–B27, 2015.

[MKK$^+$10] Jason E Miller, Harshad Kasture, George Kurian, Charles III Gruenwald, Nathan Beckmann, Christopher Celio, Jonathan Eastep, and Anant Agarwal. Graphite: A distributed parallel simulator for multicores. In *HPCA-16 2010 The Sixteenth International Symposium on High-Performance Computer Architecture*, pages 1–12. IEEE, 2010.

[MST$^+$19] George Michelogiannakis, Yiwen Shen, Min Yee Teh, Xiang Meng, Benjamin Aivazi, Taylor Groves, John Shalf, Madeleine Glick, Manya Ghobadi, Larry Dennison, and Keren Bergman. Bandwidth steering in HPC using silicon nanophotonics. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–25, 2019.

[MWT$^+$19] George Michelogiannakis, Jeremiah Wilke, Min Yee Teh, Madeleine Glick, John Shalf, and Keren Bergman. Challenges and opportunities in system-level evaluation of photonics. In *Metro and Data Center Optical Networks and Short-Reach Links II*, volume 10946, page 1094607. International Society for Optics and Photonics, 2019.

[Net17] Cisco Visual Networking. The zettabyte era–trends and analysis. Technical report, Cisco white paper, 2017.

[NFA13] Christopher J Nitta, Matthew K Farrens, and Venkatesh Akella. On-chip photonic interconnects: A computer architect's perspective. *Synthesis Lectures on Computer Architecture*, 8(5):1–111, 2013.

[NFF$^+$18] Radhakrishnan Nagarajan, Mark Filer, Yang Fu, Masaki Kato, Todd Rope, and James Stewart. Silicon photonics-based 100 Gbit/s,PAM4, DWDM data center interconnects. *Journal of Optical Communications and Networking*, 10(7):B25–B36, 2018.

[NMAPR11] Javier Navaridas, Jose Miguel-Alonso, Jose A Pascual, and Francisco J Ridruejo. Simulating and evaluating interconnection networks with INSEE. *Simulation Modelling Practice and Theory*, 19(1):494–515, 2011.

[NNLBX17] Gabriela Nicolescu, Mahdi Nikdast, Sébastien Le Beux, and Jiang Xu. *Photonic Interconnects for Computing Systems: Understanding and Pushing Design Challenges*. River Publishers, 2017.

[NPE+19] Javier Navaridas, Jose A Pascual, Alejandro Erickson, Iain A Stewart, and Mikel Luján. INRFlow: An interconnection networks research flow-level simulation framework. *Journal of parallel and distributed computing*, 130:140–152, 2019.

[NRC+15] Dessislava Nikolova, Sébastien Rumley, David Calhoun, Qi Li, Robert Hendry, Payman Samadi, and Keren Bergman. Scaling silicon photonic switch fabrics for data center interconnection networks. *Optics express*, 23(2):1159–1175, 2015.

[(OL21] Oak Ridge Leadership Computing Facility (OLCF). Frontier, direction of discovery, 2021. [Online; accessed 9-November-2021].

[OTW71] David C Opferman and Nelson T Tsao-Wu. On a class of rearrangeable switching networks part i: Control algorithm. *The Bell System Technical Journal*, 50(5):1579–1600, 1971.

[PC20] Nick Parsons and Nicola Calabretta. Optical switching for data center networks. In *Springer Handbook of Optical Networks*, pages 795–825. Springer, 2020.

[PGGH10] Subodh Prabhu, Boris Grot, P Gratz, and Jiang Hu. Ocin_tsim - DVFS aware simulator for NoCs. *Proc. SAW*, 1, 2010.

[PKK+09] Yan Pan, Prabhat Kumar, John Kim, Gokhan Memik, Yu Zhang, and Alok Choudhary. Firefly: Illuminating future network-on-chip with nanophotonics. In *Proceedings of the 36th annual international symposium on Computer architecture*, pages 429–440, 2009.

[PL+16] Lorenzo Pavesi, David J. Lockwood, et al. Silicon photonics III. *Topics in applied physics*, 122:1–36, 2016.

[PN87] Krishnan Padmanabhan and Arun Netravali. Dilated networks for photonic switching. *IEEE Transactions on Communications*, 35(12):1357–1365, 1987.

[QTC17] Lei Qiao, Weijie Tang, and Tao Chu. 32× 32 silicon electro-optic switch with built-in monitors and balanced-status units. *Scientific Reports*, 7(1):1–7, 2017.

[RB91] Cauligi S. Raghavendra and Rajendra V. Boppana. On self-routing in beneš and shuffle-exchange networks. *IEEE Transactions on Computers*, 40(9):1057–1064, 1991.

[RBW$^+$16] Sébastien Rumley, Meisam Bahadori, Ke Wen, Dessislava Nikolova, and Keren Bergman. PhoenixSim: Crosslayer design and modeling of silicon photonic interconnects. In *Proceedings of the 1st International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems*, pages 1–6, 2016.

[RNH$^+$15] Sébastien Rumley, Dessislava Nikolova, Robert Hendry, Qi Li, David Calhoun, and Keren Bergman. Silicon photonics for exascale systems. *Journal of Lightwave Technology*, 33(3):547–562, Feb 2015.

[RSS09] Rajiv Ramaswami, Kumar Sivarajan, and Galen Sasaki. *Optical Networks: a Practical Perspective*. Morgan Kaufmann, 2009.

[SB87] Ron A. Spanke and Václad E Beneš. N-stage planar optical permutation network. *Applied Optics*, 26(7):1226–1229, 1987.

[SB07] Assaf Shacham and Keren Bergman. Building ultralow-latency interconnection networks using photonic integration. *IEEE Micro*, 27(4):6–20, 2007.

[SCK$^+$12] Chen Sun, Chia-Hsin Owen Chen, George Kurian, Lan Wei, Jason Miller, Anant Agarwal, Li-Shiuan Peh, and Vladimir Stojanovic. DSENT-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling. In *2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip*, pages 201–210. IEEE, 2012.

[SHM87] T Shimoe, K Hajikano, and K Murakami. Path-independent insertion loss optical space switch. In *Optical Fiber Communication Conference*, page WB2. Optical Society of America, 1987.

[Spa86] Ron A. Spanke. Architectures for large nonblocking optical space switches. *IEEE Journal of quantum electronics*, 22(6):964–967, 1986.

[SRM$^+$19] Rick Stevens, Jini Ramprakash, Paul Messina, Michael Papka, and Katherine Riley. Aurora: Argonne's next-generation exascale supercomputer. Technical report, ANL (Argonne National Laboratory (ANL), Argonne, IL (United States)), 2019.

[SXH$^+$15] Harish Subbaraman, Xiaochuan Xu, Amir Hosseini, Xingyu Zhang, Yang Zhang, David Kwong, and Ray T. Chen. Recent advances in silicon-based passive and active optical interconnects. *Optics express*, 23(3):2487–2511, 2015.

[SZC$^+$21] Ashkan Seyedi, Segev Zarkovsky, Shai Cohen, Paraskevas Bakopoulos, and Giannis Patronas. Novel interconnects for next-gen ai systems. Presented in the Photonics in Switching and Computing (PSC), 2021.

[TAM$^+$19] Apostolos Tsakyridis, Theoni Alexoudi, Amalia Miliou, Nikos Pleros, and Christos Vagionas. 10 Gb/s optical random access memory (RAM) cell. *Optics letters*, 44(7):1821–1824, 2019.

[TFH13] Han Wui Then, Milton Feng, and Nick Holonyak. The transistor laser: Theory and experiment. *Proceedings of the IEEE*, 101(10):2271–2298, 2013.

[TLN$^+$18] Christos A Thraskias, Eythimios N Lallas, Niels Neumann, Laurent Schares, Bert J Offrein, Ronny Henker, Dirk Plettemeier, Frank Ellinger, Juerg Leuthold, and Ioannis Tomkos. Survey of photonic and plasmonic interconnect technologies for intra-datacenter and high-performance computing communications. *IEEE Communications Surveys & Tutorials*, 20(4):2758–2783, 2018.

[TOT96] Hiroshi Takahashi, Kazuhiro Oda, and Hiromu Toba. Impact of crosstalk in an arrayed-waveguide multiplexer on $n \times n$ optical interconnection. *Journal of lightwave technology*, 14(6):1097–1105, 1996.

[VSM$^+$08] Dana Vantrease, Robert Schreiber, Matteo Monchiero, Moray McLaren, Norman P Jouppi, Marco Fiorentino, Al Davis, Nathan Binkert, Raymond G Beausoleil, and Jung Ho Ahn. Corona: System implications of emerging nanophotonic technology. *ACM SIGARCH Computer Architecture News*, 36(3):153–164, 2008.

[WAD+17] Zhechao Wang, Amin Abbasi, Utsav Dave, Andreas De Groote, Sulakshna Kumari, Bernadette Kunert, Clement Merckling, Marianna Pantouvaki, Yuting Shi, Bin Tian, et al. Novel light source integration approaches for silicon photonics. *Laser & Photonics Reviews*, 11(4):1700063, 2017.

[WF80] Chuan-Lin Wu and Tse-Yun Feng. On a class of multistage interconnection networks. *IEEE transactions on Computers*, 100(8):694–702, 1980.

[WFP+18] Sebastian Werner, Pouya Fotouhi, Roberto Proietti, Xian Xiao, and SJ Ben Yoo. Towards energy-efficient high-throughput photonic NoCs for 2.5D integrated systems: A case for AWGRs. In *2018 Twelfth IEEE/ACM International Symposium on Networks-on-Chip (NOCS)*, pages 1–8. IEEE, 2018.

[WG17] Xiaomu Wang and Xuetao Gan. Graphene integrated photodetectors and opto-electronic devices—a review. *Chinese Physics B*, 26(3):034203, 2017.

[WGK+16] Zhao Wang, Yuliang Gao, Aazar S Kashi, John C Cartledge, and Andrew P Knights. Silicon microring modulator for dispersion uncompensated transmission applications. *Journal of Lightwave Technology*, 34(16):3675–3681, 2016.

[WHD14] Jian Wang, Sailing He, and Daoxin Dai. On-chip silicon 8-channel hybrid (de) multiplexer enabling simultaneous mode-and polarization-division-multiplexing. *Laser & Photonics Reviews*, 8(2):L18–L22, 2014.

[WNL17] Sebastian Werner, Javier Navaridas, and Mikel Luján. A survey on optical network-on-chip architectures. *ACM Computing Surveys (CSUR)*, 50(6):1–37, 2017.

[WVM+17] Zhechao Wang, Kasper Van Gasse, Valentina Moskalenko, Sylwester Latkowski, Erwin Bente, Bart Kuyken, and Gunther Roelkens. A III-V-on-Si ultra-dense comb laser. *Light: Science & Applications*, 6(5):e16260–e16260, may 2017.

[WWX$^+$16] Zhifei Wang, Zhehui Wang, Jiang Xu, Peng Yang, Luan Huu Kinh Duong, Zhe Wang, Haoran Li, and Rafael Kioji Vivas Maeda. Low-loss high-radix integrated optical switch networks for software-defined servers. *Journal of Lightwave Technology*, 34(18):4364–4375, 2016.

[XLDD$^+$18] Yelong Xu, Jiachuan Lin, Raphaël Dubé-Demers, Sophie LaRochelle, Leslie Rusch, and Wei Shi. A single-laser flexible-grid WDM silicon photonic transmitter using microring modulators. In *Optical Fiber Communication Conference*, pages W1I–3. Optical Society of America, 2018.

[ZCC$^+$96] Jingyu Zhou, Roberto Cadeddu, Emilio Casaccia, Carlo Cavazzoni, and Michael J O'Mahony. Crosstalk in multiwavelength optical cross-connect networks. *Journal of lightwave technology*, 14(6):1423–1435, 1996.

[ZLZ$^+$16] Shuoyi Zhao, Liangjun Lu, Linjie Zhou, Dong Li, Zhanzhi Guo, and Jianping Chen. 16×16 silicon mach–zehnder interferometer switch actuated with waveguide microheaters. *Photonics Research*, 4(5):202–207, 2016.

[ZXS$^+$17] Zhendong Zhang, Yiyuan Xie, Tingting Song, Chao He, Jiachao Li, and Yong Liu. Exploring crosstalk noise generated in the n-port router used in the WDM-based ONoC. *Optical Engineering*, 56(7):076112, 2017.

[ZYM15] Zhiping Zhou, Bing Yin, and Jurgen Michel. On-chip light sources for silicon photonics. *Light: Science & Applications*, 4(11):e358–e358, 2015.