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

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.
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.

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. 

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.
We aim to understand the scalability implications relating to ILoss which controls MZI states and receives path requests from the network. This 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” to “right.”
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 \( \lambda \)), as well as that of the routing strategies. We then use the best data rate per \( \lambda \) 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.
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 \( \lambda \). 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 32\( \lambda \) 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 \( \lambda \) 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/\( \lambda \) does not increase laser energy consumption substantially (<1%), whereas increasing from 8 to 16 Gbps/\( \lambda \) increases consumption by approx. 33%. Consequently, the least energy consumption from lasers and switching is afforded at 8Gbps/\( \lambda \), which for 32\( \lambda \) 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 \( \lambda \). 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 73W to be...
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 behaviors 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 (Bi-section, 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


