Wavelength Reconfigurable Transceiver For Multi-Interface Compute Accelerator Networks

Zhenguo Wu[1], Robert Parsons[1], Songli Wang[1], Yuyang Wang[1], and Keren Bergman[1]

[1] *Department of Electrical Engineering, Columbia University, 500 W 120th St., New York, New York, USA. 10027*

zw2542@columbia.edu

Abstract: We present a multi-port reconfigurable silicon photonic transceiver for flexible bandwidth reallocation in multi-interface architectures. We demonstrate on-chip wavelength reconfiguration on a optical testbed and show 94% job completion time improvement in large-scale network simulations. © 2024 The Author(s)

1. Introduction

To enhance system reliability and collective communication efficiency, current datacenter, HPC and AI accelerator hardware solutions are incorporating multiple communication interfaces into servers or computing units (CUs). These multi-interface networks efficiently support distributed deep learning (DDL) tasks by enabling parallel gradient exchange among multiple CUs [1]. Notable examples of such networks include Alibaba's Super Leaf Spine (connecting each server to two ToRs) [2], Nvidia's DGX GH200 (connecting each Grace Hopper chip to three NVSwitches) [3], the SiP-OCS architecture (connecting each GPU to multiple optical circuit switches (OCS)) [4], and the SiPAC architecture (connecting each CU to multiple wavelength selective switches (WSS)) [5].

In this work, we introduce a wavelength-reconfigurable multi-port silicon photonic transceiver architecture for dynamic bandwidth allocation across multiple network interfaces. We demonstrate a feasible implementation of this reconfigurable architecture using cascaded stages of balanced Mach-Zehnder Interferometers (MZI) and broadband optical interleavers. We validate the wavelength reconfigurability in a testbed experiment, demonstrating wavelength routing to either a single output port or uniform distribution across two output ports. System-level simulations show a maximum of 94% job completion time reduction in multi-tenant DDL workloads compared to the original configuration without bandwidth redistribution.

2. System Architecture

The proposed multi-port transceiver is designed to replace multiple transceivers on multi-interface CUs (or servers), serving as a reconfigurable bandwidth supplier. Fig. 1a) shows a SiPAC($l = 1$) network [5] as an example multi-interface architecture with 2 interfaces per CU. The silicon photonic (SiP) I/O embedded onto each CU augments the comb-driven DWDM transceiver in [6] by incorporating a wavelength reconfiguration structure (Fig. 1b). By default, modulated DWDM wavelengths coming into this structure are evenly distributed across the two output ports. In response to changing traffic demand or in the event of a connection failure (e.g., a switch or link failure), the transceiver can be reconfigured to allocate wavelengths flexibly across different output ports.

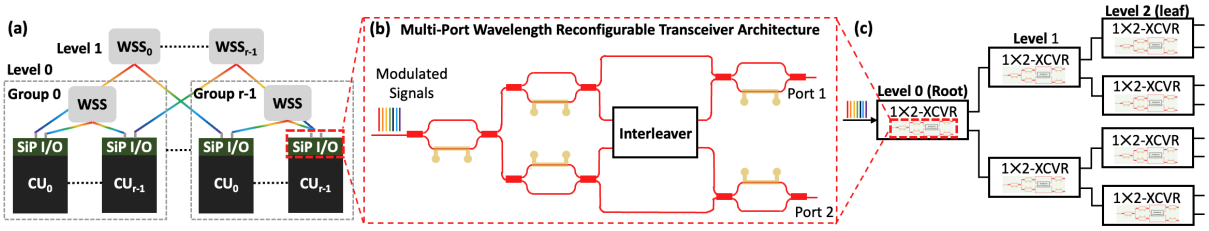

Fig. 1: (a) An example SiPAC architecture with two network interfaces per CU connected to WSSes in two levels. (b) An example of a base unit 1×2 wavelength reconfigurable transceiver architecture. (c) An example of a 1×8 wavelength reconfigurable transceiver architecture achieved by cascading three stages of the base units.

The general $1 \times N \times M\lambda$ transceiver architecture has 1 input, N outputs and a total of M DWDM wavelengths. To ensure a minimum of one wavelength per output port when wavelengths are evenly distributed, we need at least $M = N$ wavelengths. The architecture adopts a binary tree structure (Fig. 1c) where N scales as 2^{l+1}. Here $l \in [0, L-1]$ is the level index with $L = \max(l) + 1$ being the total number of levels. In total, the architecture comprises

$2^L - 1$ base units, similar to a perfect binary tree. In this binary tree hierarchy, the initial level receives modulated DWDM wavelengths as input, while the children of the leaf nodes serve as the output ports. As the architecture scales up, the wavelength allocation at each level follows the *Children Sum Property* of a binary tree structure, where the sum of all wavelengths at child nodes equals the number in their parent node. Every level l consists of 2^l nodes or base units, each adopting a 1×2 switching structure (Fig. 1b). Each base unit can be implemented with 5 balanced MZIs and a tunable band-interleaver [7]. The first and last MZIs in successive stages could be combined to reduce complexity. The balanced MZIs serve as a spatial switch to direct all incoming wavelengths to either of the two arms, while the tunable band-interleaver separates the M wavelengths into arbitrary distributions across its two output ports. For uniform wavelength distribution, the MZIs in the first two stages are tuned to route all wavelengths through the central broadband interleaver which evenly distributes the wavelengths into its two output ports. To redirect all wavelengths to one of the two output ports, the MZI in the second stage can be adjusted, directing all wavelengths through either the upper or lower branch while bypassing the broadband interleaver. We note that the typical number of required output ports (N) is small (e.g., 2 to 4), as an end host typically does not require the full switching capability as an optical switch does. We also assume that the receiving structure (or connection) at the opposite end of the link adopts a similar structure as [6] and has the capacity to support the same number of wavelengths as the transmitting end.

3. Testbed Experiment

We conduct a small-scale on-chip testbed experiment to demonstrate the feasibility of the multi-port wavelength reconfigurable transceiver, as illustrated in Fig. 2a). We emulate parallel wavelength transmission of a 100 GHz comb laser by deploying an array of distributed feedback lasers (DFB) spaced at 100 GHz intervals and modulated with a 16 Gbps PRBS31 via a linear reference modulator. As the proposed transceiver architecture is symmetrical, our demonstrated wavelength routing operations are conducted in the upper half (black dashed box in Fig. 2a), which can be mirrored symmetrically to the bottom half.

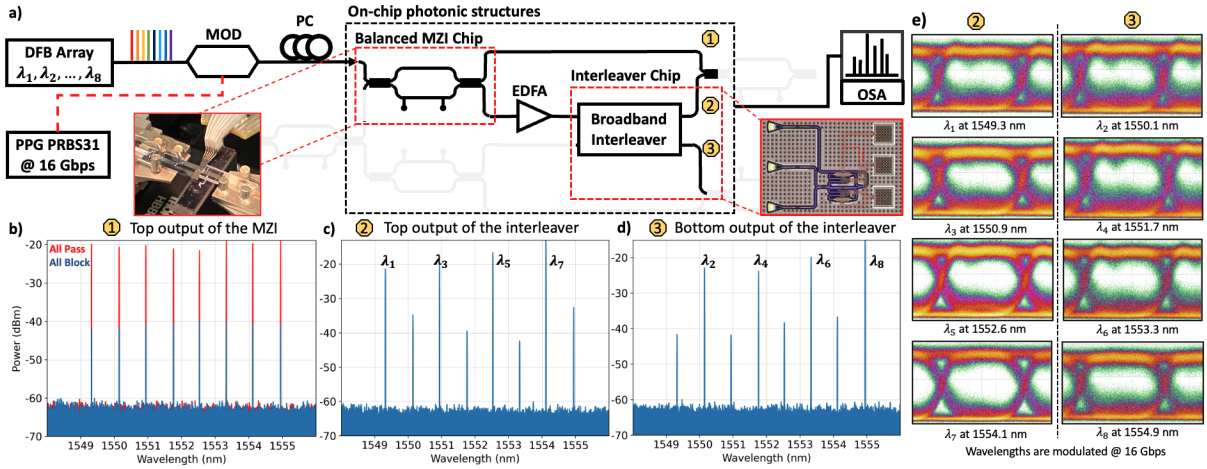

Fig. 2: (a) Schematic of experimental setup: 8 evenly spaced wavelengths generated by DFB lasers are modulated at 16 Gbps and traverse the upper section of the structure, featuring a MZI cascaded with a broadband interleaver. Optical spectra (b, c, d) depict the focused signals in the target wavelength range for various routing operations. Corresponding eye diagrams for each channel in the interleaver's two outputs are displayed in (e).

The signals are coupled into an on-chip balanced MZI via an edge-coupled fiber array. Depending on the wavelength routing operation, the MZI's output can be directly connected to the upper output port or undergo amplification by an EDFA before entering a even-odd interleaver chip. This interleaver [6] emulates the band interleaver and evenly splits the incoming wavelengths into two output ports. The outputs are connected to an optical spectrum analyzer (OSA) for optical spectrum measurements. In the first operation, all wavelengths are routed to the upper port of the MZI, allocating all bandwidth to the first port. The optical spectrum is depicted in Fig. 2b) with red lines representing the all-pass operation. In the second operation, all wavelengths are directed to the bottom MZI port, resulting in an all-block spectrum at the top MZI output, shown in blue lines in Fig. 2b). The extinction ratio of the MZI is observed to be 17 to 21 dB. Wavelengths routed from the bottom port of the MZI are coupled into the interleaver, distributing them into the upper (odd) and lower (even) outputs. The optical spectra of these two output ports are displayed in Fig. 2c) and d). We observe a crosstalk suppression ranging from 14 to 24 dB between selected and adjacent unselected channels. Open eyes are observed across both interleaver outputs (Fig. 2e), confirming the feasibility of the proposed wavelength reconfigurable multi-port transceiver architecture.

4. System-Scale Evaluation

We use a packet-level simulator called Netbench [8] to evaluate the performance of the multi-interface architectures with and without the bandwidth reconfigurable feature. We normalize each topology instance with a per-CU bandwidth of 1920 Gb/s (60 wavelengths modulated at 32 Gb/s each). The workload comprises realistic deep learning traces extracted from application communication task graphs. Multiple instances of each workload are mapped to each network to simulate multi-tenancy, and the job size distribution is set to be uniform. In baseline scenarios, workloads are mapped such that CUs communicate only in a single dimension for Flex-SiPAC, while for SiP-OCS and Super Leaf Spine, it is assumed that half of the outgoing links per CU are not usable. In current multi-interface configurations, Flex-SiPAC's link bandwidth in non-communicating dimensions as well as the bandwidth in non-functioning links of SiP-OCS and Super Leaf Spine are wasted. Using the multi-port reconfigurable transceiver, we can efficiently recover these bandwidths and redirect them to where is needed. Fig. 3 shows the percentage improvement in job completion time (JCT) for networks using the bandwidth-reconfigurable transceiver, compared to those without it, across small (16 CUs), medium (64 CUs), and large (256 CUs) cluster sizes. We observe that Flex-SiPAC, SiP-OCS, and Super Leaf Spine can achieve a maximum of 50%, 94% and 80% improvement, respectively, across topology sizes. For larger topologies, communication in Super Leaf Spine topology is bottlenecked at upper-layer switches and gains little improvement with the reconfiguration ability at the edge. The Flex-SiPAC topology is able to achieve consistent performance improvement due to its all-to-all intra-dimension connectivity.

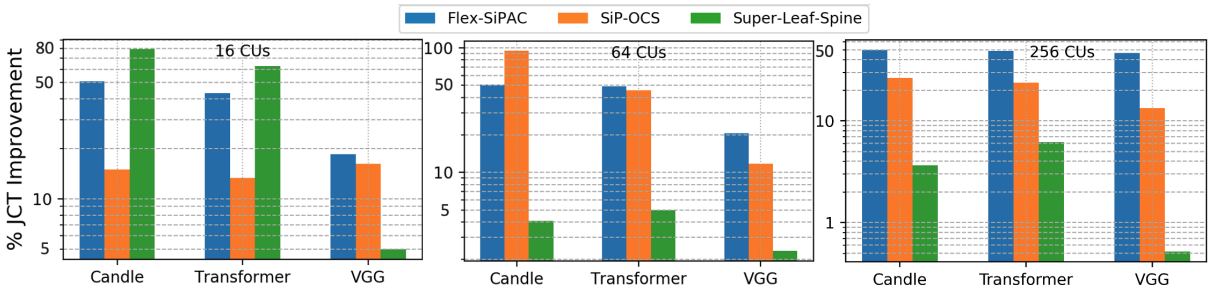

Fig. 3: Percent job completion time (JCT) improvement for networks with bandwidth-reconfigurable transceiver compared to those without it, across small (16 CUs), medium (64 CUs), and large (256 CUs) network sizes.

5. Conclusion

In this work, we propose a multi-port wavelength reconfigurable transceiver architecture. Our experimental testbed results show the feasibility of realizing this architecture on-chip. We report system-level simulation results that show a maximum of 94% communication time reduction compared to the non-reconfigurable counterparts.

Acknowledgments. This work is supported in part by the ARPA-E ENLITENED Program, the National Security Agency (NSA) Laboratory for Physical Sciences (LPS) Research Initiative, and the Center for Ubiquitous Connectivity (CUbiC), and is sponsored by Semiconductor Research Corporation (SRC) and Defense Advanced Research Projects Agency (DARPA) under the JUMP 2.0 program.

References

1. Q. Zhou, K. Wang, H. Lu, W. Xu, Y. Sun, and S. Guo, "Canary: Decentralized distributed deep learning via gradient sketch and partition in multi-interface networks," *IEEE Transactions on Parallel and Distributed Systems*, 2020.
2. F. Yan, C. Xie, J. Zhang, Y. Xi, Z. Yao, Y. Liu, X. Lin, X. Zhang, N. Calabretta *et al.*, "Network traffic characteristics of hyperscale data centers in the era of cloud applications," *Journal of Optical Communications and Networking*, 2023.
3. "Nvidia dgx gh200." [Online]. Available: https://resources.nvidia.com/en-us-dgx-gh200/technical-white-paper
4. M. Khani, M. Ghobadi, M. Alizadeh, Z. Zhu, M. Glick, K. Bergman, A. Vahdat, B. Klenk, and E. Ebrahimi, "Sip-ml: high-bandwidth optical network interconnects for machine learning training," in *ACM SIGCOMM 2021*.
5. Z. Wu, L. Y. Dai, A. Novick, M. Glick, Z. Zhu, S. Rumley, G. Michelogiannakis, J. Shalf, and K. Bergman, "Peta-scale embedded photonics architecture for distributed deep learning applications," *Journal of Lightwave Technology*, 2023.
6. A. Rizzo, A. Novick, V. Gopal, B. Y. Kim, X. Ji, S. Daudlin, Y. Okawachi, Q. Cheng, M. Lipson, A. L. Gaeta *et al.*, "Massively scalable kerr comb-driven silicon photonic link," *Nature Photonics*, vol. 17, no. 9, pp. 781–790, 2023.
7. S. Wang, A. Novick, A. Rizzo, R. Parsons, S. Sanyal, K. J. McNulty, B. Y. Kim, Y. Okawachi, Y. Wang, A. Gaeta *et al.*, "Integrated, compact, and tunable band-interleaving of a kerr comb source," in *CLEO: Science and Innovations*, 2023.
8. Netbench, https://github.com/ndal-eth/netbench.