SiP Architecture For Accelerating Collective Communication in Distributed Deep Learning

Zhenguo Wu[1,*], Liang Yuan Dai[1], Ziyi Zhu[1], Asher Novick[1], Madeleine Glick[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 silicon photonic architecture for accelerating collective communications in distributed deep learning. We demonstrate a 22% job completion time improvement in a small-scale testbed and 1.4 to 5.9× improvement in large-scale simulations. © 2023 The Author(s)

1. Introduction

The increasing requirement for more accurate Deep Learning (DL) models has driven the demand for larger model and dataset sizes. In response, Distributed Deep Learning (DDL) has been adopted. Using DDL, a training job can be distributed to multiple computing units (CUs), interconnected through an intermediary network carrying collective communication traffic. However, current hardware solutions can only provide high-bandwidth connections for a limited *group* of CUs. If a workload is scaled larger than the memory available in a *group* of CUs, bandwidth discrepancy will arise, severely limiting the communication efficiency during the training process.

We present a Silicon Photonic Accelerated Compute cluster architecture, *SiPAC*. SiPAC leverages frequency comb sources, high bandwidth DWDM links, and multi-wavelength selective switches to construct a topology that ensures low network diameter while providing high-bandwidth paths for efficient collective communications in DDL workloads. To demonstrate SiPAC's feasibility, we conducted a testbed experiment where an array of wavelengths are shuffled by a cascaded ring switch, with each ring selecting and forwarding multiple wavelengths to increase the effective communication bandwidth. We also demonstrate the construction of a 4-GPU testbed running a realistic DDL workload. Results show that SiPAC is able to achieve 22% performance improvement relative to a similarly sized leaf-spine topology. Large-scale simulations show that SiPAC improves the communication time by 3.6× to 5.3× over DGX-SuperPod [1], 1.4× to 5.9× over 2D-Torus [2] and 1.4× to 3.4× over BCube [3].

2. System Architecture

The SiPAC architecture realizes a low-diameter, multi-dimensional all-to-all topology by leveraging the multicasting capability of a novel micro-ring resonator (MRR) based multi-wavelength selective switch (WSS). Similar to a BCube [3], a general SiPAC$_l$ ($l \geq 1$) of level l is constructed from r^l r-port switches connecting r SiPAC$_{l-1}$s, totaling r^{l+1} CUs and $L = l+1$ levels of switches (Fig.1a). CUs in a SiPAC$_l$ have $L = l+1$ optical ports and are connected to a WSS in each of the L levels (the diameter L is typically small). Unlike BCube, we replace 1) each compute server with a disaggregated CU and 2) each electronic packet switch (EPS) with a WSS (Fig.1 b).

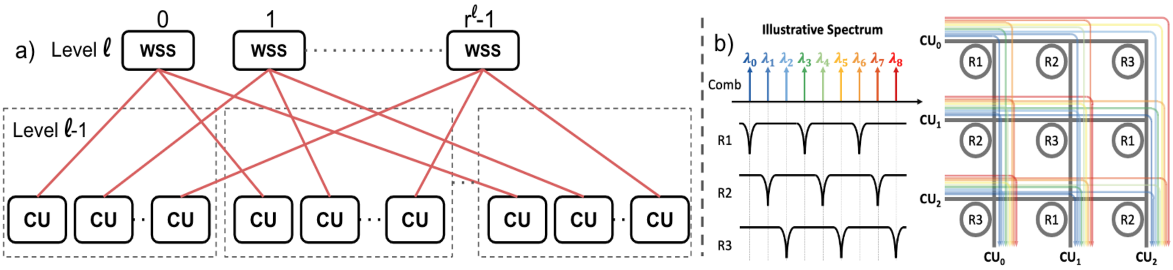

Fig. 1: a) SiPAC architecture schematic. b) Wavelength multiplexing for 9 wavelengths in a 3 × 3 WSS. 3 wavelengths are dropped at each ring to maintain sufficient bandwidth and achieve optical multi-casting.

The proposed WSS (Fig.1b) is designed to exploit the periodic property of the free spectral range (FSR) of the MRRs, extending past works that used AWGR [4]. By carefully engineering the FSRs, each MRR has the ability to drop multiple wavelengths which enables CUs connected to the same WSS to communicate with uniform high bandwidth. The cascaded ring structure separates the incoming wavelengths into subgroups and recombines the interleaved wavelengths into common output buses which contain all different wavelengths. It shuffles the input wavelengths to different outputs and effectively achieves the optical multi-casting functionality. This enables each CU to communicate directly via arbitration-free high-bandwidth light paths with $(l+1)(r-1)$ other CUs with no output port blocking. And the packet-switch-less design mitigates intermediate packet buffering and reduces in-network queuing delays.

3. Testbed Experiments

We first highlight a hardware implementation for achieving multi-wavelength selective switching via a single WSS cell, in conjunction with a high bandwidth density Kerr frequency comb source. Our testbed setup is illustrated in Fig.2a). A continuous-wavelength tunable-laser-source (CW-TLS) centered on 1561.42 nm is used to pump a silicon-nitride Kerr comb chip (Fig.2b), which generates evenly spaced lines at 201.5 GHz (≈ 1.6 nm) intervals. The outputs are filtered by an optical bandpass filter (OBF). The wavelengths are modulated with a 10 Gbps PRBS31 via a linear reference modulator, and coupled into the cascaded 1 × 8 MRR switch (Fig.2d). Each MRR has a FSR of 14.41 nm and can drop multiple channels. The rings are thermo-optically tuned to select the comb lines at 1534.07 and 1548.48 nm for R1 and 1532.48 nm and 1546.87 nm for R2. Polarization controllers (PC) are used to maximize the optical power. The optical spectrum captured at the drop port (Fig.2e&f) measured through an optical spectrum analyzer (OSA) shows that our channels of interest have a crosstalk suppression of 13.3 dB over an adjacent unselected one. The dropped signals are amplified by an Erbium Doped Fiber Amplifier (EDFA), and variable optical attenuators (VOA) are used to manage received optical power at the photo-detector (PD). Open eyes were observed in all cases (Fig.2h), confirming the feasibility of the proposed switch architecture.

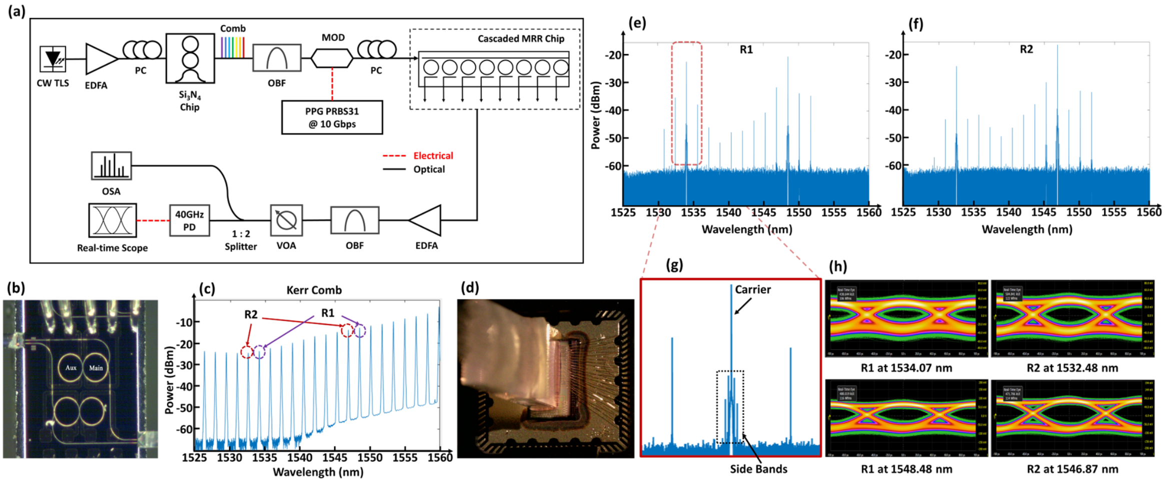

Fig. 2: (a) Schematic of the experimental setup. A Kerr frequency comb (b) is used to generate evenly spaced lines (c). The signal is modulated with a 10 Gbps PRBS31 via a linear reference modulator, and coupled into a 1×8 MRR switch (d). The optical spectra (e&f) show the focused signal in our wavelength range of interest. Open eyes, required to ensure proper operation, are observed for both wavelengths (h).

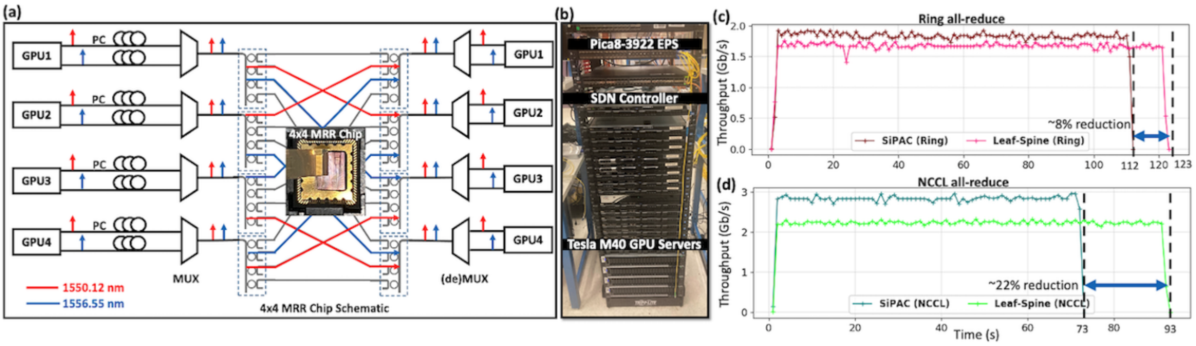

Fig. 3: (a) Schematic of the SiPAC$(r=2,l=1)$ testbed setup. (b) GPU servers connected to a EPS and a SDN controller. Throughput of the injection port under (c) ring all-reduce and under (d) NCCL all-reduce.

We then demonstrate the system-level performance of a small-scale SiPAC$(r = 2, l = 1)$ architecture using 4 NVIDIA Tesla M40 GPUs with RoCEv2 enabled Mellanox ConnectX-4 NICs (Fig.3b). The testbed setup is shown in Fig.3a). To emulate parallel wavelength transmission, each GPU is configured to have a virtual bridge equipped with two 10 Gbps SFP+ transceivers sending at two different wavelengths (1550.12 nm & 1556.55 nm). We use a separate 4×4 MRR-based WSS to realize the wavelength shuffling and recombining in the optical layer of the SiPAC topology. We use TensorFlow to run a distributed MobileNetV2 neural network using both the ring and NCCL collective algorithm and compare SiPAC's performance with a similarly sized EPS-based leaf-spine topology. We ran each training workload for two epochs with a batch size of 128. The network throughput is captured using the Ryu SDN OpenFlow monitoring program (Fig.3c&d). Under ring-based all-reduce, SiPAC is able to achieve a 8% job completion time (JCT) reduction relative to the leaf-spine architecture from shorter network paths and less packet buffering. When using NCCL all-reduce, the JCT reduction is further increased to 22% as the NCCL tree-based algorithm can better leverage the multi-port property of the SiPAC architecture.

4. System-Scale Evaluation

We use Netbench, a packet-level simulator [5] to evaluate the performance of large-scale SiPAC architectures. We compare its performance against serveral state-of-the-art DL clusters including DGX-Superpod [1], 2D-Torus [2] and BCube [3]. We normalize the per-CU bandwidth (i.e., the sum of all link bandwidths connected to a CU) in each topology to be 2048 Gb/s. Under general collective traffic patterns (i.e., incast, broadcast, all-to-all), we observe that SiPAC consistently performs well, with 3.6× to 5.3× JCT improvement over SuperPod, 1.4× to 5.9× over 2D Torus, and 1.4× to 3.4× over electronic BCube (Fig.4a). This is due to SiPAC's ability to enable simultaneous direct transmissions to and from $(l+1)(r-1)$ different endpoints without intermediate switch buffering. We then vary the per-CU bandwidth for each architecture from 128 Gb/s to 4096 Gb/s under hybrid collective communications. While the JCT of SiPAC continues to decrease as the per-CU bandwidth increases, the JCTs for the other topologies do not improve much further (Fig.4b). Taking SuperPod as an example, the communication becomes bottlenecked at the slower inter-server links. The SiPAC architecture is able to achieve much better bandwidth scaling which shows its promise for extreme high-bandwidth silicon photonic technologies.

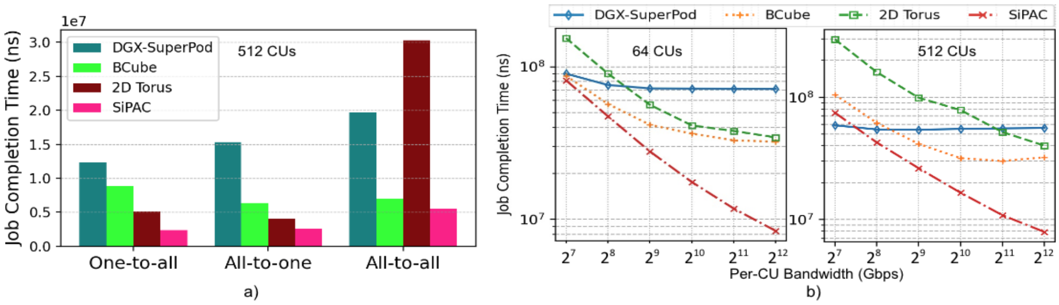

Fig. 4: a) JCT of primitive collective communications for 512 CUs at 1 MB message size. b) JCT of hybrid parallel collective communication (all-reduce & all-to-all) at 64 and 512 CUs with 100 MB message size.

5. Conclusion

In this work, we propose the SiPAC architecture for DDL acceleration. Our experimental testbed results show MRR's capability to achieve compact and high bandwidth multi-wavelength switching, demonstrating the feasibility of the SiPAC architecture. We report system-level testbed results that show a 22% performance improvement on realistic DDL workload. Large-scale simulations show that SiPAC clusters can achieve a 1.4× to 5.9× collective communication time reduction compared to current state-of-the-art architectures.

Acknowledgments. This work was supported by ARPA-E ENLITENED Program and the National Security Agency (NSA) Laboratory for Physical Sciences (LPS) Research Initiative (R3/NSA).

References

1. "Nvidia dgx superpod." [Online]. Available: https://www.nvidia.com/en-us/data-center/dgx-superpod/
2. "Google cloud tpu." [Online]. Available: https://cloud.google.com/tpu
3. C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "Bcube: a high performance, server-centric network architecture for modular data centers," in *Proceedings of the ACM SIGCOMM 2009*.
4. M. Fariborz, X. Xiao, P. Fotouhi, R. Proietti, and S. B. Yoo, "Silicon photonic flex-lions for reconfigurable multi-gpu systems," *Journal of Lightwave Technology*, vol. 39, no. 4, pp. 1212–1220, 2021.
5. Netbench, https://github.com/ndal-eth/netbench.