Efficient Multicast Communication in Silicon Photonics Enhanced DNN Acceleration

Yuan Li*, Ahmed Louri*, Avinash Karanth†
*Department of Electrical and Computer Engineering, George Washington University, Washington, DC, USA
†School of Electrical Engineering and Computer Science, Ohio University, Athens, Ohio, USA
Emails: *{liyuan5859, louri}@gwu.edu, †karanth@ohio.edu

***Abstract*—The interconnection network plays a vital role in determining the functionality and performance of a hardware deep neural network accelerator. We explore employing silicon photonic interconnects and leveraging their energy-efficient multicast and distance-independent latency properties. We discuss consequent innovations in dataflow optimization and architecture design.**

***Index Terms*—silicon photonics, neural network, accelerator**

I. Introduction

A hardware deep neural network (DNN) accelerator typically includes numerous simple processing elements (PEs) working in coordination. The interconnection network connecting these PE and the memory hierarchy plays a vital role in determining the functionality and performance of a DNN accelerator [1], [2]. Conventional metallic-based wires often create a bottleneck in scaled-up DNN accelerators as they cannot effectively support communication over increased distances without performance degradation [3]–[5]. We propose to incorporate silicon photonic interconnects in accelerator design to tackle the communication bottleneck. Among many well-recognized advantages of silicon photonic interconnects compared to metallic-based wires, we are particularly interested in the energy-efficient multicast and distance-independent latency properties [5]–[7] and their impact on dataflow optimization and architecture design choices.

Prior dataflow optimizations target maximizing data reuse in memory hierarchies close to computing [1], [2] since accessing data in lower hierarchies incurs notable overhead in terms of latency and energy consumption. Nevertheless, local data reuse inherently constrains the obtainable parallelism, and potentially creates date duplicates that fill up the valuable on-chip memory. Fig. 1 shows a unicast channel and a single-write-multiple-read (SWMR) multicast channel. They both outmatch metallic-based wires with the advanced ground-referenced signaling (GRS) technique while SWMR channel achieves 0.1 $pJ/bit/receiver$ communication, which is 17× lower than metallic-based wires due to transmitter sharing. Our proposed dataflow optimization maximizes multicast communications by performing operations with shared data in parallel on different PEs.

II. Dataflow Optimization

Operations in a DNN layer are presented as a nested loop on the following dimensions: output channel $\langle k \rangle$, input channel $\langle c \rangle$, weight kernel shape $\langle r\ s \rangle$, and output plane shape $\langle e\ f \rangle$. Our dataflow optimization includes three interactive parts, with the first and the last parts targeting maximizing local data reuse inside off-chip DRAM and on-chip PEs, respectively. We only elaborate on the second part here since it targets maximizing multicast communications on the silicon photonic interconnects between the GLB and PEs while still maintaining reasonable data reuse in the GLB. Table I illustrates the generated multicast opportunities, remaining local reuse opportunities, and incurred communication patterns of all three involved data types, namely weight kernel ($\mathcal{W}$), input feature ($\mathcal{I}$), and partial sum or output feature ($\mathcal{P}$), when enabling parallel execution in each individual dimension or a collection of dimensions. Our observations are: (1) the priority to enable parallel execution should be $\langle k \rangle$, $\langle e\ f \rangle$, $\langle c \rangle$, and $\langle r\ s \rangle$; (2) parallel execution in $\langle k\ e\ f \rangle$ dimensions yields maximum multicast of $\mathcal{W}$ and $\mathcal{I}$ while maintaining local

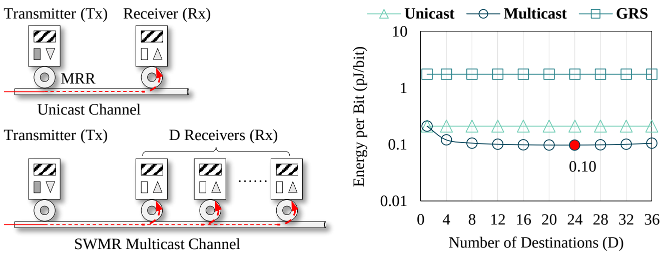

Fig. 1: Photonic communication channels (left) and their energy efficiency per receiver (right) with parameters as in [6].

TABLE I: Parallelism Exploration

Dimension	Data	Multicast	Reuse	Communication Patterns
$\langle k \rangle$	$\mathcal{W}$	✘	✔	GLB → PE Unicast Communication
	$\mathcal{I}$	✔	✔	GLB → PE Multicast Communication
	$\mathcal{P}$		✔	GLB ↔ PE Unicast Communication
$\langle e\ f \rangle$	$\mathcal{W}$	✔	✘	GLB → PE Multicast Communication
	$\mathcal{I}$	✔	✔	GLB → PE Multicast Communication*
	$\mathcal{P}$		✔	GLB ↔ PE Unicast Communication
$\langle c \rangle$	$\mathcal{W}$	✘	✔	GLB → PE Unicast Communication
	$\mathcal{I}$	✘	✔	GLB → PE Unicast Communication
	$\mathcal{P}$		✔	GLB ↔ PE Unicast Communication
$\langle r\ s \rangle$	$\mathcal{W}$	✘	✔	GLB → PE Unicast Communication
	$\mathcal{I}$	✔	✔	GLB → PE Multicast Communication
	$\mathcal{P}$		✔	GLB ↔ PE Unicast Communication
$\langle k\ e\ f \rangle$	$\mathcal{W}$	✔	✘	GLB → PE Multicast Communication
	$\mathcal{I}$	✔	✘	GLB → PE Multicast Communication
	$\mathcal{P}$		✔	GLB ↔ PE Unicast Communication
$\langle k\ e\ f\ c \rangle$	$\mathcal{W}$	✔	✘	GLB → PE Multicast Communication
	$\mathcal{I}$	✔	✘	GLB → PE Multicast Communication
	$\mathcal{P}$		✔	GLB ↔ PE Unicast Communication
$\langle k\ e\ f\ c\ r\ s \rangle$	$\mathcal{W}$	✔	✘	GLB → PE Multicast Communication
	$\mathcal{I}$	✔	✘	GLB → PE Multicast Communication
	$\mathcal{P}$		✘	GLB ↔ PE Unicast Communication

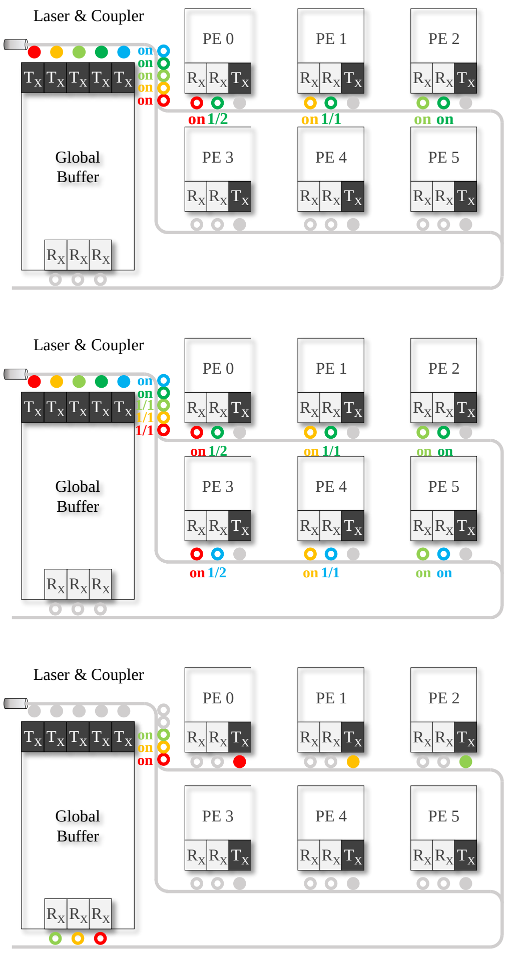

Fig. 2: Unicast & multicast (top), two-dimensional multicast (middle), and write-back unicast (bottom) communication modes in our proposed architecture design with $M \times N = 6$ PEs. M and N represent the numbers of rows and columns of the PE array, respectively.

reuse of $\mathcal{P}$; (3) the silicon photonic interconnects are expected to support simultaneous unicast & multicast communication, simultaneous two-dimensional multicast communication, and write-back unicast communication.

III. Network Architecture

We demonstrate the GLB, PEs, and interconnection network formed by silicon photonic interconnects inside our proposed accelerator architecture in Fig. 2. We use M and N to represent the numbers of rows and columns of the PE array. A laser is coupled to a waveguide that traverses the transmitters in the GLB and connects to a set of M horizontal waveguides. A set of $M + N$ microring resonators (MRRs) is responsible for splitting a proper fraction of laser power to the corresponding horizontal waveguide. The M horizontal waveguides finally merge into one waveguide which connects to the receivers in the GLB. In addition to working in on-resonant and off-resonant states to act as modulators or filters, MRRs in our architecture also work in a transient state with a biased voltage applied. An MRR in this transient state forwards α and $1 - \alpha$ fraction of laser power to drop and through ports, respectively, forming a split ratio of $\alpha/(1-\alpha)$. The interconnection network in our proposed architecture can switch dynamically among three working modes. The unicast & multicast mode is for cases where only one type of input data ($\mathcal{W}$ or $\mathcal{I}$) is multicast such as parallel execution in $\langle k \rangle$ dimension only. The two-dimensional multicast mode is for cases where both types of input data are multicast such as parallel execution in $\langle k\ e\ f \rangle$ dimensions. The write-back unicast mode is utilized to send intermediate or final computing results to the GLB. The MRRs are tuned accordingly to switch among working modes. In Fig. 2 filled circles represent MRRs as modulators while hollow circles represent MRRs as filters or splitters. The color of a circle represents the specific wavelength that it resonates while a circle in grey indicates that this MRR is inactive.

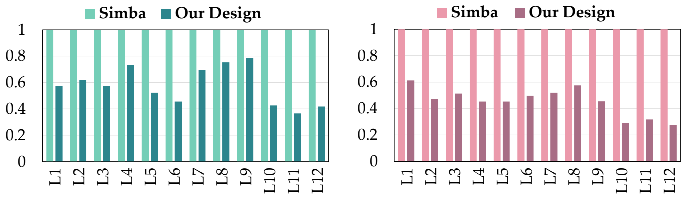

Fig. 3: Execution time (left) and energy (right) comparison between Simba and our design, normalized to Simba.

IV. Evaluation

Fig. 3 illustrates the per-layer execution time and energy comparison between Simba [1], which is a state-of-the-art DNN accelerator that only implements metallic-based wires, and our proposed design using the `VGG-16` neural network model for `ImageNet` application. Our proposed design achieves on average 51% and 67% decrease in execution time and energy, respectively. The area of a PE in our architecture is 0.72 mm^2 while the area for a transceiver is 0.0096 $mm^2/wavelength$ [8]. The area overhead is 3.9%.

Acknowledgment

This research was partially supported by NSF grants CCF-1702980, CCF-1812495, CCF-1901165, CCF-1953980, CCF-1703013, and CCF-1901192.

References

[1] Y. S. Shao *et al.* Simba: Scaling deep-learning inference with multi-chip-module-based architecture. In *MICRO*, 2019.

[2] Y. Chen *et al.* Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. In *ISCA*, 2016.

[3] Y. Li *et al.* Scaling deep learning inference with chiplet-based architecture and photonic interconnects. In *DAC*, 2021.

[4] Y. Li *et al.* SPRINT: A high-performance, energy-efficient, and scalable chiplet-based accelerator with photonic interconnects for CNN inference. *IEEE TPDS*, 2022.

[5] D. Miller. Device requirements for optical interconnects to silicon chips. *Proceedings of the IEEE*, 2009.

[6] Y. Li *et al.* SPACX: Silicon photonics-based scalable chiplet accelerator for DNN inference. In *HPCA*, 2022.

[7] Y. Li *et al.* ASCEND: A scalable and energy-efficient deep neural network accelerator with photonic interconnects. *IEEE TCAS I*, 2022.

[8] Y. Thonnart *et al.* A 10Gb/s Si-photonic transceiver with 150μW 120μs-lock-time digitally supervised analog microring wavelength stabilization for 1Tb/s/mm^2 die-to-die optical networks. In *ISSCC*, 2018.