WDM-Enabled Photonic Edge Computing

Alexander Sludds[1], Ryan Hamerly[1,2,*], Saumil Bandyopadhyay[1], Zaijun Chen[1], Zhizhen Zhong[1], Liane Bernstein[1], Manya Ghobadi[1], Dirk Englund[1]
[1] Research Laboratory of Electronics, MIT, 50 Vassar St, Cambridge, MA 02139, USA
[2] NTT Research Inc., PHI Laboratories, 940 Stewart Dr, Sunnyvale, CA 94305
[*]rhamerly@mit.edu

Abstract: *We experimentally realize photonic edge computing over an 86-km fiber link with 3 THz optical bandwidth and demonstrate DNN inference at 98.8% accuracy with optical energy consumption below 40 aJ/MAC.*
Keywords: *Edge computing, neural networks, fiber optics, WDM*

I. Introduction

Machine learning is ubiquitous in cloud computing and data centers, but recently, network and privacy constraints are pushing processing closer to the end user [1]. In this "edge computing" paradigm, processing is instead done on size, weight, and power (SWaP)-constrained smart sensors and end stations, which have difficulty running state-of-the-art deep neural networks (DNNs) due to the large power and memory requirements of the latter. This problem has persisted despite great efforts toward SWaP-constrained hardware [2] and model compression [3]. Photonic DNN accelerators aim to bridge this gap [4, 5], but chip-area constraints arising from the weight-stationarity [2] of existing architectures (and error propagation issues [6-9]) likely preclude their application to very large DNNs. As the limitations to existing designs stem from weight stationarity, where each DNN weight maps to a discrete photonic element, recently we proposed an alternative based on output-stationary coherent detection and integration [10, 11] where weights are encoded optically in the temporal domain and the chip area scales with the number of neurons, not synapses (O(N) rather than O(N^2)). This optical weight encoding suggests an edge computing architecture where weights are distributed to edge clients optically, pushing the bulk of the computational cost to the server while the computation is still performed at the edge. However, the imaging requirements of Ref. [10] rendered this particular approach impractical for such an edge-computing scenario.

II. NetCast Protocol

Recently, we proposed NetCast, an optical server-client protocol based on wavelength-division multiplexing (WDM), difference detection and integration, and optical weight delivery [12, 13]. Our protocol splits the computation over two components: a "weight server" consisting of a WDM modulator array, connected by an optical link to a SWaP-constrained client. As shown in Fig. 1, over a sequence of time steps, the weight server encodes the DNN weights as an analog signal in an optical time-frequency basis. This signal is transmitted to the client, where it is modulated and demultiplexed; the time-integrated photocurrent encodes the neuron activations for the next DNN layer.

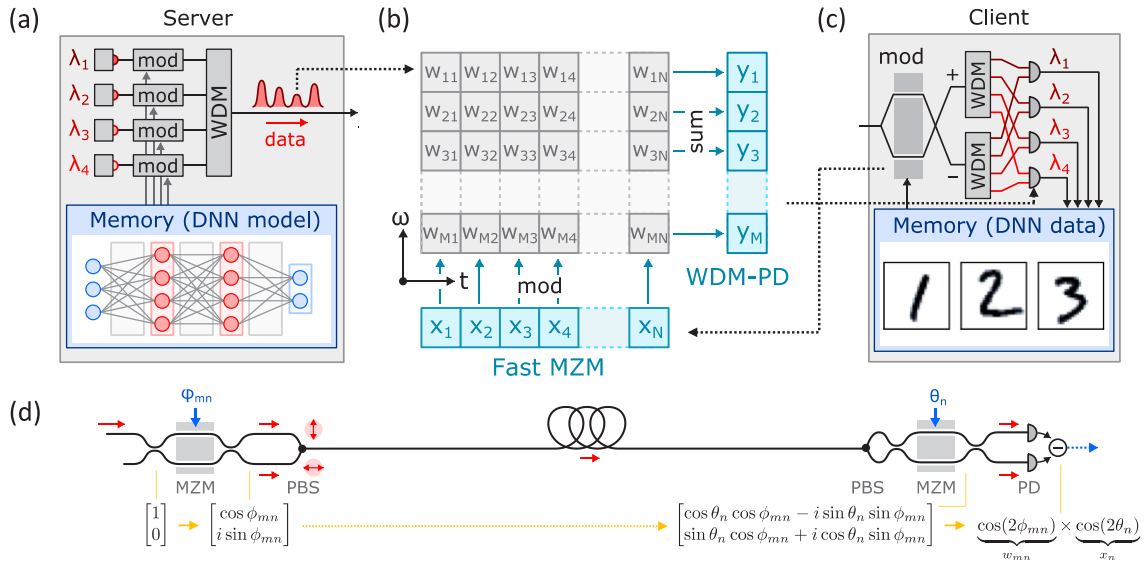

Fig. 1. NetCast architecture and dataflow. (a) DNN model weights are encoded on an optical pulse train at the server, a WDM modulator bank. (b) Matrix rows and columns mapped to a time-frequency basis. (c) Matrix-vector multiplication is performed at the client by modulation and integration detection. (d) Dataflow of a single wavelength channel.

©IEICE

Authorized licensed use limited to: MIT Libraries. Downloaded on September 10,2022 at 18:18:14 UTC from IEEE Xplore. Restrictions apply.

As linear algebra is the rate-limiting step for DNNs, here we focus on how NetCast accelerates matrix multiplication $y_m = \sum_n w_{mn} x_n$ (activations, normalization and pooling can be performed locally at the client for minimal added cost). The server (Fig. 1(a)) consists of a broadband WDM transmitter source with multiple channels that transmits the DNN weight matrices w_{mn} to the client in a time-frequency basis, with rows (resp. columns) of w_{mn} mapped to WDM channels (resp. time bins) of the optical signal (Fig. 1(b)). At the client, this signal passes through a broadband modulator encoding the activations x_j, and is demultiplexed into WDM channels for integration detection (Fig. 1(c)). To understand how this operation maps to multiplication, consider the path of a the m^{th} WDM channel, Fig. 1(d). On the server side, at the n^{th} time step, a dual-port MZM splits the input into two channels with amplitudes $\vec{a}_{mn} = [\cos(\phi_{mn}), i\sin(\phi_{mn})]$, where the signals encode the weight w_{mn} through differential signaling. A polarization beamsplitter (PBS) combines these channels onto orthogonal polarizations of a fiber (or free-space link) which connects the server to the client. At the client side, the polarizations are recombined onto a second (broadband) MZM whose coupling angle θ_n encodes the n^{th} vector element x_n. After demultiplexing, the accumulated differential current is:

$$Q_m \propto \cos(2\phi_{mn})\cos(2\theta_n)$$

In this way, with the encodings $2\phi_{mn} = \cos^{-1}(w_{mn})$ and $2\theta_n = \cos^{-1}(x_n)$, the client generates a signal proportional to $y_m = \sum_n w_{mn} x_n$, performing the desired matrix-vector product via optical modulation and detection. The key insight here is that the optical link allows us to separate the tasks of logic (client) and memory access (server), significantly reducing the cost of the computation at the client side—for large DNNs in particular, the energy and memory costs at the client are dwarfed by that at the server. This liberates the edge device from its SWaP constraints, enabling the edge deployment of whole new classes of DNNs that have heretofore been restricted to data centers.

III. Experiment

We realize the NetCast protocol using a smart transceiver, shown in Fig. 2(a), fabricated on a 220-nm silicon-photonics (SiPh) process at OpSIS/IME (now AMF). This chip consists of 48 MZMs, each capable of modulation up to 50 Gbps for a total bandwidth of 2.4 Tbps [14]. The smart transceiver supports WDM, and we demonstrated multiplexing of 16 WDM lasers simultaneously transmitting through the chip with -10 dBm (100 μW) power per wavelength. Fig. 2(b) shows an open eye diagram for on-off keying (OOK) of a single modulator channel at 50 GHz. For a field demonstration, weights are transmitted over 43 km of deployed optical fiber connecting MIT's main campus with MIT Lincoln Laboratory (MIT-LL), for a total round-trip distance of 86 km (Fig. 2(c)).

At the client, a passive wavelength demultiplexer separates each wavelength channel for detection onto an array of custom time-integrating receivers. With uniformly distributed random data, we measured an r.m.s. error of $\sigma_{rms} = 0.005$, corresponding to at least 8 bits of precision, primarily limited by calibration of modulator and detector nonlinearities, which is comfortably higher than the ~5 bits of precision required for common DNN inference [15, 16]. As proof of this, we perform image classification by running a benchmark image classification task (MNIST) on a pre-trained DNN using the NetCast hardware. The observed classification accuracy of 98.8% (Fig. 2(d)) performed using 3 THz of optical bandwidth over the deployed fiber, is statistically indistinguishable from the model's canonical accuracy, and is independent of whether the server and client are connected locally or over the 86-km link. This result shows the potential for this architecture to support edge computing at ultra-high bandwidths in real-world deployed systems using presently-available telecom components.

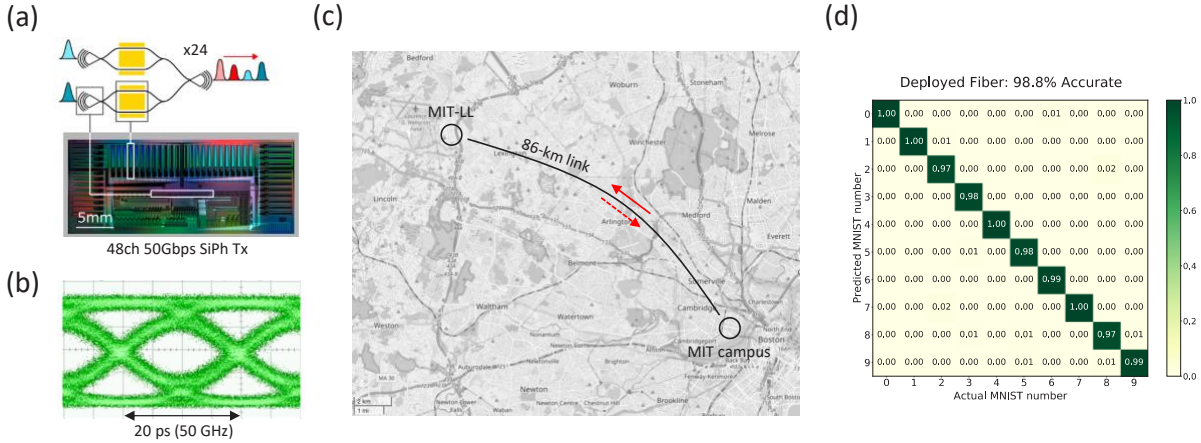

Fig. 2. Experimental demonstration of NetCast. (a) SiPh smart transceiver used as a weight server. (b) Eye diagram showing 50 Gbps OOK transmission through a single MZM. (c) Trunk link consisting of 43km of deployed optical fiber between MIT campus and MIT-LL, for a round-trip distance of 86 km. (d) MNIST classification accuracy for weight delivery over the deployed fiber.

©IEICE

Authorized licensed use limited to: MIT Libraries. Downloaded on September 10,2022 at 18:18:14 UTC from IEEE Xplore. Restrictions apply.

To evaluate the energy consumption of NetCast, we also operated the client in a photon-starved environment by attenuating the input laser power. Consistent with Ref. [10], both shot noise and thermal noise limit the performance of the system in this regime. By measuring the classification accuracy as a function of optical input power, we can derive a lower bound for the optical energy consumption required for this system. With conventional photodetectors and time-integrateing receivers, we find that high accuracy is attainable with < 40 aJ/MAC optical power, limited by Johnson-noise fluctuations in the accumulated charge $\sigma_{th} = \sqrt{kTC/q}$. Improvements to time-integrating receivers aimed at reducing the integration capacitance may reduce this figure still further, suggesting high-fidelity edge inference is possible even over very high-loss links. The electrical energy consumption at the client side is also reduced by a factor of the matrix dimension N [12].

IV. Conclusion

As computing moves to the edge, optics offers new possibilities to deliver high performance while adhering to strict SWaP constraints. In this paper, we have introduced and experimentally demonstrated NetCast, a server-client architecture that leverages unique advantages of optics—the bandwidth of fiber links, support for WDM, and analog integration detection—to split the DNN inference problem into two tasks, effectively pushing the energy- and memory-intensive tasks to the server. We used this protocol to perform DNN inference over an 86-km deployed fiber using 3 THz of optical bandwidth, and showed classification accuracy is not degraded compared to the model's canonical performance. Moreover, we operated the receiver in a photon-starved environment to show that accurate inference is possible with as low as 40 aJ/MAC, limited by thermal charge fluctuations. This demonstration suggest NetCast is a promising approach to optical information processing at the edge.

Acknowledgment

This research is funded by a collaboration with NTT Research and NSF Eager (CNS-1946976). This material is based on research sponsored by the Air Force Office of Scientific Research (AFOSR) under award number FA9550-20-1-0113, the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1007, the Army Research Office (ARO) under agreement number W911NF-17-1-0527, NSF RAISE-TAQS grant number 1936314 and NSF C-Accel grant number 2040695.

References

[1] W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, and X. Yang, "A survey on the edge computing for the internet of things," IEEE Access 6, pp. 6900–6919, 2017.
[2] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," Proceedings of the IEEE 105(12), pp. 2295–2329, 2017.
[3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
[4] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, et al., "Deep learning with coherent nanophotonic circuits," Nature Photonics 11(7), p. 441, 2017.
[5] A. N. Tait, T. F. Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Scientific Reports 7(1), p. 7430, 2017.
[6] M. Y.-S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, "Design of optical neural networks with component imprecisions," Optics Express 27(10), pp. 14009–14029, 2019.
[7] S. Bandyopadhyay, R. Hamerly, and D. Englund, "Hardware error correction for programmable photonics," arXiv preprint arXiv:2103.04993, 2021.
[8] R. Hamerly, S. Bandyopadhyay, and D. Englund, "Stability of self-configuring large multiport interferometers," arXiv preprint arXiv:2106.04363, 2021.
[9] R. Hamerly, S. Bandyopadhyay, and D. Englund, "Accurate self-configuration of rectangular multiport interferometers," arXiv preprint arXiv:2106.03249, 2021.
[10] R. Hamerly, L. Bernstein, A. Sludds, M. Soljacic, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," Physical Review X 9(2), p. 021032, 2019.
[11] L. Bernstein, A. Sludds, R. Hamerly, V. Sze, J. Emer, and D. Englund, "Freely scalable and reconfigurable optical hardware for deep learning," Scientific Reports 11(1), pp. 1–12, 2021.
[12] R. Hamerly, A. Sludds, S. Bandyopadhyay, L. Bernstein, Z. Chen, M. Ghobadi, and D. Englund, "Edge Computing with Optical Neural Networks via WDM Weight Broadcasting," in SPIE OP21 (no. 11804-63), 2021.
[13] A. Sludds, S. Bandyopadhyay, Z. Chen, Z. Zhong et al., "Wavelength Multiplexed Ultralow-Power Photonic Edge Computing," arXiv:2203.05466, 2022.
[14] M. Streshinsky, A. Novack, R. Ding, Y. Liu, A. E.-J. Lim, P. G.-Q. Lo, T. Baehr-Jones, and M. Hochberg, Journal of Lightwave Technology 32, p. 4370, 2014.
[15] T. Gokmen, M. J. Rasch, and W. Haensch, in 2019 IEEE International Electron Devices Meeting (IEDM) (IEEE, 2019) pp. 22–3.
[16] S. Garg, J. Lou, A. Jain, and M. Nahmias, arXiv preprint arXiv:2102.06365, 2021.

©IEICE

Authorized licensed use limited to: MIT Libraries. Downloaded on September 10,2022 at 18:18:14 UTC from IEEE Xplore. Restrictions apply.