Wavelength Multiplexed Photonic Edge Computing in the Output Stationary Frame

Ryan Hamerly[1,2], Alex Sludds[1], Saumil Bandyopadhyay[1], Zaijun Chen[1], Zhizhen Zhong[1], Liane Bernstein[1], Manya Ghobadi[1], and Dirk Englund[1]

[1]*MIT RLE, 50 Vassar St., Cambridge, MA 02139, USA.* [2]*NTT Research, 940 Stewart Dr., Sunnyvale, CA 94085, USA.*
rhamerly@mit.edu

Abstract: We propose a photonic edge computing architecture based on WDM, broadband modulation, and output-stationary integration. Using this scheme, we demonstrate 98.8%-accurate DNN inference over an 86-km deployed fiber link with 3 THz optical bandwidth.

OCIS codes: 200.4260, 200.4860, 060.2330

Machine learning is by now ubiquitous in cloud computing and data centers, but in recent years, privacy and network limitations are pushing processing closer to the user at the "edge" of the network. The size, weight, and power (SWaP) constraints of edge devices limit the ability to run state-of-the-art deep neural networks (DNNs) due to the significant power and memory requirements of the latter. Despite recent efforts in SWaP-constrained hardware (including photonic hardware) and model compression, to date there is no approach that satisfactorily solves both the power and memory bottlenecks of edge computing while also being scalable to large, state-of-the-art models.

This talk focuses on our recent progress towards photonic architectures for accelerated DNN inference at the edge. First, we introduce Netcast, an optical server-client protocol that relies on wavelength division multiplexing (WDM), broadband modulation, and output-stationary integration [1, 2]. Netcast divides the computation among two components: a "weight server" consisting of a WDM array of modulators (Fig. 1(a)) and a client (Fig. 1(c)). The weight server encodes the DNN weights into an analog optical signal in a time-frequency basis (Fig. 1(b)), which is then

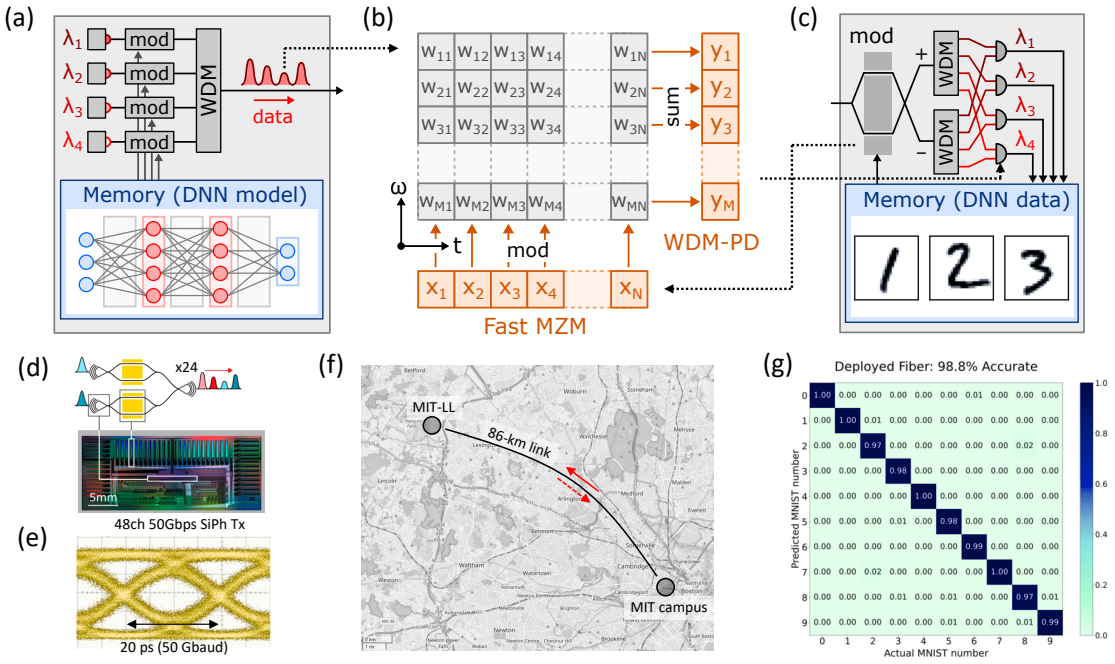

Fig. 1. Netcast concept. (a-c) Dataflow between (a) server and (c) client, in a (b) time-wavelength basis. (d-e) Experimental SiPh server. (f) MIT/MIT-LL fiber link. (g) DNN confusion matrix.

transmitted to the client, where it is modulated, demultiplexed, and detected. After time integration, the detector outputs encode the matrix-vector multiplication (MVM) $y_i = \sum_j w_{ij} x_j$ between the (optical, server-side) weights w_{ij} and the (client-side) activations x_j. This signal is then processed digitally, computing the input activations to the next layer.

We implemented Netcast using a 220-nm silicon-photonics (SiPh) smart transceiver (Tx) fabricated at OpSIS/IME (now AMF), which consists of 48 parallel 50 Gbps MZMs with an aggregate bandwidth of 2.4 Tbps (Fig. 1(d)). The SiPh Tx supports WDM, and we multiplexed 16 μITLA WDM lasers through the chip with $P = 10$ dBm/λ. Fig. 1(e) shows the OOK eye diagram, which remains open at 50 Gbaud. To prove the potential of this protocol for delocalized edge computing, the server and client were separated by 86 km of deployed optical fiber via a round-trip link between MIT's main campus and MIT Lincoln Laboratory (MIT-LL) (Fig. 1(f)). Accurate MVM was achieved with an r.m.s. error (on uniform random data) of $\sigma_{\text{rms}} = 0.005$, corresponding to a precision of ENOB $= 7.4$ bits, above the ~ 5 bits required for common inference tasks. To verify this, we performed MNIST image classification on a pre-trained DNN using Netcast, achieving 98.8% classification accuracy (Fig. 1(g)) using 3 THz of optical bandwidth over the deployed fiber. The energy efficiency was studied by varying the optical power, by which we find high accuracy is attainable with < 40 aJ/MAC optical power for conventional detectors, limited by Johnson-noise fluctuations $\sigma_{\text{th}} = \sqrt{kTC}/q$.

We also highlight work towards improving the scalability of "weight-stationary" optical neural networks based on Mach-Zehnder interferometer (MZI) meshes [3], which may be a highly efficient approach for mid-scale DNNs. Such meshes are limited by hardware (fabrication) imperfections, which cause errors in the realized matrix (Fig. 2(a-b)). Previously, we developed hardware error correction techniques to partially correct for these errors through local perturbations or self-configuration [4], although the scalability of such schemes was limited by the presence of MZI "forbidden regions". Recently, we developed a new 3-MZI architecture (Fig. 2(c)) that removes this obstacle through a Möbius transformation, improving the accuracy of hardware error correction by orders of magnitude (Fig. 2(d)), and becomes asymptotically perfect as the mesh size is increased [5].

In conclusion, we have presented two schemes to improve the scalability of photonic computation—Netcast and hardware error correction—that will open new paths for photonics in SWaP-constrained edge computing.

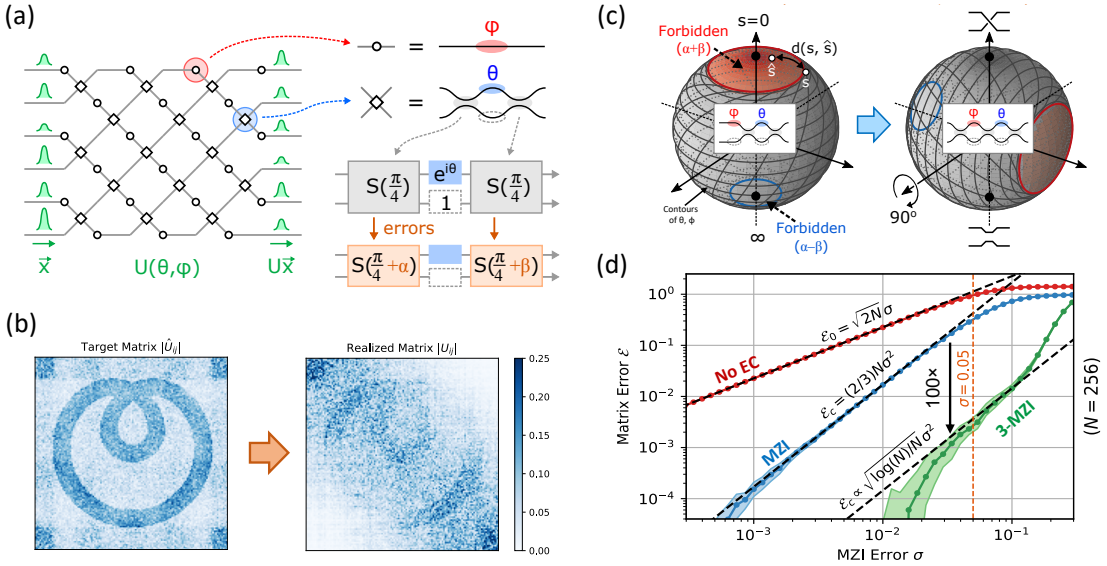

Fig. 2. Photonic error correction. (a) MZI mesh implemented with imperfect components, (b) poor resulting matrix fidelity, (c) 3-MZI scheme to cure errors, (d) simulated accuracy of MZI vs. 3-MZI.

References

1. A. Sludds, S. Bandyopahyay, Z. Chen, Z. Zhong, J. Cochrane, L. Bernstein et al., *Science* **378**, 270–276 (2022).
2. R. Hamerly, A. Sludds, S. Bandyopadhyay, Z. Chen, Z. Zhong et al., arXiv:2207.01777 (2022).
3. S. Bandyopadhyay, A. Sludds, S. Krastanov, R. Hamerly, N. Harris et al., arXiv:2208.01623 (2022).
4. S. Bandyopadhyay et al., *Optica* **8**, 1247 (2021); R. Hamerly et al., *Phys. Rev. Appl.* **18**, 024019 (2022).
5. R. Hamerly, S. Bandyopadhyay, and D. Englund, *Nat. Commun.* **13**, 6831 (2022).