Demonstration of WDM-Enabled Ultralow-Energy Photonic Edge Computing

Alexander Sludds[1,*], **Ryan Hamerly**[1,2], **Saumil Bandyopadhyay**[1], **Zhizhen Zhong**[3], **Zaijun Chen**[1], **Liane Bernstein**[1], **Manya Ghobadi**[3], **Dirk Englund**[1]

[1] *Research Laboratory of Electronics, MIT, 50 Vassar Street, Cambridge, MA, 02139, USA*
[2] *NTT Research Inc., Physics and Informatics Laboratories, 940 Stewart Drive, Sunnyvale, CA 94085, USA*
[3] *Computer Science and Artificial Intelligence Laboratory, MIT, 32 Vassar St, Cambridge, MA, 02139, USA*
[*] *e-mail address: asludds@mit.edu*

Abstract: We present experimental demonstrations of ultra-low power edge computing enabled by wavelength division multiplexed optical links and time-integrating optical receivers. Initial experimentation demonstrations show $\lesssim$ 10 fJ of optical energy per MAC.

1. Introduction

Machine learning models have revolutionized the performance of tasks in many fields such as computer vision [1] and game playing [2, 3]. These models are ubiquitously deployed in data centers where computate resources are abundant, but for many applications, computation needs to be kept close to the end user at the "edge" of the communications network. Unlike devices in data centers, edge devices are heavily size, weight, and power (SWaP) constrained. Modern machine learning models are large, possessing hundreds of millions of weighting values, which limits the ability of current digital hardware to run these models at the edge [4], a problem exacerbated by the von Neumann data-movement bottleneck in modern microprocessors. Here, we alleviate this bottleneck with an integrated photonic architecture, named Netcast, utilizing wavelength division multiplexing and time integrating receivers to shift the burden of data movement away from the edge computing devices.

2. Architecture and Experiment

Silicon photonics (SiPh) has emerged as a promising platform for reconfigurable, large-scale, high-speed, and energy-efficient systems [5]. Prior work on machine learning using SiPh has considered architectures which are "weight stationary" where weighting values are statically programmed into an array of devices [6, 7]. While these architectures are promising for datacenter deployments, they still require all weighting values to be kept at the system itself. An alternative architecture with an "output stationary" dataflow, where outputs are accumulated in the time domain at a coherent receiver, was recently proposed [8]. Output stationary architectures allow for substantial increases in chip area efficiency since inner products in vector-matrix computation can be cast into the time domain rather than spatial domain. In addition, output-stationarity allows inference at the edge to be split

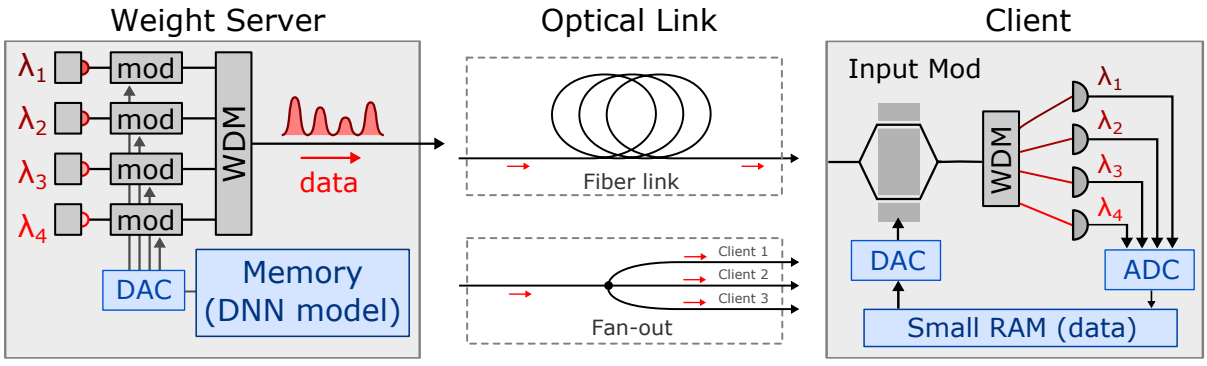

Fig. 1. A bank of lasers generate an array of distinct wavelengths. Weight data is encoded onto each wavelength through an array of optical modulators and combined onto a fiber link using wavelength multiplexing. Input activation values are applied at the client through a single broadband modulator. All wavelengths are separated using a wavelength demultiplexer before time integration and readout.

between a server and SWaP-constrained client. This is illustrated in the Netcast scheme of Fig. 1, which makes use of many frequency modes in a single-mode optical link (such a fiber) to distribute weighting values [9].

To demonstrate the potential of the proposed architecture, we have created a proof-of-principle demonstration comprised of a large SiPh transmitter, wavelength division multiplexer, and a time integrating receiver which are all shown in figure 2. The SiPh transmitter, shown in figure 2 (a), was fabricated in a CMOS pilot line at the OpSIS-IME foundry and consists of 48 high-speed Mach-Zender interferometers utilizing the free-carrier plasma dispersion effect in silicon each capable of modulation at speeds in excess of 50Gbps [10]. A bank of tunable lasers is used as the light source for the system. Light is coupled into the silicon chip using a 64-channel fiber array coupled to on-chip grating couplers. Weight data are modulated onto each wavelength of light using the SiPh transmitter before being combined through a fiber multiplexer onto a deployed fiber link. At the receiver (figure 2 (b)), a single modulator (LN fiber modulator) is used to simultaneously encode the value of a single input activation onto all wavelengths equally. A demultiplexer is then used to separate each wavelength and send it to individual integrating photodetectors.

As a first demonstration we configure a single unitcell of the SiPh transmitter driven with 4 distinct C-band wavelengths as our weight server. A 10km bus fiber is used to simulate deployment over realistic distances. A commercial traveling-wave Lithium Niobate (LN) fiber modulator (JDSU AM-150)is used as the client side modulator. Commercial fiber wavelength multiplexers are used to separate out each wavelength at the receiver. For initial characterization bulk photodiodes are used. With this setup in place we consider the computation accuracy of the system by encoding uniformly distributed random floating-point values onto a weight and input modulator. The resulting value from the analog computation is compared to the correct digital computed value and a difference distribution is created figure 2 (c). This error distribution shows an error of $\sigma \approx 0.005$ corresponding to $7-8$ bits of precision. With this measurement in place, a benchmark neural network trained on the MNIST dataset is run through the system, with the resulting confusion matrix shown in figure 2 (d). The optical hardware obtains an accuracy of 92.6%, comparable to the original digital accuracy of 94.4%. To understand the sources

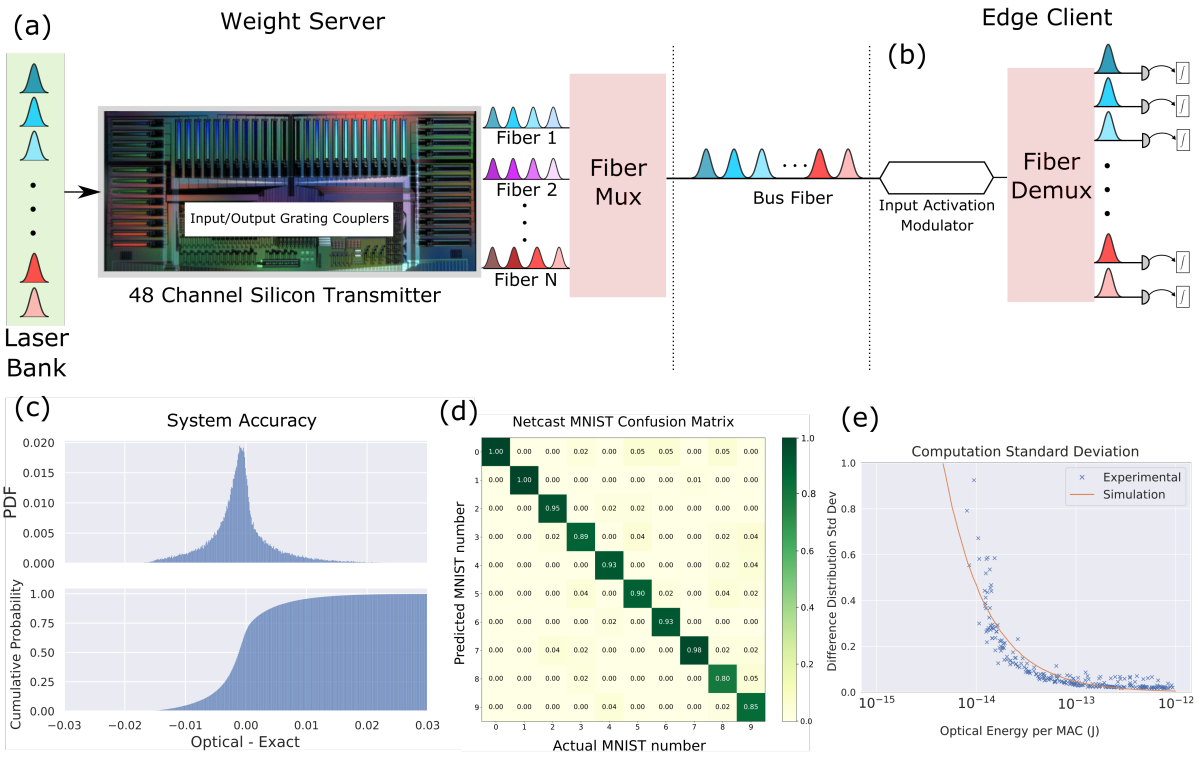

Fig. 2. (a) A bank of tunable lasers acts as the optical source for the system. A 48 channel SiPh transmitter chip is used to generate the weighting values. Individual modulated wavelengths are combined using a fiber multiplexer and deployed over a bus fiber. (b) All wavelengths are simultaneously modulated by a single input activation modulator before being demultiplexed and sent to individual time integrating receivers.(c-d) Example computation results realized by this system with only 1.8% accuracy drop on the MNIST task relative to it's digital baseline. (e) Accuracy of optical computation as a function of optical energy per multiplication

of error in a system the light at the receiver is attenuated by a programmable optical attenuator (JDSU ha9) and sent into a characterized photodiode (Thorlabs PDA10CS). The receiver attenuation is increased and the standard deviation of the error probability density function generated is calculated. This creates a scatter plot showing the error in computation as a function of the optical energy required for each multiplication and accumulation operation (MAC). Shown in orange is the expected error in computation which is calculated using the signal-to-noise ratio of incident optical power over the photodiodes output noise. This demonstration shows that non-integrating photoreceivers allow for $\sim 10\frac{\text{fJ}}{\text{multiply}}$ of optical energy.

3. Discussion

The above experimental demonstration shows the promise of the Netcast architecture to realize ultra-low energy and high speed computation. An experimental demonstration shows that off-the-shelf hardware can enable highly accuracy computation on the MNIST dataset using broadband wavelength multiplexing and accurate computation with $\sim$ 10fJ of optical energy per MAC. To understand the importance of this number we must consider the two critical figures of merit of the system: client energy consumption and optical energy per MAC. First, we consider the client energy consumption from optical and electrical components. These components are: a single optical modulator to encode input activations, a digital-to-analog converter (DAC) and analog-to-digital converter (ADC) for encoding and readout, and a small memory for storing input activations. These costs can all be heavily amortized because of frequency multiplexing and time multiplexing. A standard silicon photonic MZI uses $\sim$ 1pJ per value, but this one modulation usage is amortized over the N wavelengths at the receiver [11]. DACs and ADCs are also $\sim$ 1pJ per usage, which is amortized over the number of incident wavelengths and number of integration timesteps respectively [12, 13]. Electrical memory costs of DRAM are $\sim$ 1pJ per value [4]. Assuming a square weight matrix of size $N \times N$ near-term electrical energy consumption can approach $\sim 1\frac{\text{fJ}}{\text{multiply}}$ for $N = 1000$. In the future, the proposed system can be further scaled by making use of custom integrated electronics to create a weight server capable of outputting data over the entire C and L communication bands (95nm/ 11THz of bandwidth). Emerging technologies, such as optical frequency combs and organic electro-optic polymers, are CMOS compatible technologies which can enable large scale systems in a fully-integrated platform [14, 15].

References

1. A. Krizhevsky et al, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.
2. O. Vinyals et al, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, 2019.
3. D. Silver et al, "Mastering the game of go with deep neural networks and tree search," *Nature*, 2016.
4. V. Sze et al, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, 2017.
5. W. Bogaerts et al, "Programmable photonic circuits," *Nature*, 2020.
6. Y. Shen et al, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, 2017.
7. A. Tait et al, "Broadcast and weight: an integrated network for scalable photonic spike processing," *Journal of Lightwave Technology*, 2014.
8. R. Hamerly et al, "Large-scale optical neural networks based on photoelectric multiplication," *Physical Review X*, 2019.
9. R. Hamerly et al, "Edge computing with optical neural networks via wdm weight broadcasting," in *Emerging Topics in Artificial Intelligence (ETAI) 2021*, International Society for Optics and Photonics, 2021.
10. M. Streshinsky et al, "Silicon parallel single mode 48× 50 gb/s modulator and photodetector array," *Journal of Lightwave Technology*, 2014.
11. J. Witzens, "High-speed silicon photonics modulators," *Proceedings of the IEEE*, 2018.
12. B. Murmann, "The race for the extra decibel: A brief review of current adc performance trajectories," *IEEE Solid-State Circuits Magazine*, 2015.
13. P. Caragiulo et al, "Dac performance survey 1996-2020."
14. A. Rizzo et al, "Integrated kerr frequency comb-driven silicon photonic transmitter," *arXiv:2109.10297*, 2021.
15. U. Koch et al, "A monolithic bipolar cmos electronic–plasmonic high-speed transmitter," *Nature Electronics*, 2020.