Photonic Multiply-Accumulate Operations for Neural Networks

Mitchell A. Nahmias

implementation of multiply-accumulate operations (which take the form $\spadesuit' \leftarrow \spadesuit + w \cdot \ \ $) in various platforms in Section III, discussing the costs and benefits of digital electronics, analog electronics, and photonics. We provide a comparison of the fundamental limits of electronic crossbar arrays and photonic linear computing systems in Section IV, and analyze the performance of these models across of metrics such as energy, speed, and computational density. We consider the general performance of photonic MACs along these metrics based on practical devices that are compatible with large-scale silicon photonic foundries. In the last section, we provide a concrete comparison between fully-tunable neuromorphic photonic networks based on known photonic device models and principles with electronic state-of-the-art deep learning chips.

II. Multiply-Accumulate Operations

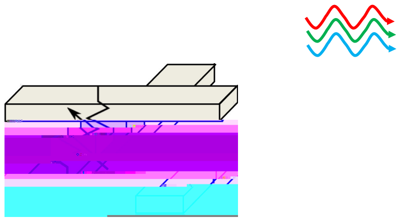
The multiply-accumulate (MAC) operation calculates the product of two numbers and adds the result to an accumulator. For a given accumulation variable $\spadesuit$ and modified state $\spadesuit'$, the operation takes the following form:

$$\spadesuit' \leftarrow \spadesuit + (w \times \quad) \tag{1}$$

MACs are constituents of a number of linear mathematical operations, including dot products, matrix multiplications, Fourier transforms, and convolutions. MACs have traditionally characterized the performance signal processing (DSP) applications [9], [10], but have become increasingly prominent in modern HPC.

We are most interested in a specific use case: the simulation of neural network models. AI applications typically divide into *training*, in which models learn to understand a data set, and *inference*, in which trained models are deployed on new data to draw conclusions or extract information. For a set of input variables and output variables , each node domain to the photonic domain and back. Waveguides can thus beat metal wires in efficiency, provided that the cost of / / conversion is less than that of charging a metal wire over the same distance.

It is not yet clear whether addressing the data movement problem alone is worthwhile—we still pay the / / cost ($\sim$

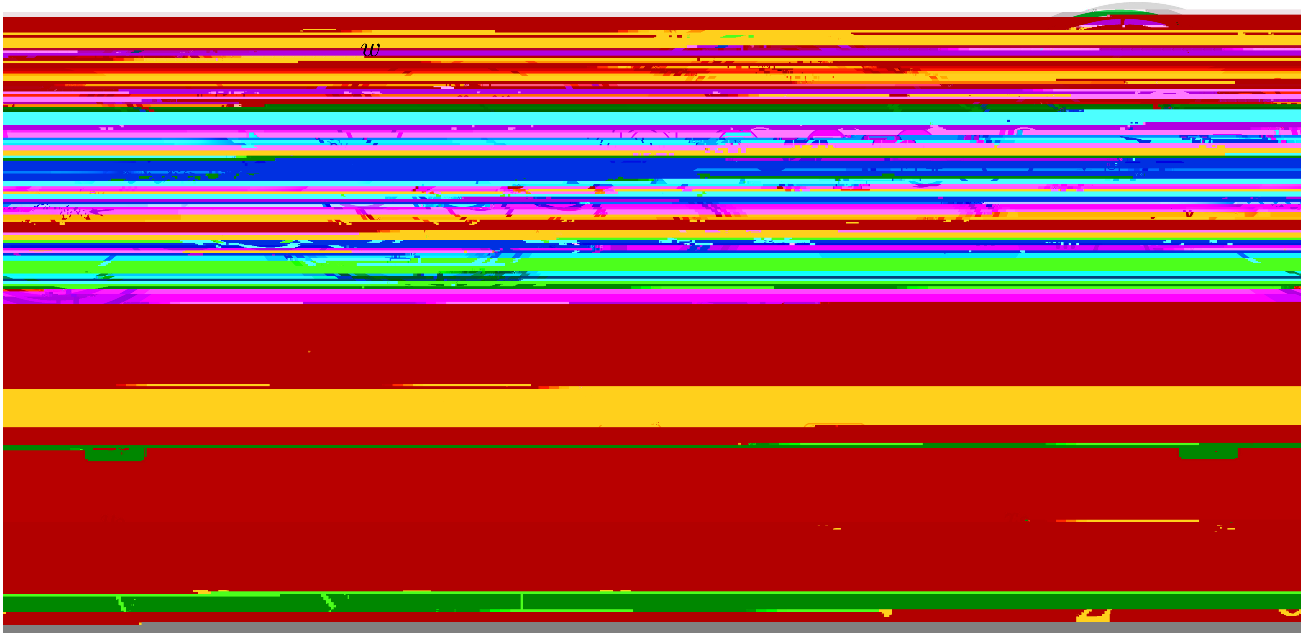


(a)

(b)

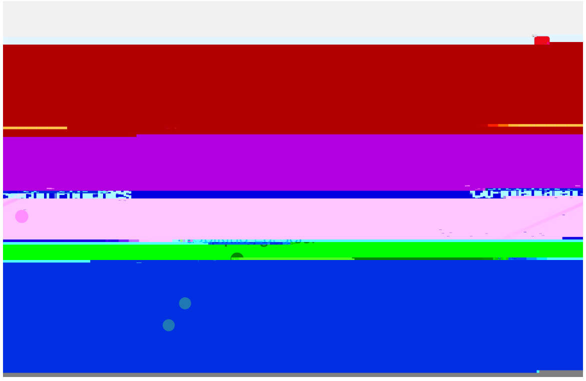
compute core must simultaneously address both processing within the core (i.e., an efficient implementation of a MAC operation $+ w \times$) and data movement *across* the core (i.e., each MAC operation requires a result from a prior MAC unit in order to perform a full dot product w at the end of each row). As we will see, the data movement constraint can bound the performance of each of the cores.

We assume that there is a tunable, resistive element at the interface between metal crossbars, and each tunable element emulates a simple resistor associated with a fixed weight w. Kirchhoff's current law performs the summation w with the weights within each matrix, determined by the relative resistance values along each wire. A standard formula for assessing the bandwidth of on-chip metal interconnects is $B \leq B \quad A/L^2$ per wire, for constant bit rate B , architecture-dependent constant B (typically $B \sim 10^{16}$ for on-chip RC interconnects [62]), cross sectional area A, and length L of the wire. Extending this analysis to crossbars, we make the simple observation that the area occupied by each resistive element is approximatelyresistivei


w

V. PHOTONIC MULTIPLY-ACCUMULATE O

TABLE II
COMPARISON OF ELECTRONIC ARCHITECTURES (TOP) WITH ESTIMATES FOR VARIOUS PHOTONIC NEURAL NETWORK (NN) APPROACHES (BOTTOM). DENSITY IS COMPUTED WITH RESPECT TO THE CORE(S) ONLY.



where η η $\eta_{\boldsymbol{w}}$ η is laser efficiency, photonic link efficiency, and photodetector efficiency, respectively; is the inverse slope of the modulator's voltage-to-transmission curve $(\quad)$; and $C_{\mathrm{mod}}, C_{\mathrm{PD}}$ are the joint capacitances of the photodetetor and modulator. In a typical foundry-model where $(C_{\mathrm{mod}} + C_{\mathrm{PD}}) \sim$ 70 fC and $\eta \sim .0$ (which includes the passive losses through the weight banks, which can be made quite small [116]), even with ρ N in fixed point systems, we arrive at a floor of approximately E

[35] S. Sukhbaatar and R. Fergus, “Learning from noisy labels with deep neural networks,” 2014, *arxiv preprint arXiv:1406.20802.3*.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[37] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, “Regularization of neural networks using dropconnect,” in *Proc. 30th Int. Conf. Mach. Learn.*, Jun. 17–19, 2013, pp. 1058–1066. [Online]. Available: http://proceedings.mlr.press/v28/wan13.html

[38] C. M. Bishop, “Training with noise is equivalent to tikhonov regularization,” *Neural Comput.*, vol. 7, no. 1, pp. 108–116, Jan. 1995. [Online]. Available: https://doi.org/10.1162/neco.1995.7.1.108

[39] A. Neelakantan *et al.*, “Adding gradient noise improves learning for very deep networks,” 2015, *arXiv:1511.06807*.

[40] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Quantized neural networks: Training neural networks with low precision weights and activations,” *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6869–6898, Jan. 2017.

[41] S. Agarwal *et al.*, “Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding,” *Frontiers Neurosci.*, vol. 9, pp. 484–493, 2016. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2015.00484

[42] S. Galal and M. Horowitz, “Energy-efficient floating-point unit design,” *IEEE Trans. Comput.*, vol. 60, no. 7, pp. 913–922, Jul. 2011.

[illegible] received the B.S. (Hons.) degree in electrical engineering with a Certificate in Engineering Physics and the M.A. degree in electrical engineering in 2012 and 2014, respectively, from Princeton University, Princeton, NJ, USA, where he is currently working toward the Ph.D. degree with the Princeton Lightwave Communications Laboratory.