Edge computing with optical neural networks via WDM weight broadcasting

PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Edge computing with optical neural networks via WDM weight broadcasting

Ryan Hamerly, Alexander Sludds, Saumil Bandyopadhyay, Liane Bernstein, Zaijun Chen, et al.

Ryan Hamerly, Alexander Sludds, Saumil Bandyopadhyay, Liane Bernstein, Zaijun Chen, Manya Ghobadi, Dirk Englund, "Edge computing with optical neural networks via WDM weight broadcasting," Proc. SPIE 11804, Emerging Topics in Artificial Intelligence (ETAI) 2021, 118041R (2 September 2021); doi: 10.1117/12.2594886

SPIE.

Event: SPIE Nanoscience + Engineering, 2021, San Diego, California, United States

Edge Computing with Optical Neural Networks via WDM Weight Broadcasting

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

[1]Research Laboratory of Electronics, MIT, 50 Vassar Street, Cambridge, MA 02139, USA
[2]NTT Research Inc., PHI Laboratories, 940 Stewart Drive, Sunnyvale, CA 94085, USA
[3]Computer Science and AI Laboratory, MIT, 32 Vassar Street, Cambridge, MA 02139, USA

ABSTRACT

We introduce an optical neural-network architecture for edge computing that takes advantage of wavelength multiplexing, high-bandwidth modulation, and integration detection. Our protocol consists of a server and a client, which divide the task of neural-network inference into two steps: (1) a difficult step of optical weight distribution, performed at the server and (2) an easy step of modulation and integration detection, performed at the edge device. This arrangement allows for large-scale neural networks to be run on low-power edge devices accessible by an optical link. We perform simulations to estimate the speed and energy limits of this scheme.

Keywords: Edge computing, split computing, neural networks, WDM

1. INTRODUCTION

Machine learning has become ubiquitous in cloud computing and data centers, but in recent years, network and privacy constraints are pushing processing closer to the end user.[1] In this "edge computing" paradigm, the majority of data processing is done on size, weight, and power (SWaP)-constrained smart sensors and end stations, rather than in the data center. Since a growing fraction of data processing relies on deep neural networks (DNNs), great effort has gone into developing SWaP-constrained hardware[2,3] and efficient models[4,5] for DNN inference at the edge. However, many state-of-the-art DNNs[6,7] are now so large that they can only be run in a data center, as both the model size and energy consumption exceed the SWaP constraints for realistic edge devices. Such DNNs cannot be run on the edge with conventional hardware, limiting their utility in situations where latency, network bandwidth, and security are paramount.

Many emerging technologies are being explored to tackle this problem. Memristive circuits[8] are a leading candidate that leverage standard electronics processes, but practical issues including device nonuniformity, bits of precision, and accurately updating resistance values remain major challenges.[9] Photonic architectures are considered promising due to the correspondence between matrix multiplication (the computational bottleneck of DNNs) and passive optical propagation. However, photonic approaches suffer from limited scalability: free-space systems[10,11] offer large numbers of modes but limited connectivity, while on-chip architectures based on interferometery[12,13] and arrays of ring resonators[14] achieve all-to-all connectivity, but chip-area constraints (and error propagation in deep circuits[15–18]) make the application to large large neural networks very challenging. As the limitations to most photonic architectures stem from their weight-stationary[2] nature, where each synaptic weight is mapped to a discrete photonic device, recently we proposed an alternative scheme based on output-stationary coherent detection and integration,[19,20] where weights are encoded optically in the time domain and the chip area scales with the number of neurons, not synapses (a scaling of $O(N)$ rather than $O(N^2)$). The encoding of synaptic weights in the optical domain suggests an edge computing architecture where weights are delivered 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 multimode imaging requirements of Ref.[19] render this particular scheme impractical for such an architecture.

Further author information: (Send correspondence to R.H.)
R.H.: E-mail: rhamerly@mit.edu

Emerging Topics in Artificial Intelligence (ETAI) 2021, edited by Giovanni Volpe, Joana B. Pereira, Daniel Brunner, Aydogan Ozcan, Proc. of SPIE Vol. 11804, 118041R · © 2021 SPIE
CCC code: 0277-786X/21/$21 · doi: 10.1117/12.2594886

This paper introduces NetCast, a new optical server-client protocol based on wavelength-division multiplexing (WDM), difference detection and integration, and optical weight delivery. Our protocol involves two components: a "weight server" consisting of a WDM bank of analog modulators, connected by an optical link to a SWaP-constrained client consisting of a single modulator, integration detectors, and passive demultiplexing optics. 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. We investigate the speed and energy limits to this protocol: crosstalk modeling reveals that the server-client bandwidths may be comparable to GPU high-bandwidth memory, while simulations of noisy DNN inference lead to estimates of power and link loss tolerance. The advantages of NetCast stem from the high bandwidth of WDM optical links and the ability to postprocess optically encoded data with minimal power overhead, effectively pushing the hard part of the computation to the server. These features enable low-power neural-network inference at the edge for DNNs of arbitrary size, unbounded by the power or memory constraints of edge devices.

Fig. 1 illustrates the concept. The architecture consists of a *server* and a *client*, connected by an optical *link*. As linear algebra is the rate-limiting step for DNNs, here we focus on how NetCast accelerates matrix multiplication $y_m = \sum_n w_{mn} x_n$ (the activations and pooling operations between layers 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)). This signal is received by the client, where it passes through a broadband modulator encoding the activations x_j, and is then 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 $\vec{a}_{mn} = \big(\cos(\phi_{mn}), i\sin(\phi_{mn})\big)$. The signals encode the weight through differential signaling: ϕ_{mn} is set so that $w_{mn} = |a_{mn,1}|^2 - |a_{mn,2}|^2$. 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,

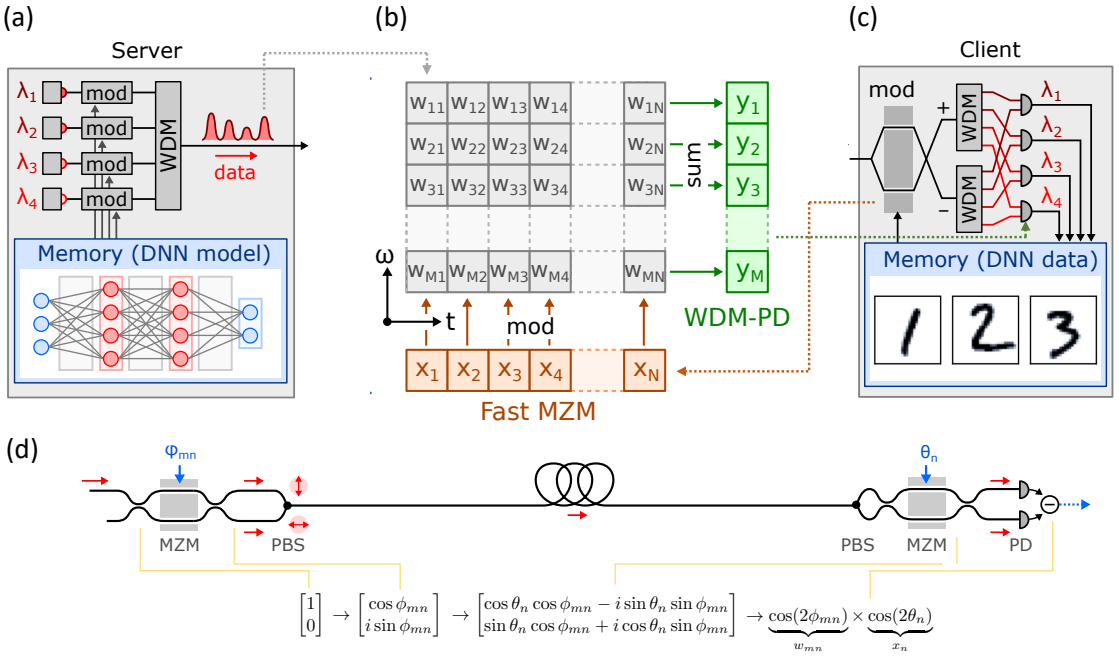

Figure 1. NetCast concept, consisting of (a) a weight server, which (b) encodes the weight matrix into a time-frequency basis. The client (c) receives this signal and uses it to compute a matrix-vector product. (d) Detailed dataflow for a Mach-Zehnder implementation, showing a single frequency channel.

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:

$$Q_m \propto \sum_n \cos(2\phi_{mn}) \cos(2\theta_n) \tag{1}$$

In this way, with the encodings $\phi_{mn} = \frac{1}{2}\cos^{-1}(w_{mn})$ and $\theta_n = \frac{1}{2}\cos^{-1}(x_n)$, the client will generate 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 not that photonics provides a means to multiply numbers, but that we can separate the tasks of logic (client) and memory access (server) using an optical link. The client and server must cooperate to perform the computation, but the workloads are not equal. For large DNNs in particular, the energy and memory costs at the client are dwarfed by that at the server:

- For an $N \times N$ fully connected layer, the number of memory reads scales as $O(N^2)$ at the server and only $O(N)$ at the client. In addition, the server must drive N modulators for N steps (an $O(N^2)$ energy cost), while the client drives only a single modulator ($O(N)$ cost) and reads out from the detectors only once ($O(N)$ cost).
- The memory requirement at the server scales as $O(N^2)$ times the number of layers, while the client needs a much smaller memory of size $O(N)$. This is consistent with the general fact that the number of synapses in a brain is much larger than the number of neurons.

By means of an optical link, NetCast pushes all the costly parts of the computation back to 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.

2. ENERGY EFFICIENCY

Since the total number of operations scales as $O(N^2)$, NetCast does not yield an improvement in *total* energy efficiency compared to running a DNN digitally. However, the client-side power consumption is reduced significantly. As discussed above, the electrical energy consumption for an $N \times N$ layer will scales as $O(N)$, which translates to an energy per MAC scaling of $O(N^{-1})$. Realistic figures based on current technology ($O(1)$ pJ/sample for modulation, DAC, and ADC[19,21–24]), suggest fJ/MAC (client side) performance with matrix sizes $N \geq 100$, which is several orders of magnitude below the current CMOS state of the art.[25,26]

It is also important to consider the optical energy consumption, as this is often tied to fundamental limits on the performance of photonic hardware.[19,27] The optical power sets the signal-to-noise ratio of the client's detectors; this noise manifests itself as a Gaussian error term in the analog matrix-vector multiplication:

$$y_m = \sum_n w_{mn}x_n + N(0, \sigma^2), \quad \sigma = \sqrt{\sigma_J^2 + \sigma_S^2} \tag{2}$$

(w_{mn}, x_n are restricted to the interval $[-1, 1]$). The terms σ_J and σ_S correspond to Johnson (kTC) and shot noise, which scale as:

$$\sigma_J^2 = \frac{(kTC/e^2)}{N_{\text{tr}}^2} P_{\text{tr}}^2, \quad \sigma_S^2 = F_{\text{tr}} \frac{N}{N_{\text{tr}}} \tag{3}$$

Here N_{tr} is the average number of transmitted photons per weight, N is the matrix size, and P_{tr} and F_{tr} are dimensionless constants that depend on the encoding scheme.

As the performance of DNNs is degraded with excess noise, one can determine an empirical lower bound to the required optical power by running benchmark neural networks on the device. Here, we simulate DNN inference on a NetCast client running 3-layer perceptrons trained on MNIST classification (the same networks used on Ref.[19]). Five encoding schemes are studied (Fig. 2(a)): "simple" incoherent encoding based on difference detection $w = |a_1|^2 - |a_2|^2$, "low-noise" encoding with an additional intensity modulator (to reduce the shot

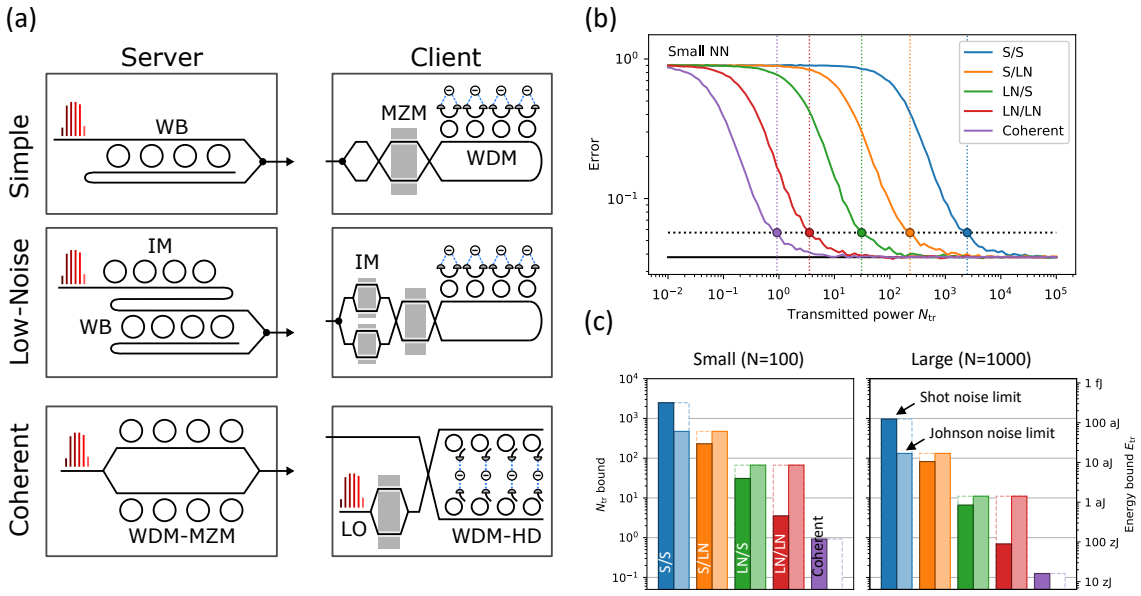

Figure 2. (a) Illustration of simple, low-noise, and coherent NetCast variants (employing rings instead of MZMs, although the result is the same for both). (b) DNN classification error as a function of transmitted power N_{tr}. (c) Energy limits set by Johnson ($C = 0.1$ pF) and shot noise.

noise for small weight values), and coherent encoding. Fig. 2(b) plots the MNIST classification error for each encoding scheme as a function of transmitted power N_{tr}. From this, we define the cutoff as the point where the error reaches 50% above the zero-noise value; this is plotted for the different schemes in Fig. 2(c).

Since the neural networks studied here have many small weights and activations ($|x_n|, |w_{mn}| \ll 1$), the low-noise and coherent schemes, which alleviate the problem of noisy difference detection of small signals, greatly reduce the required optical energy relative to a simple scheme (identical to Fig. 1(d)). Even in the absence of shot noise, the coherent scheme enjoys the lowest optical energy bound. Not surprisingly, consistent with our prior work on optical neural networks,[19] accurate inference is possible even with few or even < 1 photons per MAC, since the SNR is set by the integrated charge, not the charge accumulated by a single pulse.

While such low energy figures are not relevant for a device's power budget, they illustrate the robustness to link loss, suggesting that NetCast can perform efficient inference even in photon-starved environments such as long free-space links between the server and client.

3. THROUGHPUT LIMIT

Since time and frequency are noncommuting operators, the time-frequency bins of Fig. 1(b) are non-orthogonal, leading to inevitable crosstalk between matrix elements. This crosstalk ultimately limits the available (analog) throughput of the channel. For concreteness, focusing on ring-based modulators and multiplexers (Fig. 2(a)), the crosstalk takes two forms, which can be calculated analytically:

1. Temporal crosstalk χ_t, which arises from the finite photon lifetime of the ring modulators and their RC time constant, which can be lumped into an approximate modulator response time $\tau = \sqrt{1/\kappa^2 + (RC)^2} \approx \sqrt{2}/\kappa$ for efficient modulators ($RC \approx \kappa$). For a fixed χ_t, the symbol rate is bounded by:

$$R \leq \frac{\kappa}{\sqrt{2}\log(1/\chi_t)} \tag{4}$$

2. Frequency crosstalk χ_ω, which arises from overlapping neighboring WDM channels when the channel spacing is small. For a fixed χ_ω, one derives a minimum channel spacing:

$$\Delta\omega \geq \frac{\kappa}{2\sqrt{\chi_\omega}} \tag{5}$$

Combining these and assuming $\chi_t \approx \chi_\omega \equiv \chi$, the channel capacity is bounded by:

$$C = R\frac{2\pi B}{\Delta\omega} \leq \frac{2\pi\sqrt{2\chi}}{\log(1/\chi)} \equiv C_0 B \tag{6}$$

where B is the optical bandwidth (in Hz) and C_0 is the normalized symbol rate (units 1/Hz-s).

The maximum allowed crosstalk depends on the network and must be determined empirically. In recent work, analyzed the accuracy of both AlexNet and MNIST DNNs in the presence of crosstalk, finding minimal added error up to $\chi = 5$–10%.[28] This corresponds to a normalized symbol rate of $C_0 = 0.66$–1.22, or 3–6 Twt/s (20–40 Tbps at 8 bits/wt) if the entire C-band is used ($B = 4.4$ THz). By way of comparison, these values are larger than the data bandwidths available to the high-bandwidth memory (HBM) of top workstation GPUs.[29]

4. CONCLUSION

As computing moves to the edge, optics can open up new possibilities to deliver high performance while simultaneously adhering to strict SWaP constraints. In this paper, we have introduced NetCast, a server-client architecture that leverages unique advantages of optics—the high bandwidth of fiber links, support for wavelength multiplexing, and analog integration detection—to split the DNN inference problem into two tasks: weight retrieval and encoding at the server and lightweight optical postprocessing at the client, effectively pushing the energy- and memory-intensive tasks to the server. We analyzed the throughput and energy-efficiency bounds for this scheme using simulations based on pre-trained DNNs, which suggest that NetCast should support data bandwidths comparable to high-end GPU HBM, with optical energy consumption that can theoretically reach the sub-attojoule regime. This high theoretical performance limits suggest NetCast is a promising approach to optical information processing that merits further study.

ACKNOWLEDGMENTS

This research is funded by a grant from NTT Research Inc. and NSF EAGER no. 1946967.

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] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits* **52**(1), pp. 127–138, 2017.

[4] 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.

[5] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360* , 2016.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[7] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860* , 2019.

[8] O. Krestinskaya, A. P. James, and L. O. Chua, "Neuromemristive circuits for edge computing: A review," *IEEE transactions on neural networks and learning systems* **31**(1), pp. 4–23, 2019.

[9] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. Di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. Farinha, *et al.*, "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature* **558**(7708), pp. 60–67, 2018.

[10] E. G. Paek and D. Psaltis, "Optical associative memory using Fourier transform holograms," *Optical Engineering* **26**(5), p. 265428, 1987.

[11] N. J. New, "Reconfigurable optical processing system," Mar. 14 2017. US Patent 9,594,394.

[12] 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.

[13] M. Prabhu, C. Roques-Carmes, Y. Shen, N. Harris, L. Jing, J. Carolan, R. Hamerly, T. Baehr-Jones, M. Hochberg, V. Čeperić, *et al.*, "Accelerating recurrent ising machines in photonic integrated circuits," *Optica* **7**(5), pp. 551–558, 2020.

[14] 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.

[15] 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.

[16] S. Bandyopadhyay, R. Hamerly, and D. Englund, "Hardware error correction for programmable photonics," *arXiv preprint arXiv:2103.04993* , 2021.

[17] R. Hamerly, S. Bandyopadhyay, and D. Englund, "Stability of self-configuring large multiport interferometers," *arXiv preprint arXiv:2106.04363* , 2021.

[18] R. Hamerly, S. Bandyopadhyay, and D. Englund, "Accurate self-configuration of rectangular multiport interferometers," *arXiv preprint arXiv:2106.03249* , 2021.

[19] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," *Physical Review X* **9**(2), p. 021032, 2019.

[20] 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.

[21] D. A. Miller, "Energy consumption in optical modulators for interconnects," *Optics Express* **20**(102), pp. A293–A308, 2012.

[22] A. H. Atabaki, S. Moazeni, F. Pavanello, H. Gevorgyan, J. Notaros, L. Alloatti, M. T. Wade, C. Sun, S. A. Kruger, H. Meng, *et al.*, "Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip," *Nature* **556**(7701), p. 349, 2018.

[23] B. E. Jonsson, "An empirical approach to finding energy efficient ADC architectures," in *Proc. of 2011 IMEKO IWADC & IEEE ADC Forum*, pp. 1–6, 2011.

[24] S. Cosemans, B. Verhoef, J. Doevenspeck, I. Papistas, F. Catthoor, P. Debacker, A. Mallik, and D. Verkest, "Towards 10000TOPS/W DNN inference with analog in-memory computing–a circuit blueprint, device options and requirements," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 22–2, IEEE, 2019.

[25] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*, pp. 10–14, IEEE, 2014.

[26] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Computer Architecture (ISCA), 2017 ACM/IEEE 44th Annual International Symposium on*, pp. 1–12, IEEE, 2017.

[27] S. Garg, J. Lou, A. Jain, and M. Nahmias, "Dynamic precision analog computing for neural networks," *arXiv preprint arXiv:2102.06365* , 2021.

[28] R. Hamerly, A. Sludds, L. Bernstein, M. Prabhu, C. Roques-Carmes, J. Carolan, Y. Yamamoto, M. Soljačić, and D. Englund, "Towards large-scale photonic neural-network accelerators," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 22–8, IEEE, 2019.

[29] Z. Jia, M. Maggioni, B. Staiger, and D. P. Scarpazza, "Dissecting the NVIDIA Volta GPU architecture via microbenchmarking," *arXiv preprint arXiv:1804.06826* , 2018.