Accuracy/Energy Tradeoffs in Optical Accelerators

Amadou Latyr Ngom
ngom@mit.edu

Adriano Hernandez
adrianoh@mit.edu

Dylan Isaac
disaac@mit.edu

I. Introduction

Deep Neural Networks (DNNs) have driven remarkable progress across vision, language, and other domains, but their ever-growing sizes have placed unprecedented demands on compute and energy requirements. In modern datacenters, the cost of powering and cooling large-scale inference and training workloads is a primary bottleneck, motivating the exploration of novel hardware paradigms that can deliver cost/performance benefits beyond what CMOS alone can offer. Optical accelerators, which exploit photonic properties for efficient broadcast, multicast, and arithmetic operations, hold great promise for breaking through the limits of electronic scaling and reducing inference energy, with minimal accuracy loss.

Among silicon photonic designs, the Albireo architecture [2] stands out as a well–studied accelerator that targets strict accuracy constraints while minimizing energy. By encoding analog weights in Mach-Zehnder Modulators (MZMs) and using microring resonators (MRRs) arrays and photodiodes (PDs) to perform multiply–accumulate operations, Albireo demonstrates how careful co–design of optics, electronics, and mapping strategies can yield high accuracy and performance under tight energy budgets. Other recent systems similarly aim to push down energy per operation while mostly preserving full–precision inference accuracy.

In this work, we take a complementary perspective. Rather than fixing an accuracy target and then optimizing for energy, we systematically explore the full trade-off between energy consumption and inference quality to characterize the accuracy–energy Pareto frontier of optical accelerators. This approach gives hardware designers a tunable knob: they can choose points on the curve that prioritize ultra–low power, or conversely sacrifice some energy savings to maintain high fidelity, or settle on an intermediate balance. By quantifying these trade-offs under realistic noise and quantization configurations, we enable more flexible co-design of optics and workloads to meet the diverse constraints of next-generation DNN workloads.

Our approach begins by taking the Albireo accelerator as a baseline and re-implementing its core optical kernels as fully parameterized PyTorch modules. Each optical kernel can be dropped in for standard `nn.Linear` or `nn.Conv2d`, enabling seamless conversion of most pretrained models into an optical analog. We pay particular attention to modeling the dominant noise sources, exposing them as tunable parameters so that the impact of analog error can be explored end-to-end. The most important of these is the MRR crosstalk, which occurs when adjacent wavelength amplitudes leak into each other.

Next, we integrate these parameters with the CiMLoop framework to obtain fine-grained energy estimates at both component and aggregate levels. Through sensitivity analysis, we isolate three key hyperparameters that dominate the accuracy–energy trade-offs: ADC/DAC (analog-digital/digital-analog conversions) resolution, which governs quantization errors, the MRR cross-coupling coefficient k^2, which affects crosstalk, and the amount of parallelism in the compute units also affects crosstalk. We then perform a systematic grid sweep over these dimensions with a representative DNN workload and chart the resulting accuracy and energy metrics. From these metrics, we propose design guidelines for energy-focused, accuracy-focused, or balanced architectures under various noise sensitivity conditions. We also briefly discuss area usage, though it is not our main focus.

Our key innovations and insights are:

- **Generalizable optical kernels:** Drop-In replacements for Pytorch's compute kernels, allowing to simulate almost any model under an optical architecture. [1].
- **Pareto frontier exploration:** Combined use of CiMLoop and our optical simulations to identify and sweep the key hyperparameters governing energy–accuracy trade-offs.
- **Design guidelines:** A detailed analysis explaining how to optimize for energy/accuracy/balance targets under various noise conditions. Put succinctly, accuracy is primarily optimized by increasing resolution (lower quantization error) and reducing parallelism (lower crosstalk), whereas energy is primarily optimized by reducing resolution and increasing parallelism (higher reuse). Good balance is only possible under low quantization and/or crosstalk sensitivity. Under high sensitivity, any noticeable improvement in one leads to a noticeable degradation in the other.

II. Related Works (Graduate Only)

Optical hardware for neural computation has a long pre-history, but the deep-learning era began only recently. Early holographic neural nets in the late-1980s and early-1990s proved that optics could dense neural networks [19] [20] [21]. Photonic reservoir computers and neuromorphic computing systems begin to support recurrent and spiking models around 2014 [22] [23] [24]. To our knowledge, the first

[1] You can find our code on github: 4gatepylon/65931FinalProject

DNN–class demonstrations—systems that directly mapped multi-layer feed-forward DNNs—arrived in 2017 with the creation of a programmable nanophotonic processor featuring a cascaded array of 56 programmable Mach–Zehnder interferometers in a silicon photonic integrated circuit [13].

Optical computing promises substantial efficiency gains for deep-learning workloads: recent silicon–photonic prototypes report multiply–accumulate (MAC) energies on the order of 10-100 femtojoules and system-level efficiencies on the order of 0.5-200 TOPS per Watt while still performing in the 10-100 Gigahertz (of MACs) range and boasting compute density as high as 100's of TOPS per mm sq. [14] [17]. There is a lot of different work on aspects ranging from materials to light-source to computer architecture and hardware/software co-design, but some key branches of photonic computing research for DNN acceleration are: (1) non-coherent optical computing using microring resonators and wavelength-division multiplexing, which offers superior density and throughput versus (2) coherent optical computing using MZI devices that manipulate amplitude and phase. The research is further organized by target neural network architectures (CNNs, RNNs, Transformers, GNNs) and deployment scenarios (high-performance versus power-constrained edge/IoT environments), with cross-layer optimizations addressing challenges in variations, losses, and electro-optical conversions [25].

Albireo leverages MZI and WDM via optical crossbars, in conjunction with digital logic for other components such as activation functions, to uniquely exploit the reuse of weights in convolutional kernels for CNNs. In general, (electronic) Deep Neural Network (DNN) accelerators face a significant challenge in tackling the energy and latency costs associated with extensive data movement between memory and processing units [3] [5]. While Compute-in-Memory (CiM) architectures, particularly analog CiM (AIMC), aim to mitigate this by computing within memory arrays and controlling noise through a slew of clever interventions [3] [9], they introduce new bottlenecks. Notably, the Analog-to-Digital Converters (ADCs) required to read out analog results often dominate the system's energy consumption and area, scaling exponentially with required resolution [8] [9]. Furthermore, leveraging weight or activation parallel reuse opportunities for different DNN architectures, such as CNNs, can cost a lot of power in the electronic domain if using broadcast and multi-cast operations [2]. Silicon photonic accelerators present an alternative paradigm. Photonics inherently excel at these collective communication primitives by leveraging efficient optical signal splitting. This, combined with the potential for high bandwidth via Wavelength-Division Multiplexing (WDM) and low-energy photonic components, offers a promising path towards scalable and energy-efficient DNN acceleration.

Despite the potential, designing efficient accelerators involves navigating complex tradeoffs. A central challenge, especially for analog and mixed-signal systems including photonics, is the fundamental tradeoff between energy efficiency and computational accuracy [6] [9]. Accuracy in these systems is often limited by non-idealities such as device variations, parasitic effects, and various noise sources. In photonic systems, key limitations arise from laser noise (RIN), shot noise, thermal noise, and particularly optical crosstalk in components like microring resonators (MRRs), which dictates the achievable precision. Many architectures employ techniques like operand slicing or adaptive arithmetic to manage the ADC resolution versus fidelity challenge, alongside balancing array size against utilization [3] [9].

While optical accelerators like Albireo make great strides towards real-world optical compute deployment, they assume strict accuracy preservation constraints; they lack a systematic exploration of the tradeoffs between energy consumption and fidelity, which could enable more flexible accelerator designs. This paper seeks to close this important gap.

III. Approach

A. Simulated Optical Kernels

We begin by reconstructing Albireo's analog optical dot-product primitive in PyTorch. Each simulated kernel performs (1) Weight DAC (digital to electrical) of weights into MZMs, (2) Input DAC (digital to electrical to optical) and wavelength mux/demux with arrayed waveguide gratings (AWGs) and star couplers, (3) parallel element-wise multiplication in MZMs with selective mux/demux via MRRs, (4) optical-to-electrical accumulation in PDs, and (5) ADC (electrical to digital) conversion of the results. By parameterizing every stage, we capture the dominant error sources, namely ADC/DAC bit resolution, MRR and AWG crosstalk, thermal and shot noise, and laser RIN. Together, these steps simulate what happens with photonic locally connected groups and units (PLCGs and PLCUs).

Using the dot-product primitive, we write `OpticalConv` and `OpticalFC` kernels that can be drop-in replacements for PyTorch's fully connected and convolutional kernels, allowing us to automatically convert a wide set of models (e.g., convolutional networks, transformers) into simulated, parameterized optical models.

Most noise models are taken directly from Albireo's published information; however, crosstalk modeling required our own formulation. We assume (following Kamei *et al.* [26]) that immediately adjacent wavelengths dominate interference, so we implement crosstalk as a 1D convolution over indices $(i\pm1)$ and $(i\pm2)$. Under a lossless MRR assumption, the finesse is

$$F \;=\; \frac{\pi\sqrt{1-k^2}}{k^2},$$

and, using a Lorentzian approximation [27], the normalized crosstalk rates become

$$\left(\frac{N_{\text{distinct}}}{2F}\right)^2 \quad \text{for } i\pm1, \qquad \left(\frac{1}{2}\,\frac{N_{\text{distinct}}}{2F}\right)^2 \quad \text{for } i\pm2.$$

We can tune these rates by modifying k^2, which affects the finesse.

In summary, our framework can transform an arbitrary PyTorch model into a highly tunable optical accelerator simulation under realistic noise and conversion effects.

B. Accuracy–Energy Trade-off Exploration

The combined search space for all parameters is too large for an exhaustive search. Therefore, we isolate the key dimensions through targeted analysis. On the accuracy side, Albireo's results show that analog crosstalk is by far the dominant analog noise source. MRR crosstalk scales with the parallelism within PLCUs (e.g., number of parallel output columns). AWG crosstalk further depends on the number of parallel PLCUs within PLCGs and remains secondary unless group-level parallelism is significantly higher. Besides analog errors, quantization errors are governed directly by ADC/DAC bit resolution.

Energy profiling via CiMLoop indicates that ADC/DAC for weights and outputs the most power (affected by bit resolution and parallelism), followed by MRR operations (which depend on both parallelism and k^2) and then laser pumping (proportional to the total number of wavelength-encoded values).

Putting these insights together, we focus our systematic sweep on three key hyperparameters:

- ADC/DAC bit resolution.
- The number of parallel columns per PLCU.
- MRR coupling coefficient k^2, as proxy for finesse.

These choices let us map out the empirical accuracy–energy Pareto frontier for optical DNN inference.

IV. Experiment Setup

We evaluate our framework on ResNet18 by running a grid search with our simulated optical kernels to obtain inference accuracy, and with CiMLoop to extract fine-grained energy breakdowns. For brevity, we focus primarily on (1) top-1 accuracy and (2) energy per MAC, decomposed by major components (ADC/DAC, MRR operations, laser pumping, etc.). We also briefly gloss over cross-entropy losses and area results.

To gauge the impact of each optical configuration, we compare against three baselines: the original FP32 ResNet18, an 8-bit resolution without crosstalk, and a 4-bit resolution without crosstalk. All experiments use the Mini-ImageNet dataset—a 100-class subset of ImageNet [28]—and, due to runtime constraints, we sample 400 validation images at random, which we found sufficient to yield low-variance accuracy estimates.

Here are the hyperparameters of our grid search:

- **Resolution bits:** 4 bits, or 8 bits.
- **PLCU Parallel Columns:** 3, 5, or 7 columns.
- **MRR Crosstalk Level:** $k^2 = 0.02$ (Low) or $k^2 = 0.05$ (High).

V. Results and Discussion

A. Model Quality

fig. 1 summarizes the top-1 accuracy of each configuration on our 400-sample Mini-ImageNet set. Among the baselines, the original FP32 ResNet18 achieves 81.25%, while the noiseless 4-bit and 8-bit resolutions drop to 68.25% and 70.25%, respectively. Introducing crosstalk noise compounds

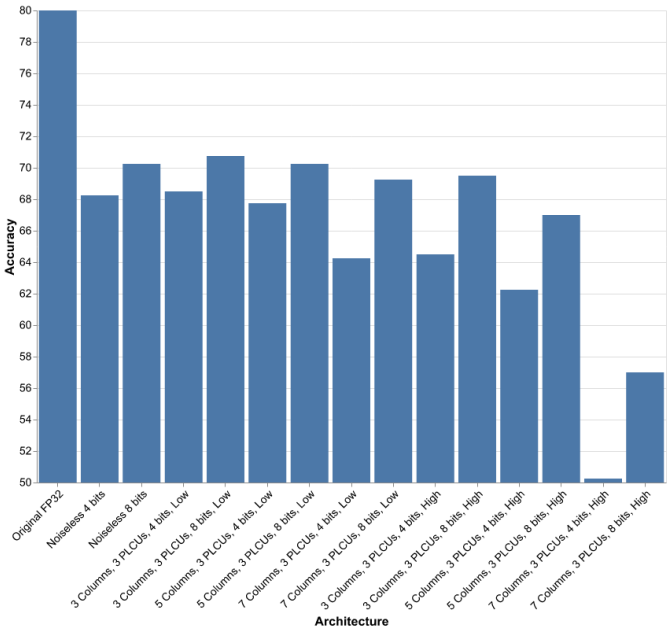


Fig. 1: ResNet18 Accuracies

these losses: as the number of parallel columns increases, analog interference degrades accuracy more severely, particularly under high sensitivity ($k^2 = 0.05$). For example, in the 8-bit, high-noise regime, the 7-column variant attains only 57% accuracy compared to 69.5% for a 3-column setup. By contrast, under low sensitivity ($k^2 = 0.02$), the same 7-column design recovers to 69.25%, which closer the 3-column's 70.75%. The worst performance (50.25%) occurs when both quantization error and crosstalk are maximized (4 bits, 7 columns, high sensitivity). Cross-entropy losses (omitted) follow the same pattern. There are some minor fluctuations that likely stem from our limited sample size.

B. Energy Breakdown

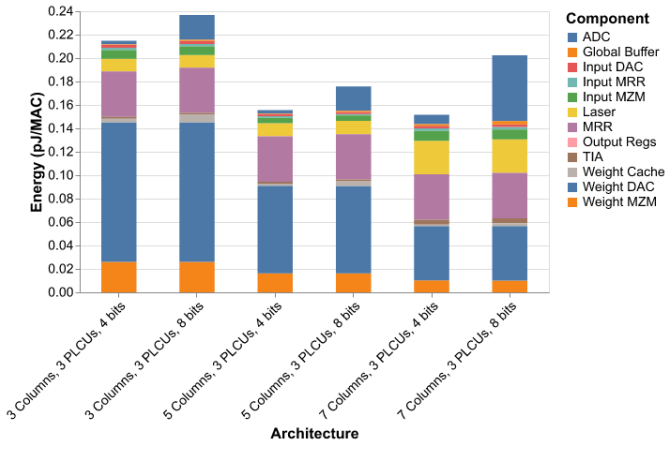
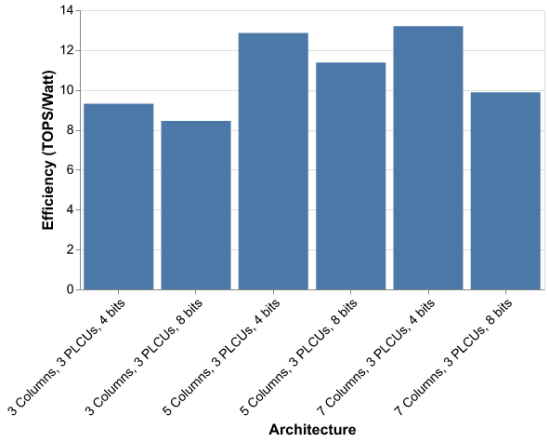
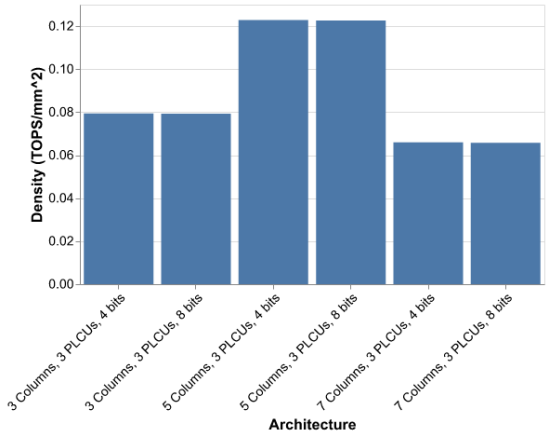


Fig. 2: ResNet18 Energy Breakdown

fig. 2 presents the fine-grained energy per MAC across major components and configurations. At low parallelism (3



(a) Energy Efficiencies



(b) Compute Densities

Fig. 3: Additional ResNet18 Measurements: energy efficiency and compute density.

columns), weight processing (DAC + MZM) dominates at 0.15 pJ/MAC. As columns increase, weight reuse drives this cost down to 0.06 pJ/MAC. However, with 8-bit resolution, output ADC processing starts to become as large: at 7 columns, it rises to 0.06 pJ/MAC versus only 0.02 pJ/MAC at 3 columns. In the 4-bit regime, ADC cost remains modest, highlighting the benefits of low-resolution conversions. Laser energy also scales with parallelism, tripling from 0.01 to 0.03 pJ/MAC between 3- and 7-column setups. MRR energy, which we could not tune via k^2, increases only slightly with parallelism.

Overall, in the 4-bit case, higher parallelism yields net energy savings as the drop in weight processing outweighs modest increases in ADC and laser costs, making the 7-column design most efficient. Conversely, at 8 bits, the rising ADC and laser overheads offset weight savings, so the 5-column balanced configuration achieves the lowest total energy.

Additional metrics are shown in fig. 3. Energy efficiency curves in fig. 3a perfectly mirror energy breakdown, so do not discuss it further. fig. 3b shows area-related results. For the sake of conciseness, we only observe that the right approach is to pack enough compute in each unit (more that 3 columns) without making each unit too large (less than 7 columns). A 5-column design seems to be the right balance. Our subsequent discussion focuses on energy, not on area.

C. Bringing It Together: Tradeoff Analysis

The combined results above illustrate that accuracy and energy are inversely affected by our key variables.

Energy optimization should favor lower ADC/DAC resolution and higher unit-level parallelism to maximize weight reuse until ADC and laser overheads dominate.

Accuracy optimization should favor higher resolution and reduced parallelism to suppress both quantization and crosstalk noise.

Good tradeoffs are possible when (1) the model is resilient to quantization errors, or (2) the MRRs have a low level of crosstalk.

- When the model is resilient to quantization errors, one can significantly reduce ADC/DAC resolution energy without significantly affecting accuracy. ResNet18 is moderately resilient to quantization errors making this a reasonable options. Our experiments with transformers show them to be very sensitive to quantization errors, making this option unavailable unless we use advanced techniques such as AWQ [29] or SmoothQuant [30].
- When the MRRs have low crosstalk sensitivity, we can decrease DAC processing energy without significantly affecting accuracy, until DAC/Laser energies become too significant.

Difficult tradeoffs exist when both quantization and crosstalk sensitivity are high. Any energy gain produces unacceptable accuracy degradation, and vice versa.

Future Directions primarily involves turning our approach into a generic tool that takes in a model, some energy and accuracy constraints, and some optimization targets, and returns the best-performing optical architecture (or any analog or digital architecture) we can find. We would just be limited to Albireo and would have much more sophisticated search strategies.

Acknowledgments

We thank and acknowledge Tanner and the other TAs for guiding us through this project.

REFERENCES

[1] T. P. Xiao, B. Feinberg, C. H. Bennett, V. Prabhakar, P. Saxena, V. Agrawal, S. Agarwal, and M. J. Marinella, "On the accuracy of analog neural network inference accelerators," arXiv preprint arXiv:2109.01262v3 [cs.AR], Feb 2022.

[2] K. Shiflett, A. Karanth, R. Bunescu, and A. Louri, "Albireo: Energy-Efficient Acceleration of Convolutional Neural Networks via Silicon Photonics," in *Proc. 48th Annual Int. Symp. on Computer Architecture (ISCA)*, pp. 860–873, 2021.

[3] P. Houshmand, J. Sun, and M. Verhelst, "Benchmarking and modeling of analog and digital SRAM in-memory computing architectures," arXiv preprint arXiv:2305.18335v1 [cs.AR], May 2023.

[4] T. Andrulis, J. S. Emer, and V. Sze, "CiMLoop: A Flexible, Accurate, and Fast Compute-In-Memory Modeling Tool," in *Proc. IEEE Int. Symp. on Performance Analysis of Systems and Software (ISPASS)*, pp. 10–23, 2024.

[5] T.-J. Yang and V. Sze, "Design Considerations for Efficient Deep Neural Networks on Processing-in-Memory Accelerators," arXiv preprint arXiv:1912.12167v1 [cs.CV], Dec 2019.

[6] S. K. Roy and N. R. Shanbhag, "Energy-Accuracy Trade-Offs for Resistive In-Memory Computing Architectures," *IEEE J. Explor. Solid-State Comput. Devices Circuits (JXCDC)*, vol. 10, pp. 22–30, Mar 2024.

[7] S. K. Roy, A. Patil, and N. R. Shanbhag, "Fundamental Limits on the Computational Accuracy of Resistive Crossbar-based In-memory Architectures," in *Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS)*, pp. 384–388, 2022.

[8] T. Andrulis, R. Chen, H.-S. Lee, J. S. Emer, and V. Sze, "Modeling Analog-Digital-Converter Energy and Area for Compute-In-Memory Accelerator Design," arXiv preprint arXiv:2404.06553v2 [cs.AR], May 2024.

[9] T. Andrulis, J. S. Emer, and V. Sze, "RAELLA: Reforming the Arithmetic for Efficient, Low-Resolution, and Low-Loss Analog PIM: No Retraining Required!," in *Proc. 50th Annual Int. Symp. on Computer Architecture (ISCA '23)*, pp. 1–16, Jun 2023.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

[11] Y. N. Wu, P.-A. Tsai, A. Parashar, V. Sze, and J. S. Emer, "Sparseloop: An Analytical Approach To Sparse Tensor Accelerator Modeling," in *Proc. 55th IEEE/ACM Int. Symp. on Microarchitecture (MICRO)*, pp. 1377–1395, 2022.

[12] A. Parashar et al., "Timeloop: A Systematic Approach to DNN Accelerator Evaluation," in *Proc. IEEE Int. Symp. on Performance Analysis of Systems and Software (ISPASS)*, pp. 304–315, 2019.

[13] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, vol. 11, pp. 441–446, Jun 2017.

[14] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Le Gallo, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, pp. 52–58, Jan 2021.

[15] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, pp. 1004–1008, Sept 2018.

[16] H. Mo, W. Xie, Y. Zhao, L. Zhou, and M. Guo, "An integrated large-scale photonic accelerator with ultralow latency," *Nature*, vol. 640, pp. 361–367, Apr 2025.

[17] M. Le Gallo, R. Khaddam-Aljameh, M. Stanisavljevic, P. Hoskins, A. Sebastian, and E. Eleftheriou, "A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference," *Nature Electronics*, vol. 6, pp. 680–693, Sept 2023.

[18] Lightmatter, "Passage™ 3D photonics engine" (product brief), miscellaneous publication, 2024.

[19] D. Psaltis, D. Brady, X.-G. Gu, and S. Lin, "Holography in artificial neural networks," *Nature*, vol. 343, pp. 325–330, Jan 1990.

[20] K.-Y. Hsu, H.-Y. Li, and D. Psaltis, "Holographic implementation of a fully connected neural network," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1637–1645, Oct 1990.

[21] C. X.-G. Gu, "Optical neural networks using volume holograms," Ph.D. dissertation, California Institute of Technology, Sept 1990.

[22] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, "Experimental demonstration of reservoir computing on a silicon photonics chip," *Nature Communications*, vol. 5, Art. 3541, Mar 2014.

[23] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and Weight: An integrated network for scalable photonic spike processing," *Journal of Lightwave Technology*, vol. 32, no. 21, pp. 3427–3439, Aug 2014.

[24] A. N. Tait, T. F. de 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*, vol. 7, Art. 7430, Aug 2017.

[25] S. Pasricha, "Optical Computing for Deep Neural Network Acceleration: Foundations, Recent Developments, and Emerging Directions," arXiv preprint arXiv:2407.21184v1 [cs.AR], Jul 2024.

[26] Kamei, S., Kaneko, A., Ishii, M., Shibata, T., Inoue, Y. & Hibino, Y. Crosstalk reduction in arrayed-waveguide grating multiplexer/demultiplexer using cascade connection. *Journal Of Lightwave Technology*. **23**, 1929-1938 (2005)

[27] Gu, L., Fang, H., Li, J., Fang, L., Chua, S., Zhao, J. & Gan, X. . *Nanophotonics*. **8**, 841-848 (2019), https://doi.org/10.1515/nanoph-2018-0229

[28] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. & Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge . *International Journal Of Computer Vision (IJCV)*. **115**, 211-252 (2015)

[29] Lin, J., Tang, J., Tang, H., Yang, S., Chen, W., Wang, W., Xiao, G., Dang, X., Gan, C. & Han, S. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings Of Machine Learning And Systems*. **6** pp. 87-100 (2024)

[30] Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J. & Han, S. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. (2024), https://arxiv.org/abs/2211.10438