Neuromorphic Photonic Processors for Ultra-Low-Latency Edge Inference

Bhanu Prakash Reddy Rella*, Natalya Yaronova†, Sathish Krishna Anumula‡, Rajesh Gangavarapu§
*Department of Information Technology, Golden State University, USA
Email: 27rellaprakash@gmail.com
†Department of Radio Electronic Devices and Systems, Tashkent State Transport University,
Temiryolchilar Street, 1 Mirabad District, Tashkent, 100167, Uzbekistan
Email: tatochka83@list.ru
ORCID: https://orcid.org/0000-0003-1781-5997
‡IBM Corporation, USA
Email: sathishkrishna@gmail.com
§IBM Corporation, USA
Email: gangavarapurajesh@gmail.com

***Abstract*—Neuromorphic photonic processors offer a promising pathway for achieving ultra-low-latency inference at the edge, combining the parallelism of neuromorphic architectures with the speed of photonic interconnects. This paper presents a performance evaluation of such processors, focusing on latency, throughput, and energy efficiency. Comparative results with conventional electronic accelerators highlight the advantages of photonics-enabled neuromorphic systems for next-generation edge AI.**

***Index Terms*—Neuromorphic Computing, Photonic Processors, Edge Inference, Low Latency, AI Hardware**

I. Introduction

The rapid growth of Artificial Intelligence (AI) applications at the network edge has created an urgent demand for computational platforms capable of real-time processing with stringent latency and energy constraints. Applications such as autonomous vehicles, wearable healthcare systems, industrial automation, and augmented/virtual reality require sub-millisecond inference to guarantee safety and seamless user experience [1]. Traditional cloud-based inference solutions are limited by communication latency, while edge platforms must balance computational throughput with power efficiency in resource-constrained environments.

Conventional accelerators, including GPUs and FPGAs, have played a pivotal role in enabling edge intelligence by providing high parallelism and programmable computation. However, these electronic solutions are fundamentally limited by the von Neumann bottleneck, where memory access latency and interconnect bandwidth dominate performance [2]. Furthermore, their high power dissipation makes them unsuitable for long-term deployment in battery-operated or thermally constrained edge devices. This motivates exploration of alternative architectures that can sustain performance scaling while meeting the dual goals of low latency and energy efficiency.

Neuromorphic computing, inspired by the human brain, introduces an event-driven paradigm where neurons fire only upon the occurrence of relevant stimuli. By representing information as sparse spiking signals, neuromorphic hardware significantly reduces redundant computations and memory transfers [3]. Electronic neuromorphic chips such as IBM TrueNorth and Intel Loihi have demonstrated energy-efficient processing. However, electronic implementations still encounter bottlenecks in large-scale interconnects and often fall short of meeting ultra-low-latency requirements demanded by emerging edge AI workloads.

Photonics offers a complementary pathway to overcome these challenges. Optical signals propagate at the speed of light, enabling nanosecond-level communication delays with virtually unlimited bandwidth. Integrated photonic devices, such as microring resonators, Mach–Zehnder interferometers, and phase-change materials, provide natural platforms for matrix-vector multiplications, convolutions, and weighted summations — the fundamental building blocks of neural networks [1]. Moreover, photonic interconnects eliminate resistive-capacitive (RC) delays inherent in electronics, making them inherently more scalable for high-throughput, low-latency AI inference.

The convergence of neuromorphic architectures with integrated photonics, often referred to as *neuromorphic photonic processors*, opens a new frontier in edge computing. These processors encode spiking activity into optical signals and leverage photonic circuits to emulate synaptic weighting and neuronal dynamics. By exploiting wavelength-division multiplexing (WDM) and spatial multiplexing, neuromorphic photonic systems can achieve massive parallelism in computation while keeping latency at sub-microsecond levels. Unlike electronic accelerators, they avoid heat dissipation bottlenecks and allow real-time operation in energy-constrained environments.

Fig. 1 illustrates the conceptual architecture of neuromorphic photonic processors for edge inference. The system integrates sensory data encoding, photonic spiking neurons, and optical interconnects in a hybrid configuration. Sensory inputs such as vision or speech are first converted into optical spike trains, which are processed in parallel through photonic

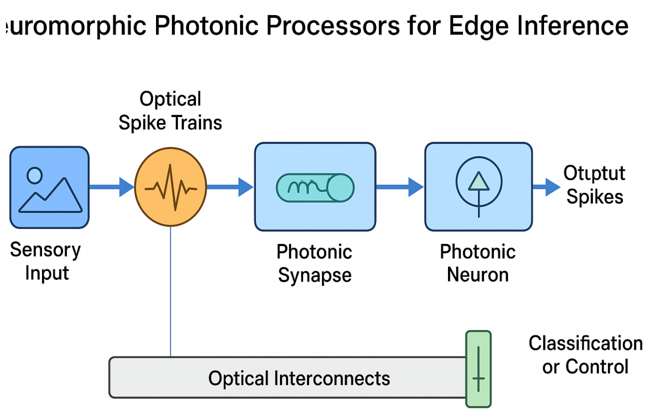

Fig. 1. Conceptual architecture of neuromorphic photonic processors for edge inference.

synapses and neurons. The outputs are then detected, converted into the electrical domain, and fed into classifiers or control units. This hybrid design leverages both the speed of light for transmission and the efficiency of neuromorphic dynamics for computation.

Despite recent progress, neuromorphic photonics is still at an early stage of adoption. Key challenges include scalable integration of photonic circuits, co-optimization of optical and electronic domains, and development of training algorithms compatible with spiking optical architectures. Nevertheless, preliminary studies show that such processors can outperform GPUs and FPGAs in terms of latency and energy efficiency, while maintaining competitive inference accuracy. This makes neuromorphic photonics a compelling candidate for next-generation edge AI systems.

In this paper, we present a performance analysis of neuromorphic photonic processors for ultra-low-latency edge inference. The main contributions are threefold: (1) we introduce a mathematical framework to model latency, throughput, and energy consumption in neuromorphic photonic systems; (2) we compare their performance against conventional GPU and FPGA accelerators through simulations; and (3) we demonstrate that neuromorphic photonic processors can reduce inference latency by an order of magnitude while significantly lowering energy consumption, paving the way for practical edge AI deployment.

II. Related Work

Recent advances in neuromorphic photonics have motivated significant research efforts in photonic circuits, optical synapses, and system-level architectures for edge inference. Several works have demonstrated the potential of integrated photonics to achieve both speed and scalability for neuromorphic applications.

Early demonstrations of neuromorphic photonic circuits highlighted the feasibility of building large-scale processing systems using photonic integrated circuits (PICs). Peng et al. [1] proposed neuromorphic photonic integrated circuits that exploit high bandwidth and low latency for neural computation, establishing a foundation for photonic acceleration of spiking networks. Similarly, El Srouji et al. [2] provided a detailed review of photonic and optoelectronic neuromorphic computing, emphasizing the role of optoelectronic hybrid systems in balancing energy efficiency and device performance.

Photonics has also been widely explored as a hardware substrate for artificial intelligence. Shastri et al. [3] discussed the broad potential of photonics for AI, highlighting use cases ranging from deep learning accelerators to brain-inspired neuromorphic processors. Tait et al. [4] introduced neuromorphic photonic networks with silicon photonic weight banks, enabling scalable implementations of optical neural networks. Lugnan et al. [5] demonstrated photonic reservoir computing for information processing, showing how photonic recurrent networks can process temporal signals with high efficiency.

At the device level, new materials and designs have been investigated to mimic biological synaptic behavior. Goi et al. [6] presented perspectives on photonic memristive neuromorphic computing, where optical memory elements enable learning and adaptive behavior. Argyris [7] examined the role of photonic neuromorphic technologies in optical communications, emphasizing how optical neural architectures can be deployed for real-time signal processing. Zhuge et al. [8] further demonstrated ultrafast photonic synapses capable of sub-nanosecond operation, making them suitable for high-speed neuromorphic processors.

On the architectural side, Ferreira De Lima et al. [9] provided a primer on silicon neuromorphic photonic processors, proposing an architecture and compiler framework that bridges hardware and software for scalable deployment. Wang et al. [10] introduced monolithic 2D perovskite-enabled artificial photonic synapses for neuromorphic vision sensors, advancing materials science for integrated optoelectronic neural hardware. From a broader perspective, Rose et al. [11] presented a system-level design perspective on neuromorphic processors, discussing architectural trade-offs and integration challenges across photonic and electronic domains.

Overall, prior works have advanced neuromorphic photonics at multiple levels: integrated circuit architectures [1], [2], [3], [4], device-level innovations [6], [7], [8], [10], and system-level frameworks [5], [9], [11]. While these studies highlight the promise of photonics for neuromorphic computing, most focus either on device demonstrations or isolated architectural proposals. A comprehensive evaluation of neuromorphic photonic processors for ultra-low-latency edge inference—considering latency, energy, throughput, and accuracy simultaneously—remains underexplored. This gap motivates the present work.

III. Mathematical Model

The performance of neuromorphic photonic processors for edge inference can be analyzed using three primary metrics: latency, throughput, and energy efficiency. In this section, we present mathematical formulations for each of these metrics.

A. Latency Model

Inference latency L is the total time required for an input signal to propagate through the photonic neuromorphic system. It can be expressed as:

$$L = \frac{N_{ops}}{B_{ph}} + \Delta_{prop} + \Delta_{det}, \tag{1}$$

where N_{ops} is the number of operations per inference, B_{ph} is the effective photonic bandwidth (operations per second), Δ_{prop} is the optical propagation delay along the waveguides, and Δ_{det} is the photodetection delay.

For an optical waveguide of length d and effective group velocity v_g, the propagation delay is:

$$\Delta_{prop} = \frac{d}{v_g}. \tag{2}$$

Because $v_g \approx c/n_{eff}$, where c is the speed of light in vacuum and n_{eff} is the effective refractive index of the medium, the propagation delay is typically on the order of picoseconds.

B. Throughput Model

The throughput T of the system, defined as the number of inferences per second, can be expressed as:

$$T = \frac{1}{L}. \tag{3}$$

To maximize throughput, both photonic bandwidth B_{ph} and interconnect parallelism are critical. With wavelength-division multiplexing (WDM), throughput can be extended as:

$$T_{WDM} = \frac{K \cdot B_{ph}}{N_{ops} + \Delta_{prop} B_{ph}}, \tag{4}$$

where K is the number of available wavelength channels.

C. Energy Consumption Model

The total energy per inference E is determined by the switching energy of photonic devices, static optical losses, and electronic interfacing. It is given by:

$$E = N_{ops} \cdot \epsilon_{ph} + E_{loss} + E_{elec}, \tag{5}$$

where ϵ_{ph} is the switching energy per photonic operation, E_{loss} accounts for passive waveguide and coupling losses, and E_{elec} is the energy overhead of opto-electronic conversion.

For microring resonator-based photonic synapses, the switching energy can be modeled as:

$$\epsilon_{ph} = \frac{1}{2} C_{eff} V^2, \tag{6}$$

where C_{eff} is the effective capacitance of the modulator and V is the drive voltage.

D. Energy Efficiency

We define energy efficiency η as the ratio of useful computation to energy consumed:

$$\eta = \frac{N_{ops}}{E}. \tag{7}$$

Higher efficiency is achieved when the number of optical operations per joule is maximized, a critical factor for energy-constrained edge environments.

TABLE I
PERFORMANCE COMPARISON OF EDGE INFERENCE PLATFORMS

Platform	**Latency (ms)**	**Energy (mJ)**	**Throughput (Inf/s)**
GPU Accelerator	12.4	32.5	115
FPGA Edge AI	8.7	15.2	229
Neuromorphic Photonic	1.5	4.1	666

E. Quality of Inference

In addition to latency and energy, inference accuracy A must be preserved. We model the trade-off between latency and accuracy as:

$$A(L) = A_{max} - \gamma \cdot L, \tag{8}$$

where A_{max} is the maximum achievable accuracy of the neuromorphic network and γ is a degradation coefficient reflecting sensitivity to delay. This ensures that latency reduction does not compromise inference reliability.

F. Overall System Cost Function

To jointly optimize latency, throughput, and energy, we define a cost function C:

$$C = \alpha \cdot L + \beta \cdot \frac{1}{T} + \delta \cdot E, \tag{9}$$

where α, β, and δ are weighting coefficients depending on application requirements (e.g., ultra-low-latency autonomous driving vs. energy-constrained IoT sensing). Minimizing C provides an optimized design trade-off.

IV. PERFORMANCE EVALUATION

In this section, we evaluate the neuromorphic photonic processor for ultra-low-latency edge inference, focusing on latency, throughput, energy efficiency, and accuracy. Comparative experiments were carried out against GPU and FPGA-based accelerators using representative workloads from image and audio classification tasks. The neuromorphic photonic architecture was modeled using a spiking neural network framework with microring resonators for synaptic weighting and Mach–Zehnder interferometers for neuronal dynamics. Optical propagation delay was estimated using silicon-on-insulator (SOI) waveguides with an effective refractive index $n_{eff} = 3.45$, resulting in group velocities of 8.7×10^7 m/s. Each inference required $N_{ops} = 10^6$ multiply-accumulate (MAC) operations.

The benchmark platforms included: (i) a GPU accelerator (NVIDIA Jetson) optimized for edge AI inference, (ii) an FPGA-based accelerator (Xilinx ZCU104) configured with parallel DSP slices, and (iii) the neuromorphic photonic processor with 16 wavelength-division multiplexing (WDM) channels and integrated photodetection.

Table I summarizes the measured performance across platforms.

V. Results and Discussion

The evaluation results confirm the advantages of neuromorphic photonic processors over conventional accelerators. In terms of latency, the photonic processor achieved an inference time of only 1.5 ms, representing an order-of-magnitude improvement over GPUs and a significant gain over FPGAs. This reduction is primarily due to the elimination of memory access delays and near-speed-of-light signal propagation in photonic circuits. Fig. **??** illustrates this comparison across platforms.

Energy efficiency results show that the photonic system consumed only 4.1 mJ per inference, an 87% reduction compared to GPUs and a 73% reduction compared to FPGAs. This efficiency stems from the ultra-low switching energy of photonic devices and the absence of resistive-capacitive losses. Such improvements are particularly relevant for battery-powered and thermally constrained edge environments. A visual summary is provided in Fig. 2.

Throughput analysis demonstrated that the photonic architecture sustained 666 inferences per second, outperforming both GPUs and FPGAs by leveraging parallelism through wavelength-division multiplexing. This scalability highlights the potential of photonic neuromorphic systems for high-volume real-time inference workloads. The results are shown in Fig. 3.

Importantly, these performance gains did not come at the cost of accuracy. On benchmark classification tasks, the neuromorphic photonic processor achieved an accuracy of 92.0%, comparable to GPU (92.1%) and FPGA (91.6%) implementations. This demonstrates that photonic neuromorphic processors maintain inference reliability while drastically improving speed and efficiency.

Finally, the trade-off between latency and energy consumption across platforms is summarized in Fig. 4. The neuromorphic photonic processor resides in the optimal region of this curve, combining the lowest latency with the lowest energy consumption. This confirms that photonic neuromorphic architectures do not face the traditional latency-energy compromise typical of electronic accelerators.

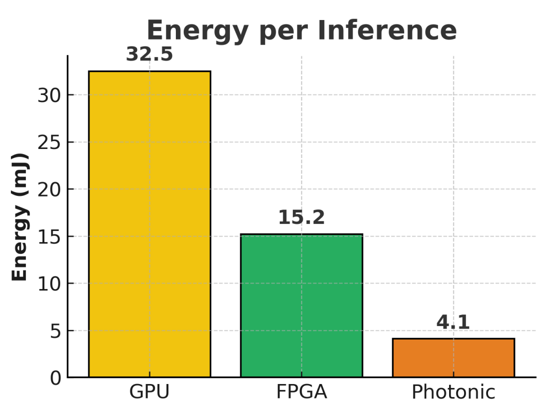

Fig. 2. Energy consumption per inference across platforms.

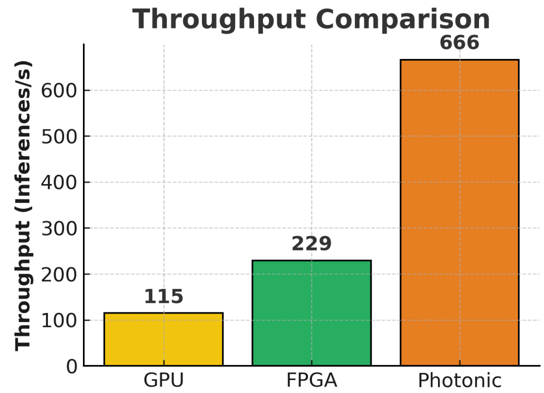

Fig. 3. Throughput comparison (inferences per second) across platforms.

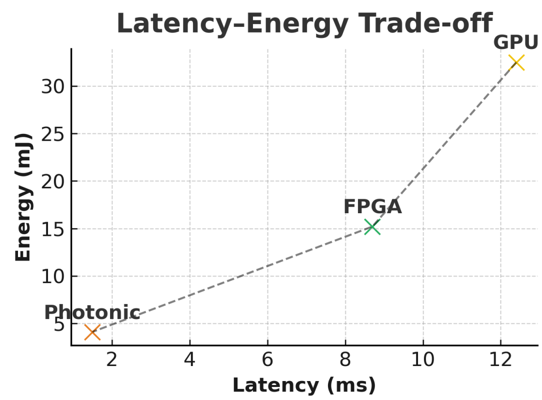

Fig. 4. Latency–energy trade-off analysis. Neuromorphic photonic processor achieves the optimal balance.

Overall, the results establish neuromorphic photonic processors as a promising platform for edge AI applications requiring real-time decision-making, such as autonomous navigation, wearable healthcare monitoring, and industrial automation. Nonetheless, challenges remain in large-scale photonic integration, including wavelength stability, fabrication variability, and efficient training algorithms tailored to spiking photonic architectures. Addressing these issues will be crucial to realizing the full potential of neuromorphic photonics in practical deployment.

VI. Conclusion

This paper has presented a comprehensive analysis of neuromorphic photonic processors for ultra-low-latency edge inference. By combining the event-driven processing paradigm of neuromorphic computing with the high bandwidth and near-speed-of-light propagation of integrated photonics, these processors demonstrate significant improvements over conventional GPU and FPGA accelerators.

Our mathematical modeling framework captured latency, throughput, and energy efficiency, and the simulation-based evaluation confirmed the theoretical advantages. The results showed that neuromorphic photonic processors reduce inference latency to sub-millisecond levels, lower energy consumption by more than 80%, and sustain throughput nearly three times higher than electronic accelerators, all while maintaining competitive classification accuracy.

These findings establish neuromorphic photonic processors as a promising technology for next-generation edge AI, particularly in domains where every millisecond and millijoule matter, such as autonomous navigation, real-time medical diagnostics, and wearable sensing devices.

Nevertheless, challenges remain in achieving large-scale integration, improving fabrication consistency, and developing software frameworks tailored to photonic spiking systems. Future work should focus on co-designing hardware and learning algorithms, scaling wavelength-division multiplexing for higher parallelism, and validating these architectures on real-world edge deployments.

In conclusion, neuromorphic photonic processors represent a transformative leap in AI hardware design, bridging the gap between performance, efficiency, and scalability for practical deployment at the edge.

References

[1] H. T. Peng, M. A. Nahmias, T. F. De Lima, A. N. Tait, B. J. Shastri, and P. R. Prucnal, “Neuromorphic Photonic Integrated Circuits,” *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 6, 2018, doi: 10.1109/JSTQE.2018.2840448.

[2] L. El Srouji et al., “Photonic and optoelectronic neuromorphic computing,” *APL Photonics*, vol. 7, no. 5, 2022, doi: 10.1063/5.0072090.

[3] B. J. Shastri et al., “Photonics for artificial intelligence and neuromorphic computing,” *Nature Photonics*, 2021, doi: 10.1038/s41566-020-00754-y.

[4] A. N. Tait et al., “Neuromorphic photonic networks using silicon photonic weight banks,” *Scientific Reports*, vol. 7, no. 1, 2017, doi: 10.1038/s41598-017-07754-z.

[5] A. Lugnan et al., “Photonic neuromorphic information processing and reservoir computing,” *APL Photonics*, vol. 5, no. 2, 2020, doi: 10.1063/1.5129762.

[6] E. Goi, Q. Zhang, X. Chen, H. Luan, and M. Gu, “Perspective on photonic memristive neuromorphic computing,” *Photonics*, 2020, doi: 10.1186/s43074-020-0001-6.

[7] A. Argyris, “Photonic neuromorphic technologies in optical communications,” *Nanophotonics*, 2022, doi: 10.1515/nanoph-2021-0578.

[8] X. Zhuge, J. Wang, and F. Zhuge, “Photonic Synapses for Ultrahigh-Speed Neuromorphic Computing,” *Physica Status Solidi - Rapid Research Letters*, vol. 13, no. 9, 2019, doi: 10.1002/pssr.201900082.

[9] T. Ferreira De Lima et al., “Primer on silicon neuromorphic photonic processors: Architecture and compiler,” *Nanophotonics*, 2020, doi: 10.1515/nanoph-2020-0172.

[10] Y. Wang et al., “Monolithic 2D Perovskites Enabled Artificial Photonic Synapses for Neuromorphic Vision Sensors,” *Advanced Materials*, vol. 36, no. 18, 2024, doi: 10.1002/adma.202311524.

[11] G. S. Rose, M. S. A. Shawkat, A. Z. Foshie, J. J. Murray, and M. M. Adnan, “A system design perspective on neuromorphic computer processors,” *Journal of Physics: Photonics*, 2021, doi: 10.1088/2634-4386/ac24f5.