

Classification of Photonic Inference Architectures

Andreas Falkenberg

Principal AI Accelerator Architect

QPI Laboratories

Version 1.0 — October 5, 2026

PREPRINT / TECHNICAL WHITE PAPER — NOT PEER REVIEWED. This version is provided for technical discussion and may be revised or submitted for peer-reviewed publication.

Executive Abstract

The global demand for artificial intelligence inference compute has entered a period of exponential growth, driven primarily by the deployment of large language models (LLMs) and multimodal foundation models at unprecedented scale. Contemporary CMOS-based accelerators, including graphics processing units (GPUs) and tensor processing units (TPUs), are approaching fundamental thermodynamic and memory-bandwidth limits that constrain both throughput and energy efficiency. In this context, photonic computing has re-emerged as a compelling alternative, offering speed-of-light signal propagation, massive wavelength-division multiplexing (WDM) parallelism, and the prospect of zero-static-power analog matrix computation — key attributes that directly address the memory wall and thermal constraints of digital silicon.

This paper presents a comprehensive taxonomy of six principal photonic inference architecture families: (i) diffractive optical networks (D^2NN), (ii) silicon photonic Mach-Zehnder interferometer (MZI) meshes, (iii) electro-optic multiply-accumulate (MAC) units, (iv) resonator-based inference engines, (v) vertical-cavity surface-emitting laser (VCSEL) array processors, and (vi) free-space optical accelerators. Each family is evaluated across five headline dimensions—inference throughput, energy efficiency, effective numerical precision, programmability, and technology readiness—and complementary implementation attributes including substrate, weight encoding, nonlinearity, and scalability. Reported results range from theoretical passive-optics efficiency limits approaching peta-OPS/W to experimentally demonstrated integrated photonic tensor-core and neuromorphic systems,

while leading CMOS accelerators provide higher precision and broader model programmability. The paper’s primary contribution is a unified, axis-aligned classification framework that supports direct cross-family comparison and guides the selection of photonic platforms for specific inference workloads.

Keywords: photonic computing, optical neural networks, diffractive optical networks, silicon photonics, MZI mesh, electro-optic, resonator inference, VCSEL arrays, free-space optical computing, AI accelerator

1. Introduction

Demand for artificial-intelligence inference is increasing data-center compute and power requirements, particularly for large language models and multimodal foundation models. Serving these models at scale requires sustained accelerator capacity and repeated movement of parameters and key-value caches between memory and compute. As CMOS scaling continues, static leakage, interconnect energy, memory bandwidth, and package-level power density increasingly constrain system efficiency alongside arithmetic throughput. These pressures motivate investigation of alternative computing substrates, including photonic accelerators, for workloads dominated by linear algebra and data movement.

The memory wall, first articulated by Wulf and McKee in 1995 [25], has become acutely relevant in the context of transformer-based inference. The attention mechanism requires repeated, bandwidth-intensive transfers of key-value caches between high-bandwidth memory (HBM) and compute arrays, with effective arithmetic intensity frequently falling below the roofline threshold. Meanwhile, HBM stacking has saturated at practical limits imposed by thermal dissipation in the stacked die, and die-to-die interconnect bandwidth is constrained by electrical signaling at the chip boundary.

Photonics addresses these constraints through physical mechanisms that differ from electronic interconnects. Guided optical signals avoid resistive conductor loss, although waveguide propagation, coupling, modulation, and detection still introduce attenuation and energy overhead. Wavelength-division multiplexing allows a single waveguide to carry multiple independent data channels simultaneously, increasing aggregate bandwidth without requiring a separate physical path for every

channel. Analog optical matrix-vector multiplication can, in principle, be performed in a single propagation pass through a passive optical medium, avoiding clocked transistor switching within the matrix transform itself. System-level energy nevertheless remains bounded by light generation, modulation, photodetection, conversion, calibration, and data movement.

The intellectual lineage of optical computing extends to the 1980s, when researchers at Bell Laboratories and MIT Lincoln Laboratory explored holographic associative memories and optical Fourier processors. Those systems were limited by the absence of suitable light sources, modulators, and detector arrays, as well as by the intractability of training algorithms prior to the backpropagation era. The modern resurgence, initiated by Shen et al. in 2017 with the demonstration of a silicon photonic neural network capable of vowel classification [3], and accelerated by Lin et al.'s 2018 diffractive deep neural network (D²NN) reported in *Science* [1], has proceeded in parallel with the explosion of deep learning and the emergence of dedicated inference ASICs as a mainstream product category.

Between 2018 and 2026, the field diversified across silicon-on-insulator, indium phosphide, lithium niobate on insulator, silicon nitride, free-space optics, and hybrid electro-photonic platforms. Phase-change materials enabled non-volatile photonic weight storage, while VCSEL arrays were investigated for reservoir computing and spiking networks [14]. Commercial development has concentrated on integrated photonic compute, optical interconnect, and memory-disaggregation architectures, with representative programs discussed in Section 12.

Despite this activity, the field lacks a unified classification framework that permits systematic, quantitative comparison across architecture families. Practitioners selecting a photonic platform face a heterogeneous literature spanning multiple physical substrates, training methodologies, and performance reporting conventions. This paper addresses that gap. The remainder of the paper is organized as follows: Section 2 provides background on optical MVM, analog precision, and integration platforms. Sections 3–9 develop the taxonomy and analyze each architecture family. Section 10 provides a cross-family comparison in a unified table and discusses the principal trade-offs. Sections 11 and 12 address scalability, future directions, and commercialization pathways. Section 13 concludes.

2. Background and Fundamentals

2.1 Optical Matrix-Vector Multiplication

Matrix-vector multiplication (MVM) is the fundamental primitive of neural network inference. In a fully connected layer, the output activation vector $\mathbf{y}$ is computed as $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$, where $\mathbf{W}$ is the weight matrix, $\mathbf{x}$ is the input activation vector, and $\mathbf{b}$ is the bias vector. For transformer models, the dominant arithmetic is similarly structured: query-key dot products, value aggregation, and feed-forward projections all reduce to batched MVM.

Photonic systems implement MVM via three principal mechanisms. In coherent waveguide systems, the amplitude and phase of optical fields encode complex-valued vector components; interference between guided modes at directional couplers and phase-shifted segments performs multiply-accumulate operations on the complex field. In incoherent free-space systems, intensity patterns encoded on spatial light modulators are multiplied by transmission masks and integrated on detector arrays, implementing positive-valued MVM. In wavelength-division multiplexed systems, multiple optical carriers at distinct wavelengths traverse a shared physical medium, with each wavelength channel providing an independent computational lane; the photodetector integrates all channels, implementing a weighted summation.

The unifying advantage is that these operations are performed at the speed of light, with energy cost scaling sub-linearly with the size of the computation for passive (non-modulated) weight implementations. The principal constraint is the mapping from complex, bipolar, floating-point weights to physical optical degrees of freedom, which introduces quantization error and limits effective precision.

2.2 Precision and Noise in Analog Photonic Systems

Analog photonic inference systems are subject to multiple noise sources that collectively determine the effective number of bits (ENOB). Shot noise, arising from the discrete quantum nature of photons, imposes a fundamental limit: the signal-to-noise ratio scales as the square root of the photon number, yielding an ENOB of approximately 6–8 bits for moderate optical power levels. In practice, ENOB is further degraded by thermal noise in photodetectors and transimpedance amplifiers (~1–2 bit degradation), fabrication imperfections in waveguide geometry and coupling ratios (~1–2 bits), and crosstalk between adjacent optical channels. Reported ENOB values for integrated photonic MVM systems range from 3 to 6 bits across the literature.

Digital correction techniques — including calibration matrices, error-compensating hardware-aware training, and in-situ optimization — have been demonstrated to partially recover lost precision. Quantization-aware training (QAT) allows neural network weights to be optimized for deployment at 4-bit precision with minimal accuracy degradation on tasks such as ImageNet classification, suggesting that the photonic precision floor may be compatible with a meaningful class of inference applications, particularly those requiring classification outputs rather than high-precision generative outputs.

2.3 Integration Platforms

Silicon-on-insulator (SOI) is the most mature photonic integration platform, leveraging CMOS-compatible fabrication at 200 mm and 300 mm foundries. SOI waveguides at 220 nm or 300 nm thickness support single-mode propagation at 1310 nm and 1550 nm telecom wavelengths, with propagation losses of 1–3 dB/cm. Indium phosphide (InP) offers monolithic integration of active optical components — lasers, amplifiers, and high-speed modulators — with modulation bandwidths exceeding 100 GHz. Lithium niobate on insulator (LNOI) provides exceptionally strong electro-optic coefficients ($r_{33} \approx 30$ pm/V), enabling sub-1 V·cm modulation with bandwidths above 100 GHz and minimal optical loss. Silicon nitride (SiN) offers ultra-low propagation loss (<0.1 dB/cm) and a wide transparency window extending into the visible, making it suitable for hybrid integration. Polymer waveguides enable large-area, low-cost fabrication for free-space coupling applications. Free-space bulk-optic systems exploit established spatial light modulator (SLM) and diffractive optical element (DOE) technology. Hybrid electro-photonic platforms — combining SOI photonics with 3D-stacked CMOS — represent the leading commercial architecture for near-term product deployment.

3. Taxonomy of Photonic Inference Architectures

We propose a classification framework for photonic inference architectures along four orthogonal axes, enabling systematic placement of any photonic system within a structured design space:

1. **Spatial encoding:** coherent waveguide (phase and amplitude), incoherent free-space (intensity only), or resonator-selective (wavelength-resolved amplitude).

2. **Compute paradigm:** analog wave-optic (coherent field interference), analog intensity (incoherent power summation), or digital-assisted hybrid (optical front-end with digital correction).
3. **Programmability:** fixed/fabricated (weights encoded at manufacturing time), runtime-tunable (weights updated via electrical actuation during operation), or hybrid (fixed optical weights with tunable digital correction).
4. **Integration level:** chip-scale (monolithic photonic integrated circuit), board-scale (multi-chip photonic module), or system-scale (rack-mounted free-space optical system).

Within this framework, six principal architecture families are identified and analyzed in the subsequent sections: (i) diffractive optical networks, (ii) MZI mesh networks, (iii) electro-optic MAC units, (iv) resonator-based inference engines, (v) VCSEL array processors, and (vi) free-space optical accelerators.

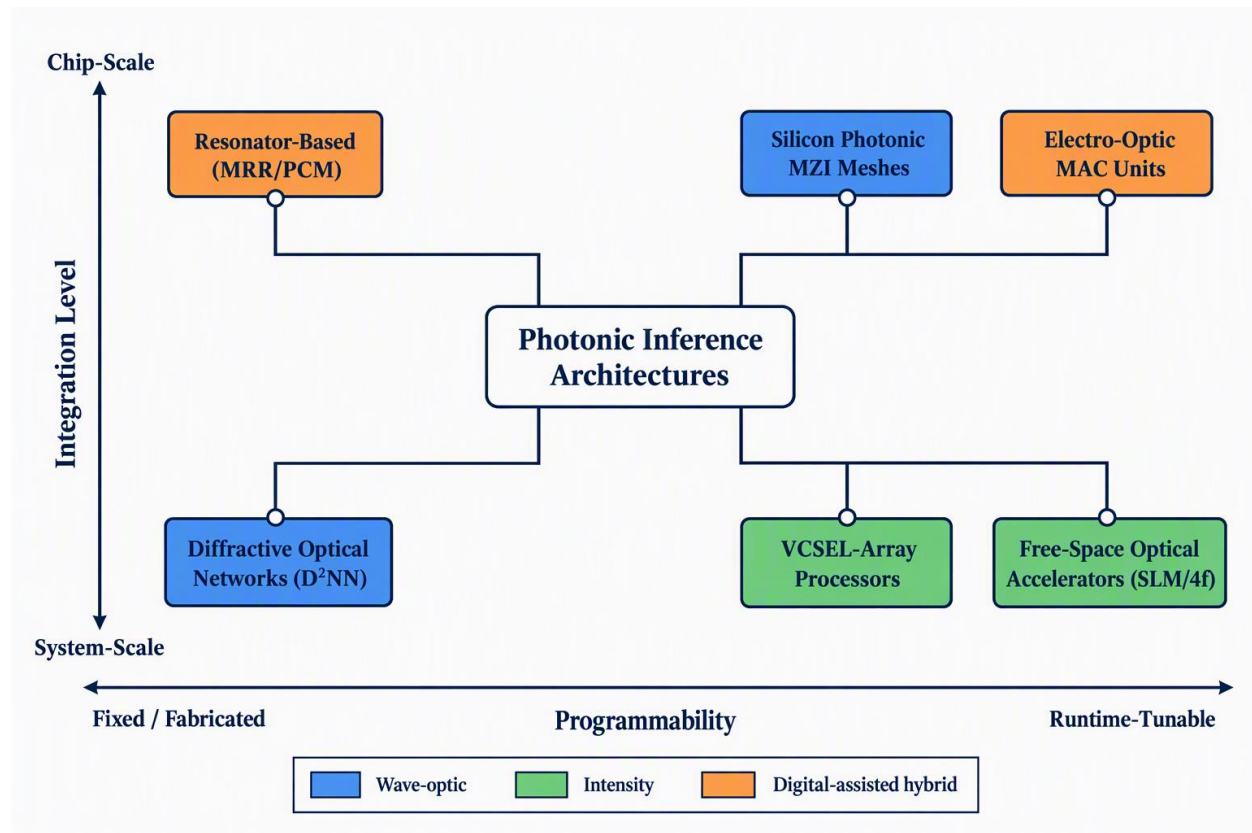


Fig. 1. Proposed taxonomy of photonic inference architectures across programmability and integration level.

4. Diffractive Optical Networks (D²NN)

4.1 Operating Principle

Diffractive deep neural networks (D²NN), introduced by Lin et al. in 2018, implement multi-layer neural network inference through the physical propagation of coherent light through a cascade of passive diffractive surfaces [1]. Each diffractive layer consists of a planar array of transmissive or reflective pixels, each imparting a locally controlled phase shift and/or amplitude attenuation to the incident wavefront. Learned weights are encoded as the phase/amplitude values of these pixels, which are determined offline via backpropagation through a differentiable physical simulation of light propagation — typically the angular spectrum method or the Rayleigh-Sommerfeld diffraction integral.

The wave optical forward model is fully differentiable with respect to the pixel phase values, enabling gradient-based optimization using standard deep learning frameworks (PyTorch, TensorFlow). Once trained, the phase masks are fabricated as static optical elements, and inference requires only the propagation of the input light field through the stack — no electronic computation, no weight memory reads, and no clock-synchronized arithmetic. The output class scores are read by a photodetector array at the terminal diffraction plane, with each detector region corresponding to a class label. D²NN implementations have been demonstrated at terahertz (THz) frequencies using 3D-printed polymer layers, at near-infrared wavelengths using photolithographically patterned SiO₂/TiO₂ metasurfaces, and in the visible using spatial light modulators for reconfigurable emulation.

4.2 Fabrication

THz D²NN layers have been fabricated by fused deposition modeling (FDM) 3D printing of polylactic acid (PLA) or polypropylene at feature sizes of approximately 0.5–1 mm, corresponding to sub-wavelength sampling at 0.1–1 THz. Near-infrared implementations employ dielectric metasurfaces with phase-shifting nanopillar arrays (TiO₂ on SiO₂ substrate), offering pixel pitches of 300–800 nm and phase control spanning 0–2 π . Reconfigurable D²NN emulators use liquid-crystal-on-silicon (LCoS) SLMs with megapixel resolution and phase modulation depths exceeding 2 π , enabling in-situ weight updates at video frame rates (60–120 Hz). Metasurface-based implementations have achieved diffraction efficiencies of 70–92%, while 3D-printed THz layers exhibit efficiencies of 55–80% due to Fresnel reflection losses.

4.3 Performance

A D²NN can expose very high spatial parallelism: a five-layer system operating on a 200×200-pixel input field performs $O(N^2L)$ weighted optical interactions per inference pass, where N is the pixel count and L is the number of layers. Published theoretical estimates report efficiencies approaching 10^{15} equivalent operations per watt for idealized passive propagation. These estimates should not be interpreted as directly measured system throughput, because practical performance also depends on source efficiency, detector bandwidth and noise, optical loss, alignment, and the convention used to count an optical interaction as an operation.

Lin et al. demonstrated 91.75% classification accuracy on the MNIST handwritten digit dataset using a five-layer THz D²NN with 200×200 pixel layers, closely matching the fully digital baseline [1]. Subsequent work demonstrated classification on the Fashion-MNIST dataset and, via deeper architectures, on subsets of CIFAR-10 [2]. The depth-vs-accuracy trade-off follows the same qualitative behavior as digital deep networks: adding diffractive layers increases model expressivity but introduces additional propagation loss and alignment sensitivity.

4.4 Limitations

The defining limitation of fabricated D²NN systems is the absence of runtime weight updates: once manufactured, the optical weights are immutable. This precludes fine-tuning, domain adaptation, or multi-task operation without physical replacement of the diffractive layers. Optical nonlinearity is absent in passive diffractive materials; nonlinear activation functions must be implemented via optoelectronic detector-modulator pairs between layers, which reintroduces electronic latency and energy cost. Fabricated systems are sensitive to sub-wavelength misalignment, vibration, and thermal expansion of the layer stack. Additionally, the diffraction efficiency budget limits the depth of usable layers before optical power falls below the detector noise floor.

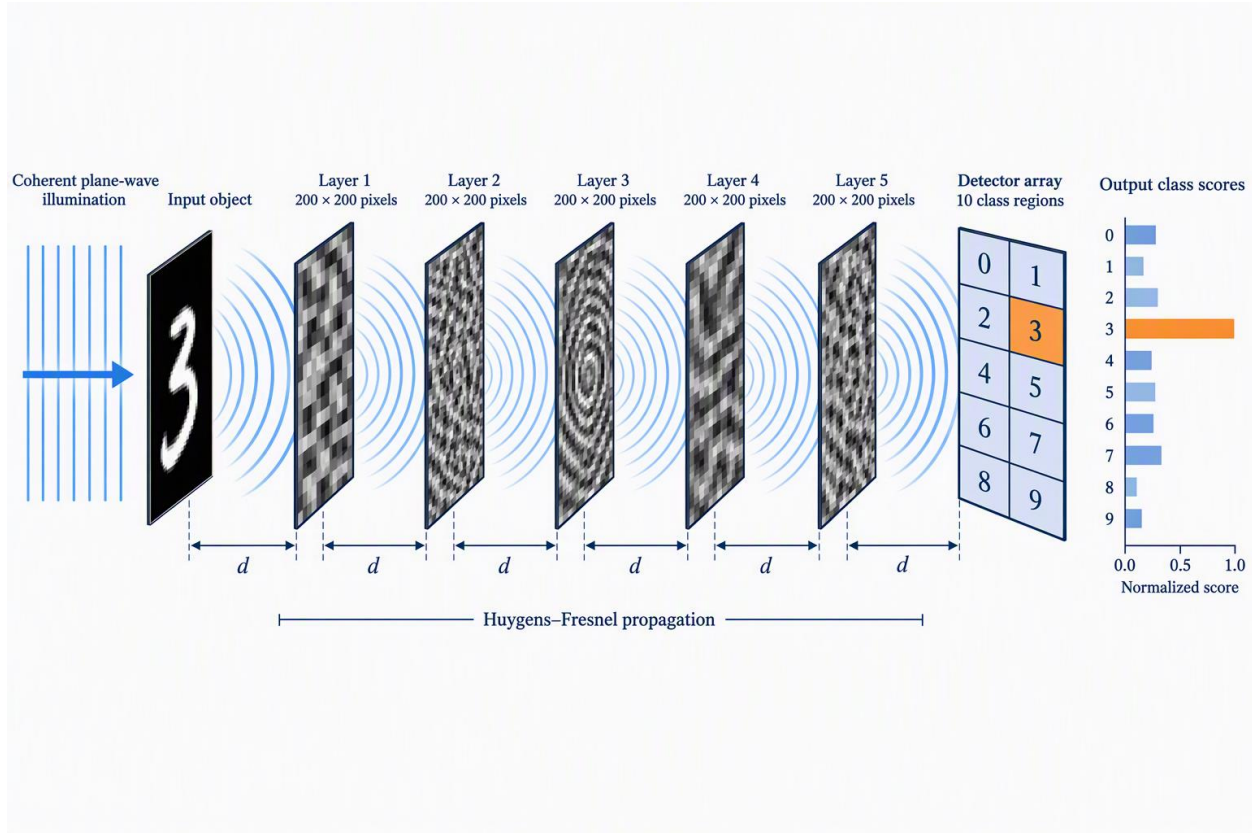


Fig. 2. Five-layer diffractive optical neural network with free-space Huygens–Fresnel propagation and spatial class detection.

5. Silicon Photonic MZI Meshes

5.1 Operating Principle

The Mach-Zehnder interferometer (MZI) mesh architecture, demonstrated for neural network inference by Shen et al. in 2017 [3], implements complex-valued unitary matrix operations through a programmable network of on-chip optical beam splitters and phase shifters. A single MZI unit cell consists of two directional couplers (50:50 beam splitters) enclosing an electro-optically or thermo-optically tunable phase segment, implementing a 2×2 unitary transformation parameterized by two angles (θ , ϕ). By cascading $N(N-1)/2$ MZI unit cells in the triangular arrangement of Reck et al. [11] or the rectangular arrangement of Clements et al. [4], any $N \times N$ unitary matrix can be realized on-chip. Singular value decomposition (SVD) decomposes an arbitrary real or complex matrix $W = U\Sigma V^\dagger$ into unitary factors (U ,

$V^\dagger$) and a diagonal singular value matrix (Σ), allowing the MZI mesh to implement arbitrary matrix-vector multiplication by concatenating two unitary meshes with an intervening variable optical attenuator (VOA) bank.

5.2 Chip Implementation

MZI mesh photonic integrated circuits are predominantly fabricated on the silicon-on-insulator (SOI) platform at 220 nm or 300 nm silicon waveguide core thickness, processed at 300 mm CMOS foundries. Thermo-optic phase shifters, which exploit the temperature-dependent refractive index of silicon ($dn/dT \approx 1.86 \times 10^{-4} \text{ K}^{-1}$), are the most common actuation mechanism, requiring approximately 10–30 mW per π radian of phase shift with microsecond-scale settling times. Electro-optic phase shifters on lithium niobate-on-silicon (LiNbO_3 -on-Si) platforms achieve sub-1 V-cm half-wave voltage-length products and picosecond switching times, enabling gigahertz-rate weight update for rapidly reconfigurable inference. Integrated optical neural chips have demonstrated complex-valued operations by encoding information in both optical phase and magnitude [20]. Dense MZI meshes scale quadratically in component count, requiring $N(N-1)/2$ interferometer cells for an $N \times N$ unitary transformation. Key research and industry contributors include MIT, Stanford Photonics Research Center, MIT Lincoln Laboratory, Lightmatter, and Luminous Computing.

5.3 Performance

Silicon photonic MZI meshes have demonstrated optical matrix-vector multiplication throughputs in the range of 10–100 tera-operations per second per watt (TOPS/W), primarily limited by the energy cost of phase shifter actuation rather than the optical computation itself. With WDM encoding of up to 64 wavelength channels on a single mesh, effective MAC throughput scales proportionally, pushing demonstrated system throughputs into the range of peta-MAC/s. The optical bandwidth of silicon waveguides exceeds 1 THz at C/L-band, supporting dense WDM without inter-channel crosstalk. Primary performance limiters include thermal drift of thermo-optic phase shifters (requiring active stabilization), phase noise from fabrication-induced coupling ratio variation, and insertion loss accumulation across the mesh depth.

5.4 Error Correction and Training

Hardware imperfections — including coupling ratio deviations of ± 1 –3%, phase shifter nonlinearity, and waveguide propagation loss variation — necessitate hardware-aware training methodologies. In-situ

gradient descent, implemented by measuring the optical output gradient directly from the chip, avoids the need for an accurate physical model. The adjoint method enables gradient computation with only two forward passes through the optical system, regardless of matrix size, making in-situ training practical for large meshes [18]. Hardware-aware training augments digital simulation with noise injection models calibrated to measured device parameters, pre-compensating for known imperfections before on-chip fine-tuning [19]. Recent single-chip photonic neural-network work has also demonstrated accelerated training in integrated hardware [21].

5.5 Limitations

The fundamental scalability constraint of MZI meshes is the accumulated optical insertion loss: each MZI unit cell introduces 0.5–2 dB of insertion loss, and an $N \times N$ mesh contains $O(N^2)$ cells, placing an approximate 20 dB system loss budget limit at $N \approx 64$ –128 for current waveguide technology. Beyond this depth, optical amplification (on-chip SOAs or off-chip EDFAs) must be interleaved, reintroducing electrical power consumption and noise. Phase shifter power — approximately 10–30 mW/ π for thermo-optic devices — dominates the system power budget for large, frequently-updated meshes, partially eroding the energy efficiency advantage over CMOS.

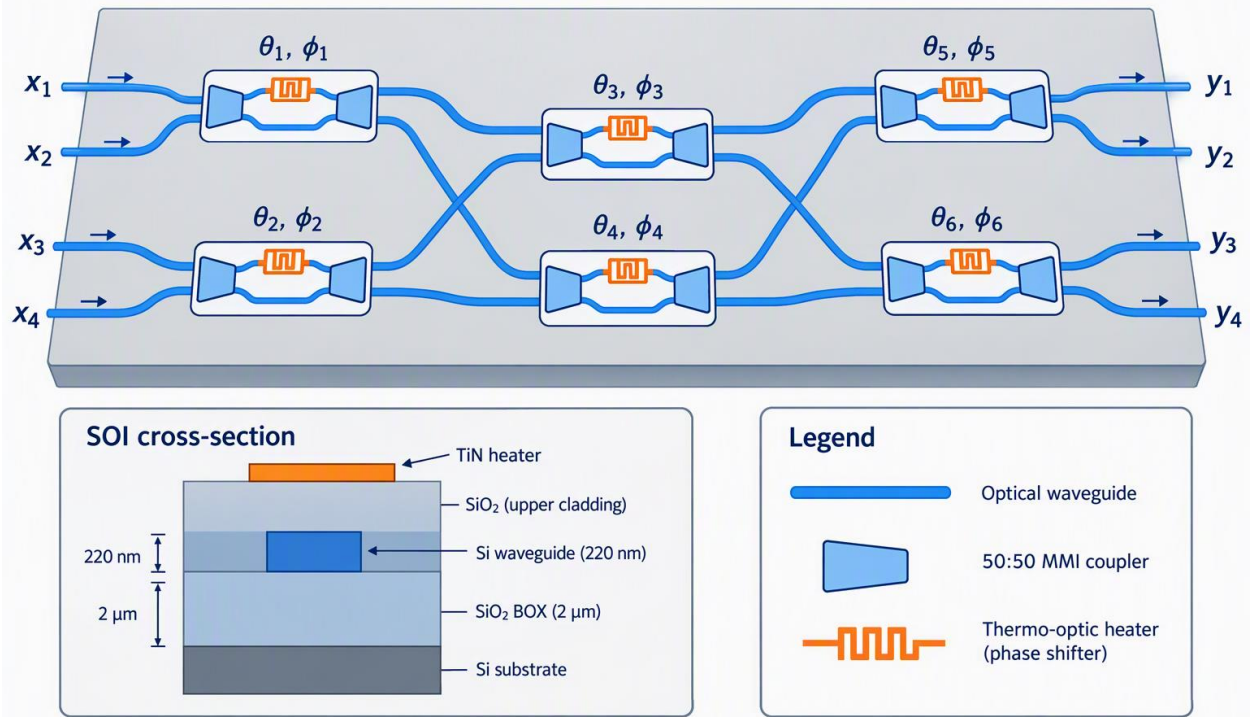


Fig. 3. A 4×4 Clements-decomposition MZI mesh on SOI with thermo-optic phase shifters.

6. Electro-Optic MAC Units

6.1 Operating Principle

Electro-optic multiply-accumulate (EO-MAC) units implement dot-product computation in the optical domain by encoding neural network weights as modulator bias voltages and input activations as optical pulse amplitudes or intensities. A modulator — typically a Mach-Zehnder modulator (MZM) or a microring modulator (MRM) — performs analog multiplication: the optical throughput power is a function of both the modulator transfer characteristic and the applied electrical bias (weight) and input signal. A balanced photodetector pair or a transimpedance amplifier array sums the optical power contributions from multiple weight-modulator-input product terms, implementing the accumulation step.

In WDM-stacked EO-MAC architectures, multiple input activations are encoded on distinct wavelength channels and broadcast along a shared optical path. A bank of wavelength-selective modulators applies

the weights associated with one output channel, and the photodetector sums the weighted optical powers to produce the corresponding dot product. Multiple detector or modulator banks can operate in parallel to generate additional output rows. This scheme increases MAC throughput in proportion to the number of usable WDM channels, typically 8–64, without requiring a separate waveguide for every input channel.

6.2 Platforms

Silicon microring modulators provide a compact, CMOS-compatible platform for electro-optic modulation [6], with typical device radii of approximately 5–10 μm and modulation bandwidths of 10–50 GHz, but they require precise thermal stabilization because of their narrow resonance linewidths (~ 0.1 nm). Thin-film lithium niobate (TFLN) MZMs provide extremely high electro-optic bandwidth (>100 GHz demonstrated), sub-1 V $\cdot\text{cm}$ $V_{\pi}L$, and low optical insertion loss (<1 dB), making them the preferred platform for ultra-high-speed EO-MAC implementations. InP Mach-Zehnder modulators offer monolithic laser integration for fully chip-integrated transmitter solutions. WDM stacking using dense wavelength-division multiplexed (DWDM) combs — generated by microresonator optical frequency combs — provides 64–256 parallel channels on a single waveguide bus.

6.3 Performance

EO-MAC units have demonstrated over 1 TOPS at energy efficiencies in the 10 fJ/MAC range in silicon photonic implementations. IBM Research and MIT have reported sub-picojoule MAC operations in prototype systems. Feldmann et al. demonstrated parallel convolutional processing with an integrated photonic tensor core [5]. Photonic tensor-core architectures and their potential performance advantages are reviewed in [17]. Related work by Feldmann et al. demonstrated all-optical spiking neurosynaptic networks with self-learning capabilities [7].

6.4 Integration with CMOS

The practical deployment of EO-MAC units requires tight co-integration with CMOS circuitry for digital-to-analog conversion (DAC) of weight signals, analog-to-digital conversion (ADC) of photocurrent outputs, and control logic. Two integration paradigms are pursued: monolithic co-integration, where photonic and electronic components are fabricated on the same silicon die using a modified CMOS process (e.g., GlobalFoundries 45SPCLO, IMEC iSiPP50G), and 3D-stacked heterogeneous integration, where a photonic die is bump-bonded to a CMOS logic/DRAM stack with sub-100 μm interconnect pitch. The ADC/DAC

bottleneck is a first-order constraint: 6-bit ADCs operating at 50 GS/s consume approximately 10–50 mW each, and a large-scale EO-MAC array requires hundreds of ADC channels. Optical I/O fiber coupling, using edge-coupled or grating-coupled approaches, imposes an additional alignment and packaging burden.

6.5 Limitations

The effective precision of EO-MAC units is principally limited by DAC/ADC resolution (typically 6–8 bits in high-speed implementations) and by the nonlinear transfer characteristic of modulators, which introduces harmonic distortion and inter-modulation products. Silicon microring modulators are acutely sensitive to chip temperature, with resonance wavelength drifting at approximately 0.1 nm/K, necessitating active thermal stabilization circuitry that consumes static power. In WDM architectures, channel crosstalk from ring modulator side-lobes and wavelength-selective crosstalk in arrayed waveguide gratings (AWGs) further limits effective ENOB.

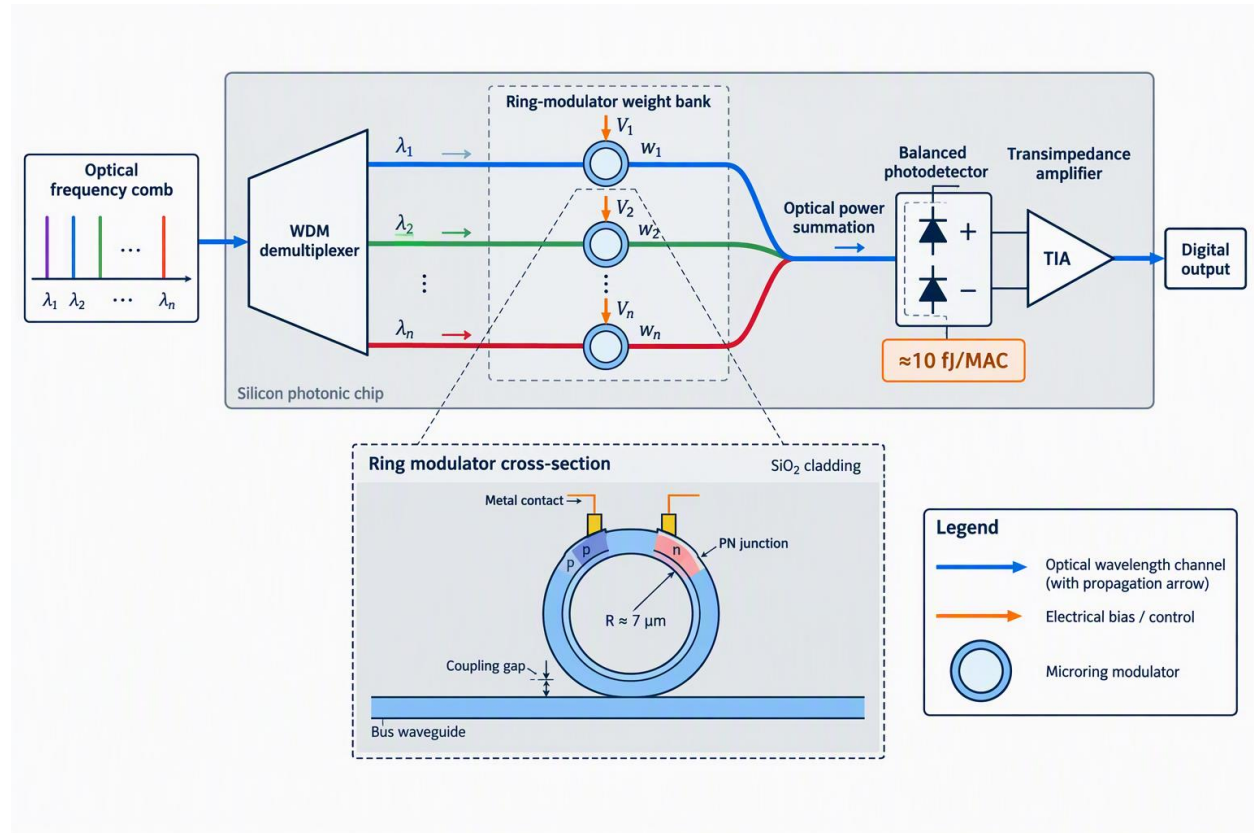


Fig. 4. WDM-parallel electro-optic MAC unit with ring-modulator weights and balanced photodetection.

7. Resonator-Based Inference

7.1 Operating Principle

Microring resonator (MRR) and photonic crystal nanocavity arrays implement wavelength-selective weighting by exploiting the sharp resonance transmission characteristic of high-Q optical cavities. When the wavelength of an incoming WDM channel coincides with the resonance wavelength of an MRR, light is coupled from the bus waveguide into the ring and either dissipated (all-pass filter configuration, implementing absorption) or redirected to a drop port (add-drop configuration, implementing selective routing). The resonance wavelength — and therefore the coupling efficiency for a given input wavelength — is tuned via the carrier plasma dispersion effect (carrier injection or depletion) or the thermo-optic effect, implementing weight programmability. An array of MRRs distributed across a WDM bus waveguide computes the elements of a weight matrix row, with each ring selectively attenuating or passing its

assigned wavelength channel. Photocurrent integration at a photodetector at the end of the bus sums contributions from all channels, completing the dot product [13].

7.2 Non-Volatile Weight Storage

A transformative capability of resonator-based photonic systems is the integration of phase-change materials (PCMs) for non-volatile optical weight storage. Materials such as GeSbTe (GST) and GeSbSeTe (GSST), when deposited on or near a microring resonator, modify the effective refractive index and absorption of the optical mode as their crystallographic state changes. For conventional GST at telecommunications wavelengths, the crystalline state generally has a higher refractive index and higher optical loss than the amorphous state; low-loss alternatives such as GSST are used when absorption must be reduced. The resulting change in resonance wavelength and transmission implements a non-volatile weight value. State transitions are induced by short optical or electrical pulses, with the required energy depending on material volume and device geometry. Integrated all-photonic phase-change memory has demonstrated non-volatile multilevel optical storage [16]. Feldmann et al. subsequently demonstrated PCM-based photonic synapses with self-learning capabilities and zero-static-power weight retention [7]. The non-volatile weight property eliminates the continuous power required by thermo-optic phase shifters in MZI meshes, representing a potentially significant system-level energy advantage for inference workloads with infrequently updated weights.

7.3 Photonic Spiking Neural Networks

Resonator-based excitable photonic devices — including VCSELs with saturable absorbers and microring lasers near lasing threshold — exhibit integrate-and-fire neuronal dynamics that directly map to spiking neural network (SNN) models. When an optical input pulse exceeds a threshold intensity, the device emits a coherent optical spike with sub-nanosecond latency and a well-defined refractory period, enabling temporal spike-rate coding of neural activations [12]. Arrays of such excitable resonators, coupled via on-chip waveguide networks with PCM-encoded synaptic weights, implement fully photonic spiking neural networks with spike propagation delays approaching the speed-of-light limit. Temporal coding schemes allow the relative spike timing to carry graded information, potentially circumventing the precision limitations of amplitude-coded analog systems. Related photonic crystal nanocavity networks have been investigated as compact physical reservoirs for neuromorphic computation [24].

7.4 Performance

Microring resonator arrays offer the highest integration density of any on-chip photonic compute element, with MRR pitch values of 5–10 μm enabling thousands of weight elements per square millimeter of chip area. Non-volatile PCM-based weight storage eliminates static power consumption for weight maintenance, leaving only the optical readout power (typically μW per MRR) and the spike generation energy (pJ per spike) as primary energy contributors. Effective precision is limited to approximately 3–5 bits by resonance wavelength variability (due to wafer-scale fabrication non-uniformity and temperature gradients), which perturbs the mapping between the intended weight value and the realized resonance detuning. Active wavelength locking via integrated heaters partially compensates fabrication variance at the cost of additional static power.

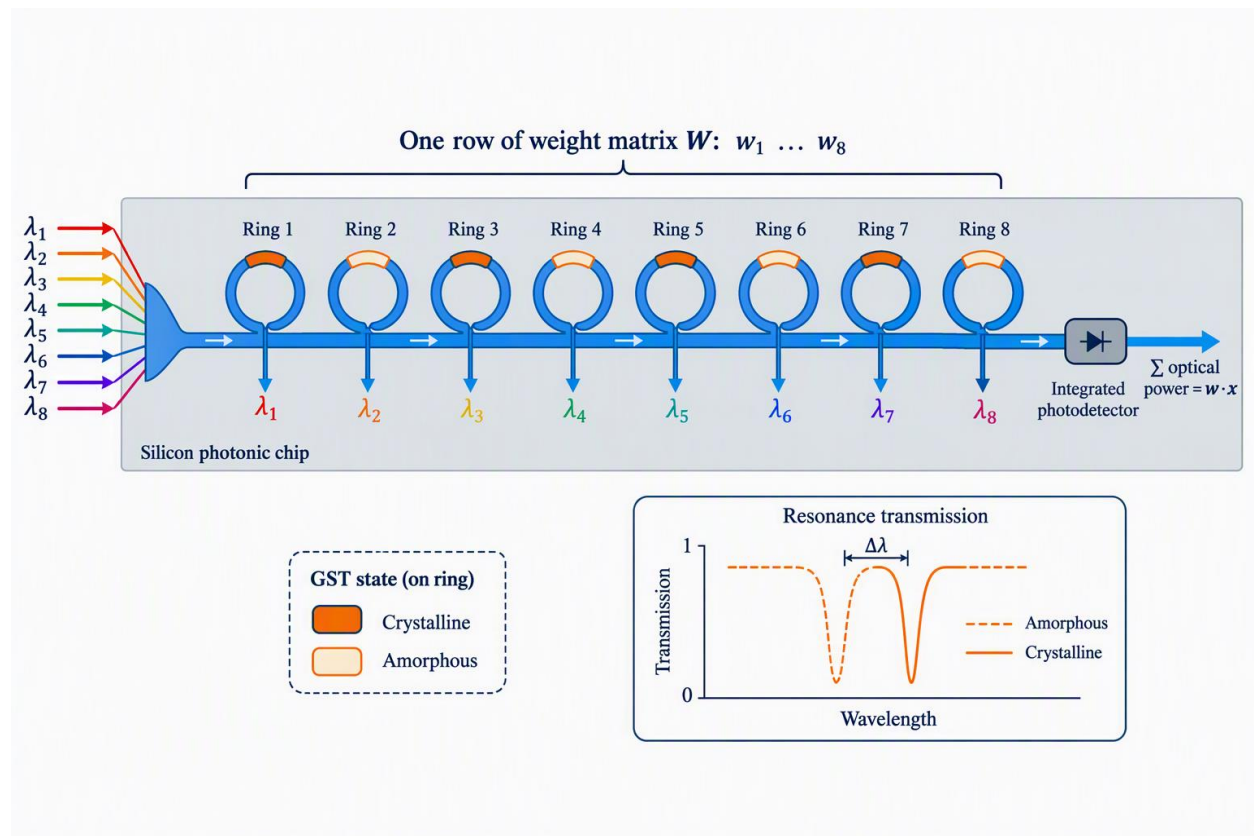


Fig. 5. Microring-resonator weight bank with GST nonvolatile weight storage on a WDM bus.

8. VCSEL-Array Processors

8.1 Operating Principle

Vertical-cavity surface-emitting laser (VCSEL) array processors exploit the two-dimensional spatial geometry of VCSEL arrays as optically-addressed neural processing planes. Each VCSEL element functions as a nonlinear optical neuron: it integrates received optical input power (from upstream neurons) and, upon exceeding the lasing threshold, emits a coherent optical output beam normal to the chip surface. The threshold nonlinearity of the VCSEL lasing process provides an inherent sigmoid-like or step activation function without requiring any separate electronic nonlinear element. Two-dimensional VCSEL arrays with pixel counts up to 1000×1000 and emitter pitches of $10\text{--}50\ \mu\text{m}$ have been fabricated on GaAs substrates using standard epitaxial growth and lithographic patterning. Free-space optical interconnects — implemented with refractive microlens arrays, holographic optical elements (HOEs), or diffractive optical elements — route the output beams of one VCSEL array plane to the input apertures of the next, implementing the synaptic weight matrix as the spatial transfer function of the interconnect optic.

8.2 Speed and Connectivity

Individual VCSEL elements support direct modulation bandwidths of $10\text{--}30\ \text{GHz}$, enabling neuron firing rates in the gigahertz range — orders of magnitude faster than biological neurons and faster than most electronic spike-encoding circuits. The critical architectural advantage of VCSEL-based free-space systems is the connectivity scaling: an $N \times N$ VCSEL plane coupled through a single $4f$ lens system to a detector array achieves all-to-all $O(N^2)$ connectivity in a single optical propagation step. This contrasts with on-chip waveguide networks, where routing $O(N^2)$ connections between N nodes requires $O(N^2)$ waveguide crossings, incurring proportional crosstalk and insertion loss. Holographic optical elements programmed with Fourier-domain phase patterns implement arbitrary complex weight matrices as the spatial transfer function of the free-space interconnect, with the hologram serving as the programmable synaptic weight store.

8.3 Reservoir Computing with VCSELs

A particularly effective application of VCSEL-based photonic systems is reservoir computing (RC), a paradigm in which a fixed, high-dimensional nonlinear dynamical system (the reservoir) is driven by the input signal, and only a linear readout layer is trained. The rich transient dynamics of a mutually coupled

VCSEL network — including relaxation oscillations, excitable spiking, and polarization switching — constitute a natural high-dimensional reservoir. Appeltant et al. demonstrated information processing with a single nonlinear node and delayed feedback [8]. Larger et al. reported an optoelectronic implementation of reservoir computing [10], and Brunner et al. demonstrated parallel photonic information processing at gigabyte-per-second data rates using transient states [15]. Subsequent VCSEL-specific demonstrations achieved word error rates below 1% on NIST spoken digit benchmarks at processing speeds of 1 GS/s. The single-node time-multiplexed RC architecture dramatically reduces hardware complexity by emulating a large virtual network through temporal multiplexing of a single physical VCSEL with a feedback delay line.

8.4 Nonlinearity

The lasing threshold of a VCSEL provides an inherent optical nonlinearity that eliminates the need for a separate activation function element — a significant advantage over MZI mesh systems and EO-MAC units, which require electronic nonlinear components between optical linear transform stages. The sharpness of the VCSEL threshold characteristic can be tuned by adjusting the bias current relative to threshold, implementing soft (sigmoid-like) or hard (step-like) nonlinearities as required by the network architecture.

8.5 Limitations

VCSEL-array processors face significant engineering challenges in scaling to large, fully programmable systems. Because the intended free-space interconnect spans only millimeter-scale distances, optical propagation and coupling losses can be far lower than in network-distance links, reducing the required launch power. Array-level consumption is therefore governed less by propagation distance than by VCSEL threshold and bias current, duty cycle, and the number of emitters active simultaneously. Near-threshold operation with sparse or event-driven activation can keep average power substantially below a simple full-array estimate; kilowatt-class delivery would apply only to the limiting case in which approximately one million emitters each draw about 1 mW continuously. Thermal cross-talk in densely packed VCSEL arrays can still induce wavelength and threshold shifts in adjacent elements, degrading computation fidelity. Free-space optical alignment stability over operational temperature ranges remains a systems engineering challenge because sub-micron misalignment at the holographic interconnect plane can cause cross-talk between neural channels. Achieving large-scale programmability — i.e., the ability to rapidly

reconfigure the inter-VCSEL coupling weights without replacing the holographic element — also remains an open research problem.

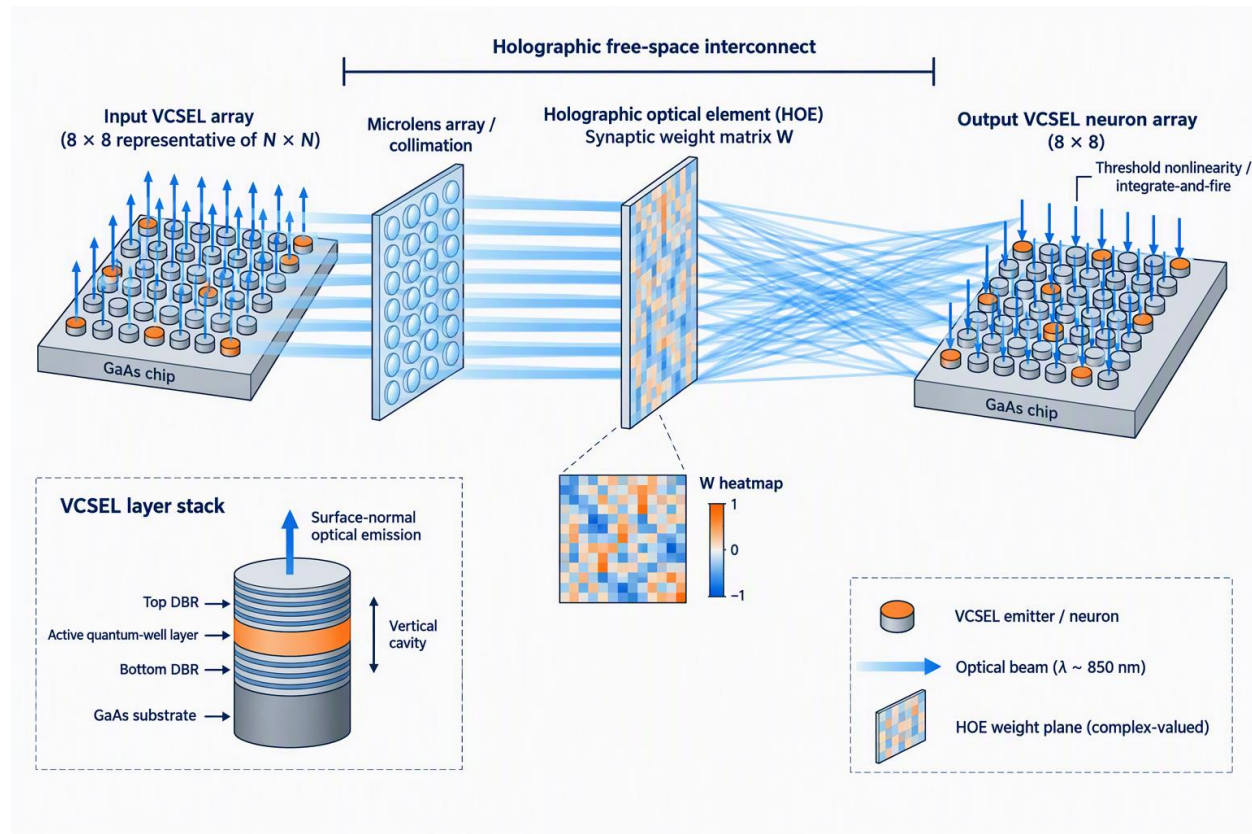


Fig. 6. VCSEL-array processor with a holographic free-space interconnect and threshold nonlinearity.

9. Free-Space Optical Accelerators

9.1 Operating Principle

Free-space optical (FSO) accelerators implement large-scale matrix operations by exploiting the spatial degrees of freedom of coherent or incoherent optical beams propagating through bulk-optic systems. In a canonical $4f$ optical architecture—two converging lenses separated by the sum of their focal lengths—an input spatial pattern at the front focal plane of the first lens is Fourier transformed at the shared focal plane. A spatial light modulator placed at that Fourier plane multiplies the spectrum by a programmable transfer function, and the second lens performs the corresponding inverse transform, producing the filtered field at the output plane, up to the system's inversion and scaling conventions. This arrangement

implements spatial-frequency filtering and therefore convolution of the input field. Convolutional neural network layers used in image recognition and segmentation can map to this operation, allowing a complete linear convolution to be executed in one optical propagation pass.

9.2 SLM-Based Inference

Liquid-crystal-on-silicon (LCoS) SLMs provide megapixel weight planes with phase modulation depth exceeding 2π at 1550 nm, refreshable at 60–120 Hz. Digital micromirror devices (DMDs) offer binary-amplitude modulation at refresh rates up to 22 kHz, enabling pulsed inference at kilohertz rates. The input activation is encoded as an intensity or complex amplitude pattern on a coherent laser beam (via a second SLM or a spatial light source array), and the output is read by a CCD or sCMOS detector array. The number of multiply-accumulate operations per inference pass scales as N^2 for an $N \times N$ weight plane, with N routinely exceeding 10^6 for megapixel SLMs, yielding theoretical throughputs in the peta-OPS/s range for a single optical pass.

9.3 Optical Fourier Neural Networks

Wetzstein et al. reviewed the framework of optical neural-network inference in 2020 [9], including lens-based systems that perform two-dimensional Fourier transforms through propagation rather than arithmetic clock cycles. Alternating Fourier-domain SLM multiplication with spatial-domain nonlinear activation can form a hybrid optical-electronic inference pipeline. In such systems, the optical stage performs the linear spatial transform, while detector readout, nonlinear activation, calibration, and correction remain electronic.

9.4 Performance

Free-space optical accelerators offer the largest spatial parallelism of any photonic inference platform — megapixel weight planes provide $O(10^6)$ parallel multiply operations per propagation pass. This massive parallelism is ideally matched to spatial inference tasks (image classification, object detection, medical imaging analysis) and to applications where the weight matrix is large, slowly updated, and applied to many input patterns. The principal throughput limitation is SLM refresh rate: at 120 Hz update rate, a megapixel LCoS SLM provides 120 matrix-vector products per second, which constrains sequential token inference throughput. Physical footprint (typically 0.1–1 m² per system) and sensitivity to mechanical vibration and ambient light restrict deployment to controlled laboratory or server-room environments.

9.5 Hybrid Digital-Optical Systems

Practical FSO inference systems generally use a hybrid architecture in which the optical Fourier front-end performs a linear spatial transformation, while a digital CMOS back-end implements nonlinear activation, residual addition, normalization, pooling, calibration, and control. This division assigns highly parallel spatial operations to the optical stage and algorithmically complex operations to electronics. The digital-to-optical interface, including laser modulation and SLM programming, and the optical-to-digital interface, including detector readout and conversion, remain primary engineering bottlenecks.

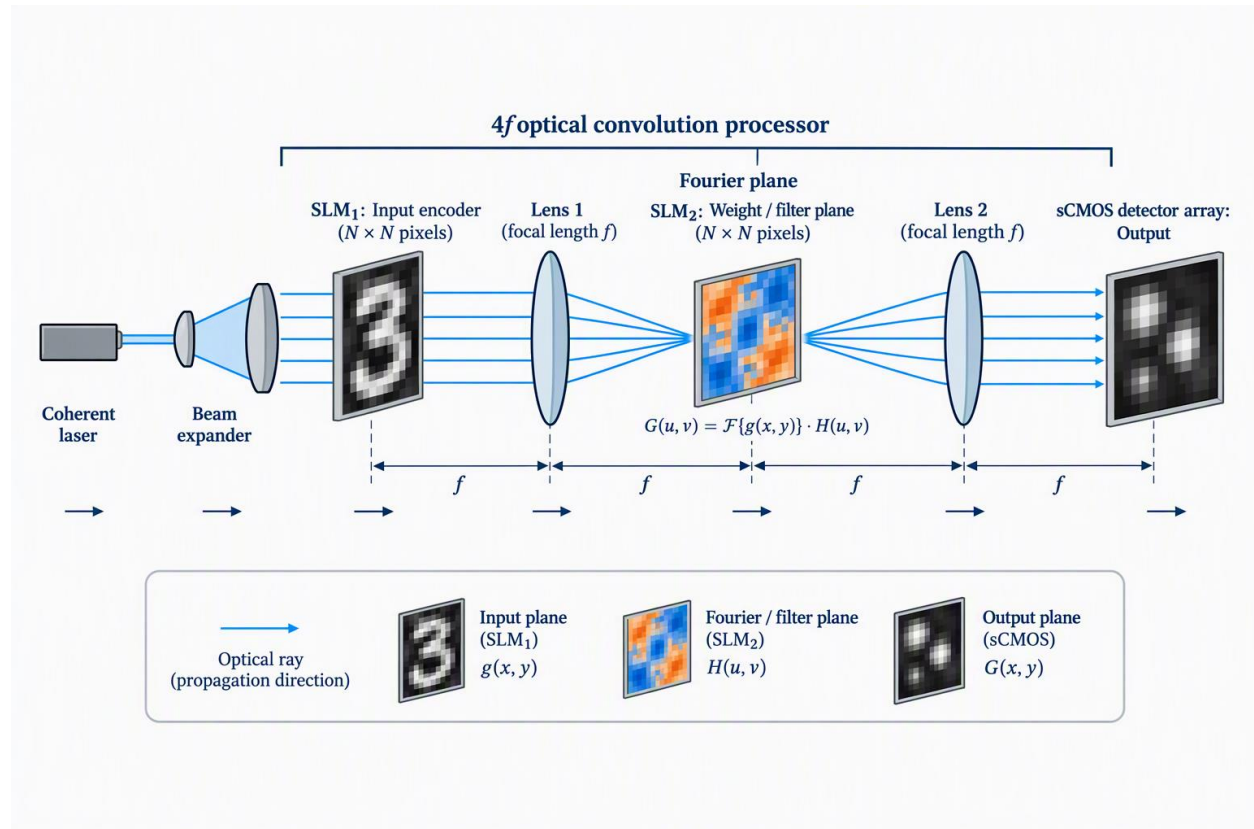


Fig. 7. A 4f free-space optical processor implementing convolution with an SLM at the Fourier plane.

10. Comparative Analysis

10.1 Head-to-Head Comparison Table

Table I compares the six photonic inference architecture families across ten dimensions, covering implementation characteristics, performance, scalability, technology readiness, and representative demonstrations.

TABLE I — Comparative Analysis of Photonic Inference Architecture Families

Dimension	D ² NN	MZI Mesh	EO-MAC Unit	Resonator (MRR/PCM)	VCSEL Array	Free-Space (SLM/4f)
Compute Substrate	Polymer, SiO ₂ /TiO ₂ metasurface, SLM	SOI, LiNbO ₃ -on-Si, SiN	SOI, TFLN, InP	SOI + GST/GSST PCM	GaAs VCSEL + free-space HOE	Bulk optics, LCoS SLM, DMD
Weight Encoding	Phase/amplitude mask (static)	MZI phase angles (θ , ϕ) via TO or EO	Modulator bias voltage (EO)	Resonance detuning / PCM state	HOE hologram / free-space coupling	SLM pixel phase / amplitude
Programmability	Fixed (fabricated)	Fast (μ s–ps)	Ultra-fast (ps–ns)	Non-volatile / Slow (ns– μ s PCM)	Fixed (HOE) / Slow (LC-SLM)	Slow (ms–s, 60–22k Hz)
Throughput (TOPS)	$\sim 10^3$ – 10^6 (theoretical, optical freq.)	10–100 TOPS (w/ WDM)	1–100+ TOPS	0.1–10 TOPS	0.01–1 TOPS	~ 10 –1000 TOPS (spatial parallel)
Energy Efficiency	$\sim 10^6$ TOPS/W (theoretical passive)	10–100 TOPS/W	~ 9 –100 TOPS/W (exp.)	~ 10 –50 TOPS/W (non-volatile)	~ 0.1 –1 TOPS/W	~ 1 –10 TOPS/W (system-level)
Effective Precision (ENOB)	4–6 bits	4–6 bits	4–6 bits	3–5 bits	3–5 bits (threshold coding)	5–7 bits (spatial, intensity)
Scalability	Large (cm ² –m ² aperture)	Medium ($N \leq 256$ on-chip)	Medium (WDM $\times$ spatial)	High density (5 μ m pitch)	Medium (1000 $\times$ 1000 VCSEL)	Very Large (megapixel SLM)
Nonlinearity Source	Optoelectronic (between layers)	Optoelectronic (electronic)	Optoelectronic (electronic)	Excitable resonator (optical spike)	VCSEL threshold (optical)	Optoelectronic (electronic)
Technology Readiness (TRL)	TRL 4–5	TRL 5–7	TRL 5–7	TRL 3–5	TRL 3–4	TRL 4–6
Representative Demonstrations / Companies	Lin et al. 2018 (Science); UCLA; MIT; Wetzstein Group	Shen et al. 2017; Lightmatter; Luminous; MIT LL; Stanford	Feldmann et al. 2021; IBM; MIT; Analog Devices; Lightmatter	Feldmann et al. 2019; Oxford; EPFL; IBM Zurich	Appeltant et al. 2011; ULB Brussels; Ghent University	Optalysys; Neuroptica (MIT); Wetzstein 2020; DARPA DESI

10.2 Energy-Throughput Trade-off Discussion

Figure 8 maps the six photonic architecture families on an energy-efficiency-versus-throughput scatter plot, with CMOS references—NVIDIA H100 and Google TPUv5—shown for context. The photonic design space separates into distinct regions: EO-MAC units and silicon MZI meshes occupy the high-throughput, high-efficiency portion of the reconfigurable category; passive D²NN and free-space SLM systems extend toward extreme theoretical throughput for fixed inference tasks; and resonator-based systems with non-volatile PCM weights occupy a high-efficiency, moderate-throughput region.

Photonics is most compelling for ultra-low-latency analog compute with infrequently updated weights. For applications such as fixed-function image classification, hyperspectral sensing, or radar signal

processing — where the same weight matrix is applied to a continuous input stream — passive optical propagation can make the incremental energy of the matrix transform very small. System-level comparisons must still include light generation, modulation, photodetection, data conversion, control, calibration, and memory movement. Photonics currently struggles to combine high precision, high programmability, and high throughput simultaneously: no existing photonic system matches the NVIDIA H100's combination of 8-bit precision, arbitrary model support, and sustained data-center-scale operation. The photonic precision ceiling at 4–6 bits ENOB remains a significant near-term barrier to transformer-based LLM inference, where FP8 precision is increasingly used in production deployments.

10.3 Precision Limits

The 4–6 bit ENOB ceiling common to analog photonic inference systems arises from the interplay of three noise sources: (1) shot noise, which scales as $N^{-1/2}$ with photon number N and imposes an approximately 6-bit SNR limit at moderate optical power levels; (2) thermal noise in receiver electronics, which contributes an additional 1–2 bit degradation; and (3) fabrication-induced structural imperfections — coupling ratio variance, waveguide width variation, and resonator frequency spread — which introduce systematic errors that appear as quantization noise. Quantization-aware training (QAT) with photonic noise injection models has been shown to mitigate much of the accuracy degradation associated with 4-bit precision, enabling competitive accuracy on image classification tasks (ImageNet top-5 accuracy within 2–3% of FP32 baseline). Digital error correction, implemented as a learned calibration matrix applied to the photonic output vector before the digital activation function, can further improve effective precision by 1–2 bits in static (fixed-weight) photonic systems.

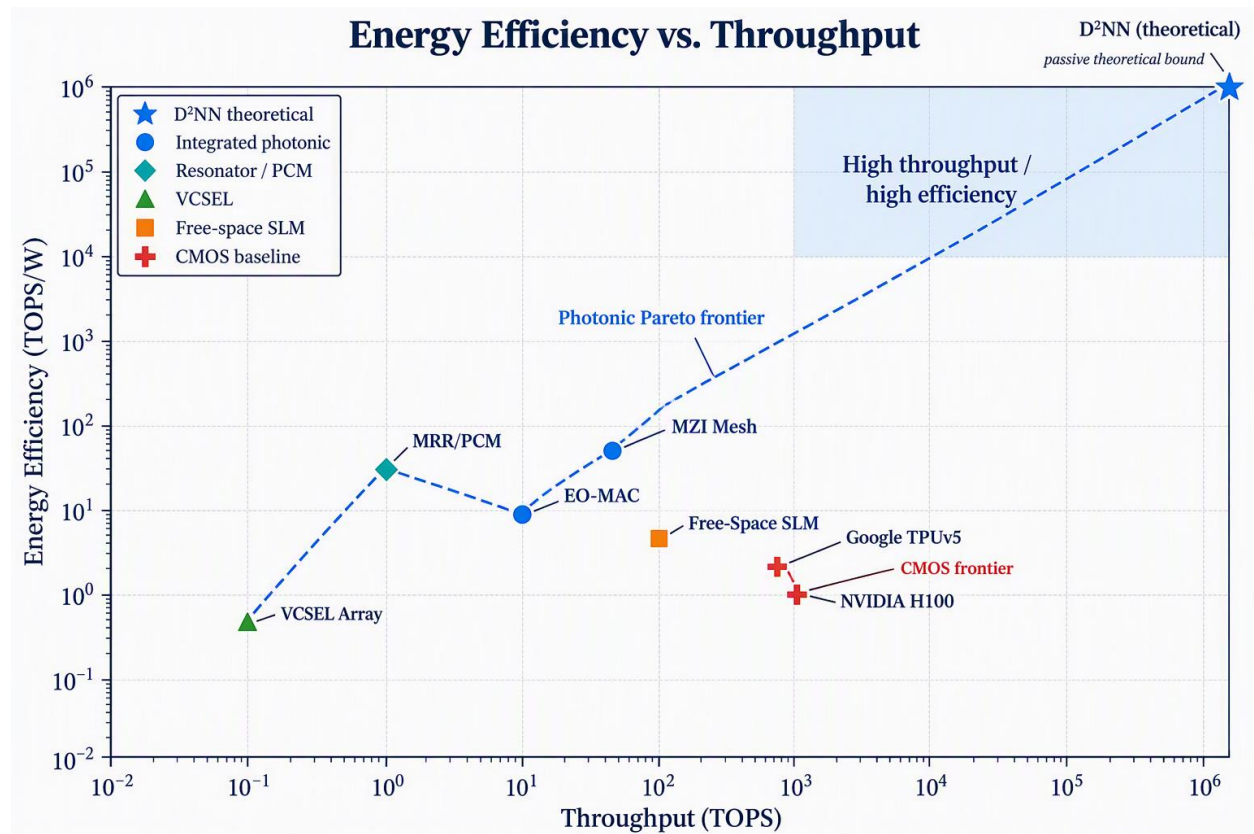


Fig. 8. Energy efficiency versus throughput for photonic architecture families and CMOS reference accelerators.

11. Scalability and Future Directions

The trajectory of photonic inference architecture scalability converges on several enabling technology developments that are expected to mature over the 2026–2035 horizon.

Photonic chiplet integration and co-packaged optics represent the nearest-term scalability pathway. Co-packaged optics (CPO) — in which photonic chiplets are bump-bonded directly onto the processor package substrate, replacing electrical I/O with optical I/O — has been adopted by hyperscalers (Meta, Microsoft, Google) for switch interconnect, and the same architecture enables high-density photonic-CMOS interface for inference accelerators. Ayar Labs' TeraPHY optical I/O chiplet demonstrates 2 Tb/s in a 12 mm² die area, providing the I/O bandwidth necessary for large photonic MAC arrays [23].

3D-stacked photonic-CMOS heterogeneous integration enables tight co-location of photonic compute layers with DRAM and CMOS logic in a single stacked package, dramatically reducing the ADC/DAC

interconnect latency and the optical I/O coupling loss that are primary overheads in 2D co-integration. MIT Lincoln Laboratory's 3D-integrated photonic-electronic platform has demonstrated sub-5 μm bonding pitch with electrical interconnect density exceeding 10^6 connections/ cm^2 .

Neuromorphic photonics at scale may prove to be the domain where photonic systems most naturally exceed CMOS performance, given the inherent analog-temporal dynamics of excitable photonic devices. Spike-based temporal coding, combined with PCM-encoded non-volatile synaptic weights, could yield fully photonic SNN inference systems with sub-picojoule per spike energy budgets, enabling event-driven inference at unprecedented energy efficiency for sparse, asynchronous input streams.

Quantum dot and 2D-material photonic devices — including InAs quantum dot lasers on Si (enabling chip-scale laser integration on SOI), MoS_2 and hBN photodetectors, and graphene electro-optic modulators — provide a materials expansion of the photonic integration toolkit beyond classical III-V and IV materials, with implications for both operating wavelength range and achievable modulation bandwidth.

Learned optical front-ends for edge inference exploit the D^2NN principle in ultra-compact form factors for always-on edge AI sensors. Integrating the first layers of a neural network directly into the camera optical system — as a diffractive or metasurface-based optical preprocessor — reduces digital computation requirements for image classification tasks at the sensor by one to two orders of magnitude, enabling sub-milliwatt inference for IoT edge devices.

A potential path toward system-level efficiencies above 100 TOPS/W combines low-energy EO-MAC devices on TFLN platforms, WDM parallelism, non-volatile weight storage, and tightly integrated CMOS for activation and digital correction. Achieving this target would require the complete system — including lasers, modulators, converters, calibration, control, and memory traffic — to remain within the stated power boundary. Device-level results suggest that 50–200 TOPS/W may be technically plausible for narrow-precision, fixed-function workloads, but this range remains a projection rather than a demonstrated end-to-end result.

12. Commercialization Pathways

The commercial landscape for photonic inference accelerators is characterized by a small number of well-capitalized startups, large established semiconductor companies with photonics divisions, and hyperscaler in-house development programs. **Lightmatter** (Boston, MA) has commercialized its Passage photonic interconnect fabric and is developing the Envisé photonic inference processor, based on MZI mesh MVM with CMOS-integrated weight control and WDM input encoding [22]. **Luminous Computing** (Sunnyvale, CA) has articulated a roadmap to a petaFLOP-scale photonic processor for LLM inference. **Celestial AI** is developing optical disaggregation fabrics that use photonics to connect memory and compute chiplets across a high-bandwidth optical bus, addressing the memory wall without requiring photonic computation itself. **Ayar Labs** focuses on optical I/O chiplets (TeraPHY) that provide the high-bandwidth chip-to-chip connectivity prerequisite for photonic compute integration. **Intel Silicon Photonics** and **IBM's photonic computing research group** pursue integrated photonic-CMOS platforms within their broader semiconductor development programs.

Market drivers for photonic inference acceleration are closely aligned with the LLM inference capacity crisis: data center operators face power density limits of 20–50 kW per rack that constrain GPU cluster scaling, creating demand for more energy-efficient inference alternatives. High-performance computing (HPC) interconnect, autonomous vehicle sensor fusion, and real-time edge AI for industrial inspection and medical imaging represent additional verticals where photonic inference provides differentiated value propositions.

Key barriers to commercialization include: (1) packaging yield — fiber-to-chip coupling and photonic die bonding remain low-yield processes relative to CMOS packaging; (2) the absence of mature photonic-electronic co-design EDA tools — designing mixed photonic-CMOS systems currently requires disparate and poorly integrated tool chains; (3) the lack of open photonic process design kits (PDKs) and standardized photonic interface specifications; and (4) volume manufacturing cost — photonic foundry capacity is an order of magnitude smaller than leading CMOS foundries, driving high per-wafer costs.

A plausible commercialization scenario is niche deployment during 2026–2028 for fixed-function inference applications such as optical Fourier processing for imaging and EO-MAC acceleration for signal processing. Broader data-center integration of photonic-CMOS hybrid accelerators may follow during

2030–2035 if packaging yield, calibration, software support, and electronic-photonic design tools mature sufficiently. Commodity edge-photonic inference would likely require further expansion of foundry capacity, standardized interfaces, and substantial reductions in packaging cost.

13. Conclusion

This paper has presented a comprehensive, axis-aligned taxonomy of six photonic inference architecture families — diffractive optical networks, silicon photonic MZI meshes, electro-optic MAC units, resonator-based inference engines, VCSEL-array processors, and free-space optical accelerators — characterized along four classification axes (spatial encoding, compute paradigm, programmability, and integration level) and compared across ten performance dimensions.

The analysis reveals that no single photonic architecture currently dominates across all performance dimensions. Passive diffractive systems and free-space SLM accelerators offer exceptional theoretical throughput and energy efficiency for fixed-function spatial inference but are limited by restricted runtime programmability and large physical footprints. Silicon MZI meshes and electro-optic MAC units are among the most mature and commercially relevant photonic inference platforms because they combine reconfigurability, high reported throughput, and compatibility with CMOS-oriented manufacturing and packaging. Resonator-based architectures with phase-change-material weight storage offer a pathway toward non-volatile, high-density inference, whereas VCSEL-array and free-space systems are best matched to reservoir-computing and fixed-pattern spatial workloads, respectively. Comparisons of TOPS/W across these families should remain workload- and boundary-aware because published values do not always include the same laser, conversion, control, and memory overheads.

The precision frontier — currently at 4–6 bits ENOB for analog photonic systems — is a critical near-term barrier to deployment in transformer-based LLM inference. Progress in quantization-aware training, in-situ calibration, and digital correction may narrow the gap to the 8-bit precision commonly used for production inference. Photonic integration, neuromorphic computing, and quantum information processing share enabling substrates in SOI, LNOI, and III-V platforms, but their convergence into deployable physical-computing systems remains prospective. Whether photonic architectures can complement or displace CMOS at data-center scale during 2025–2035 will depend on demonstrated

improvements in manufacturability, programmability, precision, packaging, and end-to-end system efficiency.

Suggested citation: A. Falkenberg, "Classification of Photonic Inference Architectures," Version 1.0, technical white paper and preprint, QPI Laboratories, October 5, 2026.

AI-Assistance Disclosure

Generative AI tools were used to assist with language refinement, organization, and editorial review. AI-generated material was not treated as an authoritative technical source. The author reviewed and revised the resulting text and is responsible for the technical content, interpretations, citations, and conclusions.

References

- [1] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, no. 6406, pp. 1004–1008, Sep. 2018.
- [2] Y. Luo, D. Mengü, N. T. Yardimci, Y. Rivenson, M. Veli, M. Jarrahi, and A. Ozcan, "Design of task-specific optical systems using broadband diffractive neural networks," *Light: Science & Applications*, vol. 8, no. 1, p. 112, Dec. 2019.
- [3] 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, no. 7, pp. 441–446, Jul. 2017.
- [4] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, "Optimal design for universal multiport interferometers," *Optica*, vol. 3, no. 12, pp. 1460–1465, Dec. 2016.
- [5] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, Jan. 2021.
- [6] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, "Micrometre-scale integrated silicon opto-electronic modulator," *Nature*, vol. 435, no. 7040, pp. 325–327, May 2005.
- [7] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, pp. 208–214, May 2019.
- [8] L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer, "Information processing using a single dynamical node as complex system," *Nature Communications*, vol. 2, p. 468, Sep. 2011.

- [9] G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. B. Miller, and D. Psaltis, "Inference in artificial intelligence with deep optics and photonics," *Nature*, vol. 588, no. 7836, pp. 39–47, Dec. 2020.
- [10] C. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer, "Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing," *Optics Express*, vol. 20, no. 3, pp. 3241–3249, Jan. 2012.
- [11] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," *Physical Review Letters*, vol. 73, no. 1, pp. 58–61, Jul. 1994.
- [12] M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, "A leaky integrate-and-fire laser neuron for ultrafast cognitive computing," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, no. 5, p. 1800212, Sep.–Oct. 2013.
- [13] A. N. Tait, T. Ferreira 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, p. 7430, Aug. 2017.
- [14] B. J. Shastri, A. N. Tait, T. F. de Lima, W. H. P. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, "Photonics for artificial intelligence and neuromorphic computing," *Nature Photonics*, vol. 15, no. 2, pp. 102–114, Feb. 2021.
- [15] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, "Parallel photonic information processing at gigabyte per second data rates using transient states," *Nature Communications*, vol. 4, p. 1364, Jan. 2013.
- [16] C. Ríos, M. Stegmaier, P. Hosseini, D. Wang, T. Scherer, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "Integrated all-photonic non-volatile multilevel memory," *Nature Photonics*, vol. 9, no. 11, pp. 725–732, Oct. 2015.
- [17] M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning," *Applied Physics Reviews*, vol. 7, no. 3, p. 031404, Sep. 2020.
- [18] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica*, vol. 5, no. 7, pp. 864–871, Jul. 2018.
- [19] M. Y. S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, "Design of optical neural networks with component imprecisions," *Optics Express*, vol. 27, no. 10, pp. 14009–14029, May 2019.
- [20] H. Zhang, M. Gu, X. D. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. H. Yung, Y. Z. Shi, F. K. Muhammad, G. Q. Lo, X. S. Luo, B. Dong, D. L. Kwong, L. C. Kwek, and A. Q. Liu, "An optical neural chip for implementing complex-valued neural networks," *Nature Communications*, vol. 12, p. 457, Jan. 2021.
- [21] S. Bandyopadhyay, A. Sludds, S. Krastanov, R. Hamerly, N. Harris, D. Bunandar, M. Streshinsky, M. Hochberg, and D. Englund, "Single chip photonic deep neural network with accelerated training," *Nature Communications*, vol. 15, p. 2847, Apr. 2024.
- [22] Lightmatter, Inc., "Enviser: A photonic AI computing platform," Technical White Paper, Cambridge, MA, 2023. [Online]. Available: <https://lightmatter.co>
- [23] Ayar Labs, Inc., "TeraPHY optical I/O chiplet: Architecture and performance," Product Brief, Santa Clara, CA, 2024. [Online]. Available: <https://ayarlabs.com>
- [24] M. Debruyne, J. Dambre, and P. Bienstman, "Reservoir computing with a photonic crystal nanocavity network," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 29, no. 2, pp. 1–10, Mar.–Apr. 2023.

[25] W. A. Wulf and S. A. McKee, "Hitting the Memory Wall: Implications of the Obvious," *ACM SIGARCH Computer Architecture News*, vol. 23, no. 1, pp. 20–24, Mar. 1995, doi: 10.1145/216585.216588.

© 2026 QPI Laboratories, Inc. All rights reserved. Preprint / Technical White Paper, Version 1.0.

Prepared October 5, 2026. Escondido, CA, United States.