International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Photonic Neural Networks and Optical AI Accelerators: A Comprehensive Review of Architectures, Material Platforms, and System-Level Challenges

Abubakar Umar Hamza
Lumentum Technology United Kingdom
Abubakar.UmarHamza@lumentum.com

Abstract- The rapid advancement of artificial intelligence (AI), particularly deep learning and large-scale neural networks, has created significant demand for high-performance and energy-efficient computing architectures. Conventional electronic processors such as GPUs and TPUs are increasingly constrained by power consumption, memory bandwidth limitations, and data movement bottlenecks. In response, photonic neural networks and optical AI accelerators have emerged as promising alternatives that exploit the properties of light to perform computation at high speed and low energy consumption. This paper presents a comprehensive systematic narrative review of photonic neural networks and optical AI accelerators, focusing on their architectures, material platforms, and key engineering challenges. The methodology employed involves structured literature collection from recent peer-reviewed studies, thematic classification of photonic architectures (including Mach–Zehnder interferometer meshes, microring resonator networks, and diffractive optical systems), and comparative analysis of material platforms and performance metrics such as energy efficiency, scalability, and computational latency. The results of the review indicate that photonic systems offer significant advantages over electronic computing, particularly in terms of energy per multiply–accumulate operation (femtojoule-level), ultra-high bandwidth (terahertz range), and low-latency computation. However, practical deployment remains limited by challenges in scalability, fabrication variability, noise sensitivity, and the lack of efficient optical training mechanisms. The analysis further shows that hybrid photonic–electronic architectures currently represent the most viable pathway toward near-term implementation, while heterogeneous material integration is essential for achieving fully functional photonic AI systems. The contribution of this research lies in providing a structured and critical synthesis of recent advancements in photonic AI hardware, identifying key technological bottlenecks, and outlining future research directions toward scalable and commercially viable optical computing systems. This work serves as a reference framework for researchers working at the intersection of photonics, electronics, and artificial intelligence.

Keywords: Photonic neural networks, optical AI accelerators, silicon photonics, integrated photonics, neuromorphic computing, optical computing, hybrid architectures.

I. INTRODUCTION

Artificial intelligence (AI) has experienced unprecedented growth over the past decade, driven by breakthroughs in deep learning, generative AI systems, and large-scale language models. Modern neural networks now contain billions of parameters and require massive computational resources for both training and inference. At the core of these workloads lies repeated matrix–vector and matrix–matrix multiplications, which account for the majority of computational complexity in neural network operations (Xu et al., 2025). As model sizes continue to scale, the demand for faster, more energy-efficient hardware has become a critical challenge for both academia and industry. Conventional electronic computing platforms, particularly Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs), have enabled the current AI revolution but are increasingly approaching fundamental physical and architectural limits (Perera et al., 2024). These systems face growing challenges related to high energy consumption, excessive heat generation, limited memory bandwidth, and the inefficiencies associated with data movement between memory and processing units. In large data centers, the power required to support AI workloads has become a major economic and environmental concern, motivating the search for alternative computing paradigms capable of sustaining continued AI growth (Sheng et al., 2026). In response to these limitations, photonic computing has emerged as a promising candidate for next-generation AI hardware. Unlike electronic systems that rely on charge transport, photonic systems use light to carry and process information. Optical signals can propagate at extremely high speeds, support massive parallelism through wavelength-division multiplexing, and perform linear operations such as interference and superposition naturally in the optical domain (Charalampous et al., 2025). These characteristics make photonics particularly well suited for implementing the linear algebra operations that dominate neural network computation. Photonic Neural Networks (PNNs) therefore represent a transformative computing

paradigm in which neural network operations are executed using photonic integrated circuits. By leveraging optical interference, phase modulation, and wavelength multiplexing, PNNs have the potential to perform large-scale computations at the speed of light while consuming orders of magnitude less energy than electronic counterparts. As a result, photonic AI accelerators are increasingly viewed as a key enabling technology for future high-performance, energy-efficient AI systems.

This research is therefore motivated by the need to critically examine the rapid developments in photonic neural networks and optical AI accelerators, synthesize existing knowledge, and identify the opportunities and challenges that will shape the future of this emerging field. Specifically, the review aims to provide a comprehensive analysis of the architectures, material platforms, and integration strategies enabling photonic AI hardware, while evaluating their performance relative to conventional electronic systems. By consolidating current progress and highlighting open research problems, this work intends to guide future research efforts and support the development of scalable, practical, and commercially viable photonic AI computing technologies.

II. LITERATURE REVIEW

Evolution of Optical Computing

The concept of optical computing dates back to the 1980s, when early research explored the use of free-space optics for performing mathematical operations such as Fourier transforms and convolution. These early systems demonstrated the inherent parallelism and high bandwidth of light; however, they suffered from major limitations including bulky optical setups, lack of scalability, and poor integration with electronic systems (McMahon, 2023). As a result, the field experienced a period of slow progress due to the absence of practical fabrication and integration technologies. The resurgence of optical computing in recent years has been largely driven by advances in silicon photonics and photonic integrated circuit (PIC) fabrication. Leveraging mature complementary metal–oxide–semiconductor (CMOS) manufacturing processes, researchers can now integrate thousands of optical components such as waveguides, modulators, and detectors onto a single chip. This technological breakthrough has enabled the development of compact, scalable, and cost-effective optical computing hardware. Consequently, photonic computing has re-emerged as a viable solution to address the growing computational demands of artificial intelligence.

Recent research has demonstrated significant progress in the realization of optical computing hardware. Notable achievements include optical matrix multiplication chips capable of performing linear algebra operations at extremely high speeds, photonic convolution processors designed for real-time image processing, and early demonstrations of transformer acceleration using optical hardware (Fu et al., 2024). These developments indicate that photonic computing is transitioning from theoretical exploration toward practical AI acceleration platforms.

Photonic Neural Network Architectures in Literature

A growing body of literature has explored different architectures for implementing neural networks using photonic integrated circuits. Among these, three dominant approaches have emerged: Mach–Zehnder interferometer meshes, microring resonator networks, and diffractive optical neural networks.

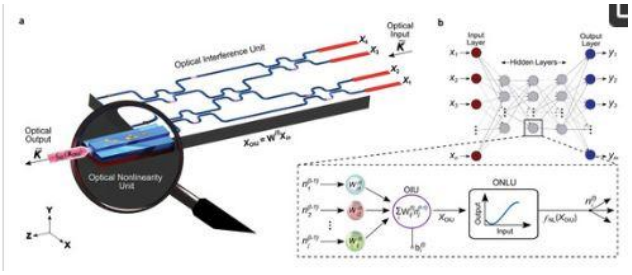

Figure 4. Illustration of a fully integrated photonic neural network based on MXene material with an optical nonlinearity unit (Silva et al., 2025).

Mach–Zehnder Interferometer (MZI) Meshes

Mach–Zehnder interferometer (MZI) meshes represent one of the most widely studied and implemented architectures for photonic neural networks. Their popularity stems from their programmability, scalability, and compatibility with silicon photonics fabrication processes (Bandyopadhyay et al., 2023). MZI meshes operate by manipulating the phase of optical signals and exploiting interference to perform matrix transformations. Through careful tuning of phase shifters, these meshes can implement arbitrary unitary matrices, which form the mathematical foundation of neural network linear layers.

Numerous studies have demonstrated the ability of MZI-based photonic processors to perform deep neural network inference with high speed and low energy consumption. The reconfigurability of these meshes makes them

particularly attractive for general-purpose optical AI accelerators (Brückerhoff-Plückelmann et al., 2025). However, the literature also highlights significant challenges, including high power consumption associated with thermal tuning and sensitivity to fabrication imperfections and environmental fluctuations. Addressing these limitations remains an active area of research.

Microring Resonator Networks

Microring resonator networks offer an alternative architecture that relies on wavelength-selective filtering and resonance effects to perform computation. These devices can exploit wavelength-division multiplexing (WDM) to process multiple data channels simultaneously, enabling extremely high computational throughput (Xia et al., 2014). Their compact footprint and energy efficiency make them attractive for dense integration and large-scale photonic processors.

Despite these advantages, microring-based systems face notable challenges. Their performance is highly sensitive to temperature variations and fabrication tolerances, which can lead to resonance shifts and crosstalk between wavelength channels. As a result, maintaining stability and calibration in large-scale microring networks remains a significant research challenge.

Diffractive Optical Neural Networks

Diffractive optical neural networks (D^2NNs) represent a fundamentally different approach that utilizes passive diffractive layers to perform computation in free space or integrated platforms (Han et al., 2025). In these systems, carefully designed diffractive elements manipulate the propagation of light to perform neural network operations without requiring active power consumption. This enables ultra-fast and energy-efficient computation, making D^2NNs attractive for applications requiring real-time inference.

However, the passive nature of diffractive networks limits their programmability and adaptability. Once fabricated, their functionality is typically fixed, making them better suited for specific inference tasks rather than general-purpose computing (Liu et al., 2022). Current research therefore focuses on improving reconfigurability and integrating diffractive approaches with programmable photonic platforms.

Materials for Photonic AI

The performance, scalability, and practicality of photonic neural networks are strongly influenced by the choice of material platforms used to fabricate photonic integrated circuits. Over the past decade, significant research has focused on identifying materials that provide low optical loss, efficient modulation, strong nonlinear effects, and compatibility with large-scale semiconductor manufacturing. The literature highlights several key platforms that are currently shaping the development of photonic AI hardware (Zhou et al., 2023).

Silicon photonics remains the most dominant and mature platform due to its compatibility with established CMOS fabrication processes. This compatibility enables high integration density and cost-effective mass production, allowing thousands of photonic components to be fabricated on a single chip. Silicon waveguides exhibit strong optical confinement, which enables compact device footprints and dense integration of interferometers, modulators, and detectors (Bogaerts & Selvaraja, 2014). However, silicon suffers from weak intrinsic optical nonlinearity and lacks efficient light emission, which necessitates hybrid integration with other materials for lasers and nonlinear functions. Silicon nitride has emerged as an important complementary platform, particularly for applications requiring ultra-low optical propagation loss.

Compared with silicon, silicon nitride waveguides exhibit significantly reduced scattering losses, making them suitable for large-scale interferometer meshes and long optical delay lines. This property is particularly valuable in photonic neural networks where signal integrity must be preserved across many interconnected components. Although silicon nitride devices typically have a larger footprint due to weaker optical confinement, their stability and low loss make them attractive for high-precision photonic computing systems (Xiang et al., 2022). Indium phosphide plays a critical role in enabling active photonic components, especially integrated laser sources. Unlike silicon, indium phosphide can efficiently generate light, making it essential for fully integrated optical systems. Hybrid integration of indium phosphide lasers onto silicon photonic platforms has become a major research direction, enabling compact and energy-efficient light sources for optical AI accelerators. Despite its advantages, indium phosphide fabrication is more complex and expensive, which presents challenges for large-scale manufacturing (Zhou, 2025).

Lithium niobate has recently gained renewed attention due to advances in thin-film lithium niobate technology. This material exhibits a strong electro-optic effect, allowing high-speed and low-power optical modulation. Such capabilities are particularly valuable for encoding neural

network inputs and dynamically tuning photonic circuits. The integration of lithium niobate with silicon photonics has therefore become an important research focus for high-performance optical AI hardware.

Graphene and other two-dimensional materials represent an emerging class of materials for photonic neural networks. These materials exhibit exceptional optical nonlinearities, ultrafast carrier dynamics, and broadband operation. Their unique properties make them promising candidates for implementing optical activation functions and nonlinear processing elements, which are essential for fully optical neural network operation. While research in this area is still developing, the integration of graphene and related materials is expected to play a significant role in future photonic AI systems.

Table 1: Comparison of Key Material Platforms for Photonic Neural Networks and Optical AI Accelerators

Platform	Key Contribution
Silicon photonics	High integration density
Silicon nitride	Ultra-low optical loss
Indium phosphide	Integrated lasers
Lithium niobate	High-speed modulation
Graphene	Optical nonlinearities

Optical AI Accelerators

Recent advances in photonic integrated circuits have accelerated the development of optical AI hardware designed specifically to address the computational demands of modern deep learning systems. Researchers have increasingly focused on building specialized photonic accelerators capable of executing core neural network operations particularly matrix multiplication and convolution at extremely high speed and low energy consumption (Tsakyridis et al., 2024). These efforts have led to the emergence of several key classes of optical AI accelerators that demonstrate the practical feasibility of photonic computing for real-world applications. One of the most notable developments is the introduction of optical tensor cores, which are photonic equivalents of the tensor processing units used in modern GPUs and AI accelerators. Optical tensor cores exploit the natural ability of light to perform interference-based linear algebra operations, enabling massively parallel matrix multiplications using wavelength-division multiplexing and interferometric networks (Zhang et al., 2025). By performing multiply–accumulate operations in the optical domain, these systems significantly reduce energy consumption and latency compared to electronic processors. Experimental demonstrations have shown that optical tensor cores can achieve orders-of-magnitude improvements in computational throughput, positioning them as a promising solution for large-scale neural network inference. In parallel, photonic convolution engines have been developed to accelerate convolutional neural networks (CNNs), which are widely used in computer vision, medical imaging, and autonomous systems. Optical convolution can be implemented using Fourier optics, integrated interferometers, or wavelength multiplexing techniques, allowing real-time image processing with extremely high bandwidth. These systems are particularly attractive for edge computing and real-time perception applications, where low latency and high energy efficiency are essential.

Another important area of progress is the use of optical interconnects in data centers and high-performance computing environments. As AI workloads continue to grow, the energy cost of data movement between processors and memory has become a major bottleneck. Optical interconnects provide high-bandwidth, low-latency communication links that can significantly reduce power consumption and improve system performance. Integrating photonic interconnects with AI accelerators enables efficient data transfer between computing nodes, supporting the development of large-scale, distributed AI systems.

Collectively, these advancements demonstrate rapid progress toward the commercialization of photonic AI technologies. The transition from laboratory prototypes to industry-driven solutions indicates that optical AI accelerators are moving closer to practical deployment in data centers, autonomous systems, and edge computing platforms.

III. Methodology

Review Design

This study adopts a systematic narrative review methodology to provide a structured and comprehensive synthesis of research on photonic neural networks and optical AI accelerators. Unlike a purely descriptive literature survey, this approach combines systematic search, thematic categorization, and comparative analysis to ensure that the reviewed works are critically evaluated and organized in a coherent framework. The methodology was designed to capture the rapid evolution of this interdisciplinary field, which spans photonics, artificial

intelligence, materials science, and semiconductor engineering.

The review process followed four major stages. First, relevant research themes and subdomains were identified through an initial scoping survey of the field. Second, the collected literature was classified according to architectural approaches and material platforms. Third, a comparative evaluation of performance metrics and technological readiness was conducted. Finally, key challenges, research gaps, and future directions were synthesized from the analyzed studies.

Review Workflow

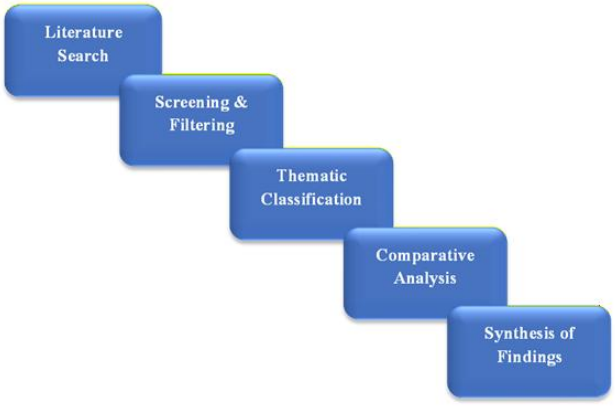

Figure 1: Review workflow adopted in the research

Search and Selection Scope

To ensure relevance and currency, the review primarily focuses on peer-reviewed journal articles, conference proceedings, and high-impact reports published within the last decade. This time frame captures the period during which photonic AI research transitioned from conceptual demonstrations to scalable integrated photonic implementations.

The literature selection concentrated on four major research domains:

- Photonic neural network architectures
- Optical AI accelerators and computing hardware
- Silicon photonics and heterogeneous integration
- Neuromorphic and hybrid photonic–electronic systems

Priority was given to studies that reported experimental demonstrations, performance evaluations, or comprehensive theoretical frameworks. This approach ensured that the review reflects both foundational research and recent technological advancements.

Screening and Classification Strategy

The selected literature was systematically categorized based on technological focus and application domain. The classification process enabled the identification of major research trends and facilitated structured comparison across different approaches. The primary classification categories included:

Photonic neural network architectures
Materials and fabrication platforms
Optical AI accelerator designs
Hybrid photonic–electronic systems

Analytical Framework

To ensure consistent and objective evaluation, the reviewed literature was analyzed using four key criteria that reflect the practical requirements of next-generation AI hardware.

Table 2: Evaluation Criteria for Systematic Analysis of Photonic Neural Networks and Optical AI Accelerators

Evaluation Criterion	Purpose of Analysis
Architecture	Assess scalability, programmability, and computational capability
Material Platforms	Evaluate performance, fabrication maturity, and integration potential
Energy Efficiency	Compare optical and electronic computing efficiency
Practical Challenges	Determine readiness for commercialization and large-scale deployment

Analysis

This section presents a technical and engineering-oriented analysis of photonic AI hardware from the perspective of electronics and communication engineering. The discussion focuses on quantitative performance comparison, architectural trade-offs, and material-level implications for scalable system design.

Comparative Performance Analysis: Electronic vs Photonic Computing

Modern electronic AI accelerators are primarily constrained by the von Neumann bottleneck, interconnect

bandwidth limitations, and thermal power density. Photonic computing, by contrast, performs linear algebra operations directly in the optical domain, thereby reducing data movement and enabling ultra-high bandwidth signal processing.

Table 3: Comparative Performance Analysis: Electronic vs Photonic Computing

Performance Metric	Electronic AI Accelerators (GPU/TPU)	Photonic AI Accelerators (PNNs/PICs)
Energy per MAC operation	picojoule (pJ) to nanojoule (nJ)	femtojoule (fJ)
Operating bandwidth	GHz range	THz range
Latency	nanoseconds (ns) to microseconds (µs)	picoseconds (ps)
Data transfer mechanism	Electrical interconnects	Optical waveguides
Parallelism capability	Limited (clock and memory bound)	Massive (WDM, spatial multiplexing)
Heat dissipation	High due to resistive losses	Very low (minimal Joule heating)
Computation principle	Digital switching and arithmetic logic	Optical interference and propagation
Scalability bottleneck	Memory bandwidth & power density	Fabrication variability & optical loss

Energy Efficiency Analysis

Energy consumption per multiply–accumulate (MAC) operation is a key performance metric for AI hardware. Electronic processors typically require picojoule–nanojoule energy per MAC due to resistive losses and repeated data movement between memory and compute units. Photonic systems perform these operations using passive interference and propagation, reducing energy consumption to the femtojoule scale.

From a communication engineering perspective, this improvement can be interpreted as a reduction in switching energy and interconnect power, which are dominant contributors to system-level energy consumption in electronic processors.

Bandwidth and Throughput Analysis

Electronic interconnects operate within GHz frequency ranges due to parasitic capacitance and resistance. Photonic interconnects, however, operate in the THz optical frequency regime and support wavelength-division multiplexing (WDM), enabling multiple data streams to be transmitted simultaneously over a single waveguide.

This results in:

- Higher channel capacity
- Reduced congestion in interconnect networks
- Increased computational throughput
- Latency and Signal Propagation

Signal propagation in electronic circuits is limited by RC delays and clock synchronization constraints. Optical propagation occurs at the speed of light with negligible resistive delay, enabling picosecond-scale latency. This feature is particularly valuable for real-time AI inference and edge computing applications.

Parallelism and Communication Scalability

Photonic systems inherently support massive parallelism via multiplexing in wavelength, phase, and spatial domains. This property aligns closely with communication engineering principles of multi-carrier transmission and parallel channel utilization, making photonics well suited for large-scale neural network deployment.

Engineering-Insight: Photonic computing effectively transforms neural network computation into a high-capacity optical communication problem, where matrix multiplication becomes a signal interference process. Photonic neural network architectures operate within a fundamentally constrained design space defined by competing requirements of programmability, energy efficiency, scalability, and system stability. Unlike electronic neural networks, where abstraction layers mask most physical device limitations, photonic systems directly encode computation into optical wave propagation. As a result, architectural performance is strongly influenced by physical phenomena such as interference, phase noise, thermal drift, and fabrication variability. This makes system-level trade-off analysis essential for understanding the practical feasibility of different photonic computing approaches.

From an electronics and communication engineering perspective, photonic neural network architectures can be interpreted as different implementations of optical signal processing systems, where neural computation is mapped

onto controllable transformations of optical fields. Each architecture therefore represents a different balance between flexibility, efficiency, and robustness under real-world operating conditions.

Mach–Zehnder Interferometer (MZI) Mesh Networks

Mach–Zehnder Interferometer (MZI) mesh networks are currently the most widely studied and experimentally demonstrated architecture for photonic neural computing. These systems consist of cascaded interferometric units arranged in a programmable mesh topology, where each unit performs controlled splitting, phase shifting, and recombination of optical signals. By tuning the phase shifters within each interferometer, the system can implement arbitrary linear transformations, making it highly suitable for representing the weight matrices in neural networks. The primary strength of MZI-based architectures lies in their high degree of programmability. They can be dynamically reconfigured to implement different neural network layers and support a wide range of computational tasks, including matrix multiplication and signal transformation. In addition, their compatibility with silicon photonics fabrication processes enables relatively high integration density and scalability compared to earlier free-space optical systems.

However, this flexibility comes at a significant engineering cost. MZI meshes rely heavily on thermo-optic phase shifters, which introduce substantial static power consumption even when the system is not actively switching. As the size of the mesh increases, thermal crosstalk and power dissipation become increasingly difficult to manage. Furthermore, optical losses accumulate as signals propagate through multiple interferometric stages, which limits the achievable depth of the network and reduces overall signal-to-noise ratio. Another critical limitation is calibration complexity; large-scale MZI networks require continuous tuning and compensation to maintain phase accuracy due to fabrication imperfections and environmental variations.

From a system engineering standpoint, MZI mesh networks represent a flexible but power-intensive programmable optical computing architecture, where scalability is primarily constrained by thermal management and control overhead rather than computational capability.

Microring Resonator Networks

Microring resonator-based architectures offer an alternative approach that prioritizes compactness, energy efficiency, and high-density integration. These systems utilize resonance phenomena in optical microcavities to selectively filter and manipulate different wavelength channels. By encoding neural network weights into resonance characteristics such as coupling coefficients, phase shifts, and resonance offsets, microring networks can perform parallel computations using wavelength-division multiplexing (WDM). A major advantage of microring architectures is their ability to support extremely dense integration, enabling a large number of computational elements to be packed into a small chip area. Additionally, their inherent compatibility with WDM allows multiple data streams to be processed simultaneously, significantly increasing throughput and improving computational efficiency. These characteristics make microring networks particularly attractive for large-scale photonic accelerators targeting data-intensive AI workloads. Despite these advantages, microring resonators exhibit significant sensitivity to environmental and fabrication variations. Even small changes in temperature or structural dimensions can shift resonance frequencies, leading to performance degradation and signal misalignment. This temperature dependence necessitates active stabilization or feedback control systems, which introduce additional system complexity. Furthermore, when multiple resonators are densely integrated, spectral overlap and channel crosstalk can occur, reducing signal fidelity and limiting network scalability.

In engineering terms, microring resonator networks can be characterized as highly compact but stability-sensitive wavelength-domain processing systems, where performance is strongly dependent on precise physical control and environmental compensation mechanisms.

Diffractive Optical Neural Networks (D^2NNs)

Diffractive optical neural networks represent a fundamentally different class of photonic computing architecture that relies on passive optical propagation through engineered diffractive layers. In these systems, computation is performed through spatial modulation of light intensity and phase as optical waves propagate through structured media. Each diffractive layer is designed to encode a fixed transformation, allowing the system to implement neural inference without active electronic control or external power input during operation.

The most significant advantage of diffractive systems is their extremely low energy consumption. Since computation is achieved through passive propagation of light, no dynamic electrical power is required during inference. Additionally, these systems operate at the physical speed of light, enabling ultra-fast signal processing with minimal latency. These properties make

diffractive networks highly attractive for ultra-low-power and high-speed inference applications.

However, this architecture also introduces severe limitations. Once a diffractive structure is fabricated, its computational behavior is fixed, meaning that it lacks reconfigurability and cannot adapt to new tasks or updated datasets. This restricts its applicability to specialized, single-purpose inference problems. Furthermore, diffractive systems are highly sensitive to optical alignment and fabrication precision, which can significantly affect performance in practical implementations.

From a communication systems perspective, diffractive neural networks function as static optical transfer systems, analogous to fixed filters in signal processing, offering high efficiency but extremely limited adaptability.

Overall System System-Level Trade-Off Perspective

When analyzed collectively, these photonic neural network architectures reveal a fundamental and persistent trade-off between programmability, efficiency, and scalability. MZI mesh networks provide the highest flexibility but suffer from significant power and thermal constraints. Microring resonator systems offer superior integration density and energy efficiency but are limited by environmental sensitivity and stability issues. Diffractive optical networks achieve unmatched energy efficiency and speed but lack the adaptability required for general-purpose computing. This trade-off landscape indicates that no single photonic architecture currently satisfies all the requirements for large-scale, general-purpose neural computation. Consequently, the future of photonic AI is likely to depend on hybrid and heterogeneous system architectures, where different photonic approaches are combined and integrated with electronic control systems to balance performance, efficiency, and adaptability.

Material Platform Analysis for Photonic AI Hardware

Material selection is a fundamental determinant of performance in photonic neural networks and optical AI accelerators. Unlike electronic systems where silicon dominates almost all computing layers, photonic systems require a heterogeneous material ecosystem to simultaneously support light generation, modulation, propagation, and detection. Each material platform contributes distinct optical, electrical, and fabrication characteristics that directly influence device efficiency, scalability, and manufacturability.

From a systems and communication engineering perspective, material selection defines the physical layer constraints of photonic AI hardware, including loss budget, modulation speed, integration density, and thermal stability. Therefore, understanding material trade-offs is essential for designing scalable photonic computing architectures.

Silicon Photonics

Silicon photonics serves as the foundational platform for large-scale photonic integration due to its compatibility with mature CMOS fabrication infrastructure. This compatibility enables the reuse of existing semiconductor manufacturing ecosystems, significantly reducing cost and accelerating industrial adoption. Silicon waveguides also exhibit a high refractive index contrast, which allows strong optical confinement and supports dense integration of photonic components such as waveguides, modulators, and interferometers on a single chip.

From an engineering perspective, these characteristics make silicon photonics highly suitable for implementing large-scale photonic neural networks, particularly MZI mesh architectures that require dense routing and compact integration.

However, silicon also presents fundamental physical limitations. Its weak intrinsic optical nonlinearity restricts its ability to support all-optical computation and nonlinear activation functions required in neural networks. Additionally, silicon is inefficient as a light emitter due to its indirect bandgap, necessitating hybrid integration with external laser sources or III–V materials.

Overall, silicon photonics acts as the structural backbone of photonic AI systems, providing scalability and integration capability but requiring complementary materials for full functionality.

Silicon Nitride (SiN)

Silicon nitride has emerged as a critical material platform for applications requiring ultra-low optical loss. Its ability to support long-distance optical propagation with minimal attenuation makes it particularly suitable for large-scale interferometric networks and delay-line-based architectures. The primary engineering advantage of silicon nitride lies in its extremely low propagation loss, which enables high-fidelity signal transmission across complex photonic circuits. This property is especially important in deep photonic neural networks where signals traverse multiple cascaded stages.

However, silicon nitride exhibits weaker optical confinement compared to silicon, resulting in larger device footprints. This introduces a fundamental trade-off between optical performance and integration density.

From a system design perspective, silicon nitride is best characterized as a low-loss propagation medium optimized for signal integrity rather than compactness, making it highly suitable for precision photonic computing applications.

Indium Phosphide (InP)
Indium phosphide is a key material for active photonic components, particularly light sources and optical amplifiers. Unlike silicon, indium phosphide possesses a direct bandgap, enabling efficient on-chip laser generation. This capability is essential for fully integrated photonic systems, where external laser coupling introduces complexity and power inefficiency. In photonic AI hardware, indium phosphide is primarily used to provide optical gain, signal amplification, and integrated laser functionality, which are critical for maintaining signal strength across large-scale photonic networks. However, the fabrication of indium phosphide devices is significantly more complex and expensive compared to silicon-based platforms. Integration with CMOS processes is also more challenging, limiting its scalability for large-volume production. From an engineering standpoint, indium phosphide represents a high-performance active photonic material, offering essential functionality at the cost of manufacturing complexity and reduced compatibility with large-scale silicon-based integration.

Lithium Niobate ($LiNbO_3$)
Lithium niobate, particularly in its thin-film form, has gained increasing attention due to its exceptionally strong electro-optic properties. This enables ultra-fast and highly efficient optical modulation, which is critical for encoding and controlling optical signals in neural network operations. The primary advantage of lithium niobate lies in its ability to achieve high-speed modulation with low optical loss, making it suitable for dynamic reconfiguration of photonic circuits. Recent advances in thin-film lithium niobate integration have further improved its compatibility with on-chip photonic systems, enabling more compact device architectures. However, integrating lithium niobate into large-scale photonic circuits remains technically challenging. Issues related to fabrication complexity, heterogeneous bonding, and device scaling limit its widespread adoption in current photonic AI systems.

From a system-level perspective, lithium niobate can be viewed as a high-speed modulation layer material, providing critical dynamic control functionality but requiring complex integration strategies

System-Level Insight
At the system level, no single material platform is sufficient to fully realize scalable and high-performance photonic neural networks. Instead, photonic AI hardware depends on a heterogeneous integration strategy, where different materials are combined to leverage their complementary strengths.

In this architecture:

- Silicon provides integration density and system scalability
- Silicon nitride ensures low-loss signal propagation
- Indium phosphide enables light generation and amplification
- Lithium niobate provides high-speed modulation capabilities

This multi-material approach introduces a new design paradigm in which photonic AI systems are no longer constrained by a single substrate but are instead engineered as layered heterogeneous systems optimized for specific functional roles.

From an electronics and communication engineering standpoint, this represents a shift toward system-on-photonics architectures, where performance is determined not only by circuit design but also by material-level co-optimization of optical, electrical, and thermal properties.

IV. RESULTS AND DISCUSSION

Synthesis of Key Findings
The analytical comparison of photonic and electronic computing, architectures, and material platforms reveals several consistent trends that define the current state of Photonic AI.

Provides Unmatched Efficiency for Linear Algebra Operations

The quantitative comparison in Section 4 demonstrates orders-of-magnitude improvements in:

- Energy per MAC: pJ–nJ → fJ
- Bandwidth: GHz → THz
- Latency: ns–µs → ps

These advantages arise from the physics of optical interference. Matrix multiplication the dominant workload in deep neural networks can be executed passively through light propagation in interferometric meshes.

This confirms that photonics is intrinsically suited to the most computationally intensive portion of AI workloads.

Engineering implication: Photonic processors are particularly advantageous for:

- Large-scale inference
- Transformer models
- Edge AI requiring ultra-low power

Hybrid Photonic–Electronic Systems Are the Most Viable Near-Term Architecture

Table 5: The architecture comparison highlighted key trade-offs

Architecture	Strength	Limitation
MZI meshes	Programmability	Thermal power overhead
Microring resonators	Compact footprint	Temperature sensitivity
Diffractive optics	Passive operation	Lack of reconfigurability

No single architecture currently satisfies scalability, programmability, and robustness simultaneously.

This strongly supports the conclusion that fully optical neural networks remain impractical in the short term.

Instead, hybrid systems leverage complementary strengths: Photonic domain → Linear algebra acceleration Electronic domain → Nonlinear activation, memory, control

Engineering implication: Future AI accelerators will likely resemble heterogeneous chiplets integrating
CMOS electronics
Silicon photonics
Optical interconnects

Material Innovation is the Central Enabler of Scalable Photonic AI
The materials analysis reveals a critical pattern:

Table 5: Functional Roles of Key Material Platforms in Photonic Neural Networks and Optical AI Systems

Material	System Role
Silicon	Integration backbone
Silicon nitride	Low-loss interconnects
Indium phosphide	On-chip light sources
Lithium niobate	High-speed modulators

Each material solves a different bottleneck, indicating that no single platform can support a complete photonic AI stack.

This confirms the emergence of heterogeneous photonic integration as the dominant research direction.

Discussion of Major Technical Challenges
Despite strong performance advantages, several engineering barriers prevent large-scale deployment.

Scalability Constraints in Large Photonic Networks
As photonic neural networks scale, performance degradation becomes significant due to:
Optical loss accumulation across cascaded components
Fabrication variability affecting phase accuracy
Thermal tuning power in MZI meshes
These factors limit current systems to small and medium-scale demonstrations.

Future Research Directions and Outlook
The trajectory of research suggests a clear evolution toward large-scale deployment.

Near-Term (0–5 years)
Photonic chiplets integrated with GPUs and AI accelerators
Optical interconnects in hyperscale data centres
Energy-efficient AI inference engines

Mid-Term (5–10 years)
- Wafer-scale photonic neural processors
- Photonic memory and non-volatile optical weights
- Standardized photonic AI design frameworks

Long-Term Vision

The long-term evolution of photonic artificial intelligence is expected to be driven by the convergence of heterogeneous material platforms, hybrid photonic–electronic architectures, and advanced scalable packaging technologies.

As these components mature and become more tightly integrated, they are likely to enable a new class of computing systems with fundamentally enhanced performance characteristics. In particular, future developments may lead to the realization of optical AI supercomputers capable of performing extremely large-scale neural computations with significantly reduced energy consumption and latency compared to conventional electronic systems. In addition, ultra-low-power edge AI devices are expected to emerge, enabling real-time intelligent processing in resource-constrained environments such as IoT systems, autonomous platforms, and wearable technologies. Furthermore, the deployment of data-centre-scale photonic acceleration platforms is anticipated, where optical interconnects and photonic processors work in synergy to overcome current bottlenecks in bandwidth, energy efficiency, and computational throughput.

Overall Interpretation

The combined evidence from architectural, material, and performance analyses indicates that photonic artificial intelligence is currently transitioning from laboratory-scale demonstrations toward early stages of commercialization. Significant progress in photonic neural network architectures, material integration, and optical computing performance demonstrates that the field is no longer purely theoretical but is beginning to show practical engineering viability. However, despite these advancements, the widespread deployment of photonic AI systems remains contingent upon overcoming several critical challenges. These include improving system scalability to support large-scale neural networks, developing efficient and practical training methodologies compatible with photonic hardware, and ensuring high levels of computational precision, stability, and reliability under real-world operating conditions. Addressing these fundamental challenges will ultimately determine the rate at which photonic AI evolves from an emerging research domain into a mature and mainstream computing paradigm.

V. Conclusion

This review has examined the rapidly developing field of photonic neural networks and optical AI accelerators, focusing on their architectures, material platforms, performance advantages, and key engineering limitations. The findings indicate that photonic computing offers a fundamentally more energy-efficient and faster alternative to electronic AI hardware, particularly for matrix-based operations that dominate neural network workloads. The comparative analysis shows that architectures such as MZI meshes, resonators and diffractive systems each provide valuable capabilities, but none fully satisfies scalability, programmability, and stability requirements simultaneously. Similarly, while silicon photonics remains the dominant integration platform, heterogeneous material systems are necessary to achieve full device functionality. However, several challenges still limit practical deployment, including scalability constraints, noise sensitivity, and the absence of efficient optical training methods. These issues currently restrict photonic AI systems to experimental and small-scale implementations. Overall, the study concludes that photonic AI is transitioning from research to early-stage application, with hybrid photonic–electronic systems representing the most realistic near-term solution. Continued advances in materials, architectures, and integration strategies will be critical to enabling large-scale, energy-efficient photonic computing systems in the future.

REFERENCES

1. Bandyopadhyay, S., Sludds, A., Krastanov, S., Hamerly, R., Harris, N., Bunandar, D., Streshinsky, M., Hochberg, M., & Englund, D. (2023). A Photonic Deep Neural Network Processor on a Single Chip with Optically Accelerated Training. https://doi.org/10.1364/CLEO_SI.2023.SM2P.2
2. Bogaerts, W., & Selvaraja, S. K. (2014). Silicon-on-insulator (SOI) technology for photonic integrated circuits (PICs) (pp. 395–434). https://doi.org/10.1533/9780857099259.2.395
3. Brückerhoff-Plückelmann, F., Meyer, L., Dijkstra, J., Tebeck, S., McRae, L., Bahr, N., Steinmeyer, D., Koptyaev, S., Bernasconi, J., Pavlov, N., Karpov, M., Jost, J., & Pernice, W. (2025). Deep Neural Network Inference on an Integrated, Reconfigurable Photonic Tensor Processor. https://doi.org/10.21203/rs.3.rs-7859321/v1
4. Charalampous, G., Chen, R., On, M., Nasirov, A., Cheng, C.-Y., AbdelGhany, M., Majumdar, A., Wang, J., Black, J., Kubendran, R., Oskay, C., Bai, Z., Palermo, S., Papp, S., & Yoo, S. (2025). Mixed Precision Photonic Computing with 3D Electronic-

Photonic Integrated Circuits. https://doi.org/10.48550/arXiv.2508.03063

5. Fu, T., Zhang, J., Sun, R., Huang, Y., Xu, W., Yang, S., Zhu, Z., & Chen, H. (2024). Optical neural networks: progress and challenges. Light, Science & Applications, 13, 263. https://doi.org/10.1038/s41377-024-01590-3
6. Han, T., Sun, J., & Yang, X. (2025). Advancements in Optical Diffraction Neural Networks. Photonics, 12, 1187. https://doi.org/10.3390/photonics12121187
7. Liu, C., Ma, Q., Luo, Z., Hong, Q., Xiao, Q., Zhang, H., Miao, L., Yu, W.-M., Cheng, Q., Li, L., & Cui, T. (2022). A programmable diffractive deep neural network based on a digital-coding metasurface array. Nature Electronics, 5, 1–10. https://doi.org/10.1038/s41928-022-00719-9
8. McMahon, P. (2023). The physics of optical computing. https://doi.org/10.48550/arXiv.2308.00088
9. Perera, S., Perera, H., & Chamika, D. (2024). Architecture of AI Accelerators. https://doi.org/10.5281/zenodo.17409544
10. Sheng, Y., Zhang, C., Zhu, Z., Xu, H., Wen, J., Wang, R., Yang, J., & Bu, S. (2026). Power for AI Data Centers: Energy Demand, Grid Impacts, Challenges and Perspectives. Energies, 19, 722. https://doi.org/10.3390/en19030722
11. Silva, L. C. B., Marciano, P. R. N., Pontes, M. J., Monteiro, M. E., André, P. S. B., & Segatto, M. E. V. (2025). Integrated Photonic Neural Networks for Equalizing Optical Communication Signals : A Review. 1–25.
12. Tsakyridis, A., Moralis-Pegios, M., Giamougiannis, G., Kirtas, M., Passalis, N., Tefas, A., & Pleros, N. (2024). Photonic neural networks and optics-informed deep learning fundamentals. APL Photonics, 9. https://doi.org/10.1063/5.0169810
13. Xia, J., Bianco, A., Bonetto, E., & Gaudino, R. (2014). On the design of microring resonator devices for switching applications in flexible-grid networks. In 2014 IEEE International Conference on Communications, ICC 2014. https://doi.org/10.1109/ICC.2014.6883842
14. Xiang, C., Jin, W., & Bowers, J. (2022). Silicon nitride passive and active photonic integrated circuits: trends and prospects. Photonics Research, 10. https://doi.org/10.1364/PRJ.452936
15. Xu, B., Banerjee, A., & Gupta, S. (2025). Hardware Acceleration for Neural Networks: A Comprehensive Survey. https://doi.org/10.48550/arXiv.2512.23914
16. Zhang, Y., Liu, X., Yang, C., Yan, H., Fu, T., Wang, K., Su, Y., Sun, Z., & Guo, X. (2025). Direct tensor processing with coherent light. Nature Photonics, 20, 102–108. https://doi.org/10.1038/s41566-025-01799-7
17. Zhou, Z. (2025). Recent development in photonic integration. https://doi.org/10.1117/12.3060174
18. Zhou, Z., Ou, X., Fang, Y.-T., Alkhazraji, E., Xu, R., Wan, Y., & Bowers, J. (2023). Prospects and applications of on-chip lasers. ELight, 3, 1. https://doi.org/10.1186/s43593-022-00027-x