International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 8.141
Peer Reviewed Bi-Monthly
Vol. 6 No. 38 September - October - 2025

Energy-efficient AI Hardware using Silicon Oxynitride Integrated Photonics

Archana Chaudhari
Assistant Professor,
Department of Computer Science, Dr. D. Y. Patil Science & Computer Science College, Akurdi, Pune-411044
Corresponding Author – Archana Chaudhari
DOI - 10.5281/zenodo.17309826

Abstract:
Artificial Intelligence (AI) workloads demand unprecedented computational power and energy efficiency. Photonic hardware has emerged as a promising alternative to conventional electronics due to its ultra-fast signal propagation, low latency, and reduced energy consumption. In this work, we explore Silicon Oxynitride (SiON) as a photonic platform for AI hardware, highlighting its low propagation loss, CMOS compatibility, and scalability. We discuss SiON waveguide-based architectures for optical neural networks, emphasizing their potential to achieve favorable energy–latency trade-offs. Our findings suggest that SiON integrated photonics offers a viable path towards next-generation energy-efficient AI accelerators.

Keywords: Silicon Oxynitride, Integrated Photonics, Optical Neural Networks, Artificial Intelligence Hardware, Energy Efficiency.

Introduction:

The exponential growth of AI applications has created an urgent need for energy-efficient hardware accelerators. Conventional CMOS-based systems face bottlenecks in terms of power consumption and latency. Integrated photonics offers a disruptive alternative by leveraging light for computation and communication. Among photonic platforms, Silicon Oxynitride (SiON) has gained attention due to its low-loss propagation, broad transparency window, and CMOS-compatible fabrication.

Evolution of AI Hardware (CMOS → GPUs → TPUs → Photonics):

1. CMOS-based CPUs: Early AI algorithms (before ~2010) ran primarily on **general-purpose CPUs**. CMOS scaling (Moore's Law) enabled increases in clock speed and transistor density. But CPUs are **serial architectures, which is** inefficient for AI workloads that require massive parallelism. This CMOS based CPU has limitation of energy bottleneck and slow training for large-scale neural networks.

2. GPUs (Graphics Processing Units): Around 2010, researchers began using GPUs for **deep learning** [5]. GPUs are massively parallel, handling thousands of threads simultaneously. They are capable of **training of deep convolutional neural networks (CNNs)** in days instead of months. But the main limitations are high power consumption (hundreds of watts per chip), memory bottlenecks, and scaling inefficiency for ever-larger AI models.

3. TPUs (Tensor Processing Units): Google introduced TPUs in 2016 as **ASICs optimized for AI workloads**. Specialized for **matrix multiplications** (core operation in neural networks). In this, we achieved **higher performance(watt) as** compared to GPUs for inference and training. But still were lagged in bounding by resistive-capacitive (RC) delay, heat dissipation, and energy costs of moving data (memory wall).

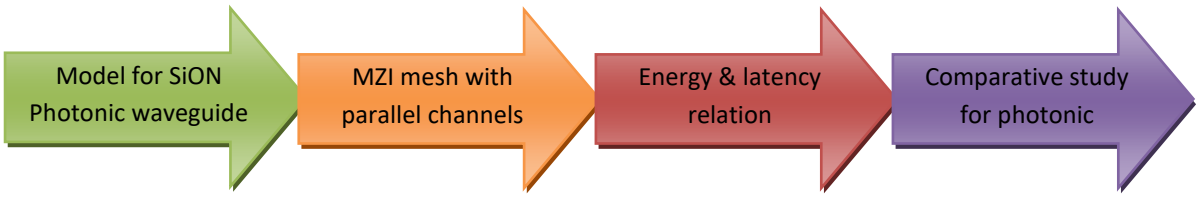
4. Integrated Photonics for AI: To overcome CMOS scaling limits, researchers started using **non-traditional hardware** like **Neuromorphic chips** (IBM TrueNorth, Intel Loihi) which are brain-inspired spiking networks. Also used the a**nalog in-memory computing** that compute inside memory arrays to reduce data transfer.

Apart from this if we use **Photonic hardware** which uses light for computation, promising **ultra-low latency** and **energy efficiency**. We can use the concept of ONN (Optical Neural Networks) for AI hardware, that perform **matrix multiplication via interference in waveguide meshes (e.g., MZIs)**.[7] This causes speed of light propagation with ultra fast inference, low energy per operation ((< 1 fJ/MAC in theory) and natural parallelism due to wavelength-division multiplexing. But the a**dvantages of Photonics over electronics are** Energy Efficiency: Optical interconnects reduce the energy cost of data movement, a major bottleneck in electronic accelerators. Low Latency: Photons travel at the speed of light with minimal delay, enabling ultra-fast inference. Scalability: Multiple wavelengths can carry parallel information streams on the same waveguide.

Among photonic materials, ***Silicon Oxynitride (SiON)*** *is promising because of Low propagation loss, Wide transparency window and CMOS-compatible fabrication.*

During **2017–2019**, integrated optical matrix multiplication were done using silicon and silicon nitride platforms. Later after **2020** Hybrid optical–electrical architectures for deep learning inference had been in use. Emergence of **programmable photonic chips** with hundreds of tunable elements are pointing towards scalable ONNs. But role of Silicon Oxynitride (SiON) in ONN are promisingly increasing. In ONN (Optical-Nano-Nitride) research, Silicon Oxynitride (SiON) (SiNxOy) is a crucial material due to its tunable optical and electrical properties, bridging the gap between silicon oxide and silicon nitride. As SiOn has a broad refractive index range for waveguides and Bragg gratings in integrated optics, a high-quality dielectric insulator for high-power electronics and MEMS by controlling its oxygen-to-nitrogen ratio, used for optoelectronic devices and surface passivation in electronics.

Methodology:



Why SiON ? : SiON is a balanced material between SiO_2, Si_3N_4. It has tunable refractive index (1.45–2.0) better for controllable confinement. **Low propagation loss** (<0.1 dB/cm achievable). **CMOS-compatible fabrication** using standard

LPCVD/PECVD processes. [3] Wide transparency window (visible (blue) to near-IR 2 µm). Thermally stable and less prone to nonlinear absorption than silicon. Most optical-AI chips use **Si** or **SiN.** SiON material mostly establish a **low-loss, CMOS-friendly platform** and characterize its **nonlinearities** which makes a solid base for SiON-for-AI contribution.

Why SiON photonic Waveguide? : Mostly **Waveguide Architectures should be Single-mode rib/strip type for** good confinement, low loss. SiON photonic waveguide provides the same. Along with this, if used as **Multimode waveguides** for mode-division multiplexing. For **Arrayed Waveguide Gratings (AWGs),** SiON is popular in telecom WDM components. **MZI meshes & reconfigurable circuits which are** suitable for optical neural network layers. Fig-1 shows the basic structure of SiON based photonic waveguide.

Why SiON Photonic waveguide for AI? : Advantages for AI Photonics includes low cumulative loss in deep meshes which is the key for large matrix multiplication circuits in ONNs. **Tunable index contrast which** balances footprint (not as bulky as SiN, not as nonlinear/unstable as Si). **Hybrid integration potential like** detectors, modulators, and lasers can be bonded. **Nonlinearities (Brillouin/Kerr) that** promises for optical nonlinear activation functions.

SiON-based Architectures for AI Hardware:

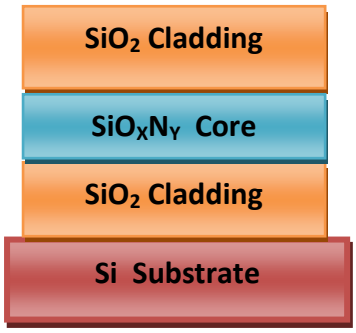


Fig1: Structure of SiON photonic waveguide

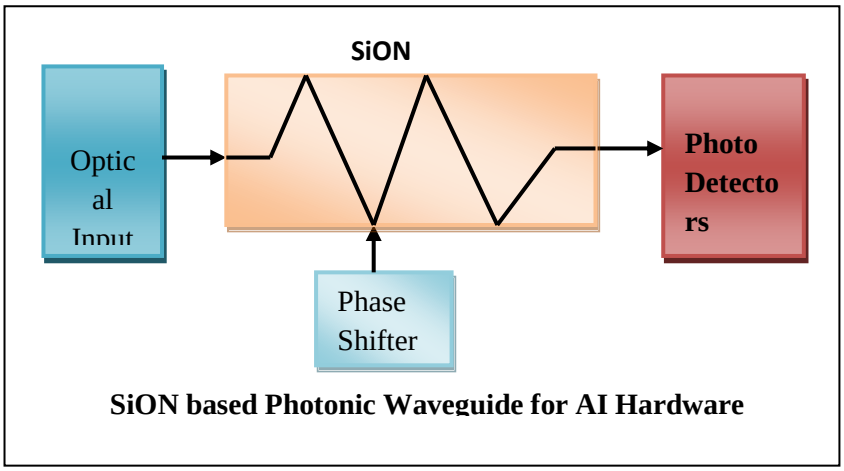


Fig2: Block diagram of SiON Photonic waveguide for AI

We reconfigure Mach–Zehnder interferometer (MZI) meshes to implement arbitrary unitary (or near-unitary) linear transforms, enabling matrix–vector multiplication (MVM) in a single optical pass. On SiON, MZIs are realized with low-loss waveguides and thermo/electro-optic phase shifters. To structure this waveguide we have to solve the decomposition like any unitary $U \in C^{N\times N}$ can be factorized into a sequence of 2×2 beam-splitter (MZI) elements and phase shifts (e.g., Reck or Clements decomposition), enabling linear layers of a neural network. SiON reduced cumulative insertion loss in deep meshes, enabling larger N before signal sinks below detector noise.

Simulation Model for energy-latency analysis:

We proposed a model **model** of SiON photonic AI hardware that uses MZI mesh with parallel channels. Here we are considering the **SiON photonic waveguide + MZI mesh** as a **matrix multiplication engine.**

Each multiplication involves the parameters like **Optical loss per MZI** (α _MZI),

Phase shifter energy E by relation E_{MZI} = CV^2

Waveguide Propagation delay by $\tau = \frac{L*n\_g}{C}$. [5]

Multiply-accumulate is given by $E_{MAC} = \frac{P_{laser}\tau_{wg}}{N_{parallel}}$

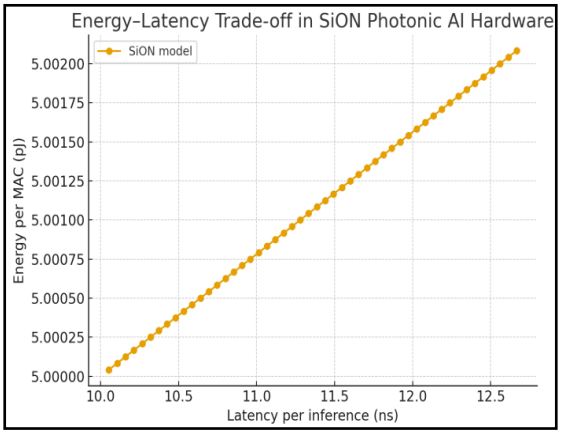


Fig3: energy–latency curve

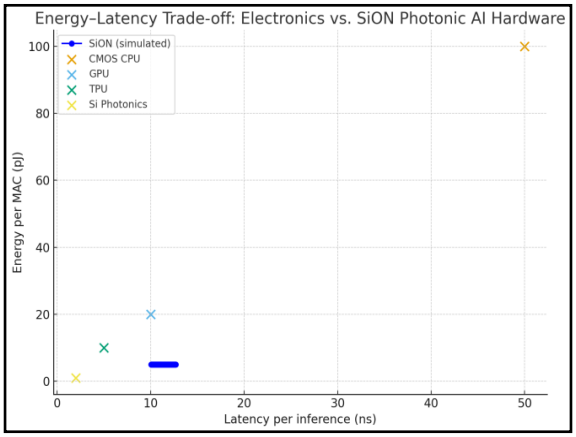


Fig 4: Comparative study

Discussion:

Fig3 curve represents the **energy–latency trade-off**. Longer waveguides, more latency in turn slower is the inference. But slightly more optical loss [8]. Phase-shifter energy dominates, giving ~5 pJ/MAC baseline.

Fig4 represents the **comparative study of energy–latency trade-off with SiON, CMOS CPU, GPU, TPU & Si.** CMOS CPU (~100 pJ, 50 ns), GPU (~20 pJ, 10 ns), TPU (~10 pJ, 5 ns), Si Photonics (~1 pJ, 2 ns) [4][1][6]. It clearly shows how **SiON can bridge the gap** between current electronics (high energy, lower latency) and future photonics (ultra-low energy, low latency). Further the **SiON curve trends towards better efficiency** as parallelism increases.

Challenges and Future Directions:

Still SiON is not as compact as Si, but more stable than SiN. SiON based ecosystem is still in developing mode as compared to Si/SiN with limited industrial-scale support. Active device integration like fast modulators, detectors is limited in case of SiON as compared to Si/SiN.

Conclusion:

We have outlined the potential of Silicon Oxynitride integrated photonics as a material platform for energy-efficient AI hardware. By enabling low-loss, scalable, and CMOS-compatible architectures, SiON waveguides represent a promising step towards next-generation optical neural

networks. SiON shows promising role because of its **tailorability and moderate loss**, making it attractive for experimental AI photonic accelerators.

References:

1. Adam Krzywaniak [0000−0003−1904−2510], Pawel Czarnul [0000−0002−4918−9196] and Jerzy Proficz [0000−0003−2975−9339]
2. Gian-Luca Bona (Swiss Federal Laboratories for Materials Science and Technology), Roland Germann (IBM), and B. J. Offrein, IBM Journal of Research & development 47(2.3):239 – 249 DOI: 10.1147/rd.472.0239
3. International Conference on Microelectronics (ICM), 2000 (IEEE Cat. No.00EX453) "Data-driven dynamic logic versus NP-CMOS logic, a comparison"
4. Jinhwi Kim, Apostolos Galanopoulos, Jude Vivek Joseph, Jeongho Kwak ICTC 10.1109/ICTC49870.2020.9289270
5. Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E. (2017-05-24). "ImageNet classification with deep convolutional neural networks" *ACM*. **60** (6): 8490. doi:10.1145/3065386. ISSN 0001-0782. S2CID 195908774.
6. Murad Qasaimeh, Kristof Denolf, Jack Lo, Kees Vissers, Joseph Zambreno, and Phillip H. Jones, IEEE10.1109/ICESS.2019.8782524
7. Rui Tang, Makoto Okano, Chao Zhang, Kasidit Toprasertpong, Shinichi Takagi, Mitsuru Takenaka, 384770862, “Waveguide multiplexed photonic matrix vector multiplication processor using multiport photodetectors.”
8. Zhiping Zhou, Bing Yin, Qingzhong Deng, Xinbai Li, and Jishi Cui Vol. 3, Issue 5, pp. B28-B46 (2015) 10.1364/PRJ.3.000B28