NEUROPHOS

Revolutionizing the AI Data Center:

Delivering 4,000 Peta Operations per Second at 1% of the Power Using Optical Systolic Arrays

Michael Shebanow, Chief Digital Architect, Hod Finkelstein, CTO, and Patrick Bowen, CEO

Neurophos Inc.
November 2025

1 Executive Summary

The rapid advancement of artificial intelligence (AI) demands unprecedented computational power, yet current hardware solutions, primarily based on silicon GPUs, face fundamental limits in energy efficiency and cost. This white paper explores the challenges of scaling AI compute and introduces an innovative optical computing approach developed by Neurophos. By leveraging analog optical systolic arrays with metasurface technology, Neurophos achieves 100x improvements in speed and energy efficiency compared to traditional GPUs. This breakthrough aligns economic and environmental incentives, enabling widespread access to AI capabilities without prohibitive costs or energy consumption.

2 Introduction

As AI developers call for 100x more compute globally, the cost of building and operating these systems threatens to consume a significant portion of global energy production, if it continues to rely on existing silicon-based solutions. The root problem is the energy inefficiency of traditional processors, which scale speed without significantly increasing power efficiency. To achieve 100x more compute, increasing the speed of a modern GPU without altering the energy efficiency would turn a 1,200 W processor into a 120 kW device. The exponentially-increasing demand for AI compute is unsustainable without a dramatic improvement in power efficiencies.

Neurophos addresses this critical challenge by focusing on AI inference hardware for data centers using optical computing. This approach prioritizes energy efficiency as the foundation for speed gains, reducing the energy-per-operation of AI hardware by 100x while minimizing environmental impact. By solving the energy problem first, Neurophos enables scalable, high-performance AI without compromising on economics or sustainability.

3 The Fundamentals of AI Compute Efficiency

AI performance, particularly for inference tasks, is governed by two key metrics: compute density (operations per square millimeter of silicon) and manufacturing efficiency (cost per square millimeter). Their product determines raw speed per dollar. However, traditional silicon scaling has diminishing returns, as Moore's Law slows and energy costs dominate.

3.1 Systolic Arrays: A Historical Foundation

The concept of systolic arrays, pioneered by H. T. Kung and Charles Leiserson in the late 1970s, revolutionized matrix operations central to AI workloads. Systolic arrays process batches of data in a 2D grid of processing elements (PEs). Data flows rhythmically through the array, like blood flow pumped from the heart, performing matrix-matrix multiplications (e.g., $A \times B = C$) in the PEs. PEs are "fixed function," meaning they are hardwired to perform one or perhaps a few different operations, thereby avoiding the need to fetch-decode-dispatch instructions entirely.

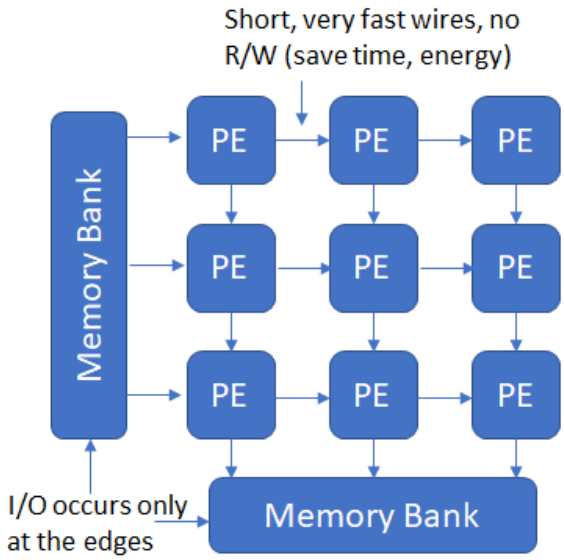


Figure 1: Systolic array operation using Processing Elements

Operationally, in a "weight stationary" mode of operation, one matrix is first loaded into the PEs, each PE having a small amount of memory. As a second step, rows of the second matrix are read from memory and streamed through the PEs. Each PE computes locally and passes data to neighbors, outputting rows of the result matrix which is in turn stored back in memory.

The systolic array architecture, resurrected in Google's Tensor Processing Units (TPUs), are very area and power efficient, making them ideal candidates for AI inference. However, digital systolic arrays plateau in efficiency around 100-200 elements due to the array's own power consumption dominating over memory savings.

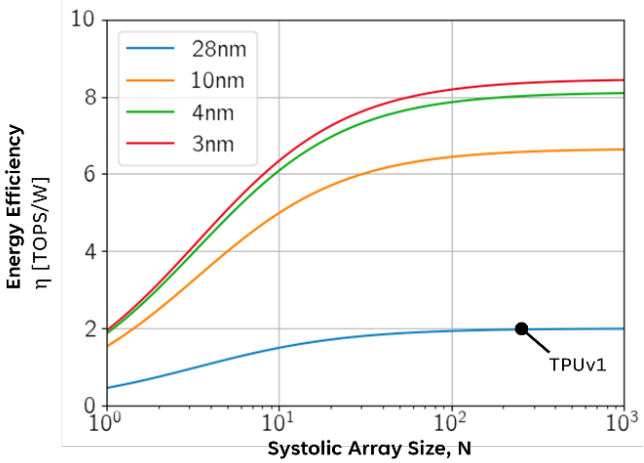


Figure 2: Energy efficiency limitation of leading semiconductor technology nodes

3.2 Digital vs. Analog In-Memory Compute

Digital systolic arrays typically operate at 1-2 GHz. Their power consumption scales with area (dimension²). Energy efficiency improves initially by amortizing memory costs but diminishes as

compute power overtakes. Analog systolic arrays replace digital PEs with linear analog counterparts driven by digital-to-analog converters (DACs) and analog-to-digital converters (ADCs). Unlike digital systolic arrays where one matrix can be kept stationary, the analog systolic array typically streams both matrices after converting then from digital-to-analog using DACs. At the output of the array, the result matrix is then streamed back from analog-to-digital using ADCs.

A benefit of implementing PEs in analog is greatly improved power efficiency. In an analog systolic array, power scales with perimeter (number of DACs/ADCs), while throughput scales with area. Larger arrays become increasingly efficient, as compute energy becomes negligible.

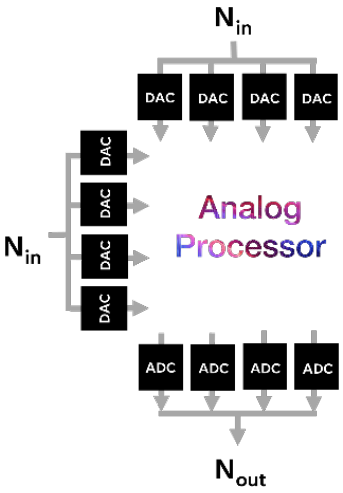


Figure 3: Number of DACs and ADCs (corresponding to power consumption) scales with the perimeter while throughput scales with area

Comparing the digital (tensor) and analog approaches, we see that the digital In-Memory Compute requires $E_{dig} = N^2 e_{op}$ per clock cycle while the analog In-Memory Compute requires $E_{analog} = N(e_{DAC} + e_{ADC})$. Thus, the analog architecture is more efficient by a factor of $\frac{E_{analog}}{E_{digital}} \propto \frac{1}{N}$. As arrays increase in size, the analog architecture becomes dramatically more efficient.

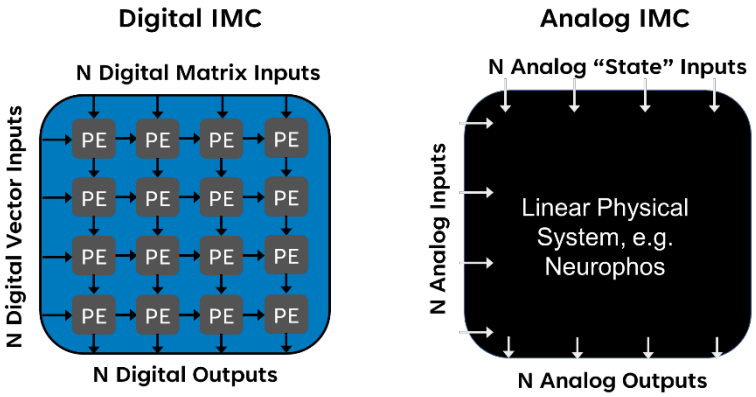


Figure 4: Illustration of a tensor and analog in-memory compute architectures

Despite the promise, analog electronic implementations (e.g., using resistors and capacitors) suffer from signal propagation RC delays proportional to area, resulting in clock speeds which are orders of magnitude slower than is needed to compete with digital arrays

4 Optical Systolic Arrays: Unlocking Speed and Efficiency

Neurophos has overcome these limitations by developing an optical in-memory compute module, where signals propagate at the speed of light and are not encumbered by RC delays, enabling clock speeds of 50 GHz, and higher, independent of array size.

4.1 From Electronics to Photonics

Optical arrays maintain analog scaling laws (power scales with perimeter (∼N), throughput scales with area (N^2), where N is the number of vector elements) but eliminate RC delays. As matrix sizes increase, so does their efficiency improvement over digital arrays, effectively without bounds.

Previous attempts to use silicon photonics to implement optical arrays failed due to the large physical dimension of Mach-Zender modulators (1 mm), limiting the size of matrices to approximately 64×64 elements, which is far smaller than required for AI inference applications.

4.2 The Ideal AI Tensor Core

Neurophos' design combines optimal features:

- **Dense**: Small modulators for massive arrays.
- **Analog**: Perimeter-based power scaling.
- **Photonic**: Speed-of-light propagation.
- **Volatile**: Can be updated at MHz rates.
- **Model-agnostic**: Compatible with diverse models.
- **Flexible**: Handles fixed/floating-point formats.
- **Manufacturable**: Integrates with standard CMOS Technologies.

5 Neurophos' Matrix Multiplication Architecture

Modern AI pipelines implement sequential matrix-matrix multiplications. An Activation Matrix, containing query data typically multiplies a Weight Matrix, containing model data. The dimensions of these matrices vary by model and operation but can grow very large. For example, in DeepSeek V3 Prefill, some activation matrices have 131,072 x 16,284 = 2.14 billion elements, and some weight matrices have 16,384 x 5,120 = 84 billion elements.

To process these huge matrices, Neurophos' first generation product parses the activation matrix into slices which are 1,024 wide and up to 56,000 long and parses the weight matrix into 1024 x 1024 tiles. Each column slice in the activation matrix is multiplied by a weights tile resulting in a column partial product, which is stored in memory. In practice, the column slice is decomposed to a stream of vectors, each of which multiplies the weights tile, decomposed to a stream of vectors, processed at 56 GHz, for 56 billion vector-matrix multiplications per second. Subsequent column slices are multiplied by subsequent vertical tiles and the results are accumulated into memory. This process repeats for each column of weight tiles.

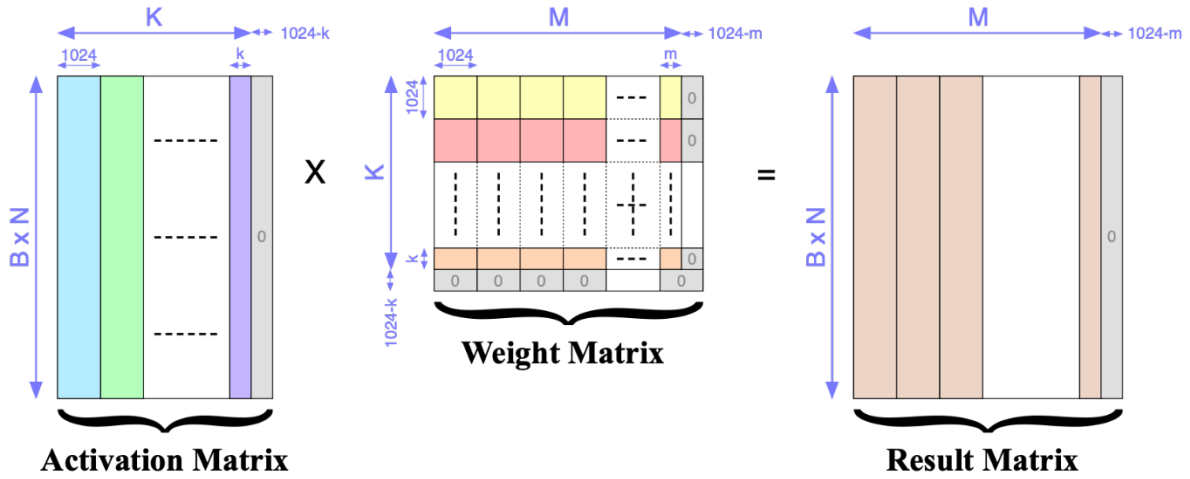
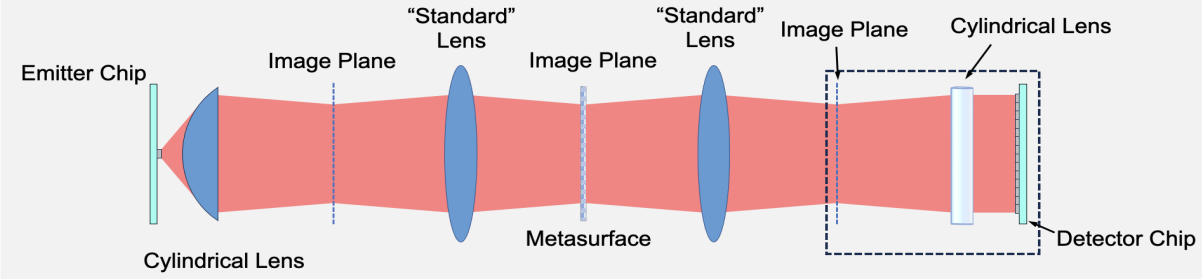


Figure 5: Neurophos' Architecture: Slicing and tiling enables efficient matrix-matrix multiplication

6 Neurophos' Metasurface Technology

Inspired by J.W. Goodman's 1970s Fourier optics vector-matrix multiplier—using lenses and holographic films for multiplication and summation—Neurophos has developed a compact and densely packaged module.

Goodman's original concept defined a method to perform vector-matrix multiplication. The values of vector elements are encoded as amplitude in a linear emitter, e.g., as a vertical set of beams. The beams are expanded orthogonally (horizontally) via a cylindrical lens and illuminate a partially-transmissive screen on the image plane. The screen encodes matrix values as opaqueness of individual elements. A second cylindrical lens sums light intensities in the orthogonal (vertical) dimension to provide the vector-matrix-multiplication resultant vector.



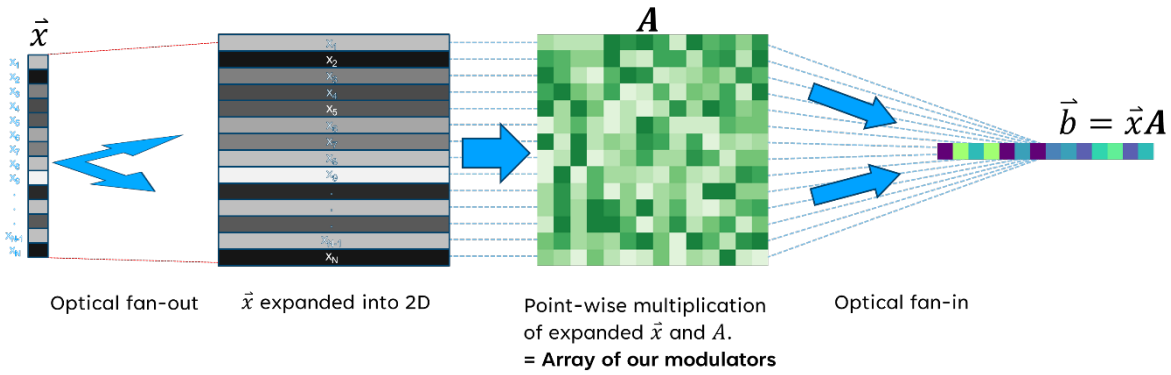


Figure 6: The Goodman architecture: An emitter chip encodes vector data as intensities; emitter beams are projected on a metasurface, which encodes matrix values,; transmitted light through the metasurface is focused in one dimension to deliver the resultant vector

Goodman's architecture has not been translated into a commercial product in the 50 years since its inception due to the lack of dense and volatile screens, as well as to the long physical distance between the emitter and detector devices, which result in system and interconnect complexity.

Neurophos' innovations: Neurophos has introduced two innovations which finally enable the productization of Goodman's architecture.

- The first is the world's densest and fastest commercially-available spatial-light-modulator array.
- The second is the optical folding of the system using beam splitters, which enable a single-plane for the input and output vectors, as well as the matrix, all of which are vertically-stacked on top of CMOS drive circuitry.

6.1 Core Innovation: Metasurface Modulators

Metasurfaces are planar devices containing subwavelength features, which manipulate the wavefront of light. Typically, metasurfaces are made of dielectric materials with etched patterns, and are used either as flat lenses or as beam-shaping elements. These devices can be considered as One-Time-Programmable (OTP) optical memories which encode a certain wavefront transfer function in a substrate material.

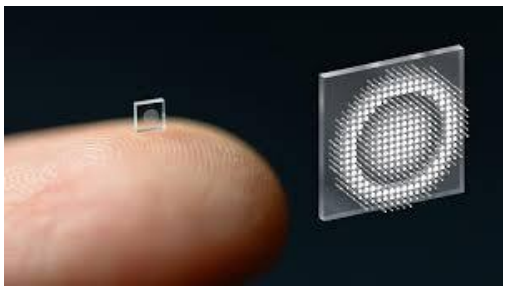


Figure 7: A legacy static metasurface can be designed to manipulate the wavefront of light to replace a traditional lens

Neurophos has developed an active or tunable metasurface platform, where this wavefront transfer function can be dynamically re-written in an array format, akin to an optical Dynamic Random Access Memory (DRAM). A Neurophos metasurface chip, which is completely CMOS-compatible (contains only standard silicon-foundry materials and uses only standard foundry tooling and processes), comprises an array of pixels. Each pixel contains tens to hundreds of cell elements, which, when combined and electrically biased, exhibits a programmable reflectivity and optical phase retardation.

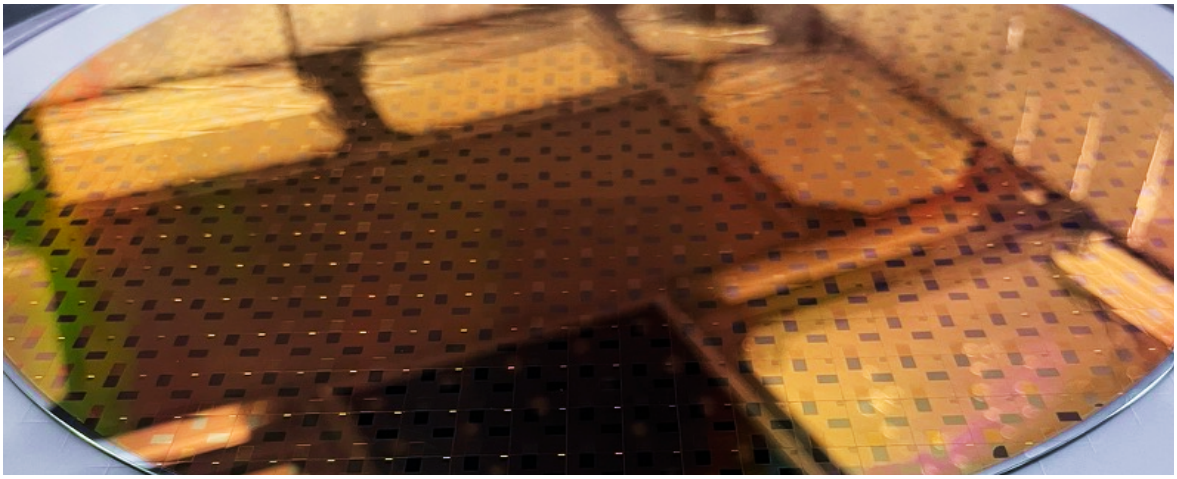


Figure 8: Neurophos' active metasurfaces are manufactured on a 300mm silicon wafer using standard fab tooling and materials

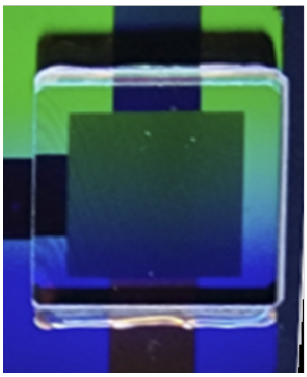


Figure 9: Neurophos' packaged and lidded metasurface chip

Neurophos has already fabricated these metasurfaces and validated their performance. Switching speeds below 1 microsecond have been demonstrated, in good agreement with modeling and simulation data. The eye diagram below shows the accumulation of many switching cycles. Specifically, the signal shown is the output of a photodetector recording the light intensity from the metasurfaces as its drive voltage is switched from high to low every 420 ns. The opening between the high photodetector output and the low output indicate that we can reliably distinguish between the low and high states of the metasurface, enabling matrix switching rates in excess of 1 MHz.

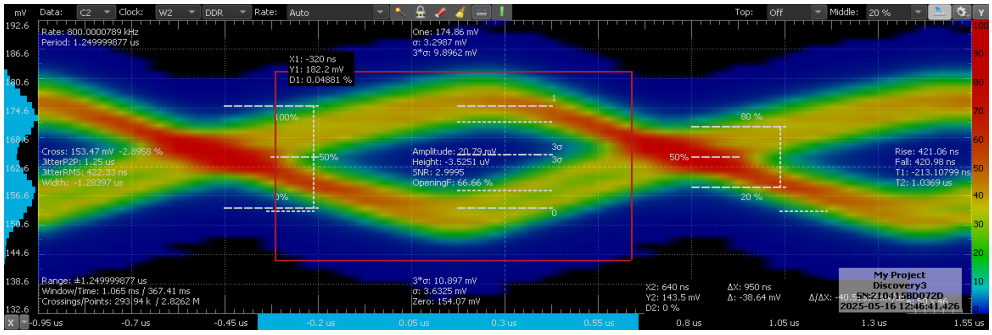


Figure 10: An eye diagram demonstrating clean at-speed switching of the metasurface

Below are spectra collected from the Neurophos metasurface array. Referring to the solid lines (averaged over multiple traces to remove measurement noise), each curve corresponds to a different bias voltage on the same metasurface device. These bias voltages control the effective index of refraction of the metasurface from 1.5 to 1.7. A spectrometer records the reflectance spectrum from the metasurface for each of these voltage bias conditions. As can be seen, the metasurface exhibits a reflection minimum, which changes position as a function of voltage.

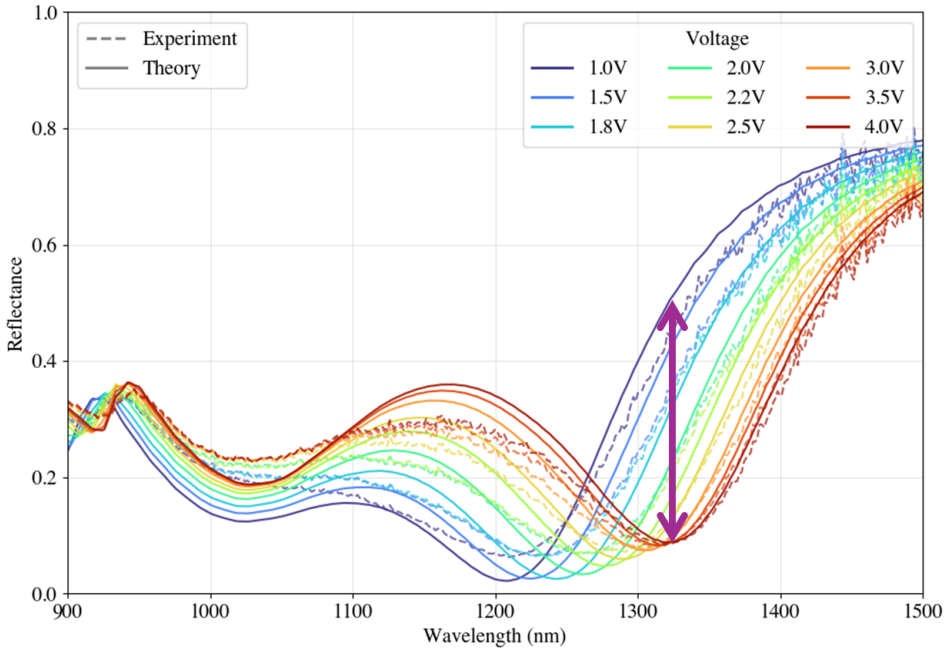


Figure 11: Spectral reflectivity of the metasurface; solid lines represent simulated responses and dotted lines are experimental data. As shown by the purple arrow, at a given wavelength, changing the metasurface pixel's voltage modifies its reflectivity, enabling optical multiplication

During module operation, monochromatic light illuminates the surface. The reflectivity of the metasurface (and its phase retardation, not shown in the plot) changes value, from 10% to 55%. This switching behavior enables the metasurface to encode the matrix values. When a beam, which encodes information in its amplitude and phase, interacts with a metasurface pixel, which encodes information in its reflectivity and phase retardation, a multiplication process takes place.

Neurophos' tiny metasurface elements offer clear advantages over traditional silicon photonic modulators:

- They are approximately 10,000 times smaller than the traditional Mach-Zehnder modulator devices, thus enabling integration of millions of such devices on a single die.
- They are drastically more robust than ring modulators, which suffer from thermal and carrier effects due to silicon's non-linear behavior, leading to instability and self-pulsation, and resulting in extreme sensitivity to temperature variations. Consequently, Neurophos' metasurfaces do not require the complex DC tuning and heaters, which increase ring modulators' power consumption and design complexity.
- Each input interacts with a section of spatial light modulators (weights) only once, avoiding sequential processing. This represents a significant improvement over existing optical and electronic analog systolic array implementations, which suffer from the extinction problem. In crossbar architectures, where signals pass sequentially through modulator arrays, losses accumulate rapidly as the signal traverses rows. For example, with a mere 5% attenuation per pass, the signal strength after 1,000 rows drops to 0.95^{1000} of its original value, rendering it effectively undetectable.

6.2 Optical Folding and System Architecture

To utilize the metasurface in a high throughput Optical Processing Unit, Neurophos has folded the Goodman optical architecture, creating a dense, manufacturable module design.

As shown schematically below, an Electronic Integrated Circuit (EIC) converts lower-speed digital signals into high-speed analog drive signals. A coherent laser source (not shown) is coupled into a photonic integrated circuit (PIC), is fanned out to N channels (N being the vector length), and the light in each of these channels is modulated by the analog drive signals from the EIC. The light from these N channels is coupled into free space using grating couplers, is expanded in one dimension and folded back onto the reflective metasurface, which contains N x N pixels. The reflected light is compressed in the orthogonal dimension and is folded back into a linear array of grating couplers on a Receive PIC, where it is detected. The EIC digitizes the analog information, and transmits it as a slower, wider data stream.

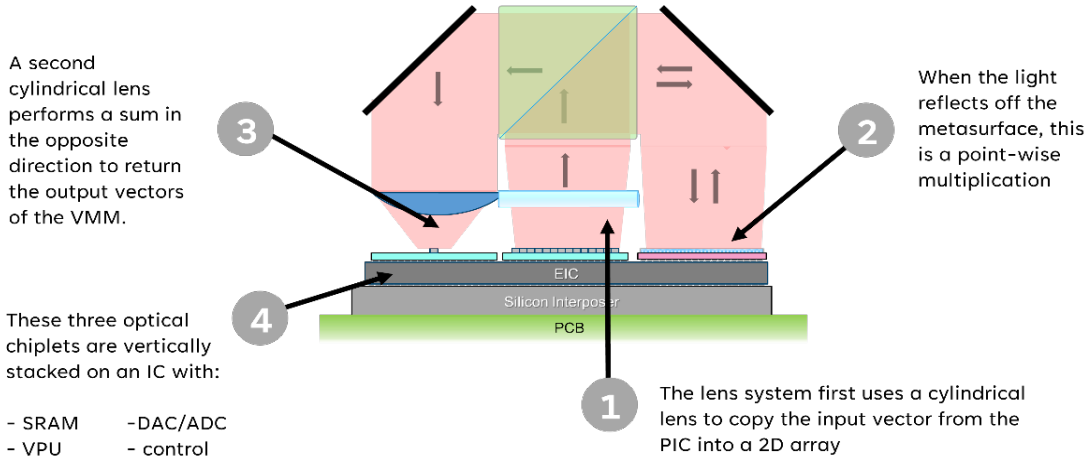


Figure 12: The Neurophos folded Goodman architecture module

7 Performance and Future Outlook

Neurophos' first generation product OPU will deliver a 1024 x 1024 pixel metasurface array, with 4 bit precision, at 56 GHz. Each vector-matrix multiplication translates to 1024^2 multiplications and 1024×1023 additions, for a total of 2.1 million operations. This product can perform two such multiplications in parallel, and at 56 GHz this yields a peak 235 Peta Operations per Second (POPS) (2.35×10^{17}) during peak operation while consuming 675 W, corresponding to 348 Tera (3.48×10^{14}) Operations per Second (TOPS) per Watt and a compute density of 285 TOPS/mm^2. In contrast, NVIDIA's B200 GPU delivers 9 POPS (dense FP4 precision) and consumes approximately 1,000 W. This corresponds to an efficiency of 9 TOPS/W with a compute density of 5.5 TOPS/ mm^2.

	NVIDIA B200	**Neurophos Gen 0**	**Ratio**
Data Throughput [POPS]	9	235	26.1
Compute Density [TOPS/mm2]	5.5	285	51.8
Compute Efficiency [TOPS/W]	9	348	38.7

Table 1: Key Performance Indicators of the NVIDIA B200 and Neurophos Gen 0 devices

In the future, integrating 9.5-million-pixel metasurface arrays at 56 GHz in a 2-chip configuration is projected to yield up to 4,228 dense FP4 POPS, 6.5 times faster than an NVIDIA GB200 NVL72 rack, which contains 72 Blackwell GPUs. This will enable hyperscalers to keep up with the exponentially growing demand for AI compute power without overwhelming our global power grid.

8 Conclusion

By addressing energy efficiency through optical systolic arrays and metasurface technology, Neurophos paves the way for dramatically faster and more power-efficient AI hardware. This not only accelerates AI innovation but aligns with sustainable growth. As the world pursues AI's full potential, Neurophos' optical processing solution will be pivotal in making compute abundant and accessible.