

## RESEARCH ARTICLE OPEN ACCESS

# Photonic Systolic Array for All-Optical Matrix–Matrix Multiplication

Jungmin Kim  | Qingyi Zhou  | Zongfu Yu

Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, Wisconsin, USA

**Correspondence:** Jungmin Kim ([jkim2325@wisc.edu](mailto:jkim2325@wisc.edu)) | Qingyi Zhou ([qzhou75@wisc.edu](mailto:qzhou75@wisc.edu))**Received:** 28 July 2025 | **Revised:** 27 October 2025 | **Accepted:** 4 November 2025**Keywords:** matrix-matrix multiplication | optical computing | systolic array

## ABSTRACT

Systolic arrays have proven to be highly efficient for parallelized matrix–matrix multiplication (MMM), utilizing synchronized, heartbeat-like data flows across an array of processing elements. While optical structures, such as waveguide crossbar arrays and Mach-Zehnder interferometer-based meshes, serve as photonic equivalents to the systolic arrays, the disparity between the two input matrices for multiplication—one using optical signals and the other with system-defined parameters—gives rise to a bottleneck in modern machine-learning tasks, such as evaluating attention scores in large language models. Here, a photonic systolic array that performs MMM entirely with optical signals is proposed, utilizing homodyne detection at each array cell. Adjoint-based design of compact on-chip freeform optical modules enables precise control of light flow without bulky waveguide coupling schemes. The operation of 4×4 and 2×2 photonic systolic arrays are numerically verified, achieving a theoretical computation density of 4.4 PMACs/mm<sup>2</sup>/s. This design marks a significant step toward practical photonic computing hardware for modern AI workloads.

## 1 | Introduction

With the recent rise of large language models [1–3] for generative artificial intelligence, which rely on the attention mechanism involving extensive matrix–matrix multiplication (MMM) operations [4], there has been an increasing demand for energy-efficient and highly parallelized computing hardware. One notable example is the systolic array (SA), which was originally introduced in 1978 [5] and recently has been revisited for its efficiency in MMM [6–11]. As implemented in Google’s tensor processing units [6, 11], this architecture performs MMM directly on a two-dimensional array of processing elements (PEs), significantly reducing the need for memory access by utilizing the heartbeat-like movement of input and/or output data streams. Importantly, SAs can be classified into two types: weight-

stationary (WSSA) and output-stationary (OSSA) [8]. In WSSA, MMM works similarly to a linear transformation, where a weight matrix ( $\mathbf{W}$ ) is preloaded onto the array of PE while an input matrix ( $\mathbf{X}$ ) is streamed in one direction, resulting in the output matrix ( $\mathbf{Y} = \mathbf{WX}$ ) streamed out to another direction, as shown in Figure 1a. On the other hand, OSSA receives two input matrices ( $\mathbf{A}$  and  $\mathbf{B}$ ) and accumulates the multiplication result ( $\mathbf{C} = \mathbf{A}^T \mathbf{B}$ ) on the spot, as shown in Figure 1b.

Leveraging photon as an information carrier [12], significant efforts have been devoted to the optical analogy of parallel computing hardware to address the needs with extreme operation speed and energy efficiency [13–22]. For instance, waveguide crossbar arrays [23–28], unitary meshes based on Mach-Zehnder interferometers (MZI) [29–34], and inverse-designed nanopho-

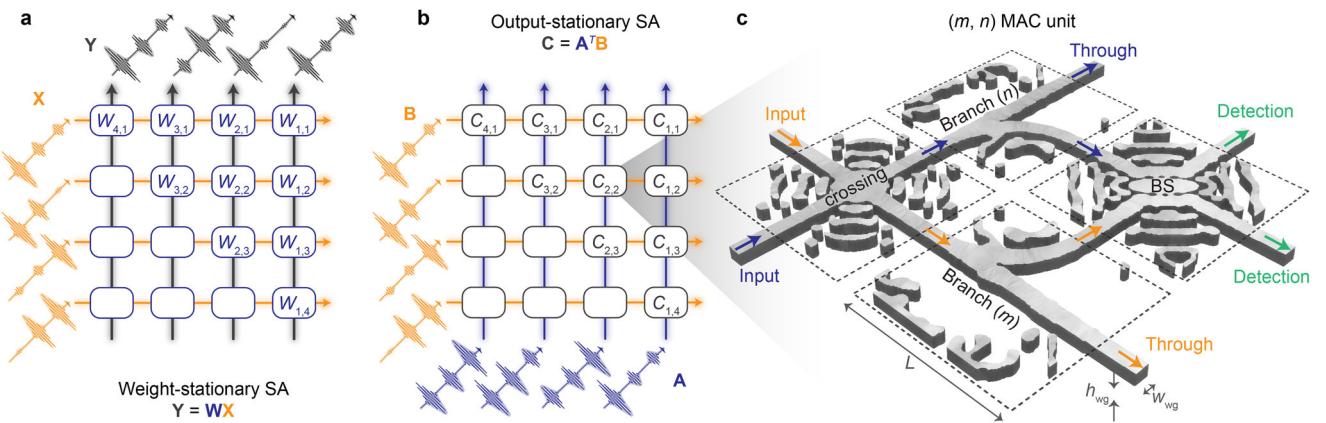
---

Jungmin Kim and Qingyi Zhou contributed equally to this work.

---

This is an open access article under the terms of the [Creative Commons Attribution](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). Laser & Photonics Reviews published by Wiley-VCH GmbH



**FIGURE 1** | Concept of photonic systolic array. (a) Weight-stationary type: input signals  $\mathbf{X}$  are transformed to the output signals  $\mathbf{Y} = \mathbf{W}\mathbf{X}$  by a system parameters  $\mathbf{W}$ . (b) Output-stationary type: two input signals ( $\mathbf{A}$  and  $\mathbf{B}$ ) are multiplied on the array of multiply-accumulate (MAC) units, resulting in the stationary output ( $\mathbf{C} = \mathbf{A}^T \mathbf{B}$ ). The array is interconnected via vertical and horizontal waveguides, carrying amplitude-modulated optical pulses for parallel MAC operations, respectively. (c) Each optical MAC unit consists of a waveguide crossing, two branches indexed by  $m$  and  $n$ , and a beam splitter for homodyne detection. The waveguides and submodules are built from a Si slab structure with uniform thickness  $h_{wg}$ , embedded within a  $\text{SiO}_2$  buried oxide layer and cladding. Parameters:  $L = 3.50$ ;  $h_{wg} = 0.22$ ;  $w_{wg} = 0.3$  [ $\mu\text{m}$ ].

tonic structures [35–37] can be regarded as WSSAs in the sense that only one of two matrices is encoded as traveling optical time signals. That is, the other matrix is not encoded optically, but as near-field (resonant) coupling constants, electro-optic modulation phases at the intersections, or complex material distribution. However, this weight-stationary approach is not ideal for certain machine-learning tasks; during training of deep neural networks, weights need to be not only programmable but also directly mappable to the corresponding site, unlike MZI meshes which requires a prior nulling process [31]. Furthermore, two multiplicands often need to be tuned dynamically, such as the multiplication of query ( $\mathbf{Q} = \mathbf{W}_Q \mathbf{X}$ ) and key ( $\mathbf{K} = \mathbf{W}_K \mathbf{X}$ ) both as a function of embeddings  $\mathbf{X}$  in transformers [4]. In this context, a photonic OSSA is particularly important, allowing for the general MMM between two optical signals on an equal footing.

Despite its potential advantages, however, photonic OSSA has been explored far less compared to WSSA. In Ref. [22], Ou et al. experimentally demonstrated a MMM between wavelength- and space-encoded signals ( $(X_{\lambda,t}, W_{t,s})$ , with the results temporally accumulated:  $Y_{\lambda,s}$ . While this approach can be categorized as OSSA in the sense that the output is obtained over time, there remains a substantial disparity between encoding schemes of the two input signals, making it difficult to treat them equally. Another OSSA prototype has been suggested and experimentally demonstrated in Refs. [38, 39], where both input signals are spatially encoded and the output is directly captured by an image sensor, resulting in a stationary and truly matrix-shaped output. In this work, we advance this architecture by proposing a small-footprint, freeform-designed, output-stationary, silicon-on-insulator photonic systolic array (PSA) for real-valued MMM between purely optical signals. By employing the adjoint state method [34, 37, 40–44], we precisely target the amplitude and phase-matching conditions necessary for even power distribution and homodyne detection [14, 19]. This approach resolves the remaining incoherency issues upon device scaling in the preliminary studies [38, 39], while significantly reducing the device footprint down to ten-micron scale per element by eliminating the need for waveguide coupling schemes. As a result, we

achieve an exceptionally high computation density, theoretically reaching 4.4 PMACs/ $\text{mm}^2/\text{s}$ . As an example, we demonstrate the temporal operation of vector outer product and matrix-matrix multiplication using  $4 \times 4$  and  $2 \times 2$  PSAs, respectively, enabled by GPU-accelerated finite-difference time-domain (FDTD) simulation [45, 46].

## 2 | Results

### 2.1 | Operation Principle

Scalar multiplication and addition are the two main arithmetic operations for MMM, which can be performed optically using homodyne detection [14, 19, 39]. This involves a beam splitter (BS) that interferes two monochromatic signals,  $a e^{-i\omega_0 t}$  and  $b e^{-i\omega_0 t}$ , of a  $\pi/2$  phase difference. When these signals impinge into a 50:50 BS, the scattering relation can be expressed as:

$$\begin{pmatrix} o_1 \\ o_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix} \begin{pmatrix} a \\ ib \end{pmatrix} \quad (1)$$

where  $a, b$ , and  $o_{1,2}$  are real-valued input amplitudes and the complex output signals, respectively. The intensity difference between the two output beams is written as  $\Delta I \equiv |o_1|^2 - |o_2|^2 = -2\text{Re}(a^*b) = -2ab$ , effectively performing scalar multiplication on the two input amplitudes up to a factor of -2. Furthermore, if the amplitudes are slowly varying over time as  $a(t)$  and  $b(t)$ , we can approximate the received photon counts difference at two detectors within a time range  $[t_s, t_e]$  as

$$\Delta N = -\frac{2A}{\hbar\omega_0} \int_{t_s}^{t_e} dt a(t)b(t) \quad (2)$$

where  $A$  is the area of detectors. These arriving photons are converted to electrons with quantum efficiency  $\eta$  and accumulated at a capacitor, resulting in the charge  $\langle a|b \rangle_t \equiv \eta q \Delta N_{12}$  as the real-valued inner product between  $a$  and  $b$  along time. It is noted that throughout the paper, units of intensity (continuous waves) and

power (or, energy for amplitude-modulated pulses) can be used interchangeably for the device output.

Now, our approach focuses on (1) designing a compact photonic MAC unit and (2) implementing it in an array to combine temporal inner products with spatial outer products [39], to achieve MMM (Figure 1b). In this scheme, every single column vector of input matrices (**A** and **B**) is encoded as a time signal  $A_m(t)$  and  $B_n(t)$  consisting of several optical pulses, where  $m$  and  $n$  are array indices. Each element of the output matrix is computed through the photonic MAC process.

The design of the MAC unit, illustrated in Figure 1c, includes four key components: a waveguide crossing, two indexed ( $m$  and  $n$ ) branches, and a BS, all of which are free-form designed within the black square area. In each  $(m, n)$  MAC unit, the input waves from neighboring cells  $(m+1, n)$  and  $(m, n+1)$  first cross without crosstalk. The waves then fork into a straight path ( $\sim 1-k^{-1}$ ) and a right-angled path ( $\sim k^{-1}$ ) with the noted split ratio, where  $k = m, n$ . The branched signals lastly encounter at the BS to return the inner product of the two time signals. Notably, this configuration allows the input signals to travel through waveguides distributing almost the same portion of their energy for detection at each MAC unit.

## 2.2 | Adjoint-Based Submodule Design

As mentioned, the primary goals of the inverse design in this scheme are compactness, equal power distribution, and precise phase matching for interference-based operation. In sharp contrast to the conventional approach relying on waveguide near-field coupling, we utilize adjoint method for the inverse design, by which we can directly assign the necessary specifications for the submodules (e.g., amplitude and phase of the  $S$ -parameter) while significantly reducing the design footprint [21].

We assume the  $TM_{00}$  mode at the frequency  $f_0 = 193.4$  THz of the rectangular waveguide as a carrier mode, which is injected from the bottom port of each submodule (black arrows, Figure 2a). Then,  $S$ -parameters for the same mode are measured at the top ( $S_{13}$ , yellow arrows) and right-side ( $S_{23}$ , teal arrows) ports for comparison with target values. For the target phase of the  $S$ -parameters, we set all output phases to converge to  $k_{\text{eff}}L_{\text{port}}$ , where  $k_{\text{eff}}$  and  $L_{\text{port}}$  are the effective wave number through waveguides and the physical port distance, respectively, with the only exception for a  $\pi/2$  phase shift at the right-side port of the BS. See Figure S1 for the effective index and mode profile of the waveguide structure.

For the target amplitude of the  $S$ -parameters, ideal values would be  $|S_{23}|^2 = 0$  for the crossing, 0.5 for the BS, and  $n^{-1}$  for the  $n$ -th branch to ensure uniform energy distribution across the array, with  $|S_{13}|^2 = 1 - |S_{23}|^2$ . However, due to practical limitations on achieving zero insertion loss for broadband operation, known as the Bode-Fano limit [47], a bit of margin for the insertion loss is incorporated into the optimizations as  $\alpha \equiv |S_{13}|^2 + |S_{23}|^2 \lesssim 1$ , in a loss-compensated fan-out design approach [38]. Based on this approach, the practical target amplitudes for branch submodules are slightly adjusted from their ideal values, as depicted by dashed horizontal lines in Figure 2b. See Note S1 for the details of loss compensation.

By specifying the target amplitude and phase for the  $S$ -parameters, the gradient-descent optimization for each submodule is available, as shown in Figure 2 (see Note S2 and Figures S2 and S3 for detailed simulation descriptions). The submodule structures and the corresponding wave profiles are shown in Figure 2a. Importantly, the output phase at ports 1 and 2 (yellow and teal arrows) are all matched, except for the  $\pi/2$  phase shift at port 2 of the BS. Figure 2b illustrates the frequency response of the designed structures, showing good agreement between the target transmission levels (dashed lines) and the actual transmission (solid lines). Notably, these transmission spectra provide a rough estimation for operation bandwidth around  $0.01f_0$ , which will be analyzed in the next section. It is also noted that, although we did not explicitly evaluate the sensitivity of the optical responses to fabrication errors during the optimization process, such as erosion or dilation of material boundaries, Figure S4 shows the robustness of our designed structures against such variations.

## 2.3 | Time-Domain Analysis

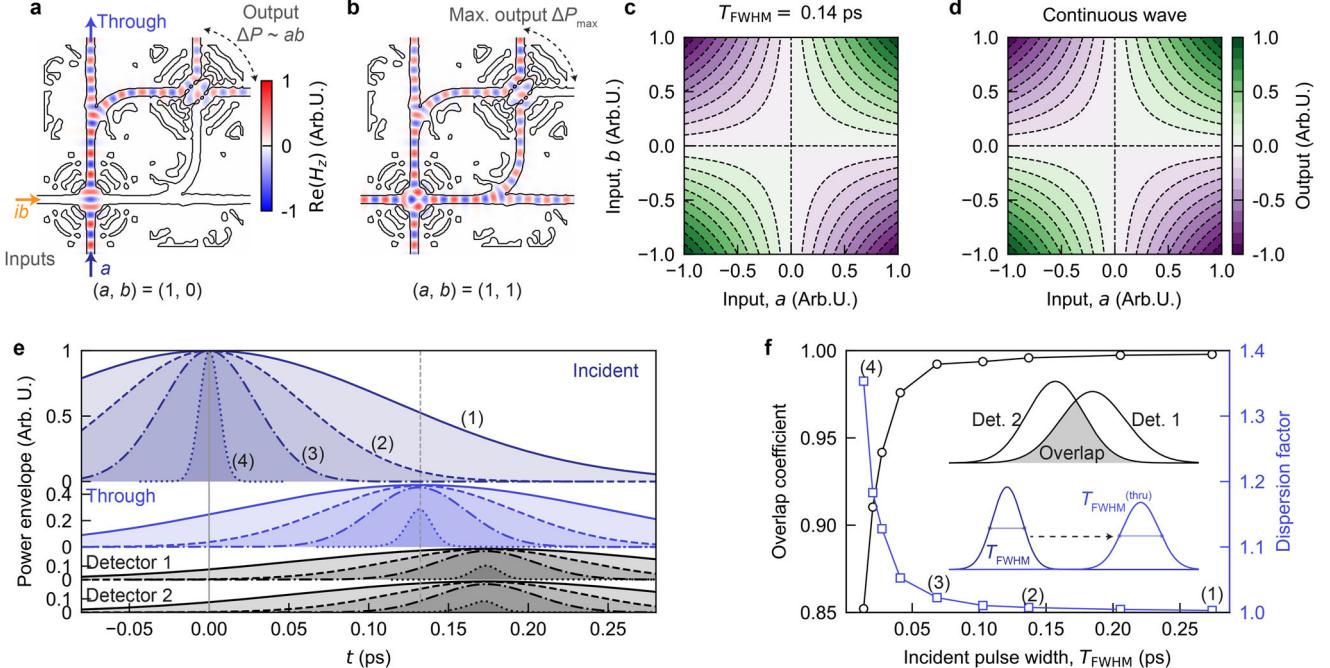
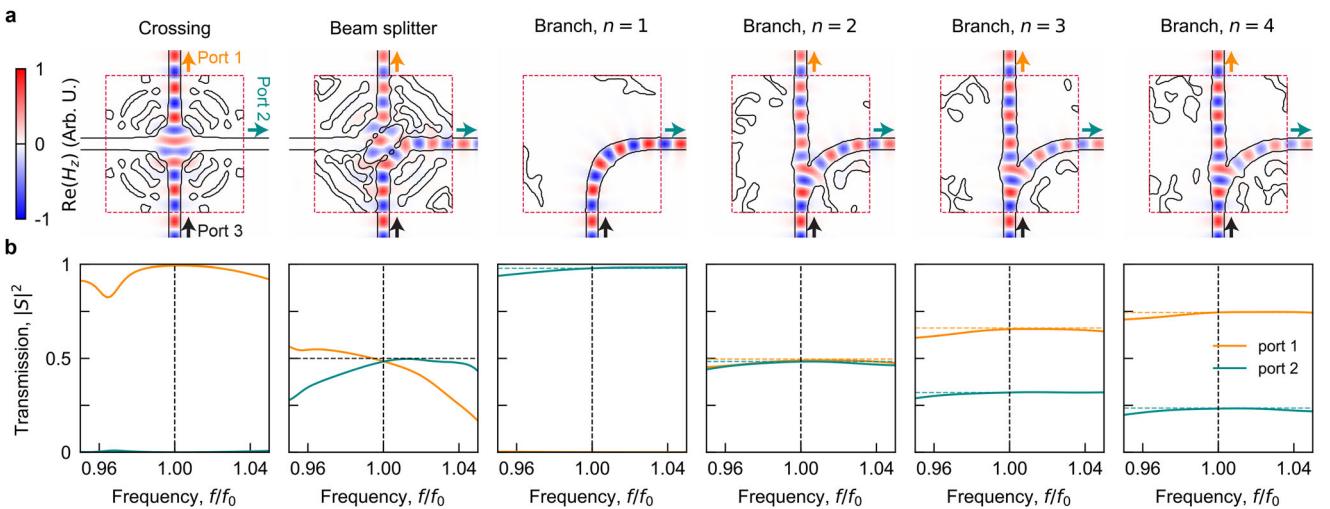
The key demonstration of this work is the operation of PSA, which relies on the time-domain streaming of data through optical pulses along waveguides and their exact synchronization. To assess device performance, it is essential to evaluate how different pulse widths affect the operation, thereby determining the maximum achievable bandwidth. We therefore test the (2,2)-MAC unit as an example, consisting of one crossing, two branches with  $n = m = 2$ , and one BS, as shown in Figure 3a (see detailed simulation setup in Figure S5). In this configuration, approximately half of the input power is expected to be passed to subsequent units, while the other half is detected as:

$$\begin{pmatrix} o_1 \\ o_2 \end{pmatrix} \approx \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} ib \\ a \end{pmatrix} \quad (3)$$

similar to Equation (1) except for the transposed inputs after passing through the branches and the factor of  $1/\sqrt{2}$ , leading to the power difference  $\Delta P = |o_1|^2 - |o_2|^2 \approx ab$ . We define the maximum power as  $\Delta P_{\max} = \Delta P(a=1, b=1)$ , as shown in Figure 3b, which is used for the normalization of the detected power to guarantee that the device output for the input (1,1) corresponds to one.

In Figures 3c and 3d, we verify that the (2,2)-MAC unit performs scalar multiplication accurately for both finite-width Gaussian pulses  $(a, ib) \exp[-(t/T)^2/2 - i2\pi f_0 t]$  with full width at half maximum (FWHM)  $T_{\text{FWHM}} \sim 0.14$  ps (Figure 3c) and continuous waves for  $T \rightarrow \infty$  (Figure 3d). The colormap displays the normalized device output  $\Delta P(a, b)/\Delta P_{\max}$ , which matches the ground truth ( $= ab$ ) represented by dashed contours.

However, due to the non-flat spectral responses shown in Figure 2b, reducing the pulse width  $T$  leads to performance degradation. To quantify this, we use Gaussian input pulses with different FWHMs from 0.274 ps to 0.014 ps, corresponding to pulse widths in frequency domain from  $f_{\text{width}} \equiv (2\pi T)^{-1} = f_0/200$  to  $f_0/10$ , respectively. Figure 3e shows that as the input pulse (top panel) narrows in time [ $T_{\text{FWHM}} = 0.274$  ps (1) to 0.014 ps (4)], the output pulses at the through port (center panel) and detection ports (bottom panel) become increasingly



dispersed, modifying both their heights and shapes. For instance, a dispersion factor, defined as the ratio of the FWHM at the through port to the incident FWHM (i.e.,  $T_{\text{FWHM}}^{(\text{thru})}/T_{\text{FWHM}}$ ), increases rapidly when the incident FWHM drops below  $0.14\text{ ps}$ , as shown by the blue curve in Figure 3f. This dispersion primarily arises from the crossing and branch submodules. Additionally, the BS's dispersion, illustrated in Figure 2b, contributes to the

discrepancy between the two output pulses at different detection ports. To evaluate the synchronization between the two detected powers,  $p_{1,2}(t) = |o_{1,2}(t)|^2$ , for the input  $(a, b) = (1, 0)$ , we define the overlap as:

$$\text{Overlap}(p_1, p_2) \equiv \frac{\int \min[p_1(t), p_2(t)]dt}{[\int p_1(t)dt \int p_2(t)dt]^{1/2}} \quad (4)$$

As shown in Figure 3f, the overlap between the two output ports rapidly decreases below one as the pulse width narrows, indicating that the device cannot ensure precise homodyne detection due to the unsynchronized signals. Based on these studies, it is safe to conclude that sufficiently long optical pulses can robustly perform scalar multiplication and inner product without suffering from signal degradation throughout the system.

## 2.4 | Vector Outer Product

We now extend the concept toward an array of the MAC units based on the previous time-domain analysis, enabling spatial parallelization of a MMM as shown in Figure 1b. In sharp contrast to the single MAC operation, however, the way of arranging practical detectors within each of elements becomes a problem, since they cannot be simply regarded as in-plane ports in practical setting. Among several options such as integrated on-chip detectors using InGaN/GaN nanowires [48] or graphene sheets [49], we utilize grating couplers [42] to extract optical energy out to  $z$ -direction in free space, so that we can acquire 2D planar images over the chip, as experimentally demonstrated in Ref. [39]. Thus, we additionally inverse design a grating coupler in a similar manner to the submodules, obtaining a freeform layout of the same module size ( $L$ ) that efficiently converts the waveguide mode into a Gaussian beam of beam waist of  $L/2$  along the  $z$ -direction. Detailed schematic and the training result is shown in Supplementary Figures S6 and S7.

Figure 4 demonstrates the outer product  $\mathbf{A} \otimes \mathbf{B}$  (i.e., a MMM only with a single temporal element) using  $4 \times 4$  array of MAC units. In detail, two grating couplers are attached to the two output ports of the BS at each element, as marked by red boxes in Figure 4a. Then, a single pulse of amplitude  $A_m$  and  $iB_n$  is injected through each waveguide from bottom and left side, respectively. Those optical signals are processed at the array of MAC units as shown in Figure 4a, and the output signals are emitted through the grating couplers, which are then recorded by CMOS sensor placed over the chip as illustrated in Figure 4b. Green and purple dashed boxes indicate the location of the pair of detectors for each MAC unit, having the strong vertical photon flux through them. See Figure S8 for the simulation setup.

While Figure 4a,b only visualizes the  $f = f_0$  component of the device operation, indeed the optical pulses with finite pulse widths should be injected in such a way that  $A_m$  and  $B_n$  encounter right at the  $(m, n)$ -MAC unit, maximizing the computation accuracy and efficiency. Figure 4c shows the time delay of pulses to ensure this operation.  $A_4$  and  $B_4$  are initially injected simultaneously, which are followed by  $m, n = 3, 2, 1$ , with certain delays corresponding to the traveling time for an optical pulse between two adjacent units. Those input powers are equally distributed over the array and processed, and then output pulses are emitted from the grating couplers turn by turn from the lower left unit ( $m + n = 8$ ) at  $t = t_8$  to upper right one ( $m + n = 2$ ) at  $t = t_2$ .

Figures 4d-4g display the outer product results. For the same input signals  $\mathbf{A} = [0.42, -0.5, 0.65, -0.5]$  and  $\mathbf{B} = [0.76, -0.27, 0.65, -0.73]$  as optical pulses with the FWHM of 0.274 ps as shown in Figure 4c, we obtain the received energy

$P_{m,n}^{\text{C,D}}$ , at detectors C and D per each  $(m, n)$ -MAC unit, respectively, and calculate the raw output by  $\Delta P_{m,n}(\mathbf{A}, \mathbf{B}) = P_{m,n}^{\text{C}} - P_{m,n}^{\text{D}}$  as a function of input  $\mathbf{A}$  and  $\mathbf{B}$ . The ground truth result and the raw output are compared by Figure 4d,e, respectively, where the raw output is represented in units of normalization constant:  $P_{\max} \equiv \max_{m,n} \Delta P_{m,n}(\mathbf{e}_m, \mathbf{e}_n)$  and  $\mathbf{e}_i$  is a one-hot vector with index  $i$ . In this way, we define the multiplication output “1” as a maximum physical energy (power) difference achievable over the array with corresponding pulse injections.

While this simple normalization already shows a quite good accuracy, quantified by the mean absolute error (MAE) of 0.017 with respect to the ground truth table, the element-wise differential normalization by  $\Delta P_{m,n}(\mathbf{e}_m, \mathbf{e}_n)$  provides a more accurate output as displayed in Figure 4f, with reduced MAE = 0.0097 (see Figure S9 for the normalization powers). This can be realized by a tailored filter between the waveguide and the detection planes or the slight adjustment of the detection time windows. Lastly, the multiplication result  $\Delta P(\mathbf{e}_2, \mathbf{e}_3)$  normalized by  $\Delta P_{m,n}(\mathbf{e}_m, \mathbf{e}_n)$  is shown in Figure 4g, showcasing the fair errors in relation to the “multiplication by 0,” despite the spread of the error signals along the waveguides due to the reflections from submodules. In addition, we also evaluate the average error of the device with a set of  $10^4$  input samples, resulting in the reduction of MAE from 0.0212 to 0.0185, for simple and differential normalization techniques as described in the above (see Figure S10 for details).

## 2.5 | Matrix-Matrix Multiplication

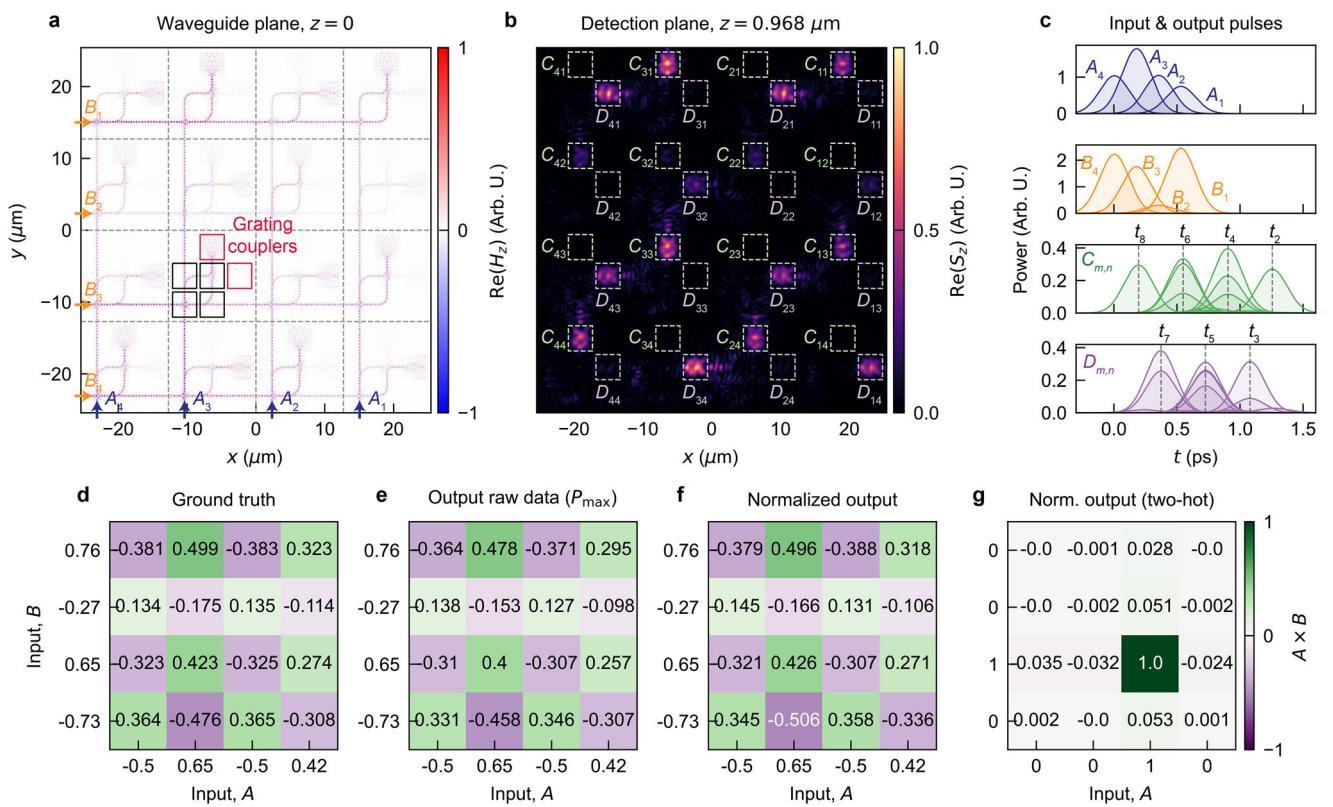
Building upon the vector outer product operation, we extend the concept to realize a full MMM using a  $2 \times 2$  PSA. While the outer product scheme verifies the fundamental spatial operation of the array, performing MMM requires precise spatiotemporal coordination among multiple pulse trains, which represent the columns of the input matrices. In particular, the pulse propagation time across a single MAC unit must match the propagation delay between adjacent trains within the array, ensuring two pulses encounter at the exact locations for MAC operations. In detail, the field injection at each input port ( $A_m$  and  $B_n$ ) is modulated with Gaussian pulse trains as:

$$a_m(t) = e^{-i2\pi f_0 t} \sum_{k=1}^K A_{k,m} \exp \left[ -\frac{1}{2} \left( \frac{t-t_{\text{delay}}(M+k-m)}{T} \right)^2 \right] \quad (5)$$

$$b_n(t) = e^{-i2\pi f_0 t} \sum_{k=1}^K iB_{k,n} \exp \left[ -\frac{1}{2} \left( \frac{t-t_{\text{delay}}(N+k-n)}{T} \right)^2 \right] \quad (6)$$

where  $f_0$ ,  $T \equiv T_{\text{FWHM}}/(2\log 2)$  and  $t_{\text{delay}}$  denote the center frequency, pulse width, and pulse interval, respectively, and  $M, N$ , and  $K$  represent the dimension of matrices  $\mathbf{A} \in \mathbb{R}^{K \times M}$  and  $\mathbf{B} \in \mathbb{R}^{K \times N}$ .

Figure 5 illustrates the full-wave simulation of the  $2 \times 2$  PSA performing MMM between two matrices  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{K \times 2}$  ( $K = 10$ ). Similar to the previous section, the PSA receives signals from four ports, two from the bottom and two from the left, each carrying column-wise information of  $\mathbf{A}$  and  $\mathbf{B}$ . Figure 5a visualizes the



**FIGURE 4** | Full-wave simulation of a vector outer product using  $4 \times 4$  PSA. (a,b) Example single-frequency field distribution along the device with the input ports ( $A_m$  and  $B_n$  with blue and yellow arrows, a) and the corresponding power emission through the output ports at the detection plane ( $C_{m,n}$  and  $D_{m,n}$  with green and purple boxes, b). Black and red boxes represent the four submodules of the MAC unit and additional grating couplers, respectively. The lattice period of the array is  $12.7 \mu\text{m}$ . (c) Input and output pulse envelopes encoded with different heights for the squared signal amplitude.  $t_{m+n} (t_8 < t_7 < \dots < t_2)$  indicate the pulse arrival time at each port  $C, D_{m,n}$ . (d–g) Outer product results: ground truth values for given inputs in (a) (d), output raw data in units of a power constant  $P_{\max}$  defined in the main text (e), and element-wise normalized output for error correction (f); and the element-wise normalized output for inputs  $\mathbf{A} = [0, 1, 0, 0]$  and  $\mathbf{B} = [0, 0, 1, 0]$  (g).

main signal paths (solid arrows) that define the intended data flow from input to designated output ports. Notably, minor leakage signals (dashed arrows) originate from the imperfection with numerical errors and are also likely expected in practical implementation. Nevertheless, their impact on other detectors outside the main path is negligible. This is because their propagation direction is opposite to the designed directionality of branch modules, preventing them from entering into grating couplers. The lattice period here is  $25.4 \mu\text{m}$ , which is twice that used in the previous  $4 \times 4$  PSA.

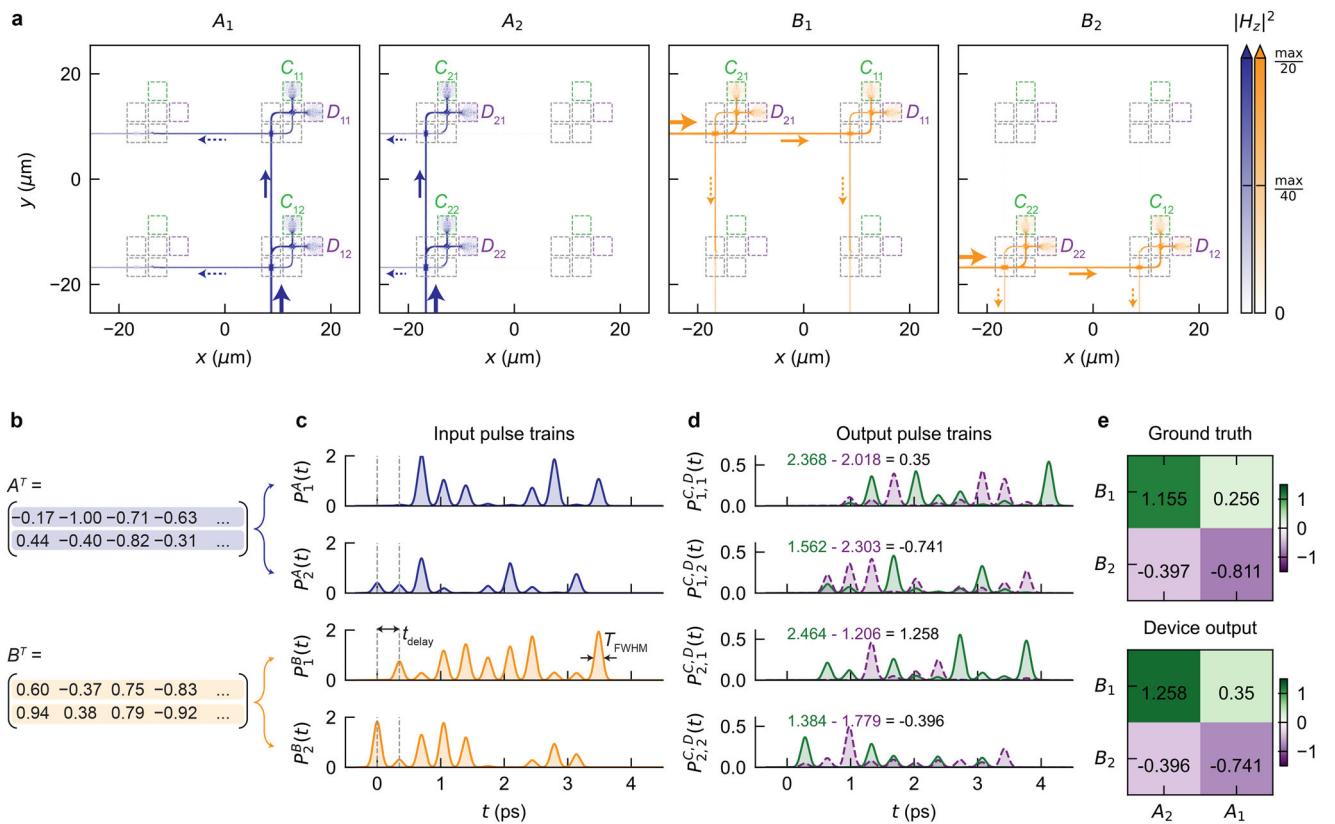
Figure 5b,c shows the temporal encoding of the input matrices as amplitude-modulated pulse trains based on Equations (5) and (6), where each matrix element is represented by the pulse amplitude proportional to the square root of the corresponding power envelope height. The resulting output pulse trains, detected at arrays  $C$  (green) and  $D$  (purple) and shown in Figure 5d, exhibit areas under envelope whose differences correspond to the computed matrix elements of output matrix, up to normalization factors. Physically, this process is realized by accumulating the photogenerated electrons over time at each detection area. Finally, Figure 5e compares the ground-truth result and the simulated device output for  $(\mathbf{A}^T \mathbf{B}) \in \mathbb{R}^{2 \times 2}$ , confirming that the PSA actually performs optical matrix–matrix multiplication through distributed pulse interactions in space and

time. See Note S3 for the analysis on device output error with respect to numerical precisions.

### 3 | Discussion

We have demonstrated the single-pulse operation of a MAC unit, its extension to  $4 \times 4$  array for vector outer product and  $2 \times 2$  array for MMM, respectively. It is noted that our system is linear time-invariant without time-varying components, in terms of the relationship between input and output fields. This property enables a straightforward generalization to indefinite length of pulse trains for the computation of inner products between time signals and matrix–matrix multiplications between two spatio-temporally encoded signals. Consequently, we achieved a single MAC operation within an area of  $(25.4 \mu\text{m})^2$  at a data rate of  $t_{\text{delay}}^{-1} = 2.87 \text{ THz}$ , given that pulse trains can be streamed every  $t_{\text{delay}}$  without overlap, as depicted in Figure 5c. This corresponds to a theoretical computation density of  $4.4 \text{ PMACs/mm}^2/\text{s}$ .

It is noted that the THz-scale rate is limited by the current operation speed of imaging devices and electro-optic modulators, typically constrained to the MHz to a few tens of GHz range. From a practical view point regarding pulse generation, a lowered modulation speed of 100 GHz would reduce the computation



**FIGURE 5** | Full-wave simulation of a matrix-matrix multiplication using  $2 \times 2$  PSA. (a) Magnetic field intensity for single-frequency injection at each of ports  $A_1$ ,  $A_2$  (from bottom side),  $B_1$ , and  $B_2$  (from left side). Solid and dashed arrows indicate the designed signal path and undesired signal leakage, respectively. The lattice period of the array is  $25.4 \mu\text{m}$ . (b) Example matrices  $\mathbf{A}$  and  $\mathbf{B} \in \mathbb{R}^{10 \times 2}$  to be processed. (c) Input pulse trains represented by power envelopes. Each column of matrices is encoded as pulse amplitudes with a width  $T_{\text{FWHM}} = 0.137 \text{ ps}$  and interval  $t_{\text{delay}} = 0.35 \text{ ps}$ . (d) Output pulse trains emitted through detector arrays  $C$  and  $D$  (green and purple regions in a). The difference in the integrated areas of the two envelopes corresponds to each element of device output matrix. (e) ground truth (top) and simulated device output (bottom) results for matrix product  $\mathbf{A}^T \mathbf{B} \in \mathbb{R}^{2 \times 2}$ .

density to about 155 TMACs/mm<sup>2</sup>/s. On the detection and imaging side, however, the speed is not a strict limitation because detection occurs once every pulse train accumulated over time. Therefore, with sufficiently long pulse trains, the device remains suitable and can benefit from reduced photon shot noise and other device-level imperfections (Figure 4f). This implies the particular advantage of the device in data processing involving very large arrays or matrices, which will further benefit from future innovations in optical modulation technologies.

We also anticipate that the computation density could be further enhanced through wavelength-division multiplexing (WDM) [22, 50]. In our context, wavelength division can be employed to perform batch computation of multiple MMM operations in parallel, using pulse trains with distinct center frequencies. Implementing such an approach would require sufficiently long pulse widths in time domain to ensure the narrow spectral bandwidth, avoiding overlap in frequency domain.

In summary, we employed the adjoint method to inverse-design the submodules (waveguide crossings, branches, and beam splitters) in a free-form shape, arranging them to constitute the photonic systolic array. The adjoint method allowed for precise targeting of both the amplitudes and phases of the

scattering parameters while minimizing the device footprint, which is a key factor for ensuring accurate operation and high computation density. By using finite-difference time-domain simulations, we demonstrated single-unit operation in the time domain and its extension to spatial arrays, achieving matrix-matrix multiplication. Our design enables multiplication between two purely optical signals, in contrast to many existing photonic matrix-matrix or matrix-vector multipliers that operate between an optical signal and electronic modulation such as MZI modulator. Consequently, our approach will significantly advance the energy-efficient machine learning acceleration, benefiting not only inference but also training, especially when dealing with models involving extremely large data sizes.

## Acknowledgements

This work was supported by the Army Research Office through a Multidisciplinary University Research Initiative program (Grant No. W911NF-22-2-0111).

## Conflict of Interest

Z.Y. has financial interest in Flexcompute Inc., which develops the software Tidy3D used in this work.

## Data Availability Statement

All code used in this study is available at GitHub: <https://github.com/jmkim93/photonic-systolic-array>. The notebook for Figure 5 is also available at Tidy3D Library: <https://www.flexcompute.com/tidy3d/community/notebooks/OpticalMatrixMultiplication>.

## References

1. W. X. Zhao, K. Zhou, J. Li, et al., “A Survey of Large Language Models,” 2024. <https://arxiv.org/abs/2303.18223>
2. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, “Large Language Models in Medicine,” *Nature Medicine* 29 (2023): 1930–1940.
3. D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes, “Autonomous chemical research with large language models,” *Nature* 624 (2023): 570–578.
4. A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30 (Curran Associates, Inc., 2017).
5. H. T. Kung and C. E. Leiserson, “Systolic arrays (for VLSI),” in *Sparse Matrix Proceedings 1978*, vol. 1, (Society for Industrial and Applied Mathematics Philadelphia, 1979), pp. 256–282.
6. N. P. Jouppi, C. Young, N. Patil, et al., “In-datacenter performance analysis of a tensor processing unit,” 2017. <https://arxiv.org/abs/1704.04760>
7. X. He, S. Pal, A. Amarnath, et al., “Sparse-tpu: adapting systolic arrays for sparse matrices,” in *Proceedings of the 34th ACM International Conference on Supercomputing*, ser. ICS ’20, (Association for Computing Machinery, 2020). <https://doi.org/10.1145/3392717.3392751>.
8. R. Xu, S. Ma, Y. Guo, and D. Li, “A survey of design and optimization for systolic array-based dnn accelerators,” *ACM Computing Surveys* 56 (2023): 20.
9. W. Ye, X. Zhou, J. Zhou, C. Chen, and K. Li, “Accelerating attention mechanism on fpgas based on efficient reconfigurable systolic array,” *ACM Transactions on Embedded Computing Systems* 22 (2023): 93.
10. J. Si, P. Zhang, C. Zhao, et al., “A carbon-nanotube-based tensor processing unit,” *Nature Electronics* 7 (2024): 684–693.
11. A. Vahdat, “Announcing Trillium, the sixth generation of Google Cloud TPU,” Accessed October 2024, <https://cloud.google.com/blog/products/compute/introducing-trillium-6th-gen-tpus>.
12. P. L. McMahon, “The physics of optical computing,” *Nature Reviews Physics* 5 (2023): 717–734.
13. X. Lin, Y. Rivenson, N. T. Yardimci, et al., “All-optical machine learning using diffractive deep neural networks,” *Science* 361 (2018): 1004–1008.
14. R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, “Large-scale optical neural networks based on photoelectric multiplication,” *Physical Review X* 9 (2019): 021032.
15. B. J. Shastri, A. N. Tait, T. Ferreira de Lima, et al., “Photonics for artificial intelligence and neuromorphic computing,” *Nature Photonics* 15 (2021): 102–114.
16. H. Zhou, J. Dong, J. Cheng, et al., “Photonic matrix multiplication lights up photonic accelerator and beyond,” *Light: Science & Applications* 11 (2022): 30.
17. T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon, “An optical neural network using less than 1 photon per multiplication,” *Nature Communications* 13 (2022): 231.
18. F. Ashtiani, A. J. Geers, and F. Aflatouni, “An on-chip photonic deep neural network for image classification,” *Nature* 606 (2022): 501–506.
19. Z. Chen, A. Sludds, R. Davis, et al., “Deep learning with coherent vcsel neural networks,” *Nature Photonics* 17 (2023): 723–730.
20. Y. Huo, H. Bao, Y. Peng, et al., “Optical neural network via loose neuron array and functional learning,” *Nature Communications* 14 (2023): 2535.
21. L. He, D. Liu, J. Gao, et al., “Super-compact universal quantum logic gates with inverse-designed elements,” *Science Advances* 9 (2023): eadg6685.
22. S. Ou, K. Xue, L. Zhou, et al., “Hypermultiplexed integrated photonics-based optical tensor processor,” *Science Advances* 11 (2025): eadu0228.
23. J. Feldmann, N. Youngblood, M. Karpov, et al., “Parallel convolutional processing using an integrated photonic tensor core,” *Nature*, 589 (2021): 52–58.
24. S. Xu, J. Wang, S. Yi, and W. Zou, “High-order tensor flow processing using integrated photonic circuits,” *Nature Communications* 13 (2022): 7970.
25. S. Aggarwal, B. Dong, J. Feldmann, N. Farmakidis, W. H. P. Pernice, and H. Bhaskaran, “Reduced rank photonic computing accelerator,” *Optica* 10 (2023): 1074–1080.
26. B. Dong, S. Aggarwal, W. Zhou, et al., “Higher-dimensional processing using a photonic tensor core with continuous-time data,” *Nature Photonics* 17 (2023): 1080–1088.
27. B. Dong, F. Brückerhoff-Plückelmann, L. Meyer, et al., “Partial coherence enhances parallelized photonic computing,” *Nature* 632 (2024): 55–62.
28. M. Moralis-Pegios, G. Giannouliannis, A. Tsakyridis, D. Lazovsky, and N. Pleros, “Perfect linear optics using silicon photonics,” *Nature Communications* 15 (2024): 5468.
29. M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, “Experimental realization of any discrete unitary operator,” *Physical Review Letters* 73 (1994): 58–61.
30. J. Carolan, C. Harrold, C. Sparrow, et al., “Universal linear optics,” *Science* 349 (2015): 711–716.
31. W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, “Optimal design for universal multiport interferometers,” *Optica* 3 (2016): 1460–1465.
32. Y. Shen, N. C. Harris, S. Skirlo, et al., “Deep learning with coherent nanophotonic circuits,” *Nature Photonics* 11 (2017): 441–446.
33. D. Pérez, I. Gasulla, L. Crudgington, et al., “Multipurpose silicon photonics signal processor core,” *Nature Communications* 8 (2017): 636.
34. T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, “Training of photonic neural networks through in situ backpropagation and gradient measurement,” *Optica* 5 (2018): 864–871.
35. E. Khoram, A. Chen, D. Liu, et al., “Nanophotonic media for artificial neural inference,” *Photonics Research* 7 (2019): 823–827.
36. M. Camacho, B. Edwards, and N. Engheta, “A single inverse-designed photonic structure that performs parallel computing,” *Nature Communications* 12 (2021): 1466.
37. V. Nikkhah, A. Pirmoradi, F. Ashtiani, B. Edwards, F. Aflatouni, and N. Engheta, “Inverse-designed low-index-contrast structures on a silicon photonics platform for vector–matrix multiplication,” *Nature Photonics* 18 (2024): 501–508.
38. N. Youngblood, “Coherent photonic crossbar arrays for large-scale matrix–matrix multiplication,” *IEEE Journal of Selected Topics in Quantum Electronics* 29 (2023): 1–11.
39. S. R. Kari, N. A. Nobile, D. Pantin, V. Shah, and N. Youngblood, “Realization of an integrated coherent photonic platform for scalable matrix operations,” *Optica* 11 (2024): 542–551.

40. C. M. Lalau-Keraly, S. Bhargava, O. D. Miller, and E. Yablonovitch, “Adjoint shape optimization applied to electromagnetic design,” *Optics Express* 21 (2013): 21693–21701.

41. T. W. Hughes, M. Minkov, I. A. D. Williamson, and S. Fan, “Adjoint method and inverse design for nonlinear nanophotonic devices,” *ACS Photonics* 5 (2018): 4781–4787.

42. A. M. Hammond, J. B. Slaby, M. J. Probst, and S. E. Ralph, “Multi-layer inverse design of vertical grating couplers for high-density, commercial foundry interconnects,” *Optics Express* 30 (2022): 31058–31072.

43. A. M. Hammond, J. B. Slaby, M. J. Probst, and S. E. Ralph, “Phase-injected topology optimization for scalable and interferometrically robust photonic integrated circuits,” *ACS Photonics* 10 (2023): 808–814.

44. T. Wu, M. Menarini, Z. Gao, and L. Feng, “Lithography-free reconfigurable integrated photonic processor,” *Nature Photonics* 17 (2023): 710–716.

45. FlexCompute, “Python-driven FDTD software: Tidy3D.” Accessed November 2025, <https://www.flexcompute.com/tidy3d/>.

46. M. Minkov, P. Sun, B. Lee, Z. Yu, and S. Fan, “Gpu-accelerated photonic simulations,” *Optics & Photonics News* 35 (2024): 44–50.

47. R. Fano, “Theoretical limitations on the broadband matching of arbitrary impedances,” *Journal of the Franklin Institute* 249 (1950): 57–83.

48. M. Tchernycheva, A. Messanvi, A. de Luna Bugallo, et al., “Integrated photonic platform based on ingan/gan nanowire emitters and detectors,” *Nano Letters* 14 (2014): 3515–3520.

49. X. Gan, R.-J. Shieue, Y. Gao, et al., “Chip-integrated ultrafast graphene photodetector with high responsivity,” *Nature Photonics* 7 (2013): 883–887.

50. X. Yu, Z. Wei, F. Sha, et al., “Parallel optical computing capable of 100-wavelength multiplexing,” *eLight* 5 (2025): 10.

51. Y. Ma, J. Kaczynski, C. Ranacher, et al., “Nano-porous aluminum oxide membrane as filtration interface for optical gas sensor packaging,” *Microelectronic Engineering* 198 (2018): 29–34.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section.

**Supporting File:** lpor70651-sup-0001-SuppMat.docx.