



- | | |
|--|---|
| (51) International Patent Classification:
<i>G06E 3/00</i> (2006.01) <i>G06N 3/067</i> (2006.01) | (74) Agent: GANGEMI, Anthony, P.; Barclay Damon LLP,
545 Long Wharf Drive, 9th Floor, New Haven, CT 06511
(US). |
| (21) International Application Number:
PCT/US2024/053575 | (81) Designated States (<i>unless otherwise indicated, for every kind of national protection available</i>): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW. |
| (22) International Filing Date:
30 October 2024 (30.10.2024) | (84) Designated States (<i>unless otherwise indicated, for every kind of regional protection available</i>): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, |
| (25) Filing Language: English | |
| (26) Publication Language: English | |
| (30) Priority Data:
63/546,305 30 October 2023 (30.10.2023) US
63/713,086 29 October 2024 (29.10.2024) US | |
| (71) Applicant: RENSELAER POLYTECHNIC INSTITUTE [US/US]; 110 8th Street, J Building, Troy, NY 12180 (US). | |
| (72) Inventor: HUANG, Zhaoran; 1 Bernini Drive, Clifton Park, NY 12065 (US). | |

(54) Title: PHOTONIC TENSOR CORE DEVICES AND SYSTEMS

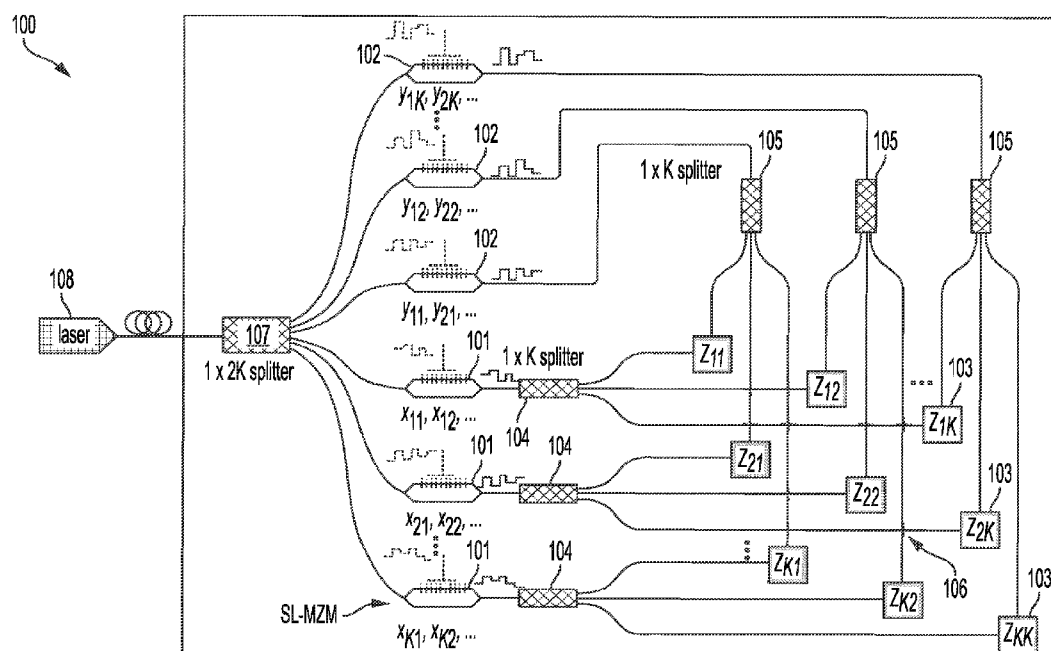


FIG. 1

(S7) Abstract: Photonic tensor cores that utilize optical modulators to perform matrix multiplication. Optical signals are encoded with matrix values via electrical signals applied by the optical modulators. Dot product engines combine the encoded optical signals and provide product photocurrent signals for conversion to digital electrical signals.

SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN,
GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

PHOTONIC TENSOR CORE DEVICES AND SYSTEMS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 63/546,305 filed October 30, 2023 and U.S. Provisional Patent Application No. 63/713,086 filed October 29, 2024, both of which are incorporated by reference as if disclosed herein in their entireties.

BACKGROUND

[0002] Photonic computing has emerged as a promising technology for high-performance and energy-efficient computing, particularly in computation-intensive artificial intelligence (AI) tasks.

Integrated photonic tensor core (PTC) designs have demonstrated ultra-fast photonic analog linear operation acceleration. Coherent PTCs that leverage interference and diffraction include Mach-Zender interferometer (MZI) arrays, butterfly-style meshes, auto-designed photonic circuits, coupler-crossbar array, star-coupler-based design, and metalens-based diffractive PTCs. In addition, to leverage the wavelength-division multiplexing (WDM) technique, there are incoherent multi-wavelength PTCs, including: microring (MRR) weight banks, Phase-change material (PCM) crossbar arrays, and micro-comb-based computing engines.

[0003] The efficiency of PTCs for use with general edge AI can be considered from the perspectives of versatility, dynamic reprogrammability, and domain-specific customization. Versatility, or universality, is one of the important features of photonic AI hardware to accelerate a variety of deep neural network (DNN) workloads. A versatile/generic photonic accelerator based on universal optical linear units is capable of realizing general matrix multiplication (GEMM) and thus directly implementing a wide spectrum of pretrained digital DNNs. Many specialized linear units are not applicable to generic tensor computation since they restrict their matrix expressivity to a subspace of specialized matrices for higher hardware efficiency.

[0004] Real-time, efficient input tensor encoding with low reconfiguration costs (*i.e.*, reprogrammability) is also advantageous for photonic computing. High weight encoding costs due to the high complexity of matrix decomposition required to encode weights can restrict photonic computing designs to only support weight-static linear operations, e.g., fully connected (FC) layers and convolutional (CONV) layers, where weights are pretrained and pre-encoded into the device/circuit transmissions. However, advanced AI models, based on attention operations where both matrix multiplication operands are dynamic, full-range, and general tensors, cannot be efficiently mapped to weight-static PTCs.

[0005] Domain-specific hardware customization is a third advantageous feature for efficient, scalable PTCs. At the device level, many optical computing hardware demonstrations are based on standard foundry process design kit (PDK) elements, which are designed for optical communications and not optimized for analog neuromorphic computing. At the circuit level, customization is important for reducing the long-lasting analog-to-digital and optical-to-electrical conversion bottlenecks. At the architecture level, due to the lack of optical memory, the large spatial footprint of photonic circuits, and the high digital memory access cost, the architecture topology and dataflow also need to be customized to fully leverage the temporal locality to reduce data movement cost and maximize hardware sharing.

[0006] What are needed, therefore, are versatile, reprogrammable, and customizable photonic computing designs to help process advanced AI tasks.

SUMMARY

[0007] Aspects of the present technology are directed to photonic tensor core devices for performing matrix multiplication. In some embodiments, the devices include sets of optical modulators for receiving optical signals and encoding matrix values onto the optical signals. In some embodiments, a set of dot product engines combines the encoded optical signals corresponding to the matrix values to provide product photocurrent signals, which are then converted to digital electric signals. In some embodiments, the optical modulators comprise Mach-Zehnder modulators that encode each optical signal via a temporal electrical signal provided to each modulator.

[0008] Other aspects of the present technology are directed to computing systems for performing matrix multiplication that comprise a light source, one or more photonic tensor core devices, and encoding circuitry configured to generate electrical signals corresponding to each matrix value to send to optical modulators for encoding onto optical signals.

[0009] Additional aspects and features of the present technology will be apparent to those of skill in the art upon consideration of the following description and the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] The drawings show embodiments of the disclosed subject matter for the purpose of illustrating the technology. However, it should be understood that the present application is not limited to the precise arrangements and instrumentalities shown in the drawings, wherein:

[0011] FIG. 1 shows a schematic view of an embodiment of a photonic tensor core device according to the present technology;

[0012] FIG. 2 shows a schematic view of an alternative embodiment of a photonic tensor core device according to the present technology;

[0013] FIG. 3 shows a schematic view of an embodiment of an optical dot-product engine according to the present technology;

5 [0014] FIG. 4 shows a schematic view of an embodiment of a temporal integrator according to the present technology;

[0015] FIG. 5 shows a schematic view of an embodiment of a computing system according to the present technology.

10 [0016] FIG. 6A shows a schematic top-view of a modulator according to an embodiment of the present technology;

[0017] FIG. 6B shows a schematic cross-section view of the modulator of FIG. 6A.

DETAILED DESCRIPTION

[0018] The following discussion relates to various embodiments of photonic sensor core systems and devices. It will be understood that the herein described versions are examples that embody
15 certain inventive concepts as detailed herein. To that end, other variations and modifications will be readily apparent to those of sufficient skill.

[0019] In some embodiments of the present technology, a time-multiplexed dynamic photonic tensor accelerator design for AI acceleration is provided. Some embodiments include ultra-compact, slow-light electro-optic modulators for input operand encoding. Some embodiments
20 include hierarchical partial product accumulation with lightweight capacitive temporal integration modules. Some embodiments include multi-core architecture to maximize sharing of data input/readout circuitry. In some embodiments, slow-light MZI modulators (SL-MZM) with enhanced light-matter interaction for size and power reduction are used. In some embodiments, the SL-MZM phase shifter length is from about 150 μm to about 200 μm with a footprint about 10
25 greater than Si MRR while an order of magnitude smaller than a typical foundry offered Si Mach-Zehnder modulator (MZM) PDK element. In some embodiments, this SL-MZM is thermally robust, with no thermal tuning/locking circuit needed, and can also tolerate large manufacturing variations. In contrast to a multiwavelength dynamic PTC design, some embodiments of the present technology simplify the spectral multi-wavelength encoding by employing high-speed
30 temporal encoding, eliminating the need for complex dispersion-engineered broadband device designs such as Si modulators, optical power splitters and directional couplers as well as remove WDM MUX/DEMUX overhead.

[0020] The present technology relates to matrix multiplication, which is an important linear operation for various information processing workloads. Some embodiments of the present technology perform matrix-matrix multiplication. Consider two input matrices, matrix X with $M \times N$ dimension and matrix Y with $N \times Q$ dimension:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{MN} \end{bmatrix}, \quad Y = \begin{bmatrix} y_{11} & \cdots & y_{1Q} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NQ} \end{bmatrix}. \quad (1)$$

5

[0021] The matrix multiplication $Z = X \cdot Y$ is:

$$\begin{aligned} Z &= \begin{bmatrix} z_{11} & \cdots & z_{1Q} \\ \vdots & \ddots & \vdots \\ z_{M1} & \cdots & z_{MQ} \end{bmatrix} \\ &= \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{MN} \end{bmatrix} \cdot \begin{bmatrix} y_{11} & \cdots & y_{1Q} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NQ} \end{bmatrix}. \end{aligned} \quad (2)$$

[0022] The resulting Z is an $M \times Q$ matrix; and its a th row, b th column element z_{ab} is obtained by calculating the dot-product of the a th row vector of X and b th column vector of Y , i.e.,

$$z_{ab} = X_a \cdot Y_b = [x_{a1} \quad \cdots \quad x_{aN}] \cdot \begin{bmatrix} y_{1b} \\ \vdots \\ y_{Nb} \end{bmatrix}. \quad (3)$$

10

[0023] Each vector dot-product operation can be mapped to a dynamic dot-product engine, explained below in reference to FIG. 3. Multiple dot-product engines can form an array structure, i.e., a tensor core, to realize parallel matrix-matrix multiplication.

[0024] Referring now to FIG. 1, a first embodiment of a photonic tensor core device 100 for multiplying a $K \times N$ portion of a first matrix X , comprising M rows and N columns, by a $N \times K$ portion of a second matrix Y , comprising N rows and Q columns, is shown schematically. In this

embodiment, the device comprises a first set 101 of K optical modulators each configured to receive an optical signal and to encode a matrix value $x_{11} \dots x_{KN}$ from matrix X onto the optical signal. The device further comprises a second set 102 of K optical modulators each configured to receive an optical signal and to encode a matrix value $y_{11} \dots y_{NK}$ from matrix Y onto the optical signal. The optical signals are provided by a light source, such as a laser, as described below.

[0025] A set 103 of $K \times K$ dot product engines, each coupled to one of the first set of modulators 101 and to one of the second set of modulators 102 is also included in this embodiment. Each dot product engine 103 is configured to combine the encoded optical signals corresponding to each of the matrix values $x_{11} \dots x_{KN}$ and $y_{11} \dots y_{NK}$ and to provide product photocurrent signals corresponding to elements $z_{11} \dots z_{KK}$ of a result matrix Z . The result matrix is the result of the multiplication of the portions of the X and Y matrices. In this embodiment, each matrix value is encoded to each optical signal via a temporal electrical signal corresponding to each matrix value provided to each modulator.

[0026] In some embodiments, the first and second sets of optical modulators 101, 102 comprise Mach-Zehnder modulators. In some embodiments, the modulators comprise slow-light Mach-Zehnder modulators. In some embodiments, Silicon-based (Si) modulators are used that utilize the carrier plasma effect for on-chip PTC. In the embodiment shown, which achieves a dot-product operation for matrices of the size of $K \times K$ requires $2K$ modulators for signal conversion. In some embodiments, the physical dimension of the Si modulators is a design parameter that impacts the scalability of matrix operation.

[0027] In some embodiments, digital electrical signals carrying the matrix information are encoded onto the optical signals represented by the amplitude and phase by the optical modulators. For example, let E_{in} be the electric field of the optical signal to the optical modulator, then the electric field of the modulator output can be expressed as $E_{in} \cos \theta$, allowing broadband mapping of both positive and negative values.

[0028] In some embodiments, a 1D dielectric photonic crystal waveguide, specifically a rectangular-shaped Bragg grating, is used as the slow-light-enabled compact modulator. In some such embodiments, the footprint of the modulator array is significantly reduced. In some embodiments, an Si slow-light MZM (SL-MZM) with a phase shifter length (LPS) of $150 \mu\text{m}$ is used. In some embodiments, the phase shifter length is in the range from about $150 \mu\text{m}$ to about $200 \mu\text{m}$. In some embodiments, the SL-MZM is fabricated under a multi-project wafer (MPW) run, for complete foundry compatibility.

[0029] FIG. 6a shows a schematic top-view diagram of an embodiment of a modulator 626. The modulator 626 comprises a Y-splitter 632 with even power splitting between a grating arm 627

and a reference arm 628 followed by a Y-combiner 633 to recombine the split signals. In this embodiment, the arms include heaters 634 to help compensate for asymmetries. FIG. 6B shows a schematic cross section view of the modulator 626. In this embodiment, the gratings 629 have a width (W_g) of 1.5 μm and a period of 282 nm, and $H_f = 220$ nm, $W_c = 400$ nm, and $H_s = 110$ nm.

To encode the input optical signal with a matrix value, voltages are applied through the vias 631.

[0030] Reflection occurring at different junctions within the modulator device, optical absorption due to carriers in waveguides, propagation loss in the Bragg grating phase shifter due to increased group indices, and mode mismatch at the Bragg grating waveguide interfaces are the primary factors contributing to the modulator insertion loss. In one embodiment, the measured total modulator insertion loss is ~ 6.4 dB for $L_{PS} = 150$ μm .

[0031] Both the electrical bandwidth and the linearity of a Si modulator help determine the high-bit resolution at a high computing clock frequency. Operating under reverse bias, the speed of a Si SL-MZM is limited by its RC time constant and photon lifetime. In some embodiments, the PN junctions are doped at an elevated level (ranging from 10^{18} cm^{-3} to 10^{19} cm^{-3}) to enhance the carrier plasma effect. As the phase shifter length is reduced in an SL-MZM, the total capacitance decreases. In some embodiments, the measured SL-MZM junction capacitance C_j was approximately ~ 0.75 pF. Depending on the doping level in the connecting Si bar from the ridge waveguide to the via contacts, the intrinsic resistance of a SL-MZM according to some embodiments ranges from 5 to 10 Ω . The estimated RC time-limited electrical bandwidth of a SL-MZM is thus in the hundreds of GHz in some embodiments. The slow-light effect can be viewed as a traveling wave resonant in its propagation direction, with the optical bandwidth determined by the Q-factor of the resonator. For the rectangular Bragg grating-shaped slow-light, an optical bandwidth of approximately ~ 26 GHz is estimated for some embodiments. In some embodiments, dispersion engineering techniques such as phase-shifted Bragg grating, dispersion compensation, and line-shift photonic crystal waveguide are used to reduce the dispersion-induced bandwidth penalty.

[0032] The embodiment of a device 100 in FIG. 1 further comprises a first set 104 of $K \times K$ optical power splitters for splitting the encoded signal corresponding to the matrix values $x_{11} \dots x_{KN}$ each coupled to one of the first set 101 of K optical modulators and coupled to K dot product engines 103. This embodiment of the device 100 further comprises a second set 105 of $K \times K$ optical power splitters for splitting the encoded signal corresponding to the matrix values $y_{11} \dots y_{NK}$ each coupled to one of the second set 102 of K optical modulators and coupled to K dot product engines 103. As shown in FIG. 1, the design in this embodiment includes $(K - 1)^2$ waveguide crossings 106.

[0033] The embodiment of FIG. 1 further comprises a multimode interferometer 107, which is coupled to each of the optical modulators 101 and 102 for providing the optical signal. The multimode interferometer 107 acts as a $1 \times 2K$ splitter to divide the incoming light signal into $2K$ optical signals to be processed by the modulators 101, 102. Accordingly, the multimode interferometer 107 is configured to be coupled to a light source 109.

[0034] Some embodiments of PTCs according to the present technology that utilize optical wave phase and amplitude in time-domain processing use a monochromatic light source for optical signal processing. Some embodiments utilize o-band operation instead of c-band components. O-band operation offers several advantages such as a smaller optical mode volume in Si/SiO₂ waveguide structure, higher mode confinement with tighter bending radius and >1.5 higher in Ge photodetector responsivity.

[0035] In some embodiments, such as the embodiment in FIG. 1, a laser module that is disposed “off-chip” is used. That is, the light source is separate from the device 100. In other embodiments, an on-chip laser diode is used as the light source. In some embodiments, an on-chip III-V integrated laser diode is used. In some embodiments, a high-power, monolithic o-band laser is used, having output power of about 150 mW. In some embodiments, a moderate laser power of 100 mW is used.

[0036] FIG. 2 shows an alternative embodiment of a photonic tensor core device 200. Similar to the device 100, device 200 includes a first set 201 of K optical modulators, a second set 202 of K of optical modulators, and a set 203 of $K \times K$ dot product engines. In this embodiment, however, the coupling arrangement between the optical modulators and the dot product engines is different. Device 200 comprises a plurality of first sets 204 of $K - 1$ optical power splitters, each set 204', 204'', 204''' is coupled in series to one of the first set 201 of K optical modulators and each coupled to a dot product engine. Device 200 also comprises a plurality of second sets 205 of $K - 1$ optical power splitters, each set 205', 205'', 205''' coupled in series to one of the second set 202 of K optical modulators and each coupled to a dot product engine. In this embodiment, the splitting ratios of the first and second sets 204 and 205 of optical power splitters are set at $1:(K - 1)$, $1:(K - 2)$, ..., $1:1$ along the series of splitters. This is a series of uneven splitters intended to reduce the number of waveguide crossings 206 as compared to the design of device 100.

[0037] FIG. 3 shows a schematic diagram of an optical dot-product engine 303 according to some embodiments of the present technology. The engine 303 comprises a 2×2 optical power splitter 309, comprising a first arm 310 and a second arm 311. In this embodiment, the first arm 310 is coupled to one of the first set (e.g., 101, 201, not shown in FIG. 3) of K modulators and the second arm 311 is coupled to one of the second set (e.g., 102, 202, not shown in FIG. 3) of K modulators. The engine 303 further comprises a $\pi/2$ phase shifter 312 coupled to the second arm 311 of the

optical power splitter; a first photodiode 313 coupled to the first arm 310 of the optical power splitter; and a second photodiode 314 coupled to the second arm 311 of the optical power splitter. In this embodiment, the first and second photodiodes are configured to convert each encoded optical signal 315, 316 to a photocurrent signal which combine to form the product photocurrent signal 317.

[0038] Thus, in this embodiment, a 2×2 50:50 optical power splitter is used to generate interference between the optical signals from two input arms. The embodiment shown in FIG. 3 utilizes a directional coupler structure to generate the 50:50 power splitting, but, in other embodiments, multimode interferometers (MMI) are used.

[0039] In some embodiments, a temporal integrator 318 is coupled to each dot product engine 303 for converting the product photocurrent signal 317 to an integrated signal. A schematic diagram of a temporal integrator 418 is shown in FIG. 4. In some embodiments, the integrator 418 comprises at least two capacitors 419 and 420 connected in parallel to the dot product engine for receiving the product photocurrent signal. In this embodiment, the capacitors 419 and 420 are arranged as a flipped capacitor pair. In some embodiments, multiple flipped capacitor pairs are connected in parallel to achieve a symmetric circuit topology. In this embodiment, input 421 receives the product photocurrent signal. In some embodiments, the integrator 418 further comprises one or more field effect transistors connected in parallel to the capacitors. As shown, two n-channel FETs 422 and two p-channel FETs 423 are shown, connected in parallel with the capacitors. In some embodiments, ten 40 nm n-channel and ten 40 nm p-channel FETs are connected in parallel with the capacitors to help with current driving capability for reset within a single baud time period. In other embodiments, integrators with an op-amp and a capacitive feedback loop are used.

[0040] In some embodiments, the devices 100 and/or 200 further comprise at least one analog-to-digital converter associated with one or more of the temporal integrators for converting the integrated signal to a digital signal.

[0041] FIG. 5 shows a schematic diagram of a computing system 524 according to another embodiment of the present technology. The system 524 comprises a monochromatic light source 508. In some embodiments, the source 508 is a laser. The system also includes one or more photonic tensor core devices 500 for multiplying a $K \times N$ portion of a first matrix X , comprising M rows and N columns, by a $N \times K$ portion of a second matrix Y , comprising N rows and Q columns. Each device 500 includes an optical power splitter 507 that receives light from the monochromatic light source 508 and splits the light into $2K$ optical signals. Each device also comprises a first set 501 of K optical modulators each configured to receive an optical signal and to encode a matrix value $x_{11} \dots x_{KN}$ from matrix X onto the optical signal and a second set 502 of

K optical modulators each configured to receive an optical signal and to encode a matrix value $y_{11} \dots y_{NK}$ from matrix Y onto the optical signal.

[0042] Further, similar to the devices described above, the devices 500 include a set of $2K$ dot product engines, where each dot product engine is coupled to one of the first set of modulators and to one of the second set of modulators, and each dot product engine is configured to combine the optical signals corresponding to each of the matrix values $x_{11} \dots x_{KN}$ and $y_{11} \dots y_{NK}$ and provide a product photocurrent signal corresponding to elements $z_{11} \dots z_{KK}$ of a result matrix Z .

[0043] The system also includes an encoding circuitry 525 that is configured to generate an electrical signal corresponding to each matrix value to send to each of the first and second optical modulators for encoding each optical signal with each matrix value. As described above, digital electrical signals carrying the matrix information are converted to optical signals represented by the amplitude and phase by the optical modulators, which comprise, e.g., slow-light Mach-Zehnder modulators. For example, let E_{in} be the electric field of the optical signal to the optical modulator, then the electric field of the modulator output can be expressed as $E_{in} \cos \theta$, allowing broadband mapping of both positive and negative values.

[0044] In this embodiment, for each tile of PTC, a set of integrators 518 integrate the photocurrent signals and a trans-impedance amplifier (TIA) and analog-to-digital converter (ADC) converts them to electronic digital signals.

[0045] In the system 524 shown in FIG. 5, a total of R tiles of C photonic tensor core devices are included, and each PTC is a crossbar of $K \times K$ dot-product engines that can finish a $K \times 1$ times $1 \times K$ vector outer product at each time step, T . In this embodiment, then, the X matrix of size $M \times N$ is partitioned into M/K horizontal strips, each with a size $K \times N$, and the matrix Y of size $N \times Q$ is partitioned into Q/K vertical strips, each with a size of $N \times K$.

[0046] Each cycle, therefore, includes (1) feeding one vector into the PTC, (2) reading out the outer product results as photocurrent, (3) converting it to the electronic domain, and (4) accumulating partial product.

[0047] The parts, components, and structural elements described herein can be combined into an integral or unitary, one-piece objects, or such parts, components, and structural elements can be distinct, removable items that are attachable to each other.

[0048] In the foregoing description, certain components or elements may have been described as being configured to mate with each other. For example, an embodiment may be described as a first element (functioning as a male) configured to be inserted into a second element (functioning as a female). It should be appreciated that an alternate embodiment includes the first element (functioning as a female) configured to receive the second element (functioning as a male). In

either such embodiment, the first and second elements are configured to mate with, fit with or otherwise interlock with each other.

[0049] Although the technology has been described and illustrated with respect to embodiments thereof, it should be understood by those skilled in the art that the foregoing and various other
5 changes, omissions and additions may be made therein and thereto, without parting from the spirit and scope of the present technology.

CLAIMS

What is claimed is:

1. A photonic tensor core device for multiplying a $K \times N$ portion of a first matrix X , comprising M rows and N columns, by a $N \times K$ portion of a second matrix Y , comprising N rows and Q columns, the device comprising:
 - a first set of K optical modulators each configured to receive an optical signal and to encode a matrix value $x_{11} \dots x_{KN}$ from matrix X onto the optical signal;
 - a second set of K optical modulators each configured to receive an optical signal and to encode a matrix value $y_{11} \dots y_{NK}$ from matrix Y onto the optical signal; and
 - a set of $K \times K$ dot product engines, wherein each dot product engine is coupled to one of the first set of modulators and to one of the second set of modulators, and each dot product engine is configured to combine the encoded optical signals corresponding to each of the matrix values $x_{11} \dots x_{KN}$ and $y_{11} \dots y_{NK}$ and to provide product photocurrent signals corresponding to elements $z_{11} \dots z_{KK}$ of a result matrix Z ;
 - wherein each matrix value is encoded to each optical signal via a temporal electrical signal corresponding to each matrix value provided to each modulator.
2. The device of claim 1, wherein each optical modulator comprises a slow-light Mach-Zehnder modulator.
3. The device of claim 2, wherein each optical modulator has a phase shifter length in the range from about 150 μm to about 200 μm .
4. The device of claim 2, further comprising a 1D dielectric photonic crystal waveguide.
5. The device of claim 1, further comprising:
 - a first set of $K \times K$ optical power splitters for splitting the encoded signal corresponding to the matrix values $x_{11} \dots x_{KN}$ each coupled to one of the first set of K optical modulators and coupled to K dot product engines; and

a second set of $K \ 1 \times K$ optical power splitters for splitting the encoded signal corresponding to the matrix values $y_{11} \dots y_{NK}$ each coupled to one of the second set of K optical modulators and coupled to K dot product engines.

5 6. The device of claim 1, further comprising:

a plurality of first sets of $K - 1$ optical power splitters, each set coupled in series to one of the first set of K optical modulators and each coupled to a dot product engine; and

a plurality of second sets of $K - 1$ optical power splitters, each set coupled in series to one of the second set of K optical modulators and each coupled to a dot product engine;

10 wherein the splitting ratios of the first and second sets of optical power splitters are set at $1:(K - 1)$, $1:(K - 2)$, ..., $1:1$ along the series of splitters.

7. The device of claim 1, wherein each of the dot product engines comprises:

15 a 2×2 optical power splitter, comprising a first arm and a second arm, wherein the first arm is coupled to one of the first set of K modulators and the second arm is coupled to one of the second set of K modulators;

a $\pi/2$ phase shifter coupled to the second arm of the optical power splitter;

a first photodiode coupled to the first arm of the optical power splitter; and

a second photodiode coupled to the second arm of the optical power splitter;

20 wherein the first and second photodiodes are configured to convert each encoded optical signal to a photocurrent signal which combine to form a product photocurrent signal.

8. The device of claim 7, further comprising a temporal integrator coupled to each dot product engine for converting the product photocurrent signal to an integrated signal,
25 comprising:

at least two capacitors connected in parallel to the dot product engine for receiving the product photocurrent signal;

one or more field effect transistors connected in parallel to the capacitors.

9. The device of claim 1, further comprising a multimode interferometer coupled to each of the optical modulators for providing the optical signal and wherein the multimode interferometer is configured to be coupled to a light source.

5

10. The device of claim 8, further comprising an analog-to-digital converter associated with one or more of the temporal integrators for converting the integrated signal to a digital signal.

11. A computing system, comprising:

10 a monochromatic light source;

one or more photonic tensor core devices for multiplying a $K \times N$ portion of a first matrix X , comprising M rows and N columns, by a $N \times K$ portion of a second matrix Y , comprising N rows and Q columns, comprising:

15 an optical power splitter that receives light from the monochromatic light source and splits the light into $2K$ optical signals;

a first set of K optical modulators each configured to receive an optical signal and to encode a matrix value $x_{11} \dots x_{KN}$ from matrix X onto the optical signal;

a second set of K optical modulators each configured to receive an optical signal and to encode a matrix value $y_{11} \dots y_{NK}$ from matrix Y onto the optical signal; and

20 a set of $2K$ dot product engines, wherein each dot product engine is coupled to one of the first set of modulators and to one of the second set of modulators, and each dot product engine is configured to combine the optical signals corresponding to each of the matrix values $x_{11} \dots x_{KN}$ and $y_{11} \dots y_{NK}$ and provide a product photocurrent signal corresponding to elements $z_{11} \dots z_{KK}$ of a result matrix Z ;

25 an encoding circuitry configured to generate an electrical signal corresponding to each matrix value to send to each of the first and second optical modulators for encoding each optical signal with each matrix value.

12. The system of claim 11, wherein each optical modulator comprises a slow-light Mach-Zehnder modulator.

13. The device of claim 12, wherein each optical modulator has a phase shifter length in the range from about 150 μm to about 200 μm .

14. The device of claim 12, further comprising a 1D dielectric photonic crystal waveguide.

15. The system of claim 11, wherein the device further comprises:

10 a first set of $K \times K$ optical power splitters for splitting the encoded signal corresponding to the matrix values $x_{11} \dots x_{KN}$ each coupled to one of the first set of K optical modulators and coupled to K dot product engines; and

a second set of $K \times K$ optical power splitters for splitting the encoded signal corresponding to the matrix values $y_{11} \dots y_{NK}$ each coupled to one of the second set of K optical modulators and coupled to K dot product engines.

16. The system of claim 11, wherein the device further comprises:

a plurality of first sets of $K - 1$ optical power splitters, each set coupled in series to one of the first set of K optical modulators and each coupled to a dot product engine;

20 a plurality of second sets of $K - 1$ optical power splitters, each set coupled in series to one of the second set of K optical modulators and each coupled to a dot product engine;

wherein the splitting ratios of the first and second sets of optical power splitters are set at $1:(K - 1)$, $1:(K - 2)$, ..., $1:1$ along the series of splitters.

25 17. The system of claim 11, wherein the device further comprises a multimode interferometer coupled to each of the optical modulators for providing the optical signal and wherein the multimode interferometer is configured to be coupled to a light source.

18. The system of claim 11, wherein each of the dot product engines comprises:

a 2×2 optical power splitter, comprising a first arm and a second arm, wherein the first arm is coupled to one of the first set of K modulators and the second arm is coupled to one of the second set of K modulators;

5 a $\pi/2$ phase shifter coupled to the second arm of the optical power splitter;

a first photodiode coupled to the first arm of the optical power splitter; and

a second photodiode coupled to the second arm of the optical power splitter;

wherein the first and second photodiodes are configured to convert each encoded optical signal to a photocurrent signal which combine to form the product photocurrent signal.

10

19. The system of claim 18, wherein the device further comprises a temporal integrator coupled to each dot product engine for converting the product photocurrent signal to an integrated signal, comprising:

15 at least two capacitors connected in parallel to the dot product engine for receiving the product photocurrent signal;

one or more field effect transistors connected in parallel to the capacitors.

20. The system of claim 19, wherein the device further comprises an analog-to-digital converter associated with one or more of the temporal integrators for converting the integrated signal to a digital signal.

20

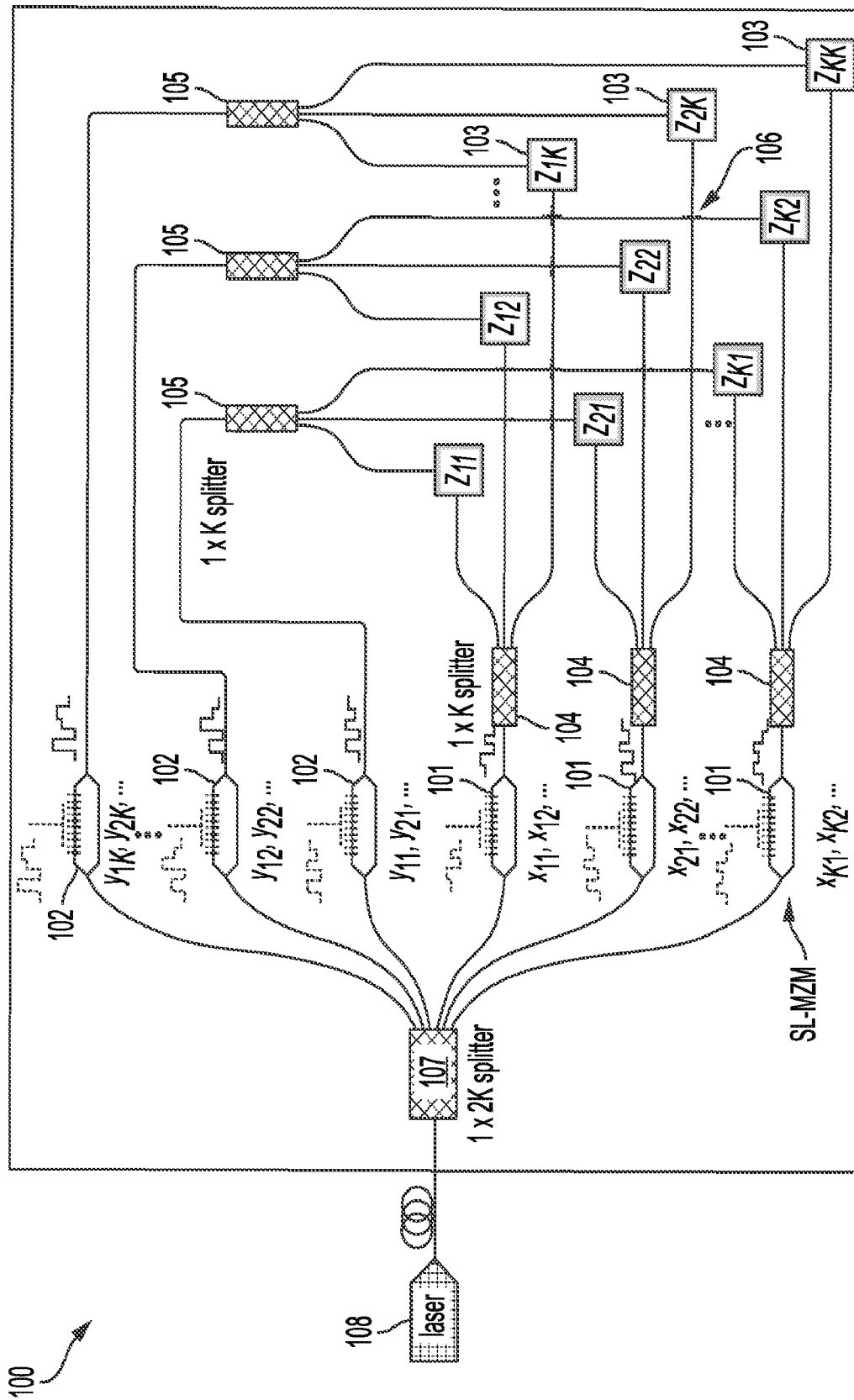


Fig. 1

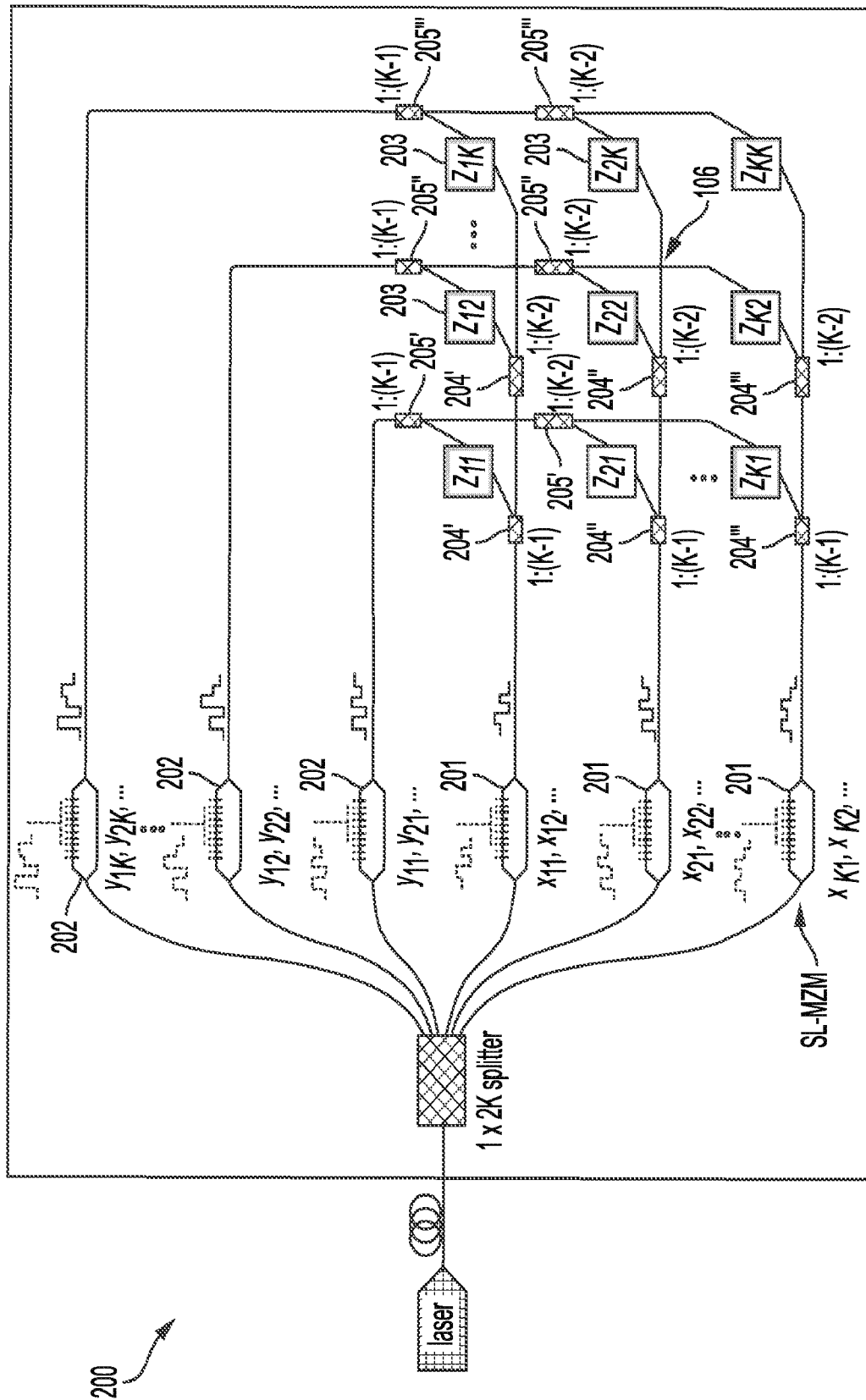


FIG. 2

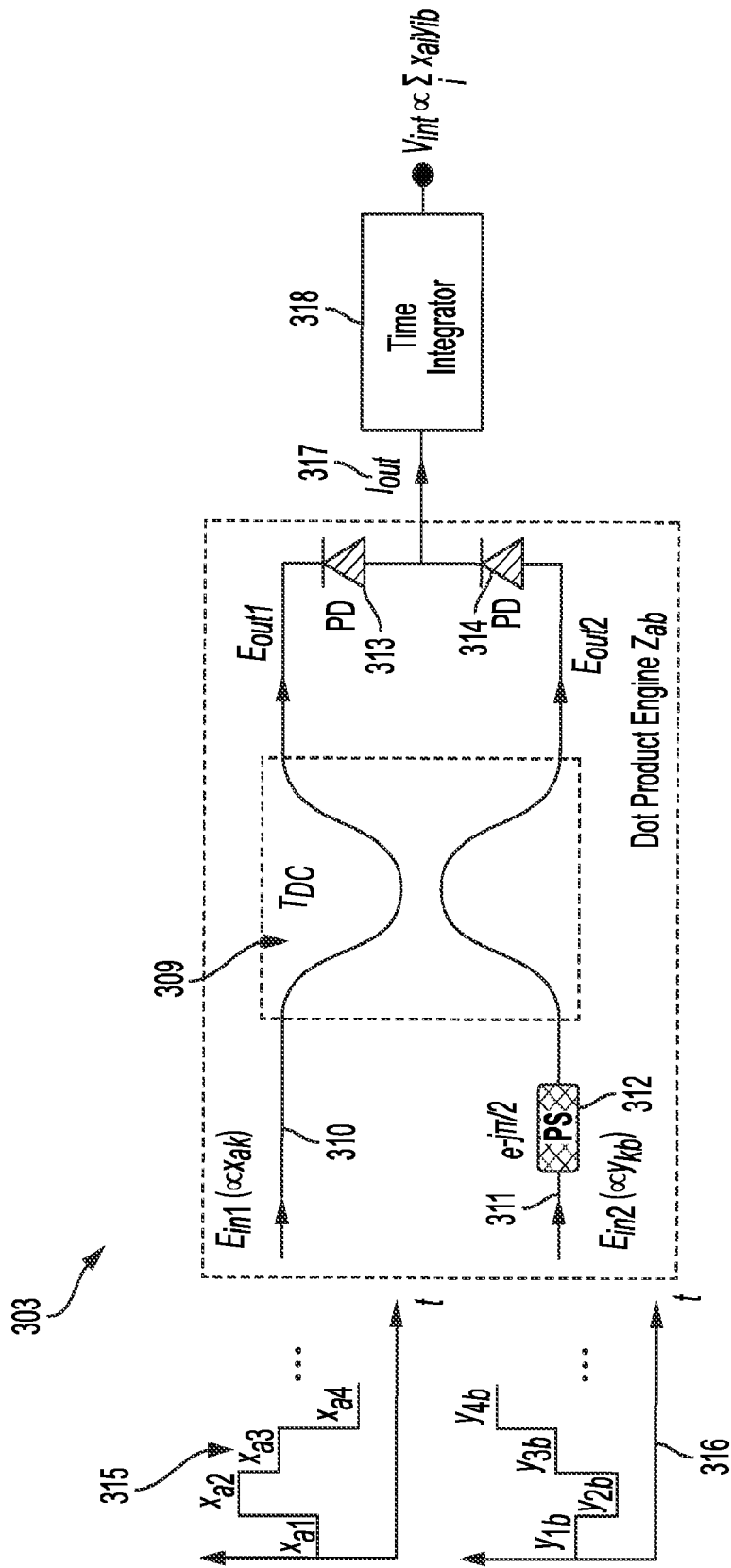


FIG. 3

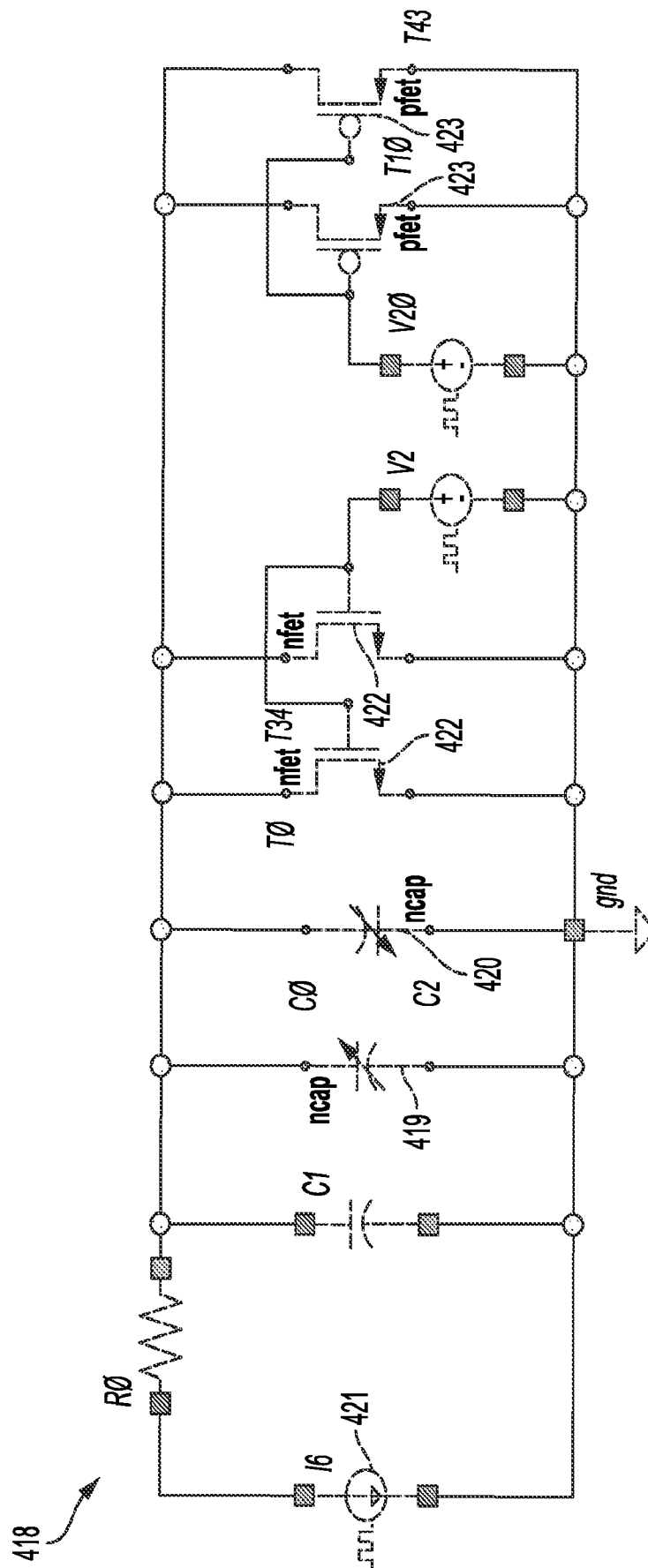
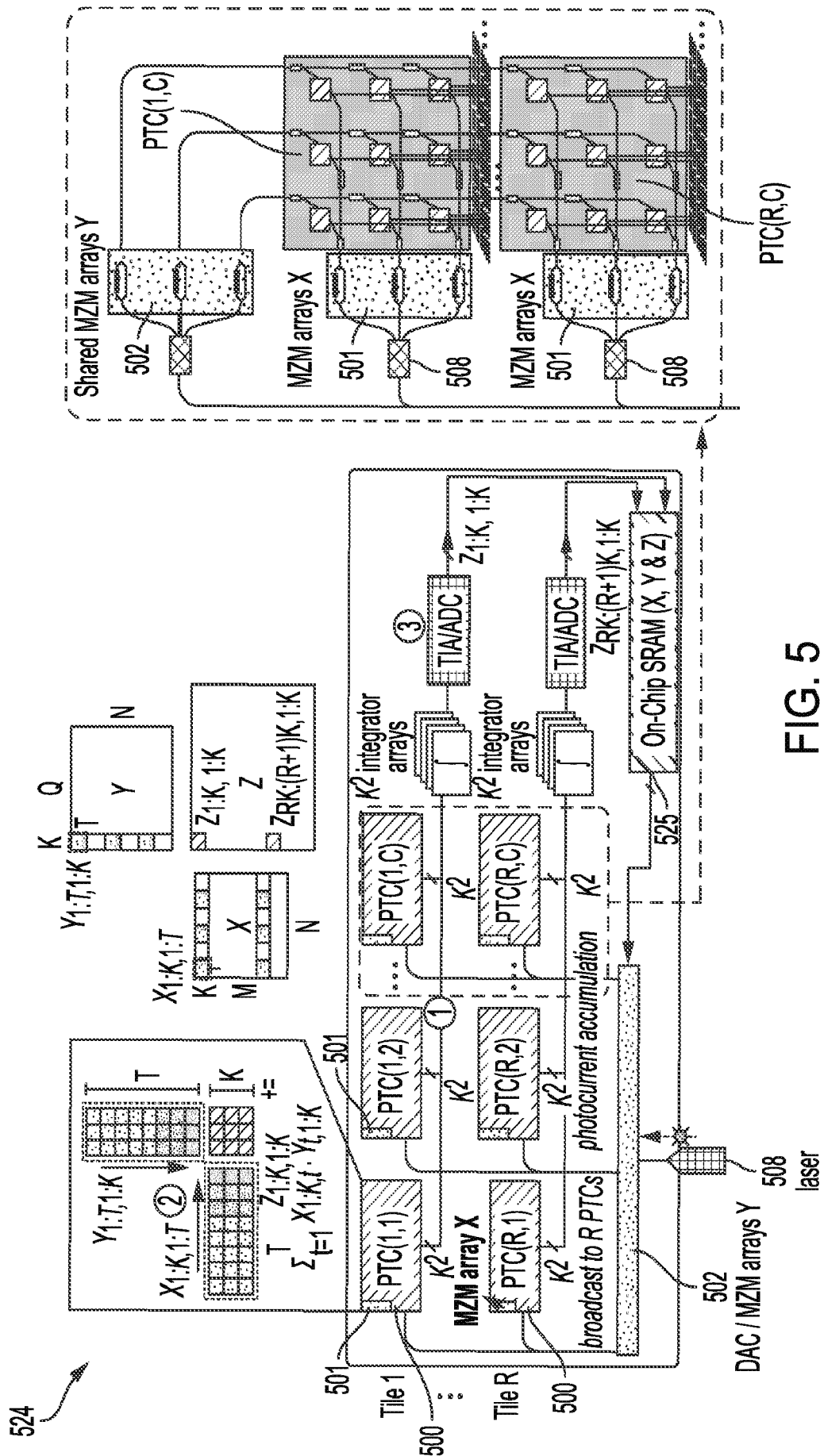


FIG. 4



5
G.
F

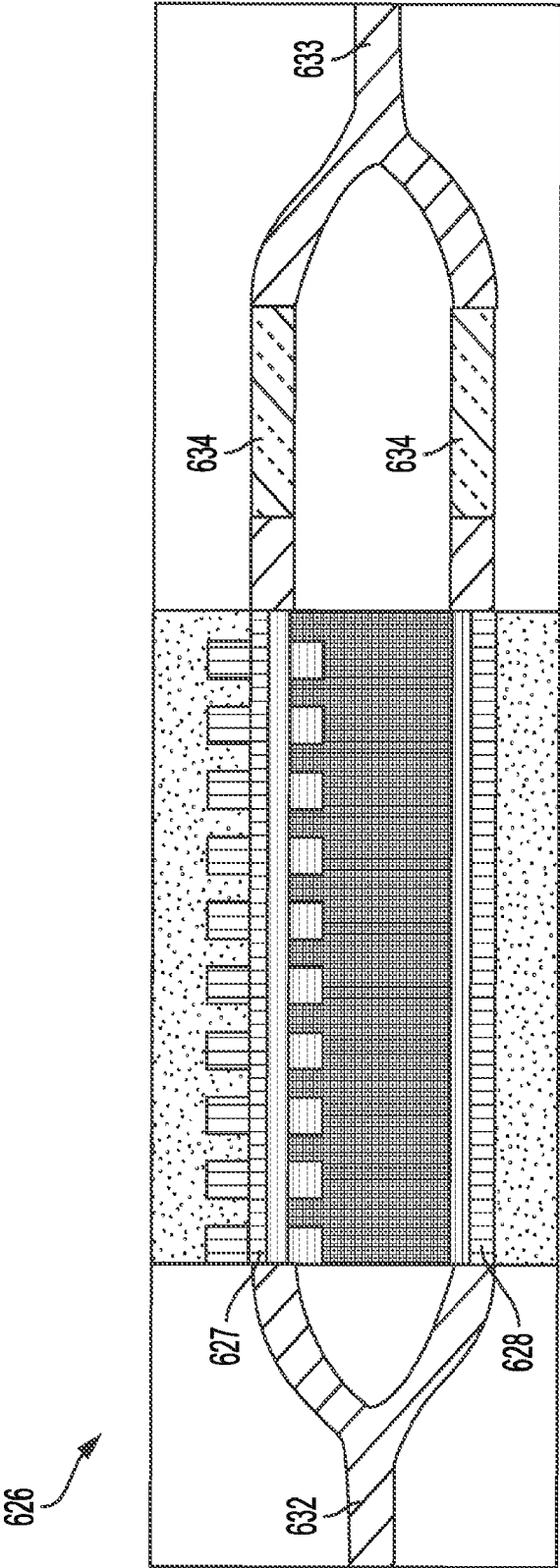


FIG. 6A

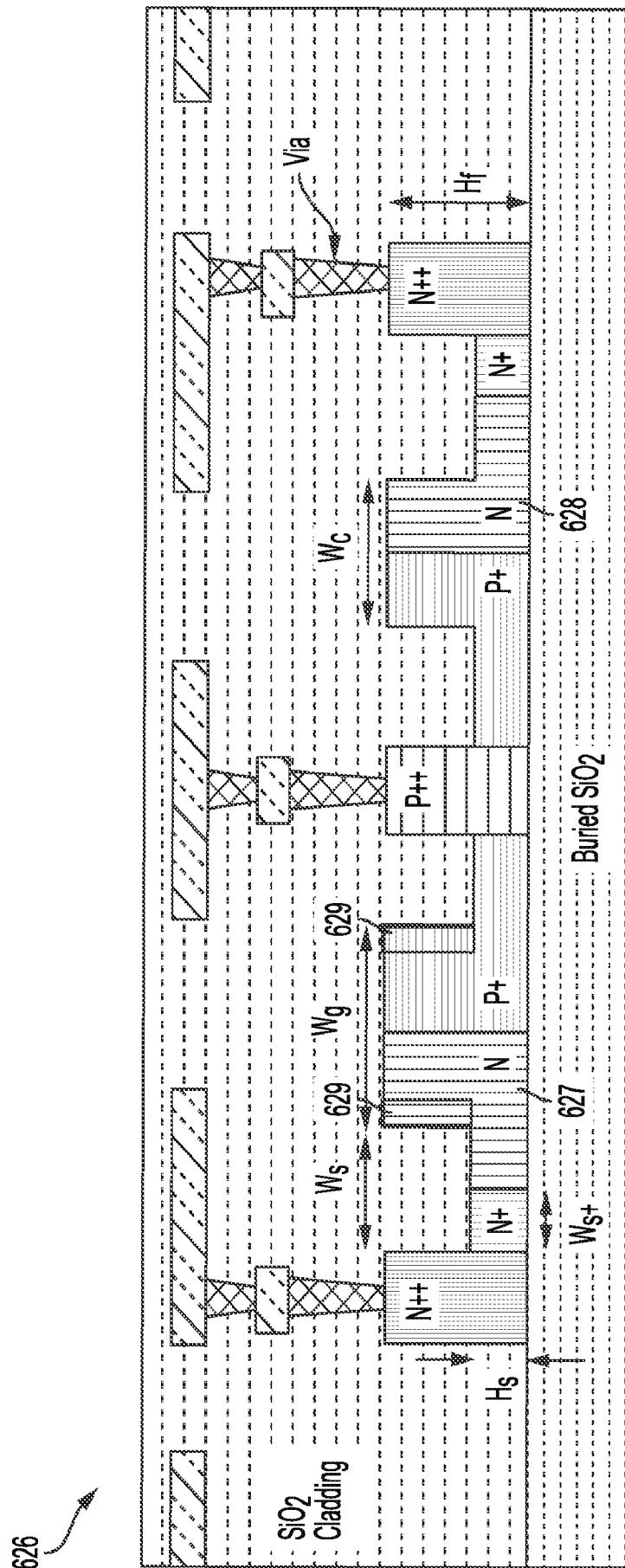


FIG. 6B

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2024/053575

A. CLASSIFICATION OF SUBJECT MATTER IPC: G06E 3/00 (2023.01); G06N 3/067 (2023.01) CPC: G06E 3/008 ; G06E 1/045; G06N 3/067 According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) See Search History Document Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched See Search History Document Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) See Search History Document		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2019/0370644 A1 (LIGHTMATTER, INC.) 05 December 2019 (05.12.2019) entire document, especially FIG. 1-17, FIG. 1-18, FIG. 1-19, para [0003], [0181]-[0182], [0192]-[0193], [0212]	1-20
A	WO 2023/043712 A1 (UNIVERSITY OF PITTSBURGH-OF THE COMMONWEALTH SYSTEM OF HIGHER EDUCATION) 23 March 2023 (23.03.2023) entire document, especially FIGS. 1A-1C, para [0051], [0058]	1-20
A	WO 2022/256905 A1 (HUAWEI TECHNOLOGIES CANADA CO., LTD.) 15 December 2022 (15.12.2022) entire document, especially FIG. 4c, FIG. 12, para [0070]	1-20
☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.		
* Special categories of cited documents: “A” document defining the general state of the art which is not considered to be of particular relevance “D” document cited by the applicant in the international application “E” earlier application or patent but published on or after the international filing date “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) “O” document referring to an oral disclosure, use, exhibition or other means “P” document published prior to the international filing date but later than the priority date claimed “T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art “&” document member of the same patent family		
Date of the actual completion of the international search 15 December 2024 (15.12.2024)		Date of mailing of the international search report 10 January 2025 (10.01.2025)
Name and mailing address of the ISA/US COMMISSIONER FOR PATENTS MAIL STOP PCT, ATTN: ISA/US P.O. Box 1450 Alexandria, VA 22313-1450 UNITED STATES OF AMERICA		Authorized officer KARI RODRIQUEZ
Facsimile No. 571-273-8300		Telephone No. PCT Help Desk: 571-272-4300