for transporting data, over long distances, and over shorter distances (e.g., within data centers). Operations performed on such optical signals often take place in the context of optical data transport, such as within devices that are used for switching or filtering optical signals in a network. Use of optical signals in computing platforms has been more limited. Various components and systems for all-optical computing have been proposed. Such systems may include conversion from and to electrical signals at the input and output, respectively, but may not use both types of signals (electrical and optical) for significant operations that are performed in computations.

SUMMARY

[0006] In general, in an embodiment, a system includes: a memory unit configured to store a dataset and a plurality of neural network weights; a digital-to-analog converter (DAC) unit configured to generate a plurality of modulator control signals and to generate a plurality of weight control signals; an optical processor including: a laser unit configured to generate a plurality of light outputs; a plurality of optical modulators coupled to the laser unit and the DAC unit, the plurality of optical modulators being configured to generate an optical input vector by modulating the plurality of light outputs generated by the laser unit based on the plurality of modulator control signals; an optical matrix multiplication unit coupled to the plurality of optical modulators and the DAC unit, the optical matrix multiplication unit being configured to transform the optical input vector into an optical output vector based on the plurality of weight control signals; and a photodetection unit coupled to the optical matrix multiplication unit and configured to generate a plurality of output voltages corresponding to the optical output vector; an analog-to-digital conversion (ADC) unit coupled to the photodetection unit and configured to convert the plurality of output voltages into a plurality of digitized optical outputs; a controller including integrated circuitry configured to perform operations including: receiving, from a computer, an artificial neural network computation request including an input dataset and a first plurality of neural network weights, wherein the input dataset includes a first digital input vector; storing, in the memory unit, the input dataset and the first plurality of neural network weights; and generating, through the DAC unit, a first plurality of modulator control signals based on the first digital input vector and a first plurality of weight control signals based on the first plurality of neural network weights.

[0025] In some implementations, the integrated circuitry can be an application specific integrated circuit.

[0026] In some implementations, the optical matrix multiplication unit can include: an array of input waveguides to receive the optical input vector; an optical interference unit, in optical communication with the array of input waveguides, to perform a linear transformation of the optical input vector into a second array of optical signals; and an array of output waveguides, in optical communication with the optical interference unit, to guide the second array of optical signals, wherein at least one input waveguide in the array of input waveguides is in optical communication with each output waveguide in the array of output waveguides via the optical interference unit.

[0027] In some implementations, the optical interference unit can include: a plurality of interconnected Mach-Zehnder interferometers (MZIs), each MZI in the plurality of interconnected MZIs including: a first phase shifter configured to change a splitting ratio of the MZI; and a second phase shifter configured to shift a phase of one output of the MZI, wherein the first phase shifters and the second phase shifters are coupled to the plurality of weight control signals.

[0028] In another embodiment, a system includes: a memory unit configured to store a dataset and a plurality of neural network weights; a driver unit configured to generate a plurality of modulator control signals and to generate a plurality of weight control signals; an optical processor including: a laser unit configured to generate a plurality of light outputs; a plurality of optical modulators coupled to the laser unit and the driver unit, the plurality of optical modulators being configured to generate an optical input vector by modulating the plurality of light outputs generated by the laser unit based on the plurality of modulator control signals; an optical matrix multiplication unit coupled to the plurality of optical modulators and the driver unit, the optical matrix multiplication unit being configured to transform the optical input vector into an optical output vector based on the plurality of weight control signals; and a photodetection unit coupled to the optical matrix multiplication unit and configured to generate a plurality of output voltages corresponding to the optical output vector; a comparator unit coupled to the photodetection unit and configured to convert the plurality of output voltages into a plurality of digitized 1-bit optical outputs; and a

controller including integrated circuitry configured to perform operations including: receiving, from a

computer, an artificial neural network computation request including an input dataset and a first plurality of neural network weights, wherein the input dataset includes a first digital input vector having a resolution of N bits; storing, in the memory unit, the input dataset and the first plurality of neural network weights; decomposing the first digital input vector into N 1-bit input vectors, each of the N 1-bit input vectors corresponding to one of the N bits of the first digital input vector; generating, through the driver unit, a sequence of N 1-bit modulator control signals corresponding to the N 1-bit input vectors; obtaining, from the comparator unit, a sequence of N digitized 1-bit optical outputs corresponding to the sequence of the N 1-bit modulator control signals; constructing an N-bit digital output vector from the sequence of the N digitized 1-bit optical outputs; performing a nonlinear transformation on the constructed N-bit digital output vector to generate a transformed N-bit digital output vector; and storing, in the memory unit, the transformed N-bit digital output vector.

[0029] In another embodiment, a method for performing artificial neural network computations in a system having an optical matrix multiplication unit configured to transform an optical input vector into an optical output vector based on a plurality of weight control signals include: receiving, from a computer, an artificial neural network computation request including an input dataset and a first plurality of neural network weights, wherein the input dataset includes a first digital input vector; storing, in a memory unit, the input dataset and the first plurality of neural network weights; generating, through a digital-to-analog converter (DAC) unit, a first plurality of modulator control signals based on the first digital input vector and a first plurality of weight control signals based on the first plurality of neural network weights; obtaining, from an analog-to-digital conversion (ADC) unit, a first plurality of digitized optical outputs corresponding to an optical output vector of the optical matrix multiplication unit, the first plurality of digitized optical outputs forming a first digital output vector; performing, by a controller, a nonlinear transformation on the first digital output vector to generate a first transformed digital output vector; storing, in the memory unit, the first transformed digital output vector; and outputting, by the controller, an artificial neural network output generated based on the first transformed digital output vector.

[0030] In another embodiment, a method includes: providing input information in an electronic format; converting at least a part of the electronic input information into an optical

input vector; optically transforming the optical input vector into an optical output vector based on

an optical matrix multiplication; converting the optical output vector into an electronic format; and electronically applying a non-linear transformation to the electronically converted optical output vector to provide output information in an electronic format.

[0031] Embodiments of the method can include one or more of the following features. For example, the method can further include: repeating the electronic-to-optical converting, the optical transforming, the optical-to-electronic converting, and the electronically applied non-linear transforming with respect to new electronic input information corresponding to the provided output information in electronic format.

[0032] In some implementations, the optical matrix multiplication for the initial optical transforming and the optical matrix multiplication of the repeated optical transforming can be the same and can correspond to the same layer of an artificial neural network.

[0033] In some implementations, the optical matrix multiplication for the initial optical transforming and the optical matrix multiplication of the repeated optical transforming can be different and can correspond to different layers of an artificial neural network.

[0034] In some implementations, the method can further include: repeating the electronic-to-optical converting, the optical transforming, the optical-to-electronic converting, and the electronically applied non-linear transforming with respect to a different part of the electronic input information, wherein the optical matrix multiplication for the initial optical transforming and the optical matrix multiplication of the repeated optical transforming are the same and correspond to a first layer of an artificial neural network.

[0035] In some implementations, the method can further include: providing intermediate information in an electronic format based on the electronic output information produced for the multiple parts of the electronic input information by the first layer of the artificial neural network; and repeating the electronic-to-optical converting, the optical transforming, the optical-to-electronic converting, and the electronically applied non-linear transforming with respect to each of different parts of the electronic intermediate information, wherein the optical matrix multiplication for the initial optical transforming and the optical matrix multiplication of the repeated optical transforming related to the different parts of the electronic intermediate information are the same and correspond to a second layer of the artificial neural network.

[0036] In another embodiment, a system includes: an optical processor including passive

[0056] In another embodiment, a method includes: 3D printing an optical matrix processing unit comprising passive diffractive optical elements, in which the passive diffractive optical elements are configured to transform an optical input vector or matrix to an optical output vector or matrix that represents a result of a matrix processing applied to an optical input vector or matrix and a predetermined vector defined by the arrangement of the diffractive optical elements.

[0057] In another embodiment, a method include: generating, using one or more laser beams, a hologram comprising passive diffractive optical elements, in which the passive diffractive optical elements are configured to transform an optical input vector or matrix to an optical output vector or matrix that represents a result of a matrix processing applied to an optical input vector or matrix and a predetermined vector defined by the arrangement of the diffractive optical elements.

[0058] In another embodiment, a system includes: an optical processor comprising passive diffractive optical elements arranged in a one-dimensional manner, in which the passive diffractive optical elements are configured to transform an optical input to an optical output that represents a result of a matrix processing applied to the optical input and a predetermined vector defined by the arrangement of the diffractive optical elements.

[0059] Implementations of the system can include one or more of the following features. For example, the matrix processing can include a matrix multiplication between the optical input and the predetermined vector defined by the arrangement of the diffractive optical elements.

[0060] In some implementations, the optical processor can include an optical matrix processing unit that includes: an input waveguide to receive the optical input, an optical interference unit comprising the passive diffractive optical elements, in which the optical interference unit is in optical communication with the input waveguide and configured to perform a linear transformation of the optical input; and an output waveguide, in optical communication with the optical interference unit, to guide the optical output.

[0061] In some implementations, the optical interference unit can include a substrate having at least one of holes or gratings, and the holes or grating elements can have dimensions in a range from 100 nm to 10 μm.

[0062] In another embodiment, a system includes: a memory unit; a digital-to-analog

converter

bit digital output vector to generate a transformed N-bit digital output vector; and storing, in the memory unit, the transformed N-bit digital output vector.

[0077] In some implementations, the memory unit can include a digital input vector memory configured to store the digital input vector and comprising at least one SRAM.

[0078] In some implementations, the laser unit can include: a laser source configured to generate light; and an optical power splitter configured to split the light generated by the laser source into the plurality of light outputs, wherein each of the plurality of light outputs have substantially equal powers.

[0079] In some implementations, the plurality of optical modulators can include one of MZI modulators, ring resonator modulators, or electro-absorption modulators.

[0080] In some implementations, the photodetection unit can include: a plurality of photodetectors; and a plurality of amplifiers configured to convert photocurrents generated by the photodetectors into the plurality of output electric signals.

[0081] In some implementations, the integrated circuitry can include an application specific integrated circuit.

[0082] In some implementations, the optical matrix processing unit can include: an array of input waveguides to receive the optical input vector; an optical interference unit, in optical communication with the array of input waveguides, to perform a linear transformation of the optical input vector into a second array of optical signals, wherein the optical interference unit comprises the passive diffractive optical elements; and an array of output waveguides, in optical communication with the optical interference unit, to guide the second array of optical signals, wherein at least one input waveguide in the array of input waveguides is in optical communication with each output waveguide in the array of output waveguides via the optical interference unit.

[0083] In another embodiment, a system includes: a memory unit; a driver unit configured to generate a plurality of modulator control signals; and an optical processor that includes: a laser unit configured to generate a plurality of light outputs; a plurality of optical modulators coupled to the laser unit and the driver unit, the plurality of optical modulators being configured to generate an optical input vector by modulating the plurality of light outputs generated by the laser unit based on the plurality of modulator control signals; an optical matrix processing unit coupled to the plurality of optical modulators and the driver unit, the

optical matrix processing unit comprising passive diffractive optical elements configured to transform the optical input vector into an optical output vector based on a plurality of weight control signals defined by the passive diffractive optical elements; and a photodetection unit coupled to the optical matrix processing unit and configured to generate a plurality of output electric signals corresponding to the optical output vector. The system also includes a comparator unit coupled to the photodetection unit and configured to convert the plurality of output electric signals into a plurality of digitized 1-bit optical outputs; and a controller including integrated circuitry configured to perform operations including: receiving, from a computer, an artificial neural network computation request comprising an input dataset, wherein the input dataset comprises a first digital input vector having a resolution of N bits; storing, in the memory unit, the input dataset; decomposing the first digital input vector into N 1-bit input vectors, each of the N 1-bit input vectors corresponding to one of the N bits of the first digital input vector; generating, through the driver unit, a sequence of N 1-bit modulator control signals corresponding to the N 1-bit input vectors; obtaining, from the comparator unit, a sequence of N digitized 1-bit optical outputs corresponding to the sequence of the N 1-bit modulator control signals; constructing an N-bit digital output vector from the sequence of the N digitized 1-bit optical outputs; performing a nonlinear transformation on the constructed N-bit digital output vector to generate a transformed N-bit digital output vector; and storing, in the memory unit, the transformed N-bit digital output vector.

[0084] Embodiments of the system can include one or more of the following features. For example, the optical matrix processing unit can include an optical matrix multiplication unit configured to transform the optical input vector into an optical output vector that represents a product of a matrix multiplication between an input vector represented by the optical input vector and a predetermined vector defined by the diffractive optical elements.

[0085] In another embodiment, a method for performing artificial neural network computations in a system having an optical matrix processing unit, the method includes: receiving, from a computer, an artificial neural network computation request comprising an input dataset comprising a first digital input vector; storing, in a memory unit, the input dataset; generating, through a digital-to-analog converter (DAC) unit, a first plurality of

modulator control signals based on the first digital input vector; transforming, by using the optical

matrix processing unit comprising an arrangement of diffractive optical elements, an optical input vector into an optical output vector, in which the optical output vector represents a result of a matrix processing applied to the optical input vector and a predetermined vector defined by the arrangement of the diffractive optical elements; obtaining, from an analog-to-digital conversion (ADC) unit, a first plurality of digitized optical outputs corresponding to the optical output vector of the optical matrix processing unit, the first plurality of digitized optical outputs forming a first digital output vector; performing, by a controller, a nonlinear transformation on the first digital output vector to generate a first transformed digital output vector; storing, in the memory unit, the first transformed digital output vector; and outputting, by the controller, an artificial neural network output generated based on the first transformed digital output vector.

[0086] Embodiments of the method can include one or more of the following features. For example, transforming the optical input vector into the optical output vector can include transforming the optical input vector into an optical output vector that represents a product of a matrix multiplication between the digital input vector and the predetermined vector defined by the arrangement of diffractive optical elements.

[0087] In another embodiment, a method includes: providing input information in an electronic format; converting at least a part of the electronic input information into an optical input vector; optically transforming the optical input vector into an optical output vector based on an optical matrix processing by an optical processor comprising passive diffractive optical elements; converting the optical output vector into an electronic format; and electronically applying a non-linear transformation to the electronically converted optical output vector to provide output information in an electronic format.

[0088] Embodiments of the method can include one or more of the following features. For example, optically transforming the optical input vector into an optical output vector can include optically transforming the optical input vector into an optical output vector based on an optical matrix multiplication between a digital input vector represented by the optical input vector and a predetermined vector defined by the passive diffractive optical elements.

[0089] In some implementations, the method can further include: repeating the electronic-to-optical converting, the optical transforming, the optical-to-electronic converting, and the

electronically applied non-linear transforming with respect to new electronic input information corresponding to the provided output information in electronic format.

[0090] In some implementations, the optical matrix processing for the initial optical transforming and the optical matrix processing of the repeated optical transforming can be the same and correspond to the same layer of an artificial neural network.

[0091] In some implementations, the method can further include: repeating the electronic-to-optical converting, the optical transforming, the optical-to-electronic converting, and the electronically applied non-linear transforming with respect to a different part of the electronic input information, wherein the optical matrix processing for the initial optical transforming and the optical matrix processing of the repeated optical transforming can be the same and correspond to a layer of an artificial neural network.

[0092] In another embodiment, a system including: an optical matrix processing unit configured to process an input vector of length N, in which the optical matrix processing unit comprises N+2 layers of directional couplers and N layers of phase shifters, and N is a positive integer.

[0093] Embodiments of the system can include one or more of the following features. For example, the optical matrix processing unit can include no more than N+2 layers of directional couplers.

[0094] In some implementations, the optical matrix processing unit can include an optical matrix multiplication unit.

[0095] In some implementations, the optical matrix processing unit can include: a substrate, and interconnected interferometers disposed on the substrate, in which each interferometer comprises optical waveguides disposed on the substrate, and the directional couplers and the phase shifters are part of the interconnected interferometers.

[0096] In some implementations, the optical matrix processing unit can include a layer of attenuators following the last layer of directional couplers.

[0097] In some implementations, the layer of attenuators can include N attenuators.

[0098] In some implementations, the system can include one or more homodyne detectors to detect outputs from the attenuators.

[0099] In some implementations, $N = 3$, and the optical matrix processing unit can include: input terminals configured to receive the input vector; a first layer of directional couplers coupled to the input terminals; a first layer of phase shifters coupled to the first layer

[00106] In some implementations, for each of the j-th layer of directional couplers in which j is an even number, the m-th directional coupler can be coupled to the $(2m)$-th and $(2m+1)$-th output of the previous layer, and m is an integer from 1 to $N/2-1$.

[00107] In some implementations, each of the i-th layer of phase shifters can include N phase shifters for which i is an odd number, and each of the j-th layer of phase shifters can include $N - 2$ phase shifters for which j is an even number.

[00108] In some implementations, N can be an odd number.

[00109] In some implementations, each layer of directional couplers can include $(N – 1) / 2$ directional couplers.

[00110] In some implementations, each layer of phase shifters can include $N – 1$ phase shifters.

[00111] In another embodiment, a system includes: a generator configured to generate a first dataset, in which the generator comprises an optical matrix processing unit; and a discriminator configured to receive a second dataset comprising data from the first dataset and data from a third dataset, the data in the first dataset having characteristics similar to those of the data in the third dataset, and classify data in the second dataset as data from the first dataset or data from the third dataset.

[00112] Embodiments of the method can include one or more of the following features. For example, the optical matrix processing unit can include at least one of (i) the optical matrix multiplication unit recited above, (ii) the passive diffractive optical elements recited above, or (iii) the optical matrix processing unit recited above.

[00113] In some implementations, the third dataset can include real data, the generator is configured to generate synthesized data that resemble the real data, and the discriminator is configured to classify data as real data or synthesized data.

[00114] In some implementations, the generator can be configured to generate datasets for training at least one of autonomous vehicles, medical diagnosis systems, fraud detection systems, weather prediction systems, financial forecast systems, facial recognition systems, speech recognition systems, or product defect detection systems.

[00115] In some implementations, the generator can be configured to generate images resembling images of at least one of real objects or real scenes, and the discriminator is

configured to classify a received image as (i) an image of a real object or real scene, or (ii) a synthesized image generated by the generator.

[00116] In some implementations, the real objects can include at least one of people, animals, cells, tissues, or products, and the real scenes comprise scenes encountered by vehicles.

[00117] In some implementations, the discriminator can be configured to classify whether a received image is (i) an image of real people, real animals, real cells, real tissues, real products, or real scenes encountered by vehicles, or (ii) a synthesized image generated by the generator.

[00118] In some implementations, the vehicles can include at least one of motorcycles, cars, trucks, trains, helicopters, airplanes, submarines, ships, or drones.

[00119] In some implementations, the generator can be configured to generate images of tissues or cells associated with at least one of diseases of humans, diseases of animals, or diseases of plants.

[00120] In some implementations, the generator can be configured to generate images of tissues or cells associated with diseases of humans, and the diseases comprise at least one of cancer, Parkinson's disease, sickle cell anemia, heart disease, cardiovascular disease, diabetes, chest disease, or skin disease.

[00121] In some implementations, the generator can be configured to generate images of tissues or cells associated with cancer, and the cancer can include at least one of skin cancer, breast cancer, lung cancer, liver cancer, prostate cancer, or brain cancer.

[00122] In some implementations, the system can further include a random noise generator configured to generate random noise that is provided as input to the generator, and the generator is configured to generate the first dataset based on the random noise.

[00123] In another embodiment, a system includes: a random noise generator configured to generate random noise; and a generator configured to generate data based on the random noise, in which the generator comprises an optical matrix processing unit.

[00124] Embodiments of the system can include one or more of the following features. For example, the optical matrix processing unit can include at least one of (i) the optical matrix multiplication unit described above, (ii) the passive diffractive optical elements described above, or (iii) the optical matrix processing unit described above.

[00125] In another embodiment, a system includes: a photonic circuit configured to perform a logic function on two input signals, the photonic circuit including: a first directional coupler having two input terminals and two output terminals, the two input terminals configured to receive the two input signals, a first pair of phase shifters configured to modify phases of the signals at the two output terminals of the first directional coupler, a second directional coupler having two input terminals and two output terminals, the two input terminals configured to receive signals from the first pair of phase shifters, and a second pair of phase shifters configured to modify phases of the signals at the two output terminals of the second directional coupler.

[00126] Embodiments of the method can include one or more of the following features. For example, the phase shifters can be configured to cause the photonic circuit to implement a rotation:

$$\mathrm{M} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \cos\pi/4 & -\sin\pi/4 \\ \sin\pi/4 & \cos\pi/4 \end{pmatrix}.$$

[00127] In some implementations, when input signals x1 and x2 are provided to the two input terminals of the first directional coupler, the phase shifters can be configured to cause the photonic circuit to implement an operation:

$$\mathrm{M} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}}(x_1 - x_2) \\ \frac{1}{\sqrt{2}}(x_1 + x_2) \end{pmatrix}.$$

[00128] In some implementations, the photonic circuit can include first photodetectors configured to generate absolute values of the signals from the second pair of phase shifters to cause the photonic circuit to implement an operation:

$$\left|\mathrm{M} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right| = \begin{pmatrix} \frac{1}{\sqrt{2}}|x_1 - x_2| \\ \frac{1}{\sqrt{2}}|x_1 + x_2| \end{pmatrix}.$$

[00143] The modules can be conveniently fabricated at a large scale by integrating photonics and electronics on a common substrate (e.g., a Silicon chip). Routing signals on the substrate as optical signals instead of electrical signals, and grouping photodetectors in a portion of the substrate may help avoid long electronic wiring and their associated challenges (e.g., parasitic capacitance, inductance, and crosstalk).

[00144] For an embodiment of the system that uses submatrix multiplication, each element of the output vector can be computed concurrently using a different device (e.g., different core, different processor, different computer, different server), helping to alleviate certain potential limitations, such as the memory wall, and helping the overall system to scale for very large matrices. In some embodiments, each submatrix can be multiplied by a corresponding sub-vector using a different device. The total sum can then be computed by collecting or accumulating summands from different devices. The intermediate results, in the form of optical signals, can be conveniently transported between devices, even if the devices are separated by relatively large distances.

[00145] Other aspects include other combinations of the features recited above and other features, expressed as methods, apparatus, systems, program products, and in other ways.

[00146] Particular embodiments of the subject matter described in this specification can be implemented to realize one or more of the following advantages. ANN computation throughput, latency, or both may be improved. Power efficiency of ANN computations may be improved.

[00147] In another embodiment, an apparatus includes: a plurality of optical waveguides, wherein a set of multiple input values are encoded on respective optical signals carried by the optical waveguides; a plurality of copying modules, and for each of at least two subsets of one or more optical signals, a corresponding set of one or more of the copying modules is configured to split the subset of one or more optical signals into two or more copies of the optical signals; a plurality of multiplication modules, and for each of at least two copies of a first subset of one or more optical signals, a corresponding multiplication module configured to multiply the one or more optical signals of the first subset by one or more matrix element values using optical amplitude modulation, where at least one of the multiplication modules includes an optical amplitude modulator including an input port and two output ports, and a pair of related optical signals is provided from the two output ports such that a difference

[00177] In another embodiment, a method includes: encoding a set of multiple input values on respective optical signals; for each of at least two subsets of one or more optical signals, using a corresponding set of one or more copying modules to split the subset of one or more optical signals into two or more copies of the optical signals; for each of at least two copies of a first subset of one or more optical signals, using a corresponding multiplication module to multiply the one or more optical signals of the first subset by one or more matrix element values using optical amplitude modulation, where at least one of the multiplication modules includes an optical amplitude modulator including an input port and two output ports, and a pair of related optical signals is provided from the two output ports such that a difference between amplitudes of the related optical signals corresponds to a result of multiplying an input value by a signed matrix element value; and for results of two or more of the multiplication modules, using a summation module configured to produce an electrical signal that represents a sum of the results of the two or more of the multiplication modules.

[00178] In another embodiment, a method includes: encoding a set of input values representing elements of an input vector on respective optical signals; encoding a set of coefficients representing elements of a matrix as amplitude modulation levels of a set of optical amplitude modulators coupled to the optical signals, where at least one of the optical amplitude modulators including an input port and two output ports provides a pair of related optical signals from the two output ports such that a difference between amplitudes of the related optical signals corresponds to a result of multiplying an input value by a signed matrix element value; and encoding a set of output values representing elements of an output vector on respective electrical signals, where at least one of the electrical signals is in the form of a current whose amplitude corresponds to a sum of respective elements of the input vector multiplied by respective elements of a row of the matrix.

[00179] Embodiments of the method can include one or more of the following features. For example, at least one of the optical signals can be provided by a first optical waveguide, and the first optical waveguide can be coupled to an optical splitter that sends a predetermined fraction of the power of an optical wave guided by the first optical waveguide to a second output optical waveguide, and sends the remaining fraction of the power of the optical wave guided by the first optical waveguide to a third optical waveguide.

[00180] In another embodiment, an apparatus includes: a plurality of optical waveguides

encoding a set of input values representing elements of an input vector on respective optical signals carried by the optical waveguides; a set of optical amplitude modulators coupled to the optical signals encoding a set of coefficients representing elements of a matrix as amplitude modulation levels, where at least one of the optical amplitude modulators including an input port and two output ports provides a pair of related optical signals from the two output ports such that a difference between amplitudes of the related optical signals corresponds to a result of multiplying an input value by a signed matrix element value; and a plurality of summation modules encoding a set of output values representing elements of an output vector on respective electrical signals, where at least one of the electrical signals is in the form of a current whose amplitude corresponds to a sum of respective elements of the input vector multiplied by respective elements of a row of the matrix.

[00181] In another embodiment, a method for multiplying an input vector by a given matrix includes: encoding a set of input values representing elements the input vector on respective optical signals of a set of optical signals; coupling a first set of one or more devices to a first set of one or more waveguides providing a first subset of the set of optical signals, and generating a result of a first submatrix of the given matrix multiplied by values encoded on the first subset of the set of optical signals; coupling a second set of one or more devices to a second set of one or more waveguides providing a second subset of the set of optical signals, and generating a result of a second submatrix of the given matrix multiplied by values encoded on the second subset of the set of optical signals; coupling a third set of one or more devices to a third set of one or more waveguides providing a copy of the first subset of the set of optical signals generated by a first optical splitter, and generating a result of a third submatrix of the given matrix multiplied by values encoded on the first subset of the set of optical signals; and coupling a fourth set of one or more devices to a fourth set of one or more waveguides providing a copy of the second subset of the set of optical signals generated by a second optical splitter, and generating a result of a fourth submatrix of the given matrix multiplied by values encoded on the second subset of the set of optical signals; wherein the first, second, third, and fourth submatrices concatenated together form the given matrix; and wherein at least one output value representing an element of an output vector corresponding to the input vector multiplied by the given matrix is encoded on electrical

signals generated by a device in communication with the first set of one or more devices and the second set of one or more devices.

[00182] Embodiments of the method can include one or more of the following features. For example, each pair of sets of the first set of one or more devices, the second set of one or more devices, the third set of one or more devices, and the fourth set of one or more devices can be mutually exclusive.

[00183] In another embodiment, an apparatus includes: a first set of one or more devices configured to receive a first set of optical signals, and to generate a result of a first matrix multiplied by values encoded on the first set of optical signals; a second set of one or more devices configured to receive a second set of optical signals, and to generate a result of a second matrix multiplied by values encoded on the second set of optical signals; a third set of one or more devices configured to receive a third set of optical signals, and to generate a result of a third matrix multiplied by values encoded on the third set of optical signals; a fourth set of one or more devices configured to receive a fourth set of optical signals, and to generate a result of a fourth matrix multiplied by values encoded on the fourth set of optical signals; and configurable connection pathways between two or more of the first set of one or more devices, the second set of one or more devices, the third set of one or more devices, or the fourth set of one or more devices, wherein a first configuration of the configurable connection pathways is configured to: (1) provide a copy of the first set of optical signals as at least one of the second set of optical signals, the third set of optical signals, or the fourth set of optical signals, and (2) provide one or more signals from the first set of one or more device and one or more signals from the second set of one or more devices to a summation module configured to produce an electrical signal that represents a sum of values encoded on signals received by the summation module.

[00184] In another embodiment, an apparatus includes: a first set of one or more devices configured to receive a first set of optical signals, and to generate a result based on optical amplitude modulation of one or more of the optical signals of the first set of optical signals; a second set of one or more devices configured to receive a second set of optical signals, and to generate a result based on optical amplitude modulation of one or more of the optical signals of the second set of optical signals; a third set of one or more devices configured to receive a third set of optical signals, and to generate a result based on optical amplitude modulation of

one or more of the optical signals of the third set of optical signals; a fourth set of one or more devices configured to receive a fourth set of optical signals, and to generate a result based on optical amplitude modulation of one or more of the optical signals of the fourth set of optical signals; and configurable connection pathways between two or more of the first set of one or more devices, the second set of one or more devices, the third set of one or more devices, or the fourth set of one or more devices, wherein a first configuration of the configurable connection pathways is configured to: (1) provide a copy of the first set of optical signals as the third set of optical signals, or (2) provide one or more signals from the first set of one or more device and one or more signals from the second set of one or more devices to a summation module configured to produce an electrical signal that represents a sum of values encoded on signals received by the summation module.

[00185] Embodiments of the apparatus can include one or more of the following features. For example, each pair of sets of the first set of one or more devices, the second set of one or more devices, the third set of one or more devices, and the fourth set of one or more devices can be mutually exclusive.

[00186] In some implementations, the first configuration of the configurable connection pathways can be configured to: (1) provide a copy of the first set of optical signals as the third set of optical signals, and (2) provide one or more signals from the first set of one or more device and one or more signals from the second set of one or more devices to the summation module configured to produce an electrical signal that represents a sum of values encoded on at least two different signals received by the summation module.

[00187] In some implementations, the first configuration of the configurable connection pathways can be configured to provide a copy of the first set of optical signals as the third set of optical signals, and a second configuration of the configurable connection pathways can be configured to provide one or more signals from the first set of one or more device and one or more signals from the second set of one or more devices to the summation module configured to produce an electrical signal that represents a sum of values encoded on signals received by the summation module.

[00188] In another embodiment, an apparatus includes: a plurality of optical waveguides, wherein a set of multiple input values are encoded on respective optical signals carried by the

optical waveguides; a plurality of copying modules, including for each of at least two subsets of one

or more optical signals, a corresponding set of one or more copying modules configured to split the subset of one or more optical signals into two or more copies of the optical signals; a plurality of multiplication modules, including for each of at least two copies of a first subset of one or more optical signals, a corresponding multiplication module configured to multiply the one or more optical signals of the first subset by one or more values using optical amplitude modulation; and one or more summation modules, including for results of two or more of the multiplication modules, a summation module configured to produce an electrical signal that represents a sum of the results of the two or more of the multiplication modules, where the results include at least one result that is encoded on an electrical signal and was derived from one of the copies of the optical signals that propagated through no more than a single optical amplitude modulator before being converted to an electrical signal.

[00189] In an aspect, the present disclosure provides a system comprising: a first unit configured to generate a plurality of modulator control signals; a processor unit comprising: a light source configured to provide a plurality of light outputs; a plurality of optical modulators coupled to the light source and the first unit, the plurality of optical modulators being configured to generate an optical input vector by modulating the plurality of light outputs provided by the light source based on the plurality of modulator control signals, the optical input vector comprising a plurality of optical signals; and a matrix multiplication unit coupled to the plurality of optical modulators and the first unit, the matrix multiplication unit being configured to transform the optical input vector into an analog output vector based on a plurality of weight control signals; a second unit coupled to the matrix multiplication unit and configured to convert the analog output vector into a digitized output vector; and a controller including integrated circuitry configured to perform operations comprising: receiving an artificial neural network computation request comprising an input dataset that comprises a first digital input vector; receiving a first plurality of neural network weights; and generating, through the first unit, a first plurality of modulator control signals based on the first digital input vector and a first plurality of weight control signals based on the first plurality of neural network weights.

[00190] Embodiments of the system can include one or more of the following features. For example, the first unit can include a digital to analog converter (DAC).

[00191] In some implementations, the second unit can include an analog to digital

accumulator can produce an output electrical signal that is encoded with more than two amplitude levels corresponding to different duty cycles of the time domain encoding over the multiple time slots.

[00257] In some implementations, the two or more of the multiplication modules each correspond to a different subset of one or more optical signals.

[00258] In some implementations, the apparatus can further include, for each copy of a second subset of one or more optical signals, different from the optical signals in the first subset of one or more optical signals, a multiplication module configured to multiply the one or more optical signals of the second subset by one or more matrix element values using optical amplitude modulation.

[00259] In another embodiment a system includes: a memory unit configured to store a dataset and a plurality of neural network weights; and a driver unit configured to generate a plurality of modulator control signals. The system includes an optoelectronic processor including: a light source configured to provide a plurality of light outputs; a plurality of optical modulators coupled to the light source and the driver unit, the plurality of optical modulators being configured to generate an optical input vector by modulating the plurality of light outputs generated by the light source based on the plurality of modulator control signals; a matrix multiplication unit coupled to the plurality of optical modulators and the driver unit, the matrix multiplication unit being configured to transform the optical input vector into an analog output vector based on a plurality of weight control signals; and a comparator unit coupled to the matrix multiplication unit and configured to convert the analog output vector into a plurality of digitized 1-bit outputs. The system includes a controller including integrated circuitry configured to perform operations including: receiving an artificial neural network computation request comprising an input dataset and a first plurality of neural network weights, wherein the input dataset comprises a first digital input vector having a resolution of N bits; storing, in the memory unit, the input dataset and the first plurality of neural network weights; decomposing the first digital input vector into N 1-bit input vectors, each of the N 1-bit input vectors corresponding to one of the N bits of the first digital input vector; generating, through the driver unit, a sequence of N 1-bit modulator control signals corresponding to the N 1-bit input vectors; obtaining, from

the comparator unit, a sequence of N digitized 1-bit outputs corresponding to the sequence of the N 1-bit modulator control

the optical input vector, a corresponding set of one or more copying modules configured to split the subset of one or more optical signals into two or more copies of the optical signals; a plurality of multiplication modules, including for each of at least two copies of a first subset of one or more optical signals, a corresponding multiplication module configured to multiply the one or more optical signals of the first subset by one or more matrix element values using optical amplitude modulation; and one or more summation modules, including for results of two or more of the multiplication modules, a summation module configured to produce an electrical signal that represents a sum of the results of the two or more of the multiplication modules.

[00266] In some implementations, at least one of the multiplication modules can include an optical amplitude modulator including an input port and two output ports, and a pair of related optical signals can be provided from the two output ports such that a difference between amplitudes of the related optical signals corresponds to a result of multiplying an input value by a signed matrix element value.

[00267] In some implementations, the matrix multiplication unit can be configured to multiply the input vector by a matrix that includes the one or more matrix element values.

[00268] In some implementations, a set of multiple output values can be encoded on respective electrical signals produced by the one or more summation modules, and the output values in the set of multiple output values can represent elements of an output vector that results from the input vector being multiplied by the matrix.

[00269] In another embodiment, a method for performing artificial neural network computations in a system having a matrix multiplication unit configured to transform an optical input vector into an analog output vector based on a plurality of weight control signals is provided. The method includes: receiving an artificial neural network computation request comprising an input dataset and a first plurality of neural network weights, wherein the input dataset comprises a first digital input vector; storing, in a memory unit, the input dataset and the first plurality of neural network weights; generating a first plurality of modulator control signals based on the first digital input vector and a first plurality of weight control signals based on the first plurality of neural network weights; obtaining a first plurality of digitized outputs corresponding to an output vector of the matrix multiplication unit, the first plurality of digitized outputs forming a first digital output vector; performing,

by a controller, a nonlinear

multiplication unit, the input vector by a matrix that includes the one or more matrix element values.

[00282] In some implementations, the method can include encoding a set of multiple output values on respective electrical signals produced by the one or more summation modules; and representing, using the output values in the set of multiple output values, elements of an output vector that results from the input vector being multiplied by the matrix.

[00283] In another embodiment, a method includes: providing input information in an electronic format; converting at least a part of the electronic input information into an optical input vector; optoelectronically transforming the optical input vector into an analog output vector based on a matrix multiplication; and electronically applying a non-linear transformation to the analog output vector to provide output information in an electronic format.

[00284] Embodiments of the method can include one or more of the following features. For example, the method can further include: repeating the electronic-to-optical converting, the optoelectronical transforming, and the electronically applied non-linear transforming with respect to new electronic input information corresponding to the provided output information in electronic format.

[00285] In some implementations, the matrix multiplication for the initial optoelectronical transforming and the matrix multiplication of the repeated optoelectronical transforming cam be the same and correspond to the same layer of an artificial neural network.

[00286] In some implementations, the matrix multiplication for the initial optoelectronical transforming and the matrix multiplication of the repeated optoelectronical transforming can be different and correspond to different layers of an artificial neural network.

[00287] In some implementations, the method can further include: repeating the electronic-to-optical converting, the optoelectronical transforming, and the electronically applied non-linear transforming with respect to a different part of the electronic input information, wherein the matrix multiplication for the initial optoelectronical transforming and the matrix multiplication of the repeated optoelectronical transforming are the same and correspond to a first layer of an artificial neural network.

[00288] In some implementations, the method can further include: providing intermediate information in an electronic format based on the electronic output information produced for

the multiple parts of the electronic input information by the first layer of the artificial neural

network; and repeating the electronic-to-optical converting, the optoelectronical transforming, and the electronically applied non-linear transforming with respect to each of different parts of the electronic intermediate information, wherein the matrix multiplication for the initial optoelectronical transforming and the matrix multiplication of the repeated optoelectronical transforming related to the different parts of the electronic intermediate information can be the same and correspond to a second layer of the artificial neural network.

[00289] In another embodiment, a system for performing artificial neural network computations is provided. The system includes: a first unit configured to generate a plurality of vector control signals and to generate a plurality of weight control signals; a second unit configured to provide an optical input vector based on the plurality of vector control signals; and a matrix multiplication unit coupled to the second unit and the first unit, the matrix multiplication unit being configured to transform the optical input vector into an output vector based on the plurality of weight control signals. The system includes a controller including integrated circuitry configured to perform operations including: receiving an artificial neural network computation request comprising an input dataset and a first plurality of neural network weights, wherein the input dataset comprises a first digital input vector; and generating, through the first unit, a first plurality of vector control signals based on the first digital input vector and a first plurality of weight control signals based on the first plurality of neural network weights; wherein the first unit, the second unit, the matrix multiplication unit, and the controller are used in an optoelectronical processing loop that is repeated for a plurality of iterations, and the optoelectronical processing loop includes: (1) at least two optical modulation operations, and (2) at least one of (a) an electrical summation operation or (b) an electrical storage operation.

In another embodiment, a method for performing artificial neural network computations is provided. The method includes: providing input information in an electronic format; converting at least a part of the electronic input information into an optical input vector; and transforming the optical input vector into an output vector based on a matrix multiplication using a set of neural network weights. The providing, converting, and transforming are performed in an optoelectronical processing loop that is repeated for a plurality of iterations using different respective sets of neural network weights and different respective input information, and the optoelectronical processing loop includes: (1) at least two optical