(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 September 2019 (19.09.2019)

(10) International Publication Number
WO 2019/177698 A1

(51) International Patent Classification:
H03M 13/11 (2006.01) *H03M 13/29* (2006.01)
H03M 13/15 (2006.01) *H03M 13/37* (2006.01)

(21) International Application Number:
PCT/US2019/015022

(22) International Filing Date:
24 January 2019 (24.01.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/642,884 14 March 2018 (14.03.2018) US
15/991,890 29 May 2018 (29.05.2018) US

(71) Applicant: **SILICON STORAGE TECHNOLOGY, INC.** [US/US]; 450 Holger Way, San Jose, CA 95134 (US).

(72) Inventors: **TRAN, Hieu Van**; 2642 Gayley Place, San Jose, CA 95135 (US). **HONG, Stanley**; 1848 Bristol Bay Common, San Jose, CA 95131 (US). **LY, Anh**; 3385 Lindmuir Drive, San Jose, CA 95121 (US). **VU, Thuan**; 431 Danna Ct., San Jose, CA 95138 (US). **PHAM, Hien**; 736/163/18 Le Duc Tho St., Ward 15, Go Vap District, Ho Chi Minh (VN). **NGUYEN, Kha**; 350/160 Nguyen Van Luong St., Ward 16, Go Vap District, Ho Chi Minh (VN). **TRAN, Han**; Room 502, Lt. A. KhangGia Apartment, Ward 14, Go Vap District, Ho Chi Minh (VN).

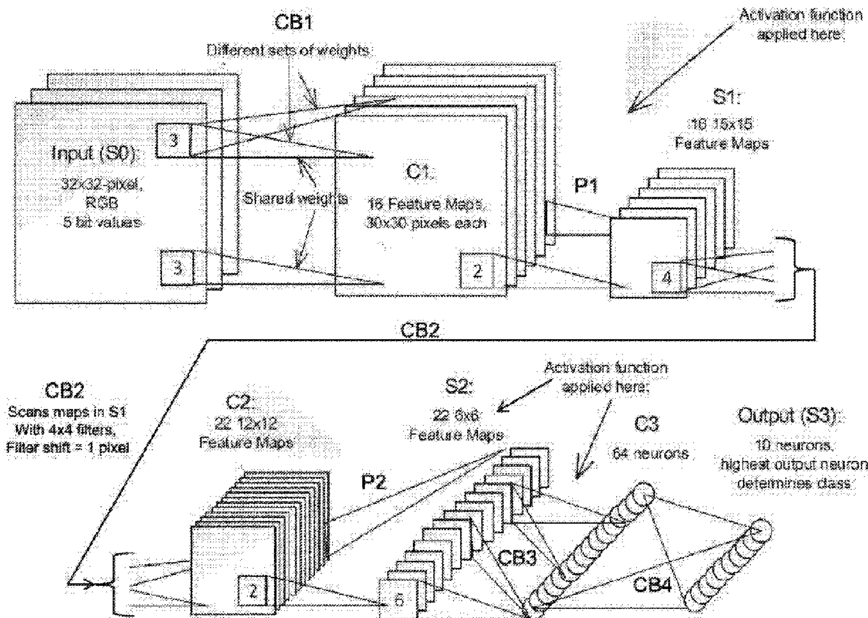
(74) Agent: **YAMASHITA, Brent**; DLA PIPER LLP US, 2000 University Avenue, East Palo Alto, CA 94303 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

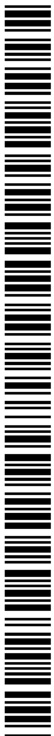
(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ,

(54) **Title:** DECODERS FOR ANALOG NEURAL MEMORY IN DEEP LEARNING ARTIFICIAL NEURAL NETWORK

FIGURE 6



(57) **Abstract:** Numerous embodiments of decoders for use with a vector-by-matrix multiplication (VMM) array in an artificial neural network are disclosed. The decoders include bit line decoders, word line decoders, control gate decoders, source line decoders, and erase gate decoders. In certain embodiments, a high voltage version and a low voltage version of a decoder is used.



WO 2019/177698 A1

[Continued on next page]

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, MK, MT, NL, NO, PL, PT, RO, RS, SE, 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))*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

DECODERS FOR ANALOG NEURAL MEMORY IN DEEP LEARNING ARTIFICIAL NEURAL NETWORK

PRIORITY CLAIMS

[0001] This application claims priority to U.S. Provisional Patent Application No. 62/642,884, filed on March 14, 2018, and titled, "Decoders for Analog Neuromorphic Memory in Artificial Neural Network," and U.S. Patent Application No. 15/991,890, filed on May 29, 2018, and titled, "Decoders For Analog Neural Memory In Deep Learning Artificial Neural Network."

FIELD OF THE INVENTION

[0002] Numerous embodiments of decoders for use with a vector-by-matrix multiplication (VMM) array in an artificial neural network are disclosed.

BACKGROUND OF THE INVENTION

[0003] Artificial neural networks mimic biological neural networks (the central nervous systems of animals, in particular the brain) which are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks generally include layers of interconnected "neurons" which exchange messages between each other.

[0004] Figure 1 illustrates an artificial neural network, where the circles represent the inputs or layers of neurons. The connections (called synapses) are represented by arrows, and have numeric weights that can be tuned based on experience. This makes neural networks adaptive to

inputs and capable of learning. Typically, neural networks include a layer of multiple inputs. There are typically one or more intermediate layers of neurons, and an output layer of neurons that provide the output of the neural network. The neurons at each level individually or collectively make a decision based on the received data from the synapses.

[0005] One of the major challenges in the development of artificial neural networks for high-performance information processing is a lack of adequate hardware technology. Indeed, practical neural networks rely on a very large number of synapses, enabling high connectivity between neurons, i.e. a very high computational parallelism. In principle, such complexity can be achieved with digital supercomputers or specialized graphics processing unit clusters. However, in addition to high cost, these approaches also suffer from mediocre energy efficiency as compared to biological networks, which consume much less energy primarily because they perform low-precision analog computation. CMOS analog circuits have been used for artificial neural networks, but most CMOS-implemented synapses have been too bulky given the high number of neurons and synapses.

[0006] Applicant previously disclosed an artificial (analog) neural network that utilizes one or more non-volatile memory arrays as the synapses in U.S. Patent Application No. 15/594,439, which is incorporated by reference. The non-volatile memory arrays operate as analog neuromorphic memory. The neural network device includes a first plurality of synapses configured to receive a first plurality of inputs and to generate therefrom a first plurality of outputs, and a first plurality of neurons configured to receive the first plurality of outputs. The first plurality of synapses includes a plurality of memory cells, wherein each of the memory cells includes spaced apart source and drain regions formed in a semiconductor substrate with a channel region extending there between, a floating gate disposed over and insulated from a first

portion of the channel region and a non-floating gate disposed over and insulated from a second portion of the channel region. Each of the plurality of memory cells is configured to store a weight value corresponding to a number of electrons on the floating gate. The plurality of memory cells is configured to multiply the first plurality of inputs by the stored weight values to generate the first plurality of outputs.

[0007] Each non-volatile memory cells used in the analog neuromorphic memory system must be erased and programmed to hold a very specific and precise amount of charge in the floating gate. For example, each floating gate must hold one of N different values, where N is the number of different weights that can be indicated by each cell. Examples of N include 16, 32, and 64.

[0008] Prior art decoding circuits (such as bit line decoders, word line decoders, control gate decoders, source line decoders, and erase gate decoders) used in conventional flash memory arrays are not suitable for use with a VMM in an analog neuromorphic memory system. One reason for this is that in a VMM system, the verify portion (which is a read operation) of a program and verify operation operates on a single selected memory cell, whereas a read operation operates on all memory cells in the array.

[0009] What is needed are improved decoding circuits suitable for use with a VMM in an analog neuromorphic memory system.

SUMMARY OF THE INVENTION

[0010] Numerous embodiments of decoders for use with a vector-by-matrix multiplication (VMM) array in an artificial neural network are disclosed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] Figure 1 is a diagram that illustrates an artificial neural network.

[0012] Figure 2 is a cross-sectional side view of a conventional 2-gate non-volatile memory cell.

[0013] Figure 3 is a cross-sectional side view of a conventional 4-gate non-volatile memory cell.

[0014] Figure 4 is a side cross-sectional side view of conventional 3-gate non-volatile memory cell.

[0015] Figure 5 is a cross-sectional side view of another conventional 2-gate non-volatile memory cell.

[0016] Figure 6 is a diagram illustrating the different levels of an exemplary artificial neural network utilizing a non-volatile memory array.

[0017] Figure 7 is a block diagram illustrating a vector multiplier matrix.

[0018] Figure 8 is a block diagram illustrating various levels of a vector multiplier matrix.

[0019] Figure 9 depicts an embodiment of a vector multiplier matrix.

[0020] Figure 10 depicts another embodiment of a vector multiplier matrix.

[0021] Figure 11 depicts another embodiment of a vector multiplier matrix.

[0022] Figure 12 depicts another embodiment of a vector multiplier matrix.

[0023] Figure 13 depicts another embodiment of a vector multiplier matrix.

[0024] Figure 14 depicts an embodiment of a bit line decoder for a vector multiplier matrix.

[0025] Figure 15 depicts another embodiment of a bit line decoder for a vector multiplier matrix.

[0026] Figure 16 depicts another embodiment of a bit line decoder for a vector multiplier matrix.

[0027] Figure 17 depicts a system for operating a vector multiplier matrix.

[0028] Figure 18 depicts another system for operating a vector multiplier matrix.

[0029] Figure 19 depicts another system for operating a vector multiplier matrix.

[0030] Figure 20 depicts an embodiment of a word line driver for use with a vector multiplier matrix.

[0031] Figure 21 depicts another embodiment of a word line driver for use with a vector multiplier matrix.

[0032] Figure 22 depicts another embodiment of a word line driver for use with a vector multiplier matrix.

[0033] Figure 23 depicts another embodiment of a word line driver for use with a vector multiplier matrix.

[0034] Figure 24 depicts another embodiment of a word line driver for use with a vector multiplier matrix.

[0035] Figure 25 depicts another embodiment of a word line driver for use with a vector multiplier matrix.

[0036] Figure 26 depicts another embodiment of a word line driver for use with a vector multiplier matrix.

[0037] Figure 27 depicts a source line decoder circuit for use with a vector multiplier matrix.

[0038] Figure 28 depicts a word line decoder circuit, a source line decoder circuit, and a high voltage level shifter for use with a vector multiplier matrix.

[0039] Figure 29 depicts an erase gate decoder circuit, a control gate decoder circuit, a source line decoder circuit, and a high voltage level shifter for use with a vector multiplier matrix.

[0040] Figure 30 depicts a word line decoder circuit for use with a vector multiplier matrix.

[0041] Figure 31 depicts a control gate decoder circuit for use with a vector multiplier matrix.

[0042] Figure 32 depicts another control gate decoder circuit for use with a vector multiplier matrix.

[0043] Figure 33 depicts another control gate decoder circuit for use with a vector multiplier matrix.

[0044] Figure 34 depicts a current-to-voltage circuit for controlling a word line in a vector multiplier matrix.

[0045] Figure 35 depicts another current-to-voltage circuit for controlling a word line in a vector multiplier matrix.

[0046] Figure 36 depicts a current-to-voltage circuit for controlling a control gate line in a vector multiplier matrix.

[0047] Figure 37 depicts another current-to-voltage circuit for controlling a control gate line in a vector multiplier matrix.

[0048] Figure 38 depicts another current-to-voltage circuit for controlling a control gate line in a vector multiplier matrix.

[0049] Figure 39 depicts another current-to-voltage circuit for controlling a word line in a vector multiplier matrix.

[0050] Figure 40 depicts another current-to-voltage circuit for controlling a word line in a vector multiplier matrix.

[0051] Figure 41 depicts another current-to-voltage circuit for controlling a word line in a vector multiplier matrix.

[0052] Figure 42 depicts operating voltages for the vector multiplier matrix of Figure 9.

[0053] Figure 43 depicts operating voltages for the vector multiplier matrix of Figure 10.

[0054] Figure 44 depicts operating voltages for the vector multiplier matrix of Figure 11.

[0055] Figure 45 depicts operating voltages for the vector multiplier matrix of Figure 12.

DETAILED DESCRIPTION OF THE INVENTION

[0056] The artificial neural networks of the present invention utilize a combination of CMOS technology and non-volatile memory arrays.

Non-Volatile Memory Cells

[0057] Digital non-volatile memories are well known. For example, U.S. Patent 5,029,130 ("the '130 patent") discloses an array of split gate non-volatile memory cells, and is incorporated herein by reference for all purposes. Such a memory cell is shown in Figure 2. Each memory cell 210 includes source region 14 and drain region 16 formed in a semiconductor substrate 12, with a channel region 18 there between. A floating gate 20 is formed over and insulated from (and controls the conductivity of) a first portion of the channel region 18, and over a portion of the source region 16. A word line terminal 22 (which is typically coupled to a word line) has a first portion that is disposed over and insulated from (and controls the conductivity of) a second portion of the channel region 18, and a second portion that extends up and over the floating gate 20. The floating gate 20 and word line terminal 22 are insulated from the substrate 12 by a gate oxide. Bitline 24 is coupled to drain region 16.

[0058] Memory cell 210 is erased (where electrons are removed from the floating gate) by placing a high positive voltage on the word line terminal 22, which causes electrons on the floating gate 20 to tunnel through the intermediate insulation from the floating gate 20 to the word line terminal 22 via Fowler-Nordheim tunneling.

[0059] Memory cell 210 is programmed (where electrons are placed on the floating gate) by placing a positive voltage on the word line terminal 22, and a positive voltage on the source 16. Electron current will flow from the source 16 towards the drain 14. The electrons will accelerate and become heated when they reach the gap between the word line terminal 22 and the floating

gate 20. Some of the heated electrons will be injected through the gate oxide 26 onto the floating gate 20 due to the attractive electrostatic force from the floating gate 20.

[0060] Memory cell 210 is read by placing positive read voltages on the drain 14 and word line terminal 22 (which turns on the channel region under the word line terminal). If the floating gate 20 is positively charged (i.e. erased of electrons and positively coupled to the drain 16), then the portion of the channel region under the floating gate 20 is turned on as well, and current will flow across the channel region 18, which is sensed as the erased or "1" state. If the floating gate 20 is negatively charged (i.e. programmed with electrons), then the portion of the channel region under the floating gate 20 is mostly or entirely turned off, and current will not flow (or there will be little flow) across the channel region 18, which is sensed as the programmed or "0" state.

[0061] Table No. 1 depicts typical voltage ranges that can be applied to the terminals of memory cell 210 for performing read, erase, and program operations:

Table No. 1: Operation of Flash Memory Cell 210 of Figure 2

	WL	BL	SL
Read	2-3V	0.6-2V	0V
Erase	~11-13V	0V	0V
Program	1-2V	1-3μA	9-10V

[0062] Other split gate memory cell configurations are known. For example, Figure 3 depicts four-gate memory cell 310 comprising source region 14, drain region 16, floating gate 20 over a first portion of channel region 18, a select gate 28 (typically coupled to a word line) over a

second portion of the channel region 18, a control gate 22 over the floating gate 20, and an erase gate 30 over the source region 14. This configuration is described in U.S. Patent 6,747,310, which is incorporated herein by reference for all purposes). Here, all gates are non-floating gates except floating gate 20, meaning that they are electrically connected or connectable to a voltage source. Programming is shown by heated electrons from the channel region 18 injecting themselves onto the floating gate 20. Erasing is shown by electrons tunneling from the floating gate 20 to the erase gate 30.

[0063] Table No. 2 depicts typical voltage ranges that can be applied to the terminals of memory cell 310 for performing read, erase, and program operations:

Table No. 2: Operation of Flash Memory Cell 310 of Figure 3

	WL/SG	BL	CG	EG	SL
Read	1.0-2V	0.6-2V	0-2.6V	0-2.6V	0V
Erase	-0.5V/0V	0V	0V/-8V	8-12V	0V
Program	1V	1μA	8-11V	4.5-9V	4.5-5V

[0064] Figure 4 depicts split gate three-gate memory cell 410. Memory cell 410 is identical to the memory cell 310 of Figure 3 except that memory cell 410 does not have a separate control gate. The erase operation (erasing through erase gate) and read operation are similar to that of the Figure 3 except there is no control gate bias. The programming operation also is done without the control gate bias, hence the program voltage on the source line is higher to compensate for lack of control gate bias.

[0065] Table No. 3 depicts typical voltage ranges that can be applied to the terminals of memory cell 410 for performing read, erase, and program operations:

Table No. 3: Operation of Flash Memory Cell 410 of Figure 4

	WL/SG	BL	EG	SL
Read	0.7-2.2V	0.6-2V	0-2.6V	0V
Erase	-0.5V/0V	0V	11.5V	0V
Program	1V	2-3μA	4.5V	7-9V

[0066] Figure 5 depicts stacked gate memory cell 510. Memory cell 510 is similar to memory cell 210 of Figure 2, except floating gate 20 extends over the entire channel region 18, and control gate 22 extends over floating gate 20, separated by an insulating layer. The erase, programming, and read operations operate in a similar manner to that described previously for memory cell 210.

[0067] Table No. 4 depicts typical voltage ranges that can be applied to the terminals of memory cell 510 for performing read, erase, and program operations:

Table No. 4 Operation of Flash Memory Cell 510 of Figure 5

	CG	BL	SL	P-sub
Read	2-5V	0.6 – 2V	0V	0V
Erase	-8 to -10V/0V	FLT	FLT	8-10V / 15-20V
Program	8-12V	3-5V	0V	0V

[0068] In order to utilize the memory arrays comprising one of the types of non-volatile memory cells described above in an artificial neural network, two modifications are made. First, the lines are configured so that each memory cell can be individually programmed, erased, and read without adversely affecting the memory state of other memory cells in the array, as further explained below. Second, continuous (analog) programming of the memory cells is provided.

[0069] Specifically, the memory state (i.e. charge on the floating gate) of each memory cells in the array can be continuously changed from a fully erased state to a fully programmed state, independently and with minimal disturbance of other memory cells. In another embodiment, the memory state (*i.e.*, charge on the floating gate) of each memory cell in the array can be continuously changed from a fully programmed state to a fully erased state, and vice-versa, independently and with minimal disturbance of other memory cells. This means the cell storage is analog or at the very least can store one of many discrete values (such as 16 or 64 different values), which allows for very precise and individual tuning of all the cells in the memory array, and which makes the memory array ideal for storing and making fine tuning adjustments to the synapsis weights of the neural network.

Neural Networks Employing Non-Volatile Memory Cell Arrays

[0070] Figure 6 conceptually illustrates a non-limiting example of a neural network utilizing a non-volatile memory array. This example uses the non-volatile memory array neural net for a facial recognition application, but any other appropriate application could be implemented using a non-volatile memory array based neural network.

[0071] S0 is the input, which for this example is a 32x32 pixel RGB image with 5 bit precision (i.e. three 32x32 pixel arrays, one for each color R, G and B, each pixel being 5 bit precision). The synapses CB1 going from S0 to C1 have both different sets of weights and shared weights,

and scan the input image with 3x3 pixel overlapping filters (kernel), shifting the filter by 1 pixel (or more than 1 pixel as dictated by the model). Specifically, values for 9 pixels in a 3x3 portion of the image (i.e., referred to as a filter or kernel) are provided to the synapses CB1, whereby these 9 input values are multiplied by the appropriate weights and, after summing the outputs of that multiplication, a single output value is determined and provided by a first neuron of CB1 for generating a pixel of one of the layers of feature map C1. The 3x3 filter is then shifted one pixel to the right (i.e., adding the column of three pixels on the right, and dropping the column of three pixels on the left), whereby the 9 pixel values in this newly positioned filter are provided to the synapses CB1, whereby they are multiplied by the same weights and a second single output value is determined by the associated neuron. This process is continued until the 3x3 filter scans across the entire 32x32 pixel image, for all three colors and for all bits (precision values). The process is then repeated using different sets of weights to generate a different feature map of C1, until all the features maps of layer C1 have been calculated.

[0072] At C1, in the present example, there are 16 feature maps, with 30x30 pixels each. Each pixel is a new feature pixel extracted from multiplying the inputs and kernel, and therefore each feature map is a two dimensional array, and thus in this example the synapses CB1 constitutes 16 layers of two dimensional arrays (keeping in mind that the neuron layers and arrays referenced herein are logical relationships, not necessarily physical relationships – i.e., the arrays are not necessarily oriented in physical two dimensional arrays). Each of the 16 feature maps is generated by one of sixteen different sets of synapse weights applied to the filter scans. The C1 feature maps could all be directed to different aspects of the same image feature, such as boundary identification. For example, the first map (generated using a first weight set, shared for all scans used to generate this first map) could identify circular edges, the second map (generated

using a second weight set different from the first weight set) could identify rectangular edges, or the aspect ratio of certain features, and so on.

[0073] An activation function P1 (pooling) is applied before going from C1 to S1, which pools values from consecutive, non-overlapping 2x2 regions in each feature map. The purpose of the pooling stage is to average out the nearby location (or a max function can also be used), to reduce the dependence of the edge location for example and to reduce the data size before going to the next stage. At S1, there are 16 15x15 feature maps (i.e., sixteen different arrays of 15x15 pixels each). The synapses and associated neurons in CB2 going from S1 to C2 scan maps in S1 with 4x4 filters, with a filter shift of 1 pixel. At C2, there are 22 12x12 feature maps. An activation function P2 (pooling) is applied before going from C2 to S2, which pools values from consecutive non-overlapping 2x2 regions in each feature map. At S2, there are 22 6x6 feature maps. An activation function is applied at the synapses CB3 going from S2 to C3, where every neuron in C3 connects to every map in S2. At C3, there are 64 neurons. The synapses CB4 going from C3 to the output S3 fully connects S3 to C3. The output at S3 includes 10 neurons, where the highest output neuron determines the class. This output could, for example, be indicative of an identification or classification of the contents of the original image.

[0074] Each level of synapses is implemented using an array, or a portion of an array, of non-volatile memory cells. Figure 7 is a block diagram of the vector-by-matrix multiplication (VMM) array that includes the non-volatile memory cells, and is utilized as the synapses between an input layer and the next layer. Specifically, the VMM 32 includes an array of non-volatile memory cells 33, erase gate and word line gate decoder 34, control gate decoder 35, bit line decoder 36 and source line decoder 37, which decode the inputs for the memory array 33. Source line decoder 37 in this example also decodes the output of the memory cell array.

Alternatively, bit line decoder 36 can decode the output of the memory array. The memory array serves two purposes. First, it stores the weights that will be used by the VMM. Second, the memory array effectively multiplies the inputs by the weights stored in the memory array and adds them up per output line (source line or bit line) to produce the output, which will be the input to the next layer or input to the final layer. By performing the multiplication and addition function, the memory array negates the need for separate multiplication and addition logic circuits and is also power efficient due to in-situ memory computation.

[0075] The output of the memory array is supplied to a differential summer (such as summing op-amp) 38, which sums up the outputs of the memory cell array to create a single value for that convolution. The differential summer is such as to realize summation of positive weight and negative weight with positive input. The summed up output values are then supplied to the activation function circuit 39, which rectifies the output. The activation function may include sigmoid, tanh, or ReLU functions. The rectified output values become an element of a feature map as the next layer (C1 in the description above for example), and are then applied to the next synapse to produce next feature map layer or final layer. Therefore, in this example, the memory array constitutes a plurality of synapses (which receive their inputs from the prior layer of neurons or from an input layer such as an image database), and summing op-amp 38 and activation function circuit 39 constitute a plurality of neurons.

[0076] Figure 8 is a block diagram of the various levels of VMM. As shown in Figure 14, the input is converted from digital to analog by digital-to-analog converter 31, and provided to input VMM 32a. The output generated by the input VMM 32a is provided as an input to the next VMM (hidden level 1) 32b, which in turn generates an output that is provided as an input to the next VMM (hidden level 2) 32b, and so on. The various layers of VMM's 32 function as

different layers of synapses and neurons of a convolutional neural network (CNN). Each VMM can be a stand-alone non-volatile memory array, or multiple VMMs could utilize different portions of the same non-volatile memory array, or multiple VMMs could utilize overlapping portions of the same non-volatile memory array. The example shown in Figure 8 contains five layers (32a,32b,32c,32d,32e): one input layer (32a), two hidden layers (32b,32c), and two fully connected layers (32d,32e). One of ordinary skill in the art will appreciate that this is merely exemplary and that a system instead could comprise more than two hidden layers and more than two fully connected layers.

Vector-by-Matrix Multiplication (VMM) Arrays

[0077] Figure 9 depicts neuron VMM 900, which is particularly suited for memory cells of the type shown in Figure 2, and is utilized as the synapses and parts of neurons between an input layer and the next layer. VMM 900 comprises a memory array 903 of non-volatile memory cells, reference array 901, and reference array 902. Reference arrays 901 and 902 serve to convert current inputs flowing into terminals BLR0-3 into voltage inputs WL0-3. Reference arrays 901 and 902 as shown are in the column direction. In general, the reference array direction is orthogonal to the input lines. In effect, the reference memory cells are diode connected through multiplexors (multiplexor 914, which includes a multiplexor and a cascoding transistor VBLR for biasing the reference bit line) with current inputs flowing into them. The reference cells are tuned to target reference levels.

[0078] Memory array 903 serves two purposes. First, it stores the weights that will be used by the VMM 900. Second, memory array 903 effectively multiplies the inputs (current inputs provided in terminals BLR0-3; reference arrays 901 and 902 convert these current inputs into the input voltages to supply to wordlines WL0-3) by the weights stored in the memory array to

produce the output, which will be the input to the next layer or input to the final layer. By performing the multiplication function, the memory array negates the need for separate multiplication logic circuits and is also power efficient. Here, the voltage inputs are provided on the word lines, and the output emerges on the bit line during a read (inference) operation. The current placed on the bit line performs a summing function of all the currents from the memory cells connected to the bitline.

[0079] Figure 42 depicts operating voltages for VMM 900. The columns in the table indicate the voltages placed on word lines for selected cells, word lines for unselected cells, bit lines for selected cells, bit lines for unselected cells, source lines for selected cells, and source lines for unselected cells. The rows indicate the operations of read, erase, and program.

[0080] Figure 10 depicts neuron VMM 1000, which is particularly suited for memory cells of the type shown in Figure 2, and is utilized as the synapses and parts of neurons between an input layer and the next layer. VMM 1000 comprises a memory array 1003 of non-volatile memory cells, reference array 1001, and reference array 1002. VMM 1000 is similar to VMM 900 except that in VMM 1000 the word lines run in the vertical direction. There are two reference arrays 1001 (at the top, which provides a reference converting input current into voltage for the even rows) and 1002 (at the bottom, which provides a reference converting input current into voltage for the odd rows). Here, the inputs are provided on the word lines, and the output emerges on the source line during a read operation. The current placed on the source line performs a summing function of all the currents from the memory cells connected to the source line.

[0081] Figure 43 depicts operating voltages for VMM 1000. The columns in the table indicate the voltages placed on word lines for selected cells, word lines for unselected cells, bit lines for

selected cells, bit lines for unselected cells, source lines for selected cells, and source lines for unselected cells. The rows indicate the operations of read, erase, and program.

[0082] Figure 11 depicts neuron VMM 1100, which is particularly suited for memory cells of the type shown in Figure 3, and is utilized as the synapses and parts of neurons between an input layer and the next layer. VMM 1100 comprises a memory array 1101 of non-volatile memory cells, reference array 1102 (providing reference converting input current into input voltage for even rows), and reference array 1103 (providing reference converting input current into input voltage for odd rows). VMM 1100 is similar to VMM 900 except VMM 1100 further comprises control line 1106 couples to the control gates of a row of memory cells and control line 1107 coupled to the erase gates of adjoining rows of memory cells. Here, the wordlines, control gate lines, and erase gate lines are of the same direction. VMM further comprises reference bit line select transistor 1104 (part of mux 1114) that selectively couples a reference bit line to the bit line contact of a selected reference memory cell and switch 1105 (part of mux 1114) that selectively couples a reference bit line to control line 1106 for a particular selected reference memory cell. Here, the inputs are provided on the word lines (of memory array 1101), and the output emerges on the bit line, such as bit line 1109, during a read operation. The current placed on the bit line performs a summing function of all the currents from the memory cells connected to the bit line.

[0083] Figure 44 depicts operating voltages for VMM 1100. The columns in the table indicate the voltages placed on word lines for selected cells, word lines for unselected cells, bit lines for selected cells, bit lines for unselected cells, control gates for selected cells, control gates for unselected cells in the same sector as the selected cells, control gates for unselected cells in a different sector than the selected cells, erase gates for selected cells, erase gates for unselected

cells, source lines for selected cells, and source lines for unselected cells. The rows indicate the operations of read, erase, and program.

[0084] Figure 12 depicts neuron VMM 1200, which is particularly suited for memory cells of the type shown in Figure 3, and is utilized as the synapses and parts of neurons between an input layer and the next layer. VMM 1200 is similar to VMM 1100, except in VMM 1200, erase gate lines such as erase gate line 1201 run in a vertical direction. Here, the inputs are provided on the word lines, and the output emerges on the source lines. The current placed on the bit line performs a summing function of all the currents from the memory cells connected to the bit line.

[0085] Figure 45 depicts operating voltages for VMM 1200. The columns in the table indicate the voltages placed on word lines for selected cells, word lines for unselected cells, bit lines for selected cells, bit lines for unselected cells, control gates for selected cells, control gates for unselected cells in the same sector as the selected cells, control gates for unselected cells in a different sector than the selected cells, erase gates for selected cells, erase gates for unselected cells, source lines for selected cells, and source lines for unselected cells. The rows indicate the operations of read, erase, and program.

[0086] Figure 13 depicts neuron VMM 1300, which is particularly suited for memory cells of the type shown in Figure 3, and is utilized as the synapses and parts of neurons between an input layer and the next layer. VMM 1300 comprises a memory array 1301 of non-volatile memory cells and reference array 1302 (at the top of the array). Alternatively, another reference array can be placed at the bottom, similar to that of Figure 10. In other respects, VMM 1300 is similar to VMM 1200, except in VMM 1300, control gates line such as control gate line 1303 run in a vertical direction (hence reference array 1302 in the row direction, orthogonal to the input control gate lines), and erase gate lines such as erase gate line 1304 run in a horizontal direction.

Here, the inputs are provided on the control gate lines, and the output emerges on the source lines. In one embodiment only even rows are used, and in another embodiment, only odd rows are used. The current placed on the source line performs a summing function of all the currents from the memory cells connected to the source line.

[0087] As described herein for neural networks, the flash cells are preferably configured to operate in sub-threshold region.

[0088] The memory cells described herein are biased in weak inversion:

$$Ids = Io * e^{(Vg - Vth)/kVt} = w * Io * e^{(Vg)/kVt}$$

$$w = e^{(-Vth)/kVt}$$

[0089] For an I-to-V log converter using a memory cell to convert input current into an input voltage:

$$Vg = k*Vt*\log [Ids/wp*Io]$$

[0090] For a memory array used as a vector matrix multiplier VMM, the output current is:

$$Iout = wa * Io * e^{(Vg)/kVt}, \text{ namely}$$

$$Iout = (wa/wp) * Iin = W * Iin$$

$$W = e^{(Vthp - Vtha)/kVt}$$

[0091] A wordline or control gate can be used as the input for the memory cell for the input voltage.

[0092] Alternatively, the flash memory cells can be configured to operate in the linear region:

$$Ids = beta * (Vgs - Vth) * Vds \; ; \; beta = u*Cox*W/L$$

$$W \;\alpha\; (Vgs - Vth)$$

[0093] For an I-to-V linear converter, a memory cell operating in the linear region can be used to convert linearly an input/output current into an input/output voltage.

[0094] Other embodiments for the ESF vector matrix multiplier are as described in U.S. Patent Application No. Application No. 15/826,345, which is incorporated by reference herein. A sourceline or a bitline can be used as the neuron output (current summation output).

[0095] Figure 14 depicts an embodiment of bit line decoder circuit 1400. Bit line decoder circuit 1400 comprises column decoder 1402 and analog neuromorphic neuron ("ANN") column decoder 1403, each of which is coupled to VMM array 1401. VMM array can be based on any of the VMM design discussed previously (such as VMM 900, 1000, 1100, 1200, and 1300) or other VMM designs.

[0096] One challenge with analog neuromorphic systems is that the system must be able to program and verify (which involves a read operation) individual selected cells, and it also must be able to perform an ANN read where all of the cells in the array are selected and read. In other words, a bit line decoder must sometimes select only one bit line and in other instances must select all bit lines.

[0097] Bit line decoder circuit 1400 accomplishes this purpose. Column decoder 1402 is a conventional column decoder (program and erase, or PE, decoding path) and can be used to select an individual bit line such as for program and program verify (a sensing operation). Outputs of the column decoder 1402 are coupled to program/erase (PE) column driver circuit for controlling program, PE verify, and erase (not shown in Fig. 14). ANN column decoder 1403 is a column decoder that is specifically designed to enable a read operation on every bit line at the same time. ANN column decoder 1403 comprises exemplary select transistor 1405 and output circuit (e.g., current summer and activation function such as tanh, sigmoid, ReLU) 1406 coupled to a bit line (here, BL0). A set of the same devices is attached to each of the other bit lines. All of the select transistors, such as select transistor 1405, is coupled to select line 1404. During an

ANN read operation, select line 1404 is enabled, which turns on each of the select transistors such as select transistor 1405, which then causes current from each bit line to be received by an output circuit such as circuit 1406 and output.

[0098] Figure 15 depicts an embodiment of bit line decoder circuit 1500. Bit line decoder circuit 1500 is coupled to VMM array 1501. VMM array can be based on any of the VMM design discussed previously (such as VMM 900, 1000, 1100, 1200, and 1300) or other VMM designs.

[0099] Select transistors 1502 and 1503 are controlled by a pair of complementary control signals (V0 and VB_0) and are coupled to a bit line (BL0). Select transistors 1504 and 1505 are controlled by another pair of complementary control signals (V1 and VB_1) and are coupled to another bit line (BL1). Select transistors 1502 and 1504 are coupled to the same output such as for enabling programming and select transistors 1503 and 1505 are coupled to the same output such as for inhibit programming. The output lines of the transistors 1502/1503/1504/1505 (program and erase PE decoding path) are such as coupled to a PE column driver circuit for controlling program, PE verify, and erase (not shown).

[00100] Select transistor 1506 is coupled to a bit line (BL0) and to output and activation function circuit 1507 (e.g., current summer and activation function such as tanh, sigmoid, ReLU). Select transistor 1506 is controlled by control line 1508.

[00101] When only BL0 is to be activated, control line 1508 is de-asserted and signal V0 is asserted, thus reading BL0 only. During an ANN read operation, control line 1508 is asserted, select transistor 1506 and similar transistors are turned on, and all bit lines are read such as for all neuron processing.

[00102] Figure 16 depicts an embodiment of bit line decoder circuit 1600. Bit line decoder circuit 1600 is coupled to VMM array 1601. VMM array can be based on any of the VMM

design discussed previously (such as VMM 900, 1000, 1100, 1200, and 1300) or other VMM designs.

[00103] Select transistor 1601 is coupled to a bit line (BL0) and to output and activation function circuit 1603. Select transistor 1602 is coupled to a bit line (BL0) and to a common output (PE decoding path).

[00104] When only BL0 is to be activated, select transistor 1602 is activated, and BL0 is attached to the common output. During an ANN read operation, select transistor 1601 and similar transistors are turned on, and all bit lines are read.

[00105] For the decoding in the Figures 14,15, and 16, for an un-selected transistor, a negative bias can be applied to reduce the transistor leakage from affecting the memory cell performance. Or a negative bias can be applied to the PE decoding path while the array is in the ANN operation. The negative bias can be from -0.1V to -0.5V or more.

[00106] Figure 17 depicts VMM system 1700. VMM system 1700 comprises VMM array 1701 and reference array 1720 (which can be based on any of the VMM design discussed previously, such as VMM 900, 1000, 1100, 1200, and 1300, or other VMM designs), low voltage row decoder 1702, high voltage row decoder 1703, reference cell low voltage column decoder 1704 (shown for the reference array in the column direction, meaning providing input to output conversion in the row direction), bit line PE driver 1712, bit line multiplexor 1706, activation function circuit and summer 1707, control logic 1705, and analog bias circuit 1708.

[00107] As shown, the reference cell low voltage column decoder 1704 is for the reference array 1720 in the column direction, meaning providing input to output conversion in the row direction. If the reference array is in the row direction, the reference decoder needs to be done on top and/or bottom of the array, to providing input to output conversion in the column direction.

[00108] Low voltage row decoder 1702 provides a bias voltage for read and program operations and provides a decoding signal for high voltage row decoder 1703. High voltage row decoder 1703 provides a high voltage bias signal for program and erase operations. Reference cell low voltage column decoder 1704 provides a decoding function for the reference cells. Bit line PE driver 1712 provides controlling function for bit line in program, verify, and erase. Bias circuit 1705 is a shared bias block that provides the multiple voltages needed for the various program, erase , program verify, and read operations.

[00109] Figure 18 depicts VMM system 1800. VMM system 1800 is similar to VMM system 1700, except that VMM system 1800 further comprises red array 1801, bit line PE driver BLDRV 1802, high voltage column decoder 1803, NVR sectors 1804, and reference array 1820. High voltage column decoder 1803 provides a high voltage bias for vertical decoding lines. Red array 1802 provides array redundancy for replacing a defective array portion. NVR (non-volatile register aka info sector) sectors 1804 are sectors that are array sectors used to store user info, device ID, password, security key, trimbits, configuration bits, manufacturing info, etc.

[00110] Figure 19 depicts VMM system 1900. VMM system 1900 is similar to VMM system 1800, except that VMM system 1900 further comprises reference system 1999. Reference system 1999 comprises reference array 1901, reference array low voltage row decoder 1902, reference array high voltage row decoder 1903, and reference array low voltage column decoder 1904. The reference system can be shared across multiple VMM systems. VMM system further comprises NVR sectors 1905.

[00111] Reference array low voltage row decoder 1902 provides a bias voltage for read and programming operations involving reference array 1901 and also provides a decoding signal for reference array high voltage row decoder 1903. Reference array high voltage row decoder 1903

provides a high voltage bias for program and operations involving reference array 1901. Reference array low voltage column decoder 1904 provides a decoding function for reference array 1901. Reference array 1901 is such as to provide reference target for program verify or cell margining (searching for marginal cells).

[00112] Figure 20 depicts word line driver 2000. Word line driver 2000 selects a word line (such as exemplary word lines WL0, WL1, WL2, and WL3 shown here) and provides a bias voltage to that word line. Each word line is attached to a select transistor, such as select iso (isolation) transistor 2002, that is controlled by control line 2001. Iso transistor 2002 is used to isolate the high voltage such as from erase (e.g., 8-12V) from word line decoding transistors, which can be implemented with IO transistors (e.g., 1.8V, 3.3V). Here, during any operation, control line 2001 is activated and all select transistors similar to select iso transistor 2002 are turned on. Exemplary bias transistor 2003 (part of wordline decoding circuit) selectively coupled a word line to a first bias voltage (such as 3V) and exemplary bias transistor 2004 (part of wordline decoding circuit) selectively coupled a word line to a second bias voltage (lower than the first bias voltage, including ground, a bias in between, a negative voltage bias to reduce leakage from un-used memory rows). During an ANN read operation, all used word lines will be selected and tied to the first bias voltage. All un-used wordlines are tied to the second bias voltage. During other operations such as for program operation, only one word line will be selected and the other word lines till be tied to the second bias voltage, which can be a negative bias (e.g., -0.3 to -0.5V or more) to reduce array leakage.

[00113] Figure 21 depicts word line driver 2100. Word line driver 2100 is similar to word line driver 2000, except that the top transistor such as bias transistor 2103 can be individually coupled to a bias voltage, and all such transistors are not tied together as in word line driver

2000. This allows all wordline to have different independent voltages in parallel at the same times.

[00114] Figure 22 depicts word line driver 2200. Word line driver 2200 is similar to word line driver 2100, except that bias transistors 2103 and 2104 are coupled to decoder circuit 2201 and inverter 2202. Thus, Figure 22 depicts a decoding sub-circuit 2203 within word line driver 2200.

[00115] Figure 23 depicts word line driver 2300. Word line driver 2300 is similar to word line driver 2100, except that bias transistors 2103 and 2104 are coupled to the outputs of stage 2302 of shift register 2301. The shift register 1301 allows by serial shifting in data (serially clocking the registers) to control each row independently, such as enabling one or more rows to be enabled at the same times depending on the shifted in data pattern.

[00116] Figure 24 depicts word line driver 2400. Word line driver 2400 is similar to word line driver 2000, except that each select transistor is further coupled to a capacitor, such as capacitor 2403. Capacitor 2403 can provide a pre-charge or bias to the word line at the beginning of an operation, enabled by transistor 2401 to sample voltage on line 2440. Capacitor 2403 acts to sample and hold (S/H) the input voltage for each wordline. Transistors 2401 are off during the ANN operation (array current summer and activation function) of the VMM array, meaning that the voltage on the S/H capacitor will serve as a (floating) voltage source for the wordline. Alternatively, capacitor 2403 can be provided by the word line capacitance from the memory array.

[00117] Figure 25 depicts word line driver 2500. Word line driver 2500 is similar to previously-described word line drivers, except that bias transistors 2501 and 2502 are connected to switches 2503 and 2504, respectively. Switch 2503 receives the output of opa (operational

amplifier) 2505, and switch 2504 provides a reference input to negative input of the opa 2505, which essentially provides the voltage stored by capacitor 2403 by action of closed loop provided by the opa 2505, the transistor 2501, the switches 2503 and 2504. In this manner, when switches 2503 and 2504 are closed , the voltage on the input 2506 is superimposed on the capacitor 2403 by the transistor 2501. Alternatively, capacitor 2403 can be provided by the word line capacitance from the memory array.

[00118] Figure 26 depicts word line driver. Word line driver 2600 is similar to previously-described word line drivers except for the addition of amplifier 2601, which will acts as a voltage buffer for the voltage on the capacitor 2604 to drive the voltage into the wordline WL0, meaning that the voltage on the S/H capacitor will serve as a (floating) voltage source for the wordline. This is for example to avoid the wordline to wordline coupling from affecting the voltage on the capacitor.

[00119] Figure 27 depicts high voltage source line decoder circuit 2700. High voltage source line decoder circuit comprises transistors 2701, 2702, and 2703, configured as shown. Transistor 2703 is used to de-select the source line to a low voltage. Transistor 2702 is used to drive a high voltage into the source line of the array and transistor 2701 is used to monitor the voltage on the source line. Transistors 2702, 2701 and a driver circuit (such as an opa) is configured in a closed loop fashion (force/sense) to maintain the voltage over PVT (process, voltage, temperature) and varied current load condition. SLE (driven source line node) and SLB (monitored source line node) can be at the one end of a source line. Alternatively SLE can be at one end and SLN at the other end of a source line.

[00120] Figure 28 depicts VMM high voltage decode circuits, comprising word line decoder circuit 2801, source line decoder circuit 2804, and high voltage level shifter 2808, which are appropriate for use with memory cells of the type shown in Figure 2.

[00121] Word line decoder circuit 2801 comprises PMOS select transistor 2802 (controlled by signal HVO_B) and NMOS de-select transistor 2803 (controlled by signal HVO_B) configured as shown.

[00122] Source line decoder circuit 2804 comprises NMOS monitor transistors 2805 (controlled by signal HVO) , driving transistor 2806 (controlled by signal HVO), and de-select transistor 2807 (controlled by signal HVO_B), configured as shown.

[00123] High voltage level shifter 2808 received enable signal EN and outputs high voltage signal HV and its complement HVO_B.

[00124] Figure 29 depicts VMM high voltage decode circuits, comprising erase gate decoder circuit 2901, control gate decoder circuit 2904, source line decoder circuit 2907, and high voltage level shifter 2911, which are appropriate for use with memory cells of the type shown in Figure 3.

[00125] Erase gate decoder circuit 2901 and control gate decoder circuit 2904 use the same design as word line decoder circuit 2801 in Figure 28.

[00126] Source line decoder circuit 2907 uses the same design as source line decoder circuit 2804 in Figure 28.

[00127] High voltage level shifter 2911 uses the same design as high voltage level shifter 2808 in Figure 28.

[00128] Figure 30 depicts word line decoder 300 for exemplary word lines WL0, WL1, WL2, and WL3. Exemplary word line WL0 is coupled to pull-up transistor 3001 and pull-down

transistor 3002. When pull-up transistor 3001 is activated, WL0 is enabled. When pull-down transistor 3002 is activated, WL0 is disabled. The function of Figure 30 is similarly to that of the Figure 21 without the isolation transistors.

[00129] Figure 31 depicts control gate decoder 3100 for exemplary control gate lines CG0, CG1, CG2, and CG3. Exemplary control gate line CG0 is coupled to pull-up transistors 3101 and pull-down transistor 3102. When pull-up transistor 3101 is activated, CG0 is enabled. When pull-down transistor 3102 is activated, CG0 is disabled. The select and de-selection function of Figure 31 is similarly to that of the Figure 30 for the control gates.

[00130] Figure 32 depicts control gate decoder 3200 for exemplary control gate lines CG0, CG1, CG2, and CG3. Control gate decoder 3200 is similar to control gate decoder 3100 except that control gate decoder 3200 contains a capacitor, such as capacitor 3203, coupled to each control gate line. These sample and hold (S/H) capacitors can provide a pre-charge bias on each control gate line prior to an operation, meaning that the voltage on the S/H capacitor will serve as a (floating) voltage source for the control gate lines. The S/H capacitor can be provided by the control gate capacitance from memory cell.

[00131] Figure 33 depicts control gate decoder 3300 for exemplary control gate lines CG0, CG1, CG2, and CG3. Control gate decoder 3300 is similar to control gate decoder 3200 except that control gate decoder 3300 further comprises a buffer 3301 (such as an opa).

[00132] Figure 34 depicts current-to-voltage circuit 3400. The circuit comprises a configured diode connected reference cell circuit 3450 and a sample and hold circuit 3460. The circuit 3450 comprises input current source 3401, NMOS transistor 3402, cascoding bias transistor 3403, and reference memory cell 3404. The sample and hold circuit consists of switch 3405, and S/H

capacitor 3406. The memory 3404 is biased in a diode connected configuration with a bias on its bit line to convert the input current into a voltage, such as for supplying the word line.

[00133] Figure 35 depicts current-to-voltage circuit 3500, which is similar to current-to-voltage circuit 3400 with the addition of amplifier 3501 after the S/H capacitor. Current-to-voltage circuit 3500 comprises a configured diode connected reference cell circuit 3550, sample and hold circuit 3470, and amplifier stage 3562.

[00134] Figure 36 depicts current-to-voltage circuit 3600, which is the same design as current-to-voltage circuit 3400 for the control gate in a diode connected configuration. Current-to-voltage circuit 3600 comprises a configured diode connected reference cell circuit 3650 and a sample and hold circuit 3660.

[00135] Figure 37 depicts current-to-voltage circuit 3700, in which a buffer 3790 is placed between reference circuit 3750 and the S/H circuit 3760.

[00136] Figure 38 depicts current-to-voltage circuit 3800 which is similar to Figure 35 with a control gate connected in a diode connected configuration. Current-to-voltage circuit 3800 comprises a configured diode connected reference cell circuit 3550, sample and hold circuit 3870, and amplifier stage 3862.

[00137] Figure 39 depicts current-to-voltage circuit 3900 which is similar to Figure 34 as applied to memory cell in Figure 2. Current-to-voltage circuit 3900 comprises a configured diode connected reference cell circuit 3950 and a sample and hold circuit 3960.

[00138] Figure 40 depicts current-to-voltage circuit 4000 which is similar to Figure 37 as applied to memory cell in Figure 2, , in which a buffer 4090 is placed between reference circuit 4050 and the S/H circuit 4060.

[00139] Figure 41 depicts current-to-voltage circuit 4100 which is similar to Figure 38 as applied to memory cell in Figure 2. Current-to-voltage circuit 4100 comprises a configured diode connected reference cell circuit 4150, sample and hold circuit 4170, and amplifier stage 4162.

[00140] It should be noted that, as used herein, the terms "over" and "on" both inclusively include "directly on" (no intermediate materials, elements or space disposed therebetween) and "indirectly on" (intermediate materials, elements or space disposed therebetween). Likewise, the term "adjacent" includes "directly adjacent" (no intermediate materials, elements or space disposed therebetween) and "indirectly adjacent" (intermediate materials, elements or space disposed there between), "mounted to" includes "directly mounted to" (no intermediate materials, elements or space disposed there between) and "indirectly mounted to" (intermediate materials, elements or spaced disposed there between), and "electrically coupled" includes "directly electrically coupled to" (no intermediate materials or elements there between that electrically connect the elements together) and "indirectly electrically coupled to" (intermediate materials or elements there between that electrically connect the elements together). For example, forming an element "over a substrate" can include forming the element directly on the substrate with no intermediate materials/elements therebetween, as well as forming the element indirectly on the substrate with one or more intermediate materials/elements there between.

What is claimed is:

1. A bit line decoder circuit coupled to a vector-by-matrix multiplication array, the vector-by matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns, wherein each column is connected to a bit line, the bit line decoder circuit comprising:

a first circuit for enabling individual bit lines during a program-and-verify operation; and

a second circuit for enabling all bit lines during a read operation.

2. The bit line decoder circuit of claim 1, wherein the second circuit comprises a select transistor and an activation function circuit coupled to each bit line.

3. The bit line decoder circuit of claim 2, wherein the gate of each select transistor is coupled to the same control line.

4. The bit line decoder circuit of claim 1, wherein a negative bias is applied to the word lines of each unselected memory cell during a program-and-verify operation or a read operation.

5. The bit line decoder circuit of claim 1, wherein each of the non-volatile memory cells is a split-gate flash memory cell.

6. The bit line decoder circuit of claim 1, wherein each of the non-volatile memory cells is a stacked-gate flash memory cell.

7. The bit line decoder circuit of claim 1, wherein each of the non-volatile memory cells are configured to operate in a sub-threshold region.

8. The bit line decoder circuit of claim 1, wherein each of the non-volatile memory cells are configured to operate in a linear region.

9. The bit line decoder circuit of claim 1, wherein a negative bias is applied to the gates of each unselected bit line decoder during a program-and-verify operation or a read operation.

10. A bit line decoder circuit coupled to a vector-by-matrix multiplication array, the vector-by matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns, wherein each column is connected to a bit line, the bit line decoder circuit comprising:

a multiplexed circuit, wherein in a first mode, the multiplexed circuit enables individual bit lines during a program-and-verify operation and in a second mode, the multiplexed circuit enabled all bit lines during a read operation.

11. The bit line decoder circuit of claim 10, wherein the multiplexed circuit comprises a select transistor and an activation function circuit coupled to each bit line.

12. The bit line decoder circuit of claim 10, wherein a negative bias is applied to the word lines of each unselected memory cell during a program-and-verify operation or a read operation.

13. The bit line decoder circuit of claim 10, wherein each of the non-volatile memory cells is a split-gate flash memory cell.

14. The bit line decoder circuit of claim 10, wherein each of the non-volatile memory cells is a stacked-gate flash memory cell.

15. The bit line decoder circuit of claim 10, wherein each of the non-volatile memory cells are configured to operate in a sub-threshold region.

16. The bit line decoder circuit of claim 10, wherein each of the non-volatile memory cells are configured to operate in a linear region.

17. The bit line decoder circuit of claim 10, wherein a negative bias is applied to the gates of each unselected bit line decoder during a program-and-verify operation or a read operation.

18. An analog neuromorphic memory system comprising:

a vector-by-matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns, wherein each column is connected to a bit line and each memory cell comprises a word line terminal and a source line terminal;

a word line decoder circuit coupled to the word line terminals of the non-volatile memory cells, wherein the word line decoder circuit is capable of applying a low voltage or a high voltage to coupled word line terminals; and

a source line decoder circuit coupled to the source line terminals of the non-volatile memory cells, wherein the source line decoder circuit is capable of applying a low voltage or a high voltage to coupled source line terminals.

19. The system of claim 18, wherein each memory cell further comprises an erase gate terminal, the system further comprising:

an erase gate decoder circuit coupled to the erase gate terminals of the non-volatile memory cells, wherein the erase gate decoder circuit is capable of applying a low voltage or a high voltage to coupled erase gate terminals.

20. A word line driver coupled to a vector-by-matrix multiplication array, the vector-by matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns, wherein each row is coupled to a word line and each word line is coupled to the word line driver, the word line driver comprising:

a plurality of select transistors, each of the plurality of select transistors comprising a first terminal, a second terminal, and a gate, wherein the gate of each of the plurality of select transistors is coupled to a common control line and the first terminal of each of the plurality of select transistors is coupled to a different word line, and wherein the second terminal of each of the plurality of select transistors is coupled to one or more bias transistors;

wherein the bias transistors coupled to each of the plurality of select transistors are capable of providing a bias voltage to a single select transistor or to all of the select transistors.

21. The word line driver of claim 20, where at least one bias transistor coupled to each of the plurality of select transistors is coupled to a common control line.

22. The word line driver of claim 20, where each bias transistor coupled to each of the plurality of select transistors is coupled to a different control line.

23. The word line driver of claim 20, wherein each of the bias transistors is coupled to a circuit for decoding a word line address

24. The word line driver of claim 20, wherein the bias transistors are coupled to a shift register.

25. The word line driver of claim 20, wherein each select transistor is coupled to a capacitor.

26. The word line driver of claim 20, wherein each bias transistor is coupled to a comparator.

27. An analog neuromorphic memory system comprising:

a vector-by-matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns, wherein each column is connected to a bit line and each memory cell comprises a word line terminal and a source line terminal;

a word line decoder circuit coupled to the word line terminals of the non-volatile memory cells, wherein the word line decoder circuit is capable of applying a low voltage through a low voltage transistor or a high voltage through a high voltage transistor to coupled word line terminals, the word line decoder circuit comprising an isolation transistor coupled to each word line to isolate the high voltage transistor from the low voltage transistor.

28. The system of claim 27, wherein a negative bias is applied to the word lines of each unselected memory cell during a program-and-verify operation or a read operation.

29. The system of claim 27, wherein each of the non-volatile memory cells is a split-gate flash memory cell.

30. The system of claim 27, wherein each of the non-volatile memory cells is a stacked-gate flash memory cell.

31. The system of claim 27, wherein each of the non-volatile memory cells are configured to operate in a sub-threshold region.

32. The system of claim 27, wherein each of the non-volatile memory cells are configured to operate in a linear region.

33. An analog neuromorphic memory system comprising:

a vector-by-matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns, wherein each column is connected to a bit line and each memory cell comprises a word line terminal and a source line terminal;

a word line decoder circuit coupled to the word line terminals of the non-volatile memory cells, wherein the word line decoder circuit is capable of applying a low voltage or a high voltage to coupled word line terminals; and

a sample and hold capacitor coupled to each word line

34. The system of claim 33, wherein a negative bias is applied to the word lines of each unselected memory cell during a program-and-verify operation or a read operation.

35. The system of claim 33, wherein the capacitor in the sample and hold capacitor is provided by an inherent capacitance of a word line.

36. The system of claim 33, wherein each of the non-volatile memory cells is a split-gate flash memory cell.

37. The system of claim 33, wherein each of the non-volatile memory cells is a stacked-gate flash memory cell.

38. The system of claim 33, wherein each of the non-volatile memory cells are configured to operate in a sub-threshold region.

39. The system of claim 33, wherein each of the non-volatile memory cells are configured to operate in a linear region.

40. The system of claim 33, wherein the S/H capacitor serves a voltage source for the wordline

41. An analog neuromorphic memory system comprising:

a vector-by-matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns, wherein each column is connected to a bit line and each memory cell comprises a word line terminal and a source line terminal; and

a source line decoder circuit coupled to the word line terminals of the non-volatile memory cells, wherein the source line decoder circuit is capable of applying a low voltage or a high voltage to coupled source line terminals, wherein said source line decoder circuit comprises a driving transistor and a monitor transistor.

42. The system of claim 41, wherein the system further comprises a force-sense circuit operating in a closed loop.

43. The system of claim 41, wherein each of the non-volatile memory cells is a split-gate flash memory cell.

44. The system of claim 41, wherein each of the non-volatile memory cells is a stacked-gate flash memory cell.

45. The system of claim 41, wherein each of the non-volatile memory cells are configured to operate in a sub-threshold region.

46. The system of claim 41, wherein each of the non-volatile memory cells are configured to operate in a linear region.

47. An analog neuromorphic memory system comprising:

a vector-by-matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns, wherein each column is connected to a bit line and each memory cell comprises a word line terminal and a source line terminal;

a control gate decoder circuit coupled to the control gate line terminals of the non-volatile memory cells, wherein the control gate decoder circuit is capable of applying a low voltage or a high voltage to coupled control gate terminals; and

a sample and hold capacitor coupled to each word line

48 The system of claim 47, wherein a negative bias is applied to the word lines of each unselected memory cell during a program-and-verify operation or a read operation.

49. The system of claim 47, wherein the sample and hold capacitor comprises control gate capacitance.

50. The system of claim 47, wherein each of the non-volatile memory cells is a split-gate flash memory cell.

51. The system of claim 47, wherein each of the non-volatile memory cells is a stacked-gate flash memory cell.

52. The system of claim 47, wherein each of the non-volatile memory cells are configured to operate in a sub-threshold region.

53. The system of claim 47, wherein each of the non-volatile memory cells are configured to operate in a linear region.

54. The system of claim 47, wherein the S/H capacitor serves a voltage source for the control gate lines

55. A current-to-voltage circuit comprising:

a reference circuit for receiving an input current and outputting a first voltage in response to the input current, the reference circuit comprising an input current source, an NMOS transistor, a cascoding bias transistor, and a reference memory cell;

a sample and hold circuit for receiving the first voltage and outputting a second voltage, the second voltage constituting a sampled value of the first voltage, the sample and hold circuit comprising a switch and capacitor.

56. The current to-voltage circuit of claim 55, further comprising an amplifier for receiving the second voltage and outputting a third voltage.

57. The current-to-voltage circuit of claim 55, wherein the second voltage is provided to a word line in a flash memory system.

58. The current-to-voltage circuit of claim 56, wherein the second voltage is provided to a word line in a flash memory system.

59. The current-to-voltage circuit of claim 55, wherein the second voltage is provided to a control gate line in a flash memory system.

60. The current-to-voltage circuit of claim 56, wherein the second voltage is provided to a control gate line in a flash memory system.

61. A current-to-voltage circuit comprising:

a reference circuit for receiving an input current and outputting a first voltage in response to the input current, the reference circuit comprising an input current source, an NMOS transistor, a cascoding bias transistor, and a reference memory cell;

an amplifier for receiving the first voltage and outputting a second voltage;

a sample and hold circuit for receiving the second voltage and outputting a third voltage, the third voltage constituting a sampled value of the second voltage, the sample and hold circuit

62. The current-to-voltage circuit of claim 61, wherein the third voltage is provided to a word line in a flash memory system.

63. The current-to-voltage circuit of claim 61, wherein the third voltage is provided to a control gate line in a flash memory system.

64. An analog neuromorphic memory system comprising:

a vector-by-matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns; and

a redundancy sector.

65. The system of claim 64, further comprising a non-volatile register for storing system information.

66. An analog neuromorphic memory system comprising:

a vector-by-matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns; and

a non-volatile register for storing system information.

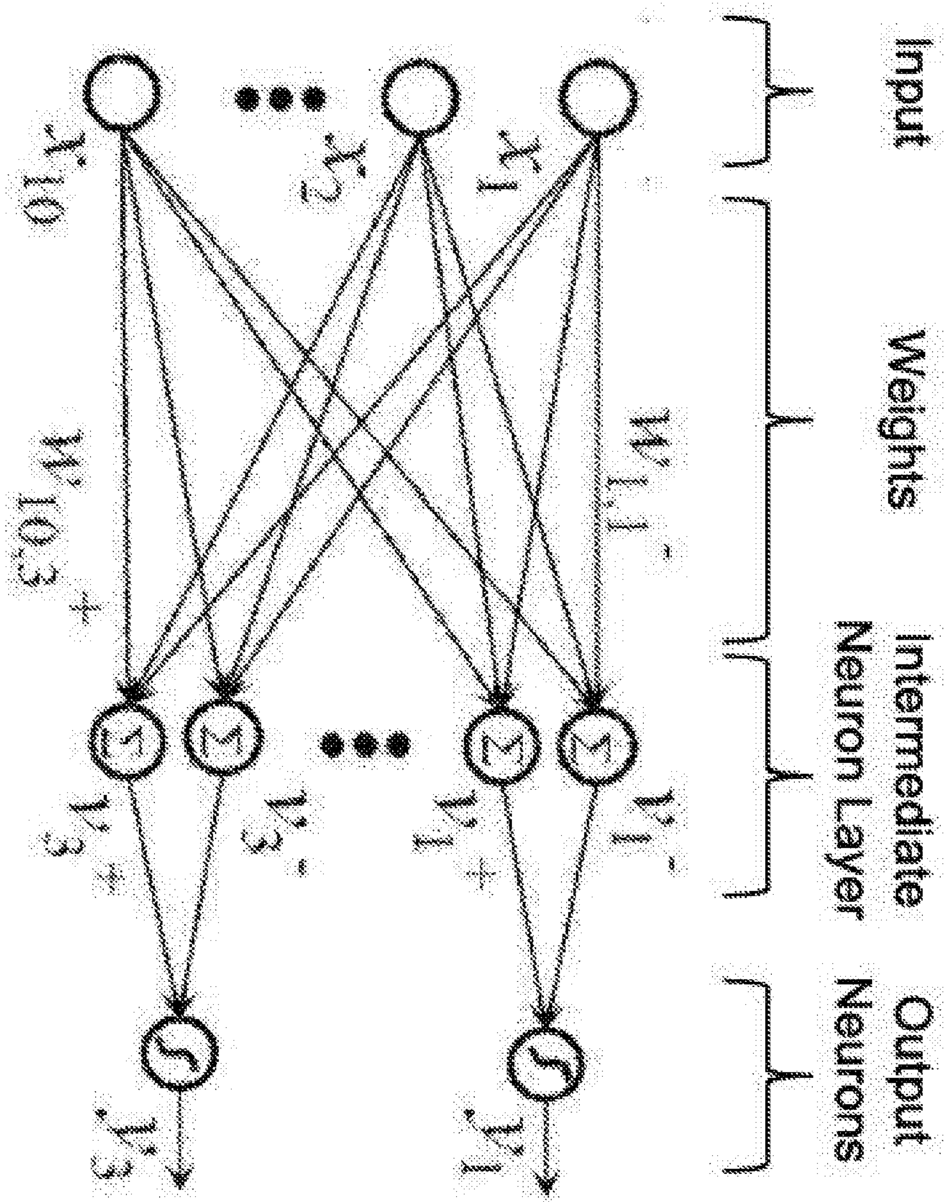
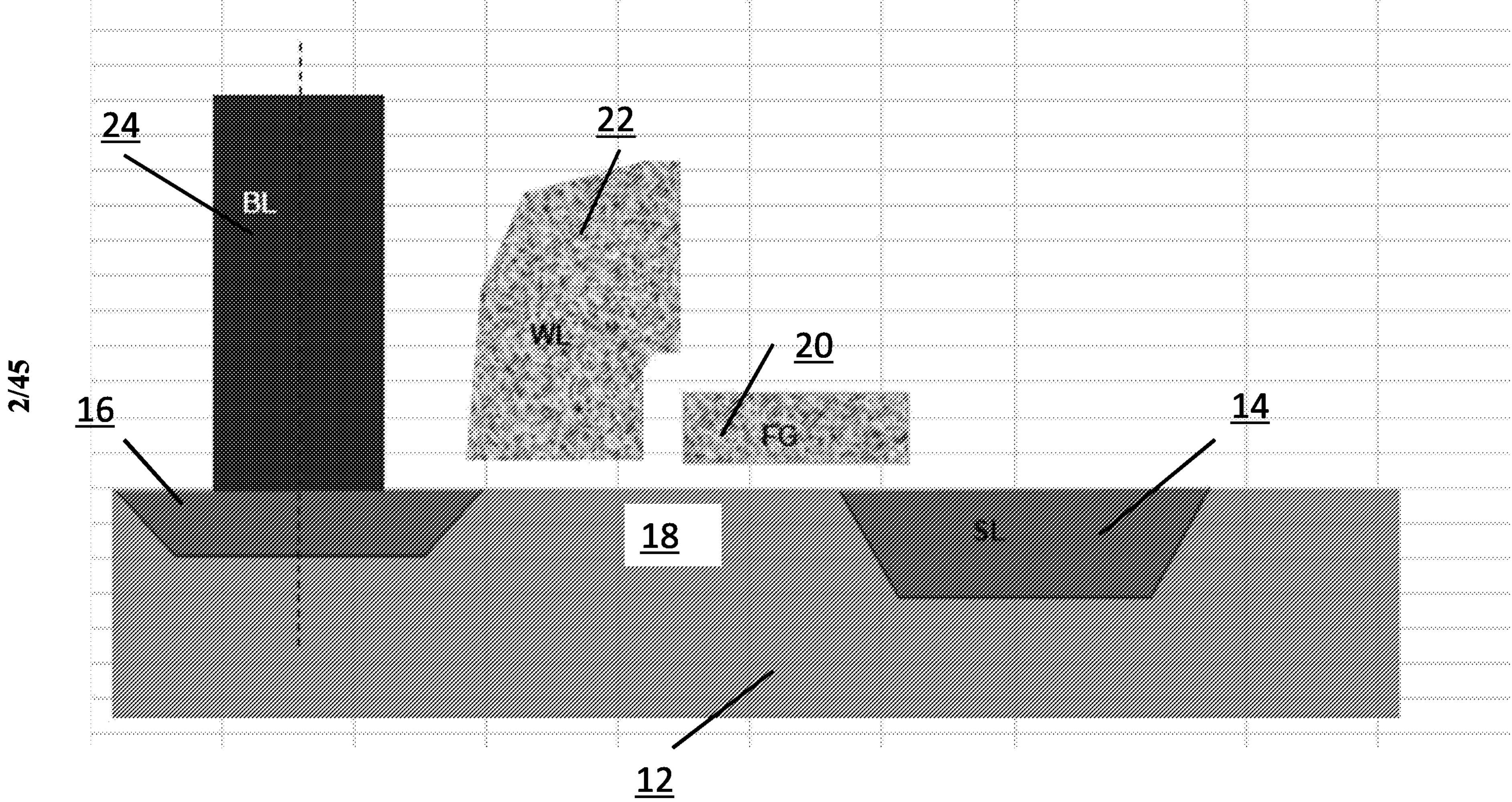


FIG. 1
(Prior Art)

FIGURE 2 (PRIOR ART)



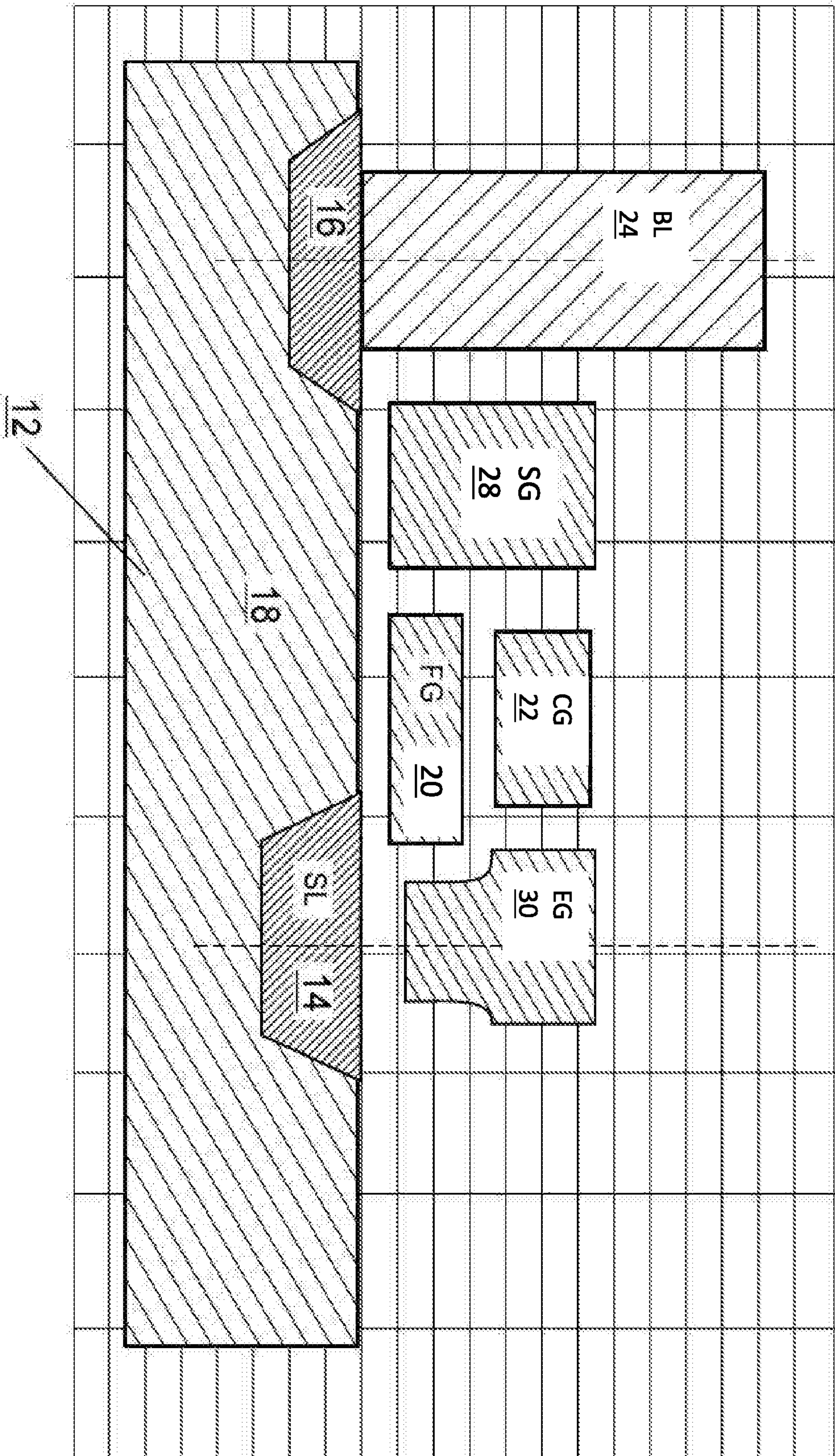


FIGURE 3 (PRIOR ART)

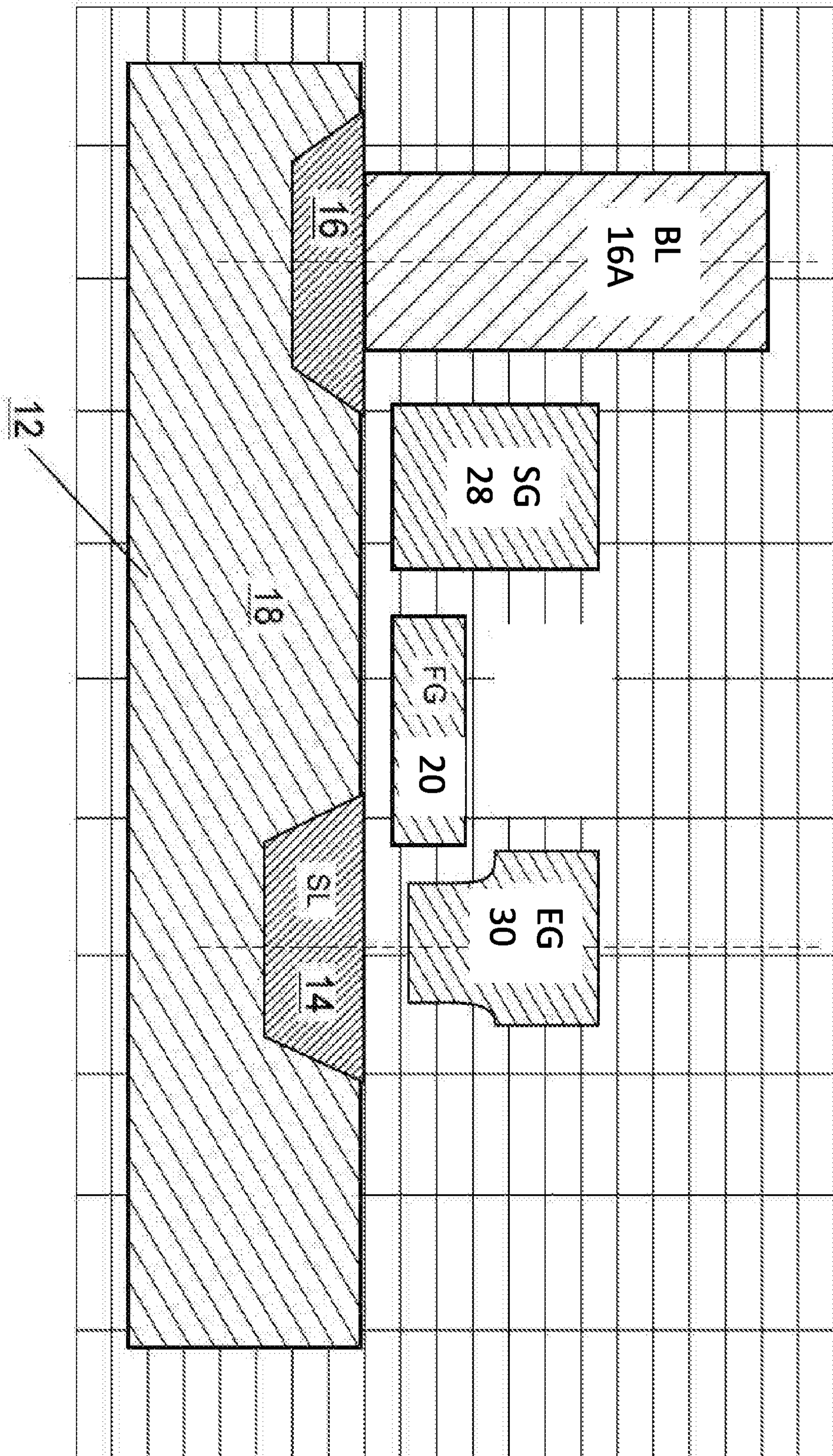


FIGURE 4 (PRIOR ART)

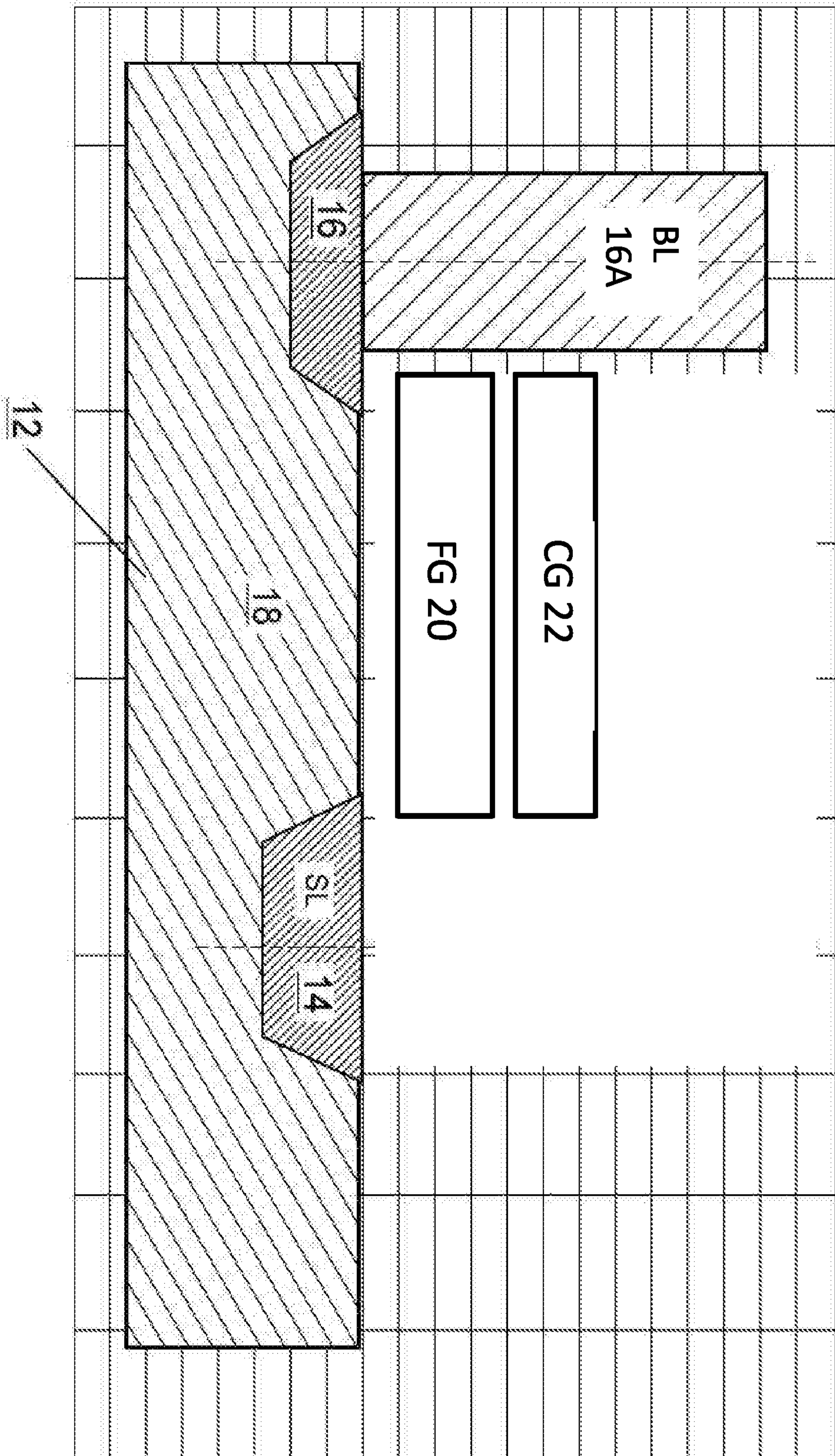


FIGURE 5 (PRIOR ART)

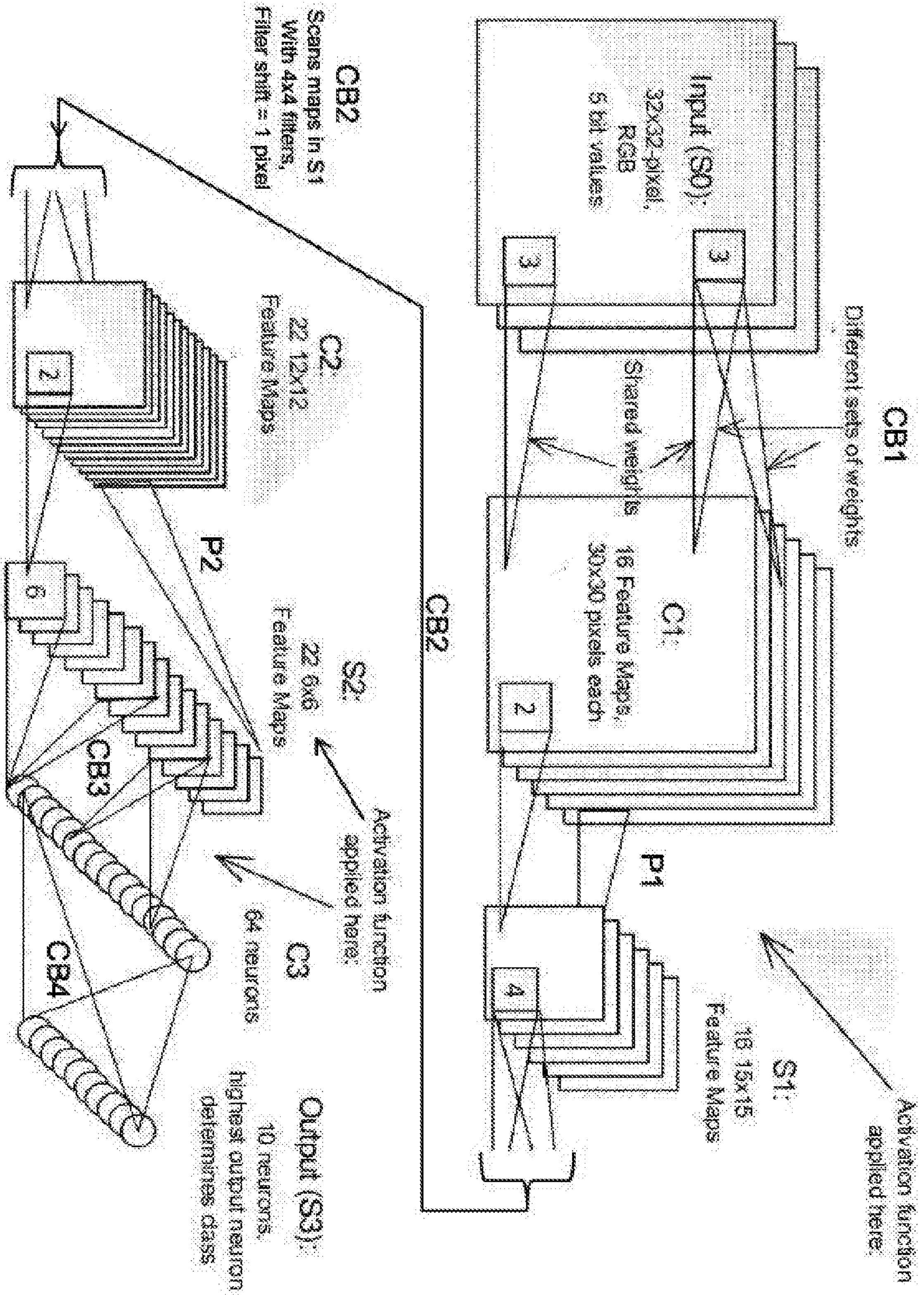


FIGURE 6

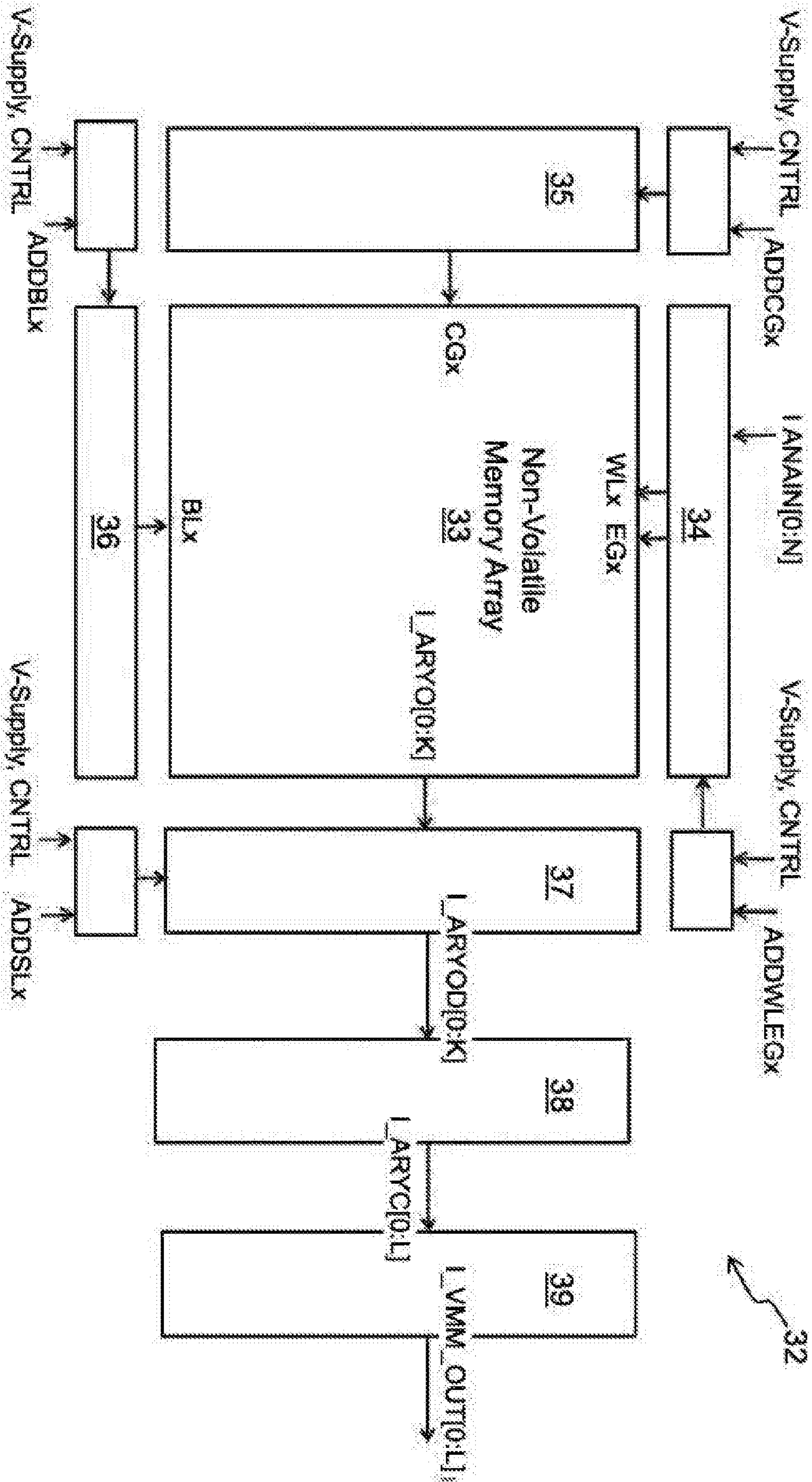


FIGURE 7

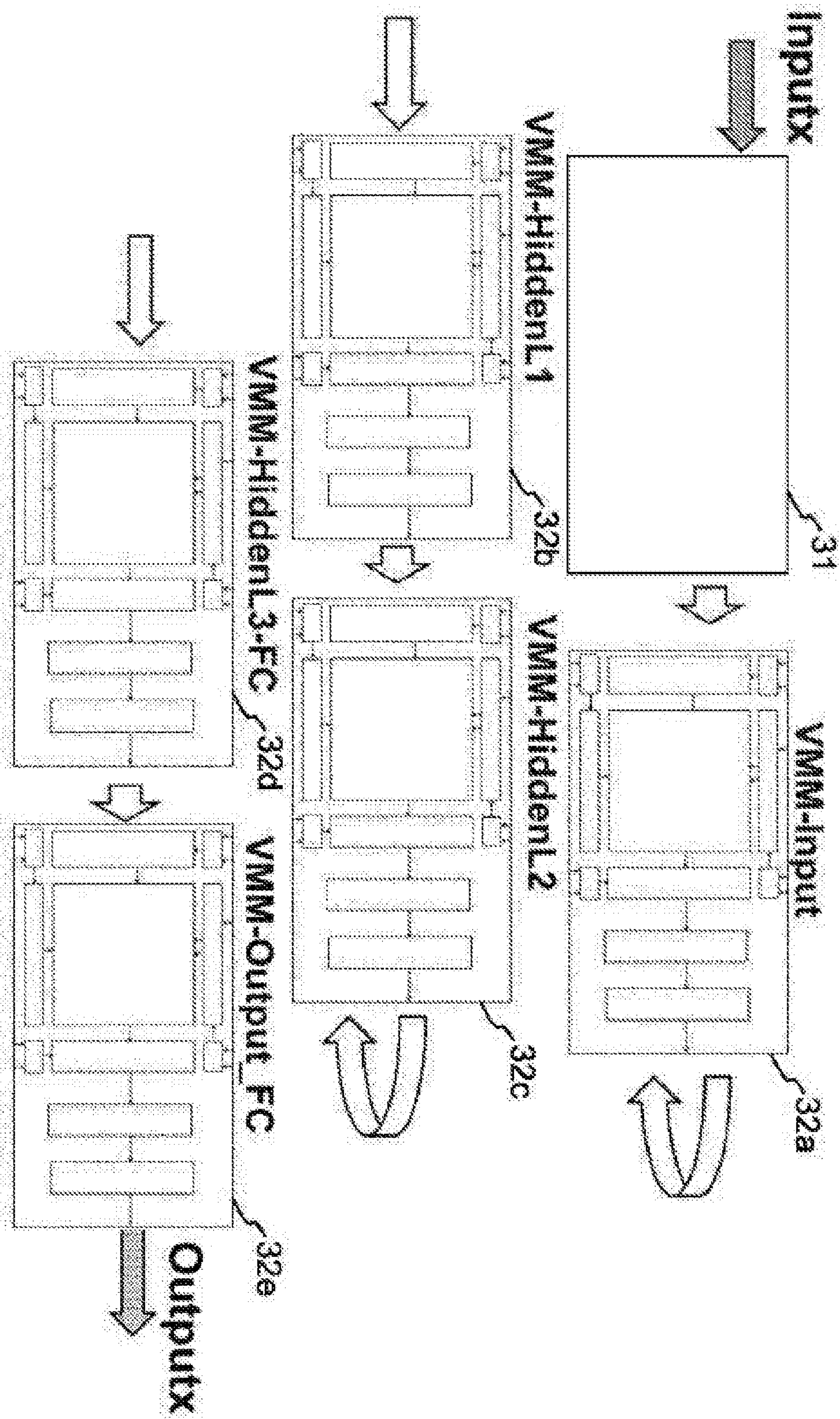


FIGURE 8
Inputx
31
VMM-Input
32a
VMM-HiddenL1
32b
VMM-HiddenL2
32c
VMM-HiddenL3-FC
32d
VMM-Output_FC
32e
Outputx

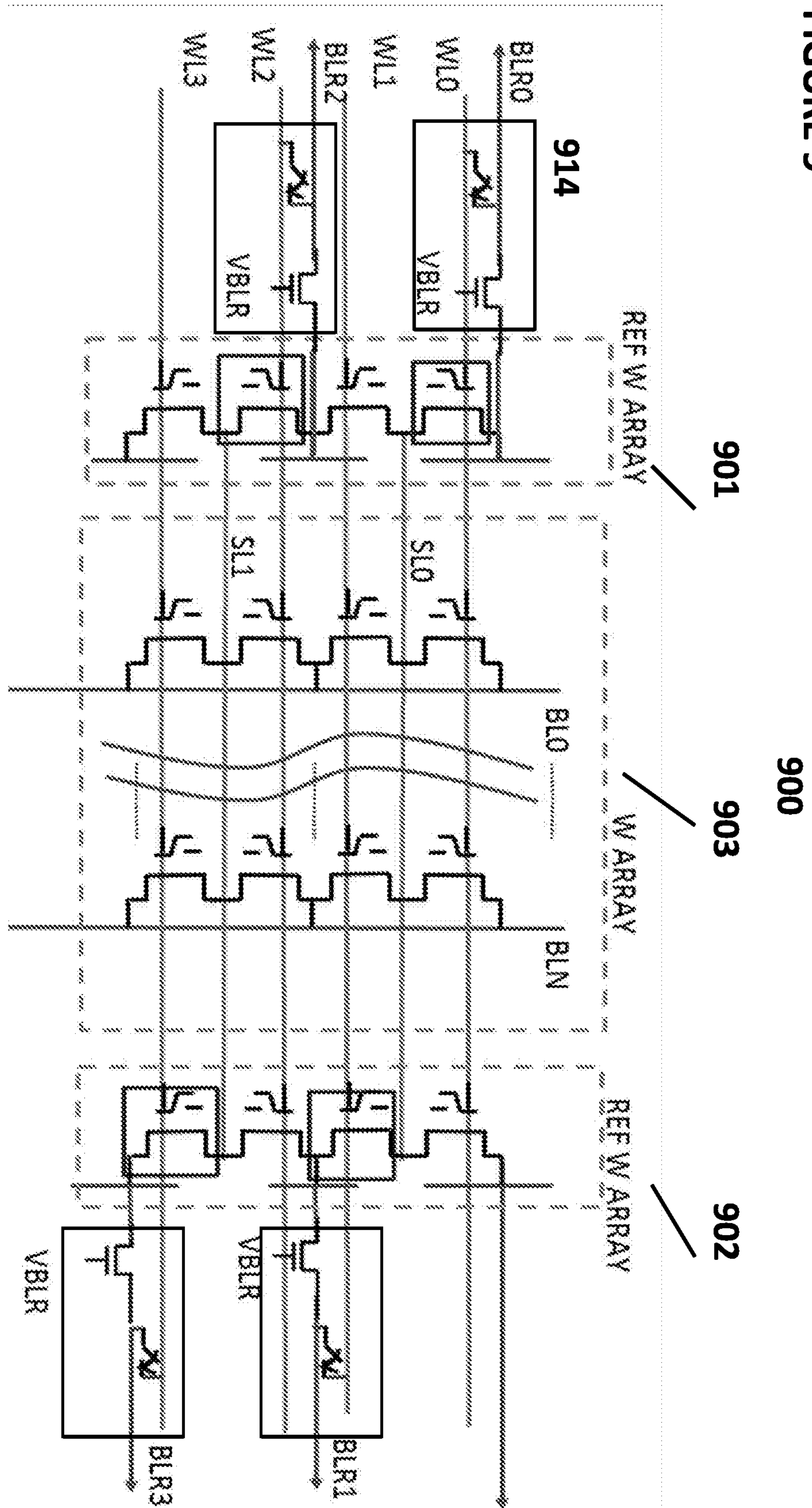


FIGURE 9
900
914
901
902
903
BLR0
WL0
WL1
BLR2
WL2
WL3
VBLR
VBLR
REF W ARRAY
SL0
SL1
BL0
BLN
W ARRAY
REF W ARRAY
VBLR
VBLR
BLR1
BLR3

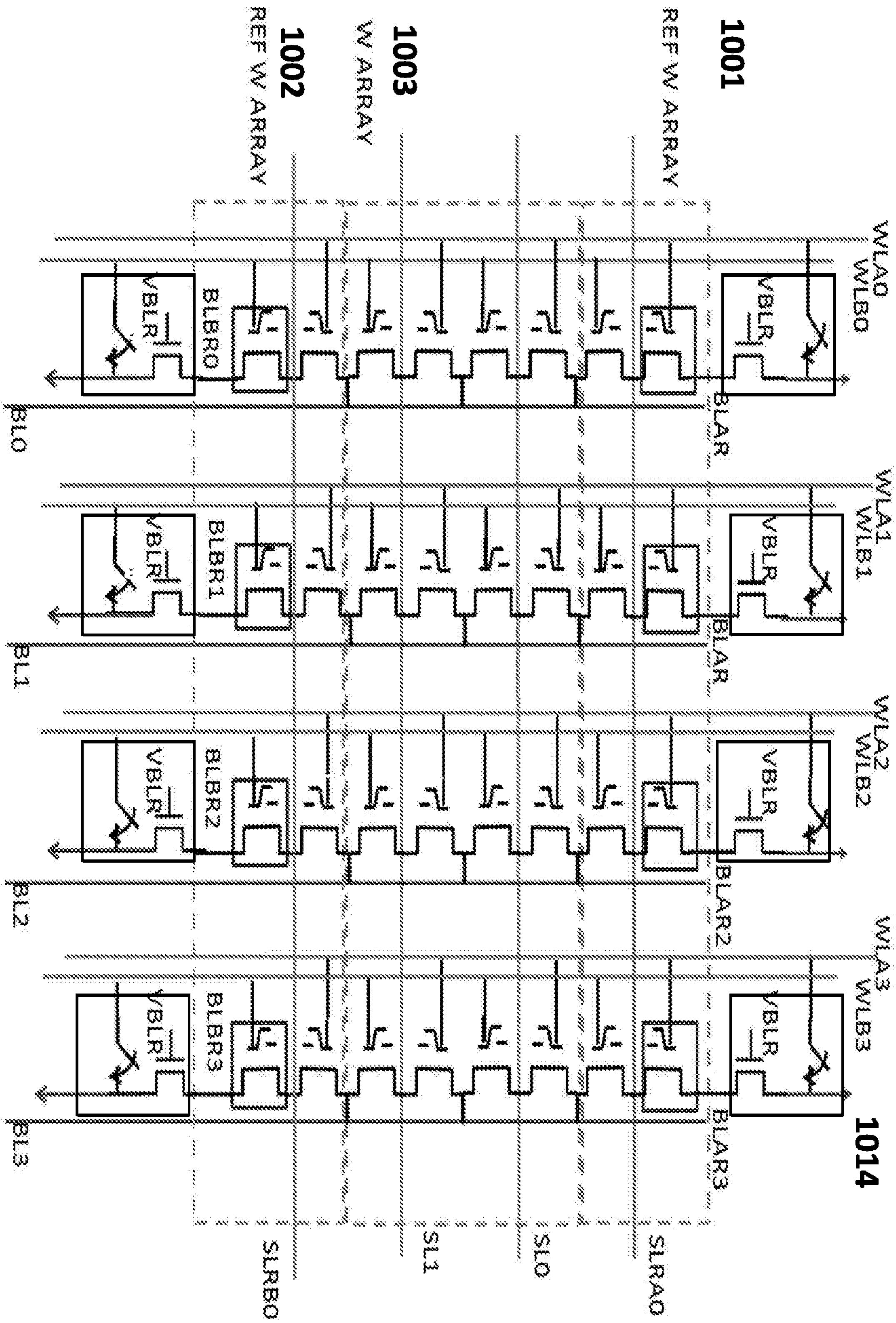


FIGURE 10
1000
1001
REF W ARRAY
1003
W ARRAY
1002
REF W ARRAY
1014
WLA0
WLB0
WLA1
WLB1
WLA2
WLB2
WLA3
WLB3
VBLR
BLBR0
BLBR1
BLBR2
BLBR3
BLAR
BLAR2
BLAR3
BL0
BL1
BL2
BL3
SLRB0
SL1
SL0
SLRA0

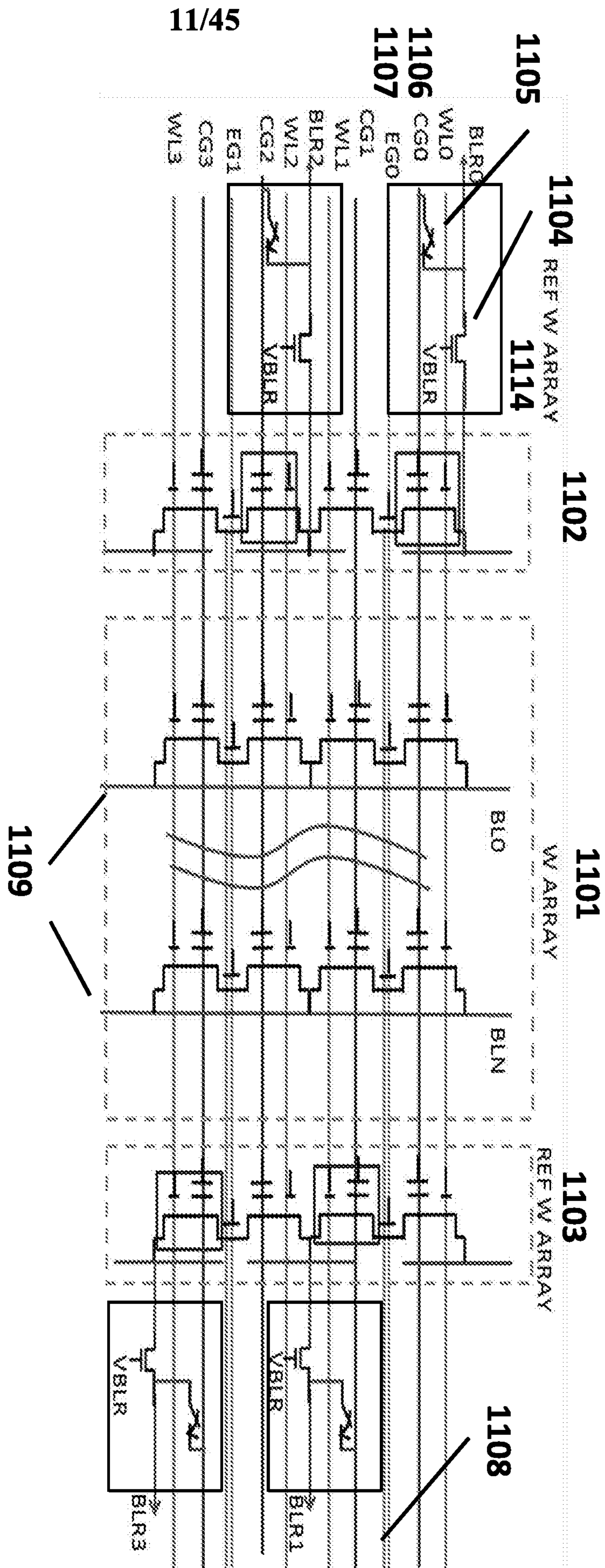


FIGURE 11
1100
1101
W ARRAY
1102
1103
REF W ARRAY
1104
REF W ARRAY
1105
1106
1107
1108
1109
1114
BLR0
WL0
CG0
EG0
CG1
WL1
BLR2
WL2
CG2
EG1
CG3
WL3
VBLR
BL0
BLN
BLR1
BLR3

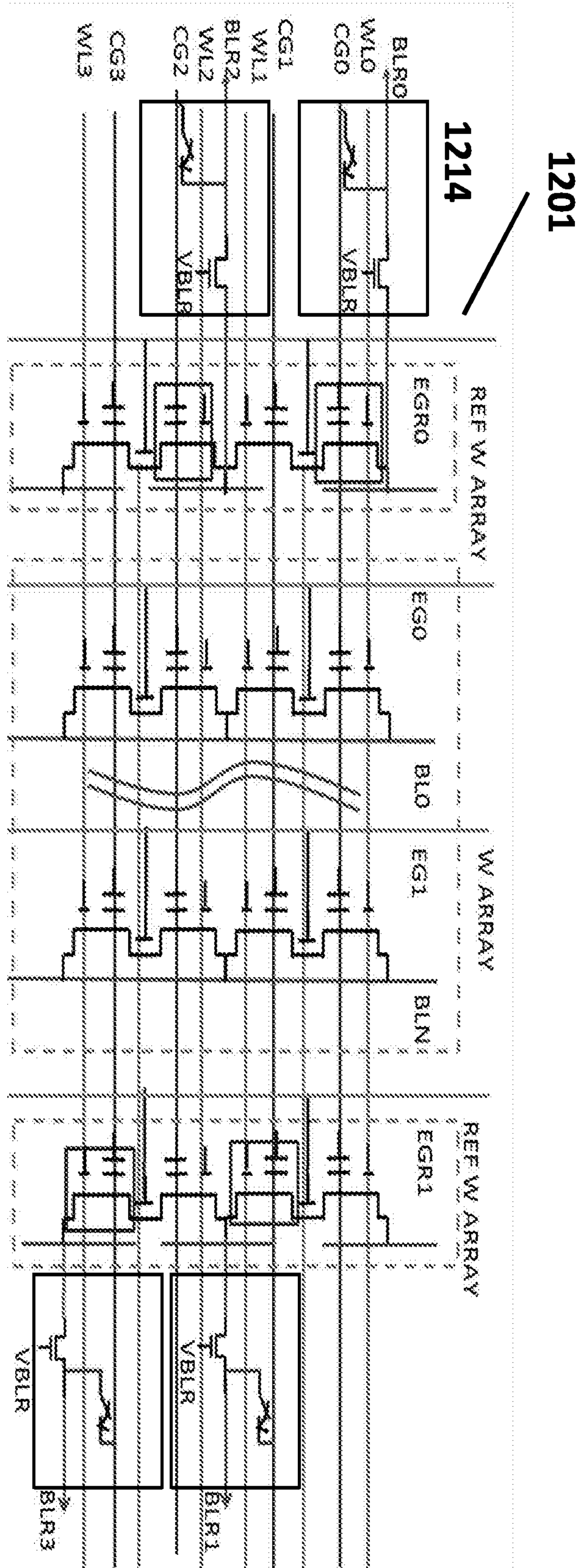


FIGURE 12
1200
1201
1214
BLR0
WL0
CG0
CG1
WL1
BLR2
WL2
CG2
CG3
WL3
VBLR
VBLR
VBLR
VBLR
REF W ARRAY
EGR0
EG0
BL0
EG1
W ARRAY
BLN
REF W ARRAY
EGR1
BLR1
BLR3

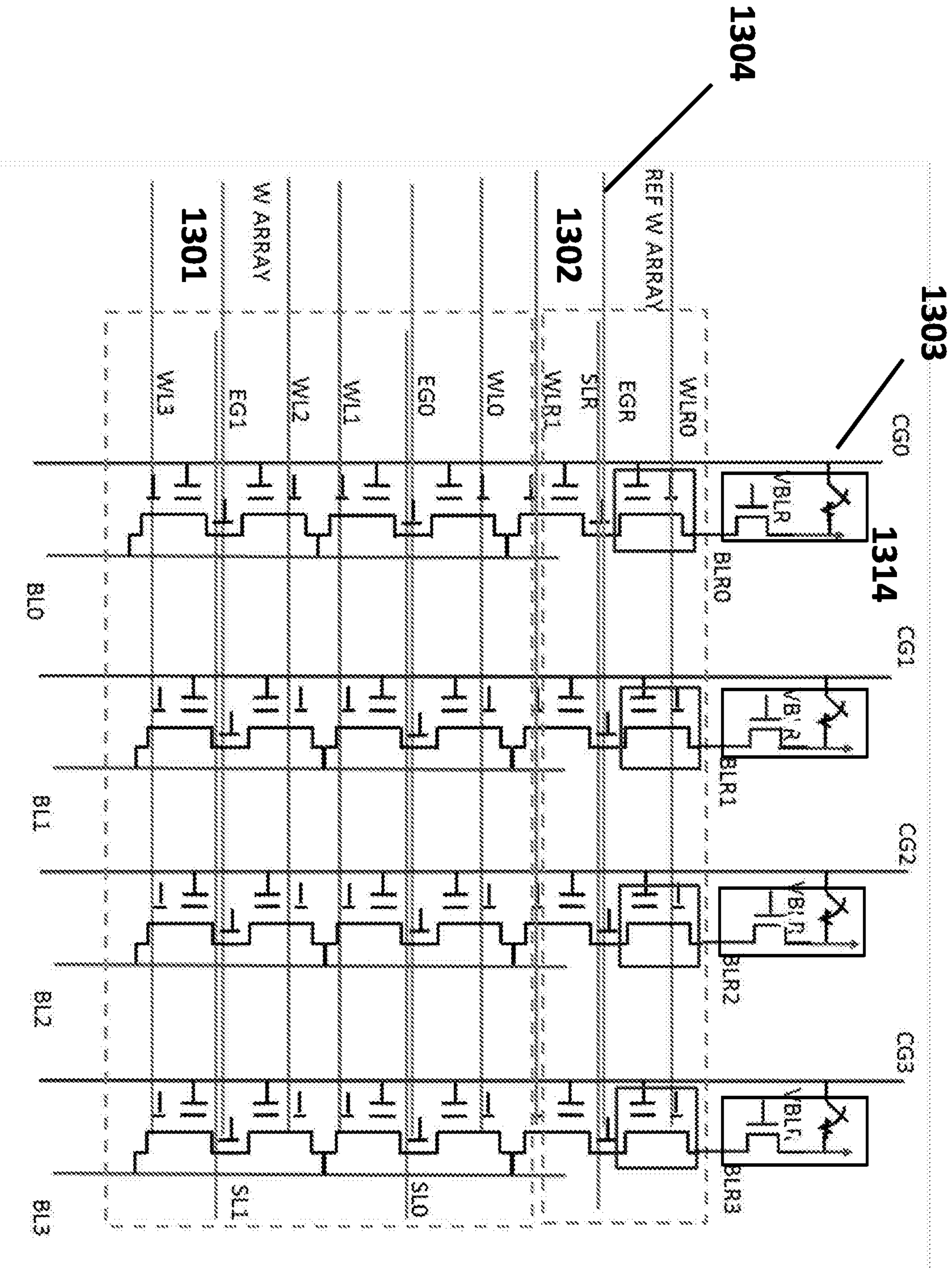


FIGURE 13

FIGURE 14

1400

1402

VMM Array
1401

1403

1404

1405

1406

V0 VB_0 BL0
V1 VB_1 BL1
V2 VB_2 BL2
V3 VB_3 BL3

f0_0 f0_1 f0_2 f0_3

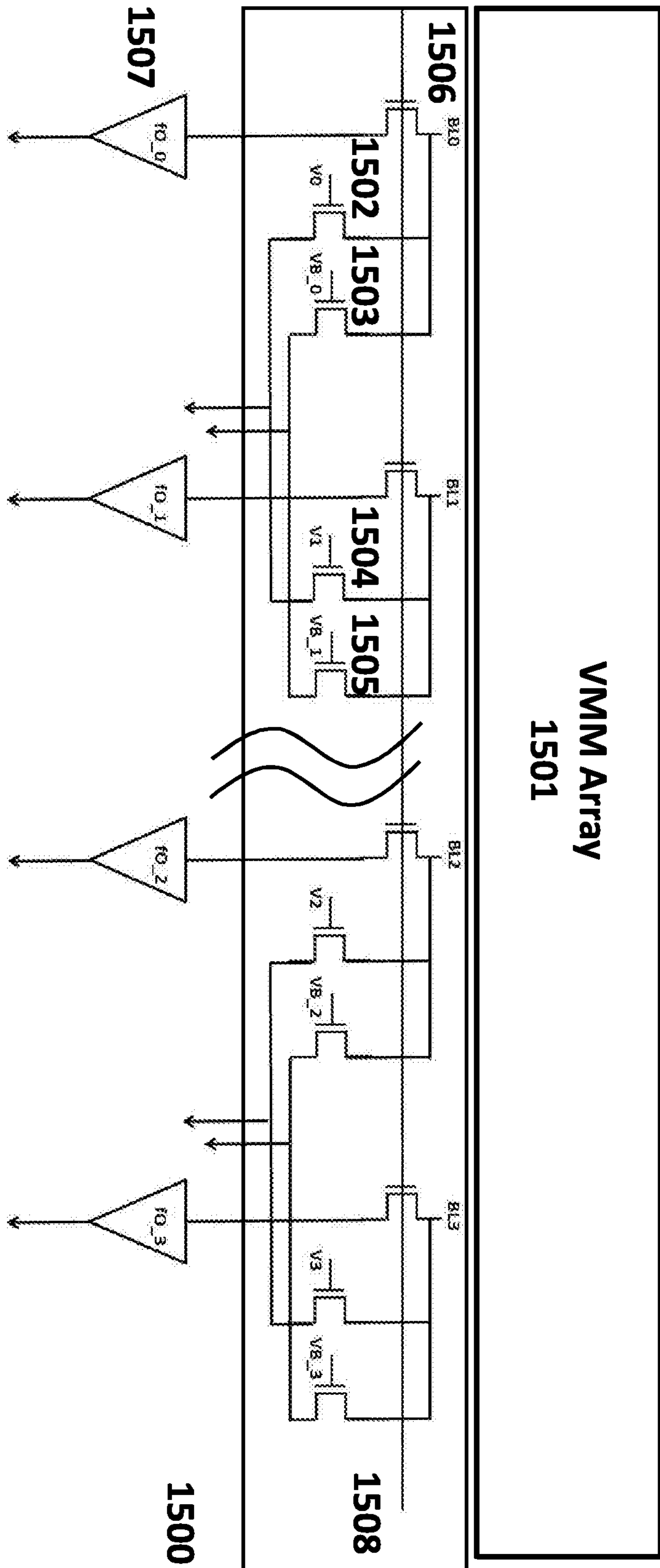


FIGURE 15

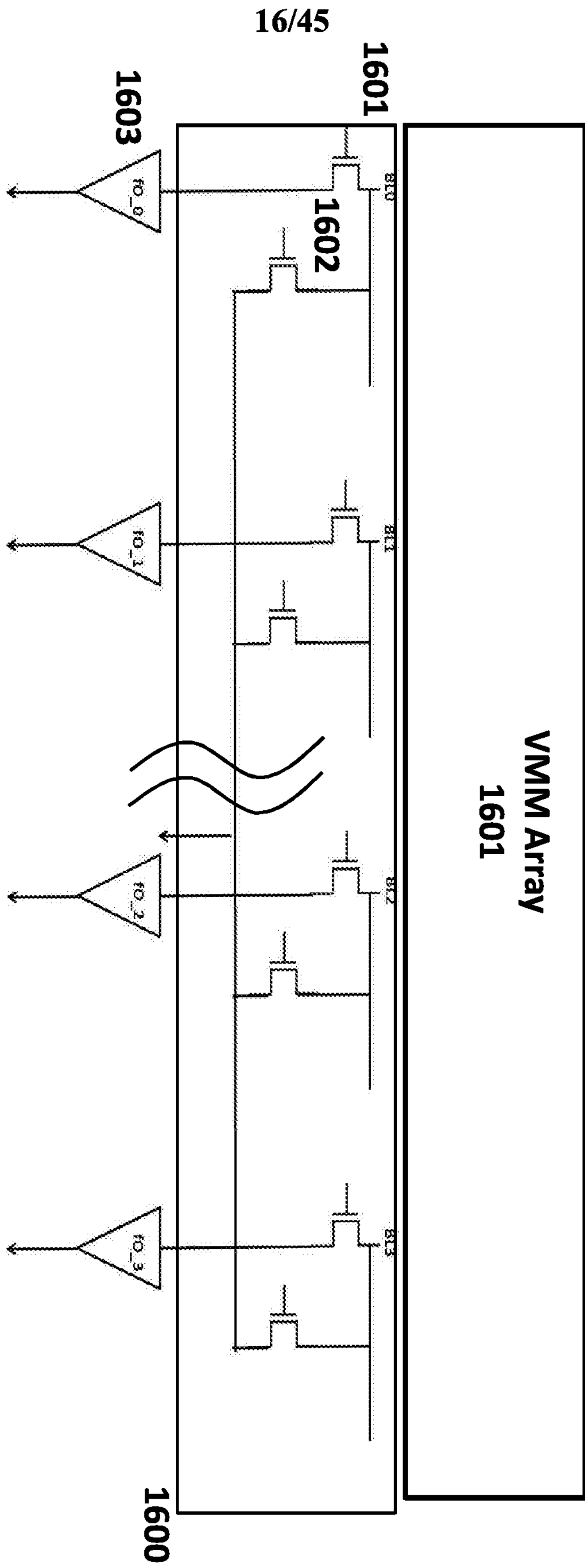


FIGURE 16

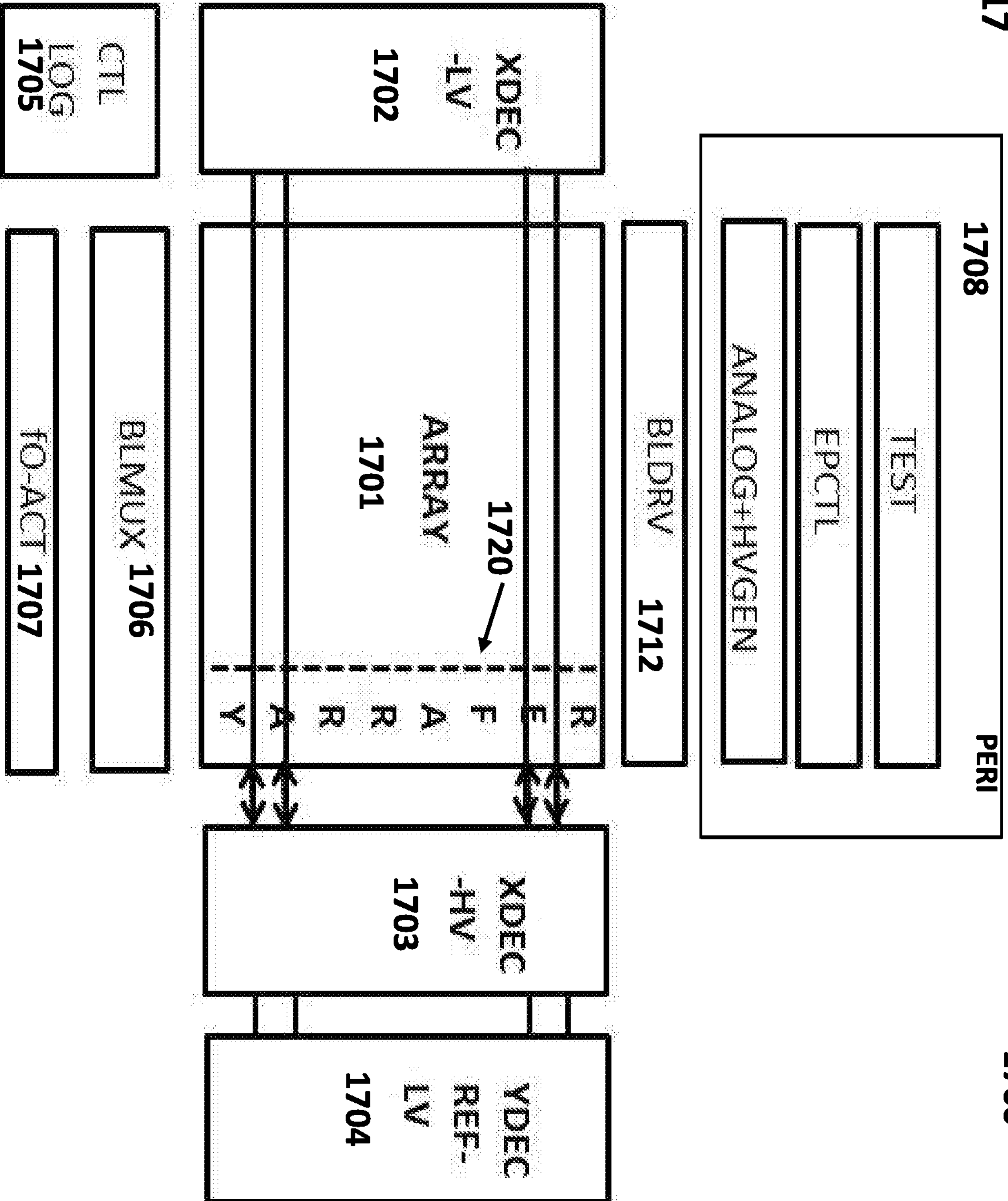
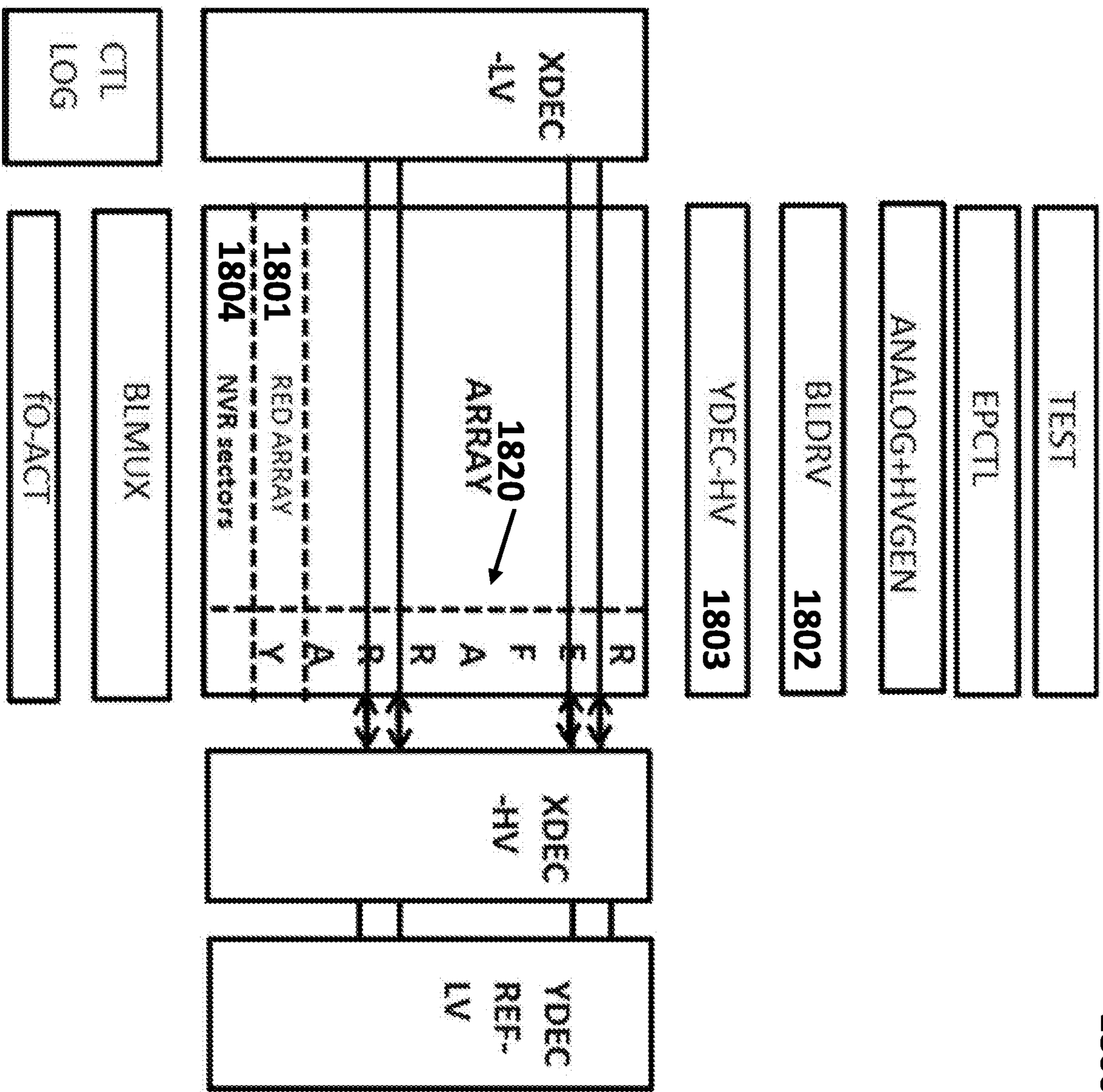


FIGURE 17
1700
XDEC
-LV
1702
CTL
LOG
1705
ARRAY
1701
1720
R
E
F
A
R
R
A
Y
BLMUX 1706
fO-ACT 1707
BLDRV
1712
1708
TEST
EPCTL
ANALOG+HVGEN
PERI
XDEC
-HV
1703
YDEC
REF-
LV
1704

FIGURE 18



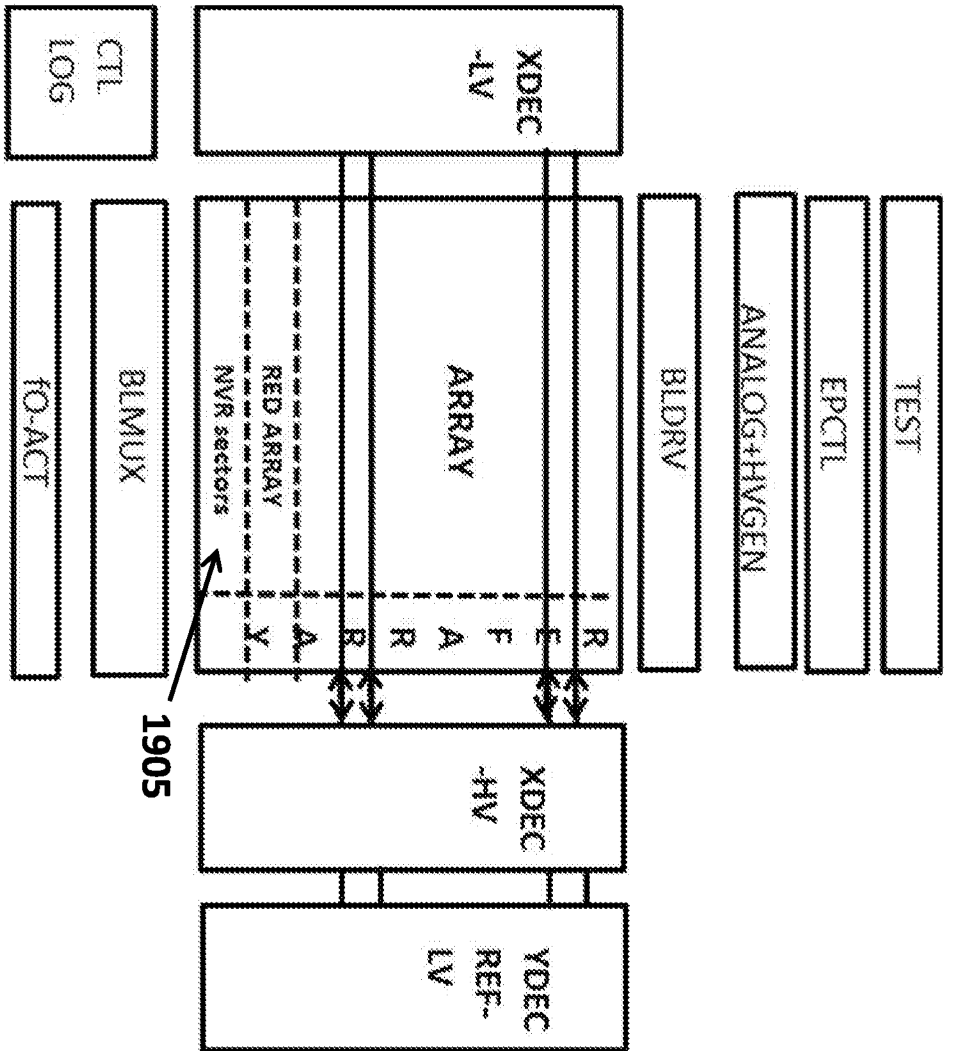


FIGURE 19
1900
XDEC -LV
CTL LOG
ARRAY
RED ARRAY
NVR sectors
REF ARRAY
1905
fO-ACT
BLMUX
BLDRV
ANALOG+HVGEN
EPCTL
TEST
XDEC -HV
YDEC REF-LV

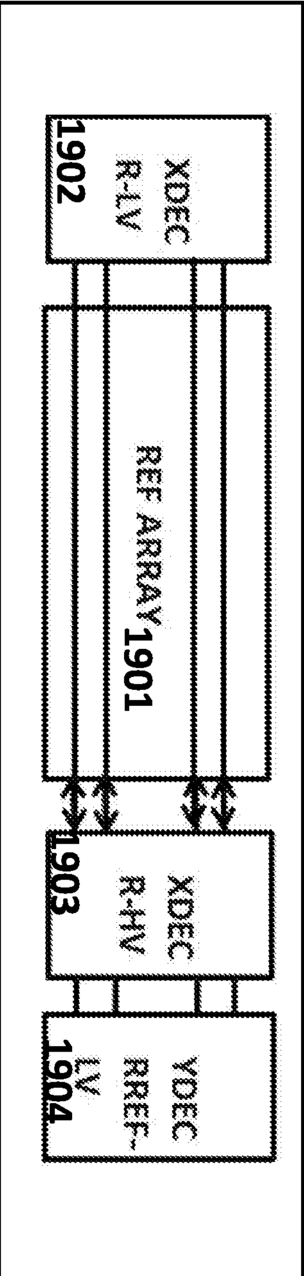
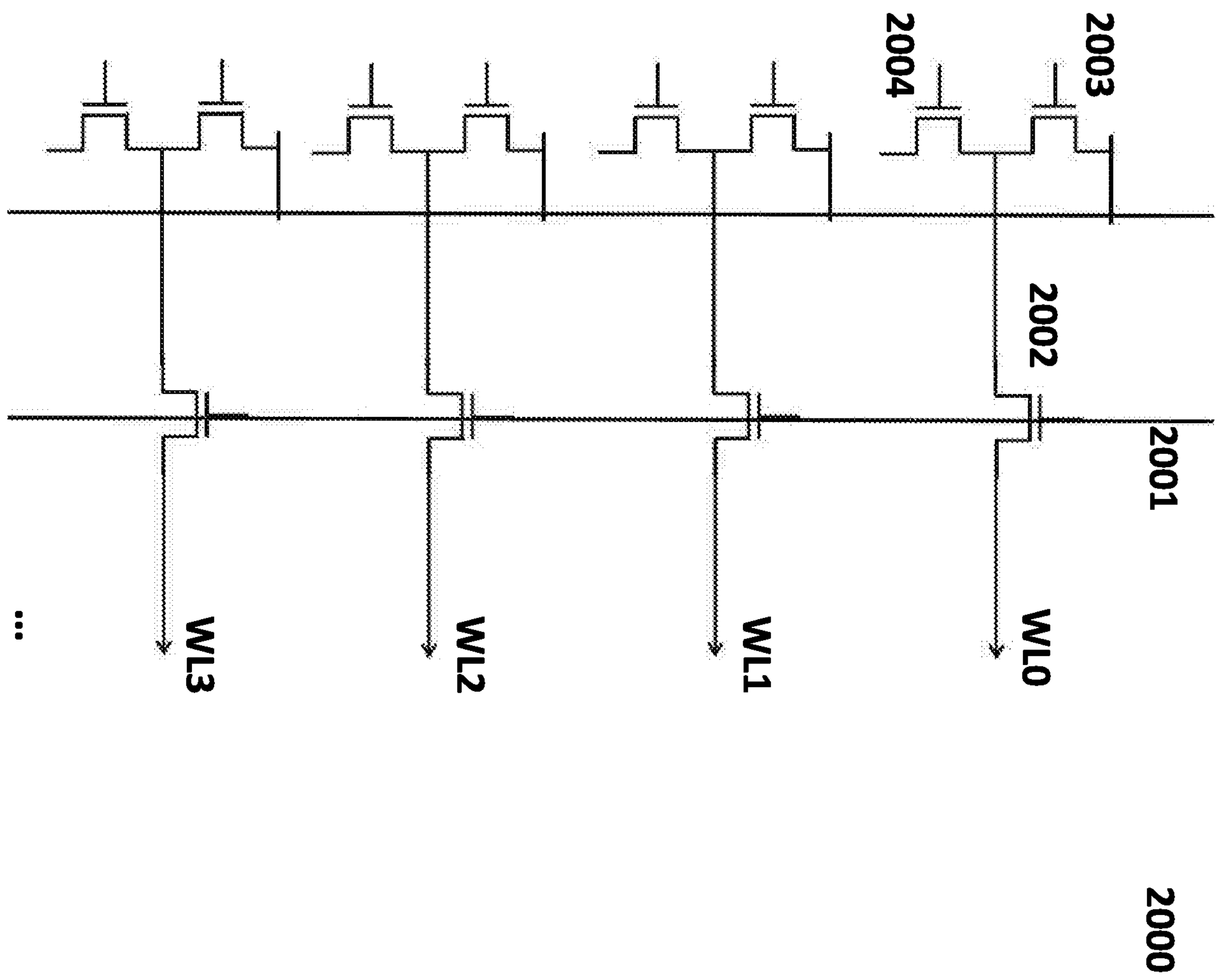

1999
XDEC R-LV
1902
REF ARRAY
1901
XDEC R-HV
1903
YDEC RREF-LV
1904

FIGURE 20



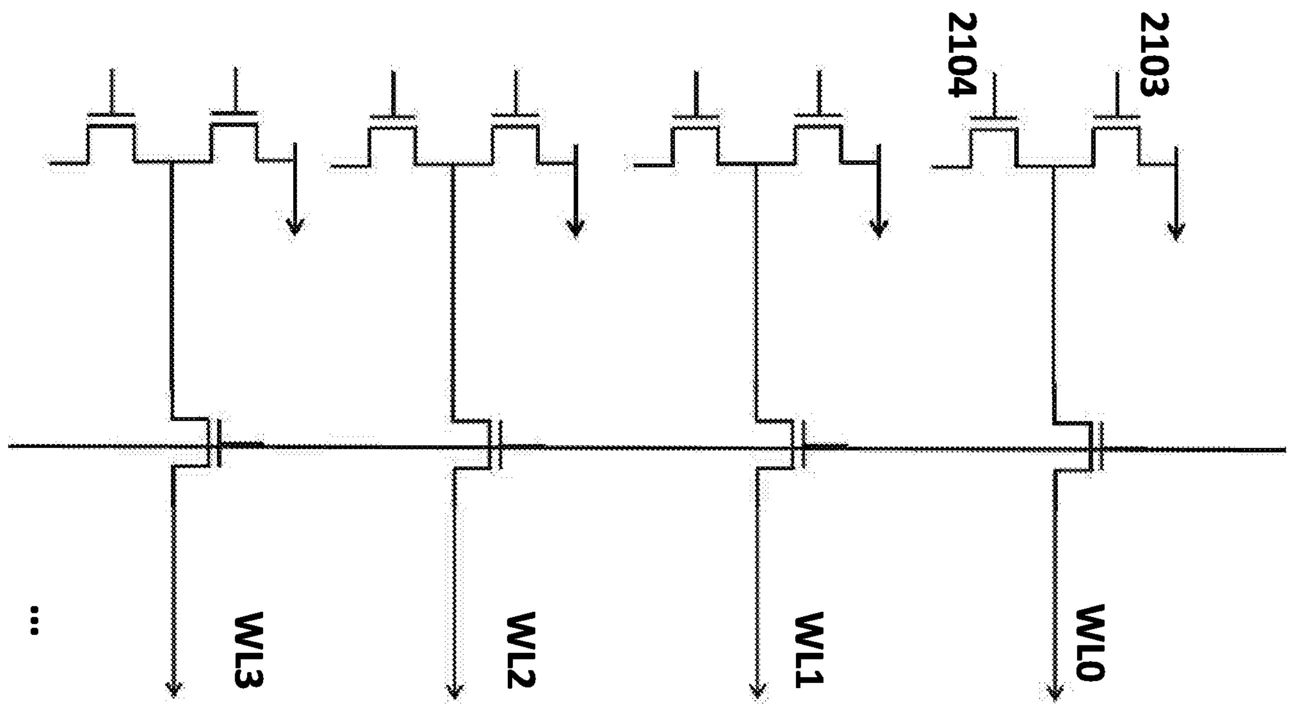


FIGURE 21

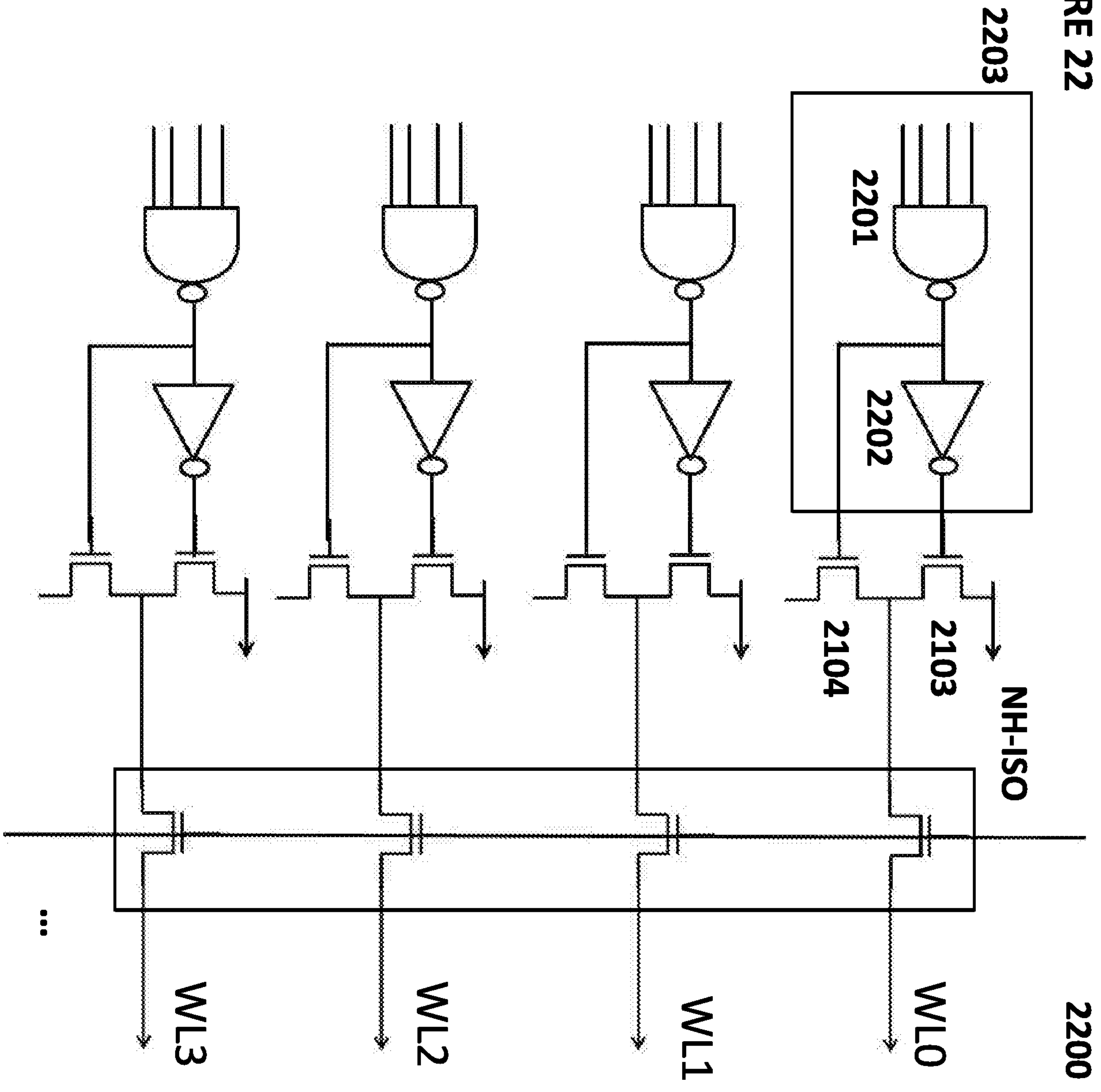


FIGURE 22

FIGURE 23

2300

2301

DIN Q
CLK Qb

2302

2103

2104

WL0

DIN Q
CLK Qb

WL1

DIN Q
CLK Qb

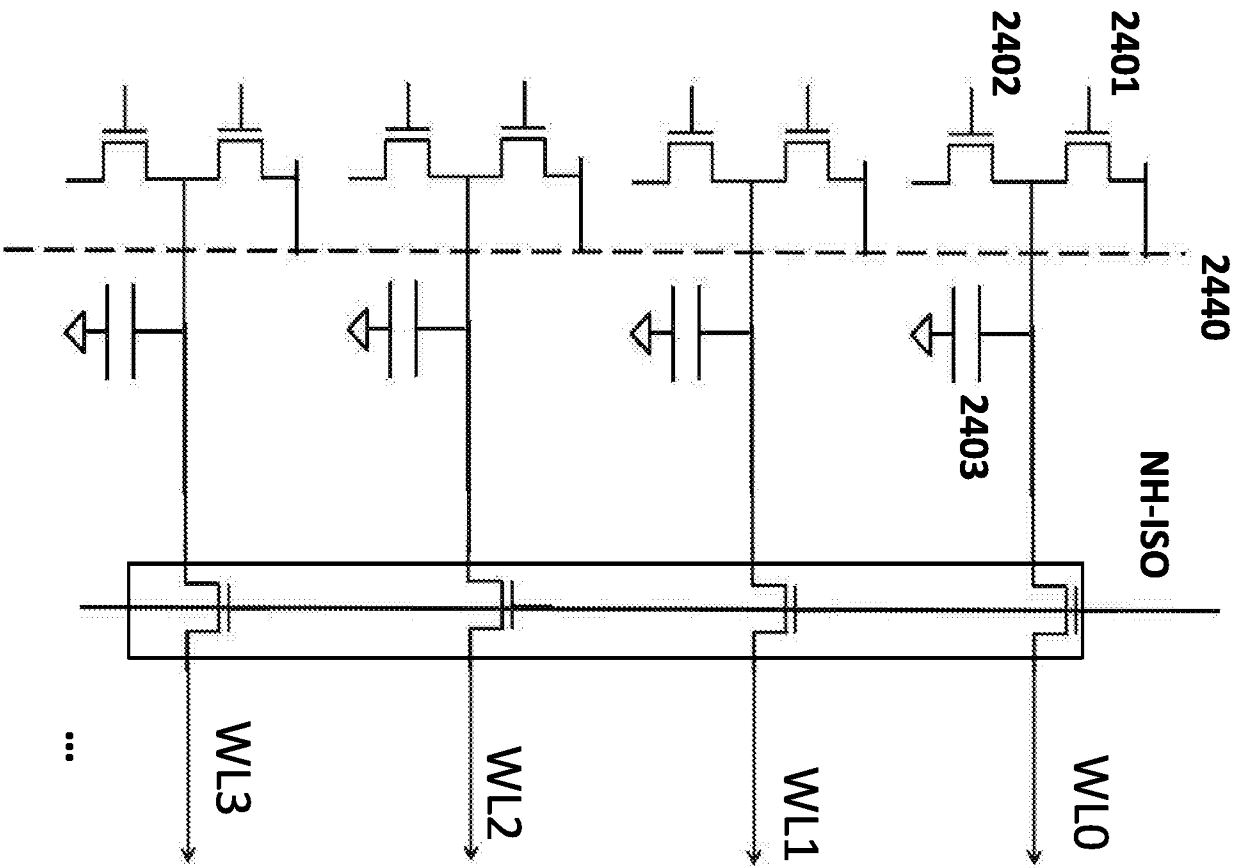
WL2

DIN Q
CLK Qb

WL3

...

FIGURE 24



2400

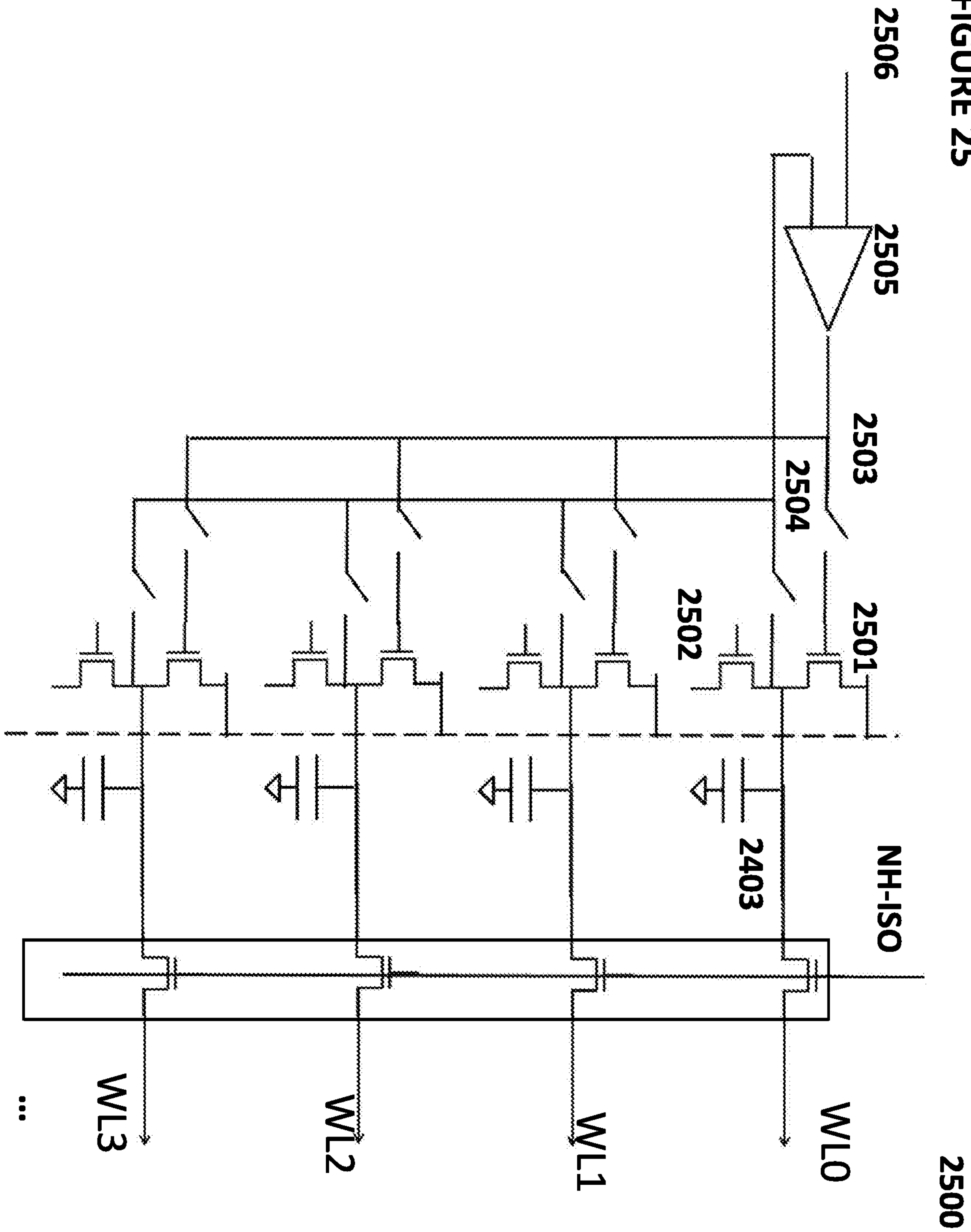


FIGURE 25
2506
2505
2503
2504
2502
2501
NH-ISO
2403
WL0
WL1
WL2
WL3
...
2500

FIGURE 26

2602
2603
2604
2601
NH-ISO
WL0
WL1
WL2
WL3
...
2600

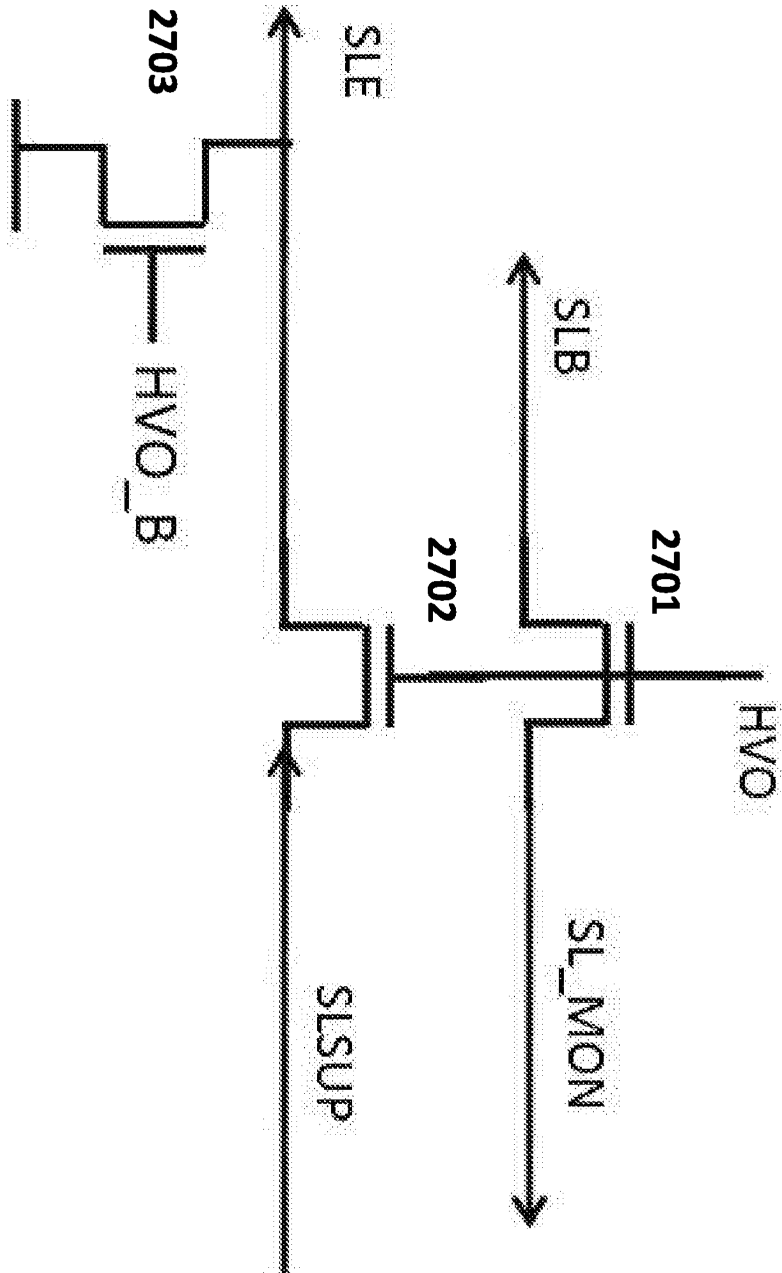


FIGURE 27

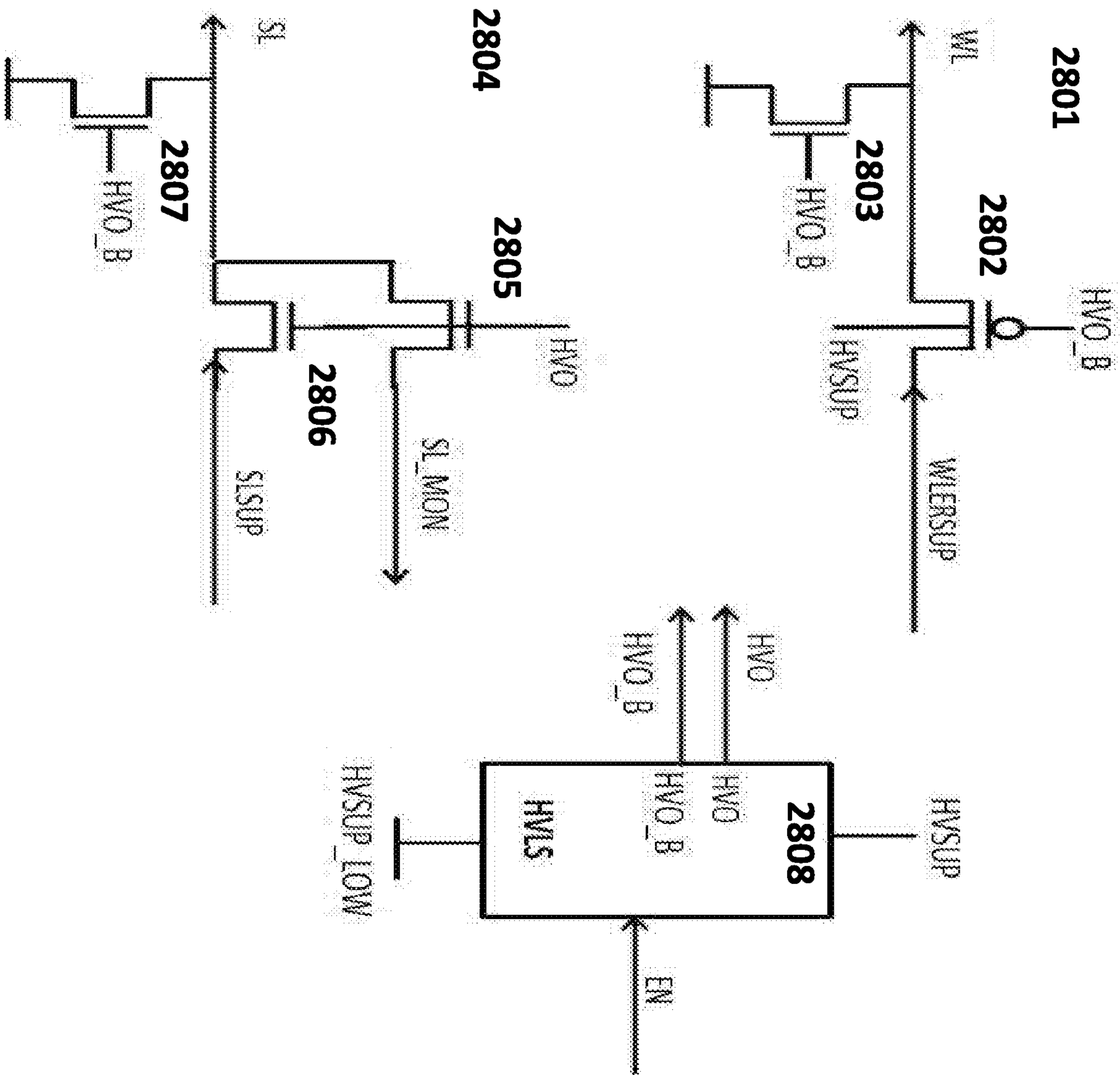


FIGURE 28

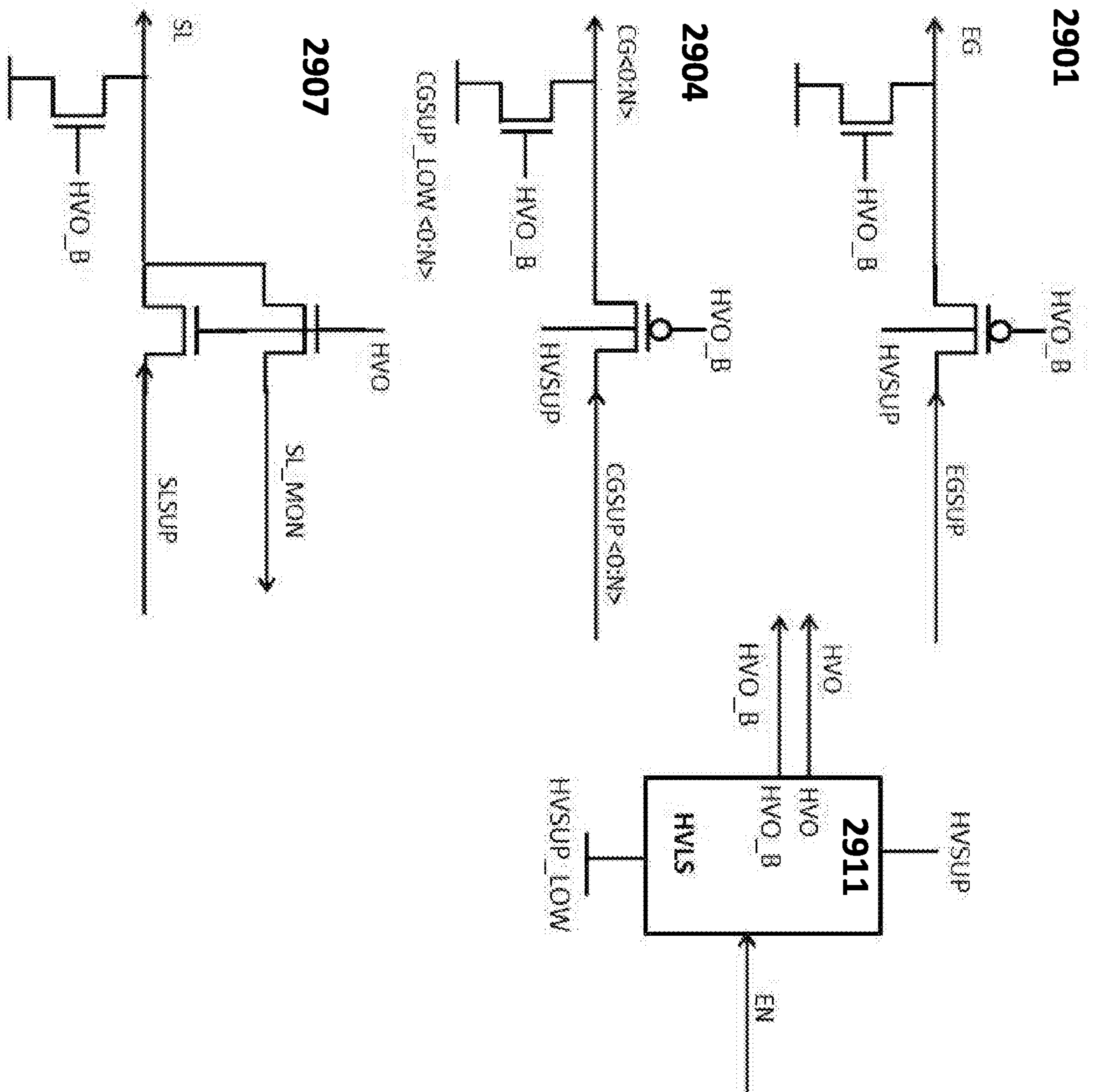
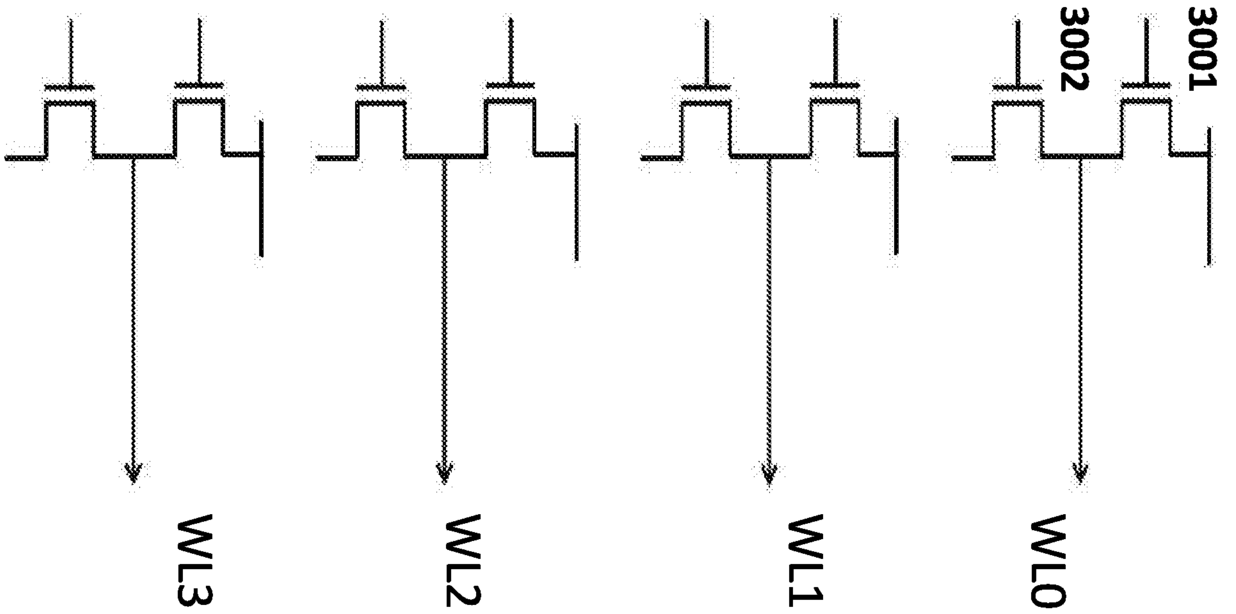


FIGURE 29

FIGURE 30



3000

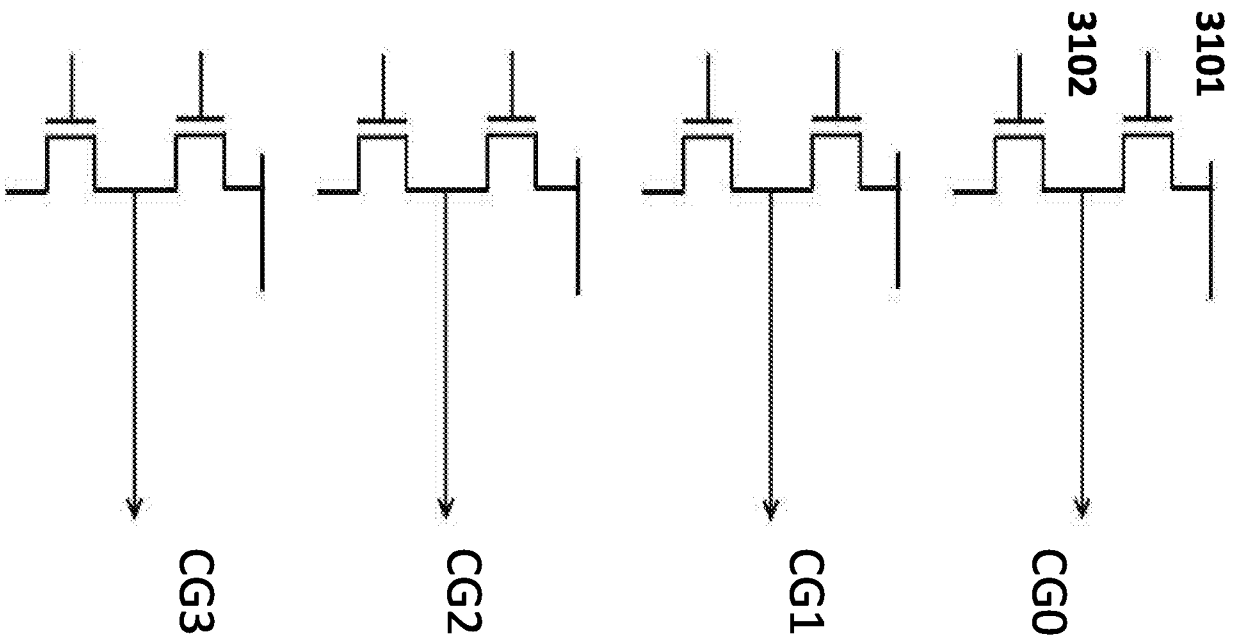


FIGURE 31

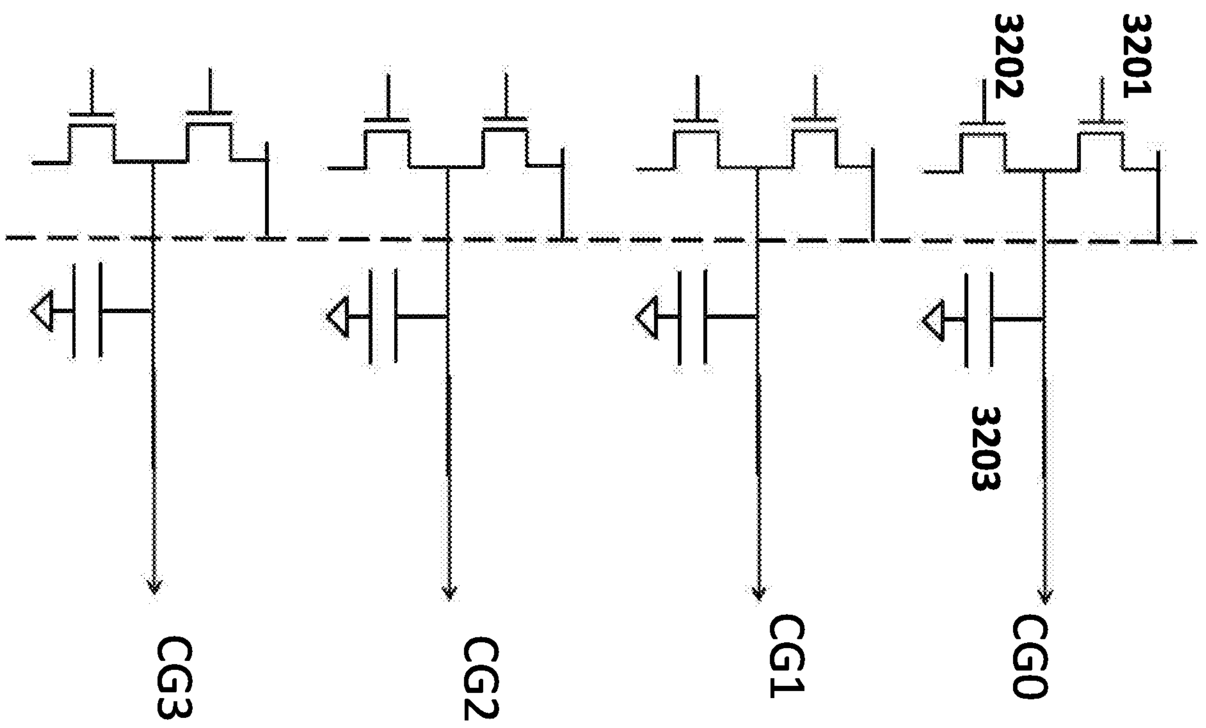
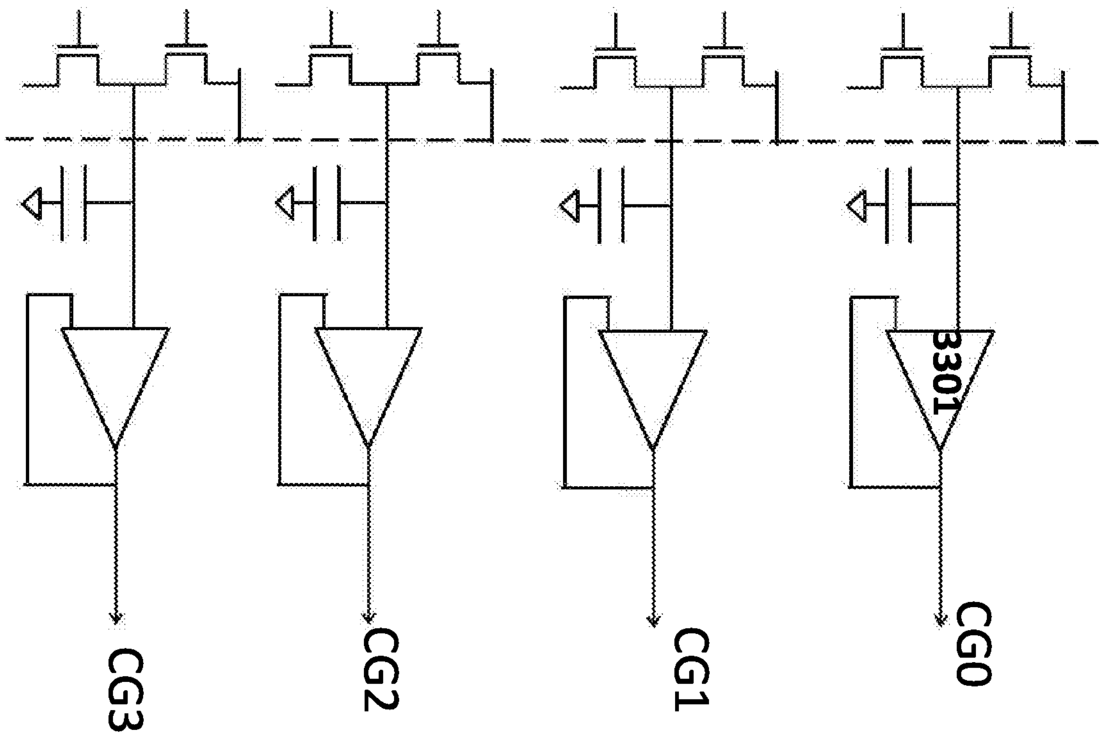


FIGURE 32

FIGURE 33



3300

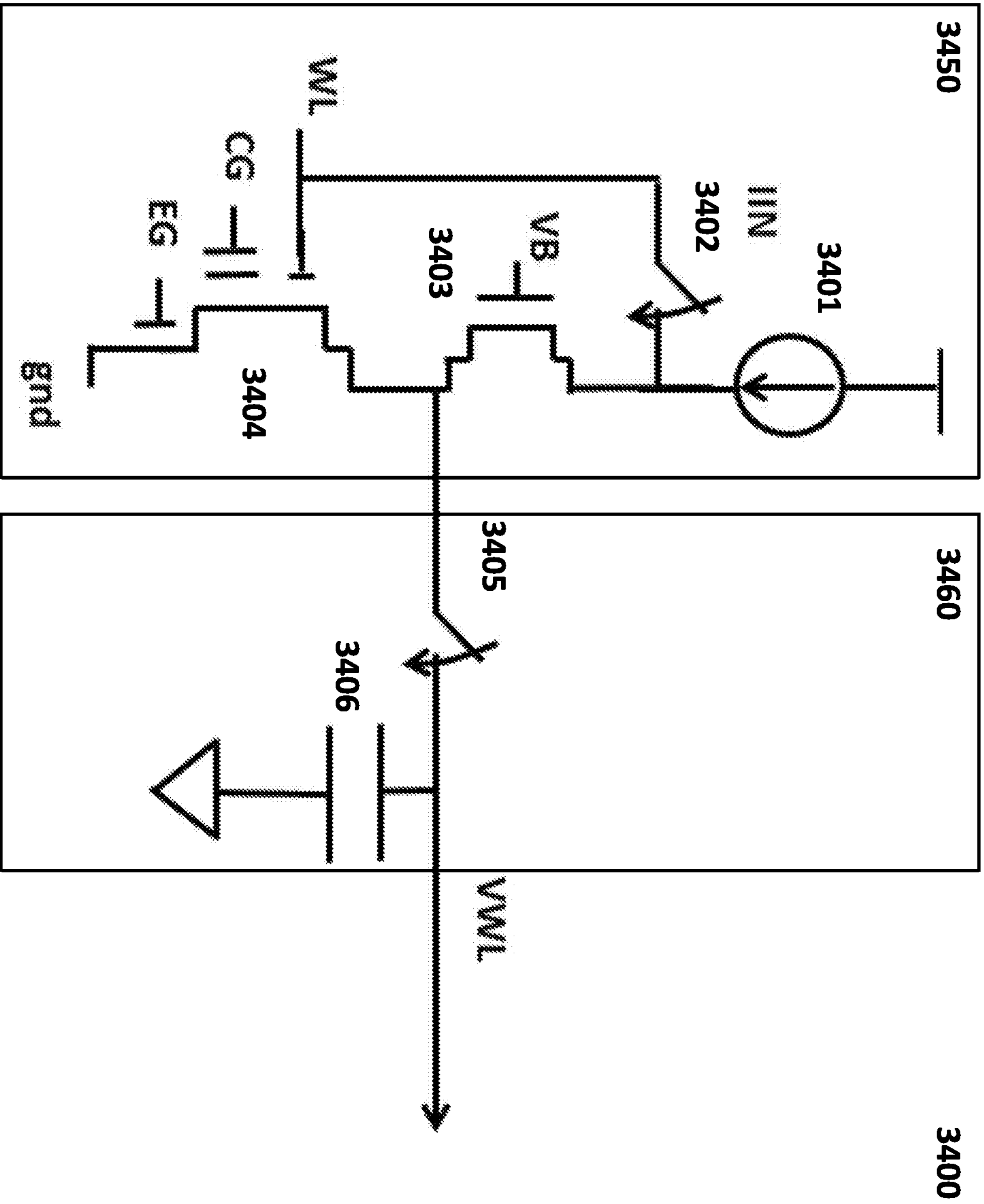


FIGURE 34
3450
WL
CG
EG
IIN
3402
3401
VB
3403
3404
gnd
3460
3405
3406
VWL
3400

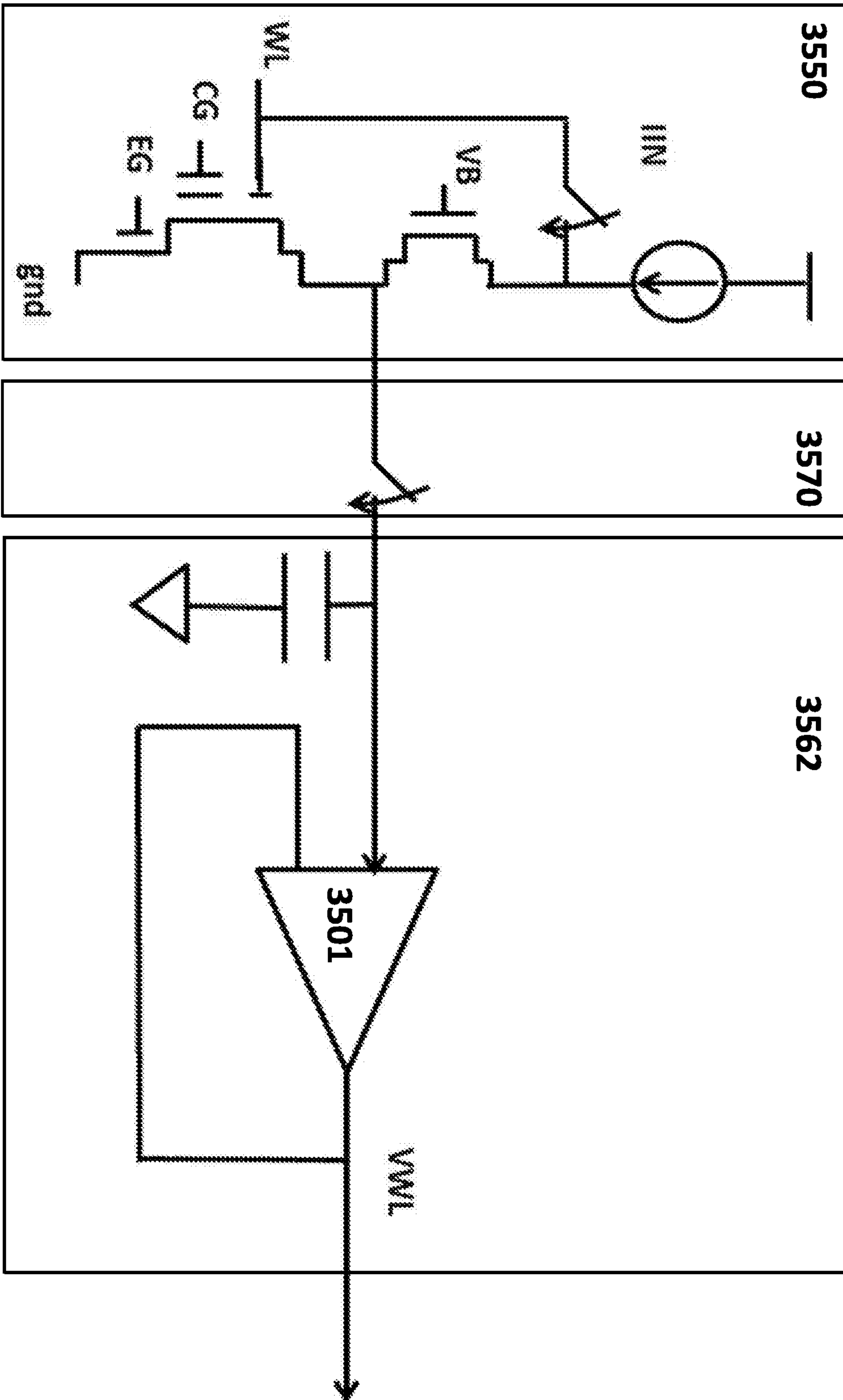


FIGURE 35

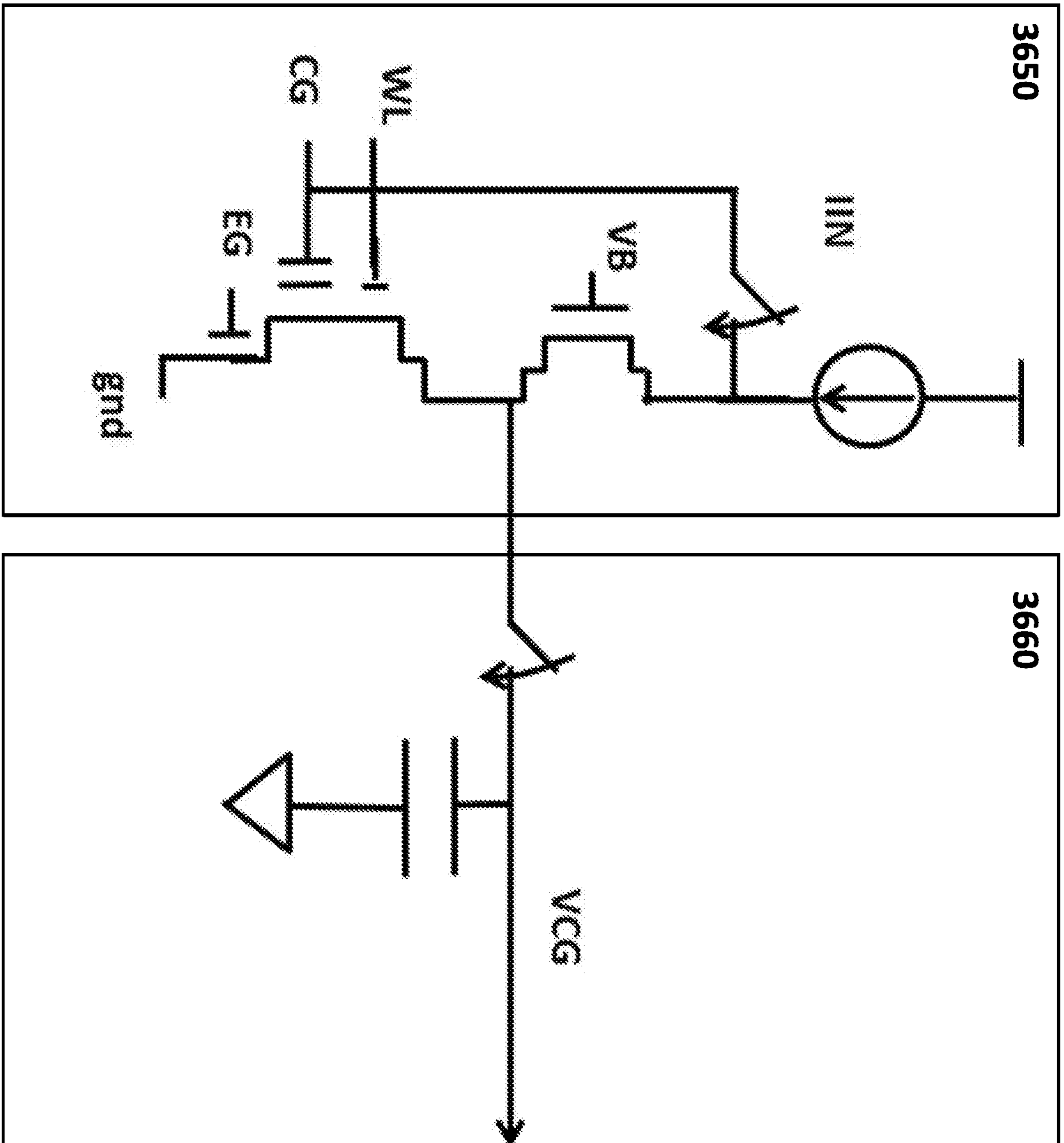


FIGURE 36
3650
CG
WL
EG
VB
IIN
gnd
3660
VCG
3600

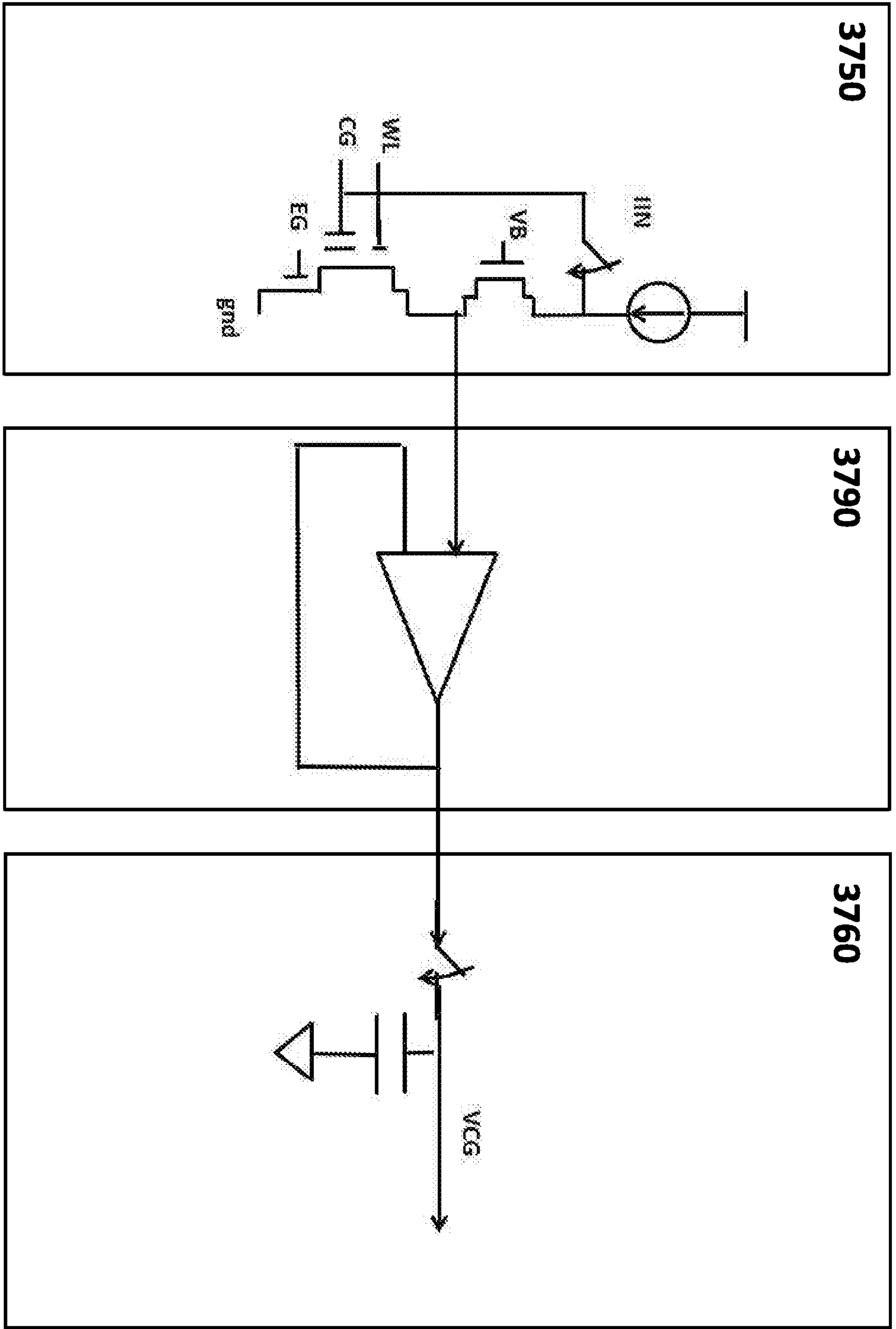


FIGURE 37

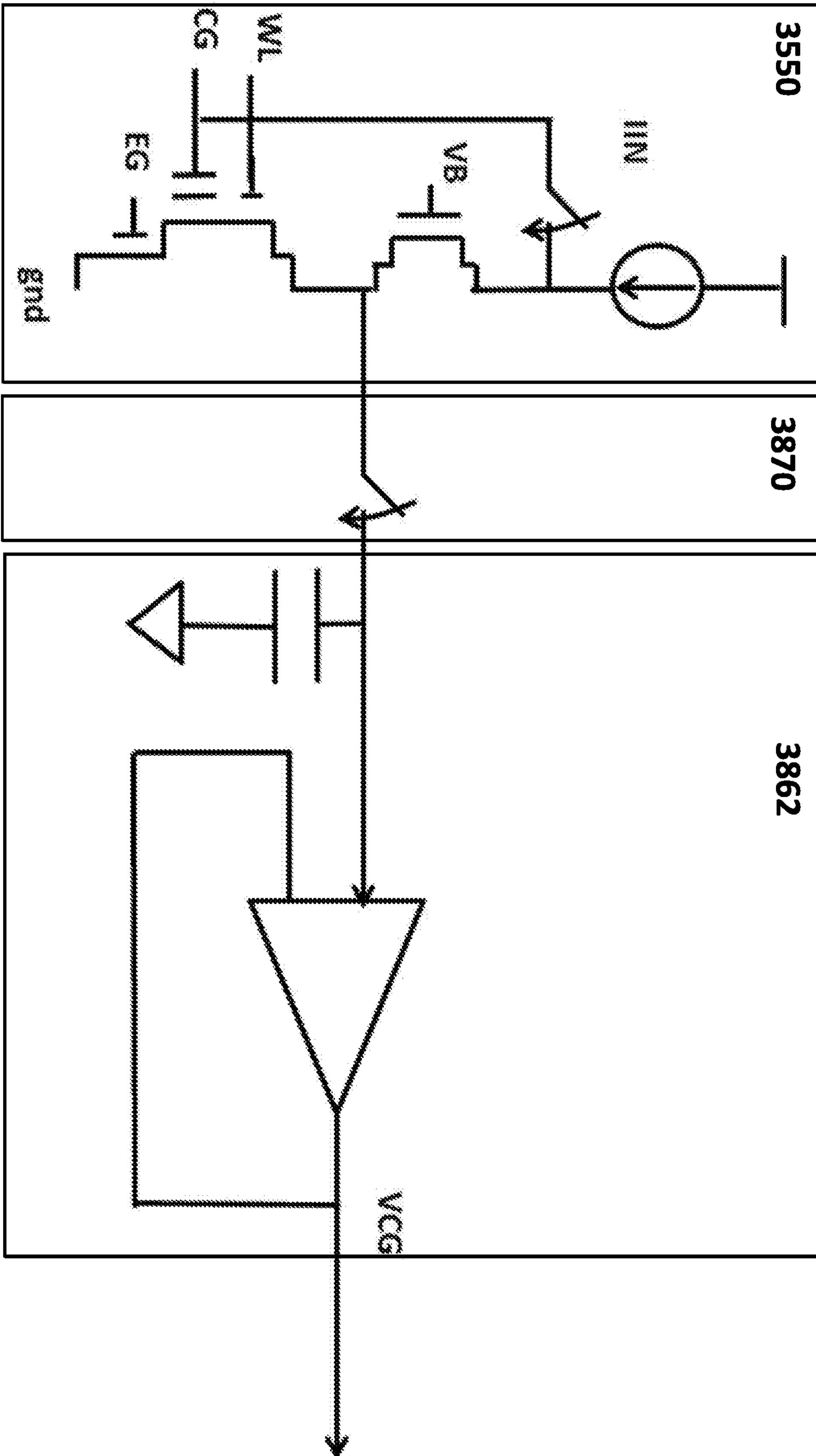
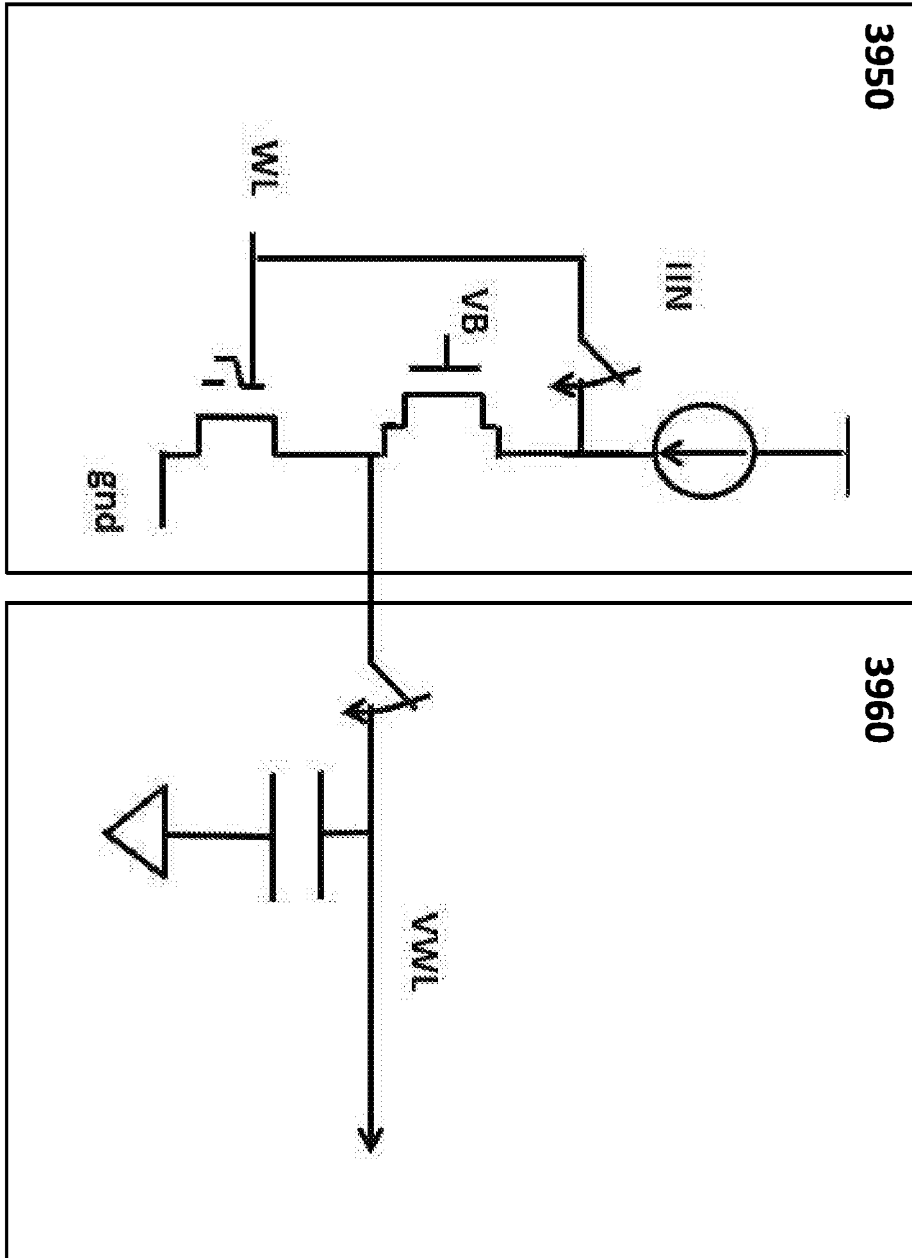


FIGURE 38



3900

FIGURE 39

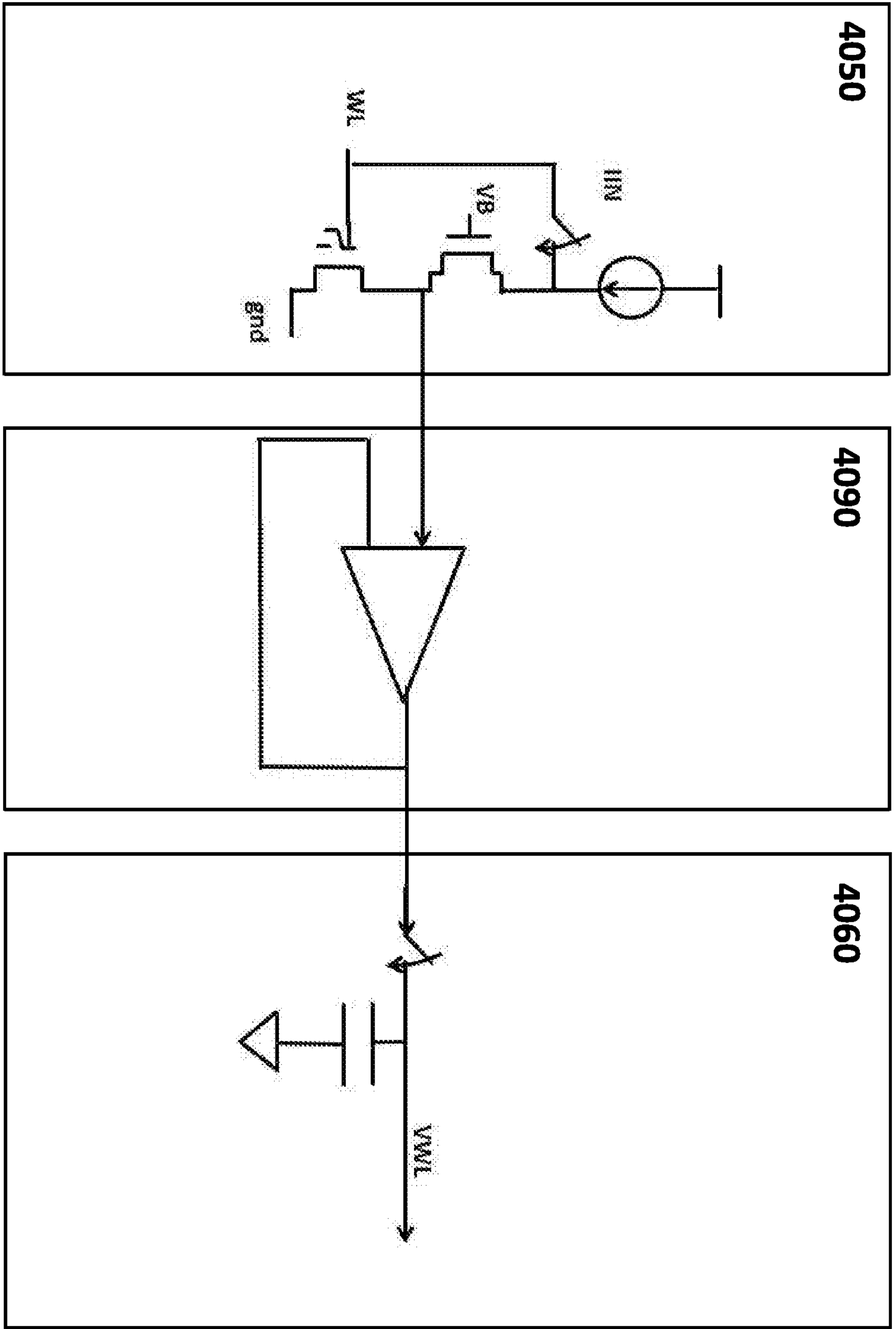


FIGURE 40

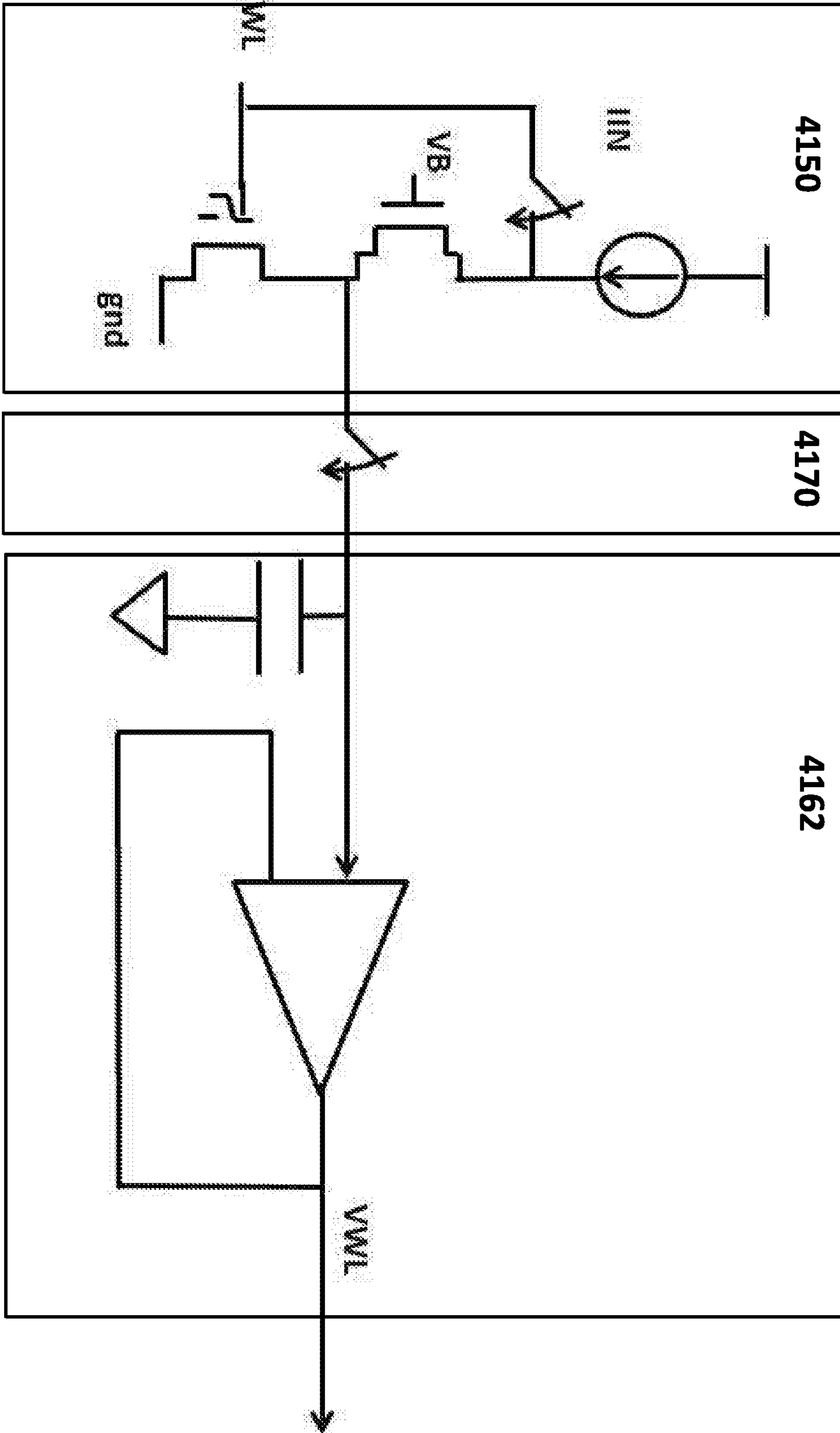


FIGURE 41

FIGURE 42

	WL	WL -unsel	BL	BL -unsel	SL	SL -unsel
Read	**1-3.5V**	-0.5V/0V	**0.6-2V (Ineuron)**	0.6V-2V/0V	**0V**	0V
Erase	**~5-13V**	0V	**0V**	0V	**0V**	0V
Program	**1-2V**	-0.5V/0V	**0.1-3 uA**	Vinh ~2.5V	**4-10V**	0-1V/FLT

FIGURE 43

	WL	WL -unsel	**BL**	BL -unsel	**SL**	SL -unsel
Read	**1-3.5V**	-0.5V/0V	**0.6-2V**	0.6V-2V/0V	**~1V (Ineuron)**	0V
Erase	**~5-13V**	0V	**0V**	0V	**0V**	**SL-inhibit** (~4-8V)
Program	**1-2V**	-0.5V/0V	**0.1-3 uA**	Vinh ~2.5V	**4-10V**	0-1V/FLT

FIGURE 44

	WL	WL -unsel	BL	BL -unsel	CG	CG -unsel same sector	CG -unsel	EG	EG -unsel	SL	SL -unsel
Read	1.0-2V	-0.5V/ 0V	0.6-2V (Ineuron)	0V	0-2.6V	0-2.6V	0-2.6V	0-2.6V	0-2.6V	0V	0V
Erase	0V	0V	0V	0V	0V	0-2.6V	0-2.6V	5-12V	0-2.6V	0V	0V
Program	0.7-1V	-0.5V/ 0V	0.1-1uA	Vinh (1-2V)	4-11V	0-2.6V	0-2.6V	4.5-5V	0-2.6V	4.5-5V	0-1V

FIGURE 45

	WL	WL -unsel	BL	BL -unsel	CG	CG -unsel same sector	CG -unsel	EG	EG -unsel	SL	SL -unsel
Read	**1.0-2V**	**-0.5V/ 0V**	**0.6-2V (Ineuron)**	0V	**0-2.6V**	0-2.6V	0-2.6V	**0-2.6V**	0-2.6V	**0V**	0V
Erase	**0V**	0V	**0V**	0V	**0V**	**CGINH (4-9V)**	0-2.6V	**5-12V**	0-2.6V	**0V**	0V
Program	**0.7-1V**	**-0.5V/ 0V**	**0.1-1uA**	Vinh (1-2V)	**4-11V**	0-2.6V	0-2.6V	**4.5-5V**	0-2.6V	**4.5-5V**	0-1V

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US19/15022

A. CLASSIFICATION OF SUBJECT MATTER

IPC - H03M 13/11, 13/15, 13/29, 13/37 (2019.01)

CPC - H03M 13/1105, 13/1575, 13/2948, 13/37

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.
X --- Y	SHAFIEE, A et al. " ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars "; 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), Seoul; Publication [online]. October 2016 [retrieved 18 July 2019]. Retrieved from the Internet: <URL: https://www.cs.utah.edu/~rajeev/pubs/isca16.pdf >; pp 1-4.	1, 10 --- 2-9, 11-17
Y	US 4810910 (SCHOELLIKOPF, J et al.) 07 March 1989; claims 2, 5	2, 3, 11
Y	US 2012/0087188 A1 (HSIEH, C et al.) 12 April 2012; paragraphs [0014], [0019], [0028], [0046]	4-7, 9, 12-15, 17
Y	US 5721702 A (BRINER, M) 24 February 1998; abstract	8, 16

☐ 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	Date of mailing of the international search report
19 July 2019 (19.07.2019)	05 AUG 2019

Name and mailing address of the ISA/US	Authorized officer
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Shane Thomas Telephone No. PCT Helpdesk: 571-272-4300

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US19/15022

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

Group I: Claims 1-17; Group II: Claims 18-19, 27-63; Group III: Claims 20-26; Group IV: Claims 64-66

-***-Continued in extra sheet-***-

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Group I: Claims 1-17

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US19/15022

-***-Continued from Box No. III - Observations where unity of invention is lacking-***-

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fee must be paid.

Group I: Claims 1-17 are directed towards a bit line decoder.
Group II: Claims 18-19 and 27-63 are directed towards an analog neuromorphic memory system comprising a word line decoder circuit, a source line decoder circuit, a sample and hold capacitor; and a current-to-voltage circuit.
Group III: Claims 20-26 are directed towards a word line driver.
Group IV: Claims 64-66 are directed towards an analog neuromorphic memory system comprising a redundancy sector and a non-volatile register.

The inventions listed as Groups I-IV do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

The special technical features of Group I include at least a first circuit for enabling individual bit lines during a program-and-verify operation; and a second circuit for enabling all bit lines during a read operation, which are not present in Groups II-IV.

The special technical features of Group II include at least a word line decoder circuit coupled to the word line terminals of the non-volatile memory cells, wherein the word line decoder circuit is capable of applying a low voltage or a high voltage to coupled word line terminals; a source line decoder circuit; a sample and hold capacitor coupled to each word line, which are not present in Groups I, III, and IV.

The special technical features of Group III include at least a plurality of select transistors, each of the plurality of select transistors comprising a first terminal, a second terminal, and a gate, wherein the gate of each of the plurality of select transistors is coupled to a common control line and the first terminal of each of the plurality of select transistors is coupled to a different word line, and wherein the second terminal of each of the plurality of select transistors is coupled to one or more bias transistors; wherein the bias transistors coupled to each of the plurality of select transistors are capable of providing a bias voltage to a single select transistor or to all of the select transistors, which are not present in Groups I, II, and IV.

The special technical features of Group IV include at least a redundancy sector and a non-volatile register for storing system information, which are not present in Groups I-III.

The common technical features shared by Groups I-V are an analog neuromorphic memory system comprising: a vector-by-matrix multiplication array, the vector-by matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns, wherein each column is connected to a bit line.

However, these common features are previously disclosed by "ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars" by SHAFIEE et al. (hereinafter "Shafiee"). Shafiee discloses an analog neuromorphic memory system (using convolutional neural networks for a memristor crossbar array; Abstract, page 1) comprising: a vector-by-matrix multiplication array (vector-matrix multiplier; Fig. 1b, page 3), the vector-by matrix multiplication array comprising an array of non-volatile memory cells organized into rows and columns (a memristor crossbar used as a vector-matrix multiplier; Fig. 1b, page 3), wherein each column is connected to a bit line (Figs. 1a and 1b, page 3).

Since the common technical features are previously disclosed by the Shafiee reference, these common features are not special and so Groups I-V lack unity.