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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
02 July 2020 (02.07.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/139895 A1

(51) International Patent Classification:
G11C 14/00 (2006.01)

(21) International Application Number:
PCT/US2019/068495

(22) International Filing Date:
24 December 2019 (24.12.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/784,636 24 December 2018 (24.12.2018) US

(71) Applicant: **THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK** [US/US]; 110 Low Memorial Library, 535 West 116th Street, New York, NY 10027 (US).

(72) Inventors: **SEOK, Mingoo**; 115 Westervelt Avenue, Tenafly, NJ 07670 (US). **JIANG, Zhewei**; 6422 217th Street, Oakland Gardens, NY 11364 (US). **SEO, Jae-sun**; 1101 E Playa Del Norte Drive #4138, Tempe, AZ 85287 (US). **YIN, Shihui**; 2433 W Main Street Apt 2015, Mesa, AZ 85201 (US).

(74) Agent: **BYRNE, Matthew, T.** et al.; Byrne Poh LLP, 11 Broadway, Suite 760, New York, NY 10004 (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,

(54) Title: CIRCUITS AND METHODS FOR IN-MEMORY COMPUTING

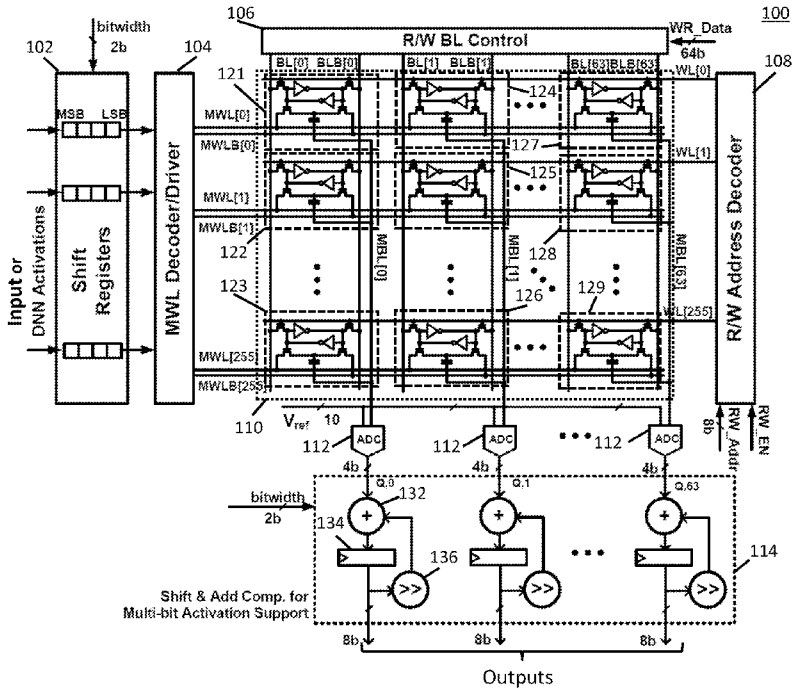


FIG. 1

(57) Abstract: In some embodiments, an in-memory-computing SRAM macro based on capacitive- coupling computing (C3) (which is referred to herein as "C3SRAM") is provided. In some embodiments, a C3SRAM macro can support array-level fully parallel computation, multi-bit outputs, and configurable multi-bit inputs. The macro can include circuits embedded in bitcells and peripherals to perform hardware acceleration for neural networks with binarized weights and activations in some embodiments. In some embodiments, the macro utilizes analog-mixed- signal capacitive-coupling computing to evaluate the main computations of binary neural networks, binary-multiply-and-accumulate operations. Without needing to access the stored weights by individual row, the macro can assert all of its rows simultaneously and form an analog voltage at the read bitline node through capacitive voltage division, in some embodiments. With one analog-to-digital converter (ADC) per column, the macro cab realize fully parallel vector-matrix multiplication in a single

WO 2020/139895 A1

[Continued on next page]

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, 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))*

cycle in accordance with some embodiments.

CIRCUITS AND METHODS FOR IN-MEMORY COMPUTING

Cross Reference to Related Application

[0001] This application claims the benefit of United States Provisional Patent Application No. 62/784,636, filed December 24, 2018, which is hereby incorporated by reference herein in its entirety.

Statement Regarding Government Funded Research

[0002] This invention was made with government support under 1652866 awarded by the National Science Foundation. The Government has certain rights in this invention.

Background

[0003] Deep learning tasks in edge computing are becoming ever more prevalent as their application is driven by the massive growth in the Internet-of-Thing paradigm. This has motivated research into low-energy hardware architectures for deep neural networks (DNN).

[0004] Recent research has investigated low precision and even binary representations of weights and activations, which allow each weight of a DNN to be mapped to one static random-access memory (SRAM) bitcell. Conventional SRAM architectures, however, still require millions or even billions of row-by-row memory accesses for DNNs, limiting improvements in energy-efficiency. To address this, recent research has demonstrated SRAM-based in-memory computing, which performs embedded multiply-and-accumulate (MAC) computation inside an SRAM without the need for explicit memory access. However, known mechanisms for SRAM-based in-memory computing are deficient in many ways.

[0005] Accordingly, it is desirable to provide new circuits and methods for in-memory computing.

Summary

[0006] In accordance with some embodiments, circuits and methods for in-memory computing are provided. In some embodiments, circuits for a bitcell are provided, the circuits comprising: a first switch having a first side, a second side, and a control input wherein the first side of the first switch is connected to a first supply voltage; a second switch having a first side, a second side, and a control input wherein the control input of the second switch is connected to the second side of the first switch, and the second side of the second switch is connected to the control input of the first switch; a third switch having a first side, a second side, and a control input wherein the first side of the third switch is connected to the second side of the first switch, the second side of the third switch is connected to a second supply voltage, and the control input of the third switch is connected to the control input of the first switch; a fourth switch having a first side, a second side, and a control input wherein the first side of the fourth switch is connected to the second side of the second switch, the second side of the fourth switch is connected to the second supply voltage, and the control input of the fourth switch is connected to the control input of the second switch; a fifth switch having a first side, a second side, and a control input wherein the first side of the fifth switch is connected to the second side of the first switch; a sixth switch having a first side, a second side, and a control input wherein the first side of the sixth switch is connected to the second side of the second switch; a seventh switch having a first side, a second side, and a control input wherein the control input of the seventh switch is connected to the second side of the first switch; an eighth switch having a first side, a second side, and a control input wherein the control input of the eighth switch is connected to the second side of the second switch; and a capacitor having a first side and a second side wherein the first side of the capacitor is connected to the first side of the seventh switch and the first side of the eighth switch.

Brief Description of the Drawings

[0007] Various objects, features, and advantages of the disclosed subject matter can be more fully appreciated with reference to the following detailed description of the disclosed subject matter when considered in connection with the following drawings, in which like reference numerals identify like elements.

[0008] FIG. 1 is an example of a schematic diagram of a C3SRAM macro in accordance with some embodiments;

[0009] FIG. 2A is an example of a schematic diagram of two bitcells in accordance with some embodiments.

[0010] FIG. 2B is an example of a timing diagram showing the operation of a bitcell in accordance with some embodiments.

[0011] FIG. 2C is an example of a schematic diagram showing the operation of a bitcell during a first step in accordance with some embodiments.

[0012] FIG. 2D is an example of a schematic diagram showing the operation of a bitcell during a second step in accordance with some embodiments.

[0013] FIG. 2E is an example of a table showing different voltage levels in a C3SRAM macro for different inputs/activations and weights in accordance with some embodiments.

[0014] FIG. 3A is an example of a schematic diagram of a portion of an analog-to-digital converter that can be used with a C3SRAM macro in accordance with some embodiments.

[0015] FIG. 3B is an example of a timing diagram showing the operation of a portion of an analog-to-digital converter that can be used with a C3SRAM macro in accordance with some embodiments.

[0016] FIG. 3C is an example of a schematic diagram showing the operation of a portion of an analog-to-digital converter that can be used with a C3SRAM macro during a second step in accordance with some embodiments.

[0017] FIG. 3D is an example of a schematic diagram showing the operation of a portion of an analog-to-digital converter that can be used with a C3SRAM macro during a third step in accordance with some embodiments.

[0018] FIG. 4 is an example of a timing diagram showing relative timing of different signals during the operation a C3SRAM macro in accordance with some embodiments.

[0019] FIG. 5 is an example of a schematic diagram showing how multi-bit weights can be implemented in a C3SRAM macro in accordance with some embodiments.

[0020] FIG. 6 is an example of a schematic diagram showing a C3SRAM macro configured to accumulate over a row with a row-wise read/write in accordance with some embodiments.

[0021] FIG. 7 is an example of a schematic diagram showing a pull-up reset transistor for bitcells in accordance with some embodiments.

[0022] FIG. 8 is an example of a schematic diagram showing an alternate configuration of a bitcell in accordance with some embodiments.

[0023] FIG. 9 is an example of a block diagram of components that can be used to implement a neural network in accordance with some embodiments.

Detailed Description

[0024] In accordance with some embodiments, circuits and methods for in-memory computing are provided.

[0025] In some embodiments, an in-memory-computing SRAM macro based on capacitive-coupling computing (C3) (which is referred to herein as "C3SRAM") is provided. In some embodiments, a C3SRAM can support array-level fully parallel computation, multi-bit outputs, and configurable multi-bit inputs.

[0026] In some embodiments, the macro is of an SRAM module with circuits embedded in bitcells and peripherals to perform hardware acceleration for neural networks

with binarized weights and activations. In some embodiments, the macro utilizes analog-mixed-signal capacitive-coupling computing to evaluate the main computations of binary neural networks, binary-multiply-and-accumulate operations. Without needing to access the stored weights by individual row, the macro can assert all of its rows simultaneously and form an analog voltage at the read bitline node through capacitive voltage division, in some embodiments. With one analog-to-digital converter (ADC) per column, the macro realizes fully parallel vector-matrix multiplication in a single cycle in accordance with some embodiments.

[0027] In some embodiments, multi-bit inputs/activations can be used improve convolutional neural network (CNN) and/or DNN accuracies over binary activations, especially for large datasets. In some embodiments, two to four-bit inputs can achieve competent accuracy in target inference tasks. In some embodiments, the input can be configurable to have a bitwidth between 1 and 4 bits. In some embodiments, multi-bit inputs can be processed across multiple cycles from LSB to MSB, where each cycle's digital MAC result is halved (right shifted) and accumulated to the next cycle's output.

[0028] FIG. 1 illustrates an example architecture 100 of a 256x64 C3SRAM macro in accordance with some embodiments. As shown, architecture 100 includes shift registers 102, MAC wordline decoder/driver 104, read/write (R/W) bitline (BL) control 106, R/W address decoder 108, array 110, analog-to-digital converters (ADCs) 112, and shift-and-add components 114.

[0029] Shift registers 102 can be any suitable shift registers (e.g., having any suitable number of bits) and any suitable number of shift registers can be included in shift registers 102. For example, in some embodiments, shift register 102 can include 256 configurable four-bit shift registers. Each of these shift registers can be configured to provide an output for a corresponding row of array 110. The configuration of the shift registers can be controlled by the two-bit (when using a four-bit shift register) bitwidth signal received at the

top of shift registers 102. The bitwidth signal can determine whether the shift register receives one, two, three, or four bits for each input or activation signal received at the left of shift registers.

[0030] MAC wordline (MWL) decoder/driver 104 can be any suitable circuit for driving MWL and MWL bar (MWLB) lines of array 110 as described further in connection with FIGS. 2A-2E below in some embodiments. For each cycle of the C3SRAM macro (as described in FIGS. 2A-2E below), MWL decoder/driver 104 receives one bit for each row of array 110 from a corresponding shift register of shift registers 102. In some embodiments, the ports of MAC wordlines (MWL and MWLB) of the bitcell can be capacitive and therefore small MWL drivers can drive all of the bitcells in a row of array 110.

[0031] In some embodiments, in order to improve accuracies of a C3SRAM macro, MWL decoder/driver 104 can have a maximum drive voltage V_{DR} that is 200mV (or any other suitable value (e.g., 150mV, 250mV, etc.) lower than the core voltage (V_{CORE}) which is used as the maximum drive voltage of R/W BL control 106 and R/W address decoder 108. Thus, in some embodiments, when V_{CORE} is equal to 1.0VDC, V_{DR} is equal to 0.8VDC.

[0032] In some embodiments, MWL decoder/driver 104 can drive MWL and MWLB lines with binary values (e.g., 0VDC and 0.8VDC). In some embodiments, MWL decoder/driver 104 can drive MWL and MWLB lines with any other suitable type of value, such as ternary values (e.g., 0VDC, 0.4VDC, and 0.8VDC).

[0033] The multi-bit input/activations received by shift registers 102 can be in the two's-complement format in some embodiments. In such instances, MWL decoder/driver 104 can produce the voltage ramping combination for -1 or 0 for the MSB cycle and +1 or 0 for all other cycles in some embodiments.

[0034] R/W BL control 106 can be any suitable read/write control circuit (as known in the art) as used in an SRAM in some embodiments. R/W BL control 106 can receive any

suitable signal (such a WR_Data (64 bits)) for controlling the bitlines (BL[0]-BL[64]) as shown in FIG. 1.

[0035] R/W address decoder 108 can be any suitable read/write address decoder (as known in the art) as used in a SRAM in some embodiments. R/W address decoder 108 can receive any suitable signals (such as RW_EN and RW_Addr (8 bits)) for decoding addresses on the wordlines (WL[0]-WL[255]) as shown in FIG. 1.

[0036] Array 110 can be an array of bitcells, such as bitcells 121-129. Any suitable number of bitcells can be included in array 110, and each bitcell can have any suitable architecture, in some embodiments. For example, as illustrated in FIG. 1, array 110 can include 256x64 (256 rows by 64 columns) bitcells, though any suitable number of rows and columns can be used in some embodiments, and each bitcell can have the architecture described in connection with, and illustrated in, FIG. 2A. Each bitcell can store one binary weight in some embodiments. These binary weights can be used to convolve the values received on the MWL and MWLB lines (MWL[0], MWLB[0]; MWL[1], MWLB[1]; … ; MWL[255], MWLB[255]) output by MWL decoder/driver 104.

[0037] ADCs 112 can be any suitable analog-to-digital converters in some embodiments, and any suitable number of ADCs 112 can be used. For example, as illustrated in FIG. 1, an ADC 112 can be provided for each column of array 110 in some embodiments. In some embodiments, ADCs 112 can be implemented with flash analog-to-digital converters that are each pitch-matched to a bitcell. In some embodiments, the ADCs can have any suitable number of levels and any suitable voltage resolution.

[0038] As shown in FIG. 1, the ADCs can receive as input the voltage levels on the MAC bitlines (e.g., MBL[0]-MBL[63]).

[0039] During operation, the ADCs can convert analog values of the dot-product of inputs/activations and weights to the digital counterparts for other DNN computations such as non-linear activations, max-pooling, etc.

[0040] Shift-and-add components 114 can be any suitable shift-and-add components in some embodiments, and any suitable number of shift-and-add components can be used in some embodiments. For example, as shown in FIG. 1, one shift-and-add component can be provided for each column, and each shift-and-add component can include an adder 132, a register 134, and a shifter 136 so that the adder can add the output of an ADC 112 to the output of the shifter, and that output can then be stored in the register. The bitwidth (e.g., 2 bits) signal received at the left side of shift-and-add components 114 can control register 134 so that it is cleared upon initialization and after the number of bits in the input or DNN activations are processed. Thus, shift-and-add components 114 can receive and accumulate the ADC outputs over multiple cycles in some embodiments.

[0041] The outputs at the bottom of shift-and-add components 114 can be the outputs for a layer of a DNN, which can then be fed back into a new layer of a DNN, or can be the outputs for the entire DNN, depending on the implementation used, in some embodiments.

[0042] In some embodiments, it may be desirable to further combine the analog signals on the MLB lines before converting the signals from the analog domain to the digital domain. In such cases, any suitable analog connections (including simple wires, analog multiplexers, analog switches, etc.) can be provided between the MLB lines.

[0043] As shown in FIG. 2A in accordance with some embodiments, a bitcell in array 110 (such as one or more of bitcells 121-129) can be formed from eight transistors (T1-T8) and one capacitor (C_C) as shown in each of bitcells 202 and 204 of FIG. 2. The transistors can be implemented using any suitable transistors, such as PMOS transistors (e.g., transistors T1 and T2) and NMOS transistors (e.g., transistors T3-T8) in some embodiments. The capacitors can be implemented using any suitable capacitor technology (e.g., using nmoscap or any other suitable technology) and have any suitable size (e.g., 3.88fF) in some embodiments.

[0044] The bitcells can compute bitwise XNORs of the values on the MAC wordlines (MWL[0], MWLB[0]; … ; MWL[255], MWLB[255]) and bitlines (BL[0]-BL[63]) using capacitive coupling as a computation mechanism in some embodiments.

[0045] In accordance with some embodiments, the steps for in-bitcell computation at a bitcell that are performed during a cycle of the C3SRAM can be as follows.

[0046] First, as shown by the left column in FIG. 2B, the voltage across capacitor C_C in the bitcell can be initialized (e.g., to ~0VDC) by driving the MAC bitline (MBL) to V_{RST} (~$0.5 \cdot V_{DR}$) using footer transistor TFT and by driving MWL and MWLB to V_{RST} using MWL decoder/driver 104. At this point, the capacitors of the bitcells are effectively arranged as shown in FIG. 2C. That is, all of the capacitors are arranged in parallel between voltages equal to V_{RST}.

[0047] Second, following this reset phase, as shown in the right column of FIG. 2B, the MBL is allowed to float by turning off the TFT. For each input/activation bit (e.g., 256 bits as shown in FIG. 1) from a corresponding shift register of shift registers 102, an input of either +1, 0, or -1 can then be applied as follows. For the input of +1, MWL decoder/driver 104 drives the MWL from V_{RST} to V_{DR} while the MWLB is driven from V_{RST} to VSS. For the input of -1, MWL decoder/driver 104 drives the MWL from V_{RST} to VSS while the MWLB is driven from V_{RST} to V_{DR}. For the input of 0, both MWL and MWLB are held at V_{RST}. When the weight is +1, the voltage ramping via T7 induces a displacement current I_C through capacitor C_C in the bitcell, whose magnitude is:

$$I_C = C_C \cdot dV_{MWL}/dt,$$

where V_{MWL} is the voltage of MWL, and t is time. If the weight is -1, the voltage ramping via T8 induces a displacement current I_C through capacitor C_C in the bitcell, whose magnitude is also:

$$I_C = C_C \cdot dV_{MWLB}/dt,$$

where V_{MWLB} is the voltage of MWL, and t is time. The amount of the charge transferred from the bitcell to MBL is then formulated as:

$Q_{Ci}=\int_0^{t_1} I_C \cdot dt = 0.5 \cdot C_C \cdot V_{DR}$,

where t_1 is the time it takes V_{MWL} to reach V_{DR}. The shared MBL potential for each column containing 256 bitcells is set to:

$V_{MBL}=C_C \cdot V_{DR} \cdot \sum_1^{256}(XNOR_i)/(256 \cdot C_C+Cp)$,

where XNORi is the XNOR output of the i-th bitcell output and the value encoded in MWL/MWLB, and Cp is the parasitic capacitance of MBL plus the input capacitance of the ADC at the bottom of the column. At this point, the capacitors of the bitcells are effectively arranged as shown in FIG. 2D. That is, n of the capacitors are arranged in parallel between V_{DR} and MBL and 256-n of the capacitor are arranged in parallel between MBL and VSS. In this arrangement, the capacitors form a capacitive voltage divider.

[0048] FIG. 2E summarizes values at different points in FIG. 2A for different inputs/activations and weights in accordance with some embodiments. As shown, when a bitcell is resetting or when the input is 0, MWL, MWLB and the voltage at node V_C (the top side of the capacitor in the bitcell) are all equal to V_{RST}. When the input/activation is +1: the MWL is V_{DR}; the MWLB is 0VDC; and V_C is V_{DR} when the weight is +1 and V_C is 0VDC when the weight is -1. When the input/activation is -1: the MWL is 0VDC; the MWLB is V_{DR}; and V_C is 0VDC when the weight is +1 and V_C is V_{DR} when the weight is -1. When the weight is +1, Q is V_{CORE} and QB is 0VDC. When the weight is -1, Q is 0VDC and QB is V_{CORE}.

[0049] Turning to FIG. 3A-3D, details of some analog-to-digital converters that can be used as ADCs 112 in some embodiments are provided. In some embodiments, each ADC can include ten double-sampling based self-calibrating single-ended comparators 300 as shown in FIG. 3A. Each comparator 300 can include an offset-cancelling capacitor 302

followed by an inverter chain 304, where the first inverter 306 acts as an amplifier and a controlled negative feedback circuit 307 is provided, in some embodiments.

[0050] In some embodiments, the ADCs can operate as follows.

[0051] First, as shown in the left column of FIG. 3B and in FIG. 3C, during step 2: MBL connects to the left side of the capacitor 302 via transistors 308 and 310 based on signals RST0 and RST0B; and transistor 312 connects V_X to V_Y based on signal RSTB (which is the opposite (bar) of RST of FIG. 2B), which places first inverter 306 in a high gain region.

[0052] Second, as shown in the right column of FIG. 3B and in FIG. 3D, during step 3: the left side of capacitor 302 is then connected to the reference voltage (VREF) by transistors 314 and 316 based on signals RST1 and RST1B; and transistor 312 disconnects V_X from V_Y based on signal RSTB (which is the opposite (bar) of RST of FIG. 2B). The voltage differential between V_{MBL} and VREF then causes charging or discharging on capacitor 302 (based on the differential), which in turn tips the inverter chain toward VDD or VSS. The gain-stage 318 of the inverter chain completes the amplification to the digital domain.

[0053] In the aforementioned three-step procedure, relevant signal transitions in step 1 (left side of FIG. 2B and FIG. 2C) and step 3 (right side of FIG. 3B and FIG. 3D) are decoupled in separate modules, meaning that while the digital output is being evaluated by the ADC, the memory array can begin computing the next batch of binary multiply and accumulate (bMAC) operations, in some embodiments. This allows a pipeline of a half-cycle in which step 1 and step 3 operate concurrently in some embodiments.

[0054] In some embodiments, the bMAC operation is timing sensitive. To minimize analog non-idealities, concurrent signal switches described above can be controlled to follow the order shown in FIG. 4. As shown, transitions from steps 1 and 3 to step 2 can follow this order: 1) the reference voltage is disconnected from capacitor 302 before MBL leaves reset; 2) the negative feedback on the inverter stage is turned on before MBL is connected to

capacitor 302; 3) MBL is connected to capacitor 302 before MBL leaves reset; and 4) MWL is not driven until MBL is floating.

[0055] As also shown, the transitions from step 2 to step 1 and 3 can follow this order: 1) MBL is disconnected from capacitor 302 before MWL drivers switch to reset voltage; 2) also, MBL is disconnected before MBL reset footer turns on; 3) also, MBL is disconnected before the negative feedback is turned off; and 4) the negative feedback is switched off before the reference voltage is connected to capacitor 302.

[0056] In some embodiments, for a confined range of bMAC values (ADC range), ten uniformly spaced reference voltages can be chosen to obtain competent accuracy for the target inference tasks. The ADC range can be determined by the statistical distribution of partial computation results of a DNN algorithm, which tend to have concentrated distribution around bMAC value of 0, in some embodiments.

[0057] Turning to FIG. 5, in some embodiments, multiple bitcells each with a different size capacitor can be connected to the same values of WL, MWL, MWLB, and MLB to realize multi-bit weights that can be controlled by two or more bitline pairs, such as: BL0[0], BLB0[0]; BL1[0], BLB1[0]; etc. Although only two bitcells are shown in FIG. 5, three or more bitcells can be used in some embodiment, with each more significant bitcell having a capacitor double the size of the capacitor in next less significant bitcell.

[0058] While the circuits of FIG. 1 and FIG. 2A accumulate over a column with a row-wise read/write, in some embodiments a C3SRAM macro can be configured to accumulate over a row with a row-wise read/write as shown in FIG. 6. Likewise, in some embodiments, C3SRAM macros can be defined to accumulate over a column with a column-wise read/write (by defining the columns in FIG. 1 and FIG. 2A as being rows and defining the rows in FIG. 1 and FIG. 2A as being columns) and can be defined to accumulate over a row with a column-wise read/write (by defining the columns in FIG. 6 as being rows and defining the rows in FIG. 6 as being columns).

[0059] While FIG. 2A illustrated an NMOS transistor that pulls the MBL line down to V_{RST}, in some embodiments, a PMOS transistor can be used to pull the MBL line up to any suitable V_{RST} as shown in FIG 7. Likewise, in some embodiments, a transmission gate or passgate can be used to pull the MBL line to any suitable voltage.

[0060] As shown in FIG. 8, in some embodiments, transistors T7 and T8 can be implemented in PMOS transistors instead of NMOS transistors as shown in FIG. 2A. In some embodiments, when implemented using PMOS transistors, the MWL and MWLB lines can be driven between voltages of 0.2VDC and 1.0VDC instead of 0.0VDC and 0.8 VDC as described above. As described herein, a 200mV offset in the low rail voltage or the high rail voltage may be beneficial for improving accuracy of the C3SRAM. However, any suitable offset value (e.g., 150mV, 250mV, etc.) or no offset value can be used in some embodiments.

[0061] In some embodiments, one or more C3SRAM macros as described herein can be used to perform the computations of convolution layers and fully connected layers of a DNN.

[0062] In some embodiments, the mapping of fully connected layer weights in C3SRAM can be implemented with weights of a layer organized column-wise, and inputs/activations can be applied at each row. In some embodiments, convolutional layer mapping can be performed as an extension of a fully connected layer mapping. For example, mapping a 3x3x256 filter from a convolution layer can be implemented in the same manner as the mapping of nine 256-neuron fully connected layer weights in some embodiments. In some embodiments, channels can be organized in column orientation, and each channel's kernel can be distributed across multiple macros. In some embodiments, partial sums produced by ADCs can be accumulated to generate a pre-activation for each neuron.

[0063] In some embodiments, each binary weight of convolution and fully connected layers can be stored in one C3SRAM bitcell. In some embodiments, the macro can compute the partial 256-input MAC operations. Accumulation of the outputs of the macro, max-pooling, and batch normalization can be performed in digital domain, in some embodiments.

[0064] As described above, the C3SRAM macro can be used to implement a DNN, a CNN, or any other suitable neural network in some embodiments. In such an application, the C3SRAM macro can be part of any suitable general-purpose computer or special-purpose computer in some embodiments. Any such general-purpose computer or special-purpose computer can include any suitable hardware in some embodiments. For example, as illustrated in example hardware 900 of FIG. 9, such hardware can include hardware processor 902, memory and/or storage 904, an input device controller 906, an input device 908, display/audio drivers 910, display and audio output circuitry 912, communication interface(s) 914, an antenna 916, and a bus 918.

[0065] Hardware processor 902 can include any suitable hardware processor, such as a microprocessor, a micro-controller, digital signal processor(s), dedicated logic, and/or any other suitable circuitry for controlling the functioning of a general-purpose computer or a special-purpose computer in some embodiments. In some embodiments, hardware processor 902 can be controlled by a program stored in memory 904.

[0066] Memory and/or storage 904 can be any suitable memory and/or storage for storing programs, data, and/or any other suitable information in some embodiments. For example, memory and/or storage 904 can include a C3SRAM macro as described herein, other random access memory, read-only memory, flash memory, hard disk storage, optical media, and/or any other suitable memory.

[0067] Input device controller 906 can be any suitable circuitry for controlling and receiving input from one or more input devices 908 in some embodiments. For example, input device controller 906 can be circuitry for receiving input from a touchscreen, from a keyboard, from one or more buttons, from a voice recognition circuit, from a microphone, from a camera, from an optical sensor, from an accelerometer, from a temperature sensor, from a near field sensor, from a pressure sensor, from an encoder, and/or any other type of input device.

[0068] Display/audio drivers 910 can be any suitable circuitry for controlling and driving output to one or more display/audio output devices 1912 in some embodiments. For example, display/audio drivers 910 can be circuitry for driving a touchscreen, a flat-panel display, a cathode ray tube display, a projector, a speaker or speakers, and/or any other suitable display and/or presentation devices.

[0069] Communication interface(s) 914 can be any suitable circuitry for interfacing with one or more communication networks. For example, interface(s) 914 can include network interface card circuitry, wireless communication circuitry, and/or any other suitable type of communication network circuitry.

[0070] Antenna 916 can be any suitable one or more antennas for wirelessly communicating with a communication network in some embodiments. In some embodiments, antenna 916 can be omitted.

[0071] Bus 918 can be any suitable mechanism for communicating between two or more components 902, 904, 906, 910, and 914 in some embodiments.

[0072] Any other suitable components can be included in hardware 900 in accordance with some embodiments.

[0073] In some embodiments, any suitable computer readable media can be used for storing instructions for performing the functions and/or processes herein. For example, in some embodiments, computer readable media can be transitory or non-transitory. For example, non-transitory computer readable media can include media such as non-transitory forms of magnetic media (such as hard disks, floppy disks, and/or any other suitable magnetic media), non-transitory forms of optical media (such as compact discs, digital video discs, Blu-ray discs, and/or any other suitable optical media), non-transitory forms of semiconductor media (such as flash memory, electrically programmable read-only memory (EPROM), electrically erasable programmable read-only memory (EEPROM), and/or any other suitable semiconductor media), any suitable media that is not fleeting or devoid of any

semblance of permanence during transmission, and/or any suitable tangible media. As another example, transitory computer readable media can include signals on networks, in wires, conductors, optical fibers, circuits, any suitable media that is fleeting and devoid of any semblance of permanence during transmission, and/or any suitable intangible media.

[0074] Accordingly, circuits and methods for in-memory computing are provided.

[0075] Although the disclosed subject matter has been described and illustrated in the foregoing illustrative embodiments, the present disclosure has been made only by way of example, and numerous changes in the details of implementation of the disclosed subject matter can be made without departing from the spirit and scope of the disclosed subject matter, which is limited only by the claims that follow. Features of the disclosed embodiments can be combined and rearranged in various ways.

What is claimed is:

1. A circuit for a bitcell, comprising:

a first switch having a first side, a second side, and a control input wherein the first side of the first switch is connected to a first supply voltage;

a second switch having a first side, a second side, and a control input wherein the control input of the second switch is connected to the second side of the first switch, and the second side of the second switch is connected to the control input of the first switch;

a third switch having a first side, a second side, and a control input wherein the first side of the third switch is connected to the second side of the first switch, the second side of the third switch is connected to a second supply voltage, and the control input of the third switch is connected to the control input of the first switch;

a fourth switch having a first side, a second side, and a control input wherein the first side of the fourth switch is connected to the second side of the second switch, the second side of the fourth switch is connected to the second supply voltage, and the control input of the fourth switch is connected to the control input of the second switch;

a fifth switch having a first side, a second side, and a control input wherein the first side of the fifth switch is connected to the second side of the first switch;

a sixth switch having a first side, a second side, and a control input wherein the first side of the sixth switch is connected to the second side of the second switch and the control input of the sixth switch is connected to the control input of the fifth switch;

a seventh switch having a first side, a second side, and a control input wherein the control input of the seventh switch is connected to the second side of the first switch;

an eighth switch having a first side, a second side, and a control input wherein the control input of the eighth switch is connected to the second side of the second switch; and

a first capacitor having a first side and a second side wherein the first side of the first capacitor is connected to the first side of the seventh switch and the first side of the eighth switch.

2. The circuit of claim 1, further comprising

a ninth switch having a first side, a second side, and a control input wherein the first side of the ninth switch is connected to the first supply voltage;

a tenth switch having a first side, a second side, and a control input wherein the control input of the tenth switch is connected to the second side of the ninth switch, and the second side of the tenth switch is connected to the control input of the ninth switch;

an eleventh switch having a first side, a second side, and a control input wherein the first side of the eleventh switch is connected to the second side of the ninth switch, the second side of the eleventh switch is connected to the second supply voltage, and the control input of the eleventh switch is connected to the control input of the ninth switch;

a twelfth switch having a first side, a second side, and a control input wherein the first side of the twelfth switch is connected to the second side of the tenth switch, the second side of the twelfth switch is connected to the second supply voltage, and the control input of the twelfth switch is connected to the control input of the tenth switch;

a thirteenth switch having a first side, a second side, and a control input wherein the first side of the thirteenth switch is connected to the second side of the ninth switch and the second side of the thirteenth switch is connected to the second side of the fifth switch;

a fourteenth switch having a first side, a second side, and a control input wherein the first side of the fourteenth switch is connected to the second side of the tenth switch, the control input of the fourteenth switch is connected to the control input of the thirteenth switch, the second side of the fourteenth switch is connected to the second side of the sixth switch;

a fifteenth switch having a first side, a second side, and a control input wherein the control input of the fifteenth switch is connected to the second side of the ninth switch;

a sixteenth switch having a first side, a second side, and a control input wherein the control input of the sixteenth switch is connected to the second side of the tenth switch; and

a second capacitor having a first side and a second side wherein the first side of the second capacitor is connected to the first side of the fifteenth switch and the first side of the sixteenth switch.

3. The circuit of claim 2, wherein the second side of the second capacitor is connected to the second side of the first capacitor and wherein during a first time period the first side of the first capacitor, the first side of the second capacitor, the second side of the first capacitor, and the second side of the second capacitor are all connected to a reset voltage.

4. The circuit of claim 2, wherein the second side of the second capacitor is connected to the second side of the first capacitor and wherein during a second time period the first side of the first capacitor is connected to a third supply voltage and the first side of the second capacitor is connected to the second supply voltage.

5. The circuit of claim 4, wherein the third supply voltage is less than the first supply voltage and greater than the second supply voltage.

6. The circuit of claim 5, wherein the first supply voltage is higher than the third supply voltage by an amount greater than a threshold voltage of the seventh transistor.

7. The circuit of claim 2, further comprising

a seventeenth switch having a first side, a second side, and a control input wherein the first side of the seventeenth switch is connected to the first supply voltage;

a eighteenth switch having a first side, a second side, and a control input wherein the control input of the eighteenth switch is connected to the second side of the seventeenth switch, and the second side of the eighteenth switch is connected to the control input of the seventeenth switch;

a nineteenth switch having a first side, a second side, and a control input wherein the first side of the nineteenth switch is connected to the second side of the seventeenth switch, the second side of the nineteenth switch is connected to the second supply voltage, and the control input of the nineteenth switch is connected to the control input of the seventeenth switch;

a twentieth switch having a first side, a second side, and a control input wherein the first side of the twentieth switch is connected to the second side of the eighteenth switch, the second side of the twentieth switch is connected to the second supply voltage, and the control input of the twentieth switch is connected to the control input of the eighteenth switch;

a twenty first switch having a first side, a second side, and a control input wherein the first side of the twenty first switch is connected to the second side of the seventeenth switch and the control input of the twenty first switch is connected to the control input of the fifth switch and to the control input of the sixth switch;

a twenty second switch having a first side, a second side, and a control input wherein the first side of the a twenty second switch is connected to the second side of the eighteenth switch and the control input of the twenty second switch is connected to the control input of the twenty first switch;

a twenty third switch having a first side, a second side, and a control input wherein the control input of the twenty third switch is connected to the second side of the seventeenth switch;

a twenty fourth switch having a first side, a second side, and a control input wherein the control input of the twenty fourth switch is connected to the second side of the eighteenth switch; and

a third capacitor having a first side and a second side wherein the first side of the third capacitor is connected to the first side of the twenty third switch and the first side of the twenty fourth switch.

8. The circuit of claim 7, wherein the second side of the second capacitor is connected to the second side of the third capacitor.

9. The circuit of claim 7, wherein the third capacitor is twice the size of the first capacitor.

10. The circuit of claim 1, further comprising a reset switch connected to the second side of the first capacitor that discharges the capacitor when on.

11. The circuit of claim 10, wherein the reset switch is an NMOS transistor.

12. The circuit of claim 10, wherein the reset switch is an PMOS transistor.

13. The circuit of claim 10, wherein the reset switch is a transmission gate.

14. The circuit of claim 1, further comprising an analog to digital converter (ADC) having an input and an output, wherein the input is connected to the second side of the first capacitor.

15. The circuit of claim 14, further comprising:

an adder having a first input, a second input, and an output, wherein the first input of the adder is connected to the output of the ADC;

a register having an input and an output, wherein the input of the register is connected to the output of the adder; and

a shifter having an input and an output, wherein the input of the shifter is connected to the output of the register and the output of the shifter is connected to the second input of the adder.

16. The circuit of claim 1, wherein the seventh transistor and the eighth transistor are NMOS transistors.

17. The circuit of claim 1, wherein the seventh transistor and the eighth transistor are PMOS transistors.

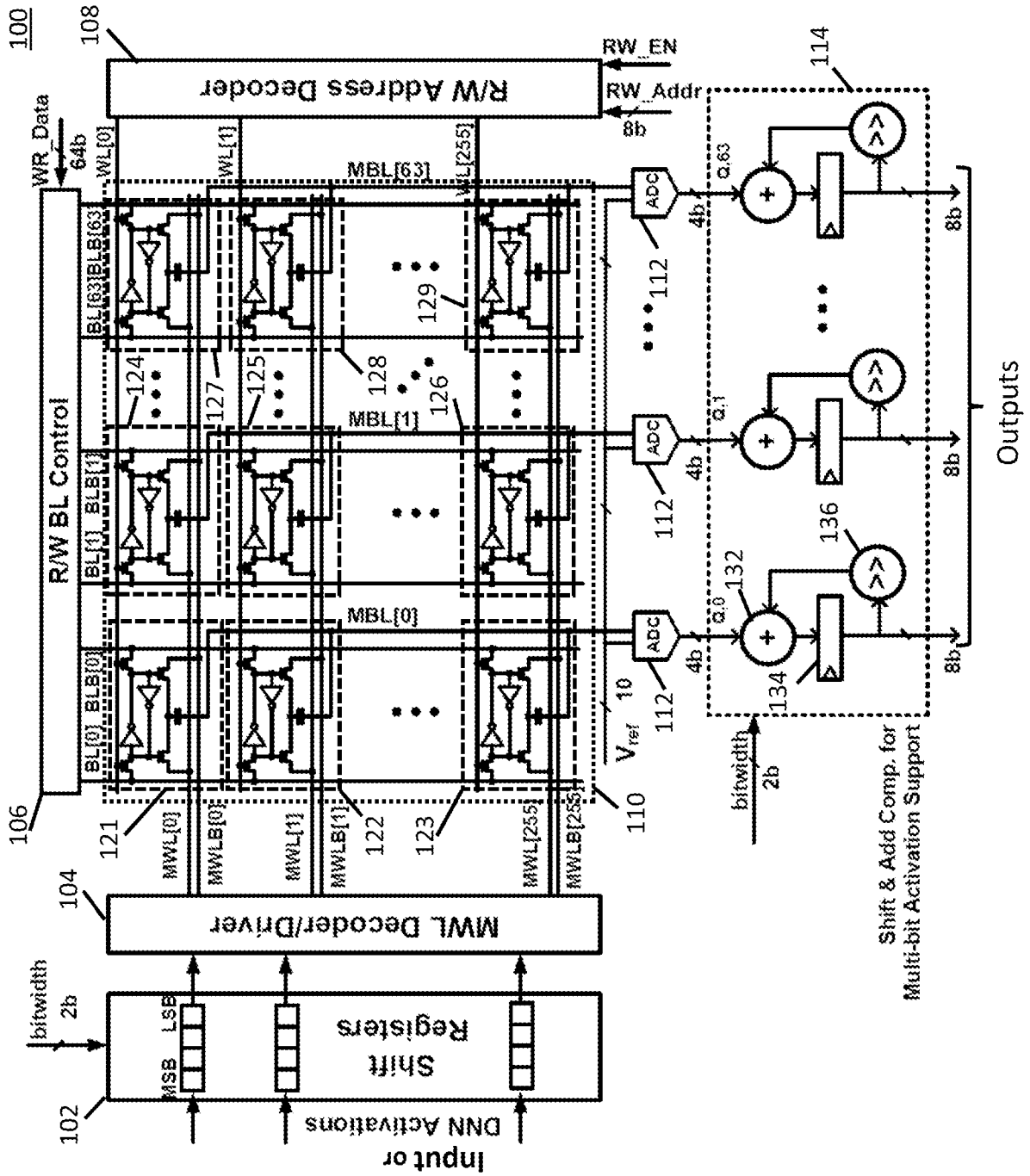

100
108
R/W Address Decoder
RW_EN
RW_Addr
8b
WR_Data
64b
WL[0]
WL[1]
WL[255]
MBL[63]
MBL[1]
MBL[0]
BL[63] BLB[63]
BL[1] BLB[1]
BL[0] BLB[0]
R/W BL Control
106
124
127
125
128
126
129
121
122
123
110
MWL[0]
MWLB[0]
MWL[1]
MWLB[1]
MWL[255]
MWLB[255]
104
MWL Decoder/Driver
bitwidth
2b
102
LSB
MSB
Shift Registers
Input or DNN Activations
ADC
112
4b
V_{ref}
10
114
Q.63
Q.1
Q.0
132
136
134
8b
Outputs
bitwidth
2b
Shift & Add Comp. for Multi-bit Activation Support

FIG. 1

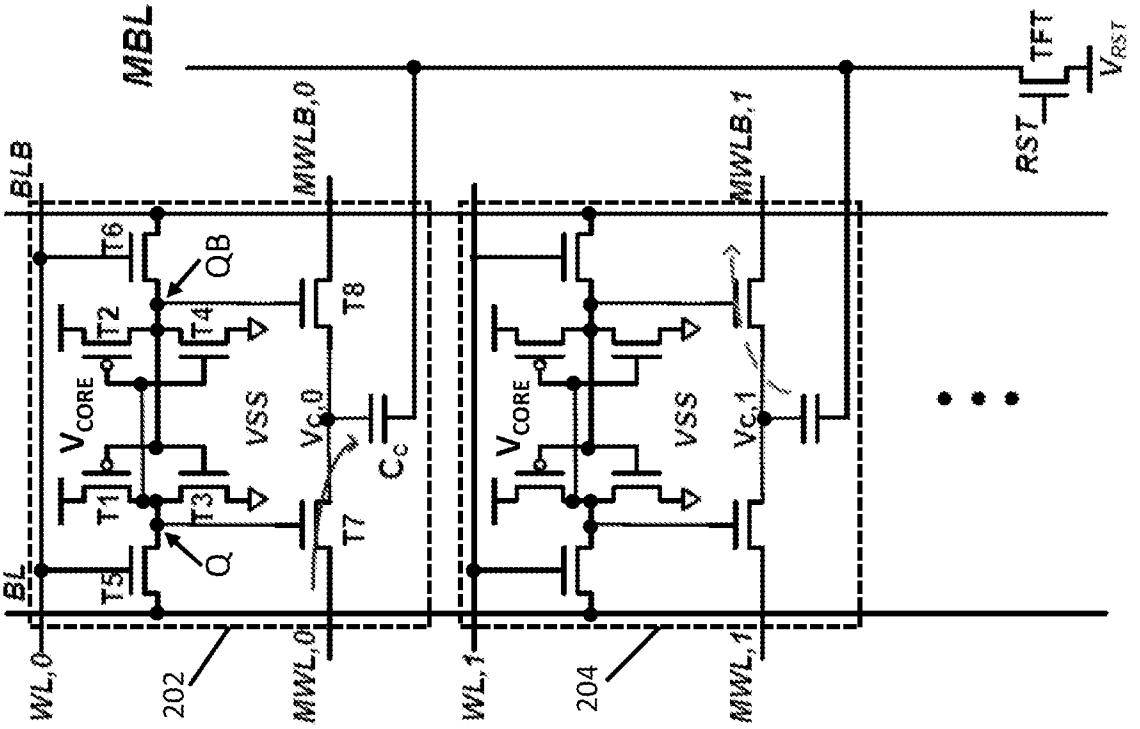

MBL
BLB
BL
MWLB,0
MWLB,1
RST
TFT
V_{RST}
T1
T2
T3
T4
T5
T6
T7
T8
QB
Q
V_{CORE}
VSS
$V_{C,0}$
$V_{C,1}$
C_C
WL,0
WL,1
MWL,0
MWL,1
202
204

FIG. 2A

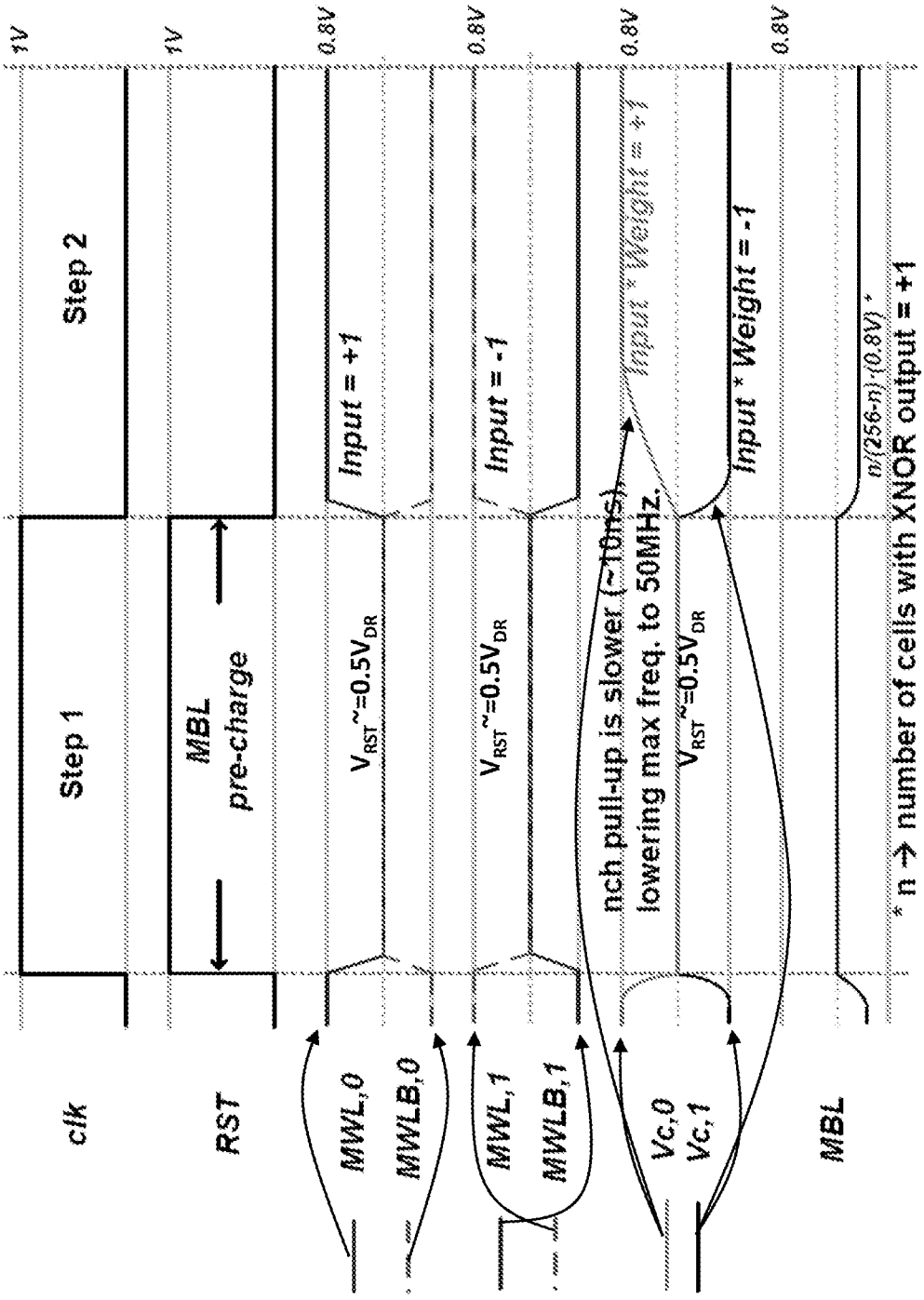

1V
1V
0.8V
0.8V
0.8V
0.8V
Step 1
Step 2
clk
RST
MBL
pre-charge
MWL,0
MWLB,0
MWL,1
MWLB,1
Vc,0
Vc,1
MBL
Input = +1
Input = -1
Input * Weight = +1
Input * Weight = -1
V_{RST}~=0.5V_{DR}
V_{RST}~=0.5V_{DR}
V_{RST}~=0.5V_{DR}
nch pull-up is slower (~10ns),
lowering max freq. to 50MHz.
n/(256-n)·(0.8V) *
* n → number of cells with XNOR output = +1

FIG. 2B

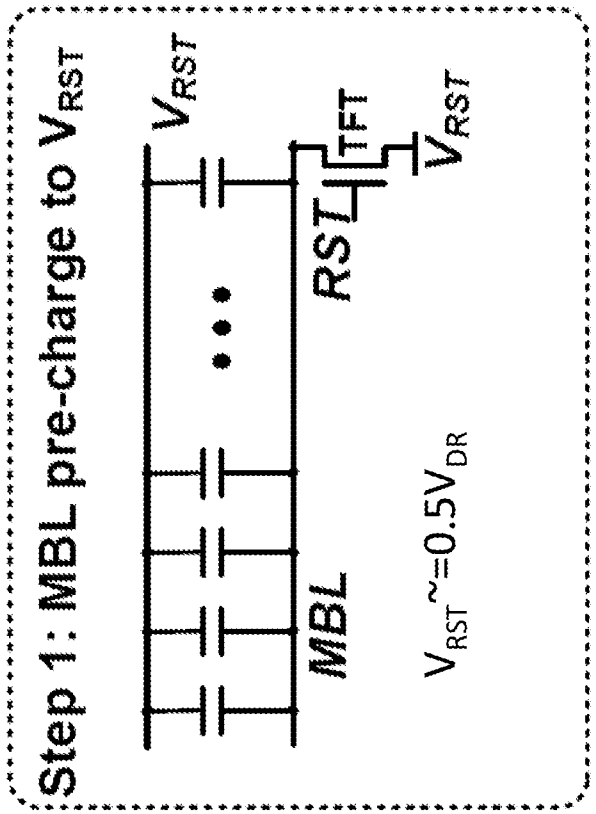


FIG. 2C

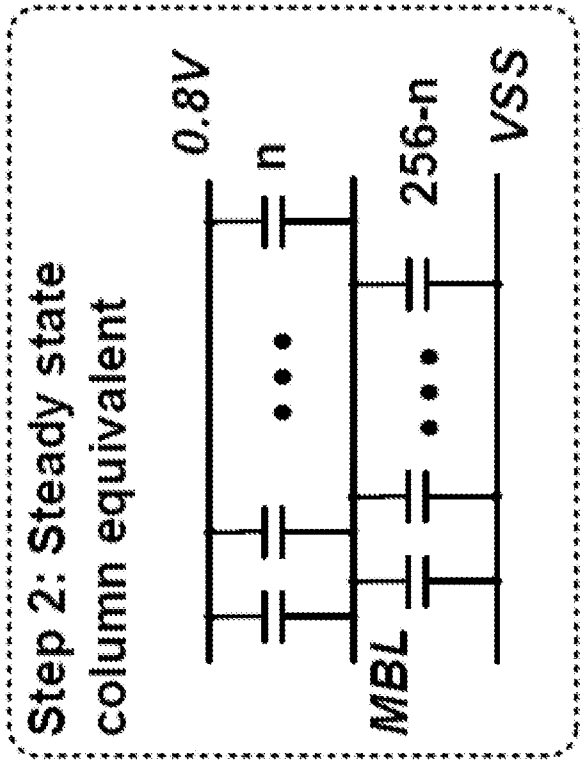


FIG. 2D

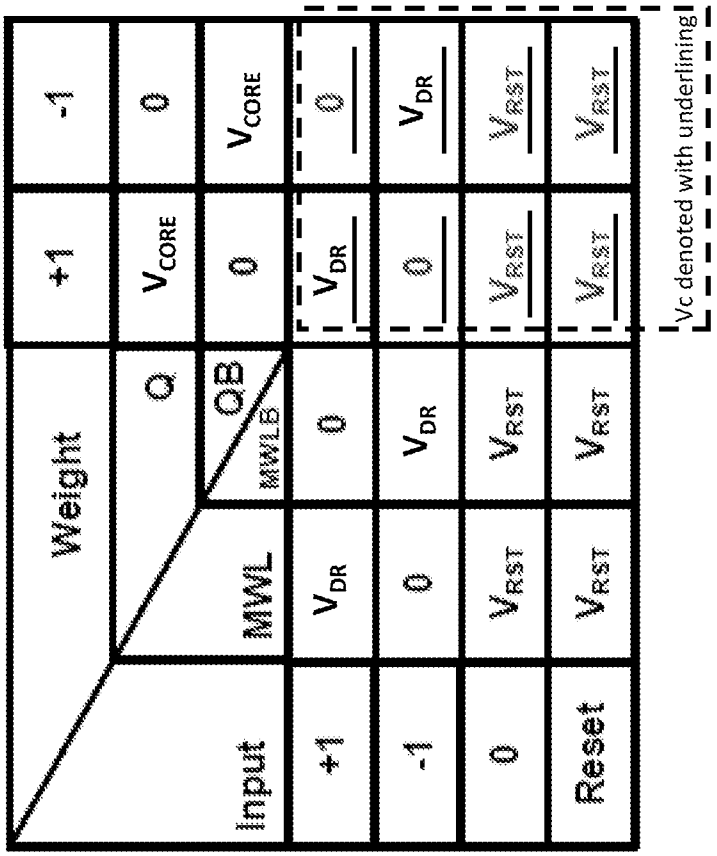

Weight
-1
+1
Q
0
V_{CORE}
QB
MWLB
0
V_{CORE}
Input
MWL
+1
V_{DR}
V_{DR}
0
-1
0
V_{DR}
0
V_{RST}
V_{RST}
0
Reset
V_{RST}
V_{RST}
Vc denoted with underlining

FIG. 2E

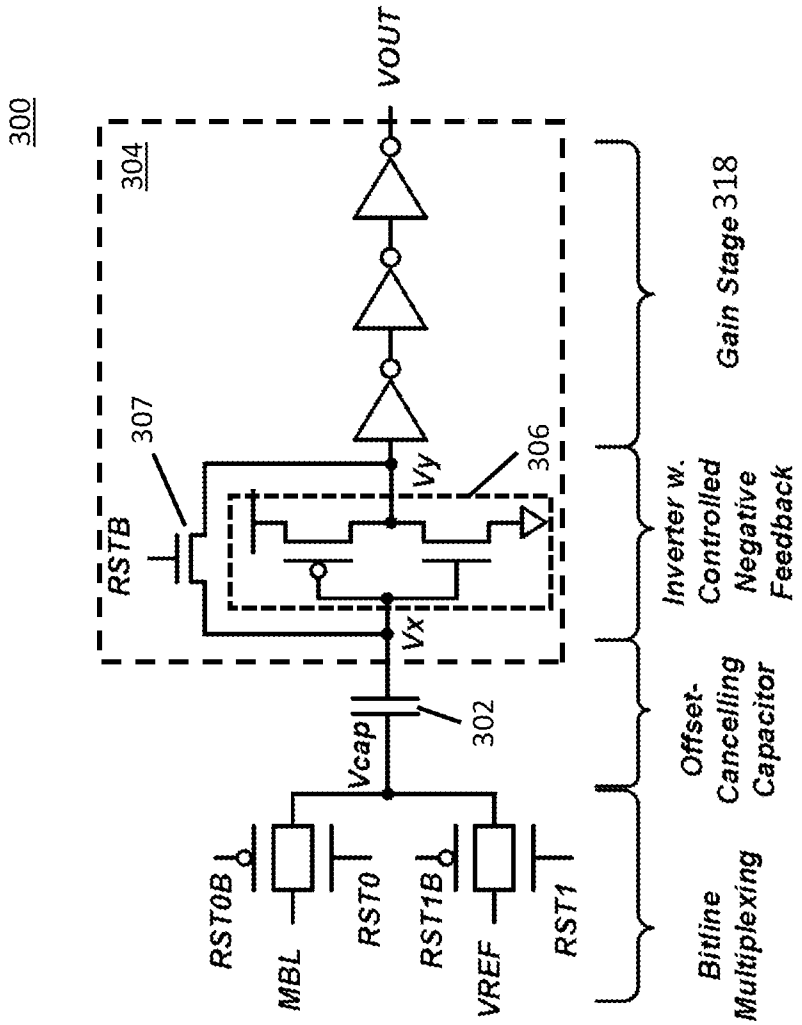

300
304
VOUT
307
RSTB
Vy
306
Vx
Vcap
302
RST0B
MBL
RST0
RST1B
VREF
RST1
Gain Stage 318
Inverter w. Controlled Negative Feedback
Offset-Cancelling Capacitor
Bitline Multiplexing

FIG. 3A

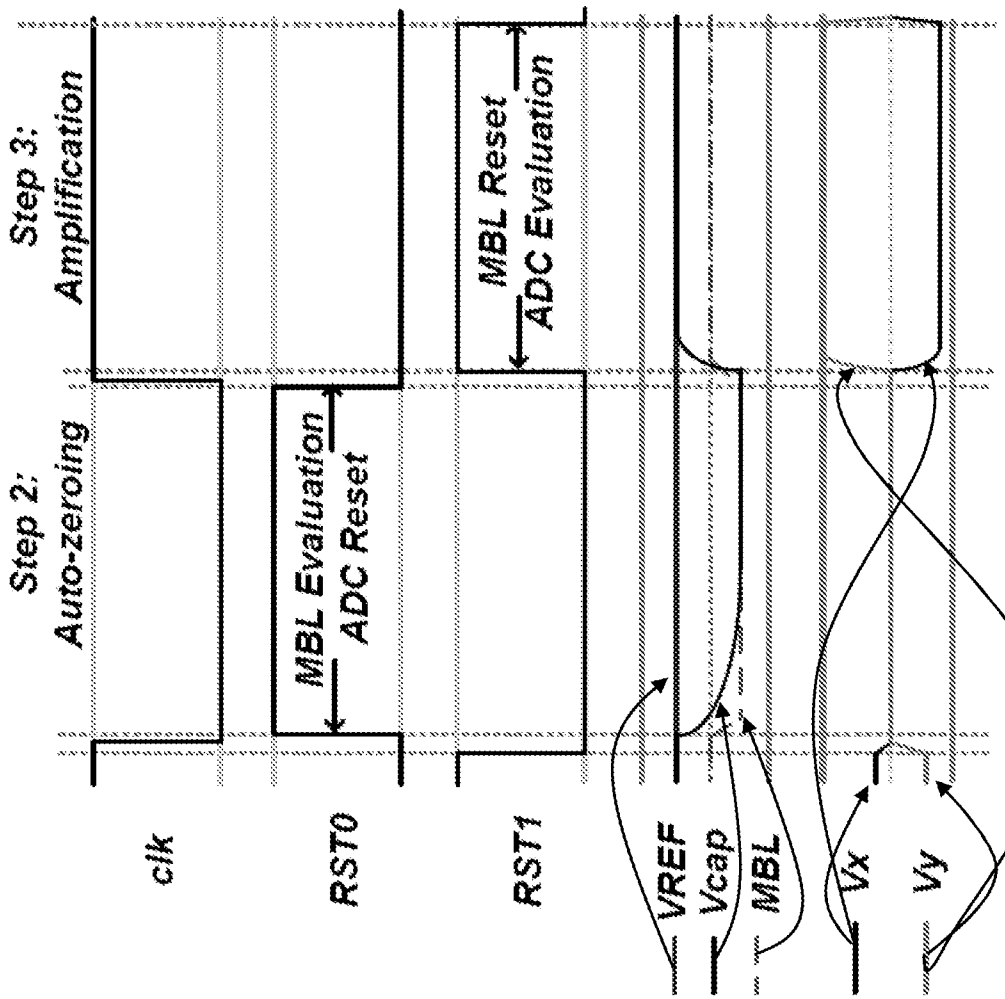

Step 2:
Auto-zeroing
Step 3:
Amplification
MBL Evaluation
ADC Reset
MBL Reset
ADC Evaluation
clk
RST0
RST1
VREF
Vcap
MBL
Vx
Vy

FIG. 3B

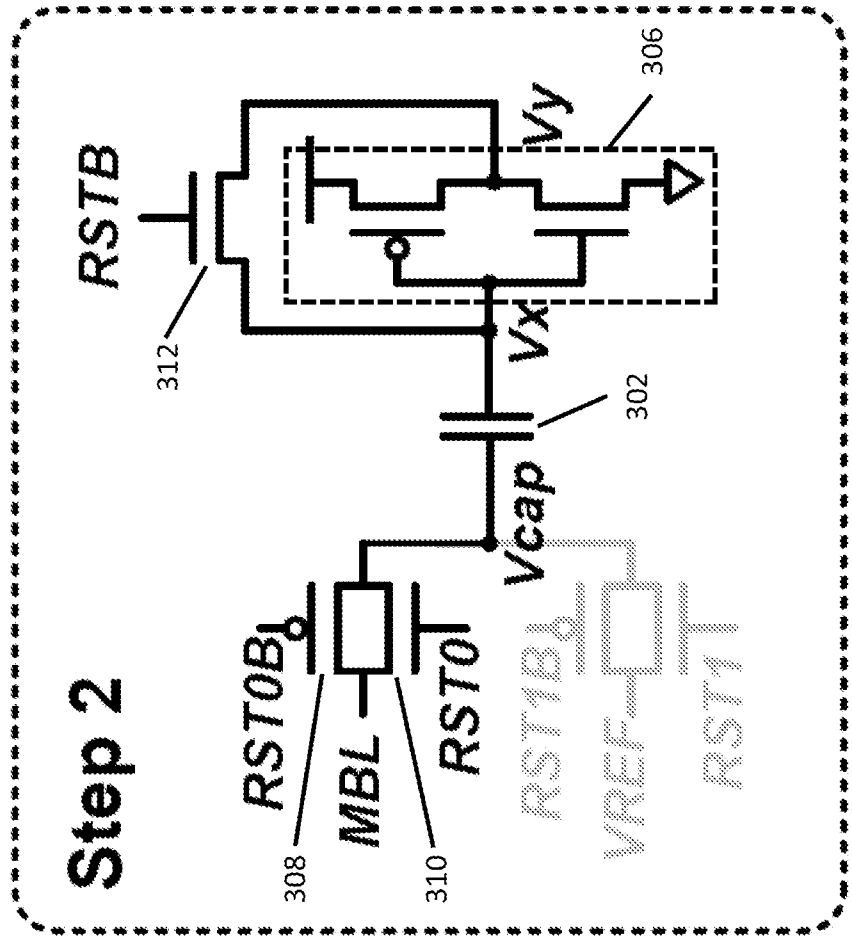

Step 2
RSTB
312
Vy
306
Vx
302
Vcap
RST0B
MBL
RST0
308
310
RST1B
VREF
RST1

FIG. 3C

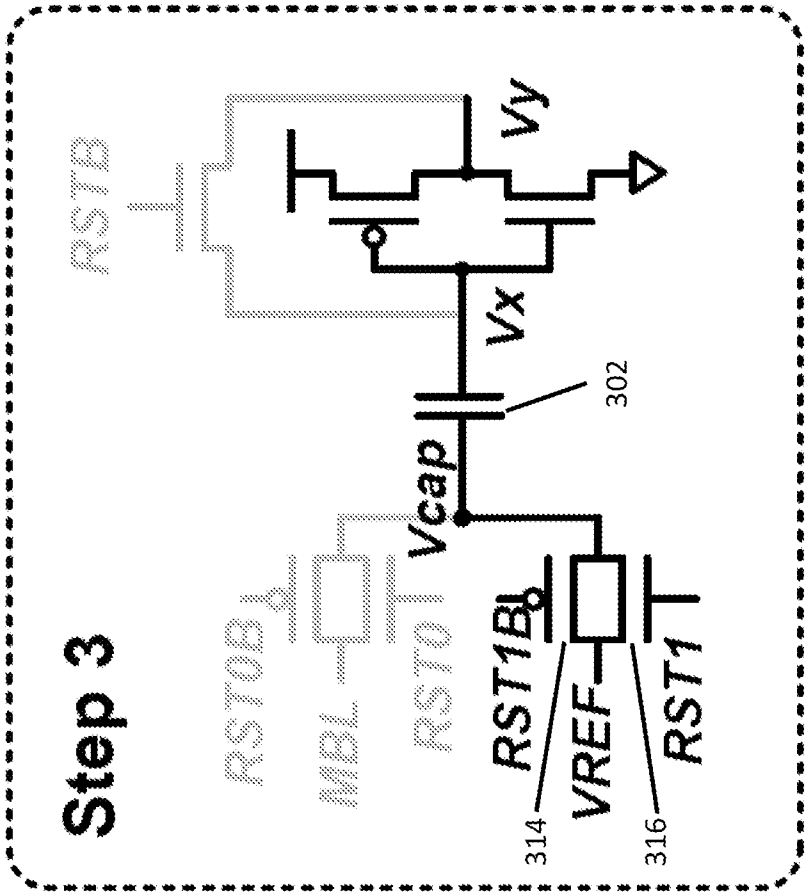

Step 3
RST0B
MBL
RST0T
Vcap
RST1B
VREF
RST1
314
316
302
Vx
Vy
RSTB

FIG. 3D

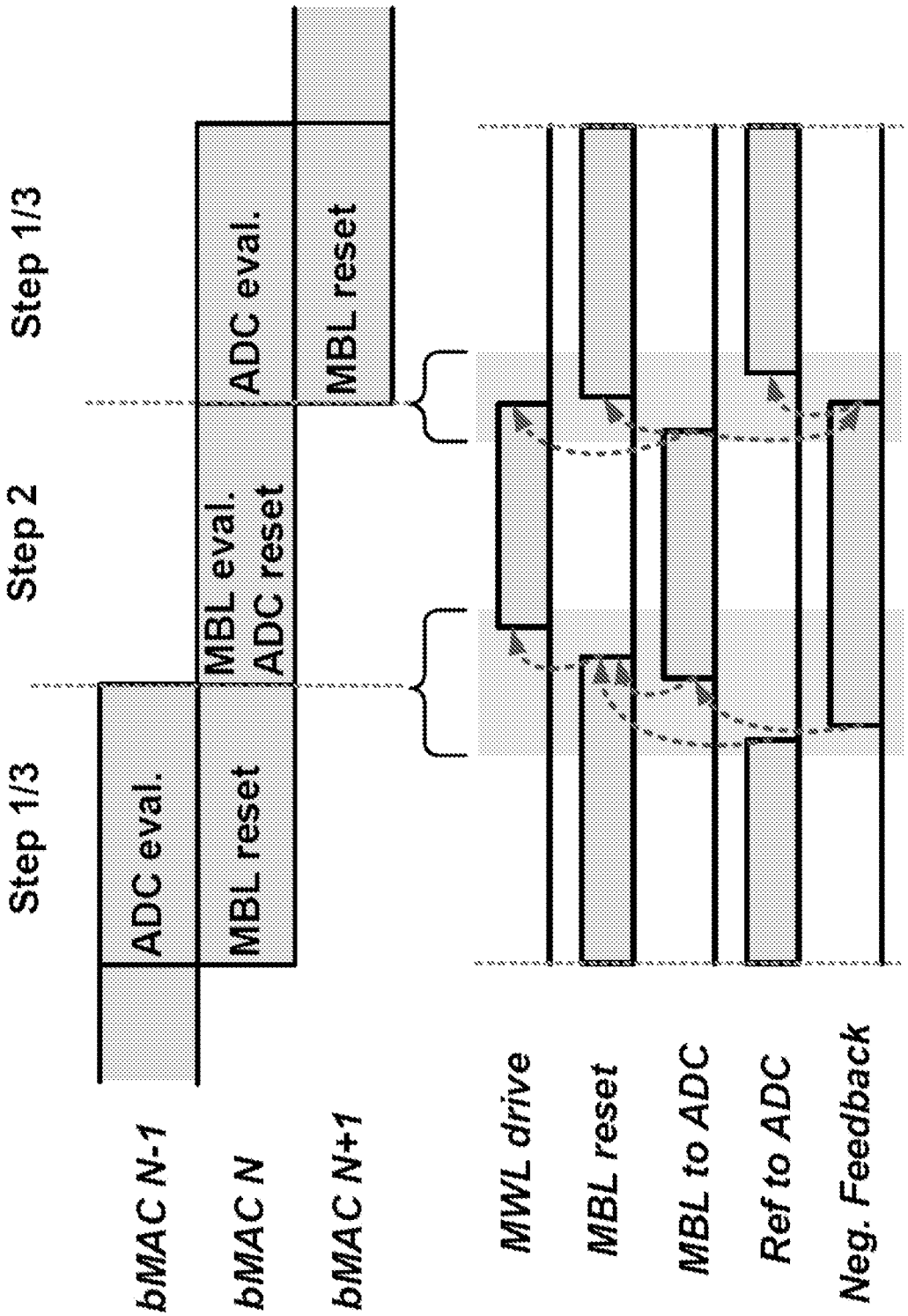

Step 1/3
Step 2
Step 1/3
ADC eval.
MBL reset
MBL eval.
ADC reset
ADC eval.
MBL reset
bMAC N-1
bMAC N
bMAC N+1
MWL drive
MBL reset
MBL to ADC
Ref to ADC
Neg. Feedback

FIG. 4

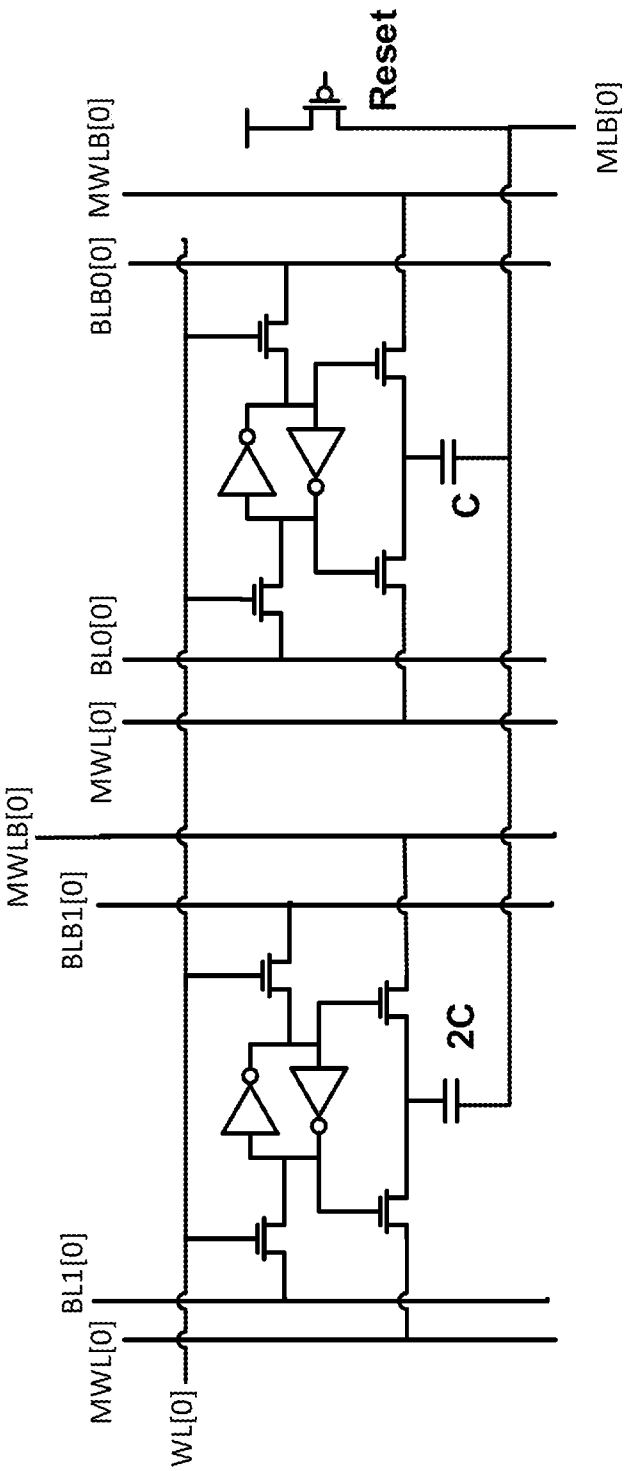

Reset
MLB[0]
MWLB[0]
BLB0[0]
C
BL0[0]
MWL[0]
MWLB[0]
BLB1[0]
2C
BL1[0]
MWL[0]
WL[0]

FIG. 5

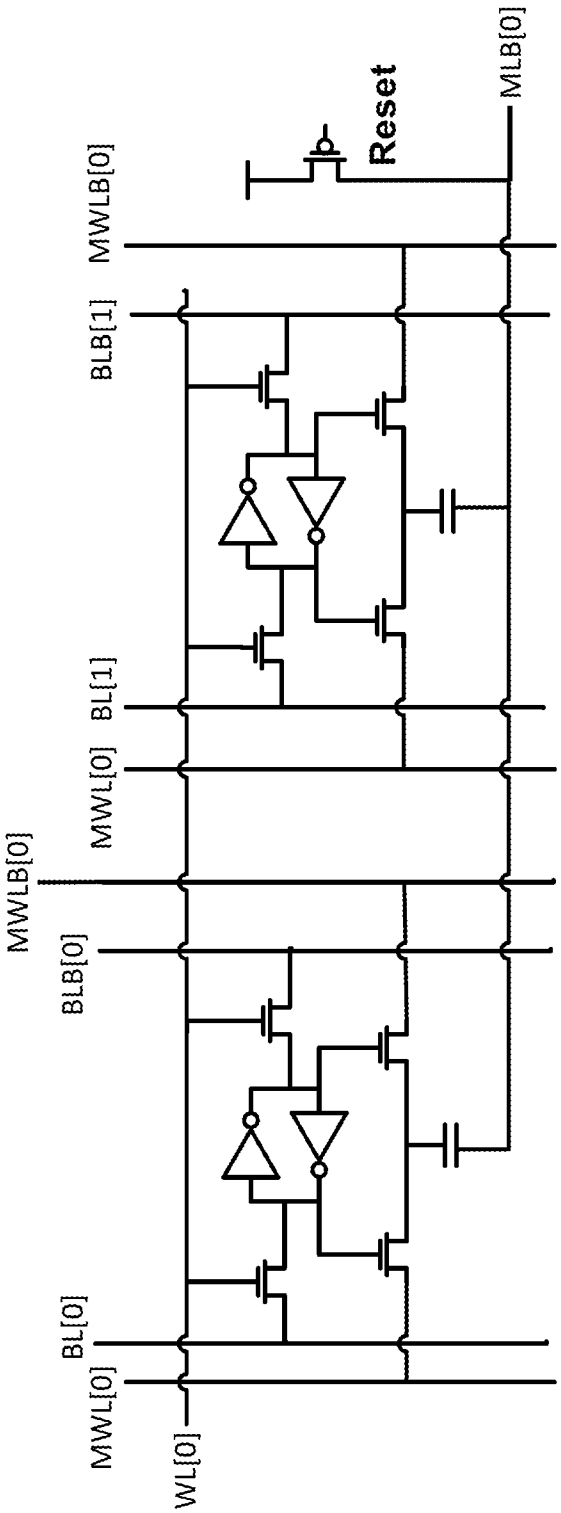


FIG. 6

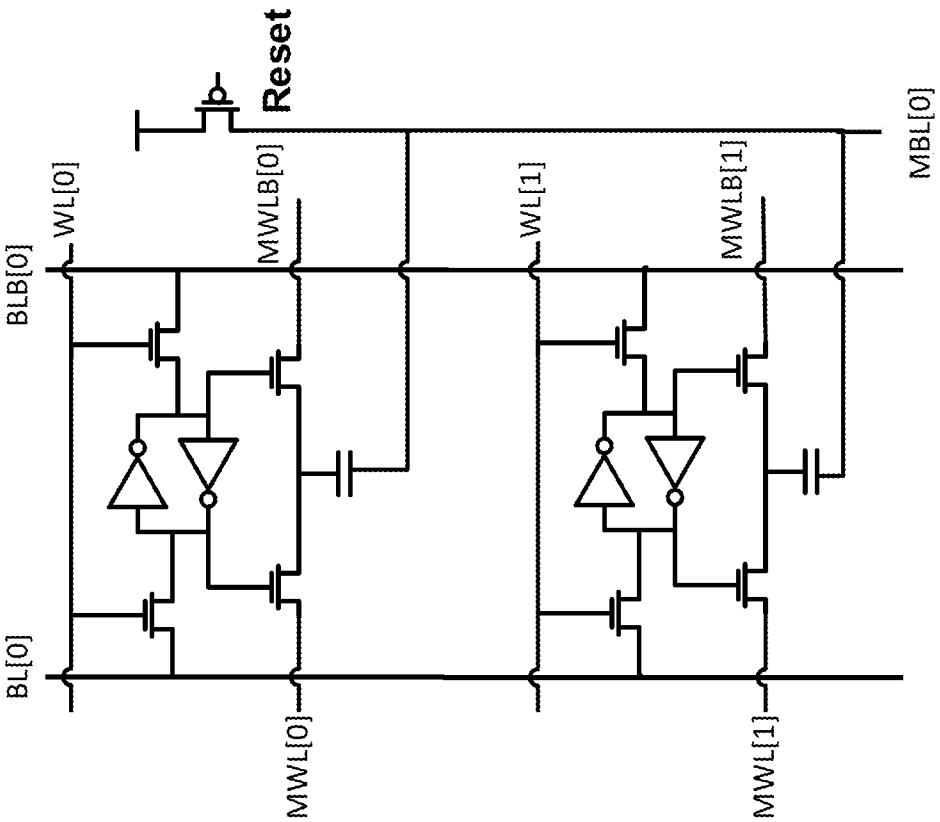


FIG. 7

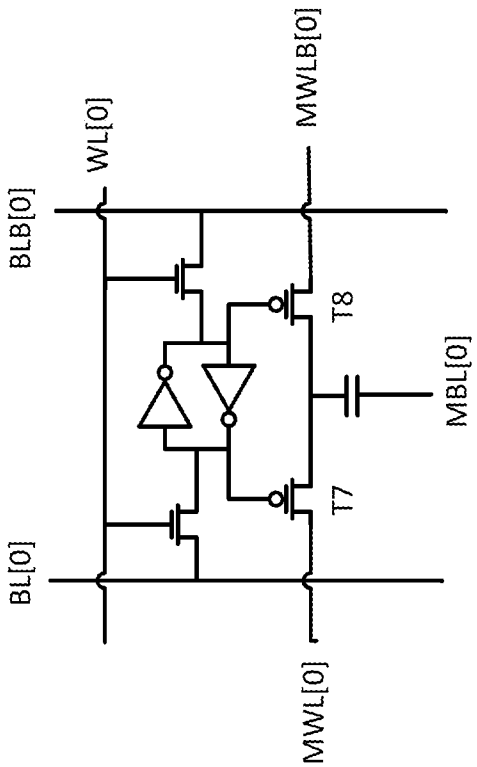


FIG. 8

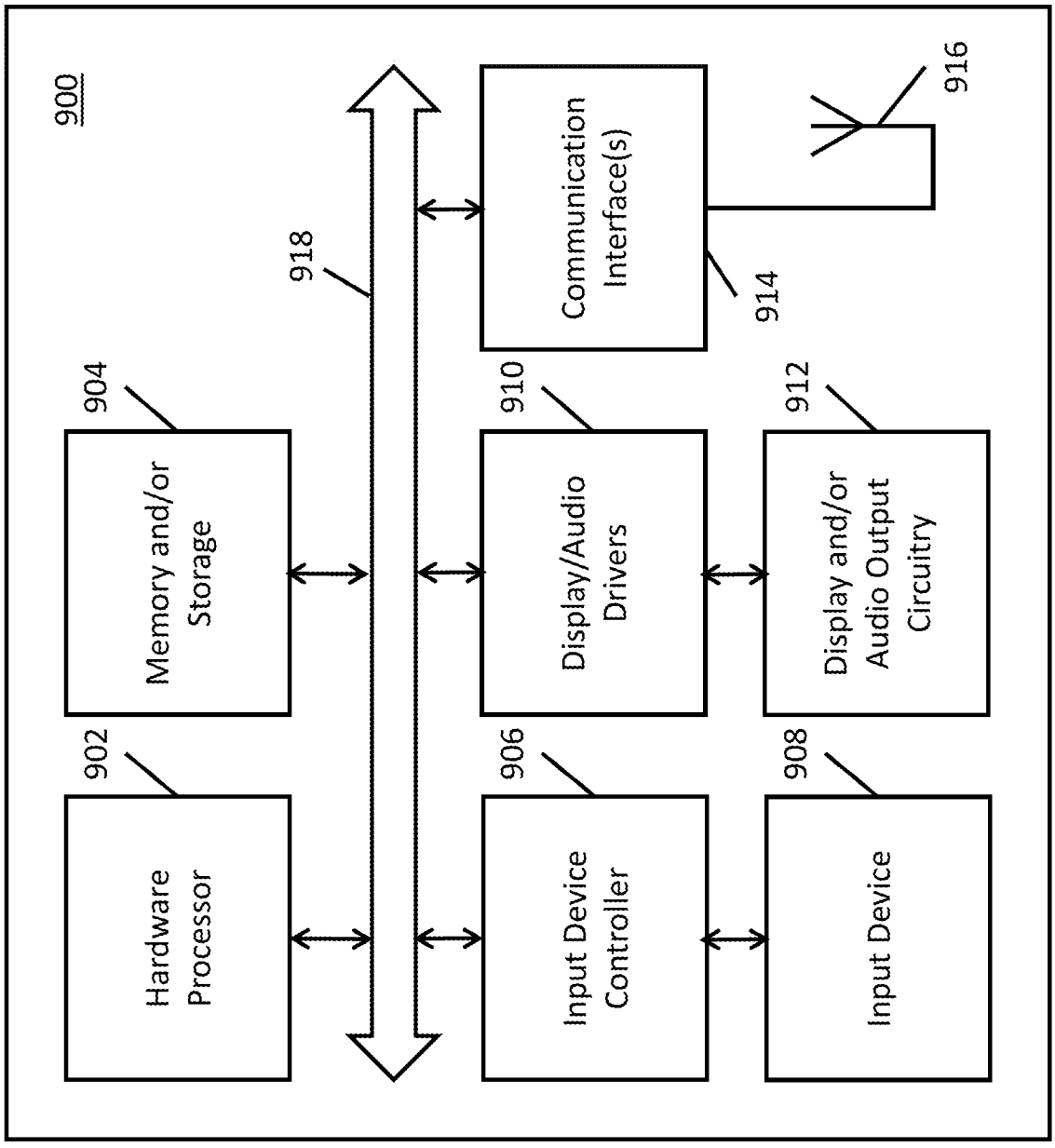

900
Hardware Processor
902
Memory and/or Storage
904
Input Device Controller
906
Display/Audio Drivers
910
Input Device
908
Display and/or Audio Output Circuitry
912
Communication Interface(s)
914
916
918

FIG. 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 19/68495

A. CLASSIFICATION OF SUBJECT MATTER
IPC - G11C 14/00 (2020.01)
CPC - G06N 3/063; G11C 11/412; G11C 14/0054

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.
P, X	JIANG et al., "C3SRAM:In-Memory Computing SRAM Macro Based on Capacitive Coupling Computing." IEEE SOLID-STATE CIRCUITS LETTERS [online], VOL. 2, NO. 9, September 2019 [online] <URL: https://ieeexplore.ieee.org/document/8877969>	1-17
A	US 6,128,207 A (LIEN et al.) 03 October 2000 (03.10.2000) Fig 3A, abstract, col 4, ln 23-39	1-17
A	US 8,787,102 B2 (ISHIZU) 22 July 2014 (22.07.2014) Fig 1, abstract, col 4, ln 47-col 6, ln 51	1-17
A	US 9,514,816 B1 (CYPRESS SEMICONDUCTOR CORPORATION) 06 December 2016 (06.12.2016) Fig 1A, abstract, col 3, ln 13-col 5, ln 67	1-17

☐ 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
27 February 2020	25 MAR 2020
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Lee Young Telephone No. PCT Helpdesk: 571-272-4300