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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 December 2020 (30.12.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/258197 A1

(51) International Patent Classification:
G11C 5/02 (2006.01)

(21) International Application Number:
PCT/CN2019/093477

(22) International Filing Date:
28 June 2019 (28.06.2019)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **YANGTZE MEMORY TECHNOLOGIES CO., LTD.** [CN/CN]; Room 7018, No.18, Huaguang Road, Guandong Science And Technology Industrial Park, East Lake Development Zone, Wuhan, Hubei 430074 (CN).

(72) Inventors: **YANG, Shengwei**; Room 7018, No.18, Huaguang Road, Guandong Science And Technology Industrial Park, East Lake High-Tech Development Zone, Wuhan, Hubei 430074 (CN). **XIA, Zhongyi**; Room 7018, No.18, Huaguang Road, Guandong Science And Technology Industrial Park, East Lake High-Tech Development Zone, Wuhan, Hubei 430074 (CN). **HAN, Kun**; Room 7018, No.18, Huaguang Road, Guandong Science And Technology Industrial Park, East Lake High-Tech Development Zone, Wuhan, Hubei 430074 (CN).

(74) Agent: **NTD UNIVATION INTELLECTUAL PROPERTY AGENCY LTD.**; 10th Floor, Tower C, Beijing Global Trade Center, 36 North Third Ring Road East, Dongcheng District, Beijing 100013 (CN).

(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: COMPUTATION-IN-MEMORY IN THREE-DIMENSIONAL MEMORY DEVICE

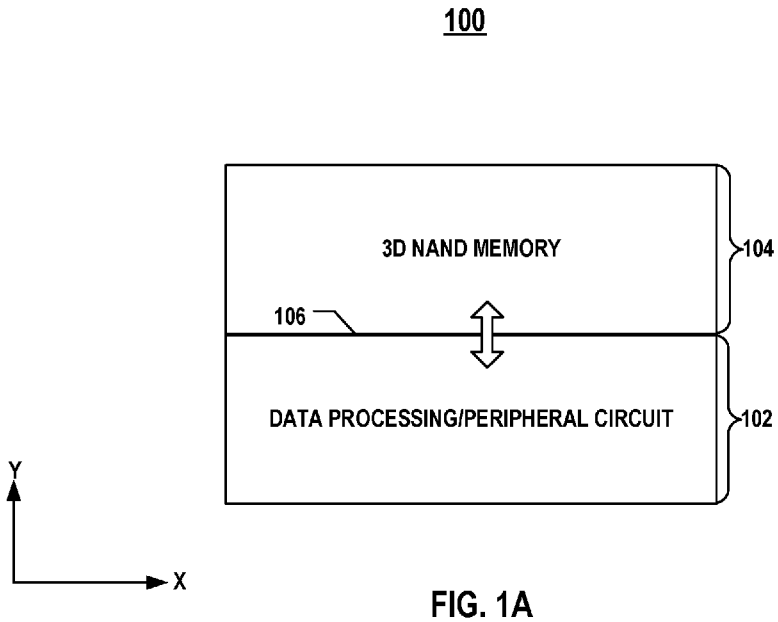

FIG. 1A

(57) Abstract: Three-dimensional (3D) memory devices are provided. An exemplary 3D memory device includes a 3D NAND memory array and an on-die data processing circuit coupled to the 3D NAND memory array on a same chip. The on-die data processing circuit is configured to receive, from an input/output (I/O) interface, control instructions for performing operations on data stored in the 3D NAND memory array. The on-die data processing circuit is also configured to retrieve the data from the 3D NAND memory array based on the control instructions and perform the operations on the retrieved data. Moreover, the on-die data processing circuit is configured to return a result of the operations to the I/O interface.

WO 2020/258197 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))*

COMPUTATION-IN-MEMORY IN THREE-DIMENSIONAL MEMORY DEVICE

BACKGROUND

[0001] Embodiments of the present disclosure relate to three-dimensional (3D) memory devices and fabrication methods thereof.

[0002] Planar memory cells are scaled to smaller sizes by improving process technology, circuit design, programming algorithm, and fabrication process. However, as feature sizes of the memory cells approach a lower limit, planar process and fabrication techniques become challenging and costly. As a result, memory density for planar memory cells approaches an upper limit.

[0003] A 3D memory architecture can address the density limitation in planar memory cells. The 3D memory architecture includes a memory array and peripheral devices for controlling signals to and from the memory array.

SUMMARY

[0004] Embodiments of 3D memory devices with on-die data processing capacity are disclosed herein.

[0005] In one example, a 3D memory device includes a 3D NAND memory array and an on-die data processing circuit coupled to the 3D NAND memory array on a same chip. The on-die data processing circuit is configured to receive, from an input/output (I/O) interface, control instructions for performing operations on data stored in the 3D NAND memory array. The on-die data processing circuit is also configured to retrieve the data from the 3D NAND memory array based on the control instructions and perform the operations on the retrieved data. Moreover, the on-die data processing circuit is configured to return a result of the operations to the I/O interface.

[0006] In another example, a system for operating a data processing unit on a 3D memory device includes a host and a storage unit having a plurality of NAND memory cells and being coupled to the host. The data processing unit is coupled to and on a same chip with the storage unit. The data processing unit is configured to receive, from the host, control instructions for performing operations on data stored in the NAND memory cells. The data processing unit is also configured to retrieve the data from the NAND memory cells based on

the control instructions and perform the operations on the retrieved data. Moreover, the data processing unit is configured to return a result of the operations to the host.

[0007] In still another example, a method for operating a logic circuit on a 3D memory device includes receiving, from an I/O interface, control instructions for performing operations on data stored in the 3D memory device. The method also includes retrieving, by the logic circuit on a same chip with the 3D memory device, the data from the 3D memory device based on the control instructions. The method further includes perform, by the logic circuit, the operations on the retrieved data and return, by the logic circuit, a result of the operations to the I/O interface.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] The accompanying drawings, which are incorporated herein and form a part of the specification, illustrate embodiments of the present disclosure and, together with the description, further serve to explain the principles of the present disclosure and to enable a person skilled in the pertinent art to make and use the present disclosure.

[0009] FIG. 1A illustrates a schematic view of a cross-section of an exemplary 3D memory device with a data processing circuit, according to some embodiments.

[0010] FIG. 1B illustrates a schematic view of a cross-section of another exemplary 3D memory device with a data processing circuit, according to some embodiments.

[0011] FIG. 2 illustrates a schematic plan view of an exemplary semiconductor structure having a peripheral circuit and a data processing circuit, according to some embodiments.

[0012] FIG. 3 illustrates a cross-section of an exemplary 3D memory device with a data processing circuit, according to some embodiments.

[0013] FIG. 4 illustrates a cross-section of another exemplary 3D memory device with a data processing circuit, according to some embodiments.

[0014] FIGs. 5A–5C illustrate a fabrication process for forming an exemplary semiconductor structure having a peripheral circuit and a data processing circuit, according to some embodiments.

[0015] FIGs. 6A and 6B illustrate a fabrication process for forming an exemplary semiconductor structure having 3D NAND memory strings, according to some embodiments.

[0016] FIGs. 7A and 7B illustrate a fabrication process for forming an exemplary 3D memory device with a data processing circuit, according to some embodiments.

[0017] FIG. 8 illustrates a schematic block diagram of an exemplary system having a 3D memory device with a data processing circuit, according to some embodiments.

[0018] FIG. 9 illustrates a schematic block diagram of an exemplary system having a 3D memory device with a data processing circuit, according to some embodiments.

[0019] FIG. 10 is a flowchart of an exemplary method for operating a 3D memory device with a data processing circuit, according to some embodiments.

[0020] FIG. 11 is a flowchart of an exemplary method for forming a 3D memory device with a data processing circuit, according to some embodiments.

[0021] Embodiments of the present disclosure will be described with reference to the accompanying drawings.

DETAILED DESCRIPTION

[0022] Although specific configurations and arrangements are discussed, it should be understood that this is done for illustrative purposes only. A person skilled in the pertinent art will recognize that other configurations and arrangements can be used without departing from the spirit and scope of the present disclosure. It will be apparent to a person skilled in the pertinent art that the present disclosure can also be employed in a variety of other applications.

[0023] It is noted that references in the specification to "one embodiment," "an embodiment," "an example embodiment," "some embodiments," etc., indicate that the embodiment described may include a particular feature, structure, or characteristic, but every embodiment may not necessarily include the particular feature, structure, or characteristic. Moreover, such phrases do not necessarily refer to the same embodiment. Further, when a particular feature, structure or characteristic is described in connection with an embodiment, it would be within the knowledge of a person skilled in the pertinent art to effect such feature, structure or characteristic in connection with other embodiments whether or not explicitly described.

[0024] In general, terminology may be understood at least in part from usage in context. For example, the term "one or more" as used herein, depending at least in part upon context, may be used to describe any feature, structure, or characteristic in a singular sense or may be used to describe combinations of features, structures or characteristics in a plural sense. Similarly, terms, such as "a," "an," or "the," again, may be understood to convey a singular usage or to convey a plural usage, depending at least in part upon context. In addition, the term "based

on" may be understood as not necessarily intended to convey an exclusive set of factors and may, instead, allow for existence of additional factors not necessarily expressly described, again, depending at least in part on context.

[0025] It should be readily understood that the meaning of "on," "above," and "over" in the present disclosure should be interpreted in the broadest manner such that "on" not only means "directly on" something but also includes the meaning of "on" something with an intermediate feature or a layer therebetween, and that "above" or "over" not only means the meaning of "above" or "over" something but can also include the meaning it is "above" or "over" something with no intermediate feature or layer therebetween (i.e., directly on something).

[0026] Further, spatially relative terms, such as "beneath," "below," "lower," "above," "upper," and the like, may be used herein for ease of description to describe one element or feature's relationship to another element(s) or feature(s) as illustrated in the figures. The spatially relative terms are intended to encompass different orientations of the device in use or operation in addition to the orientation depicted in the figures. The apparatus may be otherwise oriented (rotated 90 degrees or at other orientations) and the spatially relative descriptors used herein may likewise be interpreted accordingly.

[0027] As used herein, the term "substrate" refers to a material onto which subsequent material layers are added. The substrate itself can be patterned. Materials added on top of the substrate can be patterned or can remain unpatterned. Furthermore, the substrate can include a wide array of semiconductor materials, such as silicon, germanium, gallium arsenide, indium phosphide, etc. Alternatively, the substrate can be made from an electrically non-conductive material, such as a glass, a plastic, or a sapphire wafer.

[0028] As used herein, the term "layer" refers to a material portion including a region with a thickness. A layer can extend over the entirety of an underlying or overlying structure or may have an extent less than the extent of an underlying or overlying structure. Further, a layer can be a region of a homogeneous or inhomogeneous continuous structure that has a thickness less than the thickness of the continuous structure. For example, a layer can be located between any pair of horizontal planes between, or at, a top surface and a bottom surface of the continuous structure. A layer can extend horizontally, vertically, and/or along a tapered surface. A substrate can be a layer, can include one or more layers therein, and/or can have one or more layer thereupon, thereabove, and/or therebelow. A layer can include

multiple layers. For example, an interconnect layer can include one or more conductor and contact layers (in which interconnect lines and/or via contacts are formed) and one or more dielectric layers.

[0029] As used herein, the term "nominal/nominally" refers to a desired, or target, value of a characteristic or parameter for a component or a process operation, set during the design phase of a product or a process, together with a range of values above and/or below the desired value. The range of values can be due to slight variations in manufacturing processes or tolerances. As used herein, the term "about" indicates the value of a given quantity that can vary based on a particular technology node associated with the subject semiconductor device. Based on the particular technology node, the term "about" can indicate a value of a given quantity that varies within, for example, 10–30% of the value (e.g., ±10%, ±20%, or ±30% of the value).

[0030] As used herein, the term "3D memory device" refers to a semiconductor device with vertically oriented strings of memory cell transistors (referred to herein as "memory strings," such as NAND memory strings) on a laterally-oriented substrate so that the memory strings extend in the vertical direction with respect to the substrate. As used herein, the term "vertical/vertically" means nominally perpendicular to the lateral surface of a substrate.

[0031] In conventional 3D memory devices, the peripheral circuits that are formed outside of memory array devices on the same plane can occupy a large area of the device chip, thereby leading to poor array efficiency, large die size, and high cost. Also, the thermal budget associated with processing the memory array device limits the peripheral circuit performance requirement, making it difficult to achieve high input/output (I/O) speed of the 3D memory device. Moreover, conventional 3D memory devices rely on external devices (e.g., not on the same chip as the memory device) to perform data processing operations. As a result, data stored on the 3D memory devices need to be loaded into an external data processing device through an I/O interface (e.g., a data bus) before being processed by the external data processing device, placing a heavy burden on the I/O interface especially when the amount of data is large. Therefore, the I/O interface often becomes the bottleneck limiting the overall throughput of the host system.

[0032] Due to considerations such as volume and cost, data processing circuits are often not formed in memory packages such as an embedded multi-media card (eMMC) o universal flash storage (UFS). On-chip or in-memory data processing operations are often not enabled

in such memory packages. As a result, high-speed data processing in these memory packages can be limited. In another example, a memory device can use a considerable amount of resources (e.g., data buffers and data buses) to transmit data to an external data processing device through the I/O interface. This can slow down other operations. The overall performance of the memory device can be limited.

[0033] Various embodiments in accordance with the present disclosure provide a 3D memory device having an on-chip, in-memory data processing circuit (e.g., a logic circuit capable of performing data processing operations) with improved I/O speed, throughput, and memory density. The on-die data processing circuit can be formed on the same chip with peripheral circuits of the 3D memory device. The on-die data processing circuit can be placed in the area that is not occupied by the peripheral circuits (e.g., the spare space neighboring peripheral circuits) and thus, do not need extra space to be formed. The on-die data processing circuit can enable high-speed data processing operations on the data stored in memory cells of the 3D memory device. In an embodiment, the on-die data processing circuit is configured to perform logic operations such as searching, encryption, and preprocessing on data stored in the local memory array, and return an operation result to the I/O interface instead of loading the entirety of the underlying data to the I/O interface. The on-die data processing circuit can thus improve the I/O throughput.

[0034] FIG. 1A illustrates a schematic view of a cross-section of an exemplary 3D memory device 100 with a data processing circuit, according to some embodiments. As used herein, a data processing circuit refers to a logic circuit configured to perform data processing operations such as searching, encryption, and/or preprocessing on the underlying data. 3D memory device 100 represents an example of a non-monolithic 3D memory device. The term "non-monolithic" means that the components of 3D memory device 100 (e.g., peripheral circuit/data processing circuit and 3D NAND memory) can be formed separately on different substrates and then joined to form a 3D memory device. 3D memory device 100 can include a first semiconductor structure 102 including peripheral circuits and one or more data processing circuits. Both peripheral circuits and the data processing circuit(s) can be implemented with advanced logic processes (e.g., technology nodes of 90 nm, 80 nm, 65 nm, 55 nm, 45 nm, 40 nm, 32 nm, 28 nm, 22 nm, 20 nm, 16 nm, 14 nm, 10 nm, 7 nm, 5 nm, 3 nm, etc.) to achieve high speed. In some embodiments, the peripheral circuits and data processing

circuit(s) in first semiconductor structure 102 use complementary metal-oxide-semiconductor (CMOS) technology.

[0035] In some embodiments, the peripheral circuits include any suitable digital, analog, and/or mixed-signal peripheral circuits used for facilitating the operation of 3D memory device 100. For example, the peripheral circuits can include one or more of a page buffer, a decoder (e.g., a row decoder and a column decoder), a sense amplifier, a driver, a charge pump, a current or voltage reference, or any active or passive components of the circuits (e.g., transistors, diodes, resistors, or capacitors). A data processing circuit is integrated on the same die of the peripheral circuits, enabling in-memory computation and data processing. The data processing circuit can be embedded along with the peripheral circuits. While peripheral circuits provide functionalities more closely related to the storage-based operations of memory device 100, the data processing circuit provides computational power to implement algorithmic-based operations to the data stored in 3D memory device 100.

[0036] 3D memory device 100 can also include a second semiconductor structure 104 including an array of 3D NAND memory strings. That is, second semiconductor structure 104 can be a NAND Flash memory device in which memory cells are provided in the form of an array of NAND memory strings. In some embodiments, depending on the NAND technology (e.g., the number of levels/tiers in the memory stack), a 3D NAND memory string typically consists of 32 to 256 NAND memory cells. 3D NAND memory strings can be organized into pages which are then organized into blocks in which each 3D NAND memory string is connected to a separate line called a bit line (BL). All cells with the same position in the 3D NAND memory string can be connected through the control gates by a word line (WL). In some embodiments, a plane contains a certain number of blocks that are connected through the same bit line. Second semiconductor structure 104 can include one or more planes, and the peripheral circuits that are needed to perform all the read/write/erase operations can be included in first semiconductor structure 102.

[0037] As shown in FIG. 1A, 3D memory device 100 further includes a bonding interface 106 vertically between first semiconductor structure 102 and second semiconductor structure 104. As described below in detail, first and second semiconductor structures 102 and 104 can be fabricated separately (and in parallel in some embodiments) such that the thermal budget of fabricating one of first and second semiconductor structures 102 and 104 does not limit the processes of fabricating another one of first and second semiconductor structures 102 and 104.

Moreover, a large number of interconnects (e.g., bonding contacts via hybrid bonding) can be formed through bonding interface 106 to make direct, short electrical connections between first semiconductor structure 102 and second semiconductor structure 104, as opposed to the long-distance chip-to-chip data bus on the circuit board, thereby eliminating chip interface delay and achieving high-speed I/O throughput with reduced power consumption. Data transfer between the array of 3D NAND memory strings in second semiconductor structure 104 and the data processing circuit in first semiconductor structure 102 can be performed through the interconnects (e.g., bonding contacts via hybrid bonding) across bonding interface 106. Furthermore, by vertically integrating first and second semiconductor structures 102 and 104, the chip size can be reduced, and the memory cell density can be increased.

[0038] It is understood that the relative positions of stacked first and second semiconductor structures 102 and 104 are not limited. FIG. 1B illustrates a schematic view of a cross-section of another exemplary 3D memory device 101 with a data processing circuit, according to some embodiments. Different from 3D memory device 100 in FIG. 1A in which the second semiconductor structure 104 including the array of 3D NAND memory strings is above the first semiconductor structure 102 including the peripheral circuits and the data processing circuit, in 3D memory device 100 in FIG. 1B, the first semiconductor structure 102 including the peripheral circuits and the data processing circuit is above the second semiconductor structure 104 including the array of 3D NAND memory strings. Nevertheless, bonding interface 106 is formed vertically between first and second semiconductor structures 102 and 104 in 3D memory device 101. The first and second semiconductor structures 102 and 104 are joined vertically through bonding (e.g., hybrid bonding) according to some embodiments. Data transfer between the array of 3D NAND memory strings in second semiconductor structure 104 and the data processing circuit in first semiconductor structure 102 can be performed through the interconnects (e.g., bonding contacts via hybrid bonding) across bonding interface 106.

[0039] FIG. 2 illustrates a schematic plan view of an exemplary semiconductor structure 200 having a peripheral circuit and a data processing circuit, according to some embodiments. Semiconductor structure 200 may be one example of first semiconductor structure 102. Semiconductor structure 200 can include peripheral circuits for controlling and sensing a 3D NAND memory, including word line drivers 202, page buffers 204, and any other suitable

circuits. Semiconductor structure 200 can further include data processing circuit 206 on the same die as the peripheral circuits and fabricated using the same logic process as the peripheral circuits. FIG. 2 shows an exemplary layout of the peripheral circuits (e.g., word line drivers 202, page buffers 204) and data processing circuit 206 in which peripheral circuits (e.g., word line drivers 202, page buffers 204) and data processing circuit 206 are formed in different regions on the same plane. For example, data processing circuit 206 may be formed outside of the peripheral circuits (e.g., word line drivers 202, page buffers 204). It is understood that the layout of semiconductor structure 200 is not limited to the exemplary layout in FIG. 2. In some embodiments, the peripheral circuit (e.g., word line drivers 202 and page buffers 204) and data processing circuit 206 are formed in non-overlapping regions of the same plane. In some embodiments, on a plane, data processing circuit 206 is formed in space that is not used for the formation of the peripheral circuit. In some embodiments, the peripheral circuit (e.g., word line drivers 202 and page buffers 204) and data processing circuit 206 are stacked one over another, i.e., on different planes. For example, data processing circuit 206 (e.g., an array of logic gates) may be formed above or below the peripheral circuit (e.g., word line drivers 202, page buffers 204) to further reduce the chip size.

[0040] FIG. 3 illustrates a cross-section of an exemplary 3D memory device 300 with a data processing circuit, according to some embodiments. As one example of 3D memory device 100 described above with respect to FIG. 1A, 3D memory device 300 is a non-monolithic 3D memory device including a first semiconductor structure 302 and a second semiconductor structure 304 stacked over first semiconductor structure 302. First and second semiconductor structures 302 and 304 are joined at a bonding interface 306 therebetween. As shown in FIG. 3, first semiconductor structure 302 can include a substrate 308, which can include silicon (e.g., single crystalline silicon), silicon germanium (SiGe), gallium arsenide (GaAs), germanium (Ge), silicon on insulator (SOI), or any other suitable materials.

[0041] First semiconductor structure 302 of 3D memory device 300 can include a device layer 310 above substrate 308. It is noted that *x* and *y* axes are added in FIG. 3 to further illustrate the spatial relationship of the components in 3D memory device 300. Substrate 308 includes two lateral surfaces (e.g., a top surface and a bottom surface) extending laterally in the *x*-direction (the lateral direction or width direction). As used herein, whether one component (e.g., a layer or a device) is "on," "above," or "below" another component (e.g., a layer or a

device) of a semiconductor device (e.g., 3D memory device 300) is determined relative to the substrate of the semiconductor device (e.g., substrate 308) in the *y*-direction (the vertical direction or thickness direction) when the substrate is positioned in the lowest plane of the semiconductor device in the *y*-direction. The same notion for describing spatial relationship is applied throughout the present disclosure.

[0042] In some embodiments, device layer 310 includes a peripheral circuit 312 on substrate 308. Device layer 310 also includes a data processing circuit 314 on substrate 308 and outside of peripheral circuit 312. In some embodiments, peripheral circuit 312 includes a plurality of peripheral transistors 316 forming any suitable digital, analog, and/or mixed-signal peripheral circuits used for facilitating the operation of 3D memory device 300 including, but not limited to, a page buffer, a decoder (e.g., a row decoder and a column decoder), a sense amplifier, a driver, a charge pump, a current or voltage reference. Peripheral transistors 316 can be formed "on" substrate 308, in which the entirety or part of peripheral transistors 316 are formed in substrate 308 (e.g., below the top surface of substrate 308) and/or directly on substrate 308. Isolation regions (e.g., shallow trench isolations (STIs)) and doped regions (e.g., source regions and drain regions of peripheral transistors 316) can be formed in substrate 308 as well.

[0043] In some embodiments, data processing circuit 314 includes a plurality of transistors 318 (e.g., MOSFETs). In some embodiments, transistors 318 may form logic gates of a field-programmable gate array (FPGA). In some embodiments, transistors 318 may be part of an application-specific integrated circuit (ASIC), a microcontroller unit (MCU), or other data processing circuits capable of performing computation, data operations, and/or logic operations. In some embodiments, transistors 318 are formed "on" substrate 308, in which the entirety or part of transistors 318 are formed in substrate 308 (e.g., below the top surface of substrate 308) and/or directly on substrate 308. Isolation regions (e.g., shallow trench isolations (STIs)) and doped regions (e.g., source regions and drain regions of transistors 318) can be formed in substrate 308 as well. As shown in FIG. 3, transistors 318 and peripheral transistors 316 can be formed in different regions on the same plane, e.g., on substrate 308. That is, transistors 318 can be formed outside of the region in which peripheral circuit 312 is formed on substrate 308. For ease of illustration, FIG. 3 only depicts a limited number of transistors 318. An electrode contact 320 may connect one or more transistors 318 (e.g., via electrodes of MOSFETs) to a bonding contact 326 to facilitate data communication between

data processing circuit 314 and memory arrays in second semiconductor structure 304 through bonding interface 306. It is understood that the configuration in FIG. 3, e.g., the layout of transistors 318 and the connection between transistors 318 and bonding contact 326, do not reflect the actual layout and electrical connection between the transistors and other components.

[0044] In some embodiments, first semiconductor structure 302 of 3D memory device 300 further includes an interconnect layer 322 above device layer 310 to transfer electrical signals to and from peripheral circuit 312 and data processing circuit 314. Interconnect layer 322 can include a plurality of interconnects (also referred to herein as "contacts"), including lateral interconnect lines and vertical interconnect access (via) contacts. As used herein, the term "interconnects" can broadly include any suitable types of interconnects, such as middle-end-of-line (MEOL) interconnects and back-end-of-line (BEOL) interconnects. Interconnect layer 322 can further include one or more interlayer dielectric (ILD) layers (also known as "intermetal dielectric (IMD) layers") in which the interconnect lines and via contacts can form. That is, interconnect layer 322 can include interconnect lines and via contacts in multiple ILD layers. The interconnect lines and via contacts in interconnect layer 322 can include conductive materials including, but not limited to, tungsten (W), cobalt (Co), copper (Cu), aluminum (Al), silicides, or any combination thereof. The ILD layers in interconnect layer 322 can include dielectric materials including, but not limited to, silicon oxide, silicon nitride, silicon oxynitride, low dielectric constant (low-k) dielectrics, or any combination thereof.

[0045] As shown in FIG. 3, first semiconductor structure 302 of 3D memory device 300 can further include a bonding layer 324 at bonding interface 306 and above interconnect layer 322 and device layer 310 (including peripheral circuit 312 and data processing circuit 314). Bonding layer 324 can include a plurality of bonding contacts 326 and dielectrics electrically isolating bonding contacts 326. Bonding contacts 326 can include conductive materials including, but not limited to, W, Co, Cu, Al, silicides, or any combination thereof. The remaining area of bonding layer 324 can be formed with dielectrics including, but not limited to, silicon oxide, silicon nitride, silicon oxynitride, low-k dielectrics, or any combination thereof. Bonding contacts 326 and surrounding dielectrics in bonding layer 324 can be used for hybrid bonding.

[0046] Similarly, as shown in FIG. 3, second semiconductor structure 304 of 3D memory device 300 can also include a bonding layer 328 at bonding interface 306 and above bonding layer 324 of first semiconductor structure 302. Bonding layer 328 can include a plurality of bonding contacts 330 and dielectrics electrically isolating bonding contacts 330. Bonding contacts 330 can include conductive materials including, but not limited to, W, Co, Cu, Al, silicides, or any combination thereof. The remaining area of bonding layer 328 can be formed with dielectrics including, but not limited to, silicon oxide, silicon nitride, silicon oxynitride, low-k dielectrics, or any combination thereof. Bonding contacts 330 and surrounding dielectrics in bonding layer 328 can be used for hybrid bonding.

[0047] As described above, second semiconductor structure 304 can be bonded on top of first semiconductor structure 302 in a face-to-face manner at bonding interface 306. In some embodiments, bonding interface 306 is disposed between bonding layers 324 and 328 as a result of hybrid bonding (also known as "metal/dielectric hybrid bonding"), which is a direct bonding technology (e.g., forming bonding between surfaces without using intermediate layers, such as solder or adhesives) and can obtain metal-metal bonding and dielectric-dielectric bonding simultaneously. In some embodiments, bonding interface 306 is the place at which bonding layers 324 and 328 are met and bonded. In practice, bonding interface 306 can be a layer with a certain thickness that includes the top surface of bonding layer 324 of first semiconductor structure 302 and the bottom surface of bonding layer 328 of second semiconductor structure 304.

[0048] In some embodiments, second semiconductor structure 304 of 3D memory device 300 further includes an interconnect layer 332 above bonding layer 328 to transfer electrical signals. Interconnect layer 332 can include a plurality of interconnects, such as MEOL interconnects and BEOL interconnects. Interconnect layer 332 can further include one or more ILD layers in which the interconnect lines and via contacts can form. The interconnect lines and via contacts in interconnect layer 332 can include conductive materials including, but not limited to, W, Co, Cu, Al, silicides, or any combination thereof. The ILD layers in interconnect layer 332 can include dielectric materials including, but not limited to, silicon oxide, silicon nitride, silicon oxynitride, low-k dielectrics, or any combination thereof.

[0049] In some embodiments, second semiconductor structure 304 of 3D memory device 300 includes a NAND Flash memory device in which memory cells are provided in the form of an array of 3D NAND memory strings 338 above interconnect layer 332 and bonding layer

328. Each 3D NAND memory string 338 extends vertically through a plurality of pairs each including a conductor layer 334 and a dielectric layer 336, according to some embodiments. The stacked and interleaved conductor layers 334 and dielectric layer 336 are also referred to herein as a memory stack 333. Interleaved conductor layers 334 and dielectric layers 336 in memory stack 333 alternate in the vertical direction, according to some embodiments. In other words, except for the ones at the top or bottom of memory stack 333, each conductor layer 334 can be adjoined by two dielectric layers 336 on both sides, and each dielectric layer 336 can be adjoined by two conductor layers 334 on both sides. Conductor layers 334 can each have the same thickness or different thicknesses. Similarly, dielectric layers 336 can each have the same thickness or different thicknesses. Conductor layers 334 can include conductor materials including, but not limited to, W, Co, Cu, Al, doped silicon, silicides, or any combination thereof. Dielectric layers 336 can include dielectric materials including, but not limited to, silicon oxide, silicon nitride, silicon oxynitride, or any combination thereof.

[0050] In some embodiments, each 3D NAND memory string 338 is a "charge trap" type of NAND memory string including a semiconductor channel 342 and a memory film 340. In some embodiments, semiconductor channel 342 includes silicon, such as amorphous silicon, polysilicon, or single crystalline silicon. In some embodiments, memory film 340 is a composite dielectric layer including a tunneling layer, a storage layer (also known as "charge trap/storage layer"), and a blocking layer. Each 3D NAND memory string 338 can have a cylinder shape (e.g., a pillar shape). Semiconductor channel 342, the tunneling layer, the storage layer, and the blocking layer of memory film 340 are arranged along a direction from the center toward the outer surface of the pillar in this order, according to some embodiments. The tunneling layer can include silicon oxide, silicon oxynitride, or any combination thereof. The storage layer can include silicon nitride, silicon oxynitride, silicon, or any combination thereof. The blocking layer can include silicon oxide, silicon oxynitride, high dielectric constant (high-k) dielectrics, or any combination thereof. In one example, the blocking layer can include a composite layer of silicon oxide/silicon oxynitride/silicon oxide (ONO). In another example, the blocking layer can include a high-k dielectric layer, such as an aluminum oxide (Al_2O_3) layer, a hafnium oxide (HfO_2) layer, a tantalum oxide (Ta_2O_5) layer, etc.

[0051] In some embodiments, 3D NAND memory strings 338 further include a plurality of control gates (each being part of a word line). Each conductor layer 334 in memory stack

333 can act as a control gate for each memory cell of 3D NAND memory string 338. In some embodiments, each 3D NAND memory string 338 includes two plugs 344 and 346 at a respective end in the vertical direction. Plug 344 can include a semiconductor material, such as single-crystal silicon, that is epitaxially grown from a semiconductor layer 348. Plug 344 can function as the channel controlled by a source select gate of 3D NAND memory string 338. Plug 344 can be at the upper end of 3D NAND memory string 338 and in contact with semiconductor channel 342. As used herein, the "upper end" of a component (e.g., 3D NAND memory string 338) is the end father away from substrate 308 in the *y*-direction, and the "lower end" of the component (e.g., 3D NAND memory string 338) is the end closer to substrate 308 in the *y*-direction when substrate 308 is positioned in the lowest plane of 3D memory device 300. Another Plug 346 can include semiconductor materials (e.g., polysilicon) or conductor materials (e.g., metals). In some embodiments, plug 346 includes an opening filled with titanium/titanium nitride (Ti/TiN as a barrier layer) and tungsten (as a conductor). By covering the upper end of 3D NAND memory string 338 during the fabrication of 3D memory device 300, plug 346 can function as an etch stop layer to prevent etching of dielectrics filled in 3D NAND memory string 338, such as silicon oxide and silicon nitride. In some embodiments, plug 346 functions as the drain of 3D NAND memory string 338.

[0052] In some embodiments, second semiconductor structure 304 further includes semiconductor layer 348 disposed above memory stack 333 and 3D NAND memory strings 338. Semiconductor layer 348 can be a thinned substrate on which memory stack 333 and 3D NAND memory strings 338 are formed. In some embodiments, semiconductor layer 348 includes single-crystal silicon from which plugs 344 can be epitaxially grown. In some embodiments, semiconductor layer 348 can include polysilicon, amorphous silicon, SiGe, GaAs, Ge, or any other suitable materials. Semiconductor layer 348 can also include isolation regions and doped regions (e.g., functioning as an array common source for 3D NAND memory strings 338, not shown). Isolation regions (not shown) can extend across the entire thickness or part of the thickness of semiconductor layer 348 to electrically isolate the doped regions. In some embodiments, a pad oxide layer including silicon oxide is disposed between memory stack 333 and semiconductor layer 348.

[0053] It is understood that 3D NAND memory strings 338 are not limited to the "charge trap" type of 3D NAND memory strings and may be "floating gate" type of 3D NAND memory

strings in other embodiments. Semiconductor layer 348 may include polysilicon as the source plate of the "floating gate" type of 3D NAND memory strings.

[0054] As shown in FIG. 3, second semiconductor structure 304 of 3D memory device 300 can further include a pad-out interconnect layer 350 above semiconductor layer 348. Pad-out interconnect layer 350 include interconnects, e.g., contact pads 352, in one or more ILD layers. The pad-out interconnect layer 350 and interconnect layer 332 can be formed at opposite sides of semiconductor layer 348. In some embodiments, the interconnects in pad-out interconnect layer 350 can transfer electrical signals between 3D memory device 300 and outside circuits, e.g., for pad-out purposes.

[0055] In some embodiments, second semiconductor structure 304 further includes one or more contacts 354 extending through semiconductor layer 348 to electrically connect pad-out interconnect layer 350 and interconnect layers 332 and 322. As a result, data processing circuit 314 can be electrically connected to array of 3D NAND memory strings 338 through interconnect layers 322 and 332 as well as bonding contacts 326 and 330. Moreover, peripheral circuit 312, data processing circuit 314, and array of 3D NAND memory strings 338 can be electrically connected to outside circuits through contacts 354 and pad-out interconnect layer 350.

[0056] FIG. 4 illustrates a cross-section of another exemplary 3D memory device 400 with a data processing circuit, according to some embodiments. Similar to 3D memory device 300 described above in FIG. 3, 3D memory device 400 represents an example of a non-monolithic 3D memory device in which a first semiconductor structure 402 including 3D NAND memory strings and a second semiconductor structure 404 including peripheral circuits and data processing circuit(s) are formed separately and bonded in a face-to-face manner at a bonding interface 406. Different from 3D memory device 300 described above in FIG. 3 in which first semiconductor structure 302 including peripheral circuits and data processing circuit(s) is below second semiconductor structure 304 including 3D NAND memory strings, 3D memory device 400 in FIG. 4 includes second semiconductor structure 404 including peripheral circuits and data processing circuit(s) disposed above first semiconductor structure 402 including 3D NAND memory strings. It is understood that the details of similar structures (e.g., materials, fabrication process, functions, etc.) in both 3D memory devices 300 and 400 may not be repeated below.

[0057] First semiconductor structure 402 of 3D memory device 400 can include a substrate 408 and a memory stack 410 including interleaved conductor layers 412 and dielectric layers 414 above substrate 408. In some embodiments, an array of 3D NAND memory strings 416 each extends vertically through interleaved conductor layers 412 and dielectric layers 414 in memory stack 410 above substrate 408. Each 3D NAND memory string 416 can include a semiconductor channel 420 and a memory film 418. Each 3D NAND memory string 416 further includes two plugs 422 and 424 at its lower end and upper end, respectively. 3D NAND memory strings 416 can be "charge trap" type of 3D NAND memory strings or "floating gate" type of 3D NAND memory strings. In some embodiments, a pad oxide layer including silicon oxide is disposed between memory stack 410 and substrate 408.

[0058] In some embodiments, first semiconductor structure 402 of 3D memory device 400 also includes an interconnect layer 426 above memory stack 410 and 3D NAND memory strings 416 to transfer electrical signals to and from 3D NAND memory strings 416. Interconnect layer 426 can include a plurality of interconnects, including interconnect lines and via contacts. In some embodiments, the interconnects in interconnect layer 426 also include local interconnects, such as bit line contacts and word line contacts. In some embodiments, first semiconductor structure 402 of 3D memory device 400 further includes a bonding layer 428 at bonding interface 406 and above interconnect layer 426 and memory stack 410. Bonding layer 428 can include a plurality of bonding contacts 430 and dielectrics surrounding and electrically isolating bonding contacts 430.

[0059] As shown in FIG. 4, second semiconductor structure 404 of 3D memory device 400 includes another bonding layer 432 at bonding interface 406 and above bonding layer 428. Bonding layer 432 can include a plurality of bonding contacts 434 and dielectrics surrounding and electrically isolating bonding contacts 434. In some embodiments, second semiconductor structure 404 of 3D memory device 400 also includes an interconnect layer 436 above bonding layer 432 to transfer electrical signals. Interconnect layer 436 can include a plurality of interconnects, including interconnect lines and via contacts.

[0060] Second semiconductor structure 404 of 3D memory device 400 can further include a device layer 438 above interconnect layer 436 and bonding layer 432. In some embodiments, device layer 438 includes a peripheral circuit 442 above interconnect layer 436 and bonding layer 432 and a data processing circuit 444 above interconnect layer 436 and bonding layer 432 and outside of peripheral circuit 442. In some embodiments, peripheral circuit 442

includes a plurality of peripheral transistors 446 forming any suitable digital, analog, and/or mixed-signal peripheral circuits used for facilitating the operation of 3D memory device 400 including, but not limited to, a page buffer, a decoder (e.g., a row decoder and a column decoder), a sense amplifier, a driver, a charge pump, a current or voltage reference. Peripheral transistors 446 can be formed "on" a semiconductor layer 440, in which the entirety or part of peripheral transistors 446 are formed in semiconductor layer 440 and/or directly on semiconductor layer 440. Isolation regions (e.g., shallow trench isolations (STIs)) and doped regions (e.g., source regions and drain regions of peripheral transistors 446) can be formed in semiconductor layer 440 as well.

[0061] In some embodiments, data processing circuit 444 includes a plurality of transistors 448 (e.g., MOSFETs). In some embodiments, transistors 448 may form logic gates of an FPGA. In some embodiments, transistors 448 may be part of an ASIC, an MCU, or other data processing circuits capable of performing computation, data operations, and/or logic operations. In some embodiments, transistors 448 are formed "on" semiconductor layer 440, in which the entirety or part of transistors 448 are formed in semiconductor layer 440 and/or directly on semiconductor layer 440. Isolation regions (e.g., shallow trench isolations (STIs)) and doped regions (e.g., source regions and drain regions of transistors 448) can be formed in semiconductor layer 440 as well. As shown in FIG. 4, transistors 448 and peripheral transistors 446 can be formed in different regions on the same plane, e.g., on semiconductor layer 440. That is, transistors 448 can be formed outside of the region in which peripheral circuit 442 is formed on semiconductor layer 440. For ease of illustration, FIG. 4 only depicts a limited number of transistors 448. An electrode contact 450 may connect one or more transistors 448 (e.g., via electrodes of MOSFETs) to bonding contact 434 to facilitate data communication between data processing circuit 444 and memory arrays in first semiconductor structure 402 through bonding interface 406. It is understood that the configuration in FIG. 4, e.g., the layout of transistors and the connection between transistors 448 and bonding contact 434, do not reflect the actual layout and electrical connection between the transistors and other components.

[0062] In some embodiments, second semiconductor structure 404 further includes semiconductor layer 440 disposed above device layer 438. Semiconductor layer 440 can be a thinned substrate on which peripheral transistors 446 and transistors 448 are formed. In some embodiments, semiconductor layer 440 includes single-crystal silicon. In some embodiments,

semiconductor layer 440 can include polysilicon, amorphous silicon, SiGe, GaAs, Ge, or any other suitable materials. Semiconductor layer 440 can also include isolation regions and doped regions.

[0063] As shown in FIG. 4, second semiconductor structure 404 of 3D memory device 400 can further include a pad-out interconnect layer 452 above semiconductor layer 440. Pad-out interconnect layer 452 include interconnects, e.g., contact pads 454, in one or more ILD layers. In some embodiments, the interconnects in pad-out interconnect layer 452 can transfer electrical signals between 3D memory device 400 and outside circuits, e.g., for pad-out purposes. In some embodiments, second semiconductor structure 404 further includes one or more contacts 456 extending through semiconductor layer 440 to electrically connect pad-out interconnect layer 452 and interconnect layers 436 and 426. As a result, data processing circuit 444 can be electrically connected to array of 3D NAND memory strings 416 through interconnect layers 426 and 436 as well as bonding contacts 430 and 434. Moreover, peripheral circuit 442, data processing circuit 444, and array of 3D NAND memory strings 416 can be electrically connected to outside circuits through contacts 456 and pad-out interconnect layer 452.

[0064] FIGS. 5A–5C illustrate a fabrication process for forming an exemplary semiconductor structure having a peripheral circuit and a data processing circuit, according to some embodiments. FIGs. 6A and 6B illustrate a fabrication process for forming an exemplary semiconductor structure having 3D NAND memory strings, according to some embodiments. FIGs. 7A and 7B illustrate a fabrication process for forming an exemplary 3D memory device with a data processing circuit, according to some embodiments. FIG. 11 is a flowchart of an exemplary method 1100 for forming a 3D memory device with a data processing circuit, according to some embodiments. Examples of the 3D memory device depicted in FIGS. 5A-5C, 6A-6B, 7A-7B, and 11 include 3D memory device 300 depicted in FIG. 3 and 3D memory device 400 depicted in FIG. 4. FIGS. 5A-5C, 6A-6B, 7A-7B, and 11 will be described together. It is understood that the operations shown in method 1100 are not exhaustive and that other operations can be performed as well before, after, or between any of the illustrated operations. Further, some of the operations may be performed simultaneously, or in a different order than shown in FIG. 11.

[0065] As depicted in FIGS. 5A–5C, a first semiconductor structure including a peripheral circuit, a data processing circuit, and a first bonding layer including a plurality of first bonding

contacts is formed. As depicted in FIGS. 6A and 6B, a second semiconductor structure including an array of 3D NAND memory strings and a second bonding layer including a plurality of second bonding contacts is formed. As depicted in FIGS. 7A and 7B, the first semiconductor structure and the second semiconductor structure are bonded in a face-to-face manner, such that the first bonding contacts are in contact with the second bonding contacts at a bonding interface.

[0066] Referring to FIG. 11, method 1100 starts at operation 1102, in which the peripheral circuit and the data processing circuit are formed on a first substrate. The first substrate can be a silicon substrate. In some embodiments, to form the peripheral circuit and the data processing circuit, a plurality of transistors are formed on the first substrate. As illustrated in FIG. 5A, a plurality of transistors (e.g., peripheral transistors 504 and logic transistors 506 for forming the data processing circuit) are formed on a silicon substrate 502. Transistors 504 and 506 can be formed by a plurality of processes including, but not limited to, photolithography, dry/wet etch, thin film deposition, thermal growth, implantation, chemical mechanical polishing (CMP), and any other suitable processes. In some embodiments, doped regions are formed in silicon substrate 502 by ion implantation and/or thermal diffusion, which function, for example, as source regions and/or drain regions of transistors 504 and 506. In some embodiments, isolation regions (e.g., STIs) are also formed in silicon substrate 502 by wet/dry etch and thin film deposition.

[0067] As illustrated in FIG. 5B, electrode contacts 507 are formed as well for connecting logic transistors 506. A device layer 510 including a peripheral circuit (having peripheral transistors 504) and a data processing circuit (having logic transistors 506) are thereby formed.

[0068] Method 1100 proceeds to operation 1104, as illustrated in FIG. 11, in which a first interconnect layer is formed above the peripheral circuit and the data processing circuit. The first interconnect layer can include a first plurality of interconnects in one or more ILD layers. As illustrated in FIG. 5C, an interconnect layer 512 can be formed above device layer 510 including the peripheral circuit (having peripheral transistors 504) and the data processing circuit (having logic transistors 506). Interconnect layer 512 can include interconnects of MEOL and/or BEOL in a plurality of ILD layers to make electrical connections with device layer 510. In some embodiments, interconnect layer 512 includes multiple ILD layers and interconnects therein formed in multiple processes. For example, the interconnects in

interconnect layers 512 can include conductive materials deposited by one or more thin film deposition processes including, but not limited to, CVD, PVD, ALD, electroplating, electroless plating, or any combination thereof. Fabrication processes to form the interconnects can also include photolithography, CMP, wet/dry etch, or any other suitable processes. The ILD layers can include dielectric materials deposited by one or more thin film deposition processes including, but not limited to, CVD, PVD, ALD, or any combination thereof. The ILD layers and interconnects illustrated in FIG. 5C can be collectively referred to as interconnect layer 512.

[0069] Method 1100 proceeds to operation 1106, as illustrated in FIG. 11, in which a first bonding layer is formed above the first interconnect layer. The first bonding layer can include a plurality of first bonding contacts. As illustrated in FIG. 5C, a bonding layer 514 is formed above interconnect layer 512. Bonding layer 514 can include a plurality of bonding contacts 516 surrounded by dielectrics. In some embodiments, a dielectric layer is deposited on the top surface of interconnect layer 512 by one or more thin film deposition processes including, but not limited to, CVD, PVD, ALD, or any combination thereof. Bonding contacts 516 then can be formed through the dielectric layer and in contact with the interconnects in interconnect layer 512 by first patterning contact holes through the dielectric layer using patterning process (e.g., photolithography and dry/wet etch of dielectric materials in the dielectric layer). The contact holes can be filled with a conductor (e.g., copper). In some embodiments, filling the contact holes includes depositing a barrier layer, an adhesion layer, and/or a seed layer before depositing the conductor.

[0070] Method 1100 proceeds to operation 1108, as illustrated in FIG. 11, in which a memory stack is formed above a second substrate. The second substrate can be a silicon substrate. As illustrated in FIG. 6A, interleaved sacrificial layers (not shown) and dielectric layers 608 are formed above a silicon substrate 602. The interleaved sacrificial layers and dielectric layers 608 can form a dielectric stack (not shown). In some embodiments, each sacrificial layer includes a layer of silicon nitride, and each dielectric layer 608 includes a layer of silicon oxide. The interleaved sacrificial layers and dielectric layers 608 can be formed by one or more thin film deposition processes including, but not limited to, CVD, PVD, ALD, or any combination thereof. In some embodiments, a memory stack 604 can be formed by a gate replacement process, e.g., replacing the sacrificial layers with conductor layers 606 using wet/dry etch of the sacrificial layers selective to dielectric layers 608 and filling the resulting

recesses with conductor layers 606. As a result, memory stack 604 can include interleaved conductor layers 606 and dielectric layers 608. In some embodiments, each conductor layer 606 includes a metal layer, such as a layer of tungsten. It is understood that memory stack 604 may be formed by alternatingly depositing conductor layers (e.g., doped polysilicon layers) and dielectric layers (e.g., silicon oxide layers) without the gate replacement process in other embodiments. In some embodiments, a pad oxide layer including silicon oxide is formed between memory stack 604 and silicon substrate 602.

[0071] Method 1100 proceeds to operation 1110, as illustrated in FIG. 11, in which the array of 3D NAND memory strings extending vertically through the memory stack are formed. As illustrated in FIG. 6A, 3D NAND memory strings 610 are formed above silicon substrate 602, each of which extends vertically through interleaved conductor layers 606 and dielectric layers 608 of memory stack 604. In some embodiments, fabrication processes to form 3D NAND memory string 610 include forming a channel hole through memory stack 604 and into silicon substrate 602 using dry etching/and or wet etching, such as deep reactive-ion etching (DRIE), followed by epitaxially growing a plug 612 in the lower portion of the channel hole from silicon substrate 602. In some embodiments, fabrication processes to form 3D NAND memory string 610 also include subsequently filling the channel hole with a plurality of layers, such as a memory film 614 (e.g., a tunneling layer, a storage layer, and a blocking layer) and a semiconductor layer 616, using thin film deposition processes such as ALD, CVD, PVD, or any combination thereof. In some embodiments, fabrication processes to form 3D NAND memory string 610 further include forming another plug 618 in the upper portion of the channel hole by etching a recess at the upper end of 3D NAND memory string 610, followed by filling the recess with a semiconductor material using thin film deposition processes such as ALD, CVD, PVD, or any combination thereof.

[0072] Method 1100 proceeds to operation 1112, as illustrated in FIG. 11, in which a second interconnect layer is formed above the array of 3D NAND memory strings. The second interconnect layer can include a second plurality of interconnects in one or more ILD layers. As illustrated in FIG. 6B, an interconnect layer 620 can be formed above memory stack 604 and array of 3D NAND memory strings 610. Interconnect layer 620 can include interconnects of MEOL and/or BEOL in a plurality of ILD layers to make electrical connections with 3D NAND memory strings 610. In some embodiments, interconnect layer 620 includes multiple ILD layers and interconnects therein formed in multiple processes. For

example, the interconnects in interconnect layers 620 can include conductive materials deposited by one or more thin film deposition processes including, but not limited to, CVD, PVD, ALD, electroplating, electroless plating, or any combination thereof. Fabrication processes to form the interconnects can also include photolithography, CMP, wet/dry etch, or any other suitable processes. The ILD layers can include dielectric materials deposited by one or more thin film deposition processes including, but not limited to, CVD, PVD, ALD, or any combination thereof. The ILD layers and interconnects illustrated in FIG. 6B can be collectively referred to as interconnect layer 620.

[0073] Method 1100 proceeds to operation 1114, as illustrated in FIG. 11, in which a second bonding layer is formed above the second interconnect layer. The second bonding layer can include a plurality of second bonding contacts. As illustrated in FIG. 6B, a bonding layer 622 is formed above interconnect layer 620. Bonding layer 622 can include a plurality of bonding contacts 624 surrounded by dielectrics. In some embodiments, a dielectric layer is deposited on the top surface of interconnect layer 620 by one or more thin film deposition processes including, but not limited to, CVD, PVD, ALD, or any combination thereof. Bonding contacts 624 then can be formed through the dielectric layer and in contact with the interconnects in interconnect layer 620 by first patterning contact holes through the dielectric layer using patterning process (e.g., photolithography and dry/wet etch of dielectric materials in the dielectric layer). The contact holes can be filled with a conductor (e.g., copper). In some embodiments, filling the contact holes includes depositing a barrier layer, an adhesion layer, and/or a seed layer before depositing the conductor.

[0074] Method 1100 proceeds to operation 1116, as illustrated in FIG. 11, in which the first substrate and the second substrate are bonded in a face-to-face manner, such that the first bonding contacts are in contact with the second bonding contacts at the bonding interface. The bonding can be hybrid bonding. In some embodiments, the first substrate on which the peripheral circuit and the data processing circuit are formed (e.g., the first semiconductor structure) is disposed above the second substrate on which the 3D NAND memory strings are formed (e.g., the second semiconductor structure) after the bonding. In some embodiments, the second substrate on which the 3D NAND memory strings are formed (e.g., the second semiconductor structure) is disposed above the first substrate on which the peripheral circuit and the data processing circuit are formed (e.g., the first semiconductor structure) after the bonding.

[0075] As illustrated in FIG. 7A, silicon substrate 602 and components formed thereon (e.g., 3D NAND memory strings 610) are flipped upside down. Bonding layer 622 facing down is bonded with bonding layer 514 facing up, i.e., in a face-to-face manner, thereby forming a bonding interface 702 (as shown in FIG. 7B). In some embodiments, a treatment process, e.g., a plasma treatment, a wet treatment, and/or a thermal treatment, is applied to the bonding surfaces prior to the bonding. Although not shown in FIG. 7A, silicon substrate 502 and components formed thereon (e.g., device layer 510) can be flipped upside down, and bonding layer 514 facing down can be bonded with bonding layer 622 facing up, i.e., in a face-to-face manner, thereby forming bonding interface 702. After the bonding, bonding contacts 624 in bonding layer 622 and bonding contacts 516 in bonding layer 514 are aligned and in contact with one another, such that device layer 510 (e.g., the peripheral circuit and the data processing circuit therein) can be electrically connected to 3D NAND memory strings 610. It is understood that in the bonded device, 3D NAND memory strings 610 may be either above or below device layer 510 (e.g., the peripheral circuit and the data processing circuit therein). Nevertheless, bonding interface 702 can be formed between 3D NAND memory strings 610 and device layer 510 (e.g., the peripheral circuit and the data processing circuit therein) after the bonding as illustrated in FIG. 7B.

[0076] Method 1100 proceeds to operation 1118, as illustrated in FIG. 11, in which the first substrate or the second substrate is thinned to form a semiconductor layer. In some embodiments, the first substrate of the first semiconductor structure, which is above the second substrate of the second semiconductor structure after the bonding, is thinned to form the semiconductor layer. In some embodiments, the second substrate of the second semiconductor structure, which is above the first substrate of the first semiconductor structure after the bonding, is thinned to form the semiconductor layer.

[0077] As illustrated in FIG. 7B, the substrate on top of the bonded 3D memory device (e.g., silicon substrate 402 as shown in FIG. 7A) is thinned, so that the thinned top substrate can serve as a semiconductor layer 704, for example, a single-crystal silicon layer. The thickness of the thinned substrate can be between about 200 nm and about 5 μm, such as between 200 nm and 5 μm, or between about 150 nm and about 50 μm, such as between 150 nm and 50 μm. Silicon substrate 402 can be thinned by processes including, but not limited to, wafer grinding, dry etch, wet etch, CMP, any other suitable processes, or any combination thereof. It is understood that when silicon substrate 502 is the substrate on top of the bonded 3D

memory device, another semiconductor layer may be formed by thinning silicon substrate 502.

[0078] Method 1100 proceeds to operation 1120, as illustrated in FIG. 11, in which a pad-out interconnect layer is formed above the semiconductor layer. As illustrated in FIG. 7B, a pad-out interconnect layer 706 is formed above semiconductor layer 704 (the thinned top substrate). Pad-out interconnect layer 706 can include interconnects, such as pad contacts 708, formed in one or more ILD layers. Pad contacts 708 can include conductive materials including, but not limited to, W, Co, Cu, Al, doped silicon, silicides, or any combination thereof. The ILD layers can include dielectric materials including, but not limited to, silicon oxide, silicon nitride, silicon oxynitride, low-k dielectrics, or any combination thereof. In some embodiments, after the bonding and thinning, contacts 710 are formed extending vertically through semiconductor layer 704, for example by wet/dry etch followed by depositing conductive materials. Contacts 710 can be in contact with the interconnects in pad-out interconnect layer 706.

[0079] FIG. 8 illustrates a schematic block diagram of an exemplary system 850 having a 3D memory device with an on-die data processing system, according to some embodiments. FIG. 9 illustrates a schematic block diagram of system 920 having a 3D memory device with an on-die data processing circuit, according to some embodiments. FIG. 10 is a flowchart of an exemplary method 1000 for operating a 3D memory device with an on-die data processing circuit, according to some embodiments. The exemplary system illustrated in FIG. 9 is described together with FIG. 10. It is understood that the operations shown in method 1000 is not exhaustive and that other operations can be performed as well before, after, or between any of the illustrated operations. Further, some of the operations may be performed simultaneously, or in a different order than shown in FIG. 10.

[0080] FIG. 8 illustrates a system 850 having an on-die data processing circuit, according to some embodiments. System 850 may have a host 810, an I/O interface 802 (also referred to as "I/O" for short), a data processing circuit 804, a page buffer 806, and a 3D NAND memory (also referred to as 3D NAND memory array) 808. In some embodiments, data processing circuit 804 is formed on the same chip with page buffer 806 and 3D NAND memory 808, as described above in detail. In some embodiments, data processing circuit 804, page buffer 806, and 3D NAND memory 808 form a 3D memory device 800. Data processing circuit 804 can be referred to as an on-die data processing circuit, an in-memory computation circuit,

an on-NAND data processing circuit, or the like. Data, e.g., program data and control instructions, may be transmitted bi-directionally between host 810 and I/O interface 802, between I/O interface 802 and data processing circuit 804, between data processing circuit 804 and page buffer 806, and between 3D NAND memory 808 and page buffer 806. Data transmission 812 between host 810 and page buffer 806 through I/O interface 802 may be enabled or disabled, depending on the functions of data processing circuit 804. For example, when data processing circuit 804 performs data processing operations locally in 3D memory device 800 and returns or relays the results of the operation to host 810, data transmission 812 may be bypassed or disabled. When data processing circuit 804 is not used to perform data processing operations locally in 3D memory device 800, data transmission 812 can be bi-directional data transmission between host 810 and page buffer 806. That is, when the data processing capability of data processing circuit 804 is invoked, the bulk of data processing operations on the underlying data stored in 3D NAND memory 808 can be performed within 3D memory device 800. As a result, the burden placed on I/O interface 802 due to the transmission of the underlying data from 3D memory device 800 to host 810 can be lifted, as only the operation results are being transmitted through I/O interface 802. On the other hand, when the data processing capability of data processing circuit 804 is not invoked, host 810 may access data stored in 3D NAND memory 808 through I/O interface 802 and page buffer 806, bypassing data processing circuit 804.

[0081] Data transmission 814 between data processing circuit 804 and 3D NAND memory 808 allows data processing circuit 804 to directly access data stored in 3D NAND memory 808. In some embodiments, data transmission 814 allows data processing circuit 804 to perform operations based on programs stored in 3D NAND memory 808. The programs may be written in 3D NAND memory 808 by host 810 (e.g., through data transmission 812). In some embodiment, data processing circuit 804 may include FPGA that is programmable based on the programs stored in 3D NAND memory 808. For example, data processing circuit 804 may be programmed and/or re-programmed based on the programs stored in 3D NAND memory 808 to perform various or differing operations.

[0082] Host 810 can be any suitable devices that generate data and/or instructions, such as one or more processors. In some embodiments, host 810 includes a central processing unit (CPU), a graphics processor (e.g., graphics processing unit (GPU)), an application processor (AP), a general processor (e.g., APU, accelerated processing unit; GPGPU, general-purpose

computing on GPU), or any other suitable processor. I/O interface 802 can be a high-speed, high-throughput input/output circuit as part of the peripheral circuits. In some embodiments, host 810 includes a system controller (e.g., a controller that controls various operations of system 850) and/or a memory controller (e.g., a controller that controls various operations of 3D memory device 800). Any suitable type of data and/or instructions generated by host 810 can be transferred to data processing circuit 804 of 3D memory device 800 through I/O interface 802. For example, host 810 may generate instructions for performing certain operations on data stored in 3D NAND memory 808 and transfer the instructions to data processing circuit 804. Instead of loading all underlying data subject to the operations to host 810 through I/O interface 802, data processing circuit 804 may perform the operations locally, in an on-site and in-memory fashion, and only return the operation result to host 810. Host 810 and 3D memory device 800 can be part of any suitable apparatus, for example, a virtual reality (VR)/augmented reality (AR) device (e.g., VR headset, etc.), handheld device (e.g., dumb or smart phone, tablet, etc.), wearable device (e.g., eyeglasses, wrist watch, etc.), automobile control station, gaming console, television set, laptop computer, desktop computer, netbook computer, media center, set-top box, global positioning system (GPS), printer, or any other suitable device.

[0083] In some embodiments, data processing circuit 804 includes a plurality of transistors (e.g., logic transistors). Details of data processing circuit 804 can be found in the description of FIGS. 3, 4, and 5A-5C and thus, are not repeated herein. Data processing circuit 804 may be connected to page buffer 806, which includes a plurality of buffering sections connected to respective pages in 3D NAND memory 808.

[0084] Data processing circuit 804 may be employed as an in-memory computation unit of 3D memory device 800 to improve the speed of data operations by, for example, reducing the amount of data occupying the bandwidth of I/O interface 802, thereby improving the overall throughput of system 850. FIG. 9 illustrates a system 920 in which data processing circuit 804 is configured to perform operations on data stored in 3D memory device 800. For ease of depiction, I/O 802 interface is omitted in FIG. 9.

[0085] In some embodiments, data are stored in 3D NAND memory 808 in pages, and data processing circuit 804 may retrieve data from 3D NAND memory 808 in the unit of a page. As shown in FIG. 9, 3D NAND memory 808 may be depicted as a plurality of planes 908 (i.e., 908-1, ..., 908-M, M is a positive integer), each represents memory cells formed by a

word line and intersecting memory strings. A plane 908 may include a plurality of pages of memory cells. In operation, data stored in one page of 3D NAND memory 808 may be loaded or cached into page buffer 806 in one batch, which may then be retrieved by data processing circuit 804 as one unit of data. In some embodiments, multiple pages of data may be loaded/cached into one or more page buffers simultaneously. In other embodiments, multiple pages of data may be loaded/cached into one or more page buffers sequentially (e.g., one batch immediately after another batch). Data processing circuit 804 may retrieve cached data from page buffer 806 in parallel (e.g., retrieving multiple pages of data simultaneously) or in series (e.g., retrieving one page of data immediately after another page of data). After retrieving the data in the unit of pages, data processing circuit 804 may perform operations on the retrieved data simultaneously (e.g., processing multiple pages of data at the same time). In some embodiments, data processing circuit 804 may perform operations on the retrieved data sequentially (e.g., processing one page of data immediately after another page of data).

[0086] In some embodiments, data processing circuit 804 may access data stored in 3D NAND memory 808 directly through data transmission 814. In this case, data stored in 3D NAND memory 808 may be loaded directly into data processing circuit 804. For example, 3D NAND memory 808 may store program codes executable or configurable by data processing circuit 804 (e.g. data processing circuit 804 may include an FPGA, MCU, ASIC, or the like). The program codes may be loaded to data processing circuit 804 through data transmission 814. Data processing circuit 804 may be programmed based on the program codes to perform operations on data stored in 3D NAND memory 808. In some embodiments, data processing circuit 804 may be re-programmed using different program codes to perform different operations.

[0087] Referring to FIG. 10, method 1000 starts at operation 1002, in which 3D memory device 800 receives control instructions from host 810. The control instructions may be issued by host 810 to perform certain operations on data stored in 3D memory device 800 (e.g., stored in 3D NAND memory 808). For example, the control instructions may include performing a searching operation to search for certain record(s) from a database stored in 3D memory device 800. Conventionally, the underlying data of the database need to be transmitted to host 810, where host 810 would perform the searching operation on the data. When the amount of data is large, transmitting the underlying data of the entire database would place a heavy burden on I/O interface 802, limiting the overall system throughput. With on-die, in-

memory computation and data processing capability provided by data processing circuit 804, the bulk of the searching operation can be performed in-memory, without burdening I/O interface 802. Once the searching operation is completed, only the search result needs to be transmitted to host 810 through I/O interface 802. This in-memory computation functionality thus can greatly lower the data transmission burden placed on I/O interface 802. In addition, because data processing circuit 804 can be connected to 3D NAND memory array 808 through hybrid bonding, high data transmission rate can be achieved thanks to the short and direct physical connections between data processing circuit 804 and 3D NAND memory array 808. In-memory computation capability provided by data processing circuit 804 can be used to perform various operations within 3D memory device 800 such as searching operations discussed above, data encryption, and data preprocessing operations.

[0088] At operation 1004, data processing circuit 804 may retrieve the data from 3D NAND memory array 808 based on the control instructions. As discussed above in connection with FIGS. 8 and 9, data processing circuit 804 may retrieve raw data stored in 3D NAND memory array 808 in the unit of a page and may retrieve data stored in multiple pages simultaneously or sequentially. Take searching operations, for example, host 810 may instruct data processing circuit 804 to perform searching operations to search for certain records from data stored in 3D NAND memory 808. Instead of transmitting the raw data to host 810, data processing circuit 804 may perform the search operations within 3D memory device 800 based on, for example, a search program stored in 3D NAND memory array 808. In other words, data processing circuit 804 may be programmed to perform the search operation using program codes stored locally in 3D NAND memory array 808 to perform the search operation on raw data stored in 3D memory device 800. Data stored in pages of 3D memory device 800 can be retrieved in batches, for example, one page at a time, by data processing circuit 804. In some embodiments, data retrieval can be performed in parallel such that multiple pages of data can be loaded into data processing circuit 804 for processing simultaneously. In other embodiments, data retrieval can be performed sequentially such that data processing circuit 804 retrieves one page at a time.

[0089] At operation 1006, data processing circuit 804 may perform error checking and correcting (ECC) operations on the retrieved data. For example, data processing circuit 804 may include a built-in ECC controller to perform the ECC operations. In some embodiments, the built-in ECC controller may be formed together with other components (e.g., logic transistors)

of data processing circuit 804. In some embodiments, the ECC controller may be implemented by programming data processing circuit 804 based on, for example, program codes stored in 3D NAND memory array 808. In another example, data processing circuit 804 may connect to an external ECC controller. The external ECC controller may be part of the peripheral circuits 312/442. In some embodiments, the external ECC controller may be outside 3D memory device 800 and connected to data processing circuit 804. It is noted that ECC operations may be omitted in some applications.

[0090] At operation 1008, data processing circuit 804 may perform the operations on the retrieved data locally in-memory. In the searching operation example, data processing circuit 804 may traverse the raw data retrieved from pages of 3D NAND memory array 808 to search for the records that meet the searching criteria. Because data transmission between data processing circuit 804 and 3D NAND memory array 808 can achieve high speed thanks to the hybrid bonding technology discussed above, the searching operations can be performed very efficiently, without burdening I/O interface 802 that is designed for general data transmission among components of system 850. In another example, data processing circuit 804 may perform data encryption/decryption on the data stored in 3D NAND memory array 808 and store the encrypted/decrypted data back to the memory cells of 3D memory device 800. The encrypted/decrypted data may replace the original data or be stored in different memory cells. Similarly, data processing circuit 804 may perform data compression/decompression, data preprocessing/preconditioning, image rendering, feature leaning/model training, or any other operations requiring logic computation to a large amount of data.

[0091] At operation 1010, data processing circuit 804 may return a result of the operations to host 810 via I/O interface 802. For example, the result of searching operations may include one or more records that meet that searching criterion. In another example, the result of encryption operations may include a confirmation code, a key, a checksum, or the like. In any case, the underlying data upon which these operations are performed need not be transmitted to host 810 via I/O interface 802, relieving I/O interface 802 from carrying heavy data traffic.

[0092] 3D NAND memory 808 may include a multi-level cell (MLC) NAND memory device, in which a number of the plurality of pages corresponds to a number of bits stored in a memory cell. In some embodiments, 3D NAND memory 808 includes a triple-level cell (TLC)

NAND memory device packed in a RAM-less application environment, such as an eMMC or a UFS.

[0093] In some embodiments, a 3D memory device includes a 3D NAND memory array and an on-die data processing circuit coupled to the 3D NAND memory array on a same chip. The on-die data processing circuit is configured to receive, from an input/output (I/O) interface, control instructions for performing operations on data stored in the 3D NAND memory array. The on-die data processing circuit is also configured to retrieve the data from the 3D NAND memory array based on the control instructions and perform the operations on the retrieved data. Moreover, the on-die data processing circuit is configured to return a result of the operations to the I/O interface.

[0094] In some embodiments, the 3D NAND memory array includes a plurality of pages. The on-die data processing circuit is configured to retrieve the data from the 3D NAND memory array in a unit of a page.

[0095] In some embodiments, the on-die data processing circuit is configured to perform the operations on data retrieved from multiple pages simultaneously.

[0096] In some embodiments, the on-die data processing circuit is configured to perform the operations on data retrieved from multiple pages sequentially.

[0097] In some embodiments, the on-die data processing circuit is programmed to perform the operations based on program codes stored in the 3D NAND memory array.

[0098] In some embodiments, the on-die data processing circuit comprises at least one of a field-programmable gate array (FPGA), a microcontroller unit (MCU), or an application-specific integrated circuit (ASIC).

[0099] In some embodiments, the 3D memory device includes a controller configured to perform error checking and correcting (ECC) on the retrieved data.

[0100] In some embodiments, the 3D memory device is packaged in at least one of an eMMC or a UFS.

[0101] In some embodiments, the 3D memory device includes a peripheral circuit configured to perform data storage operations of the 3D memory device. The peripheral circuit and the on-die data processing circuit occupy non-overlapping space of the 3D memory device.

[0102] In some embodiments, a system for operating a data processing unit on a three-dimensional (3D) memory device includes a host and a storage unit having a plurality of NAND memory cells and being coupled to the host. The data processing unit is coupled to

and on a same chip with the storage unit. The data processing unit is configured to receive, from the host, control instructions for performing operations on data stored in the NAND memory cells. The data processing unit is also configured to retrieve the data from the NAND memory cells based on the control instructions and perform the operations on the retrieved data. Moreover, the data processing unit is configured to return a result of the operations to the host.

[0103] In some embodiments, the NAND memory cells are arranged in a plurality of pages. The data processing unit is configured to retrieve the data from the NAND memory cells in a unit of a page.

[0104] In some embodiments, the data processing unit is configured to perform the operations on data retrieved from multiple pages simultaneously.

[0105] In some embodiments, the on-die data processing circuit is configured to perform the operations on data retrieved from multiple pages sequentially.

[0106] In some embodiments, the data processing unit is programmed to perform the operations based on program codes stored in the NAND memory cells.

[0107] In some embodiments, the data processing unit comprises at least one of a field-programmable gate array (FPGA), a microcontroller unit (MCU), or an application-specific integrated circuit (ASIC).

[0108] In some embodiments, the system also includes a controller configured to perform error checking and correcting (ECC) on the retrieved data.

[0109] In some embodiments, the system also includes a peripheral circuit configured to perform data storage operations of the storage unit. The peripheral circuit and the data processing unit occupy non-overlapping space of the storage unit.

[0110] In some embodiments, a method for operating a logic circuit on a three-dimensional (3D) memory device includes receiving, from an input/output (I/O) interface, control instructions for performing operations on data stored in the 3D memory device. The method also includes retrieving, by the logic circuit on a same chip with the 3D memory device, the data from the 3D memory device based on the control instructions and perform, by the logic circuit, the operations on the retrieved data. In addition, the method includes return, by the logic circuit, a result of the operations to the I/O interface.

[0111] In some embodiments, the 3D memory device comprises a plurality of pages and the method includes retrieving, by the logic circuit, the data from the 3D memory device in a unit of a page.

[0112] In some embodiments, the method includes performing the operations on data retrieved from multiple pages simultaneously.

[0113] In some embodiments, the method includes performing the operations on data retrieved from multiple pages sequentially.

[0114] In some embodiments, the method includes performing, by a controller on the same chip with the 3D memory device, error checking and correcting (ECC) on the retrieved data.

[0115] In some embodiments, the logic circuit includes at least one of a field-programmable gate array (FPGA), a microcontroller unit (MCU), or an application-specific integrated circuit (ASIC).

[0116] In some embodiments, the method also includes performing, by a peripheral circuit, storage operations of the 3D memory device. The peripheral circuit and the logic circuit occupy non-overlapping space of the 3D memory device.

[0117] The foregoing description of the specific embodiments will so reveal the general nature of the present disclosure that others can, by applying knowledge within the skill of the art, readily modify and/or adapt for various applications such specific embodiments, without undue experimentation, without departing from the general concept of the present disclosure. Therefore, such adaptations and modifications are intended to be within the meaning and range of equivalents of the disclosed embodiments, based on the teaching and guidance presented herein. It is to be understood that the phraseology or terminology herein is for the purpose of description and not of limitation, such that the terminology or phraseology of the present specification is to be interpreted by the skilled artisan in light of the teachings and guidance.

[0118] Embodiments of the present disclosure have been described above with the aid of functional building blocks illustrating the implementation of specified functions and relationships thereof. The boundaries of these functional building blocks have been arbitrarily defined herein for the convenience of the description. Alternate boundaries can be defined so long as the specified functions and relationships thereof are appropriately performed.

[0119] The Summary and Abstract sections may set forth one or more but not all exemplary embodiments of the present disclosure as contemplated by the inventor(s), and thus, are not intended to limit the present disclosure and the appended claims in any way.

[0120] The breadth and scope of the present disclosure should not be limited by any of the above-described exemplary embodiments, but should be defined only in accordance with the following claims and their equivalents.

WHAT IS CLAIMED IS:

1. A three-dimensional (3D) memory device, comprising:

a 3D NAND memory array; and

an on-die data processing circuit coupled to the 3D NAND memory array on a same chip, wherein the on-die data processing circuit is configured to:

receive, from an input/output (I/O) interface, control instructions for performing operations on data stored in the 3D NAND memory array;

retrieve the data from the 3D NAND memory array based on the control instructions;

perform the operations on the retrieved data; and

return a result of the operations to the I/O interface.

2. The 3D memory device of claim 1, wherein:

the 3D NAND memory array comprises a plurality of pages; and

the on-die data processing circuit is configured to retrieve the data from the 3D NAND memory array in a unit of a page.

3. The 3D memory device of claim 2, wherein the on-die data processing circuit is configured to perform the operations on data retrieved from multiple pages simultaneously.

4. The 3D memory device of claim 2, wherein the on-die data processing circuit is configured to perform the operations on data retrieved from multiple pages sequentially.

5. The 3D memory device of any one of claims 1-4, wherein the on-die data processing circuit is programmed to perform the operations based on program codes stored in the 3D NAND memory array.

6. The 3D memory device of any one of claims 1-5, wherein the on-die data processing circuit comprises at least one of a field-programmable gate array (FPGA), a microcontroller unit (MCU), or an application-specific integrated circuit (ASIC).

7. The 3D memory device of any one of claims 1-6, further comprising a controller configured to perform error checking and correcting (ECC) on the retrieved data.

8. The 3D memory device of any one of claims 1-7, wherein the 3D memory device is packaged in at least one of an embedded multi-media card (eMMC) or a universal flash storage (UFS).

9. The 3D memory device of any one of claims 1-8, further comprising a peripheral circuit configured to perform data storage operations of the 3D memory device, the peripheral circuit and the on-die data processing circuit occupying non-overlapping space of the 3D memory device.

10. A system for operating a data processing unit on a three-dimensional (3D) memory device, comprising:

a host; and

a storage unit having a plurality of NAND memory cells and being coupled to the host;

wherein the data processing unit is coupled to and on a same chip with the storage unit, the data processing unit being configured to:

receive, from the host, control instructions for performing operations on data stored in the NAND memory cells;

retrieve the data from the NAND memory cells based on the control instructions;

perform the operations on the retrieved data; and

return a result of the operations to the host.

11. The system of claim 10, wherein:

the NAND memory cells are arranged in a plurality of pages; and

the data processing unit is configured to retrieve the data from the NAND memory cells in a unit of a page.

12. The system of claim 11, wherein the data processing unit is configured to perform the operations on data retrieved from multiple pages simultaneously.

13. The system of claim 12, wherein the on-die data processing circuit is configured to perform the operations on data retrieved from multiple pages sequentially.

14. The system of any one of claims 10-13, wherein the data processing unit is programmed to perform the operations based on program codes stored in the NAND memory cells.

15. The system of any one of claims 10-14, wherein the data processing unit comprises at least one of a field-programmable gate array (FPGA), a microcontroller unit (MCU), or an application-specific integrated circuit (ASIC).

16. The system of any one of claims 10-15, further comprising a controller configured to perform error checking and correcting (ECC) on the retrieved data.

17. The system of any one of claims 1-8, further comprising a peripheral circuit configured to perform data storage operations of the storage unit, the peripheral circuit and the data processing unit occupying non-overlapping space of the storage unit.

18. A method for operating a logic circuit on a three-dimensional (3D) memory device, comprising:

receiving, from an input/output (I/O) interface, control instructions for performing operations on data stored in the 3D memory device;

retrieving, by the logic circuit on a same chip with the 3D memory device, the data from the 3D memory device based on the control instructions;

perform, by the logic circuit, the operations on the retrieved data; and

return, by the logic circuit, a result of the operations to the I/O interface.

19. The method of claim 18, wherein:

the 3D memory device comprises a plurality of pages; and

the method comprises:

retrieving, by the logic circuit, the data from the 3D memory device in a unit of a page.

20. The method of claim 19, comprising performing the operations on data retrieved from multiple pages simultaneously.

21. The method of claim 19, comprising performing the operations on data retrieved from multiple pages sequentially.

22. The method of any one of claims 18-21, further comprising:

performing, by a controller on the same chip with the 3D memory device, error checking and correcting (ECC) on the retrieved data.

23. The method of any one of claims 18-22, wherein the logic circuit comprises at least one of a field-programmable gate array (FPGA), a microcontroller unit (MCU), or an application-specific integrated circuit (ASIC).

24. The method of any one of claims 18-23, further comprising:

performing, by a peripheral circuit, storage operations of the 3D memory device, the peripheral circuit and the logic circuit occupying non-overlapping space of the 3D memory device.

100

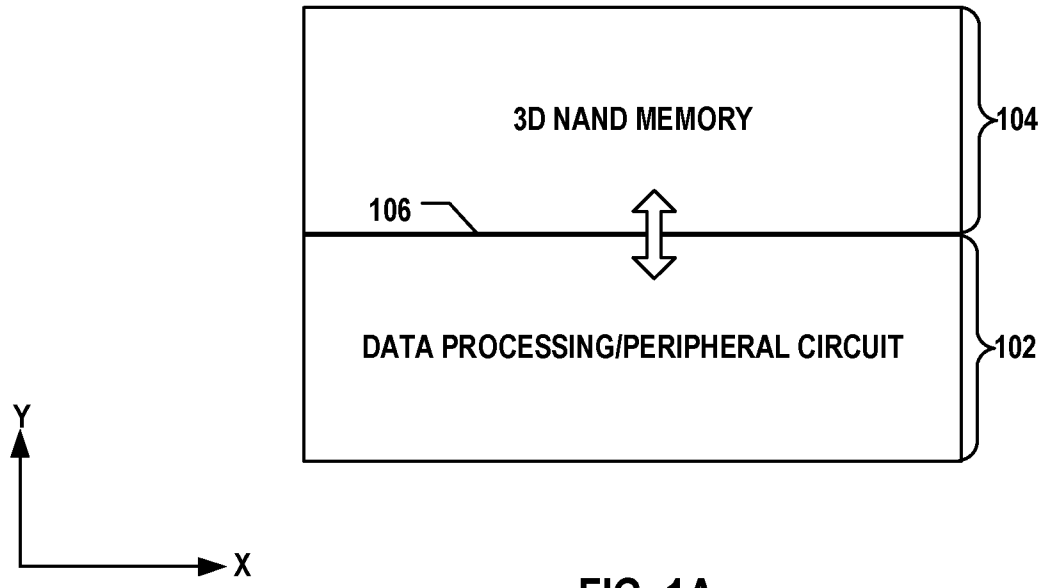

FIG. 1A

101

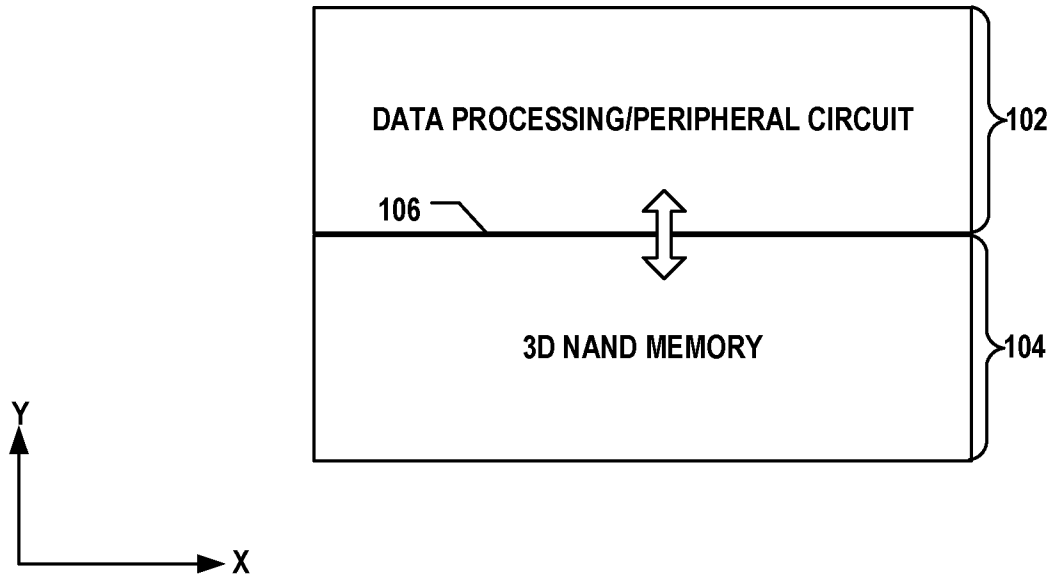

FIG. 1B

200

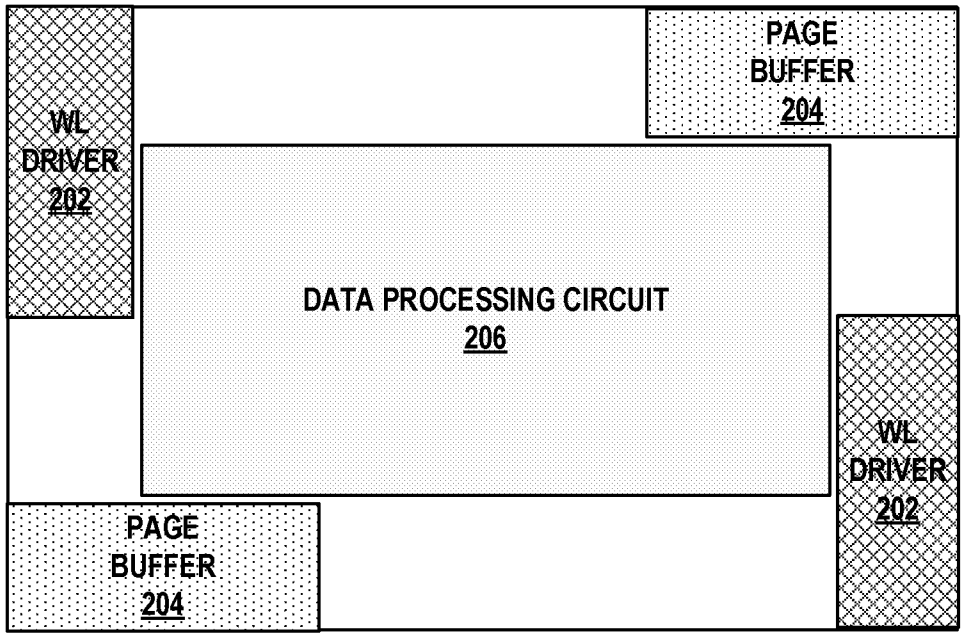
PAGE BUFFER 204
WL DRIVER 202
DATA PROCESSING CIRCUIT 206
WL DRIVER 202
PAGE BUFFER 204

FIG. 2

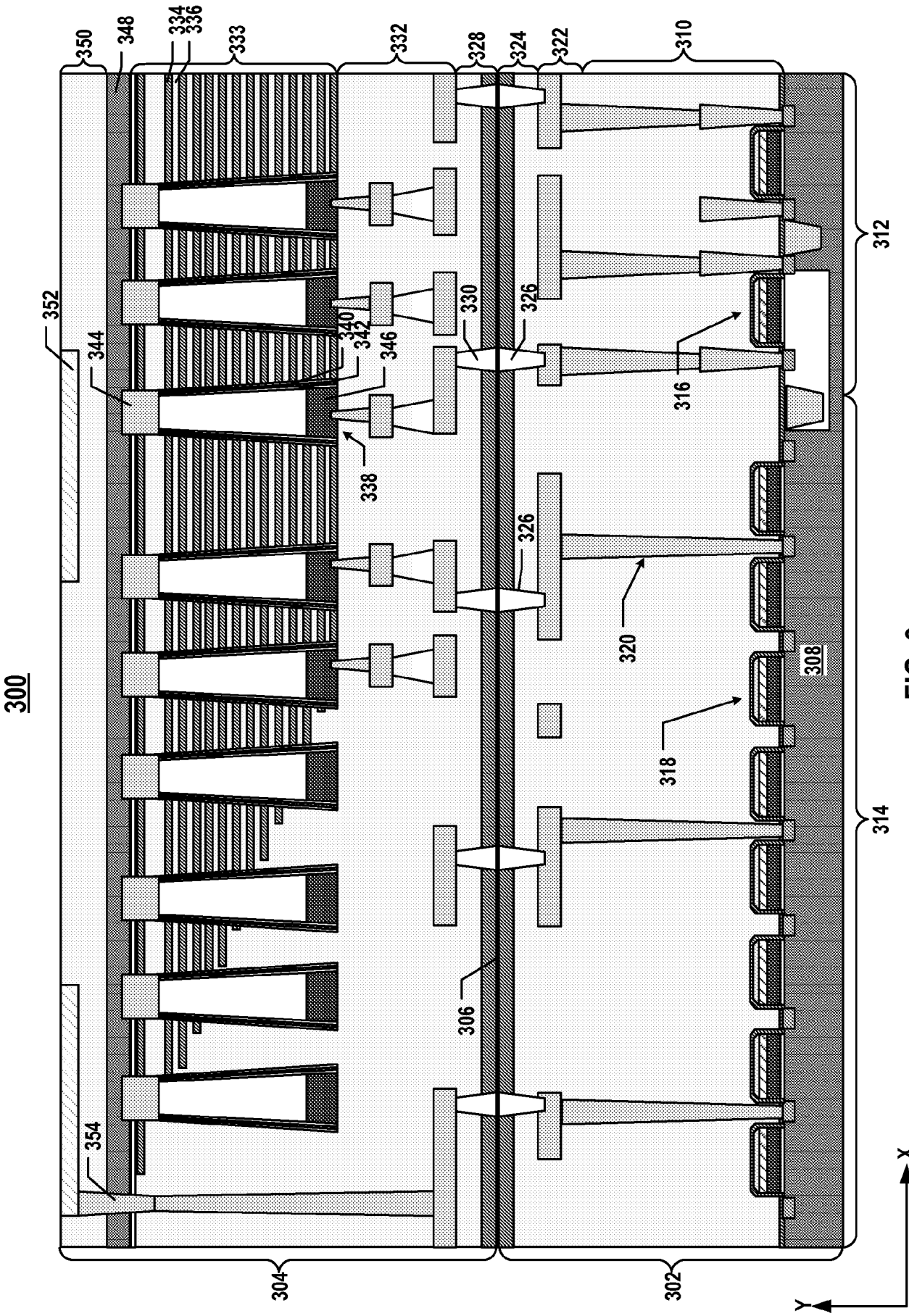
300
350
348
334
336
333
332
328
324
322
310
312
314
352
344
340
342
346
338
330
326
316
320
318
308
306
354
304
302
X
Y

FIG. 3

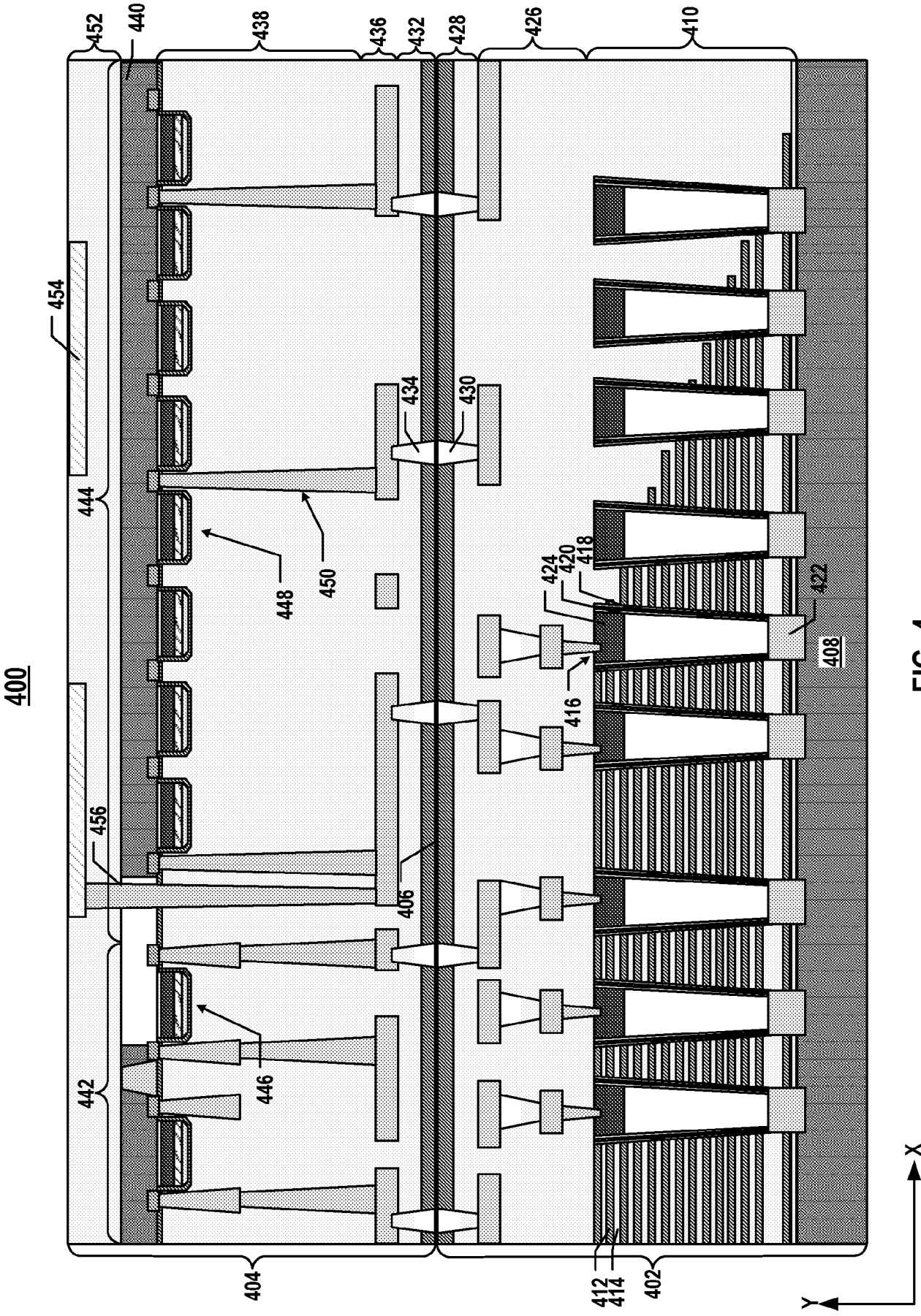
400
452
440
438
436
432
428
426
410
454
444
434
430
448
450
424
420
418
422
408
416
456
406
442
446
404
412
414
402
X
Y

FIG. 4

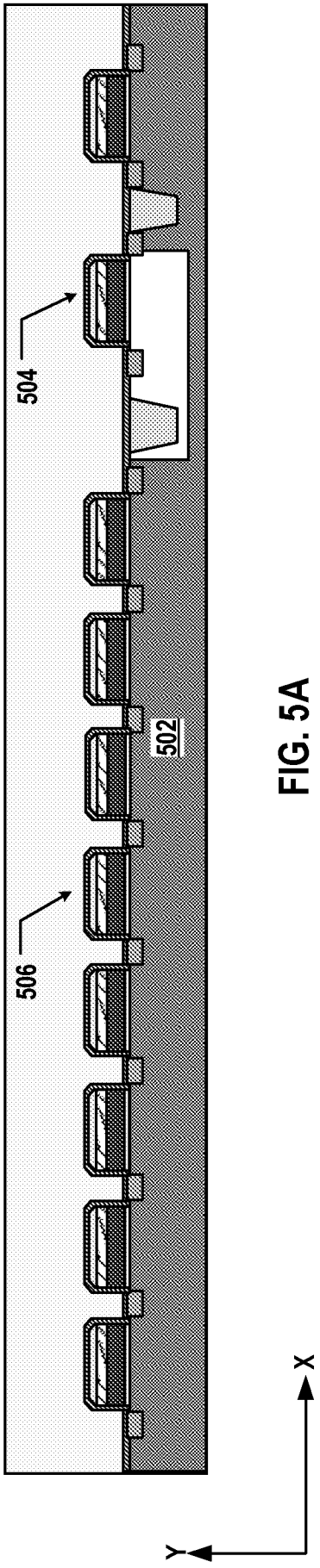

FIG. 5A

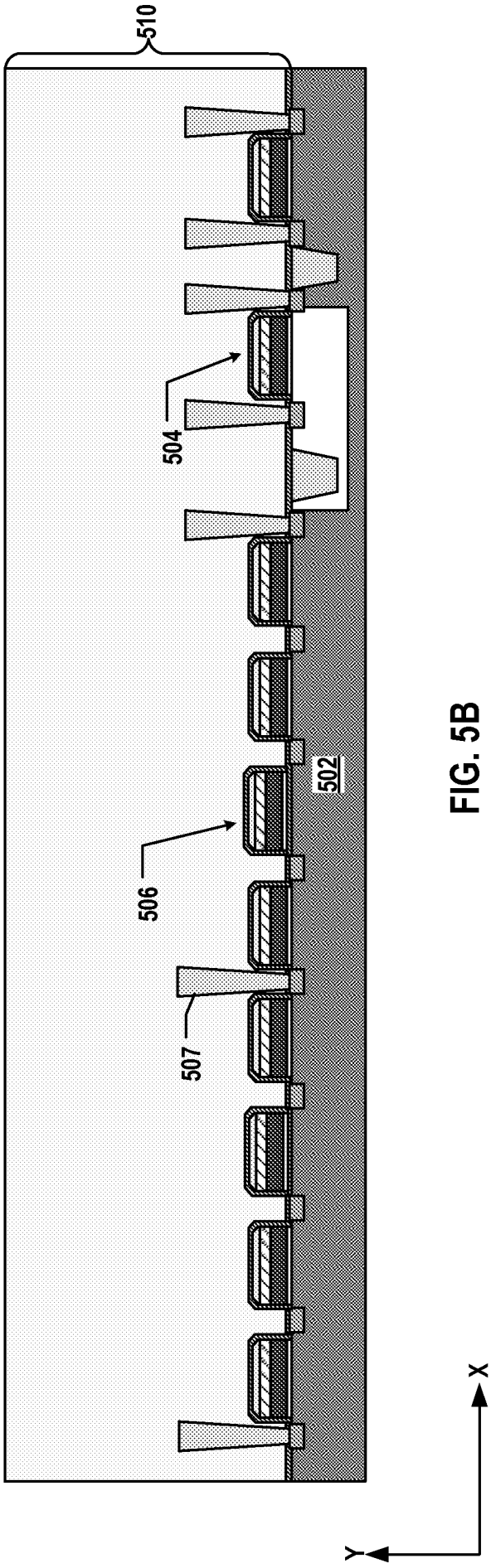

FIG. 5B

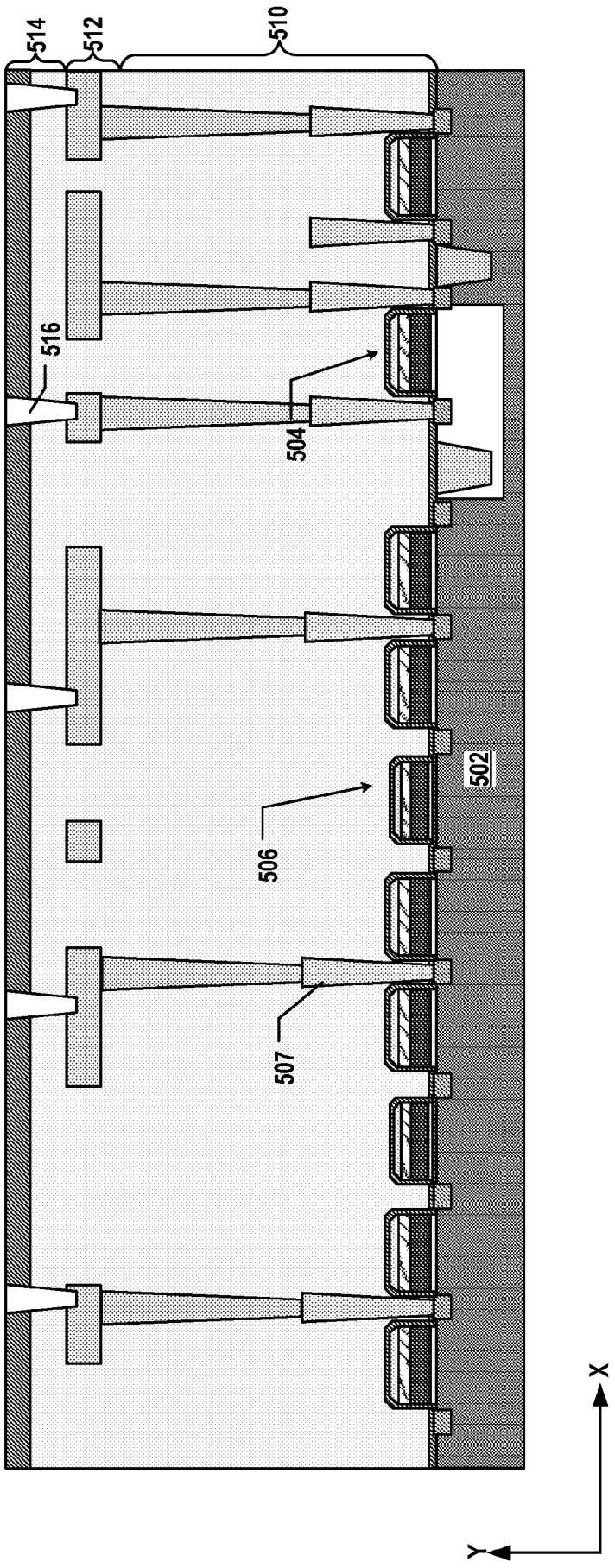

FIG. 5C

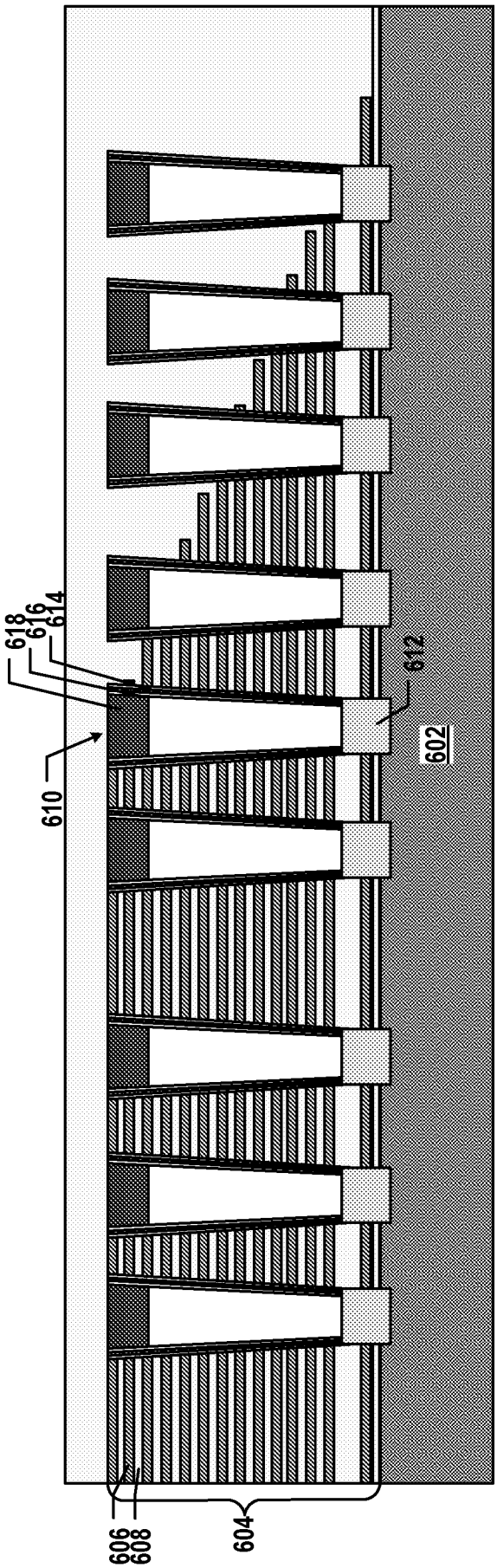

FIG. 6A

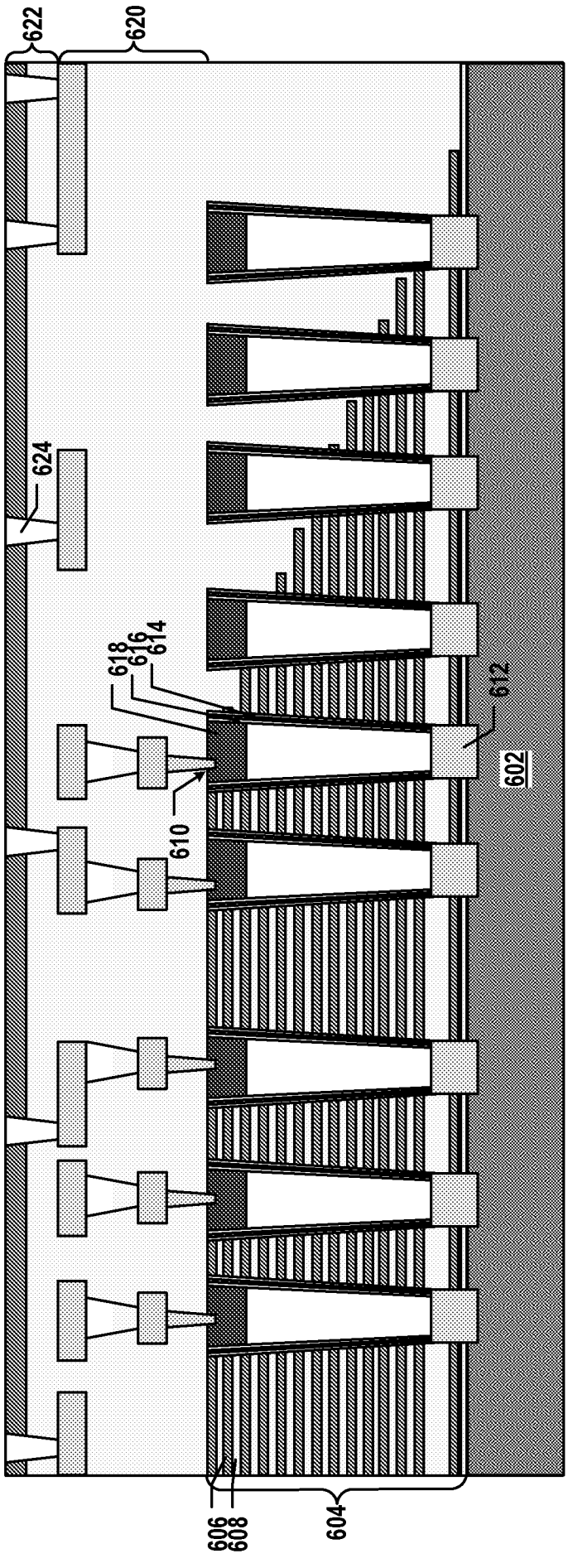

FIG. 6B

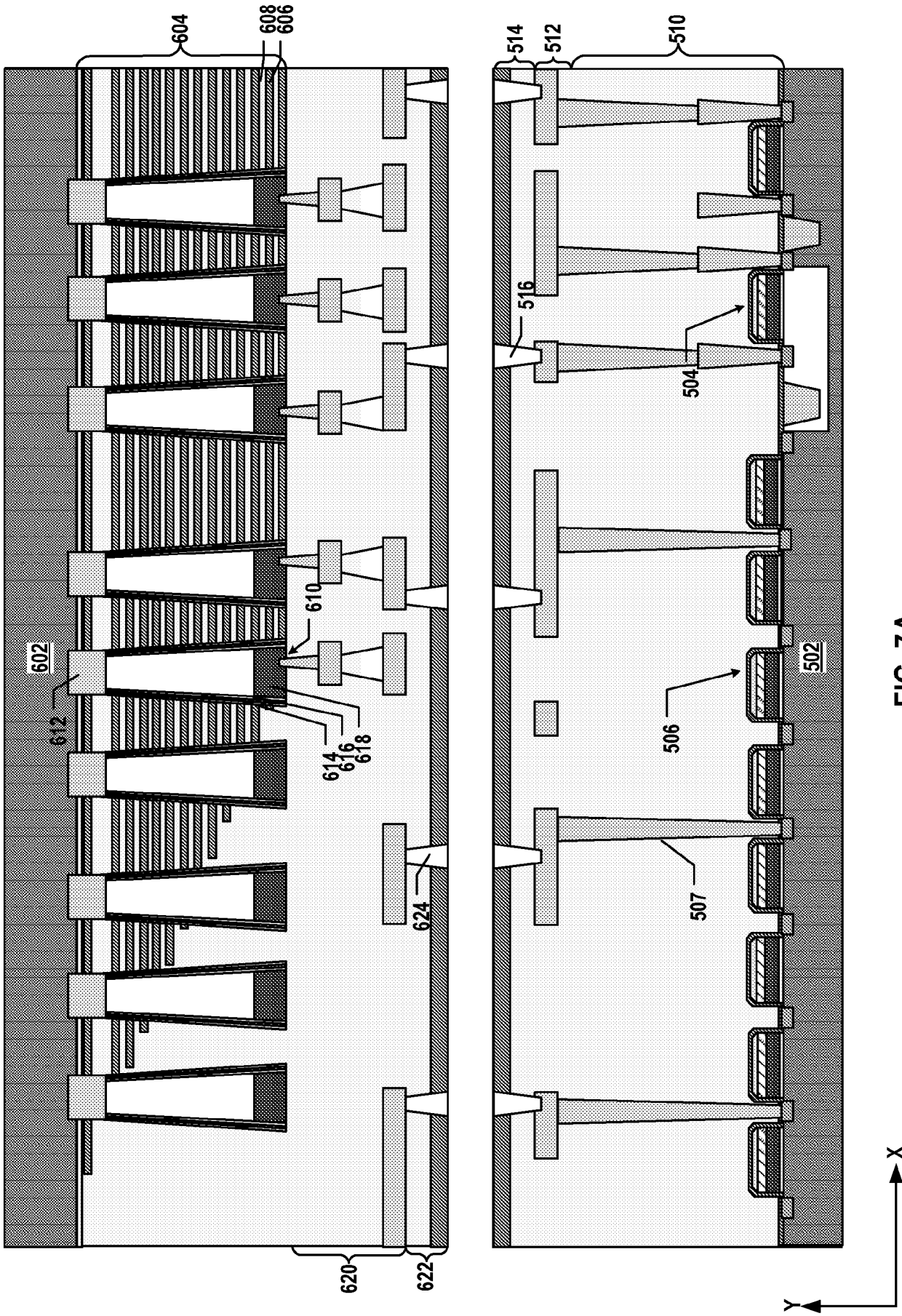
604
608
606
514
512
510
516
504
610
602
612
614
616
618
624
506
502
507
620
622
X
Y

FIG. 7A

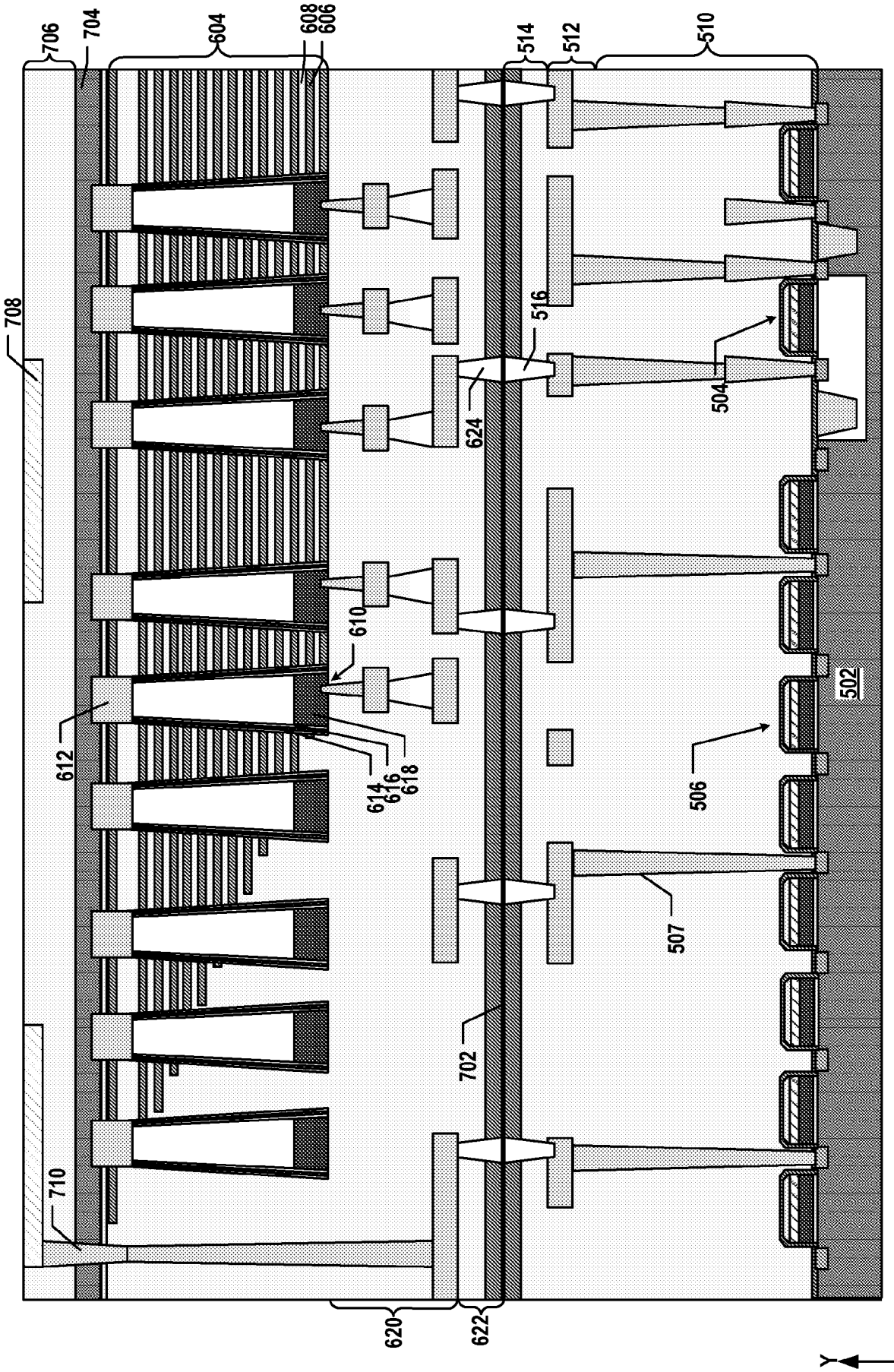
706
704
604
608
606
514
512
510
708
516
504
624
610
612
614
616
618
502
506
507
702
710
620
622
X
Y

FIG. 7B

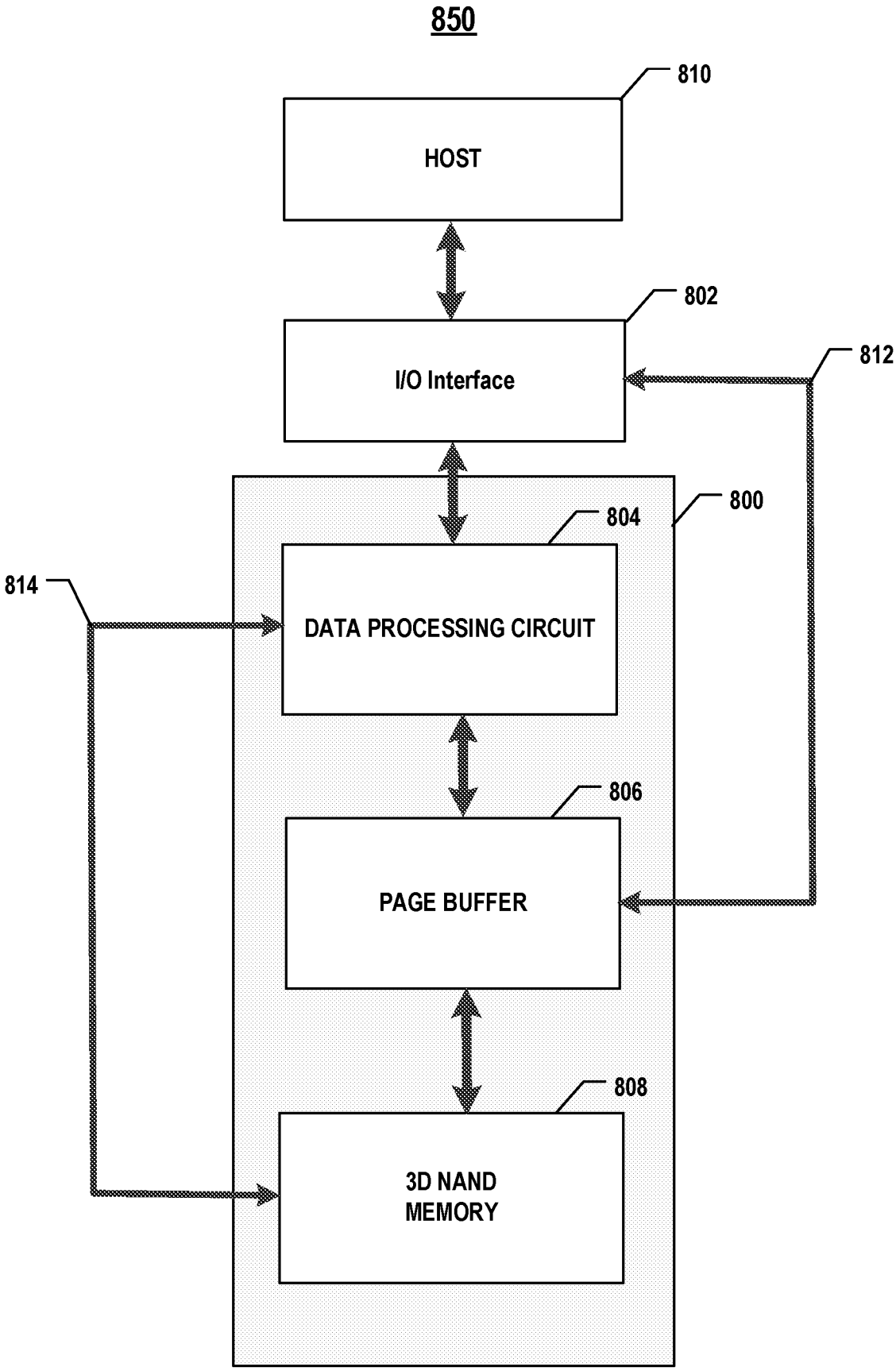
850
810
HOST
802
I/O Interface
812
800
804
814
DATA PROCESSING CIRCUIT
806
PAGE BUFFER
808
3D NAND
MEMORY

FIG. 8

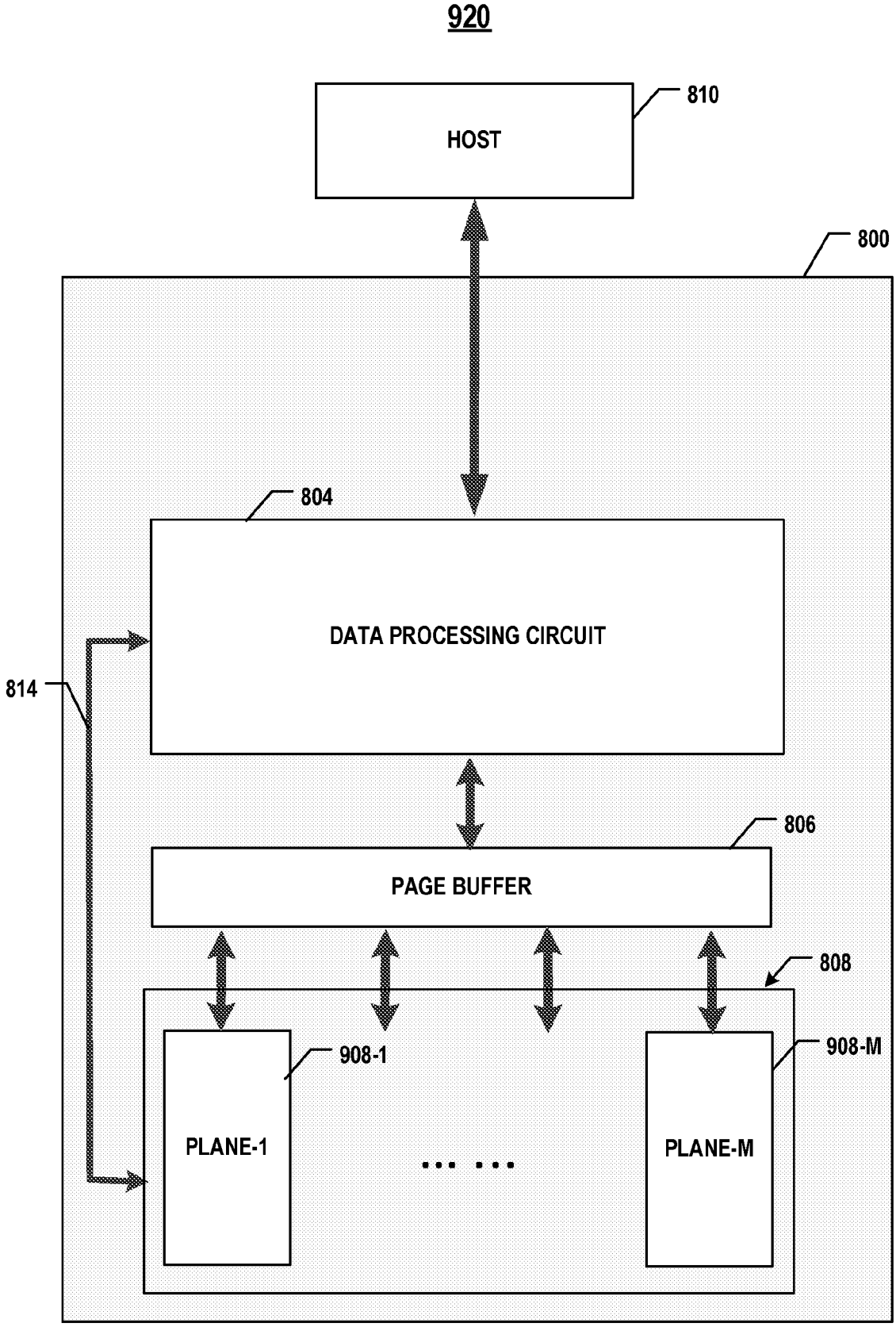
920
HOST
810
800
804
DATA PROCESSING CIRCUIT
814
806
PAGE BUFFER
808
908-1
908-M
PLANE-1
PLANE-M

FIG. 9

1000

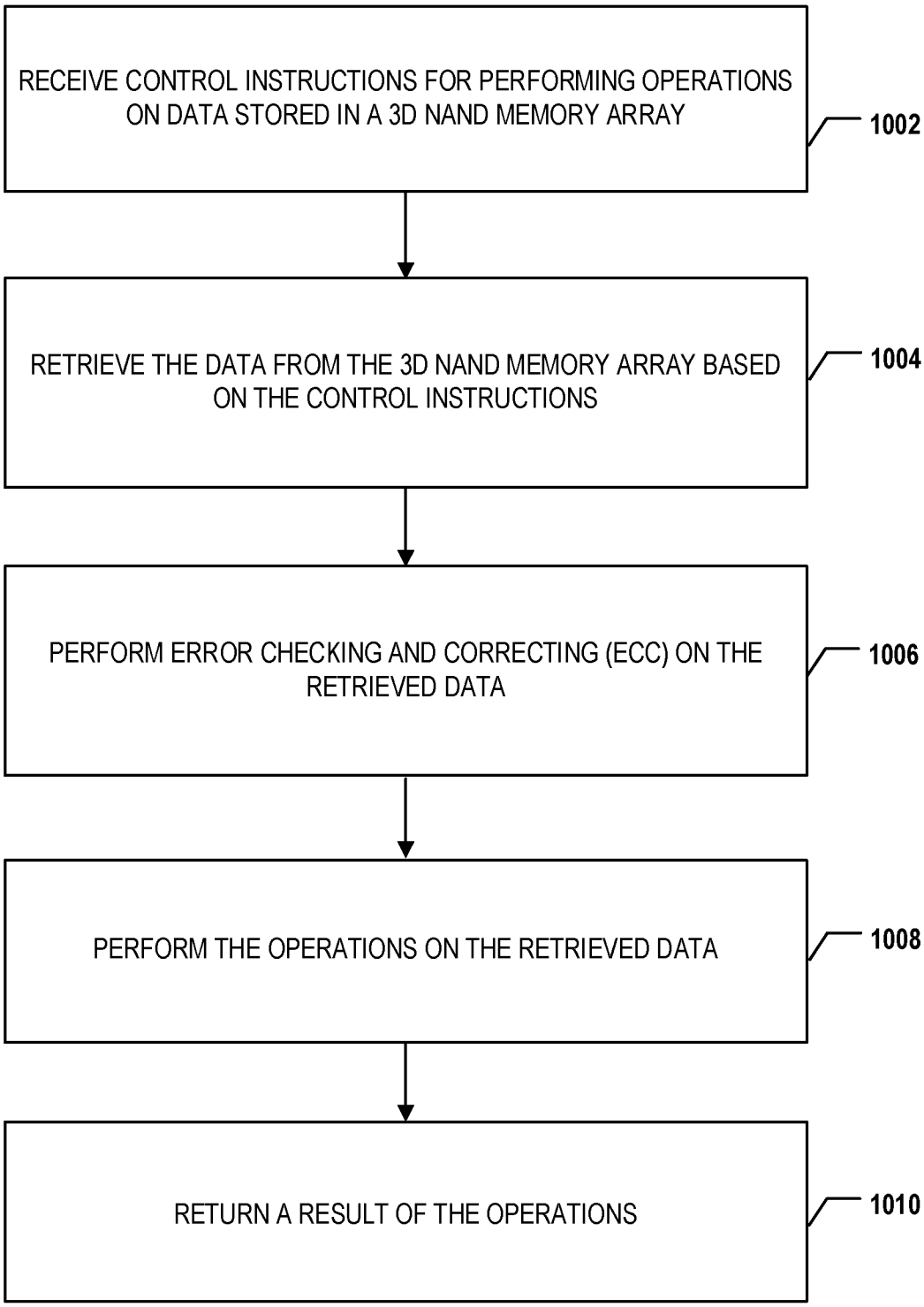
RECEIVE CONTROL INSTRUCTIONS FOR PERFORMING OPERATIONS ON DATA STORED IN A 3D NAND MEMORY ARRAY
1002
RETRIEVE THE DATA FROM THE 3D NAND MEMORY ARRAY BASED ON THE CONTROL INSTRUCTIONS
1004
PERFORM ERROR CHECKING AND CORRECTING (ECC) ON THE RETRIEVED DATA
1006
PERFORM THE OPERATIONS ON THE RETRIEVED DATA
1008
RETURN A RESULT OF THE OPERATIONS
1010

FIG. 10

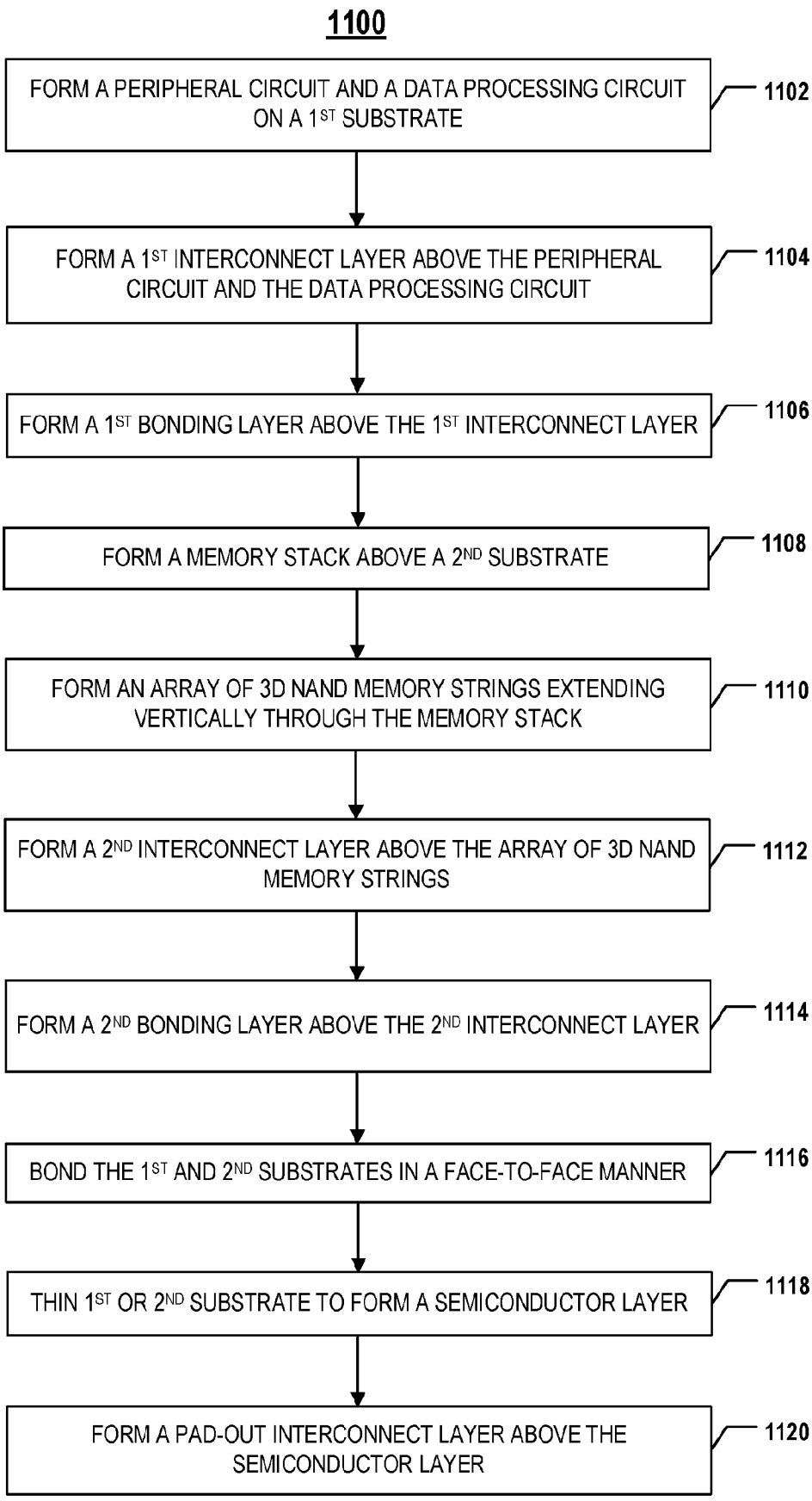
1100
FORM A PERIPHERAL CIRCUIT AND A DATA PROCESSING CIRCUIT ON A 1ST SUBSTRATE
1102
FORM A 1ST INTERCONNECT LAYER ABOVE THE PERIPHERAL CIRCUIT AND THE DATA PROCESSING CIRCUIT
1104
FORM A 1ST BONDING LAYER ABOVE THE 1ST INTERCONNECT LAYER
1106
FORM A MEMORY STACK ABOVE A 2ND SUBSTRATE
1108
FORM AN ARRAY OF 3D NAND MEMORY STRINGS EXTENDING VERTICALLY THROUGH THE MEMORY STACK
1110
FORM A 2ND INTERCONNECT LAYER ABOVE THE ARRAY OF 3D NAND MEMORY STRINGS
1112
FORM A 2ND BONDING LAYER ABOVE THE 2ND INTERCONNECT LAYER
1114
BOND THE 1ST AND 2ND SUBSTRATES IN A FACE-TO-FACE MANNER
1116
THIN 1ST OR 2ND SUBSTRATE TO FORM A SEMICONDUCTOR LAYER
1118
FORM A PAD-OUT INTERCONNECT LAYER ABOVE THE SEMICONDUCTOR LAYER
1120

FIG. 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/093477

A. CLASSIFICATION OF SUBJECT MATTER

G11C 5/02(2006.01)i

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)

G11C 5/-

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS;SIPOABS;DWPI:data, memory, stor+, array, die, chip, same, interface, input, output, periphery, three dimension, 3D, retrieve

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2017255834 A1 (ZHANG GUOBIAOet al.) 07 September 2017 (2017-09-07) description paragraphs 13, 17, 25, 26, 34	1-24
Y	US 2017170845 A1 (CHO YONGWONet al.) 15 June 2017 (2017-06-15) description paragraphs 7, 29-36, 42, claim 8	1-24
Y	CN 109145683 A (CHENGDU HAICUN IP TECHNOLOGY LLC) 04 January 2019 (2019-01-04) description paragraphs 8, 9, 25	1-24
A	US 2017308299 A1 (SAMSUNG ELECTRONICS CO LTD) 26 October 2017 (2017-10-26) whole document	1-24
A	CN 108172255 A (SHANGHAI XINCHU INTEGRATED CIRCUIT CO LTD) 15 June 2018 (2018-06-15) whole document	1-24

☐ 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
"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
11 November 2019	**07 January 2020**
Name and mailing address of the ISA/CN **National Intellectual Property Administration, PRC** **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer **LI,Yandong** Telephone No. **62411735**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/093477

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US 2017255834 A1	07 September 2017	US 2019171815 A1	06 June 2019
		CN 107169404 A	15 September 2017
		WO 2017152828 A1	14 September 2017
US 2017170845 A1	15 June 2017	US 2019036546 A1	31 January 2019
		US 10090858 B2	02 October 2018
		CN 107045462 A	15 August 2017
		KR 20170070921 A	23 June 2017
CN 109145683 A	04 January 2019	None	
US 2017308299 A1	26 October 2017	KR 20170121798 A	03 November 2017
		US 10156995 B2	18 December 2018
		CN 107393596 A	24 November 2017
CN 108172255 A	15 June 2018	None	