



US 20250266099A1

(19) **United States**

(12) **Patent Application Publication**
Prakash et al.

(10) **Pub. No.: US 2025/0266099 A1**

(43) **Pub. Date: Aug. 21, 2025**

(54) **METHODS TO IMPROVE CURRENT CONSUMPTION AND READ TIME IN SUCCESSIVE READS**

(52) **U.S. Cl.**
CPC *G11C 16/26* (2013.01); *G11C 16/08* (2013.01); *G11C 16/32* (2013.01)

(71) Applicant: **Western Digital Technologies, Inc.,**
San Jose, CA (US)

(57) **ABSTRACT**

(72) Inventors: **Abhijith Prakash**, Milpitas, CA (US);
Xiang Yang, Santa Clara, CA (US)

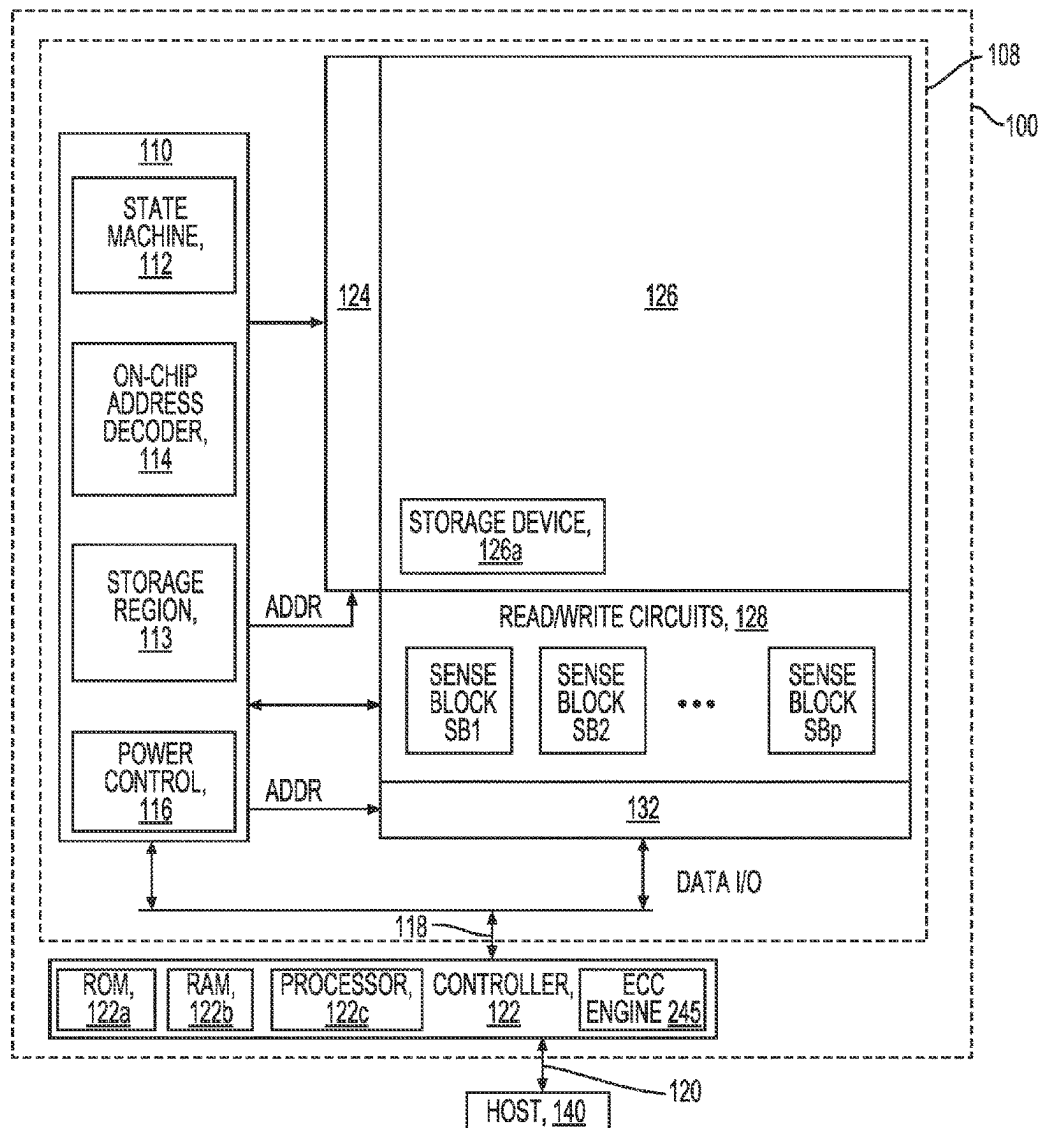
A memory apparatus includes memory cells each connected to word lines and configured to store a threshold voltage corresponding to data states. The memory apparatus also includes a control means configured to apply at least one read voltage associated with one of the data states to ones of the word lines connected to the memory cells being read in one read operation. The control means is also configured to adjust a voltage applied to the ones of the word lines during at least one of an end of the one read operation or a beginning of a subsequent read operation of the memory cells based on at least one of the subsequent read operation following the one read operation within a predetermined time or a control gate ready voltage of the ones of the word lines targeted following the one read operation.

(21) Appl. No.: **18/443,992**

(22) Filed: **Feb. 16, 2024**

Publication Classification

(51) **Int. Cl.**
G11C 16/26 (2006.01)
G11C 16/08 (2006.01)
G11C 16/32 (2006.01)



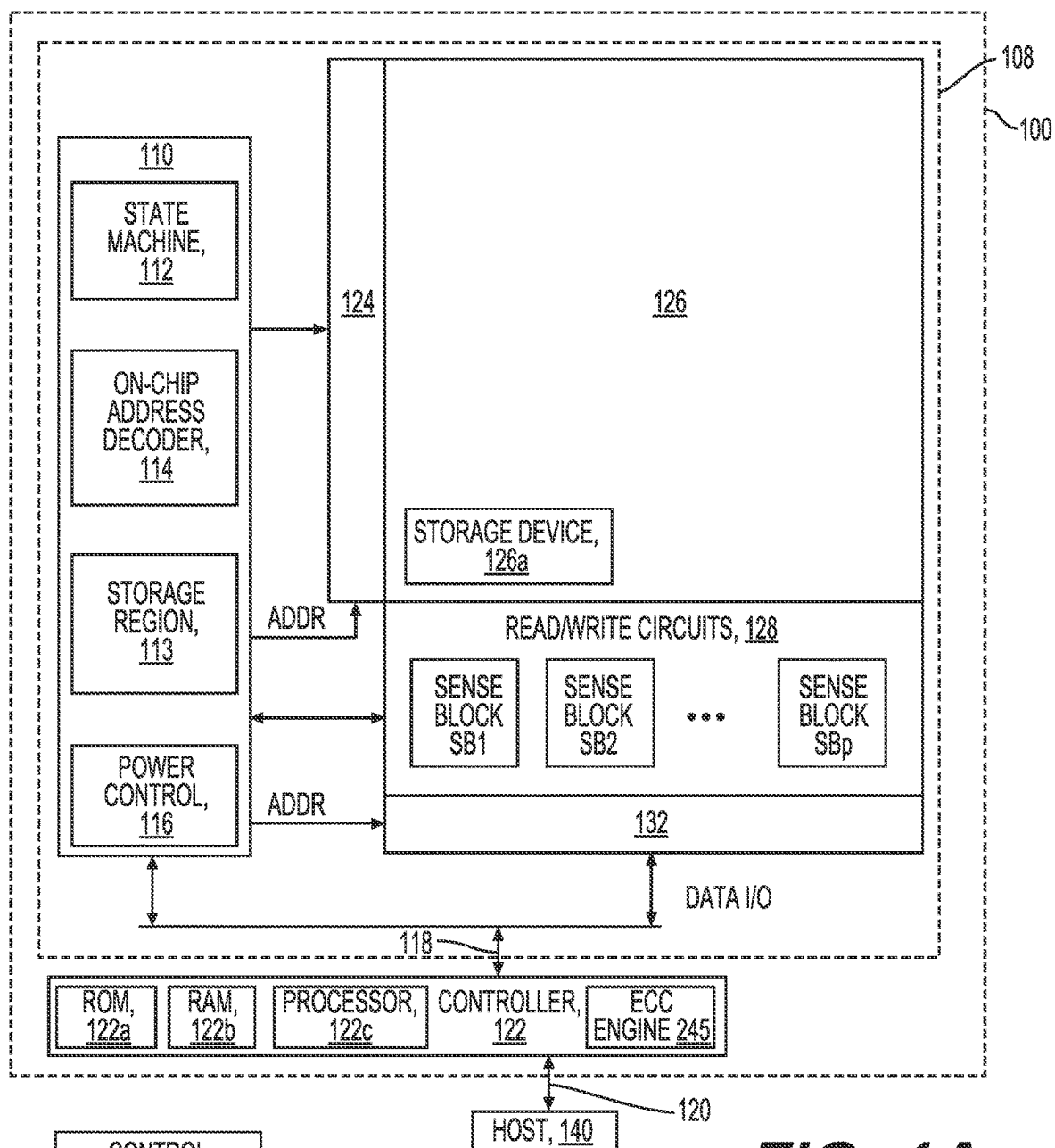


FIG. 1A

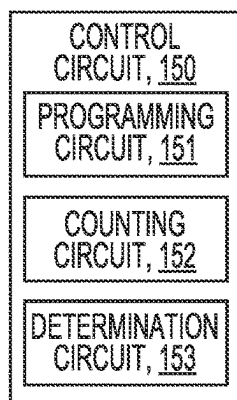
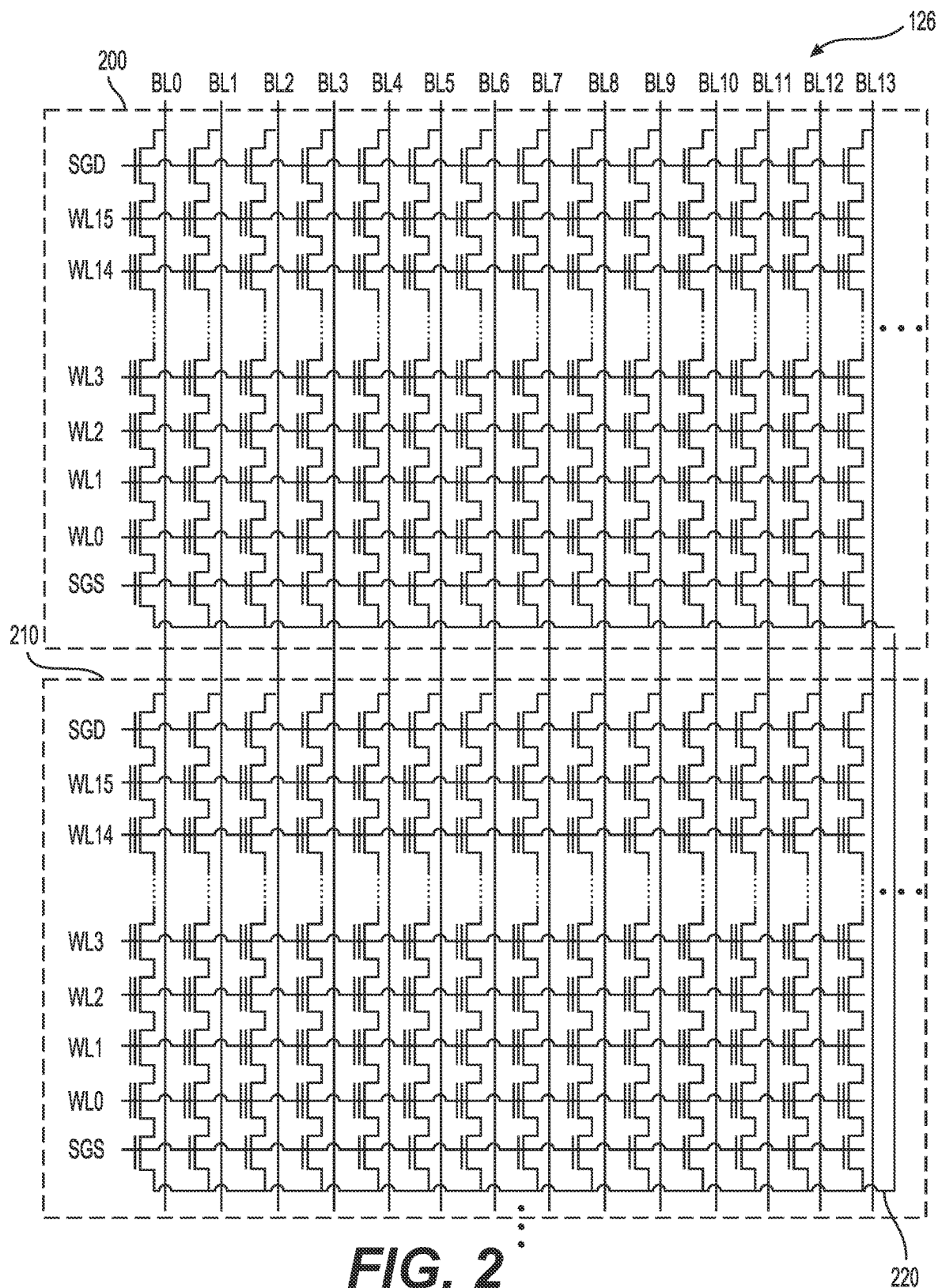


FIG. 1B



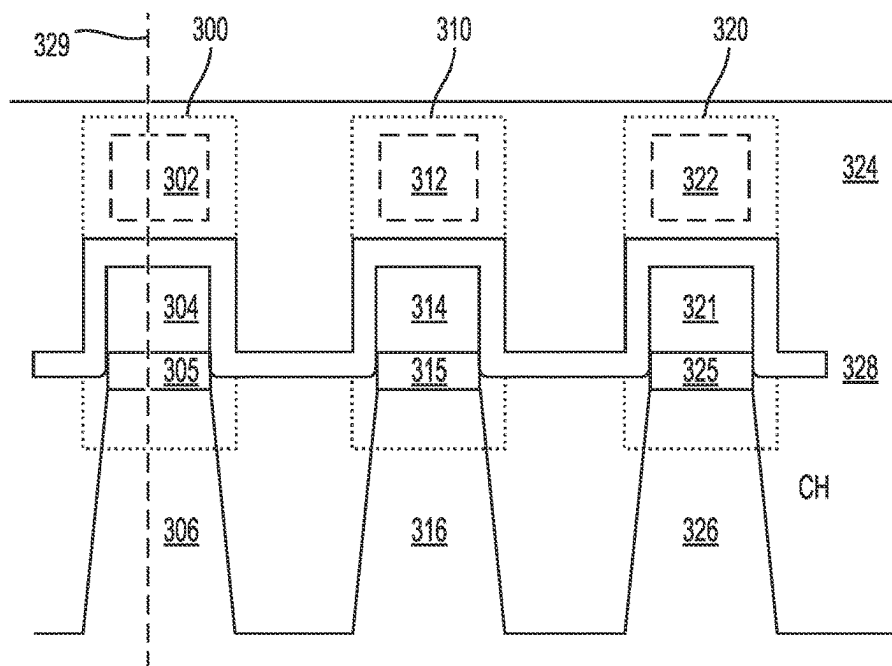


FIG. 3A

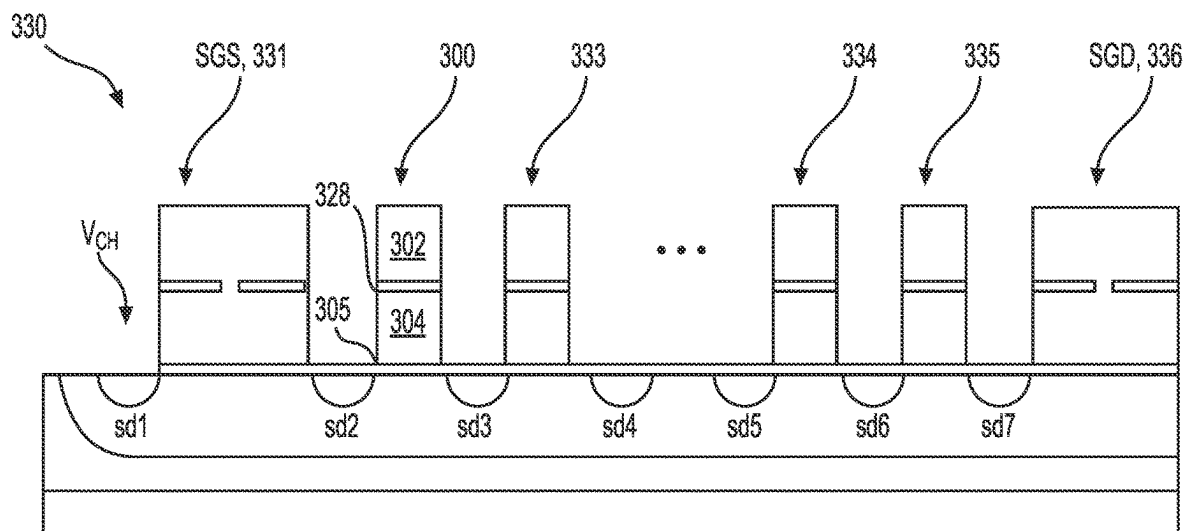


FIG. 3B

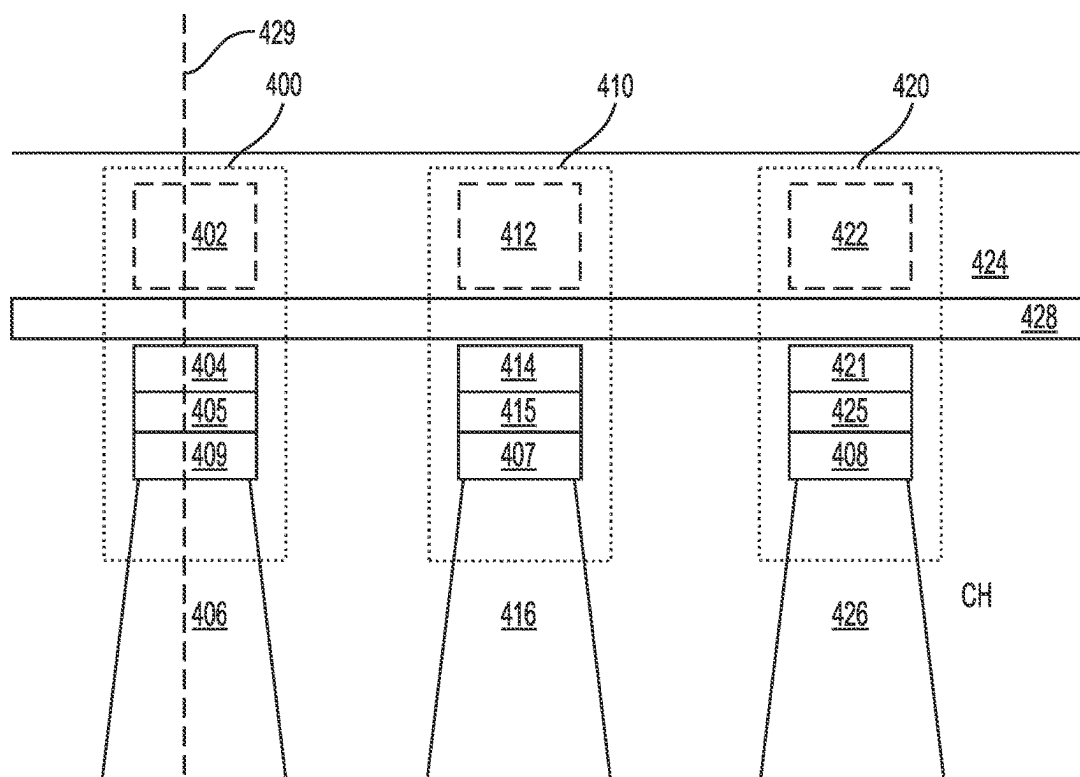


FIG. 4A

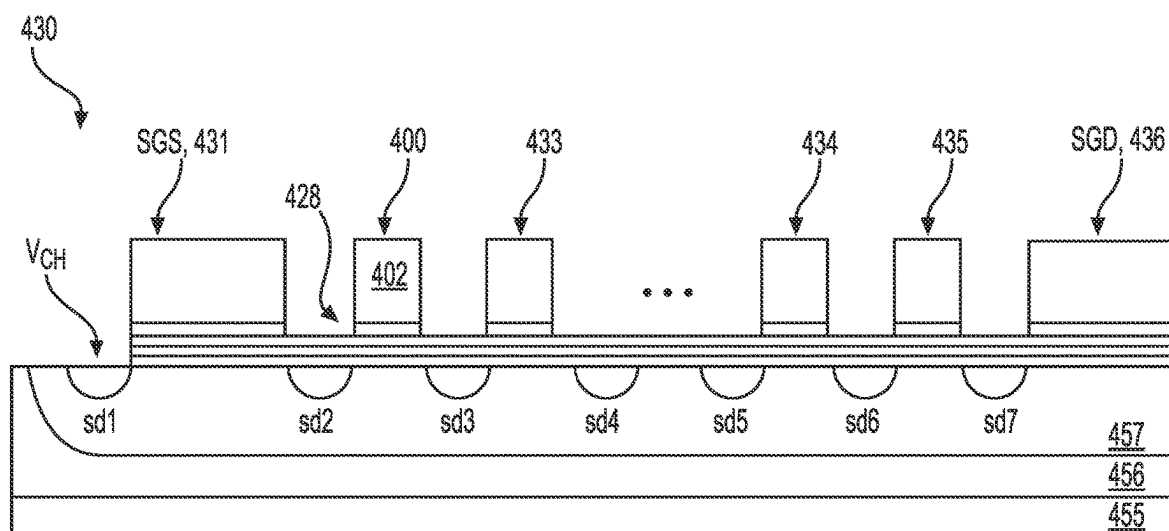


FIG. 4B

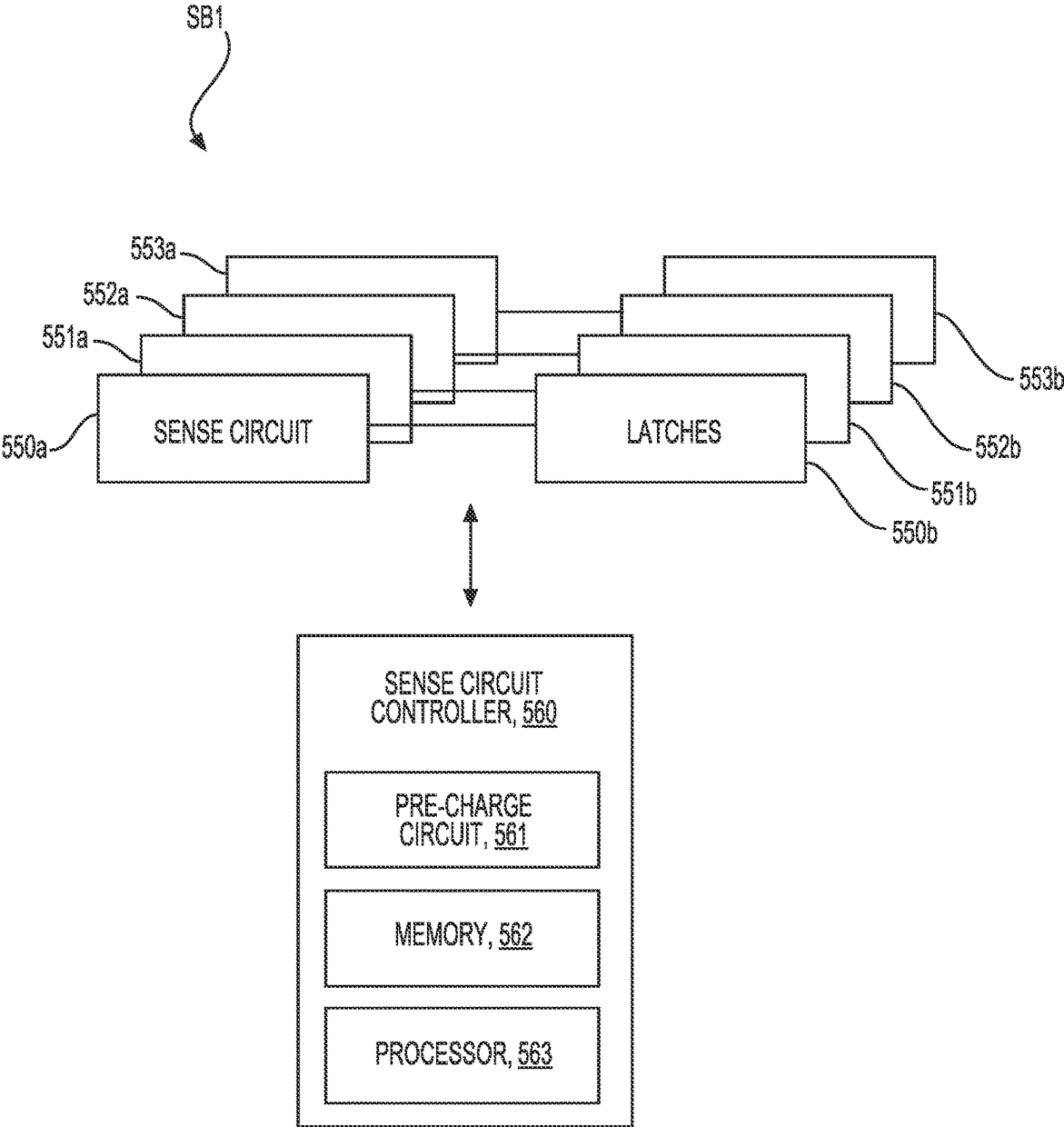


FIG. 5A

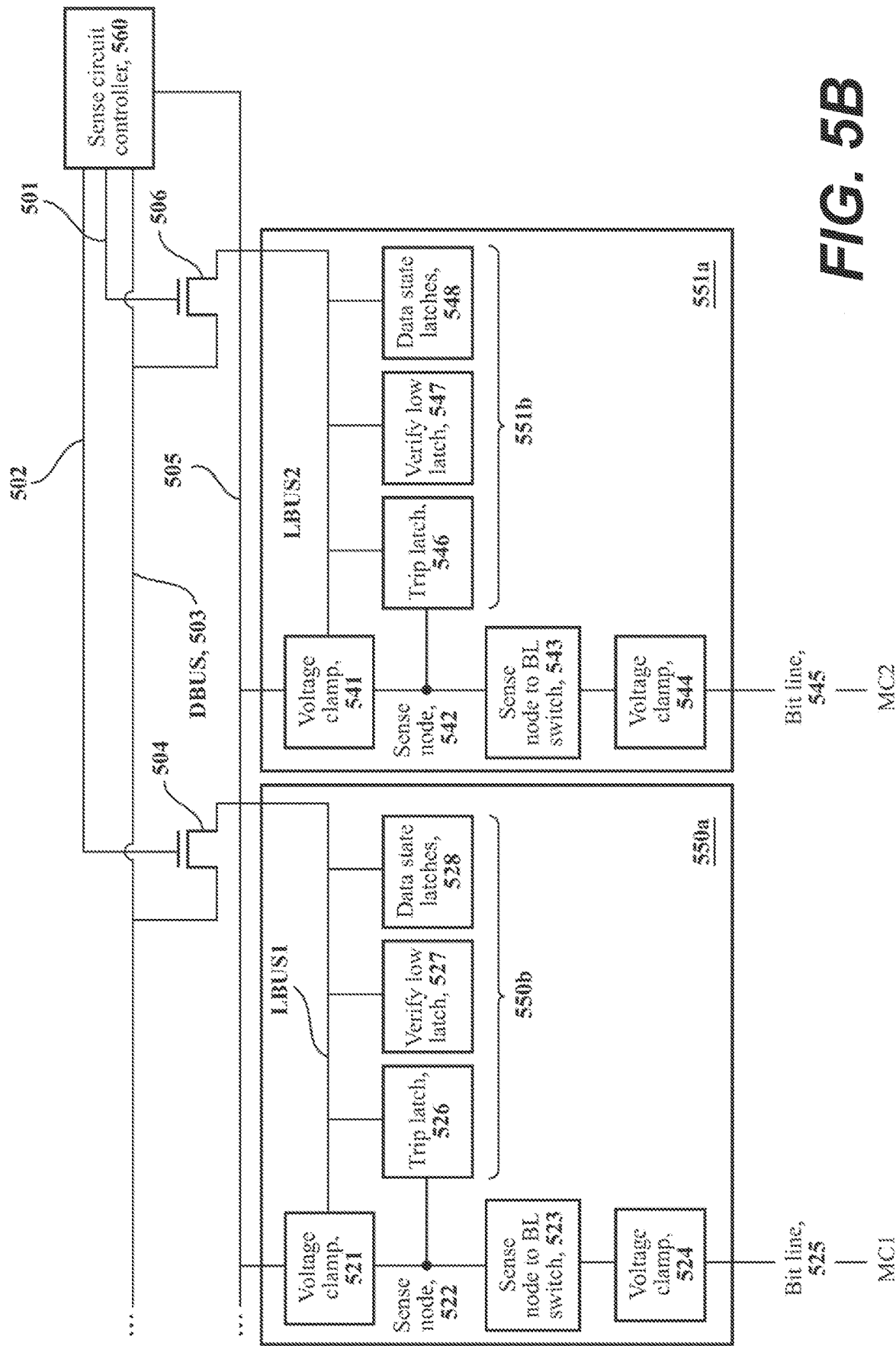


FIG. 5B

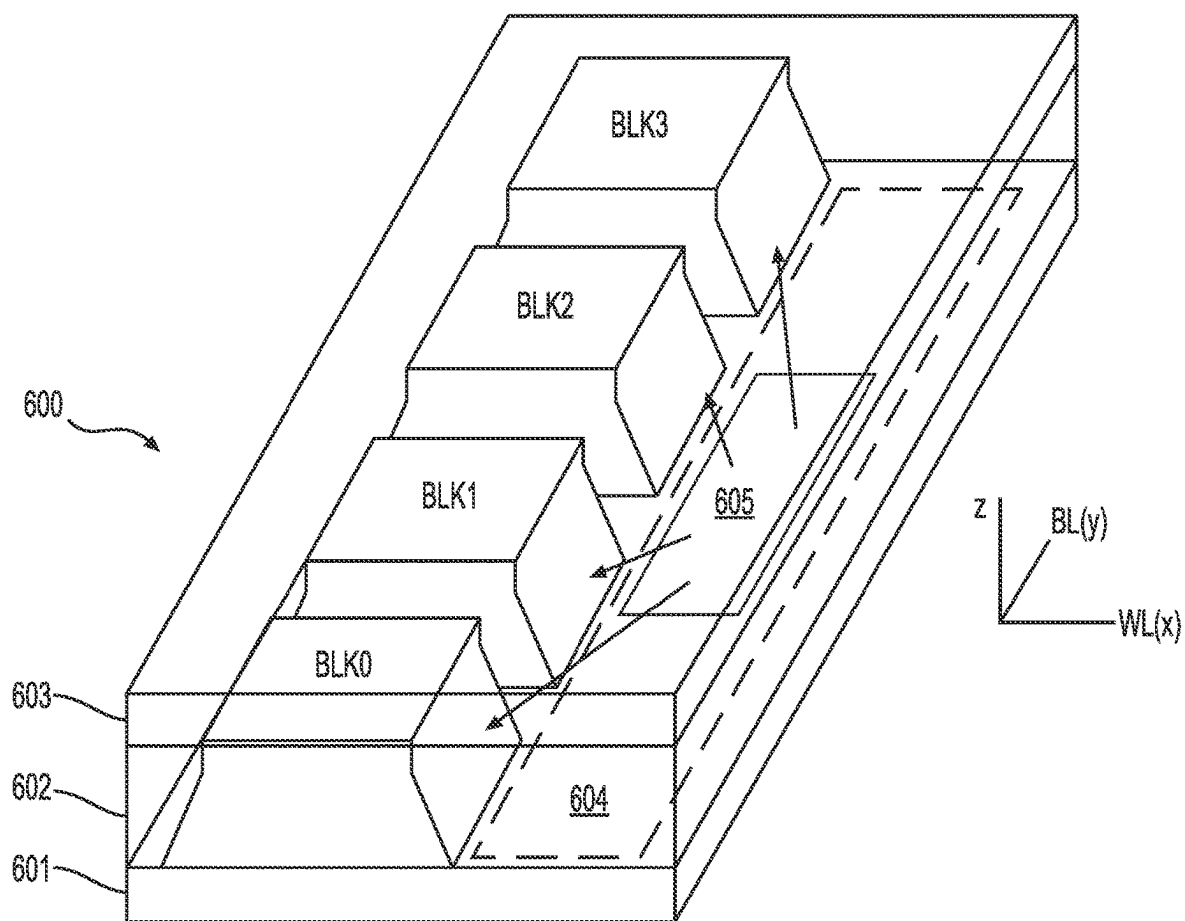


FIG. 6A

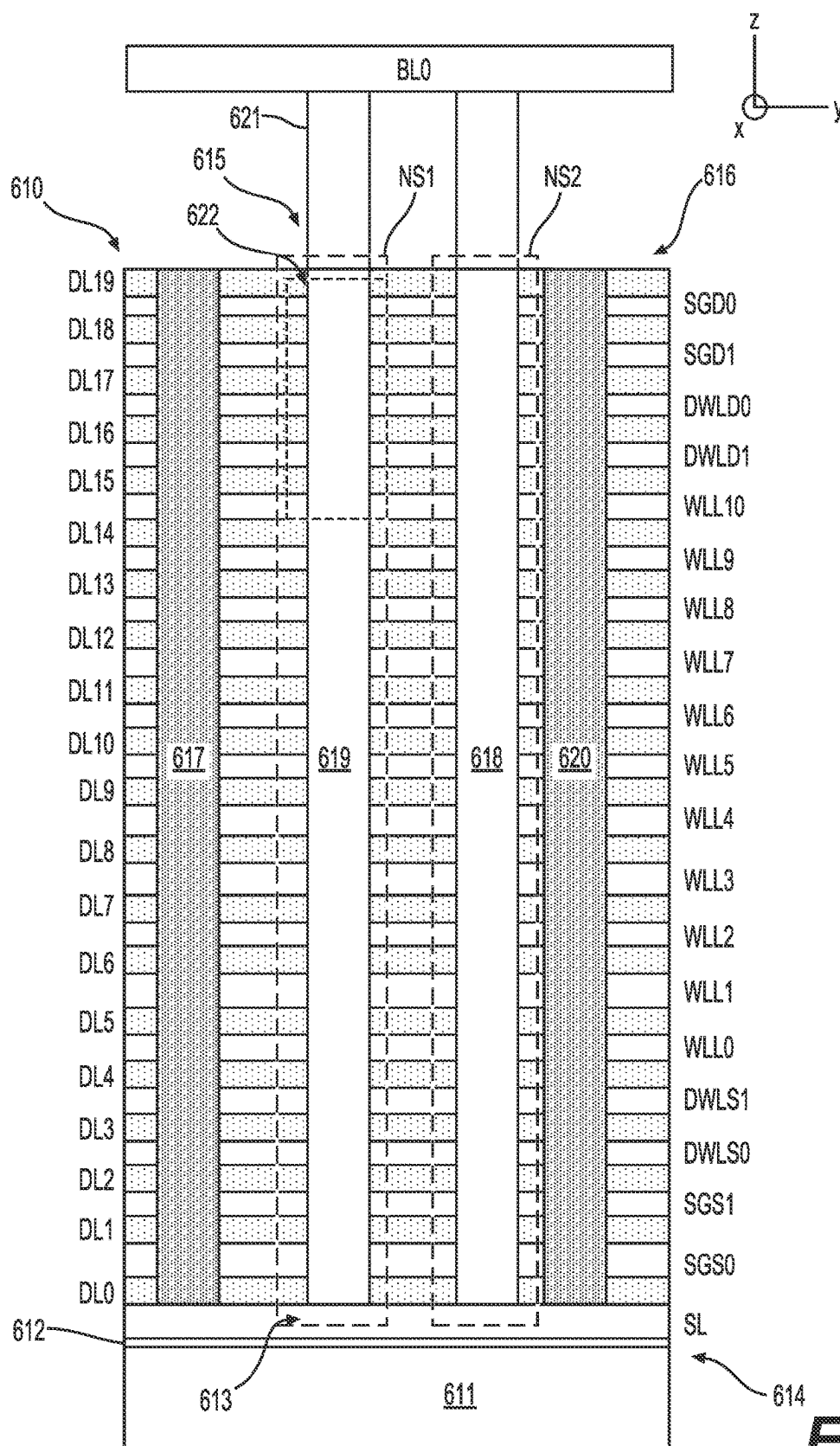


FIG. 6B

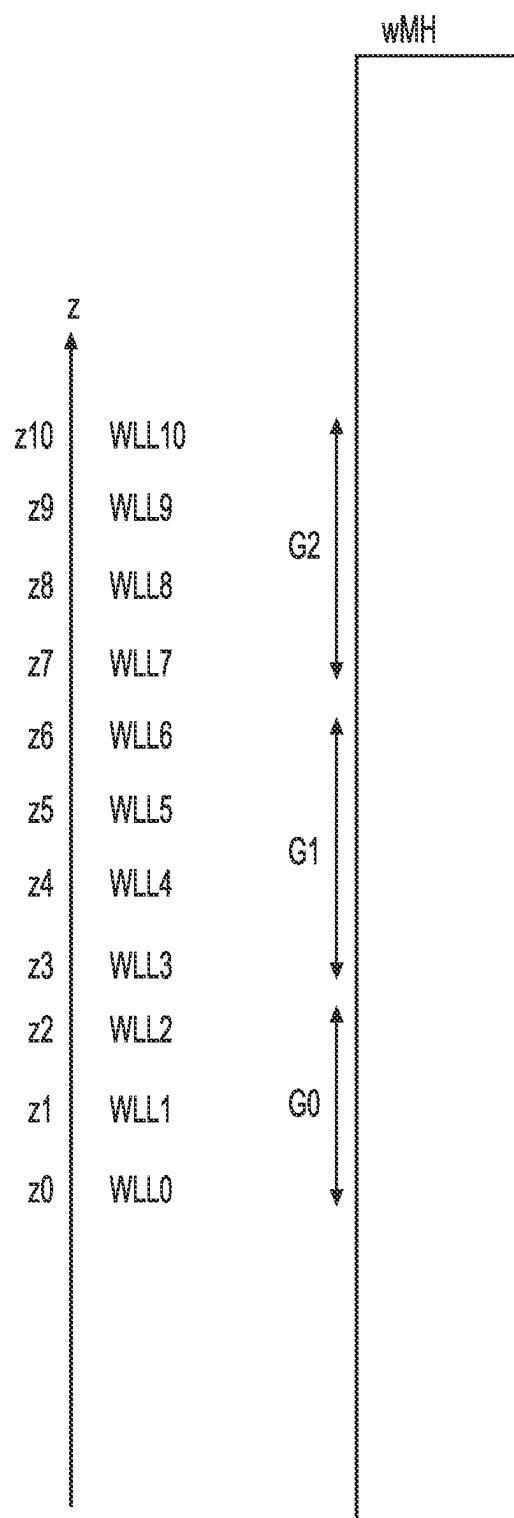


FIG. 6C

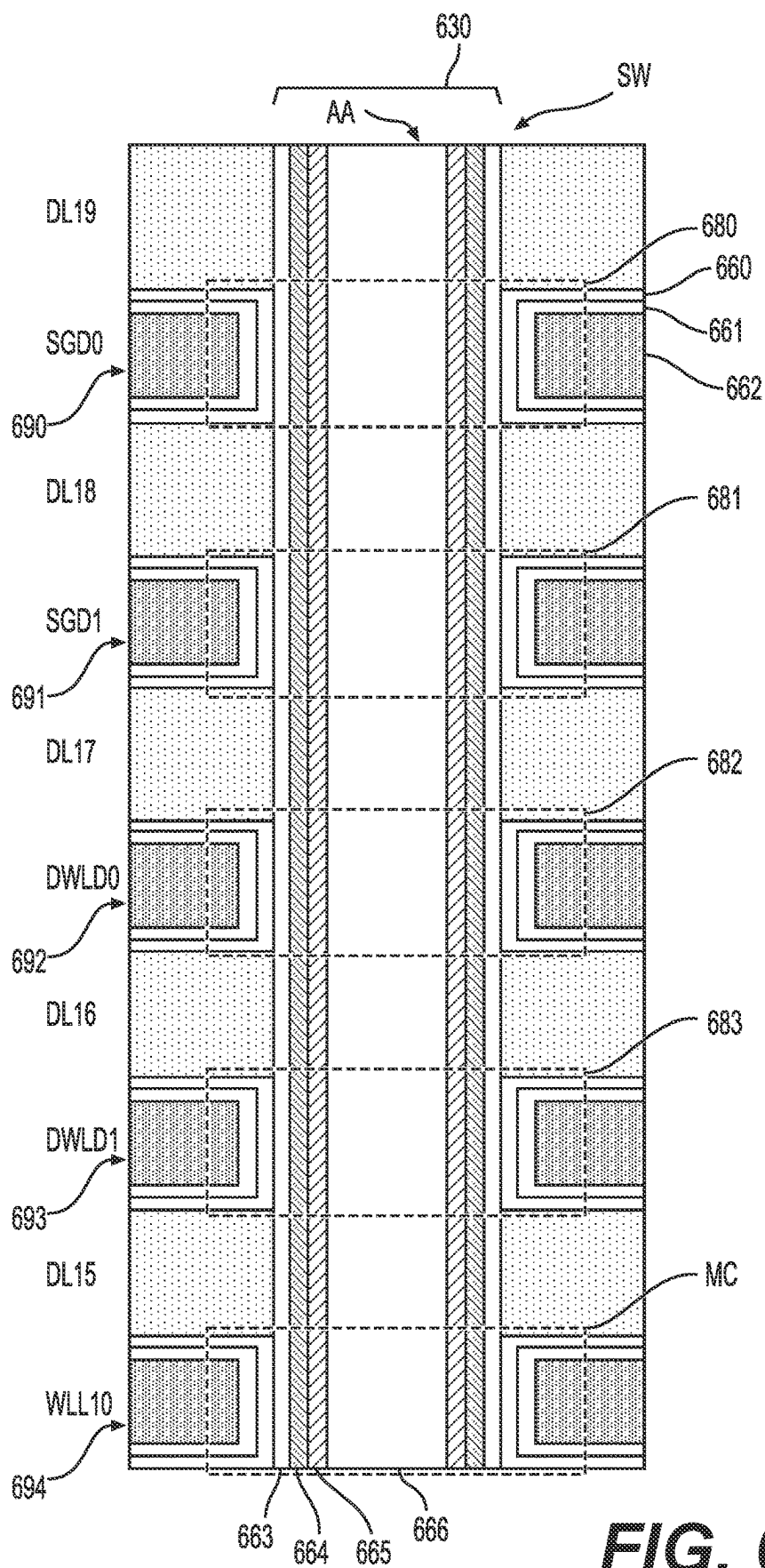


FIG. 6D

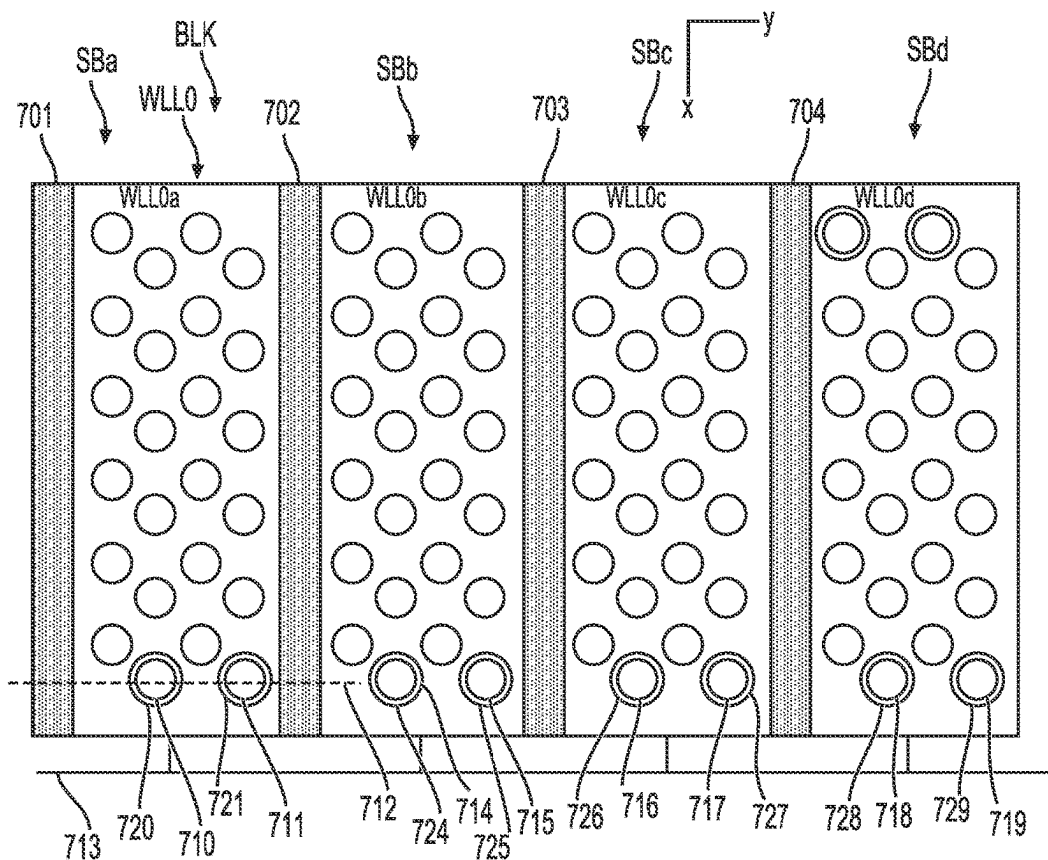


FIG. 7A

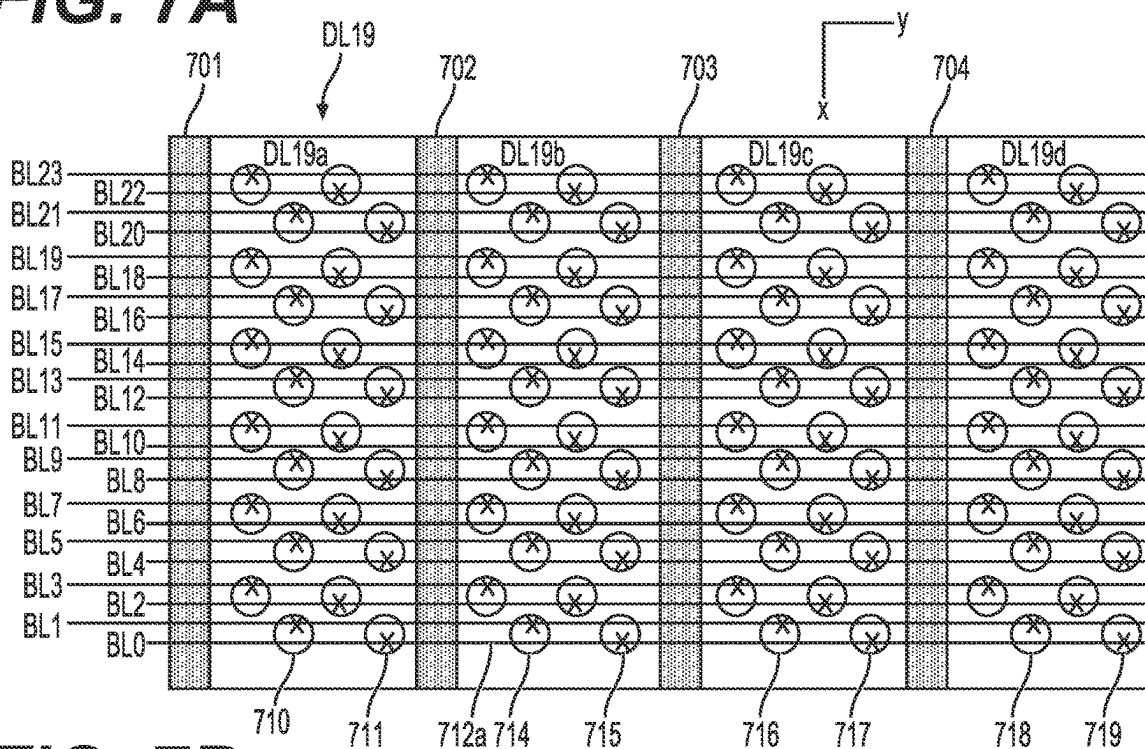


FIG. 7B

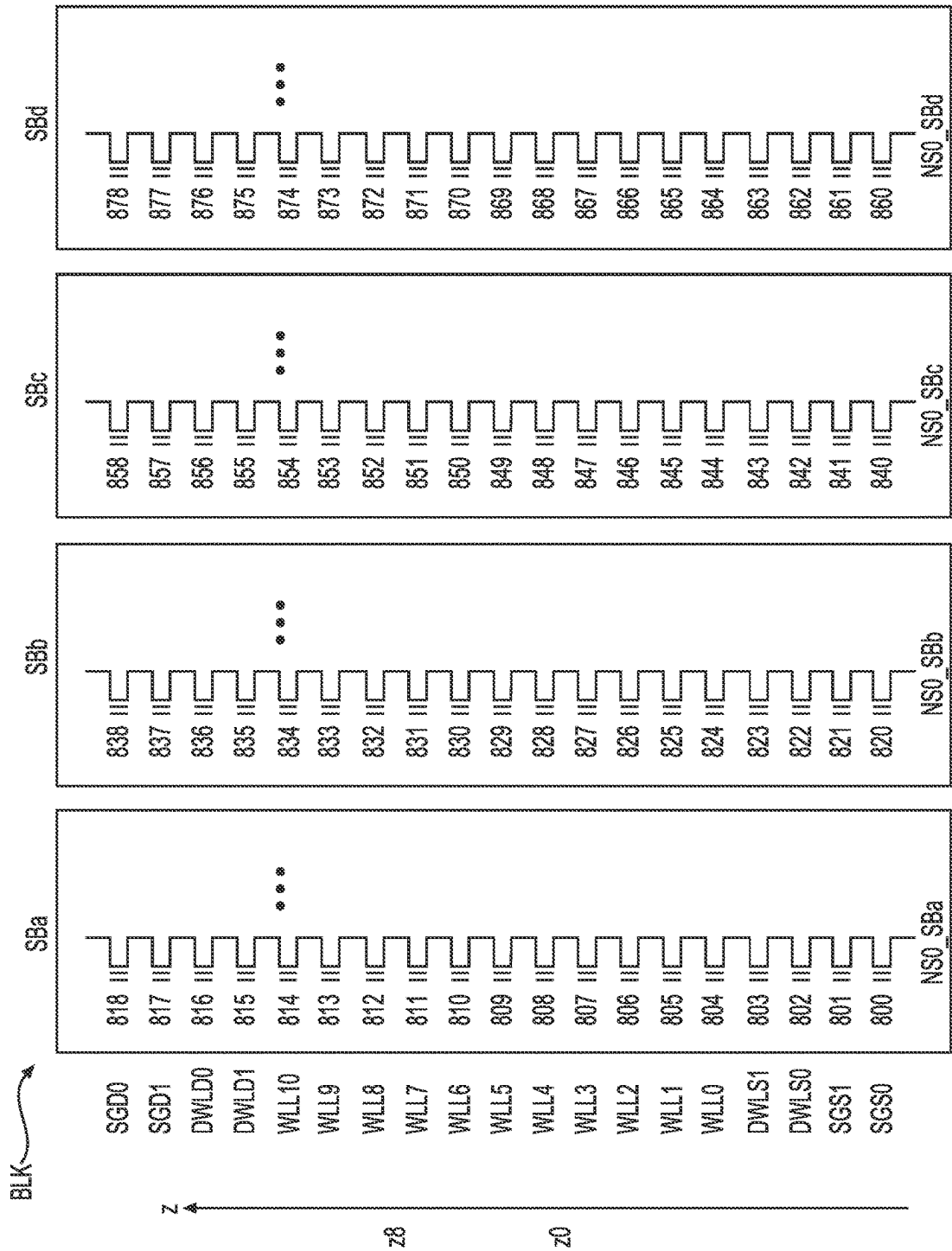


FIG. 8A

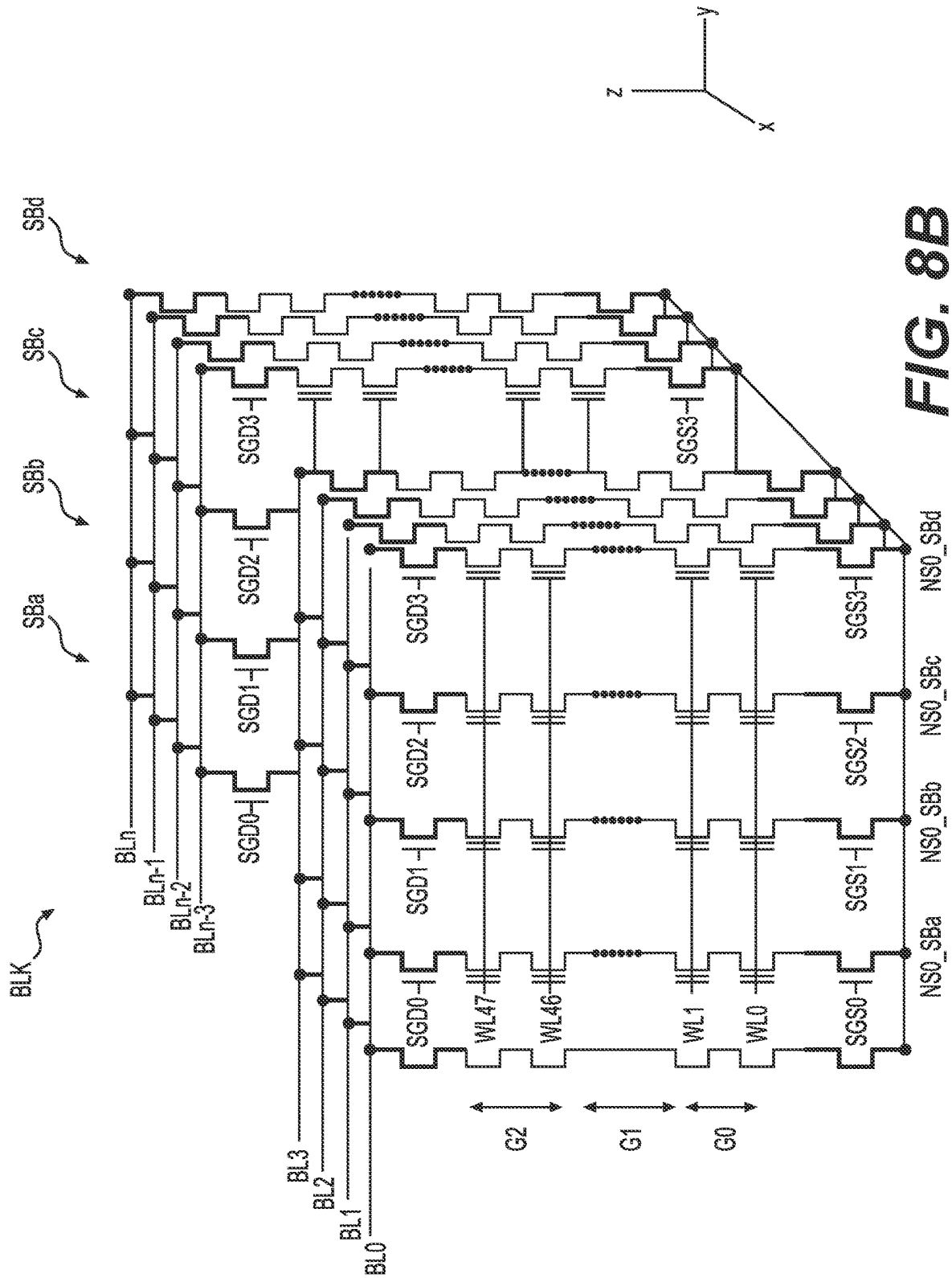


FIG. 8B

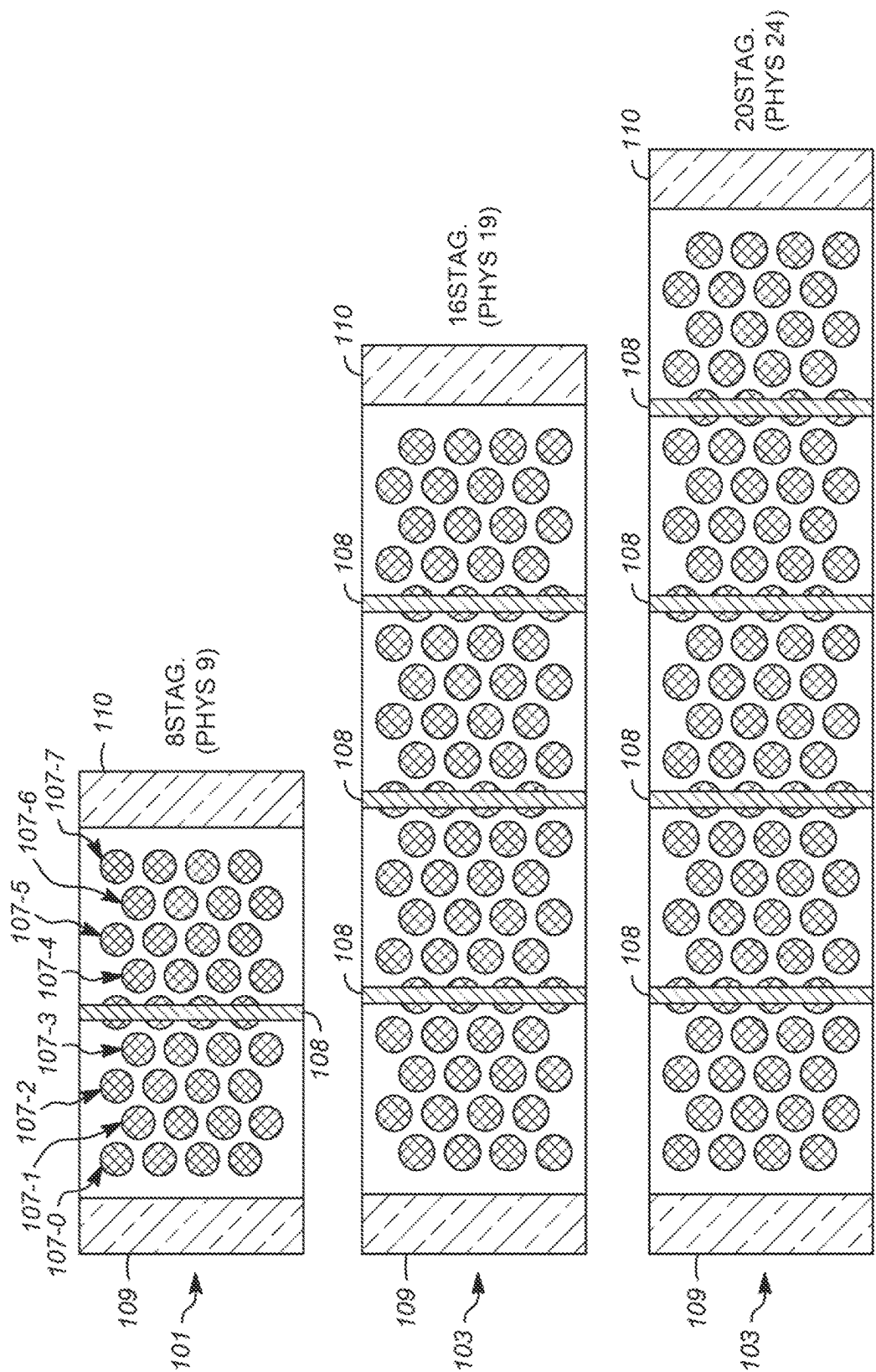


FIG. 8C

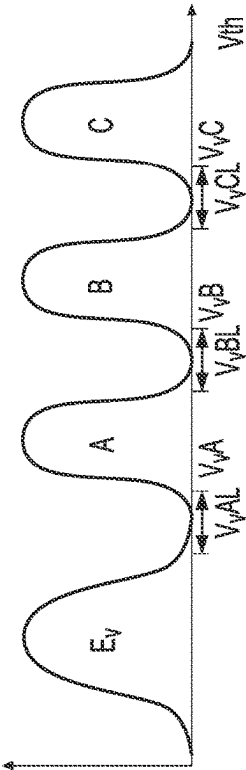


FIG. 9

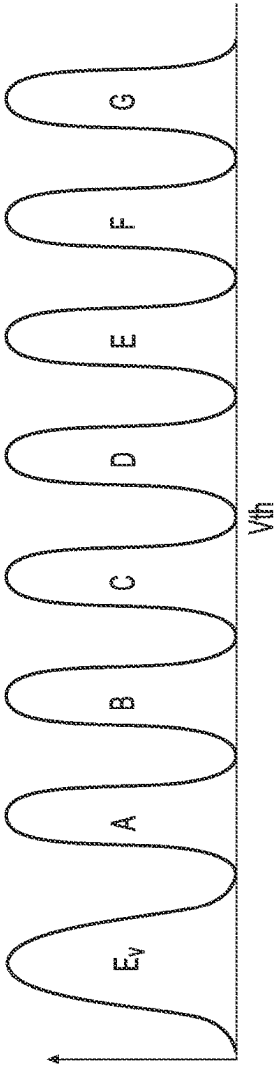


FIG. 10

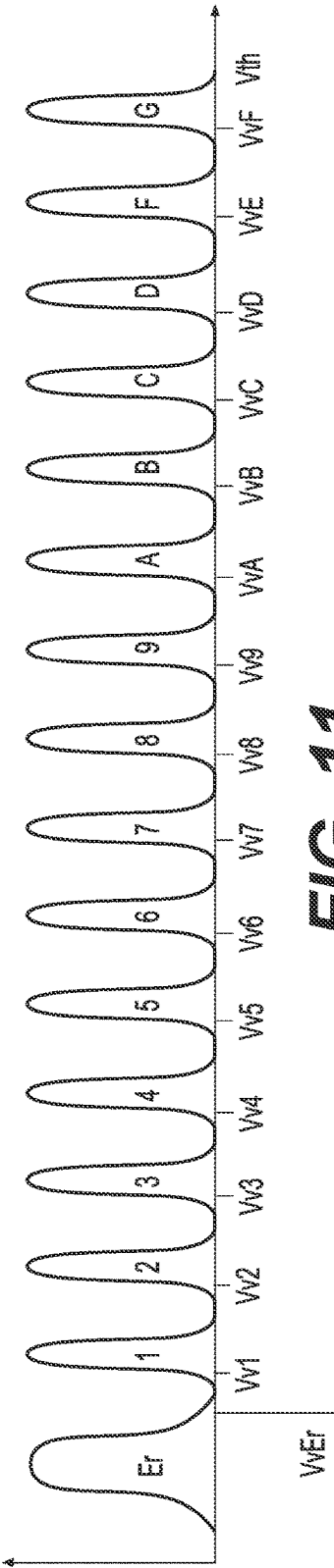


FIG. 11

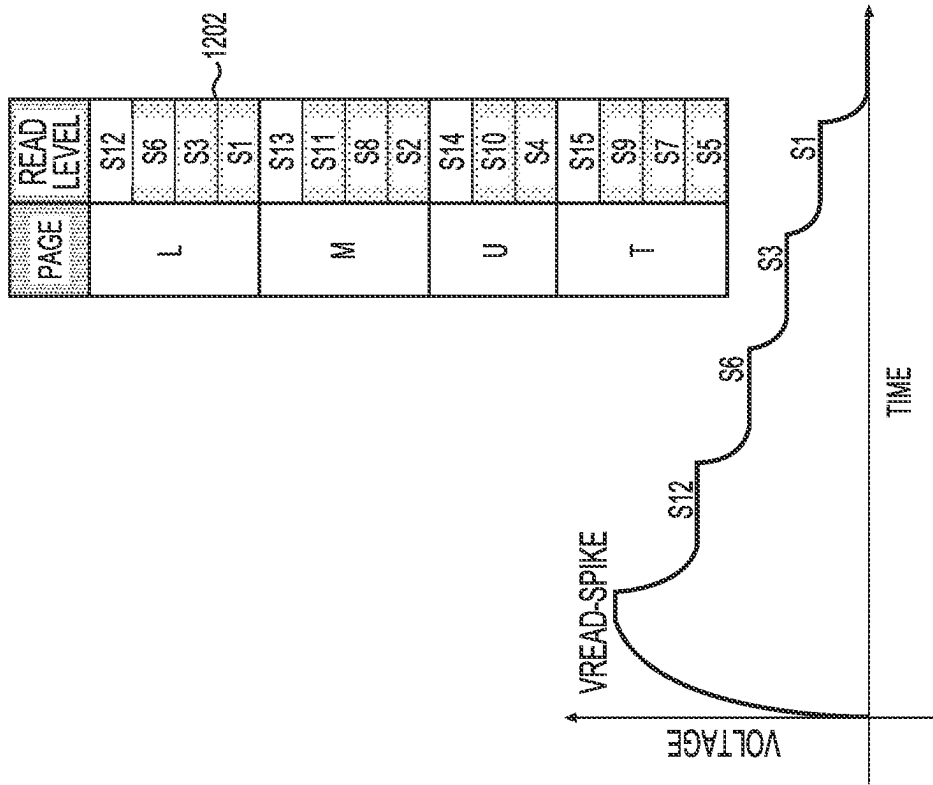


FIG. 12

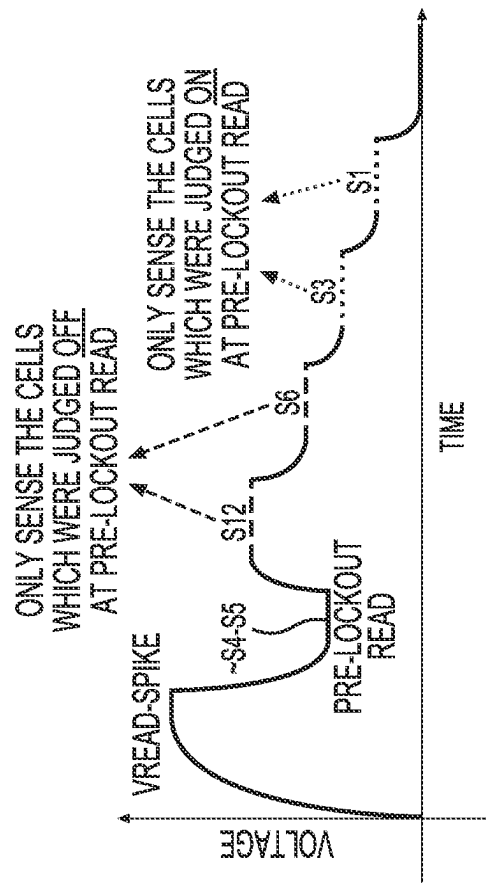


FIG. 13

PAGE	READ LEVEL
L	S12
	S6
	S3
	S1
M	S13
	S11
	S8
	S2
U	S14
	S10
	S4
	S15
T	S9
	S7
	S5

1202

ONLY SENSE THE CELLS WHICH WERE JUDGED OFF AT PRE-LOCKOUT READ

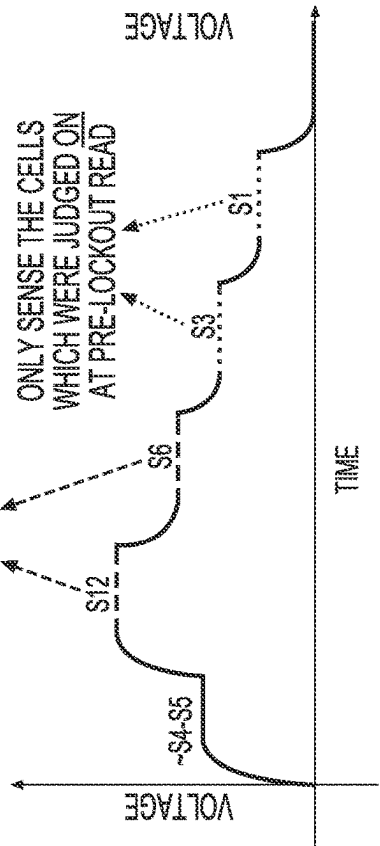


FIG. 14

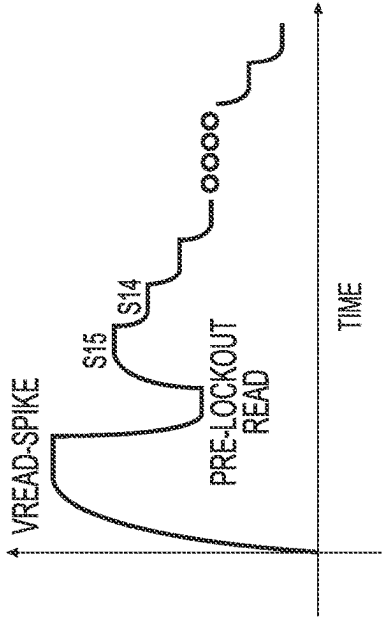
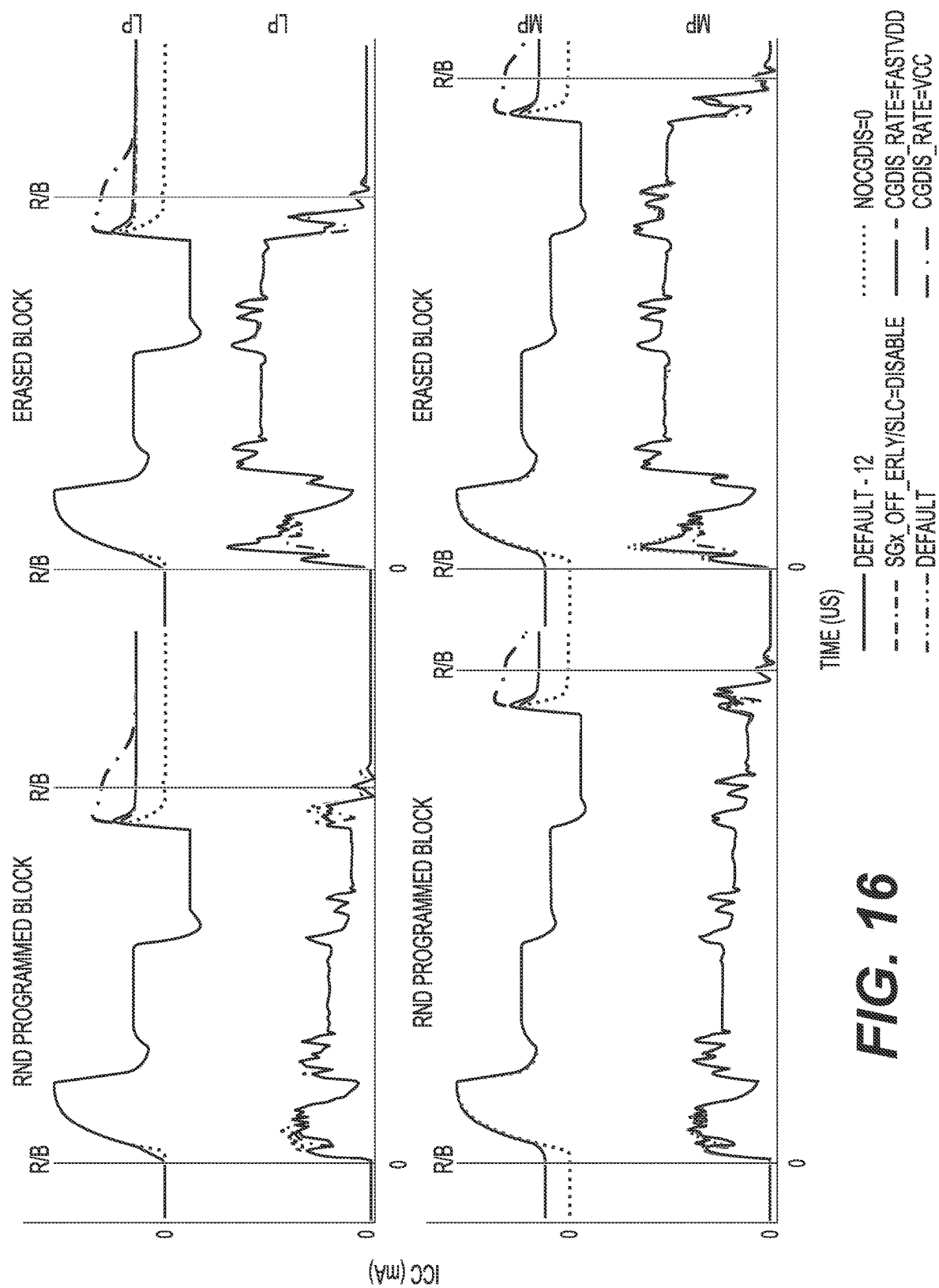
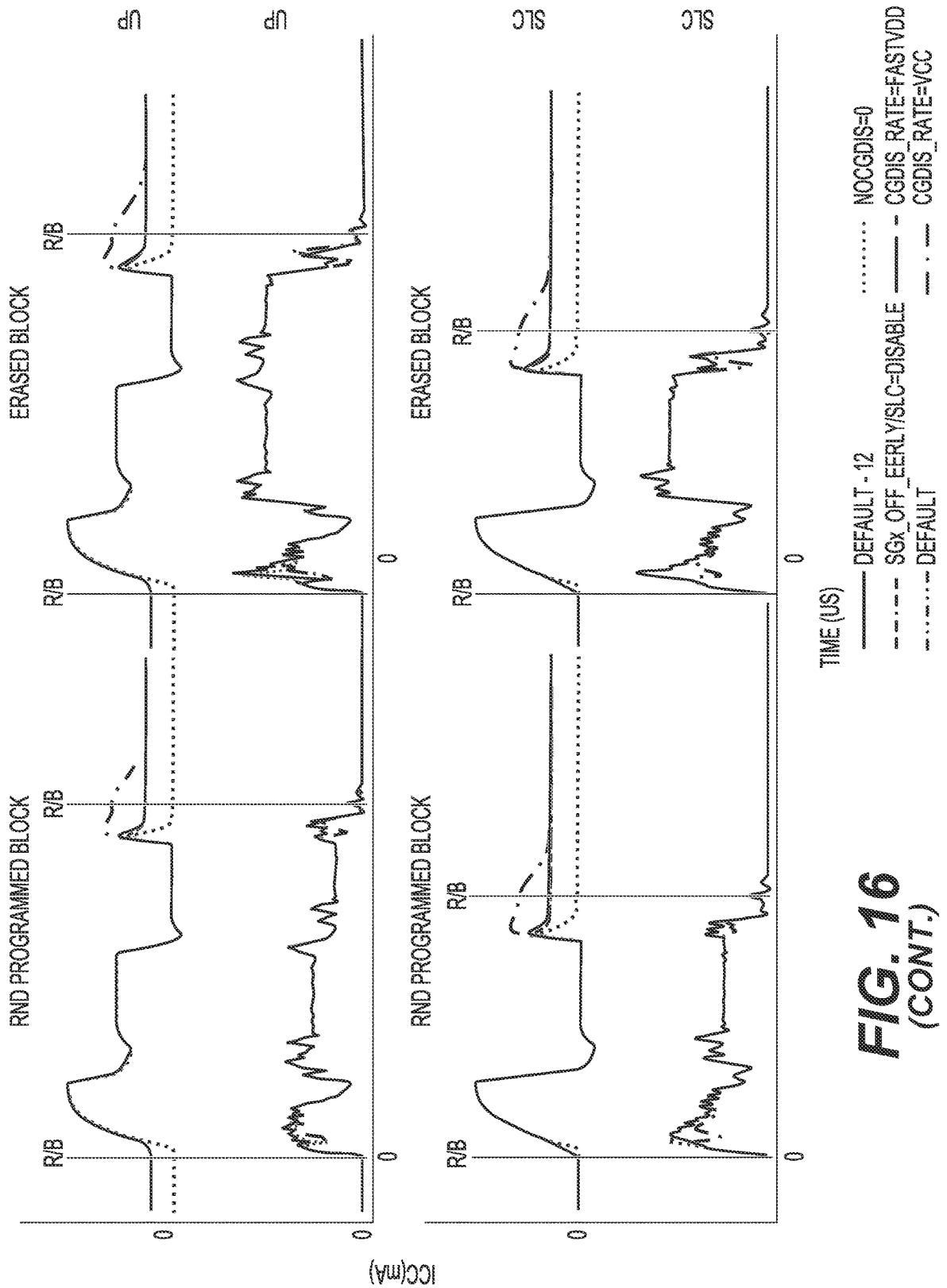


FIG. 15





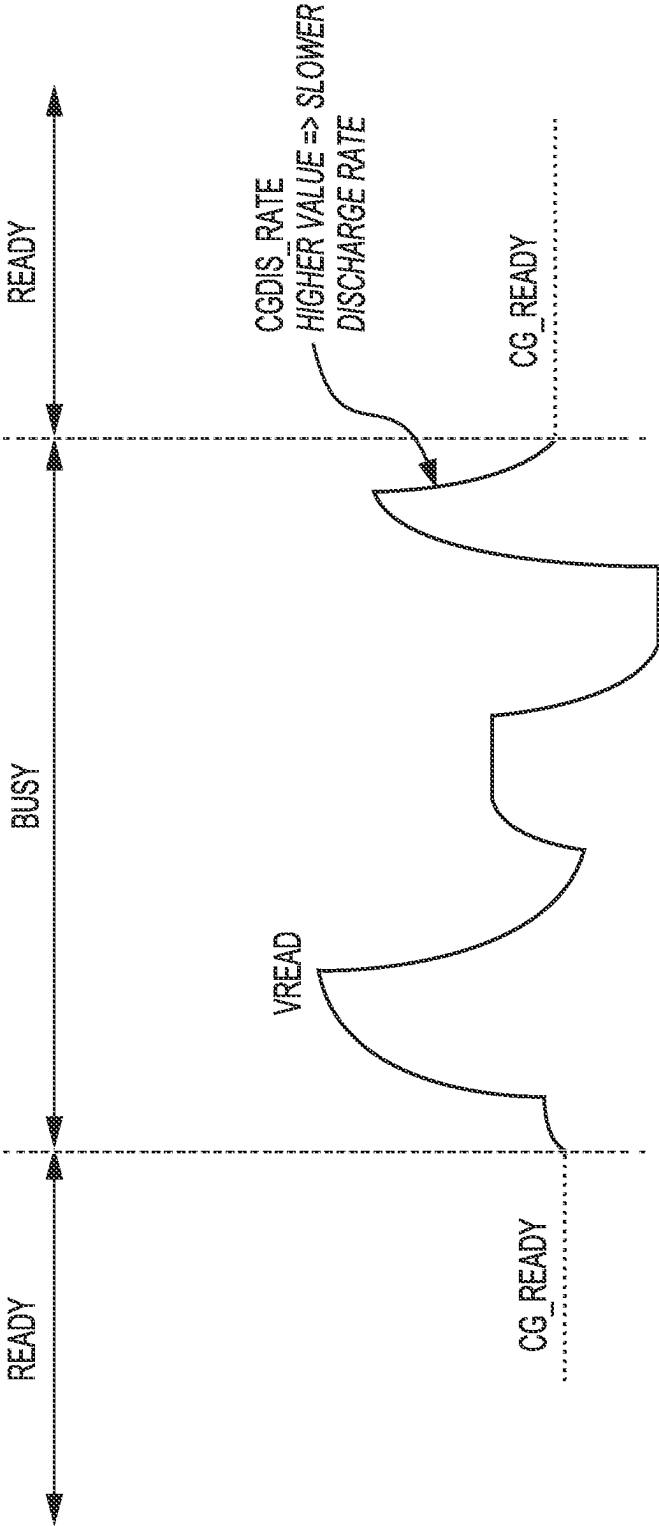


FIG. 17

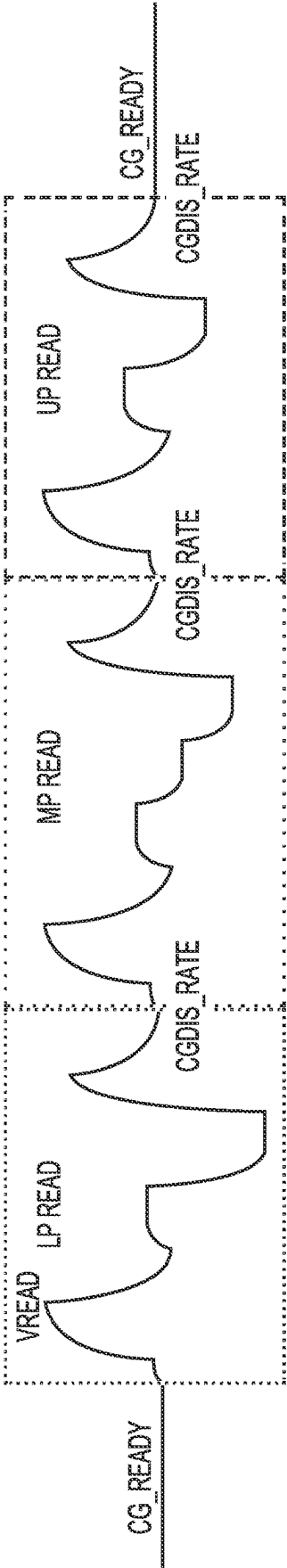


FIG. 18

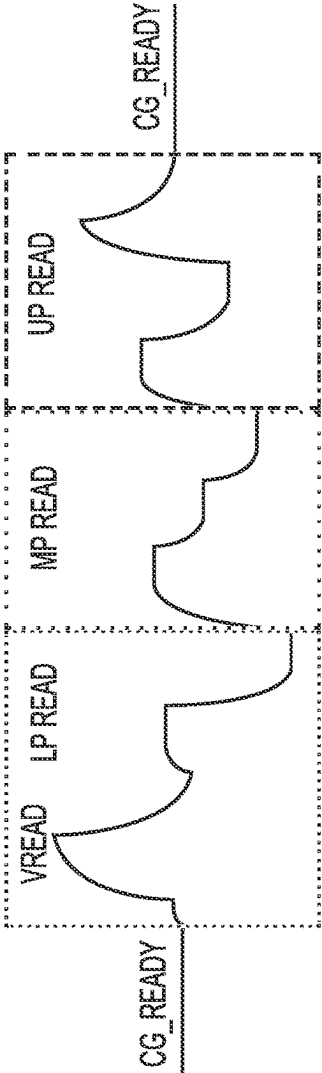


FIG. 19

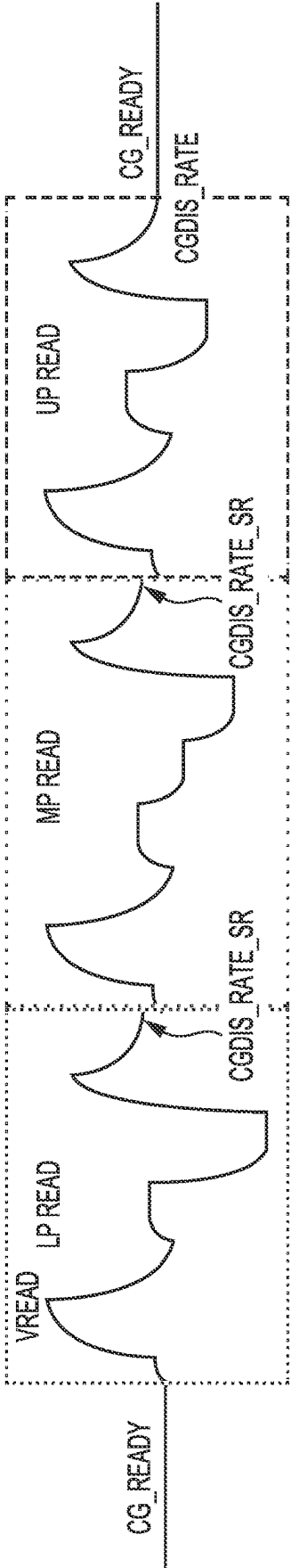


FIG. 20

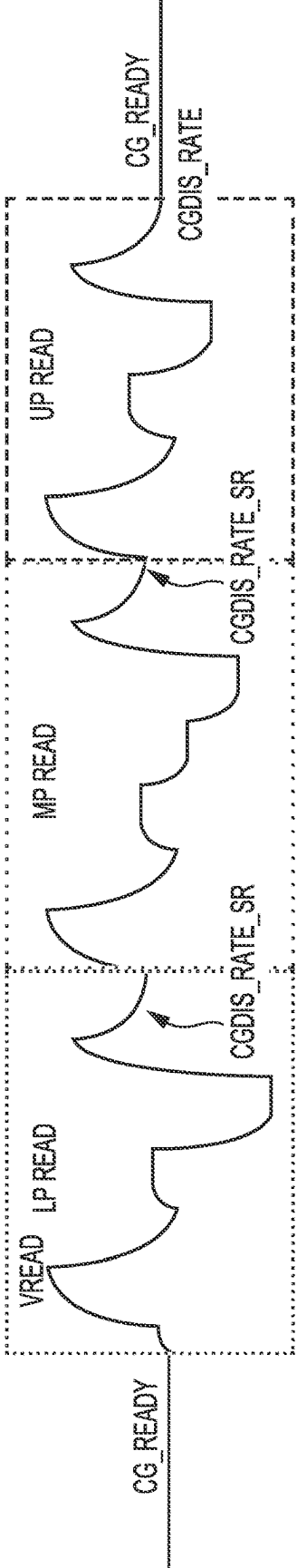


FIG. 21

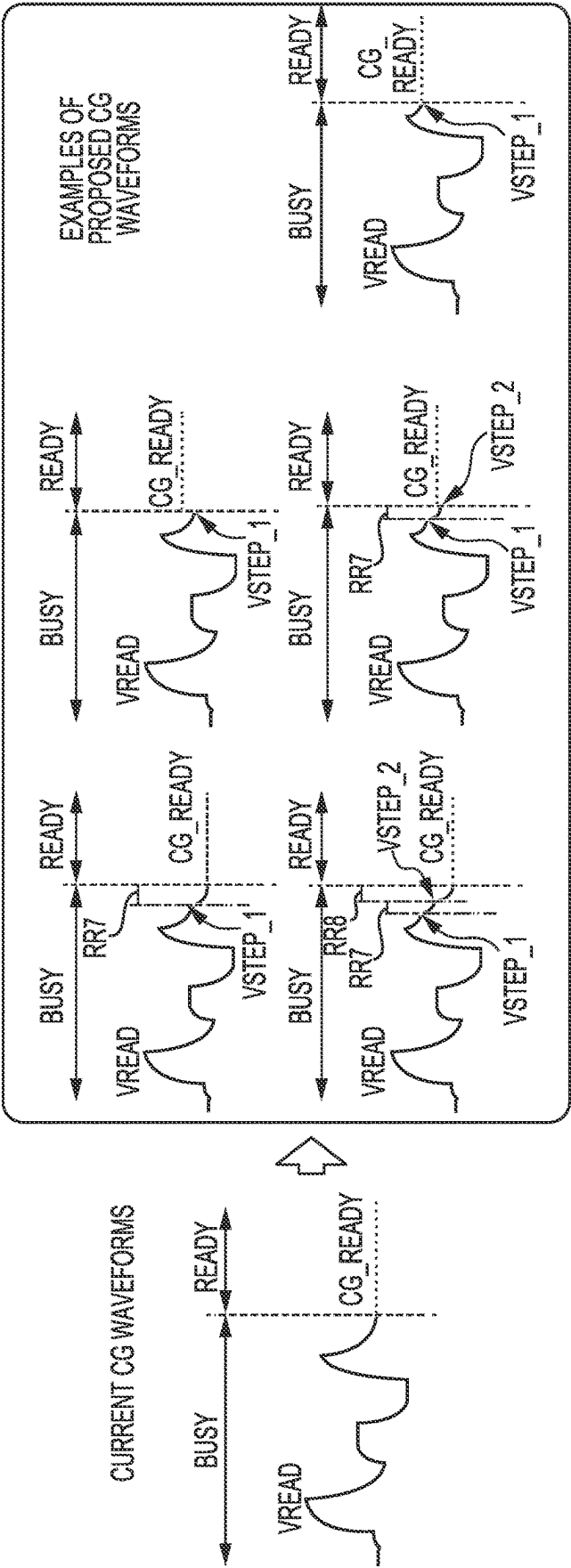


FIG. 22

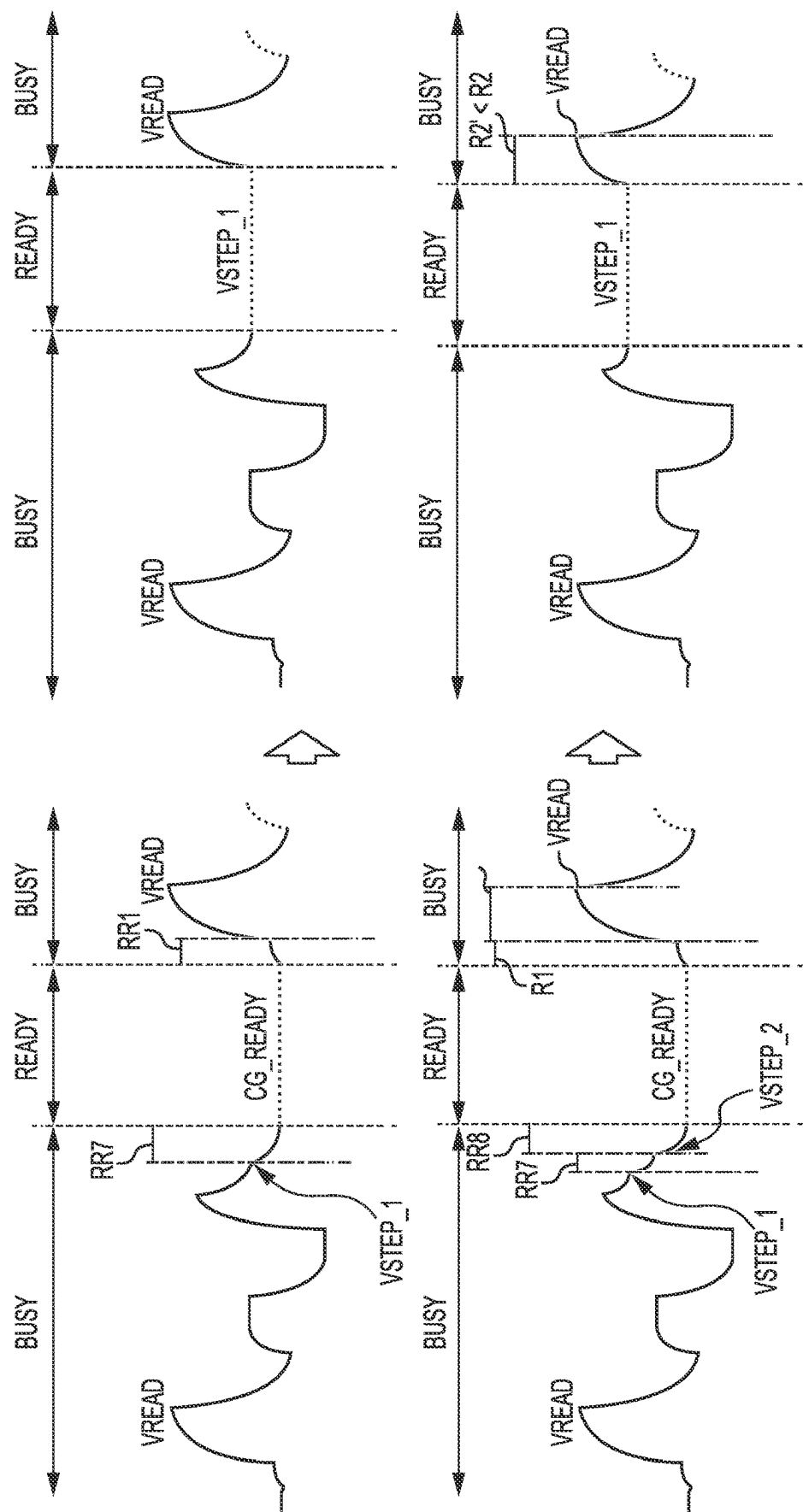
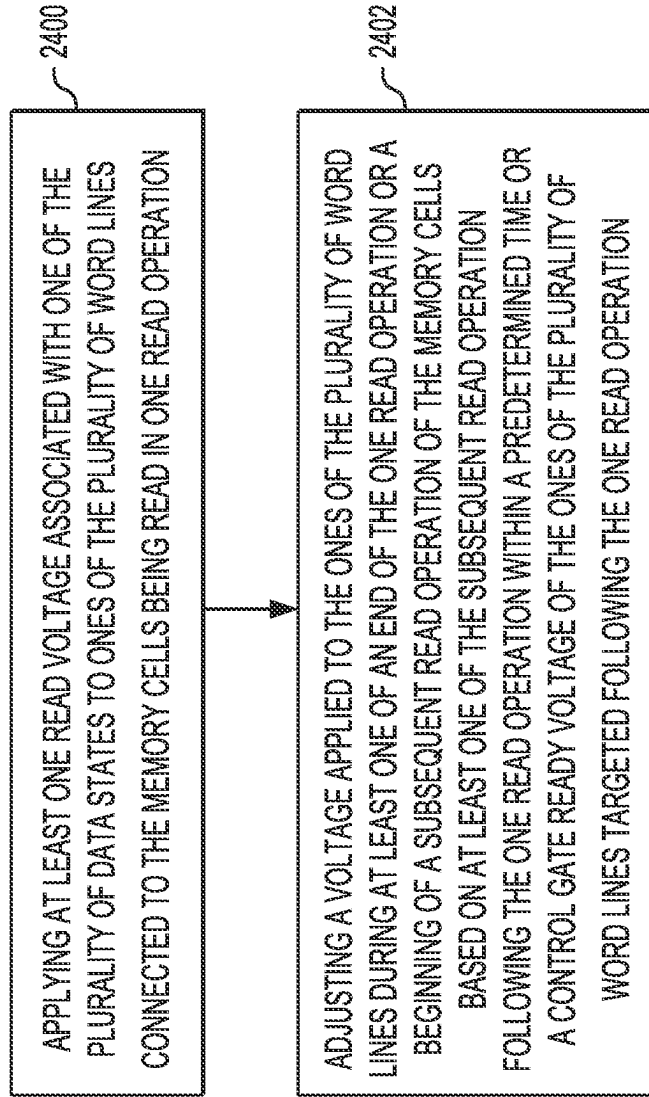


FIG. 23

**FIG. 24**

METHODS TO IMPROVE CURRENT CONSUMPTION AND READ TIME IN SUCCESSIVE READS

FIELD

[0001] This application relates to non-volatile memory apparatuses and the operation of non-volatile memory apparatuses.

BACKGROUND

[0002] This section provides background information related to the technology associated with the present disclosure and, as such, is not necessarily prior art.

[0003] Semiconductor memory apparatuses have become more popular for use in various electronic devices. For example, non-volatile semiconductor memory is used in cellular telephones, digital cameras, personal digital assistants, mobile computing devices, non-mobile computing devices and other devices.

[0004] A charge-storing material such as a floating gate or a charge-trapping material can be used in such memory apparatuses to store a charge which represents a data state. A charge-trapping material can be arranged vertically in a three-dimensional (3D) stacked memory structure, or horizontally in a two-dimensional (2D) memory structure. One example of a 3D memory structure is the Bit Cost Scalable (BiCS) architecture which comprises a stack of alternating conductive and dielectric layers.

SUMMARY

[0005] This section provides a general summary of the present disclosure and is not a comprehensive disclosure of its full scope or all of its features and advantages.

[0006] An object of the present disclosure is to provide a memory apparatus and a method of operation of the memory apparatus that address and overcome shortcomings described herein. Accordingly, it is an aspect of the present disclosure to provide a memory apparatus including memory cells each connected to one of a plurality of word lines and configured to store a threshold voltage corresponding to one of a plurality of data states. The memory apparatus also includes a control means configured to apply at least one read voltage associated with one of the plurality of data states to ones of the plurality of word lines connected to the memory cells being read in one read operation. The control means is also configured to adjust a voltage applied to the ones of the plurality of word lines during at least one of an end of the one read operation or a beginning of a subsequent read operation of the memory cells based on at least one of the subsequent read operation following the one read operation within a predetermined time or a control gate ready voltage of the ones of the plurality of word lines targeted following the one read operation.

[0007] According to another aspect of the disclosure, a controller in communication with a memory apparatus including memory cells each connected to one of a plurality of word lines is provided. The memory cells are configured to store a threshold voltage corresponding to one of a plurality of data states. The controller is configured to instruct the memory apparatus to apply at least one read voltage associated with one of the plurality of data states to ones of the plurality of word lines connected to the memory cells being read in one read operation. The controller is also

configured to instruct the memory apparatus to adjust a voltage applied to the ones of the plurality of word lines during at least one of an end of the one read operation or a beginning of a subsequent read operation of the memory cells based on at least one of the subsequent read operation following the one read operation within a predetermined time or a control gate ready voltage of the ones of the plurality of word lines targeted following the one read operation.

[0008] According to an additional aspect of the disclosure a method of operating a memory apparatus is provided. The memory apparatus includes memory cells each connected to one of a plurality of word lines. The memory cells are configured to store a threshold voltage corresponding to one of a plurality of data states. The method includes the step of applying at least one read voltage associated with one of the plurality of data states to ones of the plurality of word lines connected to the memory cells being read in one read operation. The method also includes the step of adjusting a voltage applied to the ones of the plurality of word lines during at least one of an end of the one read operation or a beginning of a subsequent read operation of the memory cells based on at least one of the subsequent read operation following the one read operation within a predetermined time or a control gate ready voltage of the ones of the plurality of word lines targeted following the one read operation.

[0009] Further areas of applicability will become apparent from the description provided herein. The description and specific examples in this summary are intended for purposes of illustration only and are not intended to limit the scope of the present disclosure.

DRAWINGS

[0010] For a detailed description of example embodiments, reference will now be made to the accompanying drawings in which:

[0011] FIG. 1A is a block diagram of an example memory device according to aspects of the disclosure;

[0012] FIG. 1B is a block diagram of an example control circuit which comprises a programming circuit, a counting circuit, and a determination circuit according to aspects of the disclosure;

[0013] FIG. 2 depicts blocks of memory cells in an example two-dimensional configuration of the memory array of FIG. 1A according to aspects of the disclosure;

[0014] FIG. 3A depicts a cross-sectional view of example floating gate memory cells in NAND strings according to aspects of the disclosure;

[0015] FIG. 3B depicts a cross-sectional view of the structure of FIG. 3A along line 329 according to aspects of the disclosure;

[0016] FIG. 4A depicts a cross-sectional view of example charge-trapping memory cells in NAND strings according to aspects of the disclosure;

[0017] FIG. 4B depicts a cross-sectional view of the structure of FIG. 4A along line 429 according to aspects of the disclosure;

[0018] FIG. 5A depicts an example block diagram of the sense block SB1 of FIG. 1A;

[0019] FIG. 5B depicts another example block diagram of the sense block SB1 of FIG. 1A according to aspects of the disclosure;

[0020] FIG. 6A is a perspective view of a set of blocks in an example three-dimensional configuration of the memory array of FIG. 1A according to aspects of the disclosure;

[0021] FIG. 6B depicts an example cross-sectional view of a portion of one of the blocks of FIG. 6A according to aspects of the disclosure;

[0022] FIG. 6C depicts a plot of memory hole diameter in the stack of FIG. 6B according to aspects of the disclosure;

[0023] FIG. 6D depicts a close-up view of the region 622 of the stack of FIG. 6B according to aspects of the disclosure;

[0024] FIG. 7A depicts a top view of an example word line layer WLL0 of the stack of FIG. 6B according to aspects of the disclosure;

[0025] FIG. 7B depicts a top view of an example top dielectric layer DL19 of the stack of FIG. 6B according to aspects of the disclosure;

[0026] FIG. 8A depicts example NAND strings in the sub-blocks SBa-SBd of FIG. 7A according to aspects of the disclosure;

[0027] FIG. 8B depicts another example view of NAND strings in sub-blocks according to aspects of the disclosure;

[0028] FIG. 8C depicts a top view of example word line layers of a stack according to aspects of the disclosure;

[0029] FIG. 9 depicts the Vth distributions of memory cells in an example one-pass programming operation with four data states according to aspects of the disclosure;

[0030] FIG. 10 depicts the Vth distributions of memory cells in an example one-pass programming operation with eight data states according to aspects of the disclosure;

[0031] FIG. 11 depicts the Vth distributions of memory cells in an example one-pass programming operation with sixteen data states according to aspects of the disclosure;

[0032] FIGS. 12-15 depict embodiments of a voltage waveform corresponding with the word line voltages applied to a selected word line within a memory block during consecutive read operations associated with the same word line according to aspects of the disclosure;

[0033] FIG. 16 shows plots of current consumption during an example read operation of a lower page, middle page, and upper page of a block of a memory apparatus with triple level cells and for a read operation of a block of a memory apparatus with single level cells that is erased and randomly programmed using different ramp rates according to aspects of the disclosure;

[0034] FIG. 17 is a reference waveform for an example read operation of a page and illustrates the importance of target discharge bias according to aspects of the disclosure;

[0035] FIG. 18 shows an example waveform of a voltage applied to word lines for consecutive reads of all pages on the same word line and string according to aspects of the disclosure;

[0036] FIG. 19 shows an example waveform of a voltage applied to word lines for consecutive reads of all pages on the same word line and string with the discharge stage and the following read pass voltage spike removed according to aspects of the disclosure;

[0037] FIG. 20 shows an example waveform of a voltage applied to word lines for consecutive reads of all pages on the same word line but different string with a different target discharge bias used for the discharge stage according to aspects of the disclosure;

[0038] FIG. 21 shows an example waveform of a voltage applied to word lines for consecutive reads of all pages on

the same word line but different string with the double ramp-up in the reads after CGDIS_RATE_SR skipped according to aspects of the disclosure;

[0039] FIG. 22 shows a waveform for an example read operation of a page along with examples of the discharge including two or more steps according to aspects of the disclosure;

[0040] FIG. 23 shows changes to a waveform for an example read operation of a page along with examples of the stepwise discharge and the corresponding additional times being eliminated according to aspects of the disclosure; and

[0041] FIG. 24 illustrates steps of a method of operating a memory apparatus according to aspects of the disclosure.

[0042] To facilitate understanding, identical reference numerals have been used, where possible, to designate identical elements that are common to the figures. It is contemplated that elements disclosed in one embodiment may be beneficially utilized on other embodiments without specific recitation.

DETAILED DESCRIPTION

[0043] In the following description, details are set forth to provide an understanding of the present disclosure. In some instances, certain circuits, structures and techniques have not been described or shown in detail in order not to obscure the disclosure.

[0044] In general, the present disclosure relates to non-volatile memory apparatuses of the type well-suited for use in many applications. The non-volatile memory apparatus and associated methods of forming of this disclosure will be described in conjunction with one or more example embodiments. However, the specific example embodiments disclosed are merely provided to describe the inventive concepts, features, advantages and objectives with sufficient clarity to permit those skilled in this art to understand and practice the disclosure. Specifically, the example embodiments are provided so that this disclosure will be thorough, and will fully convey the scope to those who are skilled in the art. Numerous specific details are set forth such as examples of specific components, devices, and methods, to provide a thorough understanding of embodiments of the present disclosure. It will be apparent to those skilled in the art that specific details need not be employed, that example embodiments may be embodied in many different forms and that neither should be construed to limit the scope of the disclosure. In some example embodiments, well-known processes, well-known device structures, and well-known technologies are not described in detail.

[0045] A programming operation for a set of memory cells of a memory device typically involves applying a series of program voltages to the memory cells after the memory cells are provided in an erased state. Each program voltage is provided in a program loop, also referred to as a program-verify iteration. For example, the program voltage may be applied to a word line which is connected to control gates of the memory cells. In one approach, incremental step pulse programming is performed, where the program voltage is increased by a step size in each program loop. Verify operations may be performed after each program voltage to determine whether the memory cells have completed programming. When programming is completed for a memory cell, it can be locked out from further programming while programming continues for other memory cells in subsequent program loops.

[0046] Each memory cell may be associated with a data state according to write data in a program command. Based on its data state, a memory cell will either remain in the erased state or be programmed to a data state (a programmed data state) different from the erased state. For example, in a one-bit per cell memory device (single-level cell (SLC)), there are two data states including the erased state and one higher data state. In a two-bit per cell memory device (multi-level cell (MLC)), there are four data states including the erased state and three higher data states referred to as the A, B and C data states (see FIG. 9). In a three-bit per cell memory device (triple-level cell (TLC)), there are eight data states including the erased state and seven higher data states referred to as the A, B, C, D, E, F and G data states (see FIG. 10). In a four-bit per cell memory device (quad-level cell (QLC)), there are sixteen data states including the erased state and fifteen higher data states referred to as the Er, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E and F data states (see FIG. 11). Each memory cell may store a data state (e.g., a binary value) and is programmed to a threshold voltage state corresponding to the data state. Each state represents a different value and is assigned a voltage window including a range of possible threshold voltages.

[0047] When a program command is issued, the write data is stored in latches associated with the memory cells. During programming, the latches of a memory cell can be read to determine the data state to which the cell is to be programmed. Each programmed data state is associated with a verify voltage such that a memory cell with a given data state is considered to have completed programming when a sensing operation determines its threshold voltage (V_{th}) is above the associated verify voltage. A sensing operation can determine whether a memory cell has a V_{th} above the associated verify voltage by applying the associated verify voltage to the control gate and sensing a current through the memory cell. If the current is relatively high, this indicates the memory cell is in a conductive state, such that the V_{th} is less than the control gate voltage. If the current is relatively low, this indicates the memory cell is in a non-conductive state, such that the V_{th} is above the control gate voltage.

[0048] The verify voltage which is used to determine that a memory cell has completed programming may be referred to as a final or lockout verify voltage. In some cases, an additional verify voltage may be used to determine that a memory cell is close to completion of the programming. This additional verify voltage may be referred to as an offset verify voltage, and may be lower than the final verify voltage. When a memory cell is close to completion of programming, the programming speed of the memory cell may be reduced such as by elevating a voltage of a respective bit line during one or more subsequent program voltages. For example, in FIG. 9, a memory cell which is to be programmed to the A data state may be subject to verify tests at V_{VAL} , an offset verify voltage of the A data state, and V_{VA} , a final verify voltage of the A data state.

[0049] After the memory cells are programmed, the data can also be read back in a read operation. A read operation can involve applying a series of read voltages to a word line while sensing circuitry determines whether cells connected to the word line are in a conductive or non-conductive state. If a cell is in a non-conductive state, the V_{th} of the memory cell exceeds the read voltage. The read voltages are set at

levels which are expected to be between the threshold voltage levels of adjacent data states

[0050] To help further illustrate the foregoing, FIG. 1A will now be described. FIG. 1A is a block diagram of an example memory device. The memory device 100 may include one or more memory die 108. The memory die 108 includes a memory structure 126 of memory cells, such as an array of memory cells, control circuitry 110, and read/write circuits 128. The memory structure 126 is addressable by word lines via a row decoder 124 and by bit lines via a column decoder 132. The read/write circuits 128 include multiple sense blocks SB1, SB2, . . . , SBp (sensing circuitry) and allow a page of memory cells to be read or programmed in parallel. Typically a controller 122 is included in the same memory device 100 (e.g., a removable storage card) as the one or more memory die 108. Commands and data are transferred between the host 140 and controller 122 via a data bus 120, and between the controller and the one or more memory die 108 via lines 118.

[0051] As described with reference to FIG. 1A, controller 122 is included in the same memory device 100 as the one or more memory die 108. For example, in some embodiments, a memory device may include NAND flash dies that are situated in columns next to each other; whereas, in some embodiments, a memory device may include NAND flash dies that are stacked on top of each other. Further, in some embodiments, the memory device may include the stacked NAND flash dies and a controller as a separate die inside the package.

[0052] The memory structure can be 2D or 3D. The memory structure may comprise one or more array of memory cells including a 3D array. The memory structure may comprise a monolithic three dimensional memory structure in which multiple memory levels are formed above (and not in) a single substrate, such as a wafer, with no intervening substrates. The memory structure may comprise any type of non-volatile memory that is monolithically formed in one or more physical levels of arrays of memory cells having an active area disposed above a silicon substrate. The memory structure may be in a non-volatile memory device having circuitry associated with the operation of the memory cells, whether the associated circuitry is above or within the substrate.

[0053] The control circuitry 110 cooperates with the read/write circuits 128 to perform memory operations on the memory structure 126, and includes a state machine 112, an on-chip address decoder 114, and a power control module 116. The state machine 112 provides chip-level control of memory operations. A storage region 113 may be provided, e.g., for verify parameters as described herein.

[0054] The on-chip address decoder 114 provides an address interface between that used by the host or a memory controller to the hardware address used by the decoders 124 and 132. The power control module 116 controls the power and voltages supplied to the word lines and bit lines during memory operations. It can include drivers for word lines, SGS and SGD transistors and source lines. The sense blocks can include bit line drivers, in one approach. An SGS transistor is a select gate transistor at a source end of a NAND string, and an SGD transistor is a select gate transistor at a drain end of a NAND string.

[0055] In some implementations, some of the components can be combined. In various designs, one or more of the components (alone or in combination), other than memory

structure **126**, can be thought of as at least one control circuit which is configured to perform the actions described herein. For example, a control circuit may include any one of, or a combination of, control circuitry **110**, state machine **112**, decoders **114/132**, power control module **116**, sense blocks **SBb**, **SB2**, . . . , **SBp**, read/write circuits **128**, controller **122**, and so forth.

[0056] The control circuits can include a programming circuit configured to program memory cells of a word line of a block and verify the set of the memory cells. The control circuits can also include a counting circuit configured to determine a number of memory cells that are verified to be in a data state. The control circuits can also include a determination circuit configured to determine, based on the number, whether the block is faulty.

[0057] For example, FIG. 1B is a block diagram of an example control circuit **150** which comprises a programming circuit **151**, a counting circuit **152** and a determination circuit **153**. The programming circuit may include software, firmware and/or hardware. The counting circuit may include software, firmware and/or hardware. The determination circuit may include software, firmware and/or hardware.

[0058] The off-chip controller **122** may comprise a processor **122c**, storage devices (memory) such as ROM **122a** and RAM **122b** and an error-correction code (ECC) engine **245**. The ECC engine can correct a number of read errors which are caused when the upper tail of a V_{th} distribution becomes too high. However, uncorrectable errors may exist in some cases. The techniques provided herein reduce the likelihood of uncorrectable errors.

[0059] The storage device comprises code such as a set of instructions, and the processor is operable to execute the set of instructions to provide the functionality described herein. Alternatively or additionally, the processor can access code from a storage device **126a** of the memory structure, such as a reserved area of memory cells in one or more word lines.

[0060] For example, code can be used by the controller **122** to access the memory structure such as for programming, read and erase operations. The code can include boot code and control code (e.g., set of instructions). The boot code is software that initializes the controller during a booting or startup process and enables the controller to access the memory structure. The code can be used by the controller to control one or more memory structures. Upon being powered up, the processor **122c** fetches the boot code from the ROM **122a** or storage device **126a** for execution, and the boot code initializes the system components and loads the control code into the RAM **122b**. Once the control code is loaded into the RAM, it is executed by the processor. The control code includes drivers to perform basic tasks such as controlling and allocating memory, prioritizing the processing of instructions, and controlling input and output ports.

[0061] In one embodiment, the host is a computing device (e.g., laptop, desktop, smartphone, tablet, digital camera) that includes one or more processors, one or more processor readable storage devices (RAM, ROM, flash memory, hard disk drive, solid state memory) that store processor readable code (e.g., software) for programming the one or more processors to perform the methods described herein. The host may also include additional system memory, one or more input/output interfaces and/or one or more input/output devices in communication with the one or more processors.

[0062] Other types of non-volatile memory in addition to NAND flash memory can also be used.

[0063] Semiconductor memory devices include volatile memory devices, such as dynamic random access memory (“DRAM”) or static random access memory (“SRAM”) devices, non-volatile memory devices, such as resistive random access memory (“ReRAM”), electrically erasable programmable read only memory (“EEPROM”), flash memory (which can also be considered a subset of EEPROM), ferroelectric random access memory (“FRAM”), and magnetoresistive random access memory (“MRAM”), and other semiconductor elements capable of storing information. Each type of memory device may have different configurations. For example, flash memory devices may be configured in a NAND or a NOR configuration.

[0064] The memory devices can be formed from passive and/or active elements, in any combinations. By way of non-limiting example, passive semiconductor memory elements include ReRAM device elements, which in some embodiments include a resistivity switching storage element, such as an anti-fuse or phase change material, and optionally a steering element, such as a diode or transistor. Further by way of non-limiting example, active semiconductor memory elements include EEPROM and flash memory device elements, which in some embodiments include elements containing a charge storage region, such as a floating gate, conductive nanoparticles, or a charge storage dielectric material.

[0065] Multiple memory elements may be configured so that they are connected in series or so that each element is individually accessible. By way of non-limiting example, flash memory devices in a NAND configuration (NAND memory) typically contain memory elements connected in series. A NAND string is an example of a set of series-connected transistors comprising memory cells and SG transistors.

[0066] A NAND memory array may be configured so that the array is composed of multiple strings of memory in which a string is composed of multiple memory elements sharing a single bit line and accessed as a group. Alternatively, memory elements may be configured so that each element is individually accessible, e.g., a NOR memory array. NAND and NOR memory configurations are examples, and memory elements may be otherwise configured.

[0067] The semiconductor memory elements located within and/or over a substrate may be arranged in two or three dimensions, such as a two dimensional memory structure or a three dimensional memory structure.

[0068] In a two dimensional memory structure, the semiconductor memory elements are arranged in a single plane or a single memory device level. Typically, in a two dimensional memory structure, memory elements are arranged in a plane (e.g., in an x-y direction plane) which extends substantially parallel to a major surface of a substrate that supports the memory elements. The substrate may be a wafer over or in which the layer of the memory elements are formed or it may be a carrier substrate which is attached to the memory elements after they are formed. As a non-limiting example, the substrate may include a semiconductor such as silicon.

[0069] The memory elements may be arranged in the single memory device level in an ordered array, such as in a plurality of rows and/or columns. However, the memory

elements may be arrayed in non-regular or non-orthogonal configurations. The memory elements may each have two or more electrodes or contact lines, such as bit lines and word lines.

[0070] A three dimensional memory array is arranged so that memory elements occupy multiple planes or multiple memory device levels, thereby forming a structure in three dimensions (i.e., in the x, y and z directions, where the z direction is substantially perpendicular and the x and y directions are substantially parallel to the major surface of the substrate).

[0071] As a non-limiting example, a three dimensional memory structure may be vertically arranged as a stack of multiple two dimensional memory device levels. As another non-limiting example, a three dimensional memory array may be arranged as multiple vertical columns (e.g., columns extending substantially perpendicular to the major surface of the substrate, i.e., in the y direction) with each column having multiple memory elements. The columns may be arranged in a two dimensional configuration, e.g., in an x-y plane, resulting in a three dimensional arrangement of memory elements with elements on multiple vertically stacked memory planes. Other configurations of memory elements in three dimensions can also constitute a three dimensional memory array.

[0072] By way of non-limiting example, in a three dimensional NAND memory array, the memory elements may be coupled together to form a NAND string within a single horizontal (e.g., x-y) memory device level. Alternatively, the memory elements may be coupled together to form a vertical NAND string that traverses across multiple horizontal memory device levels. Other three dimensional configurations can be envisioned wherein some NAND strings contain memory elements in a single memory level while other strings contain memory elements which span through multiple memory levels. Three dimensional memory arrays may also be designed in a NOR configuration and in a ReRAM configuration.

[0073] Typically, in a monolithic three dimensional memory array, one or more memory device levels are formed above a single substrate. Optionally, the monolithic three dimensional memory array may also have one or more memory layers at least partially within the single substrate. As a non-limiting example, the substrate may include a semiconductor such as silicon. In a monolithic three dimensional array, the layers constituting each memory device level of the array are typically formed on the layers of the underlying memory device levels of the array. However, layers of adjacent memory device levels of a monolithic three dimensional memory array may be shared or have intervening layers between memory device levels.

[0074] Then again, two dimensional arrays may be formed separately and then packaged together to form a non-monolithic memory device having multiple layers of memory. For example, non-monolithic stacked memories can be constructed by forming memory levels on separate substrates and then stacking the memory levels atop each other. The substrates may be thinned or removed from the memory device levels before stacking, but as the memory device levels are initially formed over separate substrates, the resulting memory arrays are not monolithic three dimensional memory arrays. Further, multiple two dimensional memory arrays or three dimensional memory arrays (mono-

lithic or non-monolithic) may be formed on separate chips and then packaged together to form a stacked-chip memory device.

[0075] Associated circuitry is typically required for operation of the memory elements and for communication with the memory elements. As non-limiting examples, memory devices may have circuitry used for controlling and driving memory elements to accomplish functions such as programming and reading. This associated circuitry may be on the same substrate as the memory elements and/or on a separate substrate. For example, a controller for memory read-write operations may be located on a separate controller chip and/or on the same substrate as the memory elements.

[0076] One of skill in the art will recognize that this technology is not limited to the two dimensional and three dimensional exemplary structures described but covers all relevant memory structures within the spirit and scope of the technology as described herein and as understood by one of skill in the art.

[0077] FIG. 2 depicts blocks of memory cells in an example two-dimensional configuration of the memory array 126 of FIG. 1A. The memory array can include many blocks. Each example block 200, 210 includes a number of NAND strings and respective bit lines, e.g., BL0, BL1, . . . which are shared among the blocks. Each NAND string is connected at one end to a drain select gate (SGD), and the control gates of the drain select gates are connected via a common SGD line. The NAND strings are connected at their other end to a source select gate which, in turn, is connected to a common source line 220. Sixteen word lines, for example, WL0-WL15, extend between the source select gates and the drain select gates. In some cases, dummy word lines, which contain no user data, can also be used in the memory array adjacent to the select gate transistors. Such dummy word lines can shield the edge data word line from certain edge effects.

[0078] One type of non-volatile memory which may be provided in the memory array is a floating gate memory. See FIGS. 3A and 3B. Other types of non-volatile memory can also be used. For example, a charge-trapping memory cell uses a non-conductive dielectric material in place of a conductive floating gate to store charge in a non-volatile manner. See FIGS. 4A and 4B. A triple layer dielectric formed of silicon oxide, silicon nitride and silicon oxide ("ONO") is sandwiched between a conductive control gate and a surface of a semi-conductive substrate above the memory cell channel. The cell is programmed by injecting electrons from the cell channel into the nitride, where they are trapped and stored in a limited region. This stored charge then changes the threshold voltage of a portion of the channel of the cell in a manner that is detectable. The cell is erased by injecting hot holes into the nitride. A similar cell can be provided in a split-gate configuration where a doped polysilicon gate extends over a portion of the memory cell channel to form a separate select transistor.

[0079] In another approach, NROM cells are used. Two bits, for example, are stored in each NROM cell, where an ONO dielectric layer extends across the channel between source and drain diffusions. The charge for one data bit is localized in the dielectric layer adjacent to the drain, and the charge for the other data bit localized in the dielectric layer adjacent to the source. Multi-state data storage is obtained by separately reading binary states of the spatially separated

charge storage regions within the dielectric. Other types of non-volatile memory are also known.

[0080] FIG. 3A depicts a cross-sectional view of example floating gate memory cells in NAND strings. A bit line or NAND string direction goes into the page, and a word line direction goes from left to right. As an example, word line 324 extends across NAND strings which include respective channel regions 306, 316 and 326. The memory cell 300 includes a control gate 302, a floating gate 304, a tunnel oxide layer 305 and the channel region 306. The memory cell 310 includes a control gate 312, a floating gate 314, a tunnel oxide layer 315 and the channel region 316. The memory cell 320 includes a control gate 322, a floating gate 321, a tunnel oxide layer 325 and the channel region 326. Each memory cell is in a different respective NAND string. An inter-poly dielectric (IPD) layer 328 is also depicted. The control gates are portions of the word line. A cross-sectional view along line 329 is provided in FIG. 3B.

[0081] The control gate wraps around the floating gate, increasing the surface contact area between the control gate and floating gate. This results in higher IPD capacitance, leading to a higher coupling ratio which makes programming and erase easier. However, as NAND memory devices are scaled down, the spacing between neighboring cells becomes smaller so there is almost no space for the control gate and the IPD between two adjacent floating gates. As an alternative, as shown in FIGS. 4A and 4B, the flat or planar memory cell has been developed in which the control gate is flat or planar; that is, it does not wrap around the floating gate, and its only contact with the charge storage layer is from above it. In this case, there is no advantage in having a tall floating gate. Instead, the floating gate is made much thinner. Further, the floating gate can be used to store charge, or a thin charge trap layer can be used to trap charge. This approach can avoid the issue of ballistic electron transport, where an electron can travel through the floating gate after tunneling through the tunnel oxide during programming.

[0082] FIG. 3B depicts a cross-sectional view of the structure of FIG. 3A along line 329. The NAND string 330 includes an SGS transistor 331, example memory cells 300, 333, . . . , 334 and 335, and an SGD transistor 336. The memory cell 300, as an example of each memory cell, includes the control gate 302, the IPD layer 328, the floating gate 304 and the tunnel oxide layer 305, consistent with FIG. 3A. Passageways in the IPD layer in the SGS and SGD transistors allow the control gate layers and floating gate layers to communicate. The control gate and floating gate layers may be polysilicon and the tunnel oxide layer may be silicon oxide, for instance. The IPD layer can be a stack of nitrides (N) and oxides (O) such as in a N—O—N—O—N configuration.

[0083] The NAND string may be formed on a substrate which comprises a p-type substrate region 355, an n-type well 356 and a p-type well 357. N-type source/drain diffusion regions sd1, sd2, sd3, sd4, sd5, sd6 and sd7 are formed in the p-type well. A channel voltage, V_{ch} , may be applied directly to the channel region of the substrate.

[0084] FIG. 4A depicts a cross-sectional view of example charge-trapping memory cells in NAND strings. The view is in a word line direction of memory cells comprising a flat control gate and charge-trapping regions as a 2D example of memory cells in the memory cell array 126 of FIG. 1A. Charge-trapping memory can be used in NOR and NAND flash memory device. This technology uses an insulator such

as an SiN film to store electrons, in contrast to a floating-gate MOSFET technology which uses a conductor such as doped polycrystalline silicon to store electrons. As an example, a word line (WL) 424 extends across NAND strings which include respective channel regions 406, 416 and 426. Portions of the word line provide control gates 402, 412 and 422. Below the word line is an IPD layer 428, charge-trapping layers 404, 414 and 421, polysilicon layers 405, 415 and 425 and tunneling layer layers 409, 407 and 408. Each charge-trapping layer extends continuously in a respective NAND string.

[0085] A memory cell 400 includes the control gate 402, the charge-trapping layer 404, the polysilicon layer 405 and a portion of the channel region 406. A memory cell 410 includes the control gate 412, the charge-trapping layer 414, a polysilicon layer 415 and a portion of the channel region 416. A memory cell 420 includes the control gate 422, the charge-trapping layer 421, the polysilicon layer 425 and a portion of the channel region 426.

[0086] A flat control gate is used here instead of a control gate that wraps around a floating gate. One advantage is that the charge-trapping layer can be made thinner than a floating gate. Additionally, the memory cells can be placed closer together.

[0087] FIG. 4B depicts a cross-sectional view of the structure of FIG. 4A along line 429. The view shows a NAND string 430 having a flat control gate and a charge-trapping layer. The NAND string 430 includes an SGS transistor 431, example memory cells 400, 433, . . . , 434 and 435, and an SGD transistor 435.

[0088] The NAND string may be formed on a substrate which comprises a p-type substrate region 455, an n-type well 456 and a p-type well 457. N-type source/drain diffusion regions sd1, sd2, sd3, sd4, sd5, sd6 and sd7 are formed in the p-type well 457. A channel voltage, V_{ch} , may be applied directly to the channel region of the substrate. The memory cell 400 includes the control gate 402 and the IPD layer 428 above the charge-trapping layer 404, the polysilicon layer 405, the tunneling layer 409 and the channel region 406.

[0089] The control gate layer may be polysilicon and the tunneling layer may be silicon oxide, for instance. The IPD layer can be a stack of high-k dielectrics such as AlOx or HfOx which help increase the coupling ratio between the control gate layer and the charge-trapping or charge storing layer. The charge-trapping layer can be a mix of silicon nitride and oxide, for instance.

[0090] The SGD and SGS transistors have the same configuration as the memory cells but with a longer channel length to ensure that current is cutoff in an inhibited NAND string.

[0091] In this example, the layers 404, 405 and 409 extend continuously in the NAND string. In another approach, portions of the layers 404, 405 and 409 which are between the control gates 402, 412 and 422 can be removed, exposing a top surface of the channel 406.

[0092] FIG. 5A depicts an example block diagram of the sense block SB1 of FIG. 1A. In one approach, a sense block comprises multiple sense circuits. Each sense circuit is associated with data latches. For example, the example sense circuits 550a, 551a, 552a and 553a are associated with the data latches 550b, 551b, 552b and 553b, respectively. In one approach, different subsets of bit lines can be sensed using different respective sense blocks. This allows the processing

load which is associated with the sense circuits to be divided up and handled by a respective processor in each sense block. For example, a sense circuit controller **560** in SB1 can communicate with the set of sense circuits and latches. The sense circuit controller may include a pre-charge circuit **561** which provides a voltage to each sense circuit for setting a pre-charge voltage. In one possible approach, the voltage is provided to each sense circuit independently, e.g., via the data base **503** and a local bus such as LBUS1 or LBUS2 in FIG. 5B. In another possible approach, a common voltage is provided to each sense circuit concurrently, e.g., via the line **505** in FIG. 5B. The sense circuit controller may also include a memory **562** and a processor **563**. As mentioned also in connection with FIG. 2, the memory **562** may store code which is executable by the processor to perform the functions described herein. These functions can include reading latches which are associated with the sense circuits, setting bit values in the latches and providing voltages for setting pre-charge levels in sense nodes of the sense circuits. Further example details of the sense circuit controller and the sense circuits **550a** and **551a** are provided below.

[0093] FIG. 5B depicts another example block diagram of the sense block SB1 of FIG. 1A. The sense circuit controller **560** communicates with multiple sense circuits including example sense circuits **550a** and **551a**, also shown in FIG. 5A. The sense circuit **550a** includes latches **550b**, including a trip latch **526**, an offset verify latch **527** and data state latches **528**. The sense circuit further includes a voltage clamp **521** such as a transistor which sets a pre-charge voltage at a sense node **522**. A sense node to bit line (BL) switch **523** selectively allows the sense node to communicate with a bit line **525**, e.g., the sense node is electrically connected to the bit line so that the sense node voltage can decay. The bit line **525** is connected to one or more memory cells such as a memory cell MC1. A voltage clamp **524** can set a voltage on the bit line, such as during a sensing operation or during a program voltage. A local bus, LBUS1, allows the sense circuit controller to communicate with components in the sense circuit, such as the latches **550b** and the voltage clamp in some cases. To communicate with the sense circuit **550a**, the sense circuit controller provides a voltage via a line **502** to a transistor **504** to connect LBUS1 with a data bus DBUS, **503**. The communicating can include sending data to the sense circuit and/or receive data from the sense circuit.

[0094] The sense circuit controller can communicate with different sense circuits in a time-multiplexed manner, for instance. A line **505** may be connected to the voltage clamp in each sense circuit, in one approach.

[0095] The sense circuit **551a** includes latches **551b**, including a trip latch **546**, an offset verify latch **547** and data state latches **548**. A voltage clamp **541** may be used to set a pre-charge voltage at a sense node **542**. A sense node to bit line (BL) switch **543** selectively allows the sense node to communicate with a bit line **545**, and a voltage clamp **544** can set a voltage on the bit line. The bit line **545** is connected to one or more memory cells such as a memory cell MC2. A local bus, LBUS2, allows the sense circuit controller to communicate with components in the sense circuit, such as the latches **551b** and the voltage clamp in some cases. To communicate with the sense circuit **551a**, the sense circuit controller provides a voltage via a line **501** to a transistor **506** to connect LBUS2 with DBUS.

[0096] The sense circuit **550a** may be a first sense circuit which comprises a first trip latch **526** and the sense circuit **551a** may be a second sense circuit which comprises a second trip latch **546**.

[0097] The sense circuit **550a** is an example of a first sense circuit comprising a first sense node **522**, where the first sense circuit is associated with a first memory cell MC1 and a first bit line **525**. The sense circuit **551a** is an example of a second sense circuit comprising a second sense node **542**, where the second sense circuit is associated with a second memory cell MC2 and a second bit line **545**.

[0098] FIG. 6A is a perspective view of a set of blocks **600** in an example three-dimensional configuration of the memory array **126** of FIG. 1A. On the substrate are example blocks BLK0, BLK1, BLK2 and BLK3 of memory cells (storage elements) and a peripheral area **604** with circuitry for use by the blocks. For example, the circuitry can include voltage drivers **605** which can be connected to control gate layers of the blocks.

[0099] In one approach, control gate layers at a common height in the blocks are commonly driven. The substrate **601** can also carry circuitry under the blocks, along with one or more lower metal layers which are patterned in conductive paths to carry signals of the circuitry. The blocks are formed in an intermediate region **602** of the memory device. In an upper region **603** of the memory device, one or more upper metal layers are patterned in conductive paths to carry signals of the circuitry. Each block comprises a stacked area of memory cells, where alternating levels of the stack represent word lines. In one possible approach, each block has opposing tiered sides from which vertical contacts extend upward to an upper metal layer to form connections to conductive paths. While four blocks are depicted as an example, two or more blocks can be used, extending in the x- and/or y-directions.

[0100] In one possible approach, the length of the plane, in the x-direction, represents a direction in which signal paths to word lines extend in the one or more upper metal layers (a word line or SGD line direction), and the width of the plane, in the y-direction, represents a direction in which signal paths to bit lines extend in the one or more upper metal layers (a bit line direction). The z-direction represents a height of the memory device.

[0101] FIG. 6B depicts an example cross-sectional view of a portion of one of the blocks of FIG. 6A. The block comprises a stack **610** of alternating conductive and dielectric layers. In this example, the conductive layers comprise two SGD layers, two SGS layers and four dummy word line layers DWLD0, DWLD1, DWLS0 and DWLS1, in addition to data word line layers (word lines) WLL0-WLL10. The dielectric layers are labelled as DL0-DL19. Further, regions of the stack which comprise NAND strings NS1 and NS2 are depicted. Each NAND string encompasses a memory hole **618** or **619** which is filled with materials which form memory cells adjacent to the word lines. A region **622** of the stack is shown in greater detail in FIG. 6D.

[0102] The stack includes a substrate **611**, an insulating film **612** on the substrate, and a portion of a source line SL. NS1 has a source-end **613** at a bottom **614** of the stack and a drain-end **615** at a top **616** of the stack. Metal-filled slits **617** and **620** may be provided periodically across the stack as interconnects which extend through the stack, such as to connect the source line to a line above the stack. The slits may be used during the formation of the word lines and

subsequently filled with metal. A portion of a bit line BL0 is also depicted. A conductive via 621 connects the drain-end 615 to BL0.

[0103] FIG. 6C depicts a plot of memory hole diameter in the stack of FIG. 6B. The vertical axis is aligned with the stack of FIG. 6B and depicts a width (wMH), e.g., diameter, of the memory holes 618 and 619. The word line layers WLL0-WLL10 of FIG. 6A are repeated as an example and are at respective heights z0-z10 in the stack. In such a memory device, the memory holes which are etched through the stack have a very high aspect ratio. For example, a depth-to-diameter ratio of about 25-30 is common. The memory holes may have a circular cross-section. Due to the etching process, the memory hole width can vary along the length of the hole. Typically, the diameter becomes progressively smaller from the top to the bottom of the memory hole. That is, the memory holes are tapered, narrowing at the bottom of the stack. In some cases, a slight narrowing occurs at the top of the hole near the select gate so that the diameter becomes slight wider before becoming progressively smaller from the top to the bottom of the memory hole.

[0104] Due to the non-uniformity in the width of the memory hole, the programming speed, including the program slope and erase speed of the memory cells can vary based on their position along the memory hole, e.g., based on their height in the stack. With a smaller diameter memory hole, the electric field across the tunnel oxide is relatively stronger, so that the programming and erase speed is relatively higher. One approach is to define groups of adjacent word lines for which the memory hole diameter is similar, e.g., within a defined range of diameter, and to apply an optimized verify scheme for each word line in a group. Different groups can have different optimized verify schemes.

[0105] FIG. 6D depicts a close-up view of the region 622 of the stack of FIG. 6B. Memory cells are formed at the different levels of the stack at the intersection of a word line layer and a memory hole. In this example, SGD transistors 680 and 681 are provided above dummy memory cells 682 and 683 and a data memory cell MC. A number of layers can be deposited along the sidewall (SW) of the memory hole 630 and/or within each word line layer, e.g., using atomic layer deposition. For example, each column (e.g., the pillar which is formed by the materials within a memory hole) can include a charge-trapping layer or film 663 such as SiN or other nitride, a tunneling layer 664, a polysilicon body or channel 665, and a dielectric core 666. A word line layer can include a blocking oxide/block high-k material 660, a metal barrier 661, and a conductive metal 662 such as Tungsten as a control gate. For example, control gates 690, 691, 692, 693 and 694 are provided. In this example, all of the layers except the metal are provided in the memory hole. In other approaches, some of the layers can be in the control gate layer. Additional pillars are similarly formed in the different memory holes. A pillar can form a columnar active area (AA) of a NAND string.

[0106] When a memory cell is programmed, electrons are stored in a portion of the charge-trapping layer which is associated with the memory cell. These electrons are drawn into the charge-trapping layer from the channel, and through the tunneling layer. The V_{th} of a memory cell is increased in proportion to the amount of stored charge. During an erase operation, the electrons return to the channel.

[0107] Each of the memory holes can be filled with a plurality of annular layers comprising a blocking oxide layer, a charge trapping layer, a tunneling layer and a channel layer. A core region of each of the memory holes is filled with a body material, and the plurality of annular layers are between the core region and the word line in each of the memory holes.

[0108] The NAND string can be considered to have a floating body channel because the length of the channel is not formed on a substrate. Further, the NAND string is provided by a plurality of word line layers above one another in a stack, and separated from one another by dielectric layers.

[0109] FIG. 7A depicts a top view of an example word line layer WLL0 of the stack of FIG. 6B. As mentioned, a 3D memory device can comprise a stack of alternating conductive and dielectric layers. The conductive layers provide the control gates of the SG transistors and memory cells. The layers used for the SG transistors are SG layers and the layers used for the memory cells are word line layers. Further, memory holes are formed in the stack and filled with a charge-trapping material and a channel material. As a result, a vertical NAND string is formed. Source lines are connected to the NAND strings below the stack and bit lines are connected to the NAND strings above the stack.

[0110] A block BLK in a 3D memory device can be divided into sub-blocks, where each sub-block comprises a set of NAND string which have a common SGD control line. For example, see the SGD lines/control gates SGD0, SGD1, SGD2 and SGD3 in the sub-blocks SBa, SBb, SBc and SBd, respectively. The sub-blocks SBa, SBb, SBc and SBd may also be referred herein as a string of memory cells of a word line. As described, a string of memory cells of a word line may include a plurality of memory cells that are part of the same sub-block, and that are also disposed in the same word line layer and/or that are configured to have their control gates biased by the same word line and/or with the same word line voltage.

[0111] Further, a word line layer in a block can be divided into regions. Each region is in a respective sub-block and can extend between slits which are formed periodically in the stack to process the word line layers during the fabrication process of the memory device. This processing can include replacing a sacrificial material of the word line layers with metal. Generally, the distance between slits should be relatively small to account for a limit in the distance that an etchant can travel laterally to remove the sacrificial material, and that the metal can travel to fill a void which is created by the removal of the sacrificial material. For example, the distance between slits may allow for a few rows of memory holes between adjacent slits. The layout of the memory holes and slits should also account for a limit in the number of bit lines which can extend across the region while each bit line is connected to a different memory cell. After processing the word line layers, the slits can optionally be filled with metal to provide an interconnect through the stack.

[0112] These figures and other are not necessarily to scale. In practice, the regions can be much longer in the x-direction relative to the y-direction than is depicted to accommodate additional memory holes.

[0113] In this example, there are four rows of memory holes between adjacent slits. A row here is a group of memory holes which are aligned in the x-direction. Moreover, the rows of memory holes are in a staggered pattern to

increase the density of the memory holes. The word line layer or word line is divided into regions WLL0 a, WLL0 b, WLL0 c and WLL0 d which are each connected by a connector **713**. The last region of a word line layer in a block can be connected to a first region of a word line layer in a next block, in one approach. The connector, in turn, is connected to a voltage driver for the word line layer. The region WLL0 a has example memory holes **710** and **711** along a line **712**. The region WLL0 b has example memory holes **714** and **715**. The region WLL0 c has example memory holes **716** and **717**. The region WLL0 d has example memory holes **718** and **719**. The memory holes are also shown in FIG. 7B. Each memory hole can be part of a respective NAND string. For example, the memory holes **710**, **714**, **716** and **718** can be part of NAND strings NS0_SBa, NS0_SBb, NS0_SBc and NS0_SBd, respectively.

[0114] Each circle represents the cross-section of a memory hole at a word line layer or SG layer. Example circles shown with dashed lines represent memory cells which are provided by the materials in the memory hole and by the adjacent word line layer. For example, memory cells **720** and **721** are in WLL0 a, memory cells **724** and **725** are in WLL0 b, memory cells **726** and **727** are in WLL0 c, and memory cells **728** and **729** are in WLL0 d. These memory cells are at a common height in the stack.

[0115] Metal-filled slits **701**, **702**, **703** and **704** (e.g., metal interconnects) may be located between and adjacent to the edges of the regions WLL0 a-WLL0 d. The metal-filled slits provide a conductive path from the bottom of the stack to the top of the stack. For example, a source line at the bottom of the stack may be connected to a conductive line above the stack, where the conductive line is connected to a voltage driver in a peripheral region of the memory device. See also FIG. 8A for further details of the sub-blocks SBa-SBd of FIG. 7A.

[0116] FIG. 7B depicts a top view of an example top dielectric layer DL19 of the stack of FIG. 6B. The dielectric layer is divided into regions DL19 a, DL19 b, DL19 c and DL19 d. Each region can be connected to a respective voltage driver. This allows a set of memory cells in one region of a word line layer to be programmed concurrently, with each memory cell being in a respective NAND string which is connected to a respective bit line. A voltage can be set on each bit line to allow or inhibit programming during each program voltage.

[0117] The region DL19a has the example memory holes **710** and **711** along a line **712a** which is coincident with a bit line BL0. A number of bit lines extend above the memory holes and are connected to the memory holes as indicated by the "X" symbols. BL0 is connected to a set of memory holes which includes the memory holes **711**, **715**, **717** and **719**. Another example bit line BL1 is connected to a set of memory holes which includes the memory holes **710**, **714**, **716** and **718**. The metal-filled slits **701**, **702**, **703** and **704** from FIG. 7A are also depicted, as they extend vertically through the stack. The bit lines can be numbered in a sequence BL0-BL23 across the DL19 layer in the -x direction.

[0118] Different subsets of bit lines are connected to cells in different rows. For example, BL0, BL4, BL8, BL12, BL16 and BL20 are connected to cells in a first row of cells at the right hand edge of each region. BL2, BL6, BL10, BL14, BL18 and BL22 are connected to cells in an adjacent

row of cells, adjacent to the first row at the right hand edge. BL3, BL7, BL11, BL15, BL19 and BL23 are connected to cells in a first row of cells at the left hand edge of each region. BL1, BL5, BL9, BL13, BL17 and BL21 are connected to cells in an adjacent row of cells, adjacent to the first row at the left hand edge.

[0119] FIG. 8A depicts example NAND strings in the sub-blocks SBa-SBd of FIG. 7A. The sub-blocks are consistent with the structure of FIG. 6B. The conductive layers in the stack are depicted for reference at the left hand side. Each sub-block includes multiple NAND strings, where one example NAND string is depicted. For example, SBa comprises an example NAND string NS0_SBa, SBb comprises an example NAND string NS0_SBb, SBc comprises an example NAND string NS0_SBc, and SBd comprises an example NAND string NS0_SBd.

[0120] Additionally, NS0_SBa include SGS transistors **800** and **801**, dummy memory cells **802** and **803**, data memory cells **804**, **805**, **806**, **807**, **808**, **809**, **810**, **811**, **812**, **813** and **814**, dummy memory cells **815** and **816**, and SGD transistors **817** and **818**.

[0121] NS0_SBb include SGS transistors **820** and **821**, dummy memory cells **822** and **823**, data memory cells **824**, **825**, **826**, **827**, **828**, **829**, **830**, **831**, **832**, **833** and **834**, dummy memory cells **835** and **836**, and SGD transistors **837** and **838**.

[0122] NS0_SBc include SGS transistors **840** and **841**, dummy memory cells **842** and **843**, data memory cells **844**, **845**, **846**, **847**, **848**, **849**, **850**, **851**, **852**, **853** and **854**, dummy memory cells **855** and **856**, and SGD transistors **857** and **858**.

[0123] NS0_SBd include SGS transistors **860** and **861**, dummy memory cells **862** and **863**, data memory cells **864**, **865**, **866**, **867**, **868**, **869**, **870**, **871**, **872**, **873** and **874**, dummy memory cells **875** and **876**, and SGD transistors **877** and **878**.

[0124] At a given height in the block, a set of memory cells in each sub-block are at a common height. For example, one set of memory cells (including the memory cell **804**) is among a plurality of memory cells formed along tapered memory holes in a stack of alternating conductive and dielectric layers. The one set of memory cells is at a particular height z0 in the stack. Another set of memory cells (including the memory cell **824**) connected to the one word line (WLL0) are also at the particular height. In another approach, the set of memory cells (e.g., including the memory cell **812**) connected to another word line (e.g., WLL8) are at another height (z8) in the stack.

[0125] FIG. 8B depicts another example view of NAND strings in sub-blocks. The NAND strings includes NS0_SBa, NS0_SBb, NS0_SBc and NS0_SBd, which have 48 word lines, WL0-WL47, in this example. Each sub-block comprises a set of NAND strings which extend in the x direction and which have a common SGD line, e.g., SGD0, SGD1, SGD2 or SGD3. In this simplified example, there is only one SGD transistor and one SGS transistor in each NAND string. The NAND strings NS0_SBa, NS0_SBb, NS0_SBc and NS0_SBd are in sub-blocks SBa, SBb, SBc and SBd, respectively. Further, example, groups of word lines G0, G1 and G2 are depicted.

[0126] FIG. 8C generally illustrates a schematic view of three versions of staggered string architecture **101**, **103**, **105** for BiCS memory, e.g., NAND. With reference the string architecture **101**, the strings are shown in rows **107-0**

through 107-7 in architecture 101. Each row is shown with four ends to the strings. A string may be connected to an adjacent string at an end (not visible beneath this view). A first group of rows 107-0 through 107-3 are shown on a left side of a dummy row 108. A second group of rows 107-4 through 107-7 are shown on a right side of the dummy row 108. The dummy row 108 separates the two groups of rows in the staggered eight row. A source line 109 is positioned at an edge of the first group and is remote from the dummy row 108. A source line 110 is positioned at an edge of the second group and is remote from the dummy row 108 and source line 109.

[0127] The staggered string architectures 103, 105 for BiCS memory are similar to that of architecture 101 except additional groups are added. Architecture 103 is double the size of architecture 101 and includes sixteen rows of strings with each group of four rows separated by a dummy row. Architecture 105 is larger than both the architecture 101 and the architecture 103. Architecture 105 includes twenty rows of strings with each group of four rows separated by a dummy row 108.

[0128] These architectures 101, 103, 105 can include a chip under array structure, e.g., the control circuitry is under the memory array that can include the groups of memory strings. With the chip under array structure, the strings may include a direct strap contact for the source line for read and erase operations.

[0129] Reverse Order Read (ROR) provides improvements to the time to perform a read operation (Tread) over Normal Order Read (NOR) (e.g., by ~5-10%). For example, one time savings provided by ROR includes VREAD spike discharge time. Additionally, ROR is compatible with Reverse Order Verify (ROV), which is mostly used for program verify (PVFY). This helps improve Vth window (e.g., up to ~50 mV gain is seen in reverse order verify (ROV) and ROR as compared to ROV and NOR).

[0130] However, ROR is not compatibility with lockout mode. Lockout mode discharges bit lines that were “Read ON” at a prior read, thus helping to reduce the total amount of cells conductive on subsequent reads (i.e., RWL_CLK reads) which saves ICC. Because ROR does higher read level sensing first, the lockout mode concept does not work for ROR.

[0131] To address the above, lockout mode can be implemented for a ROR operation. FIGS. 12-15 depict embodiments of a voltage waveform (e.g., VCG) corresponding with the word line voltages applied to a selected word line within a memory block during consecutive read operations associated with the same word line. For example, in FIG. 12, the voltage waveform depicts a ROR operation without lockout mode. In FIG. 12, a ROR operation of executed read levels, S12-S6-S3-S1, for a lower page (LP) is depicted following a read pass voltage (VREAD) spike. As shown in table 1202, each page is associated with a plurality of read levels.

[0132] FIG. 13 depicts implementation of a pre-lockout read for ROR operation with lockout mode. A pre-lockout read may be performed at a read level that is in the middle of the read levels being executed for a given page read. For example, as shown in FIG. 13, for a LP read (where read levels S12-S6-S3-S1 are executed) the pre-lockout read level may be performed at a read level that is in between S4 and S5 read level. Additionally, with continued reference to FIG. 13, any memory cell that is read as “OFF” in the

pre-lockout read is sensed in the first two read operations (e.g., S12 and S6) and all other memory cells that were read as “ON” are locked out. In contrast, in the subsequent reads (e.g., S3 and S1), the memory cells that were judged ON at the pre-lock out read are sensed while the other cells that were judged OFF are locked out. In some embodiments, in the subsequent reads (e.g., S3 and S1), all memory cells may be sensed. The same process is implemented in subsequent page reads.

[0133] By doing this, the numbers of cells conducting are reduced (especially in first two read levels, such as S12 and S6 read levels), thereby reducing ICC in a similar manner to lockout mode in NOR. Further, the pre-lockout read may be a rough sensing (e.g., shorter sensing time, BL/WL-settling time, etc.) to reduce Tread.

[0134] As represented in FIG. 14, a pre-lockout read for ROR operation with lockout mode may be implemented without a Vread spike operation (which may prevent electron injection disturb). If a Vread spike operation is not performed, then the pre-lockout read operation may be performed without further Tread penalty. FIG. 15 depicts a voltage waveform for a pre-lockout read for a sequential read. In some embodiments, an optimized state for a pre-lockout read state may be balanced among all data states. For example, for QLC, an optimized state for pre-lockout read may be S7 for a plurality of states (e.g., S1, S2, S3 . . . S15).

[0135] As discussed, the ramp rates used during the charging and discharging of word line during a read operation affect read time and current consumption of the memory device or apparatus. Current consumption (ICC) is a very important metric for power efficient 3D NAND operation. FIG. 16 shows plots of current consumption during an example read operation of a lower page, middle page, and upper page of a block of a memory apparatus with triple level cells and for a read operation of a block of a memory apparatus with single level cells that is erased and randomly programmed using different ramp rates. FIG. 17 is a reference waveform for an example read operation of a page and illustrates the importance of target discharge bias CGDIS_RATE (i.e., the discharge rate to the control gate ready voltage CG_READY during a discharge stage of a read operation). As shown, the rate of charging-up in an initial stage R clk (ramping up to a read pass voltage VREAD in FIG. 17) and the rate of discharge in a discharge stage RR clk (ramping down toward the control gate ready voltage CG_READY) contribute significantly to current consumption. Reducing the charging-up rate helps in reducing peak current consumption, which generally is seen in the initial stage R clk, whereas reducing the discharge rate improves the discharge stage RR clk peak current consumption. Improvement in R2 clk (during the initial stage R clk) peak current consumption for a programmed block is approximately 25% in SLC and approximately 9% in TLC. For an erased block (i.e., closer to open block with one word line programmed), it is approximately 39%. This also makes R2 clk peak current consumption similar between closed block (i.e., all word lines programmed) and open block (i.e., no word lines programmed). With target discharge bias CGDIS_RATE equal to a supply voltage VCC, improvement in discharge stage RR clk peak current consumption for a programmed block is approximately 25% in SLC and approximately 22% in TLC.

[0136] Consequently, described herein is a memory apparatus (e.g., memory device **100** in FIG. **1A**) including memory cells (e.g., data memory cell transistors **482-484** in FIG. **6D**) configured to retain a threshold voltage V_{th} corresponding to one of a plurality of data states (e.g., Er, A, B and C data states in FIG. **9**, Er, A, B, C, D, E, F and G data states in FIG. **10**, and Er, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E and F data states in FIG. **11**). The memory apparatus also includes a control means (e.g., control circuitry **110**, controller **122**, decoders **124**, **132**, read/write circuits **128**, and sense blocks SB1, SB2, SBp in FIG. **1A**, control circuit **150** in FIG. **1B**) configured to apply at least one read voltage (e.g., read levels S1-S15 in FIGS. **12-15**) associated with one of the plurality of data states to ones of the plurality of word lines connected to the memory cells being read in one read operation. The control means is also configured to adjust a voltage applied to the ones of the plurality of word lines during at least one of an end of the one read operation or a beginning of a subsequent read operation of the memory cells based on at least one of the subsequent read operation following the one read operation within a predetermined time or a control gate ready voltage CG_READY of the ones of the plurality of word lines targeted following the one read operation.

[0137] For successive reads which are scheduled to happen within a pre-determined threshold time duration within the same block and same string, one approach is to have the word lines discharge at the end of each read in the discharge stage RR clk and charge-up in the beginning of each read in the initial stage R clk. FIG. **18** shows an example waveform of a voltage applied to word lines for consecutive reads of all pages on the same word line and string (several kick biases are not shown). Such successive reads occur whether for two different pages or of the same page on the same word line, string or for different combinations in the same block. As discussed above, the memory cells can be disposed in memory holes (e.g., memory hole **630** in FIG. **6D**) grouped into a plurality of strings (e.g., sub-blocks SBa-SBd in FIG. **7A**), the plurality of strings comprise a block (e.g., block BLK in FIG. **7A**). Thus, the control means, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation. The control means is additionally configured to skip discharging the ones of the plurality of word lines during the end of the one read operation and a ramping up of the ones of the plurality of word lines to a read pass voltage VREAD in response to the at least one of the subsequent read operation involving the memory cells of the one of the plurality of strings of the block and occurring within the predetermined time following the one read operation. The control means is also configured to apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the subsequent read operation.

[0138] Thus, the discharge stage RR clk may be removed and the following read pass voltage VREAD spike during the initial stage R clk can also be removed for any successive reads on the same block and string. This includes situations such as reading the same page in case of optimal reads,

direct look ahead DLA read, consecutive reads of different pages on the same word line and/or string, or read transition from a word line WLn to an adjacent word line WLn+1 or vice versa, if on the same string. FIG. **19** shows an example waveform of a voltage applied to word lines for consecutive reads of all pages on the same word line and string with the discharge stage RR clk and the following read pass voltage VREAD spike removed.

[0139] The removal of the discharge stage RR clk and the following read pass voltage VREAD spike may be accompanied by a wait time (i.e., during the period indicated as ready in FIG. **17**) in between when the control gate line is disconnected from the word line. Under such circumstances, a BLKSEL signal that controls whether the word lines are connected to the control gate or not could be either kept on between the reads or could be optionally turned off and turned back on at the beginning and the end of the wait time respectively. If the BLKSEL signal is turned off and turned on, there might be disturb during the wait time. Hence, this approach is not recommended. If the BLKSEL signal is kept on, some of the selected memory holes with conducting memory cells might continue to conduct, which will decrease the improvement in current consumption. So, for this approach to be useful, wait times between read operations should be very small. Also, system can schedule immediate next read or read operation (i.e., back-to-back reads) on the same string to save both current consumption and read time t_R . For applications where the reads are immediate most of the time, such as optimal reads, DLA reads etc., this could be applied even without additional read scheduling.

[0140] For successive reads which are scheduled to happen within a pre-determined threshold time duration within the same block, but different string, a different target discharge bias CGDIS_RATE may be used for the discharge stage RR clk. FIG. **20** shows an example waveform of a voltage applied to word lines for consecutive reads of all pages on the same word line but different string with a different target discharge bias CGDIS_RATE used for the discharge stage RR clk. The end of the one read operation can comprise the discharge stage RR clk. Thus, the control means, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation. In addition, the control means is configured to ramp the voltage applied to the ones of the plurality of word lines up to a discharge kick voltage (i.e., the voltage spike applied to the word lines immediately prior to ramping down in the discharge stage) and back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted at a predetermined discharge ramp rate during the discharge stage RR clk. The control means is additionally configured to adjust the predetermined discharge ramp rate of the voltage applied to the ones of the plurality of word lines during the discharge stage RR clk in response to the at least one of the subsequent read operation involving the memory cells of another of the plurality of strings of the block and occurring within the predetermined time following the one read operation. The control means is also configured to apply the at least one read voltage associated with the one of the plurality of data states to the

ones of the word lines connected to the memory cells of the another of the plurality of strings of the block in the subsequent read operation.

[0141] An extension of such an approach is to skip the double ramp-up in the reads after CGDIS_RATE_SR and reduce the timing. FIG. 21 shows an example waveform of a voltage applied to word lines for consecutive reads of all pages on the same word line but different string with the double ramp-up in the reads after CGDIS_RATE_SR skipped. This approach will also improve read time t_R and may be accompanied by a wait time in-between when the control gate line is disconnected from the word line. Under such circumstances, control gate lines could be maintained at CGDIS_RATE_SR. The extension could also have control gate line bias not disconnected during the wait time and maintained at CGDIS_RATE_SR.

[0142] One of the ways to minimize the discharge stage RR clk current consumption is to discharge in two or more steps. These steps and corresponding timings can be dependent on the value of the control gate ready voltage CG_READY. FIG. 22 shows a waveform for an example read operation of a page along with examples of the discharge including two or more steps. Thus, according to an aspect, the control means is further configured to apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation. The control means is also configured to ramp the voltage applied to the ones of the plurality of word lines up to a discharge kick voltage and back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted in a stepwise fashion with a plurality of discharge steps lasting each of a plurality of discharge time periods during the discharge stage RR clk. At least one of the plurality of discharge steps or the plurality of discharge time periods can be dependent on the control gate ready voltage CG_READY. Accordingly, a better current consumption as well as better read time t_R are achieved, thereby improving read time t_R -current consumption trade-off.

[0143] According to additional aspects, in response to the plurality of discharge steps comprising two steps and a first discharge step voltage VSTEP_1 at which a first discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage CG_READY, the control means is further configured to ramp the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted using the first discharge step lasting a first discharge step time period and ending at the first discharge step voltage VSTEP_1 and a second discharge step of the plurality of discharge steps lasting a second discharge step time period and ending at a second discharge step voltage VSTEP_2. In other words, if the discharge is in two steps, then use the two-step discharge if the first discharge step voltage VSTEP_1 is greater than the control gate ready voltage CG_READY.

[0144] In response to the plurality of discharge steps comprising two steps and the first discharge step voltage VSTEP_1 at which the first discharge step of the plurality of discharge steps ends being less than or equal to the control gate ready voltage CG_READY, the control means is additionally configured to ramp the voltage applied to the ones of the plurality of word lines back down toward the control

gate ready voltage CG_READY of the ones of the plurality of word lines targeted using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage VSTEP_1 thereby eliminating the second discharge step time period. So, if the first discharge step voltage VSTEP_1 is less than or equal to the control gate ready voltage CG_READY, then have a single-step discharge while eliminating a portion of the discharge time RR7 (i.e., the second discharge step time period), which saves read time t_R .

[0145] In response to the plurality of discharge steps comprising three steps and the second discharge step voltage VSTEP_2 at which the second discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage CG_READY, the control means is configured to ramp the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted using the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage VSTEP_1 and the second discharge step of the plurality of discharge steps lasting the second discharge step time period and ending at the second discharge step voltage VSTEP_2 and a third discharge step of the plurality of discharge steps lasting a third discharge step time period and ending at a third discharge step voltage (e.g., the control gate ready voltage CG_READY). Thus, if the discharge is in three steps, then use the three-step discharge if the second discharge step voltage VSTEP_2 is greater than the control gate ready voltage CG_READY.

[0146] In response to the plurality of discharge steps comprising three steps and the first discharge step voltage VSTEP_1 at which the first discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage CG_READY and the second discharge step voltage VSTEP_2 at which the second discharge step of the plurality of discharge steps ends is less than the control gate ready voltage CG_READY, the control means ramps the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage VSTEP_1 and the second discharge step lasting the second discharge step time period and ending at the second discharge step voltage VSTEP_2 thereby eliminating the third discharge step time period. So, the two-step discharge is used if the first discharge step voltage VSTEP_1 is greater than the control gate ready voltage CG_READY and the control gate ready voltage CG_READY is greater than or equal to the second discharge step voltage VSTEP_2, while eliminating a portion of the discharge time RR8 (i.e., the third discharge step time period), which saves read time t_R .

[0147] In response to the plurality of discharge steps comprising three steps and the first discharge step voltage VSTEP_1 at which the first discharge step of the plurality of discharge steps ends being less than or equal to the control gate ready voltage CG_READY, the control means is also configured to ramp the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted using only the first discharge step lasting the first discharge step time period and ending at the first

discharge step voltage VSTEP_1 thereby eliminating the second discharge step time period and the third discharge step time period. In other words, if the first discharge step voltage VSTEP_1 is less than or equal to the control gate ready voltage CG_READY, then have a single-step discharge while eliminating a portion of the discharge time RR7 (i.e., the second discharge step time period), which further saves the read time tR.

[0148] Additionally, if the next command which will be executed is a read operation on the same block, then the stepwise discharge and the corresponding additional times can be eliminated. FIG. 23 shows changes to a waveform for an example read operation of a page along with examples of the stepwise discharge and the corresponding additional times being eliminated. Again, this includes situations such as reading the same page in case of optimal reads, direct look ahead DLA read, consecutive reads of different pages on the same word line and/or string, or read transition from word line WLn to an adjacent word line WLn+1 or vice versa. So, according to an aspect, the control means, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to ramp the voltage applied to the ones of the plurality of word lines up to an intermediate voltage (e.g., a supply voltage VCC or VDD) during a first beginning period R1 of a beginning of the one read operation (i.e., initial period R1 clk) and up to a read pass voltage during a second beginning period R2 of the beginning of the one read operation and back down to the at least one read voltage associated with the one of the plurality of data states in the one read operation.

[0149] According to another aspect, the stepwise biases VSTEP_1, VSTEP 2 applied to the word lines during the wait time (if any) may also be maintained. Thus, the control means is additionally configured to ramp the voltage applied to the ones of the plurality of word lines up to the discharge kick voltage and back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage VSTEP_1 thereby eliminating any other discharge step time periods other than the first discharge step time period. The control means maintains the voltage applied to the ones of the plurality of word lines at the first discharge step voltage VSTEP_1 during a wait time between the one read operation and the subsequent read operation.

[0150] According to further aspects, a portion of the initial stage R clk, R1 clk (during which the voltage applied to the word line is ramped up to the intermediate voltage) may also be eliminated if the stepwise bias is greater than the magnitude of the first ramp-up step (i.e., the intermediate voltage). Accordingly, the control means is also configured to skip ramping the voltage applied to the ones of the plurality of word lines up to the intermediate voltage during a first subsequent beginning period R1 of the beginning of the subsequent read operation and ramp the voltage applied to the ones of the plurality of word lines directly from the first discharge step voltage VSTEP_1 up to the read pass voltage during a second subsequent beginning period R2 of the beginning of the subsequent read operation thereby eliminating the first subsequent beginning period R1 of the beginning of the subsequent read operation.

[0151] According to another aspect, the timing of the second subsequent beginning period R2 can be changed to an adjusted second subsequent beginning period R2' shorter

than the second subsequent beginning period R2. So, the control means, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to adjust the second subsequent beginning period R2 of the beginning of the subsequent read operation to an adjusted second subsequent beginning period R2' such that the adjusted second subsequent beginning period R2' is shorter than the second subsequent beginning period R2.

[0152] Now referring to FIG. 24, a method of operating a memory apparatus is also provided. As discussed above, the memory apparatus (e.g., memory device 100 in FIG. 1A) includes memory cells (e.g., data memory cell transistors 482-484 in FIG. 6D) configured to retain a threshold voltage Vth corresponding to one of a plurality of data states (e.g., Er, A, B and C data states in FIG. 9, Er, A, B, C, D, E, F and G data states in FIG. 10, and Er, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E and F data states in FIG. 11). The method includes the step of 2400 applying at least one read voltage (e.g., read levels S1-S15 in FIGS. 12-15) associated with one of the plurality of data states to ones of the plurality of word lines connected to the memory cells being read in one read operation. The method also includes the step of 2402 adjusting a voltage applied to the ones of the plurality of word lines during at least one of an end of the one read operation or a beginning of a subsequent read operation of the memory cells based on at least one of the subsequent read operation following the one read operation within a predetermined time or a control gate ready voltage CG_READY of the ones of the plurality of word lines targeted following the one read operation.

[0153] As discussed above and with reference back to FIG. 19, for successive reads which are scheduled to happen within the pre-determined threshold time duration within the same block and same string, one approach is to have the word lines discharge at the end of each read in the discharge stage RR clk and charge-up in the beginning of each read in the initial stage R clk. Therefore, the method, in response to the subsequent read operation following the one read operation within the predetermined time, further includes the step of applying the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation. The method continues by skipping discharging the ones of the plurality of word lines during the end of the one read operation and a ramping up of the ones of the plurality of word lines to a read pass voltage in response to the at least one of the subsequent read operation involving the memory cells of the one of the plurality of strings of the block and occurring within the predetermined time following the one read operation. The method also includes the step of applying the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the subsequent read operation.

[0154] Again referring back to FIG. 21, for successive reads which are scheduled to happen within the pre-determined threshold time duration within the same block, but different string, the different target discharge bias CGDIS_RATE may be used for the discharge stage RR clk. Thus, the method, in response to the subsequent read operation following the one read operation within the predetermined time, further includes the step of applying the at least one

read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation. Next, ramping the voltage applied to the ones of the plurality of word lines up to a discharge kick voltage and back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted at a predetermined discharge ramp rate during the discharge stage RR clk. The method continues by adjusting the predetermined discharge ramp rate of the voltage applied to the ones of the plurality of word lines during the discharge stage RR clk in response to the at least one of the subsequent read operation involving the memory cells of another of the plurality of strings of the block and occurring within the predetermined time following the one read operation. The method also includes the step of applying the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of the another of the plurality of strings of the block in the subsequent read operation.

[0155] As above, with reference back to FIG. 22, one way to minimize the discharge stage RR clk current consumption is to discharge in two or more steps. In addition, these steps and corresponding timings can be dependent on the value of the control gate ready voltage CG_READY. Therefore, the method further includes the step of applying the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation. The method also includes the step of ramping the voltage applied to the ones of the plurality of word lines up to a discharge kick voltage and back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted in a stepwise fashion with a plurality of discharge steps lasting each of a plurality of discharge time periods during the discharge stage RR clk. The at least one of the plurality of discharge steps or the plurality of discharge time periods are dependent on the control gate ready voltage CG_READY.

[0156] Again, according to additional aspects, in response to the plurality of discharge steps comprising two steps and a first discharge step voltage VSTEP_1 at which a first discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage CG_READY, the method further includes the step of ramping the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted using the first discharge step lasting a first discharge step time period and ending at the first discharge step voltage VSTEP_1 and a second discharge step of the plurality of discharge steps lasting a second discharge step time period and ending at a second discharge step voltage VSTEP_2.

[0157] In response to the plurality of discharge steps comprising two steps and the first discharge step voltage VSTEP_1 at which the first discharge step of the plurality of discharge steps ends being less than or equal to the control gate ready voltage CG_READY, the next step of the method is ramping the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted using only the first discharge step lasting the first

discharge step time period and ending at the first discharge step voltage VSTEP_1 thereby eliminating the second discharge step time period.

[0158] Then, in response to the plurality of discharge steps comprising three steps and the second discharge step voltage VSTEP_2 at which the second discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage CG_READY, the method continues by ramping the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted using the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage VSTEP_1 and the second discharge step of the plurality of discharge steps lasting the second discharge step time period and ending at the second discharge step voltage VSTEP_2 and a third discharge step of the plurality of discharge steps lasting a third discharge step time period and ending at a third discharge step voltage (e.g., the control gate ready voltage CG_READY).

[0159] Furthermore, in response to the plurality of discharge steps comprising three steps and the first discharge step voltage VSTEP_1 at which the first discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage CG_READY and the second discharge step voltage VSTEP_2 at which the second discharge step of the plurality of discharge steps ends being less than the control gate ready voltage CG_READY, the method additionally includes the step of ramping the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage VSTEP_1 and the second discharge step lasting the second discharge step time period and ending at the second discharge step voltage VSTEP_2 thereby eliminating the third discharge step time period.

[0160] In addition, in response to the plurality of discharge steps comprising three steps and the first discharge step voltage VSTEP_1 at which the first discharge step of the plurality of discharge steps ends being less than or equal to the control gate ready voltage CG_READY, the method also includes the step of ramping the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines targeted using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage VSTEP_1 thereby eliminating the second discharge step time period and the third discharge step time period.

[0161] Once again with reference back to FIG. 23, if the next command which will be executed is a read operation on the same block, then the stepwise discharge and the corresponding additional times can be eliminated. Thus, the method, in response to the subsequent read operation following the one read operation within the predetermined time, further includes the step of ramping the voltage applied to the ones of the plurality of word lines up to an intermediate voltage during a first beginning period R1 of a beginning of the one read operation and up to a read pass voltage during a second beginning period R2 of the beginning of the one read operation and back down to the at least one read voltage associated with the one of the plurality of data states

in the one read operation. Next, ramping the voltage applied to the ones of the plurality of word lines up to the discharge kick voltage and back down toward the control gate ready voltage CG_READY of the ones of the plurality of word lines using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage VSTEP_1 thereby eliminating any other discharge step time periods other than the first discharge step time period. The method further includes the step of maintaining the voltage applied to the ones of the plurality of word lines at the first discharge step voltage VSTEP_1 during a wait time between the one read operation and the subsequent read operation. The method also includes the step of skipping ramping the voltage applied to the ones of the plurality of word lines up to the intermediate voltage during a first subsequent beginning period R1 of the beginning of the subsequent read operation and ramp the voltage applied to the ones of the plurality of word lines directly from the first discharge step voltage VSTEP_1 up to the read pass voltage during a second subsequent beginning period R2 of the beginning of the subsequent read operation thereby eliminating the first subsequent beginning period R1 of the beginning of the subsequent read operation.

[0162] As above, according to another aspect, the timing of the second subsequent beginning period R2 can be changed to the adjusted second subsequent beginning period R2' shorter than the second subsequent beginning period R2. So, the method, in response to the subsequent read operation following the one read operation within the predetermined time, further includes the step of adjusting the second subsequent beginning period R2 of the beginning of the subsequent read operation to an adjusted second subsequent beginning period R2' such that the adjusted second subsequent beginning period R2' is shorter than the second subsequent beginning period R2.

[0163] Thus, the memory apparatus and method of operation disclosed herein result in current consumption improvement. Furthermore, effective read time tR is improved as well.

[0164] Clearly, changes may be made to what is described and illustrated herein without, however, departing from the scope defined in the accompanying claims. The foregoing description of the embodiments has been provided for purposes of illustration and description. It is not intended to be exhaustive or to limit the disclosure. Individual elements or features of a particular embodiment are generally not limited to that particular embodiment, but, where applicable, are interchangeable and can be used in a selected embodiment, even if not specifically shown or described. The same may also be varied in many ways. Such variations are not to be regarded as a departure from the disclosure, and all such modifications are intended to be included within the scope of the disclosure.

[0165] The terminology used herein is for the purpose of describing particular example embodiments only and is not intended to be limiting. As used herein, the singular forms “a,” “an,” and “the” may be intended to include the plural forms as well, unless the context clearly indicates otherwise. The terms “comprises,” “comprising,” “including,” and “having,” are inclusive and therefore specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof. The method

steps, processes, and operations described herein are not to be construed as necessarily requiring their performance in the particular order discussed or illustrated, unless specifically identified as an order of performance. It is also to be understood that additional or alternative steps may be employed.

[0166] When an element or layer is referred to as being “on,” “engaged to,” “connected to,” or “coupled to” another element or layer, it may be directly on, engaged, connected or coupled to the other element or layer, or intervening elements or layers may be present. In contrast, when an element is referred to as being “directly on,” “directly engaged to,” “directly connected to,” or “directly coupled to” another element or layer, there may be no intervening elements or layers present. Other words used to describe the relationship between elements should be interpreted in a like fashion (e.g., “between” versus “directly between,” “adjacent” versus “directly adjacent,” etc.). As used herein, the term “and/or” includes any and all combinations of one or more of the associated listed items.

[0167] Although the terms first, second, third, etc. may be used herein to describe various elements, components, regions, layers and/or sections, these elements, components, regions, layers and/or sections should not be limited by these terms. These terms may be only used to distinguish one element, component, region, layer or section from another region, layer or section. Terms such as “first,” “second,” and other numerical terms when used herein do not imply a sequence or order unless clearly indicated by the context. Thus, a first element, component, region, layer or section discussed below could be termed a second element, component, region, layer or section without departing from the teachings of the example embodiments.

[0168] Spatially relative terms, such as “inner,” “outer,” “beneath,” “below,” “lower,” “above,” “upper,” “top,” “bottom,” and the like, may be used herein for ease of description to describe one element's or feature's relationship to another element(s) or feature(s) as illustrated in the figures. Spatially relative terms may be intended to encompass different orientations of the device in use or operation in addition to the orientation depicted in the figures. For example, if the device in the figures is turned over, elements described as “below” or “beneath” other elements or features would then be oriented “above” the other elements or features. Thus, the example term “below” can encompass both an orientation of above and below. The device may be otherwise oriented (rotated 90 degrees or at other orientations) and the spatially relative descriptions used herein interpreted accordingly.

What is claimed is:

1. A memory apparatus, comprising:

memory cells each connected to one of a plurality of word lines and configured to store a threshold voltage corresponding to one of a plurality of data states; and
a control means configured to:

apply at least one read voltage associated with one of the plurality of data states to ones of the plurality of word lines connected to the memory cells being read in one read operation, and

adjust a voltage applied to the ones of the plurality of word lines during at least one of an end of the one read operation or a beginning of a subsequent read operation of the memory cells based on at least one of the subsequent read operation following the one

read operation within a predetermined time or a control gate ready voltage of the ones of the plurality of word lines targeted following the one read operation.

2. The memory apparatus as set forth in claim 1, wherein the memory cells are disposed in memory holes grouped into a plurality of strings, the plurality of strings comprise a block, and the control means, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to:

apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation;

skip discharging the ones of the plurality of word lines during the end of the one read operation and a ramping up of the ones of the plurality of word lines to a read pass voltage in response to the at least one of the subsequent read operation involving the memory cells of the one of the plurality of strings of the block and occurring within the predetermined time following the one read operation; and

apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the subsequent read operation.

3. The memory apparatus as set forth in claim 1, wherein the memory cells are disposed in memory holes grouped into a plurality of strings, the plurality of strings comprise a block, the end of the one read operation comprises a discharge stage, and the control means, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to:

apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation;

ramp the voltage applied to the ones of the plurality of word lines up to a discharge kick voltage and back down toward the control gate ready voltage of the ones of the plurality of word lines targeted at a predetermined discharge ramp rate during the discharge stage;

adjust the predetermined discharge ramp rate of the voltage applied to the ones of the plurality of word lines during the discharge stage in response to the at least one of the subsequent read operation involving the memory cells of another of the plurality of strings of the block and occurring within the predetermined time following the one read operation; and

apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of the another of the plurality of strings of the block in the subsequent read operation.

4. The memory apparatus as set forth in claim 1, wherein the memory cells are disposed in memory holes grouped into a plurality of strings, the plurality of strings comprise a block, the end of the one read operation comprises a discharge stage, and the control means is further configured to:

apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation; and

ramp the voltage applied to the ones of the plurality of word lines up to a discharge kick voltage and back down toward the control gate ready voltage of the ones of the plurality of word lines targeted in a stepwise fashion with a plurality of discharge steps lasting each of a plurality of discharge time periods during the discharge stage, at least one of the plurality of discharge steps or the plurality of discharge time periods dependent on the control gate ready voltage.

5. The memory apparatus as set forth in claim 4, wherein the control means is further configured to:

in response to the plurality of discharge steps comprising two steps and a first discharge step voltage at which a first discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage, ramp the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage of the ones of the plurality of word lines targeted using the first discharge step lasting a first discharge step time period and ending at the first discharge step voltage and a second discharge step of the plurality of discharge steps lasting a second discharge step time period and ending at a second discharge step voltage;

in response to the plurality of discharge steps comprising two steps and the first discharge step voltage at which the first discharge step of the plurality of discharge steps ends being less than or equal to the control gate ready voltage, ramp the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage of the ones of the plurality of word lines targeted using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage thereby eliminating the second discharge step time period;

in response to the plurality of discharge steps comprising three steps and the second discharge step voltage at which the second discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage, ramp the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage of the ones of the plurality of word lines targeted using the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage and the second discharge step of the plurality of discharge steps lasting the second discharge step time period and ending at the second discharge step voltage and a third discharge step of the plurality of discharge steps lasting a third discharge step time period and ending at a third discharge step voltage;

in response to the plurality of discharge steps comprising three steps and the first discharge step voltage at which the first discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage and the second discharge step voltage at which the second discharge step of the plurality of discharge steps ends being less than the control gate ready voltage, ramp the voltage applied to the ones of the

plurality of word lines back down toward the control gate ready voltage of the ones of the plurality of word lines targeted using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage and the second discharge step lasting the second discharge step time period and ending at the second discharge step voltage thereby eliminating the third discharge step time period; and

in response to the plurality of discharge steps comprising three steps and the first discharge step voltage at which the first discharge step of the plurality of discharge steps ends being less than or equal to the control gate ready voltage, ramp the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage of the ones of the plurality of word lines targeted using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage thereby eliminating the second discharge step time period and the third discharge step time period.

6. The memory apparatus as set forth in claim 5, wherein the control means, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to:

- ramp the voltage applied to the ones of the plurality of word lines up to an intermediate voltage during a first beginning period of a beginning of the one read operation and up to a read pass voltage during a second beginning period of the beginning of the one read operation and back down to the at least one read voltage associated with the one of the plurality of data states in the one read operation;
- ramp the voltage applied to the ones of the plurality of word lines up to the discharge kick voltage and back down toward the control gate ready voltage of the ones of the plurality of word lines using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage thereby eliminating any other discharge step time periods other than the first discharge step time period;
- maintain the voltage applied to the ones of the plurality of word lines at the first discharge step voltage during a wait time between the one read operation and the subsequent read operation; and
- skip ramping the voltage applied to the ones of the plurality of word lines up to the intermediate voltage during a first subsequent beginning period of the beginning of the subsequent read operation and ramp the voltage applied to the ones of the plurality of word lines directly from the first discharge step voltage up to the read pass voltage during a second subsequent beginning period of the beginning of the subsequent read operation thereby eliminating the first subsequent beginning period of the beginning of the subsequent read operation.

7. The memory apparatus as set forth in claim 6, wherein the control means, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to adjust the second subsequent beginning period of the beginning of the subsequent read operation to an adjusted second subsequent beginning period such that the adjusted second subsequent beginning period is shorter than the second subsequent beginning period.

8. A controller in communication with a memory apparatus including memory cells each connected to one of a plurality of word lines and configured to store a threshold voltage corresponding to one of a plurality of data states, the controller configured to:

- instruct the memory apparatus to apply at least one read voltage associated with one of the plurality of data states to ones of the plurality of word lines connected to the memory cells being read in one read operation; and

- instruct the memory apparatus to adjust a voltage applied to the ones of the plurality of word lines during at least one of an end of the one read operation or a beginning of a subsequent read operation of the memory cells based on at least one of the subsequent read operation following the one read operation within a predetermined time or a control gate ready voltage of the ones of the plurality of word lines targeted following the one read operation.

9. The controller as set forth in claim 8, wherein the memory cells are disposed in memory holes grouped into a plurality of strings, the plurality of strings comprise a block, and the controller, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to:

- instruct the memory apparatus to apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation;

- instruct the memory apparatus to skip discharging the ones of the plurality of word lines during the end of the one read operation and a ramping up of the ones of the plurality of word lines to a read pass voltage in response to the at least one of the subsequent read operation involving the memory cells of the one of the plurality of strings of the block and occurring within the predetermined time following the one read operation; and

- instruct the memory apparatus to apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the subsequent read operation.

10. The controller as set forth in claim 8, wherein the memory cells are disposed in memory holes grouped into a plurality of strings, the plurality of strings comprising a block, the end of the one read operation comprises a discharge stage, and the controller, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to:

- instruct the memory apparatus to apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation;

- instruct the memory apparatus to ramp the voltage applied to the ones of the plurality of word lines up to a discharge kick voltage and back down toward the control gate ready voltage of the ones of the plurality of word lines targeted at a predetermined discharge ramp rate during the discharge stage;

- instruct the memory apparatus to adjust the predetermined discharge ramp rate of the voltage applied to the ones

of the plurality of word lines during the discharge stage in response to the at least one of the subsequent read operation involving the memory cells of another of the plurality of strings of the block and occurring within the predetermined time following the one read operation; and

instruct the memory apparatus to apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of the another of the plurality of strings of the block in the subsequent read operation.

11. The controller as set forth in claim **8**, wherein the memory cells are disposed in memory holes grouped into a plurality of strings, the plurality of strings comprise a block, the end of the one read operation comprises a discharge stage, and the controller is further configured to:

instruct the memory apparatus to apply the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation; and

instruct the memory apparatus to ramp the voltage applied to the ones of the plurality of word lines up to a discharge kick voltage and back down toward the control gate ready voltage of the ones of the plurality of word lines targeted in a stepwise fashion with a plurality of discharge steps lasting each of a plurality of discharge time periods during the discharge stage, at least one of the plurality of discharge steps or the plurality of discharge time periods dependent on the control gate ready voltage.

12. The controller as set forth in claim **11**, wherein the controller, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to:

instruct the memory apparatus to ramp the voltage applied to the ones of the plurality of word lines up to an intermediate voltage during a first beginning period of a beginning of the one read operation and up to a read pass voltage during a second beginning period of the beginning of the one read operation and back down to the at least one read voltage associated with the one of the plurality of data states in the one read operation;

instruct the memory apparatus to ramp the voltage applied to the ones of the plurality of word lines up to the discharge kick voltage and back down toward the control gate ready voltage of the ones of the plurality of word lines using only a first discharge step lasting a first discharge step time period and ending at a first discharge step voltage thereby eliminating any other discharge step time periods other than the first discharge step time period;

instruct the memory apparatus to maintain the voltage applied to the ones of the plurality of word lines at the first discharge step voltage during a wait time between the one read operation and the subsequent read operation; and

instruct the memory apparatus to skip ramping the voltage applied to the ones of the plurality of word lines up to the intermediate voltage during a first subsequent beginning period of the beginning of the subsequent read operation and ramp the voltage applied to the ones of the plurality of word lines directly from the first discharge step voltage up to the read pass voltage

during a second subsequent beginning period of the beginning of the subsequent read operation thereby eliminating the first subsequent beginning period of the beginning of the subsequent read operation.

13. The controller as set forth in claim **12**, wherein the controller, in response to the subsequent read operation following the one read operation within the predetermined time, is further configured to instruct the memory apparatus to adjust the second subsequent beginning period of the beginning of the subsequent read operation to an adjusted second subsequent beginning period such that the adjusted second subsequent beginning period is shorter than the second subsequent beginning period.

14. A method of operating a memory apparatus including memory cells each connected to one of a plurality of word lines and configured to store a threshold voltage corresponding to one of a plurality of data states, the method comprising the steps of:

applying at least one read voltage associated with one of the plurality of data states to ones of the plurality of word lines connected to the memory cells being read in one read operation; and

adjusting a voltage applied to the ones of the plurality of word lines during at least one of an end of the one read operation or a beginning of a subsequent read operation of the memory cells based on at least one of the subsequent read operation following the one read operation within a predetermined time or a control gate ready voltage of the ones of the plurality of word lines targeted following the one read operation.

15. The method as set forth in claim **14**, wherein the memory cells are disposed in memory holes grouped into a plurality of strings, the plurality of strings comprise a block, and the method, in response to the subsequent read operation following the one read operation within the predetermined time, further includes the steps of:

applying the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation;

skipping discharging the ones of the plurality of word lines during the end of the one read operation and a ramping up of the ones of the plurality of word lines to a read pass voltage in response to the at least one of the subsequent read operation involving the memory cells of the one of the plurality of strings of the block and occurring within the predetermined time following the one read operation; and

applying the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the subsequent read operation.

16. The method as set forth in claim **14**, wherein the memory cells are disposed in memory holes grouped into a plurality of strings, the plurality of strings comprise a block, the end of the one read operation comprises a discharge stage, and the method, in response to the subsequent read operation following the one read operation within the predetermined time, further includes the steps of:

applying the at least one read voltage associated with the one of the plurality of data states to the ones of the word

lines connected to the memory cells of one of the plurality of strings of the block in the one read operation;

ramping the voltage applied to the ones of the plurality of word lines up to a discharge kick voltage and back down toward the control gate ready voltage of the ones of the plurality of word lines targeted at a predetermined discharge ramp rate during the discharge stage;

adjusting the predetermined discharge ramp rate of the voltage applied to the ones of the plurality of word lines during the discharge stage in response to the at least one of the subsequent read operation involving the memory cells of another of the plurality of strings of the block and occurring within the predetermined time following the one read operation; and

applying the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of the another of the plurality of strings of the block in the subsequent read operation.

17. The method as set forth in claim **14**, wherein the memory cells are disposed in memory holes grouped into a plurality of strings, the plurality of strings comprise a block, the end of the one read operation comprises a discharge stage, and the method further includes the steps of:

applying the at least one read voltage associated with the one of the plurality of data states to the ones of the word lines connected to the memory cells of one of the plurality of strings of the block in the one read operation; and

ramping the voltage applied to the ones of the plurality of word lines up to a discharge kick voltage and back down toward the control gate ready voltage of the ones of the plurality of word lines targeted in a stepwise fashion with a plurality of discharge steps lasting each of a plurality of discharge time periods during the discharge stage, at least one of the plurality of discharge steps or the plurality of discharge time periods dependent on the control gate ready voltage.

18. The method as set forth in claim **17**, further including the steps of:

in response to the plurality of discharge steps comprising two steps and a first discharge step voltage at which a first discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage, ramping the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage of the ones of the plurality of word lines targeted using the first discharge step lasting a first discharge step time period and ending at the first discharge step voltage and a second discharge step of the plurality of discharge steps lasting a second discharge step time period and ending at a second discharge step voltage;

in response to the plurality of discharge steps comprising two steps and the first discharge step voltage at which the first discharge step of the plurality of discharge steps ends being less than or equal to the control gate ready voltage, ramping the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage of the ones of the plurality of word lines targeted using only the first discharge step lasting the first discharge step time period and ending

at the first discharge step voltage thereby eliminating the second discharge step time period;

in response to the plurality of discharge steps comprising three steps and the second discharge step voltage at which the second discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage, ramping the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage of the ones of the plurality of word lines targeted using the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage and the second discharge step of the plurality of discharge steps lasting the second discharge step time period and ending at the second discharge step voltage and a third discharge step of the plurality of discharge steps lasting a third discharge step time period and ending at a third discharge step voltage;

in response to the plurality of discharge steps comprising three steps and the first discharge step voltage at which the first discharge step of the plurality of discharge steps ends being greater than the control gate ready voltage and the second discharge step voltage at which the second discharge step of the plurality of discharge steps ends being less than the control gate ready voltage, ramping the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage of the ones of the plurality of word lines targeted using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage and the second discharge step lasting the second discharge step time period and ending at the second discharge step voltage thereby eliminating the third discharge step time period; and

in response to the plurality of discharge steps comprising three steps and the first discharge step voltage at which the first discharge step of the plurality of discharge steps ends being less than or equal to the control gate ready voltage, ramping the voltage applied to the ones of the plurality of word lines back down toward the control gate ready voltage of the ones of the plurality of word lines targeted using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage thereby eliminating the second discharge step time period and the third discharge step time period.

19. The method as set forth in claim **18**, wherein the method, in response to the subsequent read operation following the one read operation within the predetermined time, further includes the steps of:

ramping the voltage applied to the ones of the plurality of word lines up to an intermediate voltage during a first beginning period of a beginning of the one read operation and up to a read pass voltage during a second beginning period of the beginning of the one read operation and back down to the at least one read voltage associated with the one of the plurality of data states in the one read operation;

ramping the voltage applied to the ones of the plurality of word lines up to the discharge kick voltage and back down toward the control gate ready voltage of the ones of the plurality of word lines using only the first discharge step lasting the first discharge step time period and ending at the first discharge step voltage

thereby eliminating any other discharge step time periods other than the first discharge step time period; maintaining the voltage applied to the ones of the plurality of word lines at the first discharge step voltage during a wait time between the one read operation and the subsequent read operation; and skipping ramping the voltage applied to the ones of the plurality of word lines up to the intermediate voltage during a first subsequent beginning period of the beginning of the subsequent read operation and ramp the voltage applied to the ones of the plurality of word lines directly from the first discharge step voltage up to the read pass voltage during a second subsequent beginning period of the beginning of the subsequent read operation thereby eliminating the first subsequent beginning period of the beginning of the subsequent read operation.

20. The method as set forth in claim **19**, wherein the method, in response to the subsequent read operation following the one read operation within the predetermined time, further includes the step of adjusting the second subsequent beginning period of the beginning of the subsequent read operation to an adjusted second subsequent beginning period such that the adjusted second subsequent beginning period is shorter than the second subsequent beginning period.

* * * * *