US Patent & Trademark Office
Patent Public Search | Text View

United States Patent Application Publication 20250266095
Kind Code A1
Publication Date August 21, 2025
Inventor(s) Deng; Xiangying et al.

DYNAMIC WORD LINE RAMP UP KICK FOR MEMORY DEVICES

Abstract

The memory device includes an array of memory cells that are arranged in a plurality of word lines. The word lines of the memory block are associated with respective kick voltages. The kick voltages associated with at least some of the word lines are different than the kick voltages associated with at least some other of the word lines. In operation, circuitry sets a target voltage for at least one word line of the plurality of word lines at a magnitude that is equal to an intended voltage for the at least one word line plus the respective kick voltage that is associated with the at least one word line. After a kick duration, the circuitry proceeds reduces the target voltage for the at least one word line to the intended voltage.

Inventors: **Deng; Xiangying (Milpitas, CA), Amin; Parth (Fremont, CA)**

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

Family ID: **1000007786243**

Appl. No.: **18/443933**

Filed: **February 16, 2024**

Publication Classification

Int. Cl.: **G11C16/10** (20060101); **G11C16/08** (20060101); **G11C16/26** (20060101)

U.S. Cl.:

CPC **G11C16/102** (20130101); **G11C16/08** (20130101); **G11C16/26** (20130101);

Background/Summary

BACKGROUND

1. Field

[0001] The subject disclosure is related generally to programming and sensing techniques that allow the voltages of the many word lines in a memory block to ramp up to their respective intended voltages at more consistent ramp rates.

2. Related Art

[0002] Semiconductor memory is widely used in various electronic devices, such as cellular telephones, digital cameras, personal digital assistants, medical electronics, mobile computing devices, servers, solid state drives, non-mobile computing devices and other devices. Semiconductor memory may comprise non-volatile memory or volatile memory. A non-volatile memory allows information to be stored and retained even when the non-volatile memory is not connected to a source of power, e.g., a battery.

[0003] NAND memory devices include a chip with a plurality of memory blocks, each of which includes an array of memory cells arranged in a plurality of word lines. Programming the memory cells of a word line to retain data typically occurs in a plurality of program loops, each of which includes the application of a programming pulse to a control gate of the word line and, optionally, a verify operation to sense the threshold voltages of the memory cells being programmed. Each program loop may also include a pre-charge operation prior to the programming pulse to pre-charge a plurality of channels containing memory cells to be programmed.

SUMMARY

[0004] One aspect of the present disclosure is related to a method of performing an operation in a memory device. The method includes the step of preparing a memory block that includes an array of memory cells that are arranged in a plurality of word lines. The word lines of the memory block are associated with respective kick voltages. The kick voltages associated with at least some of the word lines are different than the kick voltages associated with at least some other of the word lines. The method continues with the step of setting a target voltage for at least one word line of the plurality of word lines at a magnitude that is equal to an intended voltage for the at least one word line plus the respective kick voltage that is associated with the at least one word line. After a kick duration, the method proceeds with the step of reducing the target voltage for the at least one word line to the intended voltage.

[0005] According to another aspect of the present disclosure, the plurality of word lines include word lines that are slow to ramp up to a given target voltage and word lines that are fast to ramp up to the given target voltage. The word lines that are slow to ramp up to the given voltage are associated with higher respective kick voltages than the word lines that are fast to ramp up to the given target voltage.

[0006] According to yet another aspect of the present disclosure, the operation is a programming operation and the intended voltage is a programming voltage or a pass voltage.

[0007] According to still another aspect of the present disclosure, the operation is a sensing operation and the intended voltage is a reference voltage.

[0008] According to a further aspect of the present disclosure, a plurality of word lines are arranged in a plurality of groups including a first group and a second group. The word lines in the first group are associated with a first kick voltage. The word lines in the second group are associated with a second kick voltage that is different than the first kick voltage.

[0009] According to yet a further aspect of the present disclosure, the plurality of groups further includes a third group. The word lines in the third group are associated with a third kick voltage that is different than the first and second kick voltages.

[0010] According to still a further aspect of the present disclosure, the kick voltage that is associated with each word line is based on a word line resistance capacitance of the word line.

[0011] According to another aspect of the invention, the method further includes the step of

measuring the word line resistance capacitance of each word line to determine the kick voltage that is associated with each word line.
[0012] According to yet another aspect of the present disclosure, for each word line, the associated kick voltage is used during both programming and sensing operations.
[0013] Another aspect of the present disclosure is related to a method of operating a memory device. The method includes the step of preparing a memory block that includes an array of memory cells that are arranged in a plurality of word lines. The method proceeds with the step of measuring a word line resistance capacitance of the plurality of word lines in the memory. The method continues with the step of establishing a plurality of groups of word lines. Each of the groups of word lines includes a plurality of word lines with similar measured word line resistance capacitances. The method proceeds with the step of assigning the plurality of groups different kick voltages based on the word line resistance capacitances of the word lines in each of the groups.
[0014] According to another aspect of the present disclosure, the step of assigning the plurality of groups different kick voltages based on the word line resistance capacitances of the word lines in each of the groups includes assigning a first group of word lines a first kick voltage and assigning a second group of word lines a second kick voltage that is different than the first kick voltage.
[0015] According to yet another aspect of the present disclosure, the method further includes the step of performing an operation wherein during a voltage ramp up process for at least one word line of the plurality of word lines, a target voltage is set at a level that is equal to an intended voltage plus the kick voltage that is assigned to the at least one word line.
[0016] According to still another aspect of the present disclosure, the operation is a programming operation.
[0017] According to a further aspect of the present disclosure, the operation is a sensing operation.
[0018] According to yet a further aspect of the present disclosure, the kick voltages that are assigned to the groups of word lines are used during both programming and sensing operations.
[0019] According to still a further aspect of the present disclosure, the plurality of word lines includes word lines that are slow to ramp up to a given target voltage and word lines that are fast to ramp up to the given target voltage. The word lines that are slow to ramp up to the given target voltage are assigned kick voltages that are greater than the word lines that are fast to ramp up to the given target voltage.
[0020] Yet another aspect of the present disclosure is related to a memory device that includes a memory block with an array of memory cells that are arranged in a plurality of word lines. The word lines of the memory block are associated with respective kick voltages. The kick voltages associated with at least some of the word lines are different than the kick voltages of at least some other of the word lines. The memory device also includes a programming and sensing means for programming and sensing the memory cells of the plurality of word lines. The programming and sensing means is configured to set a target voltage for at least one word line of the plurality of word lines at a magnitude that is equal to an intended voltage for the at least one word line plus the respective kick voltage that is associated with the at least one word line. After the kick duration, the programming and sensing means reduces the target voltage for the at least one word line to the intended voltage.
[0021] According to another aspect of the present disclosure, the plurality of word lines include word lines that are slow to ramp up to a given target voltage and word lines that are fast to ramp up to the given target voltage. The word lines that are slow to ramp up to the given voltage are associated with higher respective kick voltages than the word lines that are fast to ramp up to the given target voltage.
[0022] According to yet another aspect of the present disclosure, the programming and sensing means is configured to perform programming operation where the intended voltage is a programming voltage or a pass voltage.
[0023] According to still another aspect of the present disclosure, the programming and sensing

means is configured to perform a sensing operation where the intended voltage is a reference voltage.

Description

BRIEF DESCRIPTION OF THE DRAWINGS

[0024] A more detailed description is set forth below with reference to example embodiments depicted in the appended figures. Understanding that these figures depict only example embodiments of the disclosure and are, therefore, not to be considered limiting of its scope. The disclosure is described and explained with added specificity and detail through the use of the accompanying drawings in which:

[0025] FIG. **1**A is a block diagram of an example memory device;

[0026] FIG. **1**B is a block diagram of an example control circuit;

[0027] FIG. **1**C is a block diagram of example circuitry of the memory device of FIG. **1**A;

[0028] FIG. **2** depicts blocks of memory cells in an example two-dimensional configuration of the memory array of FIG. **1**A;

[0029] FIG. **3**A and FIG. **3**B depict cross-sectional views of example floating gate memory cells in NAND strings;

[0030] FIG. **4**A and FIG. **4**B depict cross-sectional views of example charge-trapping memory cells in NAND strings;

[0031] FIG. **5** depicts an example block diagram of the sense block SB**1** of FIG. **1**;

[0032] FIG. **6**A is a perspective view of a set of blocks in an example three-dimensional configuration of the memory array of FIG. **1**;

[0033] FIG. **6**B depicts an example cross-sectional view of a portion of one of the blocks of FIG. **6**A;

[0034] FIG. **6**C depicts a plot of memory hole diameter in the stack of FIG. **6**B;

[0035] FIG. **6**D depicts a close-up view of region **622** of the stack of FIG. **6**B;

[0036] FIG. **7**A depicts a top view of an example word line layer WL**0** of the stack of FIG. **6**B;

[0037] FIG. **7**B depicts a top view of an example top dielectric layer DL**116** of the stack of FIG. **6**B;

[0038] FIG. **8** depicts a threshold voltage distribution of a page of memory cells programmed to one bit per memory cell (SLC);

[0039] FIG. **9** depicts a threshold voltage distribution of a page of memory cells programmed to three bits per memory cell (TLC);

[0040] FIG. **10** is a waveform of the voltages applied to a selected word line during an example programming operation;

[0041] FIG. **11** is a schematic view illustrating the voltages applied to the word lines of an example memory block during an example sensing operation;

[0042] FIG. **12** is a plot of voltage versus time during a voltage ramp up process in three word lines including a slow word line, a medium word line, and a fast word line where no kick voltage is used during the ramp up process;

[0043] FIG. **13** is a plot of voltage versus time for an example word line during a voltage ramp up process where a kick voltage is used;

[0044] FIG. **14** is a table that includes word lines, the respective zones of the word lines, the kick ratios associated with the respective zones; and example word line ramp up plots for those word lines when the kick voltages are employed during ramp up processes;

[0045] FIG. **15** is a flow chart depicting the steps of dynamically establishing the kick voltages to be used during programming and sensing operations in a plurality of word lines in a memory block; and

[0046] FIG. **16** is a flow chart depicting the steps of using a dynamically adjusted kick voltage to improve programming uniformity across many word lines in a memory block.

DESCRIPTION OF THE ENABLING EMBODIMENTS

[0047] The present disclosure is related to programming and sensing techniques that result in a plurality of word lines in a memory block ramping up to respective intended voltages (such as programming voltages, pass voltages, or reference voltages) more uniformly, i.e., at similar or approximately the same ramp rates. As discussed in further detail below, according to these techniques, at the beginning of a word line ramp up process, a targeted voltage for a word line is increased from an intended voltage by a dynamically determined kick voltage. The many word lines are all associated with their own respective kick voltages, which are dynamically determined and some of which are different than others. More specifically, word lines that are slow to ramp up to a given target voltage are associated with higher kick voltages so that they can ramp up to their intended voltages much more quickly than if no kick voltage was applied. Conversely, word lines that are quick to ramp up to the given target voltage are associated with lesser kick voltages so that they can ramp up to their intended voltages only slightly more quickly than if no kick voltage was applied. By giving the slow word lines high kick voltages and the fast word lines slow kick voltages, the ramp rates across the many word lines are more uniform, thereby allowing for the timing parameters following the ramp up process to be more precisely set so that programming performance, sensing performance, and reliability can all be improved.

[0048] FIG. **1**A is a block diagram of an example memory device **100** is configured to program and sense the memory cells in the word lines of a memory block according to the programming and sensing techniques of the subject disclosure. 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 SB**1**, SB**2**, ... 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**.

[0049] The memory structure **126** can be two-dimensional or three-dimensional. The memory structure **126** may comprise one or more array of memory cells including a three-dimensional array. The memory structure **126** 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 **126** 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 **126** 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.

[0050] 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.

[0051] A storage region **113** may, for example, be provided for programming parameters. The programming parameters may include a program voltage, a program voltage bias, position parameters indicating positions of memory cells, contact line connector thickness parameters, a verify voltage, and/or the like. The position parameters may indicate a position of a memory cell within the entire array of NAND strings, a position of a memory cell as being within a particular NAND string group, a position of a memory cell on a particular plane, and/or the like. The contact line connector thickness parameters may indicate a thickness of a contact line connector, a substrate

or material that the contact line connector is comprised of, and/or the like.
[0052] 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.
[0053] In some embodiments, 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, SB**2**, . . . , SBp, read/write circuits **128**, controller **122**, and so forth.
[0054] The control circuits **150** can include a programming circuit **151** configured to perform a program and verify operation for one set of memory cells, wherein the one set of memory cells comprises memory cells assigned to represent one data state among a plurality of data states and memory cells assigned to represent another data state among the plurality of data states; the program and verify operation comprising a plurality of program and verify iterations; and in each program and verify iteration, the programming circuit performs programming for the one selected word line after which the programming circuit applies a verification signal to the selected word line. The control circuits **150** can also include a counting circuit **152** configured to obtain a count of memory cells which pass a verify test for the one data state. The control circuits **150** can also include a determination circuit **153** configured to determine, based on an amount by which the count exceeds a threshold, if a programming operation is completed.
[0055] For example, FIG. **1**B is a block diagram of an example control circuit **150** which comprises the programming circuit **151**, the counting circuit **152**, and the determination circuit **153**.
[0056] The off-chip controller **122** may comprise a processor **122***c*, storage devices (memory) such as ROM **122***a* and RAM **122***b* 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 Vt distribution becomes too high. However, uncorrectable errors may exist in some cases. The techniques provided herein reduce the likelihood of uncorrectable errors.
[0057] The storage device(s) **122***a*, **122***b* comprise, code such as a set of instructions, and the processor **122***c* is operable to execute the set of instructions to provide the functionality described herein. Alternately or additionally, the processor **122***c* can access code from a storage device **126***a* of the memory structure **126**, such as a reserved area of memory cells in one or more word lines. For example, code can be used by the controller **122** to access the memory structure **126** 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 **122** during a booting or startup process and enables the controller **122** to access the memory structure **126**. The code can be used by the controller **122** to control one or more memory structures **126**. Upon being powered up, the processor **122***c* fetches the boot code from the ROM **122***a* or storage device **126***a* for execution, and the boot code initializes the system components and loads the control code into the RAM **122***b*. Once the control code is loaded into the RAM **122***b*, it is executed by the processor **122***c*. 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.
[0058] Generally, the control code can include instructions to perform the functions described herein including the steps of the flowcharts discussed further below and provide the voltage waveforms including those discussed further below. For example, as illustrated in FIG. **1**C, the control circuitry **110**, controller **122**, control circuits **150**, and/or any other circuitry are

configured/programmed, during a programming or erasing operation, at step **160**, during a word line voltage ramp up operation, a target voltage for a word line is set at an intended voltage V_Intended plus a kick voltage V_Kick, i.e., V_Target=V_Intended+V_Kick. At step **161**, after a kick duration t_Kick, the target voltage for the word line is reduced to the intended voltage V_Intended. These steps are discussed in further detail below.

[0059] 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.

[0060] Other types of non-volatile memory in addition to NAND flash memory can also be used.

[0061] 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.

[0062] 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.

[0063] 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.

[0064] A NAND memory array may be configured so that the array is composed of multiple memory strings 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. 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.

[0065] 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 is 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.

[0066] 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.

[0067] 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).

[0068] 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.

[0069] By way of non-limiting example, in a three-dimensional array of NAND strings, 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.

[0070] 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.

[0071] 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 (monolithic or non-monolithic) may be formed on separate chips and then packaged together to form a stacked-chip memory device.

[0072] FIG. **2** illustrates memory blocks **200**, **210** of memory cells in an example two-dimensional configuration of the memory array **126** of FIG. **1**. The memory array **126** can include many such blocks **200**, **210**. Each example block **200**, **210** includes a number of NAND strings and respective bit lines, e.g., BL**0**, BL**1**, . . . which are shared among the blocks. Each NAND string is connected at one end to a drain-side select gate (SGD), and the control gates of the drain-side select gates are connected via a common SGD line. The NAND strings are connected at their other end to a source-side select gate (SGS) which, in turn, is connected to a common source line **220**. One hundred and twelve word lines, for example, WL**0**-WL**111**, extend between the SGSs and the SGDs. In some embodiments, the memory block may include more or fewer than one hundred and twelve word

lines. For example, in some embodiments, a memory block includes one hundred and sixty-four word lines. 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 or between certain data word lines. Such dummy word lines can shield the edge data word line from certain edge effects.

[0073] One type of non-volatile memory which may be provided in the memory array is a floating gate memory, such as of the type shown in FIGS. **3**A and **3**B. However, other types of non-volatile memory can also be used. As discussed in further detail below, in another example shown in FIGS. **4**A and **4**B, 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. 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.

[0074] 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.

[0075] FIG. **3**A illustrates a cross-sectional view of example floating gate memory cells **300**, **310**, **320** in NAND strings. In this Figure, 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 **300**, **310**, **320** is in a different respective NAND string. An inter-poly dielectric (IPD) layer **328** is also illustrated. The control gates **302**, **312**, **322** are portions of the word line. A cross-sectional view along contact line connector **329** is provided in FIG. **3**B.

[0076] The control gate **302**, **312**, **322** wraps around the floating gate **304**, **314**, **321**, increasing the surface contact area between the control gate **302**, **312**, **322** and floating gate **304**, **314**, **321**. 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 **300**, **310**, **320** becomes smaller so there is almost no space for the control gate **302**, **312**, **322** and the IPD layer **328** between two adjacent floating gates **302**, **312**, **322**.

[0077] As an alternative, as shown in FIGS. **4**A and **4**B, the flat or planar memory cell **400**, **410**, **420** has been developed in which the control gate **402**, **412**, **422** is flat or planar; that is, it does not wrap around the floating gate and its only contact with the charge storage layer **428** 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.

[0078] FIG. **4**A depicts a cross-sectional view of example charge-trapping memory cells **400**, **410**, **420** in NAND strings. The view is in a word line direction of memory cells **400**, **410**, **420** comprising a flat control gate and charge-trapping regions as a two-dimensional example of

memory cells **400**, **410**, **420** in the memory cell array **126** of FIG. **1**. 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 **424** extends across NAND strings which include respective channel regions **406**, **416**, **426**. Portions of the word line provide control gates **402**, **412**, **422**. Below the word line is an IPD layer **428**, charge-trapping layers **404**, **414**, **421**, polysilicon layers **405**, **415**, **425**, and tunneling layers **409**, **407**, **408**. Each charge-trapping layer **404**, **414**, **421** extends continuously in a respective NAND string. The flat configuration of the control gate can be made thinner than a floating gate. Additionally, the memory cells can be placed closer together.

[0079] FIG. **4**B illustrates a cross-sectional view of the structure of FIG. **4**A along contact line connector **429**. The NAND string **430** includes an SGS transistor **431**, example memory cells **400**, **433**, . . . **435**, and an SGD transistor **436**. Passageways in the IPD layer **428** in the SGS and SGD transistors **431**, **436** allow the control gate layers **402** and floating gate layers to communicate. The control gate **402** and floating gate layers may be polysilicon and the tunnel oxide layer may be silicon oxide, for instance. The IPD layer **428** can be a stack of nitrides (N) and oxides (O) such as in a N—O—N—O—N configuration.

[0080] 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 sd**1**, sd**2**, sd**3**, sd**4**, sd**5**, sd**6** and sd**7** are formed in the p-type well. A channel voltage, Vch, may be applied directly to the channel region of the substrate.

[0081] FIG. **5** illustrates an example block diagram of the sense block SB**1** of FIG. **1**. In one approach, a sense block comprises multiple sense circuits. Each sense circuit is associated with data latches. For example, the example sense circuits **550***a*, **551***a*, **552***a*, and **553***a* are associated with the data latches **550***b*, **551***b*, **552***b*, and **553***b*, 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 SB**1** can communicate with the set of sense circuits and latches. The sense circuit controller **560** 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 bus and a local bus. In another possible approach, a common voltage is provided to each sense circuit concurrently. The sense circuit controller **560** may also include a pre-charge circuit **561**, a memory **562** and a processor **563**. The memory **562** may store code which is executable by the processor to perform the functions described herein. These functions can include reading the latches **550***b*, **551***b*, **552***b*, **553***b* which are associated with the sense circuits **550***a*, **551***a*, **552***a*, **553***a*, setting bit values in the latches and providing voltages for setting pre-charge levels in sense nodes of the sense circuits **550***a*, **551***a*, **552***a*, **553***a*. Further example details of the sense circuit controller **560** and the sense circuits **550***a*, **551***a*, **552***a*, **553***a* are provided below.

[0082] In some embodiments, a memory cell may include a flag register that includes a set of latches storing flag bits. In some embodiments, a quantity of flag registers may correspond to a quantity of data states. In some embodiments, one or more flag registers may be used to control a type of verification technique used when verifying memory cells. In some embodiments, a flag bit's output may modify associated logic of the device, e.g., address decoding circuitry, such that a specified block of cells is selected. A bulk operation (e.g., an erase operation, etc.) may be carried out using the flags set in the flag register, or a combination of the flag register with the address register, as in implied addressing, or alternatively by straight addressing with the address register alone.

[0083] FIG. **6**A is a perspective view of a set of blocks **600** in an example three-dimensional configuration of the memory array **126** of FIG. **1**. On the substrate are example blocks BLK**0**,

BLK**1**, BLK**2**, BLK**3** of memory cells (storage elements) and a peripheral area **604** with circuitry for use by the blocks BLK**0**, BLK**1**, BLK**2**, BLK**3**. For example, the circuitry can include voltage drivers **605** which can be connected to control gate layers of the blocks BLK**0**, BLK**1**, BLK**2**, BLK**3**. In one approach, control gate layers at a common height in the blocks BLK**0**, BLK**1**, BLK**2**, BLK**3** are commonly driven. The substrate **601** can also carry circuitry under the blocks BLK**0**, BLK**1**, BLK**2**, BLK**3**, along with one or more lower metal layers which are patterned in conductive paths to carry signals of the circuitry. The blocks BLK**0**, BLK**1**, BLK**2**, BLK**3** 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 BLK**0**, BLK**1**, BLK**2**, BLK**3** comprises a stacked area of memory cells, where alternating levels of the stack represent word lines. In one possible approach, each block BLK**0**, BLK**1**, BLK**2**, BLK**3** has opposing tiered sides from which vertical contacts extend upward to an upper metal layer to form connections to conductive paths. While four blocks BLK**0**, BLK**1**, BLK**2**, BLK**3** are illustrated as an example, two or more blocks can be used, extending in the x- and/or y-directions.

[0084] 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.

[0085] FIG. **6**B illustrates an example cross-sectional view of a portion of one of the blocks BLK**0**, BLK**1**, BLK**2**, BLK**3** of FIG. **6**A. 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 DWLD**0**, DWLD**1**, DWLS**0** and DWLS**1**, in addition to data word line layers (word lines) WL**0**-WL**111**. The dielectric layers are labelled as DL**0**-DL**116**. Further, regions of the stack **610** which comprise NAND strings NS**1** and NS**2** are illustrated. Each NAND string encompasses a memory hole **618**, **619** which is filled with materials which form memory cells adjacent to the word lines. A region **622** of the stack **610** is shown in greater detail in FIG. **6**D and is discussed in further detail below. The dielectric layers can have variable thicknesses such that some of the conductive layers can be closer to or further from neighboring conductive layers. The thicknesses of the dielectric layers affects the “ON pitch,” which is a factor in memory density. Specifically, a smaller ON pitch allows for more memory cells in a given area but may compromise reliability.

[0086] The stack **610** includes a substrate **611**, an insulating film **612** on the substrate **611**, and a portion of a source line SL. NS**1** has a source-end **613** at a bottom **614** of the stack and a drain-end **615** at a top **616** of the stack **610**. Contact line connectors (e.g., slits, such as metal-filled slits) **617**, **620** may be provided periodically across the stack **610** as interconnects which extend through the stack **610**, such as to connect the source line to a particular contact line above the stack **610**. The contact line connectors **617**, **620** may be used during the formation of the word lines and subsequently filled with metal. A portion of a bit line BL**0** is also illustrated. A conductive via **621** connects the drain-end **615** to BL**0**.

[0087] FIG. **6**C illustrates a plot of memory hole diameter in the stack of FIG. **6**B. The vertical axis is aligned with the stack of FIG. **6**B and illustrates a width (wMH), e.g., diameter, of the memory holes **618** and **619**. The word line layers WL**0**-WL**111** of FIG. **6**A are repeated as an example and are at respective heights z**0**-z**111** 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 slightly wider before

becoming progressively smaller from the top to the bottom of the memory hole.
[0088] FIG. **6**D illustrates a close-up view of the region **622** of the stack **610** of FIG. **6**B. 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**, **681** are provided above dummy memory cells **682**, **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 **630**) 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 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 **630**. 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.
[0089] 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 threshold voltage Vt of a memory cell is increased in proportion to the amount of stored charge. During a sensing operation, the threshold voltage Vt is detected or measured. During an erase operation, the electrons return to the channel.
[0090] Each of the memory holes **630** can be filled with a plurality of annular layers comprising a blocking oxide layer, a charge trapping layer **663**, a tunneling layer **664** and a channel layer. A core region of each of the memory holes **630** is filled with a body material, and the plurality of layers are between the core region and the word line layer in each of the memory holes **630**. In some cases, the charge trapping layer **663** and the tunneling layer **664** are annular in shape. In other cases, as discussed in further detail below, these layers are semi-circular in shape.
[0091] 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.
[0092] FIG. **7**A illustrates a top view of an example word line layer WL**0** of the stack **610** of FIG. **6**B. As mentioned, a three-dimensional 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.
[0093] A block BLK in a three-dimensional memory device can be divided into sub-blocks, where each sub-block comprises a NAND string group which has a common SGD control line. For example, see the SGD lines/control gates SGD**0**, SGD**1**, SGD**2** and SGD**3** in the sub-blocks SBa, SBb, SBc and SBd, respectively. 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 contact line connectors (e.g., 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 contact line connectors 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 contact line connectors may allow for a few rows of memory holes between adjacent contact line connectors. The layout of the memory holes and contact line connectors 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 contact line connectors can optionally be filed with metal to provide an interconnect through the stack.

[0094] In this example, there are four rows of memory holes between adjacent contact line connectors. 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 WL**0***a*, WL**0***b*, WL**0***c* and WL**0***d* which are each connected by a contact line **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 contact line **713**, in turn, is connected to a voltage driver for the word line layer. The region WL**0***a* has example memory holes **710**, **711** along a contact line **712**. The region WL**0***b* has example memory holes **714**, **715**. The region WL**0***c* has example memory holes **716**, **717**. The region WL**0***d* has example memory holes **718**, **719**. The memory holes are also shown in FIG. **7**B. 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 NS**0**_SBa, NS**1**_SBb, NS**2**_SBc, NS**3**_SBd, and NS**4**_SBe, respectively.

[0095] 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**, **721** are in WL**0***a*, memory cells **724**, **725** are in WL**0***b*, memory cells **726**, **727** are in WL**0***c*, and memory cells **728**, **729** are in WL**0***d*. These memory cells are at a common height in the stack.

[0096] Contact line connectors (e.g., slits, such as metal-filled slits) **701**, **702**, **703**, **704** may be located between and adjacent to the edges of the regions WL**0***a*-WL**0***d*. The contact line connectors **701**, **702**, **703**, **704** 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.

[0097] FIG. **7**B illustrates a top view of an example top dielectric layer DL**116** of the stack of FIG. **6**B. The dielectric layer is divided into regions DL**116***a*, DL**116***b*, DL**116***c* and DL**116***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 being 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 during each programming, sensing, or erasing operation.

[0098] The region DL**116***a* has the example memory holes **710**, **711** along a contact line **712**, which is coincident with a bit line BL**0**. A number of bit lines extend above the memory holes and are connected to the memory holes as indicated by the "X" symbols. BL**0** is connected to a set of memory holes which includes the memory holes **711**, **715**, **717**, **719**. Another example bit line BL**1** is connected to a set of memory holes which includes the memory holes **710**, **714**, **716**, **718**. The contact line connectors (e.g., slits, such as metal-filled slits) **701**, **702**, **703**, **704** from FIG. **7**A are also illustrated, as they extend vertically through the stack. The bit lines can be numbered in a sequence BL**0**-BL**23** across the DL**116** layer in the x-direction.

[0099] Different subsets of bit lines are connected to memory cells in different rows. For example, BL**0**, BL**4**, BL**8**, BL**12**, BL**16**, BL**20** are connected to memory cells in a first row of cells at the right-hand edge of each region. BL**2**, BL**6**, BL**10**, BL**14**, BL**18**, BL**22** are connected to memory cells in an adjacent row of cells, adjacent to the first row at the right-hand edge. BL**3**, BL**7**, BL**11**, BL**15**, BL**19**, BL**23** are connected to memory cells in a first row of cells at the left-hand edge of each region. BL**1**, BL**5**, BL**9**, BL**13**, BL**17**, BL**21** are connected to memory cells in an adjacent row of memory cells, adjacent to the first row at the left-hand edge.

[0100] The memory cells of the memory blocks can be programmed to store one or more bits of data in multiple data states, each of which is associated with a respective threshold voltage Vt range and with a respective bit or series of bits. For example, FIG. **8** depicts a threshold voltage Vt

distribution of a group of memory cells programmed according to a one bit per memory cell (SLC) storage scheme. In the SLC storage scheme, there are two total data states, including the erased state (Er) and a single programmed data state (S**1**). FIG. **9** illustrates the threshold voltage Vt distribution of a three bits per cell (TLC) storage scheme that includes eight total data states, namely the erased state (Er) and seven programmed data states (S**1**, S**2**, S**3**, S**4**, S**5**, S**6**, and S**7**). Each programmed data state (S**1**-S**7**) is associated with a respective verify voltage (Vv**1**-Vv**7**), which is employed during a verify portion of a programming operation. Similarly, each programmed data state is associated with a unique read voltage that can be the same or different than the respective verify voltages. Other storage schemes are also available, such as two bits per cell (MLC) with four data states, four bits per cell (QLC) with sixteen data states, or five bits per cell (PLC) with thirty-two data states.

[0101] Programming the memory cells occurs on a word line-by-word line basis from one side of the memory block towards an opposite side of the memory block. In contrast, erase typically occurs on a block or sub-block basis. Typically, programming the memory cells of a selected word line to retain multiple bits per memory cell (for example, MLC, TLC, or QLC) starts with the memory cells being in the erased data state and includes a plurality of program loops. Each program loop includes both a programming pulse and a verify operation. FIG. **10** depicts a waveform **1000** of the voltages applied to a selected word line WLn during an example programming operation for programming the memory cells of the selected word line WLn to a greater number of bits per memory cell (e.g., TLC or QLC). As depicted, each program loop includes a programming pulse (hereinafter referred to as a VPGM pulse **1001**-**1018**) and one or more verify pulses **1020**-**1036**, depending on which data states are being programmed in a particular program loop. During each VPGM pulse, the unselected word lines in the memory block are also ramped to a pass voltage VPASS to prevent unintentional programming of the memory cells in the unselected word lines.

[0102] With reference now to FIG. **11**, during a sensing operation (for example, verify or read), a sense node on the drain side of the memory block is charged to a predetermined charged voltage. A reference voltage VCG (e.g., Vv**1**-Vv**7** in FIG. **9**) is applied to a control gate of the selected word line WLn. Simultaneously, all of the memory cells of the NAND string except the memory cell of the selected word line are “turned on” (made conductive) by the application of pass voltages VREAD, VREADK to the unselected word lines. A sense node is then discharged through the NAND string that contains the selected memory cell being sensed and, since all of the memory cells except the one of the selected word line WLn are turned on, a discharge current through the NAND string is largely dictated by the threshold voltage Vt of the memory cell being sensed. More specifically, the discharge current is dictated by whether the threshold voltage Vt is greater than or less than the reference voltage VCG. If the discharge current is low, then the reference voltage VCG failed to “turn on” (make conductive to electrons) the selected memory cell and the threshold voltage Vt of the selected memory cell is determined to be greater than the reference voltage VCG. On the other hand, if the discharge current is high, then the reference voltage VCG did “turn on” the selected memory cell, and the threshold voltage Vt is determined to be less than the reference voltage VCG. By repeating this process using multiple reference voltages VCG, the threshold voltage Vt of the selected memory cell can be determined.

[0103] During each of the above discussed operations (programming, verify, and read) the word lines in a selected memory block ramp up to various voltages (e.g., VPGM, VPASS, VCG, VREAD, VREADK) numerous times. However, it has been found that within a memory block, not all word lines in a memory block ramp to their respective target voltages at the same rate as one another, i.e., some word lines ramp faster or slower than other word lines. This inconsistency is due to a number of factors, including a distance between a respective word line and its associated sense amplifier and also the word line resistance capacitance (WLRC) between a word line (e.g., WLn) and its neighboring word lines (e.g., WLn−1, WLn+1).

[0104] FIG. **12** is a plot that illustrates the voltages received at three different word lines (a fast

word line **1200**, a medium word line **1202**, and a slow word line **1204**), which all begin at the same baseline voltage and are all set to have the same target voltage V_Target, which is equal to an intended voltage throughout this example, at the same time. As illustrated, the fast word line **1200** approaches and reaches the target voltage V_Target first; the medium word line **1202** approaches and reaches the target voltage V_Target second; and the slow word line **1204** approaches and reaches the target voltage V_Target last. The intended voltage could be, for example, VPGM, VPASS, VCG, VREAD, VREADK, or any suitable voltage depending on a particular application, e.g., a programming operation or a sensing operation. The notable difference in ramp rates between these three word lines can cause inconsistent programming and sensing across the many word lines of a memory block.

[0105] One approach to increase the ramp rate of a word line being ramped to a given intended voltage V_Intended is to, at the beginning of the ramping process, increase the target voltage from V_Intended by a bias voltage V_Kick for a kick duration t_Kick. In other words, for the duration t_Kick, the target voltage V_Target for the word line is set at V_Intended plus V_Kick, i.e., V_Target=V_Intended+V_Kick. After the kick duration t_Kick, the target voltage V_Target settles to V_Intended. As illustrated with line **1300**, this approach causes the word line to reach the intended voltage V_Intended more quickly as compared to if there is no kick voltage adjustment. While the employment of the kick voltage improves ramp rates, if the kick duration t_Kick and kick voltage V_Kick are fixed (equal) for all word lines, there still will be inconsistent ramp rates between slow, medium, and fast word lines.

[0106] According to an aspect of the present disclosure, a technique is provided to dynamically optimize the word line ramp up kick so that all word lines in a memory block ramp up at similar rates. More specifically according to the exemplary embodiments, the magnitude of the kick voltage V_Kick is dynamically set at an increased level (e.g., V_Kick_High) during voltage ramp up of slow word lines and is dynamically set at a reduced level (e.g., V_Kick_Low) during voltage ramp up of fast word lines. Thus, the voltage ramp rate to the intended voltage V_Intended for the slow word lines is significantly increased as compared to a baseline scenario with no kick voltage V_Kick and the voltage ramp rate is slightly (i.e., less than significantly) increased as compared to the baseline scenario. By evening out the word line ramp up rates, timing parameters after the word line ramp up process can be optimized and shortened, thereby improving programming performance and read performance (i.e., programming time tProg and read time tRead are reduced) and also improving reliability. This is because these timing parameters are typically set to work with word lines that have a range of different ramp up rates (both slow and fast word lines). In some alternate embodiments, the kick duration t_Kick can alternately or additionally be dynamically adjusted to improve the ramp rate uniformity across the many word lines in a memory block.

[0107] According to some embodiments, the word lines in a memory block can be grouped into any suitable number of groups that have similar baseline (non-adjusted) ramp rates and within each group, all of the word lines can be associated with the same kick voltage V_Kick. Then, in these embodiments, the different groups are associated with different kick voltages V_Kick.

[0108] For example, the Table depicted in FIG. **14** shows an example memory block with one hundred and eleven data containing word lines (WL**0**-WL**111**) that have been divided into three groups. In this example memory block, the source-side word lines are the slowest word lines and are in Group A, the middle word lines are the medium word lines and are in Group B, and the drain-side word lines are the fastest word lines and are in Group C. Each of these groups is associated with its own respective kick voltage V_Kick_A, V_Kick_B, V_Kick_C that is applied during voltage ramp to any suitable target voltage, e.g., VPGM, VPASS, VCG, VREAD, or VREADK. In this example embodiment, because Group C contains the fastest word lines, V_Kick_C is the smallest of the three kick voltages. In this example embodiment, V_Kick_B is fifty percent higher than V_Kick_C (V_Kick_B=1.5*V_Kick_C), and V_Kick_A is twice as high

as V_Kick_C (V_Kick_A=2*V_Kick_C).
[0109] As also illustrated in the table of FIG. **14**, during voltage ramp up of any of the Group A word lines, the voltage supplied to the Group A word line (line **1400***a*) is increased above the intended voltage V_Intended by V_Kick_A (V_Target=V_Intended+V_Kick_A) for the kick duration. This causes the voltage of the word line (line **1402***a*) to ramp at a given rate to the intended voltage V_Intended. During voltage ramp up of any of the Group B word lines, the voltage supplied to the Group B word line (line **1400***b*) is increased above the intended voltage V_Intended by V_Kick_B (V_Target=V_Intended+V_Kick_B) for the kick duration. This causes the voltage of the Group B word line (line **1402***b*) to ramp at approximately the same rate as the Group A word line. During voltage ramp up of any of the Group C word lines, the voltage supplied to the Group C word line (line **1400***c*) is increased above the intended voltage V_Intended by V_Kick_C (V_Target=V_Intended+V_Kick_C) for the kick duration. This causes the voltage of the Group C word line (line **1402***c*) to ramp at approximately the same rate as the Group A and Group B word lines.
[0110] In some other embodiments, the word lines of the memory block may be divided into any suitable number of groups (i.e., two groups or four or more groups) that are each associated with their own respective kick voltages V_Kick or each word line may be associated with its own respective kick voltage V_Kick. In some embodiments, the kick voltages can be further dynamically adjusted, e.g., based on which specific data state is being sensed during read or verify.
[0111] Turning now to FIG. **15**, a flow chart **1500** is provided that depicts the steps of dynamically establishing the kick voltages associated with a plurality of word lines in a memory block according to an example embodiment of the subject disclosure. These steps may be conducted during a die sorting operation following fabrication of the memory chip and does not have to be repeated. These steps could be performed by the controller; a processor or processing device or any other circuitry, executing instructions stored in memory; and/or other circuitry described herein that is specifically configured/programmed to execute the following steps.
[0112] At step **1502**, a selected word line is set at an initial word line in the memory block, e.g., WL**0**. At step **1504**, the WLRC of the selected word line WLn is measured during a die sorting operation. At step **1506**, the selected word line WLn is placed into one of a plurality of Groups based on its measured WLRC with word lines that have similar WLRCs being placed in the same groups. At decision step **1508**, it is determined if the selected word line WLn is the last word line in the memory block.
[0113] If the answer at decision step **1508** is “no,” then at step **1510**, the selected word line is incrementally advanced, i.e., WLn=WLn+1. The process then returns to step **1504**. If the answer at decision step **1508** is “yes,” then at step **1512**, the kick voltages V_Kick for the different groups are all established and stored in the memory device. The magnitudes of the kick voltages V_Kick of the various word line groups are dynamically set based on the differences between the WLRCs of the word lines in the groups. For example, if the difference between the WLRCs of the Group A word lines and the Group C word lines is large then the difference between V_Kick_A and V_Kick_C will be large and if the difference between the WLRCs of the Group A word lines and the Group C word lines is small, then the difference between V_Kick_A and V_Kick_C will be small.
[0114] Turning now to FIG. **16**, a flow chart **1600** is provided that depicts the steps of dynamically establishing the kick voltages associated with a plurality of word lines in a memory block according to an example embodiment of the subject disclosure. These steps may be conducted during a programming or sensing operation following fabrication of the memory chip. These steps could be performed by the controller; a processor or processing device or any other circuitry, executing instructions stored in memory; and/or other circuitry described herein that is specifically configured/programmed to execute the following steps.
[0115] At step **1602**, a voltage ramp up process begins for a given word line. At step **1604**, the kick

voltage V_Kick for the given word line is determined. The kick voltage V_Kick for that particular word line may be saved in the memory device. At step **1606**, the target voltage for the word line is set at an elevated level, i.e., V_Target=V_Intended+V_Kick. The intended voltage V_Intended could be any suitable voltage during a programming or sensing operation including, but not limited to, VPGM, VPASS, VCG, VREAD, or VREADK. At step **1608**, after the kick duration t_Kick, the target voltage is reduced from the elevated level to the intended voltage V_Intended. The programming or sensing operation then continues.

[0116] Various terms are used herein to refer to particular system components. Different companies may refer to a same or similar component by different names and this description does not intend to distinguish between components that differ in name but not in function. To the extent that various functional units described in the following disclosure are referred to as “modules,” such a characterization is intended to not unduly restrict the range of potential implementation mechanisms. For example, a “module” could be implemented as a hardware circuit that includes customized very-large-scale integration (VLSI) circuits or gate arrays, or off-the-shelf semiconductors that include logic chips, transistors, or other discrete components. In a further example, a module may also be implemented in a programmable hardware device such as a field programmable gate array (FPGA), programmable array logic, a programmable logic device, or the like. Furthermore, a module may also, at least in part, be implemented by software executed by various types of processors. For example, a module may comprise a segment of executable code constituting one or more physical or logical blocks of computer instructions that translate into an object, process, or function. Also, it is not required that the executable portions of such a module be physically located together, but rather, may comprise disparate instructions that are stored in different locations and which, when executed together, comprise the identified module and achieve the stated purpose of that module. The executable code may comprise just a single instruction or a set of multiple instructions, as well as be distributed over different code segments, or among different programs, or across several memory devices, etc. In a software, or partial software, module implementation, the software portions may be stored on one or more computer-readable and/or executable storage media that include, but are not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor-based system, apparatus, or device, or any suitable combination thereof. In general, for purposes of the present disclosure, a computer-readable and/or executable storage medium may be comprised of any tangible and/or non-transitory medium that is capable of containing and/or storing a program for use by or in connection with an instruction execution system, apparatus, processor, or device.

[0117] Similarly, for the purposes of the present disclosure, the term “component” may be comprised of any tangible, physical, and non-transitory device. For example, a component may be in the form of a hardware logic circuit that is comprised of customized VLSI circuits, gate arrays, or other integrated circuits, or is comprised of off-the-shelf semiconductors that include logic chips, transistors, or other discrete components, or any other suitable mechanical and/or electronic devices. In addition, a component could also be implemented in programmable hardware devices such as field programmable gate arrays (FPGA), programmable array logic, programmable logic devices, etc. Furthermore, a component may be comprised of one or more silicon-based integrated circuit devices, such as chips, die, die planes, and packages, or other discrete electrical devices, in an electrical communication configuration with one or more other components via electrical conductors of, for example, a printed circuit board (PCB) or the like. Accordingly, a module, as defined above, may in certain embodiments, be embodied by or implemented as a component and, in some instances, the terms module and component may be used interchangeably.

[0118] Where the term “circuit” is used herein, it includes one or more electrical and/or electronic components that constitute one or more conductive pathways that allow for electrical current to flow. A circuit may be in the form of a closed-loop configuration or an open-loop configuration. In a closed-loop configuration, the circuit components may provide a return pathway for the electrical

current. By contrast, in an open-looped configuration, the circuit components therein may still be regarded as forming a circuit despite not including a return pathway for the electrical current. For example, an integrated circuit is referred to as a circuit irrespective of whether the integrated circuit is coupled to ground (as a return pathway for the electrical current) or not. In certain exemplary embodiments, a circuit may comprise a set of integrated circuits, a sole integrated circuit, or a portion of an integrated circuit. For example, a circuit may include customized VLSI circuits, gate arrays, logic circuits, and/or other forms of integrated circuits, as well as may include off-the-shelf semiconductors such as logic chips, transistors, or other discrete devices. In a further example, a circuit may comprise one or more silicon-based integrated circuit devices, such as chips, die, die planes, and packages, or other discrete electrical devices, in an electrical communication configuration with one or more other components via electrical conductors of, for example, a printed circuit board (PCB). A circuit could also be implemented as a synthesized circuit with respect to a programmable hardware device such as a field programmable gate array (FPGA), programmable array logic, and/or programmable logic devices, etc. In other exemplary embodiments, a circuit may comprise a network of non-integrated electrical and/or electronic components (with or without integrated circuit devices). Accordingly, a module, as defined above, may in certain embodiments, be embodied by or implemented as a circuit.

[0119] It will be appreciated that example embodiments that are disclosed herein may be comprised of one or more microprocessors and particular stored computer program instructions that control the one or more microprocessors to implement, in conjunction with certain non-processor circuits and other elements, some, most, or all of the functions disclosed herein. Alternatively, some or all functions could be implemented by a state machine that has no stored program instructions, or in one or more application-specific integrated circuits (ASICs) or field-programmable gate arrays (FPGAs), in which each function or some combinations of certain of the functions are implemented as custom logic. A combination of these approaches may also be used. Further, references below to a “controller” shall be defined as comprising individual circuit components, an application-specific integrated circuit (ASIC), a microcontroller with controlling software, a digital signal processor (DSP), a field programmable gate array (FPGA), and/or a processor with controlling software, or combinations thereof.

[0120] Additionally, the terms “couple,” “coupled,” or “couples,” where may be used herein, are intended to mean either a direct or an indirect connection. Thus, if a first device couples, or is coupled to, a second device, that connection may be by way of a direct connection or through an indirect connection via other devices (or components) and connections.

[0121] Regarding, the use herein of terms such as “an embodiment,” “one embodiment,” an “exemplary embodiment,” a “particular embodiment,” or other similar terminology, these terms are intended to indicate that a specific feature, structure, function, operation, or characteristic described in connection with the embodiment is found in at least one embodiment of the present disclosure. Therefore, the appearances of phrases such as “in one embodiment,” “in an embodiment,” “in an exemplary embodiment,” etc., may, but do not necessarily, all refer to the same embodiment, but rather, mean “one or more but not all embodiments” unless expressly specified otherwise. Further, the terms “comprising,” “having,” “including,” and variations thereof, are used in an open-ended manner and, therefore, should be interpreted to mean “including, but not limited to . . . ” unless expressly specified otherwise. Also, an element that is preceded by “comprises . . . a” does not, without more constraints, preclude the existence of additional identical elements in the subject process, method, system, article, or apparatus that includes the element.

[0122] The terms “a,” “an,” and “the” also refer to “one or more” unless expressly specified otherwise. By way of example, “a processor” programmed to perform various functions refers to one processor programmed to perform each and every function or more than one processor collectively programmed to perform each of the various functions. In addition, the phrase “at least one of A and B” as may be used herein and/or in the following claims, whereby A and B are

variables indicating a particular object or attribute, indicates a choice of A or B, or both A and B, similar to the phrase “and/or.” Where more than two variables are present in such a phrase, this phrase is hereby defined as including only one of the variables, any one of the variables, any combination (or sub-combination) of any of the variables, and all of the variables.
[0123] Further, where used herein, the term “about” or “approximately” applies to all numeric values, whether or not explicitly indicated. These terms generally refer to a range of numeric values that one of skill in the art would consider equivalent to the recited values (e.g., having the same function or result). In certain instances, these terms may include numeric values that are rounded to the nearest significant figure.
[0124] In addition, any enumerated listing of items that is set forth herein does not imply that any or all of the items listed are mutually exclusive and/or mutually inclusive of one another, unless expressly specified otherwise. Further, the term “set,” as used herein, shall be interpreted to mean “one or more,” and in the case of “sets,” shall be interpreted to mean multiples of (or a plurality of) “one or more,” “ones or more,” and/or “ones or mores” according to set theory, unless expressly specified otherwise.
[0125] The foregoing detailed description has been presented for purposes of illustration and description. It is not intended to be exhaustive or be limited to the precise form disclosed. Many modifications and variations are possible in light of the above description. The described embodiments were chosen to best explain the principles of the technology and its practical application to thereby enable others skilled in the art to best utilize the technology in various embodiments and with various modifications as are suited to the particular use contemplated. The scope of the technology is defined by the claims appended hereto.

Claims

1. A method of performing an operation in a memory device, comprising the steps of: preparing a memory block that includes an array of memory cells that are arranged in a plurality of word lines; the word lines of the memory block being associated with respective kick voltages, the kick voltages associated with at least some of the word lines being different than the kick voltages associated with at least some other of the word lines; setting a target voltage for at least one word line of the plurality of word lines at a magnitude that is equal to an intended voltage for the at least one word line plus the respective kick voltage that is associated with the at least one word line; and after a kick duration, reducing the target voltage for the at least one word line to the intended voltage.
2. The method as set forth in claim 1, wherein the plurality of word lines include word lines that are slow to ramp up to a given target voltage and word lines that are fast to ramp up to the given target voltage, and wherein the word lines that are slow to ramp up to the given voltage are associated with higher respective kick voltages than the word lines that are fast to ramp up to the given target voltage.
3. The method as set forth in claim 1, wherein the operation is a programming operation and the intended voltage is a programming voltage or a pass voltage.
4. The method as set forth in claim 1, wherein the operation is a sensing operation and the intended voltage is a reference voltage.
5. The method as set forth in claim 1, wherein the plurality of word lines are arranged in a plurality of groups including a first group and a second group, wherein the word lines in the first group are associated with a first kick voltage, and wherein the word lines in the second group are associated with a second kick voltage that is different than the first kick voltage.
6. The method as set forth in claim 5, wherein the plurality of groups further includes a third group, and wherein the word lines in the third group are associated with a third kick voltage that is different than the first and second kick voltages.

7. The method as set forth in claim 1, wherein the kick voltage that is associated with each word line is based on a word line resistance capacitance of the word line.
8. The method as set forth in claim 7, further including the step of measuring the word line resistance capacitance of each word line to determine the kick voltage that is associated with each word line.
9. The method as set forth in claim 1, wherein for each word line, the associated kick voltage is used during both programming and sensing operations.
10. A method of operating a memory device, comprising the steps of; preparing a memory block that includes an array of memory cells that are arranged in a plurality of word lines; measuring a word line resistance capacitance of the plurality of word lines in the memory; establishing a plurality of groups of word lines, each of the groups of word lines including a plurality of word lines with similar measured word line resistance capacitances; and assigning the plurality of groups different kick voltages based on the word line resistance capacitances of the word lines in each of the groups.
11. The method as set forth in claim 10, wherein the step of assigning the plurality of groups different kick voltages based on the word line resistance capacitances of the word lines in each of the groups includes assigning a first group of word lines a first kick voltage and assigning a second group of word lines a second kick voltage that is different than the first kick voltage.
12. The method as set forth in claim 10, further including the step of performing an operation wherein during a voltage ramp up process for at least one word line of the plurality of word lines, a target voltage is set at a level that is equal to an intended voltage plus the kick voltage that is assigned to the at least one word line.
13. The method as set forth in claim 12, wherein the operation is a programming operation.
14. The method as set forth in claim 12, wherein the operation is a sensing operation.
15. The method as set forth in claim 12, wherein the kick voltages that are assigned to the groups of word lines are used during both programming and sensing operations.
16. The method as set forth in claim 12, wherein the plurality of word lines includes word lines that are slow to ramp up to a given target voltage and word lines that are fast to ramp up to the given target voltage, and wherein the word lines that are slow to ramp up to the given target voltage are assigned kick voltages that are greater than the word lines that are fast to ramp up to the given target voltage.
17. A memory device, comprising: a memory block that includes an array of memory cells that are arranged in a plurality of word lines; the word lines of the memory block being associated with respective kick voltages, the kick voltages associated with at least some of the word lines being different than the kick voltages of at least some other of the word lines; a programming and sensing means for programming and sensing the memory cells of the plurality of word lines, the programming and sensing means being configured to; set a target voltage for at least one word line of the plurality of word lines at a magnitude that is equal to an intended voltage for the at least one word line plus the respective kick voltage that is associated with the at least one word line; and after a kick duration, reduce the target voltage for the at least one word line to the intended voltage.
18. The memory device as set forth in claim 17, wherein the plurality of word lines include word lines that are slow to ramp up to a given target voltage and word lines that are fast to ramp up to the given target voltage, and wherein the word lines that are slow to ramp up to the given voltage are associated with higher respective kick voltages than the word lines that are fast to ramp up to the given target voltage.
19. The memory device as set forth in claim 17, wherein the programming and sensing means is configured to perform programming operation where the intended voltage is a programming voltage or a pass voltage.
20. The memory device as set forth in claim 17, wherein the programming and sensing means is configured to perform a sensing operation where the intended voltage is a reference voltage.