



(51) International Patent Classification:

H01L 21/28 (2006.01) *H01L 29/423* (2006.01)
H01L 29/788 (2006.01) *H01L 29/66* (2006.01)
G11C 16/04 (2006.01) *H01L 27/115* (2006.01)

(21) International Application Number:

PCT/US2012/070833

(22) International Filing Date:

20 December 2012 (20. 12.2012)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

61/580,203 24 December 2011 (24. 12.2011) US
 13/715,682 14 December 2012 (14. 12.2012) US

(71) Applicant (for all designated States except US):
SANDISK TECHNOLOGIES, INC. [US/US]; Two Leg-
 acy Town Center, 6900 North Dallas Parkway, Plano, TX
 75024 (US).

(72) Inventors; and

(71) Applicants (for US only): **DUTTA, Deepanshu** [US/US];
 601 McCarthy Boulevard, Milpitas, CA 95035 (US).
SATO, Shinji [JP/JP]; 601 McCarthy Boulevard, Milpitas,
 CA 95035 (US). **HIGASHITANI, Masaaki** [US/US]; 601
 McCarthy Boulevard, Milpitas, CA 95035 (US). **ZHAO,**
Dengtao [US/—]; 601 McCarthy Boulevard, Milpitas, CA
 95035 (US).

(72) Inventor; and

(71) Applicant : **LEE, Sanghyun** [US/US]; 601 McCarthy
 Boulevard, Milpitas, CA 95035 (US).

(74) Agent: **MAGEN, Burt**; Vierra Magen Marcus & Deniro,
 LLP, 575 Market Street, Suite 2500, San Francisco, CA
 94105 (US).

(81) Designated States (unless otherwise indicated, for every
 kind of national protection available): AE, AG, AL, AM,
 AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY,
 BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,
 DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
 HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP,
 KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD,

[Continued on nextpage]

(54) Title: NON- VOLATILE STORAGE SYSTEM WITH THREE LAYER FLOATING GATE

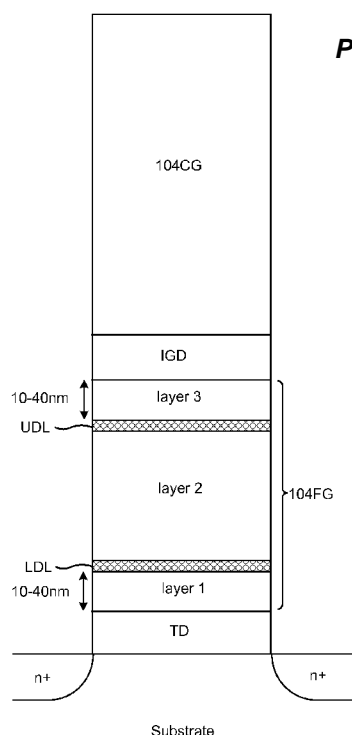


FIG. 4

(57) Abstract: A non-volatile storage system includes memory cells with floating gates (104FG) that comprises three layers separated by two dielectric layers, an upper dielectric layer (UDL) and a lower dielectric layer (LDL). The dielectric layers may be oxide layers, nitride layers, combinations of oxide and nitride, or some other suitable dielectric material. The lower dielectric layer is close to the bottom of the floating gate, near the interface between floating gate and tunnel dielectric (TD), while the upper dielectric layer is close to top of the floating gate, near the interface between floating gate and inter-gate dielectric (IGD).

- 1 -

NON-VOLATILE STORAGE SYSTEM WITH THREE LAYER FLOATING GATE

[0001] This application claim priority to U.S. Provisional Application 61/580,203, "Non-Volatile Storage System With Three Layer Floating Gate," filed on December 24, 2011, incorporated herein by reference in its entirety.

BACKGROUND

Field

[0002] The technology described herein is directed to non-volatile storage.

Description of the Related Art

[0003] Semiconductor memory has 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. Electrical Erasable Programmable Read Only Memory (EEPROM) and flash memory are among the most popular non-volatile semiconductor memories.

[0004] Both EEPROM and flash memory utilize a floating gate that is positioned above and insulated from a channel region in a semiconductor substrate. The floating gate is positioned between the source and drain regions. A control gate is provided over and insulated from the floating gate. The threshold voltage V_t of the transistor is controlled by the amount of charge that is retained on the floating gate. That is, the minimum amount of voltage that must be applied to the control gate before the transistor is turned on to permit conduction between its source and drain is controlled by the level of charge on the floating gate. Thus, a memory cell (which can include one or more

- 2 -

transistors) can be programmed and/or erased by changing the level of charge on a floating gate in order to change the threshold voltage.

[0005] Each memory cell can store data (analog or digital). When storing one bit of digital data (referred to as a binary memory cell), possible threshold voltages of the memory cell are divided into two ranges which are assigned logical data "1" and "0." In one example of a NAND type flash memory, the threshold voltage is negative after the memory cell is erased, and defined as logic "1." After programming, the threshold voltage is positive and defined as logic "0." When the threshold voltage is negative and a read is attempted by applying 0 volts to the control gate, the memory cell will turn on to indicate logic one is being stored. When the threshold voltage is positive and a read operation is attempted by applying 0 volts to the control gate, the memory cell will not turn on, which indicates that logic zero is stored.

[0006] A memory cell can also store multiple levels of information (referred to as a multi-state memory cell). In the case of storing multiple levels of data, the range of possible threshold voltages is divided into the number of levels of data. For example, if four levels of information is stored, there will be four threshold voltage ranges assigned to the data values "11", "10", "01", and "00." In one example of a NAND type memory, the threshold voltage after an erase operation is negative and defined as "11." Positive threshold voltages are used for the states of "10", "01", and "00." If eight levels of information (or states) are stored in each memory cell (e.g. for three bits of data), there will be eight threshold voltage ranges assigned to the data values "000", "001", "010", "011", "100", "101", "110" and "111." The specific relationship between the data programmed into the memory cell and the threshold voltage levels of the memory cell depends upon the data encoding scheme adopted for the memory cells. For example, U.S. Patent No. 6,222,762 and U.S. Patent Application Publication No. 2004/0255090, both of which are incorporated herein by reference in their entirety, describe various data encoding schemes for multi-

- 3 -

state flash memory cells. In one embodiment, data values are assigned to the threshold voltage ranges using a Gray code assignment so that if the threshold voltage of a floating gate erroneously shifts to its neighboring physical state, only one bit will be affected. In some embodiments, the data encoding scheme can be changed for different word lines, the data encoding scheme can be changed over time, or the data bits for random word lines may be inverted to reduce data pattern sensitivity and even wear on the memory cells. Different encoding schemes can be used.

[0007] To read multi-state memory cells, the memory system steps through various predefined control gate voltages corresponding to the various memory states supported by the memory. A sense amplifier may trip (e.g., indicating flow of current) at one or more of these voltages and the system will determine the resultant memory state by consideration of the tripping event(s) of the sense amplifier.

[0008] When programming an EEPROM or flash memory device, such as a NAND flash memory device, typically a program voltage is applied to the control gate and the bit line is grounded. Electrons from the channel are injected into the floating gate. When electrons accumulate in the floating gate, the floating gate becomes negatively charged and the threshold voltage of the memory cell is raised so that the memory cell is in a programmed state. More information about programming can be found in U.S. Patent 6,859,397, titled "Source Side Self Boosting Technique For Non-Volatile Memory," and in U.S. Patent Application Publication 2005/0024939, titled "Detecting Over Programmed Memory," both of which are incorporated herein by reference in their entirety. In many devices, the program voltage (V_{pgm}) applied to the control gate during a programming process is applied as a series of pulses in which the magnitude of the pulses is increased by a predetermined step size ΔV_{pgm} for each successive pulse. Between pulses, one or more verify

- 4 -

operations are performed to determine which memory cells have reached their target.

[0009] Memory cells are erased in one embodiment by raising the p-well to an erase voltage (e.g., 20 volts) for a sufficient period of time and grounding the word lines of a selected block while the source and bit lines are floating. In blocks that are not selected to be erased, word lines are floated. Due to capacitive coupling, the unselected word lines, bit lines, select lines, and the common source line are also raised to a significant fraction of the erase voltage thereby impeding erase on blocks that are not selected to be erased. In blocks that are selected to be erased, a strong electric field is applied to the tunnel oxide layers of selected memory cells and the selected memory cells are erased as electrons of the floating gates are emitted to the substrate side, typically by Fowler-Nordheim tunneling mechanism. As electrons are transferred from the floating gate to the p-well region, the threshold voltage of a selected cell is lowered. Erasing can be performed on the entire memory array, on individual blocks, or another unit of memory cells. One implementation of an erase process includes applying several erase pulses to the p-well and verifying between erase pulses whether the NAND strings are properly erased.

[0010] To increase the capacity of a memory devices, designers of memory systems have aggressively scaled down the size of the devices. However, such reduction in size can cause problems. For example, as the size of memory cells decrease, the amount of noise experienced during programming increases. During programming the program voltage V_{pgm} is applied to the control gates as a series of pulses that are stepped up for each pulse by a step size of ΔV_{pgm} . In response to each pulse, the threshold voltage of the memory cells, on average, increases by $\Delta V_t = \Delta V_{pgm}$. Due to quantum mechanical fluctuations, or other phenomena, the number of electrons injected into the floating gates in response to each pulse may vary across a population of memory cells at any given program pulse; or for a specific cell across different program pulses,

- 5 -

thereby causing small variations in threshold voltages. In other words, sometimes $AV_t \neq AV_{pgm}$ for some memory cells and there is a distribution of AV_t in response to a pulse of the program voltage V_{pgm} . This deviation from the ideal program behavior ($V_t = V_{pgm}$) is referred to as program noise. Program noise can be increased due to manufacturing differences, portions of the floating gate going into depletion, and other conditions. Program noise can result in over programming and/or overlapping threshold voltage distributions among the programmed data states (which lead to read errors)

BRIEF DESCRIPTION OF THE DRAWINGS

- [0011] Figure 1 is a circuit diagram of a NAND string.
- [0012] Figure 2 is a plan view of a portion of a NAND flash memory array.
- [0013] Figure 3 is an orthogonal cross-sectional view taken along line A—A of the portion of the flash memory array depicted in Figure 2.
- [0014] Figure 4 depicts one example memory cell from the portion of the flash memory array depicted in Figure 3.
- [0015] Figure 5 is a flow chart describing one embodiment of a process for fabricating non-volatile storage.
- [0016] Figures 6A-D depict non-volatile storage during various stages of fabrication.
- [0017] Figure 7 is a block diagram of a non-volatile memory system.
- [0018] Figure 8 is a block diagram depicting one embodiment of a memory array.

DETAILED DESCRIPTION

[0019] The technology described herein can be used with various types of non-volatile storage systems. One example is a flash memory system uses the NAND structure, which includes arranging multiple transistors in series, sandwiched between two select gates. The transistors in series and the select gates are referred to as a NAND string. Figure 1 is circuit diagram depicting a NAND string that includes four transistors 100, 102, 104 and 106 in series and sandwiched between a first (or drain side) select gate 120 and a second (or source side) select gate 122. Select gate 120 connects the NAND string to a bit line via bit line contact 126. Select gate 122 connects the NAND string to source line 128. Select gate 120 is controlled by applying the appropriate voltages to select line SGD. Select gate 122 is controlled by applying the appropriate voltages to select line SGS. Each of the transistors 100, 102, 104 and 106 has a control gate and a floating gate. For example, transistor 100 has control gate 100CG and floating gate 100FG. Transistor 102 includes control gate 102CG and a floating gate 102FG. Transistor 104 includes control gate 104CG and floating gate 104FG. Transistor 106 includes a control gate 106CG and a floating gate 106FG. Control gate 100CG is connected to word line WL3, control gate 102CG is connected to word line WL2, control gate 104CG is connected to word line WL1, and control gate 106CG is connected to word line WL0.

[0020] Note that although Figure 1 shows four memory cells in the NAND string, the use of four memory cells is only provided as an example. A NAND string can have less than four memory cells or more than four memory cells. For example, some NAND strings will include eight memory cells, 16 memory cells, 32 memory cells, 64 memory cells, 128 memory cells, etc. The discussion

- 7 -

herein is not limited to any particular number of memory cells in a NAND string.

[0021] A typical architecture for a flash memory system using a NAND structure will include several NAND strings. Each NAND string is connected to the source line by its source select gate controlled by select line SGS and connected to its associated bit line by its drain select gate controlled by select line SGD. Each bit line and the respective NAND string(s) that are connected to that bit line via a bit line contact comprise the columns of the array of memory cells. Bit lines are shared with multiple NAND strings. Typically, the bit line runs on top of the NAND strings in a direction perpendicular to the word lines and is connected to one or more sense amplifiers.

[0022] Relevant examples of NAND type flash memories and their operation are provided in the following U.S. Patents/Patent Applications, all of which are incorporated herein by reference: U.S. Pat. No. 5,570,315; U.S. Pat. No. 5,774,397; U.S. Pat. No. 6,046,935; U.S. Pat. No. 6,456,528; and U.S. Pat. Publication No. US2003/0002348. The discussion herein can also apply to other types of flash memory in addition to NAND, as well as other types of non-volatile memory.

[0023] A portion of a traditional NAND memory array is shown in plan view in Figure 2. BL0-BL4 represent bit line connections to global vertical metal bit lines (not shown). Four floating gate memory cells are shown in each string by way of example. Typically, the individual NAND strings include 16, 32, 64, 128 or more memory cells, forming a column of memory cells. Control gate (word) lines labeled WL0-WL3 extend across multiple strings over rows of floating gates, often in polysilicon.

[0024] Figure 3 is a cross-sectional view taken along line A—A of Figure 2, showing one NAND string. The NAND string includes four memory cells. A NAND string can have less than four memory cells or more than four memory

- 8 -

cells. For example, some NAND strings will include eight memory cells, 16 memory cells, 32 memory cells, 64 memory cells, 128 memory cells, etc. The discussion herein is not limited to any particular number of memory cells in a NAND string. The control gates (100CG, 102CG, 104CG, 106CG) are typically formed over the floating gates (100FG, 102FG, 104FG, 106FG) as a self-aligned stack, and are capacitively coupled to the floating gates through an intermediate dielectric layer (known as the inter-grate dielectric). The top and bottom of the NAND string connect to a bit line (see Metal Bit Line Contact 126) and a common source line (see Metal Source Line Contact 128) through select transistors (gates) 120 and 122, respectively. Gate 170 is controlled by selection line DSL and gate 172 is controlled by selection line SSL. The floating gate material can be shorted to the control gate for the select transistors to be used as the active gate.

[0025] Each memory cells includes a floating gate stack that comprises a floating gate, a control gate and appropriate dielectrics. The floating gate stack is implemented on the top surface of a P-Well, where the P-Well is within an N-Well and the N-Well is positioned in a P type substrate. Between floating gate stacks are n+ diffusion regions, in the P-Well. The n+ diffusion regions serve as the source/drain regions for adjacent floating gate stacks.

[0026] Capacitive coupling between the floating gate and the control gate allows the voltage of the floating gate to be raised by increasing the voltage on the control gate. An individual cell within a column is read and verified during programming by causing the remaining cells in the string to be turned on hard by placing a relatively high voltage on their respective word lines and by placing a relatively lower voltage on the one selected word line so that the current flowing through each string is primarily dependent only upon the level of charge stored in the addressed cell below the selected word line. That current typically is sensed for a large number of strings in parallel, in order to read charge level states along a row of floating gates in parallel. Examples of NAND

- 9 -

memory cell array architectures and their operation as part of a memory system are found in U.S. Pat. Nos. 5,570,315, 5,774,397 and 6,046,935.

[0027] Figure 4 depicts more details of one of the memory cells of the NAND string depicted in Figure 3. Each of the memory cells on the NAND string of Figure 3 will be implemented using the structure depicted in Figure 4.

[0028] The memory cells of Figure 4 includes n+ source/drain regions implemented in the substrate and a floating gate stack implemented on top of the substrate. The floating gate stack includes a tunnel dielectric TD on the top surface of the substrate, a floating gate 104FG on top of the tunnel dielectric layer TD, an inter-gate dielectric IGD on top of floating gate 104FG, and a control gate 104CG on top of the inter-gate dielectric IGD. In one embodiment, the tunnel dielectric TD and the inter-gate dielectric IGD comprise SiO₂; however, other dielectric materials can also be used.

[0029] The control gate can be made of any one of various suitable materials, including semiconductors and/or metals. In one embodiment, the control gate is made from polysilicon, Tungsten, Tungsten Nitride or a combination of the above-listed materials. Other materials can also be used.

[0030] The floating gate (104FG) includes three layers of floating gate material or other conductive material (including semiconductor material and/or metal). For example, Figure 4 shows floating gate (104FG) comprising layer 1, layer 2 and layer 3, all of which (in one embodiment) are made of polysilicon; however, other materials can be used (e.g., Tungsten, Tungsten Nitride or a combination of the above-listed materials). A lower dielectric layer LDL separates layer 1 from layer 2. An upper dielectric layer UDL separates layer 2 from layer 3. In one embodiment, lower dielectric layer LDL and upper dielectric layer UDL are made of SiO₂; however, other dielectric materials can also be used. In some embodiments, lower dielectric layer LDL and upper dielectric layer UDL are made of nitride. In some embodiments, lower dielectric

- 10 -

layer LDL and upper dielectric layer UDL are made of nitride and SiO₂. Thus, floating gate (104FG) includes three separate charge storing layers separated by dielectric layers.

[0031] Lower dielectric layer LDL is close to the bottom of the floating gate (the interface between the floating gate and the tunnel dielectric TD), while the upper dielectric layer UDL is close to the top of the floating gate (the interface between the floating gate and the inter-gate dielectric IGD). In one set of embodiments, the upper dielectric layer UDL is close to a top of the floating gate such that it is less than one third of the height of the floating gate down from the top of the floating gate and the lower dielectric layer LDL is close to a bottom of the floating gate such that it is less than one third of the height of the floating gate up from the bottom of the floating gate. In one example, lower dielectric layer LDL is 10nm to 40nm above the bottom interface of the floating gate, while the upper dielectric layer UDL is 10nm to 40nm below the top interface of the floating gate. Different positions of the dielectric layers may be used. The thickness of the dielectric layers LDL and UDL should be between 0.5nm to 1.5nm. If the layer is too thick, it prevents free transport of electrons/holes within the floating gate. However if the dielectric layer is too thin, it may not be sufficient to isolate one floating layer from another.

[0032] In one embodiment, layer 1 is less than one third of the height of the total floating gate, layer 2 is more than one third of the height of the total floating gate, and layer 3 is less than one third of the height of the total floating gate. In such an embodiment, upper dielectric layer UDL is close to the top interface of the floating gate such that it is less than one third of the height of the floating gate down from the top of the floating gate and lower dielectric layer LDL is less than one third of the height of the floating gate up from the bottom of the floating gate.

- 11 -

[0033] The two dielectric layers may be made of same material or different material. For example, upper dielectric layer UDL may be an oxide, while lower dielectric layer LDL may be nitride, or vice versa. Apart from the helping with generation of electron-hole pairs, the dielectric layer also helps with separating floating layers that may be doped with different specie or different dopant dose, ensuring that there is no mixing between the two adjacent floating layers. The material of choice for the dielectric layer may be dependent on the impurity that was doped. For example, boron is known to diffuse through oxide, but not nitride. Thus, to separate Boron dopant between different floating gate layers, a nitride may be used for the dielectric layer.

[0034] Another advantage of having dielectric layers within the floating gate is that it offers control over the grain size of the polycrystalline silicon. For example, a polysilicon floating gate having no dielectric-layer will have larger grain size and thus its behavior will be close to single crystal silicon. Dielectric layers restrict the growth of the grains by separating one layer of floating gate from another, resulting in a smaller grain size. A smaller grain size is preferred since it helps with electron-hole pair generation by creating of interface states between adjacent grains. Easier electron-hole pair generation, in turn helps with floating gate depletion to inversion transition, thus improving the program noise.

[0035] In one embodiment, layer 1, layer 2 and layer 3 are made of p+ polysilicon. In other embodiments p, p-, n, n- and n+ polysilicon can also be used. Here 'p' or 'n' refers to the type of dopant or conductivity, where p indicates acceptor-type (such as Boron) while n indicates donor type (such as Phosphorus). The '+' and '-' refers to the strength of doping, where '+' indicates higher doping levels, while '-' indicates lower doping levels. In some embodiments, layer 1, layer 2 and layer 3 are made from the same materials. In other embodiments, layer 1, layer 2 and layer 3 are made from different materials. In some embodiments, layer 1, layer 2 and layer 3 are made using the

- 12 -

same doping (including same conductivity type). In other embodiments, layer 1, layer 2 and layer 3 are made using different doping (including different conductivity type or same conductivity type but different amounts). The three layers of the floating gate can have different doping profiles. Even if the doping is the same type, the amount of doping may vary among the layers. For example, in one embodiment, layer 1 is p+ polysilicon, layer 2 is p polysilicon and layer 3 is p polysilicon. In another embodiment, layer 1 is p+ polysilicon, layer 2 is intrinsic polysilicon and layer 3 is n+ polysilicon. In another embodiment, layer 1 is n+ polysilicon, layer 2 is p+ polysilicon and layer 3 is p+ polysilicon. In another embodiment, layer 1 is p+ polysilicon, layer 2 is p-polysilicon and layer 3 is n+ polysilicon. Other combinations can also be used. Apart from standard donor and acceptor dopings (Boron, Phosphorus, Arsenic etc.), Nitrogen doping or Carbon doping may also be done, and combinations of them are also possible. One or more floating gate layers may be undoped (intrinsic polysilicon) or a metal (such as W, Cu etc).

[0036] In some embodiments, the floating gate can include more than three separate charge storing layers separated by dielectric layers.

[0037] One or more of the floating gate layers (layer 1, layer 2 and layer 3) may receive a dosage of certain heavy atoms (such as As) that can create defects in floating gate's polysilicon, which can help with generation of electron-hole pairs, which may help with inversion layer formation (as the dielectric layers UDL and LDL do).

[0038] It has been observed that for n+ polysilicon floating gate, the presence of the lower dielectric layer LDL reduces program noise during a program process. In other words, when the floating gates include the lower dielectric layer LDL, the change in threshold voltage AV_t of a memory cell in response to a program pulse is more likely to be equal to AV_{pgm} , thereby

- 13 -

resulting in a tighter threshold voltage distribution of programmed data states and less over programming, both of which lead to less subsequent read errors.

[0039] The presence of the upper dielectric layer UDL may reduce noise during an erase process for a n+ polysilicon floating gate. In other words, when the floating gates include the upper dielectric layer UDL, the changes in threshold voltage AV_t of the memory cell in response to an erase pulse is more likely to be uniform (or close to uniform), thereby resulting in a tighter erase threshold voltage distribution and less over erasing, both of which lead to less subsequent programming problems and improved endurance.

[0040] The presence of the upper dielectric layer UDL may reduce program noise during a programming process for a p+ polysilicon floating gate. The presence of the lower dielectric layer LDL reduces noise during an erase process.

[0041] Figure 5 is a flow chart describing one embodiment of the front end of a process for manufacturing the memory cell of Fig. 4, which covers process steps only as far as forming the source/drain regions. This flow does not cover the optional booster plates or fins, the gap fill of etched volumes between the stacks, or forming the contacts, metallizations, vias, and passivation. There are many ways to manufacture memory according to the present invention and, thus, the inventors contemplate that various methods other than that described by Fig. 5 can be used. While a flash memory chip will consist of both a peripheral circuitry, which includes a variety of low, medium, and high voltage transistors, and the core memory array, the process steps of Figure 5 are intended only to describe in general terms one possible process recipe for the fabrication of the core memory array. Many photolithography, etch, implant, diffusion and oxidation steps that are intended for the fabrication of the peripheral transistors are omitted.

- 14 -

[0042] Step 140 of Fig. 5 includes performing implants and associated anneals of the triple well. In step 142, the tunnel dielectric TD (e.g., oxide) is grown or deposited using ALD, CVD, PVD, Jet Vapor Deposition (JVD) or another suitable process. The result of step 142 is depicted in Figure 6A, which depicts the P substrate, N-well within the P-substrate, and P-Well within the N-well 22. The sidewalls of the N-well that isolate the P-wells from one another are not depicted. Also the N-well depth is typically much thicker than that of the P-well in contrast to Figure 6A. The P substrate is usually the thickest consisting of the majority of the wafer thickness. On the top surface of the P-well is the tunnel dielectric TD oxide.

[0043] Steps 144-152 include creating the three layer floating gate. In step 144, layer 1 (e.g., polysilicon) of the floating gate is deposited over tunnel dielectric TD using CVD, PVD, ALD or another suitable method. In step 146, the lower dielectric layer LDL (e.g., oxide) is grown or deposited over layer 1 using ALD, CVD, PVD, Jet Vapor Deposition (JVD) or another suitable process. In step 148, layer 2 (e.g., polysilicon) of the floating gate is deposited over lower dielectric layer LDL using CVD, PVD, ALD or another suitable method. In step 150, the upper dielectric layer UDL (e.g., oxide) is grown or deposited over layer 2 using ALD, CVD, PVD, Jet Vapor Deposition (JVD) or another suitable process. In step 152 layer 3 (e.g., polysilicon) of the floating gate is deposited over upper dielectric layer UDL using CVD, PVD, ALD or another suitable method. Figure 6B shows the memory structure after step 152, including layer 1 over tunnel dielectric TD, lower dielectric layer LDL over layer 1, layer 2 over lower dielectric layer LDL, upper dielectric layer UDL over layer 2, and layer 3 over upper dielectric layer UDL. Note that Figure 6B does not show the N-Well or P-Substrate, in order to simplify the drawing.

[0044] Step 154 of Fig. 5 includes depositing a hard mask using, for example, CVD, to deposit SiO_2 or Si_3N_4 . In step 156, photolithography is used to form strips of photoresist over what will become the NAND chains. Step 158

- 15 -

includes etching through all layers, including part of the substrate. First, the hard mask is etched through using anisotropic plasma etching, (i.e. reactive ion etching with the proper balance between physical and chemical etching for each planar layer encountered). After the hard mask layer is etched into strips, the photoresist can be stripped away and the hard mask layer can be used as the mask for etching the underlying layers. The process, then includes etching through the floating gate material, the tunnel dielectric material and into the substrate to create trenches between the NAND strings, where the bottom of the trenches are inside the top of the P-well. In step 160, the trenches are filled with SiO₂ (or another suitable material) up to the top of the hard mask using CVD, rapid ALD or PSZ STI fill as described in "Void Free and Low Stress Shallow Trench Isolation Technology using P-SOG for sub 0.1 Device" by Jin-Hwa Heo, et. al. in 2002 Symposium on VLSI Technology Digest of Technical Papers, Session 14-1. PSZ STI fill is Polysilazane Shallow trench isolation fill. The fill sequence includes spin coat by coater, and densify by furnace. Si-N bond conversion to Si-O bond enables less shrinkage than conventional SOG (Spin On Glass). Steam oxidation is effective for efficient conversion. One proposal is to use Spin-On-Glass (SOG) for the dielectric layer, which is called polysilazane-based SOG (SZ-SOG), a material used in integrating the inter layer dielectric (ILD) applications because of its excellent gap filling and planarization properties, and thermal oxide like film qualities.

[0045] In step 162, Chemical Mechanical Polishing (CMP), or another suitable process, is used to polish the material flat until reaching the floating gate poly-silicon. In step 164, the inter-gate dielectric IGD is grown or deposited using ALD, CVD, PVD, Jet Vapor Deposition (JVD) or another suitable process. Examples of materials that can be used for the inter-gate dielectric IGD include (but are not limited to) SiO₂, Si₃N₄, an alloy with a varying mole fraction as a function of depth, an alloy or nano-laminate of aluminum oxide and silicon oxide, an alloy or nano-laminate of silicon nitride

- 16 -

and silicon oxide, an alloy or nano-laminate of silicon oxide and hafnium oxide, an alloy or nano-laminate of aluminum oxide and hafnium oxide, or other suitable materials. In step 166 of Figure 5, which is an optional step, the inter-gate dielectric IGD is annealed to densify the oxide.

[0046] In step 168, the one or more layers of the control gate are deposited on the inter-gate dielectric IGD. In one embodiment, the materials deposited during step 168 include a polysilicon, while in other embodiments this layer may be a metal layer with a proper work function, thermal stability, and etch characteristics. In some embodiments, the control gate is composed of the polysilicon layer, tungsten-nitride layer, and tungsten layer, all of which are deposited in step 168. The nitride layer and tungsten layer are deposited to reduce the control gate sheet resistance and form lower resistivity word lines. These materials can be deposited in a blanket form using CVD, ALD, PVD or other suitable process. Figure 6C, which shows the Control Gate Polysilicon, WN layer and Tungsten W metal layer 40 over inter-gate dielectric IGD, depicts the device after step 168.

[0047] On top of the Tungsten layer, a hard mask of Si_3N_4 is deposited using, for example, CVD in step 170. In step 172, photolithography is used to create patterns of perpendicular strips to the NAND chain, in order to etch the multi-gate stack and form word lines (i.e. control gates) that are isolated from one another. In step 174, etching is performed using plasma etching, ion milling, ion etching that is purely physical etching, or another suitable process to etch the various layers and form the individual word lines. In one embodiment, the etching is performed until the tunnel dielectric TD is reached. In another embodiment, the process will etch all the way to the substrate. Figure 6D, which shows the floating gate stack, depicts the device after step 174.

- 17 -

[0048] In step 176, an implant process is performed to create the N+ source/drain regions by Arsenic implantation. In one embodiment, a halo implant is also used. In step 178, an anneal process is performed. There are many alternatives to the above described structures and processes within the spirit of the present invention.

[0049] In one proposed embodiment of the p+ poly silicon three layer floating gate (with p+ polysilicon control gate), a first layer which is 25nm un-doped poly Si is deposited and then 55nm 2nd un-doped polysilicon is deposited after O₂ purge. Next, O₂ purge is done again to separate 2nd and 3rd polysilicon before 25nm 3rd layer polysilicon deposition. Thus, the three layer polysilicon floating gate (three layers of polysilicon) which are separated by SiO₂ is formed. Each layers' thickness should be optimized to minimize the program noise, thereby resulting in a tighter threshold voltage distribution of programmed data states. To make these 3-layers P+ polysilicon, Boron (ex. 5keV, 3.0 E+15 cm⁻²) is implanted, using a lithography mask that covers peripheral area. After that, Phos. (ex. 5keV, 3.0 E+15 cm⁻²) for N+ polysilicon of peripheral transistors (not in the memory array) is implanted while the cell area is covered with the lithography mask. In an alternative, Boron is implanted in all areas (blanket implant), and later Phosphorous is implanted in peripheral transistors area (only) to counter the Boron dose. Such a method can reduce the number of masks involved in the process, which is a cost saving. However, it comes with the difficulty of optimizing the counter-implant and some side effects that Boron implantation in peripheral transistors area may cause. Either way, both Boron and Phosphorous implantation energy needs to be low enough to avoid Boron and Phosphorous penetration through floating gate polysilicon to tunnel dielectric and silicon substrate. After some wet cleaning and ashing processes, the top of the floating gate is oxidized by Rapid Thermal Oxidation (ex. 950C, 20sec) for Boron and Phosphorous atom's activation. Thus, a 3layer P+ polysilicon floating gate is formed.

- 18 -

[0050] In above process, the O₂ purge process is used to separate the three layers with oxide, however, N₂ or NH₃ purge process is another suitable candidate to separate the three layers with nitride. In the later case, as Boron diffusion is easily blocked by N₂ or NH₃ purged layer, 2 or 3 step Boron implantation is another method (ex, 1keV + 3keV + 5keV).

[0051] After forming the 3-layer floating gate, conventional STI and EB processes are applied and Boron doped GC-1 polysilicon and GC-2 polysilicon is formed. After that, Phosphorous is implanted to peripheral region after lithography process to cover cell area. Thus, P+ floating gate/control gate for memory cells and N+ poly gate for peripheral transistors are formed.

[0052] Figure 7 illustrates a memory system that includes memory cells of the structure described above. Memory device 210 having read/write circuits for reading and programming a page (or other unit) of memory cells (e.g., NAND multi-state flash memory or other type) in parallel. Memory device 210 may include one or more memory die or chips 212. Memory die 212 includes an array (two-dimensional or three dimensional) of memory cells 200 (fabricated as discussed above), control circuitry 220, and read/write circuits 230A and 230B. In one embodiment, access to the memory array 200 by the various peripheral circuits is implemented in a symmetric fashion, on opposite sides of the array, so that the densities of access lines and circuitry on each side are reduced by half. The read/write circuits 230A and 230B include multiple sense blocks 300 which allow a page of memory cells to be read or programmed in parallel. The memory array 200 is addressable by word lines via row decoders 240A and 240B and by bit lines via column decoders 242A and 242B. Word lines and bit lines are examples of control lines. In a typical embodiment, a controller 244 is included in the same memory device 210 (e.g., a removable storage card or package) as the one or more memory die 212; however, the controller can also be separate. Commands and data are transferred between the

- 19 -

host and controller 244 via lines 232 and between the controller and the one or more memory die 212 via lines 234.

[0053] Control circuitry 220 cooperates with the read/write circuits 230A and 230B to perform memory operations on the memory array 200. The control circuitry 220 includes a state machine 222, an on-chip address decoder 224 and a power control module 226. The state machine 222 provides chip-level control of memory operations. The on-chip address decoder 224 provides an address interface between that used by the host or a memory controller to the hardware address used by the decoders 240A, 240B, 242A, and 242B. The power control module 226 controls the power and voltages supplied to the word lines and bit lines during memory operations. In one embodiment, power control module 226 includes one or more charge pumps that can create voltages larger than the supply voltage.

[0054] In one embodiment, one or any combination of control circuitry 220, power control circuit 226, decoder circuit 224, state machine circuit 222, decoder circuit 242A, decoder circuit 242B, decoder circuit 240A, decoder circuit 240B, read/write circuits 230A, read/write circuits 230B, and/or controller 244 can be referred to as one or more managing circuits. The one or more managing circuits perform the processes to read, write and program.

[0055] Figure 8 depicts an exemplary structure of memory cell array 200. In one embodiment, the array of memory cells is divided into a large number of blocks (e.g., blocks 0 - 1023, or another amount) of memory cells. As is common for flash EEPROM systems, the block is the unit of erase. That is, each block contains the minimum number of memory cells that are erased together. Other units of eras can also be used. The memory cells of array 200 are fabricated as discussed above.

[0056] A block contains a set of NAND strings which are accessed via bit lines (e.g., bit lines BL0 - BL69623) and word lines (WL0, WL1, WL2, WL3).

Figure 8 shows four memory cells connected in series to form a NAND string. Although four cells are shown to be included in each NAND string, more or less than four can be used (e.g., 16, 32, 64, 128 or another number of memory cells can be on a NAND string). One terminal of the NAND string is connected to a corresponding bit line via a drain select gate (connected to select gate drain line SGD), and another terminal is connected to the source line via a source select gate (connected to select gate source line SGS).

[0057] One embodiment includes a non-volatile storage device, comprising: a floating gate comprising three separate charge storage layers separated by floating gate dielectric layers; a tunnel dielectric layer on a surface of a substrate, between the substrate and the floating gate; a control gate; and an inter-gate dielectric between the floating gate and the control gate.

[0058] One embodiment includes a non-volatile storage device, comprising: a substrate; a tunnel dielectric layer on a surface of the substrate; a bottom charge storing layer of a floating gate above the tunnel dielectric layer; a lower dielectric layer above the bottom charge storing layer of the floating gate; a middle charge storing layer of the floating gate above the lower dielectric layer; an upper dielectric layer above the middle charge storing layer of the floating gate; a top charge storing layer of the floating gate above the upper dielectric layer; an inter-gate dielectric layer above the top charge storing layer of the floating gate; and a control gate layer above the inter-gate dielectric layer.

[0059] One embodiment includes a method for fabricating non-volatile storage, comprising: adding a tunnel dielectric layer on a surface of a substrate; adding a bottom charge storage layer of a floating gate above the tunnel dielectric layer; adding a lower dielectric layer above the bottom charge storage layer of the floating gate; adding a middle charge storage layer of the floating gate above the lower dielectric layer; adding an upper dielectric layer above the middle charge storage layer of the floating gate; adding a top charge storage

- 21 -

layer of the floating gate above the upper dielectric layer, the middle charge storage layer is separate from the bottom charge storage layer and the top charge storage layer; adding an inter-gate dielectric layer above the top charge storage layer of the floating gate; and adding a control gate layer above the inter-gate dielectric layer. Some embodiments include adding a dosage of heavy atoms to at least one of the bottom charge storage layer, middle charge storage layer or top charge storage layer. In some embodiments, the adding the upper dielectric layer comprises adding the upper dielectric layer close to a top of the floating gate such that it is less than one third of the height of the floating gate down from the top of the floating gate; and the adding the lower dielectric layer comprises adding the lower dielectric layer close to a bottom of the floating gate such that it is less than one third of the height of the floating gate up from the bottom of the floating gate.

[0060] One embodiment includes a method for operating non-volatile storage, comprising: programming a non-volatile storage element by adding electrons to a floating gate comprising three separate charge storage layers separated by floating gate dielectric layers, the adding electrons to the floating gate changes a threshold voltage of the non-volatile storage element; and reading the non-volatile storage element by testing a threshold voltage of the non-volatile storage element.

[0061] The foregoing detailed description has been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the subject matter claimed herein to the precise form(s) disclosed. Many modifications and variations are possible in light of the above teachings. The described embodiments were chosen in order to best explain the principles of the disclosed 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. It is

- 22 -

intended that the scope of the invention be defined by the claims appended hereto.

CLAIMS

What is claimed is:

1. A non-volatile storage device, comprising:
a floating gate comprising three separate charge storage layers separated by floating gate dielectric layers;
a tunnel dielectric layer on a surface of a substrate, between the substrate and the floating gate;
a control gate; and
an inter-gate dielectric between the floating gate and the control gate.
2. The non-volatile storage device according to claim 1, wherein:
two or more of the three separate charge storage layers include a dosage of heavy atoms.
3. The non-volatile storage device according to claim 1, wherein:
one or more of the three separate charge storage layers are doped with Carbon.
4. The non-volatile storage device according to claim 1, wherein:
one or more of the three separate charge storage layers are doped with Nitrogen.
5. The non-volatile storage device according to any of claims 1-4, wherein:
the floating gate dielectric layers comprise an upper dielectric layer and a lower dielectric layer;

- 24 -

the upper dielectric layer is close to a top of the floating gate such that it is less than one third of the height of the floating gate down from the top of the floating gate; and

the lower dielectric layer is close to a bottom of the floating gate such that it is less than one third of the height of the floating gate up from the bottom of the floating gate.

6. The non-volatile storage device according to any of claims 1-5, wherein:

the three separate charge storage layers are of a same conductivity type, but different doping amounts.

7. The non-volatile storage device according to any of claims 1-6, wherein:

the three separate charge storage layers are doped differently.

8. The non-volatile storage device according to any of claims 1-5, wherein:

one of the three separate charge storage layers is n type polysilicon;
one of the three separate charge storage layers is p type polysilicon; and
one of the three separate charge storage layers is intrinsic polysilicon.

9. The non-volatile storage device according to claim 1, wherein:
at least one of the three separate charge storage layers is doped with boron.

10. The non-volatile storage device according to any of claims 1-9, wherein:

at least one of the floating gate dielectric layers comprise nitride and oxide.

- 25 -

11. The non-volatile storage device according to any of claims 1-10, wherein:

the floating gate further comprises additional separate charge storage layers separated by floating gate dielectric layers.

12. The non-volatile storage device according to any of claims 1-11, wherein:

the floating gate, tunnel dielectric, control gate and inter-gate dielectric comprise a NAND flash memory cell.

13. A method for fabricating non-volatile storage, comprising:
adding a tunnel dielectric layer on a surface of a substrate;
adding a bottom charge storage layer of a floating gate above the tunnel dielectric layer;
adding a lower dielectric layer above the bottom charge storage layer of the floating gate;
adding a middle charge storage layer of the floating gate above the lower dielectric layer;
adding an upper dielectric layer above the middle charge storage layer of the floating gate;
adding a top charge storage layer of the floating gate above the upper dielectric layer, the middle charge storage layer is separate from the bottom charge storage layer and the top charge storage layer;
adding an inter-gate dielectric layer above the top charge storage layer of the floating gate; and
adding a control gate layer above the inter-gate dielectric layer.

14. The method for fabricating non-volatile storage according to claim 13, wherein:

- 26 -

the adding the lower dielectric layer comprises growing a first oxide layer; and

the adding the upper dielectric layer comprises growing a second oxide layer.

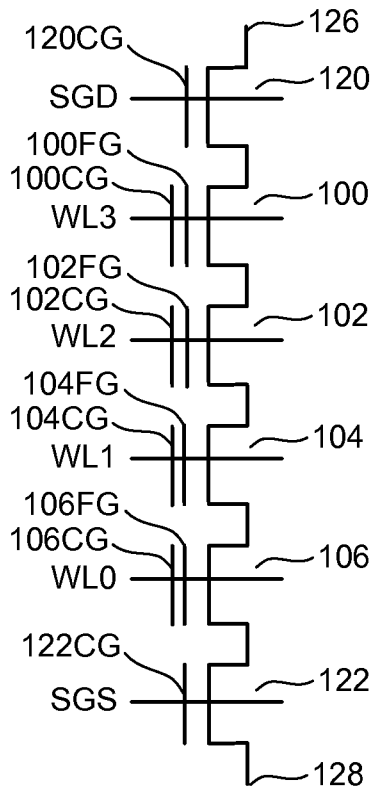
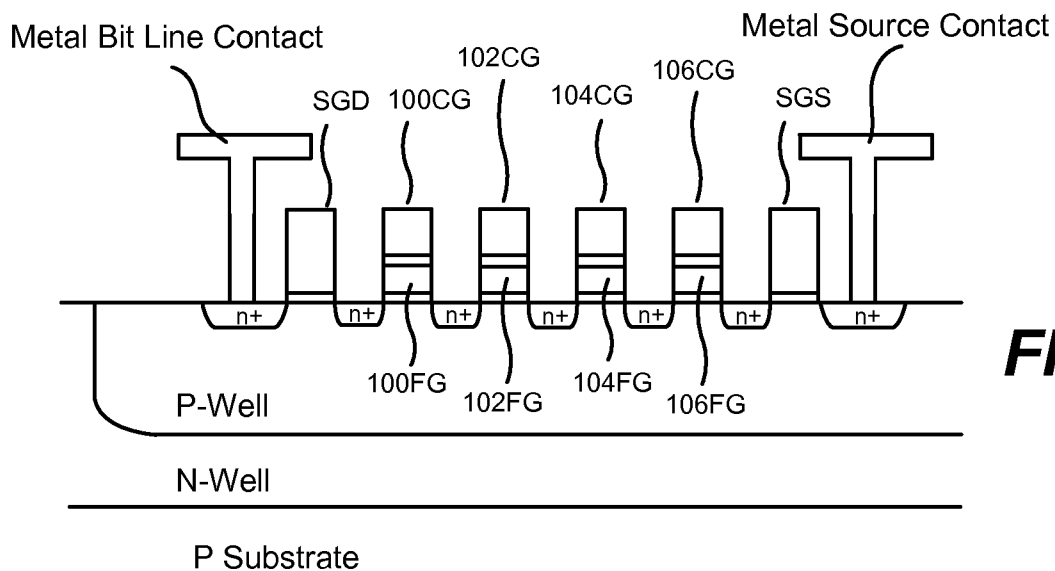
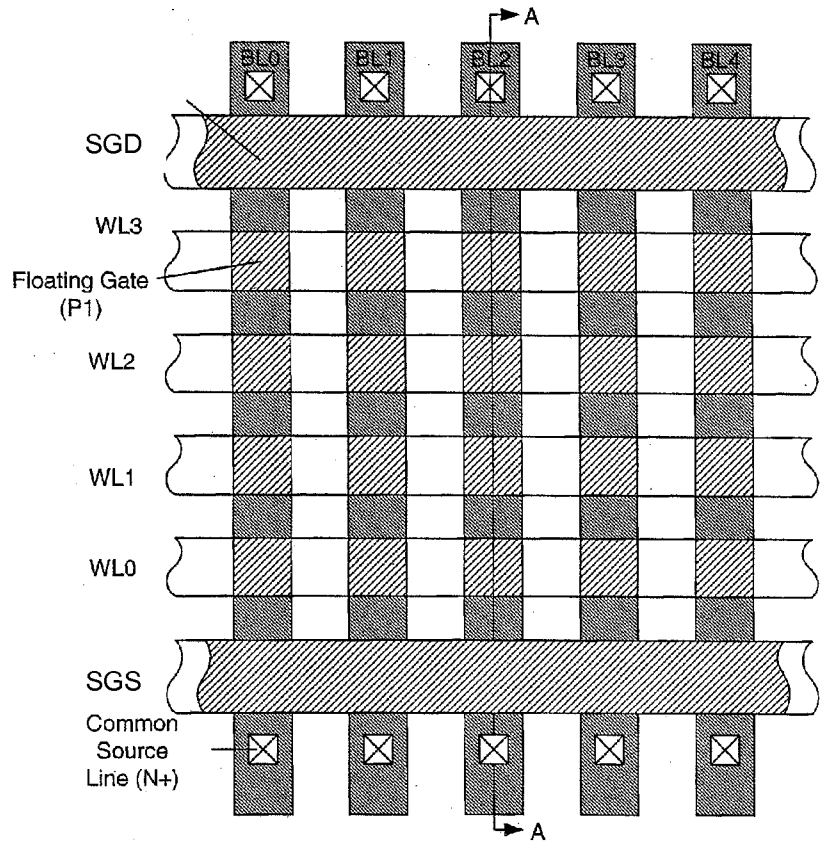
15. The method for fabricating non-volatile storage according to claims 13 or 14, further comprising:

adding a dosage of heavy atoms to at least one of the bottom charge storage layer, middle charge storage layer or top charge storage layer.

16. The method for fabricating non-volatile storage according to claims 13, 14 or 15, wherein:

the adding the upper dielectric layer comprises adding the upper dielectric layer close to a top of the floating gate such that it is less than one third of the height of the floating gate down from the top of the floating gate; and

the adding the lower dielectric layer comprises adding the lower dielectric layer close to a bottom of the floating gate such that it is less than one third of the height of the floating gate up from the bottom of the floating gate.

FIG. 1**FIG. 2****FIG. 3**

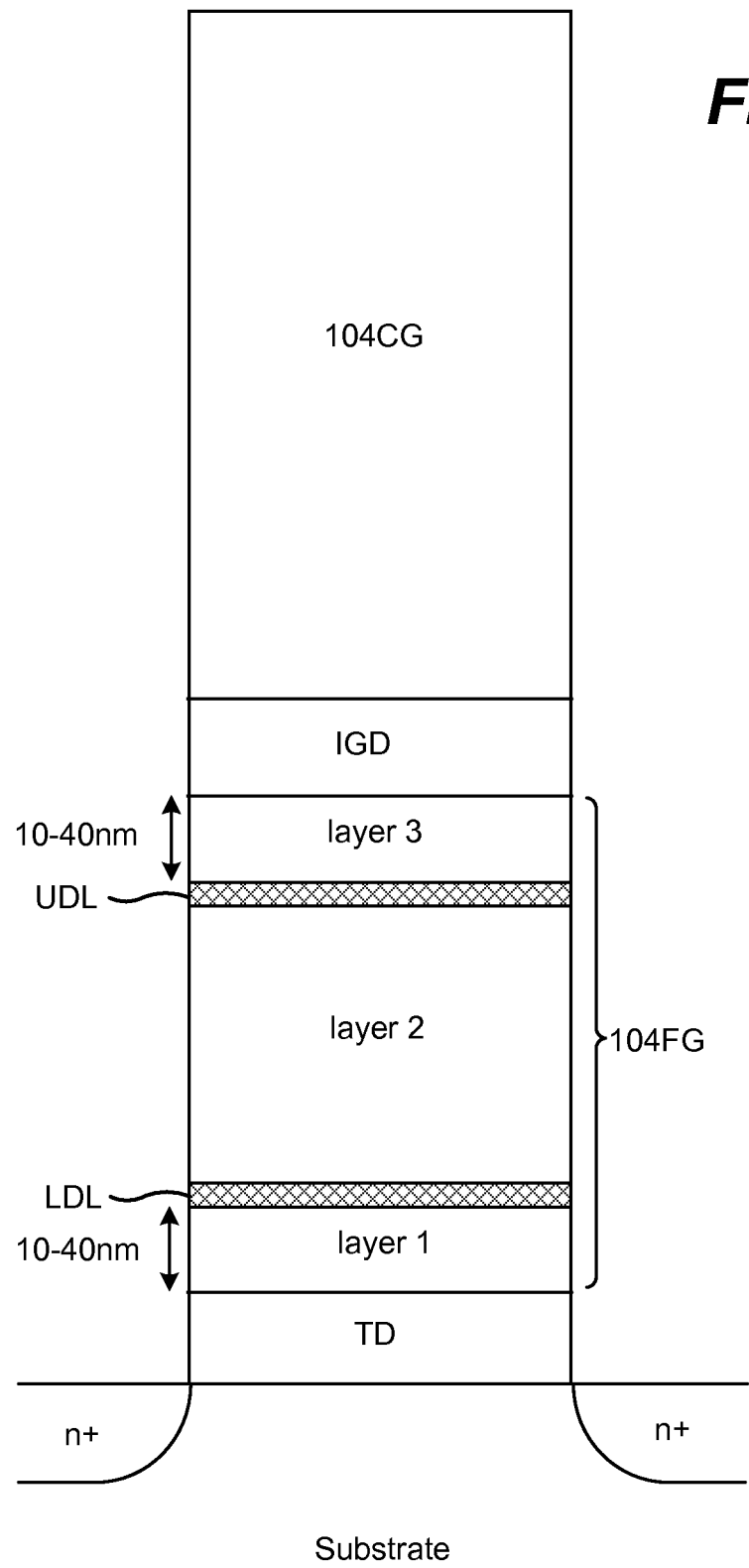


FIG. 4

Fig. 5

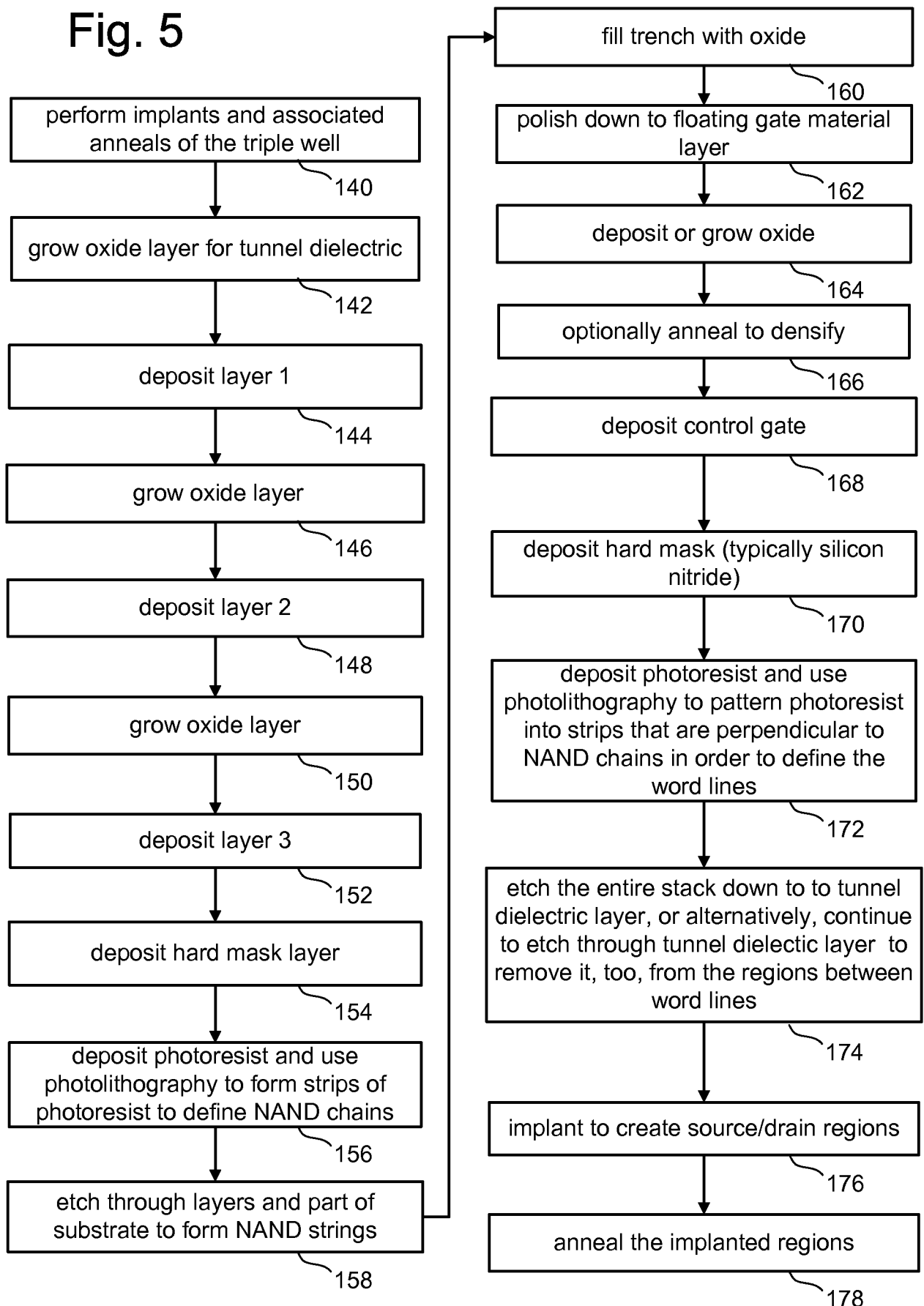


FIG. 6A

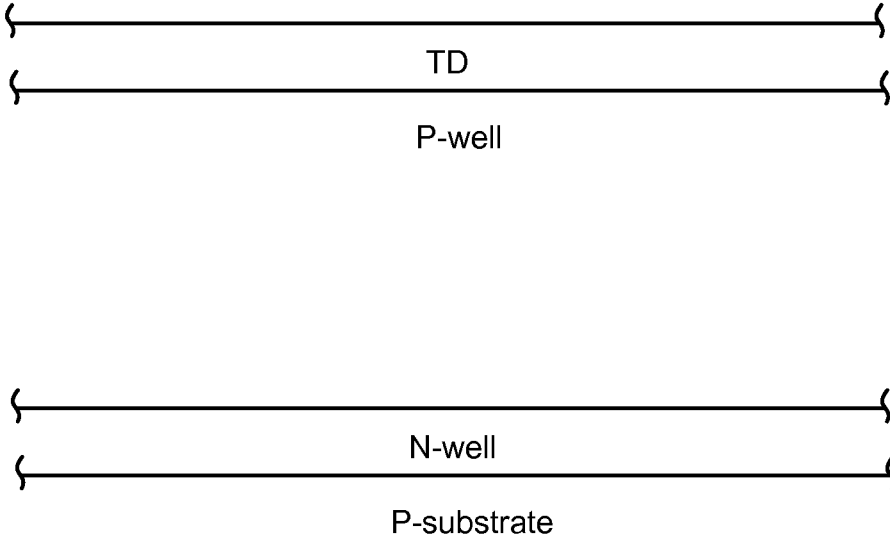


FIG. 6B

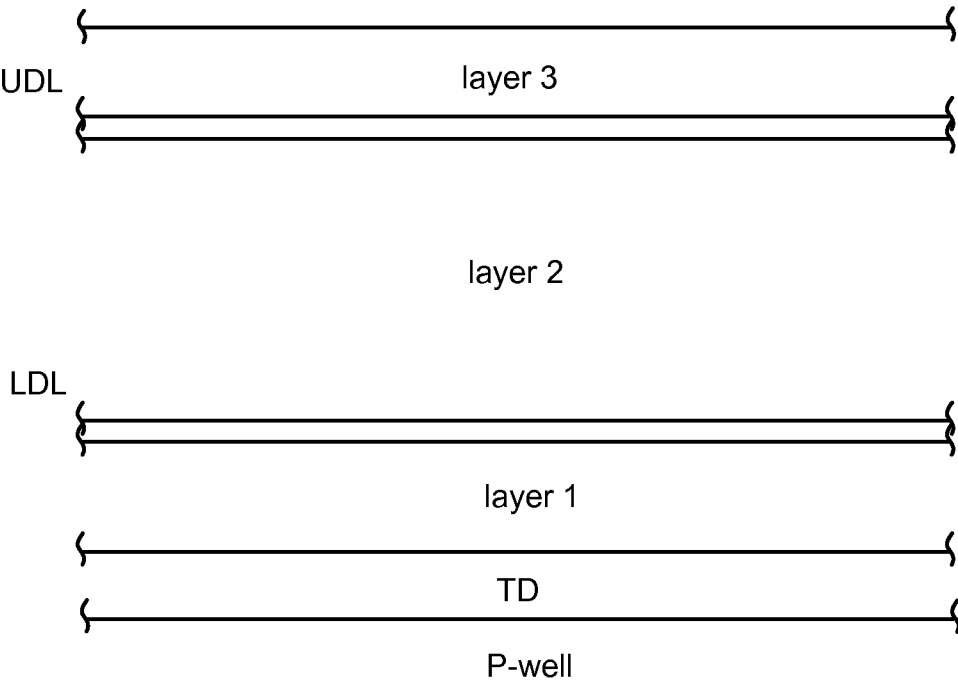
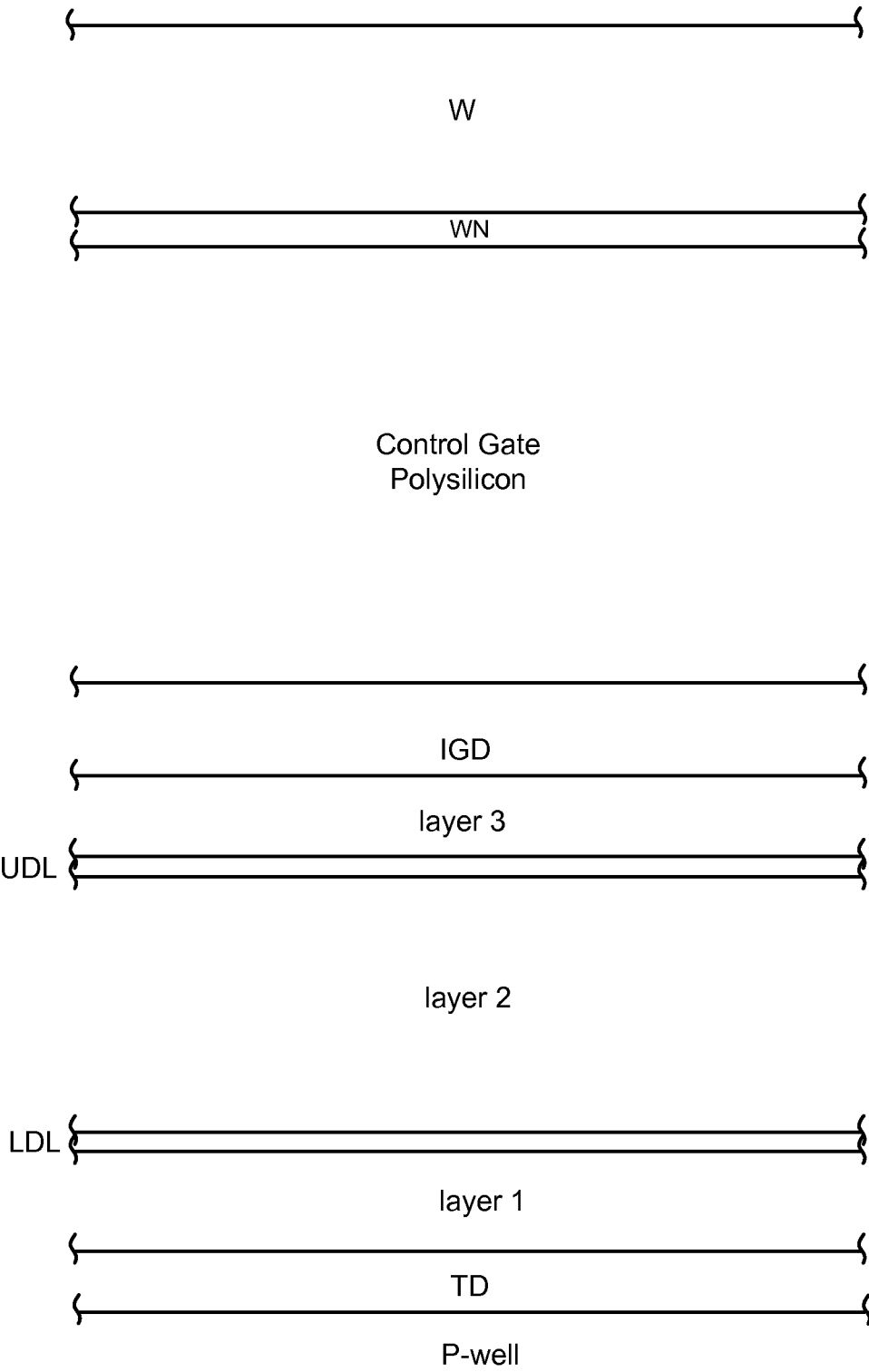


FIG. 6C



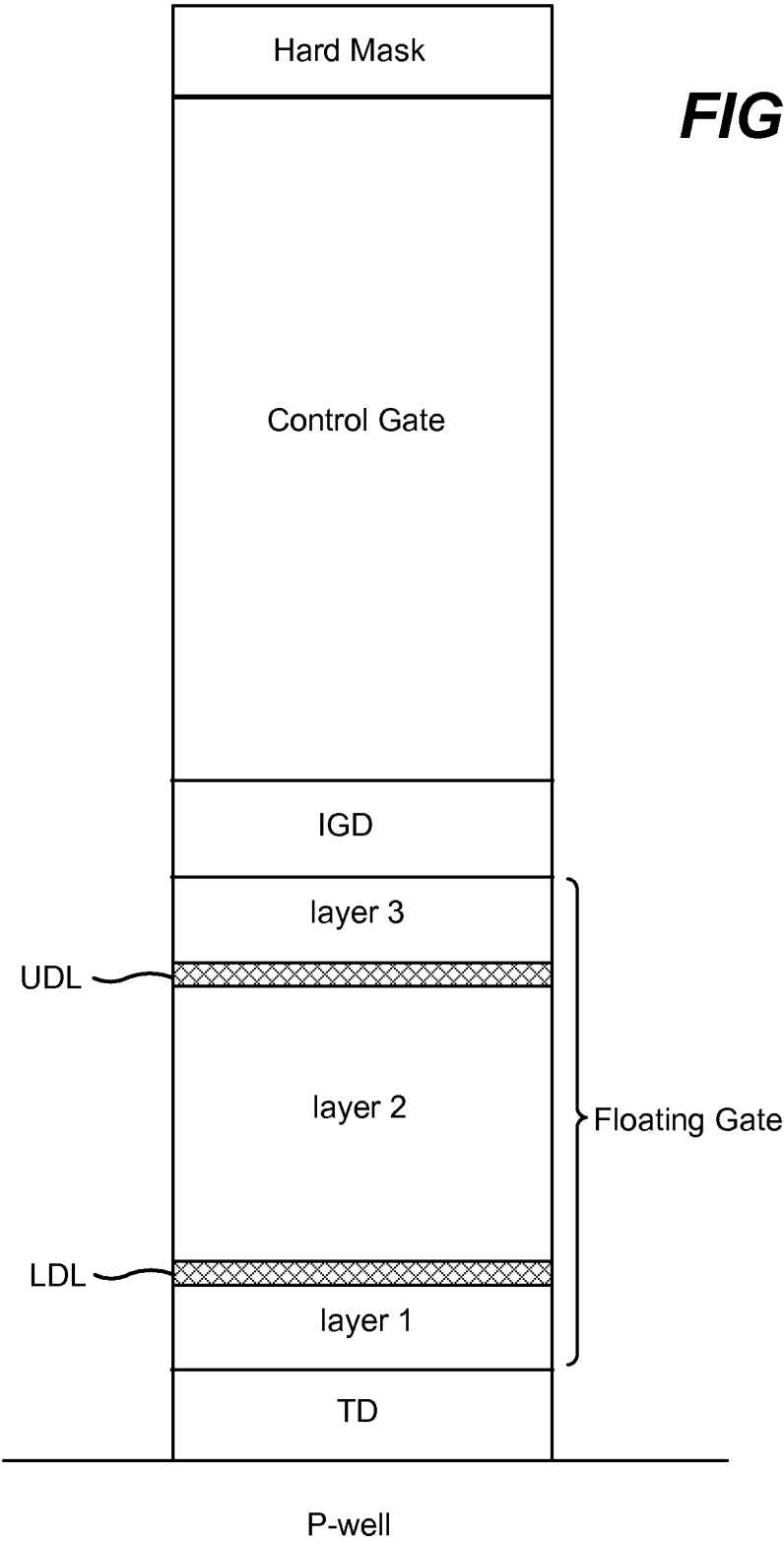


FIG. 6D

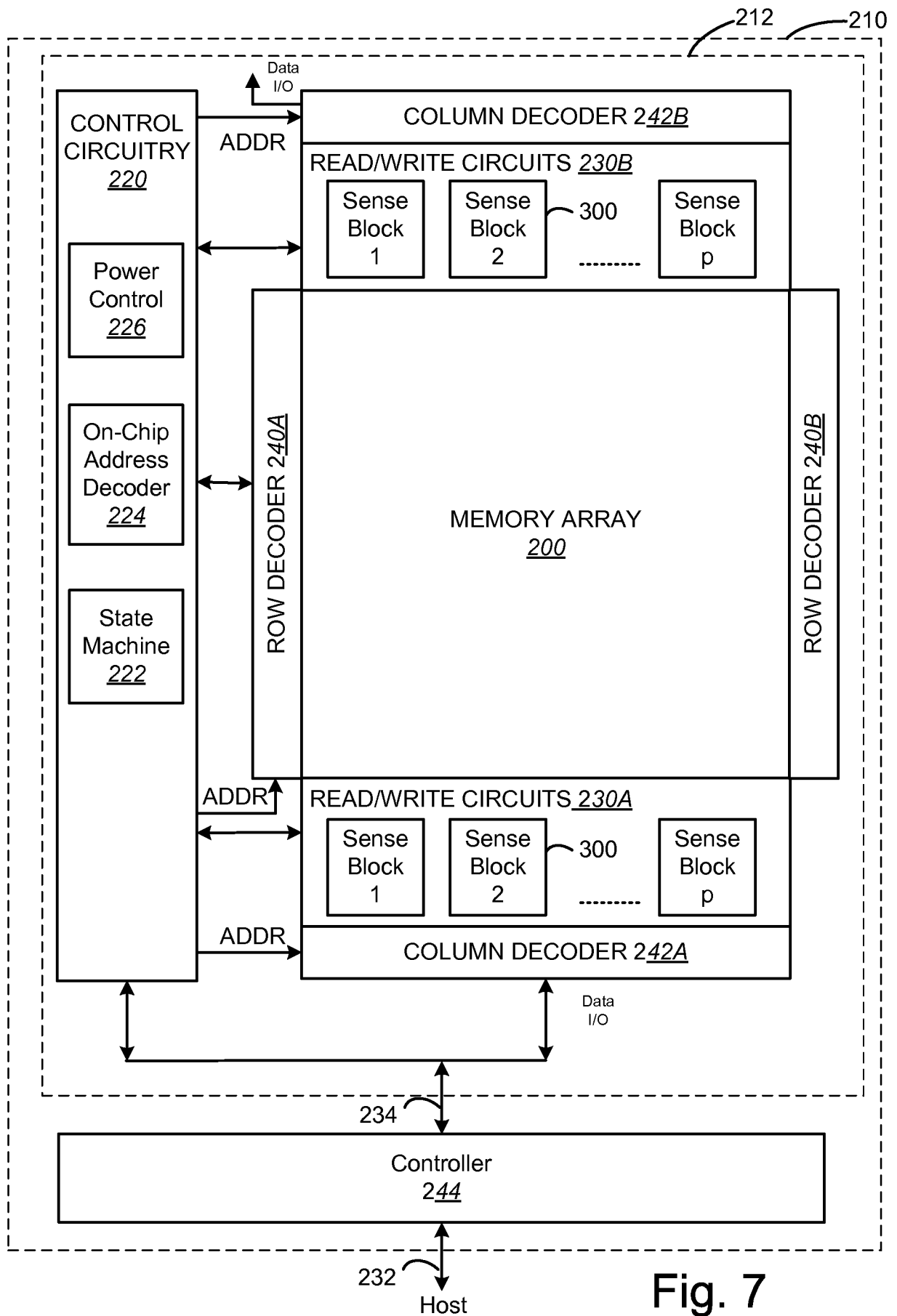
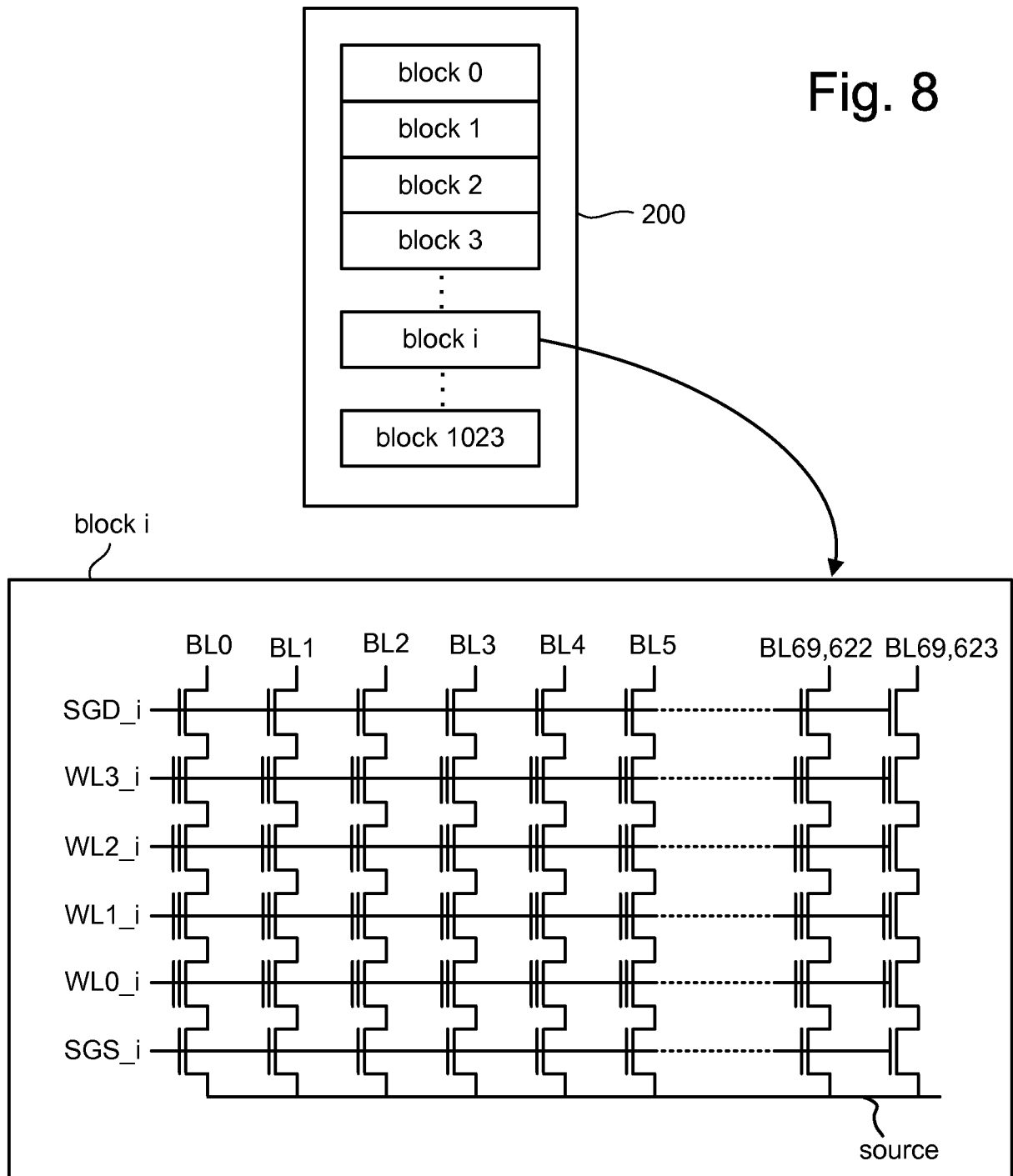


Fig. 7



INTERNATIONAL SEARCH REPORT

International application No

PCT/US2012/070833

A. CLASSIFICATION OF SUBJECT MATTER		
INV.	H01L21/28	H01L29/788 G11C16/04 H01L29/423 H01L29/66
ADD.	H01L27/115	
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED		
Minimum documentation searched (classification system followed by classification symbols)		
H01L G11C		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)		
EPO-Internal , WPI Data		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	US 2009/014774 AI (ONO MIZUKI [JP]) 15 January 2009 (2009-01-15) abstract; figures 1, 5-10,48 paragraphs [0033] - [0034] , [0061] - [0067] -----	1-9 , 11-16 10
X	US 2008/067577 AI (WANG MING-TSONG [TW] ET AL) 20 March 2008 (2008-03-20) paragraphs [0031] - [0038] ; figure 2 -----	1-16
X	US 2006/054943 AI (LI HONG-JYH [US] ET AL) 16 March 2006 (2006-03-16) paragraphs [0022] - [0031] , [0040] - [0049] ; figures 1,4-10 ----- -/- .	1,5, 11-14, 16
<input checked="" type="checkbox"/> Further documents are listed in the continuation of Box C. <input checked="" type="checkbox"/> See patent family annex.		
* Special categories of cited documents : "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search		Date of mailing of the international search report
20 March 2013		03/04/2013
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Mosi g, Karsten

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2012/070833

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2009/085094 A1 (LEE JANG-SIK [KR] ET AL) 2 April 2009 (2009-04-02)	1,13
A	paragraphs [0066] - [0069]; figure 1	2-12, 14-16
A	----- US 2007/132004 A1 (YASUDA NAOKI [JP] ET AL) 14 June 2007 (2007-06-14) abstract; figures 7-15 paragraphs [0080] - [0150]	1-16
A	----- US 2011/122698 A1 (IZUMIDA TAKASHI [JP] ET AL) 26 May 2011 (2011-05-26) abstract; figures 12-17 paragraphs [0101] - [0118] -----	1-16

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2012/070833

Patent document cited in search report		Publication date	Patent family member(s)			Publication date
US 2009014774	AI	15-01-2009	JP	2008311325	A	25-12-2008
			US	2009014774	AI	15-01-2009

US 2008067577	AI	20-03 -2008	TW	200814301	A	16-03 -2008
			US	2008067577	AI	20-03 -2008

US 2006054943	AI	16-03 -2006	NONE			

US 2009085094	AI	02-04 -2009	KR	20090032352	A	01-04 -2009
			US	2009085094	AI	02-04 -2009
			US	2010240208	AI	23-09 -2010

US 2007132004	AI	14-06 -2007	JP	4928890	B2	09-05 -2012
			JP	2007134681	A	31-05 -2007
			KR	20070041374	A	18-04 -2007
			TW	1321851	B	11-03 -2010
			US	2007132004	AI	14-06 -2007

US 2011122698	AI	26-05 -2011	JP	5039116	B2	03-10 -2012
			JP	2011114034	A	09-06 -2011
			KR	20110058630	A	01-06 -2011
			US	2011122698	AI	26-05 -2011
