US 20080111177A1

(19) United States

(12) Patent Application Publication (10) Pub. No.: US 2008/0111177 A1

Maayan et al. (43) Pub. Date: May 15, 2008

(54) NON-VOLATILE MEMORY CELL AND NON-VOLATILE MEMORY DEVICE USING SAID CELL

(76) Inventors: **Eduardo Maayan**, Kfar Saba (IL); **Boaz Eitan**, Hofit (IL)

Correspondence Address:
EMPK & Shiloh, LLP
116 JOHN ST,
SUITE 1201
NEW YORK, NY 10038 (US)

(21) Appl. No.: **12/003,704**

(22) Filed: **Dec. 31, 2007**

Related U.S. Application Data

(63) Continuation of application No. 11/979,187, filed on Oct. 31, 2007, which is a continuation of application No. 11/785,285, filed on Apr. 17, 2007, which is a continuation of application No. 11/497,078, filed on Aug. 1, 2006, which is a continuation of application No. 10/863,529, filed on Jun. 9, 2004, now Pat. No. 7,116,577, which is a continuation of application No. 10/122,078, filed on Apr. 15, 2002, now Pat. No. 6,649,972, which is a continuation of application No. 09/246,183, filed on Feb. 4, 1999, now Pat. No. 6,011,725, which is a continuation of application No. 08/905,286, filed on Aug. 1, 1997, now Pat. No. 6,768,165.

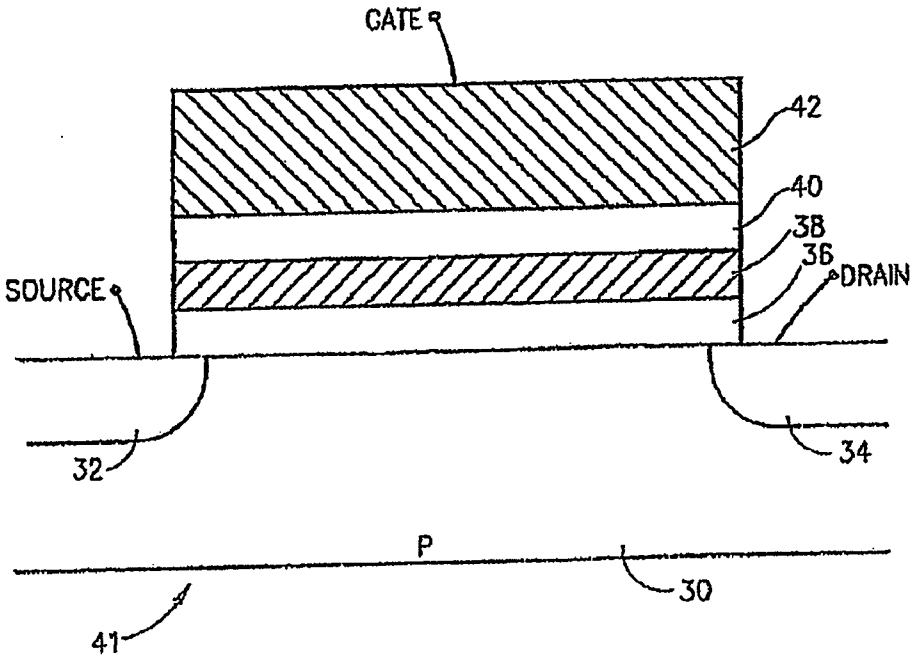
Publication Classification

(51) **Int. Cl.**
H01L 29/788 (2006.01)
H01L 21/336 (2006.01)

(52) **U.S. Cl.** **257/315**; 438/257; 257/E21; 257/E29

(57) **ABSTRACT**

A non-volatile electrically erasable programmable read only memory (EEPROM) capable of storing two bit of information having a non-conducting charge trapping dielectric, such as silicon nitride, sandwiched between two silicon dioxide layers acting as electrical insulators is disclosed. The invention includes a method of programming, reading and erasing the two bit EEPROM device. The non-conducting dielectric layer functions as an electrical charge trapping medium. A conducting gate layer is placed over the upper silicon dioxide layer. A left and a right bit are stored in physically different areas of the charge trapping layer, near left and right regions of the memory cell, respectively. Each bit of the memory device is programmed in the conventional manner, using hot electron programming, by applying programming voltages to the gate and to either the left or the right region while the other region is grounded. Hot electrons are accelerated sufficiently to be injected into the region of the trapping dielectric layer near where the programming voltages were applied to. The device, however, is read in the opposite direction from which it was written, meaning voltages are applied to the gate and to either the right or the left region while the other region is grounded. Two bits are able to be programmed and read due to a combination of relatively low gate voltages with reading in the reverse direction. This greatly reduces the potential across the trapped charge region. This permits much shorter programming times by amplifying the effect of the charge trapped in the localized trapping region associated with each of the bits. In addition, both bits of the memory cell can be individually erased by applying suitable erase voltages to the gate and either left or right regions so as to cause electrons to be removed from the corresponding charge trapping region of the nitride layer.



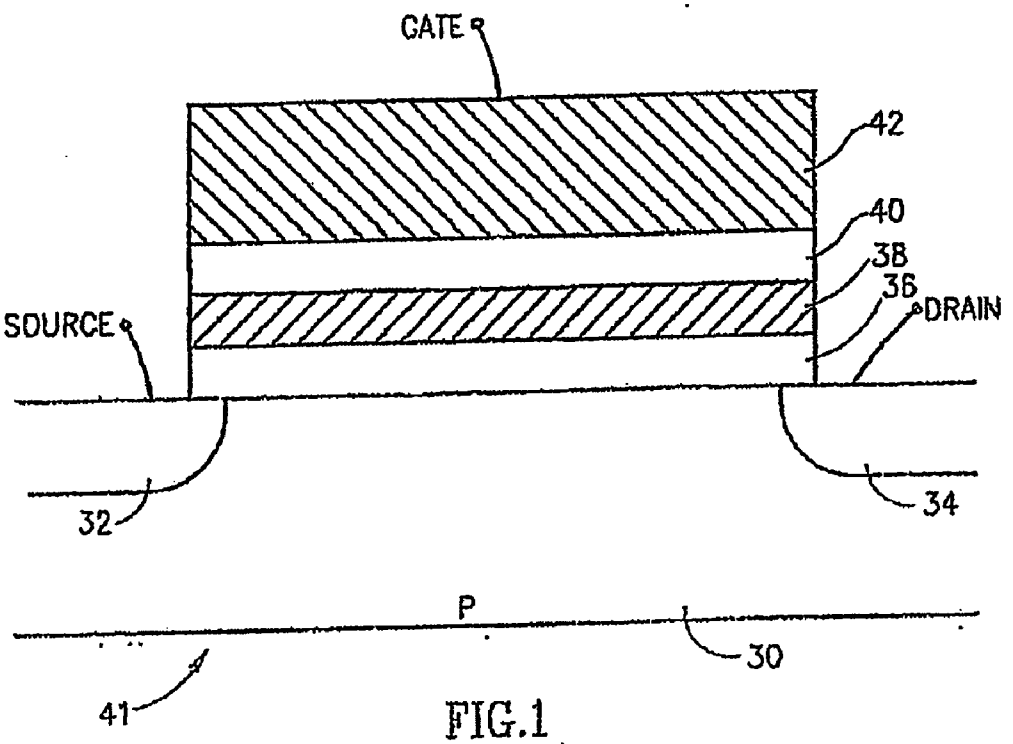

GATE
42
40
38
36
DRAIN
SOURCE
32
34
P
30
41

FIG.1

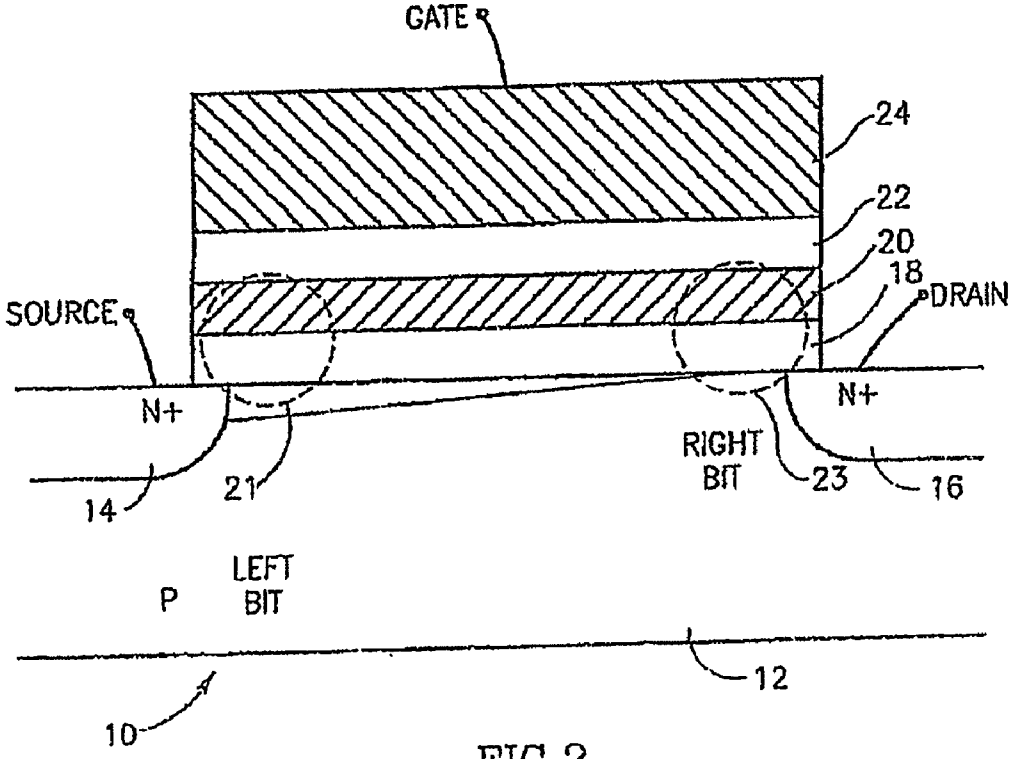

GATE
24
22
20
18
DRAIN
SOURCE
N+
N+
RIGHT
BIT
14
21
23
16
P
LEFT
BIT
12
10

FIG.2

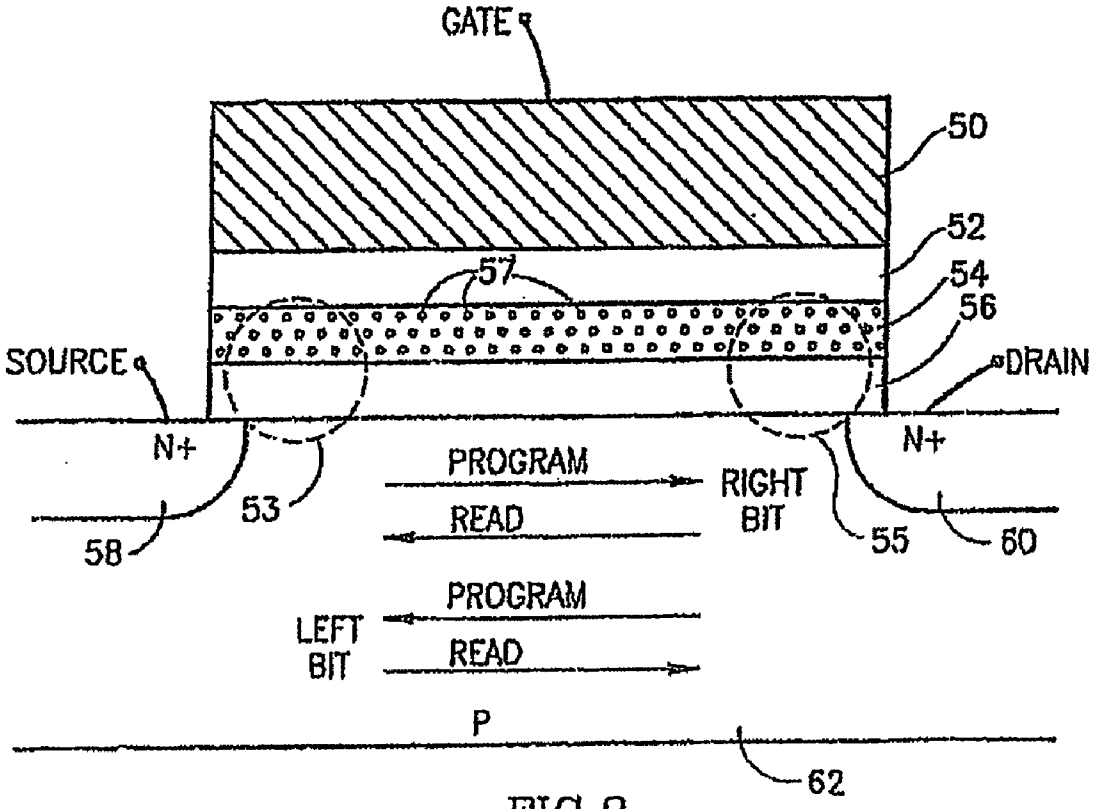

GATE
50
52
57
54
56
SOURCE
DRAIN
N+
N+
PROGRAM
RIGHT BIT
READ
53
58
55
60
PROGRAM
LEFT BIT
READ
P
62

FIG.3

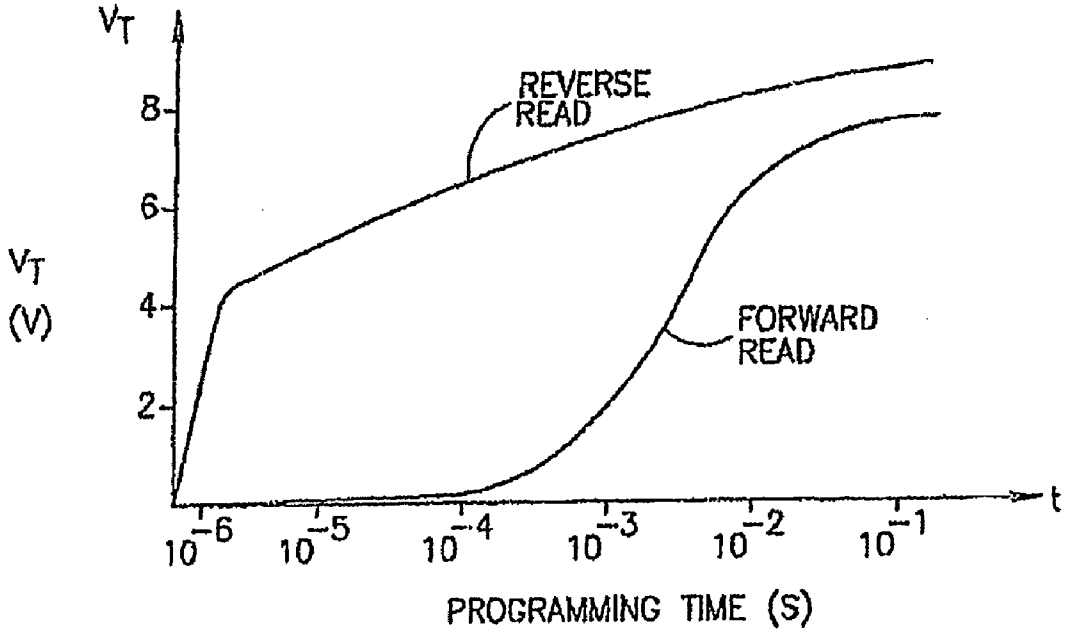

V_T
8
6
4
2
V_T (V)
REVERSE READ
FORWARD READ
t
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
PROGRAMMING TIME (S)

FIG.4

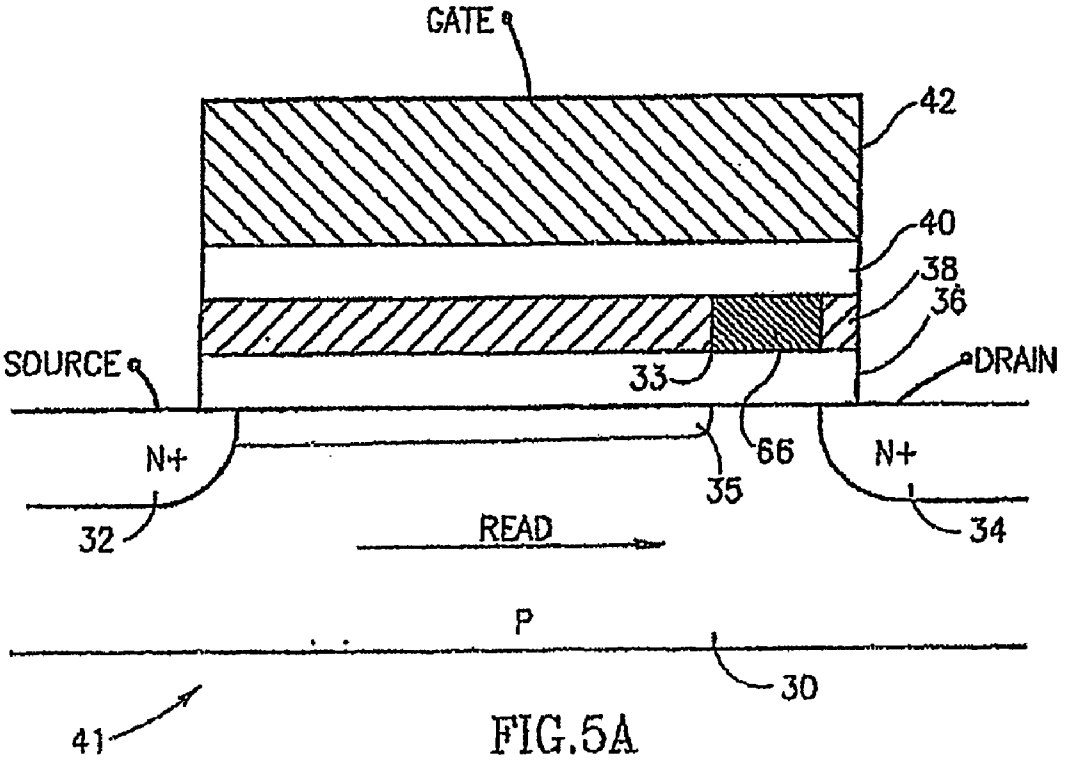

GATE
42
40
38
36
SOURCE
33
DRAIN
N+
66
N+
35
32
READ
34
P
30
41

FIG.5A

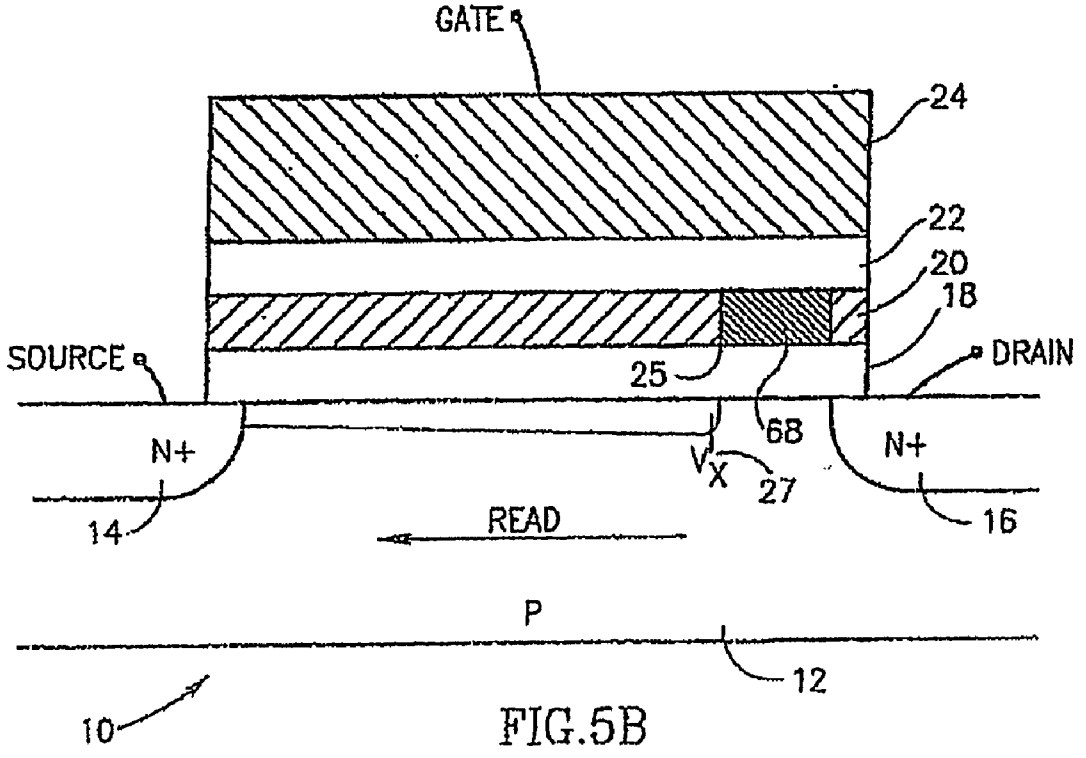

GATE
24
22
20
18
SOURCE
25
DRAIN
68
N+
N+
V_X
27
14
READ
16
P
12
10

FIG.5B

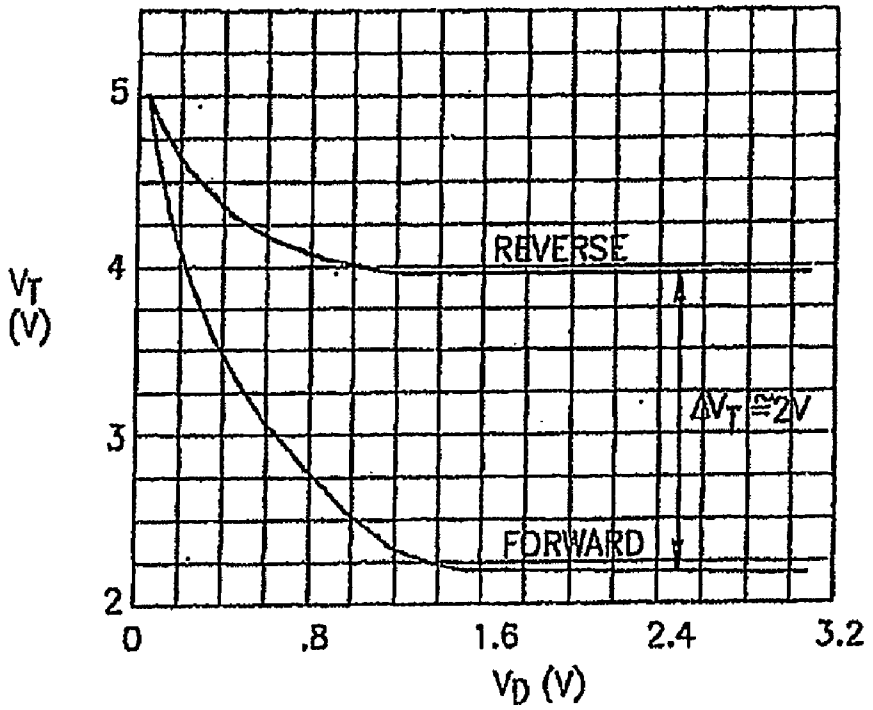

5
4
3
2
V_T (V)
REVERSE
$\Delta V_T \cong 2V$
FORWARD
0
.8
1.6
2.4
3.2
V_D (V)

FIG.6

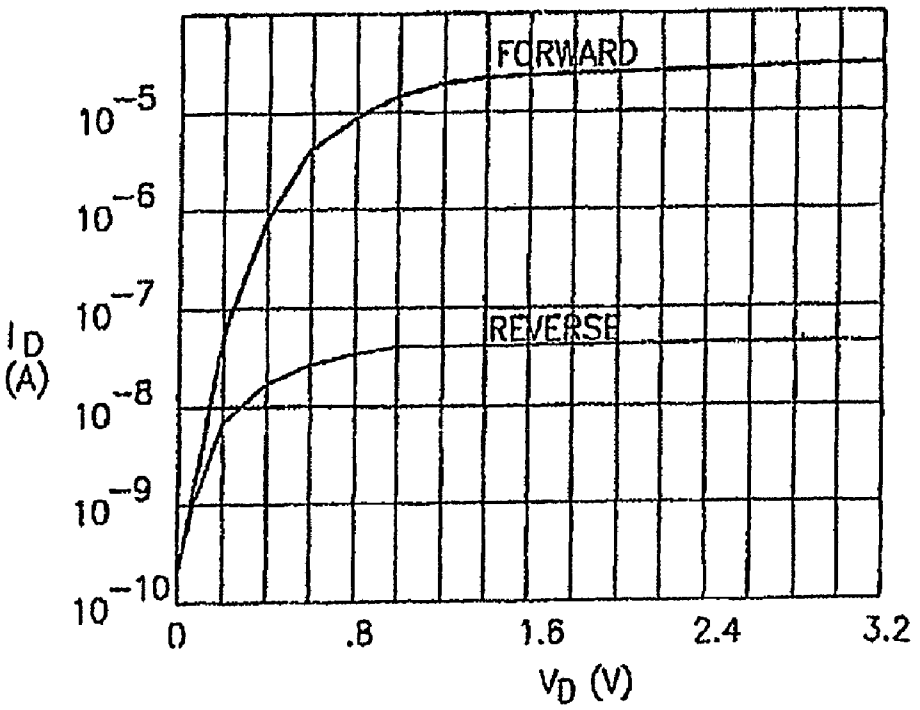

FORWARD
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
I_D (A)
REVERSE
0
.8
1.6
2.4
3.2
V_D (V)

FIG.7

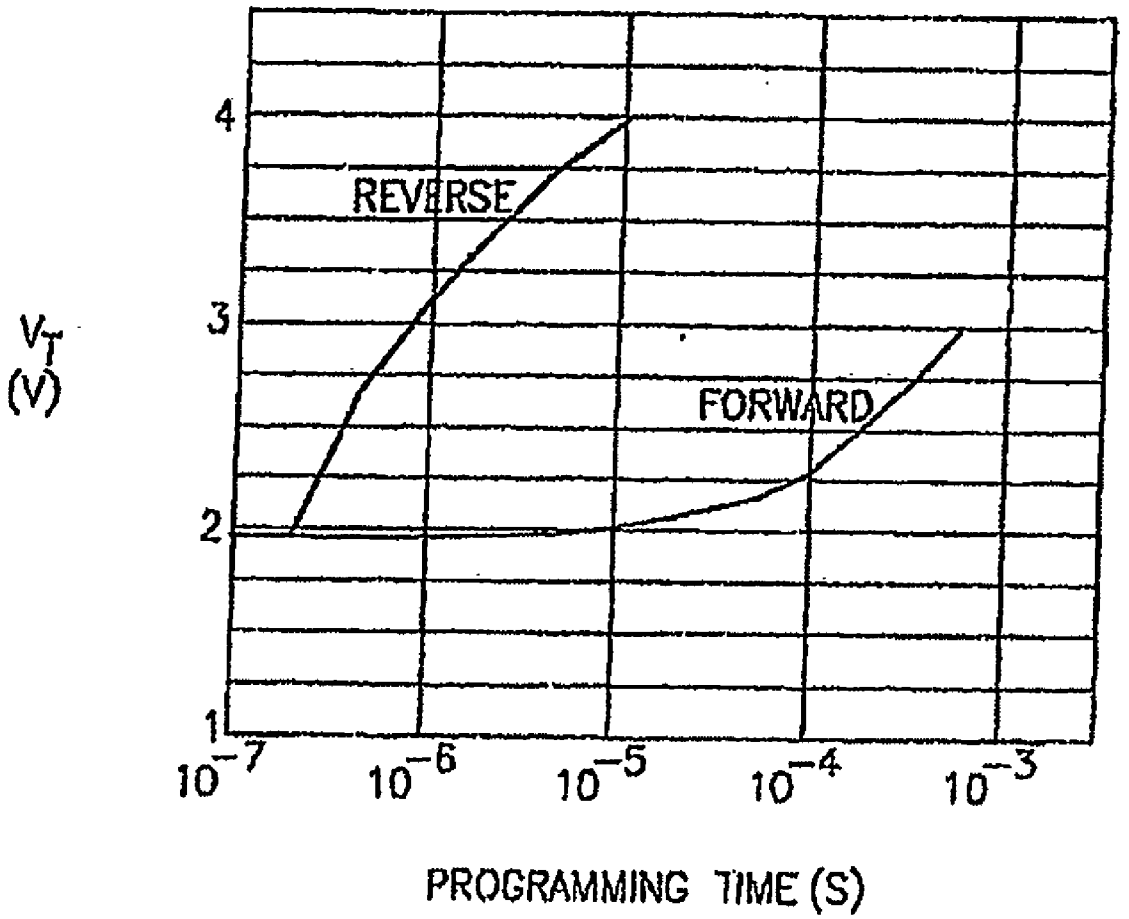

V_T (V)
4
3
2
1
REVERSE
FORWARD
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
PROGRAMMING TIME (S)

FIG.8

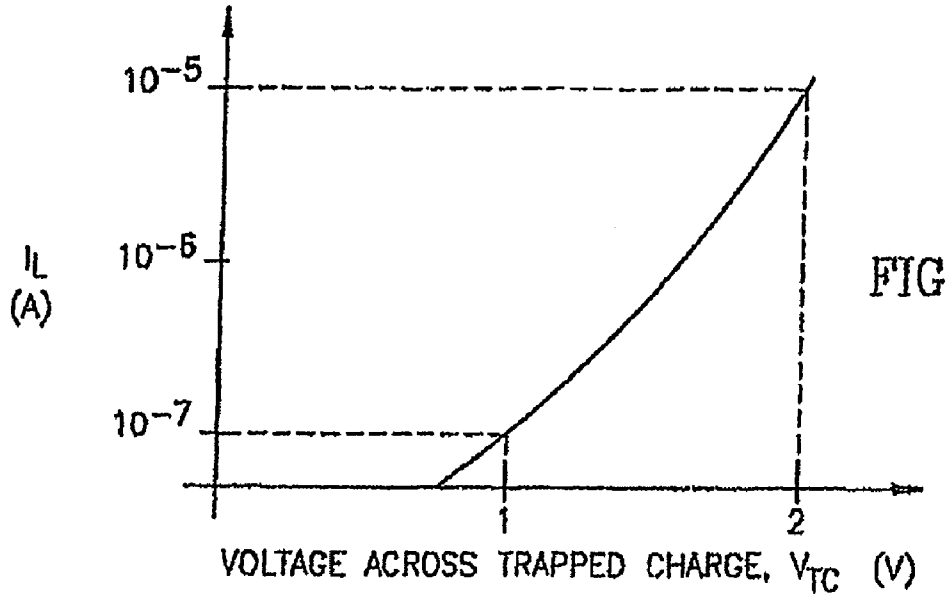

10^{-5}
10^{-6}
10^{-7}
I_L
(A)
1
2
VOLTAGE ACROSS TRAPPED CHARGE, V_{TC} (V)

FIG.9

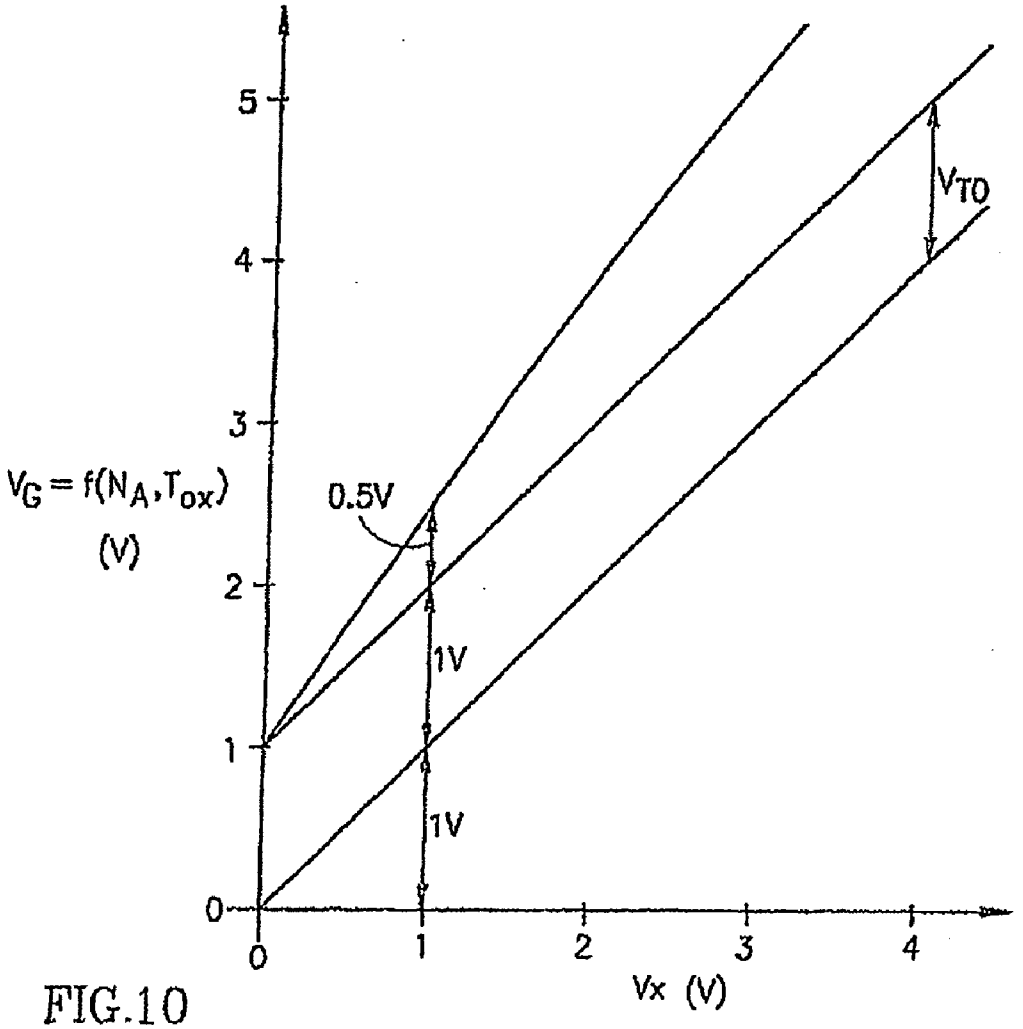

$V_G = f(N_A, T_{ox})$
(V)
0.5V
1V
1V
V_{TO}
0
1
2
3
4
5
Vx (V)

FIG.10

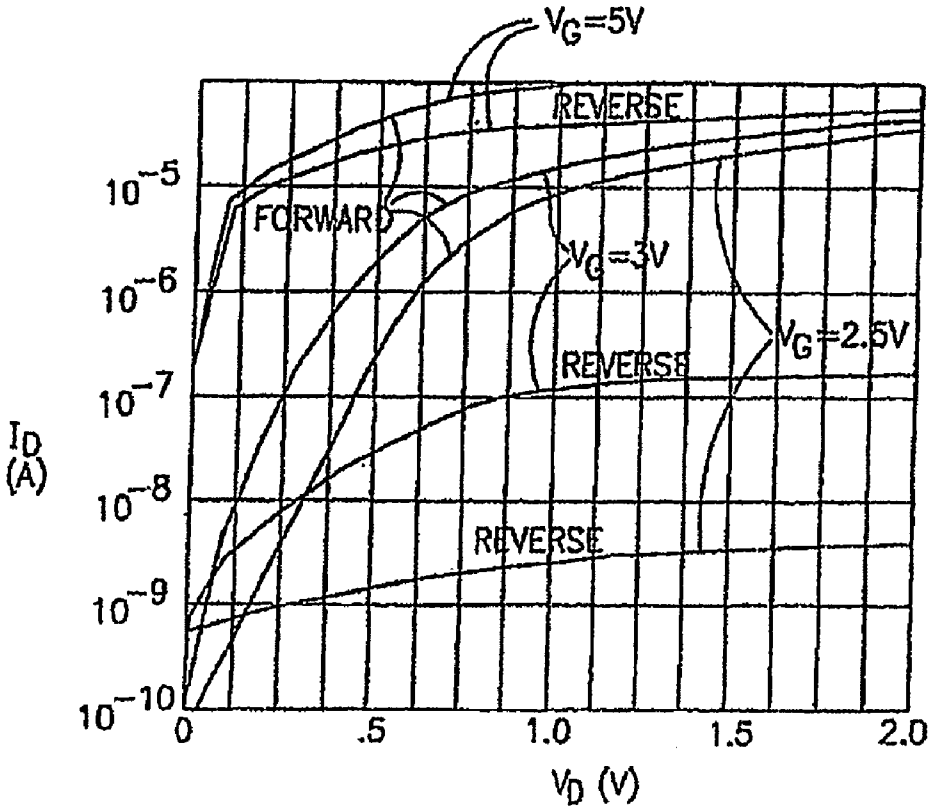

V_G=5V
REVERSE
FORWARD
V_G=3V
V_G=2.5V
REVERSE
REVERSE
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
I_D (A)
0
.5
1.0
1.5
2.0
V_D (V)

FIG.11

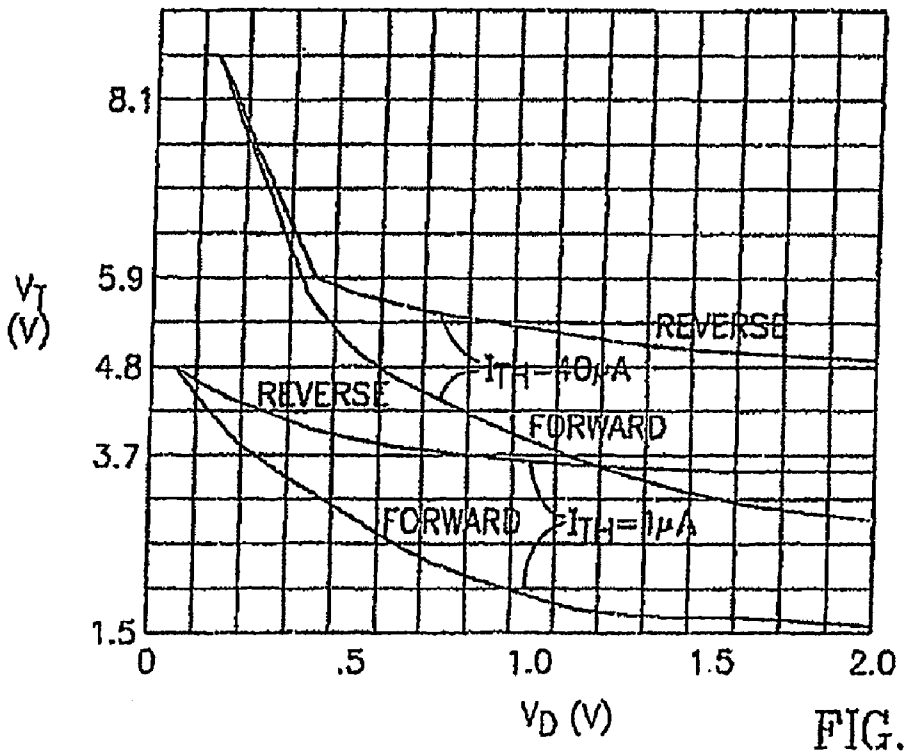

8.1
5.9
4.8
3.7
1.5
V_T (V)
REVERSE
I_{TH}=40μA
REVERSE
FORWARD
FORWARD
I_{TH}=1μA
0
.5
1.0
1.5
2.0
V_D (V)

FIG.12

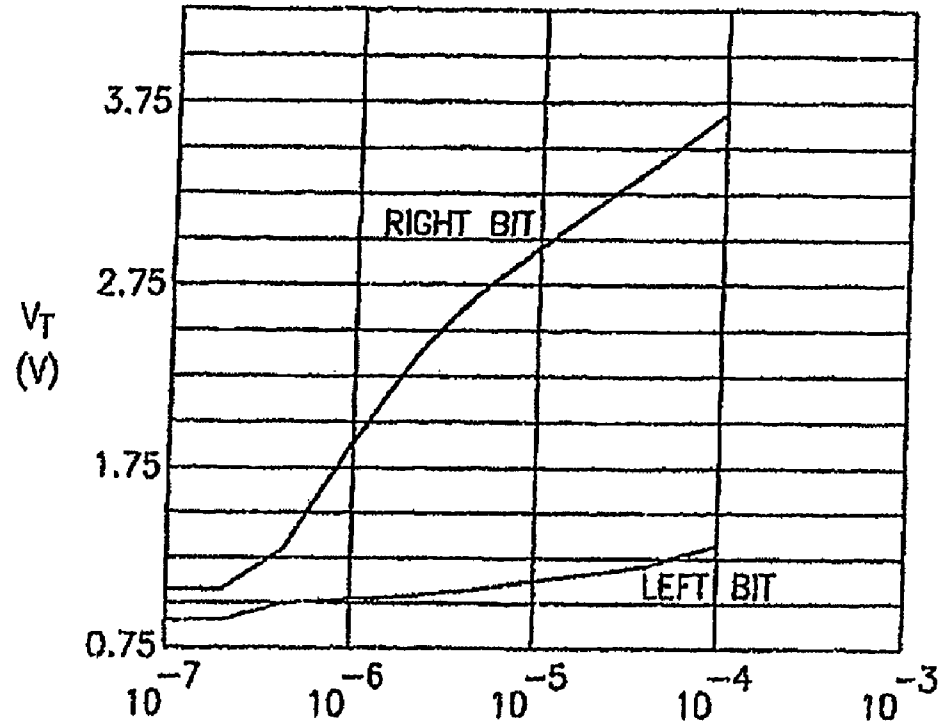

3.75
2.75
1.75
0.75
V_T
(V)
RIGHT BIT
LEFT BIT
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
PROGRAMMING TIME (S)

FIG.13

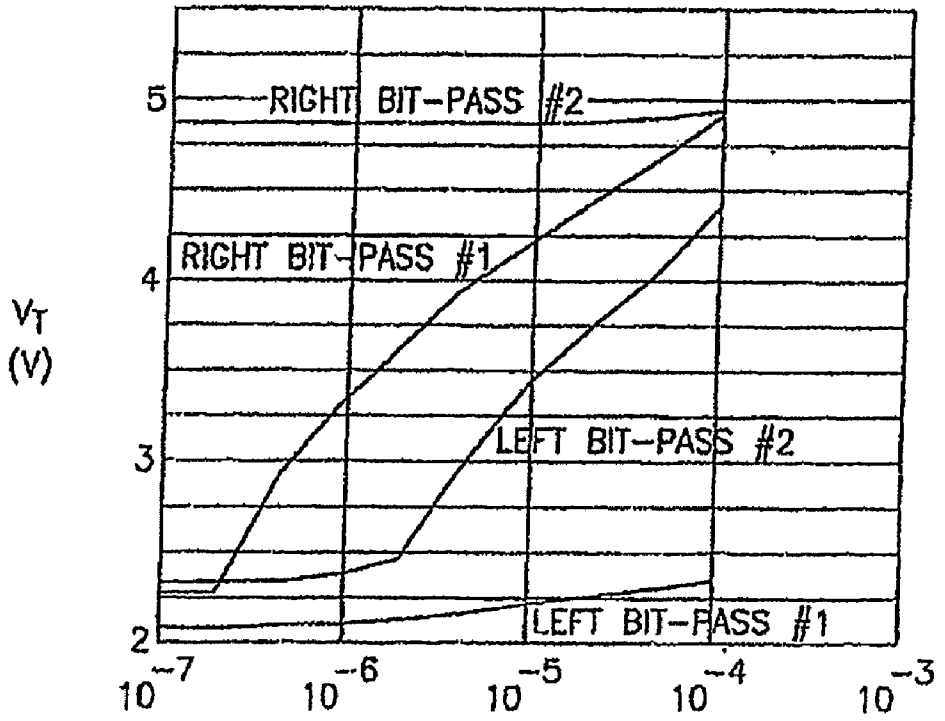

5
4
3
2
V_T
(V)
RIGHT BIT-PASS #2
RIGHT BIT-PASS #1
LEFT BIT-PASS #2
LEFT BIT-PASS #1
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
PROGRAMMING TIME (S)

FIG.14

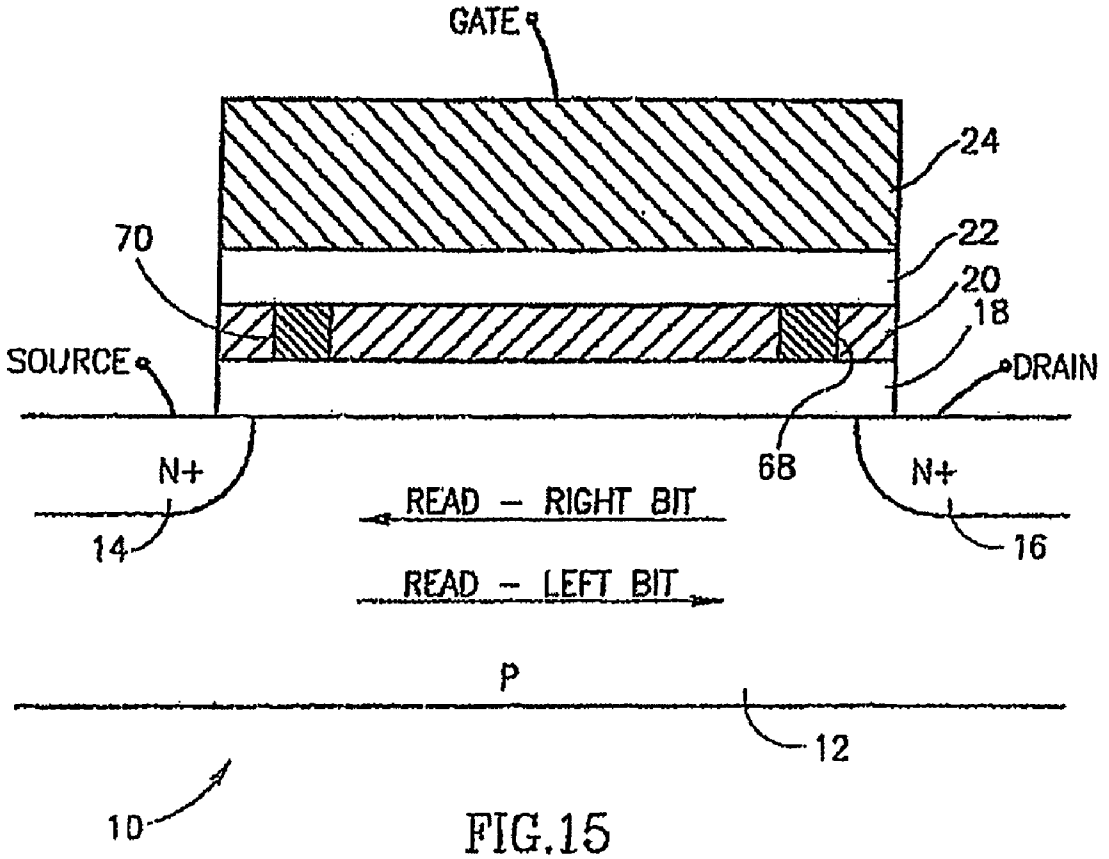

GATE
24
70
22
20
18
SOURCE
DRAIN
68
N+
N+
READ – RIGHT BIT
READ – LEFT BIT
14
16
P
12
10

FIG.15

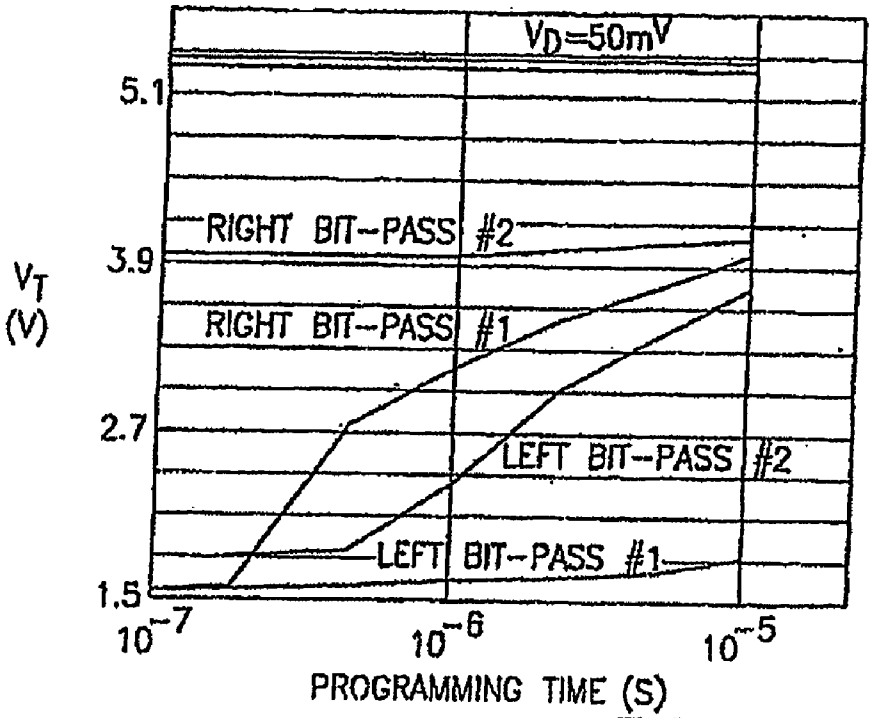

V_D=50mV
5.1
3.9
2.7
1.5
V_T (V)
RIGHT BIT-PASS #2
RIGHT BIT-PASS #1
LEFT BIT-PASS #2
LEFT BIT-PASS #1
10^{-7}
10^{-6}
10^{-5}
PROGRAMMING TIME (S)

FIG.16

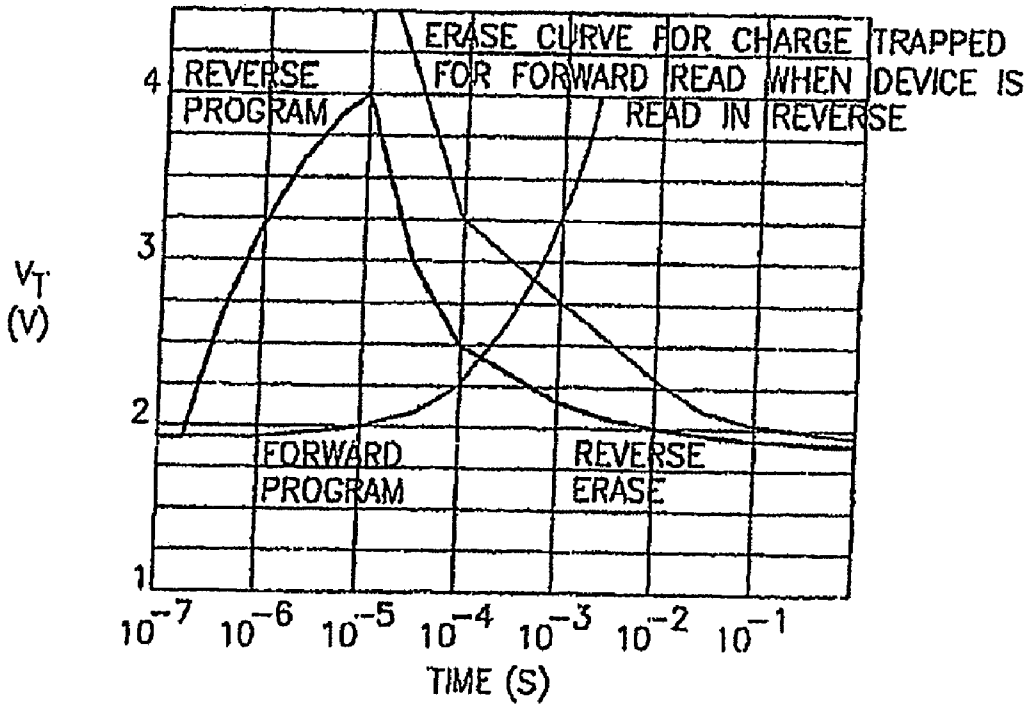

ERASE CURVE FOR CHARGE TRAPPED FOR FORWARD READ WHEN DEVICE IS READ IN REVERSE
REVERSE PROGRAM
4
3
2
1
V_T (V)
FORWARD PROGRAM
REVERSE ERASE
10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}
TIME (S)

FIG.17

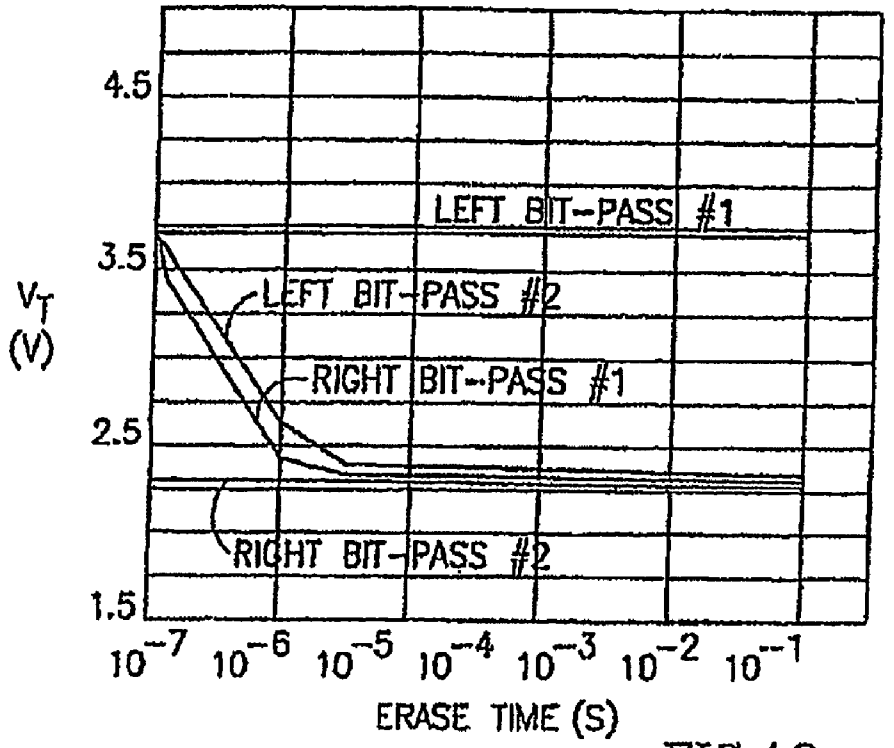

4.5
3.5
2.5
1.5
V_T (V)
LEFT BIT-PASS #1
LEFT BIT-PASS #2
RIGHT BIT-PASS #1
RIGHT BIT-PASS #2
10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}
ERASE TIME (S)

FIG.18

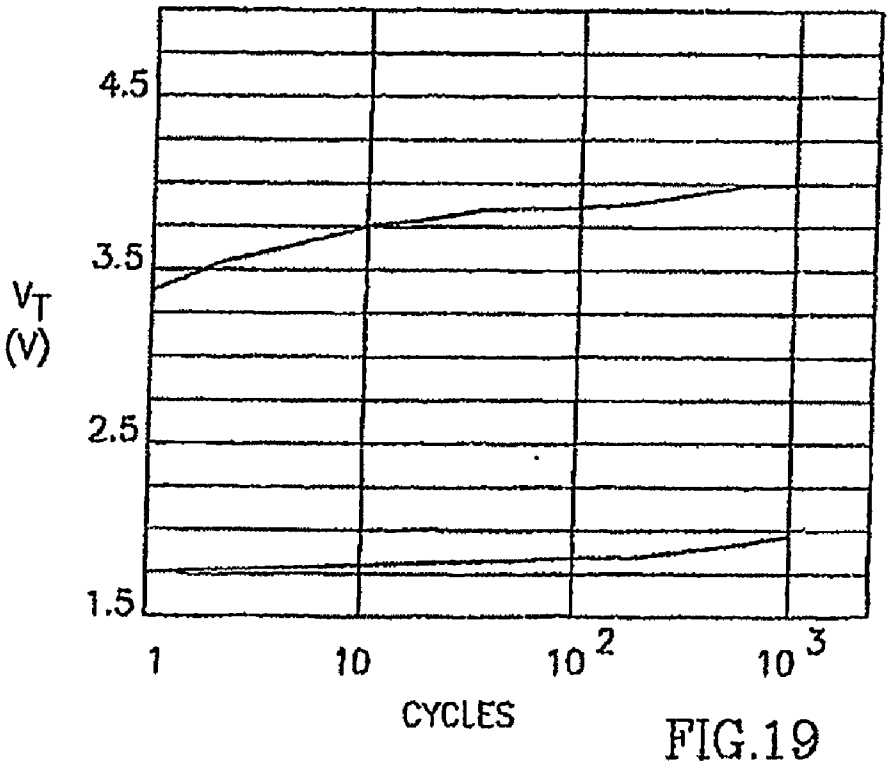

4.5
3.5
2.5
1.5
V_T (V)
1
10
10^2
10^3
CYCLES

FIG.19

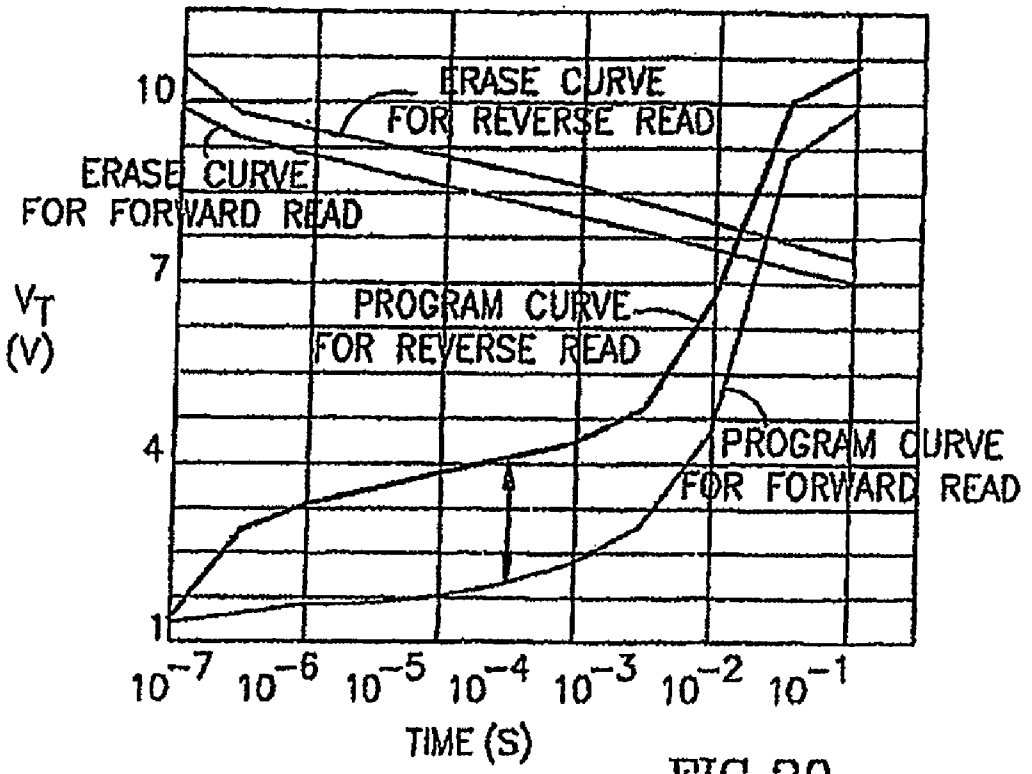

10
7
4
1
ERASE CURVE
FOR REVERSE READ
ERASE CURVE
FOR FORWARD READ
PROGRAM CURVE
FOR REVERSE READ
PROGRAM CURVE
FOR FORWARD READ
V_T
(V)
10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}
TIME (S)

FIG.20

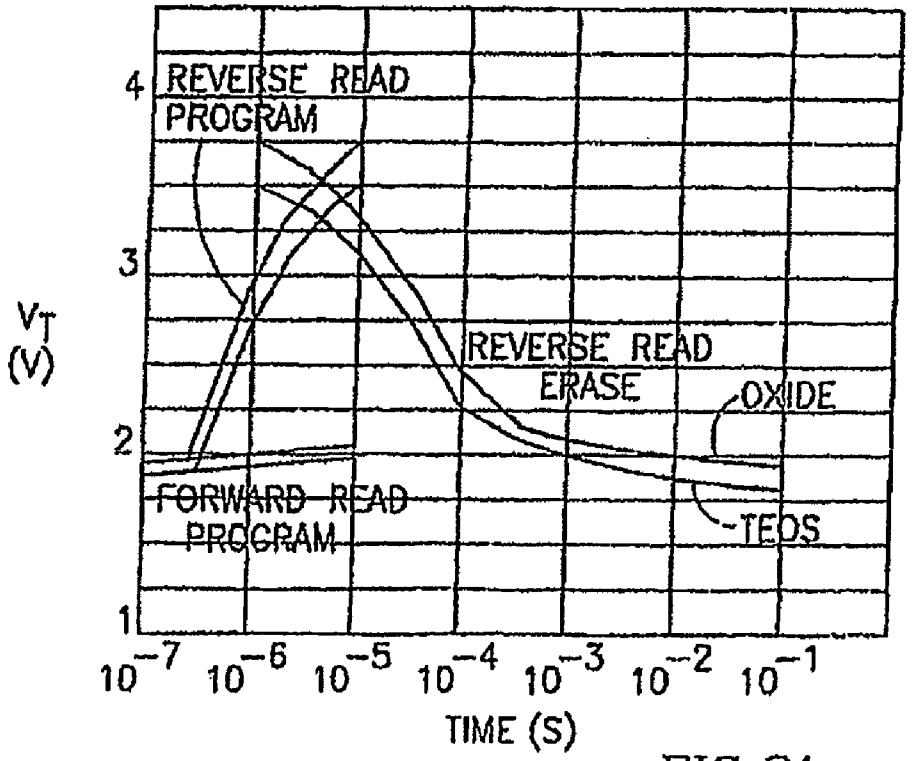

4
3
2
1
REVERSE READ
PROGRAM
REVERSE READ
ERASE
OXIDE
FORWARD READ
PROGRAM
TEOS
V_T
(V)
10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}
TIME (S)

FIG.21

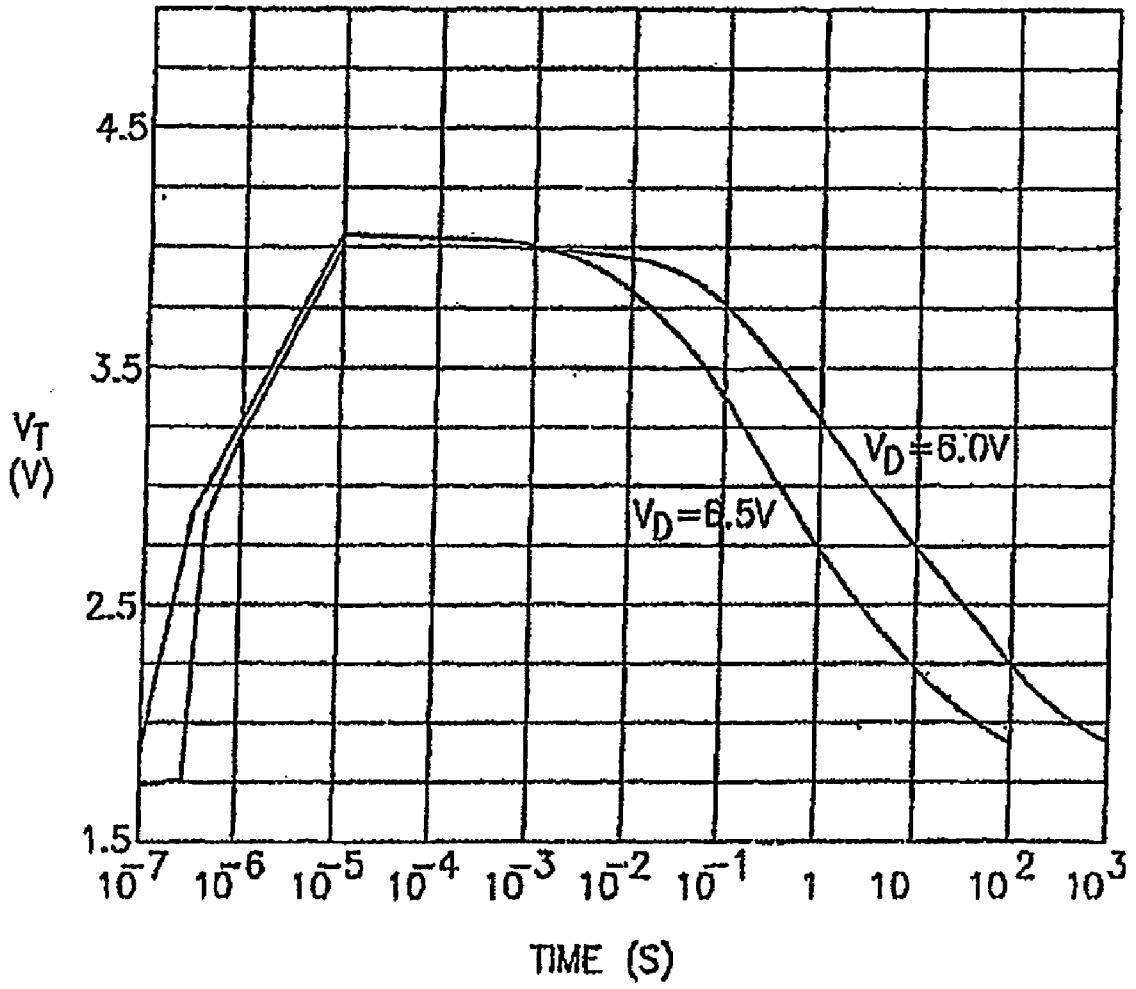

4.5
3.5
2.5
1.5
V_T (V)
V_D=6.0V
V_D=6.5V
10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 1 10 10^{2} 10^{3}
TIME (S)

FIG.22

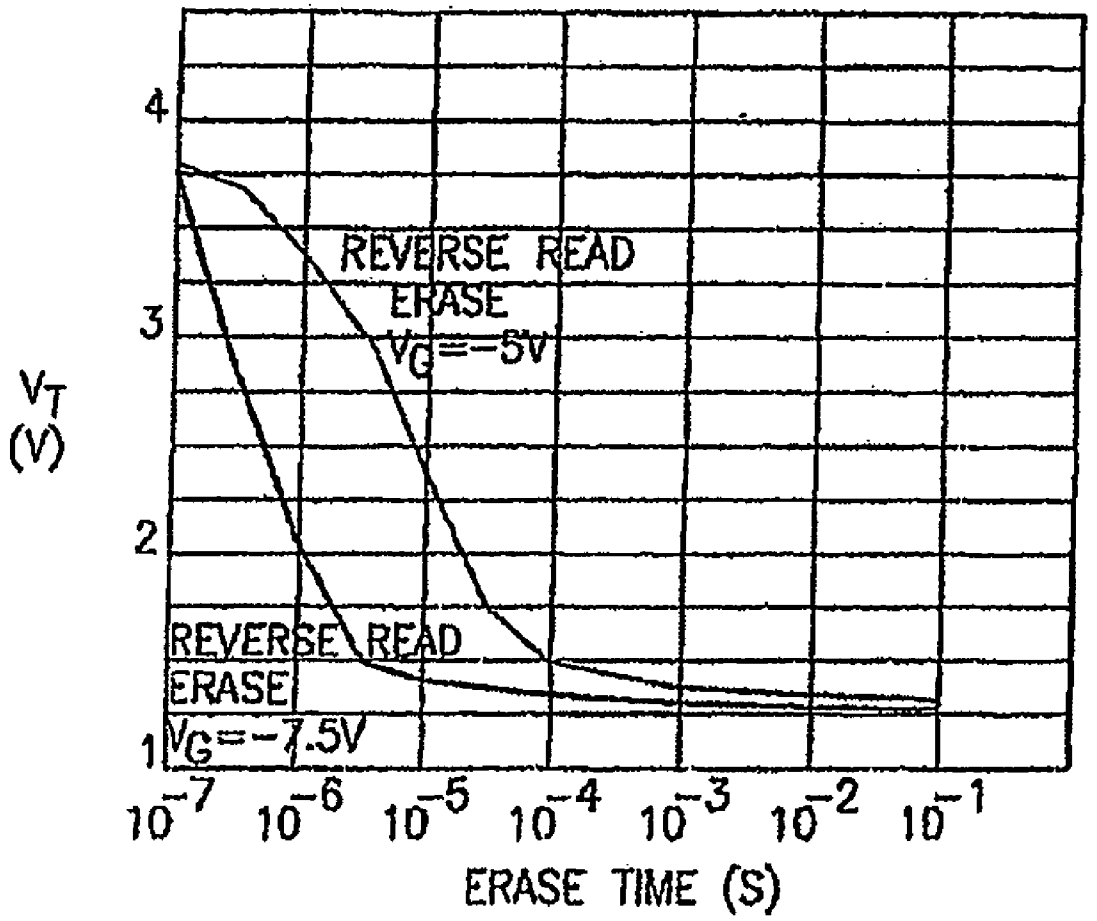

4
3
2
1
V_T (V)
REVERSE READ
ERASE
V_G=−5V
REVERSE READ
ERASE
V_G=−7.5V
10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}
ERASE TIME (S)

FIG.23

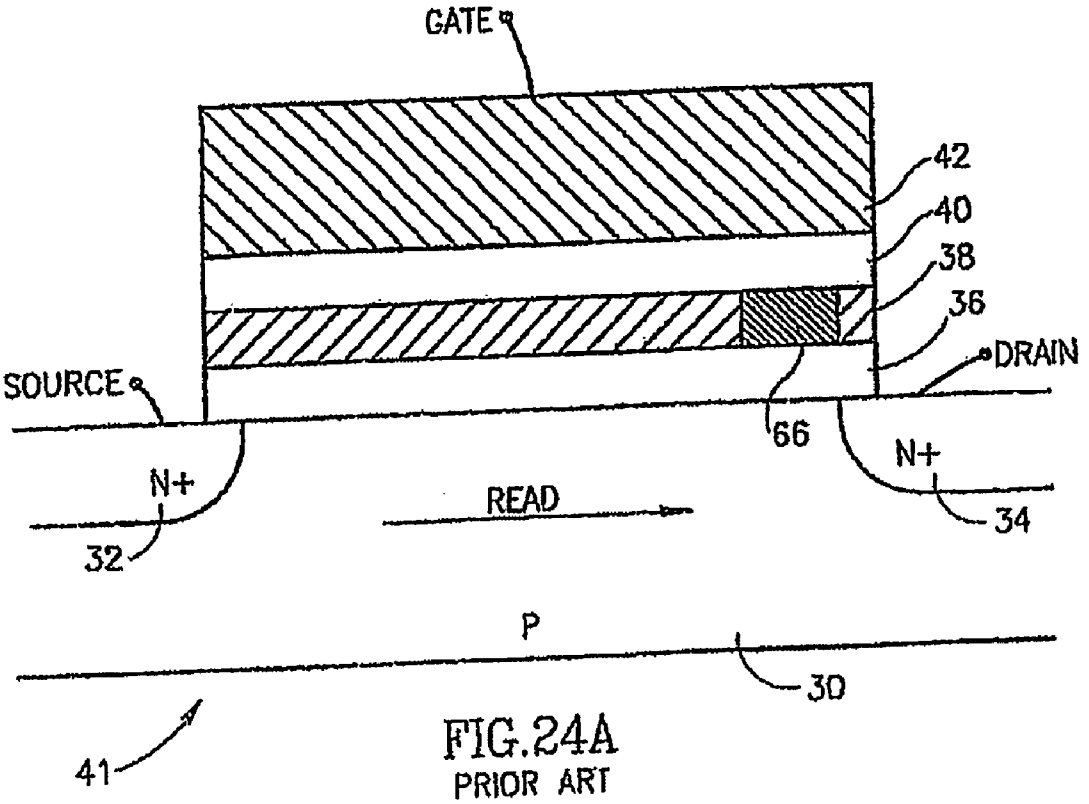


FIG.24A
PRIOR ART

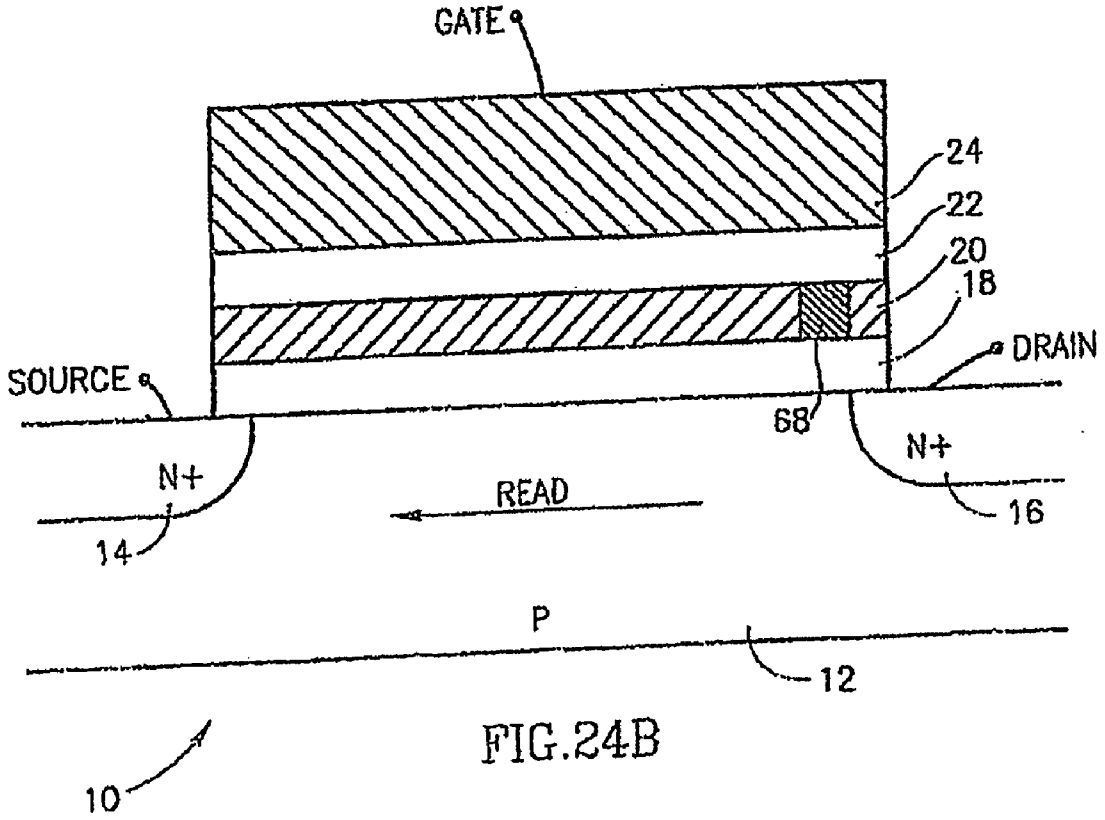


FIG.24B

NON-VOLATILE MEMORY CELL AND NON-VOLATILE MEMORY DEVICE USING SAID CELL

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application is a continuation application of U.S. patent application Ser. No. 11/979,187, filed Oct. 31, 2007 which is a continuation application of U.S. patent application Ser. No. 11/785,285, filed Apr. 17, 2007, which is a continuation of U.S. patent application Ser. No. 11/497,078, filed Aug. 1, 2006, which is a continuation of U.S. patent application Ser. No. 10/863,529, filed Jun. 9, 2004 which is a continuation of U.S. patent application Ser. No. 10/122,078, filed Apr. 15, 2002, which is a continuation of U.S. patent application Ser. No. 09/246,183 filed Feb. 4, 1999, which is a continuation of U.S. patent application Ser. No. 08/905,286, filed Aug. 1, 1997 all of which are hereby incorporated by reference.

FIELD

[0002] The present invention relates generally to semiconductor memory devices and more particularly to multi-bit flash electrically erasable programmable read only memory (EEPROM) cells that trap charge within a trapping dielectric material gate.

BACKGROUND

[0003] Memory devices for non-volatile storage of information are currently in widespread use today, being used in a myriad of applications. A few examples of non-volatile semiconductor memory include read only memory (ROM), programmable read only memory (PROM), erasable programmable read only memory (EPROM), electrically erasable programmable read only memory (EEPROM) and flash EEPROM. Further description of non-volatile memory (NVM) and related semiconductor and microelectronic device technologies may be found at "Microchip Fabrication", 1997 by Peter Van Zant (McGraw-Hill), incorporated by reference herein in its entirety.

[0004] Semiconductor ROM devices, however, suffer from the disadvantage of not being electrically programmable memory devices. The programming of a ROM occurs during one of the steps of manufacture using special masks containing the data to be stored. Thus, the entire contents of a ROM must be determined before manufacture. In addition, because ROM devices are programmed during manufacture, the time delay before the finished product is available could be six weeks or more. The advantage, however, of using ROM for data storage is the low cost per device. However, the penalty is the inability to change the data once the masks are committed to. If mistakes in the data programming are found they are typically very costly to correct. Any inventory that exists having incorrect data programming is instantly obsolete and probably cannot be used. In addition, extensive time delays are incurred because new masks must first be generated from scratch and the entire manufacturing process repeated. Also, the cost savings in the use of ROM memories only exist if large quantities of the ROM are produced.

[0005] Moving to EPROM semiconductor devices eliminates the necessity of mask programming the data but the complexity of the process increases drastically. In addition, the die size is larger due to the addition of programming circuitry and there are more processing and testing steps involved in the manufacture of these types of memory devices. An advantage of EPROMs is that they are electrically programmed, but for erasing, EPROMs require exposure to ultraviolet (UV) light. These devices are constructed with windows transparent to UV light to allow the die to be exposed for erasing, which must be performed before the device can be programmed. A major drawback to these devices is that they lack the ability to be electrically erased. In many circuit designs it is desirable to have a non-volatile memory device that can be erased and reprogrammed in-circuit, without the need to remove the device for erasing and reprogramming.

[0006] Semiconductor EEPROM devices also involve more complex processing and testing procedures than ROM, but have the advantage of electrical programming and erasing. Using EEPROM devices in circuitry permits in-circuit erasing and reprogramming of the device, a feat not possible with conventional EPROM memory. Flash EEPROMs are similar to EEPROMs in that memory cells can be programmed (i.e., written) and erased electrically but with the additional ability of erasing all memory cells at once, hence the term flash EEPROM. The disadvantage of flash EEPROM is that it is very difficult and expensive to manufacture and produce.

[0007] The widespread use of EEPROM semiconductor memory has prompted much research focusing on constructing better memory cells. Active areas of research have focused on developing a memory cell that has improved performance characteristics such as shorter programming times, utilizing lower voltages for programming and reading, longer data retention times, shorter erase times and smaller physical dimensions. One such area of research involves a memory cell that has an insulated gate. The following prior art reference is related to this area.

[0008] U.S. Pat. No. 4,173,766, issued to Hayes, incorporated herein by reference in its entirety, teaches a metal nitride oxide semiconductor (MNOS) constructed with an insulated gate having a bottom silicon dioxide layer and a top nitride layer.

[0009] U.S. Pat. No. 5,168,334, issued to Mitchell et al., incorporated herein by reference in its entirety, teaches a single transistor EEPROM memory cell.

[0010] A single transistor ONO EEPROM device is disclosed in the technical article entitled "A True Single-Transistor Oxide-Nitride-Oxide EEPROM Device," T. Y. Chan, K. K. Young and Chenming Hu, IEEE Electron Device Letters, March 1987 incorporated herein by reference in its entirety.

[0011] Multi-bit transistors are known in the art. Most multi-bit transistors utilize multi-level thresholds to store more than one bit with each threshold level representing a different state. A memory cell having four threshold levels can store two bits.

[0012] Achieving multiple thresholds in FLASH and EEPROM requires an initial erase cycle to bring all the memory cells below a certain threshold. Then, utilizing a methodical programming scheme, the threshold of each cell is increased until the desired threshold is reached. A disadvantage with this technique is that the programming process requires constant feedback, which causes multi-level programming to be slow.

[0013] In addition, using this technique causes the window of operation to decrease meaning the margins for each state are reduced. This translates to a lower probability of making good dies and a reduction in the level of quality achieved. If it is not desired to sacrifice any margins while increasing the reliability of the cell, then the window of operation must be increased by a factor of two. This means operating at much higher voltages, which is not desirable because it lowers the reliability and increases the disturbances between the cells. Due to the complexity of the multi-threshold technique, it is used mainly in applications where missing bits can be tolerated such as in audio applications.

[0014] Another problem with this technique is that the threshold windows for each state may change over time reducing the reliability. It must be guaranteed that using the same word line or bit line to program other cells will not interfere with or disturb the data in cells already programmed. In addition, the programming time itself increases to accommodate the multitude of different programming thresholds. Thus, the shifting of threshold windows for each state over time reduces the window of operation and consequently increases the sensitivity to disturbs.

[0015] The reduced margins for the threshold windows for the multiple states results in reduced yield. Further, in order to maintain quality and threshold margins, higher voltages are required. This implies higher electric fields in the channel, which contributes to lower reliability of the memory cell.

[0016] In order to construct a multi-bit memory cell, the cell must have four distinct levels that can be programmed. In the case of two levels, i.e., conventional single bit cell, the threshold voltage programmed into a cell for a '0' bit only has to be greater than the maximum gate voltage, thus making sure the cell does not conduct when it is turned on during reading. It is sufficient that the cell conducts at least a certain amount of current to distinguish between the programmed and unprogrammed states. The current through a transistor can be described by the following equation.

$$I = \frac{1}{Leff} K(V_G - V_T)$$

L_{eff} is the effective channel length, K is a constant, V_G is the gate voltage and V_T is the threshold voltage. However, in the multi-bit case, different thresholds must be clearly distinguishable which translates into sensing different read currents and slower read speed. Further, for two bits, four current levels must be sensed, each threshold having a statistical distribution because the thresholds cannot be set perfectly. In addition, there will be a statistical distribution for the effective channel length which will further widen the distribution of the read currents for each threshold level.

[0017] The gate voltage also affects the distribution of read currents. For the same set of threshold levels, varying the gate voltage directly results in a variation of the ratio between the read currents. Therefore the gate voltage must be kept very stable. In addition, since there are multiple levels of current, sensing becomes more complex than in the two level, i.e., single bit, cell.

[0018] The following art references are related to multi-bit semiconductor memory cells.

[0019] U.S. Pat. No. 5,021,999, issued to Kohda et al., incorporated herein by reference in its entirety, teaches a non-volatile memory cell using an MOS transistor having a floating gate with two electrically separated segmented floating gates.

[0020] U.S. Pat. No. 5,214,303, issued to Aoki, incorporated herein by reference in its entirety, teaches a two bit transistor which comprises a semiconductor substrate.

[0021] U.S. Pat. No. 5,394,355, issued to Uramoto et al., incorporated herein by reference in its entirety, teaches a ROM memory having a plurality of reference potential transmission lines.

[0022] U.S. Pat. No. 5,414,693, issued to Ma et al., incorporated herein by reference in its entirety, teaches a two bit split gate flash EEPROM memory cell structure that uses one select gate transistor and two floating gate transistors.

[0023] U.S. Pat. No. 5,434,825, issued to Harari, incorporated herein by reference in its entirety, teaches a multi-bit EPROM and EEPROM memory cell which is partitioned into three or more ranges of programming charge.

SUMMARY OF THE INVENTION

[0024] The present invention discloses an apparatus for and method of programming and erasing a flash electrically erasable programmable read only memory (EEPROM). The flash EEPROM memory cell is constructed having a charge trapping non-conducting dielectric layer. The non-conducting dielectric layer functions as an electrical charge trapping medium.

[0025] An aspect of the memory device is that bits are programmed in the conventional manner, using hot electron programming. For example, the bit is programmed conventionally by applying programming voltages to the gate and the drain while the source is grounded. Hot electrons are accelerated sufficiently to be injected into a region of the trapping dielectric layer near the drain.

[0026] Utilizing relatively low gate voltages, the potential drop across the portion of the channel beneath the trapped charge region is reduced. A relatively small programming region or charge trapping region is possible due to the lower channel potential drop under the charge trapping region. This permits faster programming times because the effect of the charge trapped in the localized trapping region is amplified. Programming times are reduced while the delta in threshold voltage between the programmed versus un programmed states remains the same.

[0027] Another benefit is that the memory cell is enhanced. Bits of the memory cell can be erased by applying suitable erase voltages to the gate and the drain for the bit and to the gate and the source for the bit so as to cause electrons to be removed from the charge trapping region of the nitride layer. Electrons move from the nitride through the bottom oxide layer to the drain or the source for the bits, respectively.

[0028] The erase mechanism is enhanced when the charge trapping region is made as narrow as possible.

[0029] Utilizing a thinner silicon nitride charge trapping layer helps to confine the charge trapping region to a laterally narrower region near the drain. Further, thinner top and bottom oxide sandwiching the nitride layer helps in retention of the trapped charge.

[0030] In addition, bottom and top oxide thickness can be scaled due to the deep trapping levels that function to increase the potential barrier for direct tunneling. Since the electron trapping levels are deep, thinner bottom and top oxides can be used without compromising charge retention.

[0031] Another benefit of localized charge trapping is that during erase, the region of the nitride away from the drain does not experience deep depletion since the erase occurs near the drain only. The final threshold of the cell after erasing is self limited by the device structure itself. This is in direct contrast to conventional single transistor floating gate flash memory cells which are plagued with deep depletion problems. To overcome these problems, manufacturers include complex circuitry to control the erase process in order to prevent or recover from deep depletion.

[0032] Consideration of the memory cell of the present invention is described hereinbelow. Programming the device to a low V_T, by clamping the word line voltage V_{WL}, further enhances the margin for each bit. The margin is defined as the parameters that will program one bit without affecting others.

[0033] The memory device also exhibits little or no disturb during programming. This is because during programming the drain voltage is only applied to the junction adjacent to the region where charge trapping is to occur.

BRIEF DESCRIPTION OF THE DRAWINGS

[0034] The invention is herein described, by way of example only, with reference to the accompanying drawings, wherein:

[0035] FIG. **1** illustrates a sectional view of a single bit flash EEPROM cell utilizing Oxide-Nitride-Oxide (ONO);

[0036] FIG. **2** illustrates a sectional view of a two bit flash EEPROM cell constructed in accordance with an embodiment of the present invention utilizing ONO as the gate dielectric;

[0037] FIG. **3** illustrates a sectional view of a two bit flash EEPROM cell constructed in accordance with an embodiment of the present invention utilizing a silicon rich silicon dioxide with buried islands as the gate dielectric;

[0038] FIG. **4** is a graph illustrating the threshold voltage as a function of programming time for reading in the forward and reverse directions of a selected memory cell in accordance with this invention;

[0039] FIG. **5**A illustrates a sectional view of a flash EEPROM cell of the prior art showing the area of charge trapping under the gate;

[0040] FIG. **5**B illustrates a sectional view of a flash EEPROM cell constructed in accordance with an embodiment of the present invention showing the area of charge trapping under the gate;

[0041] FIG. **6** is a graph illustrating the difference in threshold voltage in the forward and reverse directions as a function of drain voltage for a flash EEPROM cell of the present invention that has been programmed;

[0042] FIG. **7** is a graph illustrating the difference in drain current in the forward and reverse directions as a function of drain voltage for a flash EEPROM cell of the present invention that has been programmed;

[0043] FIG. **8** is a graph illustrating the threshold voltage of a flash EEPROM cell of the present invention as a function of programming time for reading in the forward and reverse directions;

[0044] FIG. **9** is a graph illustrating the leakage current through the region of trapped charge as a function of the voltage across the charge trapping region while reading in the reverse direction;

[0045] FIG. **10** is a graph illustrating the gate voltage required to sustain a given voltage in the channel beneath the edge of the region of trapped charge while reading in the reverse direction;

[0046] FIG. **11** is a graph illustrating the effect of the gate voltage applied during reading on the difference in drain current between reading in the forward versus the reverse direction;

[0047] FIG. **12** is a graph illustrating the effect of the gate voltage (as measured by threshold channel current I_{TH}) on the difference in threshold voltage between the forward read and reverse read directions;

[0048] FIG. **13** is a graph illustrating the effect programming one of the bits has on the other bit that has not been previously programmed;

[0049] FIG. **14** is a graph illustrating the effect programming one of the bits has on the other bit that has been previously programmed;

[0050] FIG. **15** is a sectional view of a two bit EEPROM cell constructed in accordance with an embodiment of the present invention showing the area of charge trapping under the gate for both the right and the left bits;

[0051] FIG. **16** is a graph illustrating the effect of a low drain voltage on the read through of a programmed bit;

[0052] FIG. **17** is a graph illustrating the effect of programming on erase for the forward and reverse directions;

[0053] FIG. **18** is a graph illustrating the separate bit erase capability of the two bit EEPROM memory cell of the present invention;

[0054] FIG. **19** is a graph illustrating the effect of cycling on the program and erase ability of the two bit EEPROM cell of the present invention;

[0055] FIG. **20** is a graph illustrating the effect of over programming on the ability to erase for the forward and reverse directions;

[0056] FIG. **21** is a graph illustrating the programming and erasing curves for using oxide versus TEOS as the material used as the top oxide;

[0057] FIG. **22** is a graph illustrating the erase curves for two different values of drain voltage while the gate is held at ground potential;

[0058] FIG. **23** is a graph illustrating the erase curve for two different values of gate voltage;

[0059] FIG. **24**A illustrates a sectional view of a flash EEPROM cell of the prior art showing the area of charge trapping under the gate after being programmed for a period of time; and

[0060] FIG. **24**B illustrates a sectional view of a flash EEPROM cell constructed in accordance with an embodiment of the present invention showing the area of charge trapping under the gate after being programmed for a sufficient time to achieve the same threshold voltage of the cell illustrated in FIG. **24**A.

DETAILED DESCRIPTION

[0061] The two bit flash EEPROM cell of the present invention can best be understood with an understanding of how single bit charge trapping dielectric flash EEPROM memory cells are constructed, programmed and read. Thus, single bit ONO EEPROM memory cells and the method used to program, read and erase them are described in some detail. Illustrated in FIG. **1** is a cross section of an ONO EEPROM memory cell similar to one discussed in the technical article entitled "A True Single-Transistor Oxide-Nitride-Oxide EEPROM Device," T. Y. Chan, K. K. Young and Chenming Hu, IEEE Electron Device Letters, March 1987, incorporated herein by reference. The memory cell, generally referenced **41**, comprises a silicon substrate **30**, two junctions between source and drain regions **32**, **34** and substrate **30**, a non conducting nitride layer **38** sandwiched between two oxide layers **36**, **40** and a conducting layer **42**.

Programming Single Bit Memory Devices

[0062] The operation of the memory cell **41** will now be described. To program or write the cell, voltages are applied to the drain **34** and the gate **42** and the source **32** is grounded. These voltages generate a vertical and lateral electric field along the length of the channel from the source to the drain. This electric field causes electrons to be drawn off the source and begin accelerating towards the drain. As they move along the length of the channel, they gain energy. If they gain enough energy they are able to jump over the potential barrier of the oxide layer **36** into the silicon nitride layer **38** and become trapped. The probability of this occurring is a maximum in the region of the gate next to the drain **34** because it is near the drain that the electrons gain the most energy. These accelerated electrons are termed hot electrons and once injected into the nitride layer they become trapped and remain stored there. Thus, the trapped charge remains in a trapping region in the nitride.

[0063] In U.S. Pat. No. 4,173,766, issued to Hayes, the nitride layer is described as typically being about 350 Angstroms thick (see column 6, lines 59 to 61). Further, the nitride layer in Hayes has no top oxide layer.

[0064] In memory cells constructed using a conductive floating gate, the charge that gets injected into the gate is distributed equally across the entire gate. The threshold voltage of the entire gate increases as more and more charge is injected into the gate. The threshold voltage increases because the electrons that become stored in the gate screen the gate voltage from the channel.

[0065] With reference to FIG. **1**, in devices with low conductivity or non conductive floating gates, the injection of hot electrons into the silicon nitride layer causes the gate threshold voltage to increase only in the trapping region. In both conductive and non conductive floating gate memory cell designs, an increase in the gate threshold voltage causes the current flowing through the channel to decrease for a given gate voltage. This reduces programming efficiency by lengthening the programming time. However, due to the electron trapping in the non conductive floating gate memory cell design, the programming time is reduced less than with the conductive floating gate memory cell design.

[0066] The method of reading flash EEPROM memory cells will now be described. The conventional technique of reading both conductive floating gate and non conductive trapping gate EEPROM or flash EEPROM memory is to apply read voltages to the gate and drain and to ground the source. This is similar to the method of programming with the difference being that lower level voltages are applied during reading than during programming. Since the floating gate is conductive, the trapped charge is distributed evenly throughout the entire floating conductor. In a programmed device, the threshold is therefore high for the entire channel and the process of reading becomes symmetrical. It makes no difference whether voltage is applied to the drain and the source is grounded or vice versa. A similar process is also used to read non conductive gate flash EEPROM devices.

[0067] The process of programming typically includes writing followed by reading. This is true for all EPROM and EEPROM memory devices. A short programming pulse is applied to the device followed by a read. The read is actually used to effectively measure the gate threshold voltage. By convention, the gate threshold voltage is measured by applying a voltage to the drain and a separate voltage to the gate, with the voltage on the gate being increased from zero while the channel current flowing from drain to source is measured. The gate voltage that provides of channel current is termed the threshold voltage.

[0068] Typically, programming pulses (i.e., write pulses) are followed by read cycles wherein the read is performed in the same direction that the programming pulse is applied. This is termed symmetrical programming and reading. Programming stops when the gate threshold voltage has reached a certain predetermined point (i.e., the channel current is reduced to a sufficiently low level). This point is chosen to ensure that a '0' bit can be distinguished from a '1' bit and that a certain data retention time has been achieved.

The Two Bit Memory Device

[0069] A sectional view of a two bit flash EEPROM cell constructed in accordance with an embodiment of the present invention utilizing ONO as the gate dielectric is shown in FIG. **2**. The flash EEPROM memory cell, generally referenced **10**, comprises a P-type substrate **12** having two buried PN junctions, one being between the source **14** and substrate **12**, termed the left junction and the other being between the drain **16** and the substrate **12**, termed the right junction. Above the channel is a layer of silicon dioxide **18**, preferably between approximately 60 to 100 Angstroms thick, which forms an electrical isolation layer over the channel. On top of the silicon dioxide layer **18** is a charge trapping layer **20** constructed preferably in the range of 20 to 100 Angstroms thick and preferably comprised of silicon nitride, Si_3N_4. The hot electrons are trapped as they are injected into the charge trapping layer. In this fashion, the charge trapping layer serves as the memory retention layer. Note that the programming, reading and erasing of the memory cell is based on the movement of electrons as opposed to movement of holes. The charge trapping dielectric can be constructed using silicon nitride, silicon dioxide with buried polysilicon islands or implanted oxide in the nitride, for example. In the third listed

alternative, the oxide can also be implanted with arsenic, for example. Thus the lifetime of the cell of this invention is extended relative to prior art MNOS devices. The memory cell **10** is capable of storing two bits of data, a right bit represented by the dashed circle **23** and a left bit represented by the dashed circle **21**.

[**0070**] It is important to note that the two bit memory cell is a symmetrical device. Therefore, the terms source and drain as used with conventional one bit devices may be confusing. In reality, the left junction serves as the source terminal and the right junction serves as the drain terminal for the right bit. Similarly, for the left bit, the right junction serves as the source terminal and the left junction serves as the drain terminal. Thus, to avoid confusion, the terms left or first junction and right or second junction are utilized most of the time rather than source and drain. When the distinction between left and right bits is not crucial to the particular discussion, the terms source and drain are utilized. However, it should be understood that the source and drain terminals for the second bit are reversed compared to the source and drain terminals for the first bit.

[**0071**] Another layer of oxide **22** is formed over the charge trapping layer, (i.e., silicon nitride layer), and is preferably between approximately 60 to 100 Angstroms thick. The oxide layer **22** functions to electrically isolate a conductive gate **24** formed over the oxide layer **22** from charge trapping layer **20**.

[**0072**] Charge trapping dielectric materials other than oxynitride/nitride may also be suitable for use as the charge trapping medium. One such material is silicon dioxide with buried islands. The silicon dioxide with polysilicon islands is sandwiched between two layers of oxide in similar fashion to the construction of the ONO memory cell in FIG. **2**. A sectional view of a two bit flash EEPROM cell constructed in accordance with a preferred embodiment of the present invention utilizing a silicon rich silicon dioxide layer **54** with buried polysilicon islands **57** as the gate dielectric is illustrated in FIG. **3**. Note that for simplicity, only a few polysilicon islands are numbered. A P-type substrate **62** has buried N+ source **58** and N+ drain **60** regions. The silicon dioxide **54** with buried polysilicon islands **57** is sandwiched between two layers of oxide **52**, **56**. Covering oxide layer **52** is gate **50**. Similar to the two bit memory cell of FIG. **2**, the memory cell of FIG. **3** is capable of storing two data bits, a right bit represented by the dashed circle **55** and a left bit represented by the dashed circle **53**. The operation of the memory cell of FIG. **3** is similar to that of the memory cell illustrated in FIG. **2**.

[**0073**] Alternatively, the charge trapping dielectric can be constructed by implanting an impurity, such as arsenic, into a middle layer **54** of silicon dioxide deposited on top of the bottom oxide **56**.

[**0074**] An aspect of the present invention lies in the manner in which the flash EEPROM memory cell **10** (FIG. **2**) is programmed. Rather than performing symmetrical programming and reading, flash EEPROM memory cell is programmed and read asymmetrically. This means that programming and reading occur in opposite directions. Thus, programming is performed in what is termed the forward direction and reading is performed in what is termed the opposite or reverse direction.

[**0075**] It is noted that throughout the discussion of the EEPROM memory cell presented below, the voltage levels discussed in connection therewith are assumed to be independent of the power supply voltage. Thus, the power supply voltages supplied to the chip embodying the EEPROM memory device may vary while the voltages applied to the gate, drain and source thereof will be supplied from regulated voltage sources.

Programming One Bit in the Forward Direction

[**0076**] As previously mentioned, the flash EEPROM memory cell **10** of FIG. **2** is programmed similarly to the flash EEPROM memory cell of FIG. **1**. Voltages are applied to the gate **24** and drain **16** creating vertical and lateral electrical fields which accelerate electrons from the source **14** along the length of the channel. As the electrons move along the channel some of them gain sufficient energy to jump over the potential barrier of the bottom silicon dioxide layer **18** and become trapped in the silicon nitride layer **20**. For the right bit, for example, the electron trapping occurs in a region near the drain **16** indicated by the dashed circle **23** in FIG. **2**. Thus the trapped charge is self-aligned to the junction between the drain **16** and the substrate. Electrons are trapped in the nitride layer **20** near but above and self-aligned with the drain region **16** because the electric fields are the strongest there. Thus, the electrons have a maximum probability of being sufficiently energized to jump the potential barrier of the silicon dioxide layer **18** and become trapped in the nitride layer **20** near the drain **16**. The threshold voltage of the portion of the channel between the source **14** and drain **16** under the region of trapped charge increases as more electrons are injected into the nitride layer **20**.

[**0077**] It is important to note that in order to be able to subsequently erase memory device **10** effectively, the programming time period must be limited. As the device continues to be programmed, the width of the charge trapping region increases. If programming continues past a certain point the charge trapping region becomes wide whereby erasing is ineffective in removing trapped charge from the nitride layer **20**.

[**0078**] If the flash EEPROM memory cell **10** is read using the conventional technique of reading in the same direction as programming, the time needed to program the device increases to achieve the same threshold voltage. Reading in the same direction as programming means the device is programmed and read in the same forward direction. During reading, voltages having levels lower than the voltages applied during programming are applied to the gate and drain and the channel current are sensed. If device **10** is programmed (i.e., a logic '0') the channel current should be very low and if the device is not programmed (i.e., a logic '1') there should be significant channel current generated. Preferably, the difference in the channel current between the '0' and '1' logic states should be maximized in order to better distinguish between the '0' and '1' logic states.

[**0079**] Illustrated in FIG. **4** is a graph showing the rise in gate threshold voltage as a function of programming time for reading in the forward direction (curve labeled FORWARD READ) and for reading in the reverse direction (curve labeled REVERSE READ). Apparent from the graph in FIG. **4** is the several orders of magnitude reduction in programming time achieved when reading in the reverse direction versus reading in the forward direction. As is described in more detail below, this reduction in programming time is due to amplification of the effect of the trapped charge injected into the nitride layer

brought about by reading the memory cell in the opposite direction from which it was programmed. However, forward or reverse read may be chosen for devices; where the device has more than one bit per gate, usually by local charge concentration such as left and right bits, some advantage is obtained by reverse read (as shown in FIG. **3**).

[**0080**] As stated above, the time needed to program the flash EEPROM memory cell increases when reading occurs in the same direction (i.e., the forward direction) as programming. The reason for this will now be explained in more detail with reference to FIGS. **5**A and **5**B. FIG. **5**A illustrates a sectional view of a flash EEPROM cell showing the area **66** of charge trapping under the gate **42**. FIG. **5**B illustrates a sectional view of a flash EEPROM cell constructed in accordance with an embodiment of the present invention showing the area **68** of charge trapping under the gate **24** for the right bit.

[**0081**] A description of what occurs during programming is presented first followed by what occurs during reading. Note that the description that follows also may pertain to the memory cell of FIG. **2** and similarly pertains to the memory cell of FIG. **3** comprising the silicon dioxide layer **54** having buried polysilicon islands **57** substituting for the nitride layer **20** of FIG. **2**. During programming, hot electrons are injected into the nitride layer **20**, as described above. Since the nitride **20** is a nonconductor, the trapped charge remains near the drain **34** (FIG. **5**A) or **16** (FIG. **5**B). The region of trapped charge is indicated by the finely hatched area **66** in FIG. **5**A and by the finely hatched area **68** in FIG. **5**B. Thus, the threshold voltage rises, for example, to approximately 4 V, in the portion of the channel under the trapped charge. The threshold voltage of the remainder of the channel under the gate remains at, for example, approximately 1 V. If the device is now read in the forward direction (i.e., voltages are applied to the gate and drain as indicated by the arrow in FIG. **5**A), electrons move off the source and begin traveling toward the drain. When a logic '0' is programmed, there can be little or no channel current through the device when it is read. Thus, only if a sufficient portion of the channel is turned off, can the electron current be stopped. If the channel cannot be completely turned off, the electrons will reach the drain. Whether the electrons reach the drain will be determined by, among other things, the length of the trapping area. If the memory cell is programmed for a sufficiently long period, eventually, the channel stops conducting when read in the forward direction. If the trapped charge region (the programmed area) **66** (FIG. **5**A) is not long enough, electrons can punch through to the drain **34** in the depletion region under the trapped charge **66**.

[**0082**] When the device is read in the forward direction, a voltage is applied to the drain and the gate, for example 2V and 3V, respectively, and the source is grounded. Inversion occurs in the channel under the area of the nitride **38** that does not have trapped charge. A vertical electric field exists in the channel that spans the length of the channel up to the region of the channel underneath the trapped charge **66**. In the inversion region, electrons travel in a linear fashion up to the edge **35** of the inversion region, which is beneath the left edge **33** of the trapped charge region **66**. This is indicated by the line shown in the channel region in FIG. **5**A that extends from the source to just beneath the edge **33** of the region of trapped charge **66**. Due to the fact that the device is in inversion (i.e., the channel is in a conductive state), the potential in the inversion layer is pinned to ground potential because the source is grounded. The voltage in the inverted channel near the trapped charge (i.e., just to the left of the right edge **35** of the channel inversion region) is approximately zero. Thus, the voltage across the region of trapped charge is close to the full drain potential of 2 V. Due to the drain potential across the channel region beneath the trapped charge **66** some of the electrons punch through across the trapped region to the drain, resulting in a channel current.

[**0083**] The diagonal line under the channel in FIGS. **2** and **5**A indicate the reduction in the number of electrons in the channel as a function of channel distance. The channel region under the trapped charge is off (i.e., not inverted) due to the high threshold voltage required to invert this region under the trapped charge. However, the channel region inside the dashed circle **23** in FIG. **2** and under the region **66** in FIG. **5**A is a depletion region because the device is in saturation (a device will be in saturation when V_{DS}, the voltage from drain to source, is higher than V_{DSAT}, the saturation voltage). Due to the voltage on the drain **34**, a lateral electric field exists in this portion of the channel under region **66**. As a result of this lateral electric field, an electron arriving at the edge of the depletion region will be swept through and pulled to the drain **34**. As described earlier, this phenomenon is called punch through. Punch through occurs if the lateral electric field is strong enough to draw electrons through to the drain, regardless of the threshold level. In order to prevent punch through from occurring during a read, the prior art memory cells require a longer programming time than does the memory cell of this invention because the prior art memory cells are read in the forward direction. As the memory device is programmed for a longer and longer time, more and more electrons are injected into the nitride, increasing the length of the programmed portion **66** (FIG. **5**A) of the channel. The memory cell must be programmed for an amount of time that yields a trapped charge region **66** of sufficient length to eliminate the punch through of electrons. When this occurs, the lateral electric field is too weak for electrons to punch through to the drain under normal operating conditions. As an example, for the threshold voltage equaling 3V during read in the forward direction, FIG. **4** shows that at programming time of approximately 3 milliseconds is required.

[**0084**] However, if the flash EEPROM memory cell **10** (FIG. **5**B) is read in the reverse direction, a different scenario exists. Reading in the reverse direction means reading in a direction opposite to that of programming. In other words, voltages are applied to the source **14** and the gate **24** and the drain **16** is grounded. Similar to the memory device of FIG. **5**A, the memory device of FIG. **5**B is programmed in the forward direction by injecting hot electrons into region **68** of the nitride layer **20**. Since nitride **20** is a nonconductor, the trapped charge remains near the drain, for example. The region of trapped charge is indicated by the finely hatched area **68** in FIG. **5**B. Thus, the threshold voltage rises, for example, to approximately 4V only in the portion of the channel under the trapped charge **68**. The threshold voltage of the remainder of the channel remains at, for example, approximately 1 V.

[**0085**] To read the right bit of the device of FIG. **5**B in the reverse direction, a voltage is applied to the source **14** and the gate **24**, for example 2V and 3V, respectively, and the drain **16** is grounded. A difference between reading in the forward direction and reading in the reverse direction is that when

reading in the reverse direction, the gate voltage required to put the channel of the memory device into inversion increases. For the same applied gate voltage of 3V, for example, there will be no inversion but rather the channel of the memory device will be in depletion. The reason for this is that the channel region next to the drain **16** (which functions as the source in read) is not inverted due to the electron charge in region **68** of the nitride **20**. The channel adjacent the source **14** (which functions as the drain in read) is not inverted because 2V is applied to the source **14** and the channel, to be inverted, must be inverted relative to 2 V. In the case of reading in the reverse direction, in order to sustain a higher voltage in the channel, a much wider depletion region must be sustained. A wider depletion region translates to more fixed charge that must be compensated for before there can be inversion. When reading in the reverse direction, to achieve a voltage drop across the charge trapping region **66** of the device shown in FIG. **5**A similar to the voltage drop achieved when reading the same device in the forward direction, a higher gate voltage is required, for example, 4 V. This is in contrast to the memory device where the source was grounded and a lower gate voltage was required to invert the channel. In the memory device, for example, as shown in FIGS. **2**, **3** and **5**B, a much higher gate voltage is required to pin the voltage in the channel to a higher voltage, i.e., the 2V that is applied to the source terminal rather than ground. In other words, device(s) recognizes and takes advantage of the fact that for the same magnitude potential across the drain and the source, the voltage across the portion of the channel under the trapped charge region **68** (FIG. **5**B) is reduced when reading occurs in a reverse direction to writing (programming) directly resulting in less punch through and greater impact of the programming charge injected in region **68** of the nitride layer **20** (FIG. **5**B) on the threshold voltage of the transistor. As an example, for the threshold voltage V_T equaling 3V during reverse read, FIG. **4** shows that a programming time of approximately 2 microseconds is required. This programming time is less than the programming time required for the same threshold voltage when the cell is read in the forward direction.

[**0086**] Memory cells utilizing ONO structure have had difficulty retaining the charge in the nitride layer. This is because such memory cells are programmed in a first forward direction and then read in the same direction. The reading of the programmed cell in the forward direction requires a significant amount of charge to be stored on the nitride to provide the desired increase in threshold voltage associated with the programmed cell. However, by reading in the reverse direction, less charge is required to be stored on the nitride to achieve the same increase in threshold voltage in a programmed cell. FIG. **4** shows the difference in charge (measured as a function of programming time required to achieve a given threshold voltage V_T) for reading in the reverse direction versus the forward direction. In the prior art, the charge retention in a localized region of the silicon nitride layer was difficult if not impossible to achieve because the lateral electric field generated by the charge dispersed the charge laterally in the nitride layer. Such dispersion particularly occurred during the high temperature retention bake required for quality control and reliability. The high temperature retention bake typically requires temperatures between 150 degrees Centigrade to 250 degrees Centigrade for at least 12 to 24 hours. The charge in the prior art devices typically dispersed through the nitride during the high temperature bake causing the performance of prior art devices using the nitride layer as a charge retention material to be less than satisfactory. Accordingly, prior art devices that use the nitride layer for charge retention are not widely used. In addition, charge stored on the nitride layer in prior art memory cells is prone to lateral diffusion and dispersion through the nitride layer in response to the retention bake due to the internal fields causing what is known as electron hopping. The phenomenon of electron hopping is dependent on the field strength. In the case of charge in the nitride layer the internally generated electric field is directly related to the amount of charge stored on the nitride layer. Because electron hopping is dependent upon the electric field strength, the additional charge required to obtain a given threshold voltage change or shift when the memory cell is read in the same direction as it was programmed causes a change in the charge distribution in the nitride layer. This change in the charge distribution degrades the threshold voltage from the intended (i.e., design) threshold voltage. Consequently, prior art ONO devices have not been successful.

[**0087**] In accordance with the present invention, by reading the memory cell in the reverse direction from which the memory cell is programmed, the amount of charge required to achieve a given threshold voltage is reduced in some cases by a factor of two or three times the amount of charge required to obtain the same threshold voltage shift when the memory cell is read in the forward direction. Accordingly, the internal electric fields generated by the charge in the nitride when the memory cell is to be read in the reverse direction are less than the internal electric fields associated with the charge stored on the nitride when the memory cell is to be read in the forward direction. Consequently electron hopping is reduced and the small amount of charge stored in the nitride does not disperse laterally through the nitride due to the internally self generated electric fields even during retention bake. Consequently, the memory cell of the present invention does not suffer the degradation in performance and reliability of prior art ONO memory cells which are programmed and read in the same direction.

Sample Flash EEPROM Device Data

[**0088**] Data obtained from flash EEPROM devices constructed in accordance with the present invention will now be presented to help illustrate the principles of operation thereof. A graph illustrating the difference in threshold voltage in the forward and reverse directions as a function of drain voltage for a flash EEPROM cell of the present invention that has been previously programmed is shown in FIG. **6**. The memory cell used to obtain the data presented in FIGS. **6**, **7** and **8** was constructed with a bottom oxide layer **18**, a top oxide **22** and a nitride layer **20**, each 100 Angstroms thick. The drawn width of the channel measures 0.6 microns and the drawn length of the channel measures 0.65 microns.

[**0089**] While reading in the forward direction, the threshold voltage is approximately the same as the threshold voltage when reading in the reverse direction for low drain voltages. At low drain voltages there is insufficient potential for punch through to occur. However, as the drain voltage increases while reading in the forward direction, the punch through region increases resulting in lower threshold voltage. At a high enough drain voltage, the entire portion of the channel under the trapped charge in region **68** of nitride layer **20** (FIG. **5**B) is punched through and the threshold voltage levels off at the original threshold voltage of the channel.

[0090] However, while reading in the reverse direction, the (V_T versus V_D) curve appears to follow the V, versus V_D curve while reading in the forward direction at low drain voltages. However, the curves rapidly diverge for higher drain voltages and the threshold voltage for reading in the reverse direction levels off at approximately 4V. At a gate voltage V_G of approximately 4V and a drain voltage V_D of 1.2V, the device has reached saturation (V_{DSAT}). At this gate voltage, any further increase in V_D cannot be transferred through the inversion layer thus establishing the maximum potential drop across the portion of the channel beneath the charge trapping region **68**. The V_T then becomes independent of further increases in V_D. For example, at a drain voltage of 1.6V, the difference in V_T between reverse and forward read is almost 2V.

[0091] A graph illustrating the difference in drain current in the forward and reverse directions as a function of drain voltage for a flash EEPROM cell of the present invention that has been programmed is shown in FIG. **7**. In FIG. **7**, rather than measure threshold voltage, the drain current is measured while keeping the gate voltage constant. In the forward direction, as expected, the drain current I_D increases as the drain voltage V_D increases. The curve labeled FORWARD also resembles the I_D curve for reading an unprogrammed cell in the reverse direction.

[0092] The drain current while reading in the reverse direction also increases with increasing drain voltage (measured at the source which functions as the drain when reading in the reverse direction) but the drain current levels off at a lower current than when reading in the forward direction. If the logic threshold for this memory cell is set to 10 μA, the forward curve can represent a logic '0' and the reverse curve a logic '1'.

[0093] The voltage V_X is defined as the voltage in the channel at a distance X from the source. Using the example presented above, the voltage V_X that exists in the channel of the memory cell of the present invention (FIG. **5**B, for example) will not be 2V because the device is in depletion rather than inversion. On the other hand, the voltage V_X must be larger than 0 because a gate voltage of only 1.5V is able to sustain approximately 0.4V in the channel. The actual voltage in the channel varies across the channel length because of the lateral electric field set up between the source and the drain. The threshold voltage, however, varies as a function of the voltage in the channel.

[0094] With reference to FIG. **5**B, the channel will be in saturation as long as the gate voltage V_G is higher than the threshold voltage V_T and the voltage V_X at any point in the channel is given by

$$V_X = V_{DSAT}$$

with

$$V_{DSAT} = V_G - V_T = V_G - V_T(V_{DSAT})$$

and

$$V_T(V_X) = V_{T0} + \Delta V_T(V_X)$$

[0095] As is shown in the above equations, the threshold voltage in the channel is equal to the threshold voltage with the source at zero potential V_{T0} plus a delta threshold voltage ΔV_T, which is itself a function of the voltage in the channel.

[0096] The leakage current through the channel under the region **68** of trapped charge, plotted as a function of the voltage V_{TC}, across the portion of the channel under the charge trapping region **68** while reading in the reverse direction, is shown in FIG. **9**. From the graph, one can see that the approximate leakage current I_L through the channel when V_{TC} is 2V is 10^{-5} A. In the case of the memory cell read in the forward direction, the voltage across the portion of the channel under region **66** of trapped charge is approximately 2V. In contrast, the voltage V_X in the channel of the memory device of the present invention at location **27** beneath the edge **25** of the region **68** of trapped charge is not 2V but something less, 1V for example. The leakage current I_L corresponding to 1V across the trapped charge region is smaller.

[0097] The edge of the region of trapped charge formed in the nitride layer during programming is the portion of the trapped charge that begins to affect the gate voltage required to invert the channel beneath that point.

[0098] A graph illustrating the gate voltage required to sustain a given voltage in the channel, V_X, spanning the distance from the drain to the edge **27** of the channel under the edge **25** of the charge trapping area for one of the two bits while reading in the reverse direction is shown in FIG. **10**. The gate voltage V_G that is required to sustain a particular V_X at the point **27** in the channel under the edge **25** of the charge trapping area **68** (FIG. **5**B) is a function of the number of acceptors N_A in the substrate and the thickness of the oxide T_{OX} and is represented by the dashed/dotted line. The solid line represents the threshold voltage in the channel that exists when the back bias effect on the threshold voltage is zero. In this case, the threshold voltage is constant along the entire channel. However, once there is a voltage in the channel, the threshold voltage is not constant along the channel. As shown in the graph, the threshold voltage increases nonlinearly as the voltage in the channel increases. The relationship between the incremental increase in threshold voltage as a function of channel voltage is well known in the art. A more detailed discussion of this relationship can be found in Chapter 2 of "The Design and Analysis of VLSI Circuits" by L. A. Glasser and D. W. Dobberpuhl, incorporated herein by reference.

[0099] It is important to emphasize that the advantages and benefits of reading in the reverse direction are achieved when combined with the use of relatively low gate voltages. For a particular drain voltage, e.g., 2V, applying a high enough V_G such as 5V, for example, causes the differences in threshold voltages between forward and reverse reading to fade. A graph illustrating the effect of the gate voltage V_G applied during reading on the difference in drain current I_D between reading in the forward direction versus reading in the reverse direction for one of the two bits is shown in FIG. **11**. The reverse V_T of the device used to generate the curves in the Figure is 3.5V. From FIG. **11** it can be seen that as V_G is increased while V_D is kept constant, the I_D curves for the reverse read begin to resemble the curves for the forward read. For example, comparing the forward and reverse read curves when V_G equals 2.5V shows the read current in the reverse direction being about four orders of magnitude lower. At a gate voltage V_G of 3V, the difference in read current between the forward and reverse directions drops to a little more than two orders of magnitude. At a gate voltage of 5V, the difference in read current is only approximately 15%. These curves clearly show that large differences in I_D between the forward and reverse read directions are obtained when V_G is chosen to be low enough. Thus, the benefits of reading in the reverse direction are achieved when suitably low gate voltages are

used for reading. There is an optimum range within which V_G should lie. If V_G is too low, insufficient current is developed in the channel. On the other hand, if V_G is chosen too high, the differences between reading in the reverse and forward directions are greatly diminished.

[0100] A graph illustrating the effect of the gate voltage on the difference in threshold voltage between the forward and reverse directions is shown in FIG. **12**. The device used to generate the curves in FIG. **12** was programmed once to a V_T of 3.5V using a V_D of 1.6V and an I_{TH} of I_{HT} of 1 μA. The V_T as a function of V_D during reading was subsequently measured. As labeled in FIG. **12**, the I_{TH} level for the lower two curves is 1 μA, and is 40 μA for the upper two curves. The effect of raising the I_{TH} is to force the V_T, measurement to be at a higher V_G level even though the amount of charge trapped in the silicon nitride layer is identical for all measurements. For the lower two curves (I_{TH} of 1 μA) the forward and reverse threshold voltages start to separate from each other at a V_D of approximately 50 mV while the V_T, for the reverse saturates at approximately 0.6 V. For the upper two curves (I_{TH} of 40 μA) the forward and reverse threshold voltages start to separate from each other at a V_D of approximately 50 mV while the V_T, for the reverse saturates at approximately 0.6V. For the upper two curves (I_{TH} of 40 μA) the forward and reverse threshold voltages start to separate from each other at a V_D of approximately 0.35V while the V_T for the reverse saturates at approximately 1.35V. Thus, these curves clearly show that the effect of the trapped charge depends heavily on the choice of V_G.

Programming the Two Bit Cell

[0101] With reference to FIG. **2**, programming the two bit EEPROM cell. In programming the two bit cell, each bit, i.e., the left and right bit, is treated as if the device was a single bit device. In other words, both the left and right bits are programmed as described in the section entitled "Programming One Bit in the Forward Direction." For the right bit, for example, programming voltages are applied to the gate **24** and drain **16** and hot electrons are injected into and trapped in the charge trapping layer **20** in the region near the drain defined by the dashed circle **23**. Correspondingly, the threshold voltage of the portion of the channel under the trapped charge increases as more and more electrons are injected into the nitride layer.

[0102] Similarly, the left bit is programmed by applying programming voltages to the gate **24** and source **14**, which now functions as the drain for the left bit. Hot electrons are injected into and trapped in the charge trapping layer **20** in the region defined by the dashed circle **21**. The threshold voltage of the portion of the channel under the trapped charge comprising the left bit increases as more and more electrons are injected into the nitride layer.

[0103] A graph illustrating the effect programming one of the bits has on the other bit, which has not been previously programmed is shown in FIG. **13**. In this particular example, the right bit is shown being programmed while the left bit is read. The threshold voltage V_T for the right bit assumes that the right bit is read in the reverse direction to the programming direction. Thus the threshold voltage for a programmed left bit will be relatively low compared to the threshold voltage for the right bit and thus the state of the right bit can be read without interference from the left bit. It is clear from the curves that during programming of the right bit, the unprogrammed left bit remains unprogrammed. This graph also illustrates the read through of the programmed right bit in order to perform a read of the left bit.

[0104] A graph illustrating the effect programming one of the bits has on the other bit, which has been previously programmed is shown in FIG. **14**. This graph was generated in two passes. Each curve is labeled either PASS #**1** or PASS #**2**. During the first pass, the right bit was programmed while reading the unprogrammed left bit, as shown by the curves labeled RIGHT BIT-PASS #**1** and LEFT BIT-PASS #**1**. These curves are similar to the curves of FIG. **13**. During the second pass, once the right bit is programmed, the left bit, previously unprogrammed, is now programmed. At the same time, the right bit is read. The second pass is represented by the curves RIGHT BIT-PASS #**2** and LEFT BIT-PASS #**2**.

[0105] As shown in FIG. **14**, during the first pass, the left bit remains unprogrammed during the programming of the right bit. Programming the right bit does not affect the unprogrammed left bit. During the second pass, the left bit is programmed and the right bit remains programmed and can still be read. The gate voltage during programming is sufficiently high (typically around 10V) that the programmed right bit does not interfere with the programming of the left bit except to increase somewhat the time required to reach a given threshold voltage relative to the time required to reach the same threshold voltage for the right bit when the right bit is programmed. The graph also shows that the right bit can be programmed during programming of the left bit. Further, the programming of the left bit does not disturb the programmed right bit. This is possible because program through (i.e. the programming of the one bit substantially without interference from the other bit when the other bit is programmed) and read through (i.e. the reading of one bit without interference from the other bit when the other bit is programmed) occurs through both the left and the right bits.

[0106] Program through and read through are possible due to the relatively low gate voltages required to turn on each programmed bit when read in the forward direction as occurs when the other bit is read in the reverse direction. Another way to look at this is that a narrow charge trapping region permits punch through to be more effective. Thus the small amount of charge **68** trapped on the right edge of charge trapping layer **20** (FIG. **15**) and self-aligned with the junction between region **16** and the substrate **12** and a comparable amount of charge **70** trapped on the left edge of charge trapping layer **20** and self-aligned with the junction between region **14** and the substrate **12** cause a narrow charge trapping region to be formed at both the right side and the left side of charge trapping layer **20** which is easy to be punched through when the bit is read in the forward direction. Thus when left bit **70** (the charge trapping region **70** is referred to as a bit because the presence or absence of charge in region **70** would represent either a zero or a one) is read in the forward direction, bit **68** is being read in the reverse direction. The punch-through under charge trap region **70** is quite easily achieved with a low gate voltage thereby allowing the charge trapped in bit **68** to control the state of the signal read out of the device. Thus for equal amounts of charge trapped in regions **70** and **68**, reading a bit in the reverse direction results in the opposite bit having no effect on the state of the signal being read.

[0107] Another reason that the bit not being programmed is not disturbed is that the programming voltage is not being

applied to the drain for the bit previously programmed. When programming the other bit, the programming voltage is applied to the drain for the bit on the other side of the device.

[0108] As discussed earlier, the programming duration must be limited for each bit in order that the other bit can still be read. For example, in the case when the right bit is programmed, i.e., a logic '0', and the left bit is not programmed, i.e., a logic '1', if the right bit was programmed for too long a time then when the left bit is read, there may be insufficient current for the sense amps to detect a logic '1' because the channel is not sufficiently conductive. In other words, if the right bit is programmed too long, a left logic '1' bit becomes slower, i.e., takes longer to read due to lower channel current, or, in the worst case, may appear to be a logic '0' because the over-programmed right bit prevents the left bit from being read. Thus, a window exists in the programming time within which a logic '0' bit must fall. One of the variable parameters is the voltage that is applied to the functional drain region during read. As the drain voltage is increased, a longer programming time, i.e., longer area of trapped charge, is required in order to avoid punch through. Thus, a longer trapping region is equivalent to increasing the programming time. The upper limit of the programming time for the window is the programming time such that a forward read does not change the read current by more than a predetermined percentage compared to the read current for a reverse read. Preferably, the percentage change to the read current should be limited to 10%. This percentage, although not arbitrary, can be optimized according to the design goals of the chip designer. For example, a designer may wish to have three orders of magnitude margin between the threshold voltage of a forward read and the threshold for a reverse read. To achieve this, the gate voltage, drain voltage and implant level are all adjusted accordingly to determine a maximum programming time.

[0109] The effect of programming one of the bits is that both programming and reading for the second bit is slowed somewhat. The second bit can be programmed as long as the gate voltage during programming is higher than the threshold voltage of the channel with the first bit programmed and sufficient voltage is placed on the drain. The channel resistance, however, is raised due to the programming of the fist bit. As long as programming parameters are tuned properly, the higher channel resistance does not prevent the second bit from being programmed and read. The higher channel resistance, however, does cause programming and reading of the second bit to take longer.

[0110] In reading the two bit cell, as in programming, each bit is treated as if the device was a single bit device. A sectional view of a two bit EEPROM cell constructed in accordance with a preferred embodiment of the present invention showing the area of charge trapping under the gate for both the right and the left bits is shown in FIG. **15**. The area of trapping for the right bit is referenced **68** and that of the left bit is referenced **70**. Also shown in FIG. **15** are two arrows labeled 'READ', one pointed in the left direction indicating the direction for reading of the right bit and one pointed in the right direction indicating the direction for reading of the left bit.

[0111] As described the right bit is read in the reverse direction by applying read voltages to the source **14** and the gate **24** and grounding the drain **16**. For example, a gate voltage of 3V and a source voltage of 2V is applied. The resulting voltage in the channel V_X will be something less than two volts in accordance with the graph in FIG. **10** and as described in detail above. Similarly, to read the left bit in the reverse direction, read voltages are applied to the gate **24** and to the drain **16** and the source **14** is grounded, e.g., 3V on the gate and 2V on the drain.

[0112] A graph illustrating the effect of a low drain voltage on the read through of a programmed bit is shown in FIG. **16**. This graph is similar to that of FIG. **14** with the addition of the top two curves above 5.1V. The four lower curves were generated using a V_D of 1.6V. The two upper curves were generated by reading the unprogrammed bit after the other bit was programmed using a V_D of 50 mV. These curves show that if V_D is made too low and the first bit is programmed, insufficient voltage exists in the channel for read through to occur. They also show that the second bit to be programmed, in this case the left bit, experiences slower programming due to the increased series resistance of the channel. Even if the second bit is unprogrammed, when the drain voltage is too low and the first bit is programmed, the second bit cannot be read properly. Insufficient voltage exists in order for punch through to occur. If punch though does not occur, the second bit looks as if it is programmed whether it really is or not.

[0113] Punch through is sensitive to the length of the trapped charge region, such as regions **68** and **70** of the structure shown in FIG. **15**. Should these regions be too wide or not self-aligned with the appropriate region **16** or **14** (depending on whether the charge represents the right bit **68** or the left bit **70**), then punch through would not be able to be guaranteed to occur.

[0114] A read of the two bit memory device of the present invention falls into one of three cases: (1) neither of the two bits are programmed (2) one of the bits is programmed and the other is not or (3) both of the bits are programmed. The first case does not require a read through. The second case requires reading through the programmed bit to read the unprogrammed bit. In this case the margin is the delta between reading a single bit in the forward direction versus the reverse direction. An example of the margin can be seen in FIGS. **6** and **7** which illustrate the difference in V_T and read current between the forward and the reverse directions for a single bit.

[0115] The third case requires read through to read both programmed bits. Programming the second bit, in fact, improves the conditions for reading the first bit. This is so because the voltage in the channel is further reduced over the case of reading a single bit. This increases the read margins between programmed and unprogrammed bits.

[0116] It is important to note that the FIG. **15** EEPROM cell stores two bits; support circuitry and concepts designed to work with single bit memory cells such as FIG. **1** and others can still be used. For example, the sense amplifier circuitry needed for the two bit memory cell is basically no different than that for the single bit memory cell. In the single bit memory cell, the sense amplifier circuitry is required to distinguish between two states, the programmed and unprogrammed states. Likewise, in the two bit memory cell, the sense amplifiers also distinguish between only two states: programmed and unprogrammed. This is in direct contrast to the prior approaches to multi-bit memory cells wherein multiple thresholds are used which require multiple current levels to be detected by the sense amps. Accurately detecting multiple current levels in a memory device is a complex and difficult task to accomplish.

[0117] In the case when one of the bits is unprogrammed, i.e., no charge injected into charge trapping layer for that bit, a read of the other bit will be unaffected by this unprogrammed bit. On the other hand, however, in the case when one bit is programmed, a read of the other bit will be affected by this other programmed bit to some extent. Depending on various process parameters, the programmed bit may cause the channel to be less conductive. However, as long as the channel is sufficiently conductive both bits can still be programmed and read correctly. This is discussed in more detail below.

[0118] With reference to FIG. **15**, the two bit memory device utilizes a punch through or read through technique to read one bit when the other bit is in a programmed state. In order to read, for example, the right bit **68**, the read current must be able to read through or punch through the left bit **70**, assuming that both the left bit and the right bit have been programmed. Thus, there is a limit on the length of the charge trapping region that can be programmed. The charge trapping region must be short enough to permit punch through of the bit not being read. If a bit is in the unprogrammed state, there is no constraint on the read current of the other bit from the unprogrammed bit.

[0119] It is important to note that when a semiconductor device is scaled, the channel lengths become shorter and short channel effects take hold. Thus, in the two bit memory cell, because each bit is stored in different areas of the transistor, short channel effects may become prevalent sooner than in the case of the single bit transistor. In order to retain the usable range of drain voltage, the two bit transistor may need to be scaled by a smaller factor.

Criteria Necessary For Two Bit Operation

[0120] A key concept associated with the two bit EEPROM memory cell of the present invention is that for the device to operate properly, both bits must be able to be written and read. If one of the bits is programmed, a reverse read on the programmed bit must sense a high V_T, i.e., a '0' and a reverse read on the unprogrammed bit must sense a low V_T, i.e., a '1'. Thus, a reverse read on the unprogrammed bit, which is equivalent to a forward read on the programmed bit, must punch through the region of trapped charge in order to generate a high enough read current. If this does not happen, the unprogrammed bit will not be able to be read as a '1', i.e., a conductive bit.

[0121] In order to achieve this goal, a sufficient margin is generated between reading in the forward and reverse directions. With reference to FIG. **11**, in order to store two bits, there must be sufficient difference between forward read of one of the bits and reverse read of the other bit. In addition, the reverse read current for one of the bits when the other bit is and is not programmed should be sufficient to distinguish between the two bits. For example, in FIG. **11**, for a gate voltage of 3V, punch through for reading in the reverse direction occurs at approximately 1V. Thus, a drain voltage of 1.6V creates a suitable safety margin ensuring that the second bit can be read when the first bit is programmed.

[0122] There are two parameters that can be used to ensure punch through of the charge trapping region. The first is the V_G, applied during reading and the second is the width of the charge trapping region. A low V_G used during reading combined with a narrow charge trapping region makes punch through more effective. The lower gate voltage produces a weaker vertical electric field which causes the lateral electric field to be stronger.

[0123] It is more important to use a low V_G during reading in the two bit memory cell than in the single bit memory cell. In the single bit case, it only had to be ensured that the reverse read was better than the forward read, meaning that the V_T of a given bit during forward reading was lower than the V_T of this bit during reverse reading. In the two bit case, however, it is not enough that the V_T drops in the forward case, it must drop sufficiently to be able to punch through when reading the other bit. If the delta V_T between the forward and reverse read is not sufficient, one bit cannot be read when the other bit is programmed.

Erasing Memory Devices

[0124] A consequence of using an oxide-nitride structure as opposed to an oxide-nitride-oxide structure is that during programming the charge gets distributed across the entire nitride layer, thus providing one form of single bit per cell device. The absence of the top oxide layer lowers the ability to control where the charge is stored in the nitride layer and allows holes from the gate to neutralize charge in the nitride layer, whether enhanced or adulterated with oxygen, or not. A thick nitride layer is required in order to generate sufficient charge retention in the device. However, the relatively thick nitride layer causes the charge trapping region to be very wide thus making erasing the cell difficult. Thus there is a tradeoff between charge retention and sufficiently large threshold voltage deltas on the one hand and the ability to erase the device on the other hand.

[0125] Some devices that use hot electron programming utilize an erase mechanism whereby the electrons previously trapped in the nitride are neutralized (i.e., erased) by transferring holes into the nitride. The information is erased by grounding the gate and applying a sufficient potential to the drain to cause avalanche breakdown. Avalanche breakdown involves hot hole injection and requires relatively high voltages on the drain for the phenomenon to occur. The hot holes are generated and caused to jump over the hole potential barrier of the bottom oxide between the channel and the nitride and recombine with the electrons in the nitride. This mechanism, however, is very complex and it is difficult to construct memory devices that work in this manner. Another disadvantage of using hot hole injection for erasing is that since the drain/substrate junction is in breakdown, very large currents are generated that are difficult to control. Further, the number of program/erase cycles that the memory cell can sustain is limited because the breakdown damages the junction area. The damage is caused by very high local temperatures generated in the vicinity of the junction when it is in breakdown.

Erasing the Two Bit Memory Cell

[0126] The erase mechanism of the two bit flash EEPROM memory cell **10** (FIG. **15**) will now be described in more detail. The mechanism used to erase the two bit flash EEPROM memory cell of the present invention involves the movement of electrons as opposed to the movement of holes. For the right bit, an erase is performed by removing electrons from the charge trapping nitride region **68** either through the gate **24** via the top oxide **22** or through the drain **16** via the bottom oxide **18**. For the left bit, an erase is performed by

removing electrons from the charge trapping nitride region **70** either through the gate **24** via the top oxide **22** or through the source **14** via the bottom oxide **18**.

[**0127**] Using the right bit as an example, one technique of erasing is to simultaneously apply a negative potential to the gate **24** and a positive potential to the drain **16** such that electron tunneling occurs from the charge trapping nitride layer **20** to the drain **16** via the bottom oxide **18**. The left bit is erased in a similar fashion except that a positive potential is applied to the source **14** rather than the drain **16**. The electron tunneling is substantially confined to a local area near the drain **16**. To facilitate the erasing of the memory cell **10** using this technique, the thickness of the bottom oxide layer **18** is suitably constructed (i.e., has a thickness of about seventy (70) Angstroms) to optimize the removal of electrons from the nitride charge trapping layer **20** into the drain **16**.

[**0128**] Using the right bit as an example, a second well known technique is to simultaneously apply a positive voltage potential to the gate **24** and zero potential, i.e., ground, to the drain **16** such that electron tunneling occurs from the charge trapping nitride layer **20** through the top oxide **22** to the gate **24**. The right bit is erased in a similar fashion with zero potential applied to the source **14**. In this case, the top oxide **22** is suitably constructed (again with a thickness of about seventy (70) Angstroms) to optimize the tunneling of electrons from the nitride charge trapping layer **20** into the gate **24** in order to facilitate the erasing of the memory cell **10**. In one embodiment, the top oxide **22** has a thickness of 50 Angstroms to 80 Angstroms for a voltage on gate **24** of 10 to 18 volts.

[**0129**] A graph illustrating the effect of programming on erase time for reading in the forward and reverse directions is shown in FIG. **17**. FIG. **17** shows the times necessary to program the device to a threshold voltage of four (4) volts for reading$_3$ in both the reverse (10^{-5} seconds) and forward ($3\times10^{}$ seconds) directions. The graph presented in FIG. **17** is based on data obtained from a memory cell. In the first pass, the device was programmed to be read in the reverse direction and then erased. In the second pass, the device was programmed to be read in the forward direction and then erased. The erase processes for the charges associated with reverse read and forward read used the same drain voltage and gate voltage, namely a V_D of 5.5V and a V_G of −8V. The thickness of the top oxide, bottom oxide and nitride layers are all 100 Angstroms. Programming for forward reading and reverse reading utilized a V_D of 5.5V and V_G of 10V. Only the programming times differ. The forward and reverse programming curves are identical to those illustrated in the graph of FIG. **8**.

[**0130**] As can be seen from FIG. **17**, even when the device is programmed to the same threshold voltage, the time to complete the reverse erase is less than the time to complete the forward erase. The forward erase (i.e. the time to remove the trapped charge associated with a given threshold voltage when the device is read in the forward direction) is slower than the reverse erase (i.e. the time to remove the trapped charge associated with a given threshold voltage when the device is read in the reverse direction). In addition, there is residual charge left in the charge trapping region as shown in the small gap between the reverse and forward erase curves at the one (1) second mark. This is due to the larger wider charge trapping region formed during the forward programming that was required to generate a threshold voltage of 4V. From the curves, the forward erase is approximately an order of magnitude slower than the reverse erase. The abrupt increase in threshold voltage for the curve labeled 'FORWARD ERASE' is due to the reverse read used to measure the threshold voltage. For the same amount of charge trapping, the equivalent threshold voltage for reverse reading is higher than that for forward reading. As can be seen in FIG. **17**, the slopes of the forward and reverse erase curves are different. Reading in the reverse direction requires trapped charge smaller than does reading in the forward direction that the erase of the trapped charge is faster. Also apparent from FIG. **17** is that the cell does not enter deep depletion. Even at the 1 second erase mark, the threshold voltage (about 2v) is no lower than that of an unprogrammed cell. An advantage over prior art memory cells especially floating gate cells where over-erase can cause a failure of the memory array due to deep depletion of the charge on the floating gate.

[**0131**] The erase mechanism in the memory cell is self limiting due to the fact that as the memory cell is erased, more and more positive charge is stored in the trapping region **68** (FIG. **15**) (for the right bit) of the nitride layer thereby neutralizing the negative charge stored there while the remainder of the nitride layer **20** remains unaffected. Thus, the threshold voltage of the channel keeps dropping until it levels off at the threshold voltage of an unprogrammed memory cell which is the threshold voltage of the larger majority of the channel closer to the source. Over-erasing the memory cell only affects (i.e., lowers) the threshold voltage of the portion of the channel under the charge trapping region **68** which is a relatively narrow region while leaving the threshold voltage of the remainder of the channel at its normal value. A graph illustrating the separate bit erase capability of the two bit EEPROM memory cell is shown in FIG. **18**. The graph was generated in two passes and initially, both the right and the left bit are programmed each with an amount of trapped charge to achieve a given threshold voltage when read in the reverse direction. During the first pass, the right bit was erased while the left bit was read, as represented by the curves labeled RIGHT BIT-PASS #**1** and LEFT BIT-PASS #**1**. During the second pass, the left bit was erased while the right bit was read, as represented by the curves labeled RIGHT BIT-PASS #**2** and LEFT BIT-PASS #**2**. The graph shows that erasing of one of the bits does not affect the other bit. This is due to the fact that the erase voltage is localized to the junction adjacent to the bit that is to be erased. The difference in location between the curve labeled "Left Bit-Pass #**2**" and the curve labeled "Right Bit-Pass #**1**" is of no significance being well within the tolerance of the measurements.

[**0132**] A graph illustrating the effect of cycling on the program and erase ability of the two bit EEPROM cell is shown in FIG. **19**. The graph shows the V_T of a bit associated with a given amount of trapped charge for reading in the reverse direction (top line) and the forward direction (bottom line). The gradual increase in threshold voltage V_T for reading in both the forward and reverse directions reflects the lack of complete erasure of all the stored charge during each erase such that the amount of trapped charge gradually increases with time after programming and erasing for 1000 cycles.

[**0133**] As explained previously, a result of reading in the reverse direction is that a narrower charge trapping region is required due to the higher efficiency of the reverse read. Since erasing is performed through the effective drain region **16** (for

trapped charge **68** and region **14** for trapped charge **70**), less charge needs to be moved off the charge trapping layer **20** and directed through the drain **16** (charge **68**) or effective drain **14** (charge **70**). Thus, reading the memory cell **10** in the reverse direction enables faster erase times. This makes the entire erase process easier. In one bit per cell trapping devices with diffuse charge (FIG. **1**) or localized charge (FIG. **5**A) with traditional read/write memory device (i.e., forward programming/forward read), the charge trapping region **66** (FIG. **5**A) was bigger and wider to achieve the desired change in threshold voltage, thus making the erase process more difficult. To erase the cell **41**, a larger amount of charge spread out over a wider trapping region **66** must be directed through the drain **34**. The danger with this lies in that if the charge trapping region **66** becomes too wide, the cell **41** may never be able to be completely erased. The charge trapping region **66** may become too wide if the device is over_programmed which is a real possibility when programming and reading in the forward direction.

[**0134**] A graph illustrating the effects associated with over programming on the ability to erase in the forward and reverse directions is shown in FIG. **20**. The graph presented in FIG. **20** was constructed using data obtained from a memory cell **10** (FIGS. **5**B and **15**). The top oxide **22** (FIG. **15**), bottom oxide **18** and nitride layer **20** are each 100 Angstroms thick for a total ONO thickness of 300 Angstroms. Programming utilized a V_D of 5.0V and V_G of 10V. Erasing utilized a V_D of 5.0V and a V_G of −8V. Note that programming and erasing are both in the forward direction. Reading, is either in the forward or reverse direction.

[**0135**] In this case, the memory cell, which has been programmed for 100 milliseconds, does not fully erase in a reasonable time (shown in FIG. **20** as 100 milliseconds) with V_T being approximately 7V after 100 milliseconds of erase for reading in both the forward and reverse directions. The cell **10** cannot be erased because it has been over programmed, meaning the charge trapping region was made too wide to effectively erase. After 100 milliseconds of programming, the charge trapping region is very wide. The 13V (V_D of 5V and V_G of −8V) that is applied across the charge trapping region **68** (FIG. **5**B) to erase the trapped charge is effective in removing the electrons that are close to the drain **16**. However, the electrons that are trapped further away from drain **16** towards the middle of the channel cannot be effectively removed because the electric field created by the 13V potential difference between the drain and the gate is weaker at that point.

[**0136**] As is apparent from FIGS. **17** and **20**, the slopes of the threshold voltage V_T versus program time curves for forward read and reverse read (labeled "forward program" and "reverse program" in FIG. **20**) are different. After approximately one millisecond, the forward program curve exhibits a higher slope than the reverse program curve. This shows that reading in the reverse direction is more tolerant to over programming than reading in the forward direction in the sense that a given uncertainty in programming time causes a bigger uncertainty in threshold voltage V_T when reading in the forward direction than when reading in the reverse direction. When reading in the reverse direction, a V_T, of about 4V is reached after approximately 100 microseconds of programming. Even if programming continues up until a millisecond, a factor of 10X, the V_T, for reading in the reverse direction is only approximately 4.5V. For reading in the forward direction, a V_T of 4V is reached only after approximately 7 milliseconds of programming. If programming is off by only 3×, the V_T increases to approximately 8.3V. At this high V_T it is not likely that the device can be erased.

[**0137**] Thus, it is important to stress that, especially for the two-bit per cell localized trapping device (e.g. FIGS. **2**, **15** and others), reading the memory device in the reverse direction does not just enable simpler and faster erasing, but in fact, if the device is to be read in the forward direction and the trapped charge is so adjusted to give the desired threshold voltage V_T, erasing is likely to be not possible at all. This is because much more charge must be trapped on the nitride **20** beneath the gate **24** to achieve a usable difference in threshold voltage V_T between the programmed and the unprogrammed state.

[**0138**] The graph of FIG. **20** also illustrates the effectiveness during erase of the voltage on the drain versus the voltage on the gate. The gate voltage is not as effective due to the distance of the gate from the trapped charge, which includes the thickness' of the top oxide **22** and the nitride layer **20**. The drain voltage is more effective since it is more proximate to the region **68** of trapped charge. However, the gate voltage is more crucial when the width of the trapped charge region **68** is narrow. In this case, the gate voltage will be effective in creating an electric field that covers the entire charge trapping region **68** making the removal of electrons more efficient. As discussed previously, if the device is read in the forward direction, the charge trapping region must be made wide enough to generate a sufficient threshold voltage to differentiate between the programming and the unprogrammed states. Charge trapped far from the drain cannot be compensated for by lowering the voltage on the gate. In addition, the drain voltage cannot be increased beyond approximately 2V due to read disturb. The read disturb refers to slow programming of the bit during read.

[**0139**] A graph illustrating the programming and erasing curves representing the use of oxide versus TEOS as the dielectric on top of the nitride is shown in FIG. **21**. The graph presented in FIG. **21** was constructed using data obtained from two memory cells constructed, one memory cell using TEOS to form the oxide layer **22** (FIG. **15**) on top of the nitride and the other memory cell using thermal oxidation of the nitride to form the top oxide layer **22**, which may also enhance or adulterate the nitride with oxygen from the oxide layer. The thickness' of the top oxide layer **22**, bottom oxide layer **18** and nitride layer **20** are 70, 100, 80 Angstroms, respectively. The width/length ratio for each memory cell channel is 0.6/0.65 microns. Programming (which is done in the forward direction) utilized a V_D, of 5.0V and a V_G of 10V. Erasing (which is also done in the forward direction) utilized a V_D of 5.0V and a V_G of −6V. This graph shows that there is little difference in the programming and erase characteristics when either oxide or TEOS is placed on top of the nitride.

[**0140**] A graph illustrating erase times for a gate voltage of zero with two different values of drain voltage is shown in FIG. **22**. Both curves were generated by first programming in the forward direction for about 10 microseconds, until the threshold voltage V_T equals about 4V and then erasing in the forward direction. For the upper curve, the gate **24** (FIG. **15**) was grounded and 6.0V applied to the drain **16**. For the lower curve, the gate **24** was grounded and 6.5V applied to the drain. For both curves, the threshold voltage is raised during pro-

gramming from nearly 1.5V to approximately 4V. Erasing then brings the VT back down to approximately 1.7V. Note that the time to erase the charge from the dielectric decreases as the drain voltage increases. The curves show that it takes about 100 seconds with a gate voltage of 6.5V to erase (i.e., remove) sufficient charge from the dielectric to bring the threshold voltage of the device down to about 1.9V and that it takes about 1000 seconds with a gate voltage of 6.0V to achieve the same threshold voltage.

[0141] A graph illustrating the erase curve for two different values of negative gate voltage is shown in FIG. **23**. The graph presented in FIG. **23** was constructed using data obtained from a memory cell constructed with a thickness of each of the top oxide **22** (FIG. **15**), bottom oxide **18** and nitride **20** layers is 100 Angstroms for a total dielectric thickness of 300 Angstroms. The channel width/length ratio is 0.6/0.65 microns. Erasing for the reverse direction utilized a constant V_D of 5.5V and a V_G of −5V versus a V_G of −7.5V. The graph shows that drain and gate voltages on the order of 5V and −5V respectively, are sufficient to enable an effective erase. The graph also shows that lowering V_G to −7.5V is effective to erase the device faster while still retaining a V_G less than 10V.

[0142] Reading the graph in FIG. **10**, one can see that to achieve a V_X equal to approximately 2V in the channel (i.e., the same conditions as the prior art memory device with 3V applied to the gate) when reading in the reverse direction, approximately 4V must be applied to the gate. When, for example, 3V is applied to the gate and the device is read in the reverse direction, only approximately 1.2V is generated in the channel. This is in contrast to reading in the forward direction wherein the potential across the trapped charge region was almost the full potential applied to the drain (i.e., 2V). A discussion of the variation in programming time as a function of various parameters, voltage and temperature is given in a paper entitled "Hot-Electron Injection Into the Oxide in n-Channel MOS Devices," B. Eitan and D. Frohman-Bentchkowsky, IEEE Transactions on Electron Devices, March 1981, incorporated herein by reference.

[0143] An advantage of reading in the opposite direction from programming is that the effect of the lateral electric field next to the charge trapping region is minimized. In addition, the gate voltage can be reduced to further minimize the potential in the channel. In fact, the gate voltage can be set to achieve the desired voltage in the channel. This was described previously, especially with reference to FIG. **10**.

[0144] The area of charge trapping necessary to program memory cell **41** is illustrated in FIG. **24**A and the area of charge trapping necessary to program memory cell **10** is illustrated in FIG. **24**B.

[0145] The erase mechanism is enhanced when the charge trapping region is made as narrow as possible. This allows for much more rapid and thus more efficient erasing of the memory cell.

[0146] Further, utilizing a trapping enhanced or a thinner silicon nitride charge trapping layer than disclosed in the prior art helps to confine the charge trapping region to a region near the drain that is laterally narrower than in the prior art. This improves the retention characteristic of the memory cell. Further, the thinner top and bottom oxide sandwiching the nitride layer helps retain the vertical electric field.

[0147] The voltage V_X in the channel is a function of the gate voltage and the impurity level in the channel. V_X is the voltage in the channel just beneath the edge of the trapped charge region above the channel (FIG. **5**B). A higher gate voltage translates to a higher voltage in the channel. When the device is N channel, the impurity in the channel region before inversion is usually boron. The voltage V_X is generally independent of the boron impurity level over a normal range of values in the forward reading mode, but V_X is dependent on the impurity level in the reverse direction, becoming smaller as the impurity level goes up. Indeed in the reverse direction the voltage V_X in the channel just beneath the edge of the trapped charge region is given by the following expression

$$V_X = V_G - (V_T + \Delta V_T)$$

where V_T is the device threshold voltage for zero substrate bias and ΔV_T is the incremental increase in threshold voltage due to substrate back bias caused by a finite value for V_X when the channel is just inverted.

[0148] Various thicknesses were tried for the second oxide layer **22** in the ONO structure of FIGS. **5**B and **24**B. The following table presents the combinations of thickness' for the ONO layers that were constructed for three embodiments of the memory cell of this invention. Note that all thickness' are in Angstroms in the table below.

Layer	Embodiment #1	Embodiment #2	Embodiment #3
Top Oxide ('O' Layer 22)	150	100	70
Nitride ('N' Layer 20)	50	50	50
Bottom Oxide ('O' Layer 18)	70	70	70
Total Thickness	270	220	190

[0149] The nitride layer **20** retains the stored localized (FIGS. **2**, **5**A, **5**B and others) or non-local (FIG. **1**) charge. By employing the reverse read as opposed to the forward read, the amount of charge required to be retained for a given shift in threshold voltage is reduced by a factor typically of two or more. By making the nitride layer **20** thinner and the top oxide layer **22** thicker, the amount of charge required to be stored on the nitride layer **20** for a given threshold voltage shift is also reduced.

[0150] It is also noted that as the thickness of the top oxide layer **22** increased, the lateral fields associated with the charge stored on the 50 Angstrom thick nitride layer **20** decreased slightly. It is also observed that as the thickness of the bottom oxide layer **18** was made thinner, the erase of the charge stored on the nitride layer **20** becomes easier. For a 70 Angstrom thick bottom oxide layer **18**, the charge stored on the nitride layer **20** is more easily erased than if the bottom oxide layer **18** is 100 Angstroms thick.

[0151] Thus, the conclusion is that the thinner the nitride the better. Nitride layers as thin as 20 Angstroms are believed possible. The thinner nitride reduces the lateral field associated with a given charge stored in the nitride layer and thus reduces the lateral dispersion of the stored charge as a result of the internally generated electric field associated with the stored charge.

[0152] In terms of optimization, three parameters can be varied to give the quickest programming time and the widest

margins. The first parameter is the channel length. A longer channel length, for a given programming time when reading in the reverse direction, increases the distance between the drain and the trapped charge (effectively, the source and drain designations are flipped). This lowers the level of the lateral electric field even lower.

[0153] The second parameter, as described previously, is the gate voltage which can be set to minimize the voltage drop in the channel across the channel region beneath the trapped charge. This further reduces the lateral electric field in the channel beneath the trapped charge. Within limits, the voltage in the channel can be 'dialed in' by varying the voltage on the gate. This allows control over the voltage drop in the channel beneath the region of trapped charge. If the gate voltage is made too low then reading a logic '1', i.e., the unprogrammed state, becomes problematic. The gate voltage for reading a logic '1' must be still high enough to generate inversion in order to produce sufficient read current for each sense amplifier. Thus, a lower limit for the gate voltage is approximately 1V above the threshold voltage. The lower limit for the gate voltage is determined by the maximum time required to sense the channel current which represents one state of the memory cell. For example, for fast access time, the maximum time would be in the range of 10 to 30 nanoseconds while for a mass storage device the maximum access time could be as high as 1 microsecond. The actual gate voltage to achieve these maximum times would depend upon the device structure, the dielectric thickness, the bit line capacitance, the doping concentration in the channel and other parameters associated with the device. An upper limit on the gate voltage is the voltage at which the voltage in the channel just beneath the edge of the region of trapped charge is just below the voltage potential applied to the source terminal during reading in the reverse direction. A too high gate voltage will cause inversion in the channel. Thus, it is not recommended to apply a gate voltage that generates such a high voltage in the channel beneath the edge of the charge trapping region because it defeats the benefits of having a lower potential across the portion of the channel beneath this charge trapping region with the accompanying reduction in leakage current and shortened programming time. In a preferred embodiment of the present invention, the gate voltage used for reading is approximately 3V which represents an optimized tradeoff between programming time and leakage current.

[0154] The third optimization method, previously described, is to vary the boron doping of the channel region under the gate. An increase in the doping concentration results in a higher threshold voltage V_T and a lower voltage generated in the channel. This is due to the reduction in the width of the depletion region formed. Thus, a higher doping concentration permits a higher gate voltage to be applied for the same voltage across the portion of the channel beneath the charge trapping region.

[0155] In addition, an increase in the N_A doping concentration for the same length trapping region will improve the punch through behavior of the device. By varying the level of boron implanted in the channel region, the width of the depletion region under the gate can be varied. An increase in the doping concentration results in a reduction in the width of the depletion region for the same applied gate voltage. The reduction in the width of the depletion region occurs because there is now more fixed charge in the substrate. Thus, varying the doping concentration can be used to limit the length of the pinchoff region under the gate. In addition, the doping concentration can be used to increase or decrease the initial threshold voltage of the device.

[0156] Optimization parameters specific to programming and reading a two bits per gate memory cell will now be described. The optimizations for programming include utilizing a longer minimum effective channel length Leer in order to physically separate the two bits better. In addition, the implant level can be reduced in the channel in order to increase the delta between forward and reverse programming. On the other hand, the implant level can be increased in the channel in order to reduce the impact of the first bit on the programming of the second bit. Thus, the implant level in the channel is a compromise between the forward and reverse delta on the one hand and the programming speed on the other hand.

[0157] The optimizations for reading include lowering the gate voltage in order to enhance the punch through during reading. As described previously, punch through is necessary to program and read the second bit. A lower implant level in the channel serves to increase punch through. Also, a higher drain voltage during read functions to increase punch through. These three optimizations relate to reading in the forward direction, which is equivalent to reading the second bit in the reverse.

[0158] In addition, a lower gate voltage reduces the number of electrons that need to be injected into the charge trapping region. This improves erasing because it eliminates residual charge remaining trapped after erasure. Any residual charge that remains in the charge trapping layer after erasure degrades cycling.

[0159] While the invention has been described with respect to a limited number of embodiments, it will be appreciated that many variations, modifications and other applications of the invention may be made.

What is claimed is:

1. A non-volatile memory cell comprising:

a dielectric charge trapping layer located substantially above a channel junction of the cell, said charge trapping layer being adapted to be charged and discharged more than 100 cycles before degrading beyond an operable state.

2. The method according to claim 1, wherein said charge trapping layer is adapted to be charged and discharged more than 500 cycles before degrading beyond an operable state.

3. The method according to claim 1, wherein said charge trapping layer is adapted to be charged and discharged more than 1500 cycles before degrading beyond an operable state.

4. The method according to claim 1, wherein said charge trapping layer is adapted to be charged and discharged more than 2000 cycles before degrading beyond an operable state.

5. The method according to claim 1, wherein said charge trapping layer is adapted to be charged and discharged more than 10,000 cycles before degrading beyond an operable state.

6. A method of fabricating a non-volatile memory cell comprising:

depositing a dielectric charge trapping layer substantially above a channel junction of the cell, said charge trapping

layer being adapted to be charged and discharged more than 100 cycles before degrading beyond an operable state.

7. The method according to claim 6, wherein dielectric charge trapping is adapted to be charged and discharged more than 500 cycles before degrading beyond an operable state.

8. The method according to claim 6, wherein dielectric charge trapping is adapted to be charged and discharged more than 1500 cycles before degrading beyond an operable state.

9. The method according to claim 6, wherein dielectric charge trapping is adapted to be charged and discharged more than 2000 cycles before degrading beyond an operable state.

10. The method according to claim 6, wherein dielectric charge trapping is Adapted to be charged and discharged more than 10,000 cycles before degrading beyond an operable state.

11. A non volatile memory cell comprising:

A charge trapping layer comprised of an impure silicon based dielectric located substantially above a channel of the cell,

wherein said silicon based dielectric contains an impurity selected from the group consisting of oxygen, boron, carbon and polycrystalline silicon,

wherein said silicon based dielectric is selected from the group consisting of silicon dioxide and silicon nitride.

12. The cell according to claim 11, wherein oxygen is introduced into said dielectric.

13. A non-volatile memory cell comprising:

a charge trapping layer at least partially including silicon nitride, wherein

said silicon nitride includes an annealed nitride portion and at least a portion of which is located substantially above a channel of the cell.

14. A non-volatile memory cell comprising:

a charge trapping layer and at least partially including silicon nitride, wherein said silicon nitride includes an annealed nitride portion and is located substantially below a gate of the cell.

15. A non-volatile memory cell comprising:

two charge trapping regions, wherein each charge trapping region is located in proximity to and above a channel junction of said cell;

wherein a thickness of said two charge trapping regions is approximately 100 angstroms or less.

16. The cell according to claim 12 further comprising an oxide layer between said charge trapping regions and a channel of said cell.

17. The cell according to claim 12, wherein a thickness of said oxide layer is 50 angstroms or greater.

18. The cell according to claim 12, wherein a ratio of oxide thickness to trapping region thickness is 0.6 and 5.

19. A non-volatile memory cell comprising:

two charge trapping regions, wherein each charge trapping region is located below a gate of said cell;

wherein a thickness of said two charge trapping regions is approximately 100 angstroms or less.

20. The cell according to claim 16, further comprising an oxide layer between said charge trapping region and said channel.

21. The cell according to claim 16, wherein a thickness of said oxide layer is 50 angstroms or greater.

22. The cell according to claim 16, wherein a ratio of oxide thickness to trapping layer thickness is between 0.6 and 5.

* * * * *