NON-VOLATILE MEMORY CELL WITH A HYBRID ACCESS TRANSISTOR

FIELD OF THE INVENTION

[0001] The present invention relates generally to integrated circuits (ICs), and more particularly to non-volatile memory devices.

BACKGROUND OF THE INVENTION

[0002] Non-volatile memory devices are able to retain stored data even when the power supply is interrupted. Non-volatile memory devices comprise flash devices which can be programmed using electrical signals. For the memory device to be byte-operational, a 2T cell structure consisting of a memory transistor and a select or access transistor in series can be provided. The memory transistor stores data programmed into the memory cell, while the access transistor selects the memory cell to be programmed or erased.

[0003] The memory transistor can be of various types including, for example, floating gate type, metal-nitride-oxide-silicon (MNOS) type, silicon-nitride-oxide-silicon (SNOS) type, metal-oxide-nitride-oxide-silicon (MONOS) type, and silicon-oxide-nitride-oxide-silicon (SONOS) type. The SONOS type memory transistor has a stacked gate structure comprising a nitride layer sandwiched between lower and upper oxide layers, and a polysilicon gate layer. The lower oxide layer is a tunnel oxide layer, the nitride layer is a memory or storage layer, and the upper oxide layer is a blocking layer for preventing the loss of stored charge. The charge-trapping ability of the SONOS stack structure allows lower programming and erase voltages to be used. The SONOS memory transistor also comprises source and drain regions formed on either side of the stack structure.

[0004] Programming is typically by hot carrier injection. Programming speed depends on the efficiency of hot carrier generation which, in turn, depends on the programming current.

Programming current is affected by various factors including, for example, channel doping level and gate width of the memory transistor. Erase can be by Fowler-Nordheim (F-N) tunneling.

[0005] Generally, it is desirable to provide a non-volatile memory device having high programming speed and small cell size to allow for high density memories.

SUMMARY OF THE INVENTION

[0006] The present invention relates generally to ICs. In particular, the present invention relates to non-volatile memory devices. In one aspect of the invention, an IC is provided. The IC comprises a thin gate well with channel doping tailored for transistors with thin gate dielectric layers and a non-volatile memory cell formed on the thin gate well. The non-volatile memory cell comprises an access transistor and a storage transistor. The access transistor comprises first, second diffusion regions and an access gate which includes a thick gate dielectric layer. The storage transistor which is coupled in series to the access transistor comprises first, second diffusion regions and a storage gate. The storage gate includes a storage layer.

[0007] In another aspect of the invention, a method of fabricating an integrated circuit is disclosed. The method comprises providing a substrate with a cell region defined thereon. The cell region is prepared with a thin gate well tailored for transistors with thin gate layers. A non-volatile memory cell having an access transistor and a storage transistor is formed in the cell region. The access transistor includes an access gate with a thick gate dielectric layer.

[0008] These and other objects, along with advantages and features of the present invention herein disclosed, will become apparent through reference to the following description and the accompanying drawings. Furthermore, it is to be understood that the features of the various embodiments described herein are not mutually exclusive and can exist in various combinations and permutations.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] In the drawings, like reference characters generally refer to the same parts throughout the different views. Also, the drawings are not necessarily to scale, emphasis instead generally being placed upon illustrating the principles of the invention. Various embodiments of the present invention are described with reference to the following drawings, in which:

[0010] Fig. 1 shows a non-volatile memory cell in accordance with one embodiment of the invention;

[0011] Fig. 2 shows a non-volatile memory array in accordance with one embodiment of the invention;

[0012] Fig. 3 shows a cross-sectional view of a non-volatile memory cell in accordance with one embodiment of the invention;

[0013] Fig. 4 shows gate voltage-drain current (V_g-I_d) curves of the hybrid access transistor according to one embodiment of the invention; and

[0014] Figs. 5a-j show a process flow for forming an IC in accordance with one embodiment of the invention.

DETAILED DESCRIPTION OF THE INVENTION

[0015] The present invention relates to non-volatile memory cells. More particularly, the present invention relates to a 2T flash type non-volatile memory device which increases programming speed while reducing cell size, enabling high density memories to be fabricated. The non-volatile memory cells can be incorporated into ICs and easily integrated into current CMOS processing technologies. The ICs can be any type of IC, for example non-volatile memories, signal processors, or system on chip devices. Other types of ICs are also useful. Such ICs are incorporated

in, for example, microcontrollers, communication systems, consumer products such as cell phones and memory cards.

[0016] Fig. 1 shows a non-volatile memory cell 110 in accordance with one embodiment of the invention. As shown, the memory cell comprises first and second series coupled transistors 120 and 140. The first transistor serves as an access transistor and the second transistor serves as a storage transistor. The transistors, in one embodiment, comprise transistors of a first polarity type. For example, the transistors comprise p-type transistors. N-type transistors are also useful. Alternatively, both n and p-type transistors are used. The first transistor includes a gate terminal 125 and first and second terminals 122 and 123; the second transistor includes a gate terminal 145 and first and second terminals 142 and 143. The gate terminal of the first transistor is coupled to an access gate line (AG) while the gate terminal of the second transistor is coupled to a control gate line (CG). The second terminals of the transistors are commonly coupled. Source and bit lines, SL and BL, are respectively coupled to the first terminals of the access and storage transistors.

[0017] In accordance with one embodiment of the invention, the access transistor comprises a hybrid transistor. The hybrid transistor includes a thick gate oxide 128 with a doped well 115 to accommodate a transistor with a thin gate oxide. The doped well comprises second polarity type dopants. For example, the doped well comprises a n-well for p-type transistors. As for the storage transistor, it includes a composite charge storing layer 148. In one embodiment, the composite charge storage layer comprises an oxide-nitride-oxide (ONO) stack. Other types of storage layers, such as nanocrystal embedded in oxide and oxide-metal(high-K)-oxide stack are also useful.

[0018] As shown in Fig. 2, a plurality of memory cells 110 can be provided to form a memory array 205. As shown, the memory cells are interconnected by control gate lines 280_{1-i}, access gate lines 282_{1-i}, bit lines 286_{1-i}, and source lines 288_{1-i}. Generally, the control and access gate lines are

arranged in a first direction while the source and bit lines are arranged in a second direction. Other configurations are also useful. Logic circuitry (not shown) can be included to program and access the memory array.

[0019] Fig. 3 shows a cross-sectional view of a memory cell 310 in accordance with one embodiment of the invention. The memory cell is formed on a substrate 311. The substrate, for example, comprises a lightly doped p-type silicon substrate. Other types of substrates are also useful. The substrate includes a cell region 306 for the memory cell. Isolation regions (not shown) are provided to isolate the cell region from other cells or devices. The isolation regions comprise, for example, shallow trench isolation regions. Other types of isolations are also useful. Typically, numerous cell regions are provided in an array region on a substrate. Logic devices, such as input/output (I/O) circuitry and control circuitry can be provided in a peripheral or logic region (not shown) on the substrate.

[0020] The cell region includes first and second transistors 320 and 340, with the first transistor serving as an access transistor and the second transistor serving as a storage transistor. The transistors, in one embodiment, comprise transistors of a first polarity type. In one embodiment, the transistors comprise p-type transistors. N-type transistors are also useful. Each transistor includes first and second diffusion regions (322, 323, 342, 343) adjacent to a gate stack (325, 345). The diffusion regions are formed by first polarity type dopants. Preferably, the transistors share a common second diffusion region 323/343. Contacts, such as silicide contacts, can be provided for the diffusion regions. A doped well 315 is provided beneath the substrate in the cell region. The doped well comprises a second polarity type doped well.

[0021] Disposed on the substrate in the cell region is the gate stack of the access transistor. The gate stack includes a gate layer 327 over a gate dielectric layer 328. The gate layer comprises, for

example, polysilicon. As for the gate dielectric layer, it can comprise various types of dielectric materials. Preferably, the gate dielectric comprises thermal oxide. In one embodiment, the gate dielectric layer comprises a thick gate dielectric layer, with a physical thickness typically about 30-80 Å. Other thicknesses may also be useful. In accordance with one embodiment of the invention, the doped well on which the access transistor is formed comprises a doped well tailored for a transistor with a thin gate dielectric layer (thin gate well), providing a hybrid access transistor. The thin gate doped well, for example, comprises a dopant concentration of about E^{17}- E^{19} cm^{-3}, and a depth of about 0.7–1.4 μm. Other concentrations and depths are also useful.

[0022] On the surface of the substrate in the cell region over the doped well is also disposed the gate stack of the storage transistor. Preferably both the access and storage transistors are formed in the same well. Forming the transistors in a common well region enables the memory cell to be reduced in size. The gate stack comprises a gate layer 347 over a storage dielectric layer 348. Typically, the gate layer comprises polysilicon. Other types of materials can also be used to form the gate layer.

[0023] In one embodiment, the storage layer comprises an oxide-nitride-oxide (ONO) trilayer, forming a SONOS gate stack structure. The ONO trilayer includes a lower (tunneling) oxide layer 331, an upper (blocking) oxide layer 333, and a nitride layer 332 sandwiched therebetween. Other dielectric materials capable of storing electrical charges may also be used. Additionally, the storage layer may comprise one, two or more layers. For example, a bilayer storage layer comprising a nitride layer and a tunneling oxide layer can be used. To facilitate tunneling and trapping of electrical charges for efficient programming and erasing of the non-volatile memory, the lower oxide layer is preferably about 24-45 Å physically thick, the nitride layer is preferably about 40-80 Å thick, and the upper oxide layer is preferably about 40-80 Å thick. More preferably, the lower

oxide layer, nitride layer, and upper oxide layer are, respectively, about 35 Å, 45 Å and 45 Å thick. Other types of flash memories can also be used.

[0024] Although not shown, an interlevel dielectric layer is provided over the memory cell. Contacts are provided, coupling the diffusion regions and gate electrodes to source lines, bit lines, access gate lines and control gate lines.

[0025] As described, a non-volatile memory cell includes a hybrid access transistor with a thick gate dielectric layer formed on a thin gate well. Thin gate wells have higher channel doping compared to thick gate wells (wells tailored for transistors with thick gate dielectric layers), resulting in higher carrier generation efficiency. Furthermore, the thick gate dielectric layer of the hybrid access transistor allows higher bias voltages to be used. Higher bias voltages result in increased programming currents, which improve programming speed.

[0026] Preferably, the gate length of the storage transistor is reduced to increase tunneling current for higher programming speed. Furthermore, the gate length of the access transistor is reduced to be the thin gate channel length, resulting in smaller memory cell sizes. The gates, however, must be of sufficient length to prevent punch-through. Since immunity to punch-through increases with increased channel doping concentration, the use of thin gate wells allows transistors to have smaller gate lengths than thick gate wells. The gate lengths of the transistors, in a preferred embodiment, are selected to be at the minimum critical dimension (CD) for a specific process. More preferably, the minimum CD of each type of transistor is used. For example, for 0.18 μm technology, the gate length of the access transistor is about 0.18 μm while that of the storage transistor is about 0.17 μm.

[0027] The memory cell can operate in various modes, such as programming, reading and erase modes. The erase mode is initiated by applying appropriate bias conditions. In one embodiment,

the erase mode is initiated by applying about 3.3 V to the gate of the access transistor, about 6 V to the thin gate n-well, and about -6.25 V to the gate of the storage transistor. Under such bias condition, data erasing is performed using a F-N tunneling current that flows through the tunneling oxide, drawing electrons from the ONO stack of the non-volatile memory transistor to the channel region. The bias voltage (V_{gb}) during erase, which is equal to the difference between the access transistor operating voltage and the n-well voltage, is about -2.7 V. Operation in this low V_{gb} region avoids the problems of oxide breakdown, trap-up or transistor degradation associated with high electric field stressing.

[0028] The use of a thick gate oxide in the access transistor allows the access transistor to operate at 3.3 V. Conversely, if a thin gate oxide were used, the access transistor can only operate at a maximum of 1.8 V if gate oxide rupture is to be avoided. In such case, V_{gb} is about -4.2 V in the erase mode since 6 V is applied to the thin gate n-well. Such a high bias voltage is liable to cause gate breakdown and reliability problems. Accordingly, the use of a hybrid access transistor is thus able to provide favorable bias conditions.

[0029] Fig. 4 shows measured gate voltage-drain current (V_g-I_d) curves of the hybrid access transistor according to one embodiment of the invention. The width, gate length, and gate oxide physical thickness of the transistor are 0.42 μm, 0.18 μm and 53 Å, respectively. V_g-I_d curves at different thick gate bias voltage (V_{gs}) values are shown. As shown, the hybrid transistor works normally at a gate bias of –3.3 V, where 3.3 V is a standard I/O voltage for 0.18 μm 1.8/3.3 V dual gate process technology. Thus, the hybrid transistor can be used as a thick gate transistor even if the well beneath is a thin gate well and gate length is a thin gate channel length.

[0030] Figs. 5a-j illustrate an exemplary process flow for forming an IC 500 in accordance with one embodiment of the invention. Referring to Fig. 5a, a substrate 511 is provided. The substrate

comprises a semiconductor substrate, such as silicon. Other types of semiconductor substrates, for example, SiGe, SiGeC, SiC, silicon-on-insulators (SOIs), SiGe-on-insulators (SGOIs), are also useful. In one embodiment, the substrate comprises a lightly doped p-type substrate. As shown, a cell region 506 is provided on the substrate on which a memory cell is formed. The cell region, for example, is a part of an array region with a plurality of cell regions on which memory cells are formed to create a memory array. Typically, the substrate also includes a logic region (not shown) on which logic circuitry is formed.

[0031] The substrate is prepared with a doped well 515 in the cell region. The doped well comprises dopants of the second polarity type. The second polarity type, for example, comprises n-type, forming a n-well. Alternatively, the second polarity type comprises p-type. In accordance with one embodiment of the invention, the doped well comprises a thin gate well having a channel doping concentration of about E^{17}- E^{19} cm^{-3}. Conventional ion implantation techniques, such as implantation with a mask can be used to form the wells. Generally, the substrate includes first and second type wells in the logic region to form, for example, CMOS devices. Separate ion implantation processes are used for different types of wells. The doped well in the cell region can be formed along with the same type of wells in the logic regions. Threshold voltage (V_T) implants are performed after well formation.

[0032] Isolation regions (not shown), such as STIs, are provided to isolate the cell region from other device regions. The STIs can be formed before or after the well is formed. In one embodiment, the STIs are formed before the wells are formed.

[0033] Referring to Fig. 5b, a storage layer is formed on the substrate surface. Typically, the storage layer is formed as a blanket layer on the substrate, covering both the array and logic regions. The storage layer, in one embodiment, comprises a composite layer stack. In one embodiment, the

composite layer stack comprises an oxide-nitride-oxide (ONO) stack 548. Other types of storage layers are also useful. Conventional techniques can be used to form the ONO stack. In one embodiment, a tunneling oxide layer 531 is first formed. The tunneling oxide layer can be formed by, for example, rapid thermal oxidation (RTO). Other techniques including, for example, low pressure chemical vapor deposition (LPCVD) and thermal oxidation are also useful. The physical thickness of the tunneling oxide layer can be about 24-45 Å, and preferably about 35 Å. Next, a nitride layer 532 is deposited. The thickness of the nitride layer can be about 40-80 Å, and preferably about 45 Å. The nitride layer can be formed by LPCVD or nitridation. Other techniques are also useful. Finally, the top or barrier oxide layer 533 is formed. The thickness of the barrier oxide layer is about 40-80 Å, and preferably about 45 Å. The barrier oxide layer may be formed using the same techniques as for the tunneling oxide layer.

[0034] Referring to Fig. 5c, a photoresist layer 572 is deposited on the substrate. As is well known in the art, an anti-reflective coating (ARC) layer can be provided under the photoresist layer. For purposes of this discussion, reference to photoresist or resist can include an ARC layer. Furthermore, removal of the photoresist or resist can include removal of the ARC layer. However, it is understood that in certain applications, the ARC layer can remain when the photoresist is removed. The photoresist layer is patterned to expose portions of the storage layer which are to be removed. At least the portion of the storage layer in the cell region where the access transistor is formed is exposed. Additionally, portions of the storage layer in the logic region are also exposed. The exposed portions of the storage layer are removed, as shown in Fig. 5d. Removal can be achieved using, for example, an anisotropic etch such as reactive ion etch (RIE). The etch preferably is selective to the substrate. As shown in Fig. 5e, the photoresist layer is removed after etch.

[0035] Referring to Fig. 5f, a thick gate dielectric layer 528 is formed on the substrate. The thick gate dielectric layer comprises, for example, silicon oxide. Other types of dielectric materials may also be useful. Preferably, the gate dielectric layer is formed by thermal oxidation or RTO. Thermal oxidation or RTO selectively forms the gate oxide on the exposed substrate. This is because in the storage transistor region, the thick gate oxide will not be formed since ONO trilayer serves as a hard mask. The physical thickness of the thick gate oxide is about 30-80 Å, and preferably about 46-60 Å. Other thicknesses may also be useful. Since the logic region is also exposed, the thick gate dielectric layer is also formed in the logic region.

[0036] In one embodiment, the array region is masked off for processing of the logic region. For example, portions of the storage layer in the logic region are removed, if applicable, and conventional dual gate processes are performed therein. For example, a thin gate dielectric layer is formed in the logic region. Since the array region is covered by mask layer during these processes, concerns about overlay or gate integrity issues are eliminated. Dual gate processes are described in, for example, U.S. Patent No. 7,029,976, which is incorporated herein by reference.

[0037] Referring to Fig. 5g, a gate layer 538 is formed on the substrate. The gate layer, for example, comprises a blanket gate layer on the substrate which covers both the array and logic regions. In one embodiment, the gate layer comprises a polysilicon gate layer. The polysilicon gate layer can be formed by chemical vapor deposition (CVD). The polysilicon layer may be doped to impart a higher conductivity to the gate. Other materials including, for example, metal or polycide, may also be used in the gate layer. The thickness of the polysilicon gate layer is, for example, about 1800-2200 Å.

[0038] Referring to Fig. 5h, the polysilicon gate layer and the layers beneath it are patterned to form gate stacks of the access and storage transistors in the cell region. Gate stacks in the logic

regions can also be patterned along with the gate stacks in the cell region. Conventional patterning techniques can be used to form the gate stacks. For example, a photoresist layer 574 is deposited on top of the polysilicon gate layer. The photoresist layer is patterned to expose portions of the polysilicon layer. The exposed polysilicon gate layer and the layers beneath it are removed to form the first and second gate stacks 525 and 545, as shown in Fig. 5i. Removal can be achieved using, for example, an anisotropic etch such as RIE. The etch is preferably selective to the substrate. The remaining portions of the photoresist layer on the top of the gate stacks are removed after etching.

[0039] Referring to Fig. 5j, first and second diffusion regions (522, 523, 542, 543) of the first polarity type are formed on the substrate adjacent to the gate stacks, creating first polarity type transistors. The diffusion regions comprise doped regions in the well. The first polarity type, for example, comprises p-type, resulting in p-type diffusion regions in the n-well. In one embodiment, the transistors share a common second diffusion region 523/543 located between the gate stacks. The diffusion regions, for example, are formed using conventional techniques, such as ion implantation. Generally, first and second type diffusion regions are formed in the logic region. In one embodiment, the diffusion regions in the array region can be formed along with the first type diffusion regions in the logic region.

[0040] The process continues by forming interconnections to the diffusion regions and gates of the transistors. For example, the diffusion regions and gates are appropriately coupled to source lines, bit lines, access lines and control lines. Additionally, the thin gate well is coupled to a bias voltage. Appropriate interconnections are also formed in the logic region. Additional processes are performed to complete the IC. These processes include, for example, final passivation, dicing, and packaging.

[0041] As described, the memory cell of the present invention can be easily integrated into conventional CMOS processes. By forming and patterning the thick gate dielectric layer in the logic and cell regions concurrently, the additional gate steps are reduced or minimized. No additional implant processes are required to form the hybrid access transistors. By providing a memory cell design which facilitates simple circuit routing along with minimum CD gate lengths, cell size is reduced and density is increased. Furthermore, masking off the array region during dual gate processes eliminates overlay or gate integrity issues in the array region. In addition, the thick gate oxide of the access transistor allows for appropriate bias conditions to increase programming speed reliably. Therefore, the present invention provides a simple, reliable, high performing and low cost solution for integrating non-volatile memories in ICs.

[0042] The invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. The foregoing embodiments, therefore, are to be considered in all respects illustrative rather than limiting the invention described herein. Scope of the invention is thus indicated by the appended claims, rather than by the foregoing description, and all changes that come within the meaning and range of equivalency of the claims are intended to be embraced therein.

CLAIMS

What is claimed is:

1. An integrated circuit (IC) comprising:

a thin gate well; and

a non-volatile memory cell formed on the thin gate well, the non-volatile memory cell comprises

an access transistor having an access gate and first and second diffusion regions, the access gate includes a thick gate dielectric layer; and

a storage transistor coupled in series to the access transistor, the storage transistor has a storage gate and first and second diffusion regions, the storage gate includes a storage layer,

wherein the thin gate well has channel doping tailored for transistors with thin gate dielectric layers.

2. The IC of claim 1 wherein the access and storage transistors comprise diffusion regions of a first polarity type and a doped well comprises a second polarity type.

3. The IC of claim 2 wherein the doped well is coupled to a bias voltage.

4. The IC of claim 2 wherein the first polarity type comprises n-type and the second polarity type comprises p-type.

5. The IC of claim 2 wherein the first polarity type comprises p-type and the second polarity type comprises n-type.

6. The IC of claim 1 wherein the storage layer comprises oxide-nitride-oxide stack.

7. The IC of claim 6 wherein the access and storage transistors comprise diffusion regions of a first polarity type and a doped well comprises a second polarity type.

8. The IC of claim 7 wherein the first polarity type comprises n-type or p-type.

9. The IC of claim 6 wherein the channel doping of the thin gate well comprises E^{17} to E^{19} cm^{-3}.

10. The IC of claim 9 wherein the access and storage transistors comprise diffusion regions of a first polarity type and a doped well comprises a second polarity type.

11. The IC of claim 10 wherein the first polarity type comprises n-type and the second polarity type comprises p-type or the first polarity type comprises p-type and the second polarity type comprises n-type.

12. The IC of claim 1 wherein the channel doping of the thin gate well comprises E^{17} to E^{19} cm^{-3}.

13. The IC of claim 12 wherein the access and storage transistors comprise diffusion regions of a first polarity type and a doped well comprises a second polarity type.

14. The IC of claim 13 wherein the first polarity type comprises n-type and the second polarity type comprises p-type or the first polarity type comprises p-type and the second polarity type comprises n-type.

15. The IC of claim 1 wherein the gate of the access transistor is coupled to an access gate line, a first terminal of the access transistor is coupled to a source line, the gate of the storage transistor is coupled to a control gate line and a first terminal of the storage transistor is coupled to a bit line.

16. The IC of claim 15 comprises a non-volatile memory array having a plurality of non-volatile memory cells interconnected by source lines, bit lines, access gate lines and control gate lines.

17. The IC of claim 15 wherein the second diffusion regions of the access transistor and the storage transistor are commonly coupled.

18. The IC of claim 15 wherein the second diffusion regions of the access transistor and the storage transistor form a common diffusion region.

19. A method of fabricating an integrated circuit (IC) comprising the steps of:

providing a substrate with a cell region defined thereon, the cell region is prepared with a thin gate well tailored for transistors with thin gate dielectric layers; and

forming a non-volatile memory cell in the cell region, the non-volatile memory cell having an access transistor and a storage transistor, wherein the access transistor includes an access gate with a thick gate dielectric layer.