US 20090179253A1

(19) United States

(12) Patent Application Publication

Levy et al.

(10) Pub. No.: US 2009/0179253 A1

(43) Pub. Date: Jul. 16, 2009

(54) OXIDE-NITRIDE-OXIDE STACK HAVING MULTIPLE OXYNITRIDE LAYERS

(75) Inventors: **Sagy Levy**, Sunnyvale, CA (US); **Krishnaswamy Ramkumar**, San Jose, CA (US); **Fredrick Jenne**, Sunnyvale, CA (US); **Sam Geha**, Cupertino, CA (US)

Correspondence Address:
CYPRESS SEMICONDUCTOR CORPORATION
198 CHAMPION COURT
SAN JOSE, CA 95134-1709 (US)

(73) Assignee: **Cypress Semiconductor Corporation**

(21) Appl. No.: **11/811,958**

(22) Filed: **Jun. 13, 2007**

Related U.S. Application Data

(60) Provisional application No. 60/931,947, filed on May 25, 2007.

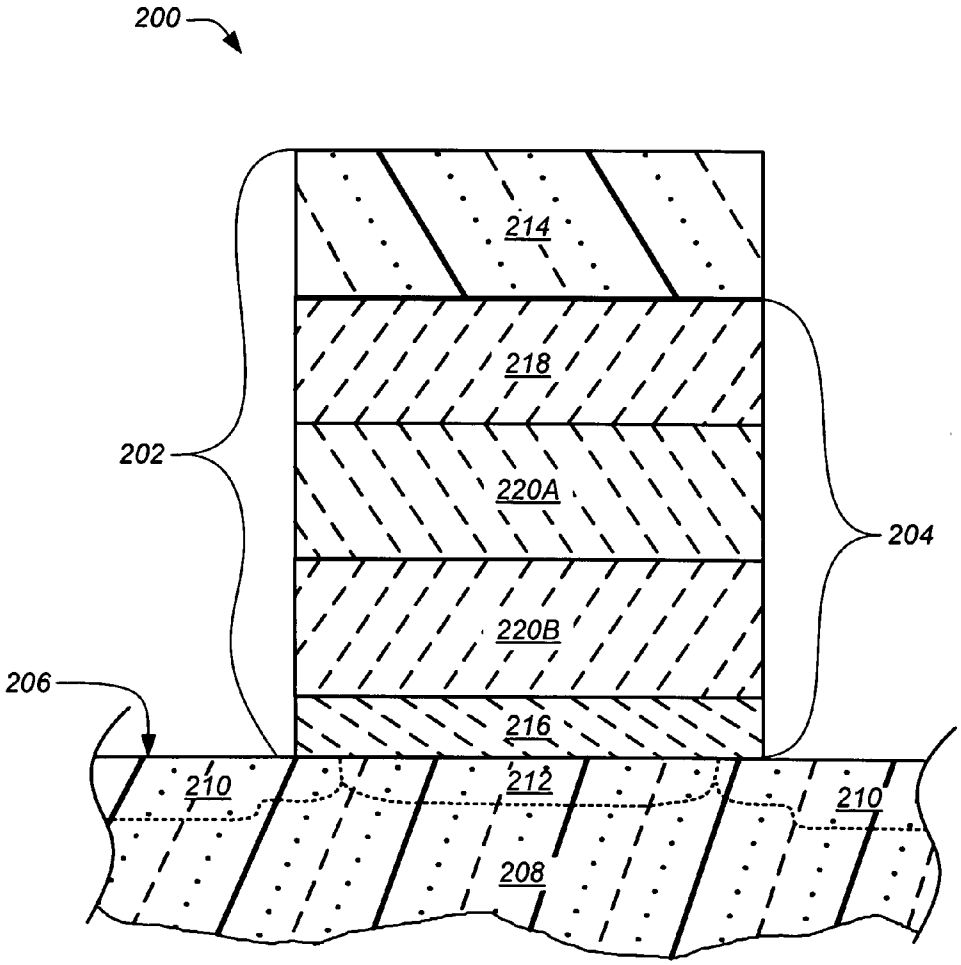
Publication Classification

(51) **Int. Cl.**
H01L 29/792 (2006.01)
H01L 21/336 (2006.01)

(52) **U.S. Cl.** .. **257/324**; 438/257; 438/770; 257/E29.309; 257/E21.209

(57) **ABSTRACT**

A semiconductor device including an oxide-nitride-oxide (ONO) structure having a multi-layer charge storing layer and methods of forming the same are provided. Generally, the method involves: (i) forming a first oxide layer of the ONO structure; (ii) forming a multi-layer charge storing layer comprising nitride on a surface of the first oxide layer; and (iii) forming a second oxide layer of the ONO structure on a surface of the multi-layer charge storing layer. Preferably, the charge storing layer comprises at least two silicon oxynitride layers having differing stoichiometric compositions of Oxygen, Nitrogen and/or Silicon. More preferably, the ONO structure is part of a silicon-oxide-nitride-oxide-silicon (SONOS) structure and the semiconductor device is a SONOS memory transistor. Other embodiments are also disclosed.



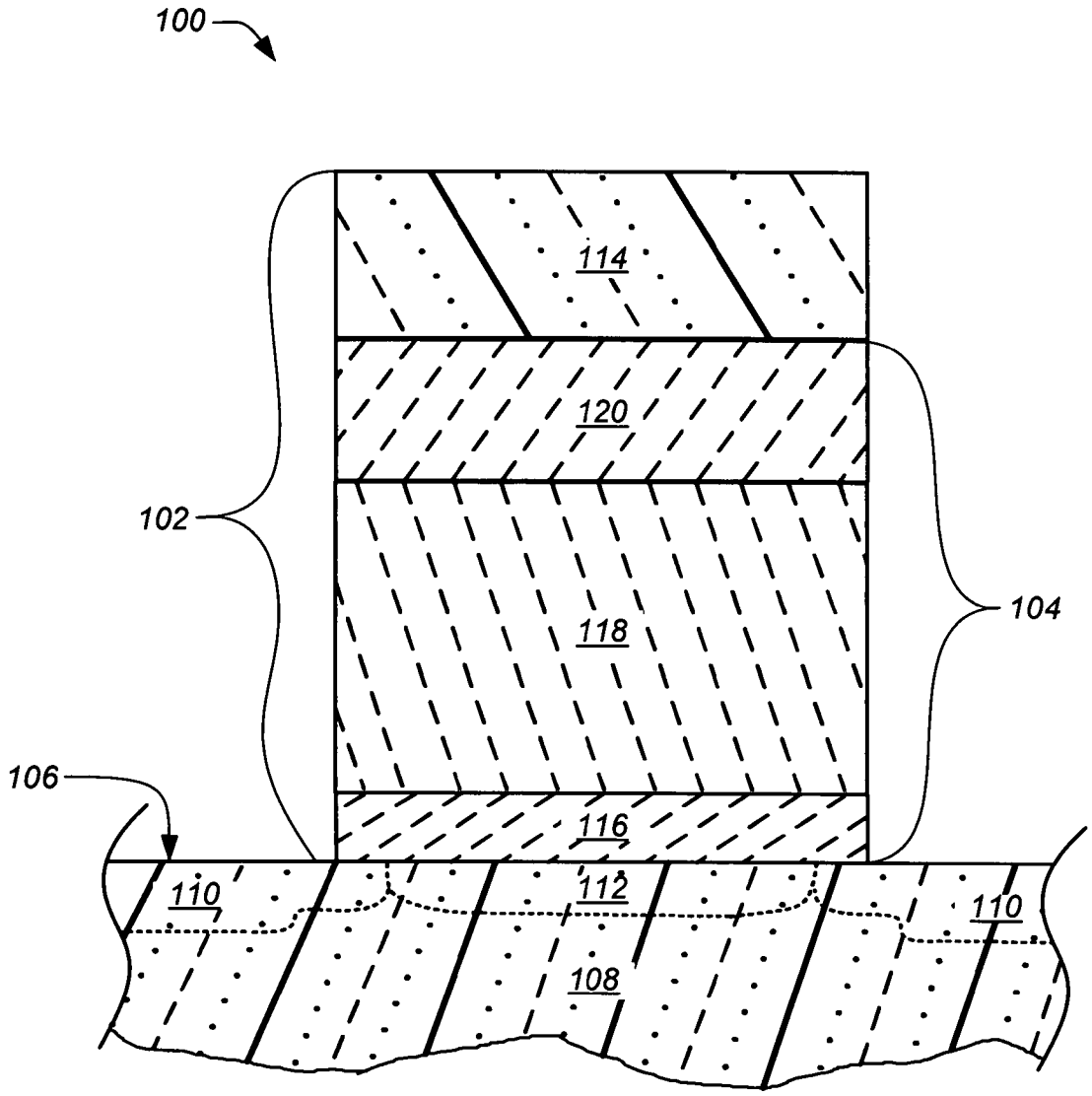


FIG. 1 (Prior Art)

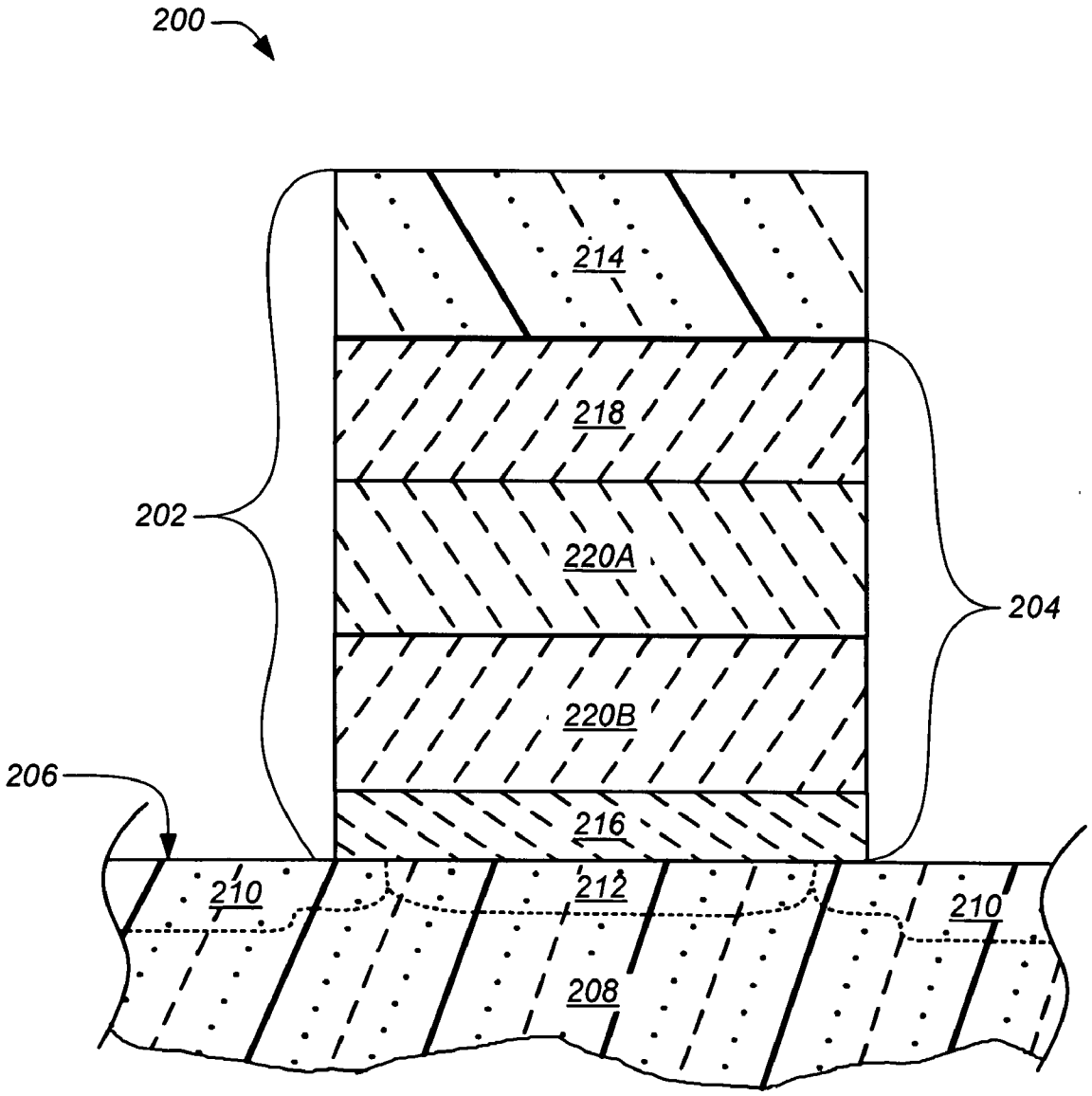


FIG. 2

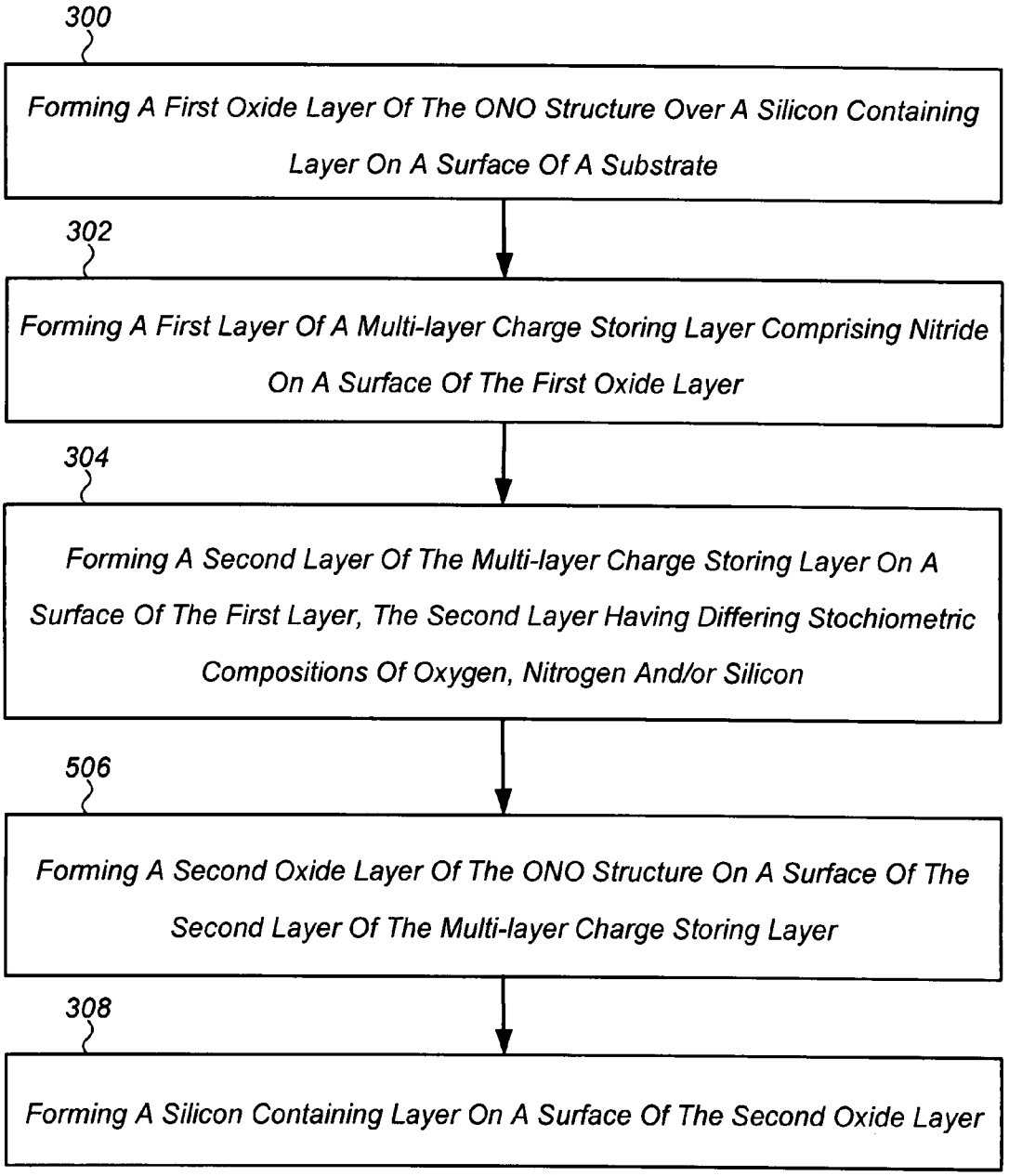

300
Forming A First Oxide Layer Of The ONO Structure Over A Silicon Containing Layer On A Surface Of A Substrate
302
Forming A First Layer Of A Multi-layer Charge Storing Layer Comprising Nitride On A Surface Of The First Oxide Layer
304
Forming A Second Layer Of The Multi-layer Charge Storing Layer On A Surface Of The First Layer, The Second Layer Having Differing Stochiometric Compositions Of Oxygen, Nitrogen And/or Silicon
506
Forming A Second Oxide Layer Of The ONO Structure On A Surface Of The Second Layer Of The Multi-layer Charge Storing Layer
308
Forming A Silicon Containing Layer On A Surface Of The Second Oxide Layer

FIG. 3

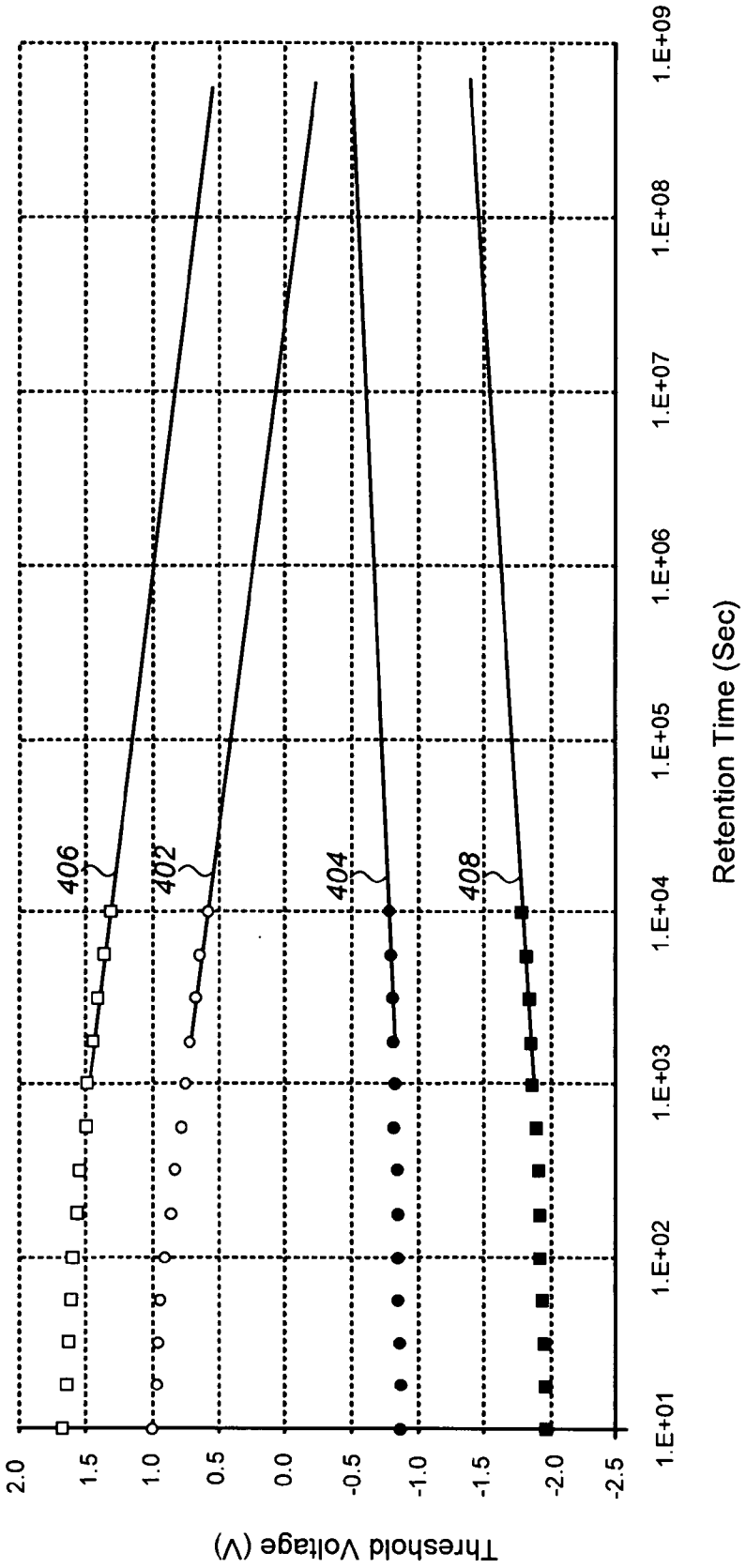

Threshold Voltage (V)
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
1.E+01
1.E+02
1.E+03
1.E+04
1.E+05
1.E+06
1.E+07
1.E+08
1.E+09
Retention Time (Sec)
406
402
404
408

FIG. 4

OXIDE-NITRIDE-OXIDE STACK HAVING MULTIPLE OXYNITRIDE LAYERS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims the benefit of priority under 35 U.S.C. 119(e) to U.S. Provisional Patent Application Ser. No. 60/931,947, entitled Oxide-Nitride-Oxide Stack Having Multiple Oxynitride Layers; which application is hereby incorporated by reference.

TECHNICAL FIELD

[0002] This invention relates to semiconductor processing and, more particularly to an oxide-nitride-oxide stack having an improved oxide-nitride or oxynitride layer and methods of forming the same.

BACKGROUND OF THE INVENTION

[0003] Non-volatile semiconductor memories, such as a split gate flash memory, typically use a stacked floating gate type field effect transistors, in which electrons are induced into a floating gate of a memory cell to be programmed by biasing a control gate and grounding a body region of a substrate on which the memory cell is formed.

[0004] An oxide-nitride-oxide (ONO) stack is used as either a charge storing layer, as in silicon-oxide-nitride-oxide-silicon (SONOS) transistor, or as an isolation layer between the floating gate and control gate, as in a split gate flash memory.

[0005] FIG. **1** is a partial cross-sectional view of an intermediate structure for a semiconductor device **100** having a SONOS gate stack or structure **102** including a conventional ONO stack **104** formed over a surface **106** of a silicon substrate **108** according to a conventional method. In addition, the device **100** typically further includes one or more diffusion regions **110**, such as source and drain regions, aligned to the gate stack and separated by a channel region **112**. Briefly, the SONOS structure **102** includes a poly-silicon (poly) gate layer **114** formed upon and in contact with the ONO stack **104**. The poly gate **114** is separated or electrically isolated from the substrate **108** by the ONO stack **104**. The ONO stack **104** generally includes a lower oxide layer **116**, a nitride or oxynitride layer **118** which serves as a charge storing or memory layer for the device **100**, and a top, high-temperature oxide (HTO) layer **120** overlying the nitride or oxynitride layer.

[0006] One problem with conventional SONOS structures **102** and methods of forming the same is the poor data retention of the nitride or oxynitride layer **118** that limits the device **100** lifetime and/or its use in several applications due to leakage current through the layer.

[0007] Another problem with conventional SONOS structures **102** and methods of forming the same is the stochiometry of the oxynitride layer **118** is neither uniform nor optimized across the thickness of the layer. In particular, the oxynitride layer **118** is conventionally formed or deposited in a single step using a single process gas mixture and fixed or constant processing conditions in an attempt to provide a homogeneous layer having a high nitrogen and high oxygen concentration across the thickness of the relatively thick layer. However, due to top and bottom effects this results in nitrogen, oxygen and silicon concentrations, which can vary throughout the conventional oxynitride layer **118**. The top effect is caused by the order in which process gases are shut off following deposition. In particular, the silicon containing process gas, such as silane, is typically shut off first resulting in a top portion of the oxynitride layer **118** that is high in oxygen and/or nitride and low in silicon. Similarly, the bottom effect is caused by the order in which process gases are introduced to initiate deposition. In particular, the deposition of the oxynitride layer **118** typically follows an annealing step, resulting in a peak or relatively high concentration of ammonia (NH_3) at the beginning of the deposition process and producing in a bottom portion of the oxynitride layer that is low in oxygen and silicon and high in nitrogen. The bottom effect is also due to surface nucleation phenomena in which that oxygen and silicon that is available in the initial process gas mixture preferentially reacts with silicon at the surface of the substrate and does not contribute to the formation of the oxynitride layer. Consequently, the charge storage characteristics, and in particular programming and erase speed and data retention of a memory device **100** made with the ONO stack **104**, are adversely effected.

[0008] Accordingly, there is a need for a memory device having an ONO stack with an oxynitride layer as a memory layer that exhibits improved programming and erase speed and data retention. There is a further need for a method or process of forming an ONO stack having an oxynitride layer that exhibits improved oxynitride stochiometry.

[0009] The present invention provides a solution to these and other problems, and offers further advantages over conventional ONO stacks or memory layers and methods of forming the same.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] These and various other features and advantages of the present invention will be apparent upon reading of the following detailed description in conjunction with the accompanying drawings and the appended claims provided below, where:

[0011] FIG. **1** (prior art) is a block diagram illustrating a cross-sectional side view of an intermediate structure for a memory device for which a method having an oxide-nitride-oxide (ONO) stack formed according to conventional method;

[0012] FIG. **2** is a block diagram illustrating a cross-sectional side view of a portion of a semiconductor device having an ONO structure including a multi-layer charge storing layer according to an embodiment of the present invention;

[0013] FIG. **3** is flow chart of a method for forming an ONO structure including a multi-layer charge storing layer according to an embodiment of the present invention; and

[0014] FIG. **4** is a graph showing an improvement in data retention for a memory device using a memory layer formed according to the present invention as compared to a memory device using a conventional memory layer.

DETAILED DESCRIPTION

[0015] The present invention is directed generally to an oxide-nitride-oxide (ONO) structure including a multi-layer charge storing layer and methods for making the same. The ONO structure and method are particularly useful for forming a memory layer in a memory device, such as a silicon-oxide-nitride-oxide-silicon (SONOS) memory transistor.

[0016] In the following description, for purposes of explanation, numerous specific details are set forth in order to

provide a thorough understanding of the present invention. It will be evident, however, to one skilled in the art that the present invention may be practiced without these specific details. In other instances, well-known structures, and techniques are not shown in detail or are shown in block diagram form in order to avoid unnecessarily obscuring an understanding of this description.

[0017] Reference in the description to "one embodiment" or "an embodiment" means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the invention. The appearances of the phrase "in one embodiment" in various places in the specification do not necessarily all refer to the same embodiment. The term "to couple" as used herein may include both to directly connect and to indirectly connect through one or more intervening components.

[0018] Briefly, the method involves forming a multi-layer charge storing layer including multiple oxynitride layers, such as silicon oxynitride (Si_2N_2O) layers, having differing concentrations of Oxygen, Nitrogen and/or Silicon. Generally, the oxynitride layers are formed at higher temperatures than nitride or oxynitride layers in conventional ONO structures, and each of the layers are formed using differing process gases mixtures and/or at differing flow rates. Preferably, the oxynitride layers include at least a top oxynitride layer and a bottom oxynitride layer. More preferably, the stoichiometric compositions of the layers is tailored or selected such that the lower or bottom oxynitride has a high oxygen and silicon content, and the top oxynitride layer has high silicon and a high nitrogen concentration with a low oxygen concentration to produce a silicon-rich nitride or oxynitride. The silicon-rich and oxygen-rich bottom oxynitride layer reduces stored charge loss without compromising device speed or an initial (beginning of life) difference between program and erase voltages. The silicon-rich, oxygen-lean top oxynitride layer increases a difference between programming and erase voltages of memory devices, thereby improving device speed, increasing data retention, and extending the operating life of the device.

[0019] Optionally, the ratio of thicknesses between the top oxynitride layer and the bottom oxynitride layer can be selected to facilitate forming of the oxynitride layers over a first oxide layer of an ONO structure following the step of forming the first oxide layer using a steam anneal.

[0020] An ONO structure and methods for fabricating the same according to various embodiments of the present invention will now be described in greater detail with reference to FIGS. **2** through **4**.

[0021] FIG. **2** is a block diagram illustrating a cross-sectional side view of a portion of a semiconductor memory device **200** having an ONO structure including a multi-layer charge storing layer according to one embodiment of the present invention. Referring to FIG. **2**, the memory device **200** includes a SONOS gate stack **202** including an ONO structure **204** formed over a surface **206** of silicon layer on a substrate or a silicon substrate **208**. In addition, the device **200** further includes one or more diffusion regions **210**, such as source and drain regions, aligned to the gate stack **202** and separated by a channel region **212**. Generally, the SONOS structure **202** includes a poly-silicon or poly gate layer **214** formed upon and in contact with the ONO structure **204** and a portion of the silicon layer or substrate **208**. The poly gate **214** is separated or electrically isolated from the substrate **208** by the ONO structure **204**. The ONO structure **204** includes a thin, lower oxide layer or tunneling oxide layer **216** that separates or electrically isolates the gate stack **202** from the channel region **212**, a top or blocking oxide layer **218**, and a multi-layer charge storing layer including multiple nitride containing layers. Preferably, as noted above and as shown in FIG. **2**, the multi-layer charge storing layer includes at least two oxynitride layers, including a top oxynitride layer **220**A and a bottom oxynitride layer **220**B.

[0022] Generally, the substrate **208** may include any known silicon-based semiconductor material including silicon, silicon-germanium, silicon-on-insulator, or silicon-on-sapphire substrate. Alternatively, the substrate **208** may include a silicon layer formed on a non-silicon-based semiconductor material, such as gallium-arsenide, germanium, gallium-nitride, or aluminum-phosphide. Preferably, the substrate **208** is a doped or undoped silicon substrate.

[0023] The lower oxide layer or tunneling oxide layer **216** of the ONO structure **204** generally includes a relatively thin layer of silicon dioxide (SiO_2) of from about 15 angstrom (Å) to about 22 Å, and more preferably about 18 Å. The tunneling oxide layer **216** can be formed or deposited by any suitable means including, for example, being thermally grown or deposited using chemical vapor deposition (CVD). In a preferred embodiment, the tunnel oxide layer is formed or grown using a steam anneal. Generally, the process involves a wet-oxidizing method in which the substrate **208** is placed in a in a deposition or processing chamber, heated to a temperature from about 700° C. to about 850° C., and exposed to a wet vapor for a predetermined period of time selected based on a desired thickness of the finished tunneling oxide layer **216**. Exemplary process times are from about 5 to about 20 minutes. The oxidation can be performed at atmospheric or at low pressure.

[0024] As noted above, the multi-layer charge storing layer generally includes at least two oxynitride layers having differing compositions of silicon, oxygen and nitrogen, and can have an overall thickness of from about 70 Å to about 150 Å, and more preferably about 100 Å. In a preferred embodiment the oxynitride layers are formed or deposited in a low pressure CVD process using a silicon source, such as silane (SiH_4), chlorosilane (SiH_3Cl), dichlorosilane (SiH_2Cl_2), tetrachlorosilane ($SiCl_4$) or Bis-TertiaryButylAmino Silane (BTBAS), a nitrogen source, such as nitrogen (N_2), ammonia (NH_3), nitrogen trioxide (NO_3) or nitrous oxide (N_2O), and an oxygen-containing gas, such as oxygen (O_2) or N_2O. Alternatively, gases in which hydrogen has been replaced by deuterium can be used, including, for example, the substitution of deuterated-ammonia (ND_3) for NH_3. The substitution of deuterium for hydrogen advantageously passivates Si dangling bonds at the silicon-oxide interface, thereby increasing an NBTI (Negative Bias Temperature Instability) lifetime of the devices.

[0025] For example, the lower or bottom oxynitride layer **220**B can be deposited over the tunneling oxide layer **216** by placing the substrate **208** in a deposition chamber and introducing a process gas including N_2O, NH_3 and DCS, while maintaining the chamber at a pressure of from about 5 millitorr (mT) to about 500 mT, and maintaining the substrate at a temperature of from about 700° C. to about 850° C. and more preferably at least about 780° C., for a period of from about 2.5 minutes to about 20 minutes. In particular, the process gas can include a first gas mixture of N_2O and NH_3 mixed in a ratio of from about 8:1 to about 1:8 and a second gas mixture of DCS and NH_3 mixed in a ratio of from about 1:7 to about

7:1, and can be introduced at a flow rate of from about 5 to about 200 standard cubic centimeters per minute (sccm). It has been found that an oxynitride layer produced or deposited under these condition yields a silicon-rich, oxygen-rich, bottom oxynitride layer **220**B, that decrease the charge loss rate after programming and after erase, which is manifested in a small voltage shift in the retention mode.

[0026] The top oxynitride layer **220**A can be deposited over the bottom oxynitride layer **220**B in a CVD process using a process gas including N_2O, NH_3 and DCS, at a chamber pressure of from about 5 mT to about 500 mT, and at a substrate temperature of from about 700° C. to about 850° C. and more preferably at least about 780° C., for a period of from about 2.5 minutes to about 20 minutes. In particular, the process gas can include a first gas mixture of N_2O and NH_3 mixed in a ratio of from about 8:1 to about 1:8 and a second gas mixture of DCS and NH_3 mixed in a ratio of from about 1:7 to about 7:1, and can be introduced at a flow rate of from about 5 to about 20 sccm. It has been found that an oxynitride layer produced or deposited under these condition yields a silicon-rich, nitrogen-rich, and oxygen-lean top oxynitride layer **220**A, which improves the speed and increases of the initial difference between program and erase voltage without compromising a charge loss rate of memory devices made using an embodiment of the inventive ONO structure **204**, thereby extending the operating life of the device.

[0027] Preferably, the top oxynitride layer **220**A is deposited sequentially in the same tool used to form the bottom oxynitride layer **220**B, substantially without breaking vacuum on the deposition chamber. More preferably, the top oxynitride layer **220**A is deposited substantially without altering the temperature to which the substrate **208** was heated during deposition of the bottom oxynitride layer **220**B. In one embodiment, the top oxynitride layer **220**A is deposited sequentially and immediately following the deposition of the bottom oxynitride layer **220**B by decreasing the flow rate of the N_2O/NH_3 gas mixture relative to the DCS/NH_3 gas mixture to provide the desired ratio of the gas mixtures to yield the silicon-rich, nitrogen-rich, and oxygen-lean top oxynitride layer **220**A.

[0028] In certain embodiments, another oxide or oxide layer (not shown in these figures) is formed after the formation of the ONO structure **204** in a different area on the substrate or in the device using a steam oxidation. In this embodiment, the top oxynitride layer **220**A and top oxide layer **218** of the ONO structure **204** are beneficially steam annealed during the steam oxidation process. In particular, steam annealing improves the quality of the top oxide layer **218** reducing the number of traps formed near a top surface of the top oxide layer and near a top surface of the underlying top oxynitride layer **220**A, thereby reducing or substantially eliminating an electric field that could otherwise form across the top oxide layer, which could result in back streaming of charge carriers therethrough and adversely affecting data or charge retention in the charge storing layer.

[0029] A suitable thickness for the bottom oxynitride layer **220**B has been found to be from about 10 Å to about 80 Å, and a ratio of thicknesses between the bottom layer and the top oxynitride layer has been found to be from about 1:6 to about 6:1, and more preferably at least about 1:4.

[0030] The top oxide layer **218** of the ONO structure **204** includes a relatively thick layer of SiO_2 of from about 30 Å to about 70 Å, and more preferably about 45 Å. The top oxide layer **218** can be formed or deposited by any suitable means including, for example, being thermally grown or deposited using CVD. In a preferred embodiment, the top oxide layer **218** is a high-temperature-oxide (HTO) deposited using CVD process. Generally, the deposition process involves exposing the substrate **208** to a silicon source, such as silane, chlorosilane, or dichlorosilane, and an oxygen-containing gas, such as O_2 or N_2O in a deposition chamber at a pressure of from about 50 mT to about 1000 mT, for a period of from about 10 minutes to about 120 minutes while maintaining the substrate at a temperature of from about 650° C. to about 850° C.

[0031] Preferably, the top oxide layer **218** is deposited sequentially in the same tool used to form the oxynitride layers **220**A, **220**B. More preferably, the oxynitride layers **220**A, **220**B, and the top oxide layer **218** are formed or deposited in the same tool used to grow the tunneling oxide layer **216**. Suitable tools include, for example, an ONO AVP, commercially available from AVIZA technology of Scotts Valley, Calif.

[0032] A method or forming or fabricating an ONO stack according to one embodiment of the present invention will now be described with reference to the flowchart of FIG. **3**.

[0033] Referring to FIG. **3**, the method begins with forming a first oxide layer, such as a tunneling oxide layer, of the ONO structure over a silicon containing layer on a surface of a substrate (step **300**). Next, the first layer of a multi-layer charge storing layer including nitride is formed on a surface of the first oxide layer (step **302**). As noted above, this first layer or bottom oxynitride layer can be formed or deposited by a CVD process using a process gas including N_2O/NH_3 and DCS/NH_3 gas mixtures in ratios and at flow rates tailored to provide a silicon-rich and oxygen-rich oxynitride layer. The second layer of the multi-layer charge storing layer is then formed on a surface of the first layer (step **304**). The second layer has a stoichiometric composition of oxygen, nitrogen and/or silicon different from that of the first layer. In particular, and as noted above, the second or top oxynitride layer can be formed or deposited by a CVD process using a process gas including DCS/NH_3 and N_2O/NH_3 gas mixtures in ratios and at flow rates tailored to provide a silicon-rich, oxygen-lean top oxynitride layer. Finally, a second oxide layer of the ONO structure is formed on a surface of the second layer of the multi-layer charge storing layer (step **306**). As noted above, this top or blocking oxide layer can be formed or deposited by any suitable means, but is preferably deposited in a CVD process. In one embodiment the top or second oxide layer is a high temperature oxide deposited in a HTO CVD process. Alternatively, the top or blocking oxide layer can be thermally grown, however it will be appreciated that in this embodiment the oxynitride thickness must be adjusted or increased as some of the top oxynitride will be effectively consumed or oxidized during the process of thermally growing the blocking oxide layer.

[0034] Optionally, the method may further include the step of forming or depositing a silicon containing layer on a surface of the second oxide layer to form a SONOS stack or structure (step **308**). The silicon containing layer can be, for example, a polysilicon layer deposited by a CVD process to form a control gate of a SONOS transistor or device.

[0035] A comparison of data retention for a memory device using a memory layer formed according to an embodiment of the present invention as compared to a memory device using a conventional memory layer will now be made with reference to FIG. **4**. In particular, FIG. **4** illustrates the change in threshold voltage of devices in an electronically erasable

programmable read-only memory (EEPROM) during programming (VTP) during erase (VTE) over device life for an EEPROM made using a conventional ONO structure and an ONO structure having a multi-layer oxynitride layer. In gathering data for this figure both devices were pre-cycled for 100K cycles at an ambient temperature of 85° C.

[0036] Referring to FIG. **4**, the graph or line **402** illustrates the change over time of a VTP for an EEPROM made using a conventional ONO structure having a single oxynitride layer without refreshing the memory after the initial writing—program or erase. Actual data points on line **402** are shown by unfilled circles, the remainder of the line showing an extrapolation of VTP to a specified end-of-life (EOL) for the EEPROM. Graph or line **404** illustrates the change over time of a VTE for the EEPROM made using a conventional ONO structure. Actual data points on line **404** are shown by filled circles, and the remainder of the line shows an extrapolation of VTE to EOL for the EEPROM. Generally, the specified difference between the VTE and VTP for an EEPROM at EOL is at least 0.5 V to be able to identify or sense the difference between the program and erase state. As seen from this figure an EEPROM made using a conventional ONO structure has a difference between VTE and VTP of about 0.35V at a specified EOL of 20 years. Thus, an EEPROM made using a conventional ONO structure and operated under the conditions described above will fail to meet the specified operating life by at least about 17 years.

[0037] In contrast, the change in VTP and VTE over time for an EEPROM made using an ONO structure having a multi-layer oxynitride layer, illustrated by lines **406** and **408** respectively, shows a difference between VTE and VTP of at least about 1.96V at the specified EOL. Thus, an EEPROM made using an ONO structure according to an embodiment of the present invention will meet and exceed the specified operating life of 20 years. In particular, graph or line **406** illustrates the change over time of VTP for an EEPROM using an ONO structure according to an embodiment of the present invention. Actual data points on line **406** are shown by unfilled squares, the remainder of the line showing an extrapolation of VTP to the specified EOL. Graph or line **408** illustrates the change over time of VTE for the EEPROM, and actual data points on line **408** are shown by filled squares, the remainder of the line showing an extrapolation of VTE to EOL.

[0038] Although shown and described above as having only two oxynitride layer, i.e., a top and a bottom layer, the present invention is not so limited, and the multi-layer charge storing layer can include any number, n, of oxynitride layers, any or all of which may have differing stoichiometric compositions of oxygen, nitrogen and/or silicon. In particular, multi-layer charge storing layers having up to five oxynitride layers each with differing stoichiometric compositions have been produced and tested. However, as will be appreciated by those skilled in the art it is generally desirable to utilize as few layers as possible to accomplish a desired result, reducing the process steps necessary to produce the device, and thereby providing a much simpler and more robust manufacturing process. Moreover, utilizing as few layers as possible also results in higher yields as it is simpler to control the stoichiometric composition and dimensions of the fewer layers.

[0039] It will further be appreciated that although shown and described as part of a SONOS stack in a SONOS memory device, the ONO structure and method of the present invention is not so limited, and the ONO structure can be used in or with any semiconductor technology or in any device requiring a charge storing or dielectric layer or stack including, for example, in a split gate flash memory, a TaNOS stack, in a 1T (transistor) SONOS cell, a 2T SONOS cell, a 3T SONOS cell, a localized 2-bit cell, and in a multilevel programming or cell, without departing from the scope of the invention.

[0040] The advantages of ONO structures and methods of forming the same according to an embodiment of the present invention over previous or conventional approaches include: (i) the ability to enhance data retention in memory devices using the structure by dividing the oxynitride layer into a plurality of films or layers and tailoring the oxygen, nitrogen and silicon profile across each layer; (ii) the ability to enhance speed of a memory device without compromising data retention; (iii) the ability to meet or exceed data retention and speed specifications for memory devices using an ONO structure of an embodiment of the present invention at a temperature of at least about 125° C.; and (iv) provide heavy duty program erase cycles of 100,000 cycles or more.

[0041] The foregoing description of specific embodiments and examples of the invention have been presented for the purpose of illustration and description, and although the invention has been described and illustrated by certain of the preceding examples, it is not to be construed as being limited thereby. They are not intended to be exhaustive or to limit the invention to the precise forms disclosed, and many modifications, improvements and variations within the scope of the invention are possible in light of the above teaching. It is intended that the scope of the invention encompass the generic area as herein disclosed, and by the claims appended hereto and their equivalents. The scope of the present invention is defined by the claims, which includes known equivalents and unforeseeable equivalents at the time of filing of this application.

What is claimed is:

1. A method of forming a charge storing layer of a semiconductor device, the method comprising steps of:

depositing on a substrate a silicon-rich nitride; and

oxidizing the silicon-rich nitride to form a silicon-rich, oxygen-rich first oxynitride layer.

2. A method according to claim **1**, further comprising the step of forming over a surface of the first oxynitride layer at least one additional layer to form a multi-layer charge storing layer.

3. A method according to claim **2**, wherein the step of forming at least a one additional layer comprises the step of forming a second oxynitride layer.

4. A method according to claim **3**, wherein the first oxynitride layer and second oxynitride layer have differing stoichiometric compositions of oxygen, nitrogen and/or silicon.

5. A method according to claim **3**, wherein the step of forming the second oxynitride layer comprises the step of forming a second oxynitride layer under conditions selected to form a silicon-rich, oxygen-lean oxynitride layer.

6. A method according to claim **5**, wherein the first oxynitride layer is formed in a chemical vapor deposition (CVD) process using a process gas comprising a dichlorosilane (SiH_2Cl_2)/ammonia (NH_3) mixture and a nitrous oxide (N_2O)/NH_3 mixture at a ratio of about 8:1, and wherein the second oxynitride layer is formed in a CVD process using a process gas comprising a N_2O/NH_3 mixture and a SiH_2Cl_2/NH_3 mixture at a ratio of about 5:1.

7. A method according to claim **6**, wherein the steps of forming the first oxynitride layer and the second oxynitride

layer are performed sequentially in a single CVD tool by changing the ratio of the N_2O/NH_3 and SiH_2Cl_2/NH_3 mixtures.

8. A method according to claim **6**, wherein at least one of the first oxynitride layer and the second oxynitride layer is formed at temperature of at least about 780° C.

9. A method of forming a semiconductor device including an oxide-nitride-oxide (ONO) structure, the method comprising steps of:

forming a first oxide layer of the ONO structure;

forming a multi-layer charge storing layer comprising nitride on a surface of the first oxide layer; and

forming a second oxide layer of the ONO structure on a surface of the multi-layer charge storing layer.

10. A method according to claim **9**, wherein, the multi-layer charge storing layer comprises at least two silicon oxynitride (Si_2N_2O) layers.

11. A method according to claim **10**, wherein the at least two oxynitride layers have differing stoichiometric compositions of oxygen, nitrogen and/or silicon.

12. A method according to claim **11**, wherein the at least two oxynitride layers include a top oxynitride layer and a bottom oxynitride layer, and wherein the top oxynitride layer is formed under conditions selected to form a silicon-rich, oxygen-lean oxynitride layer, and the bottom oxynitride layer is formed under conditions selected to form a silicon-rich, oxygen-rich oxynitride layer.

13. A method according to claim **12**, wherein the top oxynitride layer is nitrogen-rich oxynitride layer.

14. A method according to claim **12**, wherein the step of forming the first oxide layer comprises the step of forming the first oxide layer using a steam anneal, and wherein a ratio of thicknesses between the top oxynitride layer and the bottom oxynitride layer is selected to facilitate forming the multi-layer memory layer following the step of forming the tunnel oxide layer using a steam anneal.

15. A method according to claim **12**, wherein the top oxynitride layer is formed using a process gas comprising a dichlorosilane (SiH_2Cl_2)/ammonia (NH_3) mixture and a nitrous oxide (N_2O)/NH_3 mixture at a ratio of about 5:1, and the bottom oxynitride layer is formed using a process gas comprising a N_2O/NH_3 mixture and a SiH_2Cl_2/NH_3 mixture at a ratio of about 8:1.

16. A method according to claim **10**, wherein the at least two oxynitride layers are formed at temperature of at least about 780° C.

17. A semiconductor device including an oxide-nitride-oxide (ONO) structure comprising a multilayer memory layer between a first oxide layer and a second oxide layer, wherein the multilayer memory layer comprises at least two silicon oxynitride layers.

18. A semiconductor device according to claim **17**, wherein the at least two oxynitride layers have differing stoichiometric compositions of oxygen, nitrogen and/or silicon.

19. A semiconductor device according to claim **17**, wherein oxygen, nitrogen and silicon profiles across the at least two oxynitride layers are selected to meet a predetermined data retention specification at an operating temperature of at least 125° C.

20. A semiconductor device according to claim **17**, wherein the at least two oxynitride layers include a top oxynitride layer and a bottom oxynitride layer, and wherein a ratio of thickness between the top oxynitride layer and the bottom oxynitride layer is from 1 to 5.

* * * * *