(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 September 2016 (15.09.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/144397 A1

(51) International Patent Classification:
H01L 21/8234 (2006.01)

(21) International Application Number:
PCT/US2015/062490

(22) International Filing Date:
24 November 2015 (24.11.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/130,106 9 March 2015 (09.03.2015) US
14/745,217 19 June 2015 (19.06.2015) US

(71) Applicant: **CYPRESS SEMICONDUCTOR CORPORATION** [US/US]; 198 Champion Court, San Jose, California 95134 (US).

(72) Inventor: **RAMKUMAR, Krishnaswamy**; 1193 Lynbrook Way, San Jose, California 95129 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to the identity of the inventor (Rule 4.17(i))*

Published:

— *with international search report (Art. 21(3))*

WO 2016/144397 A1

(54) Title: METHOD OF ONO STACK FORMATION

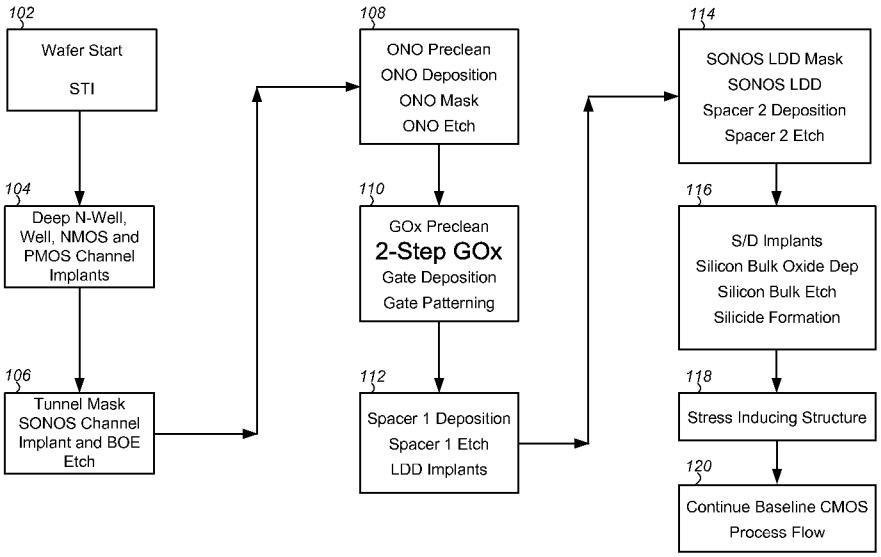


FIG. 1

(57) Abstract: A method of controlling the thickness of gate oxides in an integrated CMOS process which includes performing a two-step gate oxidation process to concurrently oxidize and therefore consume at least a first portion of the cap layer of the NV gate stack to form a blocking oxide and form a gate oxide of at least one metal-oxide-semiconductor (MOS) transistor in the second region, wherein the gate oxide of the at least one MOS transistor is formed during both a first oxidation step and a second oxidation step of the gate oxidation process

METHOD OF ONO STACK FORMATION

PRIORITY

[0001] This Application is an International Application of U.S. Application Number 14/745,217 filed June 19, 2015, which claims the priority and benefit under 35 U.S.C. § 119(e) of U.S. Provisional Application No. 62/130,106, filed on March 9, 2015, all of which are incorporated by reference herein in their entirety.

TECHNICAL FIELD

[0002] The present disclosure relates generally to semiconductor devices, and more particularly to memory cells and methods of manufacturing thereof including an embedded or integrally formed charge-trapping gate stack such as an oxide-nitride-oxide (ONO) stack into an existing complementary metal-oxide-semiconductor (CMOS) foundry logic technology.

BACKGROUND

[0003] For many applications, such as system-on-chip, it is desirable to integrate logic devices and interface circuits based upon metal-oxide-semiconductor field-effect transistors (MOSFET) and non-volatile memory (NVM) transistors on a single chip or wafer. This integration can seriously impact both the MOS transistor and NVM transistor fabrication processes. MOS transistors are typically fabricated using a standard or baseline complementary-metal-oxide-semiconductor (CMOS) process flows, involving the formation and patterning of conducting, semiconducting and dielectric materials. The composition of these materials, as well as the composition and concentration of processing reagents, and temperature used in such a

CMOS process flow are stringently controlled for each operation to ensure the resultant MOS transistors will function properly.

[0004] Non-volatile memory devices include non-volatile memory transistors, such as silicon-oxide-nitride-oxide-silicon (SONOS) based transistors, including charge-trapping gate stacks in which a stored or trapped charge changes a threshold voltage of the NVM transistor to store information as a logic 1 or 0. Charge-trapping gate stack formation involves the formation of a nitride or oxynitride charge-trapping layer(s) sandwiched between two dielectric or oxide layers typically fabricated using materials and processes that differ significantly from those of the baseline CMOS process flow, and which can detrimentally impact or be impacted by the fabrication of the MOS transistors. In particular, forming a gate oxide or dielectric of a MOS transistor may significantly degrade performance of a previously formed charge-trapping gate stack by altering a thickness or composition of the charge-trapping layer(s). In addition, this integration can seriously impact the baseline CMOS process flow, and generally requires a substantial number of mask sets and process steps, which add to the expense of fabricating the devices and can reduce yield of working devices.

[0005] Besides, it is imperative for the integrated fabrication process to be able to control the thickness of top oxide of NVM transistors in order to meet requirements such as threshold voltages V_{ts} and/or equivalent oxide thickness (EOT) requirements while satisfying gate oxide thickness requirements of MOS transistors, especially if those MOS transistors are high voltage input/output (HV I/O) transistors.

BRIEF DESCRIPTION OF THE DRAWINGS

[0006] The present disclosure is illustrated by way of example, and not by way of limitation, in the FIGS. of the accompanying drawings.

[0007] FIG. 1 is a flowchart illustrating an embodiment of a method for fabricating a memory cell including an embedded SONOS based NVM transistor and MOS transistors;

[0008] FIGS. 2A-2N are representative diagrams illustrating cross-sectional views of a portion of a memory cell during fabrication of the memory cell according to the method of FIG. 1; and

[0009] FIG. 2O is a representative diagram illustrating a cross-sectional view of a portion of a finished memory cell including an embedded SONOS based NVM transistor and MOS transistors fabricated according to the method of FIGS. 1 and 2A-2N.

DETAILED DESCRIPTION

[0010] The following description sets forth numerous specific details such as examples of specific systems, components, methods, and so forth, in order to provide a good understanding of several embodiments of the present invention. It will be apparent to one skilled in the art, however, that at least some embodiments may be practiced without these specific details. In other instances, well-known components or methods are not described in detail or are presented in a simple block diagram format in order to avoid unnecessarily obscuring the techniques described herein. Thus, the specific details set forth hereinafter are merely exemplary. Particular implementations may vary from these exemplary details and still be contemplated to be within the spirit and scope of the present invention.

[0011] Embodiments of a memory cell including an embedded non-volatile memory (NVM) transistor and a metal-oxide-semiconductor (MOS) transistor and methods of fabricating the same are described herein with reference to figures. However, particular embodiments may be practiced without one or more of these specific details, or in combination with other known methods, materials, and apparatuses. In the following description, numerous specific details are set forth, such as specific materials, dimensions, concentrations and processes parameters etc. to provide a thorough understanding of the present invention. In other instances, well-known semiconductor design and fabrication techniques have not been described in particular detail to avoid unnecessarily obscuring the present invention. Reference in the description to "an embodiment", "one embodiment", "an example embodiment", "some embodiments", and "various embodiments" means that a particular feature, structure, or characteristic described in connection with the embodiment(s) is included in at least one embodiment of the invention. Further, the appearances of the phrases "an embodiment", "one embodiment", "an example

embodiment", "some embodiments", and "various embodiments" in various places in the description do not necessarily all refer to the same embodiment(s).

[0012] The description includes references to the accompanying drawings, which form a part of the detailed description. The drawings show illustrations in accordance with exemplary embodiments. These embodiments, which may also be referred to herein as "examples," are described in enough detail to enable those skilled in the art to practice the embodiments of the claimed subject matter described herein. The embodiments may be combined, other embodiments may be utilized, or structural, logical, and electrical changes may be made without departing from the scope and spirit of the claimed subject matter. It should be understood that the embodiments described herein are not intended to limit the scope of the subject matter but rather to enable one skilled in the art to practice, make, and/or use the subject matter.

[0013] The terms "over", "under", "between", and "on" as used herein refer to a relative position of one layer with respect to other layers. As such, for example, one layer deposited or disposed over or under another layer may be directly in contact with the other layer or may have one or more intervening layers. Moreover, one layer deposited or disposed between layers may be directly in contact with the layers or may have one or more intervening layers. In contrast, a first layer "on" a second layer is in contact with that second layer. Additionally, the relative position of one layer with respect to other layers is provided assuming operations deposit, modify and remove films relative to a starting wafer without consideration of the absolute orientation of the wafer.

[0014] The NVM transistor may include memory transistors or devices implemented related to Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) or floating gate technology. An embodiment

of a method for integrating or embedding a NVM transistor into a standard or baseline CMOS process flow for fabricating one or more MOS transistors will now be described in detail with reference to FIG. 1 and FIGS. 2A through 2N. FIG. 1 is a flowchart illustrating an embodiment of a method or process flow for fabricating the memory cell. FIGS. 2A-2N are block diagrams illustrating cross-sectional views of a portion of a memory cell during fabrication of the memory cell according to the method of FIG. 1. FIG. 2O is a representative diagram illustrating a cross-sectional view of a portion of an embodiment of the finished memory cell.

[0015] Referring to FIG. 1 and FIG. 2A, the process begins with forming a number of isolation structures **202** in a substrate or wafer **204** (step **102**). The isolation structures **202** isolate the memory cell being formed from memory cells formed in adjoining areas (not shown) of the wafer **204**. Optionally and additionally, isolation structures **202** may be incorporated to isolate the NVM transistor being formed in a first region **206** of the wafer **204** from one or more of the MOS transistors being formed in a second region **208**. In one embodiment, the isolation structures **202** may include a dielectric material, such as an oxide or nitride, and may be formed by any conventional technique, including but not limited to shallow trench isolation (STI) or local oxidation of silicon (LOCOS). The wafer **204** may be a bulk wafer composed of any single crystal material suitable for semiconductor device fabrication, or may include a top epitaxial layer of a suitable material formed on a wafer. In one embodiment, suitable materials for the wafer **204** include, but are not limited to, silicon, germanium, silicon-germanium or a Group III-V compound semiconductor material.

[0016] Generally and optionally, as best shown in FIG. 2B, a pad oxide **209** may be formed over a surface **216** of the wafer **204** in both the first region **206** and the second region **208**. In one

embodiment, pad oxide **209** may be silicon dioxide (SiO_2) having a thickness of from about 10 nanometers (nm) to about 20 nm and may be grown by a thermal oxidation process or in-situ steam generation (ISSG) process.

[0017] Referring to FIG. 1 and FIG. 2B, dopants are then implanted into wafer **204** through pad oxide **209** to form wells in which the NVM transistor and/or the MOS transistors will be formed, and channels for the MOS transistors (step**104**). The dopants implanted may be of any type and concentration, and may be implanted at any energy, including energies necessary to form wells or deep wells for the NVM transistor and/or the MOS transistors, and to form channels for the MOS transistors. In a particular embodiment, illustrated in FIG. 2B as an example, dopants of an appropriate ion species are implanted to form a deep N-well **210** in the second region **208** over or in which a high-voltage (HV) MOS transistor **214**, such as a MOS input/output (I/O) transistor, will be formed. In alternative embodiments, wells or deep wells can also be formed for the NVM transistor and/or a standard or low-voltage (LV) MOS transistor, such as a MOS transistor **212**. It is further to be appreciated that the wells are formed by depositing and patterning a mask layer, such as a photoresist layer above the surface **216** of the wafer **204**, and implanting an appropriate ion species at an appropriate energy to an appropriate concentration.

[0018] In one embodiment, channels **218** for one or more of the MOS transistors **214**, **212**, are formed in the second region **208** of the wafer **204**. As with the well implant the channels **218** are formed by depositing and patterning a mask layer, such as a photoresist layer above the surface **216** of the wafer **204**, and implanting an appropriate ion species at an appropriate energy to an appropriate concentration. For example, BF_2 can be implanted at an

energy of from about 10 to about 100 kilo-electron volts (keV), and a dose of from about 1e12 cm^{-2} to about 1e14 cm^{-2} to form an N-type MOS (NMOS) transistor. A P-type MOS (PMOS) transistor may likewise be formed by implantation of Arsenic or Phosphorous ions at any suitable dose and energy. It is to be appreciated that implantation can be used to form channels **218**, in both of the MOS transistors **214**, **212**, at the same time, or at separate times using standard lithographic techniques, including a patterned photoresist layer to mask one of the channels for the MOS transistors.

[0019] Next, referring to FIG. 1 and FIG. 2C, a patterned tunnel mask **220** is formed on or overlying the pad oxide **209**, ions (represented by arrows **222**) of an appropriate, energy and concentration are implanted through a window or opening in the tunnel mask to form a channel **224** for a NVM transistor **226**, and the tunnel mask and the pad oxide in at least the second region **208** removed (step **106**). The tunnel mask can include a photoresist layer, or a hard mask formed, from a patterned nitride or silicon-nitride layer.

[0020] In one embodiment, the channel **224** for the NVM transistor **226** is a deep Indium doped channel implanted with Indium (In) at an energy of from about 50 to about 500 kilo-electron volts (keV), and a dose of from about 5e11 cm^{-2} to about 1e13 cm^{-2} to form an n-channel NVM transistor. In one embodiment, implanting Indium to form the channel **224** of the NVM transistor **226** improves the threshold voltage (V_T) uniformity of the NVM transistor from a sigma of V_T from about 150 millivolts (mV) to from about 70 to 80 mV. Optionally or additionally, a shallow doped channel is implanted with Arsenic at an energy about 20 keV and a dose of from about 5e11 cm^{-2} to about 1e13 cm^{-2} at channel **224**. Alternatively, BF_2 may be implanted to form an n-channel NVM transistor, or Arsenic or Phosphorous implanted to form a

p-channel NVM transistor. In one alternative embodiment, channel for NVM transistor **226** may also be formed concurrently with channels **218** of the MOS transistors **214**, **212**.

[0021] A photoresist tunnel mask **220** can be ashed or stripped using oxygen plasma. A hard mask can be removed using a wet or dry etch process. The pad oxide **209** is removed, for example in a wet clean process using a 10:1 buffered oxide etch (BOE) containing a surfactant. Alternatively, the wet clean process can be performed using a 20:1 BOE wet etch, a 50:1 hydrofluoric (HF) wet etch, a pad etch, or any other similar hydrofluoric-based wet etching chemistry.

[0022] Referring to FIG. 1 and FIGS. 2D-2F, the surface **209** of the wafer **204** is cleaned or precleaned, a number of dielectric layers, such as oxide-nitride-oxide or ONO layers or oxide-nitride-oxide-nitride-oxide or ONONO layers, formed or deposited, a mask formed on or overlying the dielectric layers, and the dielectric layers etched to form a dielectric gate stack **236** in the first region **206** (step **108**). The preclean can be a wet or dry process and in this embodiment is wet process using HF or standard cleans (SC1) and (SC2), and is highly selective to the material of the wafer **204**. In one embodiment, SC1 is typically performed using a 1:1:5 solution of ammonium hydroxide (NH_4OH), hydrogen peroxide (H_2O_2) and water (H_2O) at 30° C to 80° C for about 10 minutes. In another embodiment, SC2 is a short immersion in a 1:1:10 solution of HCl, H_2O_2 and H_2O at about 30° C to 80° C.

[0023] Referring to FIG. 2D, the dielectric or ONO or ONONO deposition begins with the formation of a tunnel dielectric **228** over at least the channel **224** of the NVM transistor **226** in the first region **206** of the wafer **204**, and may spread over to the second region **208** of wafer **204** where MOS transistor(s) is. The tunnel dielectric **228** may be any material and have any

thickness suitable to allow charge carriers to tunnel into an overlying charge-trapping layer under an applied gate bias while maintaining a suitable barrier to leakage when the NVM transistor is unbiased. In certain embodiments, tunnel dielectric **228** is silicon dioxide, silicon oxy-nitride, or a combination thereof and can be grown by a thermal oxidation process, using ISSG or radical oxidation.

[0024] In one embodiment a silicon dioxide tunnel dielectric **228** may be thermally grown in a thermal oxidation process. For example, a layer of silicon dioxide may be grown utilizing dry oxidation at 750°C - 800°C in an oxygen containing gas or atmosphere, such as oxygen (O_2) gas. The thermal oxidation process is carried out for a duration approximately in the range of 50 to 150 minutes to effect growth of a tunnel dielectric **228** having a thickness of from about 1.0 nanometers (nm) to about 3.0 nm by oxidation and consumption of the exposed surface of wafer.

[0025] In another embodiment, a silicon dioxide tunnel dielectric **228** may be grown in a radical oxidation process involving flowing hydrogen (H_2) and oxygen (O_2) gas into a processing chamber at a ratio to one another of approximately 1:1 without an ignition event, such as forming of a plasma, which would otherwise typically be used to pyrolyze the H_2 and O_2 to form steam. Instead, the H_2 and O_2 are permitted to react at a temperature approximately in the range of about 900° C to about 1100° C at a pressure approximately in the range of about 0.5 Torr to about 10 Torr to form radicals, such as, an OH radical, an HO_2 radical or an O diradical, at the surface of wafer. The radical oxidation process is carried out for a duration approximately in the approximate range of about 1 to about 10 minutes to effect growth of a tunnel dielectric **228** having a thickness of from about 1.0 nanometers (nm) to about 4.0 nm by oxidation and consumption of the exposed surface of wafer. It will be understood that in FIG. 2D and in

subsequent figures the thickness of tunnel dielectric **228** is exaggerated relative to the pad oxide **209**, which is approximately 7 times thicker, for the purposes of clarity. In one embodiment, a tunnel dielectric **228** grown in a radical oxidation process may be both denser and composed of substantially fewer hydrogen atoms/cm^3 than a tunnel dielectric formed by wet oxidation techniques, even at a reduced thickness. In certain embodiments, the radical oxidation process is carried out in a batch-processing chamber or furnace capable of processing multiple wafers to provide a high quality tunnel dielectric **228** without impacting the throughput (wafers/hr.) requirements that a fabrication facility may require.

[0026] In another embodiment, tunnel dielectric layer **228** is deposited by chemical vapor deposition (CVD) or atomic layer deposition and is composed of a dielectric layer which may include, but is not limited to silicon dioxide, silicon oxy-nitride, silicon nitride, aluminum oxide, hafnium oxide, zirconium oxide, hafnium silicate, zirconium silicate, hafnium oxy-nitride, hafnium zirconium oxide and lanthanum oxide. In yet another embodiment, tunnel dielectric **228** may be a bi-layer dielectric region including a bottom layer of a material such as, but not limited to, silicon dioxide or silicon oxy-nitride and a top layer of a material which may include, but is not limited to silicon nitride, aluminum oxide, hafnium oxide, zirconium oxide, hafnium silicate, zirconium silicate, hafnium oxy-nitride, hafnium zirconium oxide and lanthanum oxide.

[0027] Referring again to FIG. 2D, a charge-trapping layer is formed on or overlying the tunnel dielectric **228**. Generally, as in the embodiment shown, the charge-trapping layer is a multi-layer charge-trapping layer **230** comprising multiple layers including at least a lower or first charge-trapping layer **230***a* which is closer to the tunnel dielectric **228**, and an upper or

second charge-trapping layer **230** *b* that is oxygen-lean relative to the first charge-trapping layer and comprises a majority of a charge traps distributed in multi-layer charge-trapping layer.

[0028] The first charge-trapping layer **230** *a* of a multi-layer charge-trapping layer **230** may include a silicon nitride (Si_3N_4), silicon-rich silicon nitride or a silicon oxy-nitride ($SiO_xN_y(HO)$) layer. For example, the first charge-trapping layer **230** *a* may include a silicon oxynitride layer having a thickness of between about 2.0 nm and about 6.0 nm formed by a CVD process using dichlorosilane (DCS)/ammonia (NH_3) and nitrous oxide (N_2O)/NH_3 gas mixtures in ratios and at flow rates tailored to provide a silicon-rich and oxygen-rich oxynitride layer.

[0029] The second charge-trapping layer **230** *b* of the multi-layer charge-trapping layer **230** is then formed, either directly or indirectly, over the first charge-trapping layer **230** *a*. In one embodiment, the second charge-trapping layer **230** *b* may include a silicon nitride and silicon oxy-nitride layer having a stoichiometric composition of oxygen, nitrogen and/or silicon that is different from that of the first charge-trapping layer **230** *a*. The second charge-trapping layer **230** *b* may include a silicon oxynitride layer having a thickness of between about 2.0 nm and about 8.0 nm, and may 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 nitride layer. In one alternative embodiment, the stoichiometric composition of oxygen, nitrogen and/or silicon of first and second charge-trapping layers **230** *a&b* may be identical or approximate to one another.

[0030] In another embodiment, there may be a dielectric and/or oxide layer (not shown) formed between the first and second charge-trapping layers **230***a* and **230***b*, making the multi-layer charge trapping layer **230** an NON stack. In some embodiments, the multi-layer charge-

trapping layer **230** is a split charge-trapping layer, further including a thin, middle oxide layer (not shown) separating the first (lower) and second (upper) charge-trapping layers **230***a* and **230***b*. The middle oxide layer substantially reduces the probability of electron charge that accumulates at the boundaries of the second charge-trapping layer **230***b* during programming from tunneling into the first charge-trapping layer **230***a*, resulting in lower leakage current than for the conventional memory devices. In one embodiment, the middle oxide layer is formed by oxidizing to a chosen depth using thermal or radical oxidation. Radical oxidation may be performed, for example, at a temperature of 1000-1100°C using a single wafer tool, or 800-900°C using a batch reactor tool. A mixture of H_2 and O_2 gasses may be introduced to a process chamber at a ratio of approximately 1:1 and 10-15 Torr using a single wafer tool, or a pressure of 300-500 Torr for a batch process, for a time of 1-2 minutes using a single wafer tool, or 30 min to 1 hour using a batch process. In some embodiments, the radical oxidation process is without an ignition event, such as forming of plasma, which would otherwise typically be used to pyrolyze the H_2 and O_2 to form steam. Instead, the H_2 and O_2 is permitted to react at a surface of the first charge-trapping layer **230***a* to form radicals, such as, an OH radical, an HO_2 radical or an O diradical, to form the middle oxide layer.

[0031] As used herein, the terms "oxygen-rich" and "silicon-rich" are relative to a stoichiometric silicon nitride, or "nitride," commonly employed in the art having a composition of (Si_3N_4) and with a refractive index (RI) of approximately 2.0. Thus, "oxygen-rich" silicon oxynitride entails a shift from stoichiometric silicon nitride toward a higher weight % of silicon and oxygen (i.e. reduction of nitrogen). An oxygen rich silicon oxynitride film is therefore more like silicon dioxide and the RI is reduced toward the 1.45 RI of pure silicon dioxide. Similarly, films described herein as "silicon-rich" entail a shift from stoichiometric silicon nitride toward a

higher weight % of silicon with less oxygen than an "oxygen-rich" film. A silicon-rich silicon oxynitride film is therefore more like silicon and the RI is increased toward the 3.5 RI of pure silicon.

[0032] Referring again to FIG. 2D, the number of dielectric layers further includes a cap layer **232** that is formed on or overlying the charge-trapping layer **230**. In one embodiment, the cap layer **232** includes a silicon nitride all or part of which is subsequently oxidized to form a blocking oxide overlying the charge-trapping layer **230**. In a few embodiments, cap layer **232** may be a single layer of nitride (not shown) having a homogeneous composition, a single layer of nitride having a gradient in stoichiometric composition, or, as in the embodiment shown, may be a multi-layer cap layer including at least a lower or first cap layer **232***a* overlying the second charge-trapping layer **230***b*, and an upper or second cap layer **232***b* overlying the first cap layer **232***a*.

[0033] In one embodiment, the first cap layer **232***a* can include a silicon nitride, a silicon-rich silicon nitride or a silicon-rich silicon oxynitride layer having a thickness of between 2.0 nm and 4.0 nm formed by a CVD process using N_2O/NH_3 and DCS/NH_3 gas mixtures. Similarly, the second cap layer **232***b* can also include a silicon nitride, a silicon-rich silicon nitride or a silicon-rich silicon oxynitride layer having a thickness of between 2.0 nm and 4.0 nm formed by a CVD process using N_2O/NH_3 and DCS/NH_3 gas mixtures. Optionally, the first cap layer **232***a* and second cap layer **232***b* may comprise different stoichiometric ratio. For example, the second cap layer **232***b* may comprise a silicon or oxygen rich composition relative to the first cap layer **232***a* to facilitate removal of the second cap layer in a dry or wet clean process prior to oxidizing the first cap layer **232***a*. Alternatively, the first cap layer **232***a* may comprise a silicon

or oxygen rich composition relative to the second cap layer **232***b* to facilitate oxidation of the first cap layer **232***a*.

[0034] Referring to FIG. 2E, a sacrificial oxide layer **234** is formed on or overlying the second cap layer **232***b*. In one embodiment, the sacrificial oxide layer **234** may be formed or deposited by a chemical vapor deposition process in a low pressure chemical vapor deposition (LPCVD) chamber. For example, the sacrificial oxide layer **234** may be deposited by a CVD process using a process gas including gas mixtures of silane or dichlorosilane (DCS) and an oxygen containing gas, such as O_2 or N_2O, in ratios and at flow rates tailored to provide a silicon dioxide (SiO_2) sacrificial oxide layer. In another embodiment, the sacrificial oxide layer **234** may include a silicon dioxide layer grown by a thermal oxidation process, in-situ steam generation (ISSG), or radical oxidation, and having a thickness of approximately between 2.0 nm and 4.0 nm.

[0035] Referring to FIG. 2F, a patterned mask layer (not shown) is formed on or overlying the sacrificial oxide layer **234**. Subsequently, the sacrificial oxide layer **234**, cap layers **232***a&b*, and the charge-trapping layers **230***a&b* are etched or patterned to form a gate stack **236** overlying the channel **224** of the NVM transistor **226** and sacrificial oxide layer **234**, cap layers **232***a&b*, and charge trapping layers **230***a&b* from the second region **208** of the wafer **204** are removed. In one embodiment, the patterned mask layer (not shown) may include a photoresist layer patterned using standard lithographic techniques, and sacrificial oxide layer **234**, cap layer **232**, and charge trapping layer **230** may be etched or removed using a dry etch process including one or more separate steps to stop at tunnel dielectric layer **228** or proximate to surface **216** of wafer **204**.

[0036] Referring to FIG. 1, a gate oxide or GOX preclean is performed, gate oxides for both MOS transistors **214**, **212** formed, and a gate layer is deposited and patterned to form gates for the NVM transistor **226**, and both MOS transistors (step **110**). Referring to FIG. 2G, during the GOX preclean process, sacrificial oxide layer **234** of gate stack **236** and a portion of the cap layer **232** or substantially of all of a top most layer, such as second cap layer **232***b*, in a multi-layer cap layer **232** are removed from the gate stack **236** in a highly selective cleaning process. In certain embodiments, the highly selective cleaning process may even remove a portion of first cap layer **232***a*. In one embodiment, this cleaning process may simultaneously or concurrently further remove any residual oxide, such as an oxide tunnel dielectric **228** and pad oxide **209**, remaining in the first region **206** outside of the gate stack **236** and in the second region **208** to prepare the wafer **204** in that region for gate oxide growth. In one embodiment, the thickness of cap layer **232** is adjusted to allow a portion or substantially all of second cap layer **232** *b,* may even be a portion of first cap layer **232***a*, to be consumed by the GOX preclean. In one embodiment, sacrificial oxide layer **234** and second cap layer **232***b* are removed in a wet clean process using a 10:1 buffered oxide etch (BOE) containing a surfactant. Alternatively, the wet clean process can be performed using a 20:1 BOE wet etch, a 50:1 hydrofluoric (HF) wet etch, a pad etch, or any other similar hydrofluoric-based wet etching chemistry.

[0037] This embodiment of the GOX preclean is advantageous in that it substantially does not affect the baseline CMOS process, either in the preclean step (step **110**) or a subsequent oxidation step (step **112**), but rather uses it for the integration of the NVM transistor fabrication.

[0038] Referring to FIGS. 2H and 2I, in one embodiment, a two-step oxidation process is performed consecutively to oxidize at least the remaining portion of the cap layer **232** or the first

cap layer **232** *a* of a multi-layer cap layer, or a portion of the second charge-trapping layer **230** *b* to form a blocking oxide layer **238** overlying the second charge-trapping layer **230***b*. In one embodiment, the two-step oxidation process is adapted to oxidize the first cap layer **232** *a* to form the blocking oxide layer **238** while simultaneously or concurrently oxidizing at least a portion of the surface **216** of the wafer **204** in the second region **208** to form gate oxide **240** overlying at least the channel **218** of at least one MOS transistor. Accordingly, in general it is imperative to be able to configure the two-step oxidation process in order to grow both top oxide layer of the NVM transistor **226** and gate oxide(s) of the MOS transistor(s) **212, 214** to their respective desirable operational thicknesses. In one embodiment, gate oxide thicknesses are such that the final top oxide of NVM transistor **226** was approximately 30-45Å thick in order to meet the requirements for the reliability of the ONO stack. However, in some other embodiments, the MOS transistor(s) may be an I/O transistor in which their I/O gate oxide is required to be much thicker in order to support a higher I/O voltage. For instance, in one embodiment, one of the MOS transistors in the region **208** is a high voltage I/O transistor and thus requires a thick gate oxide (over 100Å to 200Å). In such process flows, while the I/O gate oxide may achieve its required thickness, since they are subjected to the same environment during oxidation of the gate oxide of MOS transistor(s), top oxide of the NVM transistor **226** may be grown too thick. As a result, the NVM transistor **226** may not meet the requirements for effective oxide thickness (EOT) and program/erase V_{ts}. Alternatively, the deposited top oxide may be protected and be the top oxide in the final NVM stack, the reliability of the NVM stack may however be negatively impacted by the inferior quality top oxide.

[0039] Therefore, in one embodiment, a novel two-step oxidation process is proposed to ensure desirable top oxide **238** of NVM transistor **226** and gate oxide **240** of MOS transistor(s)

thicknesses are both achieved while maintain the quality of the grown oxide layers. In one embodiment, the first step is a rapid thermal dry oxidation (RTO) and the second step is a rapid and radical wet oxidation such as in-situ steam generation (ISSG). Referring to FIG. 2H, the oxidation process starts with dry RTO performed in a batch or single wafer processing chamber with or without an ignition event such as plasma. For example, in one embodiment, the device is subjected to a rapid thermal oxidation process involving flowing oxygen (O_2) gas into a processing chamber. The O_2 gas is permitted to react at a temperature approximately in the range of 1000-1100° C. at a pressure approximately in the range of 0.5-5 Torr to form a bottom layer of first gate oxide **240***a*. In one embodiment, a bottom layer of first gate oxide **240***a* is grown, by oxidizing silicon wafer **204**, on at least a portion of the surface **216** of the wafer **204** in the second region **208** overlying at least the channel **218** of at least one MOS transistor, and in the first region **206** outside of gate stack **236**. However, in one embodiment, the dry RTO process has very little to no effect on first cap layer **232** *a* of ONO gate stack **236**, which is a nitride or oxynitride, and virtually no oxide is grown on gate stack **236**. In one alternative embodiment, dry RTO process may be substituted with a rapid molecular oxidation (dry or wet) which is a non-radical oxidation process. Since there is no radical formed during the process, the first cap layer **232** *a*, which is a nitride or oxynitride, will have virtually no oxide grown on gate stack **236** while a bottom layer of first gate oxide **240***a* is formed, on at least a portion of the surface **216** of the wafer **204** in the second region **208** overlying at least the channel **218** of at least one MOS transistor. In one embodiment, after the dry RTO oxidation process, or one of its alterative oxidation processes, a bottom layer of first gate oxide **240***a* may have a thickness of from about 85Å to about 95Å and more.

[0040] After a desirable thickness of first gate oxide bottom layer **240***a* is grown, the oxidation process may progress immediately to a separated step of a second wet rapid and radical oxidation process such as in-situ steam generation (ISSG). Referring to FIG. 2I, for example, wet rapid and radical oxidation may be performed in a batch or single wafer processing chamber with or without an ignition event such as plasma. For example, in one embodiment the blocking oxide layer **238** and a top layer of first gate oxide **240***b* may be grown in a wet radical oxidation process involving flowing hydrogen (H_2) and oxygen (O_2) gas into a processing chamber at a ratio to one another of approximately 1:1 without an ignition event, such as forming of a plasma, which would otherwise typically be used to pyrolyze the H_2 and O_2 to form steam. Instead, the H_2 and O_2 are permitted to react at a temperature approximately in the range of 1000-1100° C. at a pressure approximately in the range of 0.5-10 Torr to form radicals, such as, an OH radical, an HO_2 radical or an O diradical radicals at a surface of the cap layer **232** or the first cap layer **232***a*. The oxidation process is carried out for a duration approximately in the range of 1-5 minutes for a single wafer using an ISSG process, or 30-120 minutes for a batch furnace process to effect growth of a blocking oxide layer **238** by oxidation and consumption of the first cap layer **232***a* and may be a portion of the second charge-trapping layer **230***b*. During the same period, a top layer of first gate oxide **240***b* is grown on the first gate oxide bottom layer **240***a* in the second region **208** overlying at least the channel **218** of at least one MOS transistor. In one embodiment, after the second wet radical oxidation process such as ISSG, blocking oxide layer **238** may have a thickness of from about 30Å to about 45Å. Simultaneously, first gate oxide top layer **240*b*** is grown to complete formation of gate oxide layer **240** to a thickness of from about 105Å to about 200Å. In alternative embodiments, the second step wet rapid and radical oxidation may be substituted by processes such as chemical vapor deposition (CVD), or other

radical oxidation processes performed in a batch or single wafer processing chamber with or without an ignition event such as plasma as long as oxide will be grown or deposited both on gate stack **236** of NVM transistor and first gate oxide bottom layer **240***a* of MOS transistor(s) simultaneously. In one embodiment, after the two step oxidation process as previously described, the thickness ratio between top oxide **238** of NVM transistor and gate oxide **240** of at least one of the MOS transistor is in an approximate range of 1:2.33 (105nm/45nm) to 1:6.67 (200/30nm).

[0041] In one embodiment, by controlling the parameters in the first step dry RTO process and the second step wet ISSG process such as time duration, temperature, pressure, reactants etc., targeted thicknesses of blocking oxide layer **238** of NVM transistor in the first region **206** and gate oxide layer **240** of at least one MOS transistor in the second region **208** are achieved. The following table illustrates an example of implementation of the proposed two-step oxidation process:

Oxidation Process	Grown Oxide (Silicon) Thickness	Top Oxide (SONOS) Thickness	Gate Oxide (MOS) Thickness
Dry RTO	100Å	0Å	95Å
Wet ISSG	70Å	45Å	115Å

[0042] In this example, the dry RTO process (first oxidation step) that grows about 100Å of oxide on silicon, may grow about 95Å of oxide in the second region **208** on the wafer **204** and has very little effect on the first cap layer **232***a* which is a nitride or oxynitride. The wet ISSG process (second oxidation step), which normally grows 70Å of oxide on silicon, builds the

thickness of gate oxide **240** in the second region **208** on the wafer **204** up to approximately 115Å which may be the target for a HV I/O gate oxide in one embodiment. The top oxide of the ONNO or ONONO gate stack **236** grows only during the wet ISSG process. For 70Å oxidation on silicon, approximately 45Å oxide is grown on by consuming the nitride in first cap layer **232***a* and maybe second charge trapping layer **230***b* of gate stack **236**. In one embodiment, a target thickness of about 45Å for a top oxide of a SONOS may be desirable to meet EOT and V_{ts} requirements. In other embodiments, parameters of the rapid dry thermal oxidation and radical wet oxidation may be adjusted to attain desirable top oxide thicknesses for the NVM transistor and MOS transistor in one single process. Besides, using rapid thermal processes for both oxidations minimizes the interaction between the STI gapfill dielectric, such as dielectric in isolation structures **202** and the ONO stack. It minimizes the moisture in the gap fill dielectric from getting out and affecting the ONO thicknesses which may impact the V_{ts} of the SONOS device.

[0043] In one embodiment, the two gate oxidation steps are carried out in single wafer tools. The dry thermal oxidation and wet radical oxidation processes can be either done in two different tools such as a rapid thermal anneal (RTA) single wafer tool for the dry RTO process and subsequently an ISSG single wafer tool for the ISSG process. Alternatively, dry and wet RTO may be carried out in one ISSG tool. In this particular embodiment, the oxidation is initially performed by raising the wafer temperature to 1000-1100 ^{o}C range and flowing O_2 only for a required amount of time to finish the dry RTO process. Subsequently, H_2 is introduced in order to start the ISSG oxidation. The operation temperature of the two oxidation steps is preferably kept at the same value.

[0044] In one alternative embodiment, the sequence of the dry RTO and wet ISSG oxidation steps may be reversed. The device is first subjected to wet ISSG oxidation either in a batch or single wafer processing chamber with or without an ignition event such as plasma, wherein top oxide **238'** may be grown on both gate stack **236** of NVM transistor and first gate oxide bottom layer **240***a'* in the second region **208** overlying at least the channel **218** of at least one MOS transistor. The wet ISSG oxidation may be terminated when a desirable thickness of top oxide **238'** of NVM transistor **236** is attained. Subsequently, a dry RTO step may begin to continue growing first gate oxide top layer **240***b'* on first gate oxide bottom layer **240***a'* until a desirable combined thickness of gate oxide **240** is attained. The dry RTO step has very little to no effect on the thickness of top oxide **238'** of NVM transistor **236** that was grown during the wet ISSG oxidation.

[0045] In another alternative embodiment, both oxidation steps are done in ISSG process, hence both oxidation steps will oxidize first cap layer **232***a* of gate stack **236** of NVM transistor. In this embodiment, after the first ISSG oxidation step, a photomask is applied such that only gate stack **236** of NVM transistor regions is exposed by using photoresist. Using this photomask, a HF etch is carried out to remove only the oxide grown on and consuming a portion of the nitride in first cap layer **232***a* in NVM transistor. Following the edge process, the photoresist is removed and the wafer is subjected to a second wet ISSG process to form the rest of gate oxide **240** of MOS transistor and also form top oxide **238** of NVM transistor by consuming the rest of first cap layer **232***a* until a desirable thickness is attained. In this particular embodiment, an extra photomask may be required.

[0046] In some embodiments, as explained briefly above, the dry radical RTO step and/or wet radical ISSG oxidation step may be carried out at a batch furnace. In these embodiments, the process flow is the same but in each oxidation, a batch of wafers (100-125 wafers) are subjected to the oxidation concurrently. This embodiment assumes that a capable batch type tool for radical oxidation is available. For this scheme to be successful, the gap fill dielectric, such as dielectric in isolation structures **202** and the ONO stack, needs to be of high quality with zero or minimum moisture content.

[0047] In one embodiment, the grown gate oxide **240**, which includes a layer of first gate oxide bottom layer **240***a* grown during the first oxidation step and first gate oxide top layer **240***b* during the second oxidation step, may or may not exhibit different stoichiometric ratio and/or structures between gate oxide layers **240***a* and **240***b*.

[0048] In some embodiments, such as that shown in FIGS. 2J to 2N, the method further includes a dual gate oxide process flow to enable fabrication of both a LV MOS transistor **212** and a HV MOS transistor **214**. Referring to FIG. 2J, a patterned mask layer **242** is formed over the first and second regions **206**, **208** of the wafer **204**. The patterned mask layer **242** can be a photoresist layer patterned using standard lithographic techniques, and includes at least one opening **244** over a channel **218** in the second region **208**. The thick, first gate oxide **240** is etched in the exposed regions by using a BOE etch, under conditions similar to those described above with respect to removing the sacrificial oxide layer **234**, and the patterned mask layer **242** is then removed.

[0049] Referring to FIG. 2K, the wafer **204** is cleaned using a wet etch that does not etch oxide in order to protect the first gate oxide **240** of the HV MOS transistor **212**, and the blocking

oxide layer **238** of the gate stack **236**. The wafer **204** is then subjected to a thermal oxidation process to grow a thin, second gate oxide **246** having an appropriate thickness, such as from about 1 nm to about 3 nm. In some embodiments, the second gate oxide **246** can be overlaid with a deposited layer (not shown) such as silicon oxy-nitride, silicon nitride, aluminum oxide, hafnium oxide, zirconium oxide, hafnium silicate, zirconium silicate, hafnium oxy-nitride, hafnium zirconium oxide and lanthanum oxide.

[0050] Referring to FIG. 2L, a gate layer **248** of any conducting or semiconducting material suitable for accommodating a biasing of the NVM transistor **226** and operation of the MOS transistors **214**, **212**, is formed over the gate stack **236**, the first gate oxide **240** of the HV MOS transistor **214**, and the second gate oxide **246** of the MOS transistor **212**. In one embodiment, the gate layer **248** is formed by physical vapor deposition and is composed of a metal-containing material which may include, but is not limited to, metal nitrides, metal carbides, metal silicides, hafnium, zirconium, titanium, tantalum, aluminum, ruthenium, palladium, platinum, cobalt and nickel. In another embodiment, the gate layer is formed by a CVD process and is composed of a single doped polysilicon layer, which may then be patterned to form control gates of the NVM transistor **226** and MOS transistors **214**, **212**.

[0051] Referring to FIG. 2M, the gate layer **248** is patterned using a mask layer (not shown) and standard lithographic techniques to stop on surfaces of the blocking oxide layer **238**, the first gate oxide **240** and the second gate oxide **246**, thereby forming a gate **250** for the gate stack **236** of a NVM transistor **226**, a gate **252** for the HV MOS transistor **214**, and a gate **254** for the MOS transistor **212**.

[0052] Referring to FIG. 1 and FIG. 2N, a first spacer layer is deposited and etched to form first sidewall spacers **256** adjacent to the gates **252**, **254**, of the MOS transistors **212**, **214**, and the NVM transistor **226**, and one or more lightly-doped drain extensions (LDD **258**) are implanted adjacent to and extend under sidewall spacers **256** of one or more of the MOS transistors **212**, **214** (step **112**).

[0053] Next, a SONOS LDD mask is formed over the wafer **204** and lightly-doped drain extensions (LDD **260**) are implanted, adjacent to the NVM transistor **226**. Finally, a second spacer layer is deposited and etched to form second sidewall spacers **262** adjacent to the gate stack **236**, of the NVM transistor **226** (step **114**).

[0054] In one or more alternative embodiments, fabrication steps as illustrated and described in FIGS. 1 to 2N may be adapted or modified to manufacture a floating gate based NVM transistor, instead of or additional of the SONOS based NVM transistor **226** in an integrated baseline CMOS process.

[0055] Referring to FIGS. 1 and 2O, with the NVM transistor **226**, HV MOS transistor **214** and LV MOS transistor **212** substantially complete, source and drain implants are performed to form source and drain regions **264** for all transistors and a silicide process performed (step **116**). As depicted, silicide regions **266** may be formed on the exposed gates **250**, **252** and **254** and exposed source and drain regions **264**. The silicide process may be any commonly employed in the art, typically including a pre-clean etch, cobalt or nickel metal deposition, anneal and wet strip.

[0056] Referring FIG. 1 and FIG. 2O, optionally the method of fabricating memory cells including an embedded or integrally formed SONOS based NVM transistor and MOS transistor(s) further includes the step of forming a stress inducing layer or structure **268**, such as a stress inducing nitride layer, over the gate stack **236** of the NVM transistor **226** to increase data retention and/or to improve programming time and efficiency (step **118**). In particular, inducing stress into the charge-trapping layer **230** of the NVM transistor **226** changes energy levels of charge traps formed therein, thereby increasing charge retention of the charge-trapping layer. In addition, forming a stress inducing structure **268**, in or on the surface **216** of the wafer **204** proximal to, and preferably surrounding, a region of the wafer in which the channel **224** of NVM transistor **226** is formed will reduce the band gap, and, depending on the type of strain, increases carrier mobility. For example, tensile strain, in which inter-atomic distances in the crystal lattice of the wafer **204** are stretched, increases the mobility of electrons, making N-type transistors faster. Compressive strain, in which those distances are shortened, produces a similar effect in P-type transistors by increasing the mobility of holes. Both of these strain induced factors, i.e., reduced band gap and increased carrier mobility, will result in faster and more efficient programming of NVM transistor **226**.

[0057] The strain inducing structure **268** can include a pre-metal dielectric (PMD) layer formed using a High Aspect Ratio Process (HARP™) oxidation process, a compressive or tensile nitride layer formed using a plasma enhanced chemical vapor deposition (PECVD) or a Bis-Tertiary Butyl Amino Silane (BTBAS) nitride layer.

[0058] In certain embodiments, such as that shown in FIG. 2O, the stress inducing structure **268** may also be formed over one or more of the MOS transistors to induce strain in the channel of the MOS transistor.

[0059] Finally, the standard or baseline CMOS process flow is continued to substantially complete the front end device fabrication (step **120**), yielding the structure shown in FIG. 2O. FIG. 2O is a block diagram illustrating a cross-sectional view of a portion of a finished memory cell including an embedded SONOS based NVM transistor and MOS transistors fabricated according to the method of FIGS. **1** and **2**A-**2**N.

[0060] Thus, embodiments of memory cells including embedded or integrally formed SONOS based NVM transistor and MOS transistors and methods of fabricating the same have been described. Although the present disclosure has been described with reference to specific exemplary embodiments, it will be evident that various modifications and changes may be made to these embodiments without departing from the broader spirit and scope of the disclosure. Accordingly, the specification and drawings are to be regarded in an illustrative rather than a restrictive sense.

[0061] The Abstract of the Disclosure is provided to comply with 37 C.F.R. §1.72(b), requiring an abstract that will allow the reader to quickly ascertain the nature of one or more embodiments of the technical disclosure. It is submitted with the understanding that it will not be used to interpret or limit the scope or meaning of the claims. In addition, in the foregoing Detailed Description, it can be seen that various features are grouped together in a single embodiment for the purpose of streamlining the disclosure. This method of disclosure is not to be interpreted as reflecting an intention that the claimed embodiments require more features than

are expressly recited in each claim. Rather, as the following claims reflect, inventive subject matter lies in less than all features of a single disclosed embodiment. Thus, the following claims are hereby incorporated into the Detailed Description, with each claim standing on its own as a separate embodiment.

[0062] 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 circuit or method. The appearances of the phrase one embodiment in various places in the specification do not necessarily all refer to the same embodiment.

[0063] In the foregoing specification, the invention has been described with reference to specific exemplary embodiments thereof. It will, however, be evident that various modifications and changes may be made thereto without departing from the broader spirit and scope of the invention as set forth in the appended claims. The specification and drawings are, accordingly, to be regarded in an illustrative sense rather than a restrictive sense.

What is claimed is:

1. A method, comprising:

forming a dielectric stack on a wafer, wherein the dielectric stack includes tunneling dielectric on the wafer, a charge-trapping layer, and a cap layer overlaying the charge-trapping layer;

patterning the dielectric stack to form a non-volatile (NV) gate stack of a non-volatile memory (NVM) transistor in a first region of the wafer while concurrently removing the dielectric stack in a second region of the wafer; and

performing a two-step gate oxidation process to concurrently oxidize at least a first portion of the cap layer of the NV gate stack to form a blocking oxide and form a gate oxide of at least one metal-oxide-semiconductor (MOS) transistor in the second region, wherein the gate oxide of the at least one MOS transistor is formed during both a first oxidation step and a second oxidation step of the two-step gate oxidation process.

2. The method of claim 1, wherein after the performing the two-step gate oxidation process, a ratio between thicknesses of the blocking oxide of the NV gate stack and the gate oxide of the at least one MOS transistor is in an approximate range of 1:2.33 to 1:6.67.

3. The method of claim 1, wherein the second oxidation step of the gate oxidation process is performed immediately after the first oxidation step.

4. The method of claim 1, wherein, during the first oxidation step of the gate oxidation process, the at least the first portion of the cap layer of the NV gate stack remains substantially unchanged, both stoichiometrically and chemically.

5. The method of claim 3, wherein after the first oxidation step of the gate oxidation process, the gate oxide of the at least one MOS transistor is grown to a first thickness, and wherein during the second oxidation step of the gate oxidation process, the gate oxide continues to be grown to a second thickness, wherein the second thickness is greater than the first thickness.

6. The method of claim 1, wherein the first oxidation step of the gate oxidation process includes a dry rapid thermal oxidation (RTO) process.

7. The method of claim 1, wherein the second oxidation step of the gate oxidation process includes a wet in-situ steam generation (ISSG) process.

8. The method of claim 1, at least one of the first and the second oxidation steps is performed in a single wafer processing chamber or a batch wafer processing chamber.

9. The method of claim 3, wherein the first oxidation step of the gate oxidation process includes a wet in-situ steam generation (ISSG) process and the second oxidation step of the gate oxidation process includes a dry rapid thermal oxidation (RTO) process.

10. The method of claim 7, wherein the first oxidation step of the gate oxidation process includes a rapid molecular oxidation.

11. The method of claim 5, wherein, after completion of the first and the second oxidation steps of the gate oxidation process, the blocking oxide of NVM transistor reaches a thickness in an approximate range of 30Å to 45Å while the gate oxide of the at least one MOS transistor reaches the second thickness in an approximate range of 105Å to 200Å.

12. The method of claim 1, wherein the dielectric stack further comprises a sacrificial oxide layer deposited on top of the cap layer, wherein the cap layer includes nitrides or oxynitrides and is divided into the first portion and a second portion, wherein the second portion is deposited over the first portion, and wherein the blocking oxide of the NV gate stack is formed by consuming at least the first portion of the cap layer.

13. The method of claim 12, prior to the performing the two-step gate oxidation process, the method further comprising:

performing a gate oxide (GOX) preclean process to remove the sacrificial oxide and at least the second portion of the cap layer of the NV gate stack while simultaneously removing any remaining oxides on a surface of the wafer, and therefore exposing the surface of the wafer in the second region, wherein the second portion of the cap layer includes an oxygen rich composition relative to the first portion to facilitate removal of the second portion of the cap layer.

14. The method of claim 1, wherein prior to the forming a dielectric stack on a wafer, the method further comprising:

implanting indium to a dose between an approximate range of $5e11cm^{-2}$ to $1e13cm^{-2}$ in the first region of the wafer to form a channel of the NVM transistor.

15. The method of claim 1, wherein the first and the second oxidation steps of the gate oxidation process are both performed in an ISSG processing chamber and subjected to a temperature range of approximately 800 °C to 1100 °C, wherein only oxygen is introduced to the ISSG processing chamber during the first oxidation step, and wherein both oxygen and hydrogen is introduced during the second oxidation step.

16. The method of claim 1, wherein both the first and the second oxidation steps are ISSG process, and

wherein between performing the first and the second oxidation steps in the performing the two-step gate oxidation process, the method further comprising

patterning an opening to expose only a first blocking oxide formed from oxidizing a portion of the first portion of the cap layer of the NV gate stack during the first oxidation step, and

removing the first blocking oxide in the NV gate stack.

17. A method, comprising:

dividing a wafer into a plurality of first regions and a plurality of second regions;

forming a dielectric stack on the wafer, wherein the dielectric stack includes a bottom oxide on the wafer, a multi-layer charge-trapping film, a bottom and a top nitride cap layers overlaying the multi-layer charge-trapping film, and a top sacrificial oxide;

patterning the dielectric stack to form a non-volatile (NV) gate stack of a silicon-oxide-nitride-silicon (SONOS) transistor in each of the first regions while concurrently removing the dielectric stack in each of the second regions;

performing a two-step gate oxidation process including a dry RTO process and a wet radical ISSG process, wherein

the dry RTO process forms a first gate oxide of at least one metal-oxide-semiconductor (MOS) transistor in each of the second regions, and

the wet radical ISSG process, which is performed immediately after the dry RTO process, to concurrently oxidize at least the bottom nitride cap layer to form a top oxide for the NV gate stack in each of the first regions and continue to form oxide onto and add thickness to the first gate oxide of the at least one metal-oxide-semiconductor (MOS) transistor in each of the second regions.

18. The method of claim 17, wherein after the performing the two-step gate oxidation process, a ratio between thicknesses of the top oxide of the NV gate stack and the first gate oxide of the at least one MOS transistor is in an approximate range of 1:2.33 to 1:6.67.

19. A method of controlling a blocking oxide thickness of a SONOS transistor in an integrated complementary metal-oxide-semiconductor (CMOS) process flow, the method comprising:

dividing a wafer into a first region and a second region;

forming at least one SONOS transistor in the first region, wherein each SONOS transistor includes the blocking oxide having a first operational thickness; and

forming at least one MOS transistor in the second region, wherein the at least one MOS transistor includes a high voltage (HV) MOS transistor, wherein the HV MOS transistor includes a gate oxide having a second operational thickness,

wherein the at least one SONOS transistor and the at least one MOS transistor are formed concurrently, and

wherein the blocking oxide of each of the SONOS transistors and the gate oxide of the HV MOS transistor are formed concurrently in a two-step oxidation process, the oxidation process comprising

a dry RTO process to grow the gate oxide of the HV MOS transistor to a first thickness, and

a wet radical ISSG process to concurrently form the blocking oxide of each of the at least one SONOS transistor until its thickness approximates the first operational thickness and continue to grow oxide onto the gate oxide of the HV MOS transistor to a second thickness, wherein the second thickness is greater than the first thickness and approaches the second operational thickness of the HV MOS transistor.

20. The method of claim 19, wherein the first operational thickness of the blocking oxide of the at least one SONOS transistor is in an approximate range of 30Å to 45Å while the second operational thickness of the gate oxide of the HV MOS transistor is in an approximate range of 105Å to 200Å, and wherein a ratio between the first operational thickness of the

blocking oxide and the second operational thickness of the gate oxide is in an approximate range of 1:2.33 to 1:6.67.

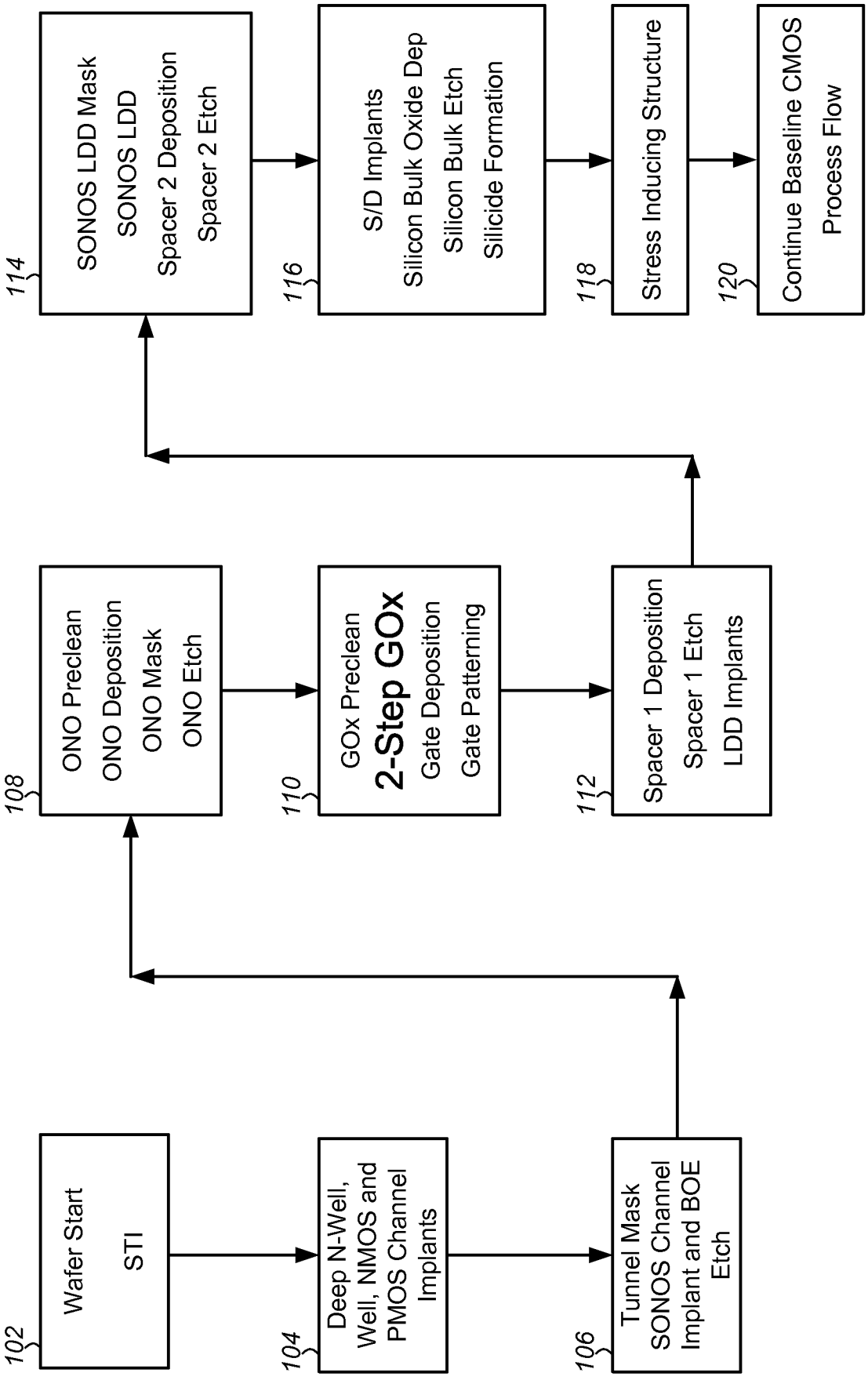

102
Wafer Start
STI
104
Deep N-Well, Well, NMOS and PMOS Channel Implants
106
Tunnel Mask SONOS Channel Implant and BOE Etch
108
ONO Preclean
ONO Deposition
ONO Mask
ONO Etch
110
GOx Preclean
2-Step GOx
Gate Deposition
Gate Patterning
112
Spacer 1 Deposition
Spacer 1 Etch
LDD Implants
114
SONOS LDD Mask
SONOS LDD
Spacer 2 Deposition
Spacer 2 Etch
116
S/D Implants
Silicon Bulk Oxide Dep
Silicon Bulk Etch
Silicide Formation
118
Stress Inducing Structure
120
Continue Baseline CMOS Process Flow

FIG. 1

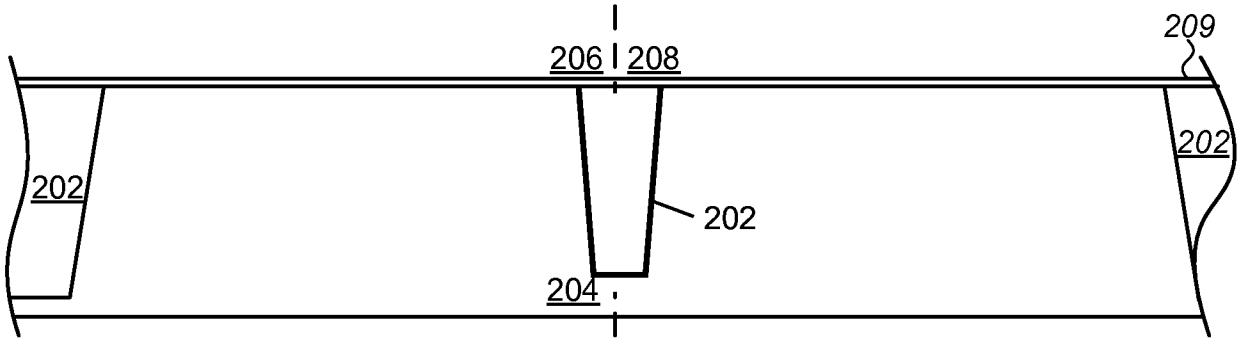


FIG. 2A

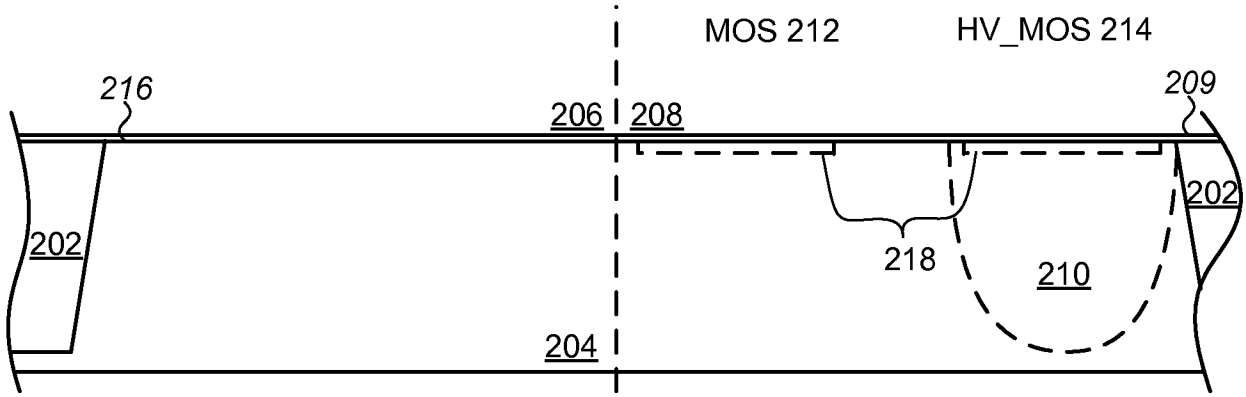


FIG. 2B

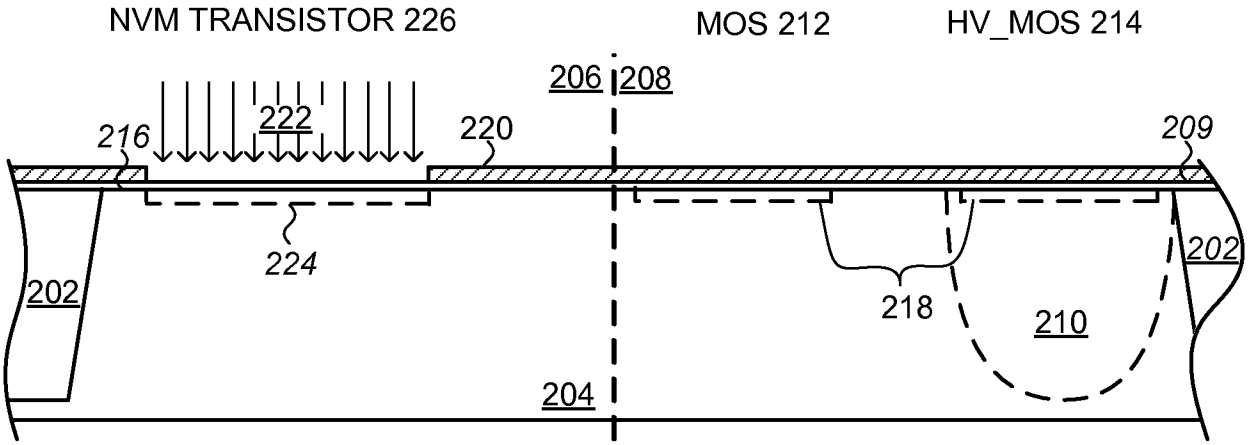


FIG. 2C

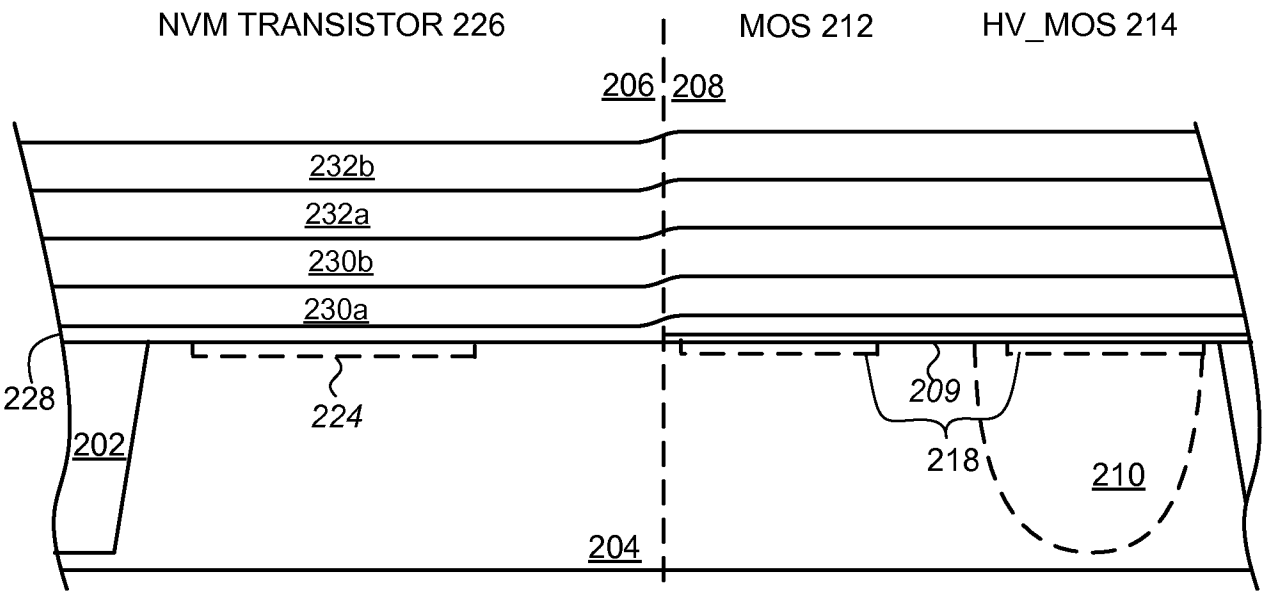

NVM TRANSISTOR 226
MOS 212
HV_MOS 214
206
208
232b
232a
230b
230a
228
224
209
202
218
210
204

FIG. 2D

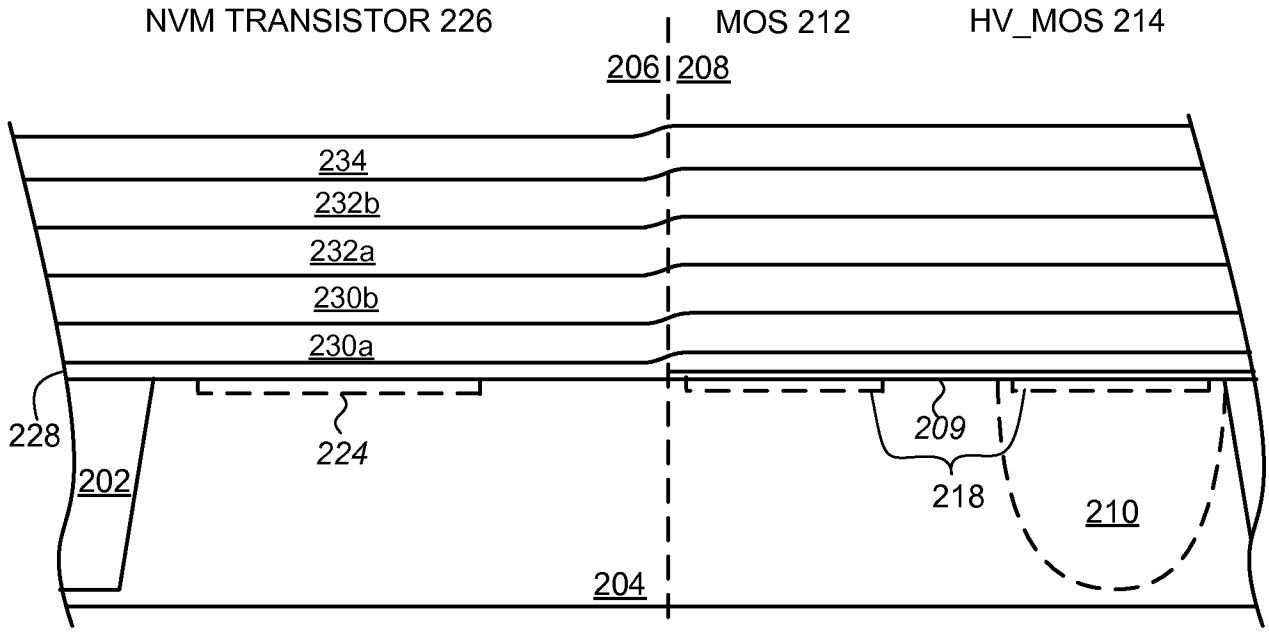

NVM TRANSISTOR 226
MOS 212
HV_MOS 214
206
208
234
232b
232a
230b
230a
228
224
209
202
218
210
204

FIG. 2E

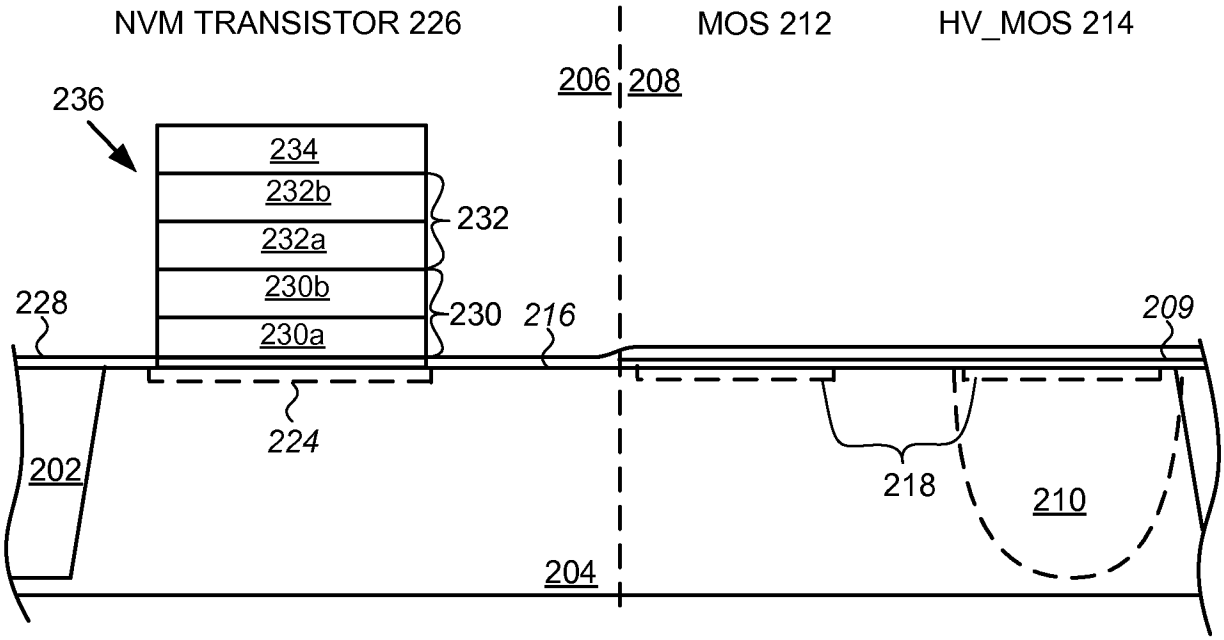

NVM TRANSISTOR 226
MOS 212
HV_MOS 214
206
208
236
234
232b
232
232a
230b
230
230a
228
216
209
224
202
218
210
204

FIG. 2F

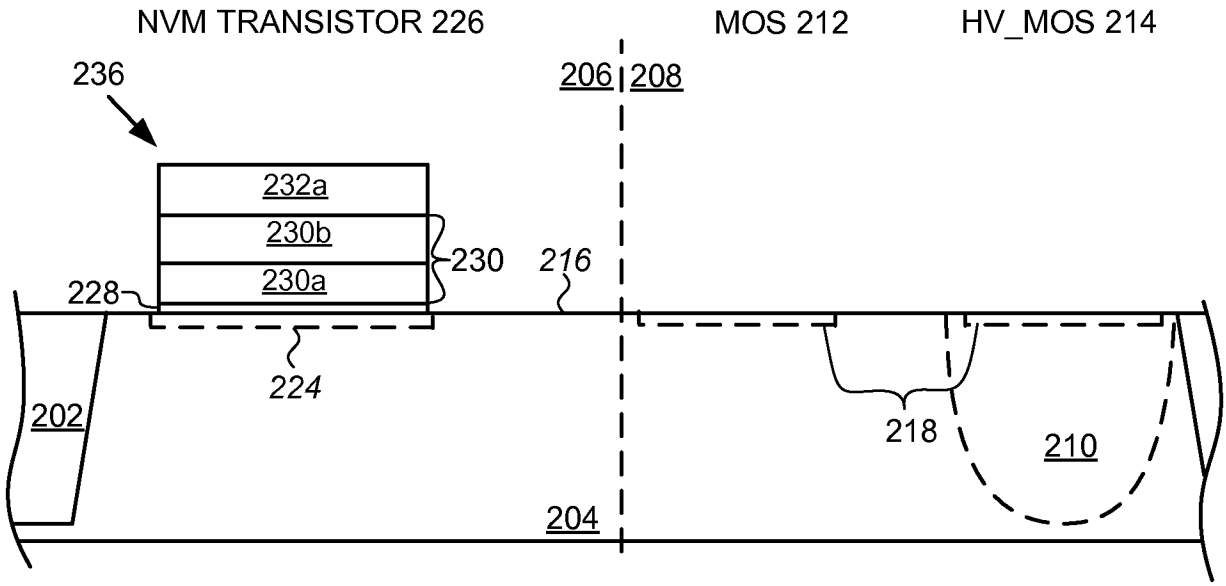

NVM TRANSISTOR 226
MOS 212
HV_MOS 214
206
208
236
232a
230b
230
230a
216
228
224
202
218
210
204

FIG. 2G

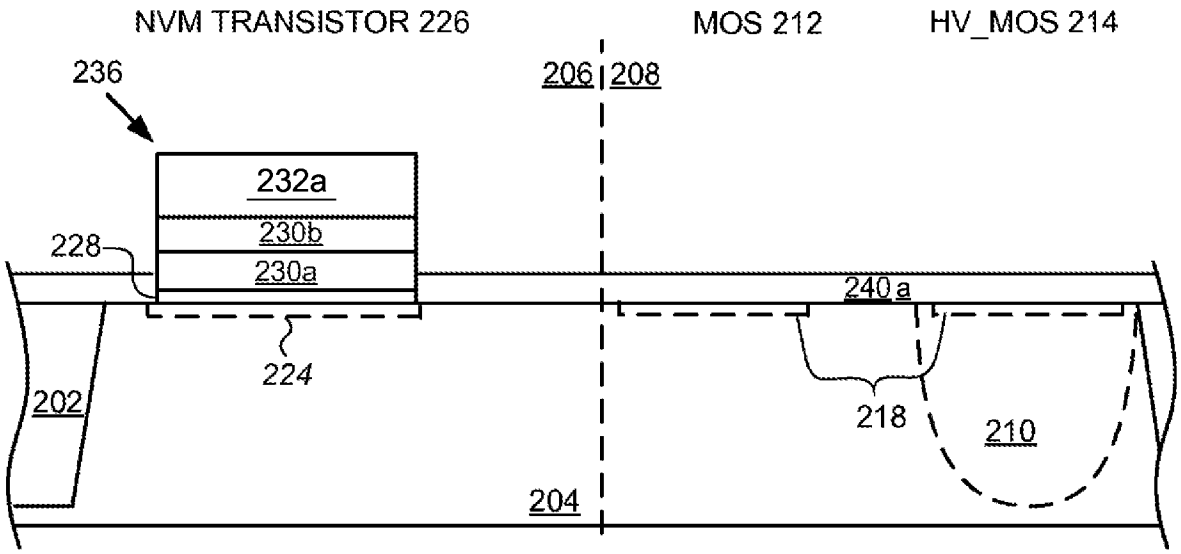

NVM TRANSISTOR 226
MOS 212
HV_MOS 214
236
206
208
232a
230b
230a
228
240a
224
218
210
202
204

FIG. 2H

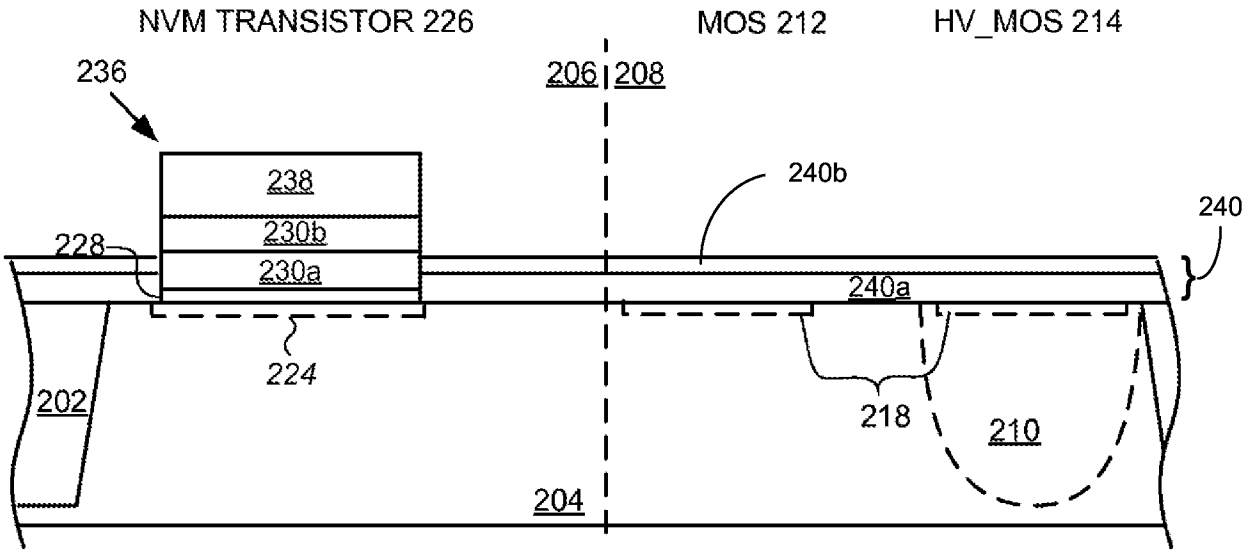

NVM TRANSISTOR 226
MOS 212
HV_MOS 214
236
206
208
238
230b
230a
228
240b
240
240a
224
218
210
202
204

FIG. 2I

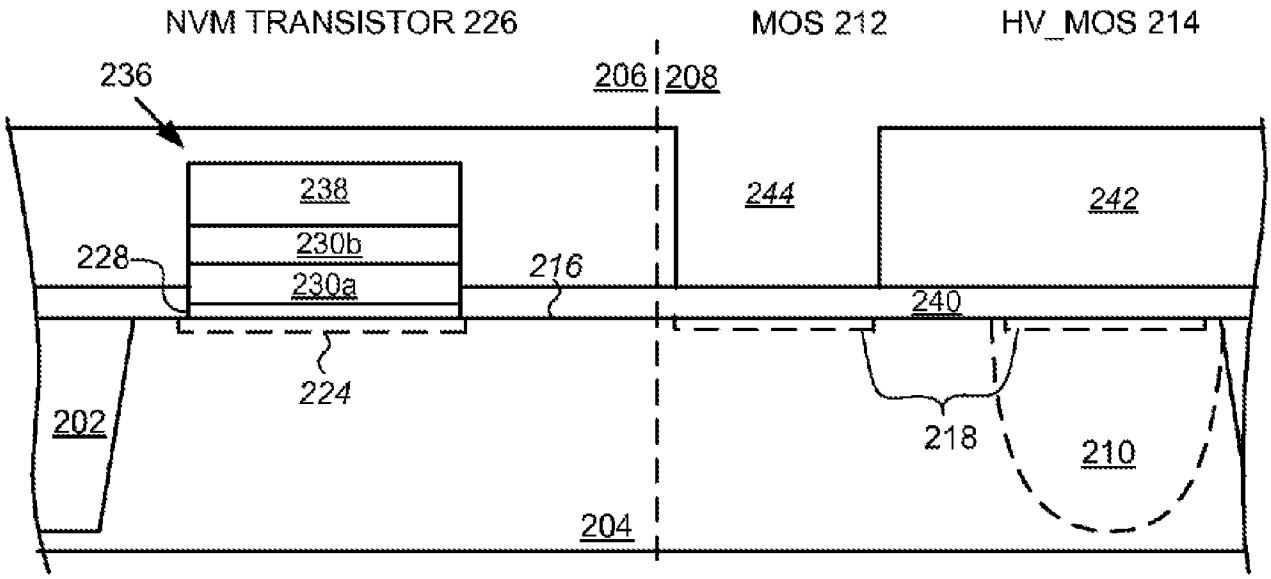

NVM TRANSISTOR 226
MOS 212
HV_MOS 214
236
206
208
238
230b
230a
228
216
244
242
240
224
218
202
210
204

FIG. 2J

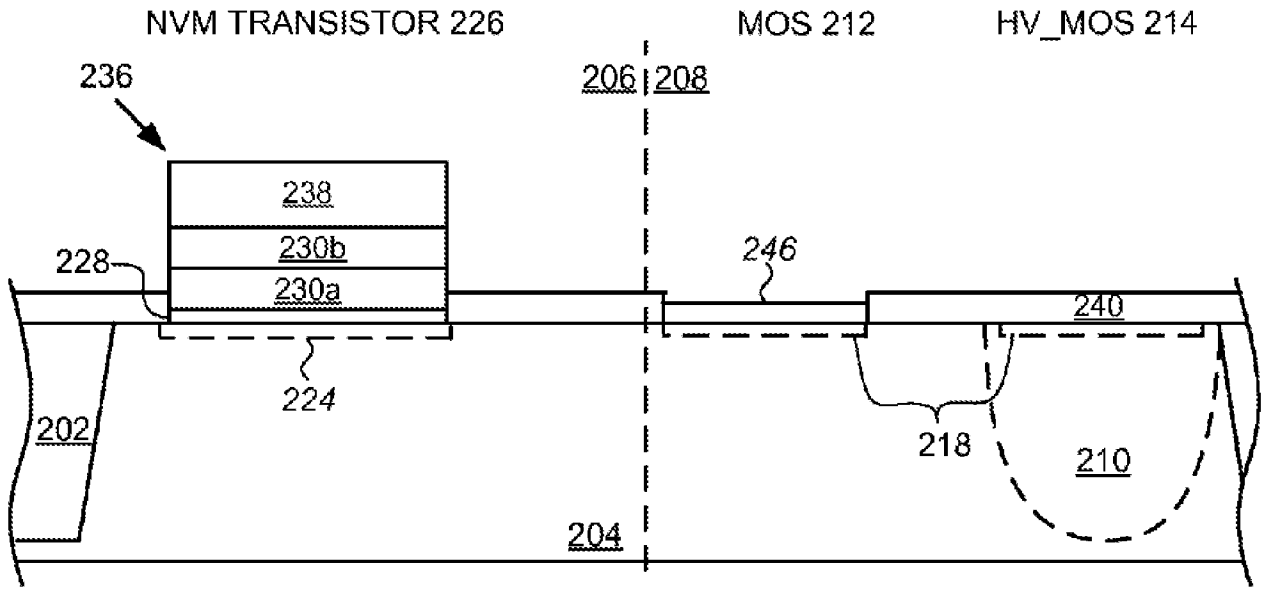

NVM TRANSISTOR 226
MOS 212
HV_MOS 214
236
206
208
238
230b
230a
228
246
240
224
218
202
210
204

FIG. 2K

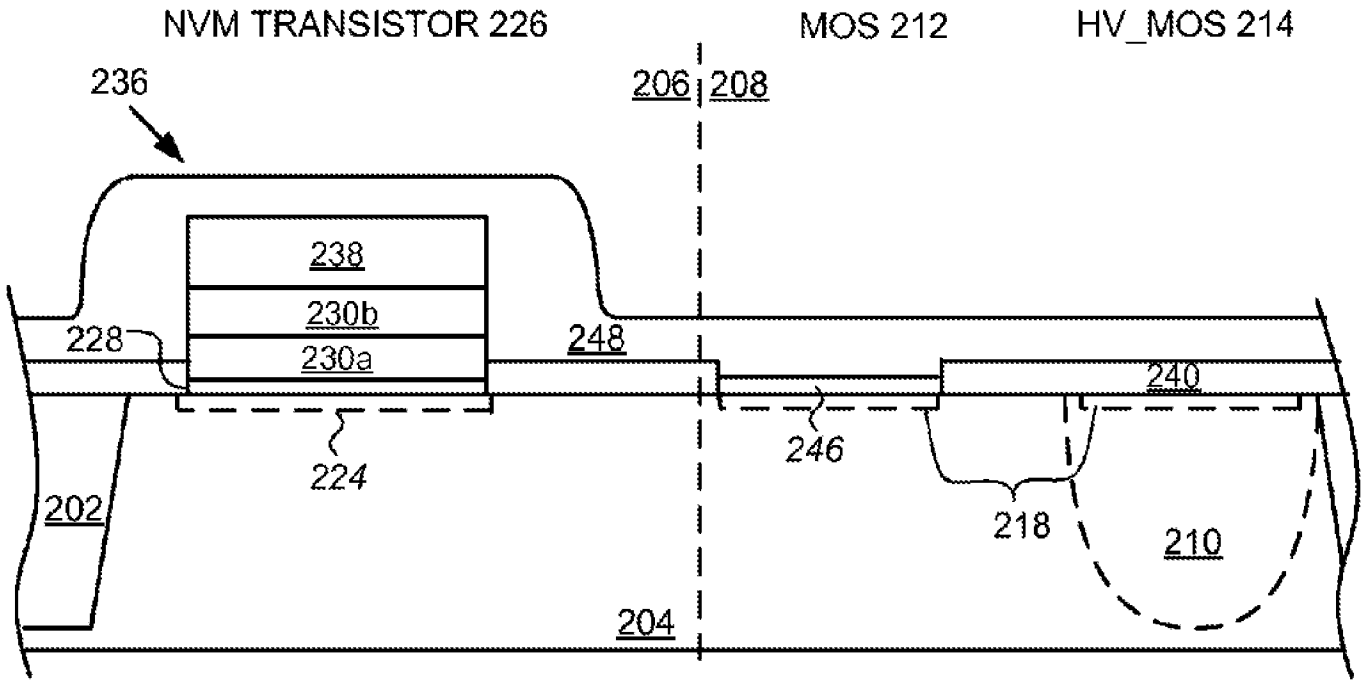

NVM TRANSISTOR 226
MOS 212
HV_MOS 214
236
206
208
238
230b
230a
228
248
240
224
246
218
210
202
204

FIG. 2L

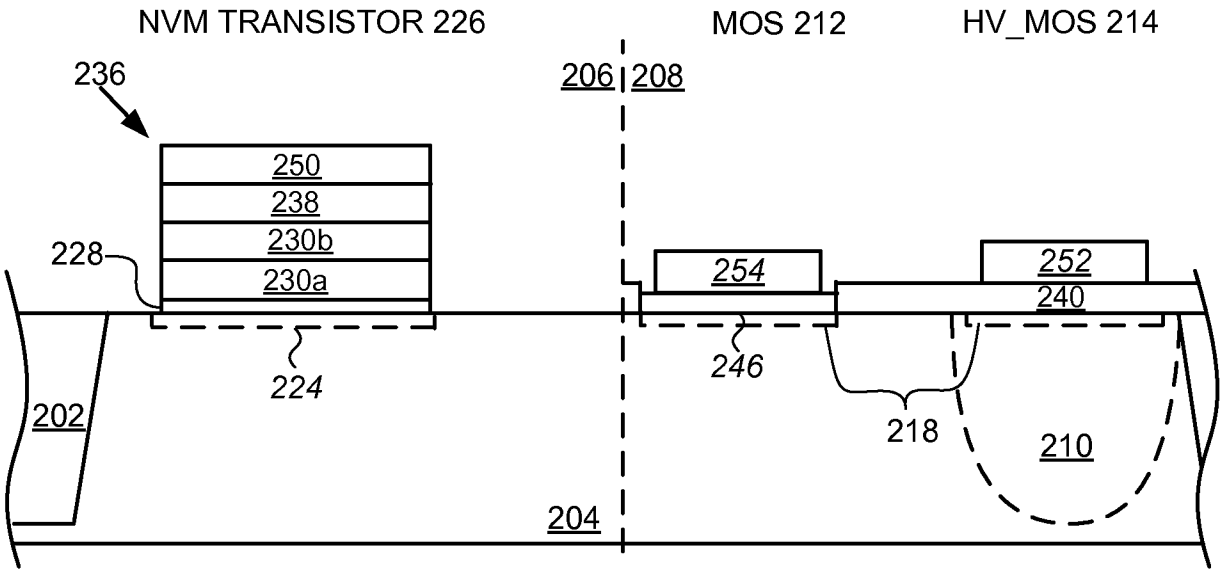

NVM TRANSISTOR 226
MOS 212
HV_MOS 214
236
206
208
250
238
230b
230a
228
254
252
240
224
246
218
210
202
204

FIG. 2M

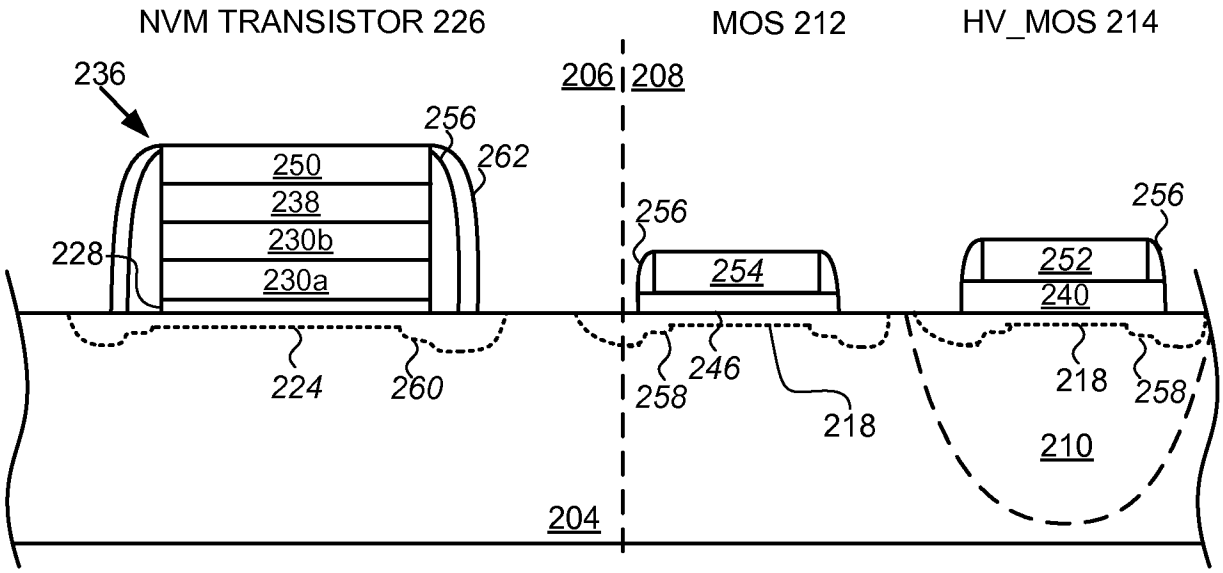

NVM TRANSISTOR 226
MOS 212
HV_MOS 214
236
206
208
256
262
250
238
230b
230a
228
254
252
240
224
260
246
258
218
210
204

FIG. 2N

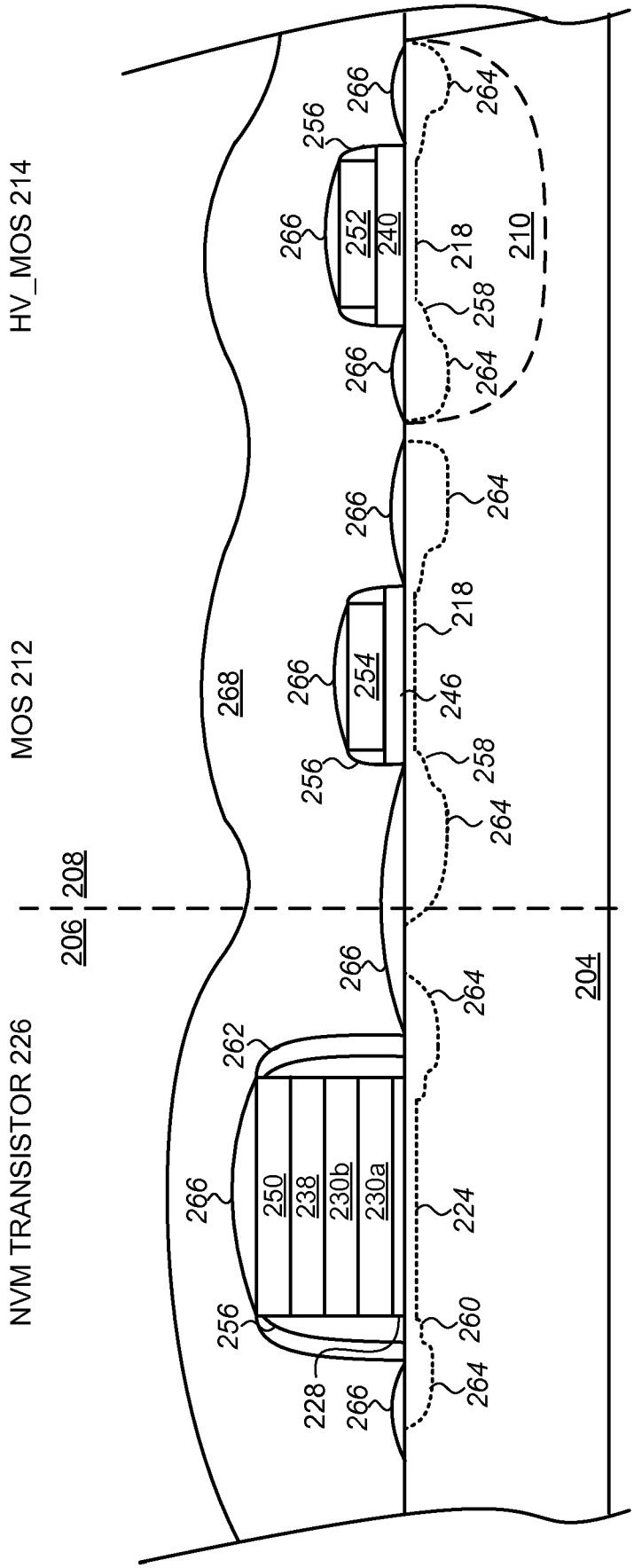

HV_MOS 214
MOS 212
NVM TRANSISTOR 226
206
208
204
210
218
224
228
230a
230b
238
240
246
250
252
254
256
258
260
262
264
266
268

FIG. 2O

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US15/62490

A. CLASSIFICATION OF SUBJECT MATTER
IPC: **H01L 21/8234**(2006.01)

USPC: 438/200
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 438/200

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2013/0210209 A1 (Ramkumar) 15 August 2013, see entire document.	1-20

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 January 2016 (28.01.2016)	03 FEB 2016
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US Commissioner for Patents P.O. Box 1450 Alexandria, Virginia 22313-1450 Facsimile No. (571) 273-3201	Authorized officer Ham Seungsook Telephone No. 571-272-2405

Form PCT/ISA/210 (second sheet) (April 2007)