US 20080293207A1

(19) **United States**

(12) **Patent Application Publication**

Koutny, JR. et al.

(10) **Pub. No.: US 2008/0293207 A1**

(43) **Pub. Date: Nov. 27, 2008**

(54) **INTEGRATION OF NON-VOLATILE CHARGE TRAP MEMORY DEVICES AND LOGIC CMOS DEVICES**

(76) Inventors: **William W.C. Koutny, JR.**, Santa Clara, CA (US); **Sam Geha**, Cupertino, CA (US); **Igor Kouznetsov**, San Jose, CA (US); **Krishnaswamy Ramkumar**, San Jose, CA (US); **Fredrick B. Jenne**, Sunnyvale, CA (US); **Sagy Levy**, Zichron (IL); **Ravindra Kapre**, San Jose, CA (US); **Jeremy Warren**, Apple Valley, MN (US)

Correspondence Address:
BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP
1279 OAKMEAD PARKWAY
SUNNYVALE, CA 94085 (US)

(21) Appl. No.: **12/125,864**

(22) Filed: **May 22, 2008**

Related U.S. Application Data

(60) Provisional application No. 60/940,148, filed on May 25, 2007, provisional application No. 60/940,137, filed on May 25, 2007.

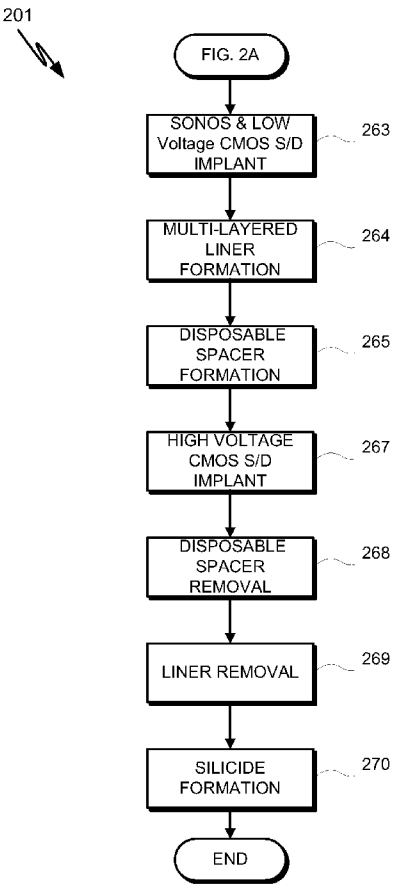
Publication Classification

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

(52) **U.S. Cl.** **438/306**; 257/E21.409

(57) **ABSTRACT**

A semiconductor structure and method to form the same. The semiconductor structure includes a substrate having a non-volatile charge trap memory device disposed on a first region and a logic device disposed on a second region. A charge trap dielectric stack may be formed subsequent to forming wells and channels of the logic device. HF pre-cleans and SC1 cleans may be avoided to improve the quality of a blocking layer of the non-volatile charge trap memory device. The blocking layer may be thermally reoxidized or nitridized during a thermal oxidation or nitridation of a logic MOS gate insulator layer to densify the blocking layer. A multi-layered liner may be utilized to first offset a source and drain implant in a high voltage logic device and also block silicidation of the nonvolatile charge trap memory device.



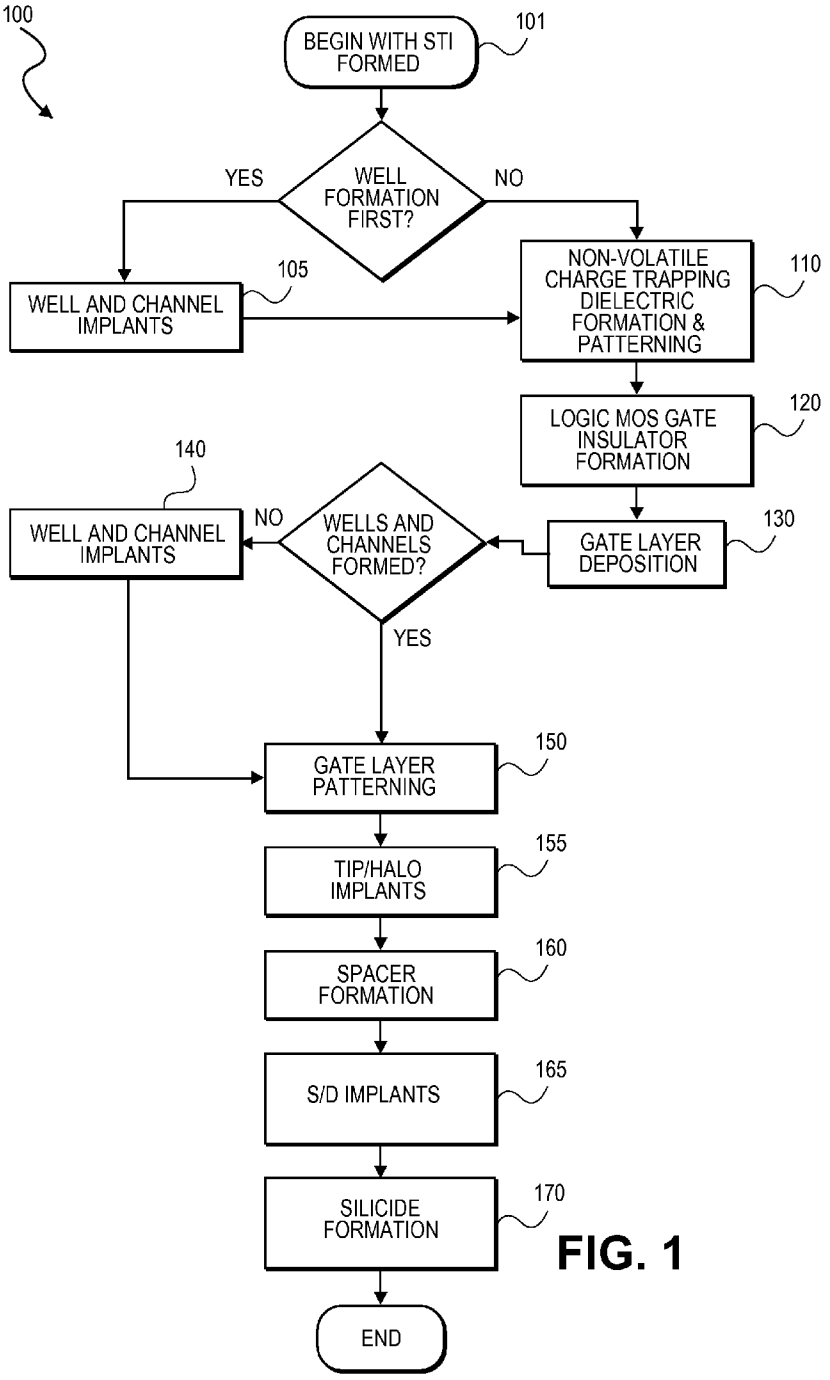

100
BEGIN WITH STI FORMED
101
WELL FORMATION FIRST?
YES
NO
WELL AND CHANNEL IMPLANTS
105
NON-VOLATILE CHARGE TRAPPING DIELECTRIC FORMATION & PATTERNING
110
LOGIC MOS GATE INSULATOR FORMATION
120
GATE LAYER DEPOSITION
130
WELLS AND CHANNELS FORMED?
NO
WELL AND CHANNEL IMPLANTS
140
YES
GATE LAYER PATTERNING
150
TIP/HALO IMPLANTS
155
SPACER FORMATION
160
S/D IMPLANTS
165
SILICIDE FORMATION
170
END

FIG. 1

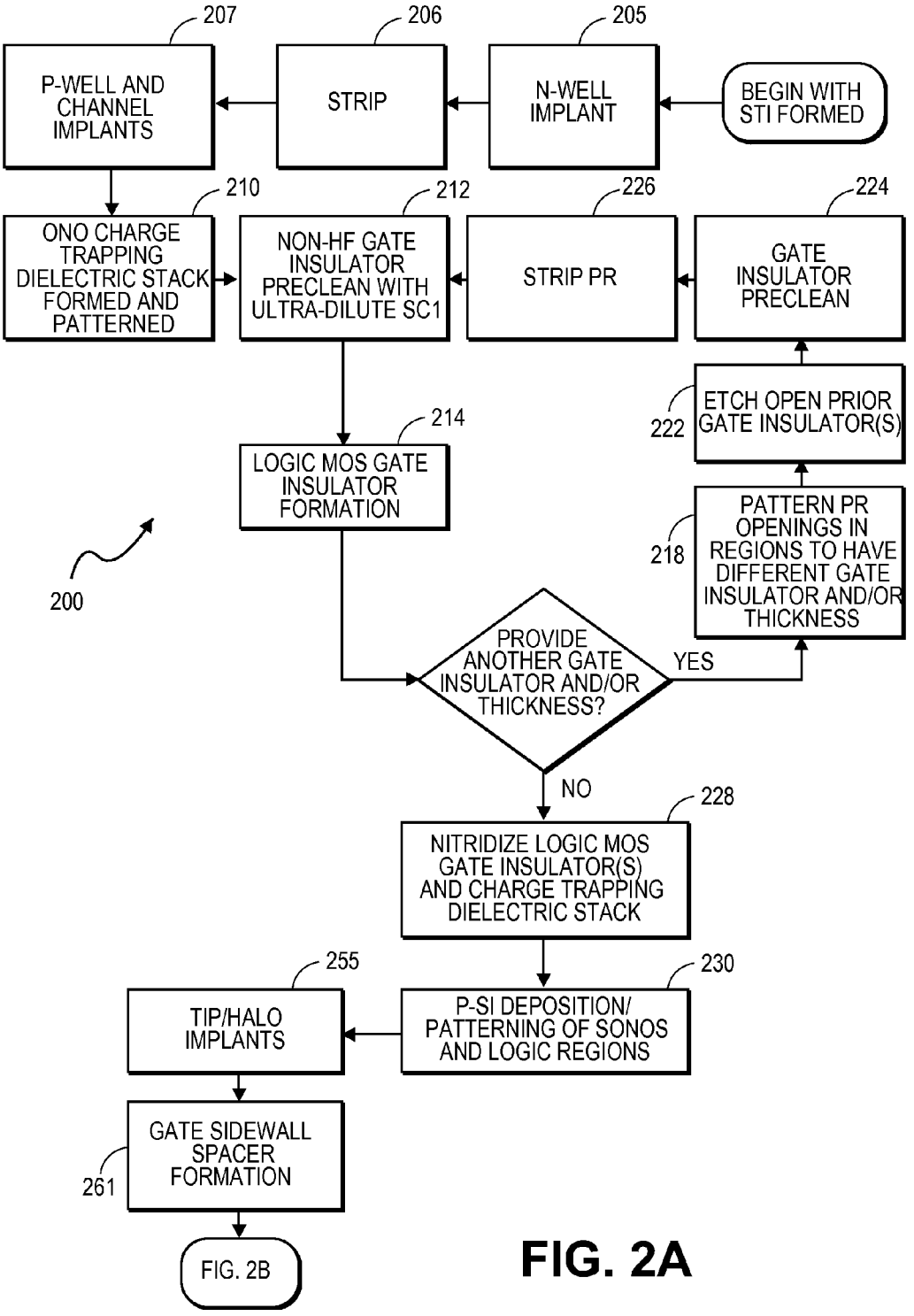

207
206
205
P-WELL AND CHANNEL IMPLANTS
STRIP
N-WELL IMPLANT
BEGIN WITH STI FORMED
210
212
226
224
ONO CHARGE TRAPPING DIELECTRIC STACK FORMED AND PATTERNED
NON-HF GATE INSULATOR PRECLEAN WITH ULTRA-DILUTE SC1
STRIP PR
GATE INSULATOR PRECLEAN
222
ETCH OPEN PRIOR GATE INSULATOR(S)
214
LOGIC MOS GATE INSULATOR FORMATION
218
PATTERN PR OPENINGS IN REGIONS TO HAVE DIFFERENT GATE INSULATOR AND/OR THICKNESS
200
PROVIDE ANOTHER GATE INSULATOR AND/OR THICKNESS?
YES
NO
228
NITRIDIZE LOGIC MOS GATE INSULATOR(S) AND CHARGE TRAPPING DIELECTRIC STACK
255
230
TIP/HALO IMPLANTS
P-SI DEPOSITION/ PATTERNING OF SONOS AND LOGIC REGIONS
GATE SIDEWALL SPACER FORMATION
261
FIG. 2B

FIG. 2A

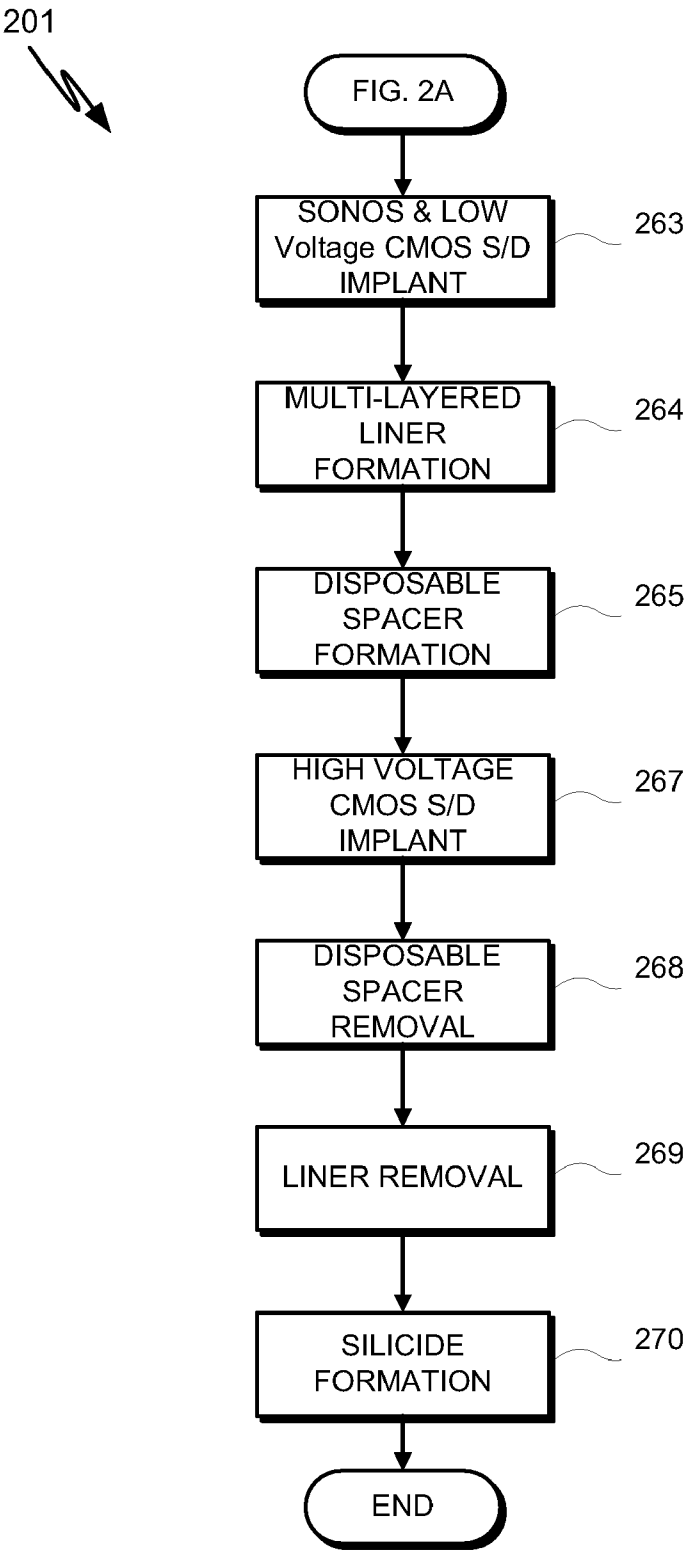

201
FIG. 2A
SONOS & LOW Voltage CMOS S/D IMPLANT
263
MULTI-LAYERED LINER FORMATION
264
DISPOSABLE SPACER FORMATION
265
HIGH VOLTAGE CMOS S/D IMPLANT
267
DISPOSABLE SPACER REMOVAL
268
LINER REMOVAL
269
SILICIDE FORMATION
270
END

FIG. 2B

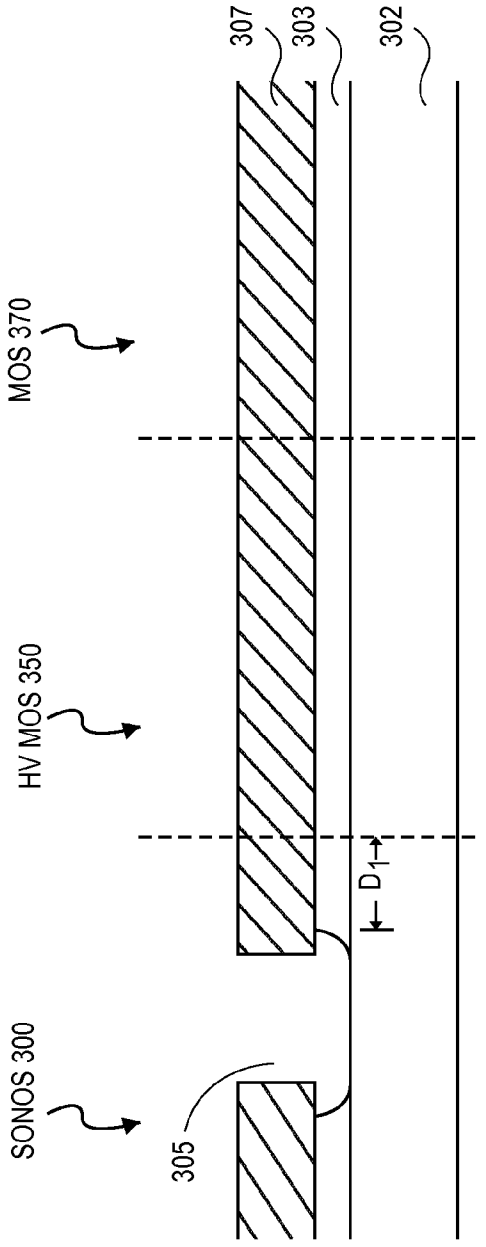

MOS 370
HV MOS 350
SONOS 300
305
307
303
302
D_1

FIG. 3A

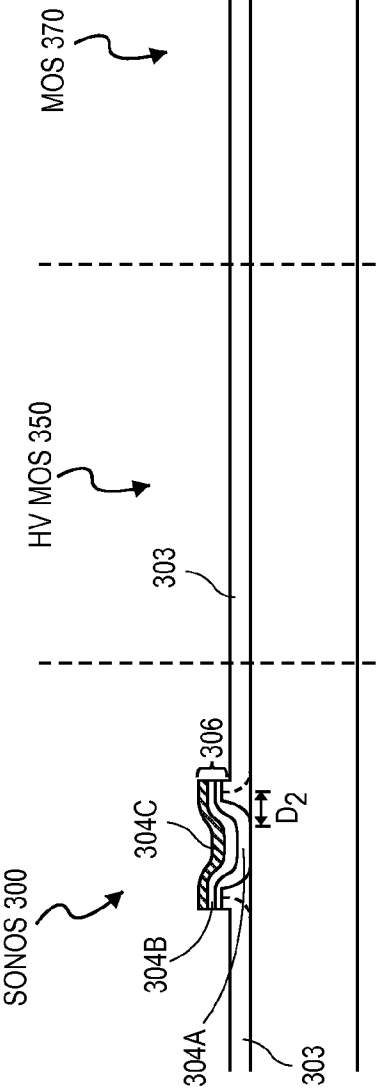

MOS 370
HV MOS 350
SONOS 300
303
306
304C
304B
304A
D_2

FIG. 3B

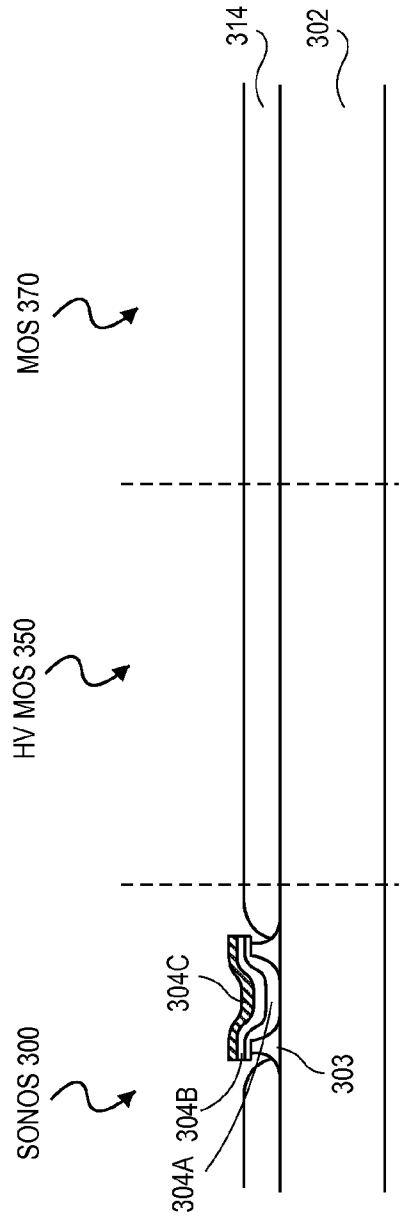

314
302
MOS 370
HV MOS 350
SONOS 300
304C
304B
304A
303

FIG. 3C

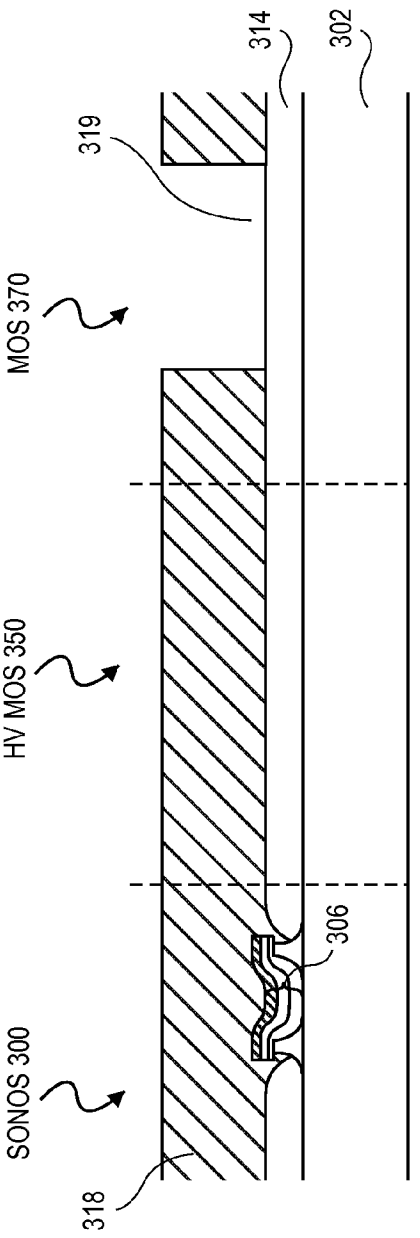

314
302
319
MOS 370
HV MOS 350
SONOS 300
306
318

FIG. 3D

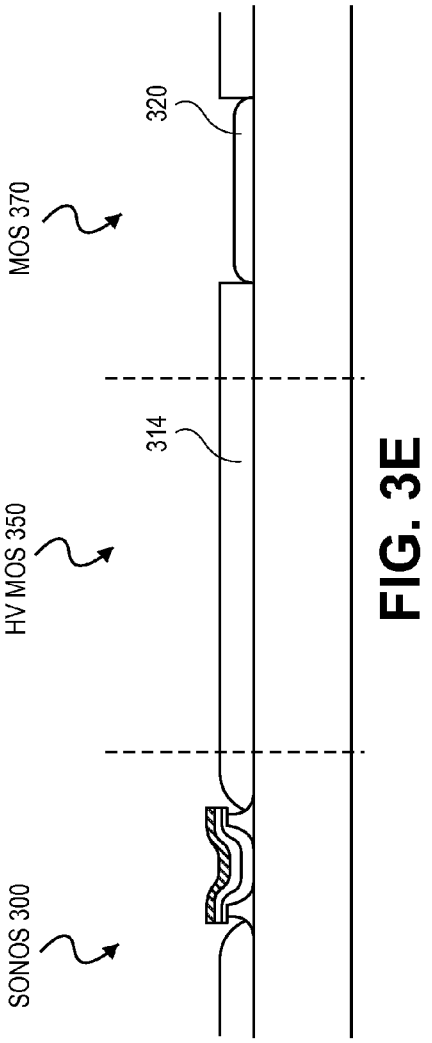

MOS 370
HV MOS 350
SONOS 300
320
314

FIG. 3E

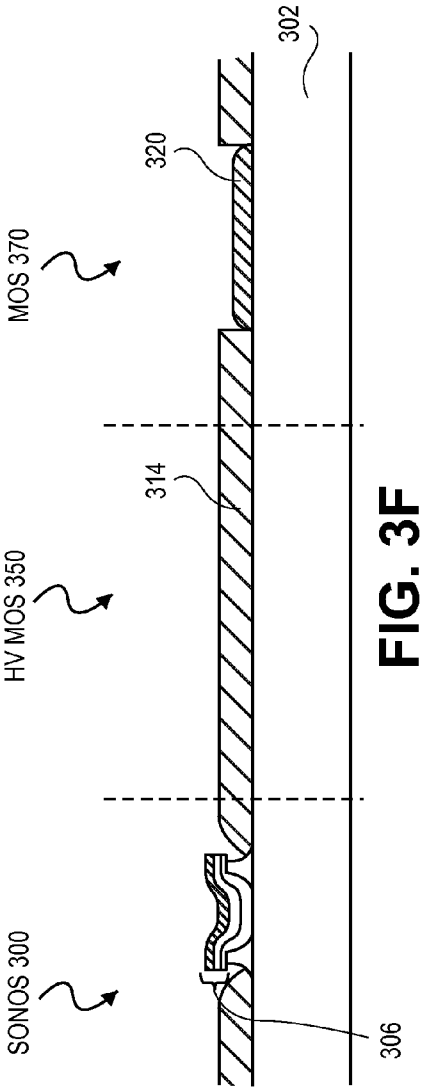

MOS 370
HV MOS 350
SONOS 300
320
314
302
306

FIG. 3F

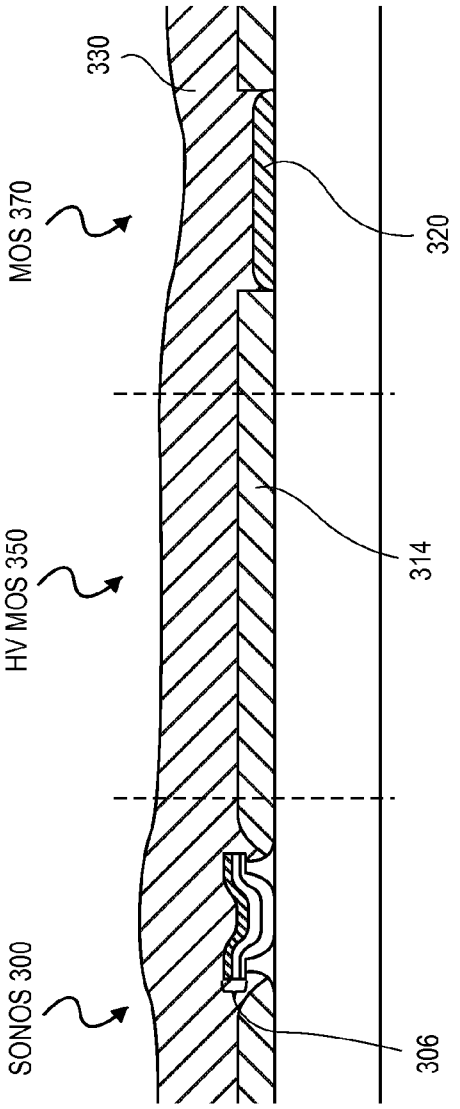

330
MOS 370
320
HV MOS 350
314
SONOS 300
306

FIG. 3G

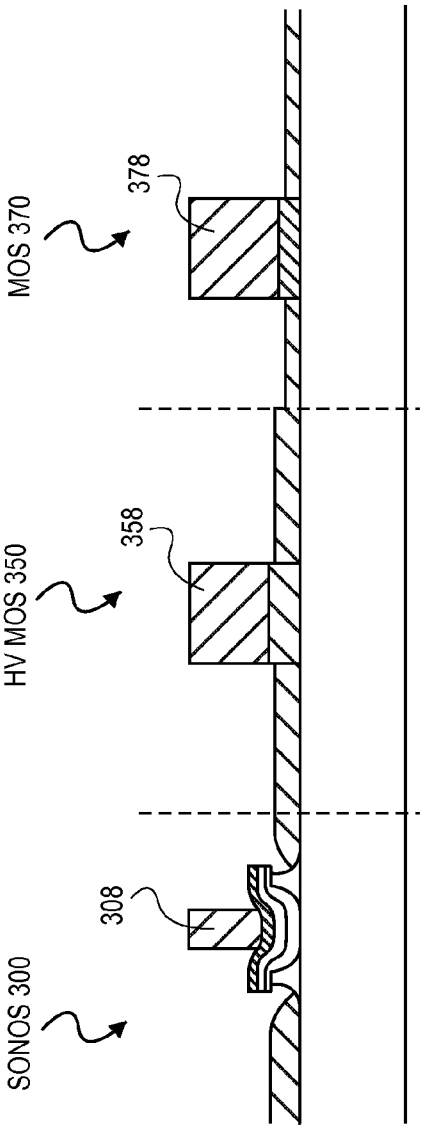

MOS 370
378
HV MOS 350
358
308
SONOS 300

FIG. 3H

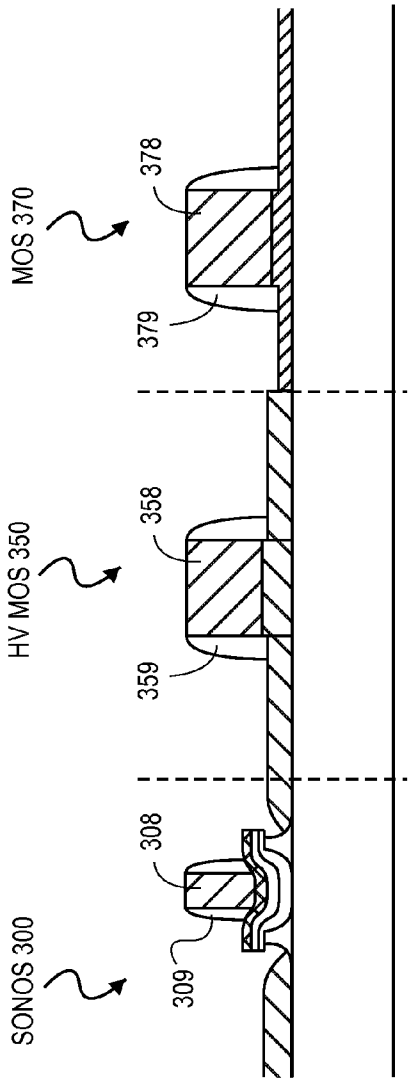

SONOS 300
309
308
HV MOS 350
359
358
MOS 370
379
378

FIG. 3I

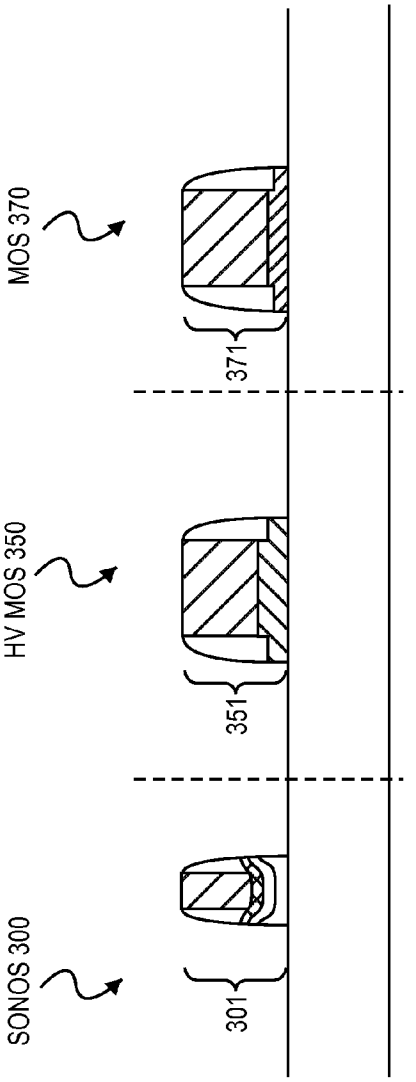

SONOS 300
301
HV MOS 350
351
MOS 370
371

FIG. 3J

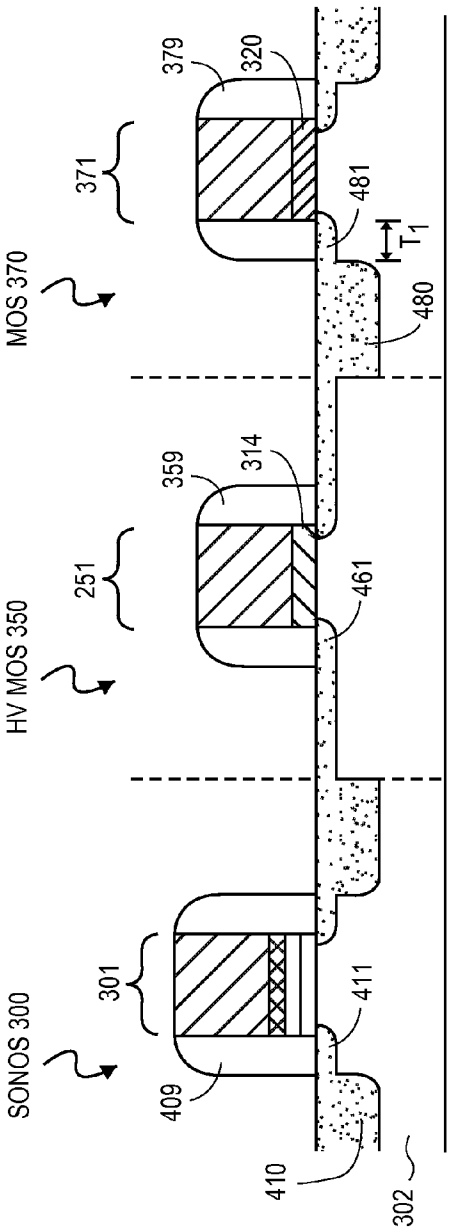

SONOS 300
HV MOS 350
MOS 370
301
251
371
409
359
379
411
314
320
461
481
480
T1
410
302

FIG. 4A

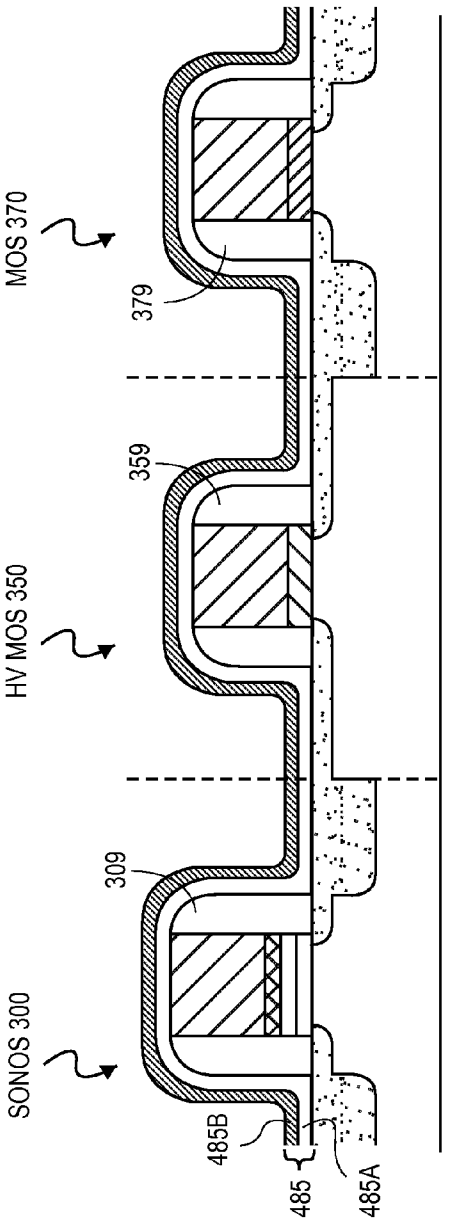

SONOS 300
HV MOS 350
MOS 370
309
359
379
485B
485
485A

FIG. 4B

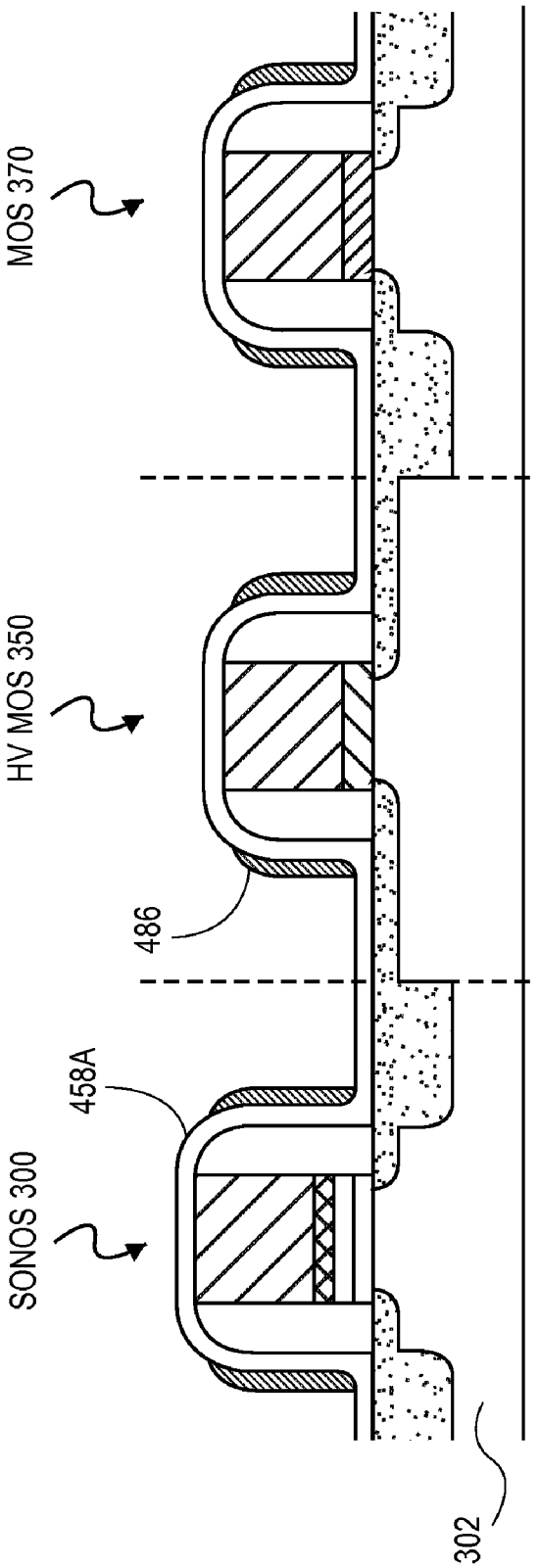


FIG. 4C

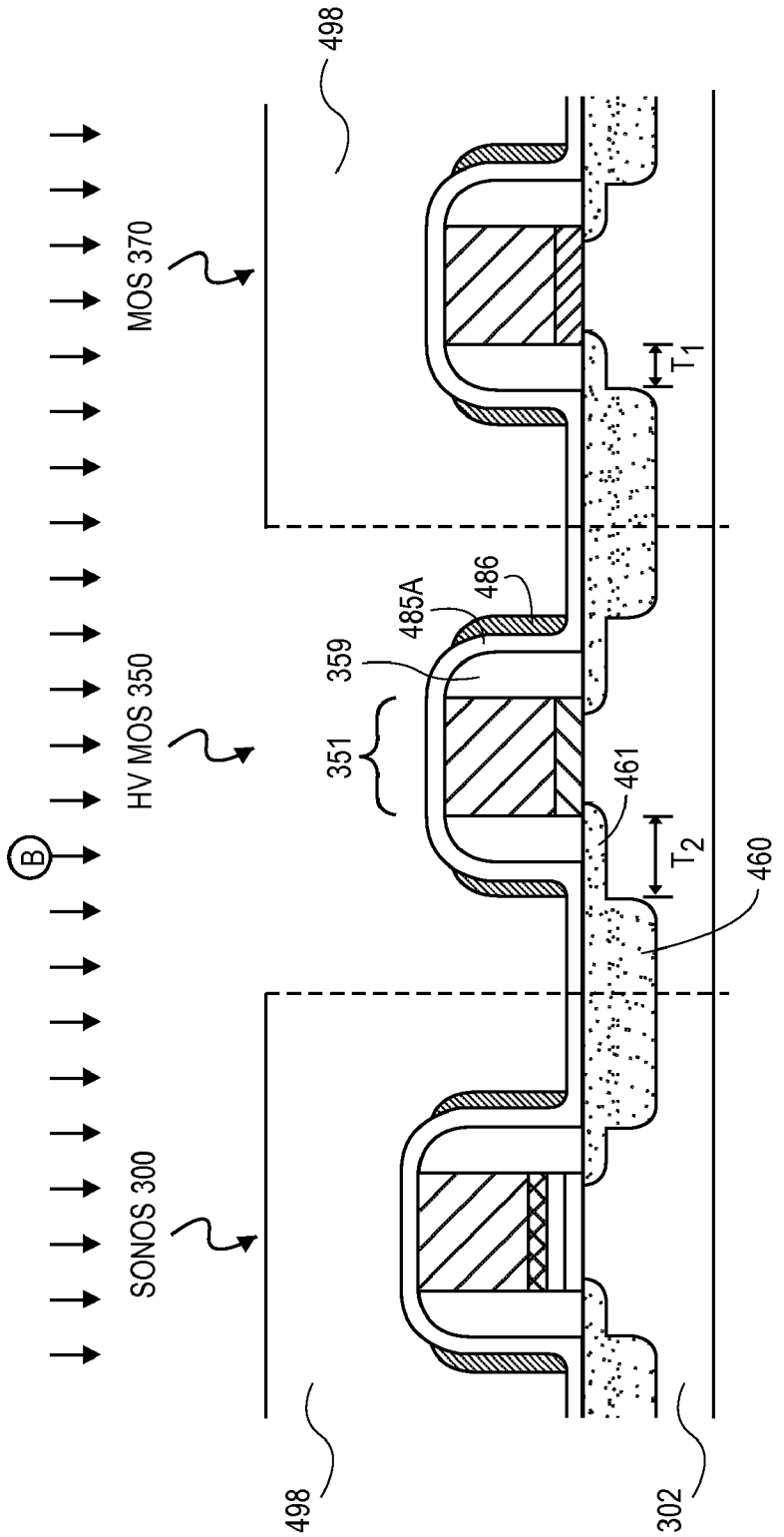

498
MOS 370
HV MOS 350
SONOS 300
B
351
359
485A
486
461
460
T1
T2
498
498
302

FIG. 4D

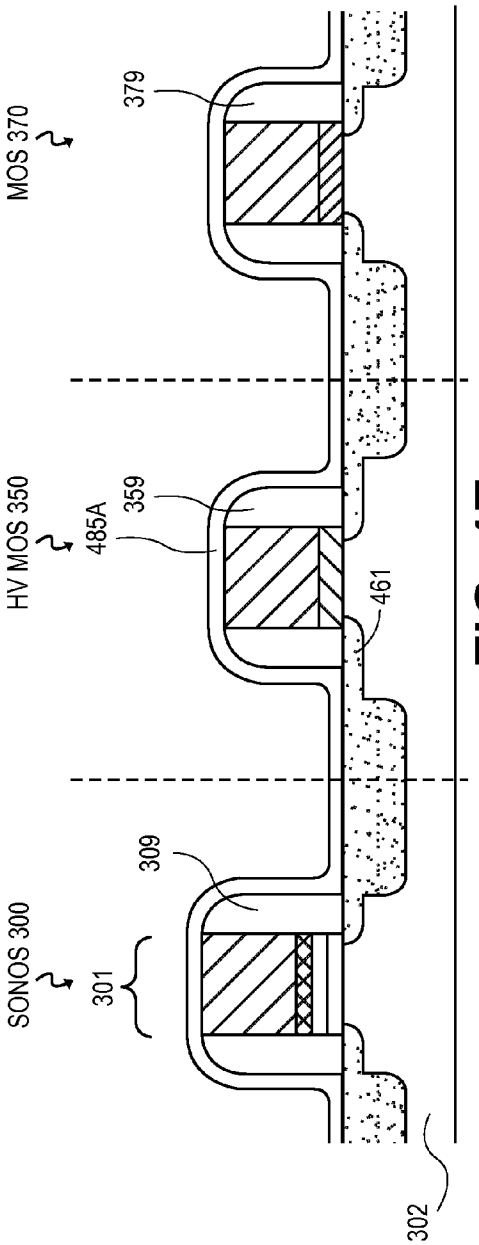

MOS 370
379
HV MOS 350
485A
359
461
SONOS 300
301
309
302

FIG. 4E

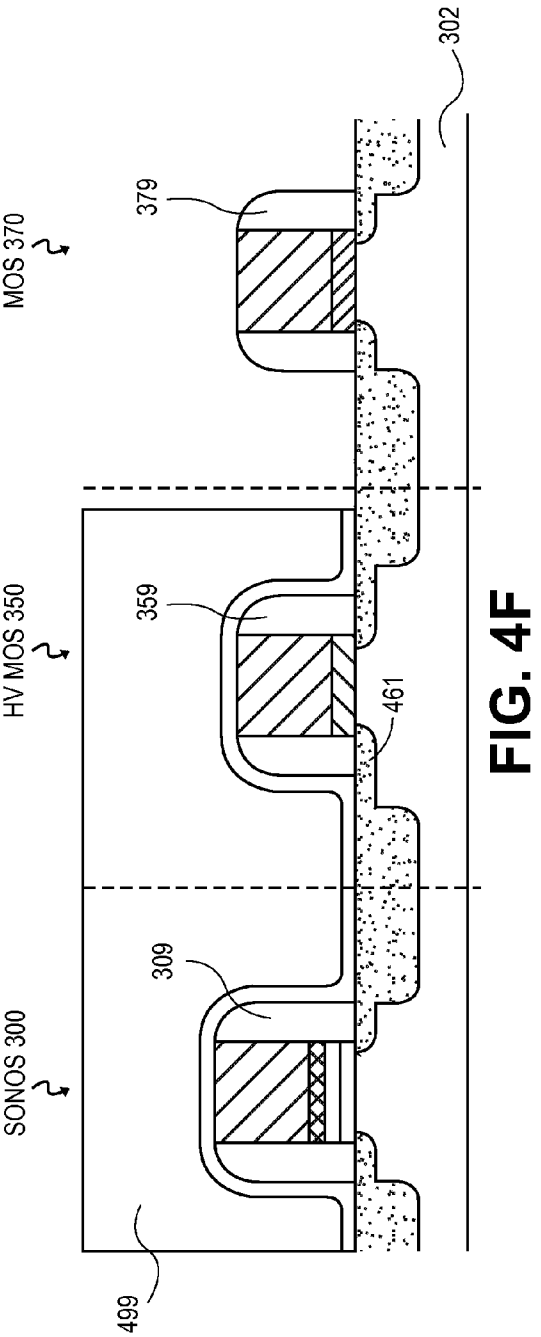

302
MOS 370
379
HV MOS 350
359
461
SONOS 300
309
499

FIG. 4F

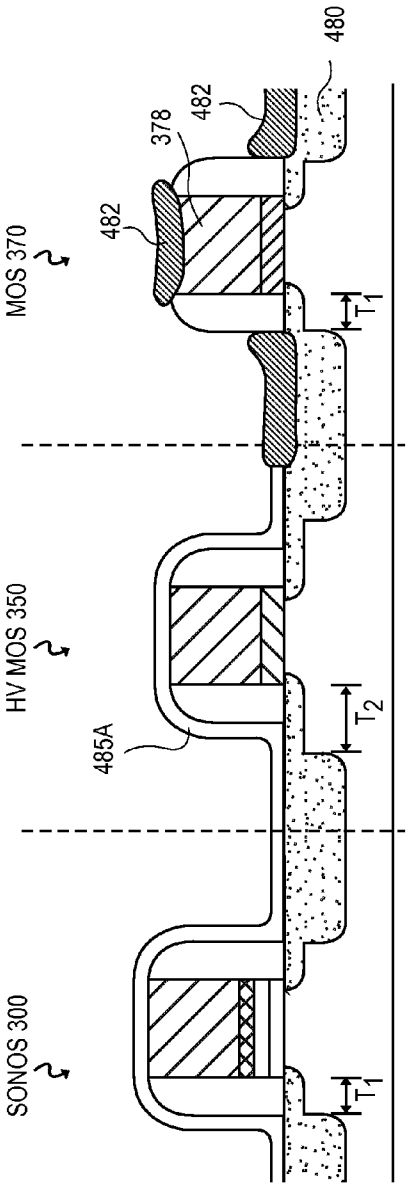

480
482
378
482
MOS 370
T1
HV MOS 350
485A
T2
SONOS 300
T1

FIG. 4G

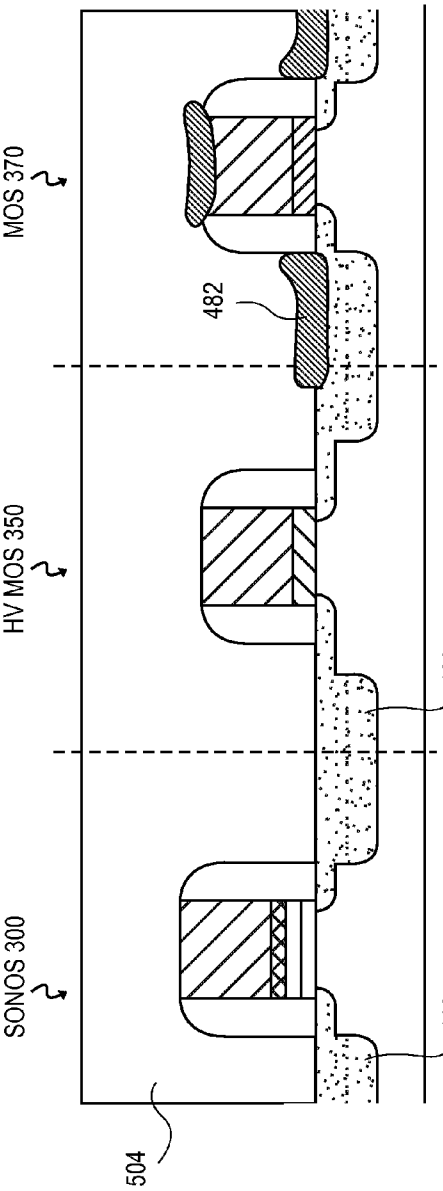

MOS 370
482
HV MOS 350
460
SONOS 300
410
504

FIG. 5

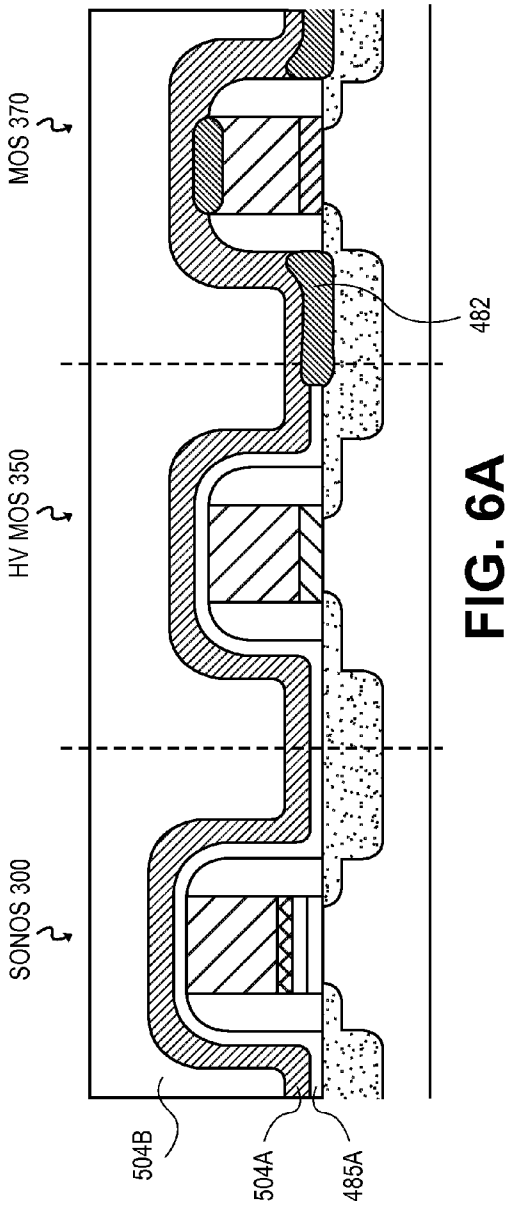

MOS 370
HV MOS 350
SONOS 300
482
504B
504A
485A

FIG. 6A

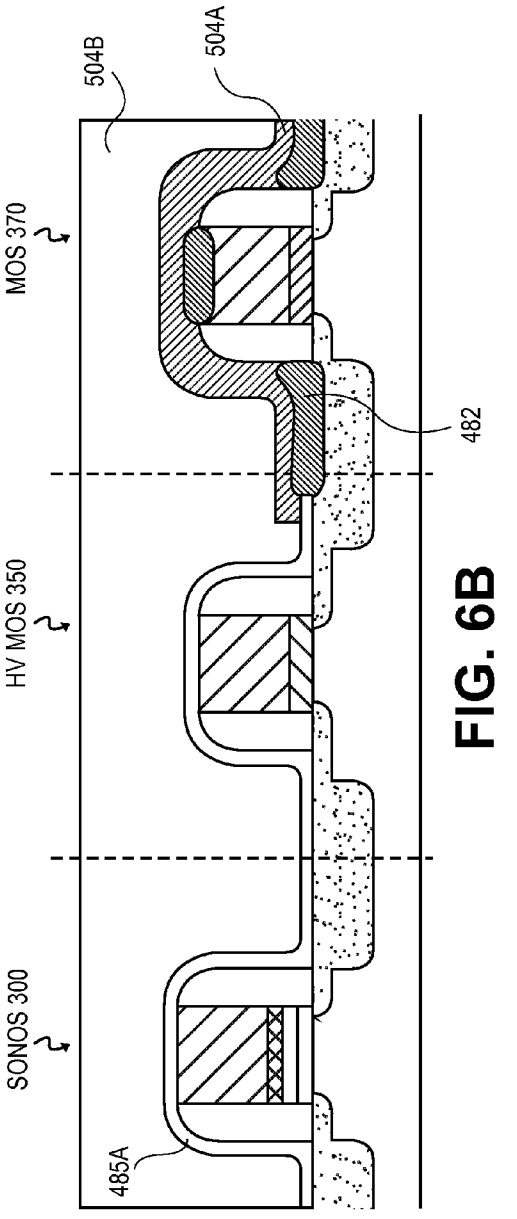

504B
504A
MOS 370
HV MOS 350
SONOS 300
482
485A

FIG. 6B

INTEGRATION OF NON-VOLATILE CHARGE TRAP MEMORY DEVICES AND LOGIC CMOS DEVICES

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 60/940,148, filed May 25, 2007, the entire contents of which are hereby incorporated by reference herein. This application also claims the benefit of U.S. Provisional Application No. 60/940,137, filed May 25, 2007, the entire contents of which are hereby incorporated by reference herein.

TECHNICAL FIELD

[0002] The invention is in the field of semiconductor devices, more specifically pertaining to non-volatile charge trap memory devices integrated with logic CMOS devices.

BACKGROUND

[0003] Feature scaling in integrated circuits is an enabler of more capable electronic devices. Scaling to smaller features increases densities of functional units in a given form factor as well as increasing device processing speeds. Device scaling, however, is not without issue. For example, optimizing the performance of smaller devices becomes increasingly difficult. This is particularly true for the scaling of non-volatile charge trap memory devices, in which data retention and sensing becomes increasingly difficult as the devices are scaled.

[0004] In addition to device scaling, system-on-a-chip type architecture also increases electronic device functionality. Such architecture may incorporate, for example, a memory device on the same substrate as a logic device to reduce the cost of fabrication as well as increase communication bandwidth between the memory and logic devices.

[0005] The integration of these dissimilar devices in a system-on-a-chip architecture is problematic because the fabrication process for the logic MOS device may hamper the fabrication process of the memory device and visa versa. Such a dilemma may occur, for example, when integrating the logic MOS gate oxide process module with the fabrication of a dielectric stack for a memory device. Also, channel and well implant processing for the logic devices may also be detrimental to the memory device dielectric stack while formation of the latter may be problematic for the former. As still another example, silicided contacts, which are advantageous for a logic transistor, may adversely affect a nonvolatile charge trap memory device.

[0006] Also, operation of a non-volatile memory device may require application of relatively high voltages (HV), typically of at least 10 V. However, the conventional processes employed in fabrication of a scaled logic device are typically optimized for device operation at 5 V or less. Such low voltage devices may lack a sufficiently high breakdown voltage to interface directly with a memory device.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] Embodiments of the present invention are illustrated by way of example, and not limitation, in the figures of the accompanying drawings in which:

[0008] FIG. **1** illustrates a flow diagram depicting sequences of particular modules employed in the fabrication process of a non-volatile charge trap memory device integrated with a logic MOS fabrication process, in accordance with particular embodiments of the present invention;

[0009] FIGS. **2**A and **2**B illustrate flow diagrams depicting sequences of particular operations in the integration of logic MOS gate fabrication with a non-volatile charge-trapping dielectric stack for implementing certain modules illustrated in FIG. **1**, in accordance with particular embodiments of the present invention; and

[0010] FIG. **3**A illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a SONOS channel implant is performed while a screening oxide is over the MOS and HV MOS regions of a substrate, in accordance with an embodiment of the present invention;

[0011] FIG. **3**B illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a SONOS charge trapping dielectric stack is formed and the MOS and HV MOS regions are cleaned in preparation for forming a first gate insulator layer, in accordance with an embodiment of the present invention;

[0012] FIG. **3**C illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a first gate insulator layer is formed over the MOS and HV MOS regions, in accordance with an embodiment of the present invention;

[0013] FIG. **3**D illustrates a cross-sectional view representing operations wherein SONOS and HV MOS device regions are masked while the first gate insulator layer in the MOS region is opened in a third region of the substrate to form a second gate insulator layer, in accordance with an embodiment of the present invention;

[0014] FIG. **3**E illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a second gate insulator layer is formed in the MOS region, in accordance with an embodiment of the present invention;

[0015] FIG. **3**F illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which the SONOS oxide blocking layer, the HV MOS gate insulator layer and the MOS gate insulator layer are nitrided, in accordance with an embodiment of the present invention;

[0016] FIG. **3**G illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a gate layer is deposited, in accordance with an embodiment of the present invention;

[0017] FIG. **3**H illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a gate electrode is formed in accordance with an embodiment of the present invention;

[0018] FIG. **3**I illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a sidewall spacer is formed in accordance with an embodiment of the present invention; and

[0019] FIG. **3**J illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which charge trap dielectric and gate dielectric is removed adjacent to sidewall spacers to complete definition of gate stacks, in accordance with an embodiment of the present invention.

[0020] FIG. **4**A illustrates a cross-sectional view representing operations in the formation of a semiconductor structure having a SONOS gate stack with adjacent sidewall spacers as wells as HV MOS and MOS device gate stacks with adjacent

UMC 1434
UMC v. AICP

sidewall spacers on a single substrate, in accordance with an embodiment of the present invention;

[0021] FIG. **4**B illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a multi-layered liner is deposited over the SONOS and logic devices, in accordance with an embodiment of the present invention;

[0022] FIG. **4**C illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which the top layer of the multi-layered liner is etched to form a disposable spacer, in accordance with an embodiment of the present invention;

[0023] FIG. **4**D illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a HV MOS device receives a source and drain implant while the SONOS and MOS devices are masked, in accordance with an embodiment of the present invention;

[0024] FIG. **4**E illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which the disposable spacer is removed from the SONOS and logic devices, in accordance with an embodiment of the present invention;

[0025] FIG. **4**F illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which the bottom layer of the multi-layered liner is removed from the MOS device but retained over the SONOS and HV MOS devices, in accordance with an embodiment of the present invention;

[0026] FIG. **4**G illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a silicide is formed on the MOS device but blocked by the bottom layer of the multi-layered liner over the SONOS and MOS devices, in accordance with an embodiment of the present invention;

[0027] FIG. **5** illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which an interlayer dielectric (ILD) layer is formed on the sidewalls of the sidewall spacers adjacent to the MOS SONOS and HV MOS gate stacks, in accordance with an embodiment of the present invention;

[0028] FIG. **6**A illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a stress inducing ILD layer is formed on the sidewall spacers adjacent to the MOS gate stack and formed on a bottom layer of the multi-layered liner over the SONOS and HV MOS gate stacks, in accordance with an embodiment of the present invention; and

[0029] FIG. **6**B illustrates a cross-sectional view representing operations in the formation of a semiconductor structure in which a low-stress ILD layer is formed on a bottom layer of the multi-layered liner covering the sidewall spacers adjacent to the SONOS and HV MOS gate stacks and formed on a stress inducing ILD layer over the MOS device, in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION

[0030] Embodiments of a non-volatile charge trap memory device integrated with logic devices 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 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 throughout this specification to "an embodiment" means that a particular feature, structure, material, or characteristic described in connection with the embodiment is included in at least one embodiment of the invention. Thus, the appearances of the phrase "in an embodiment" in various places throughout this specification are not necessarily referring to the same embodiment of the invention. Furthermore, the particular features, structures, materials, or characteristics may be combined in any suitable manner in one or more embodiments.

[0031] 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 substrate without consideration of the absolute orientation of the substrate.

[0032] In an embodiment, a nonvolatile charge trap dielectric stack, such as a SONOS stack, is formed in a first region of a substrate after at least some of the well and channel implants of logic MOS devices are formed in a second region of the substrate.

[0033] In another embodiment, a nonvolatile charge trap dielectric stack is formed prior to any logic MOS gate oxidation processing. In one such embodiment, a SONOS stack is removed from the second region of the substrate, and a thermal oxidation forms a first gate insulator layer over the second region of the semiconductor substrate and thermally reoxidizes a blocking layer of the SONOS stack. In a further embodiment, a nitridation process nitridizes the first gate insulator layer and the blocking layer simultaneously.

[0034] In another embodiment, a nonvolatile charge trap memory device without silicide contacts is integrated with a logic device having silicide contacts. Such an embodiment may advantageously improve the reliability of the nonvolatile charge trap memory device by reducing silicide-related stress in the memory device.

[0035] In a further embodiment, at least one of the logic devices has a longer lightly doped source and drain (i.e. offset source and drain) than at least another one of the logic devices to allow for HV operation (e.g. breakdown voltage greater than 10 V). In one such embodiment, wherein the logic devices include a HV PMOS device and a n-type MOS (NMOS) device, the NMOS device has a smaller source and drain offset than does the HV PMOS device. In another such embodiment, wherein the logic devices include a HV PMOS device and a PMOS device, the PMOS device has a smaller source and drain offset than does the HV PMOS. In a particular embodiment, the lightly doped source and drain of the HV MOS device is a length greater than the thickness of a sidewall spacer adjacent to a sidewall of a gate stack of the MOS device.

[0036] In another embodiment, a multi-layered liner is employed to offset the HV MOS source and drain and also protect the nonvolatile charge trap memory device from sili-

UMC 1434
UMC v. AICP

cidation. In one such embodiment, wherein the multi-layered liner includes at least a top and bottom layer, a top layer is formed into a disposable spacer to offset the HV MOS source and drain and the bottom layer is used to mask the nonvolatile charge trap memory device during a silicidation of one or more of the logic devices. In another embodiment, the bottom layer is additionally used to mask the HV MOS device during silicidation of one or more of the logic devices. In a particular embodiment, the disposable spacer is removed selectively to the bottom layer of the multi-layered liner after the HV MOS source and drain are implanted. In a further embodiment, the bottom layer of the multi-layered liner is retained over the nonvolatile charge trap memory device as an ILD layer, covered with another ILD layer and then etched through during contact formation. In another embodiment, the bottom layer of the multi-layered liner is retained over the nonvolatile charge trap memory device and the HV MOS device as an ILD layer. In one such embodiment, the stress in the bottom layer of the multi-layered liner is of opposite sign than that of a stress inducing ILD layer deposited over the bottom liner layer. In one particular embodiment, the bottom layer of the multi-layered liner induces compressive stress on the underlying device while the stress inducing ILD layer induces tensile stress on the underlying device.

[0037] FIG. **1** illustrates a flow diagram depicting sequences of particular modules employed in the fabrication process **100** of a non-volatile charge trap memory device integrated with a logic MOS device, in accordance with particular embodiments of the present invention. The methods begin with formation of isolation regions at module **101**. Isolation regions may be formed by any conventional technique, such as, but no limited to shallow trench isolation (STI) or local oxidation of silicon (LOCOS).

[0038] After the isolation regions are formed at module **101** the process flow may either proceed with well and/or channel implants at module **105** or delay the formation of the wells and/or channels until after formation of the non-volatile charge trapping dielectric stack and/or gate layer deposition.

[0039] In an advantageous embodiment, a non-volatile charge trapping dielectric stack is formed on a first region of a substrate at module **110** after at least some of the well and channel implants for the logic MOS transistors are formed at module **105**. It has been found that approximately 0.5 nm of silicon dioxide may be removed during a conventional post-implant resist strip process. The amount removed is greater if the silicon dioxide is a deposited oxide rather than a thermally grown oxide or if the silicon dioxide received an implant (e.g. 1.0 nm of silicon dioxide removed/strip process). Because there is typically a number of well and channel implants in a CMOS process flow, between 1.5 nm and 2.5 nm of silicon dioxide may be removed prior to performing the RCA cleans in preparation for a logic MOS gate dielectric. Similarly, the well and channel implant strips may also etch a non-volatile charge trapping dielectric stack (which may include silicon dioxide). While the nominal etch rate of the implant strip processes is quite small, it has also been found to form pin holes, or localized defects in the non-volatile charge trapping dielectric stack which may reduce the charge retention of a SONOS-type memory device. Thus, inserting the module forming the non-volatile charge trapping dielectric stack after the well and channel implant modules of a logic CMOS flow results in the least disruption to the non-volatile charge trapping dielectric stack.

[0040] To include a logic MOS transistor on the same substrate as the non-volatile charge trap device, a gate insulator layer is formed on the second region of the substrate at module **120**. As will be discussed in more detail subsequently, this sequence of forming the non-volatile charge trapping dielectric layer prior to forming the MOS gate insulator layer advantageously utilizes the subsequent thermal treatments forming the MOS gate insulator layer to improve the quality of the non-volatile charge trapping dielectric stack, particularly a blocking layer. Logic MOS transistor degradation from thermal processing associated with formation of the non-volatile charge trapping dielectric layers is also avoided by forming the non-volatile charge trapping dielectric stack prior to forming the logic MOS gate insulator layer.

[0041] Following module **120**, a gate layer is deposited over both the MOS gate insulator layer and over the non-volatile charge trapping dielectric stack at module **130**. Next, if the well and/or channel implants were not performed at module **105**, the well and/or channel implants may be performed at module **140**, after module **130**. In such an embodiment, the well and channel implants may advantageously dope the gate layer formed at module **130** in addition to forming the wells and/or channels. With the well and channels in place (either as a result of module **105** or module **140**), the gate layer may be then be patterned into gate electrodes at module **150**. Gate electrode patterning may occur simultaneously for both a non-volatile charge trap memory device in the first region of the substrate and a MOS device in the second region of the substrate.

[0042] With the gate stacks of the two devices substantially complete at module **150**, tip and/or HALO implants may be formed for all devices at module **155** and sidewall spacers formed for all devices at module **160**. Source and drain implants may then be formed for all devices at operation **165**. As described elsewhere herein, a multi-layered liner and disposable spacer process may be performed at these operations to provide a high voltage CMOS transistor. Then at operation **170**, a silicide process may be performed to substantially complete the front end device fabrication. As further described herein, a multi-layered liner may be utilized to provide silicidation of logic CMOS without silicidation of the non-volatile charge trap memory device (i.e. selective silicidation). Backend metallization, as is conventional in the art, may then be performed to fabricate an integrated semiconductor structure comprising a non-volatile charge trap memory device and a MOS device on a single substrate.

[0043] FIG. **2**A illustrates a flow diagram depicting fabrication process **200** including particular modules integrating formation of a charge-trapping dielectric stack with logic MOS gate insulator formation, in accordance with particular embodiments of the present invention. Thus, FIG. **2**A depicts particular process modules employed in certain implementations of the modules **105**, **110**, **120** and **130** of FIG. **1**. FIGS. **3**A through **3**J further illustrate a cross-section of a SONOS memory device, a high voltage MOS device and a low voltage MOS device as the modules in the process flow of FIG. **2**A are implemented.

[0044] FIG. **2**B illustrates a flow diagram depicting fabrication process **201** including particular modules integrating formation of a charge-trapping dielectric stack with a high voltage MOS transistor and with selective contact silicidation, in accordance with particular embodiments of the present invention. Thus, FIG. **2**B depicts particular modules employed in certain implementations of the modules **155**,

UMC 1434
UMC v. AICP

160, **165** and **170** of FIG. **1**. FIGS. **4**A through **6**B further illustrate a cross-section of the non-volatile charge trap memory device, a high voltage MOS device and a low voltage MOS device as the modules in the process flow of FIG. **2**B are implemented.

[0045] Referring to FIG. **2**A, process **200** begins with STI formed in a substrate. The substrate may be a bulk substrate comprised of a single crystal of a material which may include, but is not limited to, silicon, germanium, silicon/germanium or a III-V compound semiconductor material. In another embodiment, the substrate is comprised of a bulk layer with a top epitaxial layer. In a specific embodiment, the bulk layer is comprised of a single crystal of a material which may include, but is not limited to, silicon, germanium, silicon/germanium, a III-V compound semiconductor material and quartz, while the top epitaxial layer is comprised of a single crystal layer which may include, but is not limited to, silicon, germanium, silicon/germanium and a III-V compound semiconductor material. In another embodiment, the substrate is comprised of a top epitaxial layer on a middle insulator layer which is above a lower bulk layer. The top epitaxial layer is comprised of a single crystal layer which may include, but is not limited to, silicon (i.e. to form a silicon-on-insulator (SOI) semiconductor substrate), germanium, silicon/germanium and a III-V compound semiconductor material. The insulator layer is comprised of a material which may include, but is not limited to, silicon dioxide, silicon nitride and silicon oxy-nitride. The lower bulk layer is comprised of a single crystal which may include, but is not limited to, silicon, germanium, silicon/germanium, a III-V compound semiconductor material and quartz.

[0046] At module **205**, a first well implant, such an n-well implant is performed. Module **205** will typically include forming a patterned photomask on a screening sacrificial dielectric layer, such as a silicon dioxide layer. The n-well implant is then performed in a region of the substrate, such as the region for MOS transistor **370**. In certain embodiments, the n-well implant includes implanting a phosphorus species at concentrations and energies conventional for MOS devices. A single n-well implant may be performed for PMOS transistors, PMOS HV transistors and p-type SONOS devices.

[0047] At module **206**, a dry and/or wet strip is performed to remove the well implant photomask. Conventional plasma strips, such as oxygen, forming gas, and the like may be employed. Similarly, conventional wet strips, such as piranha clean and ozone clean may be used. Because the charge trapping dielectric stack of the non-volatile memory device has not yet been formed, the silicon dioxide etch rate of the strip module **206** is of little concern.

[0048] A module **207**, a p-well implant is performed. Here too, conventional implant species, such as boron, may be employed at typical doses and energies. The p-well implant may be, but is not necessarily, a patterned implant such as the n-well implant of module **205**. If patterned, any of those strip processes of module **206** may be repeated. In one embodiment, the p-well implant is performed in another area of the substrate, adjacent to an n-well region in preparation for an NMOS transistor. In alternative embodiments, the p-well implant is an unmasked implant.

[0049] Any number of channel implants may also be performed at module **207** to adjust threshold voltages for specific device applications. For example, an n-channel implant may be performed in a region of the substrate where a NMOS transistor channel will be located, thereby setting a threshold voltage. The n-channel implant may be of any conventional species (e.g. BF_2), dose and energy for a particular device type. A channel implant for a non-volatile charge trap memory device may also be performed in a first region of the substrate **302**, such as the region for SONOS device **300** of FIG. **3**A. Similarly, a channel implant for a high voltage MOS transistor may be performed in the substrate region of HV MOS transistor **350**. A p-channel implant may likewise be performed, for example in the substrate region of MOS transistor **370**.

[0050] In the exemplary embodiment depicted in FIG. **3**A, in preparation for a SONOS channel implant, a window **305** defined by photoresist **307** is formed in the sacrificial dielectric layer **303**. The window **305** may be of sub-micron dimension, for example, approximately 0.2 um in length and width. In one such an embodiment, an oxygen plasma clean is performed to descum photoresist residue from the corners of window **305**. A sacrificial silicon oxide layer, which in one exemplary implementation is between 10 and 30 nm thick, may then be removed with a buffered oxide etchant (BOE) containing a surfactant, again to ensure window **305** is opened completely.

[0051] The isotropic etch of the screening sacrificial dielectric layer **303** can be expected to undercut the photoresist **307** by an amount D_1. The undercut amount D_1 is important when window **305** is proximate to a logic device, such as in the region for HV MOS transistor **350** because logic device implants performed through the screening sacrificial dielectric layer **303** may have a different implant profile within the undercut region. Therefore, certain embodiments downsize the dimensions of window **305**. For example a 0.2 um drawn size may be downsized to 0.18 um to compensate for an undercut of 0.01 um on a side. In further embodiments, because the undercut of window **305** may become very close to an adjacent logic device, critical layer lithography tools are employed to reduce misregistration tolerances.

[0052] After formation of the window **305**, the channel implant may be performed and the photoresist **307** may be stripped. With the well and channel implants formed, an anneal may be performed to complete module **207**. In one such embodiment, a rapid thermal anneal is performed after implanting both the n-well and p-well. The rapid thermal anneal may be any known in the art to be suitable for MOS transistor applications.

[0053] With the well and channel implants completed, the non-volatile charge trapping dielectric stack is formed at module **210**. Referring back to FIG. **2**A, a non-volatile charge trapping dielectric stack, such as an ONO charge trapping dielectric stack is then formed and patterned to remain only in memory cell areas at module **210**. In one embodiment depicted in FIG. **3**B, a ONO charge trapping dielectric stack **306** is comprised of a tunneling layer **304**A, a charge trapping layer **304**B and a blocking layer **304**C. The tunneling layer **304**A may be any material and have any thickness allowing charge carriers to tunnel into the charge-trapping layer under a high gate bias condition while maintaining a suitable barrier to leakage under conditions of low gate bias. In certain embodiments, tunneling layer **304**A is a commonly known dielectric layer, such as silicon dioxide (SiO_2), a silicon oxynitride ($SiO_xN_y(H_z)$), a silicon dioxide that is subsequently nitridized, or a stack dielectric made of silicon dioxide and silicon nitride (Si_3N_4) or silicon oxynitride, having a physical thickness of between about 1.5 nm and 3.0 nm. In

UMC 1434
UMC v. AICP

another embodiment, tunneling layer **304**A is comprised of a dielectric layer having a dielectric constant greater than that of silicon nitride which may include, but is not limited to, hafnium oxide, zirconium oxide, hafnium silicate, hafnium oxy-nitride, hafnium zirconium oxide and lanthanum oxide.

[0054] The charge trapping layer **304**B of the SONOS device **300** may further include any commonly known charge trapping material and have any thickness suitable to store charge and, modulate the threshold voltage of the devices. In certain embodiments charge trapping layer **304**B is silicon nitride, silicon-rich silicon nitride, or silicon oxynitride. In one particular embodiment, the trapping layer **304**B has a non-uniform stoichiometry across the thickness of trapping layer. For example, the charge trapping layer **304**B may further include at least two silicon oxynitride layers having differing compositions of silicon, oxygen and nitrogen. In one particular embodiment, a bottom oxynitride within charge trapping layer **304**B has a first composition with a high silicon concentration, a high oxygen concentration and a low nitrogen concentration to provide an oxygen-rich oxynitride.

[0055] 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 wt % 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 wt % 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.

[0056] The bottom silicon oxynitride may have a physical thickness between 2.5 nm and 4.0 nm corresponding to an EOT of between 1.5 nm and 5.0 nm. The charge trapping layer **304**B may further include a top silicon oxynitride with a high silicon concentration, a high nitrogen concentration and a low oxygen concentration to produce a silicon-rich silicon oxynitride. This second silicon oxynitride may have a physical thickness of 4.0 to 6.0 nm for a charge trapping layer **304**B with a net physical thickness of 9 to 11 nm. The oxygen-rich stoichiometry of the first silicon oxynitride, being more like silicon dioxide relative to silicon nitride, provides a good quality interface with tunneling layer **304**A. In one such embodiment, the composition of the oxygen-rich oxynitride results in an RI in the range of 1.7 and 1.9 and preferably about 1.8. In a further embodiment, the composition of the silicon-rich oxynitride results in an RI in the range of 1.8 and 2.0 and preferably about 1.9.

[0057] In one embodiment, multiple silicon nitride or silicon oxynitride charge trapping layers are formed in a low pressure CVD process using a silicon source, such as silane (SiH_4), dichlorosilane (SiH_2Cl_2), tetrachlorosilane ($SiCl_4$) or Bis-TertiaryButylAmino Silane (BTBAS), a nitrogen source, such as N_2, NH_3, N_2O or nitrogen trioxide (NO_3), and an oxygen-containing gas, such as O_2 or N_2O.

[0058] In one exemplary implementation, a silicon oxynitride charge trapping layer can be deposited over a tunneling layer by placing the substrate in a deposition chamber and the flow rate of ammonia (NH_3) gas and nitrous oxide (N_2O) as mixed with a silicon precursor, such as dichlorosilane (SiH_2Cl_2), to provide the desired gas ratios to form first an oxygen-rich oxynitride film and then a silicon-rich oxynitride film. Just as the terms "oxygen-rich" and "silicon-rich" are relative to a stoichiometric Si_3N_4 film, formation of these films may also be characterized based on the 3:1 volumetric flow rate ratio, SiH_2Cl_2:NH_3, commonly employed to produce a stoichiometric (Si_3N_4) with a CVD method. The oxygen-rich oxynitride film is therefore formed with a relatively higher volumetric flow rate of oxidizer (e.g. N_2O) than used for the silicon-rich oxynitride film while the both the oxygen-rich and silicon-rich oxynitride films are formed with a relatively higher volumetric flow rate of silicon precursor (e.g. SiH_2Cl_2).

[0059] In a specific batch process embodiment, an oxygen-rich oxynitride film is formed by introducing a process gas mixture including N_2O, NH_3 and SiH_2Cl_2, while maintaining the chamber at a pressure approximately in the range of 5-500 mTorr, and maintaining substrate **400** at a temperature approximately in the range of 700-850° C., for a period approximately in the range of 2.5-20 minutes. In an exemplary embodiment, the process gas mixture includes N_2O and NH_3 at a high volumetric flow rate ratio of about 1:1 to about 3:1 N_2O:NH_3 while the SiH_2Cl_2 to NH_3 is also at a high volumetric flow rate ratio from about 3.5:1 to 8:1 SiH_2Cl_2:NH_3. In a preferred embodiment, the N_2O:NH_3 ratio is about 2:1. while the SiH_2Cl_2:NH_3 is at a ratio of about 6:1. In certain embodiments, the gases are introduced at a flow rate approximately in the range of 5-200 standard cubic centimeters per minute (sccm).

[0060] In a further embodiment, a silicon-rich oxynitride film is then formed by introducing a process gas mixture including N_2O, NH_3 and SiH_2Cl_2, while maintaining the chamber at a pressure approximately in the range of 5-500 mTorr, and maintaining substrate **400** at a temperature approximately in the range of 700-850° C., for a period approximately in the range of 2.5-20 minutes in a batch furnace. The process gas mixture includes N_2O and NH_3 at a volumetric flow rate ratio from about 1:8 to about 1:4 (N_2O:NH_3) with SiH_2Cl_2 and NH_3 at a volumetric flow rate ratio from about 3.5:1 to 5:1 (SiH_2Cl_2:NH_3). In a preferred embodiment, the N_2O and NH_3 are provided at a volumetric flow rate ratio of about 1:5 (N_2O:NH_3) while the SiH_2Cl_2 and NH_3 are at a volumetric flow rate ratio of about 4:1 (SiH_2Cl_2:NH_3). In certain embodiments, the gases are introduced at a flow rate approximately in the range of 5 to 200 sccm.

[0061] Completing the ONO charge trapping dielectric stack **306**, the blocking layer **304**C of the SONOS device **300** may be any commonly known material with any thickness suitable to maintain a barrier to charge leakage without significantly decreasing the capacitance of the gate stack. In one embodiment, blocking layer **304**C comprises a dielectric layer having a higher dielectric constant than silicon nitride which may include, but is not limited to, hafnium oxide, zirconium oxide, hafnium silicate, hafnium oxy-nitride, hafnium zirconium oxide and lanthanum oxide. In another embodiment, the blocking layer **304**C is silicon dioxide layer, silicon oxynitride layer, or a silicon dioxide and silicon nitride stack, with a physical thickness between about 3.0 nm and about 5.0 nm.

[0062] Blocking layer **304**C can be formed by any suitable means including, for example, thermal oxidation or deposition with CVD techniques. In a preferred embodiment, the blocking layer is a deposited film formed with a high-temperature CVD process. Generally, the deposition process

UMC 1434
UMC v. AICP

involves providing a silicon source, such as SiH_4, SiH_2Cl_2, or $SiCl_4$ 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. Preferably, the blocking layer is deposited sequentially in the same processing tool employed to form the charge trapping layer(s) **304**B. More preferably, the blocking layer is formed in the same processing tool as is both the charge trapping layer(s) **304**B and the tunneling layer **304**A without removing the substrate between operations.

[0063] With the ONO charge trapping dielectric layers **304**A, **304**B and **304**C formed, they are then patterned into the ONO charge trapping dielectric stack **306** in the SONOS device **300**, as depicted in FIG. **3**B. Conventional lithography and etching techniques may be employed to remove the charge trapping dielectric layers from other regions of the substrate, such as the HV MOS region **350** and MOS region **370**. In a particular embodiment, a combination of dry and wet etch is performed to achieve a good stack sidewall profile. In one such embodiment, an inorganic spin-on anti-reflective coating (ARC), the blocking layer **304**C, and the dielectric layers **304**A and **304**B are dry etched, with the dry etch process stopping on the sacrificial dielectric layer **303**. In a subsequent wet etch operation, an etchant, such as BOE, is employed to clear sacrificial dielectric layer **303**. Here too, as discussed elsewhere herein in reference to opening of window **305** of FIG. **3**A, the isotropic wet etch may undercut the masked region. In this instance, undercutting the ONO charge trapping dielectric stack **306**, as denoted by the dashed line in FIG. **3**B, reduces the overlap between ONO charge trapping dielectric stack **306** and the sacrificial dielectric layer **303** to the amount D_2. If the overlay of the ONO charge trapping dielectric stack **306** and the window **305** is not adequate (e.g. D2 becomes zero) the substrate region of SONOS device **300** may be rendered non-functional by subsequent processes. Thus, for this reason too, the dimensions and alignment of window **305** and ONO charge trapping dielectric stack **306** are important.

[0064] Upon completion of module **210**, the method of FIG. **2**A proceeds to module **212**, wherein the substrate **302** is cleaned of organic residues left on wafer from photoresist strip or etch by-products in preparation for the formation of a gate insulator layer in the HV MOS region **350** and MOS region **370**. In the particular embodiment depicted in FIG. **2**A, module **212** includes a non-HF gate insulator pre-clean. Hydrofluoric acid (HF) cleans, while conventionally performed in logic CMOS processes to remove any native or chemical oxides from the substrate **302** prior to forming a gate insulator, are disadvantageous when non-volatile charge trapping dielectric layers have already been formed and remain substantially unprotected.

[0065] Conventional HF-based gate insulator pre-cleans will etch or otherwise degrade the quality of the ONO charge trapping dielectric stack **306**, particularly when the stack includes a CVD formed blocking layer **304**C. Therefore, in the depicted embodiment, module **212** includes cleaning operations which are substantially free of HF. For such embodiments, the substrate **302** may retain a native or chemical oxide after the cleaning operations employed in the module **212**. It should be appreciated, that this concern of HF-based cleans attacking thin and critical dielectric layers is not present for standard logic MOS processes and is also not to be found in flash memory processes that protect such layers (e.g. tunnel oxide layer of flash memory device), with a polysilicon floating gate layer prior to the HF-based MOS gate insulator pre-clean.

[0066] The non-HF pre-clean of module **212** may include cleaning regimes known in the art to remove organic residues, such as, but not limited to piranha cleans, ozone cleans, and plasma cleans comprising O_2 or forming gas. The non-HF pre-clean may also include a RCA Standard Clean 1 (SC1) clean comprising a mixture of water, hydrogen peroxide and ammonium hydroxide (H_2O:H_2O_2:NH_4OH). It will be appreciated that a blocking layer **304**C formed by CVD may be particularly susceptible to dielectric etchants because, for example, a deposited oxide blocking layer is typically of poorer quality than a thermally formed oxide layer. The poorer quality, be it from film stress, porosity, stoichiometry or otherwise, is associated with elevated etch rates relative to thermally grown dielectric layers. Therefore, the processes employed in the pre-clean module **212** should not be too aggressive.

[0067] For example, SC1 mixtures in logic CMOS are typically employed at a ratio of 5:1:1 H_2O:H_2O_2:NH_4OH, however it has been found that this chemistry may etch a CVD silicon dioxide blocking layer **304**C at an average rate of approximately 0.2 to 0.3 nm/minute. Perhaps more of a concern than this nominal etch rate, is the capacity for the SC1 chemistry to roughen the blocking layer **304**C. This roughness may be characterized with RMS roughness measurements. Also a concern associated with the SC1 chemistry is formation of pinhole defects in the top oxide which may be found at a low enough density to remain undetectable with RMS roughness measurements but nonetheless decrease the quality of the blocking layer **304**C. It has been found these difficulties are avoidable or at least substantially mitigated by employing an ultra-dilute SC1 clean at module **212**. An ultra-dilute SC1 is substantially more dilute than 5:1:1. For example, in one advantageous embodiment the ultra-dilute SC1 comprises approximately 0.001% NH4OH, and 0.1% H_2O_2 in H_2O. Following the ultra-dilute SC1, the non-HF pre-clean depicted in module **212**, may further include an RCA Standard Clean 2 (SC2) clean comprising a mixture of H_2O:H_2O_2:HCl in a ratios known in the art.

[0068] In an alternative embodiment, the non-HF pre-clean of module **212** may include an ozonated water cleaning regime. For such an embodiment, the SC1 clean may be replaced by the ozonated water to remove the organics and etch residues. With elimination of the SC1 clean, few metals will be left on the substrate surface and the SC2 is therefore unnecessary.

[0069] Following the non-HF pre-clean of module **212**, a logic MOS gate insulator layer may be formed on the substrate **302**. The logic MOS gate insulator layer may comprise any of the dielectric materials described elsewhere herein for any of the charge trapping dielectric layers **304**A, **304**B and **304**C, but in a particular embodiment, includes a thermally grown oxide as the gate insulator layer **314** of FIG. **3**C. For particular embodiments employing a silicon substrate **302**, the gate insulator layer **314**, thermally grown, comprises silicon oxygen bonds. As noted elsewhere herein, the integrated process flow depicted in FIG. **1** advantageously sequences the formation of the ONO charge trapping dielectric stack **306** prior to formation of the logic MOS gate insulator so that formation of the logic MOS gate insulator with a thermal process has the advantage of additionally serving to reoxidize

UMC 1434
UMC v. AICP

the ONO charge trapping dielectric stack **306**. Reoxidation of the blocking layer **304**C may have the effect of densifying a CVD formed blocking oxide layer and improving the quality of the blocking oxide and thereby improving non-volatile charge trap memory device performance (e.g. reduced back injection). The reoxidation of the blocking layer **304**C is depicted by the addition of field lines in FIG. **3**C. In a further embodiment, formation of the gate insulator layer **314** may further oxidize or reoxidize a portion or all of the charge trapping layer, such as a portion or all of the charge trapping layer **304**B shown in FIG. **3**C, to achieve a graded band gap in the charge trapping layer **304**B. Such a graded band gap may further improved non-volatile charge trap memory device performance. Reoxidation for this purpose after the deposition of the blocking layer **304**C may enable a more controlled diffusion of oxidizer to controllably oxidize or reoxidize the thin charge trapping layer **304**B.

[0070] Generally, the formation of the gate insulator layer **314** may include any conventional gate oxidation process whereby the substrate **302** is heated in the presence of an oxidizing gas such as, oxygen (O_2), nitrous oxide (N_2O), nitric oxide (NO), ozone (O_3), and steam (H_2O). In one embodiment, the gate oxidation process is performed at a higher temperature than the temperature at which the blocking layer **304**C is deposited. In a particularly advantageous embodiment, a dilute wet oxidation is employed to form the gate insulator layer **314**. The dilute wet oxidation is distinct from a conventional wet oxidation in that the H_2:O_2 ratio is between 1 and 1.3. In one specific embodiment, a dilute oxidation with an H_2:O_2 ratio of approximately 1.2 is performed at a temperature of between 800° C. and 900° C. In a further embodiment, the duration of the dilute oxidation is sufficient to grow between 5.0 nm and 15.0 nm of silicon dioxide where substrate **302** is silicon. In one such embodiment, the duration is sufficient to for an approximately 10 nm to 1.1 nm silicon dioxide layer to be formed on a silicon substrate. Such a dilute wet oxidation process advantageously reoxidizes a deposited blocking layer **304**C and may further oxidize or reoxidize a portion of the charge trapping layer **304**B.

[0071] Where a native oxide or a chemical oxide remains on the substrate **302** after the pre-clean module **212**, a thermal oxidation forms a gate insulator layer **314** comprising silicon dioxide by consuming some of the silicon below the native or chemical oxide in the substrate **302**. Therefore, where multiple MOS gate insulator layer thicknesses are to be employed, for example one thickness in the region for HV MOS transistor **350** and a second thickness in the region of MOS transistor **370**, it may be advantageous to form the thickest gate insulator layer at module **214** of FIG. **2**A prior to additional gate insulator layers of lesser thickness so that any native or chemical oxide formed since the formation of the ONO charge trapping dielectric stack **306** is completely consumed and the electrical impact of the native oxide on the resulting MOS device is reduced by the relatively greater gate insulator layer thickness.

[0072] In one embodiment depicted in FIG. **2**A, if another gate insulator layer of differing composition and/or thickness is to be formed in the process, then the method **200** proceeds to module **218**. At module **218** a photoresist layer **318** of FIG. **3**D is deposited and patterned to have an opening **319** formed over a region of the substrate **302** that is to have the next insulator material and/or insulator layer thickness. At module **222**, any previously formed gate insulator layers, such as gate insulator layer **314**, depicted in FIG. **3**D, are selectively removed to expose the substrate **302**. Conventional lithography and etch techniques may be employed at modules **218** and **222**, such as those described in reference to module **205**.

[0073] Following the removal of the gate insulator layer(s), a pre-clean may be performed on the substrate **302**. For example, in the module **224**, while the photoresist layer **318** protects the ONO charge trapping dielectric stack **306** a clean which would be detrimental to the ONO charge trapping dielectric stack **306** if it were not protected by the photoresist layer **318** may be performed at this time. While in certain embodiments the clean in module **224** is not performed, in either case, the photoresist layer **318** is stripped at module **226**, for example with conventional piranha clean and/or plasma ash operations, subsequent to the selective removal of the gate insulator layer(s).

[0074] With the ONO charge trapping dielectric stack **306** patterned and the gate insulator layer **314** patterned, the non-HF pre-clean module **212** may then be repeated in preparation of forming another gate insulator layer. Any of the processes described elsewhere herein for module **212**, such as an ultra-dilute SC1 clean, may be performed at this time to clean the substrate **302** in preparation for formation of an additional gate insulator layer in the opening **319**. Following module **212**, another gate insulator layer may be formed at module **214**, such as gate insulator layer **320**. Gate insulator layer **320** may be any of the materials described in reference to gate insulator layer **314** and not necessarily the same material as gate insulator layer **314**. In one particular embodiment, gate insulator layer **320** is a thermally grown layer comprising silicon dioxide. In a further embodiment, the gate insulator layer **320** is formed over a third region of the substrate **302**, such for MOS transistor **370**, and is thinner than the gate insulator layer **314** formed over a second region of the substrate **302**, such as for HV MOS transistor **350**. In one such embodiment, the gate insulator layer **320** comprising silicon dioxide is formed to a thickness between approximately 3.0 nm and 8.0 nm while the gate insulator layer **314** is between 5 and 15 nm. Any of the processes described elsewhere herein for module **214** in reference to the formation of gate insulator layer **314** may also be employed to form the gate insulator layer **320**. Additionally, the blocking layer **304**C and charge trapping layer **304**B may be reoxidized during the formation of the gate insulator layer **320**, much as described in reference to the formation of the gate insulator layer **314**. It should be appreciated such as reoxidation may be to a lesser extent than what occurs during the formation of the gate insulator layer **314**, particularly where the gate insulator layer **320** is formed thinner than the gate insulator layer **314** or where the gate insulator layer **320** is formed with a process other than the dilute steam oxidation described for one embodiment of the gate insulator layer **314**.

[0075] If desired, modules **218**, **222**, **224**, **226**, **212** and **214** may be repeated any number of times to provide more than the two gate insulator layer thicknesses described in the embodiment depicted in FIGS. **3**C-**3**E. In this manner, successively thinner gate insulator layers may be formed with each iteration. For example, a third gate insulator layer may be formed to between 2.0 nm and 3.5 nm, thinner than the gate insulator layer **314** and thinner than the gate insulator layer **320**.

[0076] After forming at least one MOS gate insulator layer, such as the gate insulator layer **314**, the embodiment depicted in FIG. **2**A proceeds to module **228**. At module **228**, the gate insulator layer **314** and the ONO charge trapping dielectric

UMC 1434
UMC v. AICP

stack **306** are nitrided or nitridized. Beyond nitriding the MOS gate insulator, this nitridation process serves to incorporate nitrogen into the ONO charge trapping dielectric stack **306** and improve the quality of the interfaces in the stack (e.g. between the dielectric layers **304**C and **304**B). This nitriding process, in certain embodiments, may incorporate approximately 4-10 wt % nitrogen into the blocking layer **304**C. In a particular embodiment, the nitridation process includes heating substrate **302** in an atmosphere including nitrogen at a temperature approximately in the range of 900-1100° C.

[0077] In one embodiment, nitridation of the ONO charge trapping dielectric stack **306** is performed as part of forming the gate insulator layer (e.g. gate insulator layer **314** or gate insulator layer **320**). Thus, modules **214** and **228** of FIG. **2**A need not be performed in separate process equipment, but rather merely a separate step of a single process recipe. This nitriding process, in certain embodiments, may incorporate approximately 4-10 wt % nitrogen into the blocking layer **304**C and approximately 4-10 wt % nitrogen into the gate insulator layer **314** and/or gate insulator layer **320**. In one such embodiment, a CVD furnace is employed for the nitridation of module **228** and the duration of the nitridation may be for between 5 minutes and 10 minutes. In another embodiment, a single wafer tool may be employed for the nitridation of module **228**, exposing the gate insulator layer(s) and ONO charge trapping dielectric stack **306** to a nitrogen-containing environment for a duration in the range of approximately 30 seconds to approximately 60 seconds.

[0078] In an embodiment, the atmosphere including nitrogen is composed of a gas such as, but not limited to nitrogen (N_2), nitrous oxide (N_2O), nitrogen dioxide (NO_2), nitric oxide (NO) and ammonia (NH_3). In still other embodiments, the nitrogen environment further includes deuterium through an introduction of gases in which hydrogen has been replaced by deuterium, including, for example, the substitution of ND_3 for NH_3. The substitution of deuterium for hydrogen may advantageously passivate Si dangling bonds at the substrate interface, thereby increasing non-volatile charge trap memory device parametrics, such as NBTI (Negative Bias Temperature Instability) lifetime.

[0079] In another particular embodiment, nitridation in module **228** is performed only once after the last gate insulator layer is formed, for example, after gate insulator layer **314** and gate insulator layer **320** have been formed. The single nitridation process therefore nitridizes the ONO charge trapping dielectric stack **306**, the gate insulator layer **314** and the gate insulator layer **320**, as depicted by the field lines of FIG. **3**F. In such an embodiment, the single nitridation provides the benefits described herein while minimizing the thermal budget of the integrated process **200**.

[0080] In another embodiment, the nitridation process of module **228** is performed only once after the first gate insulator is formed, for example, after gate insulator layer **314**. In such an embodiment, the ONO charge trapping dielectric stack **306** is nitrided along with the gate insulator layer **314**. Nitridation only after the first gate insulator may allow for some of the logic MOS devices fabricated on substrate **302**, such as in the region for HV MOS transistor **350**, to have a nitrided gate insulator layer while others do not, such as in the region for MOS transistor **370**.

[0081] Embodiments employing nitridation immediately after the first gate insulator is formed may also improve the ability of the ONO charge trapping dielectric stack **306** to withstand exposure to a subsequent HF-based clean performed prior to forming a subsequent gate insulator layer, such as gate insulator layer **320**. In one particular embodiment, a silicon dioxide gate insulator layer **314** is formed to a thickness of approximately 5.0 nm to 15.0 nm (consuming silicon below a native oxide on the substrate **302**), the nitridation of module **228** is performed, a dilute HF-based clean is performed with the nitrided ONO charge trapping dielectric stack **306** exposed and then a silicon dioxide gate insulator layer **320** of a thickness between 2.0 and 8.0 nm is formed without consuming any significant thickness of native or chemical oxide on the substrate **302**. In this particular embodiment, the nitridation operation of module **228**, which may be part of a thick MOS gate insulator formation process, enables a dilute HF-based pre-clean chemistry to be employed for the more critical thin MOS gate insulator formation with minimal detrimental effect on the ONO charge trapping dielectric. In still other embodiments, each successive gate insulator formed includes the nitridation process of module **228** such that the ONO charge trapping dielectric stack **306** is exposed to a plurality of nitrogen anneals.

[0082] Following the nitridation of module **228** in FIG. **2**A, the method **200** proceeds with module **230**. At module **230**, a gate layer **330** is deposited on both the ONO charge trapping dielectric stack **306** and on the gate insulator layers **314** and **320**, as depicted in FIG. **3**G. The gate layer may be formed with any process conventionally known in the art. The gate layer may be any conductor or semiconductor material employed for gate layers in the art. In one embodiment, the gate layer **330** contains a metal, such as, but not limited to, hafnium, zirconium, titanium, tantalum, aluminum, ruthenium, palladium, platinum, cobalt and nickel, their silicides, their nitrides and their carbides.

[0083] In another embodiment, the gate layer **330** is poly-silicon (p-silicon). In a further embodiment, the poly-silicon gate layer **330** may be dual-doped to have N+ conductivity over a first portion of the first and/or second gate insulator layer (**314**, **320**) to form a HV NMOS and/or NMOS transistor, respectively, while having P+ conductivity over a second portion of the first and/or second gate insulator layer (**314**, **320**) to form a HV PMOS and/or PMOS transistor, respectively. In a further embodiment, the poly-silicon gate layer **330** may be doped to have either N+ or P+ conductivity in the SONOS device **300**. Thus, the dual-doped poly-silicon may form a P+ poly-silicon gate on an N-type SONOS memory device. Because the P+ poly-silicon gate has a Fermi level approximately 1 eV higher that an N+ poly-silicon gate, the larger workfunction of a P+ poly-silicon gate on an N-type SONOS device channel may improve reliability by reducing the number of charge carriers entering the ONO charge trapping dielectric stack **306** relative to an N-type SONOS device having an N+ poly-silicon gate.

[0084] As further shown in FIG. **3**H, a SONOS gate electrode is patterned over the substrate region of SONOS device **300**, while a HV MOS gate electrode **358** and MOS gate electrode **378** are patterned over the HV MOS transistor **350** and MOS transistor **370** substrate regions, respectively. In a particular embodiment, the patterning of SONOS gate electrode is performed with a dry etch stopping on the ONO charge trapping dielectric stack to protect the substrate semiconductor of the SONOS device **300**.

[0085] Conventional tip and/or HALO implant process may then be performed at module **255** of FIG. **2**A to form lightly doped drains (not pictured). Subsequently, at module **261** and as further depicted in FIG. **3**I, a sidewall spacer **309**

UMC 1434
UMC v. AICP

is then formed adjacent to a sidewall of the SONOS gate electrode **308** and on the ONO charge trapping dielectric stack **306**. Sidewall spacer **309**, for example, may be comprised of silicon dioxide, silicon oxynitride, or silicon nitride and may also be patterned selectively to the ONO charge trapping dielectric stack **306**. The ONO charge trapping dielectric stack **306** may then be subsequently etched to be self-aligned with sidewall spacer **309** to complete the formation of a SONOS gate stack **301** as depicted in FIG. **3**J. Similar processes may also form spacers **359** and **379** adjacent to HV MOS gate stack **351** and a MOS gate stack **371**, respectively.

[0086] FIG. **4**A illustrates a cross-sectional side view of devices following the source/drain implant module **263** of FIG. **2**B. Illustration of gate stacks **301**, **351** and **371** is simplified relative to those of FIG. **3**J merely for clarity. SONOS device **300** now includes source and drain **410** in substrate **302** having a conductivity opposite to the channel region. For example, in accordance with an embodiment of the present invention, source and drain **410** are N-type doped while channel region of substrate **302** is P-type doped. In one embodiment, substrate **302** is comprised of boron-doped single-crystal silicon having a boron concentration in the range of 1×10^{15}-1×10^{19} atoms/cm^3. In another embodiment, source and drain **410** are comprised of phosphorous- or arsenic-doped regions having a concentration of N-type dopants in the range of 5×10^{16}-1×10^{20} atoms/cm^3. In a specific embodiment, source and drain **410** have a depth in substrate **302** in the range of 80-200 nanometers. In accordance with an alternative embodiment of the present invention, source and drain **410** are P-type doped while the channel region of substrate **302** is N-type doped. As further shown, lightly doped source and drain (LDD) **411**, formed at module **255** of FIG. **2**A, extend under sidewall spacer **309**.

[0087] Both the HV MOS transistor **350** and MOS transistor **370** also include an LDD **461** and **481**, respectively. MOS transistor **370** further includes a source and drain **480** adjacent to the sidewall spacer **379** and a distance T_1 away from below the sidewall of the gate stack **371**. The source and drain **480** has an n-type conductivity and may, in certain embodiments, have substantially the same dopant concentration as that of the source and drain **410**. However, as depicted in FIG. **4**A, at module **263**, the HV MOS transistor **350** lacks a source and drain analogous to the source and drain **410** and **480**. Thus, during implantation of the source and drain **410** and **480**, the HV MOS transistor **350** may be masked with a photosensitive mask or with a commonly known non-photosensitive hardmask, such as, but not limited to, amorphous carbon, that was previously patterned.

[0088] A multi-layered liner is then formed at module **264** of FIG. **2**B. As depicted in FIG. **4**B, a liner **485** is formed over the substrate **302**, covering the SONOS device **300**, the HV MOS transistor **350** and MOS transistor **370**. In the particular embodiment depicted, the liner **485** is a multi-layered liner comprising a bottom liner layer **485**A and a top liner layer **485**B. The bottom liner layer **485**A and top liner layer **485**B may be any commonly employed materials. In a preferred embodiment, the top liner layer **485**B may be anisotropically etched selectively to the bottom liner layer **485**A. In one such embodiment, the bottom liner layer **485**A is silicon dioxide while the top liner layer **485**B is a silicon nitride. In an alternate embodiment, the bottom liner layer **485**A is a silicon nitride while the top liner layer **485**B is silicon dioxide. Other embodiments may include a top or bottom layer of silicon oxy-nitride, carbon-doped silicon nitride or boron-doped silicon nitride. While the multi-layered liner **485** has particular integration advantages discussed elsewhere herein, certain embodiments may also utilize a single layer liner comprised of, for example, a silicon dioxide layer or a silicon nitride layer.

[0089] A thin bottom liner layer **485**A advantageously reduces the lateral thickness deposited on the sidewalls of the sidewall spacer **309**, **359** and **379**, which may be in close proximity to sidewalls of other devices. A thin bottom liner layer **485**A may further reduce the amount of thickness variation in the film across different regions of the substrate **302**, the advantage of which is discussed elsewhere herein. In one such embodiment, a silicon nitride bottom liner layer **485**A is formed to a thickness of between about 2 nm about 15 nm, preferably between about 5 nm and about 8 nm. The thickness of the top liner layer **485**B may be selected to provide the multi-layered liner with a desired thickness on the sidewall of the spacer **359**, as discussed further elsewhere herein. In one embodiment, a silicon dioxide top liner layer **485**B is formed to a thickness of between about 10 nm and 40 nm, preferably between about 20 nm and 30 nm.

[0090] The layers of the multi-layered liner **485** may be deposited with any commonly known techniques, such as, but not limited to, thermal oxidation, low pressure CVD (LPCVD) plasma enhanced CVD (PECVD) and ALD processes known to those of skill in the art. For example, a nitride bottom liner layer **485**A may be deposited with a nitrogen precursor, such as NH_3, and a silicon precursor, such as silane (SiH_4), dichlorosilane (SiH_2Cl_2), or bis(tertiary-butylamino) silane (BTBAS). The deposition may be performed at a substrate temperature, for example between approximately 550° C. and approximately 850° C., and at a deposition chamber pressure between approximately 100 millitorr (mT) and approximately 700 mT, to form a film having a thickness anywhere within the ranges previously described.

[0091] An oxide layer top liner layer **485**B may be similarly formed by thermal or chemical oxidation of the bottom liner layer **485**A or a deposition process, such as an LPCVD employing any commonly known precursors, to form a film having a thickness anywhere within the ranges previously described. In a particular embodiment, either or both of the bottom liner layer **485**A and top liner layer **485**B may be deposited with techniques known to result in highly stressed films. Such stressed-film embodiments may make subsequent removal of either liner layer significantly faster and/or more selective to each other or underlying layers.

[0092] At module **265** and as further depicted in FIG. **4**C, the top liner layer **485**B is anisotropically etched to form disposable sidewall spacer **486** along sidewalls of topography present under the multi-layered liner **485**. The top liner layer **485**B is etched selectively to the bottom liner layer **485**A (i.e. the bottom liner layer **485**A provides an etch stop). In one such embodiment, the bottom liner layer **485**A remains a substantially continuous film over the substrate **302** after the top liner layer **485**B is formed in to discrete disposable sidewall spacer **486**.

[0093] The process selected to anisotropically etch the top liner layer **485**B to form disposable sidewall spacer **486** is dependent on the materials chosen. In the particular embodiment employing a silicon nitride bottom layer **485**A and a silicon dioxide top liner layer **485**B a commonly known plasma etch process may be used, such as one with a fluorine chemistry like carbon tetrafluoride (CF_4), having a high

UMC 1434
UMC v. AICP

enough selectivity to silicon nitride to stop prior to etching through the bottom liner layer **485**A. In an alternate embodiment employing a silicon dioxide bottom layer **485**A and a silicon nitride top layer **485**B, any commonly known plasma etch process may be used, such as one utilizing a fluorine-based chemistry, like nitrogen trifluoride (NF_3), or one utilizing a chlorine-based chemistry.

[0094] Subsequent to the formation of the disposable sidewall spacer **486**, a source and drain may be formed for the HV MOS transistor at module **267**. In the embodiment further depicted in FIG. **4**D, the source and drain **460** is formed with a p-type implant after a mask **498** is formed over the SONOS device **300** and MOS transistor **370**. Mask **498** may be any commonly known photosensitive mask material (i.e. photoresist) or non-photosensitive mask, such as amorphous carbon, that was previously patterned. The p-type dopant may be any commonly employed in the art, such as a Boron species. Other embodiments include n-type dopants for HV NMOS transistors.

[0095] The implantation, is self-aligned to the gate stack **351** and offset from a sidewall of the gate stack **351** by a distance T_2. The distance T_2 is approximately equal to the sidewall thickness of the sidewall spacer **359** added to the sidewall thickness of the bottom liner layer **485**A added to the sidewall thickness of the disposable sidewall spacer **486**. In the embodiment shown in FIG. **4**D, the distance T_2 for the HV MOS transistor **350** is greater than the distance T_1 for the MOS transistor **370**. In this manner, the source and drain **460** is offset by the distance T_2, greater than T_1, to increase the length of the LDD **461**. Thus, the thickness of the top liner layer **485**B deposited in the operation depicted in FIG. **4**B may be predetermined to provide a disposable sidewall spacer **486** with the appropriate lateral width (thickness).

[0096] The relatively greater offset represented by T_2 may increase the breakdown voltage by reducing the encroachment of p-type dopant diffusion from the source and drain **460** into the channel region of the HV MOS transistor **350** during subsequent thermal processing. In the embodiment depicted, the p-type implant is made through the bottom liner layer **485**A. As previously described, particular embodiments employ an advantageously thin bottom liner layer **485**A to improve the uniformity of the implant profile across the substrate **302**.

[0097] Following the formation of the source and drain **460**, the mask **498** may be removed to expose the bottom liner layer **485**A covering the SONOS device **300** and MOS transistor **370**. At module **268** and as further shown in FIG. **4**E, the disposable sidewall spacer **486** may then be removed. Removal of the disposable sidewall spacer **486** may advantageously increase the space between adjacent logic and non-volatile charge trap memory devices to enable a higher packing density of devices (i.e. smaller device pitch). This is particularly advantageous for SONOS device **300** which may be part of a closely spaced array of SONOS devices, such as in a memory cell array. Removal of the disposable sidewall spacer **486** may also improve the step coverage of subsequently deposited ILD layers over high density SONOS and logic devices.

[0098] Removal of the disposable sidewall spacer **486** may be with a masked process, whereby the disposable sidewall spacer **486** is removed, for example, from the SONOS device **300**, but retained on the HV MOS transistor **350**. However, in the embodiment depicted, the disposable sidewall spacer **486** is removed from the entire substrate **302** with an unmasked etch process. As shown, the etch process is selective to the bottom liner layer **485**A (i.e. the bottom liner layer **485**A acts as an etch stop for the etch process employed to remove the disposable sidewall spacer **486**. With the protection of bottom liner layer **485**A, substrate semiconductor and substrate insulator layers, such as shallow trench isolation (STI) are protected from the process employed to remove the disposable sidewall spacer **486**. Because it has been found that processing of the corner (not pictured, but is out of the plane of FIG. **4**E) formed where the STI meets the width of the gate stack **301** can greatly effect the performance of the SONOS device **300**, it is advantageous not to expose this region to the process employed for disposable spacer removal.

[0099] With the bottom liner **485**A serving as an etch stop layer, the substrate **302**, the SONOS gate electrode **308**, the HV PMOS gate layer **358** and the NMOS gate layer **378**, as well as the sidewall spacers **309**, **359** and **379** remain protected during the removal of the disposable sidewall spacer **486**. With such features protected, the material composition of the disposable sidewall spacer **486** (i.e. top liner layer **485**B) is independent of the materials in the gate electrodes **308**, **358** and **378** and the sidewall spacers **309**, **359** and **379**.

[0100] Disposable sidewall spacer **486** may be removed with commonly known wet chemical or dry etch processes, depending on the materials employed in the particular implementation. In one embodiment, wherein the disposable sidewall spacer **486** comprises silicon dioxide (i.e. a silicon dioxide top liner layer **485**B), a hydrofluoric acid (HF) based wet chemical etch may be performed to remove the disposable sidewall spacer **486** selectively to a silicon nitride bottom liner layer **485**A. In an alternate implementation, an isotropic dry etch process, such as one commonly known to have a high selectivity over the silicon nitride bottom liner layer **485**A may be employed. In another embodiment, wherein the disposable sidewall spacer **486** comprises silicon nitride (i.e. a silicon nitride top liner layer **485**B), a hot phosphoric acid (H_3PO_4) based wet chemical etch may be performed to remove the disposable sidewall spacer **486** selectively to a silicon dioxide bottom liner layer **485**A. In an alternate implementation, an isotropic dry etch process, such as one commonly known to have a high selectivity over the silicon dioxide bottom liner layer **485**A may be employed.

[0101] Subsequent to the removal of the disposable sidewall spacer **486**, the bottom liner layer **485**A may be removed at module **269** to expose the source and drain regions of either or both of the SONOS and logic devices in preparation for a silicidation or salicidation (self-aligned silicidation) process at module **270**. In one embodiment, a blanket strip of the bottom liner layer **485**A may be performed to expose the source and drain regions of all devices. In such an embodiment, the strip process is preferably selective to the STI corner, the gate electrodes **308**, **358** and **378** and the sidewall spacers **309**, **359** and **379**. FIG. **4**F, however, depicts an alternate embodiment utilizing a patterned etch of the bottom liner layer **485**A. The patterned etch exposes only the source and drains of those devices for which silicide is desired. The bottom liner layer **485**A may thereby further provide for device-dependent silicidation.

[0102] As previously discussed, because the silicidation process can induce stress, silicide may be detrimental to the performance and reliability of the SONOS device **300**. Therefore, a device-dependent silicidation process may be advantageous for integrating a logic device, such as MOS transistor **370**, having silicide contacts with a non-volatile charge trap

UMC 1434
UMC v. AICP

memory device, such as SONOS device **300**, having silicide-free contacts. Similarly, a HV MOS device, such as HV MOS transistor **350**, may include either silicide or silicide-free contacts. As shown in FIG. **4**E, the LDD region **461** extends beyond the sidewall spacer **359** and bottom liner **485**A (i.e. below where the disposable sidewall spacer **486** was removed) and formation of silicide over this exposed LDD region may not be desirable.

[0103] In one embodiment, as shown in FIG. **4**F, a mask **499** is formed over the SONOS device **300** and HV MOS transistor **350**. Mask **499** may be any commonly known photosensitive mask material (i.e. photoresist) or non-photosensitive mask, such as amorphous carbon, which is first patterned. The bottom liner layer **485**A may then be etched to expose the regions of the MOS transistor **370** for subsequent silicidation. Removal of the bottom liner layer **485**A may be done by any commonly known means dependent on the material composition. Advantageously, the removal process should be selective to the semiconductor substrate **302** and the STI (not shown). It may further be advantageously selective to the gate layer **378**, and the sidewall spacer **379**. In one particular embodiment employing a silicon nitride bottom liner layer **485**A, a phosphoric acid-based wet chemical etch is utilized. In an alternate embodiment employing a silicon dioxide bottom liner layer **485**A, an HF-based wet chemical etch may be used. Because the thickness of bottom liner layer **485**A is relatively small, the etch and overetch time may be kept short to avoid eroding underlayers. Furthermore, a wet chemical etch rate may be greatly enhanced for those embodiments previously described having a highly stressed bottom liner layer **485**A. In still other embodiments, a dry plasma etch employing commonly known process parameters may also be used to remove the unmasked portion of the bottom liner layer **485**A. Mask **499** may then be removed.

[0104] As shown in FIG. **4**G, a silicide process may then be performed on those areas with exposed silicon. 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. As depicted, silicide region **482** may be formed on the exposed gate layer **378** and exposed source and drain region while blocked by the bottom liner layer **485**A from the SONOS device **300** and HV device (e.g. MOS transistor **350**) regions.

[0105] In one embodiment, subsequent to the operations depicted in FIG. **4**G, processing proceeds, as shown in FIG. **5**, with a removal of the bottom liner layer **485**A and deposition of ILD **504**. Such an embodiment has the advantage of simplifying a subsequent contact etch because etching of ILD **504** will expose the source and drain **410**, the source and drain **460** and the silicide region **482** of the MOS transistor **370**. In this embodiment, the devices with silicide may be masked with any commonly known photosensitive mask material (i.e. photoresist) or non-photosensitive mask, such as amorphous carbon, which is first patterned. Masking of the silicided devices (e.g. MOS transistor **370**) is advantageous if the process employed to remove the bottom liner layer **485**A is nonselective to the silicide. Otherwise, a blanket strip of the bottom liner layer **485**A may be performed. A backend interconnect process may then begin with a deposition of ILD **504** over non-silicided SONOS device **300**, non-silicided HV MOS transistor **350** and silicided MOS transistor **370**.

[0106] In an alternative embodiment shown in FIG. **6**A, subsequent to the operations depicted in FIG. **4**G, processing proceeds with deposition of an ILD over the bottom liner layer **485**A and over the silicide region **482**. In such an embodiment, the bottom liner layer **485**A is incorporated as part of the backend ILD and subsequently removed during contact etch with an etch step selective over the silicide region **482**. In a particular embodiment employing a silicon nitride bottom liner layer **485**A, the bottom liner layer **485**A may be further utilized in a self-aligned contact (SAC) etch. The SAC etch, employing an etch recipe highly selective to silicon nitride may reduce the contact dimension from that lithographically printed to the physical space between adjacent devices covered with the bottom liner layer **485**A.

[0107] As further depicted in FIG. **6**A, the ILD layer deposited over the bottom liner layer **485**A and silicide region **482** may include a stress-inducing layer **504**A. Stress-inducing layer **504**A may be composed of any material and have any thickness suitable to exert a stress on channel region of a logic device. Stress-inducing layer **504**A may advantageously increase the carrier mobility and drive currents of a logic device, such as MOS transistor **370**. In accordance with an embodiment of the present invention, stress-inducing layer **504**A is disposed directly on MOS transistor **370**. In one embodiment, stress-inducing layer **504**A is deposited to a thickness approximately in the range of 20-100 nanometers and is composed of a material such as, but not limited to, silicon nitride, silicon oxy-nitride, carbon-doped silicon nitride or boron-doped silicon nitride. In a specific embodiment, stress-inducing layer **504**A is a tensile stress-inducing layer.

[0108] Because the stress inducing layer **504**A may induce an undesirable stress in the SONOS device **300**, causing performance and reliability degradation, the bottom liner layer **485**A may have been deposited under conditions to induce a stress opposing that of stress inducing layer **504**A. In a particular embodiment, the bottom liner layer **485**A may induce a compressive stress opposing a tensile stress in the stress inducing layer **504**A. In one such embodiment, the stress of the bottom liner layer **485**A reduces the cumulative stress on the SONOS device **300** to an amount less than half that induced by the stress inducing layer **504**A in absence of the bottom liner layer **485**A. The bottom liner layer **485**A may therefore provide both selective silicidation and selective stress induction for integration of the SONOS device **300** and a logic device, such as MOS transistor **370**. As further shown in FIG. **4**, ILD layer **504**B may then be deposited over stress inducing layer **504**A and planarized as part of a conventional backend interconnect process.

[0109] In still another embodiment, as depicted in FIG. **6**B, the bottom liner layer **485**A may provide an etch stop for removal of the stress-inducing layer **504**A from over the SONOS device **300** and HV MOS transistor **350**. The portion of stress-inducing layer **504**A above a nonvolatile charge trap memory device (e.g. SONOS device **300**) may be removed through a lithography and etch process selective to the bottom liner layer **485**A. In one embodiment, the portion of stress-inducing layer **504**A above MOS transistor **370** is first masked with a patterned photo-resist layer and the portion of stress-inducing layer **504**A above SONOS device **300** is then removed by a technique such as, but not limited to, a wet etch process using hot phosphoric acid or a conventional dry etch process.

[0110] In embodiments where the bottom liner layer **485**A is a silicon dioxide, the stress inducing layer **504**A may be removed with high selectively to the bottom liner layer **485**A. In other embodiments employing a silicon nitride bottom

UMC 1434
UMC v. AICP

liner layer **485**A, the higher stress of the stress inducing layer **504**A may provide selectivity to the bottom liner layer **485**A. The bottom liner layer **485**A therefore may provide protection to the underlying structures during patterning of the stress inducing layer **504**A. In alternate embodiments, removal of the stress inducing layer **504**A also removes the bottom liner layer **485**A.

[0111] As shown in FIG. **6**B, if the stress inducing layer **504**A is removed selectively to the bottom liner layer **485**A (e.g. to prevent a stress inducing layer **504**A from detrimentally impacting performance of SONOS device **300** or HV MOS transistor **350**), the ILD layer **504**B may then be deposited over both the bottom liner layer **485**A and over the stress inducing layer **504**A in preparation for further backend interconnect processing.

[0112] Thus, a semiconductor structure integrating charge trap memory devices with logic devices and method to form the same has been disclosed. Although the present invention has been described in language specific to structural features or methodological acts, it is to be understood that the invention defined in the appended claims is not necessarily limited to the specific features or acts described. The specific features and acts disclosed are to be understood as particularly graceful implementations of the claimed invention in an effort to illustrate rather than limit the present invention.

What is claimed is:

1. A method of fabricating a semiconductor structure, comprising:

implanting a first dopant species into a first region of a semiconductor substrate to form an n-well for a PMOS transistor; and

forming a non-volatile charge trapping dielectric stack over a second region of a semiconductor substrate after forming the PMOS transistor n-well, the non-volatile charge trapping dielectric stack including a blocking layer on a charge trapping layer over a tunneling layer formed on the semiconductor substrate.

2. The method of claim **1**, further comprising:

performing an implant photomask strip after implanting the first dopant species and prior to forming the non-volatile charge trapping dielectric stack.

3. The method of claim **1**, further comprising:

forming a gate oxide of the PMOS transistor over the first region of the semiconductor substrate after forming the non-volatile charge trapping dielectric stack.

4. The method of claim **3**, further comprising:

removing the non-volatile charge trapping dielectric stack over the first region of the semiconductor substrate;

performing a thermal oxidation to form a first gate insulator layer over the first region of the semiconductor substrate and to thermally reoxidize the blocking layer; and

performing a nitridation process to nitridize the first gate insulator layer and the blocking layer simultaneously.

5. The method of claim **1**, further comprising:

forming a first and second MOS gate stack over the first region of the semiconductor substrate;

forming a first sidewall spacer adjacent to the first MOS gate stack and a second sidewall spacer adjacent to the second MOS gate stack;

implanting a source and drain into the first region adjacent to the first sidewall spacer;

depositing a multi-layer liner over the second sidewall spacer, the multi-layer liner including a top layer and a bottom layer;

etching the top layer selective to the bottom layer to form a disposable sidewall spacer separated from the second sidewall spacer by at least the bottom layer;

implanting a source and a drain offset from the second sidewall spacer by the disposable sidewall spacer to increase a breakdown voltage; and

removing the disposable sidewall spacer selectively to the bottom layer.

6. The method of claim **1**, further comprising:

implanting a second dopant species into a third region of a semiconductor substrate to form a p-well for an NMOS transistor prior to forming the non-volatile charge trapping dielectric stack over the second region of a semiconductor substrate.

7. The method of claim **6**, further comprising:

performing an n-channel implant after forming the n-well and prior to forming the non-volatile charge trapping dielectric stack over the second region of a semiconductor substrate.

8. The method of claim **7**, further comprising:

performing a rapid thermal anneal of the substrate prior to forming the MOS gate oxide, after forming the n-well and p-well and after performing the n-channel implant, wherein the rapid thermal anneal is performed prior to forming the non-volatile charge trapping dielectric stack.

9. The method of claim **1**, wherein the substrate is single crystalline silicon, the tunneling layer comprises at least one of silicon dioxide or silicon oxy-nitride, the charge trapping layer comprises at least one of silicon nitride or silicon oxy-nitride, and the blocking layer comprises at least one of silicon dioxide or silicon oxy-nitride to form a non-volatile charge trapping dielectric stack of a SONOS memory device.

10. A method of fabricating a semiconductor structure, comprising:

implanting a first dopant species into a first region of a semiconductor substrate to form an n-well for a PMOS transistor;

implanting a second dopant species into a second region of the semiconductor substrate to form a p-well for an NMOS transistor;

forming a non-volatile charge trapping dielectric stack over a third region of a semiconductor substrate after forming the n-well and p-well, the non-volatile charge trapping dielectric stack including a blocking layer on a charge trapping layer over a tunneling layer, wherein the charge trapping layer further comprises a silicon nitride charge trapping layer comprising a non-homogenous silicon concentration across the thickness of the silicon nitride layer with a lower silicon concentration proximate to the tunneling layer;

forming a gate oxide of the MOS transistors over the first region of the semiconductor substrate after forming the non-volatile charge trapping dielectric stack; and

forming a gate electrode layer on the non-volatile charge trapping dielectric stack and on the gate oxide.

11. The method of claim **10**, further comprising:

performing an implant photomask strip after implanting the first or second dopant species and prior to forming the non-volatile charge trapping dielectric stack.

UMC 1434
UMC v. AICP

12. The method of claim **10**, further comprising:

performing an n-channel implant after forming the n-well and prior to forming the non-volatile charge trapping dielectric stack over the second region of a semiconductor substrate; and

performing a rapid thermal anneal of the substrate prior to forming the gate oxide, after forming the n-well and p-well and after performing the n-channel implant.

13. The method of claim **12**, wherein the rapid thermal anneal is performed prior to forming the non-volatile charge trapping dielectric stack.

14. The method of claim **12**, wherein the substrate is single crystalline silicon, the tunneling layer comprises at least one of silicon dioxide or silicon oxy-nitride, the charge trapping layer comprises at least one of silicon nitride or silicon oxy-nitride, and the blocking layer comprises at least one of silicon dioxide or silicon oxy-nitride to form a non-volatile charge trapping dielectric stack of a SONOS memory device.

15. A method of fabricating a SONOS memory device with CMOS devices, comprising:

performing all well and channel implants of the CMOS devices and SONOS memory device prior to forming a charge trapping dielectric stack of the SONOS memory device; and

forming the charge trapping dielectric stack prior to forming a gate insulator layer of the CMOS devices.

16. The method of claim **15**, wherein the non-volatile charge trapping dielectric stack includes a blocking layer on a charge trapping layer over a tunneling layer on a semiconductor substrate.

17. The method of claim **15**, further comprising:

performing a rapid thermal anneal process after performing all well and channel implants and prior to forming the charge trapping dielectric stack.

18. The method of claim **15**, wherein the substrate is single crystalline silicon, the tunneling layer comprises at least one of silicon dioxide or silicon oxy-nitride, the charge trapping layer comprises at least one of silicon nitride or silicon oxy-nitride, and the blocking layer comprises at least one of silicon dioxide or silicon oxy-nitride to form a non-volatile charge trapping dielectric stack of a SONOS memory device.

19. The method of claim **15**, wherein performing all well and channel implants of the CMOS devices and SONOS devices further comprises performing an n-well implant, p-well implant and at least one of an n-channel or p-channel implant.

20. The method of claim **15**, wherein all implant photomask strips associated with the well and channel implants are performed prior to forming the non-volatile charge trapping dielectric stack.

* * * * *