(74) Agent: SARLET, Steven; Holidaystraat 5, B-1831 Diegem (BE).


(84) Designated States (unless otherwise indicated, for every kind of regional protection available): AR, BE, BG, BH, BR, BY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LI, LU, LV, MC, MT, NL, NO, PL, PT, RO, RS, SL, SK, ST, TO, TR, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(54) Title: METHOD FOR FORMING A PLANAR STACKED GATE NONVOLATILE SEMICONDUCTOR MEMORY DEVICE HAVING A FLOATING GATE ELECTRODE AND DEVICES OBTAINED THEREOF.

(57) Abstract: A method of fabricating a stacked gate nonvolatile memory device, comprising the steps of forming a plurality of conductive floating gate structures above channel regions of the substrate, isolated from each other by a dielectric, the floating gate structure having a larger coupling ratio with control gates than with channel regions in the substrate. This comprises the steps of: (a) first forming the dielectric with cavities exposing the substrate at the channel regions, the cavities having predetermined shapes for shaping the floating gate structures such that their top sides have a larger area than their bottom sides, and (b) afterwards forming the floating gate structures in the cavities. A stacked gate nonvolatile memory device, comprising: a semiconductor substrate; a dielectric formed on the semiconductor substrate, the dielectric comprising cavities exposing the semiconductor substrate at channel regions; and homogenous floating gate structures in the cavities having a larger coupling ratio with the control gates than with channel regions in the substrate.
Method for forming a planar stacked gate nonvolatile semiconductor memory device having a floating gate electrode and devices obtained thereof

Field of the invention
[0001] The present invention relates to nonvolatile semiconductor memory devices and methods for fabricating such devices. In particular the present invention relates to planar stacked gate nonvolatile semiconductor memory devices whereby charge is stored on a floating gate electrode.

Background
[0002] A flash memory device is a nonvolatile semiconductor memory device comprising a stack of a charge storage gate and a control gate electrode, formed on a semiconductor substrate. The charge storage gate is separated from a channel region in the substrate by a dielectric stack, typically labeled tunnel dielectric, and separated from the control gate electrode by another dielectric stack, typically labeled interpoly dielectric. Self aligned to the charge storage gate contact regions, typically labeled junction regions, are formed for contacting the channel region. If a conductive material is used to store charge on, then the charge storage gate is labeled floating gate electrode. In such floating gate memory cell, the layer stack of control gate electrode, floating gate electrode and channel region separated by dielectric layers can be considered as a capacitive divider consisting of two capacitors in series.

[0003] By appropriate biasing of the control gate, and of the substrate and junction regions, mobile charge carriers are stored on the charge storage gate for programming the memory cell or removed from the charge storage gate for erasing the memory cell.

[0004] The programming and erasing voltage conditions and the efficiency with which mobile charge carriers can be transported to and from such floating gate depends on the coupling ratios between the floating gate electrode and the control gate electrode on the one hand, and between the floating gate electrode and the channel region on the other hand. The more voltage is coupled from the control gate electrode to the floating gate electrode, the more efficient the memory cell can be programmed and
erased. The coupling ratio between the control and floating gate electrodes must be larger than the coupling ratio between the floating gate electrode and the channel region to ensure sufficient voltage coupling between the control gate electrode and the floating gate electrode.

[0005] The coupling ratio between the control gate electrode and the floating gate electrode is function of the dielectric properties of the interpoly dielectric stack and of the overlap area between both electrodes. Likewise the coupling ratio between the floating gate electrode and the channel region will be function of the dielectric properties of the tunnel dielectric stack and of the overlap area between the floating gate electrode and the channel region.

[0006] The larger this overlap area and/or the larger the overall dielectric constant of the dielectric stack, the larger the coupling ratio will be. Therefore dielectric materials, having a dielectric constant higher than this of silicon oxide, are used to form the interpoly dielectric stack. These dielectric materials are also known as high-k dielectrics. Increasing the overlap area between the control floating gate electrodes can be obtained by having the control gate electrode overlying more than one surface of the floating gate electrode. However having the control gate electrode not only overlapping an upper surface of the floating gate electrode, but also sidewalls, requires a more complex, non-planar, fabrication process. In particular the patterning of the control gate electrode becomes more difficult. Moreover the spacing between adjacent memory devices, thus fabricated, has to be sufficiently large to accommodate for the control gate electrode overlying the sidewalls of the floating gate electrode of each memory device.

[0007] United States patent application US 6,878,985 discloses in some of its embodiments a method for forming a planar stacked, floating gate, flash memory device whereby the overlap area between the control and floating gate electrodes is larger than the overlap area between the floating gate electrode and the channel region. This difference in overlap area is obtained by forming a floating gate electrode being wider near the control gate electrode than near the channel region.

[0008] In these embodiments the floating gate is formed as a stack of a first and a second conductive layer. The first conductive layer is patterned and used to etch trenches in the semiconductor substrate thereby forming active areas self aligned to this
first conductive layer. The width of the floating gate electrode near the channel region is defined by the width of the self aligned first conductive layer and active area. These trenches are then filled with dielectric material. By extending this dielectric material above the level of the first conductive layer a recess is formed in this dielectric layer self aligned to this first conductive layer. This recess is widened by an isotropic etch. This widened recess is then filled with the second conductive layer unto the level of the dielectric material to complete the floating gate electrode. By increasing the width of the recess beyond the width of the first conductive layer, the width of the floating gate electrode near the floating gate electrode is also increased.

Although, the methods disclosed in US 6,878,985 provide a difference in overlap area between the floating gate electrode and the control gate electrode and channel region respectively, without introducing topography when forming the interpoly dielectric stack and the control gate electrode, these methods show several deficiencies. The tunnel dielectric and the first conductive layer of the floating gate electrode are exposed to several process steps which may have a detrimental effect on the quality of these layers. Such process steps are dry etching of the trenches, thermal processing and deposition of dielectric material in these trenches. The overlap area between the control and floating gate electrodes is determined inter alia by the isotropic etch widening the recess in the dielectric material. However when widening the recess, care must be taken not expose the complete sidewall of the first conductive layer during the isotropic etch. As the maximum width of the recess after the isotropic etch is determined inter alia by the height of the first conductive layer, optimization of the coupling ratios of the floating gate electrode requires modifying of at least the height of the first conductive layer. Modifying this height may impact the patterning process and overall height of the nonvolatile memory device. The provided fabrication methods only allow forming a floating gate electrode as a stack of at least two conductive layers whereby an interface region may be present between subsequent conductive layers. This interface region may impact the charge storage and retention properties of the floating gate electrode.

Disclosure of the invention
It is an object of the invention to provide a method for fabricating a nonvolatile memory device which does not suffer from at least one deficiency of the prior art.

This aim is achieved according to the invention with the method of the first independent claim. Preferred embodiments of the method of the invention are expressed in the claims depending from the first independent claim.

It is another object of the invention to provide a stacked gate flash memory device with improved characteristics with respect to the prior art.

This aim is achieved according to the invention with the method of the second independent claim. Preferred embodiments of the device of the invention are expressed in the claims depending from the second independent claim.

According to a first aspect of the invention, a method is disclosed for fabricating a stacked gate nonvolatile memory device. The method comprising the steps of forming a plurality of conductive floating gate structures above channel regions of said substrate, isolated from each other by a dielectric, said floating gate structures having bottom sides aligned with said channel regions and top sides aligned with control gate structures which are formed above said floating gate structures. Said top sides have a larger area than said bottom sides, such that the floating gates have a larger coupling ratio with said control gates than with said channel regions. The step of forming the conductive floating gate structures isolated from each other by said dielectric comprises the steps of (a) first forming said dielectric with cavities exposing said substrate at said channel regions, said cavities having predetermined shapes for shaping said floating gate structures such that their top sides have a larger area than their bottom sides, and (b) afterwards forming said floating gate structures in said cavities.

An advantage of embodiments of the method of the invention is that the impact of the fabrication process on physical and electrical properties of the tunnel dielectric and/or floating gate electrode is/are reduced.

An advantage of embodiments of the method of the invention is that the coupling ratio between the control and floating gate electrode and between the channel region and the floating gate electrode can be controlled independently of the height of the floating gate electrode.
[0017] An advantage of embodiments of the method of the invention is that a scalable method is provided for fabricating a nonvolatile memory device comprising a floating gate electrode.

[0018] The method of the invention is particularly suitable for fabricating a planar stacked gate non-volatile memory device with homogenous floating gate structures. However, this does not exclude applicability of the method of the invention for fabricating other stacked gate non-volatile memory devices, e.g. non-planar devices and/or devices without homogenous floating gate structures, e.g. multi-layer floating gate structures.

[0019] According to a second aspect of the invention, a stacked gate nonvolatile memory device is disclosed, comprising a semiconductor substrate; a dielectric formed on the semiconductor substrate, the dielectric comprising cavities exposing the semiconductor substrate at channel regions; and floating gate structures having bottom sides aligned with said channel regions and top sides aligned with control gate structures which are formed above said floating gate structures, wherein said top sides have a larger area than said bottom sides, such that the floating gates have a larger coupling ratio with said control gates than with said channel regions. The floating gate structures are homogenous, i.e. despite the difference in area between the top and bottom sides, they are homogeneously formed in the same material.

[0020] In preferred embodiments, the device is a planar stacked gate flash memory device in which the width of the floating gate electrode varies from the channel region towards the control gate electrode. Preferably the width of the floating gate electrode is the largest near the control gate electrode.

[Brief description of the drawings]

[0021] Exemplary embodiments of the invention are illustrated in referenced figures of the drawings. The embodiments and figures disclosed herein must be considered illustrative rather than restrictive. The drawings described are only schematic and are non-limiting. In the drawings, the size of some of the elements may be exaggerated and not drawn on scale for illustrative purposes and need not to correspond
to actual reductions to practice of the invention. Same numerals are used to refer to corresponding features in the drawings.

[0022] Figure 1 shows a schematic cross-sectional view of a nonvolatile memory device according to a preferred embodiment.

[0023] Figure 2 shows a schematic cross-sectional view of a nonvolatile memory device according to an embodiment.

[0024] Figures 3a-f illustrate, by means of schematic cross-sectional views, process steps of a fabrication method according to an embodiment.

[0025] Figures 4a-k illustrate, by means of schematic cross-sectional views, process steps of a fabrication method according to a preferred embodiment.

[0026] Figure 5 shows a cross-sectional Transmission Electron Microscopy (TEM) picture of a nonvolatile memory device fabricated according to a preferred embodiment.

[0027] Figure 6 shows a cross-sectional Scanning Electron Microscopy (SEM) picture of another nonvolatile memory device fabricated according to a preferred embodiment.

**Modes for carrying out the invention**

[0028] The terms first, second, third and the like in the description and in the claims, are used for distinguishing between similar elements and not necessarily for describing a sequence, either temporally, spatially, in ranking or in any other manner. It is to be understood that the terms so used are interchangeable under appropriate circumstances and that the embodiments of the invention described herein are capable of operation in other sequences than described or illustrated herein.

[0029] The terms top, bottom, over, under and the like in the description and the claims are used for descriptive purposes and not necessarily for describing relative positions. It is to be understood that the terms so used are interchangeable under appropriate circumstances and that the embodiments of the invention described herein are capable of operation in other orientations than described or illustrated herein.

[0030] It is to be noticed that the term “comprising”, used in the claims, should not be interpreted as being restricted to the means listed thereafter; it does not
exclude other elements or steps. It is thus to be interpreted as specifying the presence of the stated features, integers, steps or components as referred to, but does not preclude the presence or addition of one or more other features, integers, steps or components, or groups thereof. Thus, the scope of the expression “a device comprising means A and B” should not be limited to devices consisting only of components A and B. It means that with respect to the present invention, the only relevant components of the device are A and B.

[0031] Reference throughout this specification to “one embodiment” or “an embodiment” means that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, appearances of the phrases “in one embodiment” or “in an embodiment” in various places throughout this specification are not necessarily all referring to the same embodiment, but may. Furthermore, the particular features, structures or characteristics may be combined in any suitable manner, as would be apparent to one of ordinary skill in the art from this disclosure, in one or more embodiments.

[0032] Similarly it should be appreciated that in the description of exemplary embodiments of the invention, various features of the invention are sometimes grouped together in a single embodiment, figure, or description thereof for the purpose of streamlining the disclosure and aiding in the understanding of one or more of the various inventive aspects. This method of disclosure, however, is not to be interpreted as reflecting an intention that the claimed invention requires more features than are expressly recited in each claim. Rather, as the following claims reflect, inventive aspects lie in less than all features of a single foregoing disclosed embodiment. Thus, the claims following the detailed description are hereby expressly incorporated into this detailed description, with each claim standing on its own as a separate embodiment of this invention.

[0033] Furthermore, while some embodiments described herein include some but not other features included in other embodiments, combinations of features of different embodiments are meant to be within the scope of the invention, and form different embodiments, as would be understood by those in the art. For example, in the following claims, any of the claimed embodiments can be used in any combination.
[0034] In the description provided herein, numerous specific details are set forth. However, it is understood that embodiments of the invention may be practiced without these specific details. In other instances, well-known methods, structures and techniques have not been shown in detail in order not to obscure an understanding of this description. Such methods are known to a person skilled in the art.

[0035] In the framework of the present invention, terms such “planar” or “planar technology” refers to methods for fabricating semiconductor devices whereby, at least during the steps of patterning layers, the surface of the substrate at the processing side is substantially planar, i.e. at substantially the same level. As essentially no topography is present on the substrate the process window of the patterning steps, in particular the lithographic part thereof, becomes less stringent. Moreover, as the substrate is essentially flat, layers can be formed on such flat substrate without concern for step coverage or layer conformality.

[0036] Figure 1 shows a schematic cross-sectional view of a nonvolatile memory device according to a preferred embodiment of the invention. The memory device 1 comprises a floating gate electrode 6 dielectrically coupled to a channel region, formed in an active area 4 of a semiconductor substrate 2, by a tunnel dielectric 5. The floating gate electrode 6 is also dielectrically coupled to a control gate electrode 8 by an interpoly dielectric 7. The active area 4, the tunnel dielectric 5 and the floating gate electrode 6 are embedded in an insulating structure 3. This isolating structure 3 is preferably a Shallow Trench Isolation (STI) structure comprising a trench formed in the substrate 2, the trench being filled with a dielectric material. The interpoly dielectric 7 and the control gate electrode 8 overlap the floating gate electrode 6 and, at least partially, the insulating structure 3. Typically this control gate electrode 8 is part of a word line, common to other memory devices 1. The cross-section of figure 1 is made along the direction along which this word line 8 runs when the memory device 1 is part of a memory matrix.

[0037] Not shown in the cross-sectional view of figure 1 are the contact regions (source and drain) formed in the active area 4 for contacting opposite sides of the channel region of this active area 4. These contact regions are located in planes parallel to
and at opposite sides of the plane of figure 1. These contact regions will become part of a bit line when the memory device 1 is part of a memory matrix.

[0038] The insulating structure 3 in which the active area 4, the tunnel dielectric 5 and the floating gate electrode 6 are embedded, extends from the level of the active area 4 up to the level of the floating gate electrode 6. The floating gate electrode 6 and the insulating structure 3 thereby form a substantially planar surface as the upper surfaces of both elements 3, 6 are substantially leveled as shown in figure 1. Hence the interface between the interpoly dielectric 7 and the floating gate electrode 6, insulating structure 3 is substantially planar.

[0039] The width W of the floating gate electrode 6 is not constant. The width of the floating gate electrode 6 is defined as its dimension along the direction of the word line. Near the active area 4 the width W of floating gate electrode 6 is the smallest, while near the control gate electrode the width is the largest. In the preferred embodiment shown in figure 1, the floating gate electrode 6 is formed of a lower part, i.e. the part near the active area 4, and an upper part, i.e. the part near the control gate electrode 8. The lower part of the floating gate electrode 6 is self aligned to the active area 4. The lower part and the upper part have a respective width W1 and W2 whereby W1 < W2. Although in figure 1 the width of the bottom and upper parts remains constant over their respective cross-sections, the width of the upper part of the floating gate electrode 6 needs not to be constant over the cross-sections. It can gradually increase from the width W1 of the bottom part to the width W2 near the upper surface of the floating gate electrode 6 as shown in figure 2.

[0040] The floating gate electrode 6 of figures 1 and 2 is homogeneously formed. The bottom and upper parts of the floating gate electrode 6 are formed of the same conductive material in a continuous way without an interface region between both parts. In alternative embodiments, different materials may be used for the bottom and upper parts, or the same material with an interface region between both parts.

[0041] Figures 3a-f illustrate a method for manufacturing a nonvolatile floating gate memory device according to embodiments of the invention. For the purpose of the teaching the invention it is shown in these figures how a NAND configuration of two memory devices can be fabricated. However the invention is not limited to this
memory configuration, but can be applied to any memory configuration of a planar stacked gate nonvolatile memory device comprising a floating gate electrode.

[0042] A semiconductor substrate 2 is provided. This semiconductor substrate can be a bulk wafer or a semiconductor-on-insulator substrate such as a silicon-on-insulator (SOI) or germanium-on-insulator (GOI) substrate. On this semiconductor substrate a first hardmask layer 9 having a thickness H₁ is formed. This first hardmask layer 9 can be a single layer or a stack of layers. The material(s) of the first hardmask layer 9 are chosen to allow selective removal from the semiconductor material of the substrate 2. This first hardmask layer 9 is patterned to define active areas 4 in the semiconductor substrate 2. Self-aligned to the patterned first hardmask layer 9 trenches 3 are etched in the semiconductor substrate 2 thereby outlining the active areas 4. These trenches 3 are filled with a first dielectric material up to the level of the first hardmask layer 9 so as to form a substantially planar surface of the substrate as shown in figure 3a. The thickness H₁ of the first hardmask layer 9 determines to which extent the first dielectric material of the filled trench 3 extends above the upper surface of the active area 4. Also illustrated in figure 3a is the separation between adjacent memory devices provided by the filled trenches 3.

[0043] Overlying the filled trench 3 and the exposed patterned first hardmask layer 9 a second hardmask layer 10 having a thickness H₂ is formed. Also the material(s) of the second hardmask layer 10 are chosen to allow selective removal from the semiconductor material of the substrate 2. Preferably this second hardmask layer 10 is formed of the same material(s) as the first hardmask layer 9. This second hardmask layer 10 is patterned so as to fully overlap the patterned first hardmask layer 9. The layout of the patterned second hardmask layer 10 thus corresponds to the layout of the patterned first hardmask layer 9, whereby the dimensions W₂ of the patterned second hardmask layer 10 are larger than the dimensions W₂ of the patterned first hardmask layer 9: W₁ < W₂. The overlap (W₂ − W₁) of the patterned second hardmask layer 10 over the patterned first hardmask layer 9 is shown in figure 3b. The dimensions W₁, H₁ and W₂, H₂ of respectively the first and the second hardmask layer can be chosen independently from each-other as long as the patterned second hardmask layer 10 overlaps the patterned first hardmask layer 9, i.e. as long as the requirement W₁ < W₂ is met.
The topography created by the patterned second hardmask layer 10 is planarized by further filling the trenches 3 with a second dielectric material to the level of the second hardmask layer 10 so as to form again a substantially planar surface of the substrate as shown in figure 3c. The thickness $H_2$ of the second hardmask layer 9 determines the extent to which the second dielectric material of the filled trench 3 extends above the upper surface of the first hardmask layer 9.

The patterned second 10 and first 9 hardmask layers are then removed selectively with respect to the dielectric materials of the trench 3 thereby exposing the active area 4 as shown in figure 3d. By removing these patterned hardmask layers 9, 10 a cavity 11 is created in the dielectric materials of the trench 3. Thanks to the overlap ($W_2 - W_1$) of the patterned second hardmask layer 10 over the patterned first hardmask layer 9 this cavity will have a width that is smallest near the active area 4 and largest near the upper surface of the dielectric materials of the trench 3.

On the active area 4, exposed at the bottom of the cavity 11, the tunnel dielectric 5 is formed. This tunnel dielectric 5 can be a single layer of dielectric material, such as for example silicon oxide, or can be a stack of layers of dielectric material(s) such as for example a combination of silicon oxide and/or high-k dielectries. The cavity 11 is then filled with conductive material(s) up to the level of the dielectric materials of the trench 3 so as to form a substantially planar surface of the substrate as shown in figure 3e. Semiconductive materials such as silicon, silicon germanium can be used to fill the cavity 11. These conductive material(s) constitute the floating gate electrode 6. Preferably a single conductive material is used to fill the cavity 11. This single conductive material can then be formed in a continuous way such that the floating gate electrode 6 is formed homogeneously. The active area 4, the tunnel dielectric 5 and the floating gate electrode 6 are embedded in the first and second dielectric materials used to fill the trench 3. They form an insulating structure 3 surrounding the active area 4, the tunnel dielectric 5 and the floating gate electrode 6.

The floating gate electrode 6 has a thickness $H_3$ which depends inter alia on the thickness $H_1$, $H_2$ of the first 9 and second 10 hardmask layers respectively. The bottom part of the floating gate electrode 6, i.e. near the active area 4, is self aligned thereto. As the dimensions of the floating gate electrode 6 are defined by the dimensions
of the cavity 11, the width of the floating gate will increase from the active area 4 upwards. The bottom part of the floating gate electrode 6, i.e. near the active area 4, has a width \( W_1 \) while the upper part of the floating gate electrode 6 has a width \( W_2 \) whereby \( W_1 < W_2 \).

[0048] On the substantially flat surface of the substrate the dielectric material(s) of the interpoly dielectric 7 is formed. This interpoly dielectric 7 can be a single layer of dielectric material, such as for example silicon oxide, or can be a stack of layers of dielectric material(s) such as for example a combination of silicon oxide and/or high-k dielectrics. Preferably the interpoly dielectric 7 comprises a high-k layer. Preferably this high-k layer comprises an oxide of Aluminum and/or Hafnium. Overlying the interpoly dielectric 7 the conductive material(s) of the control gate electrode 8 are formed. This conductive material can be a metal or a silicide, or a semiconductive material such as polycrystalline silicon or a combination thereof. The stack of dielectric and conductive materials is then patterned to form the control gate electrode 8 and the interpoly dielectric 7 in a self aligned manner. Thanks to fact that the control gate electrode 8 only overlaps the floating gate electrode 6 on the upper surface thereof, the spacing between the two adjacent memory cells can be reduced. Hence the density of a memory matrix comprising nonvolatile memory device according to embodiments can be increased. As shown in figure 3f the control gate electrode 8 runs over more than 1 memory device thereby forming a word line. Using the control gate electrode 8 as a hardmask, the exposed parts of the floating gate electrode 6 are selectively removed (not shown in figure 3f). The floating gate electrode 6 is now also self aligned to the control gate electrode 8.

[0049] In the embodiment illustrated by figure 3e the width of each of the two parts of the floating gate electrode 6 is constant over the respective cross-sections. This corresponds to the nonvolatile memory device 1 illustrated by figure 1. However the width of the upper part of the floating gate electrode 6 need not be constant over its width. In an alternative embodiment the cavity 11 is formed in two steps. After leveling the topography of the patterned second hardmask layer 10 by further filling the trench 3 with the second dielectric materials as shown in figure 3e, the patterned second hardmask layer 10 is removed down to the patterned first hardmask layer 9. If different materials
are used to form respectively the first 9 and second 10 hardmask layers, this patterned second hardmask layer 10 can be selectively removed with respect to the patterned first hardmask layer 9. If the same or similar materials are used, then a timed etch can be possible to at least remove a part of the patterned second hardmask layer 10. Then an isotropic etch is performed thereby slanting the sidewalls of the partially formed cavity 11. By selectively removing the rest of the hardmask layer(s) the cavity 11 is completed. Processing can continue as shown by figures 3d to 3f thereby yielding the nonvolatile memory device of figure 2.

[0050] Figures 4a-k illustrate a method for manufacturing a nonvolatile floating gate memory device according to another preferred embodiment of the invention. For the purpose of the teaching the invention it is shown in these figures how a NAND configuration of two memory devices can be fabricated. However the invention is not limited to this memory configuration, but can be applied to any memory configuration of a planar stacked gate nonvolatile memory device comprising a floating gate electrode.

[0051] A bulk silicon substrate 2 is provided. The substrate is thermally oxidized to form a thin oxide layer, also known as padoxide (not shown) On this thin oxide layer the first hardmask layer 9 having a thickness H₁ is formed. This first hardmask layer 9 is for example a silicon nitride layer formed by Plasma Enhanced Chemical Vapor deposition (CVD) having a thickness H₁ of 115 nm. This first hardmask layer 9 is patterned using 193nm optical lithography. Self aligned to the patterned first hardmask layer 9 trenches 3 in the semiconductor substrate 2 are dry etched thereby outlining the active areas 4. In a 130nm technology the minimal width of the active area 4 and hence of the patterned first nitride layer 9 is W₁ = 150nm. These trenches 3 are filled with an oxide as first dielectric material. Oxide is also formed over the patterned nitride layer 9 as shown in figure 4a. This oxide is deposited using High Density Plasma Chemical Vapor deposition (HDP-CVD) technique providing a good fill of the trenches 3. The thickness of the HDP oxide layer in the trench 3 is 500 nm.

[0052] The substrate is then planarized using Chemical Mechanical Polishing (CMP) using the nitride of the first hardmask layer 9 as a stopping layer. Essentially no oxide is to remain on the first nitride layer 9. During this polishing process the oxide above the level of the patterned first nitride layer 9 is removed resulting in a
substantial planar surface of the substrate as shown in figure 4b. The thickness $H_1$ of the
now exposed nitride layer 9 determines to which extent the first dielectric material of the
filled trench 3 extends above the upper surface of the active area 4.

[0053] Overlying the filled trench 3 and the exposed patterned first
hardmask layer 9 a second hardmask layer 10 is formed as shown in figure 4c. This
second hardmask layer 10 is preferably also a silicon nitride layer formed by Plasma
Enhanced Chemical Vapor deposition (CVD) having a thickness $H_2$ of 115 nm. This
second nitride layer 10 is patterned using 193nm optical lithography, whereby the layout
of the second nitride layer 10 is an oversized version of the layout of the first nitride layer
9. The overlap of the patterned second nitride layer 10 over the patterned first nitride
layer 9 should be larger than the minimal overlay of the photolithographic process which
accounts for any misalignment errors: $W_2 > W_1 + 2$ overlay. In 130nm technology this
minimal overlay is 30nm. Layer 12 is a photosensitive layer which is used during the
optical lithography patterning process. This second nitride layer 10 is thus patterned so as
to fully overlap the patterned first nitride layer 9. The overlap $(W_2 - W_1)$ of the patterned
second nitride layer 10 over the patterned first nitride layer 9 is shown in figure 4d.

[0054] A second layer of dielectric material is then deposited overlying the
substrate as shown in figure 4e. Preferably about 200nm of oxide is deposited. This oxide
is deposited using the High Density Plasma Chemical Vapor deposition (HDP-CVD)
technique. The substrate is again planarized using Chemical Mechanical Polishing (CMP)
using the nitride of the second hardmask layer 10 as stopping layer. Essentially no oxide
is to remain on the second nitride layer 10. During this polishing process the oxide above
the level of the patterned second nitride layer 10 is removed resulting in a substantially
planar surface of the substrate as shown in figure 4f. The thickness $H_2$ of the second
hardmask layer 9 determines now the extent to which the second dielectric material of the
filled trench 3 extends above the upper surface of the first nitride layer 9.

[0055] The first 9 and second 10 nitride layers form a cavity in the dielectric
materials of the trench 3. The height of this cavity is determined by the combined height
$H_1 + H_2$ of the first 9 and second 10 nitride layers. By replacing later on the process this
nitride by polycrystalline silicon, a floating gate electrode 6 is formed within the
dielectrics of the trench 3. If the total height of the floating gate is to be reduced, the
oxide of the trench 3 can be etched back thereby reducing the height extension of the trench 3 dielectric with respect to the active area 4. A wet HF etch can be used to selectively remove part of the oxide in the trench 3. The height of the cavity is now \( H_3 \) whereby \( H_3 < H_1 + H_2 \) as shown in figure 4g.

5 [0056] The patterned second 10 and first 9 nitride layers are then removed selectively with respect to the oxide of the trench 3 thereby exposing the active area 4 as shown in figure 4h. By removing these patterned nitride layers 9, 10 a cavity 11 is created in the dielectric materials of the trench 3. Thanks to the overlap \( (W_2 - W_1) \) of the patterned second nitride layer 10 over the patterned first nitride layer 9 this cavity will have a width that is smallest near the active area 4 and largest near the upper surface of the dielectric materials of the trench 3. The height of the cavity is now \( H_3 \) whereby \( H_3 \leq H_1 + H_2 \), depending on whether or not an additional oxide etch is performed after polishing the second layer of trench dielectric. Preferably an additional oxide etch is performed resulting in a thickness \( H_3 \) of 125 nm. When sealing the nonvolatile memory device, this thickness \( H_3 \) is preferably also sealed. Through the padoxide at the bottom of this cavity 11, ions are implanted in the active area 4 to form doped regions, known as well regions, therein. The padoxide is then removed typically using a wet oxide etch such as an HF etch. By removing the padoxide part, the dielectric in the trench 3 will also be removed whereby the width \( W_1 \) increases to about 180nm. Hence \( W_2 > W_1 + 2 \) overlay = 180 + 2 *30nm \( \equiv \) 240nm. The cavity 11 will also overlap now the active area 4.

[0057] On the exposed active area 4, exposed at the bottom of the cavity 11, the tunnel dielectric 5 is then formed. This tunnel dielectric 5 can be a single layer of dielectric material, such as for example silicon oxide, or can be a stack of layers of dielectric material(s) such as for example a combination of silicon oxide and/or high-k dielectrics. Then a layer 13 of polycrystalline silicon is deposited in a continuous way overlying the substrate and at least filing the cavity 11 as shown in figure 4i. Silicon above the level of the trench 3 is removed by Chemical Mechanical Polishing (CMP) as to form a substantially planar surface of the substrate as shown in figure 4j. By replacing the nitride of the hardmask layers 9, 10 with polycrystalline silicon a homogeneous floating gate electrode 6 is formed within the dielectrics of the trench 3. The active area 4, the tunnel dielectric 5 and the floating gate electrode 6 are embedded in the HDP oxide
of the trench 3. This HDP oxide forms an insulating structure 3 surrounding sidewalls of the active area 4, the tunnel dielectric 5 and the floating gate electrode 6.

[0058] The bottom part of the floating gate electrode 6, i.e. near the active area 4, is self aligned thereto. As the dimensions of the floating gate electrode 6 are defined by the dimensions of the cavity 11, the width of the floating gate will increase from the active area 4 upwards. The bottom part of the floating gate electrode 6, i.e. near the active area 4, has a width W₁ while the upper part of the floating gate electrode 6 has a width W₂ whereby W₁ < W₂.

[0059] On the substantially flat surface of the substrate the dielectric material(s) of the interpoly dielectric 7 is formed. This interpoly dielectric 7 preferably comprises a high-k layer, such as for example AlₓOᵧ or HfₓAlₓOᵧ with x, y, z being integers and x+y+z = 1. Overlying the interpoly dielectric 7 the conductive material(s) of the control gate electrode 8 are formed. Preferably this control gate electrode 8 is formed of polycrystalline silicon. Optionally a thin metal layer such as a TiN layer is formed inbetween the interpoly dielectric 7 and the silicon of the control gate electrode 8. The stack of dielectric and conductive materials is then patterned to form the control gate electrode 8 and the interpoly dielectric 7 in a self aligned manner. As shown in figure 4k the control gate electrode 8 runs over more than 1 memory device thereby forming a word line. Using the control gate electrode 8 as a hardmask the exposed parts of the floating gate electrode 6 are selectively removed (not shown in figure 3f). The floating gate electrode 6 is now also self aligned to the control gate electrode 8.

[0060] Further process steps, known in semiconductor process technology, are then performed to complete the nonvolatile memory device. These further process steps may comprise forming by ion implantation the contact regions to the channel region, silicidation of exposed silicon of the contact regions and the control gate electrode 8, forming the interconnect structure comprising layers of dielectric and metal materials preferably using damascene technology to interconnect the individual nonvolatile memory devices in the memory matrix and the peripheral logic circuitry.

[0061] Figure 5 shows a cross-sectional Transmission Electron Microscopy (TEM) picture of a nonvolatile memory device fabricated according to the preferred embodiment. The cavity 11 containing the polycrystalline silicon floating gate electrode 6
overlaps the active area 4 by about 9.1nm on each side. This overlap is created by the wet etch of the padoxide prior to the formation of the tunnel oxide 5 on the active area 4. The bottom part of the floating gate electrode 6 has a width \( W_1 \) of about 180nm, while the top part has a width of \( W_2 \) of about 231.8nm. One can see that the patterned second nitride layer 10 was slightly misaligned over the patterned first nitride layer 9, as the overlap \( (W_2-W_1) \) is not completely symmetrical, i.e. about 32.6nm on the left side and about 17.9nm on the right side. Overlying the substantially flat surface provided by the homogeneous floating gate electrode 6 and the insulating structure 3, is the interpoly dielectric 7 of about 12nm Al₂O₃. On top of the interpoly dielectric 7 is the polycrystalline control gate electrode 8, which is partly removed to expose the top surface of the interpoly dielectric 7.

[0062] Figure 6 shows a cross-sectional Scanning Electron Microscopy (SEM) picture of another nonvolatile memory device fabricated according to the preferred embodiment. The arrow indicates the 30nm field overlap of the floating gate 6. The total thickness of the floating gate is about 100nm (80nm of the self-aligned part). In this figure, it can be seen that the planarity of the structure is maintained when introducing the T-shaped floating gate.
CLAIMS

1. A method of fabricating a stacked gate nonvolatile memory device, the method comprising the steps of forming a plurality of conductive floating gate structures above channel regions of said substrate, isolated from each other by a dielectric, said floating gate structures having bottom sides aligned with said channel regions and top sides aligned with control gate structures which are formed above said floating gate structures, wherein said top sides have a larger area than said bottom sides, such that the floating gates have a larger coupling ratio with said control gates than with said channel regions, characterized in that the step of forming the conductive floating gate structures isolated from each other by said dielectric comprises the steps of:
   a) first forming said dielectric with cavities exposing said substrate at said channel regions, said cavities having predetermined shapes for shaping said floating gate structures such that their top sides have a larger area than their bottom sides, and
   b) afterwards forming said floating gate structures in said cavities.

2. The method of claim 1, wherein step a) comprises the following steps:
   e) forming a patterned first hardmask layer on the substrate,
   d) forming a first layer of said dielectric on the substrate in between portions of said first hardmask layer such that the first dielectric layer and the patterned first hardmask layer form a substantially flat top surface,
   e) forming a patterned second hardmask layer aligned to and fully overlapping the patterned first hardmask layer,
   f) forming a second layer of said dielectric on the first layer of dielectric in between portions of said second hardmask layer such that the second dielectric layer and the patterned second hardmask layer form a substantially flat surface, and
   g) creating said cavities in said dielectric by removing the patterned first and second hardmask layers.
3. The method of claim 2, wherein said first and second hardmask layers are formed in
different materials such that said second hardmask layer is selectively removable from
said first hardmask layer and wherein step g) comprises the steps of:
h) removing said second hardmask layer, thereby forming upper parts of said cavities,
i) widening said upper parts of said cavities,
j) removing said first hardmask layer, thereby forming lower parts of said cavities.

4. The method of claim 2, wherein said portions of said first and second hardmask layers
respectively have substantially continuous first and second widths, said second width
being larger than said first width.

5. The method of claim 4, wherein said first and second hardmask layers are formed in
the same material and are removed in step g) a single process step.

6. The method of any one of the claims 1-5, wherein step b) comprises filling said
cavities with conductive material such that the floating gate structures and the dielectric
form a substantially flat surface.

7. The method of any one of the claims 1-6, further comprising the steps of forming
tunnel dielectrics between said floating gate structures and said channel regions and
interpoly dielectrics between said floating gate structures and said control gate structures.

8. The method of claim 7, wherein said tunnel dielectrics are formed between steps a) and
b).

9. A stacked gate nonvolatile memory device, comprising
   - a semiconductor substrate,
   - a dielectric formed on the semiconductor substrate, the dielectric comprising cavities
     exposing the semiconductor substrate at channel regions,
   - floating gate structures in said cavities, having bottom sides aligned with said
     channel regions and top sides aligned with control gate structures which are formed
above said floating gate structures, wherein said top sides have a larger area than said bottom sides, such that the floating gates have a larger coupling ratio with said control gates than with said channel regions, characterized in that said floating gate structures are homogenous.

10. A stacked gate nonvolatile memory device according to claim 9, wherein said floating gate structures and said dielectric form a substantially flat surface.

11. A stacked gate nonvolatile memory device according to claim 9 or 10, wherein said floating gate structures have a width which gradually increases from bottom side to top side.

12. A stacked gate nonvolatile memory device according to claim 9 or 10, wherein said floating gate structures each comprise a lower part upwards from said bottom side and an upper part downwards from said top side, said lower and upper parts respectively having substantially continuous first and second widths, said second width being larger than said first width.

13. A stacked gate nonvolatile memory device according to any one of the claims 9-12, further comprising tunnel dielectrics between said floating gate structures and said channel regions and interpoly dielectrics between said floating gate structures and said control gate structures.
**INTERNATIONAL SEARCH REPORT**

**A. CLASSIFICATION OF SUBJECT MATTER**

INV. H01L21/28  H01L21/8247  H01L27/115  H01L29/788

According to International Patent Classification (IPC) or to both national classification and IPC:

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

H01L

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

Electronic data base consulted during the International search (names of data base and, where practical, search terms used)

EPO-Internal

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
</table>

☐ Further documents are listed in the continuation of Box C.  

☐ See patent family annex.

* Special categories of cited documents:

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

  *E* earlier document but published on or after the international filing date

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

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

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

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

  *X* document of particular relevance; the claimed invention cannot be considered to be obvious over such document(s) taken alone

  *Y* document of particular relevance; the claimed invention cannot be considered to be obvious over such document(s), such combination being obvious to a person skilled in the art.

  *A* document member of the same patent family

**Date of the actual completion of the International search**

4 September 2008

**Date of mailing of the International search report**

12/09/2008

**Name and mailing address of the ISA**

European Patent Office, P.B. 5816 Patentlaan 2 NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-0400, Tx. 31 651 epo nl,  
Fax. (+31-70) 340-3015

**Authorized officer**

Franche, Vincent
<table>
<thead>
<tr>
<th>Patent document cited in search report</th>
<th>Publication date</th>
<th>Patent family member(s)</th>
<th>Publication date</th>
</tr>
</thead>
<tbody>
<tr>
<td>US 2002195650 A1</td>
<td>26-12-2002</td>
<td>NONE</td>
<td></td>
</tr>
<tr>
<td>US 5923063 A</td>
<td>13-07-1999</td>
<td>NONE</td>
<td></td>
</tr>
<tr>
<td>US 2006046410 A1</td>
<td>02-03-2006</td>
<td>NONE</td>
<td></td>
</tr>
</tbody>
</table>