

VERTICAL MEMORY DEVICES

BACKGROUND

[0001] Semiconductor manufactures developed vertical device technologies, such as three dimensional (3D) NAND flash memory technology, and the like to achieve higher data storage density without requiring smaller memory cells. In some examples, a 3D NAND memory device includes an array region (also referred to as core region) and a staircase region. The array region includes a stack of alternating gate layers and insulating layers. The stack of alternating gate layers and insulating layers is used to form memory cells that are stacked vertically in the array region. The staircase region includes the respective gate layers in the stair-step form to facilitate forming contacts to the respective gate layers. The contacts are used to connect driving circuitry to the respective gate layers for controlling the stacked memory cells.

SUMMARY

[0002] Aspects of the disclosure provide a semiconductor device. The semiconductor device includes gate layers and insulating layers that are stacked alternately along a direction perpendicular to a substrate of the semiconductor device and form a stack upon the substrate. The semiconductor device includes an array of channel structures that are formed in an array region of the stack. Further, the semiconductor device includes a first staircase formed of a first section of the stack in a connection region upon the substrate, and a second staircase formed of a second section of the stack in the connection region upon the substrate. In addition, the semiconductor device includes a dummy staircase formed of the first section of the stack and disposed between the first staircase and the second staircase in the connection region.

[0003] According to some aspects of the disclosure, the semiconductor device includes first contact structures formed on the first staircase and connected to the gate layers in the first section of the stack, and includes second contact structures formed on the second staircase and connected to the gate layers in the second section of the stack.

[0004] In some examples, the first section of the stack and the second section of the stack have a same number of gate layers.

[0005] In some embodiments, the first staircase is formed of first group stair steps that go down in a first direction, and the second staircase is formed of second group stair steps that go down in the first direction.

[0006] In some examples, the dummy staircase is formed of group stair steps that go down in a second direction that is opposite to the first direction.

[0007] In some embodiments, corresponding group stairs steps in the first staircase and the second staircase are of a same height.

[0008] In an example, a sidewall of the dummy staircase to the second staircase has a same height as the first staircase.

[0009] According to an aspect of the disclosure, the semiconductor device also includes a third staircase formed of a third section of the stack that corresponds to gates of top select transistors in the channel structures.

[0010] In some embodiments, each stair step of the first group stair steps and the second group stair steps is formed of multiple division stair steps. The multiple division stair steps go down in a third direction that is perpendicular to the first direction, in an example.

[0011] Aspects of the disclosure provide a method for fabricating a semiconductor device. The method includes stacking sacrificial gate layers and insulating layers alternately along a direction perpendicular to a substrate of the semiconductor device to form a stack upon the substrate, and shaping the sacrificial gate layers and the insulating layers in a first section of the stack in a connection region to form a first staircase, a second staircase and a dummy staircase in the first section of the stack in the connection region upon the substrate. The dummy staircase is disposed between the first staircase and the second staircase. Further, the method includes removing a

number of the sacrificial gate layers and the insulating layers of the second staircase to shift the second staircase into a second section of the stack.

[0012] According to some aspects of the disclosure, the method further includes forming channel structures in the stack in an array region, and replacing the sacrificial gate layers with gate layers. Then, the method includes forming first contact structures on the first staircase and second contact structures on the second staircase. The first contact structures are connected to the gate layers in the first section of the stack and the second contact structures are connected to the gate layers in the second section of the stack.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] Aspects of the present disclosure are best understood from the following detailed description when read with the accompanying figures. It is noted that, in accordance with the standard practice in the industry, various features are not drawn to scale. In fact, the dimensions of the various features may be arbitrarily increased or reduced for clarity of discussion.

[0014] Fig. 1 shows a top-down view of a semiconductor device according to some embodiments of the disclosure.

[0015] Fig. 2 shows a top-down view of a block portion in the semiconductor device according to some embodiments of the disclosure.

[0016] Fig. 3 shows a cross-sectional view of the block portion according to some embodiments of the disclosure.

[0017] Fig. 4 shows a close up view of a portion in the block portion according to some embodiments of the disclosure.

[0018] Fig. 5 shows a cross-sectional view of the portion according to some embodiments of the disclosure.

[0019] Fig. 6 shows a flow chart outlining a process example for fabricating a semiconductor device according to some embodiments of the disclosure.

[0020] Fig. 7 shows an example of a top-down view of the block portion of the semiconductor device during fabrication according to some embodiments of the disclosure.

[0021] Fig. 8 shows an example of a top-down view of the block portion of the semiconductor device during fabrication according to some embodiments of the disclosure.

[0022] Fig. 9 shows an example of a top-down view of the block portion of the semiconductor device during fabrication according to some embodiments of the disclosure.

[0023] Fig. 10 shows an example of a top-down view of the block portion of the semiconductor device during fabrication according to some embodiments of the disclosure.

[0024] Fig. 11 shows an example of a top-down view of the block portion of the semiconductor device during fabrication according to some embodiments of the disclosure.

[0025] Fig. 12 shows an example of a cross-sectional view of the block portion of the semiconductor device during fabrication according to some embodiments of the disclosure.

[0026] Fig. 13 shows an example of a perspective view of the block portion of the semiconductor device during fabrication according to some embodiments of the disclosure.

[0027] Fig. 14 shows an example of a cross-sectional view of the block portion of the semiconductor device during fabrication according to some embodiments of the disclosure.

[0028] Fig. 15 shows an example of a perspective view of the block portion of the semiconductor device during fabrication according to some embodiments of the disclosure.

DETAILED DESCRIPTION

[0029] The following disclosure provides many different embodiments, or examples, for implementing different features of the provided subject matter. Specific examples of components and arrangements are described below to simplify the present disclosure. These are, of course, merely examples and are not intended to be limiting. For example, the formation of a first feature over or on a second feature in the description that follows may include embodiments in which the first and second features are formed in direct contact, and may also include embodiments in which additional features may be formed between the first and second features, such that the first and second features may not be in direct contact. In addition, the present disclosure may repeat reference numerals and/or letters in the various examples. This repetition is for the purpose of simplicity and clarity and does not in itself dictate a relationship between the various embodiments and/or configurations discussed.

[0030] Further, spatially relative terms, such as “beneath,” “below,” “lower,” “above,” “upper” and the like, may be used herein for ease of description to describe one element or feature’s relationship to another element(s) or feature(s) as illustrated in the figures. The spatially relative terms are intended to encompass different orientations of the device in use or operation in addition to the orientation depicted in the figures. The apparatus may be otherwise oriented (rotated 90 degrees or at other orientations) and the spatially relative descriptors used herein may likewise be interpreted accordingly.

[0031] Three-dimensional (3D) semiconductor memory devices can be formed on a substrate that includes an array region (also referred to as core region in some examples) for forming an array of memory cells and a connection region for forming connections to the memory cells. For example, the memory cells are formed in the array region as an array of vertical memory cell strings. The vertical memory cell strings are formed of

gate layers and insulating layers that are stacked alternately. At the connection region, the stack of gate layers and the insulating layers are patterned into stair steps to provide contact pad regions for connecting the gate layers of the memory cells to word lines.

[0032] According to some examples, trim-etch process is used to form stair steps. The trim-etch process repetitively performs a trim step and an etch step based on a mask layer. During the trim step, the mask layer is trimmed to additionally expose a new step region on a stack of alternately stacked (sacrificial) gate layers and insulating layers. During the etch step, the stack is etched based on the mask layer to create a new step. In some examples, the trim-etch process is performed using reactive-ion etching and has a relatively low productivity, such as a relatively low wafer per hour (WPH), and trim-etch process can be high cost process for 3D memory device fabrication. In addition, when the number of stair steps is relatively large, to etch the lower stair steps, the upper stair steps and the lower stair steps have relatively large height difference. Because of the large height difference, thicker mask layer, such as thicker photoresist layer is required for the trim-etch process for the lower stair steps, and the requirement of thicker mask layer can cause, for example, difficulties in lithography process.

[0033] Generally, stair steps are formed of treads and risers. In an example, a tread is the part that is disposed horizontally between a top edge of a lower riser and a bottom edge of an upper riser, and a riser is the part that is disposed vertically between an inner edge of a lower tread and an outer edge of an upper tread. The tread is the part that can be configured into a contact pad for one or more contact structures to land on. The riser is the sidewall of a stack of layers, such as alternately disposed (sacrificial) gate layers and insulating layers. In some examples, a stair step is composed of a tread and a lower riser of the tread. The stair step is measured by depth and width of the tread and height of the lower riser. The depth of the tread is the distance from the outer edge to the inner edge of the tread. The width of the tread is the distance from one side of the tread to the other side. The height of the riser is the vertical distance of the sidewall between the lower tread and the current tread. In the present disclosure, the height of the riser can be measured in term of layer pairs. For example, a layer pair is a thickness

sum of a (sacrificial) gate layer and an insulating layer. In some examples, when a stair step has a height of multiple layer pairs, such as four layer pairs, five layer pairs, six layer pairs, the stair step is referred to as a group stair step; when a stair step has a height of one layer pair, the stair step is referred to as a division stair step.

[0034] According to some aspects of the disclosure, the stack of alternately disposed gate layers and the insulating layers can be divided into sections. Each section of the stack is further divided into groups. Each group is then divided into divisions. Each division includes a layer pair. In some embodiments, the stair steps to the different sections can be formed at the same time (e.g., in the same trim-etch cycles), and then chop processes are used to remove layers and shift stair steps of the different sections to the appropriate section layers. Thus, the total number of trim-etch cycles can be reduced. For example, when two sections are used, the total number of the trim-etch cycles can be reduced by half, and the height difference of the upper stair steps to the lower stair steps in the trim-etch process can be reduced by half for example. In another example, when three sections are used, the total number of trim-etch cycles can be reduced by 2/3, the height difference of the upper stair steps to the lower stair steps in the trim-etch process can be reduced by 2/3. Because the height difference of the upper stair steps to the lower stair steps in the trim-etch process is reduced, the trim-etch process can be performed with ease. Because the total number of trim-etch cycles is reduced, the processing efficiency is improved.

[0035] Fig. 1 shows a top-down view of a semiconductor device 100 according to some embodiments of the disclosure. The semiconductor device 100 includes a memory portion 110 that are formed of three-dimensional (3D) memory cells. The memory portion 110 can include one or more memory planes 120, and each of memory planes 120 can include a plurality of memory blocks 130. In some examples, concurrent operations can take place at the memory planes 120. In some embodiments, each of the memory blocks 130 is the smallest unit to carry out erase operations. In the Fig. 1 example, the memory portion 110 includes four memory planes 120 and each of the memory planes 120 includes six memory blocks 130. Each of the memory blocks 103 can include a plurality of memory cells, and each memory cell can be addressed through

interconnections, such as bit lines and word lines. In some examples, the bit lines and word lines can be laid out perpendicularly, forming an array of metal lines. For example, the word lines extend in the X direction, and the bit lines extend in the Y direction.

[0036] Further, each memory block 130 can be divided into block portions 140 according to stair division patterns. The block portions 140 have identical or equivalent stair division patterns. The details of the block portions 140 will be described with reference to Fig. 2-Fig.5.

[0037] It is noted that the semiconductor device 100 can be any suitable device, for example, memory circuits, a semiconductor chip (or die) with memory circuits formed on the semiconductor chip, a semiconductor wafer with multiple semiconductor dies formed on the semiconductor wafer, a stack of semiconductor chips, a semiconductor package that includes one or more semiconductor chips assembled on a package substrate, and the like.

[0038] It is also noted that, the semiconductor device 100 can include other suitable circuitry (not shown), such as logic circuitry, power circuitry, and the like that is formed on the same substrate, or other suitable substrate, and is suitably coupled with the memory portion 110. Generally, the memory portion 110 includes the memory cells and peripheral circuitry (e.g., address decoder, driving circuits, sense amplifier and the like).

[0039] Fig. 2 shows a top-down view of a block portion 140 according to some embodiments of the disclosure, Fig. 3 shows a cross-sectional view of the block portion 140 at line A-A', Fig. 4 shows a top down view of details of a portion 245 in the block portion 140, and Fig. 5 shows a cross-sectional view of the portion 245 at line B-B'. In some examples, the top-down views in Fig. 2 and Fig. 4 are views in X-Y plane, the cross-sectional view in Fig. 3 is a view in X-Z plane, and the cross-sectional view in Fig. 5 is a view in Y-Z plane.

[0040] In the Fig. 2 and Fig. 3 example, the block portion 140 includes portions 240(A)-240(D) that have identical patterns or mirrored patterns for stair divisions, and the portions 240(A)-240(D) are referred to as stair division pattern (SDP) portions 240(A)-(D). Each SDP portion 240 includes an array region 250 and a connection region 260. The array region 250 includes an array of memory strings 251 (as shown in Fig. 4), and each memory string 251 includes a plurality of stacked memory cells connected in series with one or more top select transistors and one or more bottom select transistors. The connection region 260 includes a top select gate (TSG) connection region 261, a memory cell gate (MCG) connection region 269. The TSG connection region 261 includes a staircase structure and contact structures for connecting metal wires to the gates of the top select transistors to control the top select transistors. The MCG connection region 269 includes staircase structures and contact structures for connecting word lines to the gates of the memory cells.

[0041] It is noted that, the connection region 260 may also include a bottom select gate (BSG) connection region (not shown) that includes a staircase structure and contact structures for connecting metal wires to the gates of the bottom select transistors to control the bottom select transistors.

[0042] According to some aspects of the disclosure, the MCG connection region 269 is configured according to a multi-level staircase architecture, such as a three-level staircase architecture. As shown in Figs. 2-5 example, each memory string 251 includes 108 memory cells, and the three-level stair architecture can be used to provide connections between word lines and the gates of the 108 memory cells in each memory string. For example, the 108 memory cells of a memory string 251 are referred to, in a consecutive order, as M1-M108 with M1 being the first memory cell next to the top select transistors, and M108 being the last one in the sequence. The 108 memory cells are divided into two sections, such as a first section of M1-M54 and a second section of M55-M108. Memory cells in each section are grouped into nine groups, and each group includes six consecutive memory cells.

[0043] It is noted that, in some embodiments, the strings of memory cells in an array are formed in a stack of alternately disposed gate layers and insulating layers. The gate layers form gates of the top select transistors, the memory cells (such as M1-M108 in a string) and the bottom select transistor(s). In some contexts, M1-M108 are used to refer to the gate layers (sometimes sacrificial gate layers) for the corresponding memory cells.

[0044] Specifically, in some embodiments, the three-level staircase architecture includes a section level, a group level and a division level. At the section level, in the Fig. 2-5 example, the three-level staircase architecture includes a first staircase section 270 for providing connections to the first section (S1) of M1-M54, and includes a second staircase section 290 for providing connections to the second section (S2) of M55-M108. At the group level, each staircase section includes nine group stair steps G1-G9, and each group stair step has a height of six layer pairs. At the division level, each group stair step includes 6 division stair steps D1-D6, and each division stair step has a height of one layer pair. In some examples, the group stair steps G1-G9 in each staircase section go up/down in a first direction, such as in X direction (or -X direction). Further, in some examples, division stair steps D1-D6 that go up/down in a second direction, such as Y direction (or -Y direction) that is perpendicular to the first direction.

[0045] Additionally, in the Figs. 2-5 example, the MCG connection region 269 includes a dummy staircase section 280 that is disposed between the first staircase section 270 and the second staircase section 290.

[0046] In some embodiments, the first staircase section 270, the second staircase section 290 and the dummy staircase 280 are formed by the same trim-etch process, thus the first and second staircase sections 270 and 290 and the dummy staircase 280 are of similar group stair steps. For example, the section staircases 270 and 290 and the dummy staircase 280 have the same number of group stair steps, and corresponding group stair steps are of the same group stair step height and the same group stair step depth. The first and second staircase sections 270 and 290 have the same step-down

direction, and the step-down direction of the dummy staircase 280 is the opposite direction of the step-down direction of the first and second staircase sections 270 and 290.

[0047] According to some aspects of the disclosure, a chop process is used to shift the second staircase section 290 down (e.g., -Z direction) to appropriate layers. In the Figs. 2-5 example, the first staircase section 270 and the dummy staircase section 280 are disposed with stair steps formed in the layers for the memory cells M1-M54, and the second staircase section 290 is disposed with stair steps formed in the layers for the memory cells M55-M108.

[0048] It is noted that, while in the Figs. 2-5 example, the three-level staircase architecture includes two staircase sections, the number of staircase sections in the three-level staircase architecture should not be limited and can be any suitable number, such as three, four, five and the like. It is noted that, while in the Figs. 2-5 example, each staircase section includes 9 group stair steps, the number of group stair steps in a staircase section should not be limited and can be any suitable number, such as six, seven, eight, ten and the like. It is noted that, while in the Figs. 2-5 example, each group stair step includes 9 division stair steps, the number of division stair steps in a group stair step should not be limited and can be any suitable number, such as two, three, four, five, seven and the like.

[0049] In some embodiments, the gate-last fabrication technology is used, thus slit structures are formed to assist the removal of sacrificial gate layers, and the formation of the real gates. In the Figs. 2-5 example, slit structures 211, 212(A), 212(B), 213(A), 213(B) and 214 are formed in the SDP portion 240(C) as shown in Fig. 4. The slit structures 211, 212(A), 212(B), 213(A), 213(B) and 214 extend in the X direction, and parallel to each other. The slit structures 211 and 214 separate the SDP portion 240(C) from neighboring SDP portions 240(B) and 240(D) in an example. The slit structures 212(A) and 213(A) are disposed in the array region 250 and can divide the array of memory cell strings in the SDP portion 240(C) into three finger structures 241, 242 and

243. The slit structures 212(B) and 213(B) are disposed in the connection region 260 and can divide the connection region 260 into multiple portions.

[0050] In an example, the slit structures 211 and 214 are continuous slit structures that are filled with insulating layers to electrically insulate the gate layers of the SDP portion 240(C) from neighboring SDP portions 240(B) and 240(D) for example.

[0051] In some examples, the number of the slit structures in connection region 260 is same as the number of slit structures in the array region 250. In the Figs. 2-5 example, the slit structures 212(B) and 213(B) are aligned with the slit structures 212(A) and 213(A). However, the slit structures 212(B) and 213(B) are broken from the slit structures 212(A) and 213(A) and are not continuous parts of the slit structures 212(A) and 213(A), thus the gate layers in the three fingers 241-243 are connected.

[0052] It is noted, in another example, the slit structures 212(B) and 213(B) are not aligned with the slit structures 212(A) and 213(A). In another example, the number of slit structures in the connection region 260 is not the same as the number of slit structures in the array region 250.

[0053] In some embodiments, at least some slit structures can function as the common source contact for an array of memory strings 251 in the array regions 250.

[0054] In the Figs. 2-5 example, and as shown in Fig. 4, top select gate cuts 215 can be disposed in the middle of each finger portion to divide a top select gate (TSG) layer of the memory finger into two portions, and thereby can divide a memory finger portion into two separately programmable (read/write) pages. While erase operation of a 3D NAND memory can be carried out at memory block level, read and write operations can be carried out at memory page level. In some embodiments, dummy channel structures 222 can be disposed at suitable places for process variation control during fabrication and/or for additional mechanical support.

[0055] It is noted that, in some examples, the top select gate cuts 215 do not cut the memory cell gate layers and the bottom select gate layers.

[0056] In the TSG connection region 261, a stair structure is formed. The stair structure has multiple stair steps to expose a portion of gate layers of the top select transistors, and the exposed portions can be configured as contact pads. Then, contact structures can be formed on the contact pads for connecting metal wires to the gates of the top select transistors to control the top select transistors. In the Figs. 2-5 example, and shown in Fig. 4, the stair structure at the TSG connection region 261 has two stair steps 262 and 263. In an example, each of the two stair steps 262 and 263 has a height of one layer pair. In the Fig. 2-5 example, the dashed lines show edges of treads. In an example, a memory string includes a first gate select transistor and a second gate select transistor. The gate of the first gate select transistor is connected with a contact structure 264 on the first stair step 262, and the gate of the second gate select transistor is connected with a contact structure 265 on the second stair step 263.

[0057] Details of the first staircase section 270 are shown in Fig. 4. The first staircase section 270 exposes a portion of gate layers to the memory cells M1-M54 in each memory string 251 as contact pads, and contact structures can be formed on the contact pads to connect the gate layers of the memory cells M1-M54 in each memory string 251 to word lines.

[0058] For example, the tread of division stair step D6 in the region of group stair step G9 provides contact pad for M1. The tread of division stair step D5 in the region of group stair step G9 provides contact pad for M2. The tread of division stair step D4 in the region of group stair step G9 provides contact pad for M3. The tread of division stair step D3 in the region of group stair step G9 provides contact pad for M4. The tread of division stair step D2 in the region of group stair step G9 provides contact pad for M5. The tread of division stair step D1 in the region of group stair step G9 provides contact pad for M6.

[0059] Similarly, the tread of division stair step D6 in the region of group stair step G8 provides contact pad for M7. The tread of division stair step D5 in the region of group stair step G8 provides contact pad for M8. The tread of division stair step D4 in the region of group stair step G8 provides contact pad for M9. The tread of division

stair step D3 in the region of group stair step G8 provides contact pad for M10. The tread of division stair step D2 in the region of group stair step G8 provides contact pad for M11. The tread of division stair step D1 in the region of group stair step G8 provides contact pad for M12.

[0060] Similarly, the tread of division stair step D6 in the region of group stair step G7 provides contact pad for M13. The tread of division stair step D5 in the region of group stair step G7 provides contact pad for M14. The tread of division stair step D4 in the region of group stair step G7 provides contact pad for M15. The tread of division stair step D3 in the region of group stair step G7 provides contact pad for M16. The tread of division stair step D2 in the region of group stair step G7 provides contact pad for M17. The tread of division stair step D1 in the region of group stair step G7 provides contact pad for M18.

[0061] Similarly, the tread of division stair step D6 in the region of group stair step G6 provides contact pad for M19. The tread of division stair step D5 in the region of group stair step G6 provides contact pad for M20. The tread of division stair step D4 in the region of group stair step G6 provides contact pad for M21. The tread of division stair step D3 in the region of group stair step G6 provides contact pad for M22. The tread of division stair step D2 in the region of group stair step G6 provides contact pad for M23. The tread of division stair step D1 in the region of group stair step G6 provides contact pad for M24.

[0062] Similarly, the tread of division stair step D6 in the region of group stair step G5 provides contact pad for M25. The tread of division stair step D5 in the region of group stair step G5 provides contact pad for M26. The tread of division stair step D4 in the region of group stair step G5 provides contact pad for M27. The tread of division stair step D3 in the region of group stair step G5 provides contact pad for M28. The tread of division stair step D2 in the region of group stair step G5 provides contact pad for M29. The tread of division stair step D1 in the region of group stair step G5 provides contact pad for M30.

[0063] Similarly, the tread of division stair step D6 in the region of group stair step G4 provides contact pad for M31. The tread of division stair step D5 in the region of group stair step G4 provides contact pad for M32. The tread of division stair step D4 in the region of group stair step G4 provides contact pad for M33. The tread of division stair step D3 in the region of group stair step G4 provides contact pad for M34. The tread of division stair step D2 in the region of group stair step G4 provides contact pad for M35. The tread of division stair step D1 in the region of group stair step G4 provides contact pad for M36.

[0064] Similarly, the tread of division stair step D6 in the region of group stair step G3 provides contact pad for M37. The tread of division stair step D5 in the region of group stair step G3 provides contact pad for M38. The tread of division stair step D4 in the region of group stair step G3 provides contact pad for M39. The tread of division stair step D3 in the region of group stair step G3 provides contact pad for M40. The tread of division stair step D2 in the region of group stair step G3 provides contact pad for M41. The tread of division stair step D1 in the region of group stair step G3 provides contact pad for M42.

[0065] Similarly, the tread of division stair step D6 in the region of group stair step G2 provides contact pad for M43. The tread of division stair step D5 in the region of group stair step G2 provides contact pad for M44. The tread of division stair step D4 in the region of group stair step G2 provides contact pad for M45. The tread of division stair step D3 in the region of group stair step G2 provides contact pad for M46. The tread of division stair step D2 in the region of group stair step G2 provides contact pad for M47. The tread of division stair step D1 in the region of group stair step G2 provides contact pad for M48.

[0066] Similarly, the tread of division stair step D6 in the region of group stair step G1 provides contact pad for M49, and a contact structure C1 (shown in Fig. 5) is formed on the contact pad. The tread of division stair step D5 in the region of group stair step G1 provides contact pad for M50, and a contact structure C2 (shown in Fig. 5) is formed on the contact pad. The tread of division stair step D4 in the region of group

stair step G1 provides contact pad for M51, and a contact structure C3 (shown in Fig. 5) is formed on the contact pad. The tread of division stair step D3 in the region of group stair step G1 provides contact pad for M52, and a contact structure C4 (shown in Fig. 5) is formed on the contact pad. The tread of division stair step D2 in the region of group stair step G1 provides contact pad for M53, and a contact structure C5 (shown in Fig. 5) is formed on the contact pad. The tread of division stair step D1 in the region of group stair step G1 provides contact pad for M54, and a contact structure C6 (shown in Fig. 5) is formed on the contact pad.

[0067] It is noted that, in some examples, slit structures, such as shown by the slit structures 211, 212(B), 213(B) and 214 in Fig. 5 are filled with an insulating layer 530, and a conductive material 540. The insulating layer 530 insulates the conductive material 540 from the gate layers. The conductive material 540 can be used for forming common source contact.

[0068] Fig. 6 shows a flow chart outlining a process example 600 for fabricating a semiconductor device, such as the semiconductor device 100 according to some embodiments of the disclosure. The process starts at S601 and proceeds to S610.

[0069] At S610, sacrificial gate layers and insulating layers are stacked alternately on a substrate to form an initial stack. The substrate can be any suitable substrate, such as a silicon (Si) substrate, a germanium (Ge) substrate, a silicon-germanium (SiGe) substrate, and/or a silicon-on-insulator (SOI) substrate. The substrate may include a semiconductor material, for example, a Group IV semiconductor, a Group III-V compound semiconductor, or a Group II-VI oxide semiconductor. The Group IV semiconductor may include Si, Ge, or SiGe. The substrate may be a bulk wafer or an epitaxial layer. In some examples, the insulating layers are made of insulating material(s), such as silicon dioxide, and the like, and the sacrificial layers are made of silicon nitride.

[0070] At S620, stair steps to gates of the top select transistors are formed. The stair steps to the gates of the top select transistors can be formed by any suitable process. In an example, the stair steps to the gates of the top select transistors can be formed by

applying a repetitive etch-trim process using a mask layer. The details of a repetitive etch-trim process will be described with reference to S630.

[0071] Fig. 7 shows an example of a top-down view of the block portion 140 of the semiconductor device 100 after the formation of the stair steps to the gates of the top select transistors. As shown in Fig. 7, stair steps are formed in the TSG connection region 261.

[0072] Referring back to Fig. 6, at S630, division stair steps of stair division patterns are formed in a connection region. In some examples, a mask layer is used and trimming process is applied on the mask layer to form the etch masks for forming the division stair steps.

[0073] Fig. 8 shows an example of top-down view of the block portion 140 of the semiconductor device 100 with SDP portions 240 (A)-(D) that are covered by a mask layer 810. The mask layer 810 is used to form the division stair steps in the SDP portions 240 (A)-(D). The SDP portions 240 (A)-(D) have identical SDP or mirrored SDP. The mask layer 810 covers the array region 250 and a portion of the connection region 260 adjacent to the array region 250 and the TSG connection region 261. In some embodiments, the mask layer 810 can include a photoresist or carbon-based polymer material, and can be formed using a patterning process such as lithography. In some embodiments, the mask layer 810 can also include a hard mask, such as silicon oxide, silicon nitride, TEOS, silicon-containing anti-reflective coating (SiARC), amorphous silicon, or polycrystalline silicon. The hard mask can be patterned using etching process such as reactive-ion-etching (RIE) using O₂ or CF₄ chemistry. Furthermore, the mask layer 810 can include any combination of photoresist and hard mask.

[0074] In some embodiments, the division stair steps can be formed by applying a repetitive etch-trim process using the mask layer 810. The repetitive etch-trim process includes multiple cycles of an etching process and a trimming process. During the etching process, a portion of the initial stack with exposed surface can be removed. In an example, the etch depth equals to a layer pair that is the thickness of a sacrificial

gate layer and an insulating layer. In an example, the etching process for the insulating layer can have a high selectivity over the sacrificial layer, and/or vice versa.

[0075] In some embodiments, the etching of the stack is performed by an anisotropic etching such as a reactive ion etch (RIE) or other dry etch processes. In some embodiments, the insulating layer is silicon oxide. In this example, the etching of silicon oxide can include RIE using fluorine based gases such as carbon-fluorine (CF₄), hexafluoroethane (C₂F₆), CHF₃, or C₃F₆ and/or any other suitable gases. In some embodiments, the silicon oxide layer can be removed by wet chemistry, such as hydrofluoric acid or a mixture of hydrofluoric acid and ethylene glycol. In some embodiments, a timed-etch approach can be used. In some embodiments, the sacrificial layer is silicon nitride. In this example, the etching of silicon nitride can include RIE using O₂, N₂, CF₄, NF₃, Cl₂, HBr, BCl₃, and/or combinations thereof. The methods and etchants to remove a single layer stack should not be limited by the embodiments of the present disclosure.

[0076] The trimming process includes applying a suitable etching process (e.g., an isotropic dry etch or a wet etch) on the mask layer 810 such that the mask layer 810 can be pulled back (e.g., shrink inwardly) laterally in the x-y plane from edges. In some embodiments, the trimming process can include dry etching, such as RIE using O₂, Ar, N₂, etc. In some embodiments, a pull-back distance of the mask layer 810 corresponds to the depth of a division stair step.

[0077] After trimming the mask layer 810, one portion of the topmost level of the initial stack corresponding to a division is exposed and the other portion of the topmost level of the initial stack remains covered by the mask layer 810. The next cycle of etch-trim process resumes with the etching process.

[0078] In some embodiments, the topmost level of the initial stack can be covered by an insulating layer. In some embodiments, the topmost level of the initial stack can further be covered by other dielectric materials. A process step of removing the insulating layer and/or the other dielectric materials can be added to the etching process of each etch-trim cycle to form the division stair steps.

[0079] After forming the division stair steps, the mask layer 810 can be removed. The mask layer 810 can be removed by using techniques such as dry etching with O₂ or CF₄ plasma, or wet etching with resist/polymer stripper, for example solvent based chemicals.

[0080] Fig. 9 shows an example of top-down view of the block portion 140 in the semiconductor device 100 after the mask layer 810 is removed. As shown in Fig. 9, division stair steps D1-D6 are formed.

[0081] Referring back to Fig. 6, at S640, group stair steps for the multiple staircase sections, such as the first staircase section 270, the second staircase section 290 and the like, in the connection region are formed in the upper section layers, such as the layers for M1-M54. In some examples, a mask layer is used and trimming process is applied on the mask layer to form the etch masks for forming the group stair steps.

[0082] Fig. 10 shows an example of top-down view of the block portion 140 of the semiconductor device 100 that is covered by a mask layer 1010 that is used to form the group stair steps. The mask layer 1010 is disposed over the array region 250 and a portion of the connection region 260. As shown in Fig. 10, the mask layer 1010 has a first portion 1010(A) and a second portion 1010(B). The first portion 1010(A) covers the array region 250 and a portion of the first staircase section 270, the second portion 1010(B) covers a portion of the second staircase section 290 and a portion of the dummy staircase section 280. The mask layer 1010 can be made of a similar material as the mask layer 810 and can be formed using a similar technique.

[0083] In some embodiments, the group stair steps can be formed by applying repetitive etch-trim process using the mask layer 1010, similar to the repetitive etch-trim process to form the division stair steps. In this example, the group stair steps of the first staircase section 270 can be formed by trimming the left edge of the first portion 1010(A) in X direction. The group stair steps of the second staircase section 290 can be formed by trimming the left edge of the second portion 1010(B) in X direction. The group stair steps of the dummy staircase section 280 can be formed by trimming the right edge of the second portion 1010(B) in -X direction.

[0084] In some embodiments, each group stair step includes multiple layer pairs, such as 9 layer pairs in an example. Then, an etching process etches suitable layers corresponding to the height of a group stair step, such as nine layer pairs of alternating sacrificial layers and insulating layers.

[0085] After forming the group stair steps, the mask layer 1010 can be removed. The mask layer 1010 can be removed by using techniques such as dry etching with O₂ or CF₄ plasma, or wet etching with resist/polymer stripper, for example solvent based chemicals.

[0086] Fig. 11 shows an example of top-down view of the block portion 140 in the semiconductor device 100 after the mask layer 1010 is removed. The dashed lines show the edges of treads for the group stair steps. As shown in Fig. 11, group stair steps G1-G9 are formed.

[0087] Fig. 12 shows an example of a cross-sectional view of the block portion 140 at line A-A' after the mask layer 1010 is removed. Fig. 13 shows a perspective view of the block portion 140 after the mask layer 1010 is removed. As shown in Fig. 12 and Fig. 13, group stair steps G1-G9 of the first staircase section 270 and the second staircase section 290 are formed in the layers for M1-M54.

[0088] Referring back to Fig. 6, at S650, a chop process is performed at different staircase sections to shift the staircase sections to the appropriate section layers. In an example, the second staircase section 290 is suitably exposed, and a chop process is performed to shift the second staircase section 290 to the layers for M55-M108. For example, a mask layer is disposed to cover the semiconductor device 100, and then the portion of the mask layer that covers the second staircase section 290 is suitably removed to expose the second staircase section 290. Then, etch process is performed to remove 54 layer pairs at the second staircase section 290.

[0089] In some embodiments, the etching of a lay pair (including an insulating layer and a sacrificial gate layer) at the second staircase section 290 is performed by an anisotropic etching such as a reactive ion etch (RIE) or other dry etch processes. In

some embodiments, the insulating layer is silicon oxide. In this example, the etching of silicon oxide can include RIE using fluorine based gases such as carbon-fluorine (CF₄), hexafluoroethane (C₂F₆), CHF₃, or C₃F₆ and/or any other suitable gases. In some embodiments, the silicon oxide layer can be removed by wet chemistry, such as hydrofluoric acid or a mixture of hydrofluoric acid and ethylene glycol. In some embodiments, a timed-etch approach can be used. In some embodiments, the sacrificial gate layer is silicon nitride. In this example, the etching of silicon nitride can include RIE using O₂, N₂, CF₄, NF₃, Cl₂, HBr, BCl₃, and/or combinations thereof. The methods and etchants to remove a single layer stack should not be limited by the embodiments of the present disclosure.

[0090] Fig. 14 shows an example of a cross-sectional view of the block portion 140 at line A-A' after the chop process and the mask layer is removed. Fig. 15 shows a perspective view of the block portion 140 after the chop process and the mask layer is removed. As shown in Fig. 14 and Fig. 15, group stair steps G1-G9 of the second staircase section 290 are shifted in the layers for M55-M108.

[0091] It is noted that, when more than two sections are used, the chop process can be repetitively used on other sections.

[0092] Referring back to Fig. 6, at S660, channel structures are formed. In an example, suitably planarization process is performed to obtain a relatively flat surface. Then, photo lithography technology is used to define patterns of channel holes and dummy channel holes in photoresist and/or hard mask layers, and etch technology is used to transfer the patterns into the stack of sacrificial layers and insulating layers. Thus, channel holes are formed in the array region 250 and the dummy channel holes are formed in the connection region.

[0093] Then, channel structures are formed in the channel holes, and dummy channel structures are formed in the dummy channel holes. In some embodiments, dummy channel structures can be formed with the channel structures, thus the dummy channel structures are formed of the same materials as the channel structures. In some

embodiments, the dummy channel structures are formed differently from the channel structures.

[0094] At S670, gate line slits (also referred to as slit structures in some examples) are formed. In some embodiments, the gate line slits are etched as trenches in the stack. In some examples, the gate line slits in the connection region have the same pitch as the gate line slits in the array region.

[0095] At S680, real gates are formed. In some embodiments, using the gate line slits, the sacrificial layers can be replaced by the gate layers. In an example, etchants to the sacrificial layers are applied via the gate line slits to remove the sacrificial layers. In an example, the sacrificial layers are made of silicon nitride, and the hot sulfuric acid (H_2SO_4) is applied via the gate line slits to remove the sacrificial layers. Further, via the gate line slits, gate stacks to the transistors in the array region are formed. In an example, a gate stack is formed of a high-k dielectric layer, a glue layer and a metal layer. The high-k dielectric layer can include any suitable material that provide the relatively large dielectric constant, such as hafnium oxide (HfO_2), hafnium silicon oxide (HfSiO_4), hafnium silicon oxynitride (HfSiON), aluminum oxide (Al_2O_3), lanthanum oxide (La_2O_3), tantalum oxide (Ta_2O_5), yttrium oxide (Y_2O_3), zirconium oxide (ZrO_2), strontium titanate oxide (SrTiO_3), zirconium silicon oxide (ZrSiO_4), hafnium zirconium oxide (HfZrO_4), and the like. The glue layer can include refractory metals, such as titanium (Ti), tantalum (Ta) and their nitrides, such as TiN, TaN, W₂N, TiSiN, TaSiN, and the like. The metal layer includes a metal having high conductivity, such as tungsten (W), copper (Cu) and the like.

[0096] At S690, further process(es) can be performed on the semiconductor device. For example, the gate-last process continues to, for example, fill the gate line slits with spacer material (e.g., silicon oxide) and common source material (e.g., tungsten) to form the slit structure. Further, contacts structures can be formed and metal traces can be formed.

[0097] The foregoing outlines features of several embodiments so that those skilled in the art may better understand the aspects of the present disclosure. Those skilled in the

art should appreciate that they may readily use the present disclosure as a basis for designing or modifying other processes and structures for carrying out the same purposes and/or achieving the same advantages of the embodiments introduced herein. Those skilled in the art should also realize that such equivalent constructions do not depart from the spirit and scope of the present disclosure, and that they may make various changes, substitutions, and alterations herein without departing from the spirit and scope of the present disclosure.

WHAT IS CLAIMED IS:

1. A semiconductor device, comprising:
 - gate layers and insulating layers that are stacked alternately along a direction perpendicular to a substrate of the semiconductor device and form a stack upon the substrate;
 - an array of channel structures being formed in an array region of the stack;
 - a first staircase formed of a first section of the stack in a connection region upon the substrate;
 - a second staircase formed of a second section of the stack in the connection region upon the substrate; and
 - a dummy staircase formed of the first section of the stack and disposed between the first staircase and the second staircase in the connection region.
2. The semiconductor device of claim 1, further comprising:
 - first contact structures formed on the first staircase and connected to the gate layers in the first section of the stack; and
 - second contact structures formed on the second staircase and connected to the gate layers in the second section of the stack.
3. The semiconductor device of claim 1, wherein the first section of the stack and the second section of the stack have a same number of gate layers.
4. The semiconductor device of claim 1, wherein the first staircase is formed of first group stair steps that go down in a first direction, and the second staircase is formed of second group stair steps that go down in the first direction.
5. The semiconductor device of claim 4, wherein the dummy staircase is formed of group stair steps that go down in a second direction that is opposite to the first direction.
6. The semiconductor device of claim 4, wherein corresponding group stairs steps in the first staircase and the second staircase are of a same height.
7. The semiconductor device of claim 4, wherein a sidewall of the dummy staircase to the second staircase has a same height as the first staircase.
8. The semiconductor device of claim 1, further comprising:

a third staircase formed of a third section of the stack that corresponds to gate top select transistors in the channel structures.

9. The semiconductor device of claim 4, wherein each stair step of the first group stair steps and the second group stair steps is formed of multiple division stair steps.

10. The semiconductor device of claim 9, wherein the multiple division stair steps go down in a third direction that is perpendicular to the first direction.

11. A method for fabricating a semiconductor device, comprising:
stacking sacrificial gate layers and insulating layers alternately along a direction perpendicular to a substrate of the semiconductor device to form a stack upon the substrate;
shaping the sacrificial gate layers and the insulating layers in a first section of the stack in a connection region to form a first staircase, a second staircase and a dummy staircase in the first section of the stack in a connection region upon the substrate, the dummy staircase being disposed between the first staircase and the second staircase; and
removing a number of the sacrificial gate layers and the insulating layers of the second staircase to shift the second staircase into a second section of the stack.

12. The method of claim 11, further comprising:
forming channel structures in the stack in an array region;
replacing the sacrificial gate layers with gate layers; and
forming first contact structures on the first staircase and second contact structures on the second staircase, the first contact structures being connected to the gate layers in the first section of the stack and the second contact structures being connected to the gate layers in the second section of the stack.

13. The method of claim 12, wherein the first section of the stack and the second section of the stack have a same number of gate layers.

14. The method of claim 11, wherein shaping the sacrificial gate layers and the insulating layers in the connection region to form the first staircase, the second staircase and the dummy staircase further comprises:

shaping the sacrificial gate layers and the insulating layers in the connection region into the first staircase with first group stair steps that go down in a first direction, and into the second staircase with second group stair steps that go down in the first direction.

15. The method of claim 14, further comprising:
shaping the sacrificial gate layers and the insulating layers in the connection region into the dummy staircase with dummy group stair steps that go down in a second direction that is opposite to the first direction.
16. The method of claim 14, wherein corresponding group stair steps in the first staircase and the second staircase are of a same height.
17. The method of claim 14, wherein removing the number of the sacrificial gate layers and the insulating layers of the second staircase to shift the second staircase into the second section of the stack comprise:
removing the number of the sacrificial gate layers and the insulating layers that are of a same height as the first staircase.
18. The method of claim 12, further comprising:
shaping the sacrificial gate layers and the insulating layers to form stair steps in a third section of the stack corresponding to top select transistors of the channel structures.
19. The method of claim 14, wherein each stair step of the first group stair steps and the second group stair steps is formed of multiple division stair steps.
20. The method of claim 19, wherein the multiple division stair steps go down in a third direction that is perpendicular to the first direction.