Highly scalable vertical gate 3-D NAND


Editor’s note: This work was first presented at the 2012 IEEE International Electron Devices Meeting (IEDM) and appears here courtesy of the IEEE. For more information about IEDM 2013 (Washington DC; December 9-11), click here.

Abstract
We demonstrate an 8-layer 3D Vertical Gate NAND Flash with write line (WL) half pitch =37.5nm, bit line (BL) half pitch=75nm, 64-WL NAND string with 63% array core efficiency. This is the first time that a 3D NAND Flash can be successfully scaled to below 3Xnm half pitch in one lateral dimension, thus an 8-layer stack device already provides a very cost effective technology with lower cost than the conventional sub-20nm 2D NAND. Our new VG architecture has two key features: (1) To improve the manufacturability a new layout that twists the even/odd BL’s (and pages) in the opposite direction (split-page BL) is adopted. This allows the island-gate SSL devices [1] and metal interconnections be laid out in double pitch, creating much larger process window for BL pitch scaling; (2) A novel staircase BL contact formation method using binary sum of only M lithography and etching steps to achieve $2^M$ contacts. This not only allows precise landing of the tight-pitch staircase contacts, but also minimizes the process steps and cost. We have successfully fabricated an 8-layer array using TFT BE-SONOS charge-trapping device. The array characteristics including reading, programming, inhibit, and block erase are demonstrated.

Introduction
3D stackable NAND Flash is forecasted to continue NAND Flash scaling below 15nm node [2]. In general, 3D NAND Flash uses minimal processing steps to pattern the multi-layer stacks only once, thus greatly reduces the bit cost. However, most 3D NAND architectures have relatively larger lateral half pitch (>60nm [3]) due to many limitations. The large cell pitch must be compensated by very large stack number (>32) in order to compete with sub-20nm 2D NAND. This greatly reduces the bit cost advantage and threatens the future prospect.

3D vertical gate (VG) architecture [1,4-6] was considered as the most pitch scalable 3D NAND architecture. Like conventional NAND, it uses WL/BL patterning thus the lateral pitch can be scaled in a similar way as conventional 2D NAND. However, the decoding method for 3DVG NAND is more difficult than the vertical channel (VC) NAND [3] because the BL’s are horizontal, parallel to the multi-layers and thus cannot be simply connected to metal BL’s as in the conventional 2D NAND.

The first proposed VG architecture uses plural rows of normally-on SSL devices [4] to decode the BL’s within the NAND string. As the stacked layer increases, the required row number of SSL’s increases accordingly, greatly reduces the array efficiency. Later, an island-gate SSL device [1,5] was proposed to separate channel BL for decoding, as shown in Fig. 1(a). This avoids the issue in [4] and the array efficiency can be kept constant as stack layer increases. However, fabrication of the island-gate SSL within each channel BL is difficult when BL pitch is scaled.
Figure 1: (a) Previously proposed 3D Vertical Gate (3DVG) NAND architecture [1]. Island-gate SSL is used to separate channel BL’s, while staircase BL contacts are used to decode the memory array. The island-gate SSL has the same pitch as the channel BL, thus makes BL pitch scaling difficult. (b) 3D overview of the proposed Twisted BL (Split-page) VG architecture in this work. We split the even/odd island gate SSL in the opposite direction, allowing island-gate SSL devices be laid out in double pitch, providing much larger process window for BL pitch scaling.

Twisted BL (split-page) 3-D VG architecture

In this work, we propose a new architecture to provide much better scalability. Figure 1(b) shows the 3D overview of the structure. In Fig. 2(a), we design VG in a twisted layout, where even/odd BL’s are arranged in opposite directions. This allows the island-gate SSL be laid out at double pitch of the BL, thus offering much larger process window. The NAND strings are therefore divided into even/odd pages accordingly, where string current flows in the opposite direction. The SEM top view of our fabricated array is shown in Figure 2(b). WL half pitch=37.5nm, BL half pitch=75nm. Each island-gate SSL controls one channel BL, while a BL pad groups a total of 16 pages (8 for even, and 8 for odd in the opposite sides). In each BL pad, there are a total of 8 (=N, where N is the stacking number) BL contacts, corresponding to the 8 (=N) different memory layers. Staircase BL contacts are fabricated for the array decoding. Note that the BL contacts are also at double pitch of the channel BL, thus allowing larger process window.

The staircase BL contacts are connected to a ML3 BL toward page buffer for sensing. Inside the array, ML1 and ML2 are used to decode the 16 (=2*N) SSL devices. The source contact is made at the line end of each channel BL, and is directly connected to a local ML1 common source line (CSL). Each memory cell is accessed by selecting the corresponding WL, ML3 BL (corresponding to the memory layer), SSL (corresponding to one channel BL). Page programming and reading are performed by simultaneously operating the many SSL’s in parallel in different units for a larger bandwidth. The page size is equal to the total ML3 BL number.

Figure 2: (a) Schematic diagram of the Twisted BL (Split-page) VG architecture. The island-gate SSL devices are split into even/odd pages in the opposite direction, giving a Twisted BL layout. GSL’s also have two pairs (even and odd) accordingly for correct string selection. Each island-gate SSL corresponds to one page during page program or page read.
operation. In BL direction, each BL pad (called a unit) groups a total of 2N (N: stack layer) channel BL’s, where each side has N (even or odd) SSL devices or N pages. Inside the array, ML1 and ML2 are used to connect the SSL gates to the decoder. Page operation is defined by selecting one SSL (“Page 0” to “Page 15” for 8-layer device) in one unit but parallely selecting all units together. The staircase BL contacts are carried out in the BL pad to connect to different memory layers. The staircase BL contacts have double pitch of channel BL, thus allowing more process window. (b) The SEM top view of our fabricated device. WL half pitch = 37.5nm (using self-aligned double patterning, SADP), BL half pitch = 75nm. The island-gate SSL and BL pad are clearly illustrated. The array has 63% core efficiency (= memory cell area / total area, not including decoder and peripheral circuits). (c) The 3D bird’s eye view on our fully-integrated VG array.

Figure 2(c) shows the SEM bird’s eye view of our fully integrated device structure. In our process, we use the poly plug to realize the staircase BL contact and source contact. Figure 3(a) shows the BL cross-sectional view. The 8-layer device is fabricated with excellent BL profile. The inset shows a zoom-in view, where each device is a double-gate TFT BE-SONOS charge-trapping device. Figure 3(b) shows the fabricated island-gate SSL for the 8-layer NAND. The island-shape layout of SSL is designed during the self-aligned double patterning process of the WL’s without additional mask. Figure 3(c) shows the WL profile. A high aspect ratio (>25) has been achieved.

Figure 3: (a) BL cross-sectional view of the 8-layer 3DVG device. Each poly and oxide thickness is 30nm. The inset shows the zoom-in view. Each device is a double-gate TFT BE-SONOS charge-trapping memory device. (b) The cross-sectional view of the island-gate SSL device. The island-gate SSL has double-pitch of channel BL. It is fabricated together with SADP processing of WL’s without additional mask. (c) The SEM cross-sectional view of the 37.5nm half-pitch WL’s. A very high aspect ratio (>25) has been achieved with excellent profile. The WL’s has 60nm-thick WSix to reduce the WL RC delay.

Advantages of 3-D VG NAND
There are several advantages of this split-page 3D VG NAND:

(1) **Pitch scalability**: In this work 3Xnm WL, the smallest in 3D NAND so far, is already demonstrated. Further scaling to 2D NAND limit is likely.

(2) **Double-pitch of island-gate SSL, staircase BL contacts, and metal interconnect**: These
allow larger process window for BL pitch scalability.

(3) **Low WL resistance:** It is easy to fabricate silicide on top of the WL for WL RC delay reduction. In this work, we use conventional WSix.

(4) **Low CSL resistance:** Each source contact is directly connected to the local ML1 CSL. A low resistance CSL is critically important for NAND Flash design because the page read demands large current at CSL.

(5) **Constant array efficiency at higher stacking layers:** As stack number (N) increases, we can simply adjust the BL pad size (shared by 2N pages) but there is no need to expand the array area. Outside the array, only SSL decoder number increases, while WL and BL decoders are kept the same.

**MiLC: Minimal Incremental Layer Cost for Staircase Contacts**

Staircase contact is a fundamental element for all 3D memories. In a previous proposal [3], multiple trim and etch of photo resist was proposed. However, the PR trim/etch process is not accurate thus not suitable for the tight-pitch staircase contacts. Without new innovation, to make N staircase contacts may require N lithography and etching steps, which would greatly increase the cost. Overlay between lithography steps is also a concern.

We propose a novel method to greatly simplify the staircase BL contacts, as explained in Fig. 4(a). In this work we only use three masks (LA1,2,3, corresponding to 1, 2, 4 poly/oxide etch, respectively) to define a total of 8-layer stack. Sums of binary nodes of 0, 20, 21,..., $2^{M-1}$ can generate any integer 0, 1, 2...$2^{M-1}$. To precisely define the contact position we first use an additional mask to pattern the hard mask and define the contact location precisely. The subsequent LA1, 2, 3 steps carry out the etching of 1, 2, 4 P/O respectively. A layout example is shown in Fig. 4(b). Figure 4(c) shows the fabricated staircase contacts using this novel process. Figures 4(d) and (e) shows the final BL poly plug contacts by this novel “MiLC” process. It shows excellent landing on each memory layer. SiN spacer is used to isolate poly plug and BL pad. Since MiLC allows $2^M$ contacts using only M masks each doubling of layer number only requires one more mask.

![MiLC: Minimal Incremental Layer Cost for Staircase Contacts](image)

**Figure 4:** (a) The MiLC concept. It uses binary sum of a few etching steps, each
corresponds to 2n in depth, to carry out multiple staircase contacts with minimized incremental layer cost. (b) The layout example. Three masks: LA1 for 1-P/O ETCH, LA2 for 2-P/O ETCH, and LA3 for 4 P/O ETCH are used to carry out 8-layer contacts. (c) Before poly plug, a sidewall lateral recess is carried out to enhance the isolation between poly plugs and BL pad. (d) The final TEM cross-sectional view. (e) Zoom-in views of MiLC.

Electrical performances of the 8-layer 3D VG NAND array

(1) Island-gate SSL device and Cell Initial I-V: The IdVg curves of our 8-layer island-gate SSL devices are shown in Fig. 5. It shows excellent S.S. behaviors and low leakage, which is important for the NAND string operation. The typical cell initial IdVg curves are shown in Fig. 6.

![Island-Gate SSL Devices](image)

Figure 5: Typical IdVg curves of the 8-layer island-gate SSL device.

![Array IV Curves](image)

Figure 6: Typical IdVg curves of the 8-layer memory cells.

(2) Programming performance and inhibit of the 8-layer device: Figure 7 shows the ISPP programming and self-boosting inhibit of the 8-layer device. Every memory layer (from PL1 to PL8) can be successfully programmed with memory window greater than 6V, leaving other layers well inhibited.
Figure 7: Typical programming and self-boosting inhibit characteristics of the 8-layer memory cells measured in our 3D VG array. Within one page (SSL) and a selected center WL, we program only one layer, leaving other memory cells inhibited. (a) Program PL8 only. (b) Program PL7 only. (c) Program PL6 only. (d) Program PL5 only. (e) Program PL4 only. (f) Program PL3 only. (g) Program PL2 only. (h) Program PL1 only. All 8-layer devices are successfully programmed. The self-boosting program inhibit is also successful.

Electrical performances (cont.)

(3) Block erasing performance: The block erase method is to apply a large positive bias (~+13V) at CSL and BL’s, while keeping all WL’s=0V. The SSL’s and GSL’s are applied a moderate positive bias (~+6V) to offer suitable GIDL-induced ease with minimized disturb to SSL and GSL. The erasing performance is shown in Fig. 8.

Figure 8: Block erase operation. All memory cells can be erased by WL’s=0V, CSL=BL’s=+13V. SSL’s = GSL’s =Vpass2= ~ +6V. All memory layers and all WL’s can be erased together.

(4) Number of programming (NOP) stress: For 3D VG NAND Flash, every WL has 2N pages. Thus to complete the programming for one WL, every page must endure a total NOP = 2N-1 programming stress. The typical NOP performance of our split-page VG NAND device is shown in Fig. 9. It shows the capability to sustain a high NOP = 64 stressing.
Figure 9: (a) Equivalent circuit to show the program-inhibit method. For example, to program page 0, we turn-on SSL0 by Vcc, while other SSL’s are applied a slightly negative voltage to guarantee turn-off. GSL (even) is turned-off, while GSL (odd) is turned-on. CSL is applied +Vcc to provide necessary inhibit. Page 2 (SSL2) are inhibited by floating the NAND string entirely (Mode-2). On the other hand, Page 1/3 are inhibited by precharge the channel via CSL and GSL (Mode-1). (b) NOP program inhibit performance of Mode-1, where channel is pre-charged by CSL-GSL. One NOP stress corresponds to one maximum program bias inhibit stress. (c) NOP stress performance of Mode-2. Our 3D VG NAND array has excellent program inhibit performances after NOP=64 stress.

(5) Vpass disturb: Vpass disturb is critically important for 3D NAND Flash. Since every WL has 2N pages and there are 64 WL’s, to complete a block operation the Vpass disturb stress is over 200msec for programming. Likewise, for a 100K read disturb criterion, the total stress time is over 10000 sec. Figure 10 shows the Vpass stress performance of our VG NAND at various biases. It shows enough window to sustain Vpass stress during programming. Figure 11 compares the Vpass stress of planar SONOS (VG) and nano-wire SONOS (VC). Although the curved channel of VC has field enhancement (FE) effect that allows faster speed at lower voltage, however, the Vpass disturb is also enhanced by the same FE effect, which limits the Vpass window. On the other hand, VG NAND has planar ONO that minimizes the Vpass disturb and is more reliable.

Figure 10: Vpass stress characteristics of 3D VG NAND device. After block erase, a 200msec or 2sec stress is applied at various bias voltages. For a typical 200msec Vpass stress criterion, 3D VG shows small Vpass disturb at Vpass<11V.
Figure 11: Simulated Vpass stress of planar SONOS (3D VG) and nano-wire SONOS (3D VC, [3]) at channel diameter=20nm. The nano-wire SONOS has field enhancement factor (FE) of around 1.7, and it shows much larger Vpass disturb than the planar SONOS. Vpass disturb is critically important in 3D NAND because one WL would have many pages, thus Vpass disturb is much higher than 2D NAND for both programming and reading and should be carefully minimized.

Summary
Successful 8-layer tight-pitch 3DVG NAND has been demonstrated. Figure 12 shows that VG NAND can provide 1Tb memory at 25nm half-pitch with only 32 stacked layers. For a vertical channel (VC) NAND nearly 100 layers are needed to reach the same memory density.

Figure 12: The available memory density for 3D VG at various technology nodes and stacked layer number within 100mm² area. For the split-page 3D VG, we assume a constant 60% array core efficiency. A vertical channel architecture with 43nm 6F² cell size and 90% array core efficiency is compared. We assume MLC (2b/c) for all devices.

References:

