Extended Abstracts of the 2018 International Conference on Solid State Devices and Materials, Tokyo, 2018, pp143-144

Progress of Single-Gate Vertical Channel (SGVC) 3D NAND Technology and Introduction of 3D AND-type NVM

Hang-Ting Lue

Macronix International Co., Ltd., 16 Li-Hsin Road, Hsinchu Science Park, Hsinchu, Taiwan (e-mail: htlue@mxic.com.tw)

Abstract- In this work, we first introduce the architecture design of our Single-Gate Vertical Channel (SGVC) 3D NAND Flash. Different from the usual Gate-all-around (GAA) device types in 3D NAND, SGVC produces two physical bits (double density) at each trench, thus providing an efficient way for producing high-density memory. We've successfully developed a 16-layer SGVC 3D NAND with 128Gb MLC capacity. The relatively large planar device provides excellent read disturb and is suitable for read-intensive applications. Second, we introduce a novel 3D AND-type architecture, which is a structure variation of SGVC, but with a vertical buried diffusion line that connects all transistors in parallel in an AND-type array. The 3D AND-type NVM is capable of in-memory computing and has potential to provide high-density and low-power hardware solutions to accelerate the computation of artificial neuro network.

I. Progress of SGVC 3D NAND:

3D NAND has gradually replaced conventional 2D FG NAND in recent years. Although the 1st mass produced 3D NAND was 32 layers [1], the 64-layer TLC product starts to produce lower bit cost than 2D NAND. These 3D NANDs adopt a GAA device with dimensions much larger than 2D NAND. To compensate for the large device size, it must be built with high-stacking layers that inevitably cause high processing cost. Therefore, 3D NAND requires TLC (or even QLC) to achieve low bit cost. However, TLC and QLC produce lower bit cost at the penalty of much degraded product reliability and performance.

We proposed an SGVC 3D NAND [2] architecture to provide better scaling capability by producing double-density devices in each deep trench. **Figure 1** illustrates the schematic 3D structure of SGVC 3D NAND. It's a "U-turn" NAND string, where both bitline (BL) and sourceline (SL) contacts are arranged on top. Although the U-turn NAND string has the penalty of lower read current, it has the advantage of smaller array overhead because of no source strapping area overhead.

Both top and bottom have several dummy WL's. The major purpose of the top dummy WL's is to alleviate hot-carrier disturbance during self-boosting programming inhibit. On the other hand, there is no hot-carrier issue in the bottom part, and the thick IG (inversion gate) is mainly to provide a suitable etching process window control.

The structure is arranged in a twisted layout of BL contact like other 3D NAND so that it enjoys the tight-pitch metal BL to allow higher bandwidth.

In **Fig. 2**, the structure and layout of GAA and SGVC are briefly compared. SGVC is more like a 2D planar device but arranged in 3D vertically. In each deep trench, there are two bits at both sides which are independently controlled so that SGVC provides an efficient double-density memory. The layout and design rule are compared. SGVC can be manufactured in a 0.1um * 0.22um unit area with double density, while GAA device is produced in a ~0.16um pitch. Thus SGVC has much smaller effective cell area.

We've successfully developed a 16-layer SGVC 3D NAND. **Figure 3** shows the die photo of our 128Gb MLC SGVC 3D NAND. The die size is close to 80mm^2, which is significantly smaller than the 2D 1ynm MLC (~120mm^2) for the same 128Gb MLC.

One important technology feature of SGVC is that we adopt a metal-strapped WL (at M3) instead of metal gate, as shown in **Fig. 4**. In 3D NAND, each WL will share plural strings (corresponding to the SSL for each string) horizontally, thus enabling a sizable WL pad area for each block. We've developed a special staircase process that can be fit in the WL pad area. The poly gate WL's are connected to the top low-resistance M3 WL per ~200um length, thus giving very low WL RC delay. **Figure 5** shows the typical MLC distributions. The fail-bit count (FBC) distributions can be well controlled within the range allowed by the regular BCH error correction code (ECC), even after reliability test. The device can directly pass 120M single-page read aging test without the need to refresh the data.

II. 3D AND-type NVM for AI

By slightly changing the array architecture but keeping most of the 3D processes, SGVC can be turned into a new device for AI application. In AI application, the deep-learning artificial neuro network (ANN) is the central part, and it requires huge memory (mainly to store "weights" for vector matrix multiplication, also called "MAC"). The conventional von-Neumann architecture needs buses that communicate memory with LOGIC circuits and this is the bottleneck of throughput and power consumptions. "In-memory computing" is highlighted as a solution that saves the cost and power of data movements for ANN [4].

Right now in-memory computing mostly relies on SRAM-based design, which suffers from limited density (~Mb level). A high-density (>1Gb) on-chip NVM which supports the in-memory computing is very desirable to boost the system performance (no bus) with low power (much less data movements) and low cost (to minimize SRAM and DRAM).

Currently the most popular in-memory computing using NVM devices are NOR-type 1T1R structure with ReRAM, or FG NOR Flash architecture. However, scaling capability is quite limited, far below 1Gb for these architectures. Moreover, there are challenges that connect layer to layer for deep learning, because the data transfer from output of previous layer to the input of next layer requires several digital to analog transformations and also charge pumping for WL voltages.

On the other hand, NAND Flash (including 3D NAND) is good for high-density storage, but probably not good for high-speed computing. The main reason is the small read current of serially-connected devices that causes long latency (>> 1usec) for read. The array loading effect (string current is affected by other unselected WL's) also causes poor accuracy for the summed current of plural BL's.

Considering the need for AI computing memory, we modify our SGVC 3D NAND and achieved a novel 3D AND-type architecture [5], as shown in **Fig. 6**. Different from SGVC 3D NAND, we introduce a vertical buried diffusion line that connects all transistors in parallel in an AND-type array, where both BL's and SL's are parallel. **Figure 7** briefly illustrates the process structure of this architecture. This new architecture has the advantages of high-density, high-bandwidth design, low-power bit-alterable +/- FN Program/Erase, and supports near analog operation (Type I) with multi-level input and multi-level storage.

There are two ways to design an AI 3D AND. In Type-I (**Fig. 8**), the summed current is carried out in plural BL's. Both input and output are at the same BL side, thus we can design an efficient data transfer from output of the previous layer to the input of the next layer at the sense amplifier (SA) circuits, and adopt the time domain design for deep learning using the same array without the need of too many tiles for neuro network. Type-I has potential to serve for the many convolution layers that requires extensive computing.

In Type-II (**Fig. 9**) design, the WL serves as the input. Summation is carried out in multiple WL's. Type II is only suitable for binary mode operation. Due to the high parallelism of computing, Type–II is suitable for ~~the~~ fully-connected (FC) layers.

The 3D AND-type design has potential to enable a high-density (>>1Gb) device for the AI computing era.

References:

[1] K.T. Park, et al, ISSCC, Session 19.5, 2014.
[2] H.T. Lue, et al, IEDM, Session 19-1, pp. 461-464, 2017.
[3] K. C. Ho, et al, IEEE transactions on very large scale integration (VLSI) systems, vol. 24, NO. 4, pp. 1293-1304.
[4] VLSI Symposia 2018, Friday Forum for AI.
[5] H. T. Lue, et al, VLSI 2018, pp. 177-178.

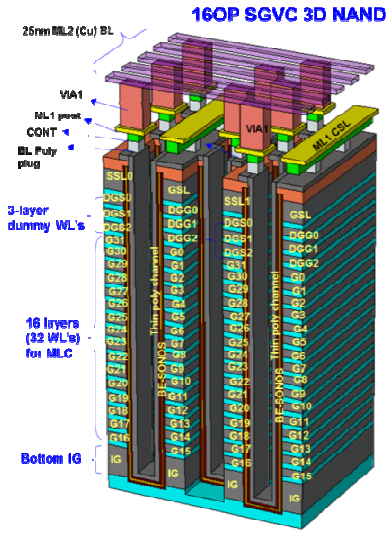


Fig. 1 The 16-layer SGVC 3D NAND architecture. It's a U-turn NAND string, with both BL and SL contacts on top. Some dummy WL's are adopted at top and bottom parts.

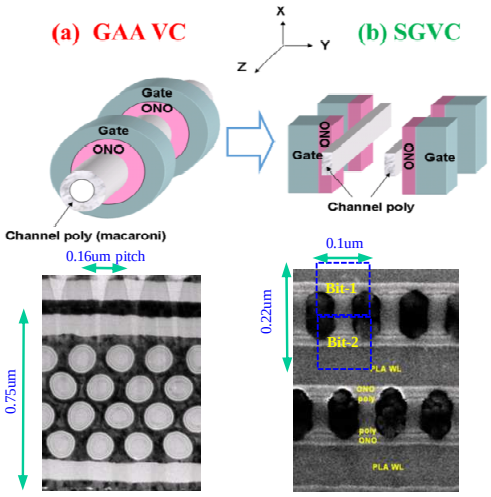


Fig. 2 Comparison of GAA and SGVC. **(a)** GAA device. The typical design rule is ~0.16um pitch for hole. Some overhead needed for the WL slit for source strapping. **(b)** SGVC device. It's double density in each trench, with flat device that has more pitch scaling capability.

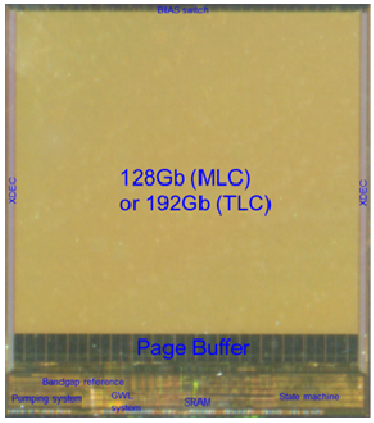
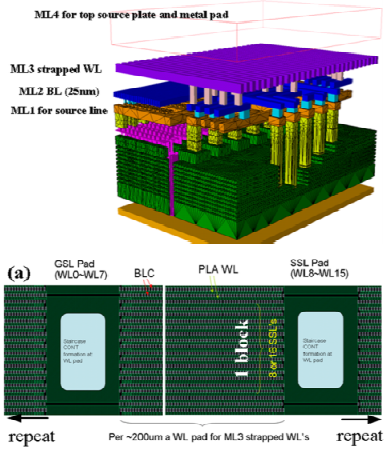


Fig. 3 A128Gb MLC product of SGVC 3D NAND designed for the mass production. The die size is close to 80mm^2, which is significantly smaller than 2D NAND (~120mm^2 for 1ynm MLC).



One block has plural SSL's in parallel (such as 8 or 16 SSL's) shared by the same WL pad. The block boundary has only one additional PLA overhead.

Fig. 4 Schematics of metal-strapped WL for SGVC. The high-resistance poly gate are connected to a low-resistance M3 lines at the WL pad, where a special staircase processing are produced to allow the WL contact at different layers. The layout overhead including WL pad and source connection is <6%.

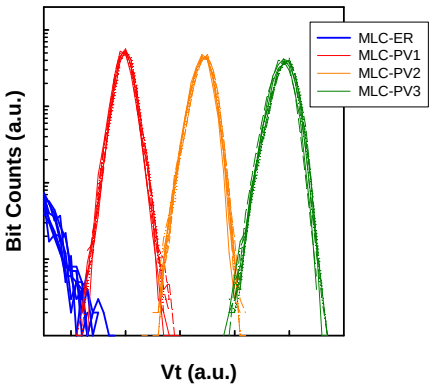


Fig. 5 The Vt distribution after full-block MLC programming, collected by wafer sort for many-die distributions. The memory window is enough for MLC, with fail-bit count (FBC) within the BCH correctable regime, after reliability test.

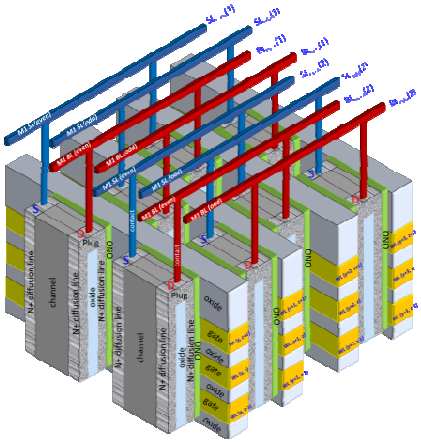


Fig. 6 A novel 3D AND-type NVM device [5] proposed to serve for the high-density and low-power in-memory computing device. The structure resembles SGVC 3D NAND, but with vertical buried diffusion lines introduced that connect all memory transistors in parallel. The sum-of-product (vector matrix multiplication) is readily provided.

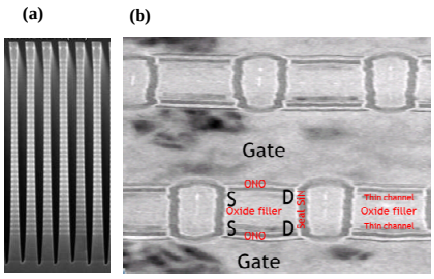


Fig. 7 Structure illustration of 3D AND. It resembles SGVC 3D NAND, but with certain modification. **(a)** The gate stack trench etching. Nearly ~3um depth is produced to provide 32-tier feasibility. **(b)** The structure plane-view TEM. In each trench, there are twin cells, where the two-side ONO are independently controlled by different gates, similar to SGVC. Charge-trapping (BE-SONOS) device is utilized for storage, and thin body (Tsi<10nm) poly-silicon channel is adopted. At the edge of the thin body, a sidewall N+ diffusion line is formed, which forms the vertical buried diffusion line that connects all transistors in the same trench of the 3 AND-type NVM.

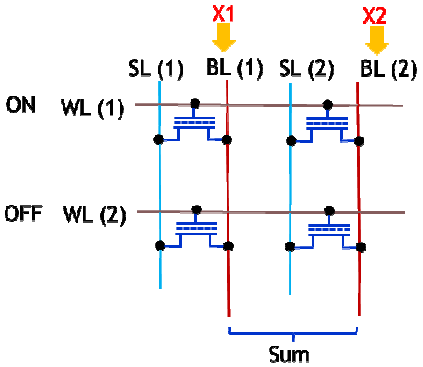


Fig. 8 Type-I operation method of summed product. Turn-on one WL each time, with multiple BL's as input. (Xi). Summed current is carried out for plural BL's. Both input and output are at the same BL side, thus we can quickly transfer the output data to become the input for the next-phase of a convolution neuro network. BL's can have multi levels, while the stored weights (conductance) can be also multi-levels. This provides near-analog (multi-level) computing.

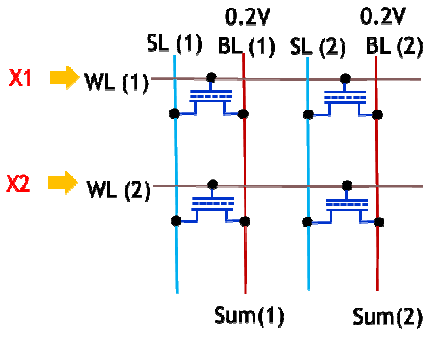


Fig. 9 Type-II operation method of summed product. Turn-on multiple WL's as the input (Xi), and assign constant BL bias for sensing. Summed current is provided over plural WL's. It shows larger parallelism by the multiplication of multi WL's and multi BL's. It's good for the fully-connect (FC) layer operation.