A Novel 3D AND-type NVM Architecture Capable of High-density, Low-power In-Memory Sum-of-Product Computation for Artificial Intelligence Application

Hang-Ting Lue, Weichen Chen, Hung-Sheng Chang, Keh-Chung Wang, and Chih-Yuan Lu

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

Abstract

An AND-type stackable 3D NVM architecture is proposed to provide an ultra-high density AI computing memory with low power. The advantages are: (1) All memory transistors in the 3D array are connected in parallel, thus enable the sum-of-product operation. (2) The 3D NAND like architecture is possible to stack to > 64 layers, thus provides ultra-high density (>128Gb) AI memory. (3) Many bit lines (>1KB) can operate in parallel for high bandwidth. (4) Uses low-power +/- FN programming/erasing which allows high parallelism, and is bit-alterable thus is ideal for training or transfer learning. (5) Excellent linearity of output current with respect to bitline bias, thus enabling ideal analog computation. (6) Adequate sensing current of the summed product thus permits fast access read for inference device. The proposed memory architecture can achieve TOPS/W>10, which is 10X greater than the conventional von Neumann architecture.

I. Introduction

The conventional "von Neumann architecture" carries out computing in a CPU, which is connected to memory devices (DRAM) though buses. Such design has limited bandwidth and high power consumptions. A memory capable of in-memory computation is highly desired because it can directly carry out the computation in the memory devices to improve the efficiency of computing [1,2].

The most popular computing memory design is to use a NOR-type array, where the transistor is often connected with a resistor memory such as ReRAM or PCM. Such 1T1R structure has some disadvantages. It is not scalable and has limited memory density much below 1Gb. The programming current of ReRAM or PCM is high (>50uA for each cell) thus does not allow highly parallel programming. The resistor devices of ReRAM or PCRAM are often highly non-linear with respect to the applied bias which makes analog computing difficult.

In principle, the 1T NOR Flash without resistor memory is also feasible for AI memory design. However, the conventional NOR Flash has limited scaling capability thus is not cost effective. Also, NOR Flash does not allow bit-alterable operation. Unfortunately, high-density NAND Flash (and 3D NAND) may not be suitable for AI computing because it naturally has slow latency (small read current), and the serially connected NAND string cannot carry out current summation.

II. Structure Explanation of 3D AND-type NOR Flash

Figure 1(a) illustrates the proposed architecture. The structure resembles the SGVC 3D NAND [3], but we introduce vertical N^+ buried diffusion lines that connect all memory transistors in parallel. **Figure 1(b)** illustrates the top view layout. Both bitline (connected to drain) and source lines (connected to source) are arranged in parallel direction, thus it should be classified as an "AND-type NVM". It is actually quite similar to the conventional NOR Flash, but AND-type array with parallel BL's and SL's can prevent sneak paths during programming, yet allow +/-FN bit-alterable operation in the 3D array.

Similar to 3D NAND, the memory cells are arranged in a twisted layout so that the density of metal bitline (BL) and source line (SL) are doubled to increase the bandwidth. There are two memory cells inside each trench, which share the same BL and SL by the common connection of a poly plug, but they do not share the same WL. Meanwhile, the memory cell in the nearby adjacent trench shares the same WL, but fortunately they do not share the same BL/SL, thanks to the twisted layout structure that automatically separates the BL/SL for them.

Figure 1 also illustrates an example (type I) of sum-of-product computation. To select a memory cell is to apply bias voltages at the corresponding WL and BL's, while de-selected WL's and BL's=0V. Each memory cell store the data information of weighting factor (W(x,y,z)), which is the conductance (Id/Vd) of the memory cell. The conductance can be programmed/erased to adjust the value. We may parallelly select plural BL's with various BL biases to carry out summed product. The currents are summed in the SL's (selected by SL decoder) for sensing.

The zoom-in structure is illustrated in **Fig. 2(a)**. There are two bits (Bit-1 and Bit-2) inside each trench, which are controlled by the two-side gates separately. The read current flows from the buried diffusion line vertically, and then go into the memory cell horizontally as indicated in the figure. Like 3D NAND, thin-body poly-silicon TFT device is utilized. Lg is the channel length, while Tsi is the thin-body thickness. W is the OP stack poly thickness, which is equal to the effective channel width.

Figure 2(b) and (c) illustrates the processing feasibility. High stacking process of > 64 OP layers like SGVC 3D NAND is doable.

Figure 3 briefly explains the process flow. After a trench etching, the ONO (for charge-trapping device) and thin poly channel are filled-in, followed by the oxide fill-in and top poly plug formation. A hole-type etching is carried out for isolation. Next, a plasma doping method is carried out to dope the edge of the channel into Drain and Source for the N^+ buried diffusion line. The metal BL and SL are connected on top of the N^+ diffusion. Optimized p-type doping at the thin channel and poly plug are necessary to prevent from punch-through leakage. Sufficient N^+ diffusion doping is necessary to reduce the buried diffusion resistance.

Figure 4 illustrates the possible memory design layout schematics for a tile in a chip. Various biases, which stand for the input signals, are applied to the bit lines. Multiple BL's can be operated simultaneously for improving the computing bandwidth. SL's are connected to the sense amplifier via a source line decoder which is designed to allow flexible selection of the address to carry out summed product. A typical input number for a summation unit is 8 (or16) BL's, where the 8 source line currents are summed together for sensing.

III. Electrical Performances and Simulations

Figure 5(a) shows the typical poly silicon TFT device's IdVg characteristics measured in a simplified process. Read current for a single cell can be ~5uA at maximal, with adjustable conductance after programming. **Figure 5(b)** shows that the Id is very linear with respect to Vd, which is important to support the analog computation, where BL's can be applied various bias instead of only digital (binary mode).

Figure 6(a) illustrates the programming simulation for the array. To select the cell, we can apply +FN ISPP voltages on the corresponding WL. The selected BL's=0V to allow programming, while de-selected BL's= +6V for inhibit. **Figure 6(b)** shows the simulated +FN programming, which has excellent program inhibit for de-selected cells.

Similar operations can be applied for erase, but with reverse polarity as shown in **Fig. 7(a)**. Unlike conventional Flash memory, bit-alterable erase is possible because the N^+ diffusion can directly pass through the BL bias even when other unselected WL's=0V, without the need of pass gate WL's operation in NAND. **Figure 7(b)** shows that erase is indeed selectable. The bit-alterable operation together with high parallelism of FN P/E offer efficient high-density memory training.

Figure 8 shows the estimated conductance with MLC distribution, which is emulated from the SGVC 3D NAND Vt distribution. Multi-level storage together with multi-level BL input voltages improve the throughputs of computing. **Figure 9** illustrates the range of summed current for an 8-input unit. The summed current range is ~40uA, which is adequate for NOR Flash like sensing to meet the fast random read for inference device. The summed current is almost linear to BL voltages, possible to support analog computation.

IV. Summary:

Figure 10 proposes two types of design method to produce the sum-of-product for 3D AND NVM. Type I (BL input) is suitable for the high-resolution "convolution" operation with analog input, while type II (WL input) is suitable for high-density and high-bandwidth "fully-connected" operation with binary mode. Optimized design may produce TOPS/W ranging from 5~40, which is 10 times greater than conventional von-Neumann architecture.

References: [1] G. Burr, IEDM Tutorial 2, 2017. [2] W. H. Chen, et al, IEDM 2017, session 28-2. [3] H. T. Lue, et al, IEDM session 19-1, 2017.

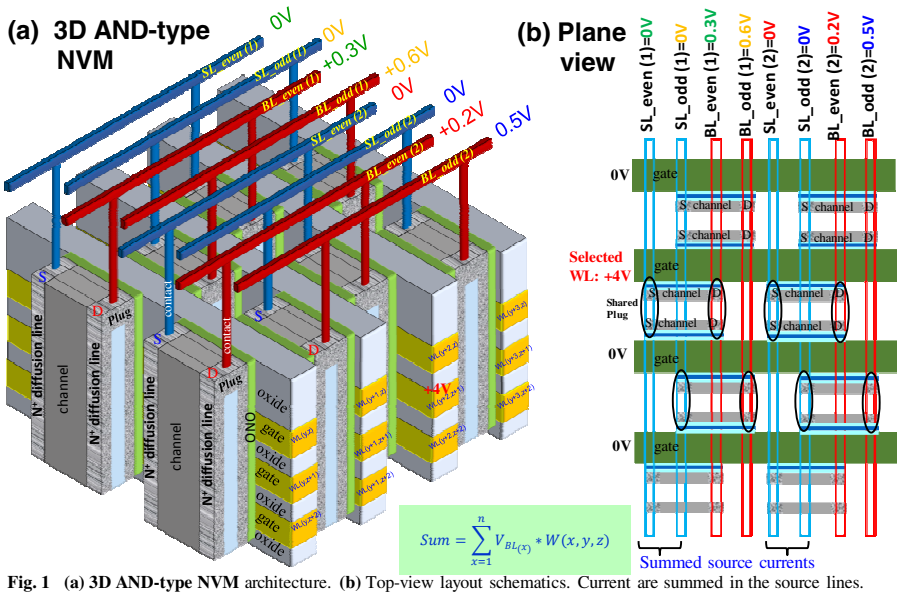


Fig. 1 **(a) 3D AND-type NVM** architecture. **(b)** Top-view layout schematics. Current are summed in the source lines.

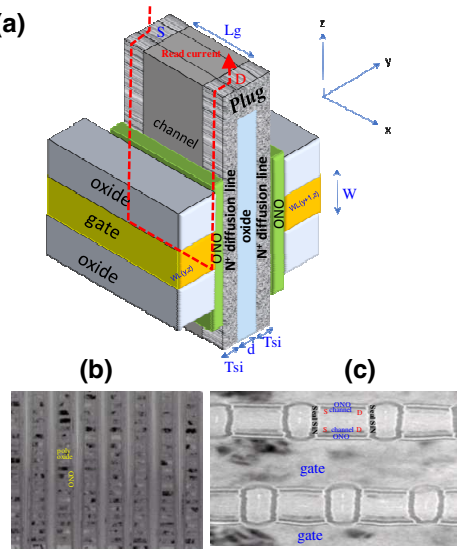


Fig. 2 **(a)** Zoom-in view of the device. **(b)** Cross-sectional view. **(c)** Plane-view of the device structure.

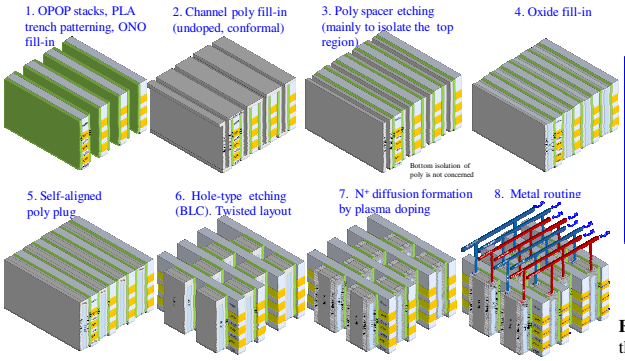


Fig. 3 Process flow to fabricate 3D AND-type NVM. After PLA trench etching, ONO and poly channel fill-in, a BLC hole etching is used to isolate device. Next, a plasma doping method is used to dope the sidewall to form the N+ diffusion line vertically, followed by metal routing.

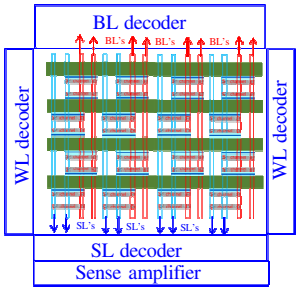


Fig. 4 Schematic showing how to design the 3D AND-type NVM. BL's are applied various voltages (input), and the current is sensed through the source lines. Address selection for summation is arranged by the source line decoder.

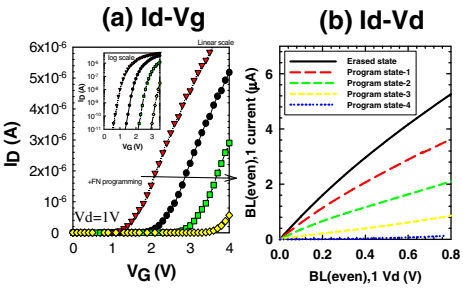


Fig. 5 The experimental data of IdVg and IdVd curves of a typical poly silicon TFT BE-SONOS device fabricated in a simplified 3D process (Lg=80nm, W=40nm, Tsi=6nm). **(a)** Each single cell can provide ~5uA current at sufficient gate overdrive. **(b)** The IdVd shows excellent linearity, where Id is almost linear to Vd.

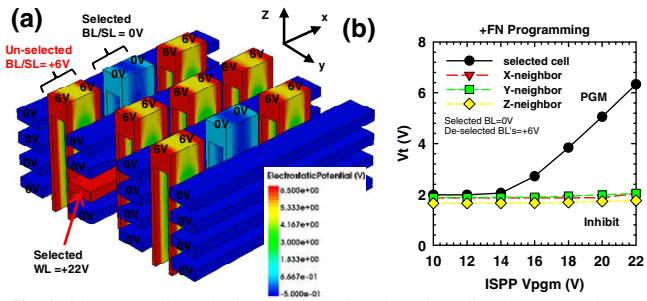


Fig. 6 **(a)** Programming selection method. Selected WL is applied ISPP +FN voltages. The selected BL's =0V, while de-selected BL's are +6V for program inhibit. The +6V bias can directly pass through the buried diffusion line to inhibit the bottom layers. **(b)** The selected cell can be programmed, while the X/Y/Z neighbor cells are well inhibited. **Y-neighbor device (back-to-back cell) is free from interference because the channel has shielding effect.**

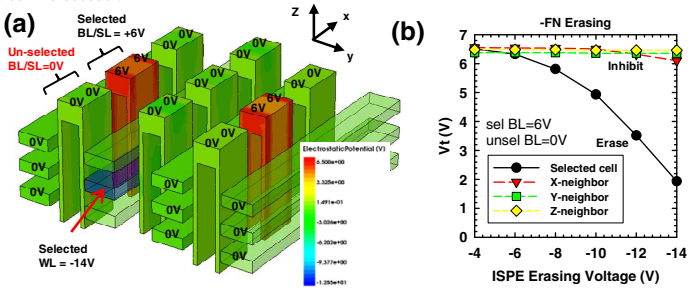


Fig. 7 **(a)** Erasing selection method. Selected WL is applied ISPE -FN voltages. The selected BL's =6V for erasing, while de-selected BL's are applied 0V for erase inhibit. The +6V bias can directly pass through the buried diffusion line for selected cell to erase. **(b)** The selected cell can be erased, while the X/Y/Z neighbor cells are well inhibited. This provides the bit-alterable erase operation.

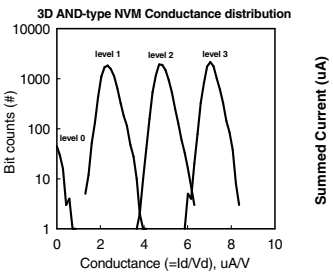


Fig. 8 The conductance (Id/Vd) distribution of 3D AND-type NVM, emulated from the MLC Vt distribution of SGVC 3D NAND chip.

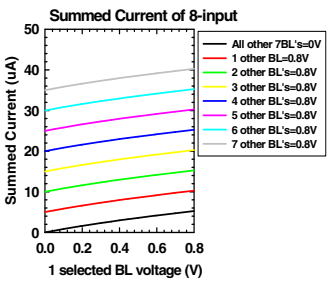
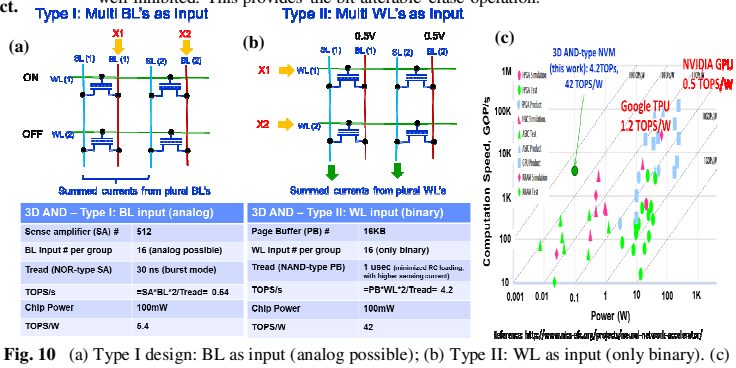


Fig. 9 The summed current of the 8-input unit. The maximal range is around 40uA when all 8 BL's are applied 0.8V. It's linear with respect to the BL voltage.



3D AND – Type I: BL input (analog)	
Sense amplifier (SA) #	512
BL input # per group	16 (analog possible)
Tread (NOR-type SA)	30 ns (burst mode)
TOPS/s	=SA*BL*2/Tread= 0.64
Chip Power	100mW
TOPS/W	6.4

3D AND – Type II: WL input (binary)	
Page Buffer (PB) #	16KB
WL input # per group	16 (only binary)
Tread (NAND-type PB)	1 usec (minimized RC loading, with higher sensing current)
TOPS/s	=PB*WL*2/Tread= 4.2
Chip Power	100mW
TOPS/W	42

Fig. 10 (a) Type I design: BL as input (analog possible); (b) Type II: WL as input (only binary). (c) Benchmark of TOPS/W. 3D AND may provide TOPS/W ranging from 5-40, greater than conventional von-Neumann architecture.