Study of Analog Weights Based Computing-in-Memory (CIM) using a Highly-Reliable and Small-Noise Vertical-Channel Gate-All-Around Split-Gate Floating Gate NOR Flash for Vector Matrix Multiplication (VMM) Accelerator

Hang-Ting Lue, Chung-Hao Fu, Tzu-Hsuan Hsu, Ming-Liang Wei, Teng-Hao Yeh, 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)

2022 IEEE Silicon Nanoelectronics Workshop (SNW) | 978-1-6654-5979-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/SNW56633.2022.9889068

Abstract

We develop a novel vertical-channel gate-all-around (GAA) split-gate floating gate (FG) NOR Flash memory device that has very small RTN noise (<1.5%), tunable and tight Icell (weight) with extremely small Ioff ~pA, excellent retention and read-disturb free reliability in order to enable the analog computing-in-memory (CIM) for high-bandwidth VMM accelerator in various AI applications. To support 4-bit resolutions in edge computing, we can directly provide 4-bit weight of 0~15 Icell levels (0~3uA) by means of the combinations of 3 cells with 5 levels in each cell, controlled by the bitline switches (BLT). We can produce sigma/mean of 4-bit analog weights in between 4-9% to support sufficient accuracy for deep neural network (DNN). However, we find difficulty in producing analog WL inputs because of the challenges in the transconductance (g_m) mismatch (variation) of memory cells. We therefore distribute the 4-bit inputs to plural WL's (select gate (SG) in our design) with single-level bias for accuracy concern. Because of extremely small OFF state current (~pA) and moderate ON-state Icell, our vertical GAA FG NOR device can support hundreds of inputs signal in one VMM computing with a steady computing bandwidth (several TOPS at 4bit) for the stationary big data. This novel Flash memory VMM accelerator can complement the Von-Neuman digital computing in saving the large stationary data movements in DRAM I/O. Applications of analog VMM accelerator include DNN inference, cosine similarity search, and digital annealing, etc.

I. Introduction

Non-volatile computing in memory (nvCIM) for VMM accelerator has been studied for years [1-3]. **Figure 1(a)** represent a general picture although it refers to our device in this work. Ideally, if we can produce analog inputs (in WL direction) and analog weights (in multi-level Icell), and if we can sum many WL's together in a BL, the summed current directly represent VMM value without the need of conventional digital circuits. Most importantly, we do not need to move the stationary data (such as weights) and this should save system power significantly.

The original idea of analog VMM accelerator was to compete with the conventional Von-Neumann digital computing (GPU + DRAM). However, through years of studies we found that it's better to complement the Von-Neumann approach, especially for pre-processing the big (stationary) data before detail digital computing. VMM are usually the most "memory-centric" computing operations that are gated by DRAM I/O bandwidth (ex: up to 6.4GB/s for DDR5). If we develop a high-density Flash memory device that can compute VMM at higher effective computing bandwidth (~TOP/s) than DRAM I/O, this already makes sense to improve the system performance. We do not need to compete with LOGIC peak performances but instead to compete with DRAM I/O for the stationary data computing.

Flash memory is for sure the best commercialized high-density non-volatile memory. In particular, NOR Flash for VMM accelerator is straightforward because the memory devices are connected in parallel (like other cross-bar emerging memories) so that the summed currents by multiple WL's are readily produced, as shown in **Fig. 1(a)**. Usually people consider Flash memory device is much slower than DRAM in both read latency and bandwidth. In VMM accelerator, the important advantage of Flash memory is that we can turn on hundreds of WL's together to boost the VMM bandwidth that far exceeds DRAM I/O performance. This creates an opportunity of Flash memory to replace part of DRAM for the stationary data computing of VMM.

For 4-bit weight, we can produce 5 levels of Icell (200nA~1uA), and parallelly connect 3 cells (by BLT switches in standard NOR Flash circuit design) to produces a total of 15 levels. The OFF state current must be very low (~pA in our device) so that we can sum hundreds of WL inputs without the background leakage issue.

In this work, we develop a novel vertical-channel gate-all-around split-gate floating-gate NOR Flash (GAA FG NOR) device to best optimize the VMM performances. The structures are illustrated in **Fig. 1(b) to 1(d)**. The device consists of bottom select gate (SG) to control the string, top memory gate (MG) with FG storage. We use gate replacement process like 3D NAND to manufacture the device. Selective epi Si is used as the channel (body) of the GAA device.

The split-gate Flash allows low-voltage read (SG=+1V, BL=+0.2V, MG=0V) to facilitate low-power and high-speed computing. The vertical GAA Flash structure has excellent areal scaling capability than 2D planar cell (>4X memory density than eFlash). Compared with our counterpart of charge-trapping device in the similar vertical GAA structure [3], we found that the FG storage have superior retention and smaller noise that are very good to realize analog VMM accelerator.

II. Device Performances

Figure 2(a) shows the typical IdVg curves. SG has nice subthreshold slope and can be completely turned off to <1pA at SG=0V. MG (w/ FG) can possess large P/E window (>7V). For VMM, we just need around <2V total Vt window to generate Icell=200nA ~ 1uA at MG=0V, as shown in **Fig. 2(b)**. For OFF state, we program to a higher Vt to get very small leakage of ~pA. The device merits including small leakage, large ON/OFF ratio of 6 orders, and the moderate Icell enable the possibility to sum hundreds of WL's, with VMM current range below 128uA, suitable for ADC design. **Figure 2(c)** shows that the RTN is extremely small (<1.5% at 1E-4 probability), indicating the excellent GAA transistor. **Figure 2(d)** shows the programming by source-side injection (SSI) method can gradually tune for various Icell.

Figure 3(a) shows the Icell distribution ranging from 200nA to 1uA for 5 levels. The distribution sigma/mean are all less than 9%, where larger Icell (corresponding to most significant bit) has even tighter distribution of ~4%. This allow analog weight with acceptable accuracy. **Figure 3(b)** shows the excellent retention of various Icell even we bake at 150C. The device is completely read-disturb free because MG=0V. FG device has very outstanding retention without tiny Vt drift, thus enabling analog operation. This differentiates it from many emerging memories.

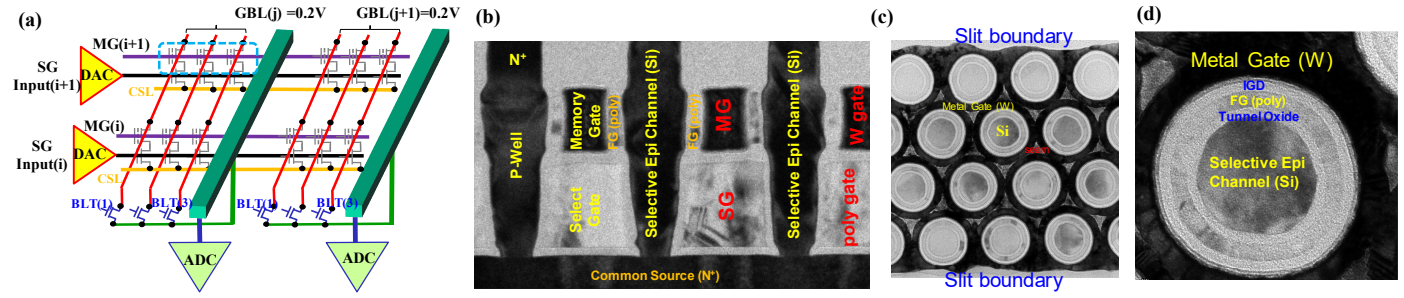
On the other hand, we've found challenges in producing analog inputs. In **Fig. 4(a)**, we can trim for a tight Icell distribution at a fixed read bias (MG=0V). However, due to the finite cell mismatch (variations) in g_m, we can not well control the Icell at various MG bias (**Fig. 4(b)**). We think that g_m mismatch is a general problem for all high-density memory devices. Variations often goes higher for high-density memory.

VMM accelerator has several applications. **Figure 5(a)** shows the example of edge computing inference with VGG7 image network. We train for the 4-bit resolution network. Since our device support 1I4W (1bit input, 4bit weight), we need to distribute each 4bit input to 4 different tiles with shifter and adder. Deeper convolution layers often require thousands of inputs as shown in the table of **Fig. 5(b)**. **Figure 5(c)** shows the network accuracy with various weight and input noise. It is found that the network can tolerate up to 10% (sigma/mean) weight noise, while less tolerance of input noise (<6% sigma/mean). Thus analog weight is far more feasible than analog inputs. **Figure 5(d)** shows the computing bandwidth (TOP/s, at 4bit) with various scenarios of maximum allowed SG number in each VMM computing. The system simulation indicates the required chip memory density is 172Mb (parallel 8 cores), with 4096 ADC. Each ADC latency is 100ns (8bit resolution, with weights shifted to positive). Only 64KB SRAM is needed to store the feature maps (meta data). Due to the SG decoder layout limitations and summed current restrictions, the maximum allowed SG number is in the range of a few hundreds. For input number >1000, we need to divide VMM into multiple steps. A practical performance is in the range of several TOP/s (@4bit), with chip power less than 1W for Flash memory design. We think that the analog Flash VMM accelerator does not far exceed the Von-Neumann brick wall of TOPS/W~10 (@>4bit), but the major benefit is to save stationary data movements and reduce DRAM usage, instead of competing with advanced GPU. Flash VMM accelerator aims for high-density stationary data computing to complement Von-Neumann computing.

III. For AI applications and Summary

Figure 6 illustrates two more examples of VMM applications, including cosine similarity search in face recognition, and the digital annealing. Both require intensive high-bandwidth VMM computing with stationary big data, and could be useful target applications.

Reference: **[1]** M. F. Chang, VLSI 2019, Short Course. **[2]** W. H. Chen et al, session 31-4, pp.494-495, ISSCC 2018. **[3]** T. H. Hsu, et al, IEDM 2020, pp. 111-114



$$VMM = IGBL(j) = \sum_{i=1}^{N\,(SG's)} \sum_{k=1}^{3\,(BLT's)} g_m(i,k) * V_{SG}(i)$$

Icell

1I4W: 1bit input (SG), 4-bit weight (5 levels of Icell, 3 BLT)

Fig. 1 **(a)** Schematic of analog computing-in-memory (CIM) for VMM accelerator. In WL-direction, select gates (SG) serve as the inputs, with digital-to-analog converter (DAC). In each global BL (GBL), we sum plural cell currents and use analog-to-digital converter (ADC) to calculate VMM. For "1I4W" (1-bit input, 4-bit weight), input is from each SG, while we directly produce 4-bit weight by using 5 levels of Icell, plus 3 BLT's to produce total of 15 levels (0~15). **(b)** The cross-sectional view of novel GAA FG NOR device. Bottom select gate (SG) is for string selection, while top memory gate (MG) has FG storage. **(c)** Plane view of the array. **(d)** Zoom-in view of the device. The tunnel oxide ~8nm, IGD (inter gate dielectric) is composed of oxide and Al_2O_3, GAA Si diameter~70nm.

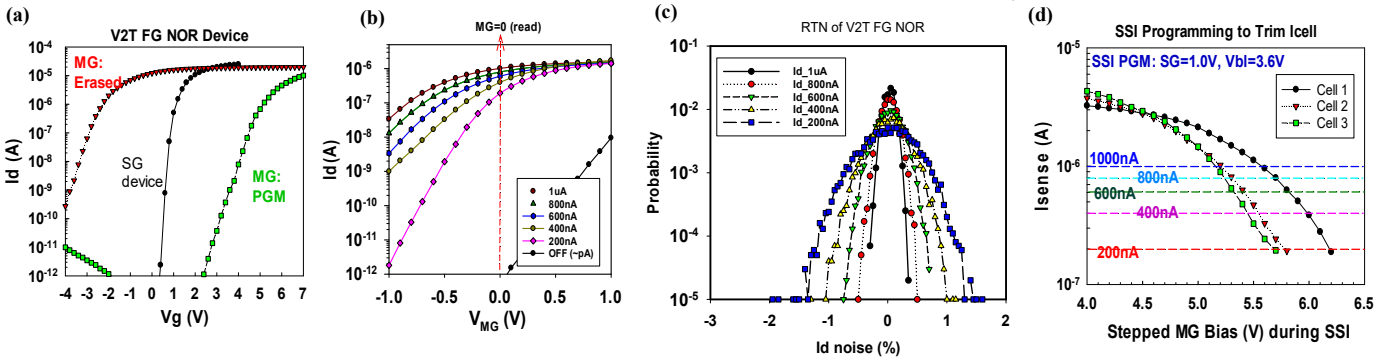


Fig. 2 **(a)** Typical Id-Vg curves of GAA FG NOR device. The SG has tight Vt distribution and good slope. SG=0 can completely cut-off the current. MG (w/ FG) device has a large P/E Vt window of more than 7V. Programming adopts source-side injection (SSI), while erasing uses –FN tunneling. **(b)** The IdVg curves show total of 6 programmed levels (Icell=0 (OFF), 200nA, 400nA, 600nA, 800nA, and 1uA). MG=0, SG=1V during reading. **(c)** The random telegraph noise (RTN) at various Icell. RTN are all smaller than +/- 1.5%. **(d)** The programming method of SSI with finer MG stepping bias at lower V_{BL}=3.6V possess slowly programming to trim for various Icell ranging from 1uA to 200nA for multi-level weights.

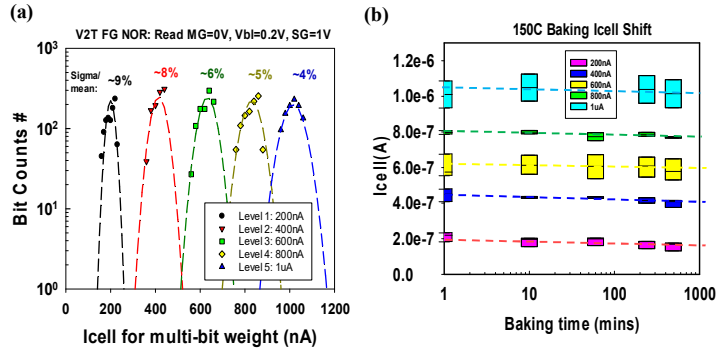


Fig. 3 **(a)** Multi-level Icell distribution for level 1~5, corresponding to 200nA to 1uA, respectively. (level 0 with small ~pA not shown). The distribution sigma/mean are all <10%. Higher Icell have smaller sigma/mean of only 4%. **(b)** The 5 levels of Icell show good retention even after 150C baking test. This provides a highly reliable analog CIM device.

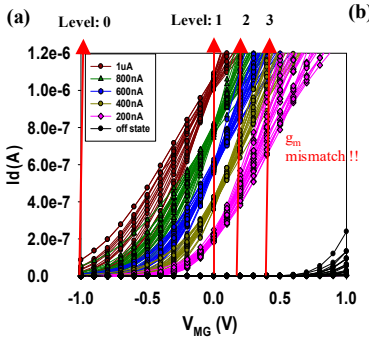
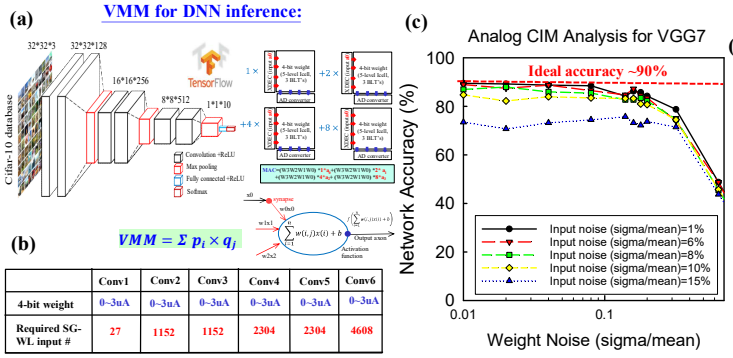


Fig. 4 Study of the feasibility of multi-bit input by various MG bias. **(a)** Gallery of many IdVg curves that are trimmed for various Icell ranging from 200nA to 1uA, at a fixed read bias of level-1 input at MG=0V. We found the memory cell g_m mismatch (variations) causes a high dispersions at level 2 and 3. **(b)** The collected Icell distribution with analog input MG bias. If we target to trim level 1, the level 2 and 3 are much more widely distributed than level 1 due to the g_m mismatch of memory cells. The noise sigma/mean will exceed 10% and not acceptable.



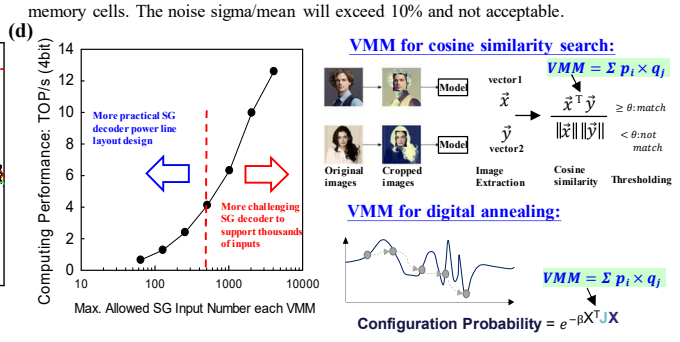


Fig. 5 (a) A 4-bit VGG7 to study the analog CIM. **(b)** 1I4W scenario: We produce 4ibt weights by means of 3 memory cells (5 levels each). 1-bit input in SG's. Thus to represent 4-bit input, we use shifter and adder. According to the convolution layer operations, the required SG input number ranges from 27 to 4608, from Conv1 to Conv 6. **(c)** The network accuracy with respect to weight noise and input noise according to 1I4W model of CIM. To make CIM close to ideal software accuracy, it is suggested to control **weight noise within 10%, while input noise within 6%.** **(d)** The computing bandwidth performance (TOP/s at 4bit) in a chip, assuming ADC read latency=100ns, total of 4096 ADC (8 parallel cores with repeated memory cells), at various maximum allowed SG input number per computing. Higher SG-WL input parallelism boosts the bandwidth, but there is a practical layout limitation to support >1000 power lines.

Fig. 6 Applications of VMM accelerator, including "cosine similarity search" in face recognition, and digital annealing methods. A high-density and high-bandwidth VMM accelerator using 3D Flash memory device can process big (stationary) data and save DRAM I/O data movements.