A Computing-in-Memory Macro for Low Bit Quantization Convolutional Neural Network and Transformer

Dong-Gu Choi and Jong-Hyeok Yoon
Department of EECS, DGIST, Daegu

I. Introduction

Deploying deep neural networks (DNNs) on edge devices demands high energy efficiency, conventional von Neumann architectures suffer from costly data movement between memory and compute cores during multiply-and-accumulate (MAC) operations. Computing-in-memory (CIM) mitigates this bottleneck by executing parallel MAC operations directly on the memory. While SRAM-based CIM offers high accuracy, its large cell area limits density; MRAM provides non-volatility and high density but has a limited ON/OFF ratio. To optimize accuracy, efficiency, and density, we propose a hybrid CIM macro integrating SRAM and MRAM banks in Samsung 28nm FD-SOI technology.

II. Hybrid CIM Architecture

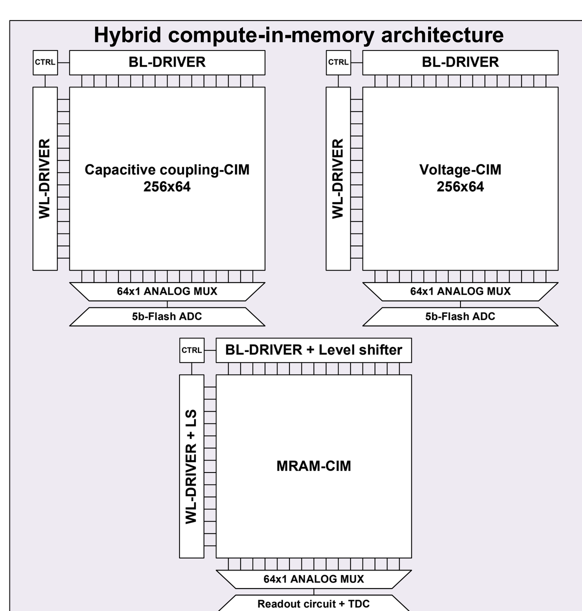

Fig. 1. Top block diagram of the hybrid CIM architecture.

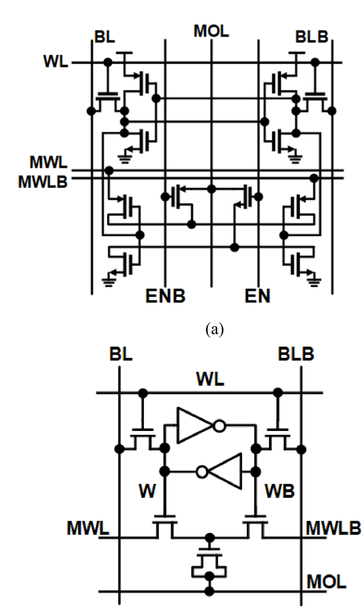

Fig. 2. Schematics of voltage-mode and charge-mode bitcell.

The hybrid CIM macro (H-CIM) comprises a voltage-mode SRAM CIM (VS-CIM), a charge-mode SRAM CIM (CS-CIM), and an MRAM CIM (M-CIM). The SRAM-CIM arrays, each with a 256×64 configuration, are tailored for transformer workloads demanding high computational accuracy. Two complementary sensing schemes—voltage-mode and charge-mode—are implemented to evaluate the impact of sensing on MAC accuracy. Except for the driver circuits and the customized bit-cell architectures, the rest of the readout circuitry was kept identical to ensure consistent experimental conditions. The MRAM array with a 4×16 configuration is optimized for lightweight, low-bit precision CNNs such as LeNet-5, ResNet-20, and VGG-style networks.

The VS-/CS-SRAM-CIM macro supports three operations: write, read, and MAC. In write operation, the WL driver asserts the word line, turning on the pass transistors, while the BL driver applies input data to BL and BLB. In read operation, BL and BLB are pre-charged to VDD, then the WL driver activates the pass transistors, allowing one BL to discharge according to the stored bit. In VS-CIM, each bitcell performs a bitwise XNOR or AND with its stored weight, and bitcells sharing the same MAC output line (MOL) form a voltage divider producing the analog MAC output. In CS-CIM, the MOL is precharged to half-VDD and charge variation is capacitively coupled, yielding the MAC output through a capacitive divider. Both outputs are digitized by a 5-bit flash ADC with an 8-bit current-steering DAC (IDAC) for reference generation.

In the voltage-sensing SRAM (VS-SRAM), the PU/PD strength ratio is equalized (PU/PD = 1) to maximize the voltage gain at the bit line. In the charge-sensing SRAM (CS-SRAM), the sensing margin is enlarged by replacing conventional metal-oxide-metal (MOM) capacitors with MOS capacitors, providing larger unit capacitance within the same footprint. Each VS-SRAM and CS-SRAM bitcell occupies 1.02 μm² and 1.30 μm², respectively. The MRAM CIM macro consists of a 4×16 array and readout peripherals. The readout block comprises a 64-to-1 MUX, a 5-bit flash ADC, and a current-adder-based reference-current generator with a 4-bit configurable current summer enabling a broad dynamic range.

III. Results and Chip Specification

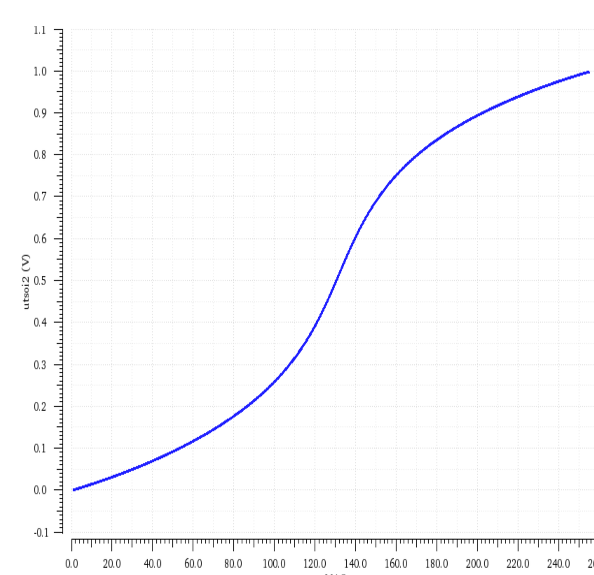

Fig. 3. Transfer curve of the VS-CIM macro.

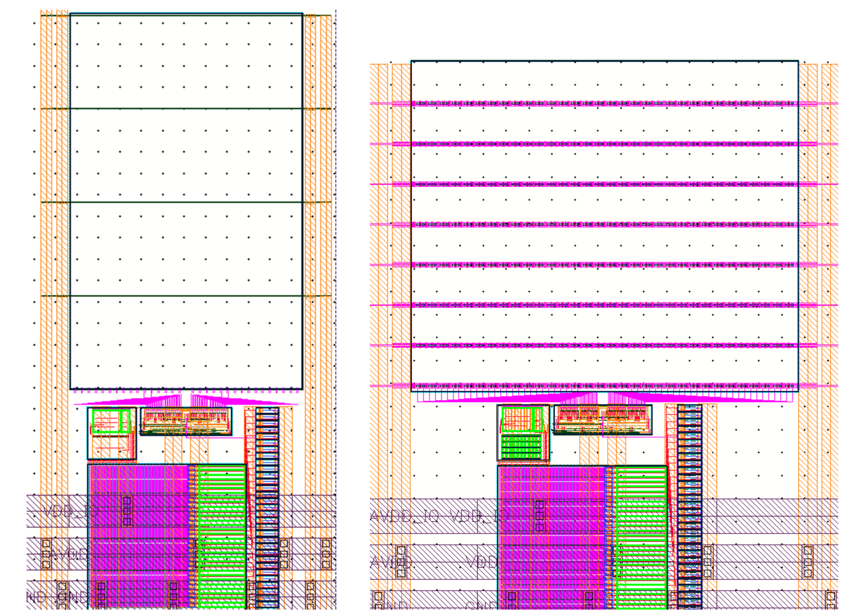

Fig. 4. Layout of the VS/CS-CIM macro.

Spectre transient simulations characterize the linearity of the VS-CIM macro. The VS-CIM resistive divider produces a non-linear response with a steeper mid-range slope. The flash-ADC reference grid is compressed in the steeper section and widened elsewhere, achieving uniform code density and lower DNL/INL. In contrast, the CS-CIM charge-redistribution path yields an almost linear transfer curve, hence uniformly spaced reference levels are used, simplifying calibration circuitry. The IDAC reference-current generator is a 4-bit current summer (LSB = 6 μA, range 6–96 μA). Biased at 48 μA, it delivers reference voltages from 0.3 V to 1.0 V with a 1.5 mV step size. The linear window of interest (300 mV to 700 mV) is fully covered, so the residual non-linearity at the extremes has no appreciable impact on MAC accuracy. The VS-CIM and CS-CIM macros occupy 0.038 mm² and 0.052 mm², respectively.

TABLE I. Chip specification table

	Hybrid-CIM macro
Technology	Samsung 28nm FD-SOI
Die Area	4 x 4 mm²
Main Frequency	100MHz
Bit-cell organization	12T / 8T-1C / 1T-1R
Input bit	1
Weight bit	1
Supply voltage	1 V

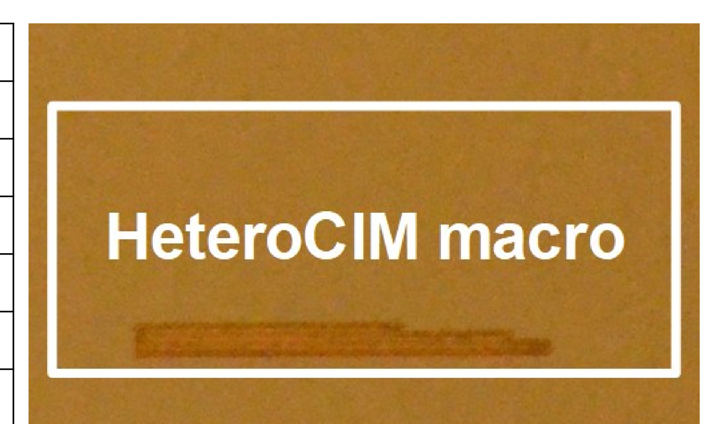

Fig. 5. Die photograph in the 28nm FD-SOI

The hybrid CIM architecture is implemented using Samsung 28nm FD-SOI process. It operates at a frequency of 100 MHz with a supply voltage of 0.7 V. The macro supports 1b precision MAC operation which can be extended using bit-serial-based multibit MAC operation.

IV. Conclusion

hybrid SRAM/MRAM CIM macro in Samsung 28nm FD-SOI integrates two 256×64 SRAM-CIM banks with a 4×16 MRAM-CIM sub-array, targeting both low-bit CNN and transformer inference. These results demonstrate the feasibility of FD-SOI-based hybrid CIM as a scalable edge-AI accelerator platform.