Gain-Cell CIM: Leakage and Bitline Swing Aware 2T1C Gain-Cell eDRAM Compute in Memory Design with Bitline Precharge DACs and Compact Schmitt Trigger ADCs

Shanshan Xie[1], Can Ni[1], Pulkit Jain[2], Fatih Hamzaoglu[2], and Jaydeep P. Kulkarni[1]
[1] The University of Texas, Austin, TX, [2] Intel Corporation, Hillsboro, OR, USA

Abstract: We present a leakage and read bitline (RBL) swing aware Compute-in-Memory (CIM) design leveraging a promising high-density gain-cell embedded DRAM bitcell and the intrinsic RBL capacitors to perform CIM computations within the limited RBL swing available in a 2T1C eDRAM. The CIM D/A converters (DAC) are realized intrinsically with variable RBL precharge voltage levels. A/D converters (ADC) are realized using Schmitt Triggers (ST) as compact and reconfigurable Flash comparators. A 65nm CMOS prototype achieves energy efficiency of 7.4-236 TOPS/W, 13.1-411 GOPS/mm^2 for the CIFAR-10 dataset with ResNet-20 and improves the defined FoM by 2.3-4.3X over prior CIM designs.

Motivation: Among the prevalent semiconductor memories, eDRAM (embedded-DRAM) offers a dense bitcell footprint, low pJ/bit access energy, and high-performance [1], making it a promising candidate for CIM designs [2]. However, due to the inherent destructive read operation in 1T1C (T=Transistor, C=Capacitor) eDRAM bitcells, this approach cannot perform in-situ multiply-accumulate-averaging (MAV) computations directly on local bitlines (BLs) [2], resulting in increased intra-memory data movement and energy overhead due to numerous read/write and data duplication operations. On the other hand, a dense 2T1C gain-cell eDRAM bitcell (Fig. 1), with decoupled read/write ports, can be a promising alternative for realizing an energy-efficient eDRAM CIM design using intrinsic RBL capacitors for computations and storage with improved data reuse. The key bottlenecks of using 2T1C eDRAM for BL computing for a CIM design are limited bitline voltage swing availability, unselected bitcells leakage, and the high area/energy cost of CIM data conversion steps. In this work, we propose a unique bitline-swing and leakage-aware MAV computation approach while minimizing data converter overheads in a 2T1C gain-cell eDRAM based CIM design.

Gain-cell eDRAM CIM Architecture and Design: Fig. 2 describes the overall computational data flow for the gain-cell eDRAM CIM design. In **step 1**, 2b inputs are fed to the BL precharge bias generator to discharge the RBLs within a certain input-dependent pulse duration to reach variable precharge voltage levels, and 1b weights are stored in the 2T1C eDRAM bitcells in a complementary form. The input-dependent RBL precharge voltage levels intrinsically achieve the D/A conversion. For weight-stationary computations, in **step 2**, a read operation is performed to multiply 2b inputs and 1b weights. When the weight is '0', the RBL is discharged. However, as the RBL starts to discharge towards V_{DD}-V_{TH} (threshold voltage of gain-cell read port transistor), the unselected bitcells on the same column which are storing '1's are weakly turned on, thus providing leakage sneak paths from V_{DD} (unselected RWL port) to RBL in the same column, as described in Fig. 3. Therefore, in **step 3**, a self-detect voltage clipper circuit (Fig. 3) is devised for mitigating the impact of RBL leakage current by utilizing the available sense amplifier (SA) for comparing the RBL with a predefined reference level (V_{REF}). If the V_{RBL} is lower than V_{REF} (corresponds to the weight being '0'), SA fires and brings the RBL back to V_{ZERO}, which is selected to be above V_{DD}-V_{TH} to mitigate the sneak path leakage current. As the RBL voltage is now higher than the V_{DD}-V_{TH} while computing the dot product with '0' weight bit, the leakage sneak path from unselected bitcells is cut off, thus eliminating the leakage-dependent RBL evaluation and improving the computation accuracy. In **step 4**, a charge share operation utilizing the intrinsic RBL capacitors is performed for accumulation and averaging. In **step 5**, the analog charge-shared voltage (V_{MAV}) is converted to a 2b digital output using Schmitt Trigger (ST) [3] based reconfigurable comparators for the ADC design. The ST-based flash ADC is developed to eliminate the need for extra reference generation circuits, to reduce the data converter area overhead, and to achieve fast ADC conversion. As shown in Fig. 4, ST-ADC is realized by embedding the reference voltage as the intrinsic ST switching threshold (V_{SW}). The input of the ST-ADC is reset to V_{SS} before conversion by asserting the discharge (DISCHAR) signal to make sure that ST always samples the input along a specific direction of the comparator hysteresis characteristics. Next, V_{MAV} is applied to the ST-ADC input through M_{P2}. Multiple stacked inverter devices (M_{P1}-M_{N1-2}), which are sized in a binary scaled fashion, are connected in parallel for configuration and calibration purposes. ST-ADC outputs (V_{OUT}) are resolved instantly depending on V_{MAV} value and corresponding V_{SW} level in each ST, resulting in fast data conversion within a half clock cycle. To mitigate the short circuit current in ST comparator, DISCHAR feedback is formed to reset $V_{IN\text{-}ADC}$ back to V_{SS} through M_{N4} right after the ST-ADC output is resolved completely. The entire data flow is performed in parallel for the remaining weight bits in the other gain-cell eDRAM arrays. 2b inputs are applied sequentially and the 2b digital results are grouped using a digital combiner for multibit MAV computation (Fig. 2).

Measurement Results: 65nm CMOS test-chip (Fig. 7) measurements for ST-ADC demonstrate less than 1LSB differential non-linearity (DNL) across spatial and temperature variations (Fig. 4). The measured Top-1 inference accuracy for CIFAR-10 dataset using ResNet-20 model drops by 0.86% compared to the INT8 software accuracy (Fig. 5a) due to non-ideality in BL precharge bias and ST-ADC. The 2b ST-ADC lowers the area by 1.9X and improves energy by 1.7X compared to a 2b in-eDRAM flash ADC design (Fig. 5b) which can generate the reference voltage using the write port of the 2T1C eDRAM gain-cell and the bit storage capacitors [2]. In comparison with prior works, the proposed 2T1C gain-cell eDRAM design improves the CIM FoM (input prec.*weight prec.*output prec.*energy efficiency) by 2.28X for 4bINx4bW and 2.26X for 8bINx8bW configuration (Fig. 5c). In summary, the proposed gain-cell eDRAM CIM approach lowers the data converter area/energy overheads and reduces intra-memory data movement while improving throughput, energy efficiency, and GOPS/mm^2 metrics (Fig. 6). It can pave a road for future CIM designs using high-density 2T1C eDRAM gain-cells.

Acknowledgments: This research is supported in parts by Intel rising star faculty award and Micron foundation faculty awards.

References:[1]C. Berry, ISSCC, pp54, 2020.[2]S. Xie, ISSCC,pp 248, 2021.[3]J. Kulkarni, JSSC, pp2304, 2007.[4]J. Su, ISSCC, pp 240, 2020.[5]X. Si, ISSCC, pp246, 2020.[6]A. Biswas, JSSC, pp217, 2019.

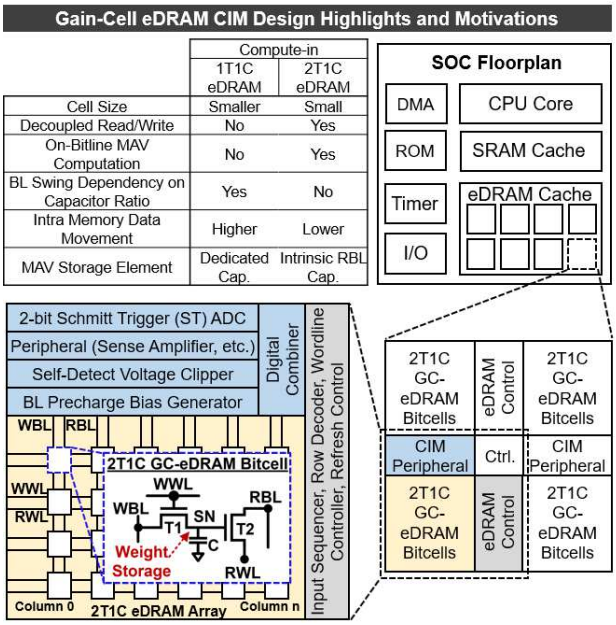

Fig. 1 Compute-in-eDRAM bitcell comparison, design highlights, array architecture, and 2T1C eDRAM bitcell circuit.

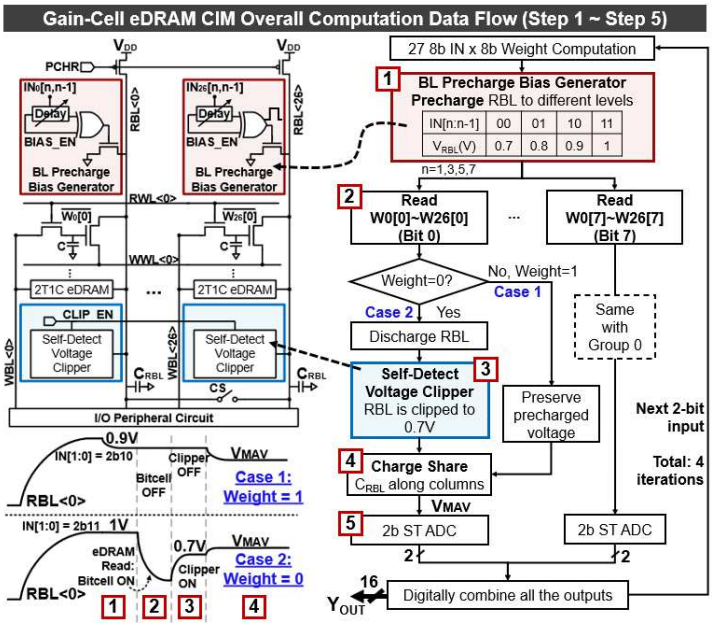

Fig. 2 Gain-cell eDRAM CIM circuit diagram, RBL waveforms for 2 cases (weight=1/0) and data flow for 27 MAVs (8b input x 8b weight) computation.

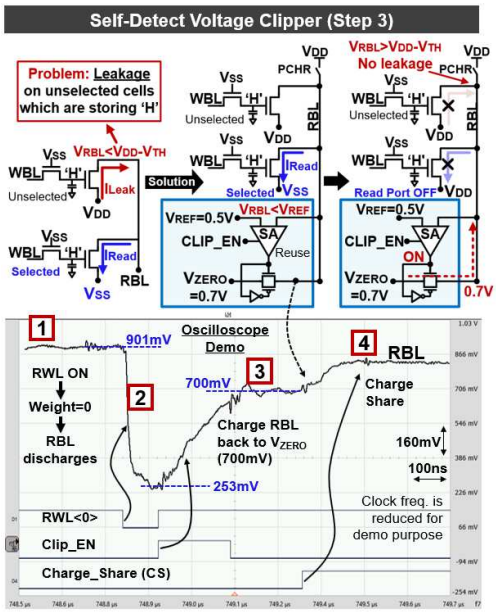

Fig. 3 Self-detect voltage clipper operation and RBL oscilloscope waveforms from step 1 to step 4.

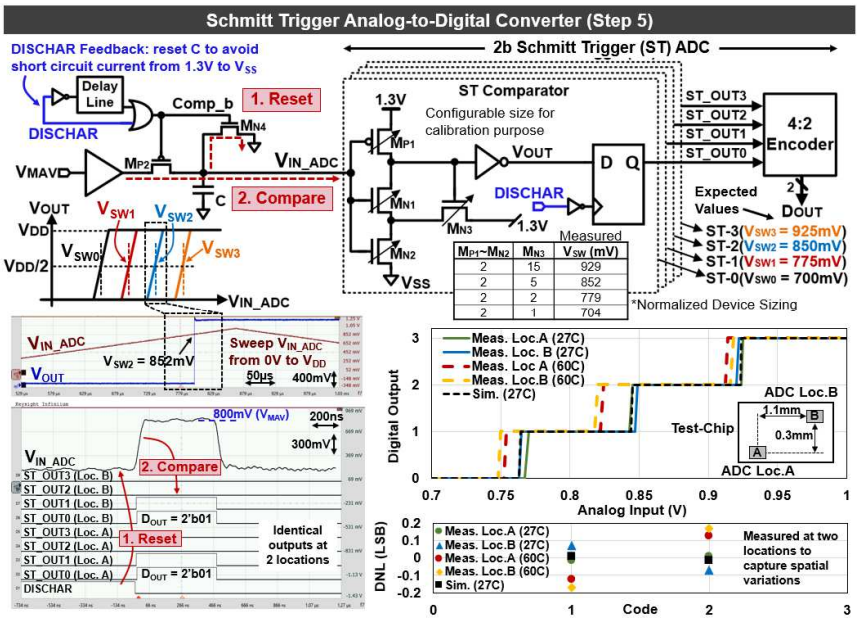

Fig. 4 ST-ADC circuit with oscilloscope waveforms showing ADC characteristics and DNL at two locations on the same die and two temperature conditions (27C and 60C).

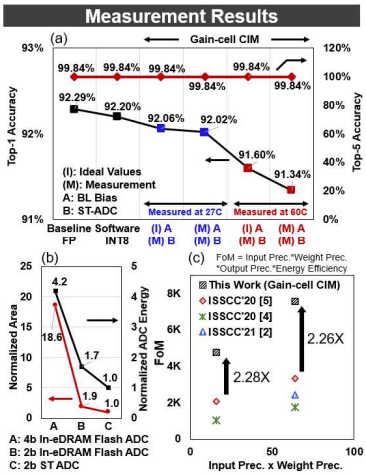

Fig. 5 (a)Meas. accuracy; (b) ADC energy & area; (c) FoM comparison.

Comparison to Prior Works

	This work	ISSCC'21 [2]	ISSCC'20 [4]	ISSCC'20 [5]
Technology	65nm	65nm	28nm	28nm
Single Level Cell Bitcell Topology	2T1C Gain-Cell eDRAM	1T1C eDRAM	6T SRAM	6T + Local Computing SRAM
Array Capacity	32Kb	16Kb	64Kb	64Kb
Input Precision (bit)	8	8	8	8
Weight Precision (bit)	8	8	8	8
Output Precision (bit)	16	8	20	20
Supply Voltage (V)	1	1~1.2	0.85~1.0	0.7~0.9
Dataset	CIFAR-10	CIFAR-10	CIFAR-10	CIFAR-10
Model	ResNet-20	CNN: 4 CONV + 2 Pooling + 2 FC	ResNet-20	ResNet-20
Measured Accuracy	92.02% (Top-1) 99.84% (Top-5) (5K-Img Tested)	80.1% (Top-1), 98.1 % (Top-5) (5K-Img Tested)	[1]91.91%	[1]92.02%
Throughput (GOPS)	22	4.71	N/A	N/A
Average Energy Efficiency (TOPS/W)	7.39	4.76	7.3	14.08
GOPS/mm²	13.1	8.26	N/A	N/A
[2]FoM (scaled to 65nm)	7567	2437	1728	3340

[1]Top-1 or Top-5 is not mentioned [2]FoM = input precision x weight precision x output precision x energy efficiency (scaled to 65nm, assume energy ∝ (Tech.)² [6]

Fig. 6 65nm CMOS prototype design metrics comparison with prior works

Testchip Summary

Design Details	
Technology	65nm CMOS
Supply Voltage (V)	1
Neural Network Parameters	
Input (IN) Precision	2b/4b/8b (Configurable)
Weight (W) Precision	1b~8b (Configurable)
Output Precision	2b~16b
Dataset	CIFAR-10
CNN Model	ResNet-20
Software Classification Accuracy (INT8) (Top-1, Top-5)	92.2%, 99.84%
Measured Inference Accuracy (5K Image) (Top-1, Top-5)	92.02%, 99.84% (27C) 91.34%, 99.84% (60C)
Measurement Results	
GOPS/mm²	2b IN x 1b W: 411.3 4b IN x 4b W: 51.2 8b IN x 8b W: 13.1
Access Time (ns)	2b IN x 1b W: 5 4b IN x 4b W: 10 8b IN x 8b W: 20
Throughput (GOPS) *1MAV = 2 operations	2b IN x 1b W: 691 4b IN x 4b W: 86 8b IN x 8b W: 22
Energy Efficiency (TOPS/W)	2b IN x 1b W: 236 4b IN x 4b W: 29.6 8b IN x 8b W: 7.39

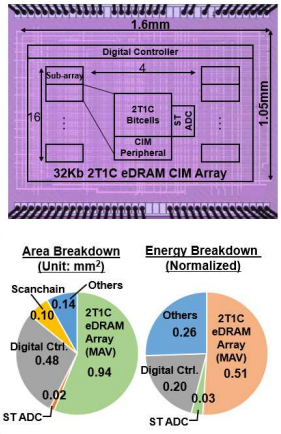

Fig. 7 Test-chip summary, die photo, area/energy split-up, design metrics for various bit precision operands.