31.5 A 65nm 4Kb Algorithm-Dependent Computing-in-Memory SRAM Unit-Macro with 2.3ns and 55.8TOPS/W Fully Parallel Product-Sum Operation for Binary DNN Edge Processors

Win-San Khwa[1,2], Jia-Jing Chen[1], Jia-Fang Li[1], Xin Si[3], En-Yu Yang[1], Xiaoyu Sun[4], Rui Liu[4], Pai-Yu Chen[4], Qiang Li[3], Shimeng Yu[4], Meng-Fan Chang[1]

[1]National Tsing Hua University, Hsinchu, Taiwan; [2]TSMC, Hsinchu, Taiwan
[3]University of Electronic Science and Technology of China, Sichuan, China
[4]Arizona State University, Tempe, AZ

For deep-neural-network (DNN) processors [1-4], the product-sum (PS) operation predominates the computational workload for both convolution (CNVL) and fully-connect (FCNL) neural-network (NN) layers. This hinders the adoption of DNN processors to on the edge artificial-intelligence (AI) devices, which require low-power, low-cost and fast inference. Binary DNNs [5-6] are used to reduce computation and hardware costs for AI edge devices; however, a memory bottleneck still remains. In Fig. 31.5.1 conventional PE arrays exploit parallelized computation, but suffer from inefficient single-row SRAM access to weights and intermediate data. Computing-in-memory (CIM) improves efficiency by enabling parallel computing, reducing memory accesses, and suppressing intermediate data. Nonetheless, three critical challenges remain (Fig. 31.5.2), particularly for FCNL. We overcome these problems by co-optimizing the circuits and the system. Recently, researches have been focusing on XNOR based binary-DNN structures [6]. Although they achieve a slightly higher accuracy, than other binary structures, they require a significant hardware cost (i.e. 8T-12T SRAM) to implement a CIM system. To further reduce the hardware cost, by using 6T SRAM to implement a CIM system, we employ binary DNN with 0/1-neuron and ±1-weight that was proposed in [7]. We implemented a 65nm 4Kb algorithm-dependent CIM-SRAM unit-macro and in-house binary DNN structure (focusing on FCNL with a simplified PE array), for cost-aware DNN AI edge processors. This resulted in the first binary-based CIM-SRAM macro with the fastest (2.3ns) PS operation, and the highest energy-efficiency (55.8TOPS/W) among reported CIM macros [3-4].

Figure 31.5.2 presents the CIM-SRAM unit macro. In inference operations input data (IN) is converted into multiple WL activations. The weights (W) of each $n\times n$ CNVL kernel or m-weight FCNL are stored in consecutive n^2/m cells on the same BL. When WL=IN=1, the read current (I_{MC}) of each activated memory-cell (MC) represents its input-weight-product (IN×W); the resulting BL voltage (V_{BL}) is the sum of IN×W. Unlike the BL-discharge approach in [3] and typical SRAM, we adopted a voltage-divider (VD) approach to read PS results. For a typical 6T CIM-SRAM the charge ($I_{MC\text{-}C}$) and discharge ($I_{MC\text{-}D}$) cell currents both develop on BL/BLB, since both pass-gates (PGL/PGR) are activated. A large number of WL activations (N_{WL}) result in high BL current ($I_{BL}= n^2(I_{MC\text{-}C}+ I_{MC\text{-}D})$ or $m(I_{MC\text{-}C}+ I_{MC\text{-}D})$). Since m (i.e. 64) is usually much larger than n^2 (i.e. 9 for a 3x3 kernel), the CIM-SRAM for FCNL faces more difficult challenges in circuit designs than that for CNVL. Thus, this work focuses on CIM-SRAM for FCNLs.

The second challenge is inefficient binary-PS result or winner detection in FCNL. The regular and last-layer in FCNLs use a reference voltage (V_{REF}) to identify the PS result (±1) or winner (V_{1ST}) from the second (V_{2ND}) candidates. In single-ended SRAM sensing, the V_{REF} for sense amplifiers (SA) is a fixed value. For However, simulations of PS result distributions for V_{1ST} and V_{2ND} using the MNIST database show that the ideal V_{REF} covers a wide range (>0.4V). Even with perfect SA, 5-to-6-sensing iterations are needed to approach the accuracy limit of the binary algorithm.

The third challenge is small voltage sensing margins (V_{SM}) across different PS results on FCNLs. Three techniques are proposed to overcome these difficulties: (1) algorithm-dependent asymmetric control (ADAC), (2) dynamic input-aware V_{REF} generation (DIARG), and (3) a common-mode-insensitive (CMI) small-offset voltage-mode sense amplifier (VSA).

Data pattern analysis of MNIST test images revealed an intriguing asymmetry between the number of IN×W=+1 (N_{+1}) and (IN×W)=-1 (N_{-1}) on a BL in the last two FCNLs (i.e. the q-1 and qth layers). This is a generic characteristic among various applications, because IN×W results in the last layer are already polarized to have a single candidate that is most probable. Also, the N_{+1}/N_{-1} asymmetry is opposed in the last two FCNLs. We used these characteristics to reduce I_{BL} and macro power consumption using an newly proposed ADAC scheme combining the previous split-WL DSC6T [8] cell. This allows for different WL/BL access modes for two layers using the same CIM-SRAM unit-macro.

The ADAC scheme (Fig. 31.5.3) comprises of an asymmetric-flag (AF), BL-selection switches (BLSW), WL-selection switches (WLSW), dual-path output-drivers (DPOD), BL-clamping (BLC) and DSC6T cells. AF can be pre-defined during training, or configured by an application, to specify whether to use WLL-BLL or WLR-BLR for sensing. It is determined by N_{+1} and N_{-1} of all the BLs in the macro (N_{+1_M} and N_{-1_M}). For ($N_{+1\_M}>N_{-1\_M}$), AF is asserted for WLL-BLL sensing. WLSW activates BLL sensing by asserting only the WLLs of the selected rows (IN=1), while all WLRs are grounded. Each BLL is connected to its corresponding VSA through BLSW, while BLR=VDD is isolated from the VSA. The VSA detects V_{BLL} and directs its output (SAOUT) through the non-inversion path of DPOD to DOUT. For AF=0 ($N_{+1}< N_{-1}$), WLR-BLR sensing is selected, the roles of WLR-BLR and WLL-BLL are switched, and the SAOUT is sent through the inversion path of DPOD to DOUT. Thus, ADAC+DSC6T consumes less I_{BL} and macro power is reduced compared to a typical 6T cell due to (a) less parasitic load on WLL/WLR (1T per cell), (b) less I_{BL} on the selected BL, and (c) no I_{BL} from unselected BL. In MNIST simulations, ADAC+DSC6T scheme consumed 61.4% less current than a conventional 6T SRAM. To avoid the write disturbance, we also employed BL-clamping (BLC), which prevents V_{BL} from dropping below the cell-write threshold voltage.

Figure 31.5.4 presents the DIARG scheme, where V_{REF} generation is based on N_{WL} for CNVL and FCNL modes of binary DNN. DIARG includes columns (RC1 and RC2) of fixed-zero (Q=0) reference-cells (F0RC), a BL-header (BLH), a WL-combiner (WLCB), a reference-WL-tuner (RWLT), and a replica BL-selection switch (RBLSW). WLCB combines the WLL/WLR of a regular array with a reference WL (RC1WL) for RC1, such that RC1WL is asserted when WLL=1 or WLR=1. The BLL and BLR of RC1 are shorted together; therefore, when N_{WL} rows are activated, RC1 always provides the V_{REF} resulting from $N_{WL}(I_{MC\text{-}D}+ I_{MC\text{-}C})$. The reference WL (RC2WL) for RC2 is controlled by RWLT to adjust I_{MC_D} and I_{MC_C} on BLL2/BLR2 for multiple-level or multi-iteration sensing. With RBLSW connecting BLL2/BLR2 to BLL1/BLR1, the required V_{REF} is a function of N_{WL}. In our example, RC1 alone (RC2 is decoupled) provides the required last-layer FCNL V_{REF} for MNIST applications. DIARG provides winner-detection accuracy of 97.3% in the first iteration, whereas conventional fixed-V_{REF} requires four iterations. This reduces latency and energy overhead by over 4×.

We use a CMI-VSA to tolerate a small BL signal margin (V_{SM}) against a wide V_{BL} common-mode ($V_{BL\text{-}CM}$) range across various PS results (Fig. 31.5.5). The CMI-VSA includes two cross-coupled inverters (INV-L, INV-R), two capacitors (C1, C2), and eight switches (SW1-SW8) for auto-zeroing and margin enhancement. The CMI-VSA provides a 2.5× improvement in offset over a conventional VSA and a constant sensing delay across the $V_{BL\text{-}CM}$ range. In standby mode, CMI-VSA latches the previous result. In phase-1 (PH1), control switches enable the two inverters (INV-L and INV-R) to auto-zero at their respective trigger points ($V_{TRP\text{-}L}$ and $V_{TRP\text{-}R}$), making the node voltages V_{INL} and V_{INR} equal to V_{BL} and V_{REF}. In PH2, (V_{BL}-V_{REF}) and (V_{REF}-V_{BL}) are respectively coupled to V_{CL} and V_{CR}. This ideally increases the difference in voltage (V_{INV}) between V_{CL} and V_{CR} to 2(V_{BL}-V_{REF}). In PH3, INV1 and INV2 amplify V_{INV} to generate full swing for V_{CL} and V_{CR}.

Figure 31.5.6 presents the measured results from a test chip with multiple 64×64 CIM-SRAM unit-macros, integrated with the last-two FCNLs and test-modes. For the last FCNLs, the macro access time ($t_{AC\text{-}M}$) is 2.3ns for MINST winner detection at V_{DD}=1V. For the integrated last-two FCN layers, the $t_{AC\text{-}M\text{-}2Layers}$=4.8ns for MNIST image identification. In a shmoo test, ADAC and CIM-VSA enabled support for V_{WL}=0.8V at V_{DD}=1V to suppress I_{BL}. The CIM macro achieved the fastest PS operations and the highest energy efficiency among CIM-SRAMs; i.e. over 4.17× faster than the previous CIM-SRAM [3]. Figure 31.5.7 presents the die photograph.

Acknowledgements:
The authors would like to thank TSMC-JDP, MTK-JDP, MOST-Taiwan for their support.

References:

[1] K. Bong, et al., "A 0.62mW Ultra-Low-Power Convolutional-Neural-Network Face-Recognition Processor and a CIS Integrated with Always-On Haar-Like Face Detector", *ISSCC*, pp. 344-346, 2017.
[2] M. Price, et al., "A Scalable Speech Recognizer with Deep-Neural-Network Acoustic Models and Voice-Activated Power Gating," *ISSCC*, pp. 244-245, 2017.
[3] J. Zhang, et al., "A Machine-learning Classifier Implemented in a Standard 6T SRAM Array," *IEEE Symp. VLSI Circuits*, 2016.
[4] F. Su, et al., "A 462GOPs/J RRAM-Based Nonvolatile Intelligent Processor for Energy Harvesting IoE System Featuring Nonvolatile Logics and Processing-In-Memory," *IEEE Symp. VLSI Circuits*, pp. 260-261, 2017.
[5] M. Courbariaux, et al., "Binarynet: Training deep neural networks with weights and activations constrained to +1 or-1," *ArXiv: 1602.02830,* 2016, https://arxiv.org/abs/1602.02830.
[6] M. Rastegari, et al., "XNOR-net: ImageNet classification using binary convolutional neural networks," *ArXiv: 1603.05279,* 2016, https://arxiv.org/abs/1603.05279.
[7] M. Kim. et al., "Bitwise neural networks," *Int. Conf. on Machine Learning Workshop on Resource-Efficient Machine Learning*, 2015.
[8] M.-F. Chang, et al., "A 28nm 256Kb 6T-SRAM with 280mV Improvement in VMIN Using a Dual-Split-Control Assist Schemem" *ISSCC*, pp. 314-315, 2015.

 978-1-5090-4940-0/18/$31.00 ©2018 IEEE

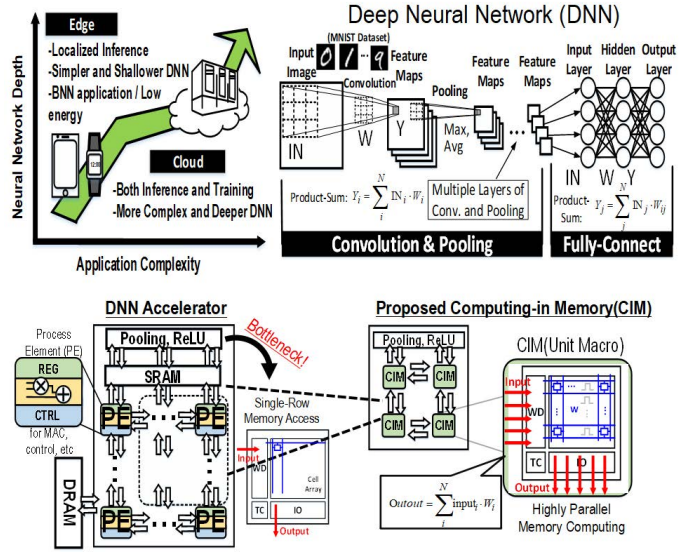


Figure 31.5.1: CIM-SRAM for AI edge processors.

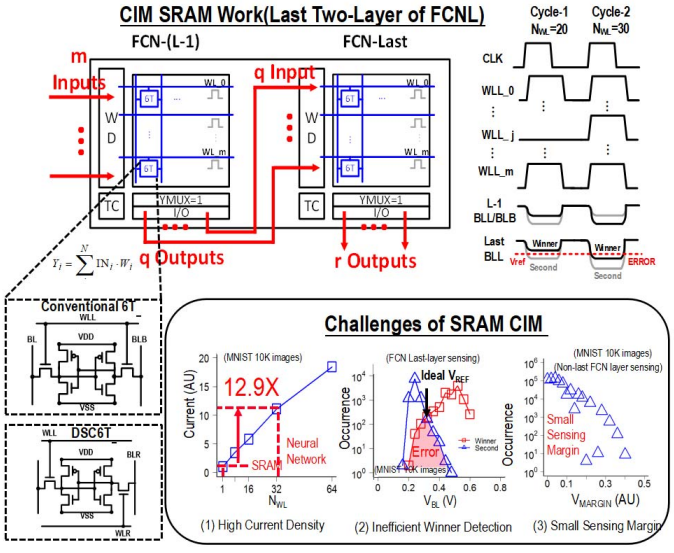


Figure 31.5.2: Proposed CIM-SRAM structure and challenges.

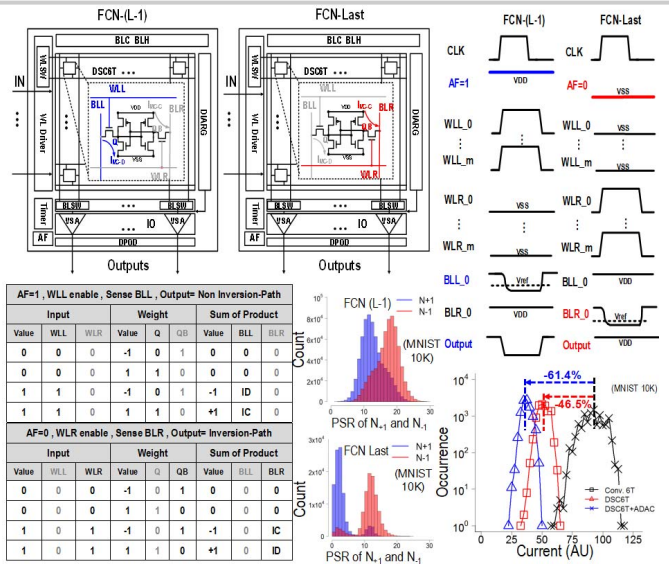


Figure 31.5.3: Algorithm-dependent asymmetric control (ADAC) scheme.

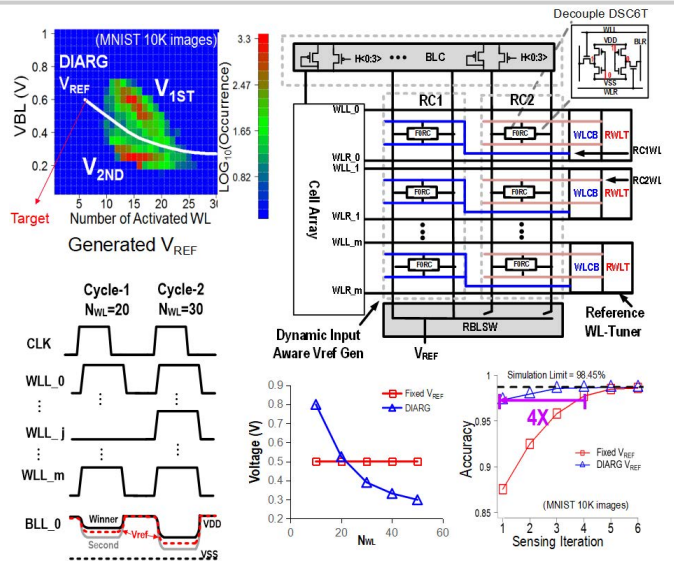


Figure 31.5.4: Dynamic input-aware V_{REF} generation (DIARG) scheme.

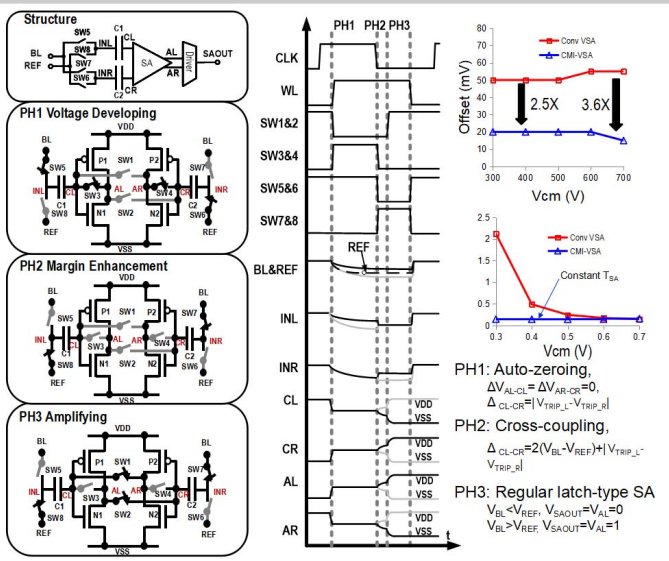


Figure 31.5.5: Operation of CMI-VSA.

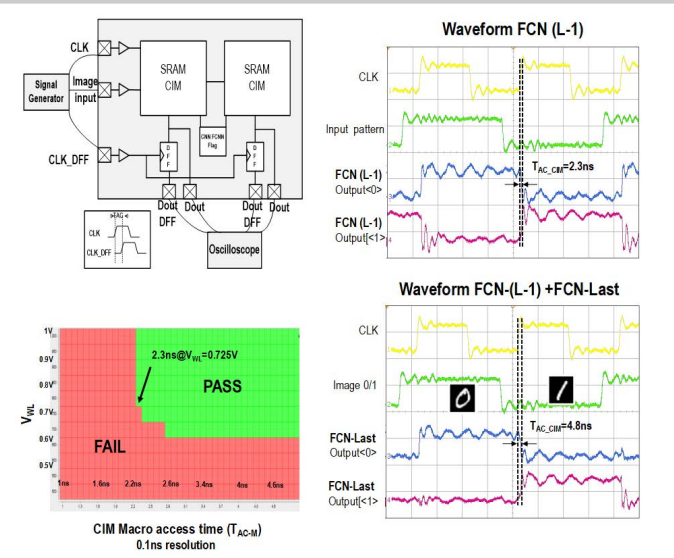
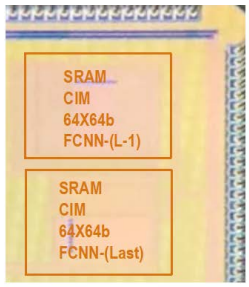
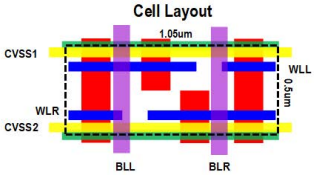


Figure 31.5.6: Measured results.



Cell Layout



CHIP SUMMARY	
Technology	65nm CMOS
Unit-Marco Size	4Kb(64x64b)
Bit cell Size	0.5μm X 1.05μm (modified compact cell)
Simulated Accuracy (MNIST)	97.5%
Inference Speed	2.3ns per layer
Measured Energy (CIM one layer)	73.4pJ
Energy efficiency (CIM one layer)	55.8TOPS/W*

*1 operation =1 multiply-count operation. Generally, if multiply-count is counted as separate operations, the energy efficiency could be 111.6 (55.8×2) TOPS/W

Figure 31.5.7: Die photo and summary table.

978-1-5090-4940-0/18/$31.00 ©2018 IEEE