7.6 A 70.85–86.27TOPS/W PVT-Insensitive 8b Word-Wise ACIM with Post-Processing Relaxation

Sung-En Hsieh[1], Chun-Hao Wei[1], Cheng-Xin Xue[1], Hung-Wei Lin[1], Wei-Hsuan Tu[1], En-Jui Chang[1], Kai-Taing Yang[1], Po-Heng Chen[1], Wei-Nan Liao[1], Li Lian Low[2], Chia-Da Lee[1], Allen-CL Lu[1], Jenwei Liang[1], Chih-Chung Cheng[1], Tzung-Hung Kang[1]

[1]MediaTek, Hsinchu, Taiwan
[2]MediaTek, Singapore, Singapore

Tiny-machine learning (TinyML) and artificial intelligence-of-things (AIoT) present new opportunities for machine-intelligent applications with stringent energy constraints. To conserve system energy, high-power devices stay dormant and are woken up only when an event is detected by a low-power always-on detector: which can be implemented by analog compute in memory (ACIM). However, there is a tradeoff between the inference accuracy and analog nonideality for ACIM designs, limiting the applicable AIoT applications. Current-domain [1] and time-domain [2] MACs show attractive energy efficiency since the area and parasitic capacitance are minimized from the compactly arranged unit cells. However, due to the large mismatch and nonlinearity of the unit cell, overall linearity is limited and is sensitive to PVT. The charge-sharing MAC [3] achieves PVT-insensitive linear 1bIN 1bW accumulation. However, expanding the IN and W precision from multiple 1bIN 1bW MACs and digital bit shifting which is also known as bit-wise causes large INL and noise discontinuities, due to a lack of inter-MSB/LSB conversion's error correction margin. Besides, energy consumption is multiplied by reusing the 1b hardware. To overcome random noise, this work implements an error desensitization algorithm by training the VWW detector with the ACIM noise model. When the noise-aware error desensitization algorithm is used in a visual-wake-word (VWW) detector, the noise requirement for ACIM is relaxed by 2×, while still achieving the same inference accuracy. To overcome static gain, offset, and linearity a PVT-insensitive 8b word-wise ACIM is proposed. Compared to a conventional bitwise 1bIN 1bW MAC design, the proposed binary-weighted data-selection (BWDS) MAC and segmentation buffer (SB) achieve 8b analog signal's convolution and 8b digital-to-analog (D2A) operation, respectively, without digital bit shifting. This achieves a 70.85 – 86.27TOPS/W energy efficiency and >10b linearity. The 10b linearity and self-calibration circuits allow for single pre-silicon training, using an ideal (IN × W = OUT) transfer curve assumption without software post-processing, as shown in Fig. 7.6.1.

Figure 7.6.1 shows the architecture of the proposed ACIM, which consists of 64 D2A converters, an SRAM, a BWDS MAC, 4:2 multiplexers, analog adders, a pseudo-differential (PD) 8b SAR ADC, and a calibration engine. Since 4–7b and 8b systems show moderate and negligible inference accuracy degradation in circuit and algorithm co-simulation, respectively, the 8bIN 8bW system is selected; however, a 4b MAC is still supported for scalability. With the proposed SB, 64 D2As directly convert 64 8b $D_{ink}[7:0]$ to 64 analog voltage (V_{8bink}) per each 8bIN 8bW cycle. Dynamic energy is reduced by using a one-time D2A conversion. This reduces the activation frequency and the analog swing of the MAC, compared to the digital bit shifting. The number of macros is doubled to fit algorithm requirements, achieving a twofold data throughput without compromising energy efficiency. One macro consists of 8 banks, and each bank has 4 binary-weighted capacitor arrays, but only 2 capacitor arrays are enabled depending on the selected weight address. To optimize mismatch and energy, a total of 240fF capacitance is selected for each capacitor array. Due to the large area of 240fF, each capacitor array (64× 8/4/2/1C) is shared by 256×8 SRAM cells (8× 64× 4b-w_k) to optimize area and parasitic routing RC. The capacitor array connects to the voltage input (V_{8bink}) or DC voltage, selected by the activated weight (64× 4b-w_k) in SRAM (MSB for 8C; MSB-3 for 1C). By connecting the top plate of 64 8/4/2/1C, 2 accumulations of 2 8bIN 4bW ($w_{64\sim1}[7:4]$ and $w_{64\sim1}[3:0]$) are separately computed from the 2 capacitor array's top plates, which is linear and PVT insensitive from the charge sharing characteristic. With the analog adder, 2 8bIN 4bW are added with the 16:1 weighting to recover the voltage domain 8bIN 8bW MAC. Without the digital bit shift, the analog charge-domain summing shows high accuracy for gain, offset, and linearity. At the end of the ACIM 8bIN 8bW cycle, 8b D_{OUT} is retrieved from the 8b ADC. Compared to the full output precision, the 8b output resolution shows negligible accuracy degradation (<0.5%) and an optimized energy efficiency based on the circuit/algorithm co-simulation and effective-number-of-bit spice simulation, respectively.

Figure 7.6.2 shows the proposed 8b SB. One SB has one push-pull buffer and 16/1C performing the signal driving and voltage addition, respectively. In the reset phase, the bottom and top plates of 16/1C sample the V_{CM} and push-pull V_{BIAS}, respectively. A PVT track replica shared by 64 D2As generates V_{BIAS} without the driving requirement due to the no switching energy during the reset phase. In the compute phase, 16C and 1C bottom plates are connected to $D_{ink}[7:4] \times V_{REF}/16$ and $D_{ink}[3:0] \times V_{REF}/16$, respectively. The voltage step ($V_{REF}/16$) is generated from a shared 4b resistive-DAC (RDAC). At the end of the compute phase, the converted voltage ($V_{8bink} = D_{ink}[7:0] \times V_{REF}/(16 \times 17)$) is buffered into the SRAM array with an >10b of linearity for the worst FF 125°C corner. By the 16× reduction of the global V_{REF} routing (from 8b 256 nodes to 16 nodes), the area is 16× smaller. Without the multi-conversions of MSB/LSB parts and digital bit shifting, the clock complexity, gain, offset, linearity, and system energy are also dramatically improved.

8bIN 8bW is finished by 2 8bIN 4bW MAC for this work, Figure. 7.6.3 shows a proposed 8bIN 4bW BWDS MAC example. An 8/4/2/1C set for IN W multiplication is shared by 8 4bW, while only one 4bW is active, enabled by the signal W_ADD_EN1. The 8/4/2/1C bottom plates are orderly controlled by 4 SRAM cells ($w_k[3:0]$). Each computing bitcell is comprised of 6T SRAM cell, 2T weight switches that enable or disable the V_{8bink} for the 8/4/2/1C, and 1T cell switch which achieves eightfold energy efficiency by avoiding the unneeded toggle of parasitic C_{par_cell} from disabled bit cells. For 8/4/2/1C, 8 global switches also achieve an eightfold improvement on the parasitic C_{par_in} dynamic energy. After 64 8/4/2/1C top plates are connected, the 8bIN 4bW convolution is achieved with low complexity and eightfold energy efficiency. By enabling 2 capacitor arrays (64 8/4/2/1C) for MSB ($V_{8bink} \times w_k[7:4]$) and LSB ($V_{8bink} \times w_k[3:0]$) parts, the final 8bIN 8bW output is computed with the analog summation of the proposed SB, as shown in Fig. 7.6.4. The share of 64 8bIN D2A buffers for multiple MAC further improves the energy efficiency and complexity.

Figure 7.6.4 shows the proposed PD 8b SAR ADC. At the ADC sampling phase, double plate sampling is applied for relaxing the noise by 2×. For the ADC conversion phase, the PD switching improves PSRR for low voltage operation (0.48V for all PVT cases). At the end of the SAR conversion, a compact, linear, PVT-insensitive, and energy-efficient 8bIN 8bW 8bOUT MAC is achieved.

Figure 7.6.5 shows the error cancellation schemes. The calibration engine infrequently aligns the starting and ending points of the linear transfer curve by tuning the ADC and RDAC with negligible energy overhead. The auto-zero technique stores the offset of 64 D2A at the reset phase. Therefore, the ideal transfer curve assumption of the single pre-silicon training is achieved for the post-processing relaxation. By triggering the worst INL test pattern (swap 10000000 and 0111111), the measured INL is smaller than 0.52LSB and shows a random characteristic without INL discontinuity.

A prototype chip was fabricated in a 12nm FinFET technology. Feature summaries and measurement results are shown in Fig. 7.6.6 and 7.6.7. The proposed VWW detector of this work uses MobileNet-v1-0.25 with all layers using ACIM for VWW classification, and only single pre-silicon training is performed without software post-processing. By the co-simulation of the circuit and algorithm, the 18% accuracy drop is monitored with 0.2LSB of gain, offset, and linearity error. For the typical environment, measured accuracy is 80%. With the actual applicated environment, input pictures are real-time classified by a complete system, and a stable 78.5 – 81% of accuracy is measured for P (2× samples), V (±40mV), and T (25/60°C) cases. This verifies <±0.09LSB of gain, offset, and linearity control ability from the proposed linear 8b word-wise ACIM and calibration engine. The supply voltage is 0.85V (DAC) and 0.53V (ADC). With the proposed energy-efficient 8bIN 8bW ACIM, the 70.85 and 86.27TOPS/W (1.806 and 1.48pJ per 64×2 OP) are measured with 0% and 90% of sparsity, respectively, including always-on ADC, SB, BWDS, SRAM, clock buffer, bias, RDAC, calibration engine, and reference buffers. Overall, a low-energy, PVT-insensitive, and linear ACIM is achieved without INL discontinuity and software post-processing.

References:

[1] Q. Dong et al., "A 351TOPS/W and 372.4GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine-Learning Applications," *ISSCC*, pp. 242-244, 2020.
[2] P.-C. Wu et al., "A 28nm 1Mb Time-Domain Computing-in-Memory 6T-SRAM Macro with a 6.6ns Latency, 1241GOPS and 37.01TOPS/W for 8b-MAC Operations for Edge-AI Devices," *ISSCC*, pp. 190-192, 2022.
[3] H. Jia et al., "A Programmable Neural-Network Inference Accelerator Based on Scalable In-Memory Computing," *ISSCC*, pp. 236-238, 2021.
[4] B. Yan et al., "A 1.041-Mb/mm2 27.38-TOPS/W Signed-INT8 Dynamic-Logic-Based ADC-less SRAM Compute-in-Memory Macro in 28nm with Reconfigurable Bitwise Operation for AI and Embedded Applications," *ISSCC*, pp. 188-190, 2022.

2023 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-6654-9016-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISSCC42615.2023.10067335

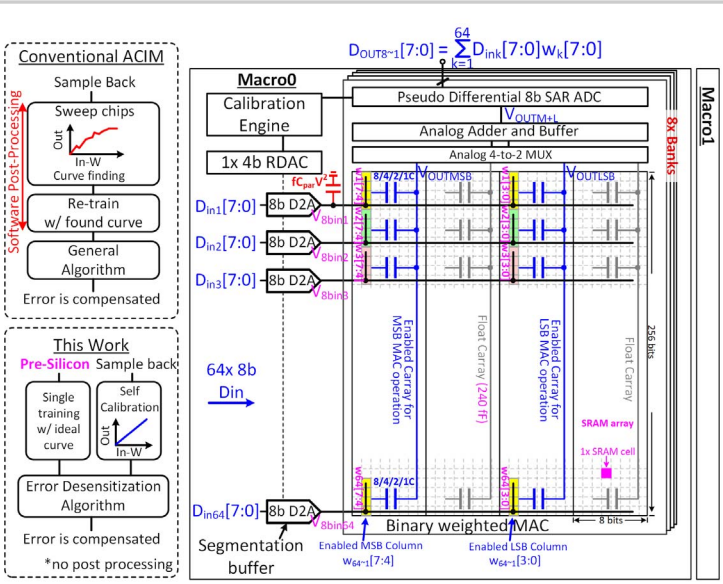


Figure 7.6.1: Design flow and architecture of the proposed binary-weight ACIM.

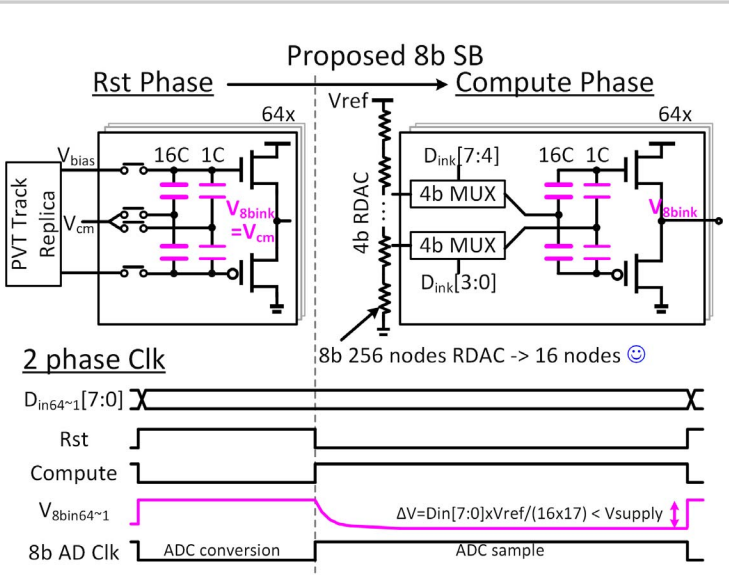


Figure 7.6.2: The proposed segmentation buffer and 2-phase clock.

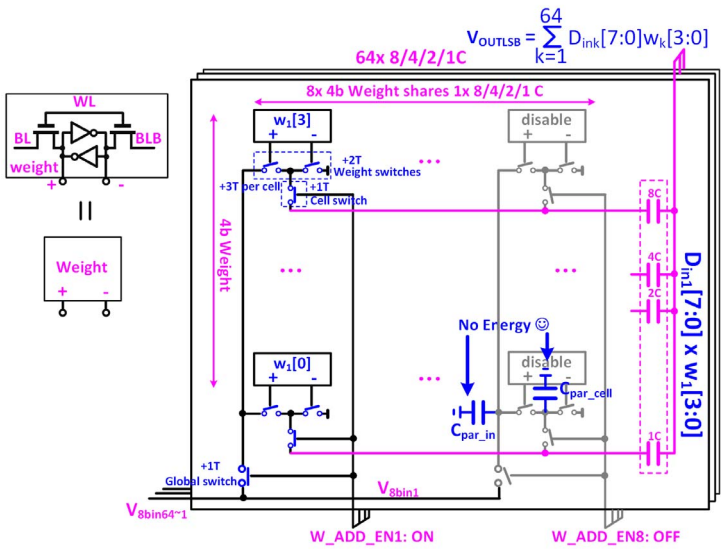


Figure 7.6.3: A 8bIN 4bW BWDS MAC operation with energy-reduction switches in an enabled capacitor array.

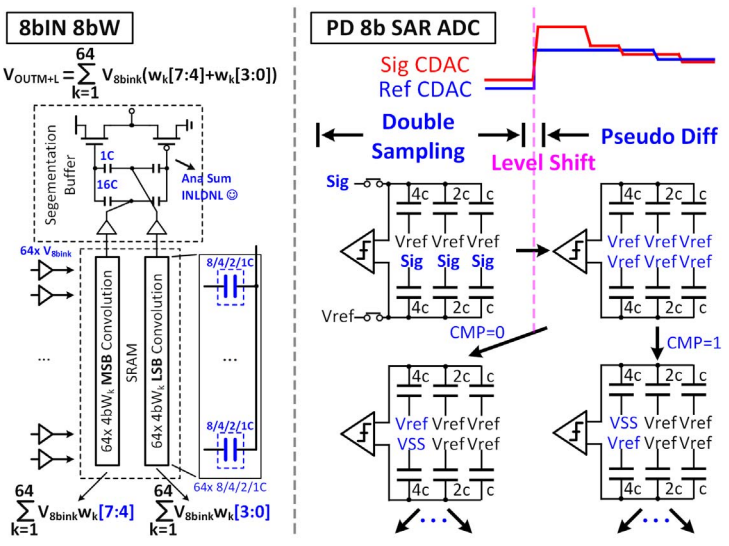


Figure 7.6.4: 8bIN 8bW operation with the output segmentation buffer and the pseudo-differential SAR ADC.

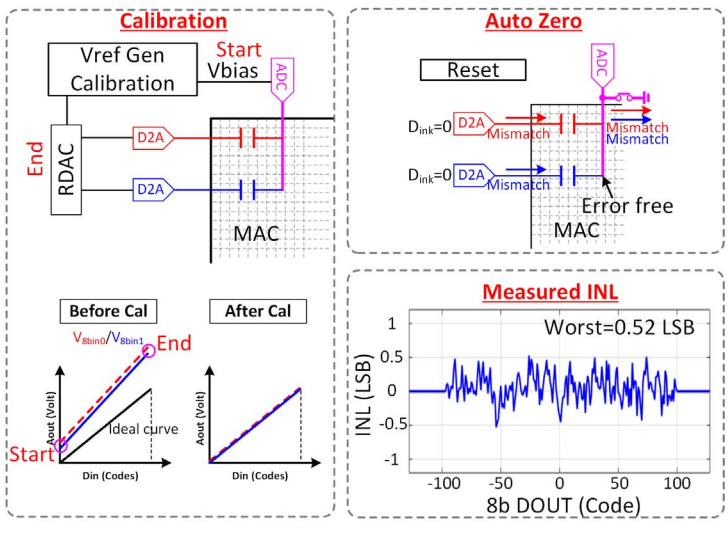
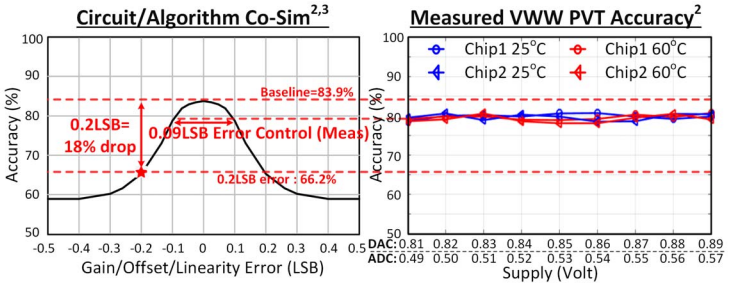


Figure 7.6.5: Error cancellation techniques and measured INL.



	This Work	
Dataset	CIFAR100	VWW[1]
Model	ResNet-20[5]	MobileNet-v1-0.25
Baseline Accuracy	68.8%	83.9%
Accuracy	67.9%	80.0%[2,4]
PVT Test	Yes	

[1]CIFAR-100 : a static single object Maximized in a box ; VWW : objects in complex everyday scenes (commercialization)
[2]All CIM Layers using ACIM, excluding depthwise layer (no software post-processing for all layers)
[3]Spice: monte-carlo, PVT linearity, Thermal noise fmax=100GHz
[4]10x averaged accuracy for benchmark (filter peak/optimistic data)
[5]https://github.com/chenyaofo/pytorch-cifar-models

Figure 7.6.6: Accuracy co-simulation and PVT measurement.

	ISSCC-21 [3]	ISSCC-22 [4]	ISSCC-22 [2]	This Work
Technology	16 nm	28 nm	28 nm	12 nm
Memory type	SRAM	SRAM	SRAM	SRAM
Macro size	4.5 MB	32 kb	1 Mb	128 kb
Input precision (bit)	8	8	8	8
Weight precision (bit)	8	8	8	8
Number of input channels	N/A	32	64	64
Output precision (bit)	N/A	21	22	8
Supply voltage (V)	0.8	0.8	0.65	0.5/0.85
INL[1] (LSB[2])	<1*	N/A	N/A	<0.52
Throughput (GOPS)	2950*	N/A	1050	102.4
Energy efficient (TOPS/W) (Sparsity)	30.25* (N/A)	27.38 (N/A)	27.75~37.01 (50~90%)	**70.85~86.27 (0~90%)**

[1]INL=(measured transfer curve-ideal transfer curve)/LSB, INL ↓ =linearity ↑
[2] VLSB ↓ with bit ↑
*estimate or normalize to 8b

Figure 7.6.7: Chip micrograph (90° rotation) and comparison table.

978-1-6654-9016-0/23/$31.00 ©2023 IEEE