16.6 A Nonvolatile AI-Edge Processor with 4MB SLC-MLC Hybrid-Mode ReRAM Compute-in-Memory Macro and 51.4-251TOPS/W

Wei-Hsing Huang*[1], Tai-Hao Wen*[1,2], Je-Min Hung*[1], Win-San Khwa*[2], Yun-Chen Lo[1], Chuan-Jia Jhang[1,2], Hung-Hsi Hsu[1], Yu-Hsiang Chin[1], Yu-Chiao Chen[1], Chung-Chuan Lo[1], Ren-Shuo Liu[1], Kea-Tiong Tang[1], Chih-Cheng Hsieh[1], Yu-Der Chih[3], Tsung-Yung Chang[3], Meng-Fan Chang[1,2]

[1]National Tsing Hua University, Hsinchu, Taiwan
[2]TSMC Corporate Research, Hsinchu, Taiwan, [3]TSMC, Hsinchu, Taiwan
*Equally Credited Authors (ECAs)

Low-power AI edge devices should provide short-latency (T_{WK-RP}) and low-energy (E_{WK-RP}) wakeup responses from power-off mode to handle event-triggered computing tasks with high inference accuracy (IA), which requires high-capacity nonvolatile memory (NVM) to store high-precision weight data in power-off and high bit-precision multiply-and-accumulate (MAC) operations with high energy efficiency. SRAM computing-in-memory (CIM) and digital processors suffer large E_{WK-RP} and long T_{WK-RP} due to the movement of weight data from off-chip NVM to the on-chip buffer and processing unit after wakeup. Thus, on-chip nonvolatile CIM (nvCIM) is preferred for AI-edge processors by combining NVM storage and computing tasks on the same macro. Among nvCIM structures, in-memory compute (IMC) [1] provides short computing latency and high energy efficiency, but suffers from low computing yield. Near-memory compute (NMC) [2-4] provides high computing yield, but suffers long computing latency and low energy efficiency.

Figure 16.6.1 shows the challenges hindering existing AI-edge processors, including (1) lack of a nvCIM-friendly computing flow and chip architectures for sufficient inference accuracy and energy efficiency, (2) a tradeoff between single/multi-level-cell (SLC/MLC) NVM devices vs. process variations, computing yield and area overhead, (3) long computing latency and low energy efficiency imposed by wordwise-input-sparsity schemes and a uniform number of accumulations (N_{ACCU}) per cycle, and (4) small signal margin (SM) between neighboring MAC values (MACV) and large bitline current (I_{BL}) when N_{ACCU} is high.

This work addresses these challenges by developing a hybrid-mode ReRAM nvCIM macro (hmRe-nvCIM), a nvCIM-friendly chip structure and corresponding data flow. This hmRe-nvCIM macro includes (a) configurable circuits supporting both NMC and IMC modes within a macro, (b) storage modes configurable for SLC, SLC-MLC-hybrid and MLC modes to manage storage density vs. computing yield, (c) a dynamic-accumulation-aware current quantization (DACQ) to suppress I_{BL} fluctuations and energy consumption in the ReRAM cell-array, and (d) a current-voltage-hybrid analog-to-digital converter (CVH-ADC) to enlarge SM and computing yield in IMC modes. The proposed processor includes (a) an nvCIM-friendly dataflow and architecture for the efficient utilization of hmRe-nvCIM across NMC-IMC/SLC-MLC modes to balance NVM capacity, energy efficiency and compute accuracy across multiple NN-layers, and (b) a bitwise input-sparsity and place-value aware dynamic accumulation (BIS-PVA-DA) scheme to shorten computing latency, increase energy efficiency and computing yield. Among reported nonvolatile AI-edge processors, the proposed 22nm AI-edge processor with 4MB MLC ReRAM-nvCIM achieved the highest energy efficiency (51.4-251TOPS/W) across datasets/NN-models/precisions with the shortest T_{WK-RP} (472.7μs for one-shot CIFAR-100 inference).

Figure 16.6.2 shows the dataflow and architecture of the proposed processor, which integrates 4MB on-chip hmRe-nvCIM macros with DACQ, CVH-ADC and 4 configurable adders (CA), a multimode nvCIM engine controller (mmCIM-EC), 4 BIS-PVA-DA units, a 512KB activation SRAM buffer and other neural network functions (e.g., a global output integrator, ReLU, pooling, etc.). The use of nvCIM eliminates the need for a weight SRAM buffer. The 4 hmRe-nvCIM macros comprise 32 sub-banks with 1024 8b-MAC readout channels (up to 24b output precision / channel), each of which can be activated separately according to the workload of the NN-layers. The mmCIM-EC and hmRe-nvCIM employ 6 modes: (1) NMC using 8 SLC (NMC-SLC); (2) NMC using 4 MLC (NMC-MLC); (3) IMC using 8 SLC (IMC-SLC); (4) IMC using 4 SLC and 2 MLC (IMC-Hybrid-I); (5) IMC using 2 SLC and 3 MLC (IMC-Hybrid-II); and (6) IMC using 4 MLC (IMC-MLC). Users can select among the six modes to implement different layers in the NN model to balance storage density, computing yield and energy efficiency. IMC modes are used for NN-layers that are less sensitive to computing yield, but require massive numbers of MAC operations with higher energy efficiency and throughput. NMC modes are used for NN-layers that require ultra-high computing yield, such as the first few and fully-connected layers.

Figure 16.6.3 presents the weight mapping strategy (DNNs or workloads) for a multi-array chip and the proposed BIS-PVA-DA, which is meant to shorten computing latency and improve energy efficiency for hmRe-nvCIM without compromising inference accuracy. The BIS-PVA-DA slices incoming 8b wordwise inputs into 8 bitwise input-groups based on their place values (BW-IN[7], …, and BW-IN[0]). For each group, BIS-PVA-DA rearranges the bitstream sequence by grouping the "bit-1" into multiple bit-1 sub-groups (INSG1) and a "bit-1-0" hybrid group (INSG10), and grouping all "bit-0" into a bit-0 group (INSG0). Each INSG1 varies in the number of N_{ACCU} (16, 32, 64, or 128), and INSG1 in a higher place (significance) value group (e.g., BW-IN[7]) has fewer N_{ACCU} to maintain high-accuracy MAC operations, while INSG1 in a lower place value group (e.g., BW-IN[0]) has more N_{ACCU} for high-throughput MAC operations with minimal impact on inference accuracy. The INSG0 groups will not be input to the hmRe-nvCIM to shorten computing latency and improve energy efficiency. For instance, an NN-layer with 512 accumulations using N_{ACCU}=16 for BW-IN[7] with 140 "1s", the INSG group has 8×INSG1, 1×INSG10, and 1×INSG0. Each INSG1 of BW-IN[7] has 16 bit-1, the INSG10 has 12 bit-1 and 4 bit-0, and the INSG0 has 368 bit-0. The BIS-PVA-DA can also improve computing yield by suppressing pattern (input × weight)-dependent variations. For example, a conventional nvCIM performs a MAC value (MACV) =3 operation with 16 unknown bitwise inputs, the BL current (I_{BL}) ranges from $3{\times}I_{LRS}+0{\times}I_{HRS}$ (3WLs on) to $3{\times}I_{LRS}+13{\times}I_{HRS}$ (16WLs on), resulting in a small SM with neighboring MACVs (4 and 2) due to non-negligible I_{HRS}. In BIS-PVA-DA, the I_{BL} for MACV=3 of INSG1s is fixed at $3{\times}I_{LRS}+13{\times}I_{HRS}$.

Figure 16.6.4 presents a sub-bank (512 rows × 1024 columns) in hmRe-nvCIM, including DACQ (for pre-quantizing I_{BL} and suppressing I_{BL} fluctuations and energy consumption) and CVH-ADC (for enhancing SM and computing yield). In MAC operations, the wide range of N_{ACCU} results in large fluctuations in I_{BL}. DACQ stabilizes the maximum I_{BL} (I_{BL-MAX}) by regulating the BL voltage (V_{BL}) according to N_{ACCU}. For example, when N_{ACCU}=16, each BL is biased at V_{BL-16} leading to $I_{BL-MAX} = 16{\times}I_{LRS}$. When N_{ACCU}=32, V_{BL}=0.5×V_{BL-16} and I_{BL-MAX} is also at 16×I_{LRS} (=32×0.5×I_{LRS}). Using a 5b CVH-ADC to convert I_{BL} to a digital partial MAC value (pMAC) can lead to various quantization cases. Cases where N_{ACCU}=16 and N_{ACCU}=32 both have full-precision readout, while the SM in N_{ACCU}=16 mode is double that of the N_{ACCU}=32 mode to increase computing yield. In cases where N_{ACCU}=64 and N_{ACCU}=128, pMAC is quantized linearly to 5b output. The CVH-ADC comprises a hybrid-mode SA (HMSA), a current mirror, an MSB-aware I-to-V converter and an I_{MSB} generator. The 5b readout operation of the CVH-ADC has three phases (P1-P3). In P1, SW3-4 are on to switch the HMSA into current mode for higher accuracy readout. The SAEN toggles twice and HMSA converts I_{BL} into 2b digital pMACVs [4:3] (MSBs). In P2, the MSB-aware I-to-V converter generates LSB current (I_{LSB}) as I_{BL}-I_{MSB} (I_{MSB} = current amplitude of pMACV [4:3]) and then converts I_{LSB} into V_{LSB} by charging C_{LSB}. The maximum I_{LSB} is 1/4× of the I_{BL-MAX}; therefore, the voltage difference of 8 pMACV [2:0] states generated by I_{LSB} charging C_{LSB} is 4× larger than it would be if using I_{BL} directly to charge C_{LSB}, which results in a 4× SM enhancement of V_{LSB} for given V_{DD} headroom. In P3, SW1-2 are on to switch the HMSA into voltage mode to improve the speed, the SAEN toggles three times, and HMSA converts V_{LSB} into 3b digital pMACV[2:0] (LSBs).

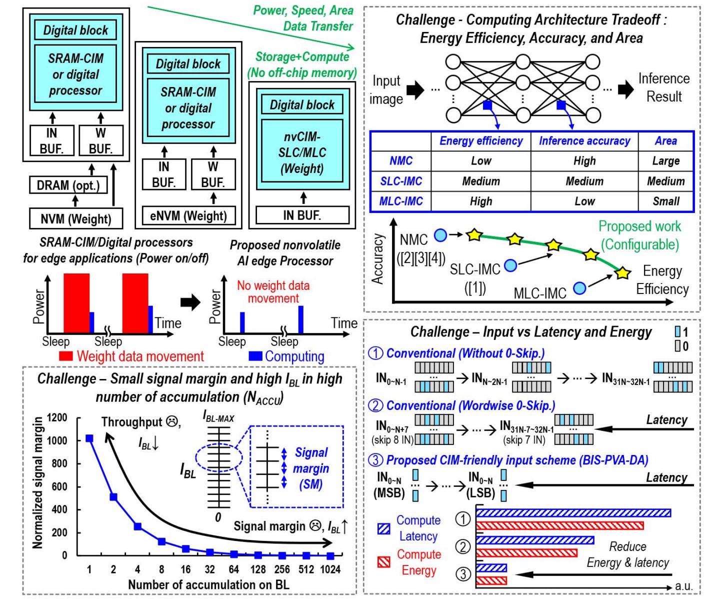
Figure 16.6.5 presents the measurement results of nonvolatile AI-edge processor and hmRe-nvCIM macro. Using the MobileNet-v2 model with 8b precision, the nonvolatile AI-edge processor achieved 25.1-51.4TOPS/W across various datasets (ImageNet, CIFAR-100, CIFAR-10). The BIS-PVA-DA reduced operation cycle counts during inference by 2.26-19.6× and 1.79-8.56× compared to conventional no-zero-skipping and wordwise-skipping schemes, respectively. When using different N_{ACCU}, DACQ reduced the energy consumption of the hmRe-nvCIM macro by 1.67-5.67× across different modes. In IMC-SLC and IMC-MLC modes with N_{ACCU}=32, the usage of CVH-ADC and BIS-PVA-DA reduced computing yield degradation by 6.33× and 6.47×, respectively, compared to that of using a conventional 5b flash ADC.

Figure 16.6.6 presents the measurement results of the AI-edge processor and comparison table. The wakeup-to-response latency was 472.71μs when applying ResNet-20 to the CIFAR-100 dataset for one-shot inference. Using ResNet-20 and MobileNet-v2 models with 8b precision, the degradation in top-1 inference accuracy was <0.67% from CIFAR-10 to ImageNet. The area overhead of reconfigurability (6 modes) is 1.43× compare to an hmRe-nvCIM macro which supports only IMC-SLC mode. Fig. 16.6.7 presents a die photo, chip summary and ReRAM device summary table.

Acknowledgement:
The authors would like to thank MOST-Taiwan, TSRI, NTHU-TSMC JDP for financial and manufacturing support.

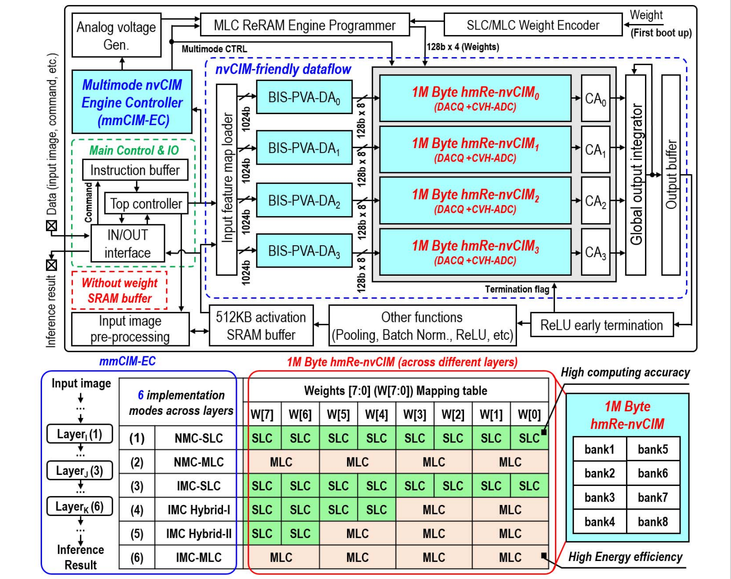
References:
[1] M. Chang et al., "A 40nm 60.64TOPS/W ECC-Capable Compute-in-Memory/Digital 2.25MB/768KB RRAM/SRAM System with Embedded Cortex M3 Microprocessor for Edge Recommendation Systems," *ISSCC*, pp. 270-271, 2022.
[2] M. Giordano et al., "CHIMERA: A 0.92 TOPS, 2.2 TOPS/W Edge AI Accelerator with 2 MByte On-Chip Foundry Resistive RAM for Efficient Training and Inference," *IEEE Symp. VLSI Circuits*, 2021
[3] D. Rossi et al., "A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7μW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode," *ISSCC*, pp. 60-61, 2021.
[4] V. Jain et al., "TinyVers: A 0.8-17 TOPS/W, 1.7 μW-20 mW, Tiny Versatile System-on-chip with State-Retentive eMRAM for Machine Learning Inference at the Extreme Edge," *IEEE Symp. VLSI Circuits*, pp. 20-21, 2022.

2023 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-6654-9016-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISSCC42615.2023.10067610


Power, Speed, Area
Data Transfer
Storage+Compute
(No off-chip memory)
Digital block
SRAM-CIM or digital processor
IN BUF.
W BUF.
DRAM (opt.)
NVM (Weight)
eNVM (Weight)
nvCIM-SLC/MLC (Weight)
SRAM-CIM/Digital processors for edge applications (Power on/off)
Proposed nonvolatile AI edge Processor
No weight data movement
Power
Time
Sleep
Weight data movement
Computing
Challenge - Computing Architecture Tradeoff : Energy Efficiency, Accuracy, and Area
Input image
Inference Result
Accuracy
NMC ([2][3][4])
SLC-IMC ([1])
MLC-IMC
Proposed work (Configurable)
Energy Efficiency
Challenge – Small signal margin and high I_BL in high number of accumulation (N_ACCU)
Throughput ☹, I_BL↓
I_BL-MAX
I_BL
Signal margin (SM)
Signal margin ☹, I_BL↑
Normalized signal margin
Number of accumulation on BL
1200
1000
800
600
400
200
0
1 2 4 8 16 32 64 128 256 512 1024
Challenge – Input vs Latency and Energy
① Conventional (Without 0-Skip.)
IN0~N-1 → IN N~2N-1 → ··· → IN31N~32N-1
② Conventional (Wordwise 0-Skip.)
IN0~N+7 (skip 8 IN) → ··· → IN31N-7~32N-1 (skip 7 IN)
Latency
③ Proposed CIM-friendly input scheme (BIS-PVA-DA)
IN0~N (MSB) → ··· → IN0~N (LSB)
Compute Latency
Compute Energy
Reduce Energy & latency
a.u.

	Energy efficiency	Inference accuracy	Area
NMC	Low	High	Large
SLC-IMC	Medium	Medium	Medium
MLC-IMC	High	Low	Small

Figure 16.6.1: Challenges hindering low-power nonvolatile AI-edge processors.



6 implementation modes across layers		W[7]	W[6]	W[5]	W[4]	W[3]	W[2]	W[1]	W[0]
		Weights [7:0] (W[7:0]) Mapping table							
(1)	NMC-SLC	SLC	SLC	SLC	SLC	SLC	SLC	SLC	SLC
(2)	NMC-MLC	MLC		MLC		MLC		MLC	
(3)	IMC-SLC	SLC	SLC	SLC	SLC	SLC	SLC	SLC	SLC
(4)	IMC Hybrid-I	SLC	SLC	SLC	SLC	MLC		MLC	
(5)	IMC Hybrid-II	SLC	SLC	MLC		MLC		MLC	
(6)	IMC-MLC	MLC		MLC		MLC		MLC	

Figure 16.6.2: Proposed architecture of ReRAM-based nonvolatile AI-edge processor and SLC-MLC hybrid with multimode.

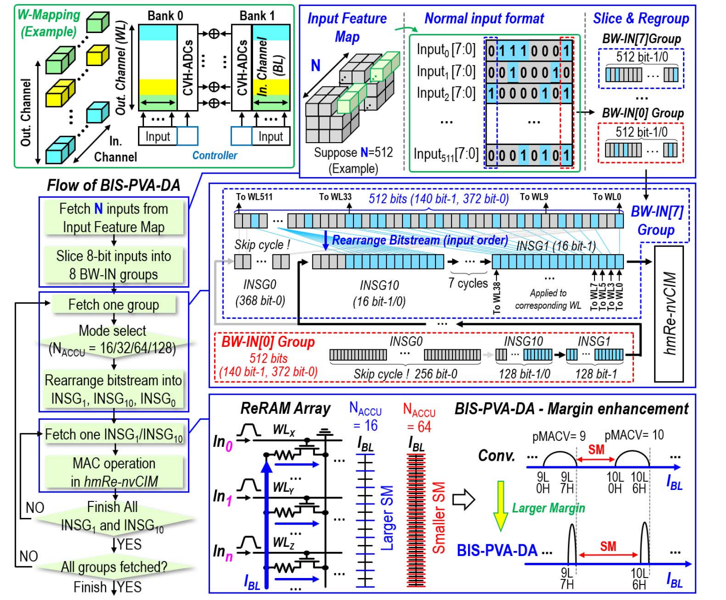
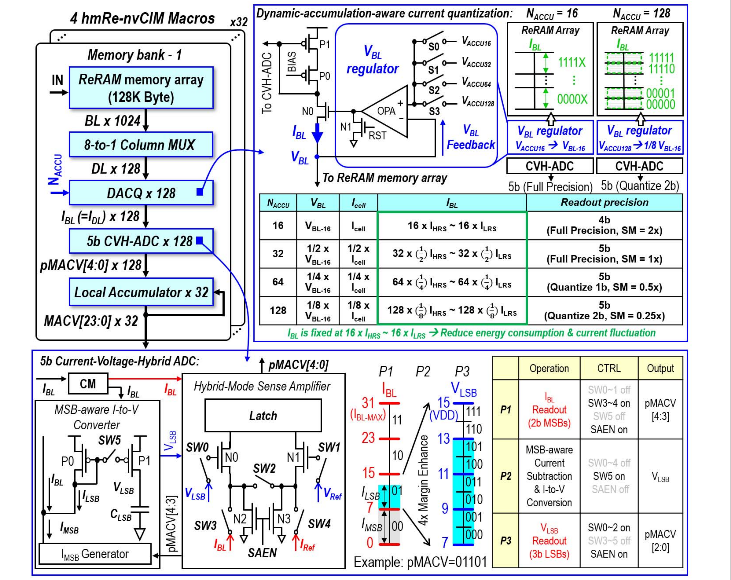


Figure 16.6.3: Flow chart and illustrations of proposed BIS-PVA-DA.

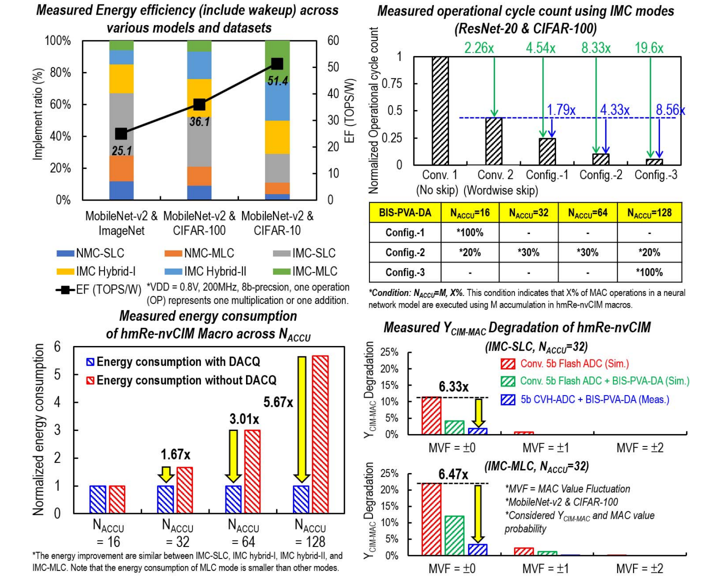


N_ACCU	V_BL	I_cell	I_BL	Readout precision
16	V_{BL-16}	I_{cell}	$16 \times I_{HRS} \sim 16 \times I_{LRS}$	4b (Full Precision, SM = 2x)
32	1/2 x V_{BL-16}	1/2 x I_{cell}	$32 \times (\frac{1}{2}) I_{HRS} \sim 32 \times (\frac{1}{2}) I_{LRS}$	5b (Full Precision, SM = 1x)
64	1/4 x V_{BL-16}	1/4 x I_{cell}	$64 \times (\frac{1}{4}) I_{HRS} \sim 64 \times (\frac{1}{4}) I_{LRS}$	5b (Quantize 1b, SM = 0.5x)
128	1/8 x V_{BL-16}	1/8 x I_{cell}	$128 \times (\frac{1}{8}) I_{HRS} \sim 128 \times (\frac{1}{8}) I_{LRS}$	5b (Quantize 2b, SM = 0.25x)

I_{BL} is fixed at $16 \times I_{HRS} \sim 16 \times I_{LRS}$ → Reduce energy consumption & current fluctuation

	Operation	CTRL	Output
P1	I_{BL} Readout (2b MSBs)	SW0~1 off SW3~4 on SW5 off SAEN on	pMACV [4:3]
P2	MSB-aware Current Subtraction & I-to-V Conversion	SW0~4 off SW5 on SAEN off	V_{LSB}
P3	V_{LSB} Readout (3b LSBs)	SW0~2 on SW3~5 off SAEN on	pMACV [2:0]

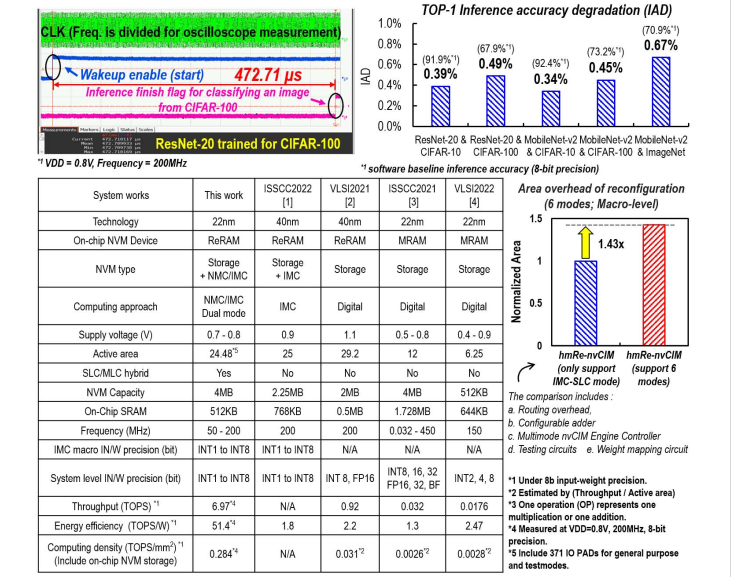
Figure 16.6.4: Illustration of dynamic-accumulation-aware current quantization (DACQ) and current-voltage-hybrid analog-to-digital converter (CVH-ADC).



BIS-PVA-DA	N_{ACCU}=16	N_{ACCU}=32	N_{ACCU}=64	N_{ACCU}=128
Config.-1	*100%	-	-	-
Config.-2	*20%	*30%	*30%	*20%
Config.-3	-	-	-	*100%

*Condition: N_{ACCU}=M, X%. This condition indicates that X% of MAC operations in a neural network model are executed using M accumulation in hmRe-nvCIM macros.

Figure 16.6.5: Measurement results of proposed nonvolatile AI-edge processor and hmRe-nvCIM macro.

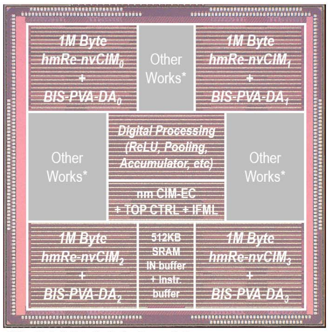


System works	This work	ISSCC2022 [1]	VLSI2021 [2]	ISSCC2021 [3]	VLSI2022 [4]
Technology	22nm	40nm	40nm	22nm	22nm
On-chip NVM Device	ReRAM	ReRAM	ReRAM	MRAM	MRAM
NVM type	Storage + NMC/IMC	Storage + IMC	Storage	Storage	Storage
Computing approach	NMC/IMC Dual mode	IMC	Digital	Digital	Digital
Supply voltage (V)	0.7 - 0.8	0.9	1.1	0.5 - 0.8	0.4 - 0.9
Active area	24.48*5	25	29.2	12	6.25
SLC/MLC hybrid	Yes	No	No	No	No
NVM Capacity	4MB	2.25MB	2MB	4MB	512KB
On-Chip SRAM	512KB	768KB	0.5MB	1.728MB	644KB
Frequency (MHz)	50 - 200	200	200	0.032 - 450	150
IMC macro IN/W precision (bit)	INT1 to INT8	INT1 to INT8	N/A	N/A	N/A
System level IN/W precision (bit)	INT1 to INT8	INT1 to INT8	INT 8, FP16	INT8, 16, 32 FP16, 32, BF	INT2, 4, 8
Throughput (TOPS) *1	6.97*4	N/A	0.92	0.032	0.0176
Energy efficiency (TOPS/W) *1	51.4*4	1.8	2.2	1.3	2.47
Computing density ($TOPS/mm^2$) *1 (Include on-chip NVM storage)	0.284*4	N/A	0.031*2	0.0026*2	0.0028*2

The comparison includes :
a. Routing overhead,
b. Configurable adder
c. Multimode nvCIM Engine Controller
d. Testing circuits e. Weight mapping circuit

*1 Under 8b input-weight precision.
*2 Estimated by (Throughput / Active area)
*3 One operation (OP) represents one multiplication or one addition.
*4 Measured at VDD=0.8V, 200MHz, 8-bit precision.
*5 Include 371 IO PADs for general purpose and testmodes.

Figure 16.6.6: Measured wakeup-to-response latency, inference accuracy and comparison table.



*Other Works include circuits for others projects

Chip summary		
Technology	22nm CMOS logic process (Ultra low leakage)	
ReRAM	Foundry provided 1T1R ReRAM	
Test chip area	30.6mm^2 (Including IO pads, testmodes, and works for other projects)	
Application	CNN/FC	
Computing approach	SLC-IMC, MLC-IMC, SLC-NMC, MLC-NMC	
Active area	24.48 mm^2 (Including IO pads)^{*6}	
Frequency	50 - 200MHz	
ReRAM memory	4M Byte	
SRAM buffer	512K Byte	
Supply voltage	0.7 - 0.8V	
Input / Weight precision	1b to 8b (INT)	
Chip-level Performance		
Throughput (TOPS)	INT 4	24.88 ^{*1}
	INT 8	6.97 ^{*1}
Energy efficiency (TOPS/W)	INT 4	197 ^{*2} - 251 ^{*3}
	INT 8	51.4 ^{*2} – 68.9 ^{*3}
Compute density (TOPS/mm^2)	INT 4	1.101 ^{*1}
	INT 8	0.284 ^{*1}
Macro-level Performance (4 macros; 32 sub-banks)		
Throughput^{*5} (TOPS)	INT 4	26.97 ^{*1}
	INT 8	8.21 ^{*1}
Energy efficiency^{*5} (TOPS/W)	INT 4	241.8 ^{*2} - 287.6 ^{*3}
	INT 8	67.2 ^{*2} - 76.5 ^{*3}
Compute density^{*5} (TOPS/mm^2)	INT 4	2.66 ^{*1}
	INT 8	0.81 ^{*1}

^{*1} Peak throughput measured under 0.8V, 200 MHz.
^{*2} Measured under 0.8V, 200 MHz using MobileNet-v2 trained for CIFAR-10.
^{*3} Measured under 0.7V, 50 MHz using MobileNet-v2 trained for CIFAR-10.
^{*4} One operation (OP) represents one multiplication or one addition.
^{*5} Include overhead of nvCIM-friendly dataflow.
^{*6} Active area = Test chip area – Other works.

ReRAM device characteristics	
Operation – SET/RESET (including MLC states and forming)	1.2V - 3.6V
Operation - read voltage	< 0.3V
Cell size	53F^2

Figure 16.6.7: Die photo, chip summary and ReRAM device summary.