A 28nm Nonvolatile AI Edge Processor using 4Mb Analog-Based Near-Memory-Compute ReRAM with 27.2 TOPS/W for Tiny AI Edge Devices

Tai-Hao Wen[1], Je-Min Hung[1], Hung-Hsi Hsu[1], Yuan Wu[1], Fu-Chun Chang[1], Chung-Yuan Li[1], Chih-Han Chien[1], Chin-I Su[2], Win-San Khwa[2], Jui-Jen Wu[2], Chung-Chuan Lo[1], Ren-Shuo Liu[1], Chih-Cheng Hsieh[1], Kea-Tiong Tang[1], Mon-Shu Ho[3], Yu-Der Chih[2], Tsung-Yung Jonathan Chang[2], Meng-Fan Chang[1,2]

[1]National Tsing Hua University, Taiwan; [2]TSMC, Taiwan, [3]National Chung Hsing University, Taiwan

*Email: mfchang@ee.nthu.edu.tw

Abstract

Tiny AI edge processors prefer using nvCIM to achieve low standby power, high energy efficiency (EF), and short wakeup-to-response latency (T_{WR}). Most nvCIMs use in-memory computing for MAC operations; however, this imposes a tradeoff between EF and accuracy, due to MAC accumulation-number (N_{ACU}) versus signal margin and readout quantization. To achieve high EF and high accuracy, we developed a system-level nvCIM-friendly control scheme and a nvCIM macro with two analog near-memory computing schemes. The proposed 28nm nonvolatile AI edge processor with 4Mb ReRAM-nvCIM achieved high EF (27.2 TOPS/W), short T_{WR} (3.19ms), and low accuracy loss (<0.5%) The EF of the ReRAM-nvCIM macro was 38.6 TOPS/W.

Keywords: multiply-and-accumulation (MAC), nonvolatile memory (NVM), nonvolatile compute-in-memory (nvCIM)

Introduction

Most AI processors [2]-[4] without nvCIM suffer low EF and long T_{WR} (Fig. 1), due to the need to move weight data from off-chip NVM to on-chip processing units after wakeup. Therefore, AI edge processors [1] prefer using nvCIM [5]-[6] to reduce T_{WR} and improve EF by integrating NVM storage and computation functionalities within the same macro (nvCIM).

The challenges of designing nvCIM-based processors (Fig.2) include (1) limited N_{ACU} in nvCIM due to small and non-linear signal margin associated with the read scheme and parasitic resistance on bitlines (BL); (2) a macro-level tradeoff between ADC precision versus read-yield/energy; (3) a system-level tradeoff between EF/latency versus inference accuracy.

Overview of proposed Nonvolatile AI Edge Processor

Our objective was to develop a nonvolatile AI processor with high EF, short T_{WR}, and high accuracy. This was achieved by developing two macro-level nvCIM schemes and one system-level scheme, including (1) near-memory horizontal accumulation (NMHA) to achieve high N_{ACU} and robust/linear signal margin; (2) non-linear quantization readout (NLQR) to increase precision and sensing margin (read-yield) without increasing ADC energy; (3) MACV-aware LSB truncation (MALT) nvCIM-control scheme to increase system throughput and EF without compromising inference accuracy. The proposed processor (Fig. 3) includes 4Mb ReRAM-nvCIM, a system controller, a 512Kb SRAM buffer, and other neural network functions.

nvCIM: Near-Memory Horizontal Accumulation

The proposed ReRAM-nvCIM macro (Fig. 4) includes 16 output channels, each of which comprises one ReRAM sub-array, one NMHA unit with 64 global computing cells (GCC), and one NLQR unit. The NMHA operation comprise three phases. Phase-1: When IN=1, each GCC precharges (PRE_EN=1) BL to V_{BL} and pulls CN to V_{CN}. When IN=0, the BL and CN remain at VSS. Phase-2: PRE_EN=0 and a wordline (WL) pulse is activated. If the weight stored in a ReRAM cell is 0 (LRS), then the memory-cell current discharges the BL and CN and then turns off N3. If the weight stored in a ReRAM-cell is 1 (HRS), then the BL and CN maintain their voltage and N3 is on (I_{N3}>0), indicating that the multiplication result (INxW) of this GCC is “1”. Phase-3: NMHA accumulates all I_{N3} horizontally from GCC_0 to GCC_{63} and generates dataline current (I_{DL}) as the current-domain partial MACV of 64 accumulations ($\text{pMACV}=\sum_{k=0}^{63}\text{IN}_k\text{xW}_k$).

nvCIM: Non-linear Quantization Readout (NLQR)

The proposed NLQR (Fig. 5) uses 5 phases to determine the pMAC precision and sensing margin of the ADC according to the detected pMACV zone. Phase-1: The 2b detector readouts I_{DL} to detect the ZONE of pMACV ($Z_{pMACV}[1:0]$). Phase-2: An adjustable current-mirror (ACM) modifies the mirror-ratio (R_{CM}) according to Z_{pMACV}. Phase-3: An ACM converts I_{DL} to a voltage (V_{C0}) on capacitor C0. Note that the signal margin on V_{C0} is proportional to R_{CM}. Phase-4: The ADC reads out 4 bits from V_{C0} and generates scaled-pMACV ($pMACV_S$). Phase-5: The place-value of $pMACV_S$ is recovered according to Z_{pMACV}.

System Scheme: MACV-aware LSB Truncation (MALT)

The proposed MALT (Fig. 6) dynamically truncates the MAC operation for the input-LSBs of nvCIM according to predicted-MACV zone. In each 8b-MAC operation, 64 (N_{ACU}) inputs are sent from the input feature map to the nvCIM macro, bitwise from MSB to LSB. The MALT counts the number of “1” ($N_{IN[7]}=\sum_{k=0}^{63}\text{IN}_k[7]$) in input-MSB, and then compared it with the threshold values ($N_{TH0\sim3}$) to determine whether to skip the MAC operation for input-LSBs. If $N_{IN[7]}>N_{TH3}$, the predicted-MACV is in a large-value zone and the MAC operation for IN[3:0] in nvCIM is skipped with small MACV deviation-%. If $N_{IN[7]}<N_{TH0}$, the predicted-MACV is small and the full INPUT is applied to nvCIM without skipping to obtain a matched-MACV without loss of system-level accuracy.

Measurement Results and Conclusion

A nonvolatile AI edge processor chip is fabricated using foundry provided 28nm ReRAM technology and a demo system is used to verify the chip, as Fig. 7 shows. When applying the ResNet-20 model to the CIFAR-100 dataset for one-shot inference, the measured T_{WR} was 3.19ms. Under 8b precision, the top-1 inference accuracy degradation was 0.45%, while the measured EFs of the processor and macro were 27.2 TOPS/W and 38.6 TOPS/W, respectively. As shown in Table I, the EF of the proposed processor was 15x higher than that of previous nvCIM-based processors.

References

[1] M. Chang, ISSCC, 2022, pp. 1-3. [2] M. Giordano, VLSI, 2021, pp. 1-2. [3] D. Rossi, ISSCC, 2021, pp. 60-62. [4] V. Jain, VLSI, 2022, pp. 20-21. [5] J. -M. Hung, ISSCC, 2022, pp. 1-3. [6] J. -M. Hung, Nat Electron 4, pp. 921–930, 2021.

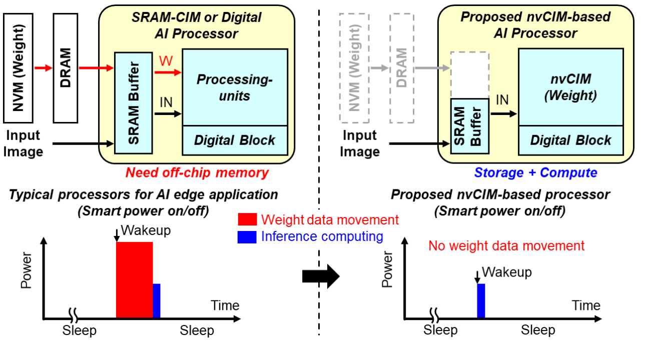


Fig.1 Architectures and behavior of typical and nvCIM processor.

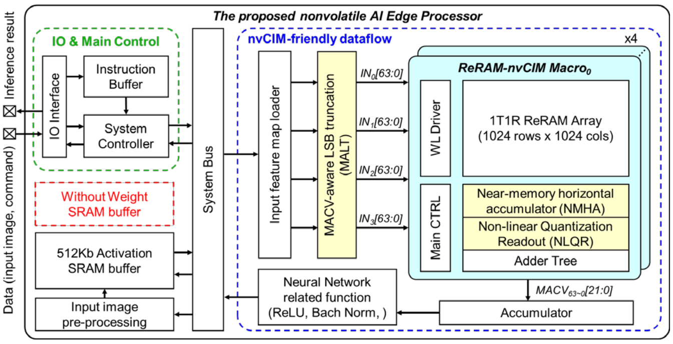
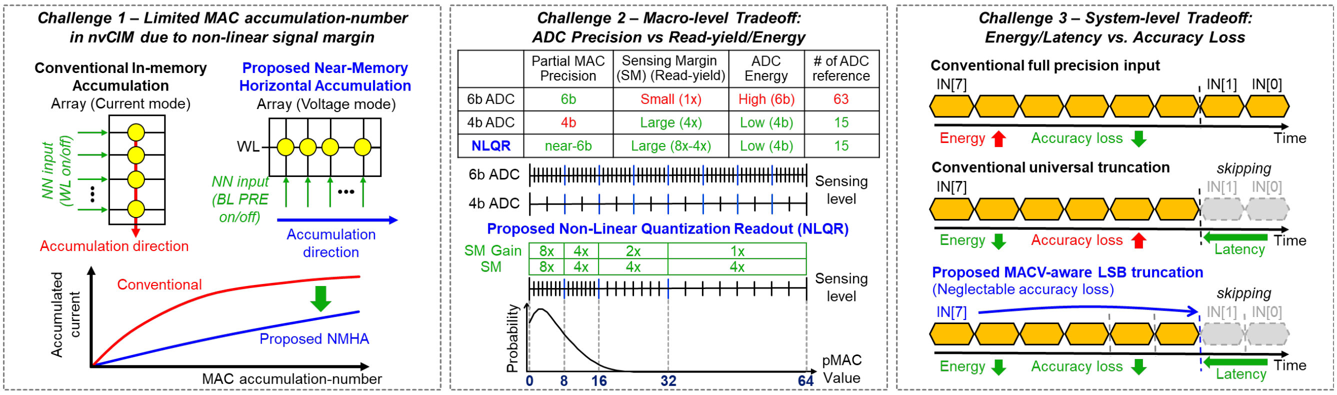
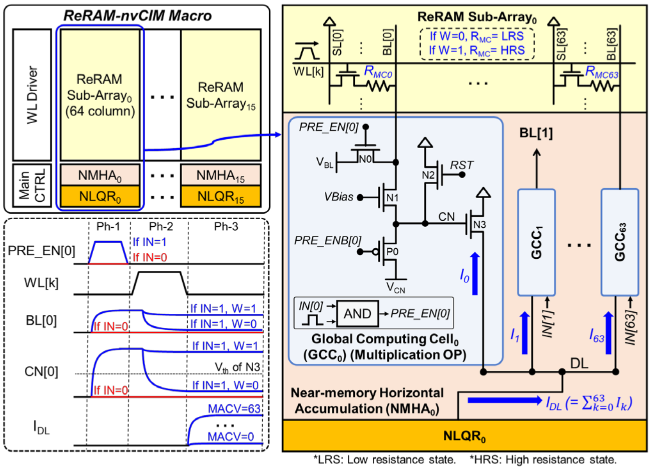

The proposed nonvolatile AI Edge Processor
IO & Main Control
IO Interface
Instruction Buffer
System Controller
Inference result
Data (input image, command)
Without Weight SRAM buffer
512Kb Activation SRAM buffer
Input image pre-processing
System Bus
nvCIM-friendly dataflow
Input feature map loader
MACV-aware LSB truncation (MALT)
IN0[63:0]
IN1[63:0]
IN2[63:0]
IN3[63:0]
x4
ReRAM-nvCIM Macro0
WL Driver
1T1R ReRAM Array (1024 rows x 1024 cols)
Main CTRL
Near-memory horizontal accumulator (NMHA)
Non-linear Quantization Readout (NLQR)
Adder Tree
MACV63~0[21:0]
Neural Network related function (ReLU, Bach Norm,)
Accumulator

Fig. 3 Architecture of proposed nonvolatile AI edge processor.


Challenge 1 – Limited MAC accumulation-number in nvCIM due to non-linear signal margin
Conventional In-memory Accumulation
Array (Current mode)
NN input (WL on/off)
Accumulation direction
Proposed Near-Memory Horizontal Accumulation
Array (Voltage mode)
WL
NN input (BL PRE on/off)
Accumulation direction
Accumulated current
Conventional
Proposed NMHA
MAC accumulation-number
Challenge 2 – Macro-level Tradeoff: ADC Precision vs Read-yield/Energy
6b ADC
4b ADC
Sensing level
Proposed Non-Linear Quantization Readout (NLQR)
SM Gain
SM
8x
4x
2x
1x
8x
4x
4x
4x
Probability
0
8
16
32
64
pMAC Value
Challenge 3 – System-level Tradeoff: Energy/Latency vs. Accuracy Loss
Conventional full precision input
IN[7]
IN[1]
IN[0]
Energy
Accuracy loss
Time
Conventional universal truncation
skipping
Latency
Proposed MACV-aware LSB truncation (Neglectable accuracy loss)

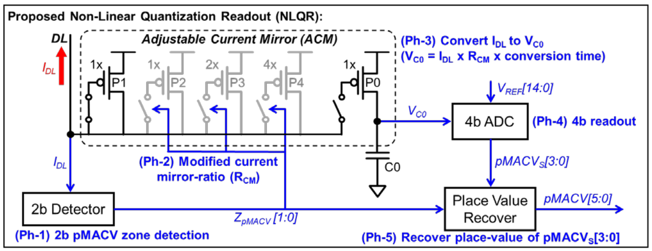
	Partial MAC Precision	Sensing Margin (SM) (Read-yield)	ADC Energy	# of ADC reference
6b ADC	6b	Small (1x)	High (6b)	63
4b ADC	4b	Large (4x)	Low (4b)	15
NLQR	near-6b	Large (8x-4x)	Low (4b)	15

Fig. 2 Challenges in developing an nvCIM-based processor and the proposed schemes.



*LRS: Low resistance state. *HRS: High resistance state.

Fig. 4 Structure of ReRAM-CIM macro and near-memory horizontal accumalation (NMHA) scheme.


Proposed Non-Linear Quantization Readout (NLQR):
DL
IDL
Adjustable Current Mirror (ACM)
1x
P1
1x
P2
2x
P3
4x
P4
1x
P0
(Ph-3) Convert IDL to VC0 (VC0 = IDL x RCM x conversion time)
VREF[14:0]
VC0
4b ADC
(Ph-4) 4b readout
C0
pMACVS[3:0]
(Ph-2) Modified current mirror-ratio (RCM)
2b Detector
ZpMACV [1:0]
Place Value Recover
pMACV[5:0]
(Ph-1) 2b pMACV zone detection
(Ph-5) Recover place-value of pMACVS[3:0]

Z_{pMACV} [1]	Z_{pMACV} [0]	R_{CM}	V_{C0} Margin Enhancement	pMACV[5]	pMACV[4]	pMACV[3]	pMACV[2]	pMACV[1]	pMACV[0]
1	1	1/8	1x	$pMACV_S[3]$	$pMACV_S[2]$	$pMACV_S[1]$	$pMACV_S[0]$	0	0
1	0	1/4	2x	0	$pMACV_S[3]$	$pMACV_S[2]$	$pMACV_S[1]$	$pMACV_S[0]$	0
0	1	1/2	4x	0	0	$pMACV_S[3]$	$pMACV_S[2]$	$pMACV_S[1]$	$pMACV_S[0]$
0	0	1	8x	0	0	0	$pMACV_S[3]$	$pMACV_S[2]$	$pMACV_S[1]$

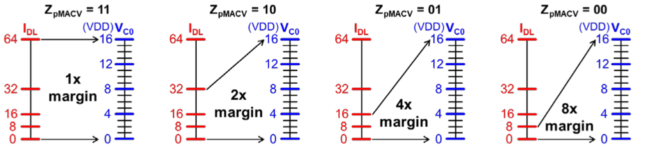


Fig.5 Circuitry and operation of proposed non-linear quantization readout (NLQR) scheme.

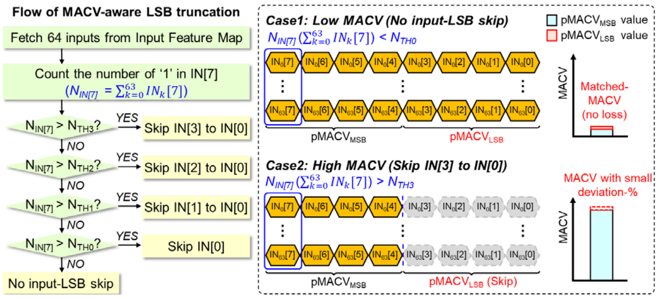

Flow of MACV-aware LSB truncation
Fetch 64 inputs from Input Feature Map
Count the number of '1' in IN[7]
(NIN[7] = Σ63k=0 INk[7])
NIN[7] > NTH3?
YES
Skip IN[3] to IN[0]
NO
NIN[7] > NTH2?
Skip IN[2] to IN[0]
NIN[7] > NTH1?
Skip IN[1] to IN[0]
NIN[7] > NTH0?
Skip IN[0]
No input-LSB skip
Case1: Low MACV (No input-LSB skip)
NIN[7] (Σ63k=0 INk[7]) < NTH0
pMACVMSB value
pMACVLSB value
MACV
Matched-MACV (no loss)
pMACVMSB
pMACVLSB
Case2: High MACV (Skip IN[3] to IN[0])
NIN[7] (Σ63k=0 INk[7]) > NTH3
MACV with small deviation-%
pMACVLSB (Skip)

Fig. 6 Flow chart and illustration of proposed MACV-aware LSB truncation (MALT) scheme.

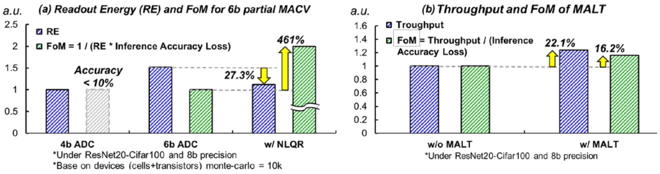
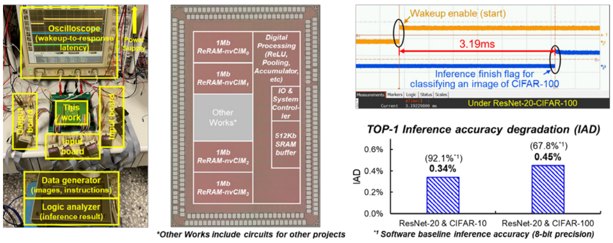


Fig. 7 Simulation results of proposed schemes: (a) NLQR decreased energy by 27.3% and increase FoM by 461%; (b) MALT increased throughput by 22.1% and FoM by 16.5%.


Oscilloscope (wakeup-to-response latency)
This work
Input board
Data generator (images, instructions)
Logic analyzer (inference result)
1Mb ReRAM-nvCIM0
1Mb ReRAM-nvCIM1
1Mb ReRAM-nvCIM2
1Mb ReRAM-nvCIM3
Other Works*
Digital Processing (ReLU, Pooling, Accumulator, etc)
IO & System Controller
512Kb SRAM buffer
Wakeup enable (start)
3.19ms
Inference finish flag for classifying an image of CIFAR-100
Under ResNet-20-CIFAR-100
TOP-1 Inference accuracy degradation (IAD)
IAD
0.6%
0.4%
0.2%
0.0%
(92.1%*1)
0.34%
(67.8%*1)
0.45%
ResNet-20 & CIFAR-10
ResNet-20 & CIFAR-100

*Other Works include circuits for other projects

*1 Software baseline inference accuracy (8-bit precision)

Fig.8 Die photo, demo system, and measured results.

System works	This work	ISSCC2022 [1]	VLSI2021 [2]	ISSCC2021 [3]	VLSI2022 [4]
Technology	28nm	40nm	40nm	22nm	22nm
On-chip NVM Device	ReRAM	ReRAM	ReRAM	MRAM	MRAM
NVM type	Storage + Compute	Storage + Compute	Storage	Storage	Storage
Computing approach	Analog NMC	IMC	Digital	Digital	Digital
Chip area	6*5	25	29.2	12	6.25
Supply voltage (V)	0.7 - 0.8	0.9	1.1	0.5 - 0.8	0.4 - 0.9
Frequency (MHz)	50 - 200	200	200	0.032 - 450	150
System level IN/W precision	INT1 to INT8	INT1 to INT8	INT 8, FP16	INT8, 16, 32, FP, BF	INT2, 4, 8
Throughput (TOPS) *1	0.34*4	N/A	0.92	0.032	0.0176
Energy efficiency (TOPS/W)*1	27.2*4	1.8	2.2	1.3	2.47
Computing density (TOPS/mm^2)*1 (Include on-chip NVM storage)	0.056^4	N/A	0.031*2	0.0026*2	0.0028*2

*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 277 IO PADs for general purpose and testmode.

Table I Comparison table of nonvalatile AI processors.