11.4 An 8-Mb DC-Current-Free Binary-to-8b Precision ReRAM Nonvolatile Computing-in-Memory Macro using Time-Space-Readout with 1286.4 - 21.6TOPS/W for Edge-AI Devices

Je-Min Hung[1], Yen-Hsiang Huang[1], Sheng-Po Huang[1], Fu-Chun Chang[1], Tai-Hao Wen[1], Chin-I Su[2], Win-San Khwa[2], Chung-Chuan Lo[1], Ren-Shuo Liu[1], Chih-Cheng Hsieh[1], Kea-Tiong Tang[1], Yu-Der Chih[2], Tsung-Yung Jonathan Chang[2], Meng-Fan Chang[1,2]

[1]National Tsing Hua University, Hsinchu, Taiwan
[2]TSMC, Hsinchu, Taiwan

2022 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-6654-2800-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISSCC42614.2022.9731715

Battery-powered edge-AI devices require nonvolatile computing-in-memory (nvCIM) macros for nonvolatile data storage and multiply-and-accumulate (MAC) operations. High inference accuracy requires MAC operations with high input (IN), weight (W), and output (OUT) precisions. A high energy efficiency (EF_{MAC}) and a short computing latency (t_{AC}) are also required. Most existing silicon-verified nvCIM macros use current-mode signal generation; using current [1-3] or hybrid current-voltage readout schemes [4-5] for multibit MAC operations to compensate for the small BL-voltage swing and signal margin resulting from the low read-disturb-free voltage (V_{RD}).

As shown in Fig. 11.4.1, current-mode nvCIMs face various challenges: (1) a limited EF_{MAC} due to the use of DC current in the NVM cell array and peripheral readout circuits; (2) a limited output precision and ratio (R_{OUT}: actual output precision/full-precision) due to the limited signal margin imposed by a low V_{RD}; (3) a long computing latency per output bit ($t_{AC-OUT} = t_{AC}$/OUT-precision) imposed by the analog signal development time in large cell arrays, and/or multi-phase small-signal readout by the ADC. These challenges are addressed by developing: (1) a DC-current-free time-space based in-memory computing (DCFTS-IMC) for in-array signal generation and peripheral readout operations to eliminate the need for a DC-current and to reduce power consumption; (2) an integration-based voltage-to-time converter (IVTC) that increases the time-step signal margin to tolerate a higher number of accumulations and enhance readout accuracy; (3) a hidden-latency time-to-MACV conversion (HLTMC) scheme to improve t_{AC-OUT} by hiding multibit MAC readout latency within the development time of the analog MAC signal. The proposed 22nm 8Mb ReRAM nvCIM macro achieves a high memory capacity with a 0.76ns/b t_{AC-OUT} and a 1286.4 – 21.6TOPS/W EF_{MAC} from binary to 8b IN - 8b W - 19b OUT; a 24b output combines outputs from 32 channels in the same macro.

Each 256kb memory bank performs 8b IN - 8b W MAC operations: using 8-channel accumulation with 2's-complement 8b-weight data stored in 8 SLC ReRAM cells across 8 columns on the same row (WL). Each set of columns comprises memory cells, column multiplexors (YMUX), a BL pre-charge circuit, an IVTC, and a HLTMC. A reconfigurable digital shift-and-add (DSA) circuit is shared by the eight columns.

Figure 11.4.2 shows the signal generation of partial MAC values (pMACV) using the proposed DCFTS-IMC scheme. Each clock cycle includes a pipelined bitwise-input computing phase (BIC-P) and a DSA phase (DSA-P), completing N iterations for an N-bit input. For example, an 8b-input (IN[7:0]) requires at most 8 WL pulses across 8 short BIC-Ps, each of which includes 3 sub-phases (SP-0~SP-2).

In SP-0, all peripheral circuits are rapidly reset to the initial state. During SP-1, the selected BL is pre-charged to a target voltage (V_{BLP}), not exceeding V_{RD}. During SP-2, 8 WLs are activated for the 8 corresponding bitwise-inputs with a given place-value: e.g., 8 MSB, $IN_0[7]$ - $IN_7[7]$, for BIC-P1. The cell-current (I_{CELL}) is the result of multiplying the WL (IN) by the cell-value (W), as in previous nvCIMs. The BL current (I_{BL}), which is the sum of 8 I_{CELL} values for 1b IN × 1b W with 8 accumulations, discharges the BL parasitic capacitance (C_{BL}) by varying the rate at which the BL voltage (V_{BL}) reduces in accordance to pMACV. A larger I_{BL} indicates that a larger number of low-resistance cells (LRS) are accessed (i.e., higher pMACV) and V_{BL} discharge rate is steeper, which reduce the BL discharge latency (t_{pMACV}) to meet the target BL voltage (V_{BL-TH}). A smaller I_{BL} indicates a lower pMACV and a more gradual V_{BL} discharge rate, resulting in a longer t_{pMACV}.

Figure 11.4.3 illustrates the operation of the proposed IVTC, which converts the V_{DL} (V_{BL} selected by YMUX) discharge rate into a delay-time rising signal (DTS), while increasing the time-step between DTSs of neighboring pMACV pairs. The IVTC comprises of a coupling capacitor (C_0), a sampling capacitor (C_1), an integration capacitor (C_2), four initial transistors (N_0, N_1, N_3, P_1), one PMOS (P_0) as an integration current source, four switches (SW_1-SW_4), one feedback transistor (P_2), one threshold-trigger NMOS (N_2), and one output inverter (INV_1).

IVTC performs three tasks: circuit initialization during SP-0, V_{t-P0} sampling during SP-1, and integration with launch during SP-2. During SP-0, SW_1 is on and resets node DC to V_{BLP}, N_0, N_1, and SW_3 are on and reset SAMPLE and SENSE nodes to 0V, and P1 is on and resets node OUTB to V_{DD}. During SP-1, SW_3 is on and SW_4 is off to store the threshold voltage of P_0 ($V_{SAMPLE} = V_{DD} – V_{t-P0}$) on C_1; thereby, suppressing the V_t induced offset variation. During SP-2, SW_1 and SW_3 are off, SW_2 and SW_4 are on, and the WL is on to discharge the selected BL and data lines (DL) via ReRAM cells; ΔV_{DL} ($V_{BLP} – V_{DL}$) is continuously AC coupled to the SAMPLE node via C_0. The larger ΔV_{DL} swing can increase the integration current (I_{charge}) provided by P_0; I_{charge} charges C_2, causing the SENSE node voltage (V_{SENSE}) to rise. When the charge integrated on C_2 is high and V_{SENSE} exceeds the threshold voltage of N_2 (V_{t-N2}), then OUTB is pulled down to launch a rising signal to IVTCO, while P_2 is switched on to pull-up the SENSE node. A larger transistor is used for N_2 to suppress its V_{t-N2} variation, and the N_2-P_2 feedback mechanism reduces the transient current and the energy consumed by INV_1. The IVTCO time-step (Δt_{IVTC}) is kept larger than Δt_{BL} by converting the small DL voltage swing to a larger SENSE voltage, where the amplification ratio ($\Delta t_{IVTC}/\Delta t_{BL}$) is determined by the size of P_0 and the C_2 matching capacitance.

Figure 11.4.4 illustrates HLTMC operation: it converts the IVTCO into digital pMACV[2:0] values, while performing the time-space readout concurrently to the analog MACV development time. Each HLTMC comprises of a time-to-digital converter (TDC) and a timing calibration table (TCT). The number of TDC cells can be increased beyond 16 if higher time-space resolution is required. HLTMC is enabled by the WL. Each TDC takes 16 reference timings (t_{REF}[15:0]) as inputs to detect the timing of the rising IVTCO signals and thereby generates 16b time-space codes (pMACV-TC[15:0]): pMACV-TC[m] is 0 if $t_{IVTCO} > t_{REF-m}$ and is 1 otherwise. pMACV-TC[15:0] is then mapped by TCT entries to generate a 3b pMACV[2:0] output. Note that TCT compensates for the near-far WL effects and for process variation in each readout path. In each DSA-P, the 1st-level of DSA (DSA-L1) combines pMACV[2:0] from HLTMC[0] to HLTMC[7] to generate an 11b pMACV value ($DSA_{L1}[10:0] = IN_0[7]·W_0[7:0] + IN_1[7]·W_1[7:0] + … + IN_7[7]·W_7[7:0]$). The 2nd-level DSA (DSA-L2) combines DSA_{L1}[10:0] from DSA-P1 to DSA-P8 and thereby generates a 19b MACV (DSA_{L2}[18:0]), where $DSA_{L2}[18:0] = IN_0[7:0]·W_0[7:0] + IN_1[7:0]·W_1[7:0] + … + IN_7[7:0]·W_7[7:0]$).

Figure 11.4.5 summarizes the performance of the proposed schemes. Using 8b IN - 8b W MAC operations, the proposed DCFTS-IMC scheme reduces the array's energy consumption (E_{ARRAY}) by 2.06 - 16.5× compared to a current-mode signal-generation scheme with a varying number of activated WLs (accessed cells). Using the ResNet-20 model trained for the CIFAR-100 dataset shows that the DCFTS -IMC reduces E_{ARRAY} by 5.15× on average. Using the proposed IVTC, Δt_{IVTC} provides a 1.58× signal margin enhancement, than that of $\Delta t_{BL.}$ The proposed HLTMC scheme improves t_{AC-OUT} by 1.36 - 8.3×, compared to previous work using conventional current/voltage mode readout schemes.

Figure 11.4.6 presents the measurement results of proposed macro, which is fabricated using foundry provided ReRAM devices. The measured waveforms confirm that each BIC-P is 1.59ns using a 0.8V supply for a 1b IN - 1b W - 3b OUT pMACV. In 8b IN - 8bW - 19b OUT operation, the Shmoo results indicate a 14.4ns t_{AC} using a 0.8V supply for 8b precision using 8 BIC-Ps. Using 8 accumulations the average EF_{MAC} is 21.6TOPS/W and the peak EF_{MAC} is 28.74TOPS/W using a 0.8V supply with a ResNet-20 model applied to the CIFAR-100 dataset; using 16 accumulations the peak EF_{MAC} is 61.84TOPS/W using a 0.75V supply with a 90% input sparsity. In binary operation, the peak energy efficiency is 1.28POPS/W using 16 accumulations and a 90% input sparsity. Compared to previous work, the proposed scheme improves FoM (EF_{MAC} × input-precision × weight-precision × output-ratio × capacity) by 276.7 - 6.18× for binary to 8b IN - 8b W configurations. The system level inference accuracy is shown to achieve 91.74% and 67.11% when applied to CIFAR-10 and CIFAR-100 datasets using a ResNet-20 model with 8b IN - 8b W precision. Figure 11.4.7 presents a summary table and die photo of the proposed macro.

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

References:

[1] C.-X. Xue et al., "A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN Based AI Edge Processors," *ISSCC*, pp. 388-389, 2019.
[2] Q. Liu et al., "A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing," *ISSCC*, pp. 500-501, 2020.
[3] C.-X. Xue et al., "A 22nm 2Mb ReRAM Compute-in-Memory Macro with 121-28TOPS/W for Multibit MAC Computing for Tiny AI Edge Devices," *ISSCC*, pp. 244-245, 2020.
[4] C.-X. Xue et al., "A 22nm 4Mb 8b-Precision ReRAM Computing-in-Memory Macro with 11.91 to 195.7TOPS/W for Tiny AI Edge Devices," *ISSCC*, pp. 245-247, 2021.
[5] J.-H. Yoon et al., "A 40nm 64Kb 56.67TOPS/W Read-Disturb-Tolerant Compute-in-Memory/Digital RRAM Macro with Active-Feedback-Based Read and In-Situ Write Verification," *ISSCC*, pp. 404-406, 2021.

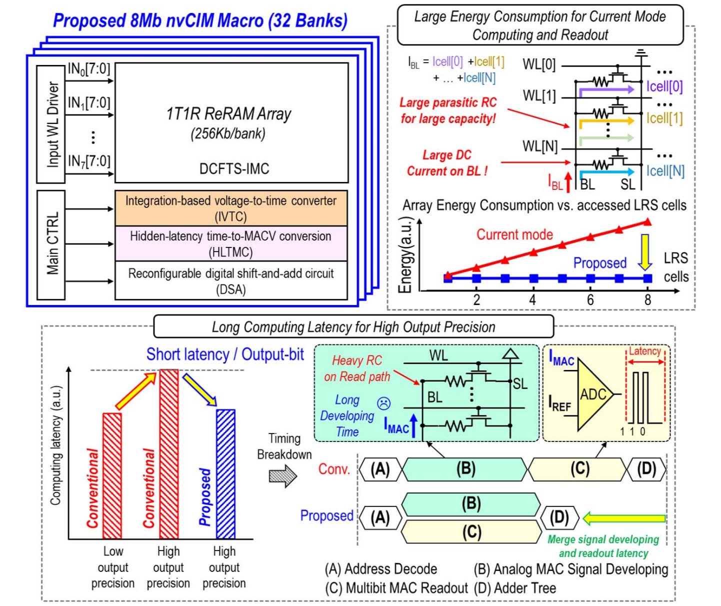
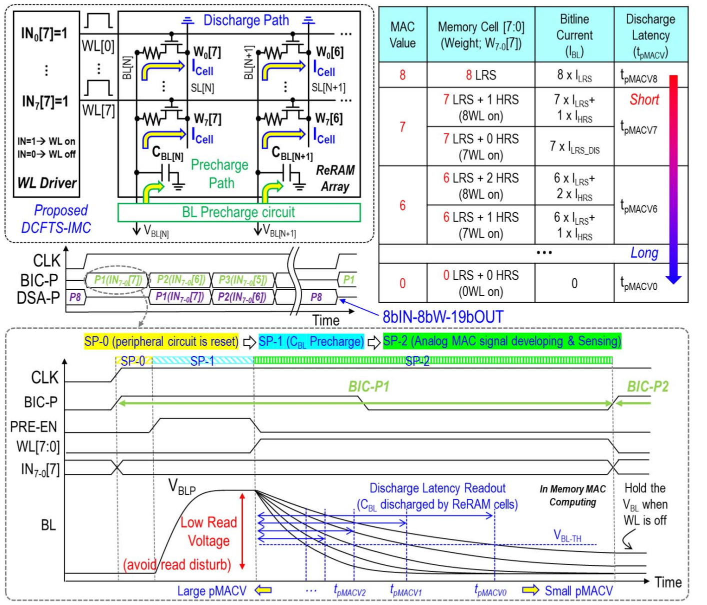

Proposed 8Mb nvCIM Macro (32 Banks)
Input WL Driver
Main CTRL
1T1R ReRAM Array (256Kb/bank)
DCFTS-IMC
Integration-based voltage-to-time converter (IVTC)
Hidden-latency time-to-MACV conversion (HLTMC)
Reconfigurable digital shift-and-add circuit (DSA)
Large Energy Consumption for Current Mode Computing and Readout
Large parasitic RC for large capacity!
Large DC Current on BL !
Array Energy Consumption vs. accessed LRS cells
Current mode
Proposed
LRS cells
Long Computing Latency for High Output Precision
Short latency / Output-bit
Computing latency (a.u.)
Conventional
Proposed
Low output precision
High output precision
Heavy RC on Read path
Long Developing Time
Timing Breakdown
Conv.
Proposed
Merge signal developing and readout latency
(A) Address Decode (B) Analog MAC Signal Developing
(C) Multibit MAC Readout (D) Adder Tree

Figure 11.4.1: Challenges and proposed scheme for multi bit nvCIM.



MAC Value	Memory Cell [7:0] (Weight; $W_{7\text{-}0}[7]$)	Bitline Current (I_{BL})	Discharge Latency (t_{pMACV})
8	8 LRS	8 x I_{LRS}	t_{pMACV8}
7	7 LRS + 1 HRS (8WL on)	7 x I_{LRS}+ 1 x I_{HRS}	Short t_{pMACV7}
	7 LRS + 0 HRS (7WL on)	7 x I_{LRS_DIS}	
6	6 LRS + 2 HRS (8WL on)	6 x I_{LRS}+ 2 x I_{HRS}	t_{pMACV6}
	6 LRS + 1 HRS (7WL on)	6 x I_{LRS}+ 1 x I_{HRS}	
	…		Long
0	0 LRS + 0 HRS (0WL on)	0	t_{pMACV0}

Figure 11.4.2: Multibit MAC operations using DCFTS-IMC.

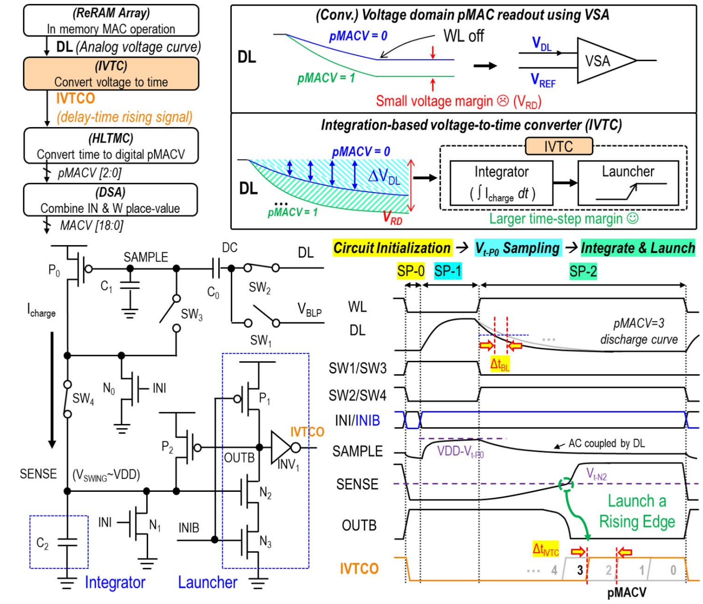

(ReRAM Array) In memory MAC operation
DL (Analog voltage curve)
(IVTC) Convert voltage to time
IVTCO (delay-time rising signal)
(HLTMC) Convert time to digital pMACV
pMACV [2:0]
(DSA) Combine IN & W place-value
MACV [18:0]
(Conv.) Voltage domain pMAC readout using VSA
WL off
Small voltage margin ☹ (V_RD)
Integration-based voltage-to-time converter (IVTC)
Integrator
Launcher
Larger time-step margin ☺
Circuit Initialization → V_t-P0 Sampling → Integrate & Launch
SP-0 SP-1 SP-2
pMACV=3 discharge curve
AC coupled by DL
Launch a Rising Edge
Integrator
Launcher

Figure 11.4.3: Integration-based voltage-to-time converter (IVTC) operation.

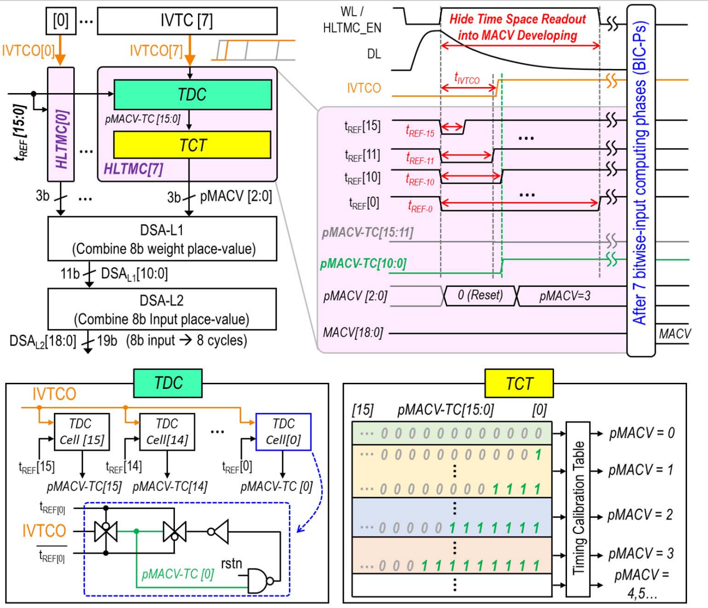

IVTC [7]
TDC
TCT
HLTMC[7]
DSA-L1 (Combine 8b weight place-value)
DSA-L2 (Combine 8b Input place-value)
19b (8b input → 8 cycles)
Hide Time Space Readout into MACV Developing
After 7 bitwise-input computing phases (BIC-Ps)
0 (Reset)
pMACV=3
TDC Cell [15]
TDC Cell[14]
TDC Cell[0]
Timing Calibration Table
pMACV = 0
pMACV = 1
pMACV = 2
pMACV = 3
pMACV = 4,5…

Figure 11.4.4: Hidden-latency time-to-MACV conversion (HLTMC) operation.

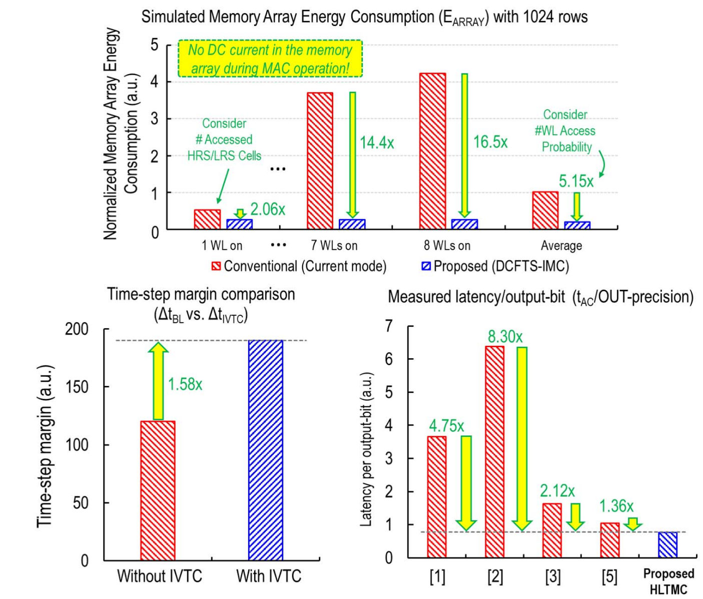


Figure 11.4.5: Simulated performance of the proposed scheme.

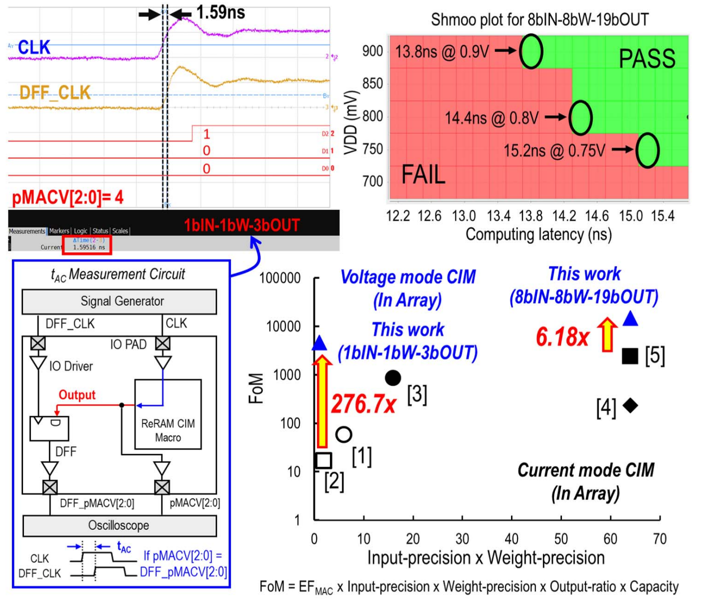
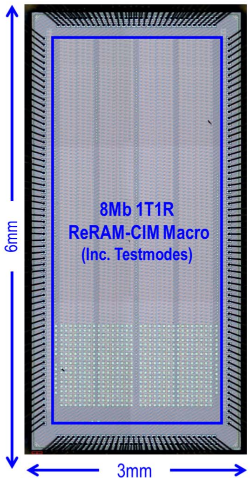


Figure 11.4.6: Measurement results and FoM comparison.



Chip Summary		
Technology	22nm CMOS Logic Process	
ReRAM	Foundry 1T1R SLC ReRAM	
Testchip Size	6mm x 3mm (Inc.IO pad and testmodes)	
Capacity	8Mb (32 Sub-banks)	
Sub-bank	1024 rows x 256 columns	
Supply Voltage	0.8V	
Input precision	1 bit ~ 8 bits	
Weight precision	1 bit ~ 8 bits	
Output precision	3 bits ~ 19 bits (24 bits for combing 32 banks output)	
Performance		
CIM-mode Computing Latency (ns)	8bIN-8bW-19bOUT	14.4*1
	1bIN-1bW-3bOUT	3.2*1
Throughput (GOPS)	8bIN-8bW-19bOUT	142.2*1
	1bIN-1bW-3bOUT	5120*1
Energy Efficiency (TOPS/W)	8bIN-8bW-19bOUT	21.6*1 ~ 61.8*2
	1bIN-1bW-3bOUT	416.5*1 ~ 1286.4*2
Inference Accuracy	CIFAR-10	91.74%*3
	CIFAR-100	67.11%*4

*1. Average performance measured @ 0.8V VDD with 8 accumulations.
*2. Peak performance measured @ 0.75V VDD with 16 accumulations and 90% input sparsity.
*3. Using ResNet-20 model and the software baseline was 92.2% (-0.46%).
*4. Using ResNet-20 model and the software baseline was 68.03% (-0.92%)

Figure 11.4.7: Die micrograph and chip summary.