24.1 A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN-Based AI Edge Processors

Cheng-Xin Xue, Wei-Hao Chen, Je-Syu Liu, Jia-Fang Li, Wei-Yu Lin, Wei-En Lin, Jing-Hong Wang, Wei-Chen Wei, Ting-Wei Chang, Tung-Cheng Chang, Tsung-Yuan Huang, Hui-Yao Kao, Shih-Ying Wei, Yen-Cheng Chiu, Chun-Ying Lee, Chung-Chuan Lo, Ya-Chin King, Chorng-Jung Lin, Ren-Shuo Liu, Chih-Cheng Hsieh, Kea-Tiong Tang, Meng-Fan Chang

National Tsing Hua University, Hsinchu, Taiwan

Embedded nonvolatile memory (NVM) and computing-in-memory (CIM) are significantly reducing the latency (t_{MAC}) and energy consumption (E_{MAC}) of multiply-and-accumulate (MAC) operations in artificial intelligence (AI) edge devices [1,2]. Previous ReRAM CIM macros demonstrated MAC operations for 1b-input, ternary-weighted, 3b-output CNNs [1] or 1b-input, 8b-weighted, 1b-output fully-connected networks with limited accuracy [2]. To support higher-accuracy convolution neural network heavy applications NVM-CIM should support multibit inputs/weights and multi-bit output (MAC-OUT) for CNN operations. One way to achieve multibit weights is to use a multi-level ReRAM cell to store the weight. However, as shown in Fig. 24.1.1, multibit ReRAM CIM faces several challenges. (1) a tradeoff between area and speed for multibit input/weight/MAC-OUT MAC operations; (2) sense amplifier's high input offset, large area, and high parasitic load on the read-path due to large BL currents (I_{BL}) from multibit MAC; (3) limited accuracy due to a small read/sensing margin (I_{SM}) across MAC-OUT or variation in cell resistance (particularly MLC cells). To overcome these challenges, this work proposes, (1) a serial-input non-weighted product (SINWP) structure to optimize the tradeoff between area, t_{MAC} and E_{MAC}; (2) a down-scaling weighted current translator (DSWCT) and positive-negative current-subtractor (PN-ISUB) for short delay, a small offset and a compact read-path area; and (3) a triple-margin small-offset current-mode sense amplifier (TMCSA) to tolerate a small I_{SM}. A fabricated 55nm 1Mb ReRAM-CIM macro is the first ReRAM CIM macro to support CNN operations using multibit input/weight MAC-OUT. This device achieves the shortest CIM-MAC-access time (t_{AC}) among existing ReRAM-CIMs (t_{MAC}=14.6ns with 2b-input, 3b-weight with 4b-MAC-OUT) and the best peak E_{MAC} of 53.17 TOPS/W (in binary mode).

Figure 24.1.2 presents the structure of the proposed ReRAM-CIM macro, using a 1T1R SLC ReRAM array, a current-aware BL clamper (CABLC), DSWCT, SINWP sampler-and-combiner (SINWP-SC), PN-ISUB, and TMCSA. This work places both positive and negative weights in the same array, but in different columns, unlike [1]. Each 3b-signed weight (W=1b-sign+2b-data) is stored in the same row of either the 2-column positive (PWG) or negative (NWG) group. In each PWG/NWG, the even BLs (BLM) represent the MSB and the odd BLs (BLL) represent the LSB. In other words, 4 SLC ReRAM cells are used to store 3b signed weights. To support configurable 1b/2b input (IN), this work translates 2b-input into two sequential single-bit (IN_0 and IN_1) WL pulses in one CIM clock cycle. For an n-by-n CNN kernel, n^2 weights are stored in n^2 consecutive rows. CABLC clamps the BL voltage (V_{BLC}) for current-mode sensing. A BL current (I_{BL}) that is equal to the sum of n^2 cell currents (I_{MC}) is generated via binary multiplication (IN×W) between a WL (IN) and a 1b-weight (W), as in [1]. The I_{BL} of BLM ($I_{BL\text{-}MSB}$) and BLL ($I_{BL\text{-}LSB}$) do not include their place-value processed in array (non-weighted I_{BL}). DSWCT and SINWP-SC then combine the $I_{BL\text{-}MSB}$ and $I_{BL\text{-}LSB}$ of a PWG/NWG to generate a weighted data-line (DL) current (I_{DL}) for the MAC value (MACV) using 2b-inputs and 2b-weights. PN-ISUB outputs the difference (I_{SUB}) between the I_{DL} of PWG ($I_{DL\text{-}P}$) and NWG ($I_{DL\text{-}N}$) as well as the sign bit ($DOUT_{SIGN}$=0 when $I_{DL\text{-}P} > I_{DL\text{-}N}$), where $I_{SUB}=|I_{DL\text{-}P} - I_{DL\text{-}N}|$. The current reduction provided by DSWCT and PN-ISUB allows for an I_{SUB} that is much smaller than the total current ($I_{DL\text{-}MSB} + I_{DL\text{-}LSB}$) of the accessed PWG and NWG. This allows TMCSA to use smaller transistors (achieving a lower parasitic load), which results in faster response times, smaller area and less power compared to schemes without DSWCT and PN-ISUB. Finally, each IO repeats the operation of TMCSA in 3 sequential phases to detect the same I_{SUB} with 3 different I_{REF} currents (I_{REF1} to I_{REF3}), and then outputs a 4b MACV with 3b data from TMCSA and a sign bit from PN-ISUB.

Figure 24.1.3 shows the operations of DSWCT, SINWP-SC, and PN-ISUB. DSWCT uses a current-mirror (CM) to translate $I_{DL\text{-}MSB}$ and $I_{DL\text{-}LSB}$ into lower weighted DL currents (I_{WDL}), $I_{WDL\text{-}MSB}$= $(1/p)\cdot 2\cdot I_{DL\text{-}MSB}$ and $I_{WDL\text{-}LSB}$=$(1/p)\cdot I_{DL\text{-}LSB}$. Choice of the reduced-amount ratio, p (p=4 for this work) is based on maintaining sufficient I_{SM}, while reducing power consumption and the transistor size (area and parasitic load) required for the remaining read-path. SINWP-SC operations occur in two phases. (1) For the 1st WL pulse (IN_0), INSW=1 and the gate-voltage ($V_{GCM\text{-}MSB}/V_{GCM\text{-}LSB}$) of CM_{MSB}/CM_{LSB} of DSWCT is sent to N3/N6 and stored in capacitor CM/CL to sample the $I_{WDL\text{-}MSB}/I_{WDL\text{-}LSB}$ resulting from IN_0×W. (2) For the 2nd WL pulse (IN_1), INSW=0 and $V_{GCM\text{-}MSB}/V_{GCM\text{-}LSB}$ is sent to N4/N7. The size of N4/N7 is 2× that of N3/N6. SINWP-SC then combines the drain currents of N3, N4, N6 and N7 to output a weighted I_{DL}, as follows:

$$I_{DL\_LSB[0]}\cdot W_{DL\text{-}LSB[0]}/8+I_{DL\_MSB[0]}\cdot W_{DL\text{-}MSB[1]}/4+I_{DL\_LSB[1]}\cdot W_{DL\text{-}LSB[0]}/4+I_{DL\_MSB[1]}\cdot W_{DL\text{-}MSB[1]}/2$$

In PN-ISUB, a comparator compares I_{DL_P} with I_{DL_N} to output the sign bit ($DOUT_{SIGN}$=0 when $I_{DL\text{-}P} > I_{DL\text{-}N}$), and $DOUT_{SIGN}$ enables PN-ISUB to connect the larger/lower current path of DL to the high/low-current input (I_{HC} / I_{LC}) terminal of a current-subtractor. If $DOUT_{SIGN} = 1$, $I_{HC} = I_{DL\text{-}N}$, and $I_{LC} = I_{DL\text{-}P}$, then the current-subtractor generates the PN-ISUB output current I_{SUB} (=$I_{HC} - I_{LC}$, =$I_{DL\text{-}N} - I_{DL\text{-}P}$ if $DOUT_{SIGN} = 1$). Compared to schemes that convert $I_{DL\text{-}P}$ / $I_{DL\text{-}N}$ to MACV separately before the digital combiner as in [1], PN-ISUB removes current leakage from HRS cells ($n^2\cdot I_{HRS}$) when $I_{DL\text{-}N} >> I_{DL\text{-}P}$, or enable larger current reduction (or small I_{SUB}) when $I_{DL\text{-}N}$ ~=$I_{DL\text{-}P}$ to improve sensing yield in the CSA.

Figure 24.1.4 shows the operation of TMCSA, which comprises two pairs of PMOS transistors (P1:P2 and P3:P4), four pairs of switches (SW1-SW4), two overdrive-coupling capacitors (C1, C2), four discharge NMOSs (DN1-DN4), an NMOS latch (N1-N3) and two current inputs (I_{IN} and I_{REF}). In stand-by mode, SW3 and SW4 are on, and SW1, SW2, and VDD_SA are off. DSD and CHD are high and DN1-DN4 are turned on to hold nodes DP1, DP2, LQ and LQB at 0V. In phase-1 (PH1, V_{TH} sampling), DSD, SW1 and SW2 are off and VDD_SA is on. The threshold voltages (V_{TH1} to V_{TH4}) of diode-connected P1 to P4 are stored on their gates (i.e. $V_{G1} = V_{DD} - V_{TH1}$, $V_{G3} = V_{DD} - V_{TH3}$). In phase-2 (PH2, V_{OV} sampling and coupling), SW4 is off, SW1 is on, and I_{IN} / I_{REF} flows through P1/P2. This results in $V_{G1} = V_{DD} - V_{TH1} - V_{OV\text{-}IN}$ and $V_{G2} = V_{DD} - V_{TH2} - V_{OV\text{-}REF}$, where $V_{OV\text{-}IN}$ / $V_{OV\text{-}REF}$ is the overdrive voltage of P1/P2 for sampling I_{IN} / I_{REF} as in [3]. In the meantime, V_{OV1} / V_{OV2} is coupled to V_{G4} / V_{G3} via C1/C2, so that $V_{G4} = V_{DD} - V_{TH4} - V_{OV1}$ and $V_{G3} = V_{DD} - V_{TH3} - V_{OV2}$. Ideally, $I_{P4} = 2\cdot I_{IN}$ and $I_{P3} = 2\cdot I_{REF}$ since transistors P3 and P4 are twice as big (2-finger style) as P1 and P2. In phase-3 (PH3, ΔI amplifying), SW2 is on and SW1 and SW3 are off. Thus, the current at node LQ is $I_{P3} - I_{IN} = 2I_{REF} - I_{IN}$. The current at node LQB is $I_{P4} - I_{REF} = 2I_{IN} - I_{REF}$. For a given period ($T_{PH3}$), the voltage difference ($V_{LQ\text{-}LQB}$) between node LQ and LQB is proportional to $[(2I_{REF} - I_{IN}) - (2I_{IN} - I_{REF})]\cdot T_{PH3} = 3\cdot(I_{IN} - I_{REF})\cdot T_{PH3}$, which is 3× the I_{SM} (=$I_{IN} - I_{REF}$) in conventional CSA. In phase-4 (PH4, latch), SAEN is high and N1-N3 are enabled to detect $V_{LQ\text{-}LQB}$ and generate a digital output at SAOUT.

Figure 24.1.5 shows the performance of the proposed schemes. For 3×3-CNN kernels DSWCT reduces the worst-case current by 3.6×. The combination of DSWCT and PN-ISUB schemes enables a 3.6–3.8× reduction in MACV-current. The SINWP scheme achieves an FoM (SM / (Energy × Area)) 1.4–6× better than those of sequential-input-parallel-weight (SIPW) and parallel-input-parallel-weight (PIPW) structures. TMCSA enabled 6× and 1.7× reductions in input offset, compared to conventional CSA and DR-CSA [1], using a 70μA input current.

Figure 24.1.6 presents the measurement results from a 1Mb ReRAM-CIM macro fabricated using 1T1R SLC ReRAM in a 55nm CMOS process. A demo system was built using our ReRAM-CIM with an FPGA host. For CNN operations using 3×3 kernels and 2b-input and 3b-weight, the captured waveforms confirm that t_{MAC} for a 4b MAC output (1b sign and 3b data) is 14.6ns, excluding the path-delay. Using 3×3-CNN kernels with 1b-input and 3b-weight, shmoo results confirm a t_{MAC} of 11.75ns per cycle at typical V_{DD}, and the system test achieve an 88.52% inference accuracy on the CIFAR-10 dataset. The measured peak energy-efficiency is 53.17TOPS/W in binary mode (1b-input, 3b-weight, 4b-MAC-OUT) and 21.9TOPS/W in multibit mode (2b-IN, 3b-weight, 4b-MAC-OUT) using CIM peripheral circuits and a reference generator. This work also achieves a 3.4× improvement in energy-efficiency and a 1.3× faster t_{MAC}, compared to [1]. Figure 24.1.7 presents the die micrograph and chip summary.

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

References:
[1] W.-H. Chen, et al., "A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors," *ISSCC*, pp. 494-495, 2018.
[2] R. Mochida, et al., "A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture," *VLSI*, pp. 175-176, 2018.
[3] M.-F. Chang, et al., "An offset tolerant current-sampling-based sense amplifier for sub-100nA-cell-current nonvolatile memory," *ISSCC*, pp. 206-207, 2011.

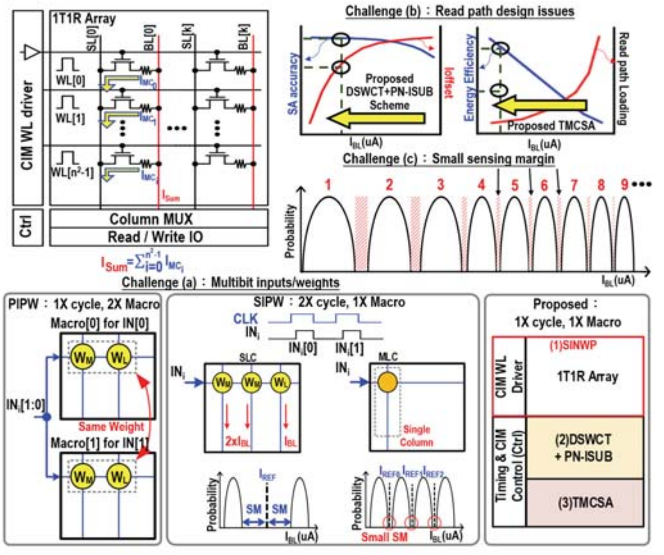
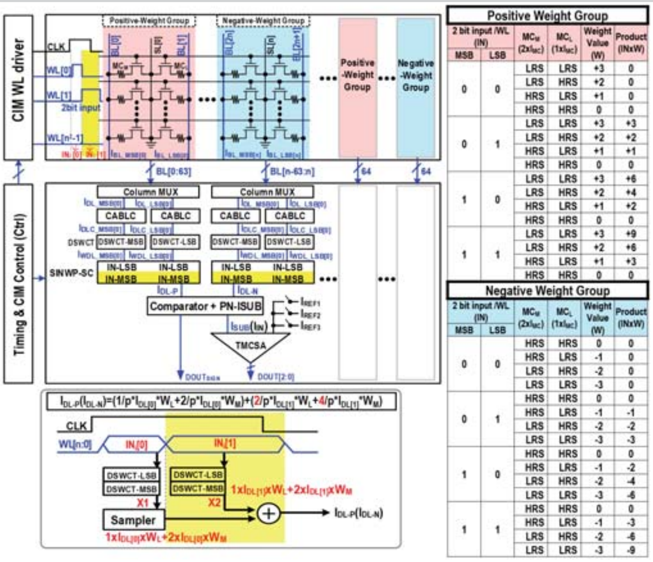


Figure 24.1.1: Multi-bit computation in nonvolatile memory.



Positive Weight Group					
2 bit input /WL (IN)		MC_M ($2xI_{MC}$)	MC_L ($1xI_{MC}$)	Weight Value (W)	Product (INxW)
MSB	LSB				
0	0	LRS	LRS	+3	0
		LRS	HRS	+2	0
		HRS	LRS	+1	0
		HRS	HRS	0	0
0	1	LRS	LRS	+3	+3
		LRS	HRS	+2	+2
		HRS	LRS	+1	+1
		HRS	HRS	0	0
1	0	LRS	LRS	+3	+6
		LRS	HRS	+2	+4
		HRS	LRS	+1	+2
		HRS	HRS	0	0
1	1	LRS	LRS	+3	+9
		LRS	HRS	+2	+6
		HRS	LRS	+1	+3
		HRS	HRS	0	0

Negative Weight Group					
2 bit input /WL (IN)		MC_M ($2xI_{MC}$)	MC_L ($1xI_{MC}$)	Weight Value (W)	Product (INxW)
MSB	LSB				
0	0	HRS	HRS	0	0
		HRS	LRS	-1	0
		LRS	HRS	-2	0
		LRS	LRS	-3	0
0	1	HRS	HRS	0	0
		HRS	LRS	-1	-1
		LRS	HRS	-2	-2
		LRS	LRS	-3	-3
1	0	HRS	HRS	0	0
		HRS	LRS	-1	-2
		LRS	HRS	-2	-4
		LRS	LRS	-3	-6
1	1	HRS	HRS	0	0
		HRS	LRS	-1	-3
		LRS	HRS	-2	-6
		LRS	LRS	-3	-9

Figure 24.1.2: Proposed ReRAM-CIM macro.

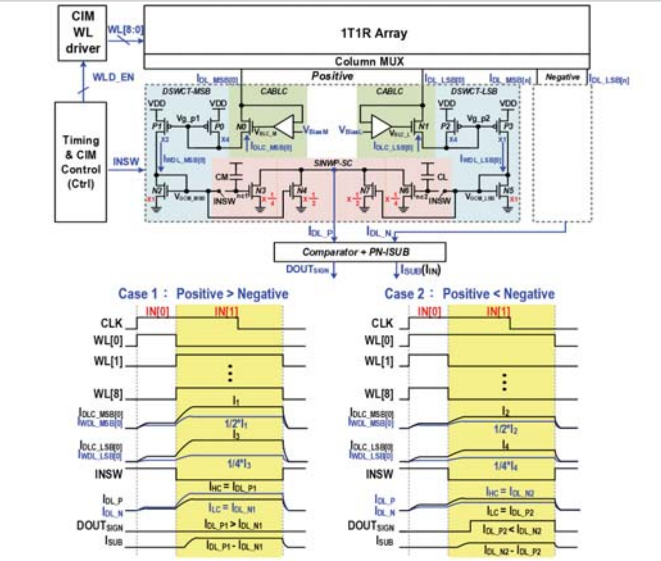
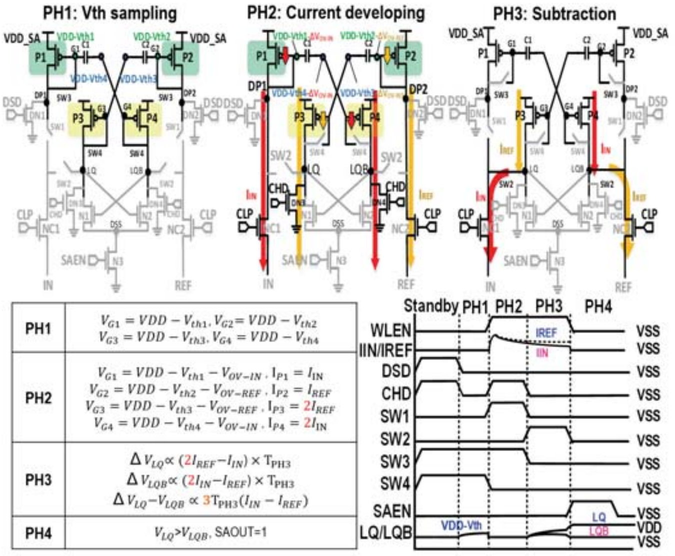


Figure 24.1.3: Sequential input parallel weight structure.



PH1	$V_{G1}=VDD-V_{th1}, V_{G2}=VDD-V_{th2}$ $V_{G3}=VDD-V_{th3}, V_{G4}=VDD-V_{th4}$
PH2	$V_{G1}=VDD-V_{th1}-V_{OV-IN}, I_{P1}=I_{IN}$ $V_{G2}=VDD-V_{th2}-V_{OV-REF}, I_{P2}=I_{REF}$ $V_{G3}=VDD-V_{th3}-V_{OV-REF}, I_{P3}=2I_{REF}$ $V_{G4}=VDD-V_{th4}-V_{OV-IN}, I_{P4}=2I_{IN}$
PH3	$\Delta V_{LQ}\propto(2I_{REF}-I_{IN})\times T_{PH3}$ $\Delta V_{LQB}\propto(2I_{IN}-I_{REF})\times T_{PH3}$ $\Delta V_{LQ}-V_{LQB}\propto 3T_{PH3}(I_{IN}-I_{REF})$
PH4	$V_{LQ}>V_{LQB}$, SAOUT=1

Figure 24.1.4: Structure and operation of TMCSA.

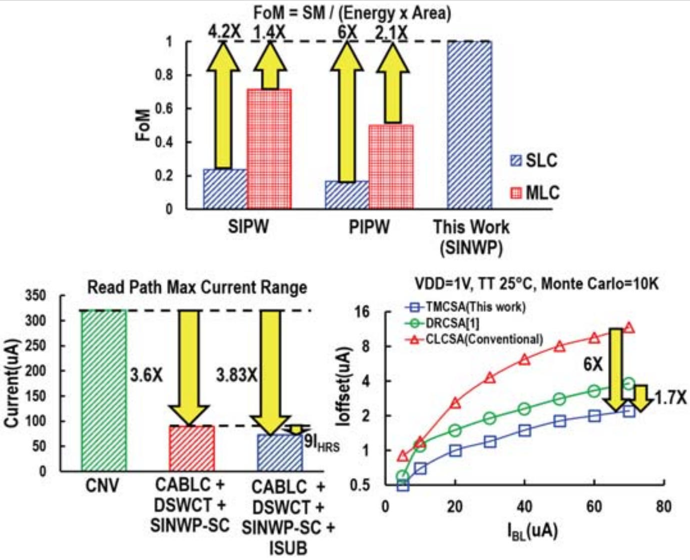


Figure 24.1.5: Performance comparison of this work vs prior-art.

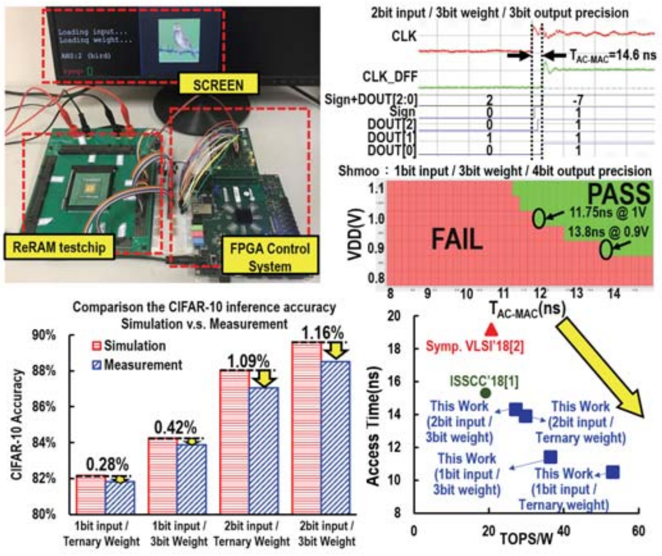
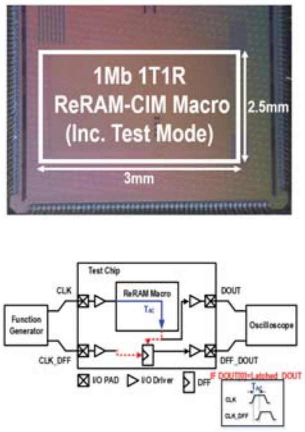


Figure 24.1.6: Measurement results.



Technology	55nm CMOS Logic Process
ReRAM	Logic-process ReRAM
Cell Size(1T1R)	0.2025um^2
ReRAM Mode	Memory/CIM
CIM Mode	CNN
Capacity	1Mb (8 Sub-array)
Sub-array	256rows x 512columns
Read Delay @ VDD =1V	Memory Mode: 3.16ns
	CNN Mode: 11.75ns (1bit input / 3bit weight) (3bit DOUT precision)
	CNN Mode: 14.6ns (2bit input / 3bit weight) (3bit DOUT precision)
Energy Efficiency @ VDD =1V	CNN Mode: 53.17 (TOPS/W) (1bit input / 3bit weight) (3bit DOUT precision)
	CNN Mode: 21.9 (TOPS/W) (2bit input / 3bit weight) (3bit DOUT precision)

Figure 24.1.7: Die micrography and summary Table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE