A 16Mb Dual-Mode ReRAM Macro with Sub-14ns Computing-In-Memory and Memory Functions Enabled by Self-Write Termination Scheme

Wei-Hao Chen[1], Wen-Jang Lin[1], Li-Ya Lai[1], Shuangchen Li[2], Chien-Hua Hsu[3], Huan-Ting Lin[1], Heng-Yuan Lee[3], Jian-Wei Su[3], Yuan Xie[2], Shyh-Shyuan Sheu[3], and Meng-Fan Chang[1]
[1]National Tsing Hua University, Hsinchu, Taiwan, email: mfchang@ee.nthu.edu.tw
[2]UC Santa Barbara, Santa Barbara, CA, USA, [3]ITRI, Hsinchu, Taiwan

***Abstract*—**Recent ReRAM devices enable the development of computing-in-memory (CIM) for beyond von Neumann structure. However, wide distribution in ReRAM resistance (R) causes low yield for CIM operations. This work proposes a dual-mode computing (DMc) ReRAM macro structure with a dual-function voltage-mode self-write termination (DV-SWT) scheme to achieve both memory and fundamental CIM functions (AND, OR and XOR operations) with high yield. The DV-SWT increases the read margin for CIM operations by suppressing the R-variations caused by macro-level IR-drop and process variations. A 16Mb DMc-ReRAM full-function macro was fabricated using 1T1R HfO ReRAM devices and 0.15um CMOS process. The measured delay of the CIM operations is less than 14ns, which is 86+x faster than previous ReRAM-based CIM works. This work also represents the first CIM ReRAM macro with ReRAM device and CIM-peripheral circuits fully integrated on the same die.

I. Introduction

Memory has proven a major bottleneck in the development of energy-efficient chips for IoT and artificial intelligence (AI) using von Neumann structure, as shown in Fig. 1(a). To suppress the overhead in delay and power consumption due to multiple memory accesses and data-transfer between CPU and memory via the bus, a computing/process-in-memory (CIM/PIM) concept was proposed [1][2], as shown in Fig. 1(b).

Recent ReRAM devices not only serve as nonvolatile memory macros, but have also enabled the development of nonvolatile logics (nvLogics) [3] and are considered a good candidate for CIM beyond von Neumann structure. Two types of previous ReRAM-based CIM works have been demonstrated using on-chip ReRAM array with off-chip measurement equipment/chip (i.e. source meter and microcontroller) to emulate fundamental (i.e AND, OR, XOR) and complicate (adder/multiplier) CIM operations. These CIM types (Fig. 2) include in-memory inputs [4] and a look-up-table with external inputs [5]. Both approaches employ voltage-divider operations to set up a temporary voltage for switching extra ReRAM cells for storing outputs within one or multiple write cycles. However, the extra ReRAM cells reduce cell-array efficiency and the extra cell-switch operations delay CIM operation. These works have not discussed the macro-level implementation against resistance variation.

In this work, we propose a Dual-Mode ReRAM structure with SWT scheme to enable CIM operations in a large-capacity ReRAM macro against cell resistance variations. A 16Mb 1T1R CIM-memory dual-mode ReRAM macro with full integration of cell array and CIM-peripheral circuits is demonstrated in silicon with a computing speed 86+x faster than previous works.

II. Proposed Dual-Mode Intelligent RERAM Macro Structure and Design Challenges

A. Proposed CIM-Mmeory Dual-Mode ReRAM Structure

Fig. 3 presents the proposed DMc-ReRAM macro structure, which performs AND, OR and XOR operations. This DMc-ReRAM macro includes dual-mode wordline (DM-WLD) drivers, multiple-logics (ML) current-mode sense amplifier (CSA), self-write-termination (SWT) circuits for SET (SWT-SET) and RESET (SWT-RESET) operations. The DM-WLD can either turn on one wordline (WL) for memory operation or two WLs simultaneously for CIM operation. The ML-CSA, which consumes only 1.3x area of typical CSA, employed two input reference currents ($I_{REF\text{-}OR}$ and $I_{REF\text{-}AND}$), two comparison branches, one input current-mirror, one data-latch, and one output driver to support both memory and CIM operations. Unlike previous CIM, the input data are taken from the memory array, and there is no extra cells are required for switching and read-out operations. Table I compares recent CIM works.

B. Proposed CIM Operations (AND, OR, XOR)

Fig. 4 shows the concept of proposed CIM operations, in which two selected rows (WL[*i*] and WL[*j*]) are turned on according to input addresses sent from CPU. Each BL (BL[*k*]) can perform one-bit (bitwise) logic operations. In this work, HRS represents logic "1", and LRS represents logic "0".With a BL clamping circuit and current-mode read scheme, each accessed memory cell MC[i,k]/MC[j,k] outputs a current (I_H/I_L) to the BL according to its stored data. Then the summed current on an accessed BL ($I_{BL}[k]=I[i,k]+I[j,k]$) is equal to $2I_L$ (two LRS cells, 2L) or I_L+I_H (1 HRS and 1 LRS cells, 1H1L) or $2I_H$ (2 HRS cells, 2H). When R-ratio is high and the distribution of cell resistance (R_H/R_L) is narrow, there is no overlap between the three current states (I_L, I_L+I_H, $2I_H$).

In an AND operation (Fig. 5), by setting the I_{REF_AND} at a value between $2I_H$ and (I_L+I_H), the ML-CSA can distinguish $2I_H$ from (I_L+I_H)/$2I_L$ and output AND/NAND. In an OR

operation (Fig. 6), by setting the I_{REF_OR} at a value between $2I_L$ and (I_L+I_H), the ML-CSA can distinguish $2I_L$ from $(I_L+I_H)/2I_H$ and output OR/NOR. In an XOR operation (Fig. 7), two parallel current comparison operations are executed by ML-CSA with both I_{REF_AND} and I_{REF_OR} as input reference.

C. . Employed HfO ReRAM Cells and Write Behavior

Fig. 8(a) shows the 1T1R HfO ReRAM cell [6] used in this study. SET operation switches a ReRAM cell from high resistance (HRS, R_H) to low resistance (LRS, R_L) with its bit-line (BL) at the SET-voltage (V_{SET}) and source-line (SL) at 0V for a period of T_{SET}. RESET operation switches a ReRAM cell from LRS to HRS, with SL=V_{RESET} and BL=0V for a period of T_{RST}. In practice, ReRAM cells suffer wide distribution in SET/RESET time (T_{SET}/T_{RST}), as shown in Fig. 8(b). Fig. 8(c) shows the conventional one-pulse write scheme, which requires long pulse width to cover the T_{SET}/T_{RST} of tail-bits. However, fast switching cells suffer excessive energy waste, long stress-time, endurance degradation, over-write issues, and large DC-current consumption. Due to limited SL/BL driver size and parasitic RC in a macro, particularly large-capacity macro, the high DC-current causes voltage-drop on V_{SET}/V_{RST} in a macro and induced write failure [7] or wider distribution in cell resistance. Fig. 8(d) presents a conventional program-verify (PV) scheme, which switches the SL/BL and WL voltages between write and verify modes during each PV cycle/period (T_{PVC}). In a large-capacity ReRAM macro, limited charge-pump output current and large parasitic RC cause long rising/falling time on SL/BL/WL voltage switching. This causes high power penalty, long overall PV latency (T_{PV}), and insufficient support for fine resolution (short) T_{PV}.

D. Design Challenges of CIM vs. Process Variation

Fig. 9 presents the measured (256 cells) and predicted (6 σ) current distribution of fabricated 16Mb macro for CIM operations. The limited R-ratio (higher I_H) and wide distribution in cell resistance cause overlap between tail-bits of 1H1L (logic "0,1" or "1,0") and 2L cases. Therefore, a novel write scheme to tackle the wide distribution in R_H/R_L caused by process variation and macro-level systematic voltage degradation (large DC-current/IR drop) [7], [8] is needed.

III. Proposed Self-Write Termination Circuit

To suppress the number of tail bits in R_H/R_L caused by process variation and macro-level V_{SET}/V_{RESET} degradation due to large DC current (IR drop), this work proposes a dual-operation voltage-mode SWT (DV-SWT) scheme. To suppress area overhead, one pair of SWT-SET/SWT-RST circuits are shared by all columns in the same IO. Table II compares the proposed DV-SWT and previous SWT schemes.

In SWT-SET (Fig. 10), when a ReRAM cell switches from HRS to LRS and is below the target threshold R_L value ($R_{L\text{-}TH}$), the SETEND switches to high and turns off P2 to disconnect DL from V_{SET}, resulting in a drop in DL voltage and termination of the SET operation. In SWT-RST (Fig. 11), when a ReRAM cell switches from LRS to HRS and exceeds the target threshold R_H value ($R_{H\text{-}TH}$), the RSTEND switches to low and turns off N1 to disconnect DL from V_{RESET} (controlled by N2), resulting in a rise in DL voltage and termination of the RESET operation.

By accurately setting the values of $R_{L\text{-}TH}$ and $R_{H\text{-}TH}$ and applying small-offset technique to SWT circuits, the DV-SWT achieves (1) elimination of tail bits due to DC-current (IR drop) induced insufficient V_{SET}/V_{RESET} for slow-switching cells, (2) reduction of stress and avoidance of over-write for fast-switching cells, and (3) resistance tuning for tail-bits.

IV. Comparisons and Measurement Results

A 16Mb DMc ReRAM macro, with full CIM-peripheral circuits and test-modes, was fabricated using 0.15um CMOS process and BEOL HfO ReRAM devices. Fig. 12 presents the die photo, testchip structure, and chip summary.

Fig. 13 shows the captured waveform of the SWT operation. The SETEND and RESTEND signals are buffered out in the testchip for monitoring the SWT operations. As expected, the rising of SETEND (Fig. 13(a)) and the falling of RESTEND (Fig. 13(b)) of the fast-switching cells occur earlier than those of slow-switching cells. Fig. 14 shows the measured cell resistance after applying DV-SWT scheme. Even with the limited number of measured cells (256 cells), the SWT-SET already achieved a 25% and 7.5 % reduction in the tail I_L while the SWT-RESET achieved a 50% and 48.3% reduction in the tail I_H. The improvement in tail-bit current enabled the CIM operation to increase in sensing windows compared to that without DV-SWT scheme (Fig. 15). Fig. 16 shows the captured waveforms of the CIM operations of the 16Mb DMc-ReRAM testchip on a test-board. Note that the measured testchip delay time ($T_{AC\text{-}TC}$) includes the macro-level delay (T_{AC}) and path-delay (T_{PATH}=2.6ns) due to IO-pads and parasitic load on PCB-board. The extracted T_{AC} for AND, OR, XOR operations are 13.1ns, 13.7ns and 13.9ns, respectively.

I. Conclusions

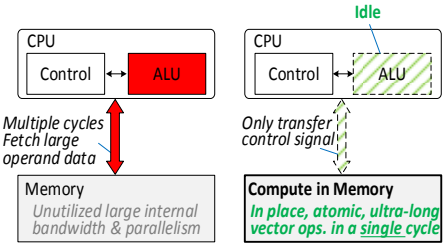
The design of ReRAM-based CIM macros presents challenges in terms of limited R-ratio and large resistance variations, which affect speed, power, and yield. This work proposes a DMc ReRAM macro structure with DV-SWT scheme to tackle the abovementioned challenges. A 16Mb DMc-ReRAM full-function macro was fabricated. SWT features were confirmed and a sub-14ns access time for CIM (AND/OR/XOR) operations was achieved. This access time is 86+x faster than previous ReRAM-based CIM works. This paper represents the first successful demonstration of a CIM ReRAM macro with ReRAM device and peripheral circuits fully integrated on the same testchip.

Acknowledgment

The authors gratefully acknowledge the financial and technical support of MOST-Taiwan, CIC and ITRI.

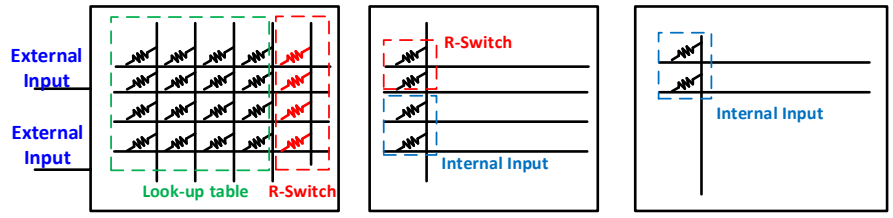
References

[1] S. Li, et al., *DAC*, p1-6, 2016. [2] F. Su, et al., VLSIC, pp. 260-261, 2017. [3] Y. Liu, et al., *ISSCC*, pp. 84-85, 2016. [4] [4] H. Li, et al., *IEDM*, p. 16.1.1-16.1.4, 2016 [5] B. Chen, et al., *IEDM*, p17.5.1-17.5.4, 2015. [6] S.-S. Sheu, et al., *ISSCC*, pp. pp. 200-201, 2016. [7] M.-F. Chang et al, *ISSCC*, pp. 332-333, 2014. [8] X.-Y. Xue, et al., VLSI Symp., pp. 42-43, 2012.



(a) (b)

Fig. 1. (a) von Neumann structure and (b) Computing-In-Memory enabled beyond von Neumann structure.



(a) (b) (c)

Fig. 2. CIM structures: (a) external inputs with ReRAM look-up table and output cells, (b) in-memory inputs with operand-setting cells, and (c) in-memory inputs without extra cells (proposed).

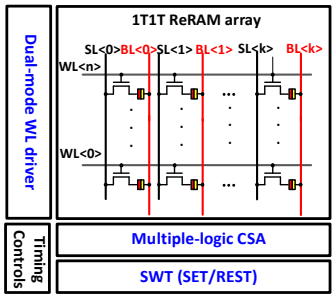


Fig. 3. Proposed Dual-Mode Computing ReRAM macro structure, including dual-mode WL driver, multiple-logic current-mode SA and SWT circuits (SWT-SET and SWT-RESET).

	This work	[4]	[5]
Capacity	16Mb	256 cells	128 cells
Computing type	Computing-in-memory	Computing-in-memory	Look-up table like
Input	Internal	Internal	External
R-switching	NO	YES	YES
Extra cell	NO	YES	YES
Area	1X	2X	10X
Function	AND/NAND、OR/NOR、XOR	XOR、XNOR	NOR AND(NOR base3 cycle) OR (NOR base2 cycle) XOR(NOR base4 cycle)
ON chip Peripheral Circuit	YES	NO (Device+ external testing equipment)	NO (Device+ PFGA)
Computing speed	<14ns (1X)	1.2us (86+X)	500s (3.5x10^{10}+X)

Table I. compares recent CIM works.

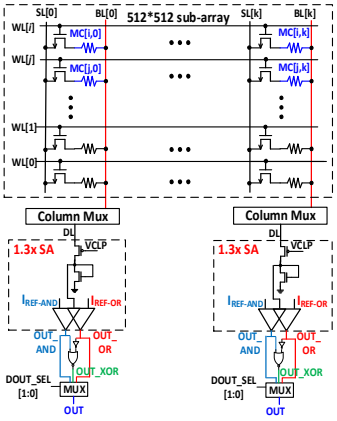


Fig. 4. Proposed CIM circuits.

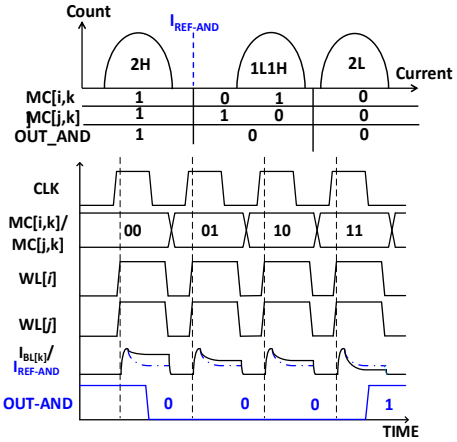


Fig. 5. (a) Concept and (b) waveform of AND operation in DMc macro.

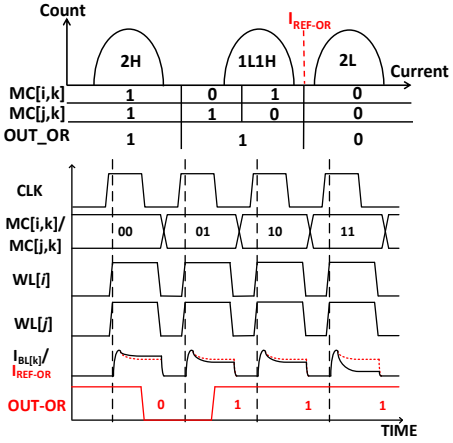


Fig. 6. (a) Concept and (b) waveform of OR operation in DMc macro.

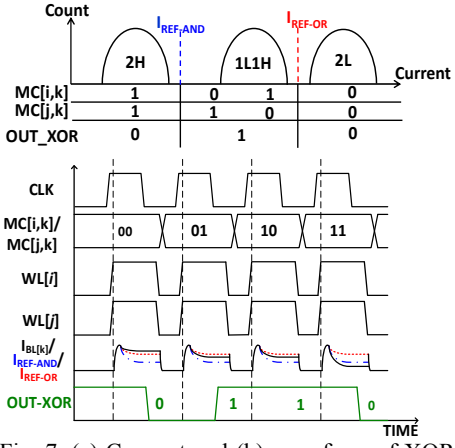
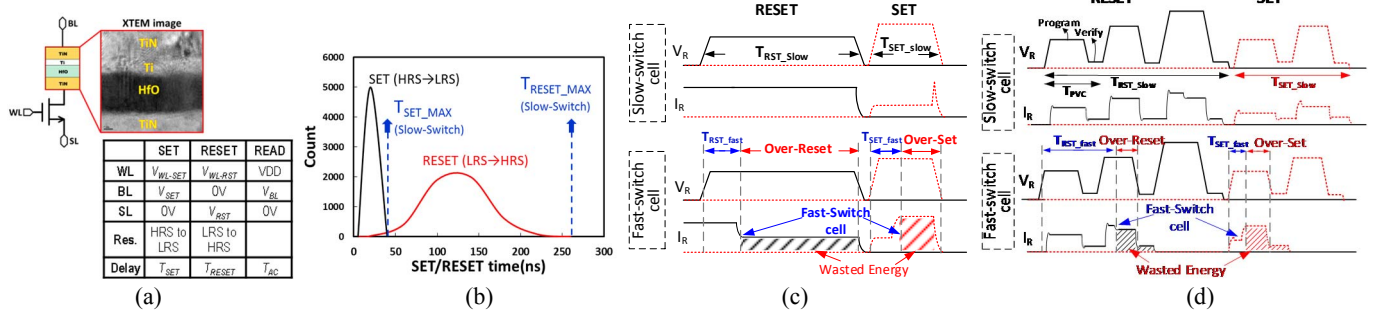


Fig. 7. (a) Concept and (b) waveform of XOR operation in DMc macro.



(a) (b) (c) (d)

Fig. 8. Employed bipolar 1T1R HfO ReRAM cell used in this work. (a) Structure and operation table. (b) Measured SET (T_{SET}) and RESET (T_{RESET}) time. (c) Macro-level write behavior using conventional one-pulse approach (worst-case coverage).(d) conventional program-verify scheme with switching in SL/BL and WL voltages between write and verify modes, which causes long switching time, large power overhead, and insufficient support for fine resolution in program-verify period.

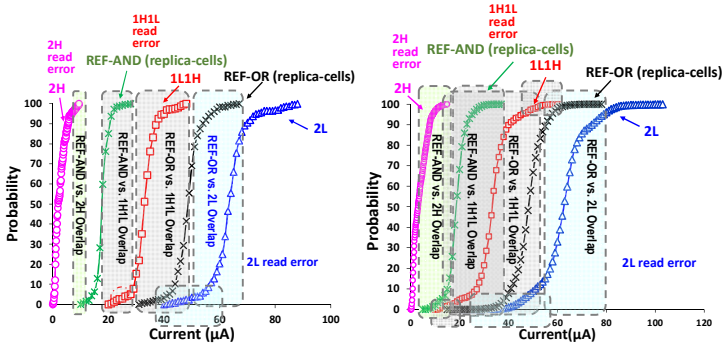


Fig. 9. (a) Measured (256 cells) and (a) predicted (6σ) current distribution of fabricated 16Mb macro for CIM operation without using DV-SWT scheme.

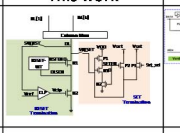
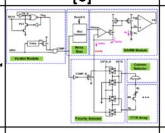
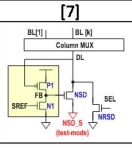
	This work	[8]	[7]
Schematic			
RESET- Termination	Voltage-mode	Current-mode	No
SET-Termination	Voltage-mode	Current-mode	Voltage-mode
Structure	RESET: 1OP+4SW+6T SET:5T	2OP+R+30T+ DelayUnit+ others	4T
Area	Middle(3x)	Large (>15X)	Small (1X)
DC-current during Write	RESET: Middle(2x) SET: no DC current	Large (>5x)	Small (1x)

Table II: Comparison between proposed DV-SWT and previous SET schemes.

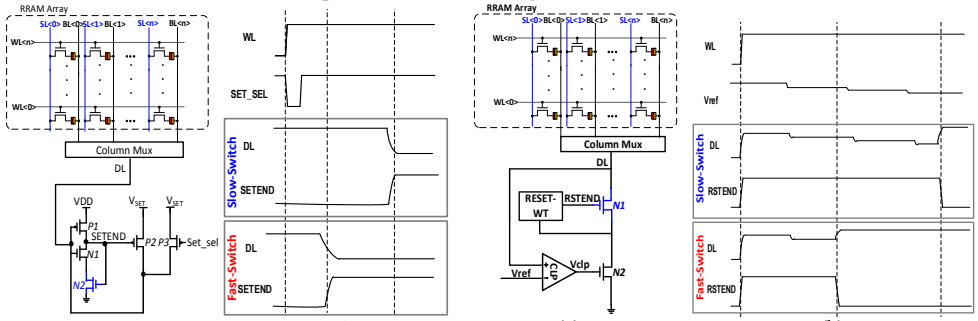


Fig. 10. Proposed DV-SWT scheme for SET operation (DV-SWT-SET): (a) circuit and (b) waveform.

Fig.11. Proposed DV-SWT scheme for RESET operation (DV-SWT-RESET): (a) circuit and (b) waveform.

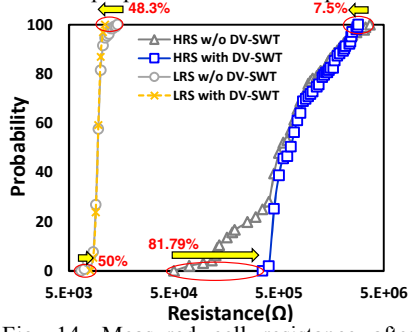


Fig. 14. Measured cell resistance after application of DV-SWT scheme.

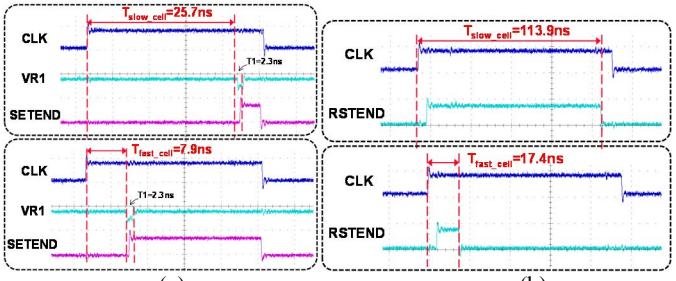


Fig. 13. Captured waveform of DV-SWT operation: (a) SET operation: the rising of SETEND of fast-switching cells occurs earlier than that of slow-switching cells, and (b) RESET operation: the falling of RSTEND of fast-switching cells occurs earlier than that of slow-switching cells. T1 is circuit response time.

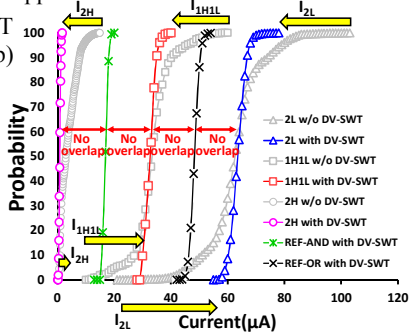


Fig. 15. BL current distribution for CIM operation using DV-SWT scheme

Process	150 nm low power	
Supply voltage	1.8V/3.3V	
Memory cell	HfO based RRAM	
Memory Capacity	16Mb	
Sub-array	Capacity: 256Kb BL Length=512 WL Length=512	
CIM access time (T_{AC}) for BL-length=512 @ VDD=1.8V	AND	13.1ns
	OR	13.7ns
	XOR	13.9ns

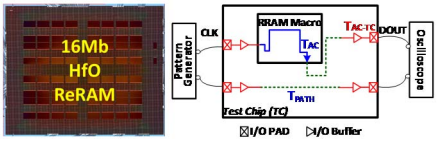


Fig.12. Chip photo and summary.

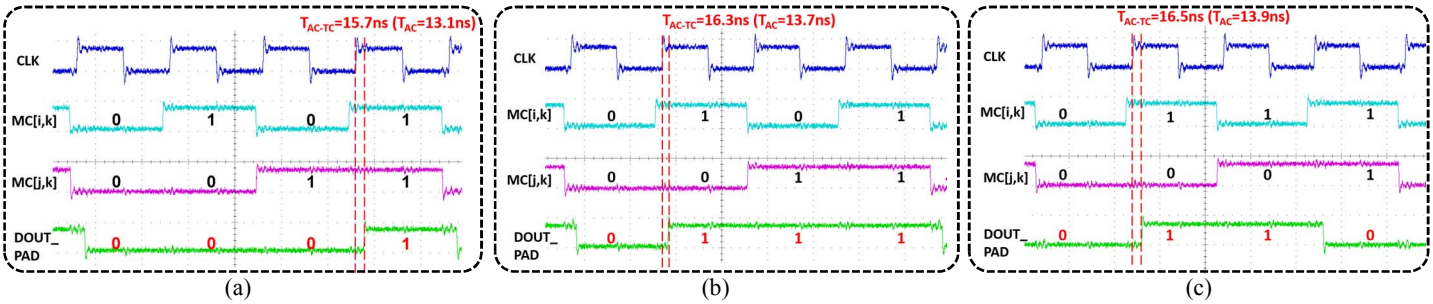


Fig. 16. Captured waveforms of CIM operation, including path-delay: (a) AND, (b) OR, and (c) XOR.