IEEE NVMSA 2021

The 10th IEEE Non-Volatile Memory Systems and Applications Symposium
August 18-20, 2021, Virtual Conference

A Non-volatile Computing-in-Memory ReRAM Macro using Two-bit Current-Mode Sensing Amplifier

Qiqiao Wu[a], Wenhao Sun[a], Junpeng Wang[a], Xuefei Bai[a], Feng Zhang[b], Song Chen[a] and Yi Kang[a]

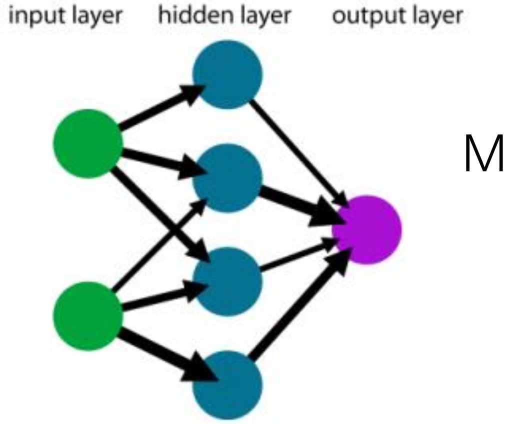
[a] School of Microelectronics, University of Science and Technology of China, Hefei, China

[b] Institute of Microelectronics of the Chinese Academy of Sciences, Beijing, China

Outline

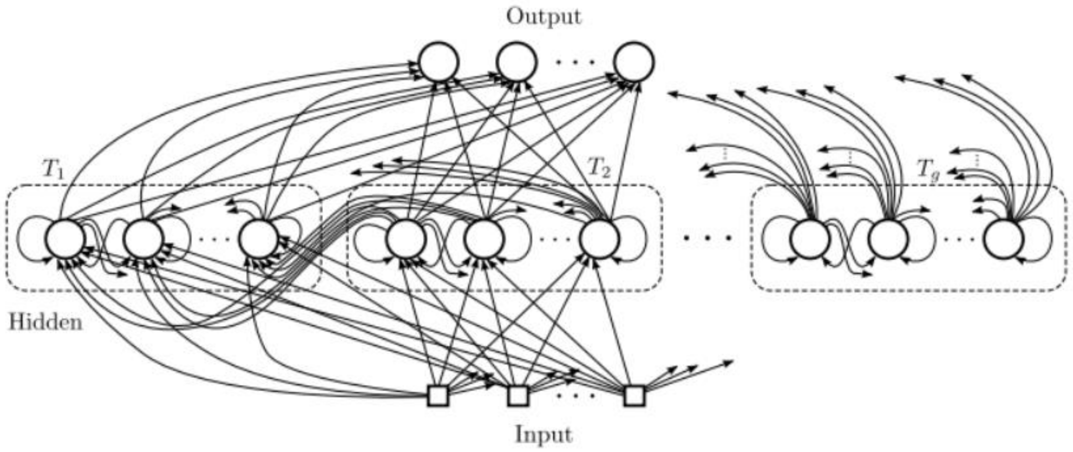
- Introduction
- Preliminary
- Architecture and circuits
- Layout and simulation
- Conclusion

Introduction

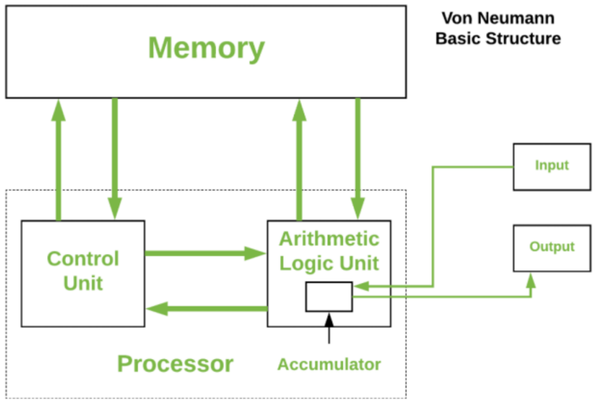


Multilayer perceptron

More data-intensive



Recurrent neural network



Traditional Von Neumann Structure

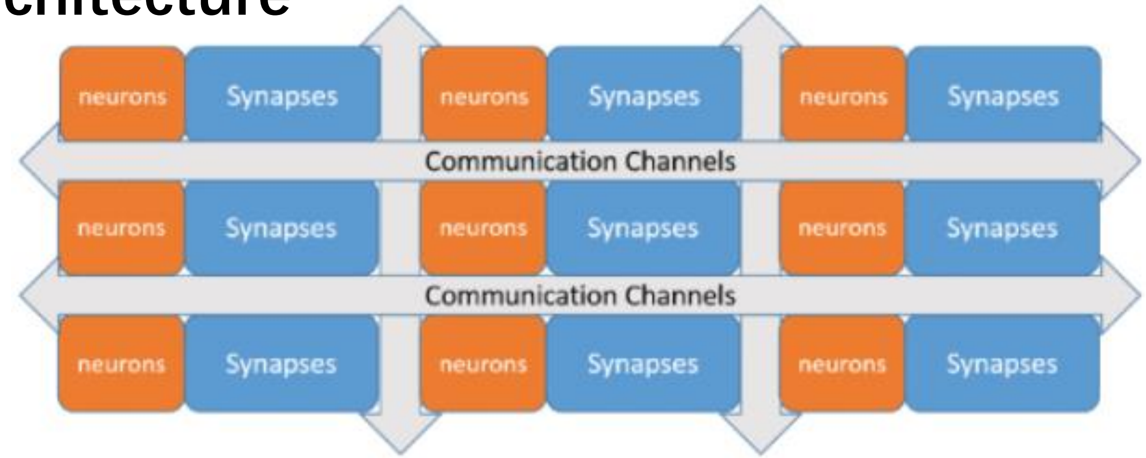
How to solve

memory wall problem

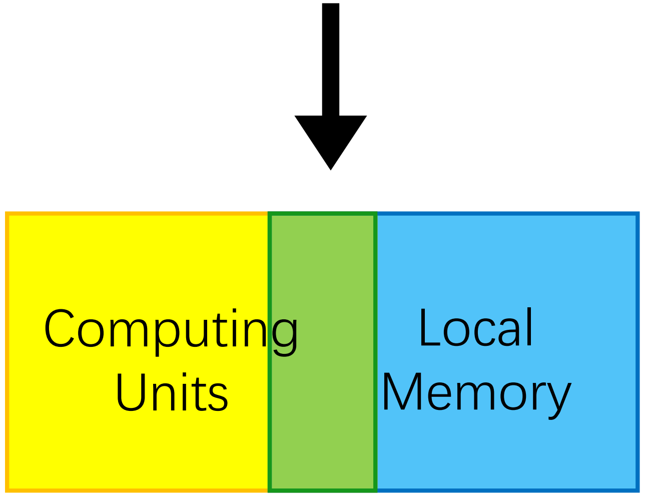
?

Introduction

Non-von Neumann architecture



Neuro-inspired Architecture [2]



Computing Units

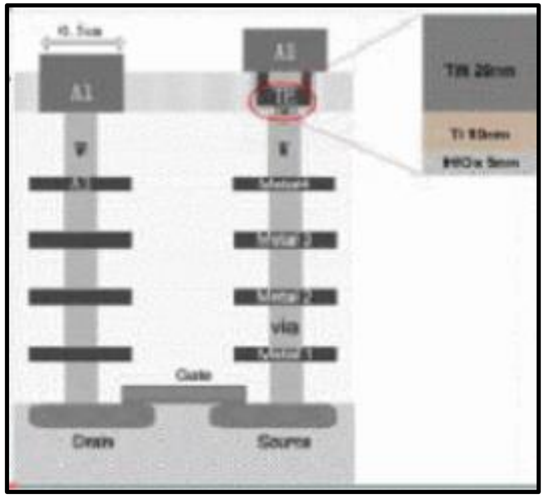
Local Memory

Non-volatile Memory

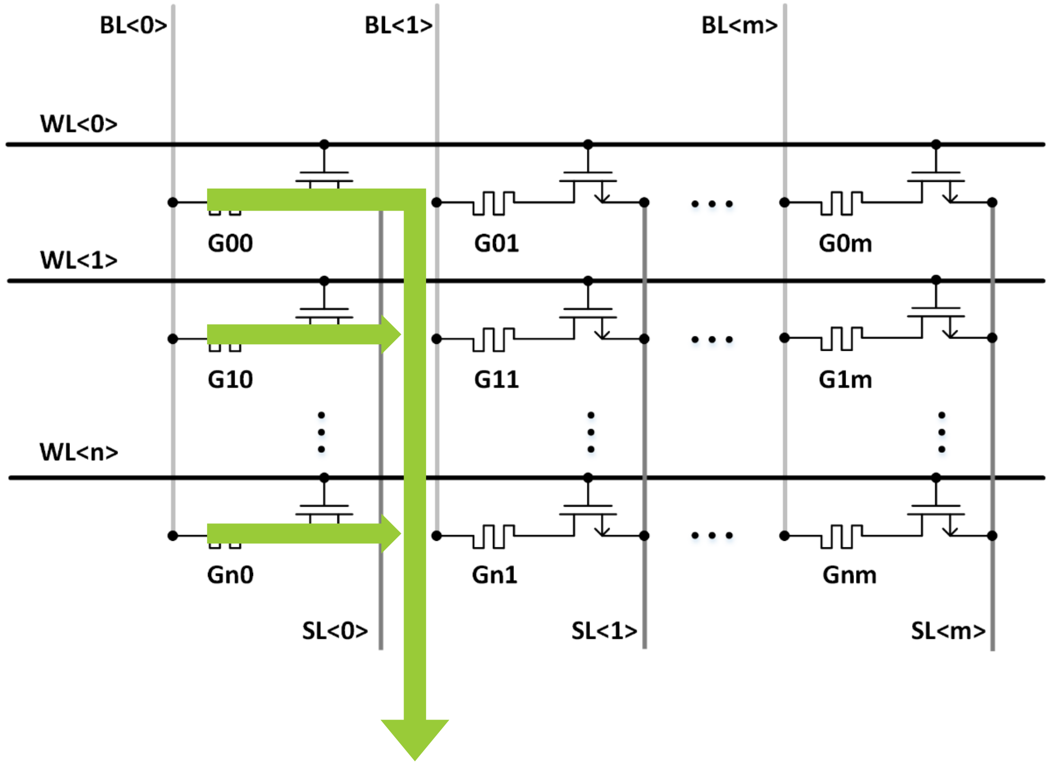
fuse together

Introduction

Resistance random access memory



Embedded ReRAM,
CMOS Technology



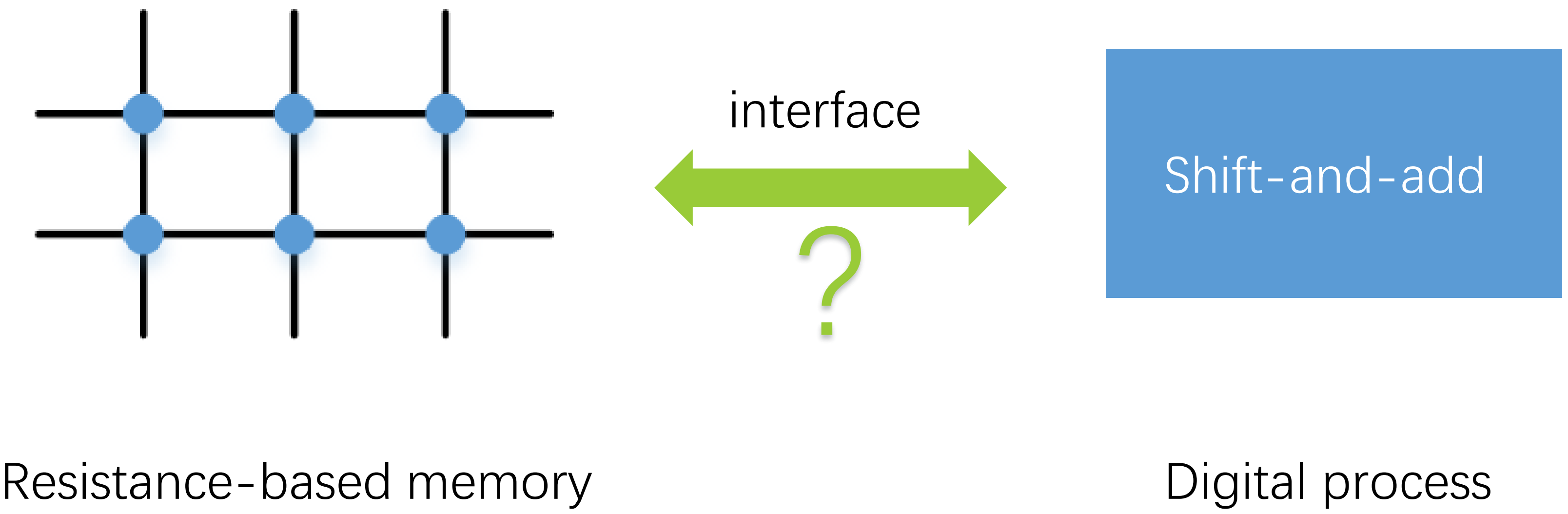
1T1R structure
Multiply and accumulate

- CNN inference [4]
- Binary DNN inference [5]
- Neural network training [6]
- ……

Application

Introduction

main challenge



interface

?

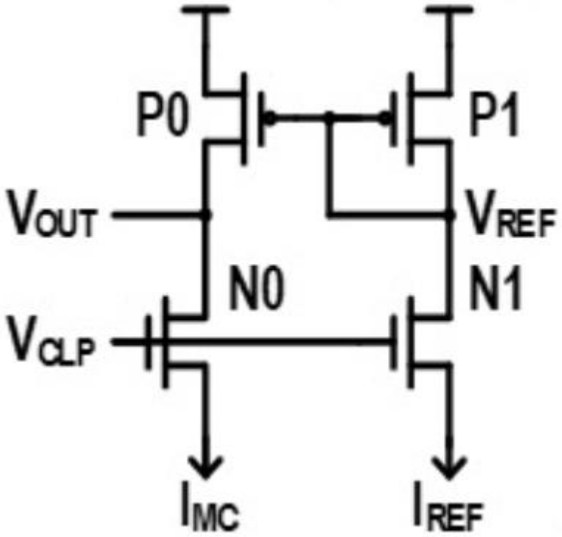
Shift-and-add

Resistance-based memory

Digital process

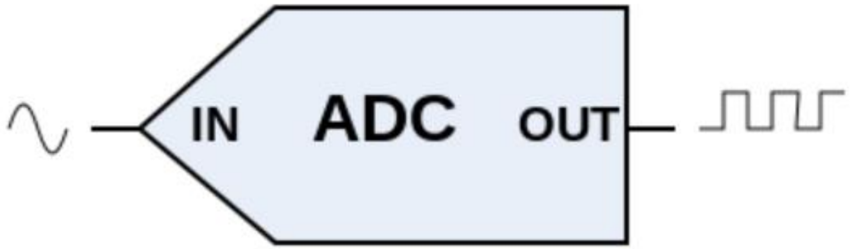
Introduction

current-mode sensing amplifier



Conventional CSA
(SRAM, binary nvRAM)

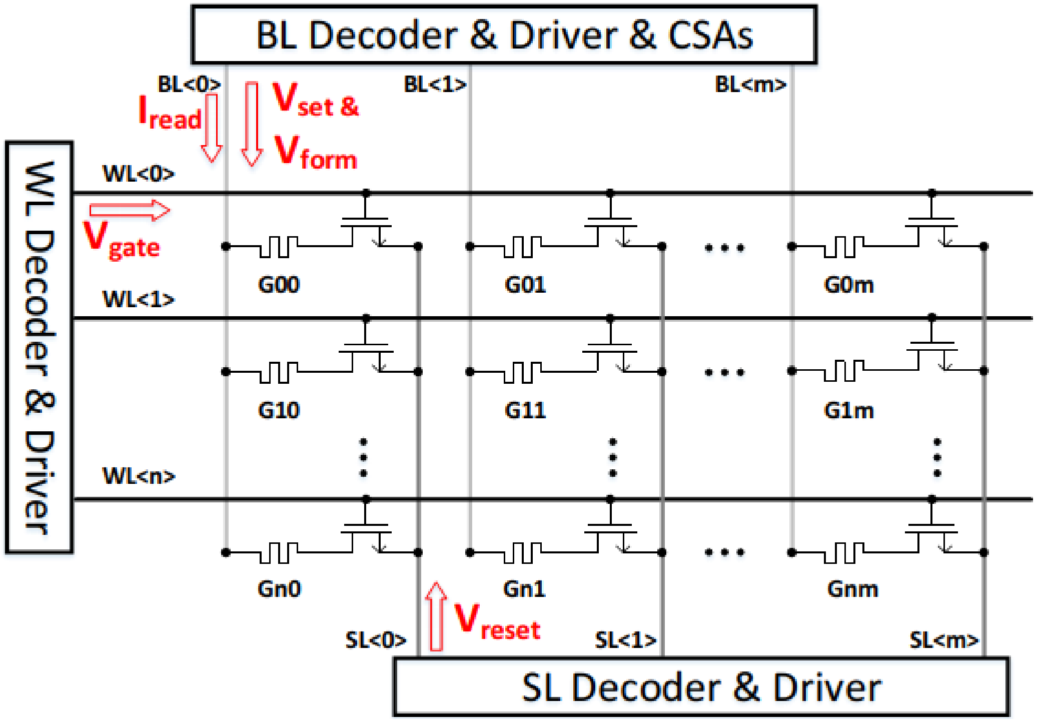
Analog-to-digital converter



(Multi-value nvRAM)

Preliminary

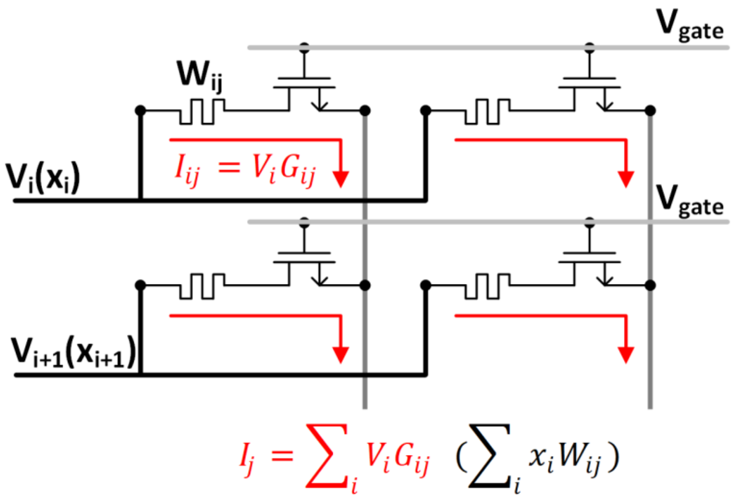
1T1R architecture with decoders, drivers and CSAs



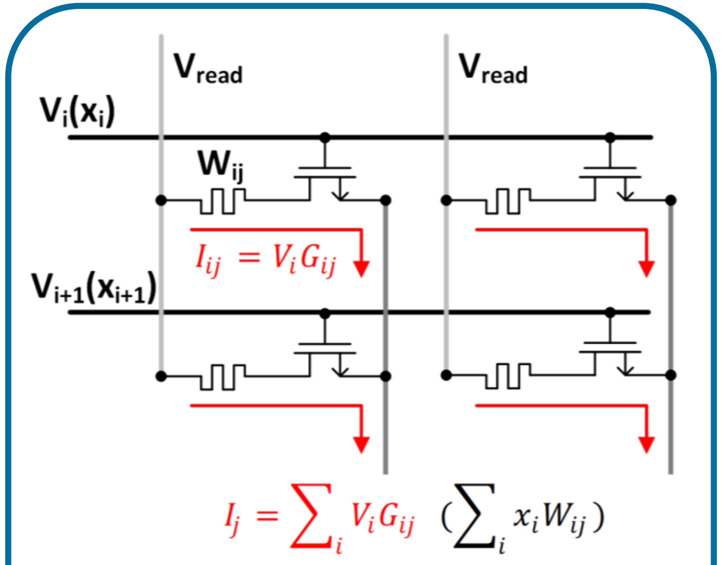
A smaller array size will ensure ReRAM device operation voltage does not exceed the limit voltage range of the CMOS technology node

Preliminary

Different topology for CIM



(a) conventional array topology



(b) low-power array topology

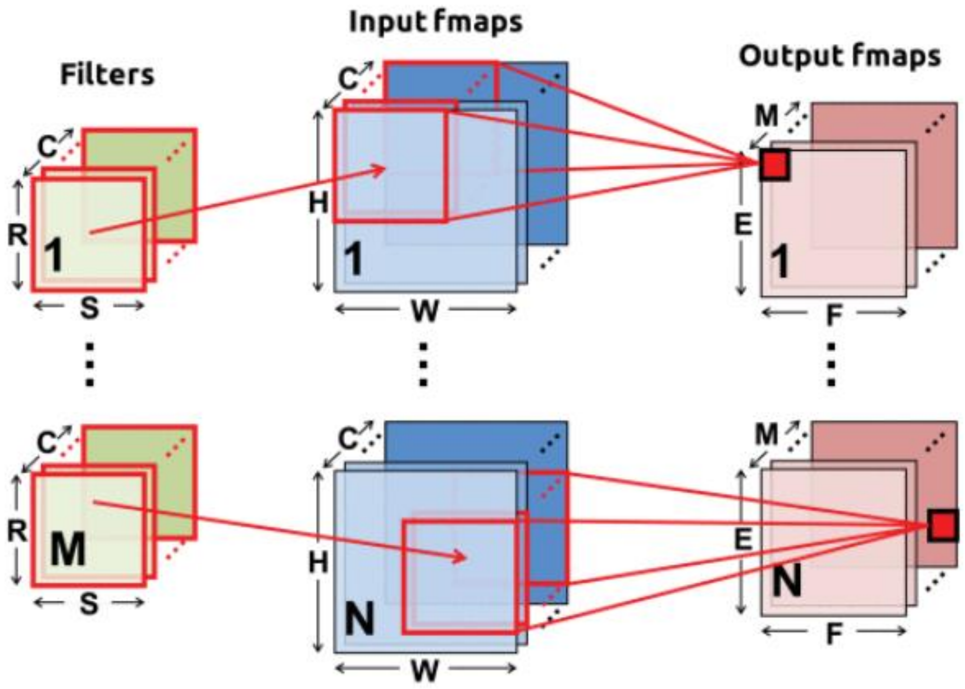
Used in combination with ADC/DAC

Used in combination with CSA

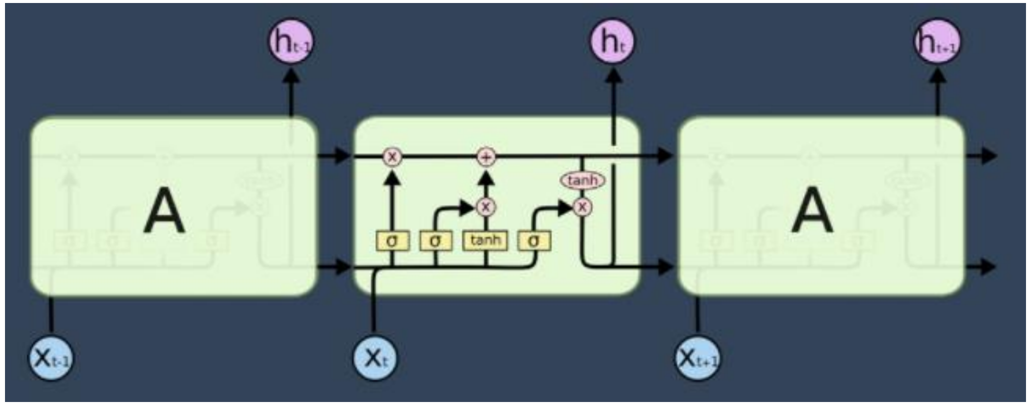
Preliminary

CNN acceleration

LSTM acceleration



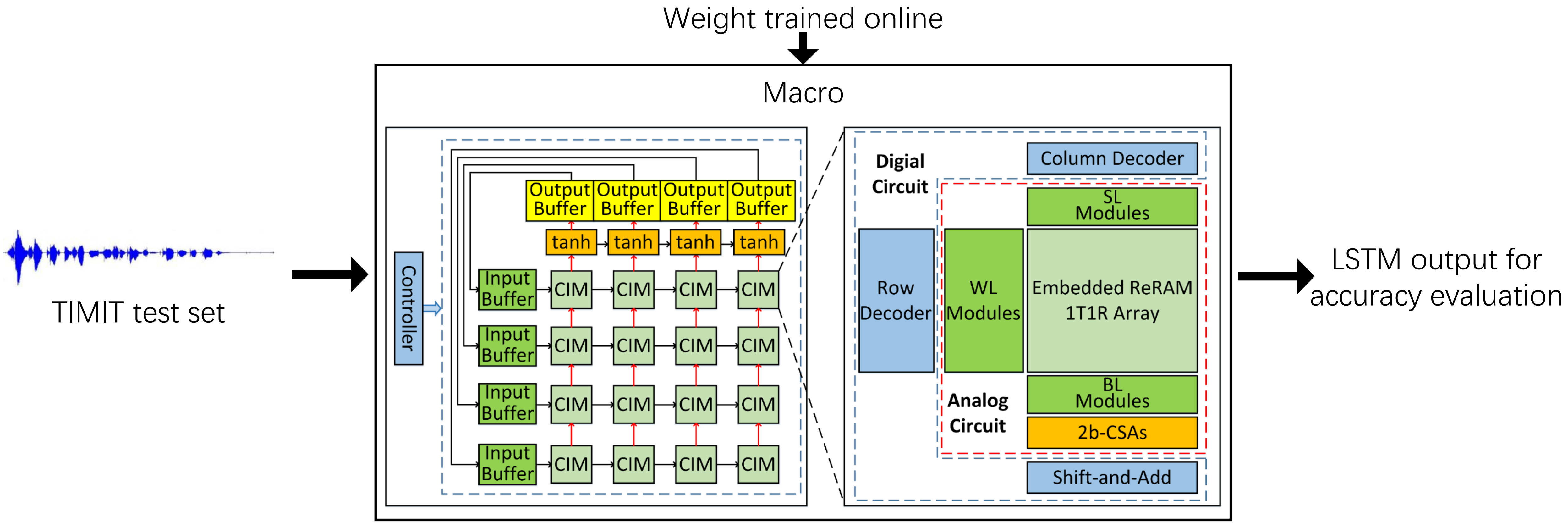
Data reuse



Require： more memory bandwidth
Performance： restricted by memory performance

Architecture and circuits

Overall structure



Weight trained online

Macro

TIMIT test set

Controller

Output Buffer
Output Buffer
Output Buffer
Output Buffer

tanh
tanh
tanh
tanh

Input Buffer
CIM
CIM
CIM
CIM

Input Buffer
CIM
CIM
CIM
CIM

Input Buffer
CIM
CIM
CIM
CIM

Input Buffer
CIM
CIM
CIM
CIM

Digial Circuit

Column Decoder

SL Modules

Row Decoder

WL Modules

Embedded ReRAM 1T1R Array

BL Modules

Analog Circuit

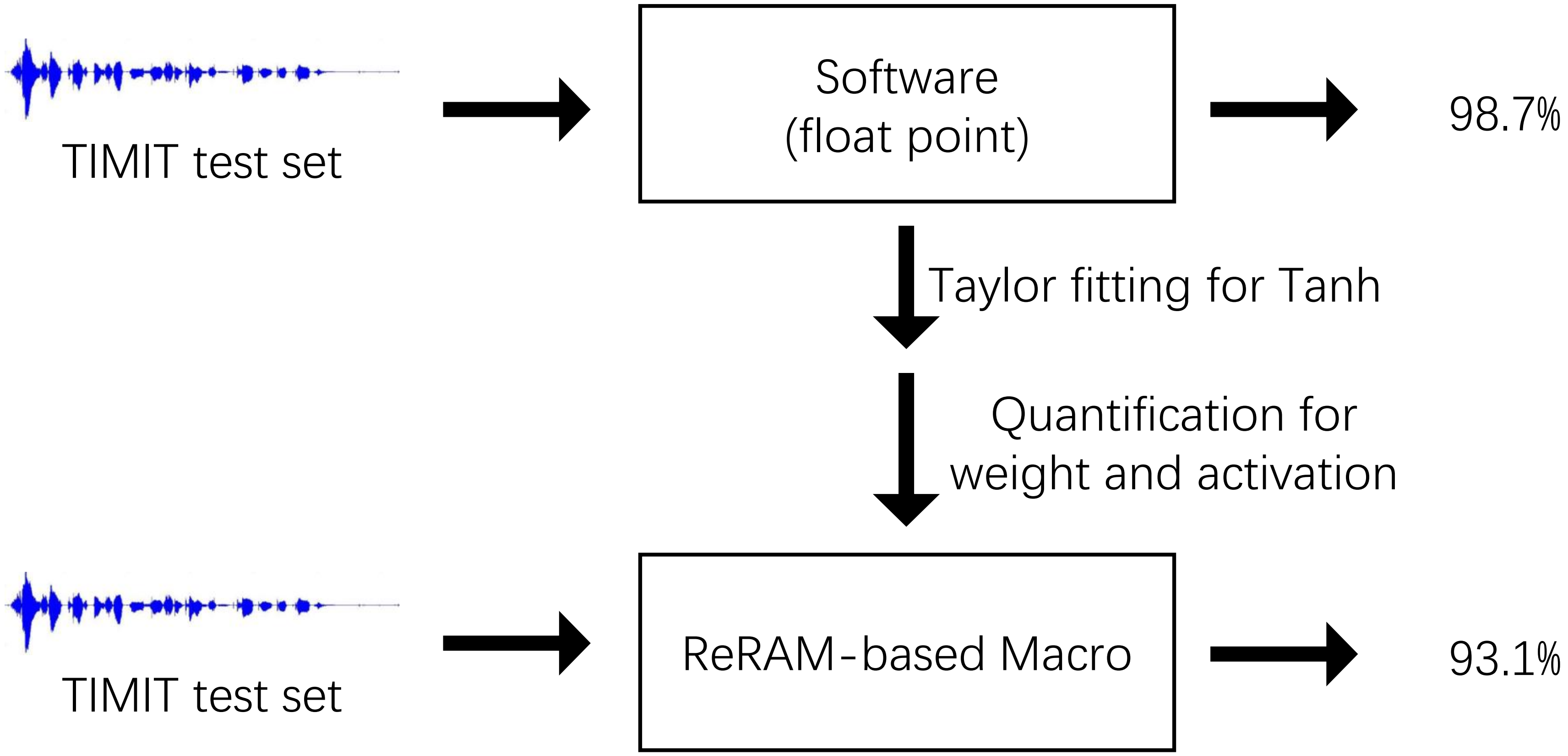
2b-CSAs

Shift-and-Add

LSTM output for accuracy evaluation

Architecture and circuits

Accuracy evaluation



TIMIT test set

Software
(float point)

98.7%

Taylor fitting for Tanh

Quantification for
weight and activation

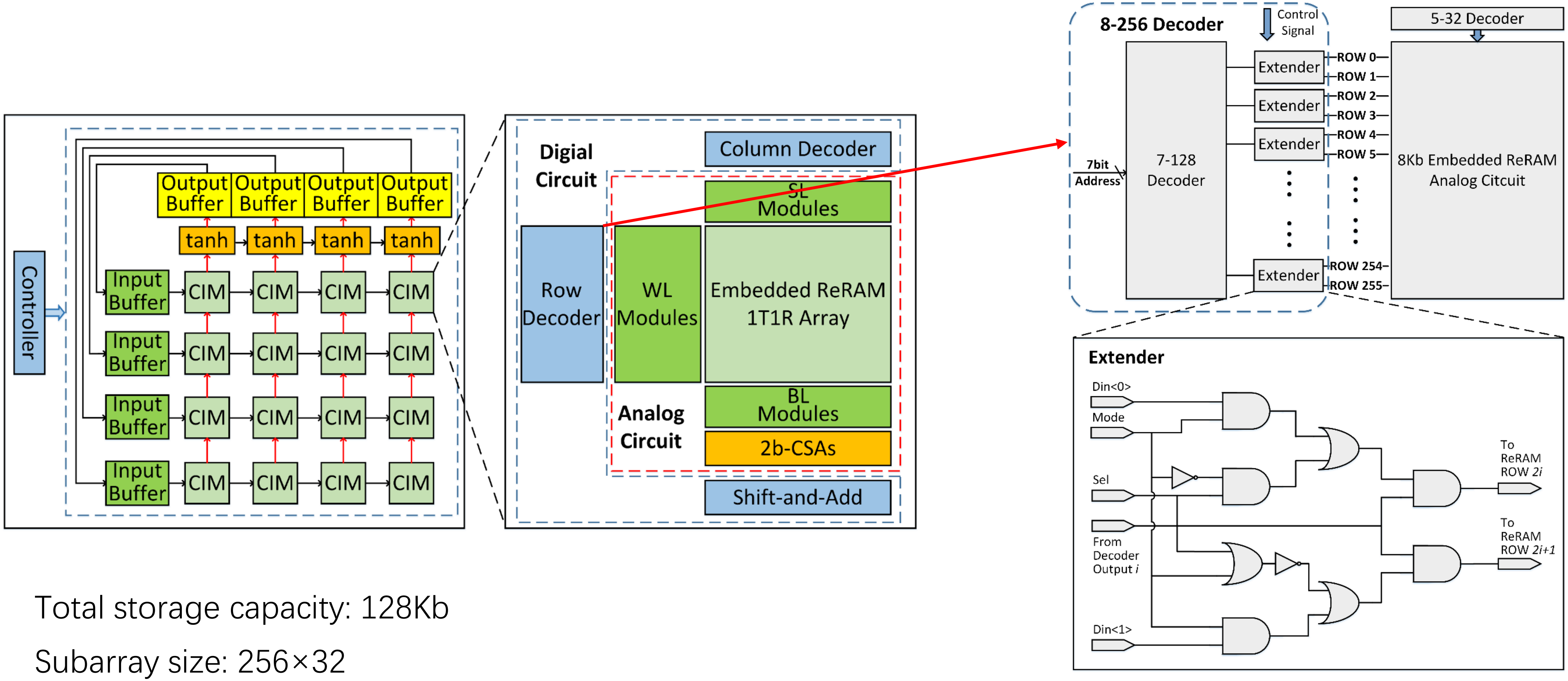
TIMIT test set

ReRAM-based Macro

93.1%

Architecture and circuits

Overall ReRAM-based architecture

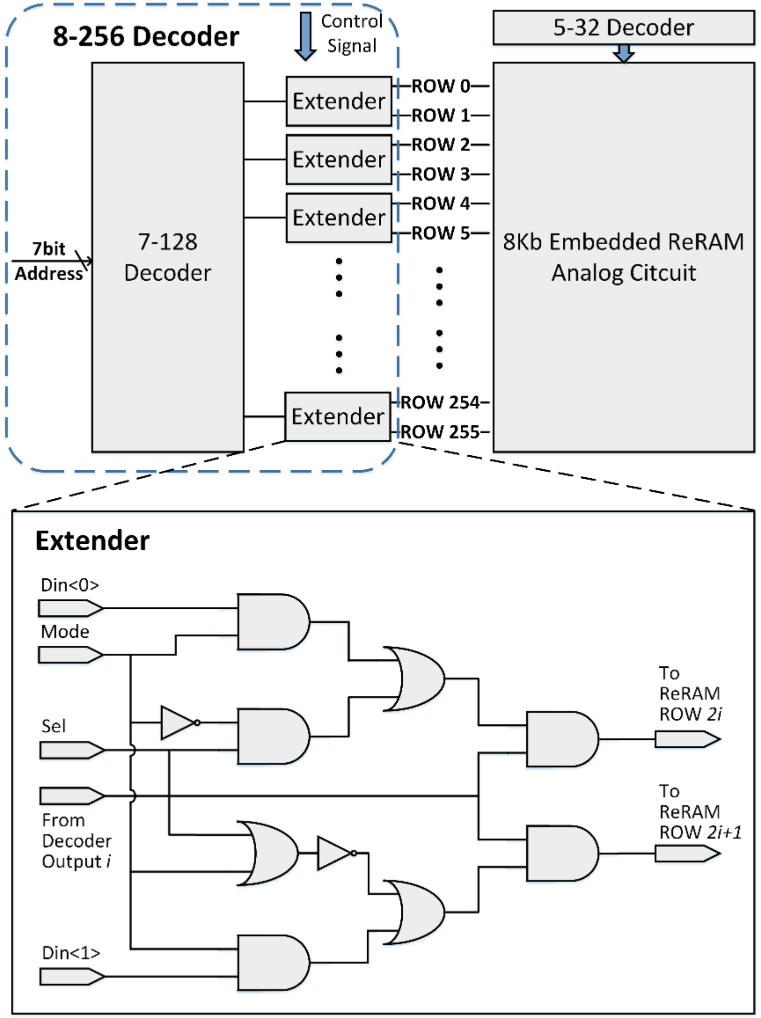


Total storage capacity: 128Kb

Subarray size: 256×32

Architecture and circuits

Decoder with extenders

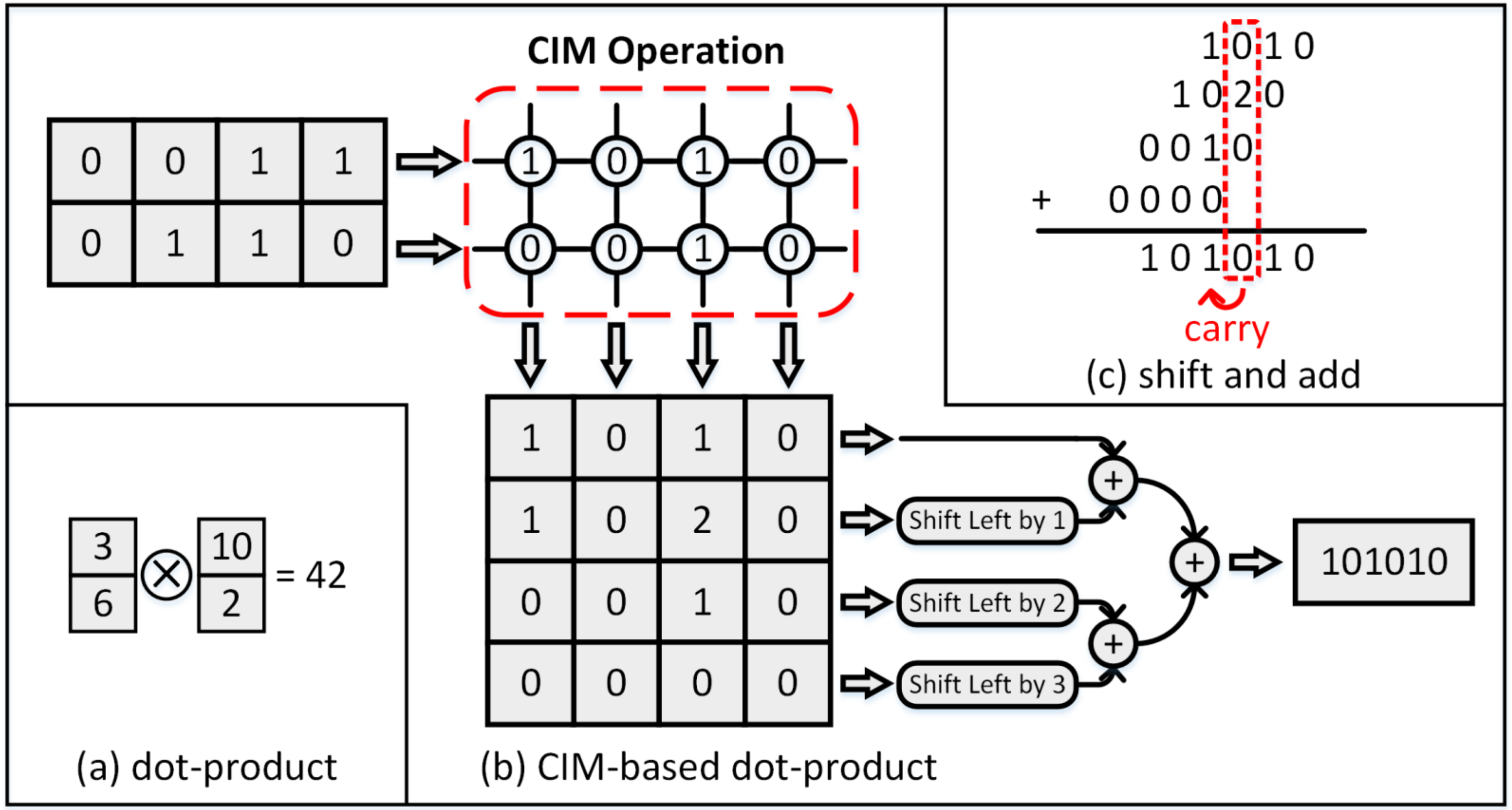


CIM mode function table

Din<1>	Din<0>	ROW 2i	ROW 2i+1
0	0	off	off
0	1	off	on
1	0	on	off
1	1	on	on

Architecture and circuits

A computation example based on CIM operation



(a) dot-product

(b) CIM-based dot-product

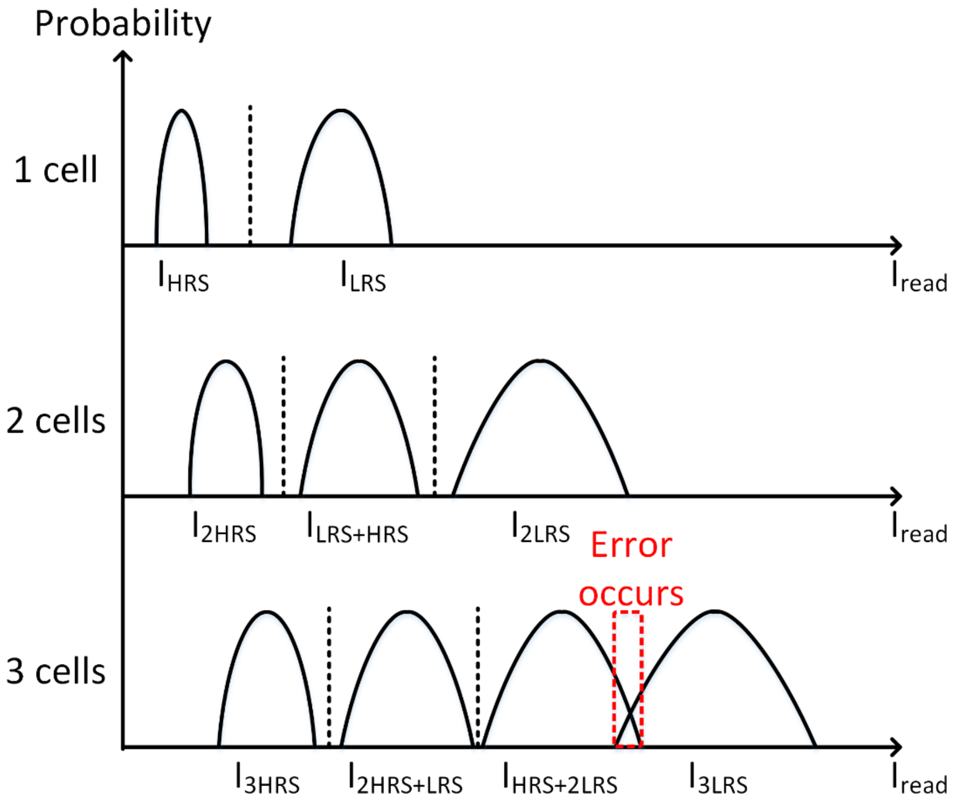
(c) shift and add

Basic formula: $3*10 + 6*2 = 42$

CIM formula: $2^0*10 + 2^1*(10+2) + 2^2*2 + 2^3*0 = 42$

Two-bit CSA

Read current analysis with ReRAM device dispersion

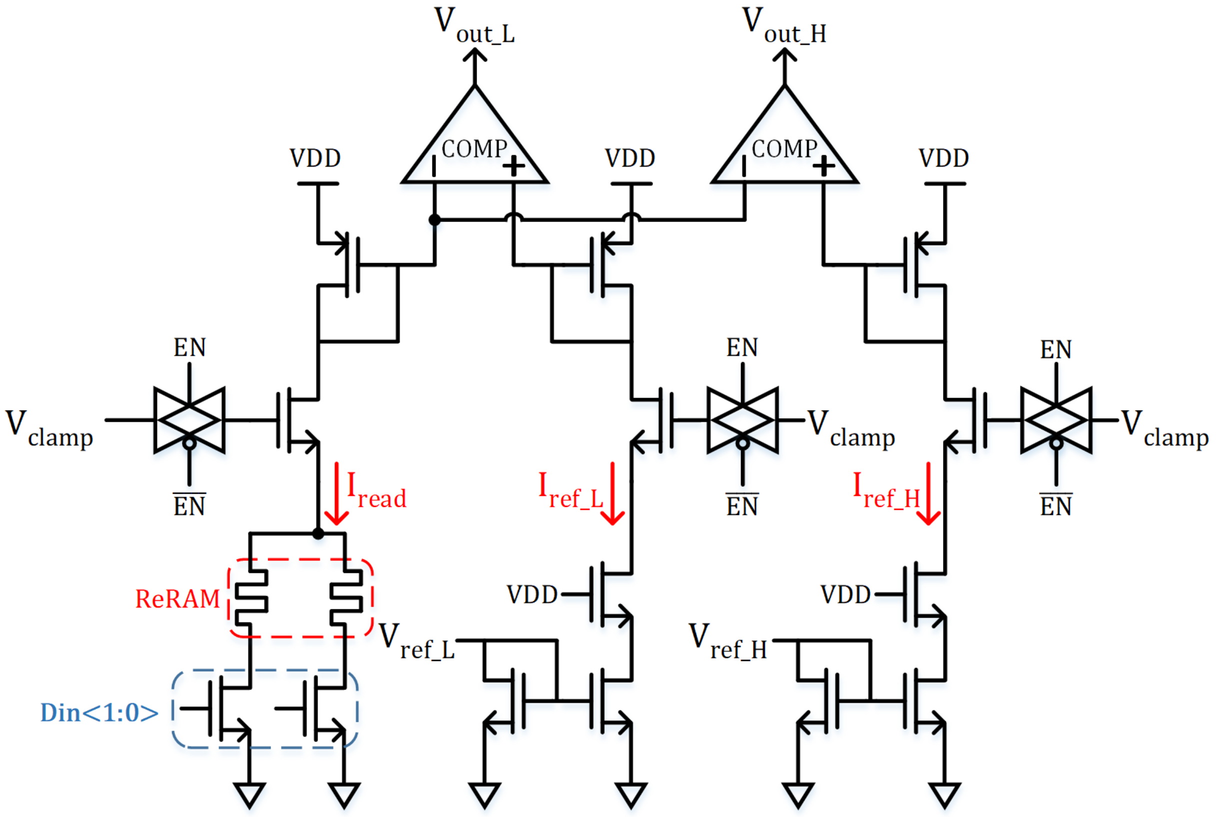


➡ More than two cells select on

The read current fluctuation of low resistance state device at 72K ohms is 22.9%
The read current fluctuation of high resistance state device at 530K ohm will reach 43.7% [17]

Two-bit CSA

Sensing circuit of 2b-CSA

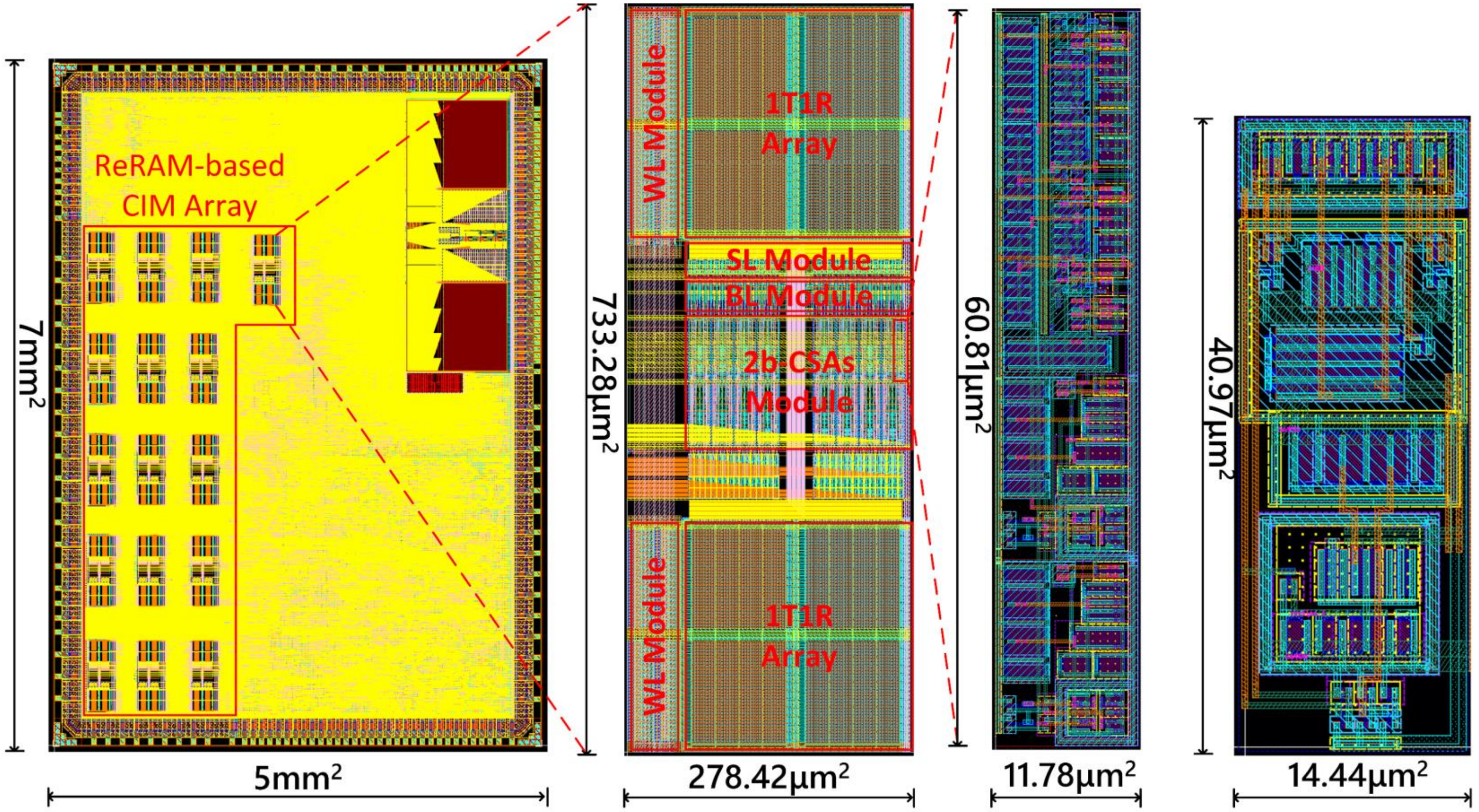


2b-CSA output coding scheme

V_{out_H}	V_{out_L}	Encoded data
0	0	0
0	1	1
1	0	invalid
1	1	2

Layout and simulation

Layout design and area comparison between 2b-CSA with traditional CSA



(a) The entire chip layout

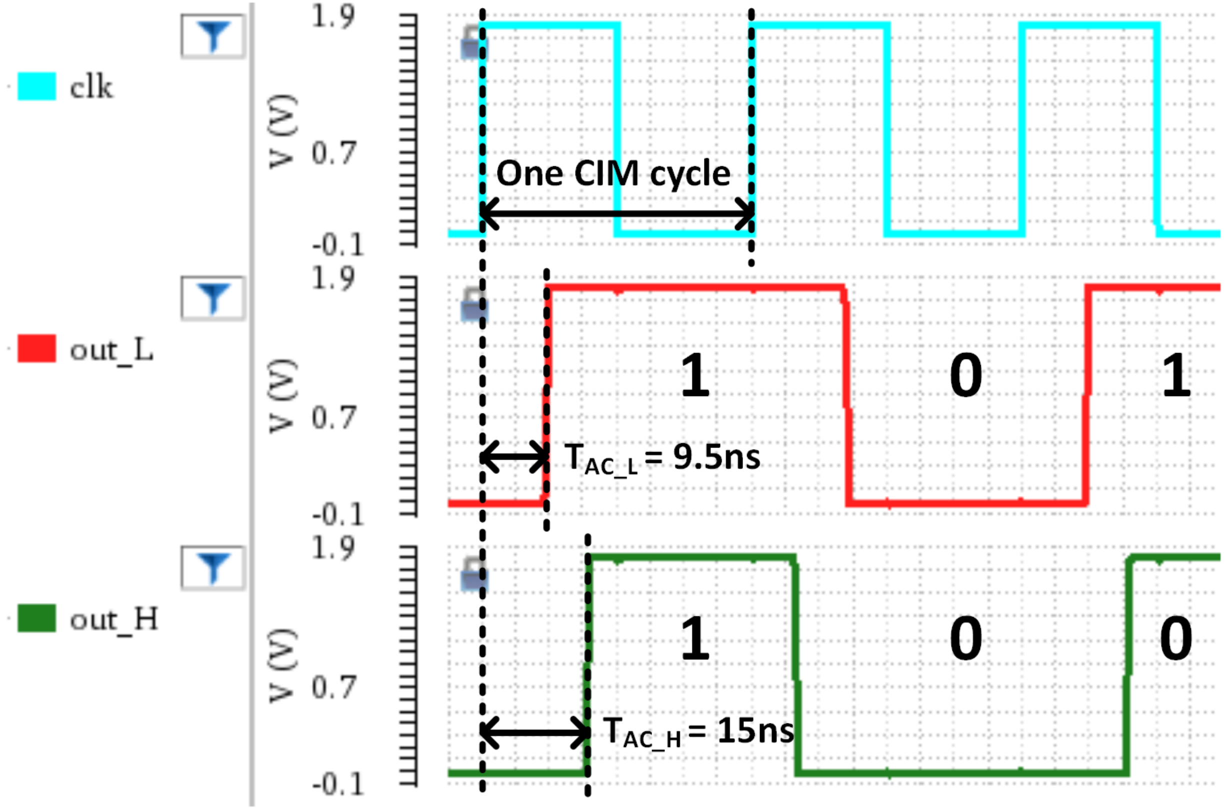
(b) ReRAM CIM layout

(c) 2b-CSA layout

(d) 1b-CSA layout referred to[8]

Layout and simulation

ReRAM-based CIM function simulation



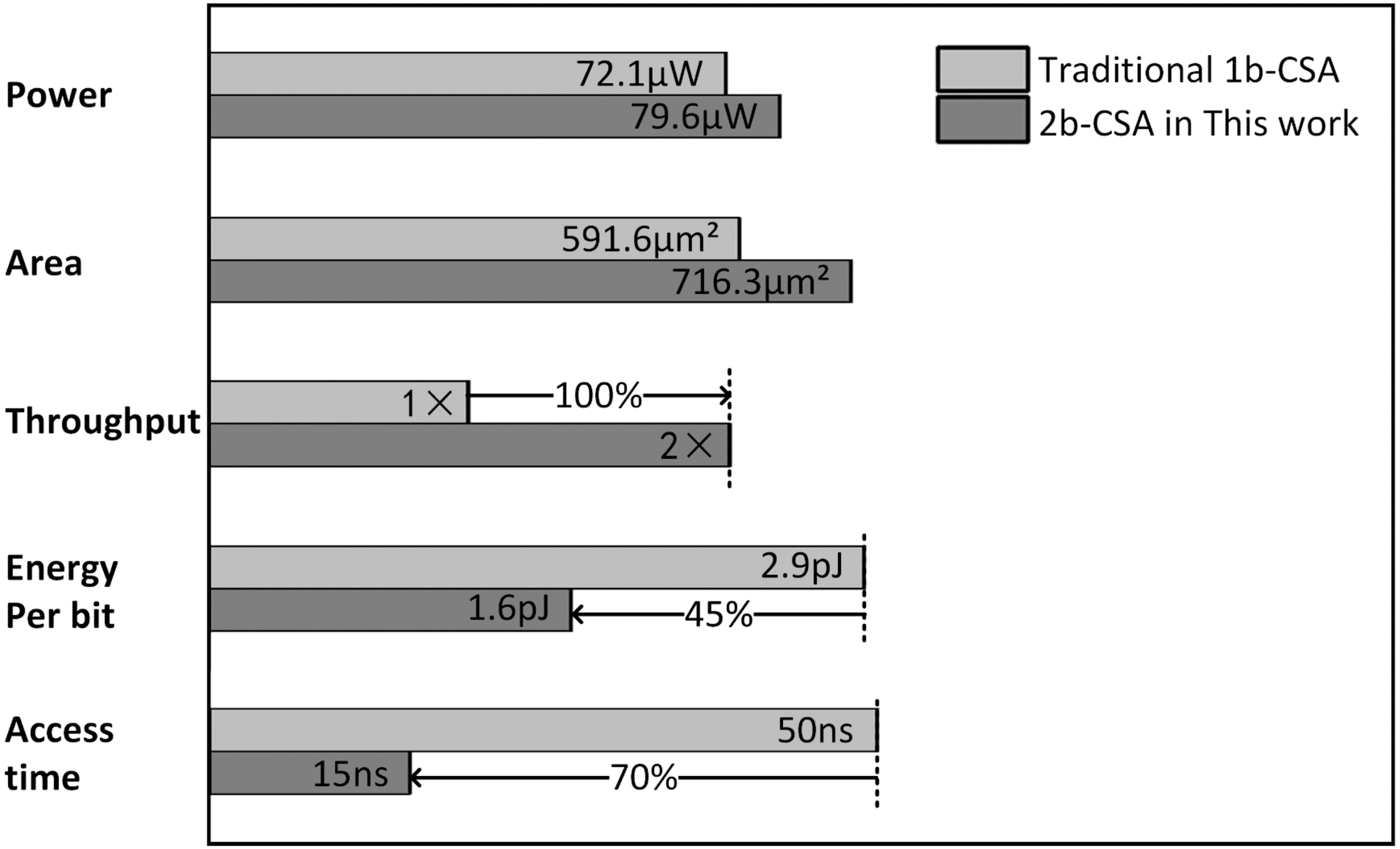
First CIM cycle: (1,1) dot product (1,1), and simulation result is (1,1)
Second CIM cycle: (1,1) dot product (0,0), the result of the simulation is (0,0)

Layout and simulation

IEEE NVMSA 2021

The 10th IEEE Non-Volatile Memory Systems and Applications Symposium
August 18-20, 2021, Virtual Conference

Power&Area&Throughput&Energy per bit analysis



Conclusion

- We consider the dispersion of ReRAM devices and realize the bit-vector matrix multiplication in the two-row mode, propose a ReRAM-based CIM architecture, including the decoder with extenders and 2b-CSA

- Compared with 1b-CSA, 2b-CSA in this work improves throughput, dramatically reduces operating energy consumption per bit and access time with a minor increase in power consumption and area

References

[2] S. Yu, “Neuro-inspired computing with emerging nonvolatile memory,” *Proceedings of the IEEE*, vol. 106, no. 2, pp. 260 – 285, 2018.

[4] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, “Fully hardware-implemented memristor convolutional neural network,” *Nature*, vol. 577, no. 7792, pp. 641 – 646, 2020.

[5] W. H. Chen, K. X. Li, W. Y. Lin, K. H. Hsu, and M. F. Chang, “A 65nm 1mb nonvolatile computing-in-memory reram macro with sub-16ns multiply-and-accumulate for binary dnn ai edge processors,” in *2018 IEEE International Solid – State Circuits Conference – (ISSCC)*, 2018.

[6] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, and G. W. Burr, “Equivalent-accuracy accelerated neural-network training using analogue memory,” *Nature*, vol. 558, no. 7708, 2018.

[8] Z. Feng, D. Fan, D. Yuan, L. Jin, and M. F. Chang, “A 130nm 1mb hfox embedded rram macro using self-adaptive peripheral circuit system techniques for 1.6x work temperature range,” in *2017 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2017.

[17] D. Dong, L. Jing, Y. Wang, X. Xu, and M. Liu, “The impact of rtn signal on array level resistance fluctuation of resistive random access memory,” *IEEE Electron Device Letters*, vol. PP, no. 99, pp. 1 – 1, 2018.

Thank You

Qiqiao Wu (wuqiqiao@mail.ustc.edu.cn)

Wenhao Sun (wh1997@mail.ustc.edu.cn)

Junpeng Wang (wjp97@mail.ustc.edu.cn)

Xuefei Bai (baixf@ustc.edu.cn)

Feng Zhang (Zhangfeng_ime@ime.ac.cn)

Song Chen (songch@ustc.edu.cn)

Yi Kang (ykang@ustc.edu.cn)