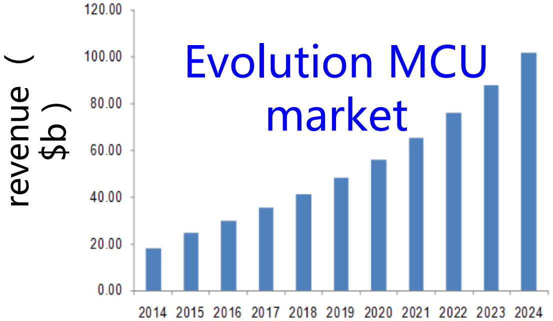
The future of RRAM : From Embedded Application to In Memory Computing and Beyond

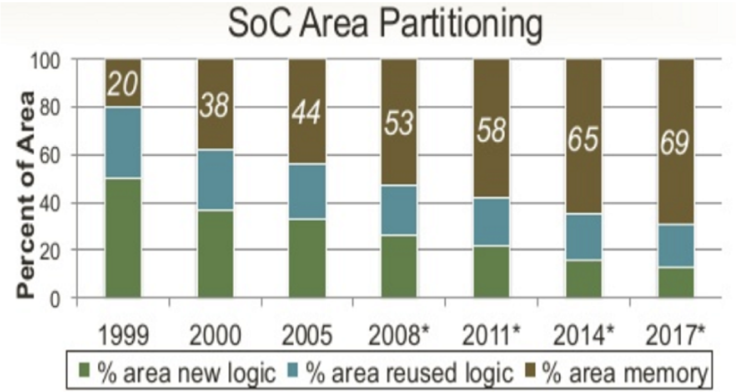
Jianguo Yang

Key Laboratory of Microelectronics Devices and Integrated Technology, Institute of Microelectronics of the Chinese Academy of Sciences, China

Embedded Memory Application Scenarios

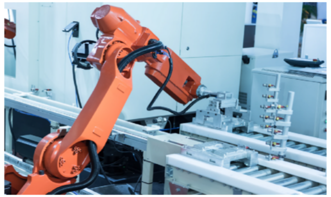
- The System-on-a-chip (SOC) is widely used in IoT, industrial, Intelligent Edge Devices etc.
- Embedded memory is a basic component of the SOC, accounting for more than 70% area.



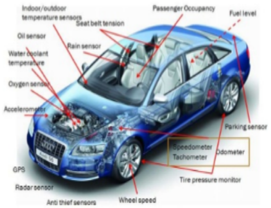


IoT

Industrial



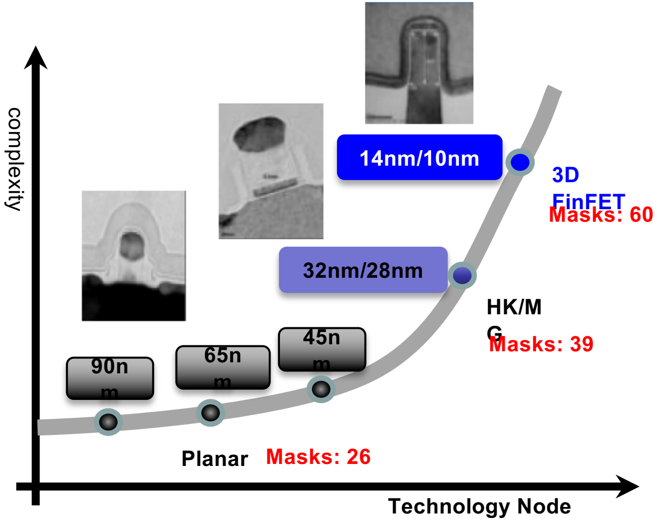
Autonomous Cars

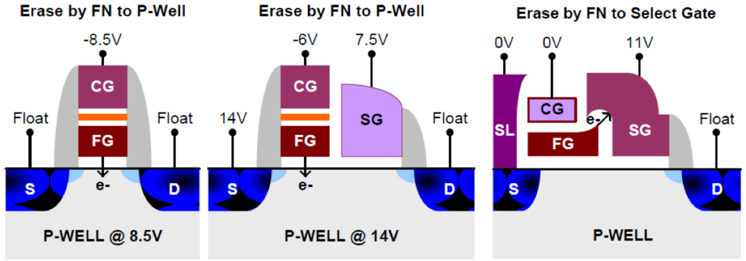


Consumer Electronics

Scaling Challenges

- eFlash is facing major scaling challenges due to rising fabrication complexity/costs for technology nodes ≤ 28nm
- In high-end processors and mobile AP will occur later due to more strict scalability requirements (≤ 14nm).





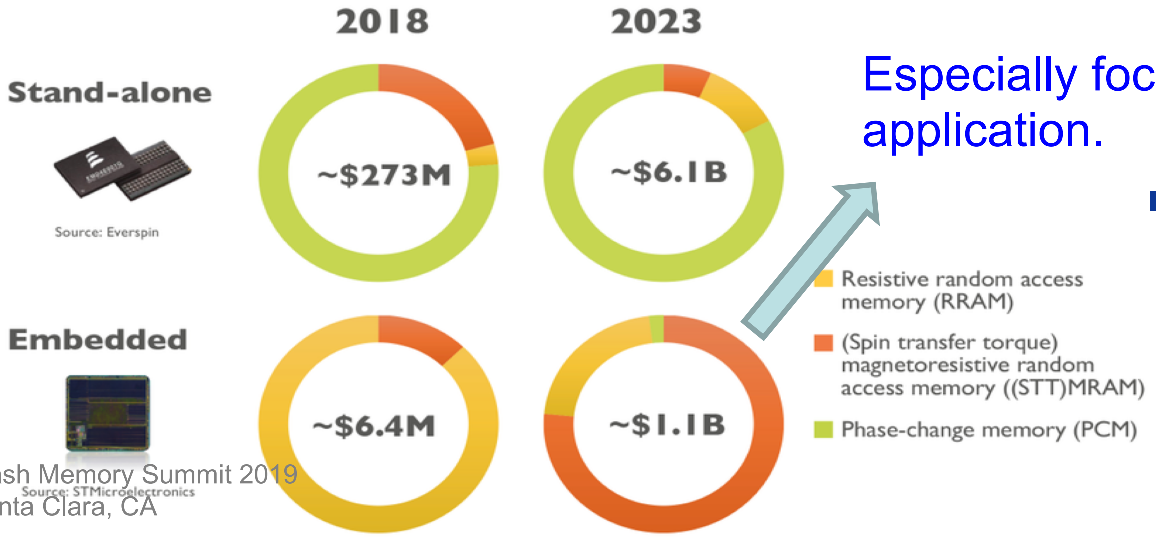
eFlash scaling challenges in 28nm and below :

- Extra 9-12 masks , high cost
- Scaling lead low reliability
- Hard to integrate with logic process

➡ New embedded memory technologies at advanced process nodes are needed!

Evolution of the Emerging Non-volatile Memory Market

- Compared to stand alone, the embedded emergingNVM market is relatively small,
- A few RRAM-based microcontrollers (MCUs) are available on the market
- All top foundries are now getting gready with 28/22nm technology processes for STT-MRAM



Source: Everspin

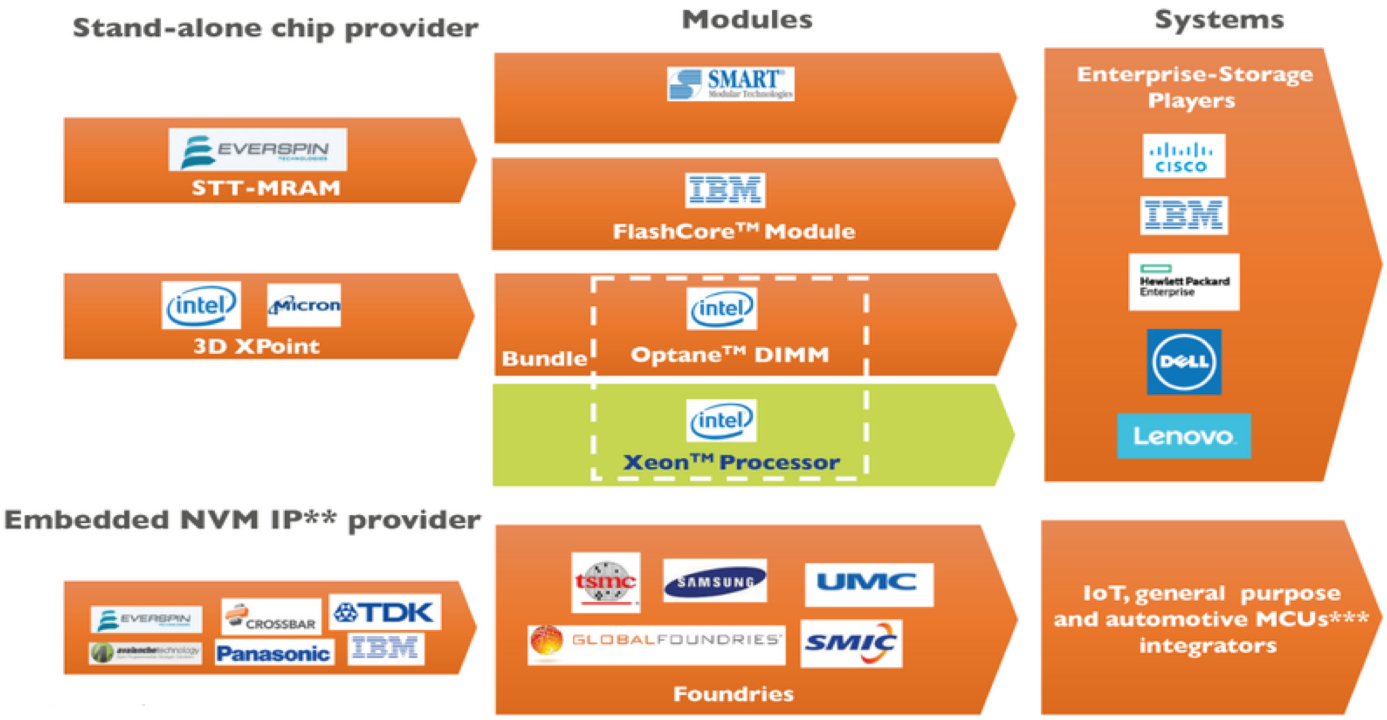
Source: STMicroelectronics

Especially focus on MCU application.

- STT-MRAM will be the first to take-off in the coming years and will lead the embedded emerging NVM market segment, then RRAM

Source: Yole development

Market Entry Strategies



Stand-alone chip provider

STT-MRAM

3D XPoint

Modules

FlashCore™ Module

Bundle

Optane™ DIMM

Xeon™ Processor

Systems

Enterprise-Storage Players

Embedded NVM IP** provider

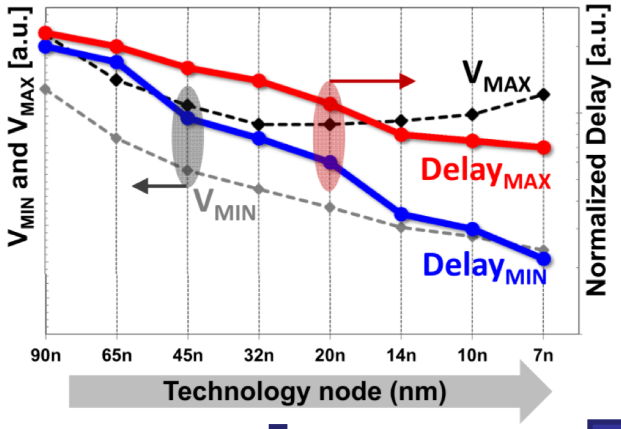
Foundries

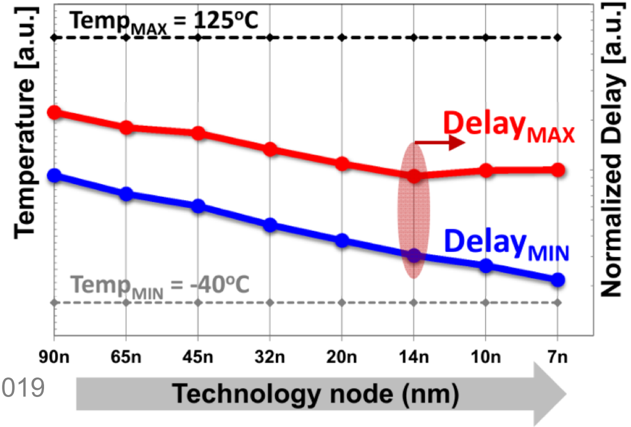
IoT, general purpose and automotive MCUs*** integrators

Source: Yole 2018

- Both small companies and big companies are focusing on emerging embedded memory
- In the embedded business, the top foundries including big IDMs are the key players

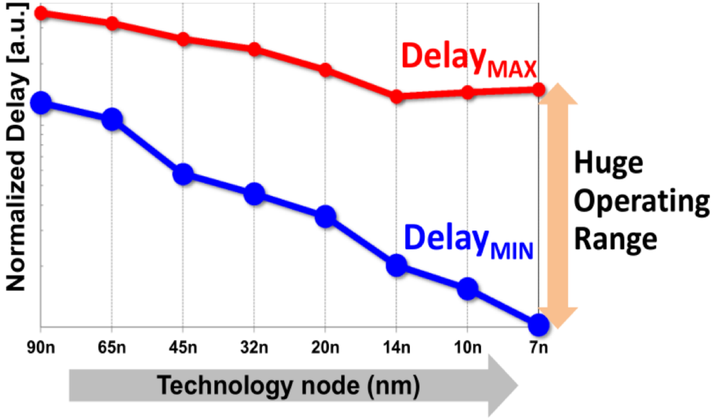
SOC Voltage and Temperature Trend





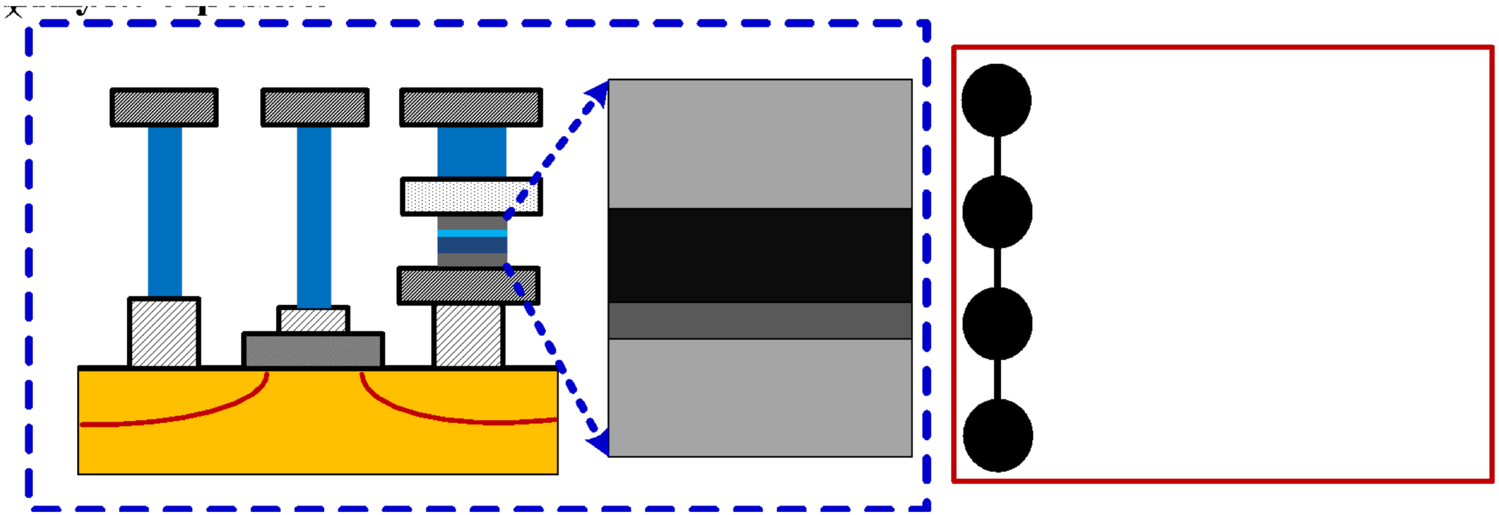
Source:ISSCC2019, Inhak Lee

- As technology scales down, the trend of voltage and temperature of SOC makes circuit performance variation larger



- Embedded memory must meet SOC trend
- Voltage and temperature aware design is required
- The variation of the emerging memory cells make circuit complicated

The Structure of RRAM Device



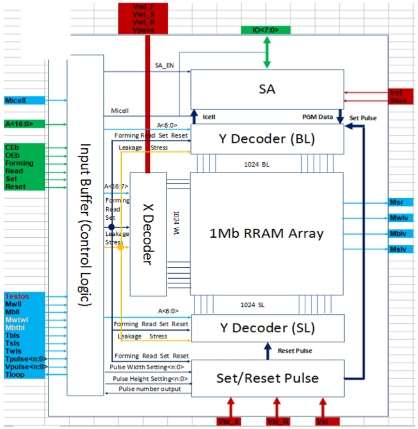
□ Simple structure based on backend

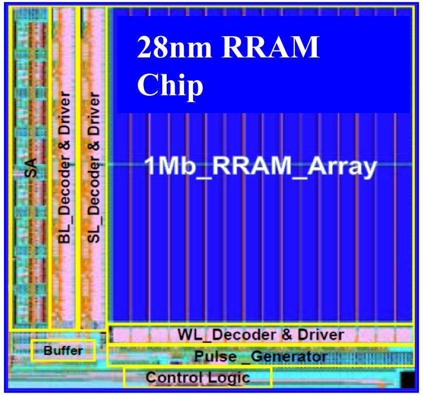
□ Strong scalability

□ CMOS compatibility

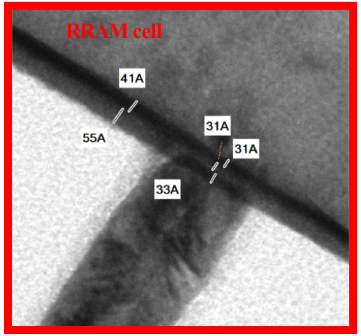
□ 3D feasibility

28 nm ReRAM Chip

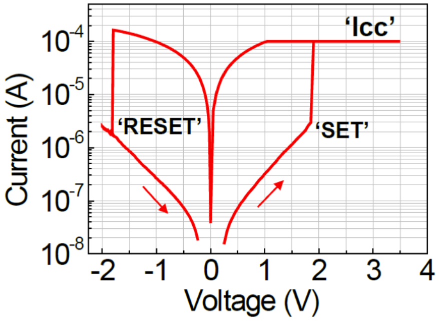




1T1R RRAM Cell



IV of RRAM Cell

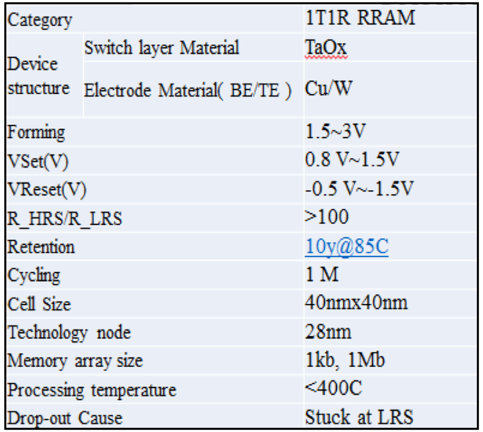


IMECAS & SMIC

Density : 1Mb
Tech node : 28nm

Function :

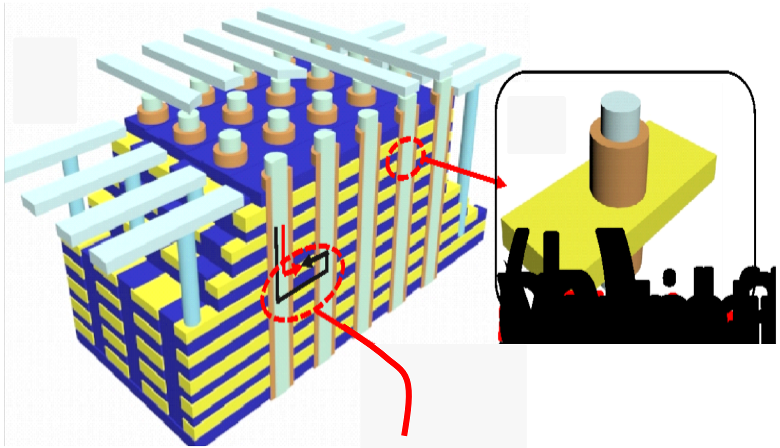
- ? read & write access to single byte with verify
- ? read & write access to single bit with DMA mode

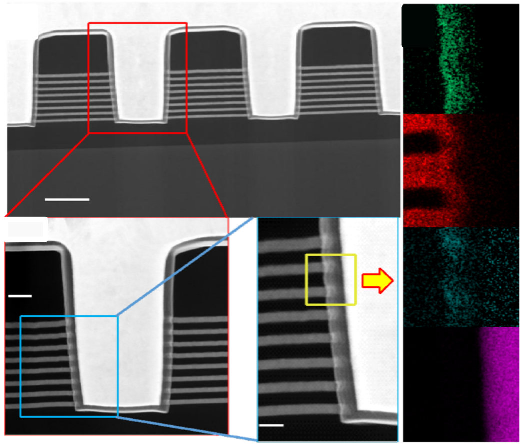


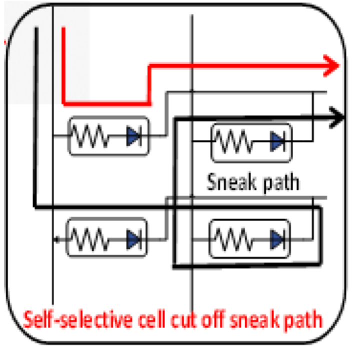
Category		1T1R RRAM
Device structure	Switch layer Material	TaOx
	Electrode Material(BE/TE)	Cu/W
Forming		1.5~3V
VSet(V)		0.8 V~1.5V
VReset(V)		-0.5 V~-1.5V
R_HRS/R_LRS		>100
Retention		10y@85C
Cycling		1 M
Cell Size		40nmx40nm
Technology node		28nm
Memory array size		1kb, 1Mb
Processing temperature		<400C
Drop-out Cause		Stuck at LRS

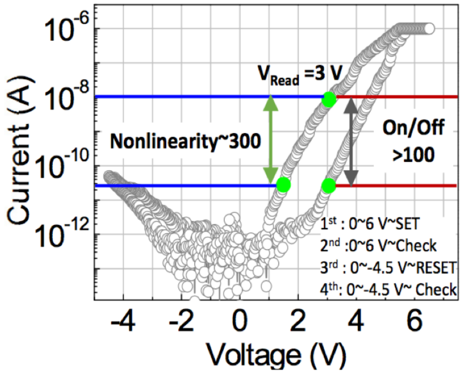
IEDM 2017, 2.4

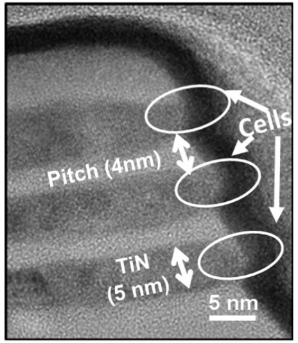
3D RRAM





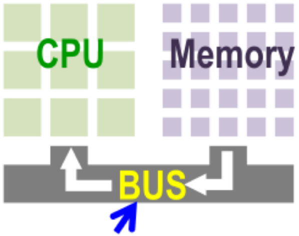




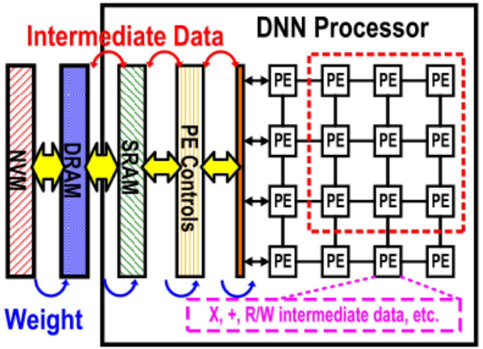


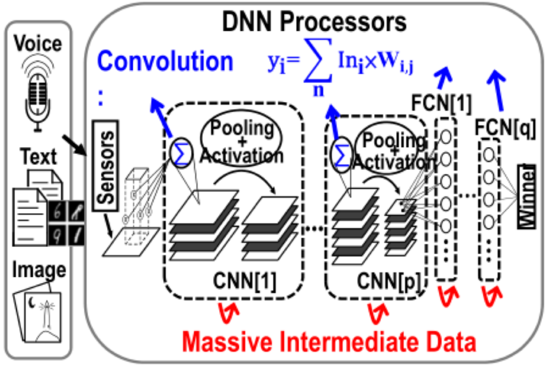
IEDM 2017, IEDM 2015, p245; IEDM 2015, p253, IEDM 2016, p302, VLSI 2016, p84; IEEE EDL, 36, 129 (2015);

Embedded Memory for Computing



Von Neumann
“Bottleneck”



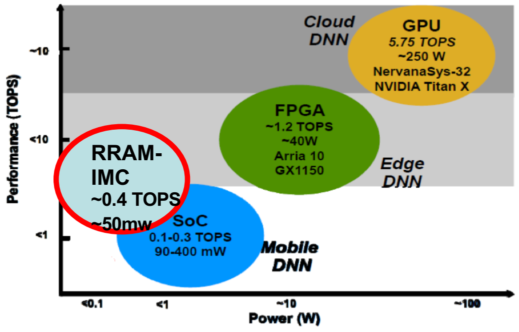


VGG-16 Recognition on LFW*	
Classes	5760
Accuracy	92.5%
Complexity	15.4 **GMACs**
Model Size	15 **MB**
Processing Energy / frame @ 1 TOPS/W	**~ 30 mJ/f** **~ 900 mW** **@ 30 fps**

AAA

1200mAh - 1.5V

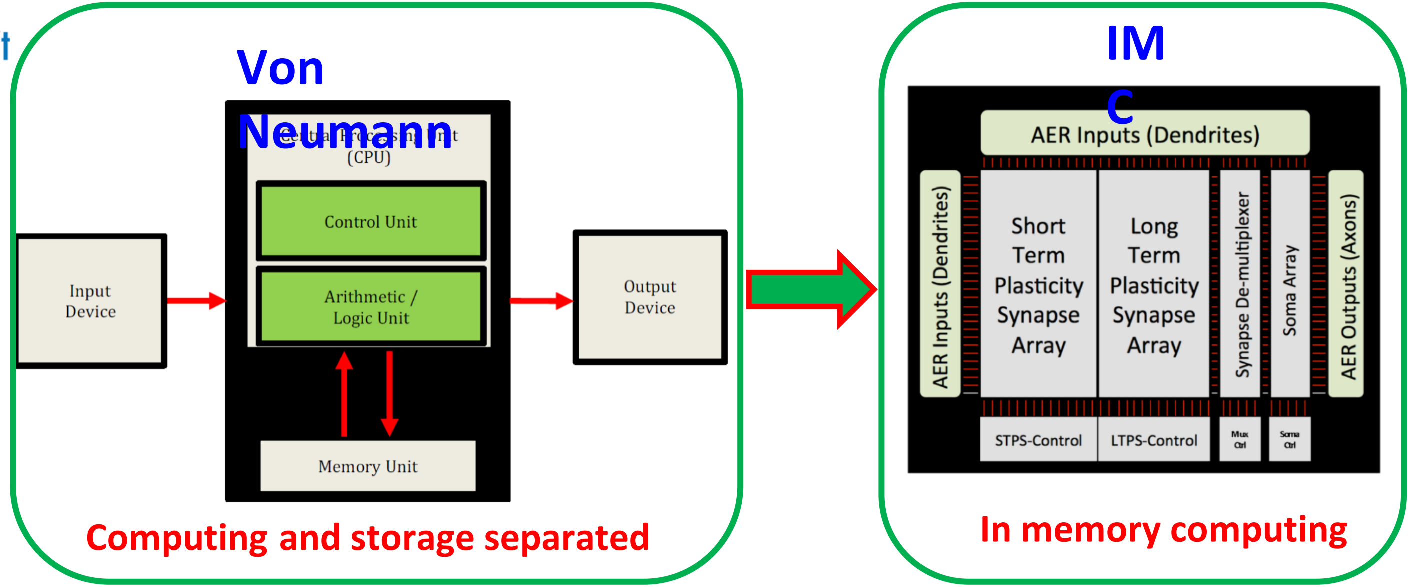
Drains in **2h**



Source:ISSCC2019, ISSCC 2018 , Chang M.F.

- High performance embedded memory is required
- Beyond Von Neumann (new) architecture in embedded application is required
- Always-on application requirement high energy efficiency-low power memories

In Memory Computing



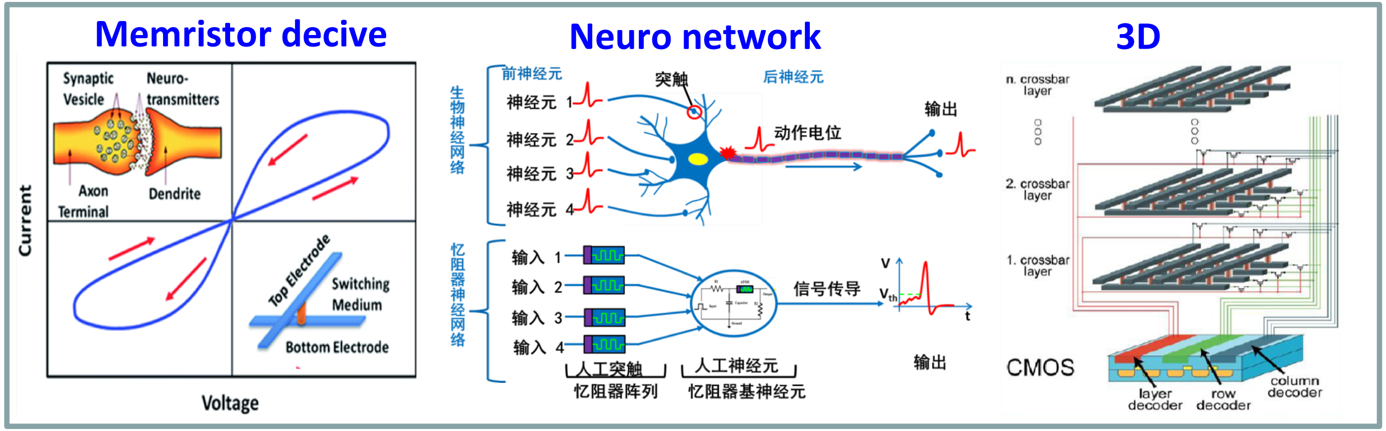
Computing and storage separated

In memory computing

- The traditional information system adopts the architecture of separation of computing and storage. The data is transmitted between the CPU and the memory, and the power consumption is large and the speed is slow.
- The IMC adopts the architecture of storage and computing together, which eliminates the data transmission process and greatly improves the information processing efficiency。

Memristor - ideal neuromorphic biomimetic device

Memrisitor: **M-I-M structure, the resistance can be tuning under the applied voltage, its resistance value is non-volatile.**



Memristor decive

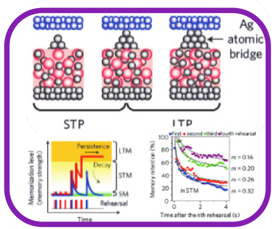
Neuro network

3D

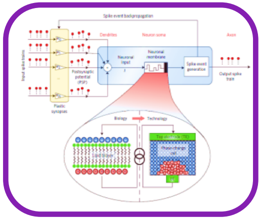
- **Storage and computation fusion: the resistance state is related to the excitation history and is non-volatile;**
- **High parallelism: cross-array structure for easy interconnection;**
- **High energy efficiency: speed ~ ns, energy consumption <pJ;**
- **High-density integration: scaled down to the nm scale and easily integrated in 3D.**

Development and Challenge of Neuro Computing Based on Memristor

The University of Michigan first proposed the use of memristors to simulate synaptic-related functions

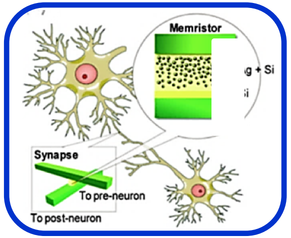


University of California implements a perceptron using a 12 x 12 memristor array

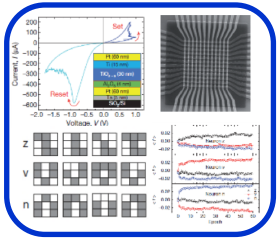


University of Michigan uses 1kb memristor array combined with sparse coding algorithm for image recognition

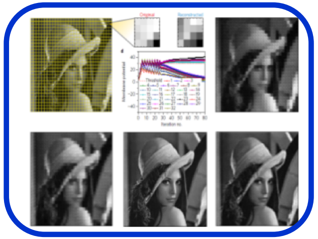
2010 2011 2015 2016 2017



Japan National Materials Research Institute uses the simulation of synaptic forgetting and learning processes

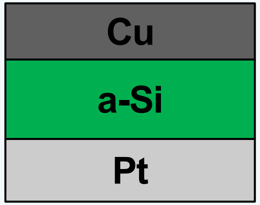
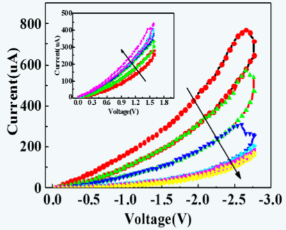
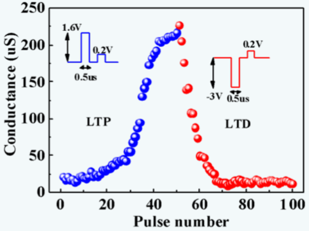
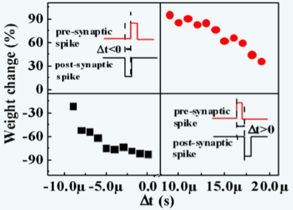
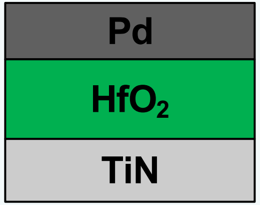
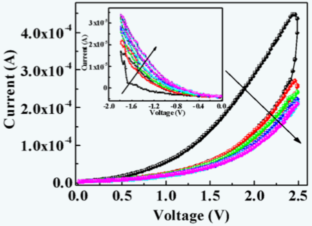
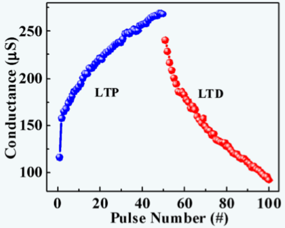
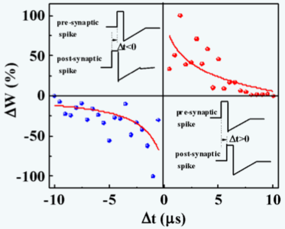
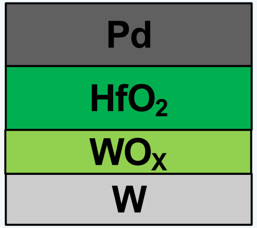
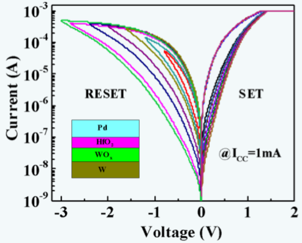
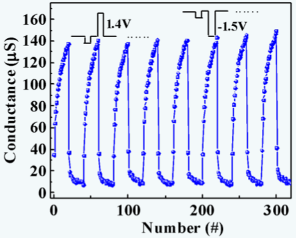
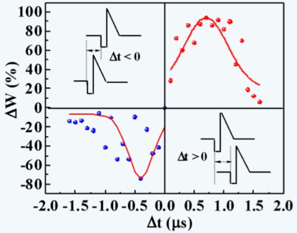


IBM Labs implements neuron IF simulation using a memristor



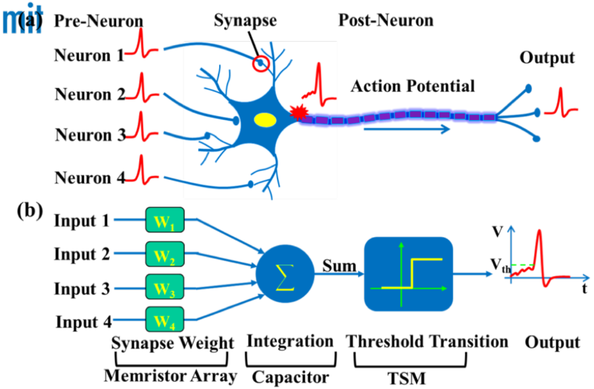
Nano Lett., 2010; Nat. Mater., 2011; Nature, 2015; Nat. Mater., 2016; Nat. Nanotechnol., 2016; Nat. Nanotechnol., 2017

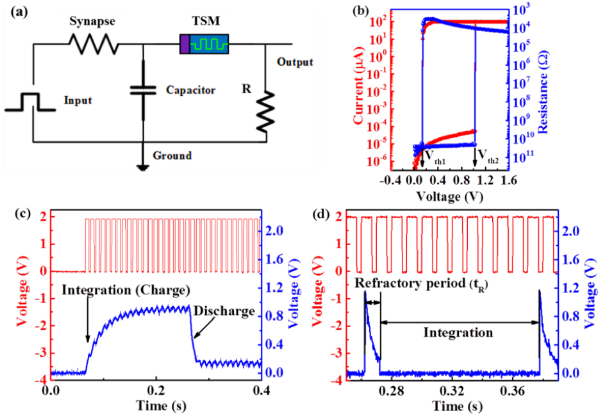
Synaptic realization based on memristor

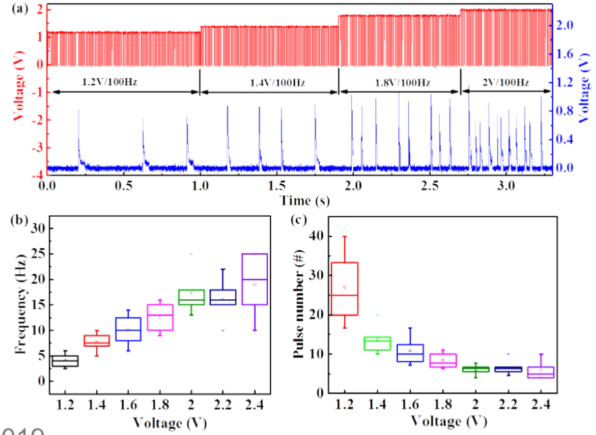
<table>
<tr><th>DEVICE STUCTURE</th><th>I-V</th><th>LTP&LTD</th><th>STDP(overlapping)</th></tr>
<tr>
<td>
Cu
a-Si
Pt</td>
<td>
</td>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>
Pd
HfO_2
TiN</td>
<td>
</td>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>
Pd
HfO_2
WO_X
W</td>
<td>
</td>
<td>
</td>
<td>
</td>
</tr>
</table>

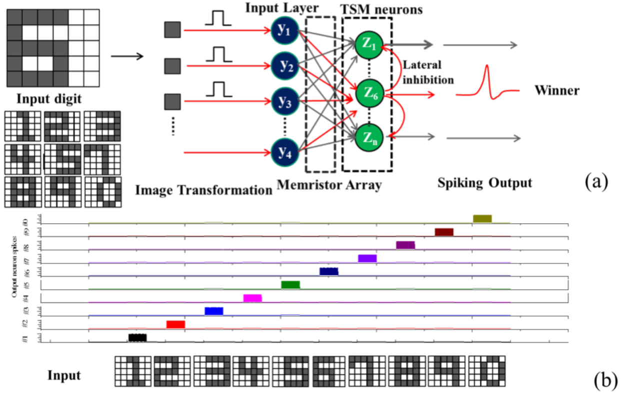
Adv. Funct. Mater., 28, 1705320 (2018); IEEE Electron Device Lett., 38 (9), 1208 (2017); Nanoscale, 9, 14442 (2017); Nano Research, 9, 18908 (2017).

Membrane biomimetic realization based on memristor



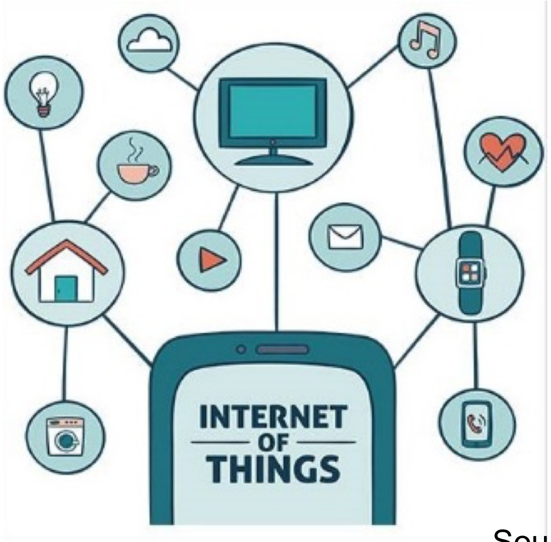


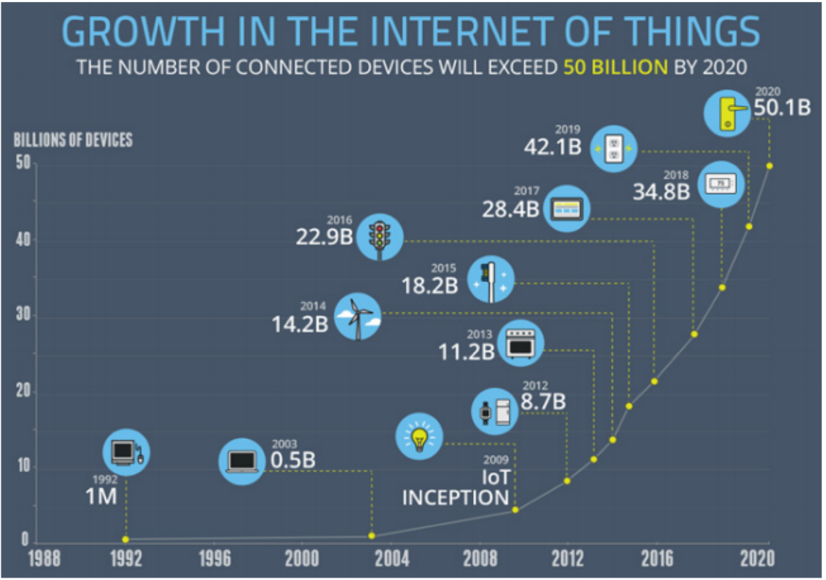




IEEE EDL, 39(2), 308, 311, 2018.

Do Not Forget Security

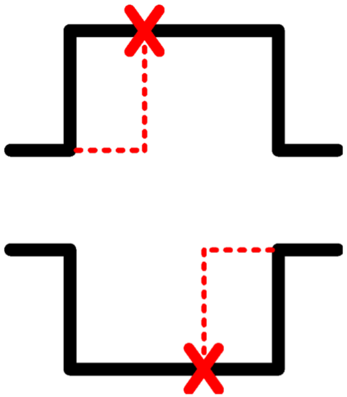




Source: CISCO / National Cable & Telecommunications Association

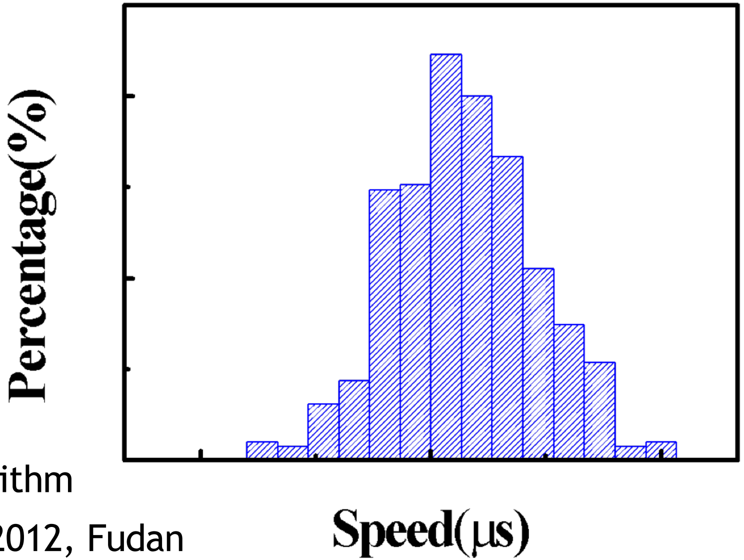
- Billions of devices connected
- Strong demand for hardware secure communication
- Embedded memory play a key role in security
- TRNG & PUF with emerging memory is hot

Write Speed Variation of RRAM



Self adaptive write algorithm

Ref: Xiaoyong Xue et al., VLSI2012, Fudan

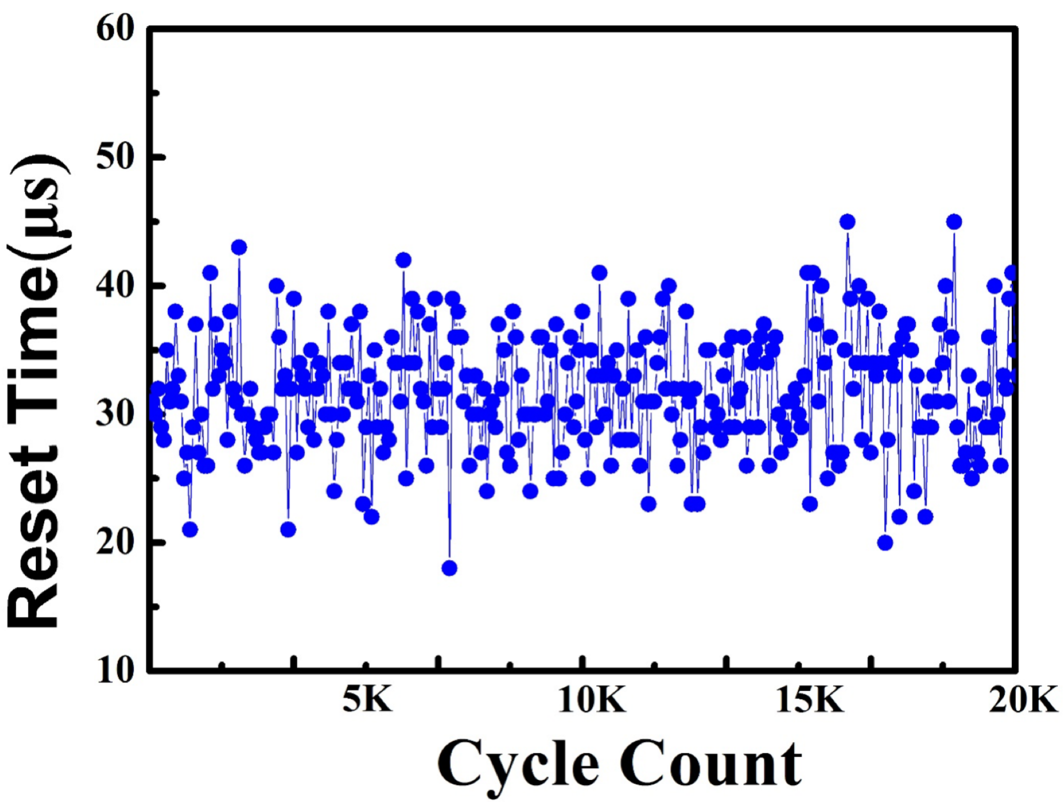


□ Both set and reset have large speed variation

□ Using reset speed variation as entropy source

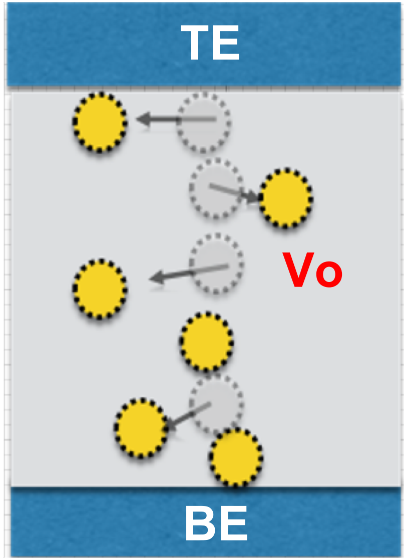
□ Given long time to generate more response bits

Reset Time Variation

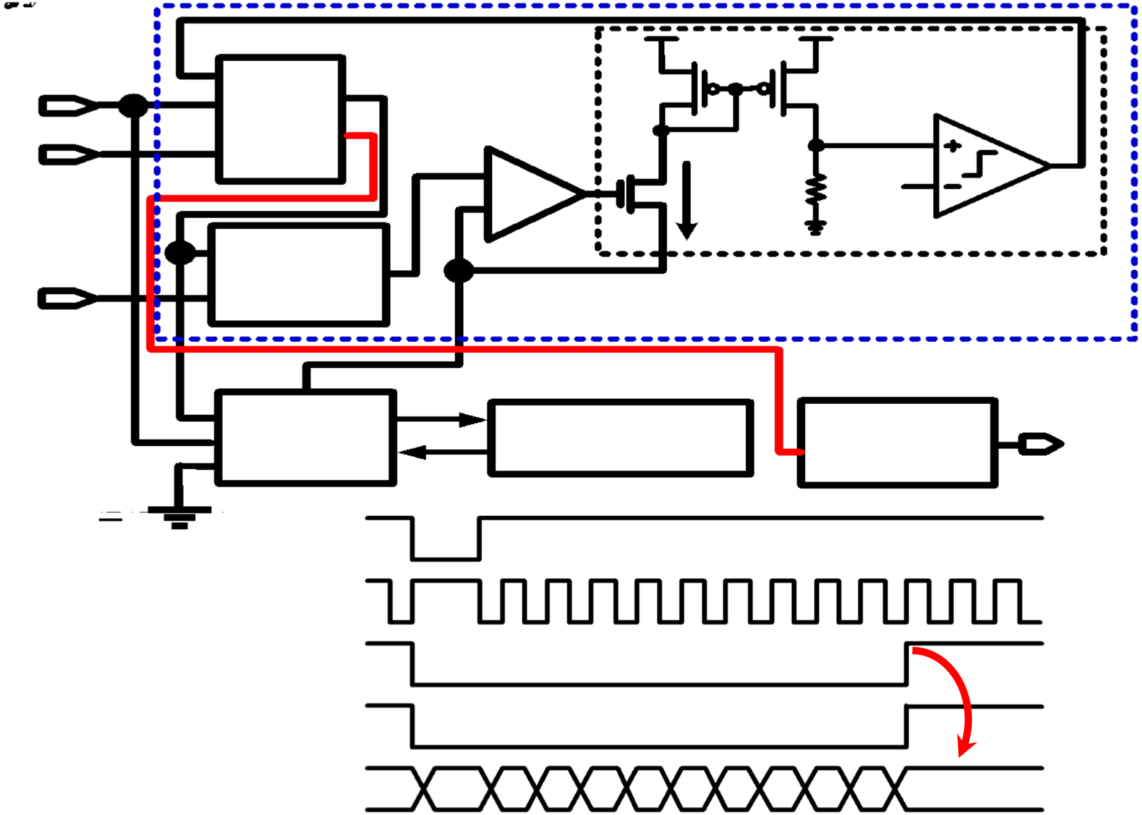


The Mechanism of Speed Variation

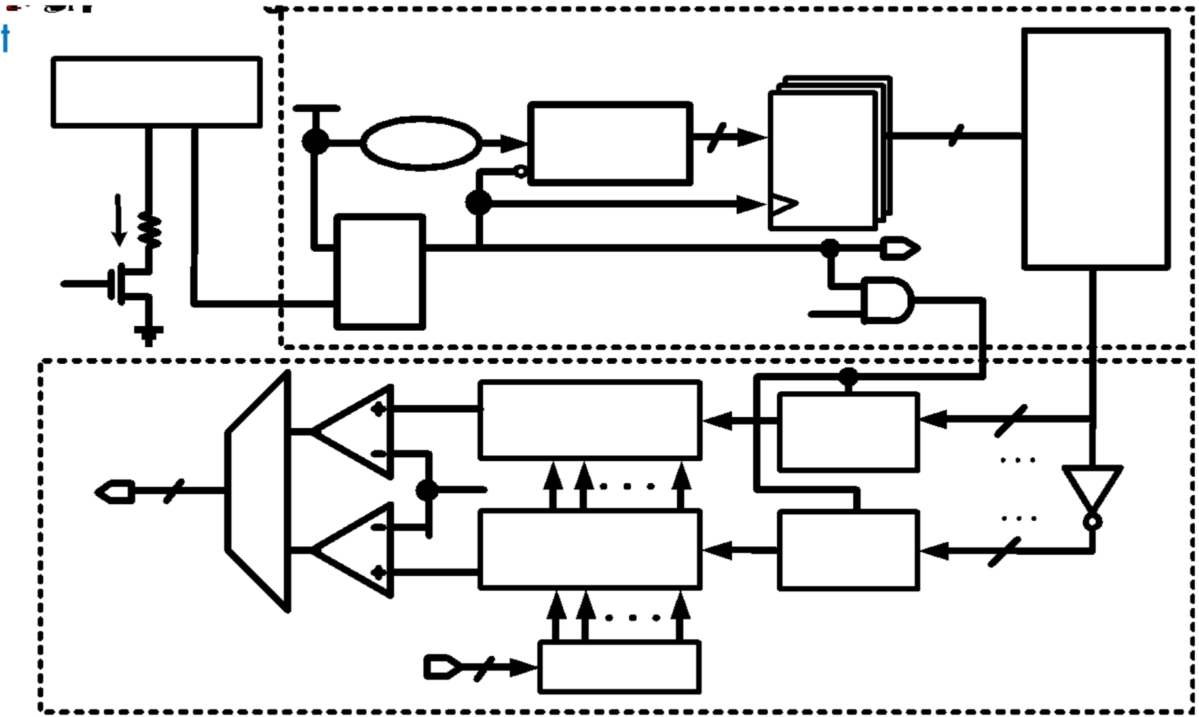
- The fluctuation of Vo trap and de-trap.
- For oxide-based RRAM, the Vo traps line up to form a CF (conductive filament) .
- Set and reset operation cause recombination (de-trap) and generation (trap) of Vo at the interface, which further leads to the connection and rupture of CF, respectively.
- The Vo quantity after trap and de-trap is sensitive to PVTA variation among cycles and locations.



Self-adaptive write driver

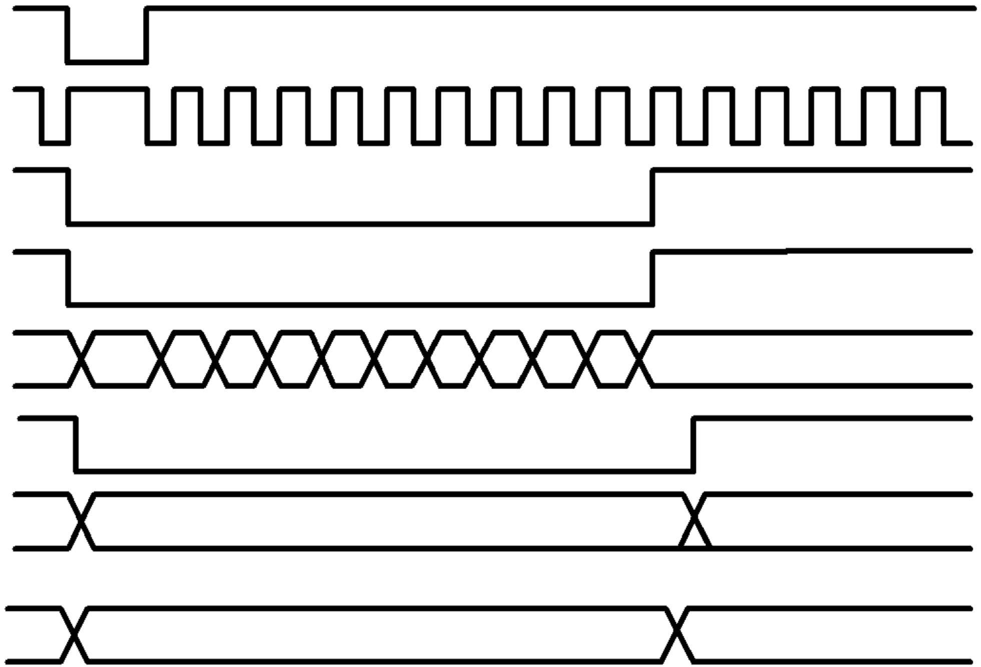


PUF Circuit Implementation



- □ The counter stop at the reset end point.
- □ The RESET speed variation was translated into a digital response

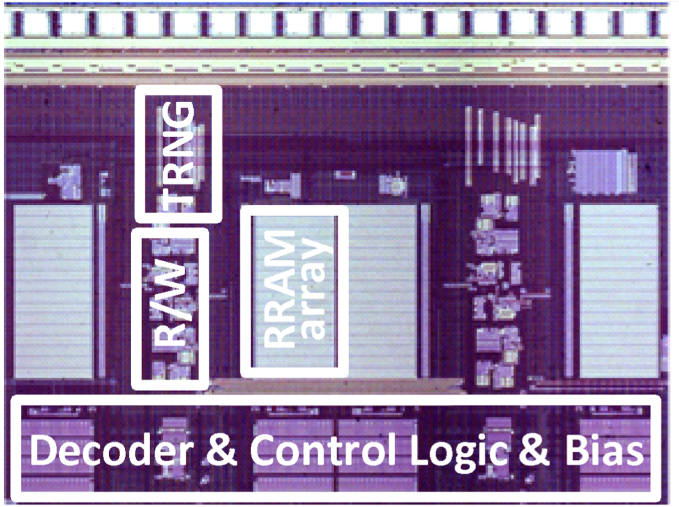
Timing waves of PUF circuit



□ The speed variation was translated to a 16 bits digital outputs

□ The digital bits were written back into arrays

RRAM Embedded Memory with TRNG

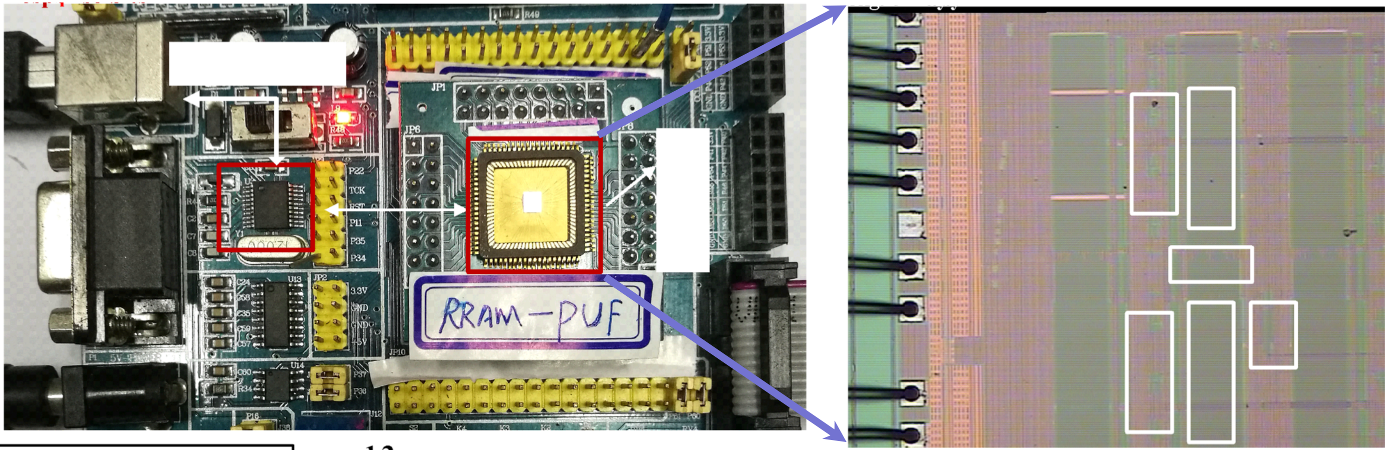


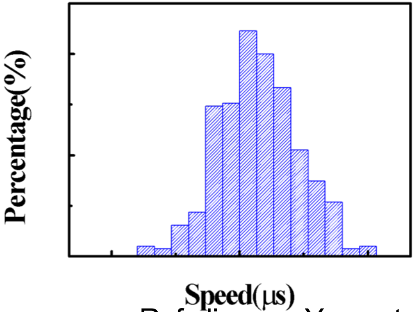
NIST TEST	P-value	Result
Frequency	0.412	PASS
Block Frequency	0.153	PASS
Cumulative Sums	0.551	PASS
Runs	0.743	PASS
Longest Runs of ones	0.583	PASS
FFT	0.514	PASS
Rank	0.397	PASS
Universal Statistical	0.093	PASS
Approximate Entropy	0.166	PASS
Linear complexity	0.045	PASS

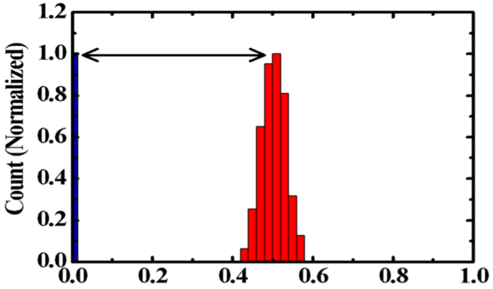
Ref: 1. Jianguo Yang etc., *ISCAS 2017* ;2. Jianguo Yang, *ASICON 2015*

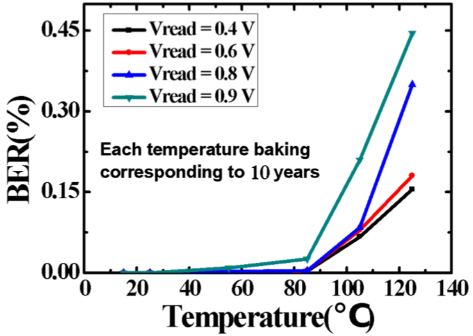
- The physical characteristics of the RRAM itself have security features
- Embedded memory is also used as TRNG source

RRAM Embedded Memory with PUF





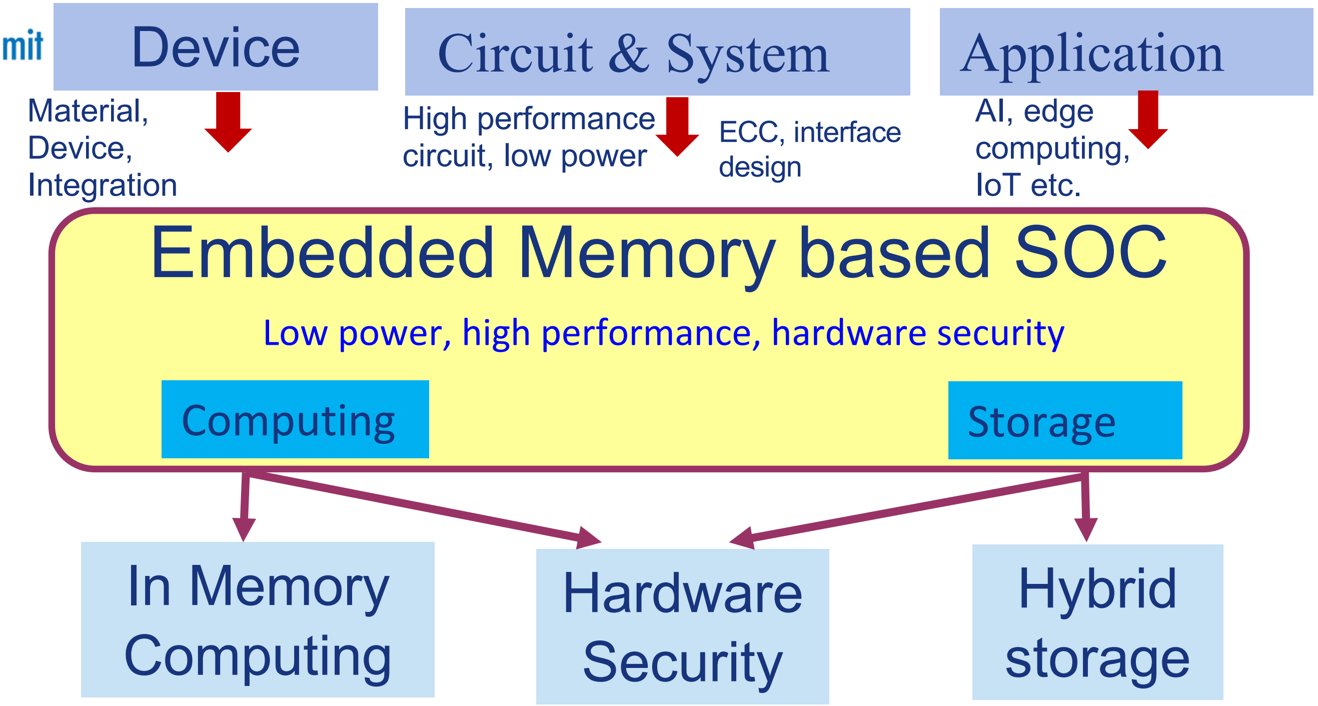




Ref: Jianguo Yang etc., *ESSCIRC & ESSDERC 2018*

- The PUF is integrated with 256Kb embedded RRAM based on 0.13 µm process.
- Embedded memory is also used as hardware security module

The Future of Embedded memory



Device

Material, Device, Integration

Circuit & System

High performance circuit, low power

ECC, interface design

Application

AI, edge computing, IoT etc.

Embedded Memory based SOC

Low power, high performance, hardware security

Computing

Storage

In Memory Computing

Hardware Security

Hybrid storage

1. Device
2. Chip and algorithm
3. Application driven

1. TRNG
2. PUF
3. System level anti-attack

1. Hybrid architecture for embedded applications
2. Device, circuit, system, packaging co-design

Thank you!

The future of RRAM : From Embedded Application to In Memory Computing and Beyond

Jianguo Yang

Key Laboratory of Microelectronics Devices and Integrated Technology, Institute of Microelectronics of the Chinese Academy of Sciences, China