清華大學
Tsinghua University

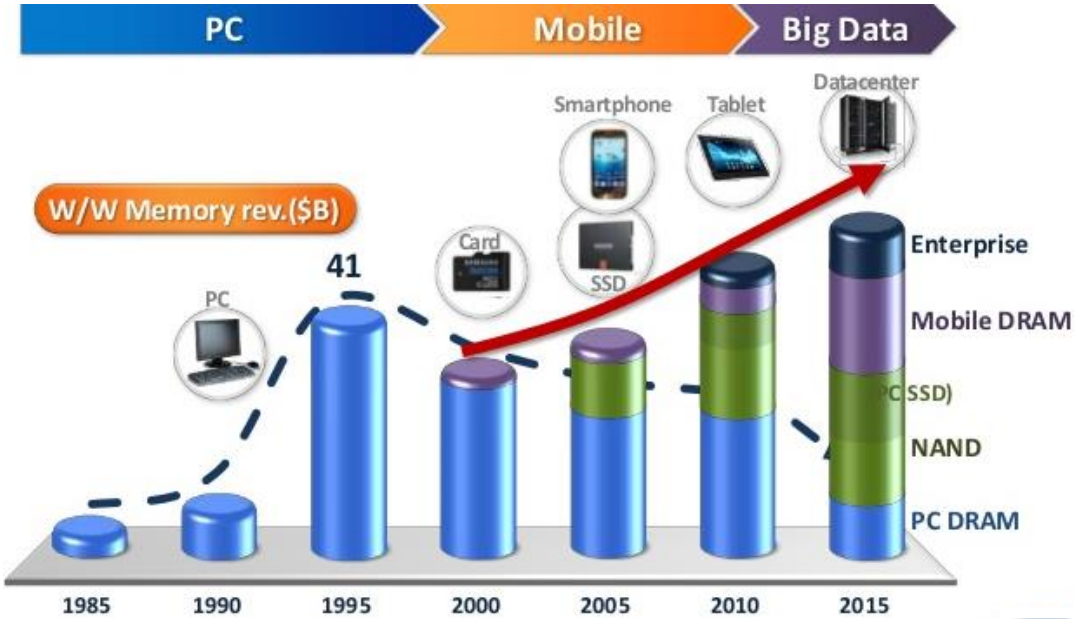
阻变存储器
可靠性与表征

高滨

清华大学微纳电子系

背景

- 存储器被广泛应用于电子产品、互联网、国防、航天等各领域。
- 存储器年产值超过700亿美元，占比超过集成电路总市场的1/4。



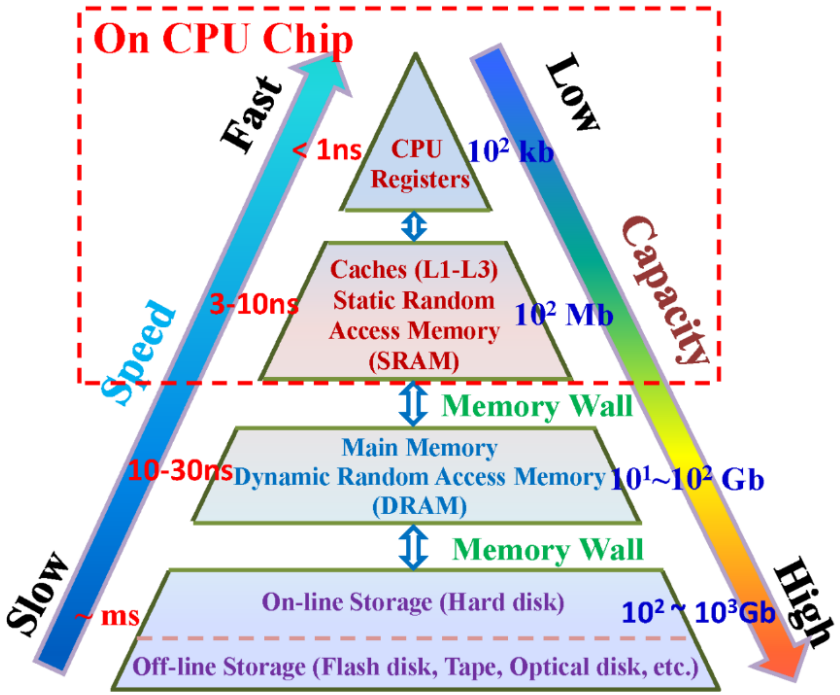
存储器的需求逐年增长

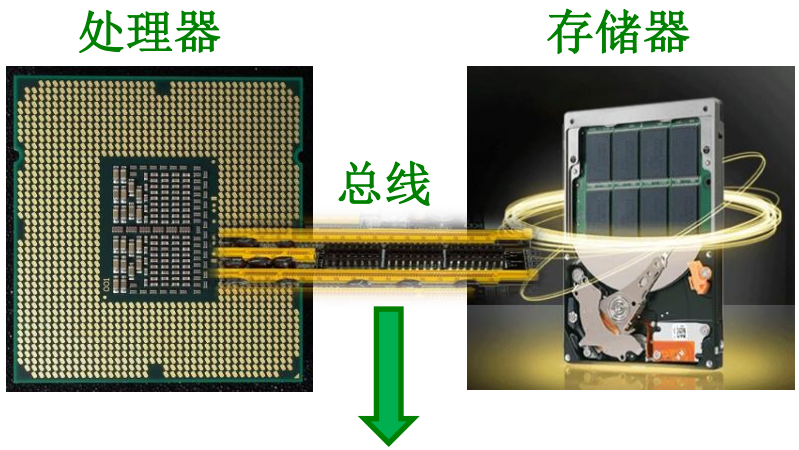
大数据时代信息技术，存储器重要性更加突出

背景

传统电荷型存储器性能已无法满足需求：

- Flash与DRAM速度差距大，导致“存储墙”、“功耗墙”等问题
- 信息存储与计算分离，成为大数据处理实时性的瓶颈

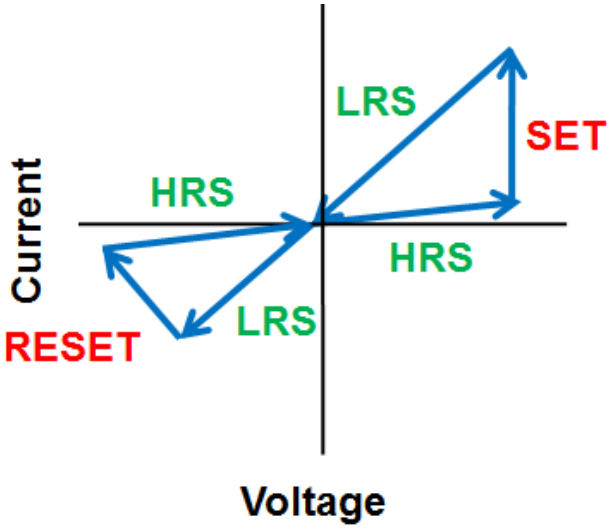


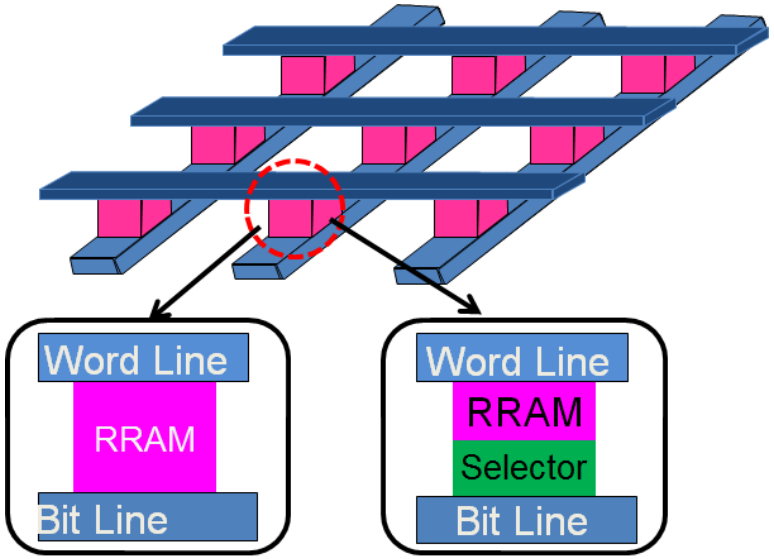


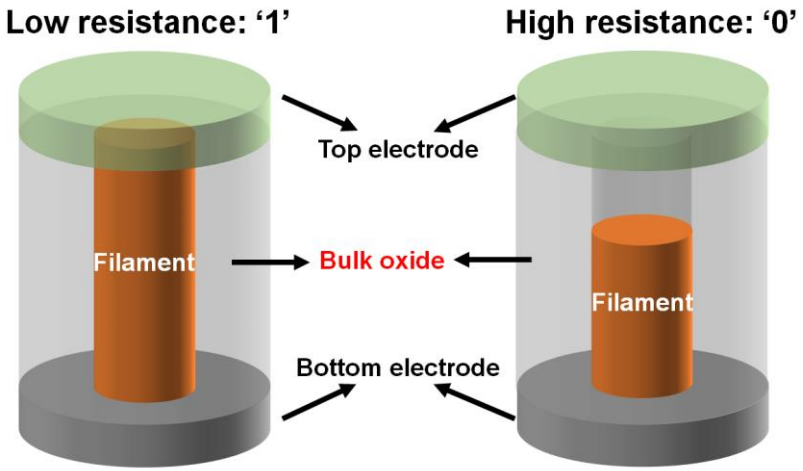
总线上的速度延迟、
额外功耗问题

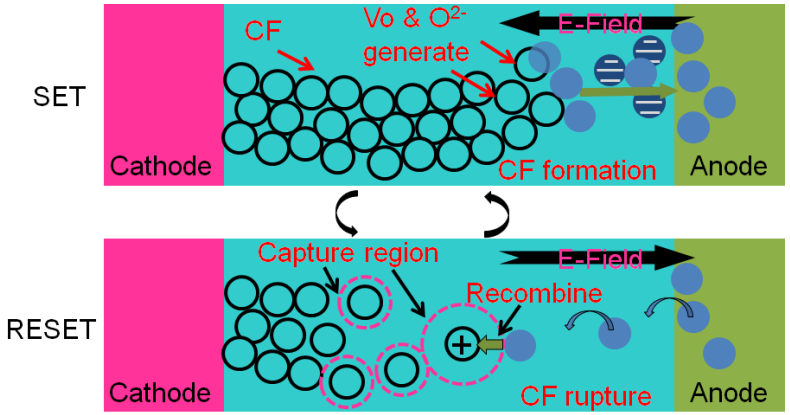
需要探索高速、可融合计算与存储的新型器件

阻变存储器（RRAM）


LRS
SET
HRS
Current
HRS
RESET
LRS
Voltage


Word Line
RRAM
Bit Line
Word Line
RRAM
Selector
Bit Line


Low resistance: '1'
High resistance: '0'
Top electrode
Filament
Bulk oxide
Filament
Bottom electrode


SET
CF
Vo & O2-
generate
E-Field
Cathode
CF formation
Anode
RESET
Capture region
E-Field
Recombine
Cathode
CF rupture
Anode

阻变存储器特点

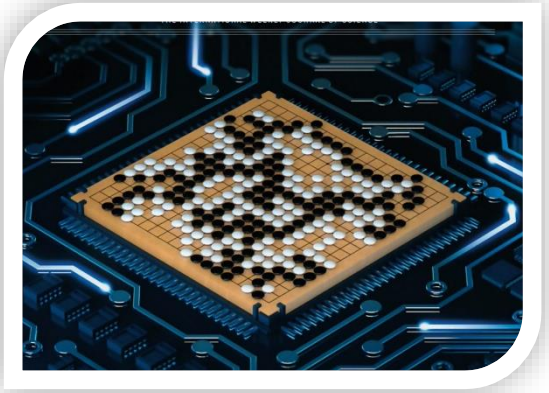
- 单器件特性
  - 速度快：转变时间 <1 ns
  - 可靠性高：循环次数>10^{12}，保持时间>10年/@200°C
  - 非挥发、多值存储
  - 高密度：器件尺寸<10 nm，可三维集成

- 电路特性
  - 同时兼备高速度、高密度和非挥发性
  - 融合计算与存储
  - 常关（normal off），功耗低
  - 非常适合神经形态计算

21st Century: Intelligent Era

Automatically driving a car

Alphago



Smart robot

Voice Assistant

Scale up requires energy efficiency

Power hungry

High efficiency

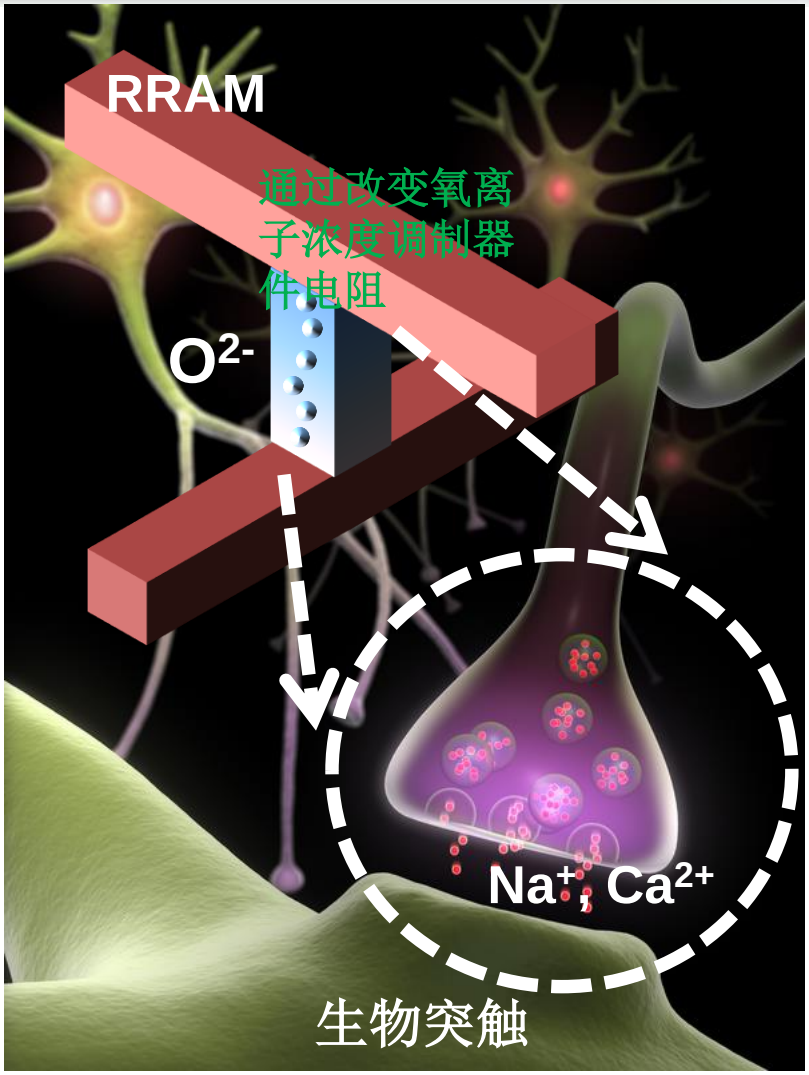
Challenge

- 10^{15} synapses
- Power consumed: 10-20 W
  - Less than a light bulb

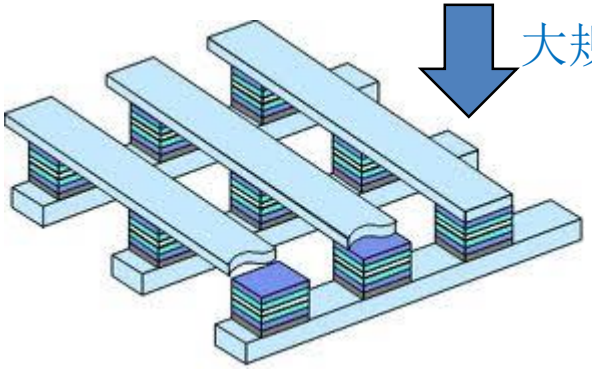
	Application	Hardware used	Estimated power consumption
Large scale	**Emulating 4.5% of human brain**: **10^{13} synapses**, 10^{9} neurons	Blue Gene/P: 36,864 nodes, 147,456 cores	**2.9 MW** (LINPACK)
	Deep sparse autoencoder: **10^{9} synapses**, 10M images	1,000 CPUs (16,000 cores)	**~100 kW** (cores only)

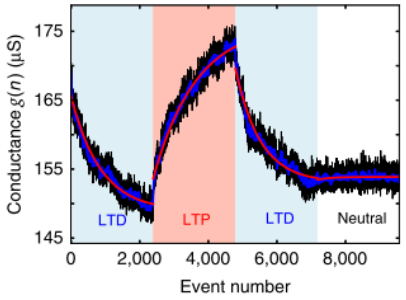
Stanford, S. B. Eryilmaz et al., IEDM 2015

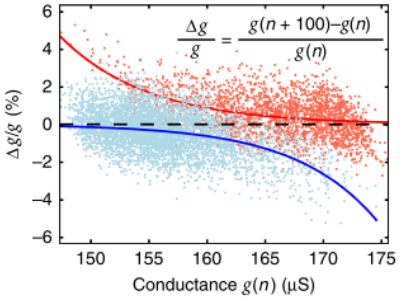
神经形态器件——阻变存储器



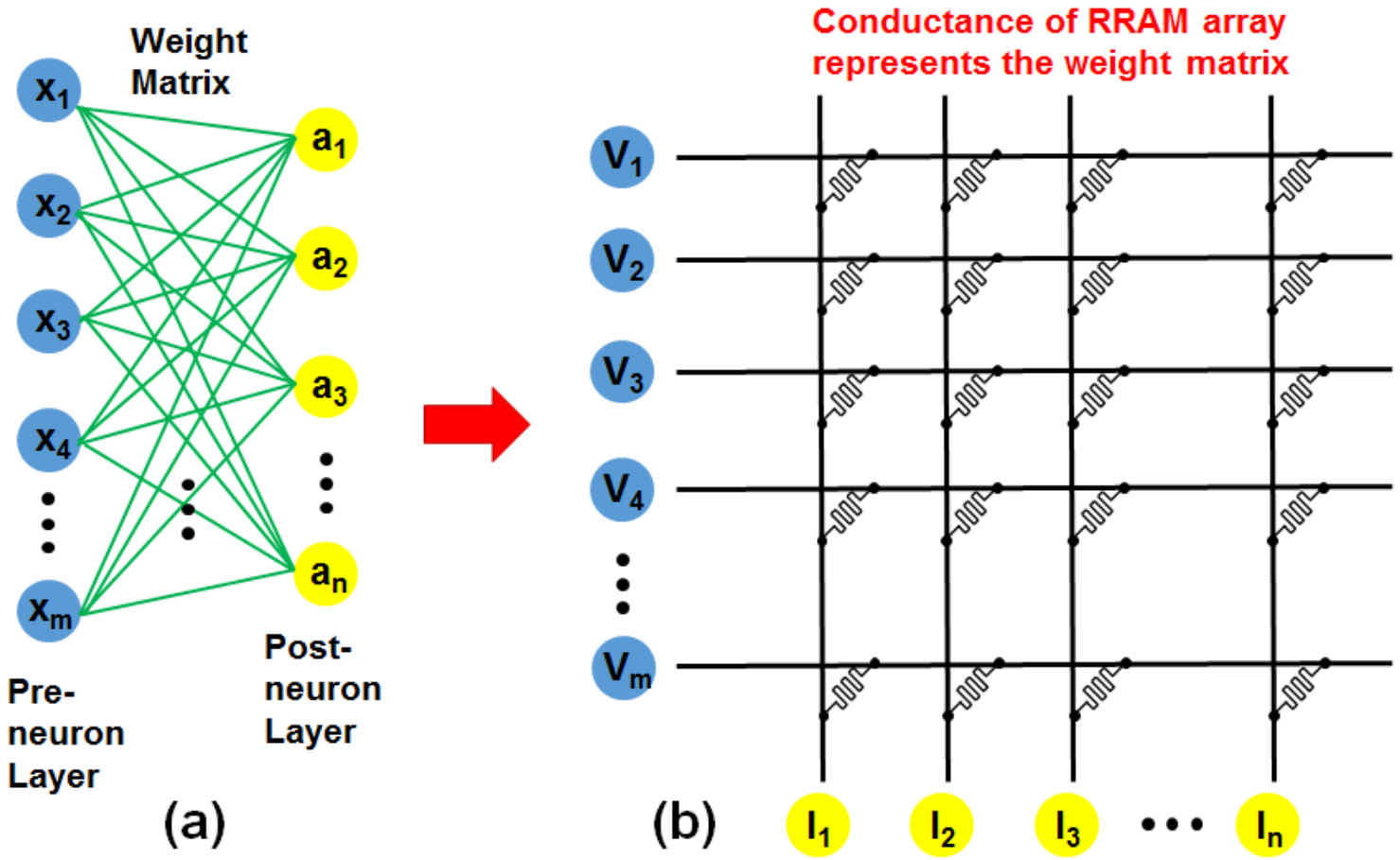
大规模并行





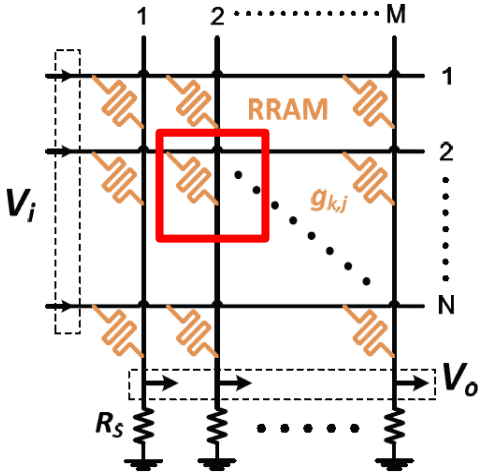


忆阻器阵列与神经网络

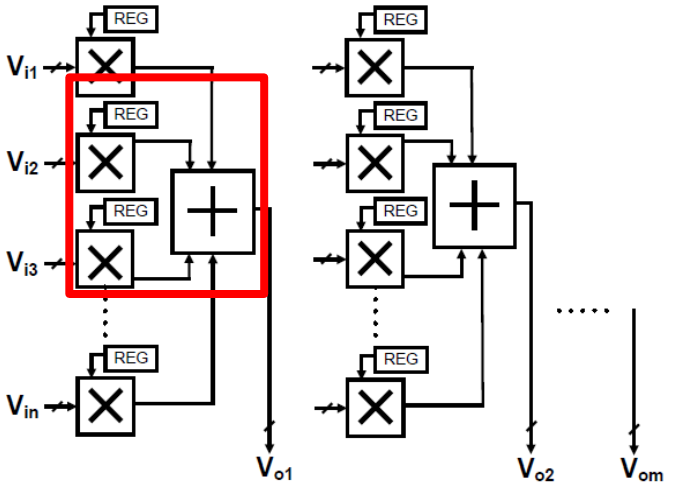

Weight Matrix
x1
x2
x3
x4
xm
a1
a2
a3
an
Pre-neuron Layer
Post-neuron Layer
(a)
Conductance of RRAM array represents the weight matrix
V1
V2
V3
V4
Vm
I1
I2
I3
In
(b)

规模和速度的提升

8-bit N x M
的向量乘法



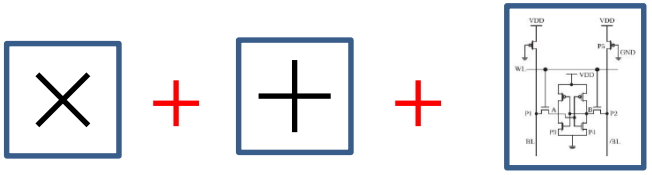
忆阻器阵列
[ASP-DAC, 2015]



CMOS电路
[Field-Programmable Technology, 2002]

1个忆阻器
（64 levels）

≈



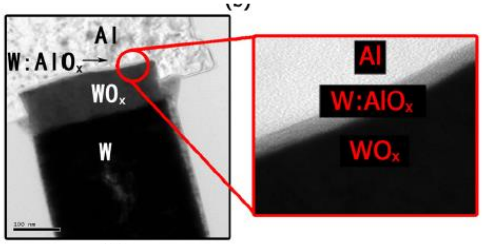
8-bit乘法器 + 8-bit加法强 + 8-bit SRAM存储器

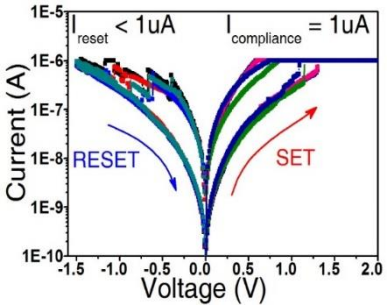
RRAM面临的主要问题

- 机理不清
- 涨落大
- 可靠性不足
- 工艺集成问题
- 模拟阻变特性优化

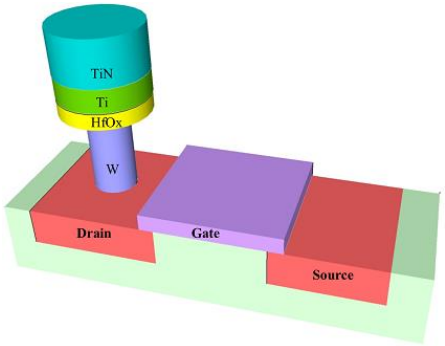
RRAM cell optimization

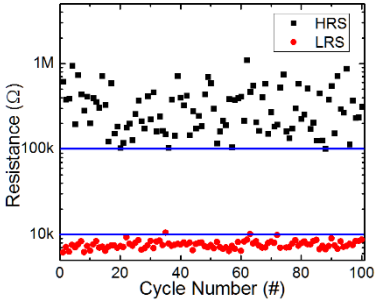
Material selection -- Developed several categories of RRAM devices



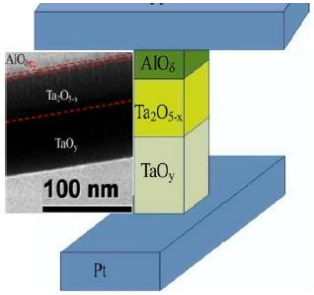


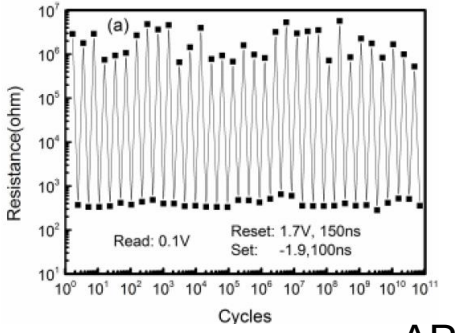
WO_x based RRAM shows low RESET current and multi-level storage behavior





HfO_x based RRAM fabricated with CMOS compatible process. Bit yield >99.9%.



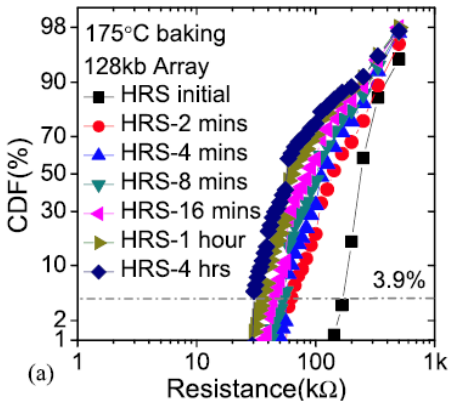


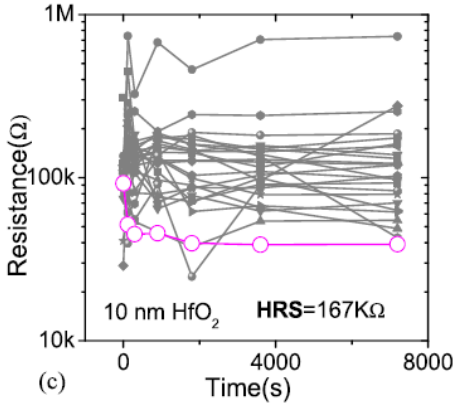
TaO_x based RRAM shows excellent endurance performance.

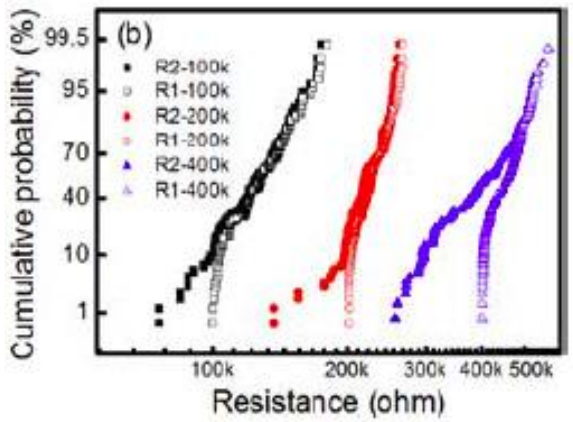
APL 2013; APL 2014; Nanotechnology 2016

RRAM cell optimization

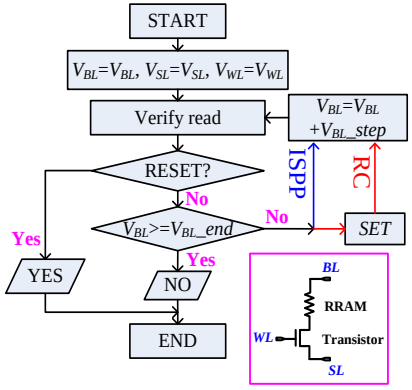
Retention improvement – suppressing relaxation effect



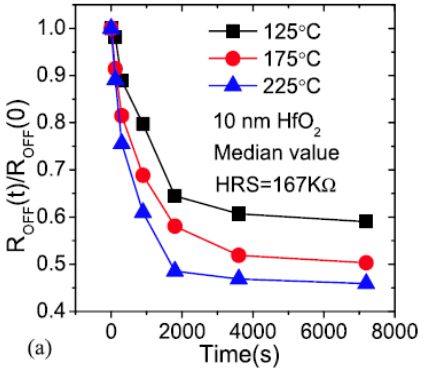


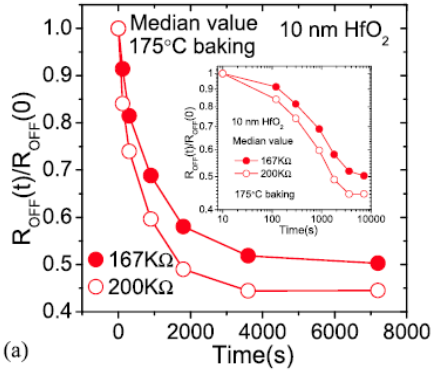


Quick and stochastic retention loss after program is observed in some of the bits



Optimized operation scheme to suppress the relaxation effect is provided

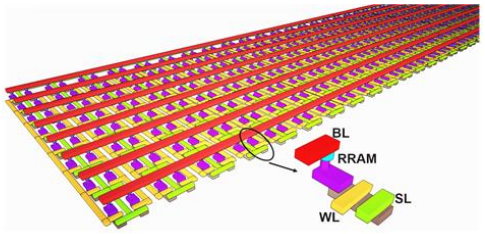


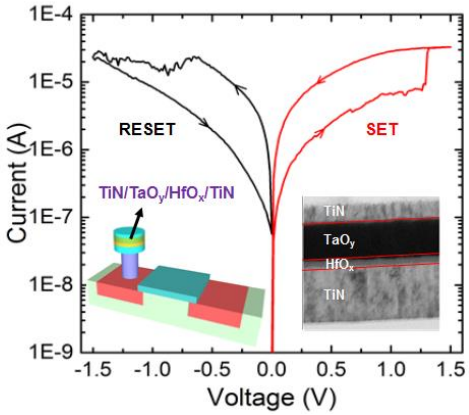


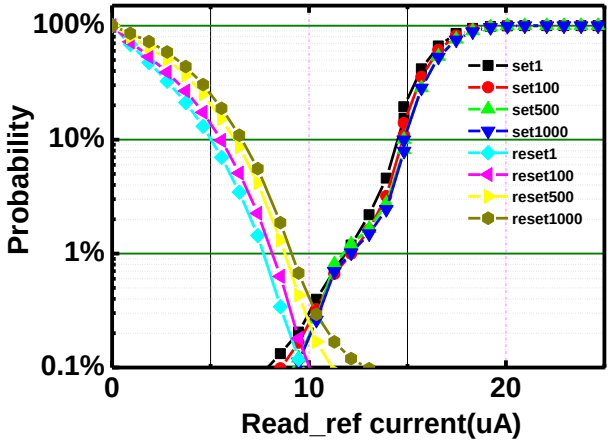
Various characterization methods are utilized to explore the mechanism of relaxation effect

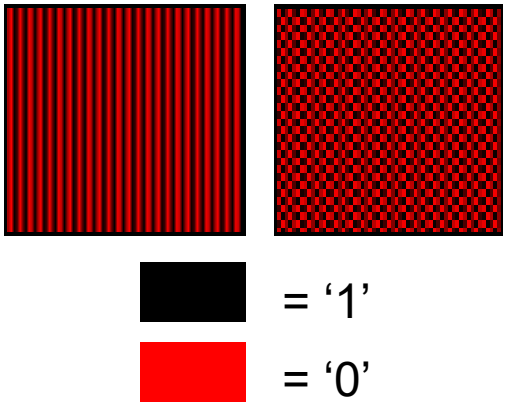
EDL 2016; EDL 2017; IMW 2016; APE 2016

Testing results on 16M RRAM chip





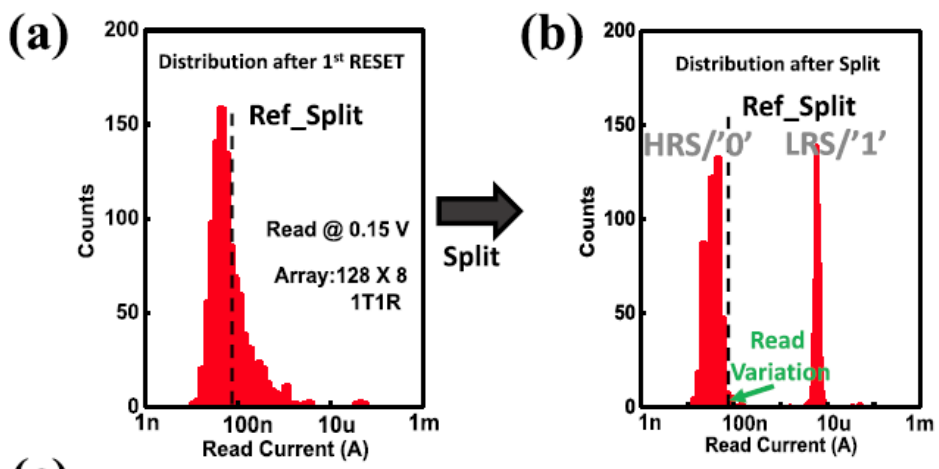


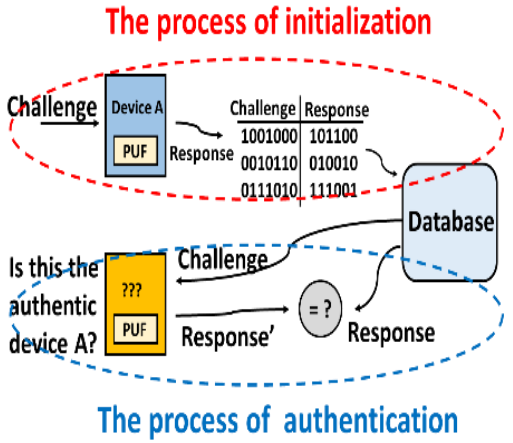


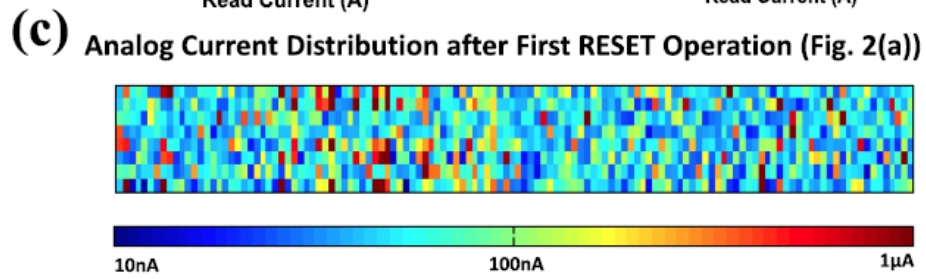
- 1T1R array with HfOx/TaOx stack as RRAM cells
- >99.9% bit yield; >99% bit yield after 10^4 cycles
- Develop a foundary+lab platform for future production and application

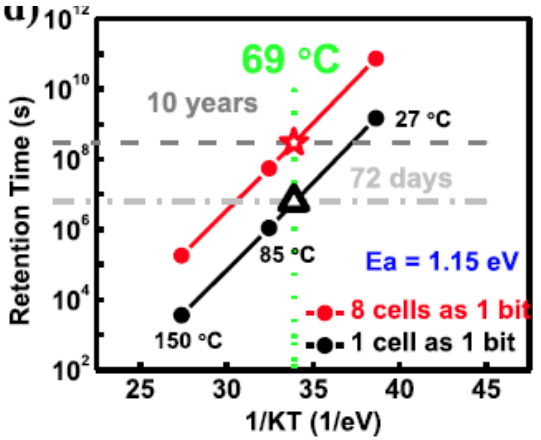
NVMTS 2015; IEDM 2016; IMW 2017; Submit to IEDM 2017

Security Application









Utilize the variability of HRS resistance, the random key can be generated

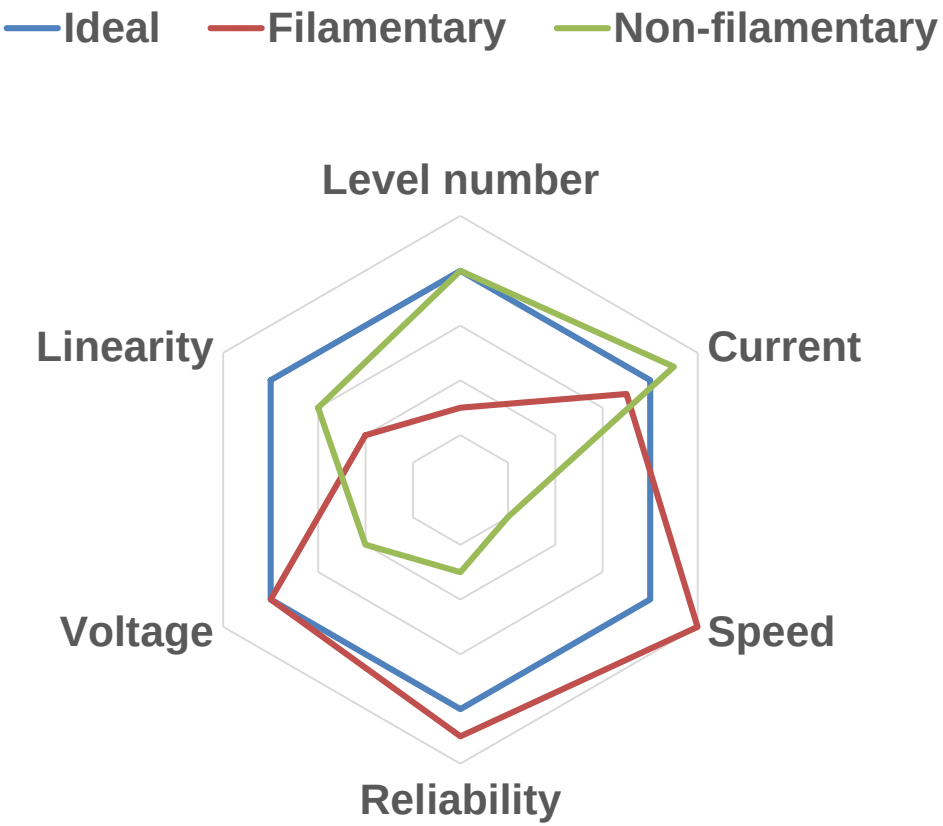
Good reliability is demonstrated on our RRAM based PUF

EDL 2017; VLSI-TSA 2017

Requirement of synaptic device

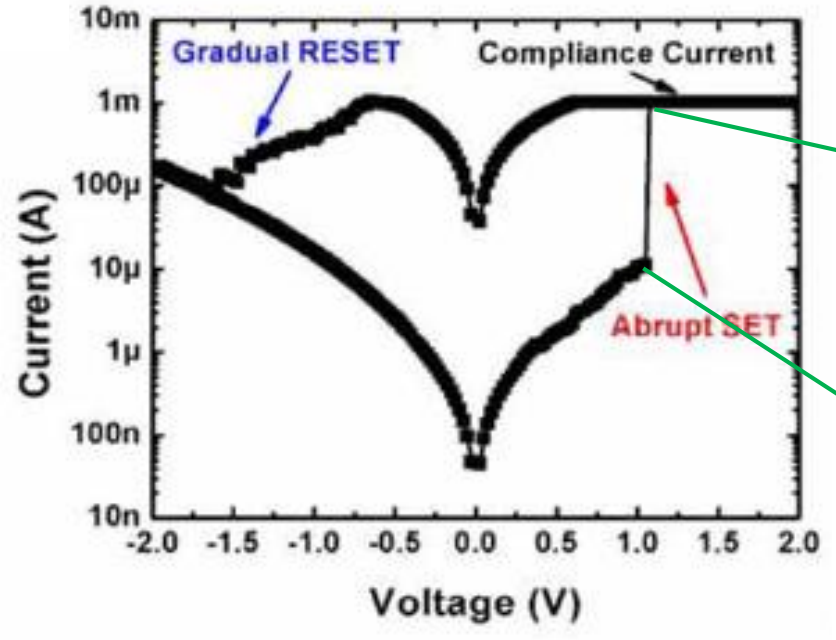
Ideal synaptic device：

- Enough level number
- Low current
- Fast speed
- Low voltage
- Good reliability
- Conductance linearity
- ……



✓ **There are some defects for both filamentary device and non-filamentary device. Device engineering is needed.**

Challenge of filamentary RRAM



I-V curve of HfOx based filamentary RRAM

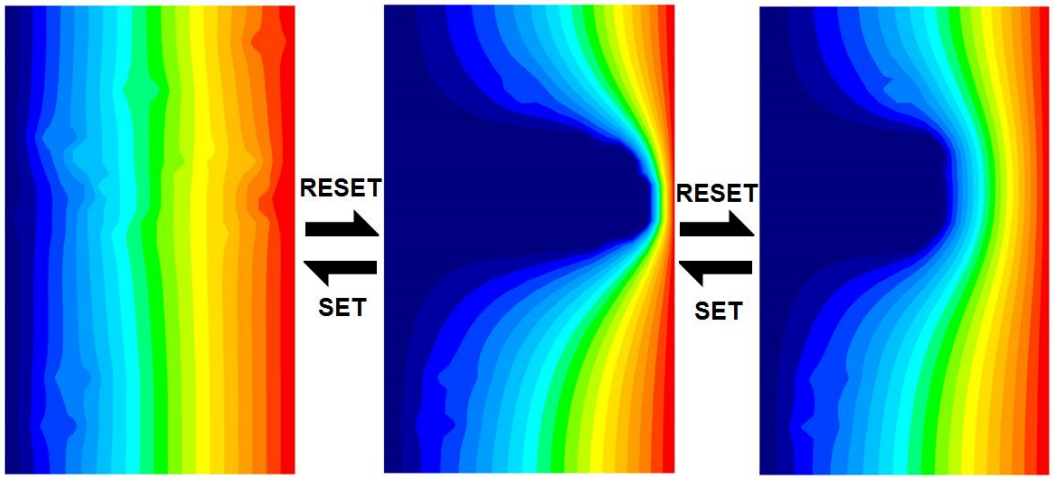
Filament connect

Filament rupture

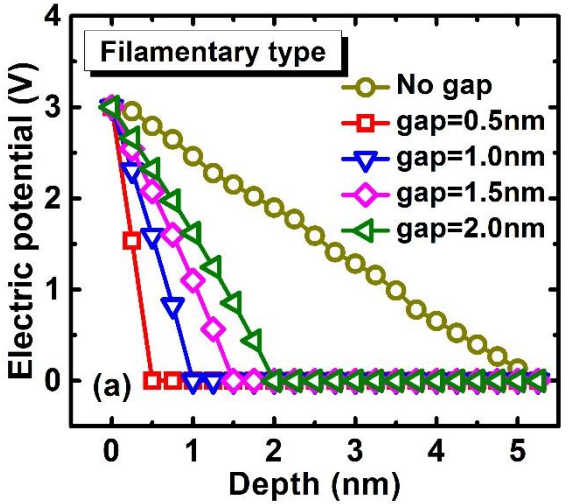
Simulated oxygen vacancy distribution

✓ **Abrupt SET transition is always observed in filamentary RRAM**

Origin of abrupt SET



Simulated 2D map of electric distribution of different switching phases



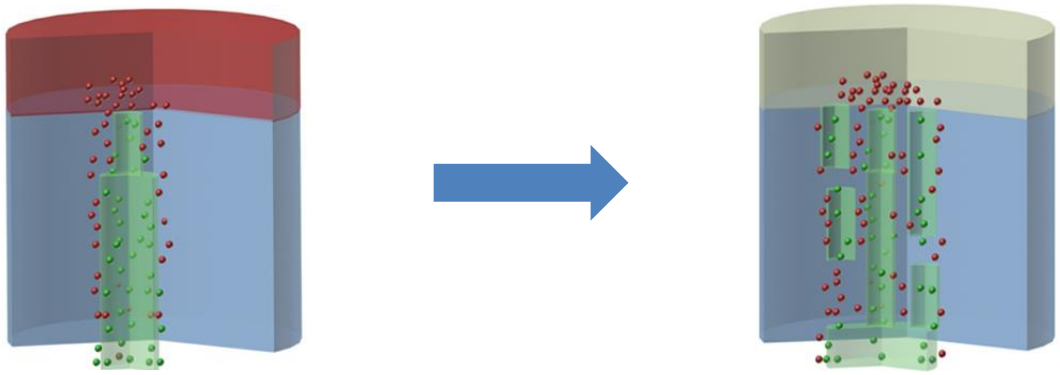
Simulated electric potential distribution along filament region

✓ **Electric field increases continuously as filament growing during SET process, resulting positive feedback which accelerate SET process**

B. Gao, et al., IEDM 2016, 176

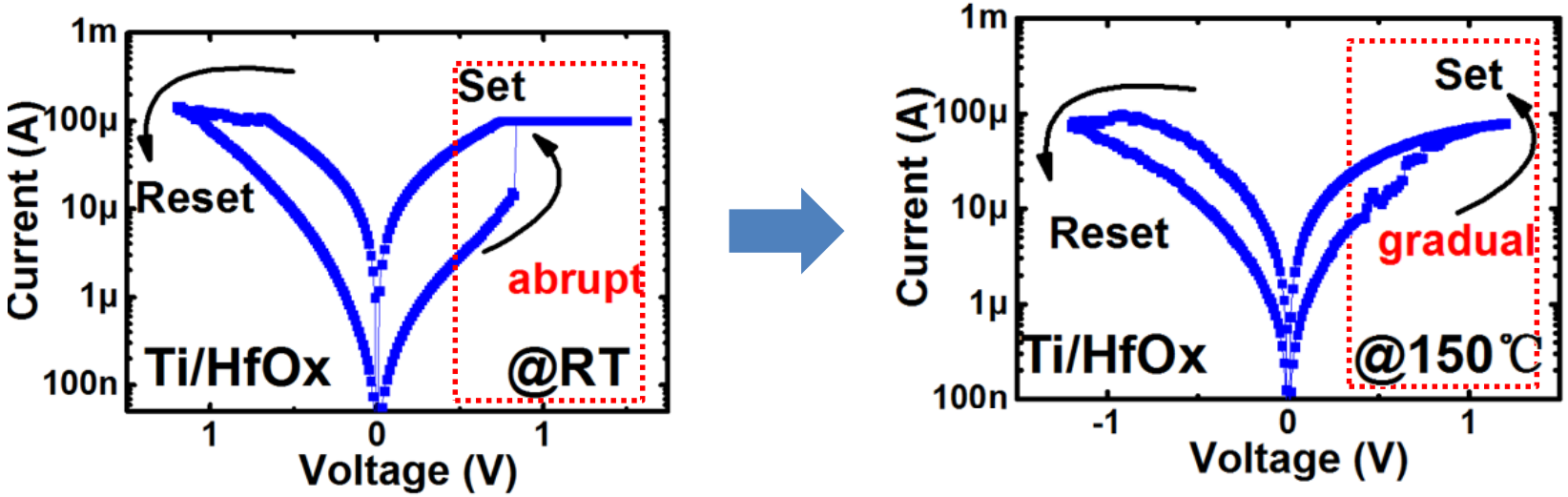
Improve analog switching

- Reduce the positive feedback of electric field during SET process
- Avoid forming single strong conductive filament
  - Disperse oxygen vacancies to a broad region
  - Forming multiple weak conductive filament



Local temperature is a key factor

$$P = \exp[(\gamma qE - \varepsilon_f)/kT]$$

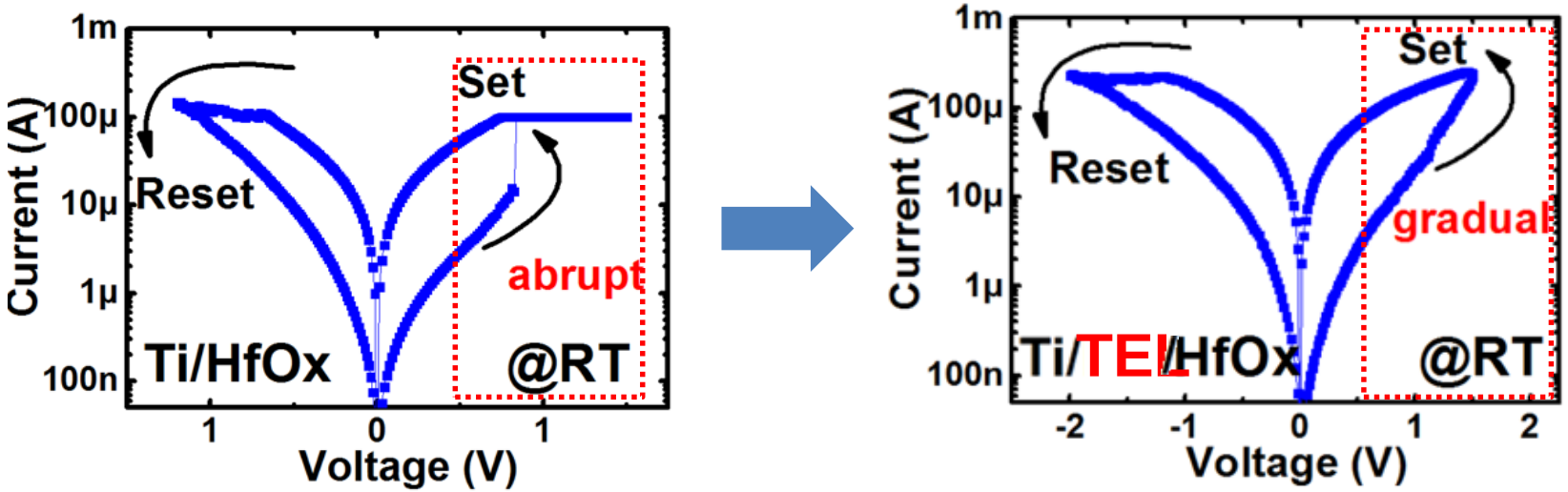


- ✓ **Ti/HfOx RRAM device exhibits abrupt switching at room temperature, but shows analog switching at high temperature**
- ✓ **High temperature reduce the influence of electric field during CF formation**

EDL 2017

Insert thermal enhanced layer (TEL)

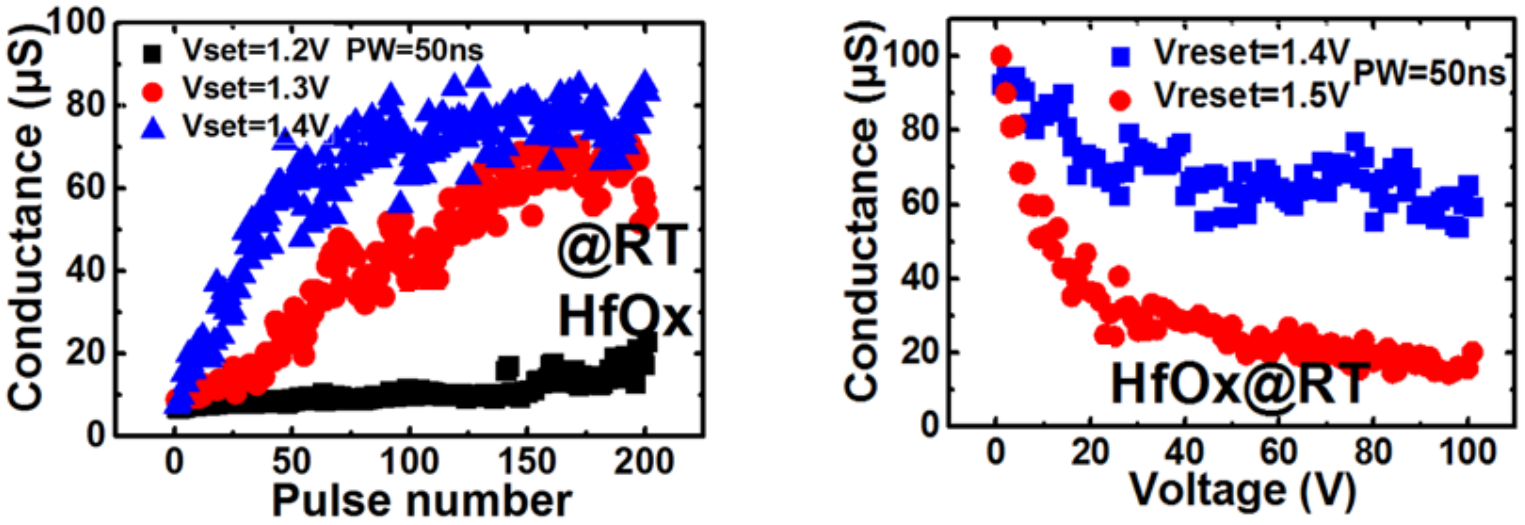
$$P = \exp[(\gamma qE - \varepsilon_f)/kT]$$



✓ **HfOx RRAM shows excellent analog switching in both SET and RESET by inserting the TEL**

EDL 2017

Insert thermal enhanced layer (TEL)



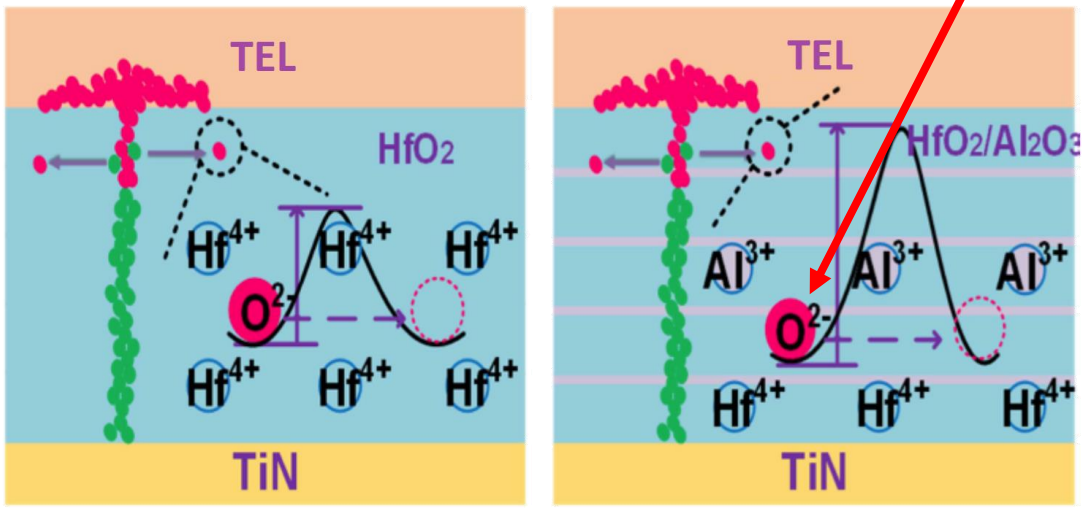
Conductance modulation under identical pulse train after inserting TEL

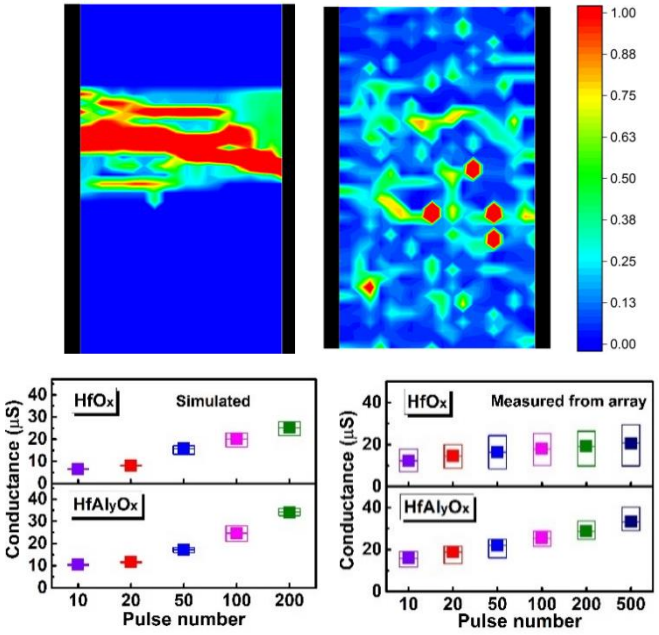
- ✓ **The conductance increases/decreases gradually by applying identical SET/RESET pulses for Ti/TEL/HfOx device.**

EDL 2017

Doping Method

$$P = \exp[(\gamma qE - \varepsilon_f)/kT]$$

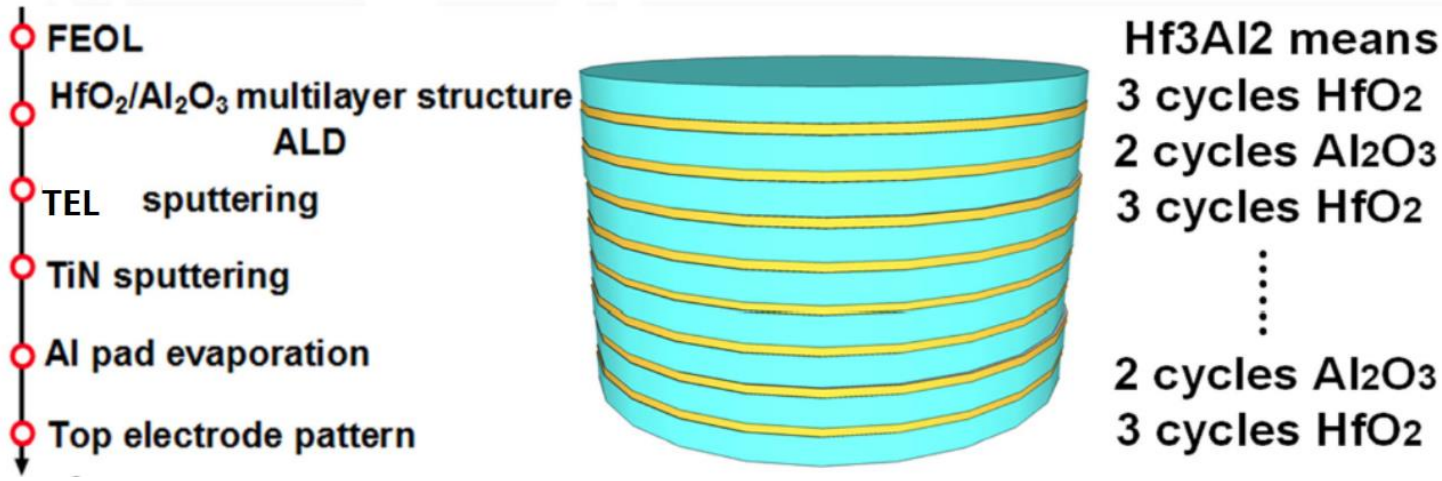




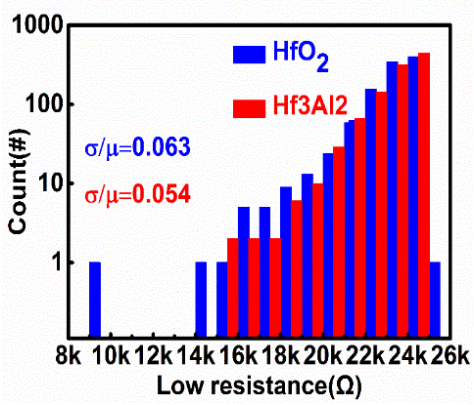
- ✓ **We fabricated HfAlyOx RRAM with ALD multilayer deposition**
  --3 cycles HfOx layer and 1 (or 2) cycles AlOx layer

- ✓ **Doping Al could localize the formation oxygen vacancies (Vo) and avoid crystallization**

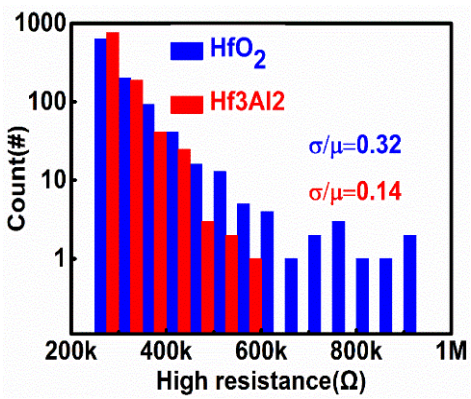
Doping Method



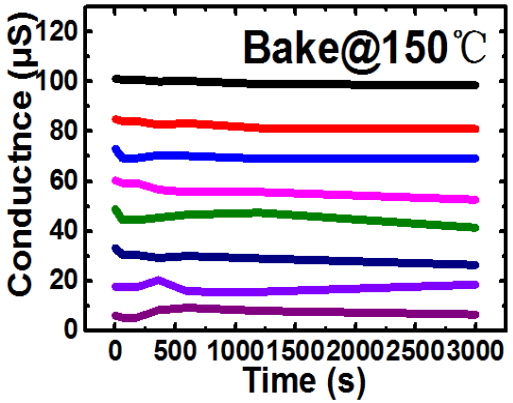
Doping Process



LRS distribution



HRS distribution

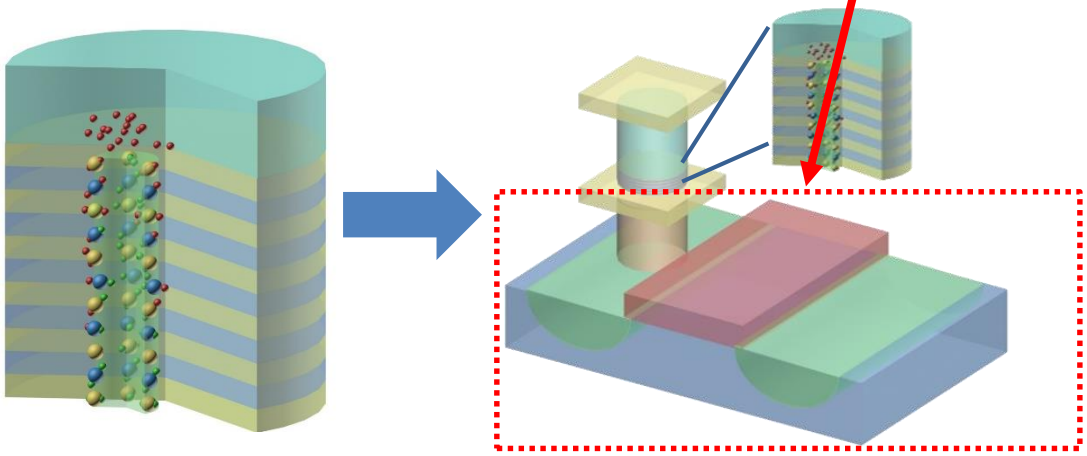


Retention

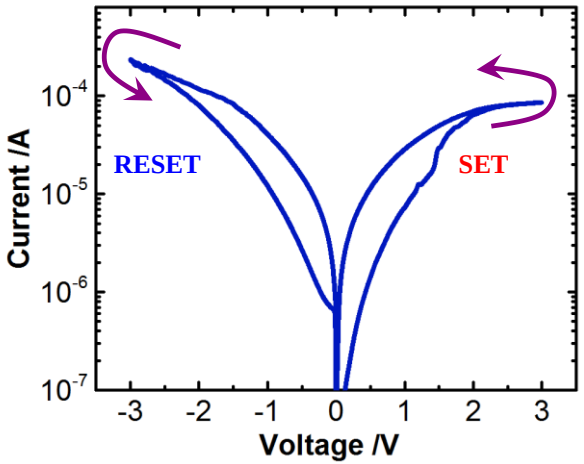
✓ **Uniformity and retention are improved by Al doping**

Connect to a transistor

$$P = \exp[(\gamma qE - \varepsilon_f)/kT]$$



1T1R cell structure is used



Typical I-V curve under DC sweep

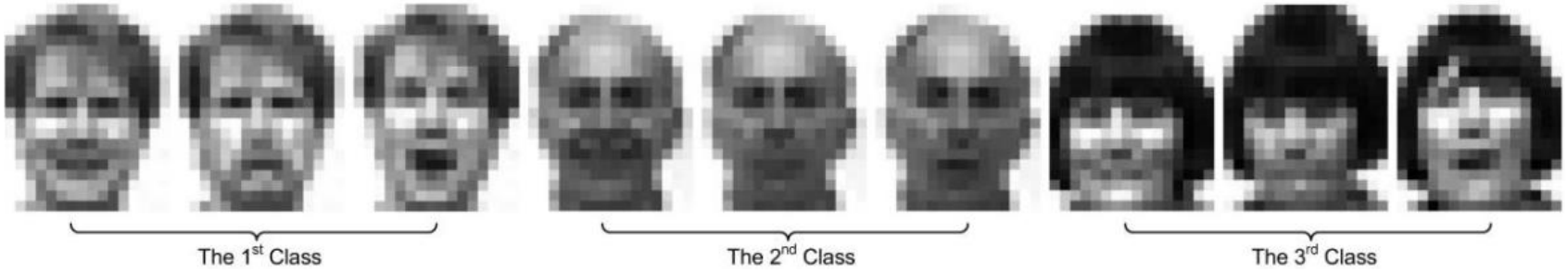
- ✓ **Transistor is used to limit electric field and avoid overshoot**
- ✓ **Reliable bi-directional analog switching behavior of 1T1R cell is observed**

Summary

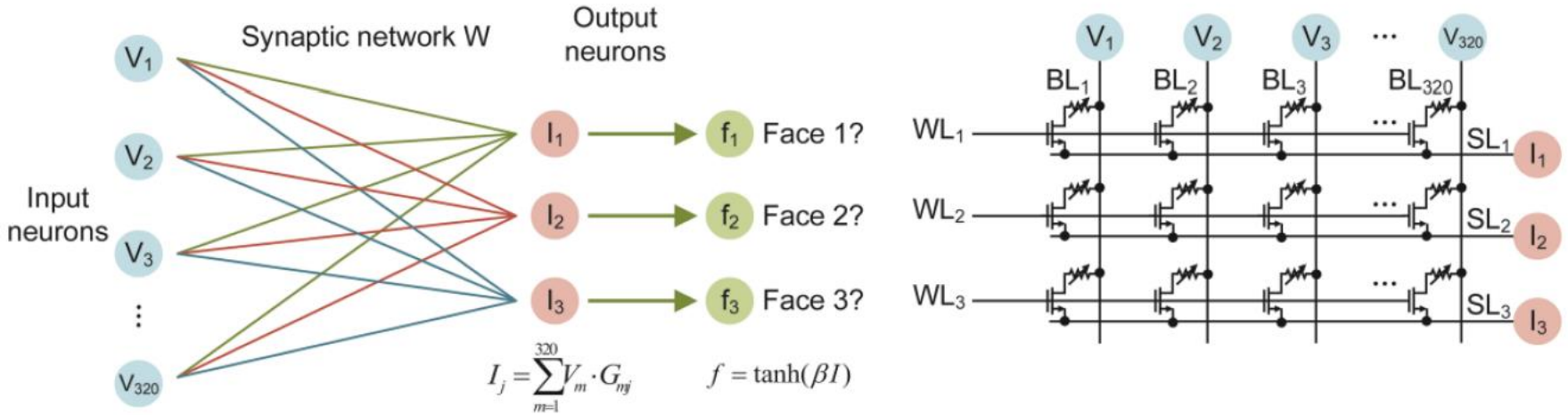
Structure	Analog behavior	Operation condition	Retention	Ratio	Array size
TaO_x/TiO_2[4] [7]	Bi-analog	9V/50μs	>10ks@RT	>2	Single cell
Mo/PCMO[8]	Bi-analog	3V/10ms	>30ks@120℃	>3	11k-bit array
Al_2O_3/TiO_{2-x}[9]	Bi-analog	1.1/500μs	>50ks	<2	12x12
$Pd/WO_x/W$[10][11]	Bi-analog	3V/400μs	$10^2 s$	>2	-
TiN/HfOx/Pt[12]	RESET analog	1.6V/100ns	-	>100	24X24
$Ag/GeS_2/W$[13]	Binary	2V/10μs	-	$>10^4$	8X8
PCM[14]	SET anglog	1V/1μs	-	>2	10x10
This work	Bi-Analog	1.6V/50ns	MLC >3ks@150℃	>10	1k-bit array

- ✓ **With material engineering, we demonstrate reliable bidirectional analog RRAM with fast speed, low voltage, and good retention**

Grey-scale real face recognition



Training set from Yale Face Database for experimental face classification



The schematic of the perceptron implemented by the 1K bit array

- ✓ **A perceptron is demonstrated experimentally for grey face pattern classification based on the 1K bit array to verify the capability of the device and the two operation schemes**

P. Yao, et al. Nat. Commun. 8, 15199, 2017

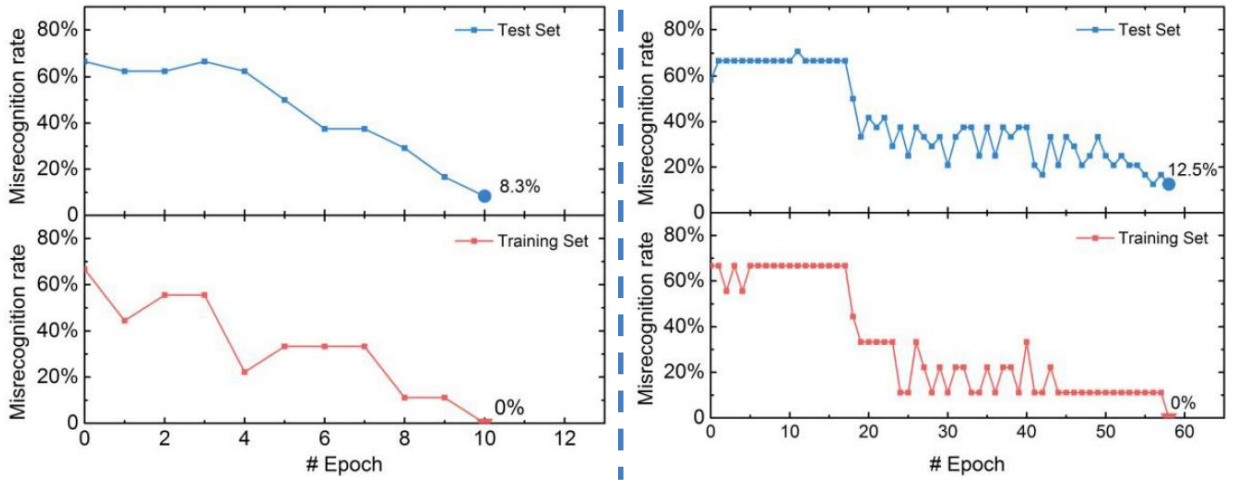
Experimental test process

Set 1: The 24 unseen test images

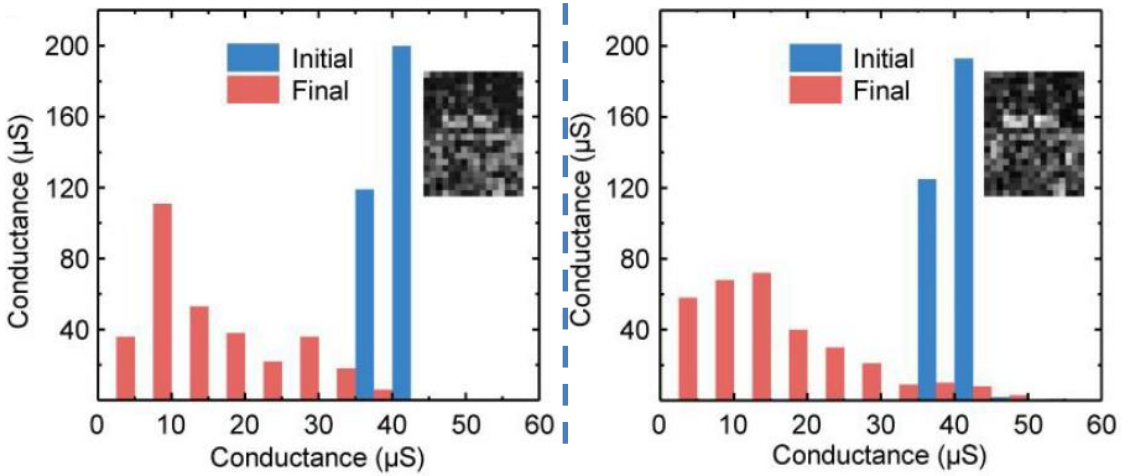
Set 2: Augmented noisy test patterns

- ✓ **Set 1: The test database from the Yale Face Database**
- ✓ **Set 2: Augmented noisy pattern set consists of 9,000 images by introducing noise to the training images. Noise patterns are generated by randomly choosing some pixels and assigning them a random value.**

Experimental training process

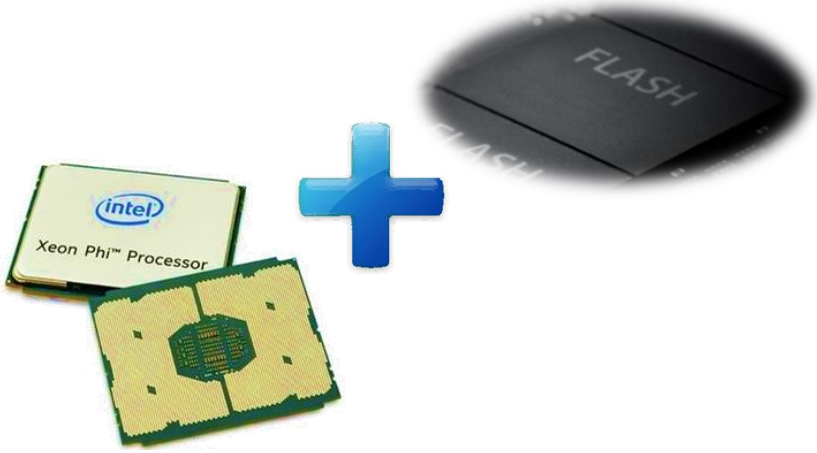


The activation function output value of the first class versus the iteration number

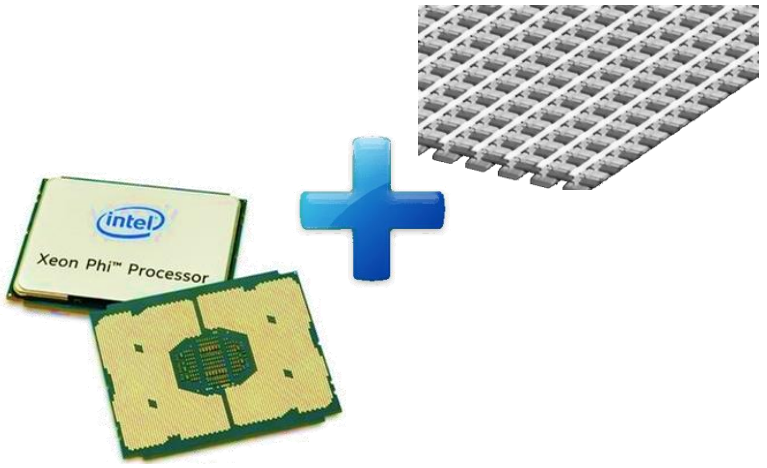


Initial and final conductance distribution comparison of the first row @ initial state LRS. Inset shows the final conductance map

Energy consumption saving



1000x energy saving than a Intel Xeon Phi processor with off-chip memory system



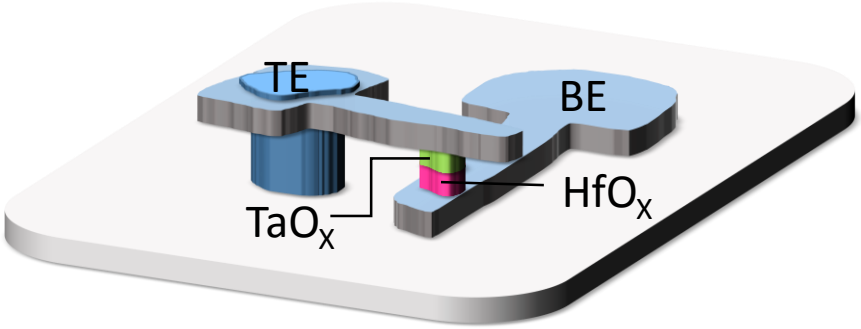
20x energy saving than a hypothetical Intel Xeon Phi processor with on-chip digital RRAM memory system

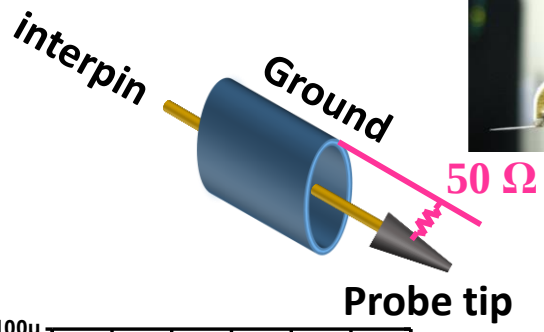
- ✓ **The two methods have similar energy consumption per iteration during training process (around 30 nJ / iteration). This shows a remarkable energy saving than the traditional way.**

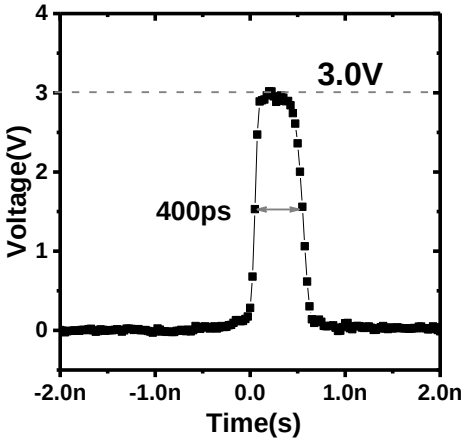
面临的挑战？

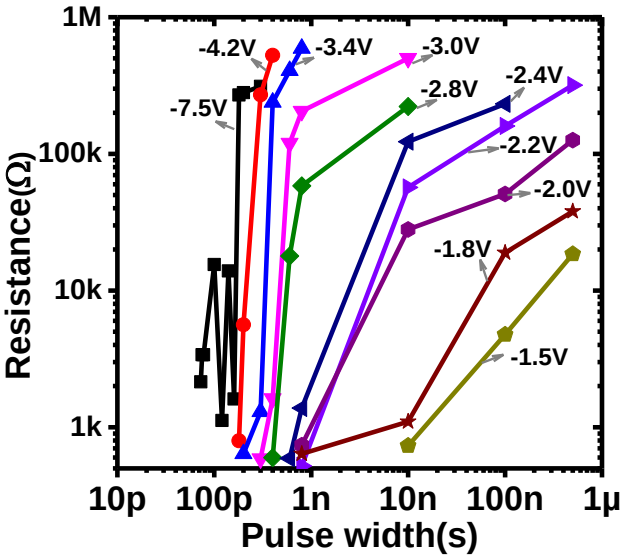
- 我们未来关于**RRAM**可靠性与表征的研究还存在诸多问题及挑战
  - **RRAM**器件
    - ✓ 瞬态测量
    - ✓ 循环次数测试
    - ✓ 微观原位表征
  - **RRAM**阵列
    - ✓ 自动测试方法
    - ✓ 读取速度

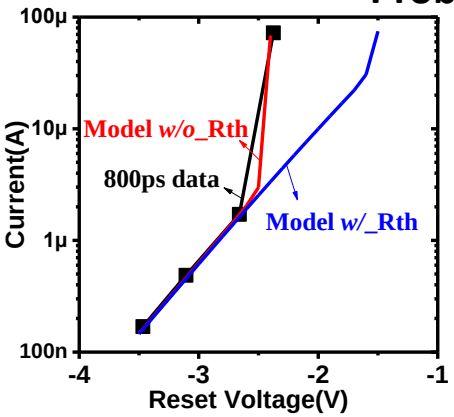
阻变速度的极限







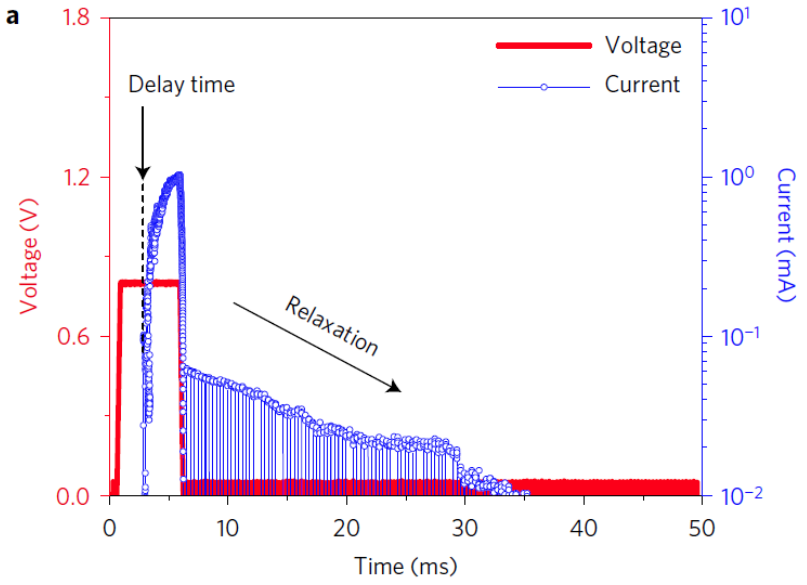


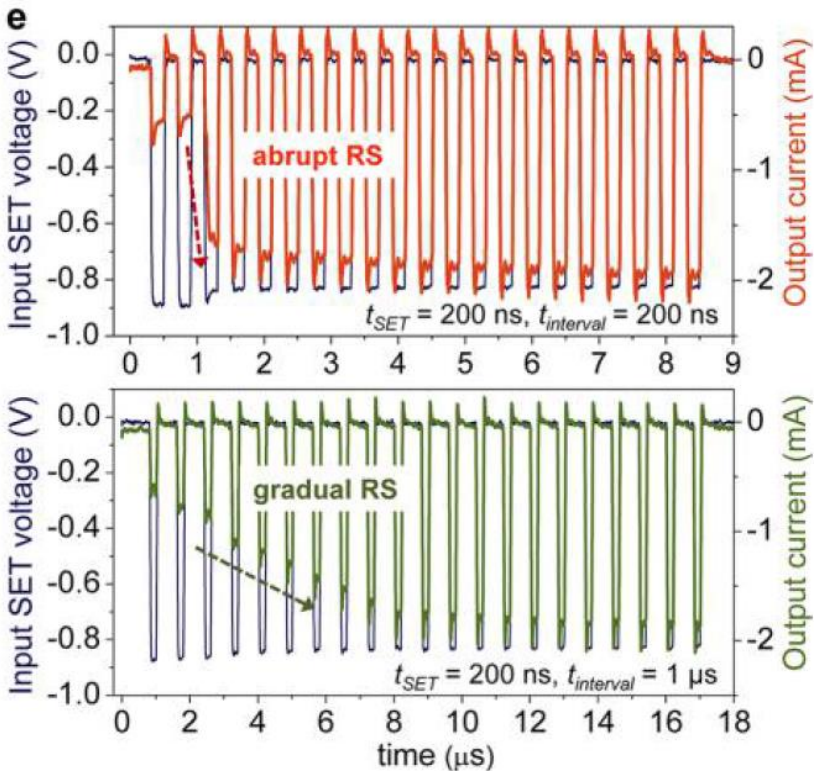


通过高速脉冲可以屏蔽热效应对阻变过程的影响

Adv. Ele. Mater. 2017

阻变动力学过程

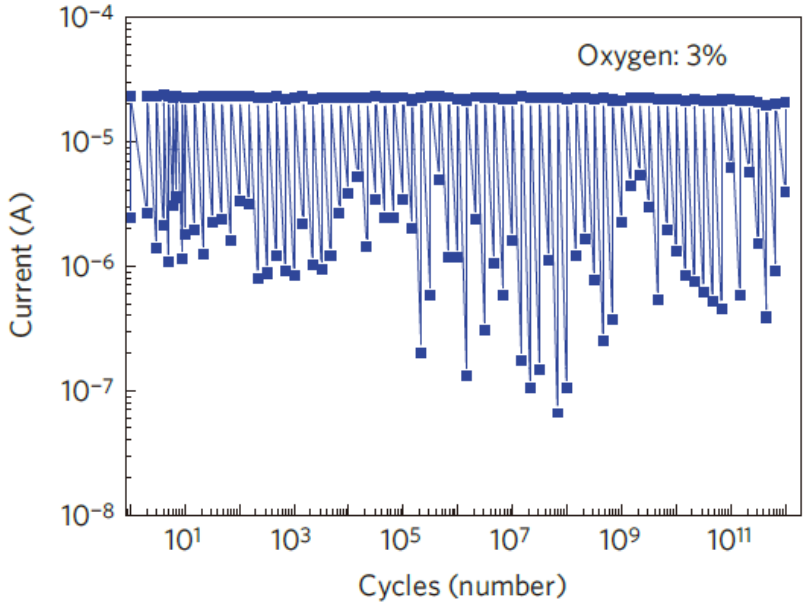




通过瞬态电流的测量可以了解阻变动力学过程，获得神经形态特性的调控方法

Z. Wang, et al, Nature Materials, 2017
C. Du et al, Nano Letters, 2015, 15, 2203

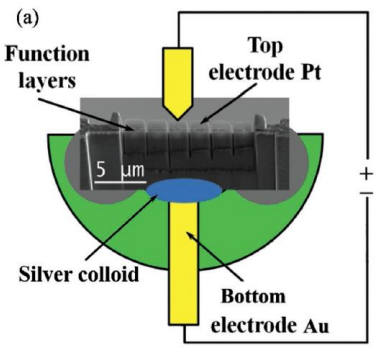
循环次数的极限

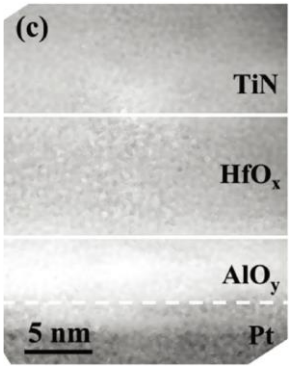


每次读写切换消耗大量的时间。
要准确评估器件的循环次数（**endurance**），需要对每次擦写后的电阻进行读取，单器件需要10^{12}以上的切换。
阵列上则需要10^{18}以上的读写次数。

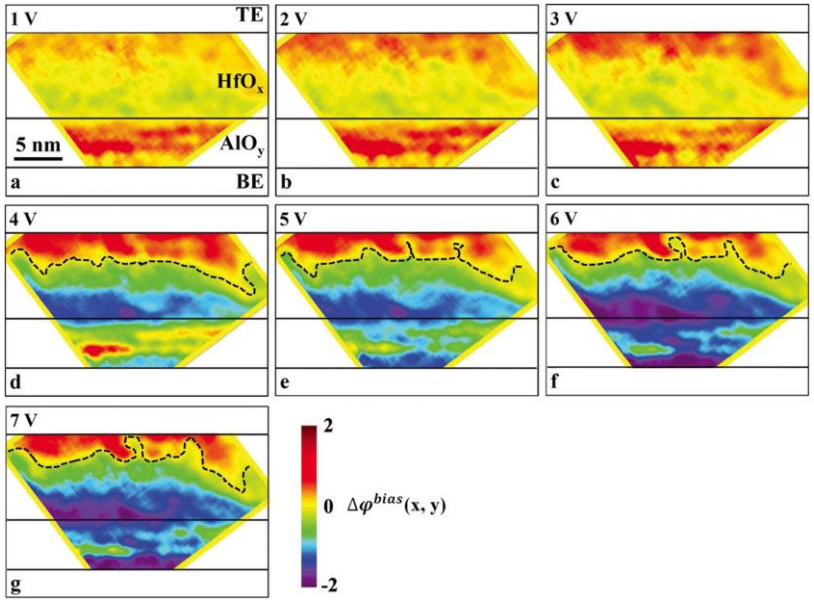
M.-J. Lee, et al, Nature Materials, 2011, 10, 625

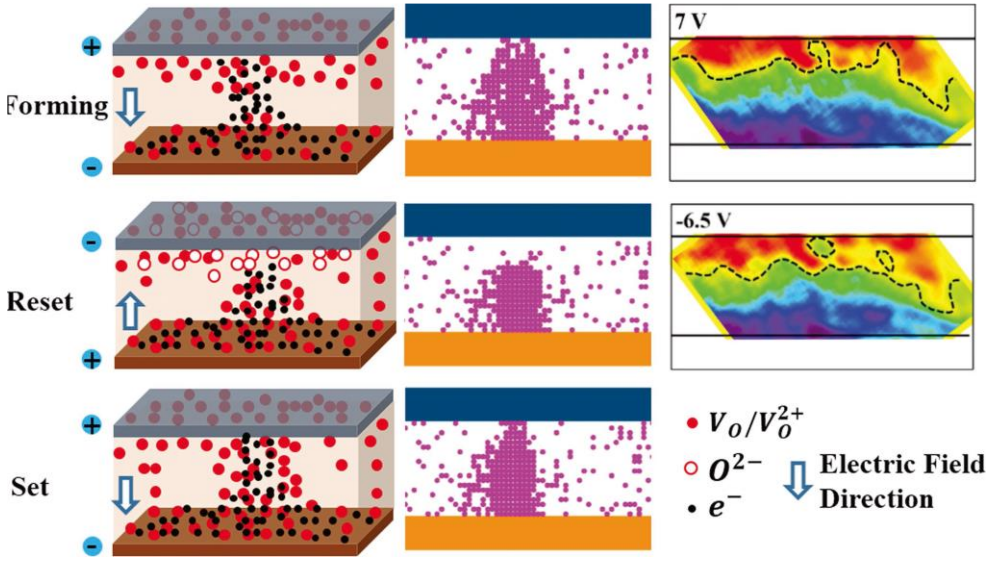
氧空位分布的原位观测





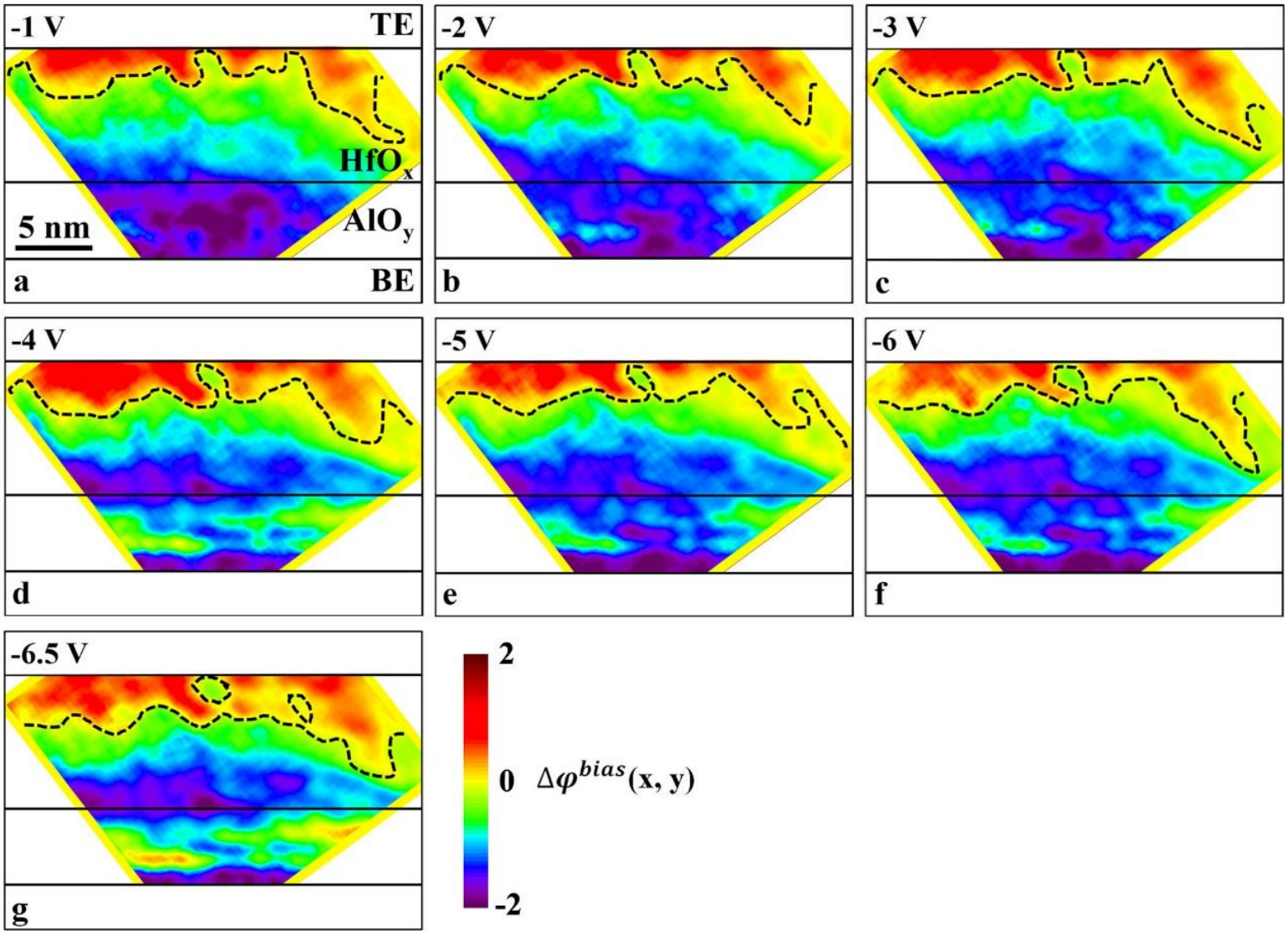
要理解阻变微观机理，了解氧空位的分布及演化情况，需要原子级的原位表征手段，对氧空位导电通道的形貌进行三维成像，并可以随时调控器件的状态。



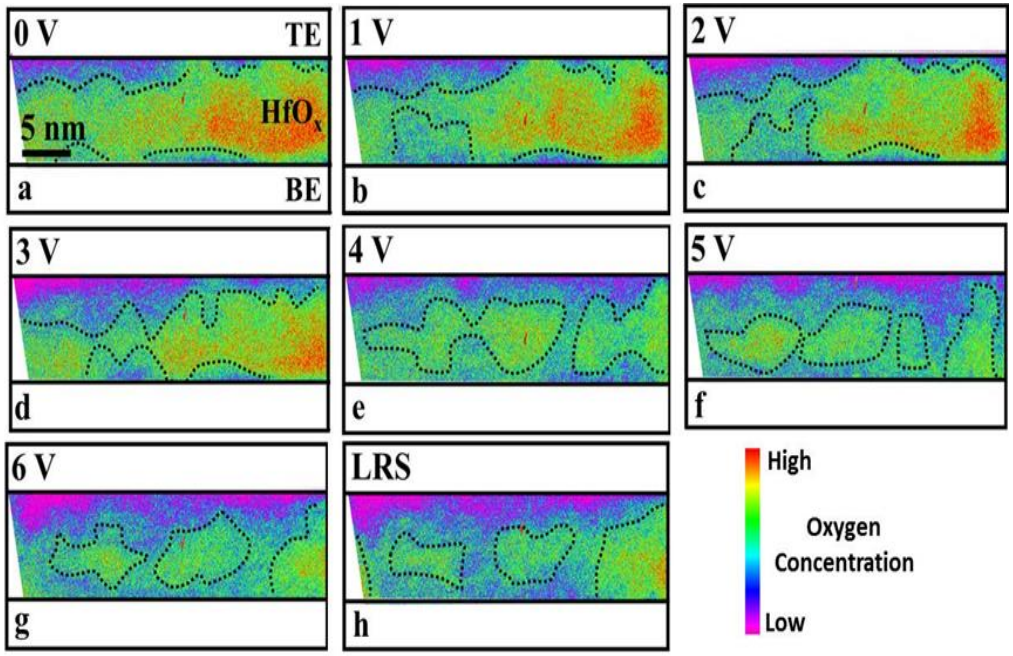


Adv. Mater. 2017

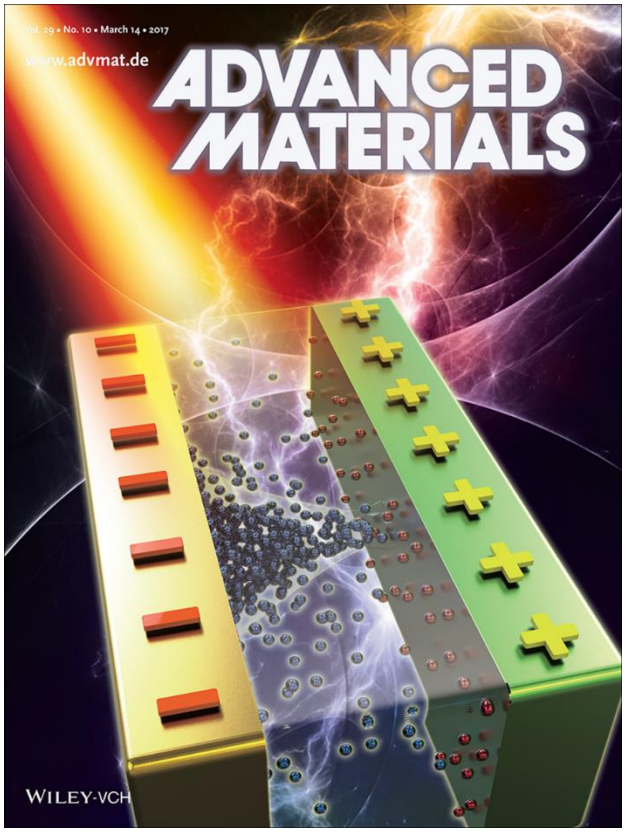
CF rupture during RESET



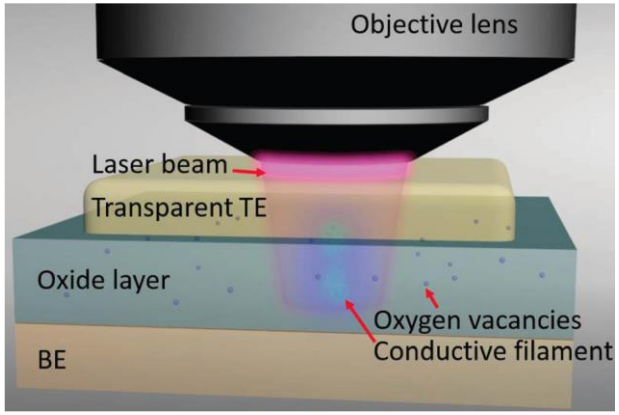
CF formation during forming

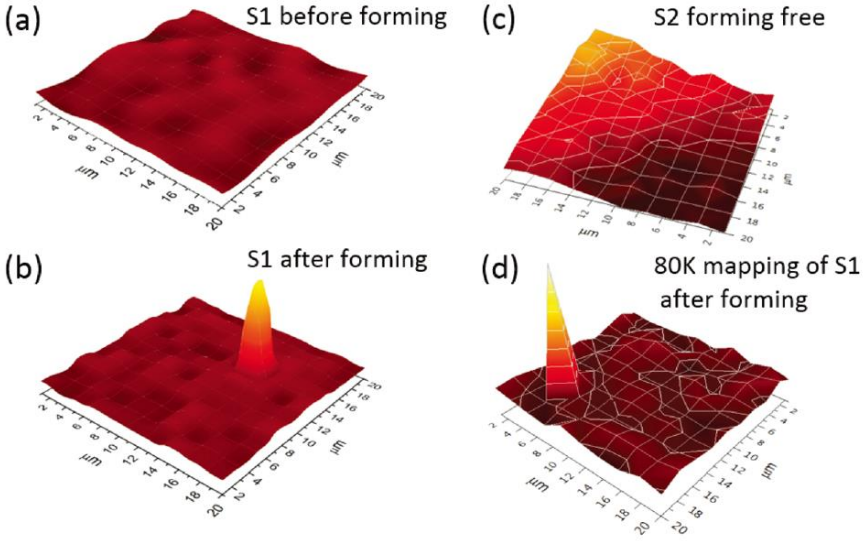


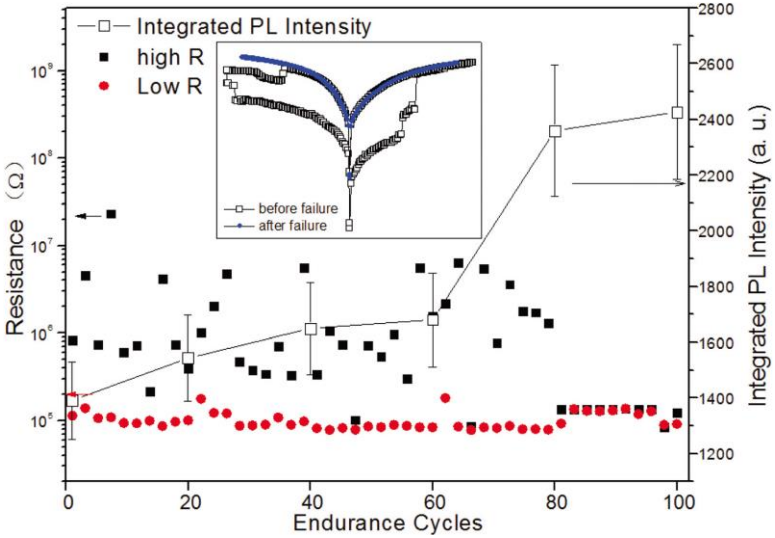
EELS mapping



氧空位分布的非破坏性观测

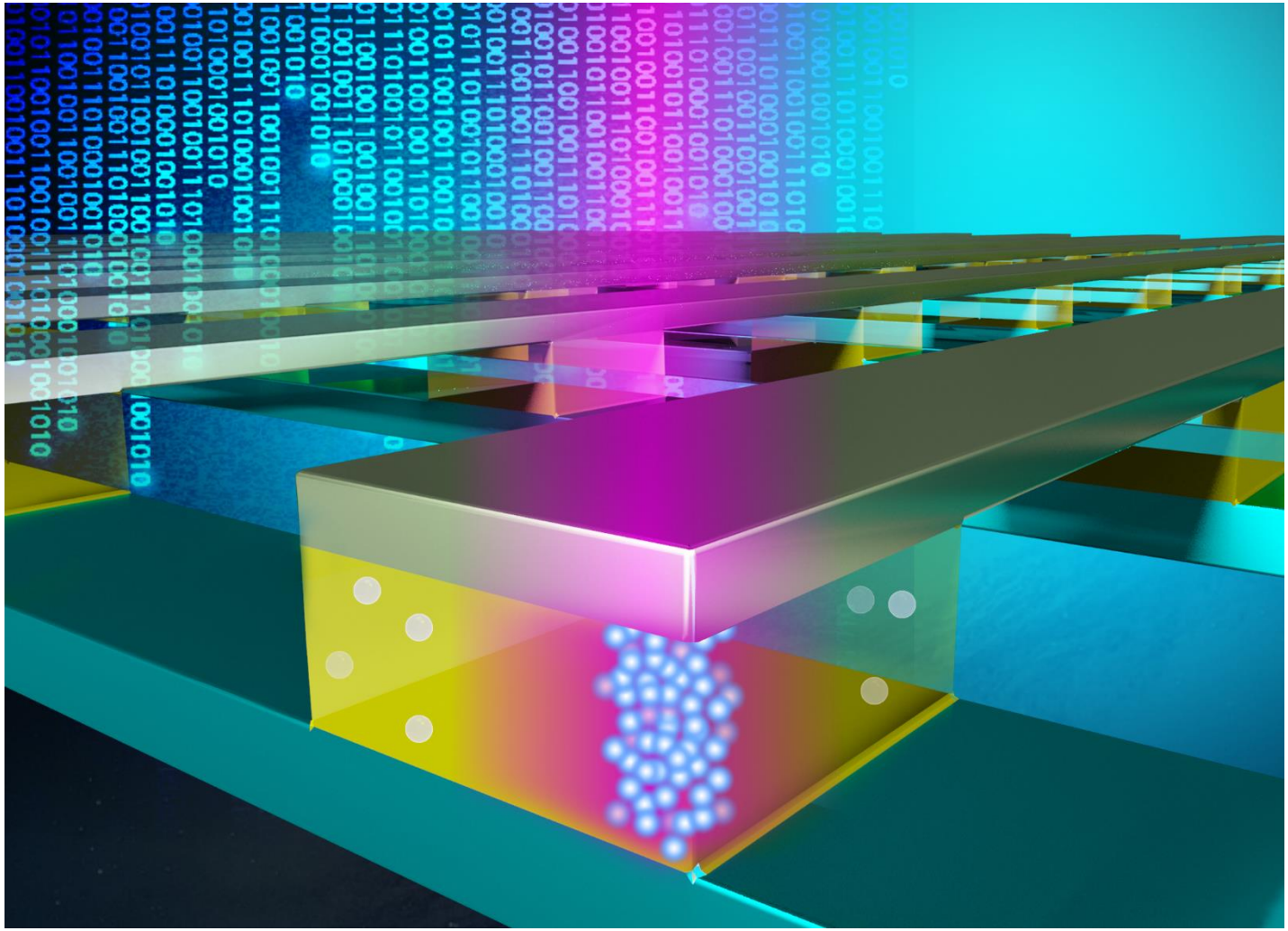




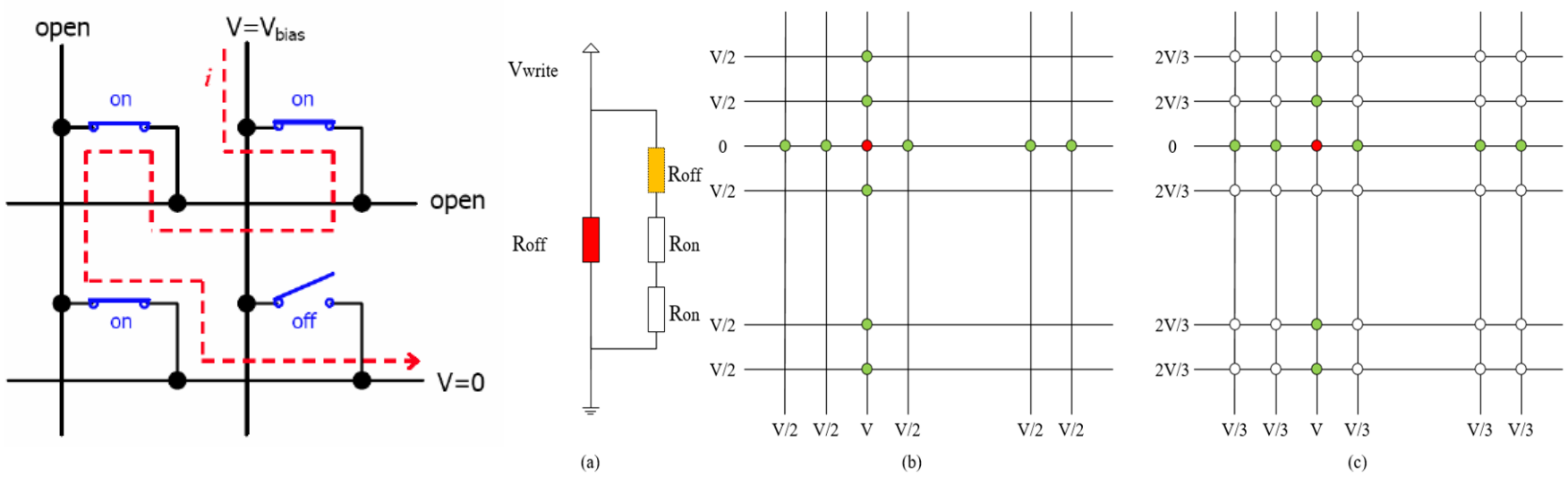


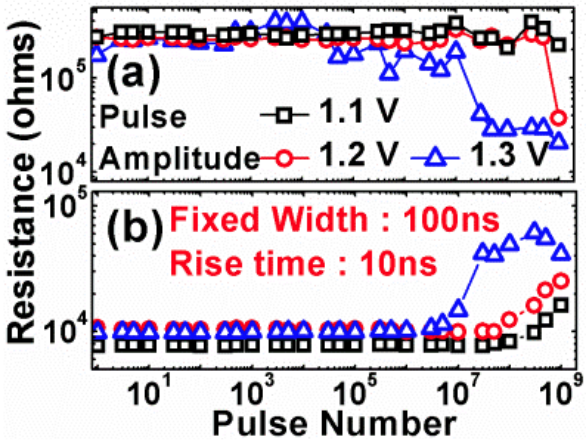
要理解氧空位分布与实际器件可靠性之间的关系，需要非破坏性的物性表征手段，在实际器件上实时观测氧空位的分布变化。

Nanoscale 2017



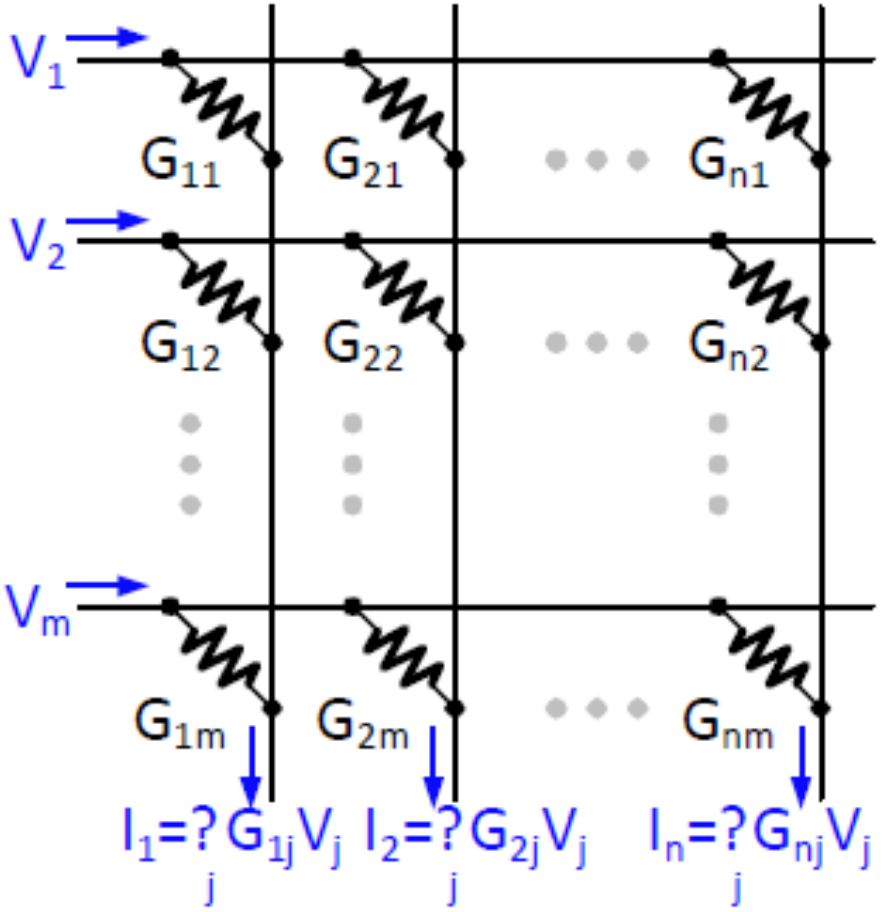
交叉阵列中的串扰





阻变存储器阵列在操作中存在多种串扰效应，为抑制这些效应，需要对每条BL、WL同时加电压；受器件特性的影响，所加电压可能不同。因此，需要开发灵活的阵列测试模块，可以多通道同时加电压。

神经网络并行处理

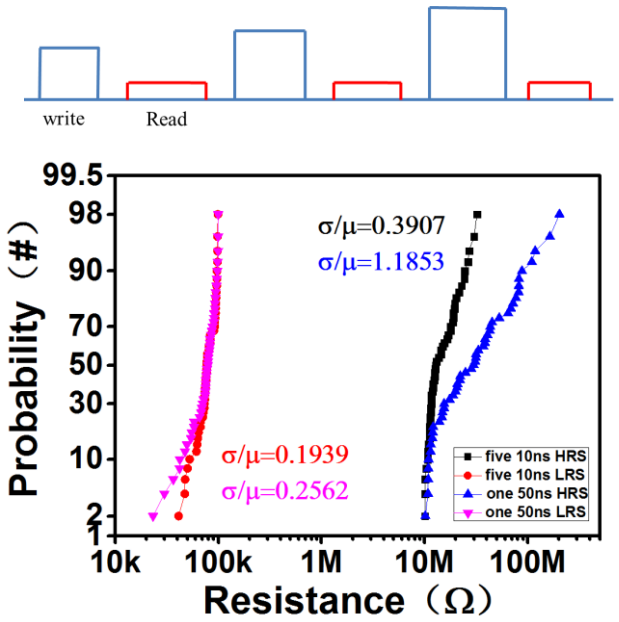


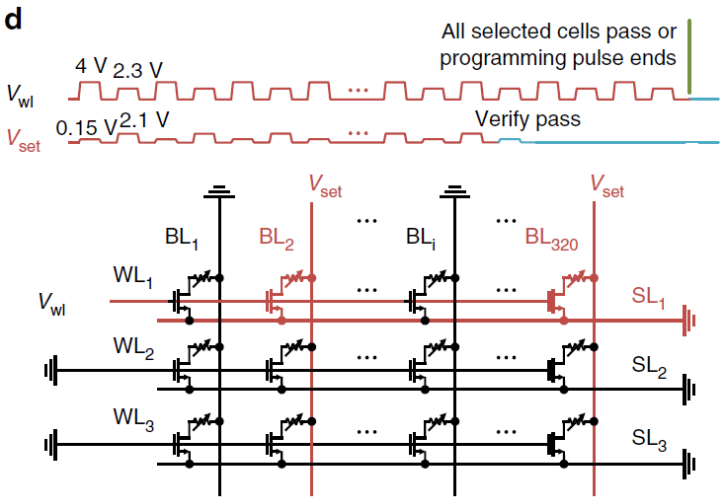
基于交叉阵列的神经网络单元研究将是未来相当长时间内的重要方向。近期关于阵列操作的文章均发表在Nature及其子刊上。

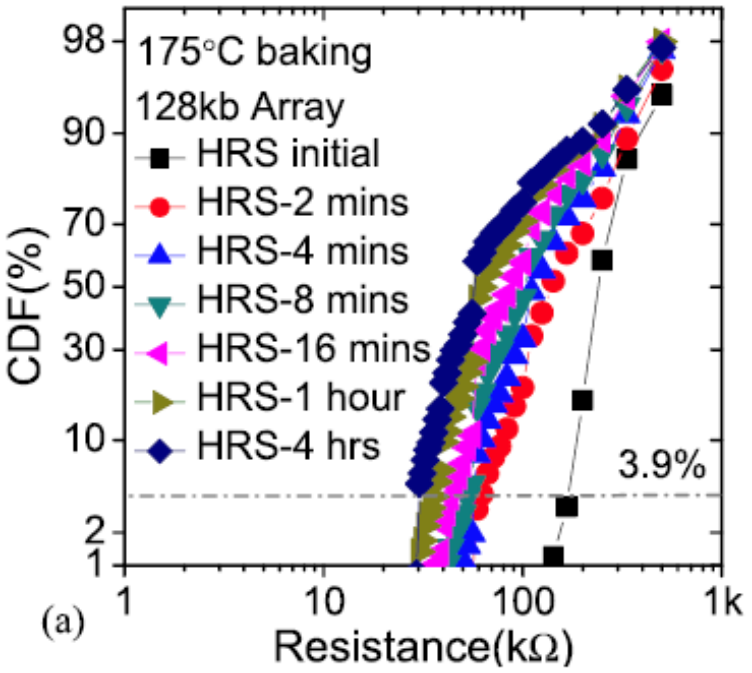
神经网络最基本的操作是在所有WL上同时加不同幅值的电压，同时在所有BL上读取电流，需要输入多通道的模拟电压值，同时能够多通道同时读取模拟电流值。

目前类似的过程都是靠电路或者FPGA完成的，灵活性差、周期长，不适合研发。通用阵列测试设备将大大提升研发效率。

自动校验与弛豫







阻变存储器阵列在擦写过程中需要对每个单元反复校验，以抑制涨落和弛豫效应对tail bits的影响。单器件的测试程序很容易开发，但是阵列级的verify自动测试方法则是未来急需的工具。

总结

- 新型存储器将是未来集成电路领域重要的发展方向
  - 兼备高速度、高密度和非挥发性，简化存储器系统
  - 颠覆传统冯诺依曼架构，融合计算与存储，神经形态计算
- 阵列测试将成为未来RRAM研究中的主要课题
  - 阵列操作模式、可靠性、一致性、神经计算等
- RRAM表征技术需要向极端化发展
  - 原子级的原位表征
  - ps级脉冲擦写及信号捕捉
  - 快速读写阵列单元
  - 大规模阵列的自动测试

谢谢!

高滨 助理教授/博士生导师

13718817295, 62771394

gaob1@tsinghua.edu.cn