Tektronix®

清華大學 Tsinghua University

忆阻器存算一体芯片与类脑计算

高滨

清华大学集成电路学院

报告内容

清华大学 Tsinghua University

- 人工智能与芯片算力
- 存算一体技术
- 研究进展
- 器件测试的新挑战

人工智能带来新机遇

Alibaba Google Baidu IBM Microsoft

图形识别

AI Cloud

AI 医疗健康

翻译

智能家居

语音识别

人工智能的三大要素

- 以深度学习为核心技术的人工智能有三大关键要素
  - ✓ 算法
  - ✓ 数据
  - ✓ 算力

算力需求的爆炸式增长

- 2020年GPT-3一炮走红
  - 超越普通NLP，写小说、编剧本无所不能
  - 5亿美元的超算中心，数千个GPU
  - 训练用了三个多月的时间，电费几千万美元
  - 训练过程中发现程序bug怎么办？忍着……

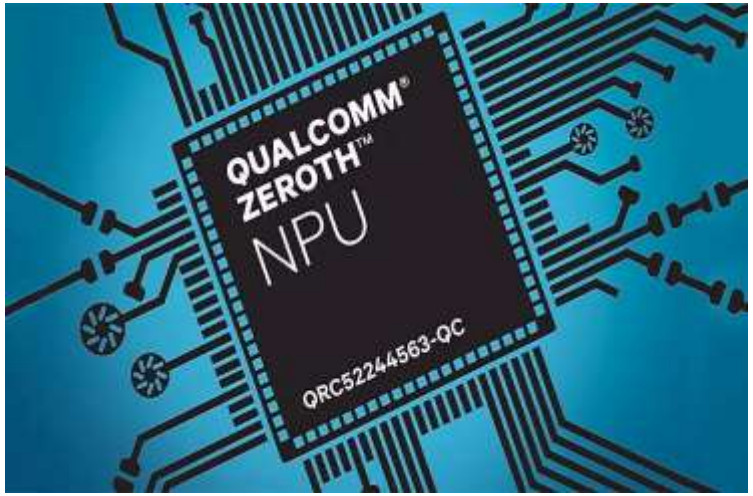
模型	发布时间	参数量
GPT	2018 年 6 月	1.17 亿
GPT-2	2019 年 2 月	15 亿
GPT-3	2020 年 5 月	1750 亿

算力需求的爆炸式增长

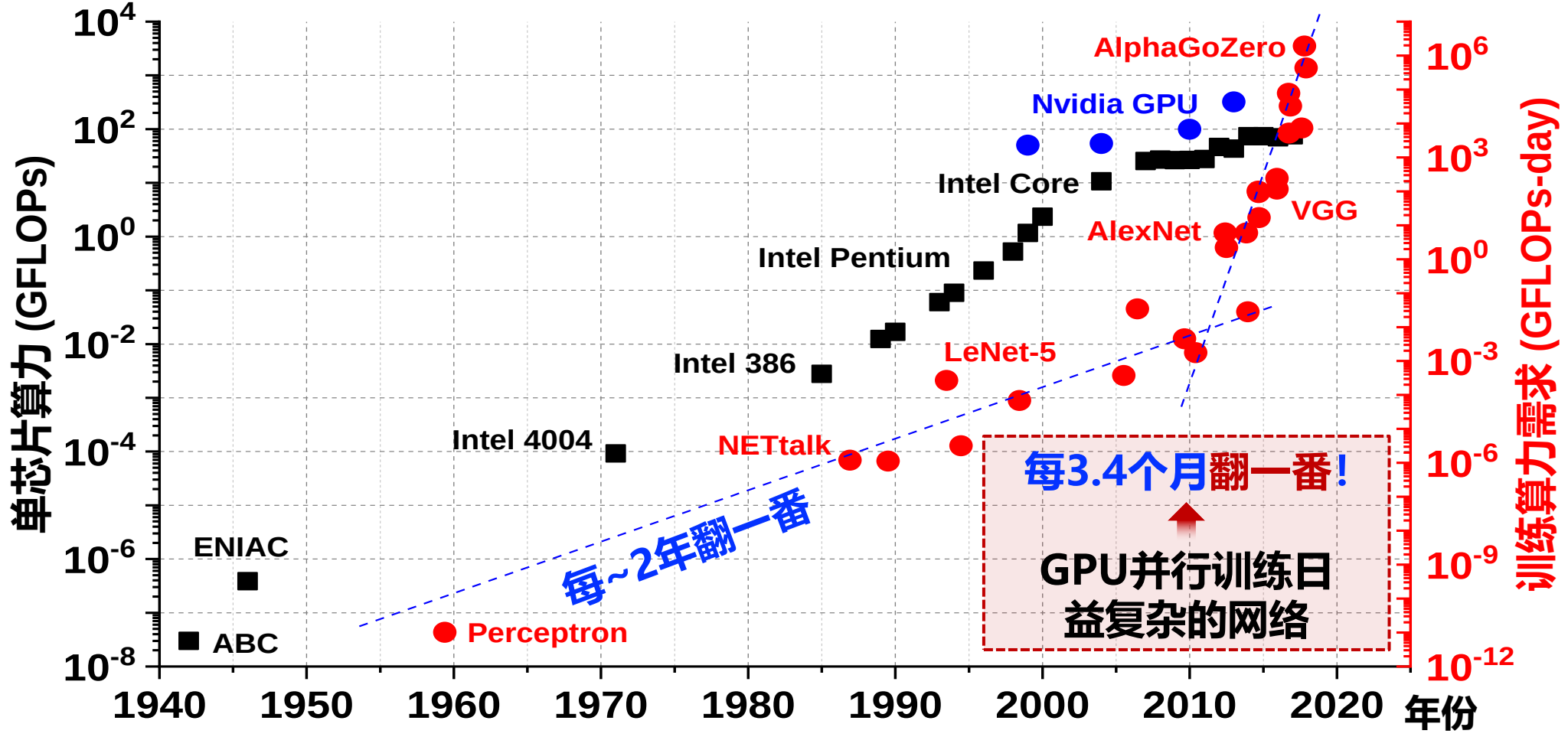
- 自动驾驶汽车
  - 2017年从1颗GPU升级为2颗
  - 2020年升级为4颗
  - 2021年？买不到了……

算力需求的爆炸式增长

- 手机拍照
  - 2017年开始全面嵌入NPU
  - NPU从单核变为多核
  - 4个算法只能处理半个……



芯片算力 vs. 人工智能高需求的矛盾



算力需求快速增长 **尖锐矛盾** **算力提升放缓**

数据来源:
Intel
NVIDIA
OpenAI

急需颠覆性新技术

人工智能的硬件平台面临

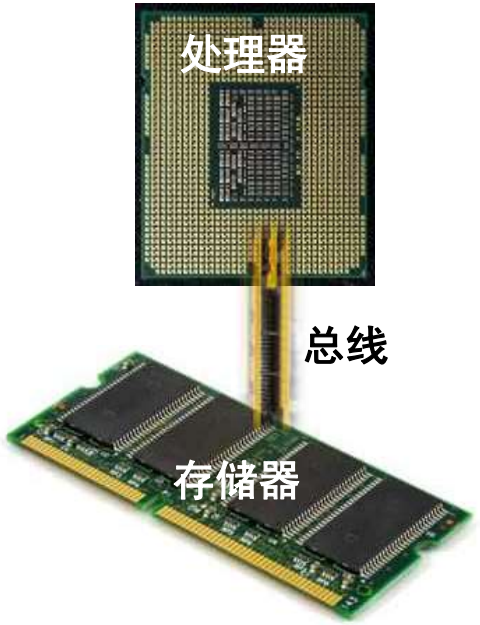
1. 算力不足

2. 能效过低

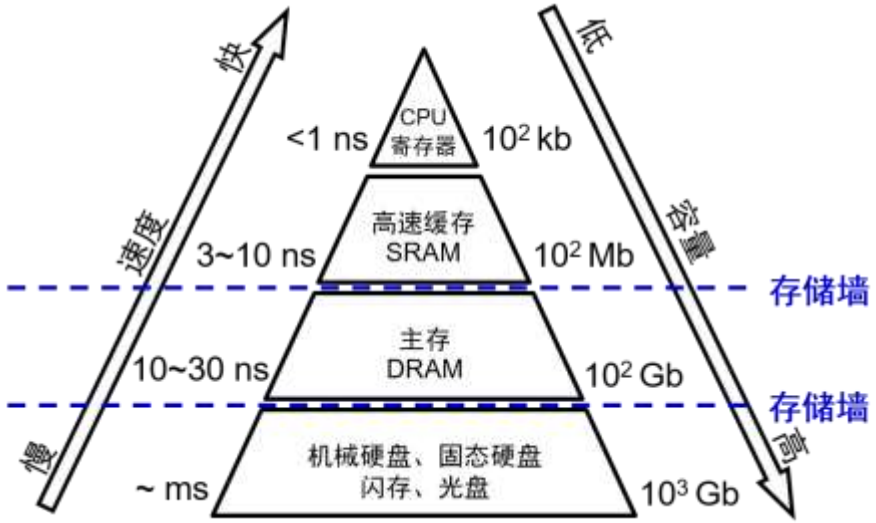
两大艰巨挑战

原因：“存算分离”架构的瓶颈

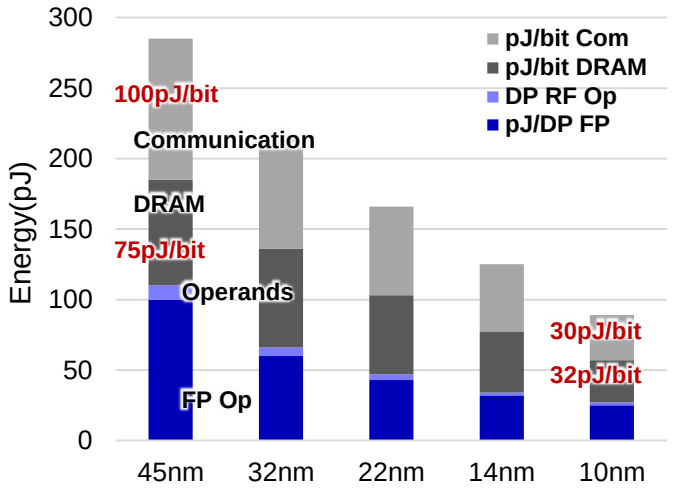
现有计算系统普遍采用存储和运算分离的架构，存在“存储墙”与“功耗墙”瓶颈，严重制约了系统算力和能效的提升。



冯诺依曼架构



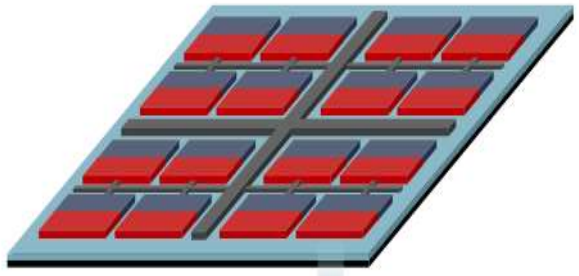
存储速度远低于计算速度



10nm工艺下，数据传输和访问功耗占比超过69%

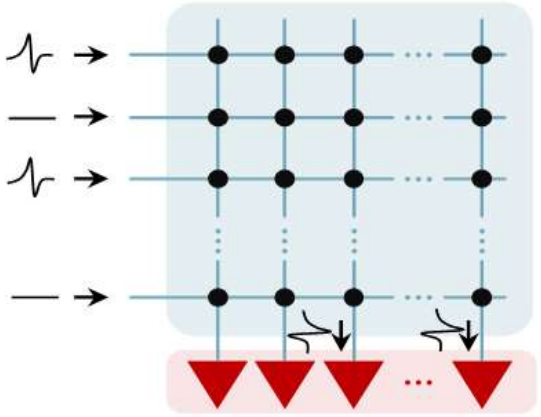
向大脑学习

类脑芯片：学习大脑的结构和/或工作机理



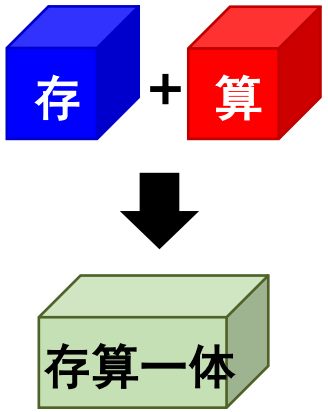
神经元-突触结构

高并行



脉冲事件驱动

低功耗



存算一体

高算力

高滨, Nature Electronics, (2020)

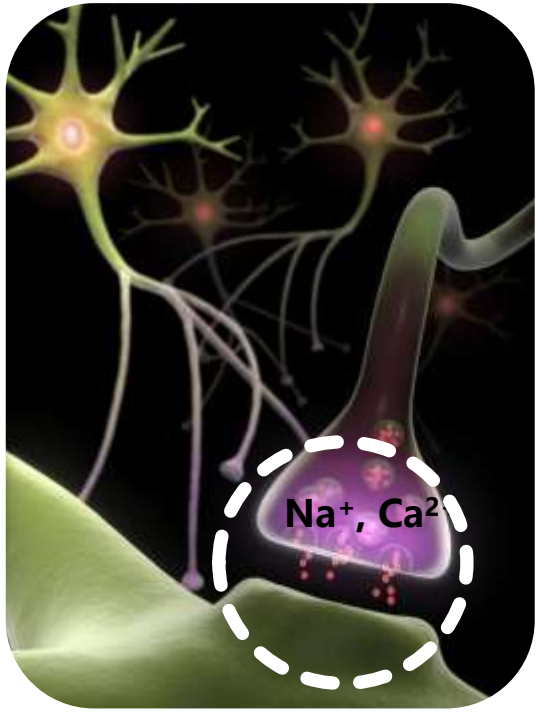
报告内容

清華大學 Tsinghua University

- 人工智能与芯片算力
- **存算一体技术**
- **研究进展**
- **器件测试的新挑战**

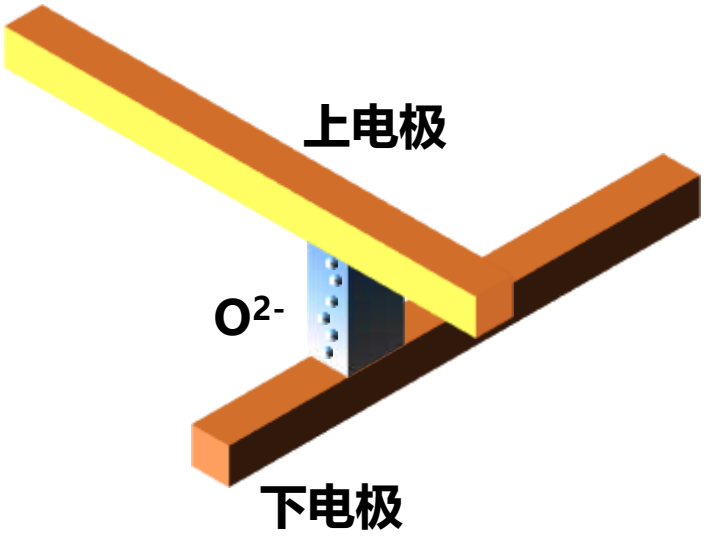
新器件：存算合一的电子突触——忆阻器(RRAM)

生物突触

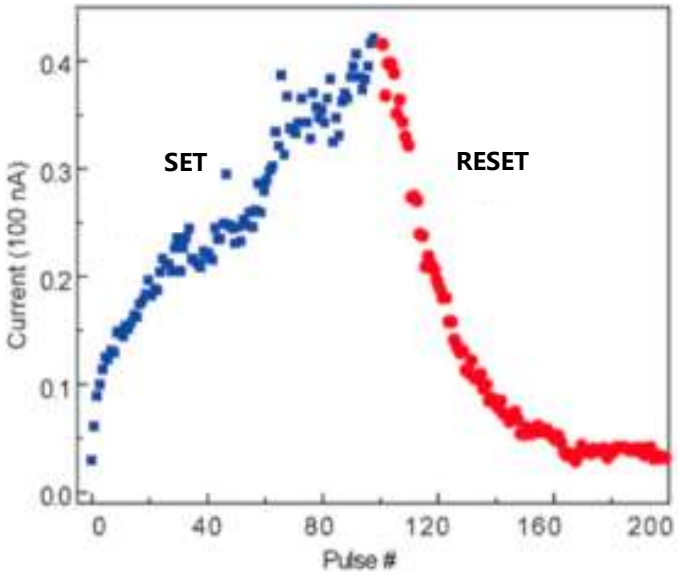


生物中 Na^{+}, CA^{2+} 离子的移动造成生物突触的强弱。

RRAM 电子突触 (Resistive Random Access Memory)



RRAM中氧离子的移动和分布导致器件阻值的变化。



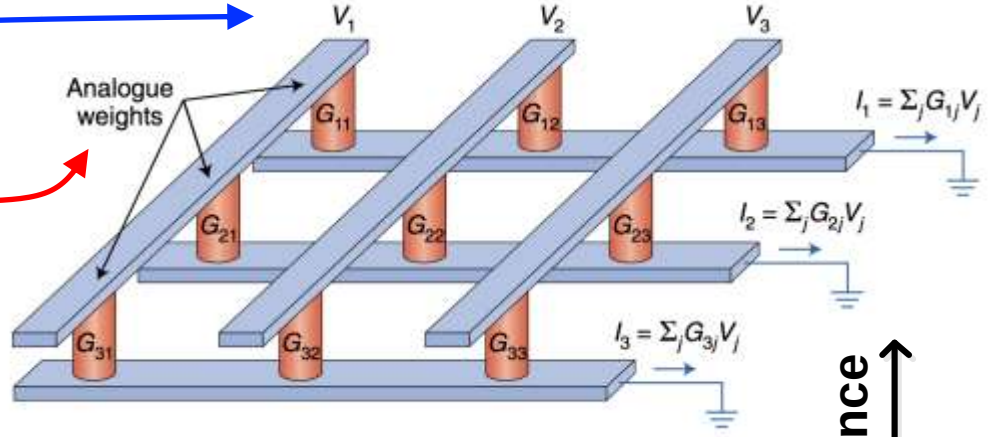
忆阻器与神经网络

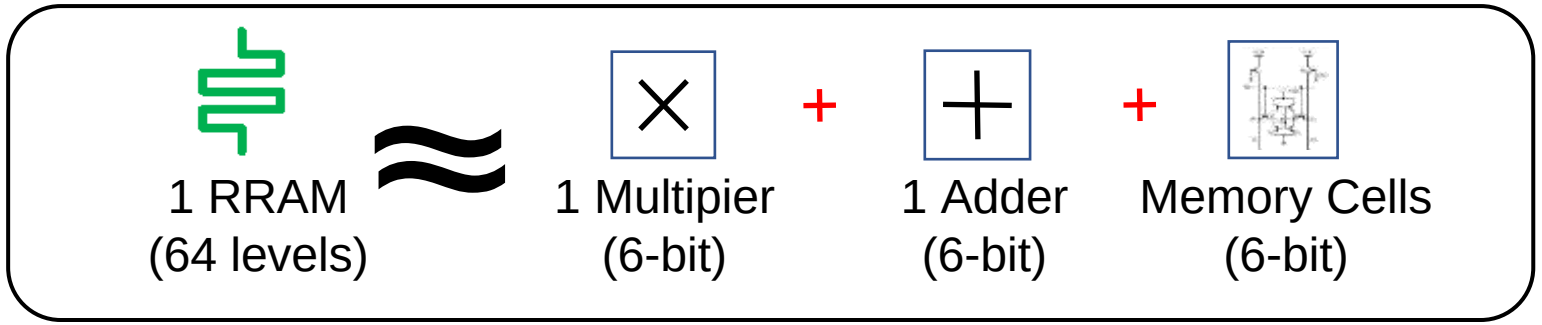
Ohmic laws

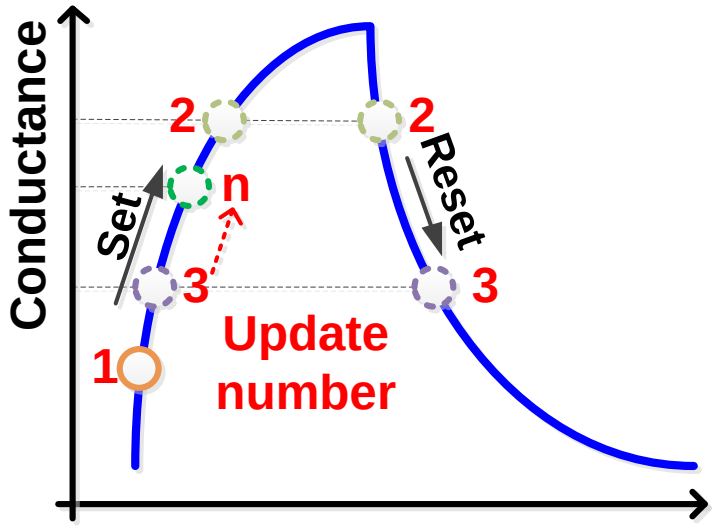
Kirchhoff's laws

$$[x_1 \quad x_2 \quad x_3] * \begin{bmatrix} G_{11} & G_{21} & \dots & G_{n1} \\ G_{12} & G_{22} & \dots & G_{n2} \\ G_{13} & G_{23} & \dots & G_{n3} \end{bmatrix}$$

$$= [\Sigma G_{1j}x_j \quad \Sigma G_{2j}x_j \quad \Sigma G_{3j}x_j]$$

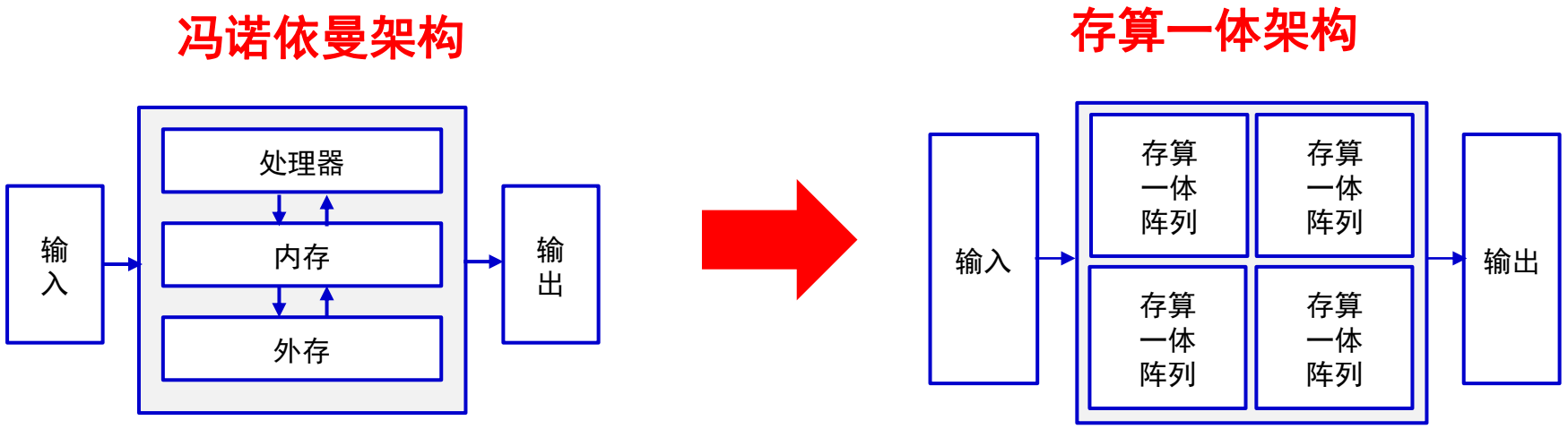






D. Ielmini and H.-S. P. Wong, *Nat. elec.*, 2018
L. Xia *et al.*, *DATE*, 2016

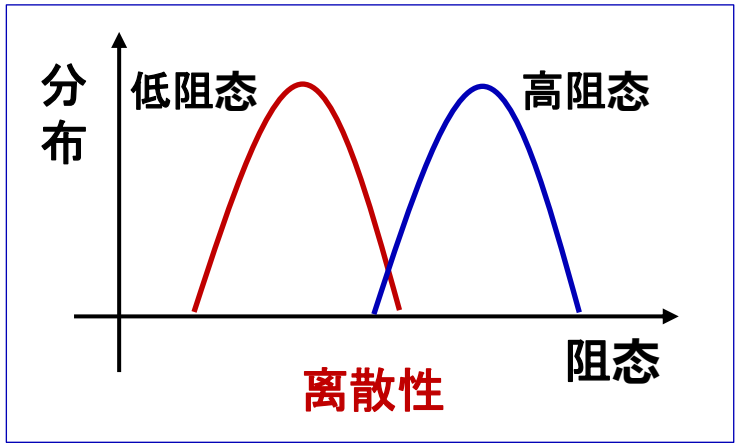
基于忆阻器存算一体技术

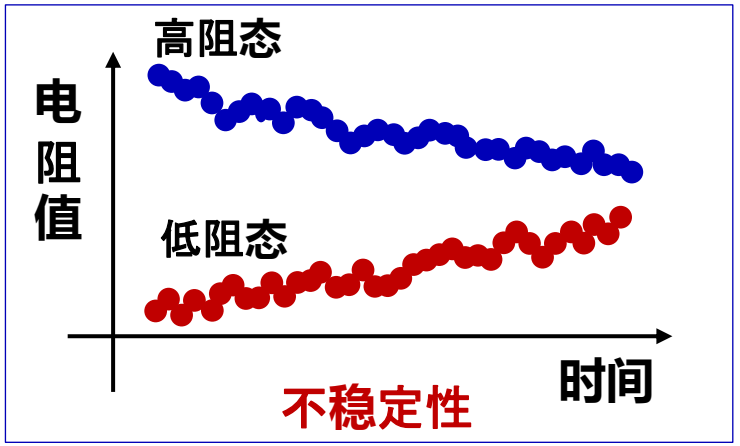


	计算器件	计算范式	算子	架构
传统计算系统	场效应晶体管	布尔逻辑数字计算	与门，非门	存算分离
存算一体计算系统	忆阻器	物理定律模拟计算	乘法，加法	存算一体

核心挑战1：器件非理想特性

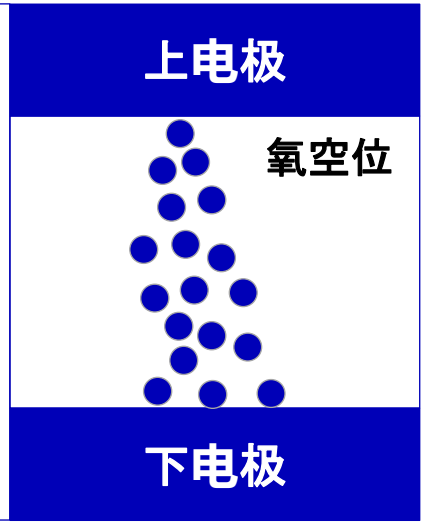
忆阻器件性能存在**离散性**和**不稳定性**的挑战，严重影响计算精度。





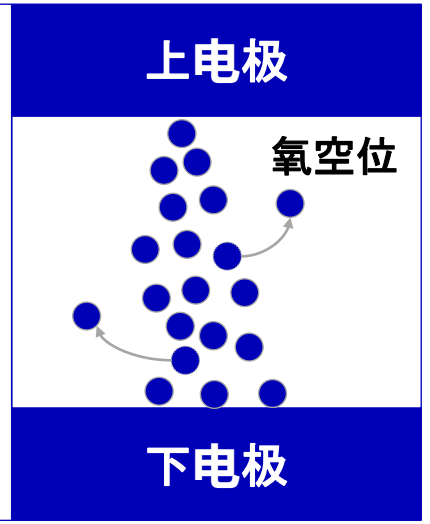
离散性：

器件中离子运动的随机性



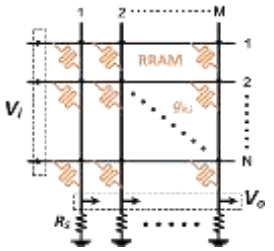
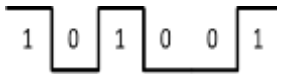
不稳定性：

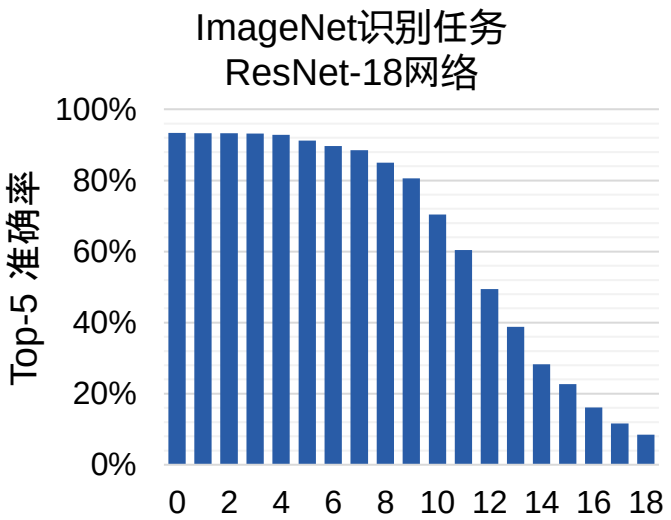
器件中离子分布形貌的弛豫效应



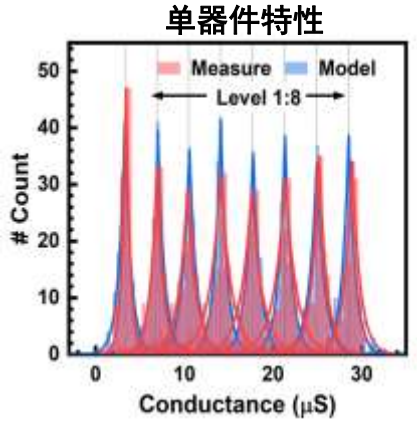
核心挑战2：模拟计算的误差累积

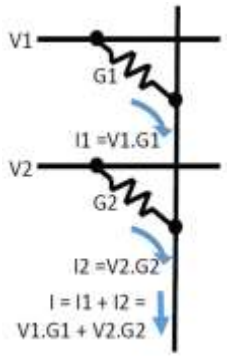
关键挑战

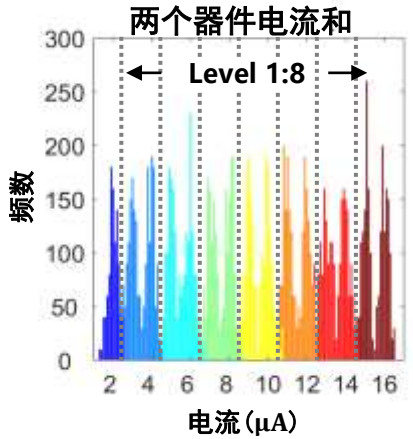
计算原理	逻辑计算	模拟计算
计算目标	7.71-2.69=5.02 8.24-7.63=0.61 8.06-6.96=1.1 8.99-8.84=0.15 5.75-1.24=4.51	
数据类型		

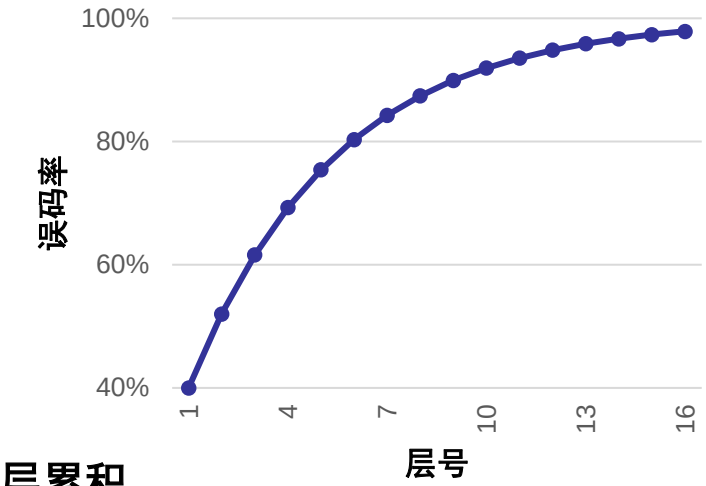


前n层引入1%误差





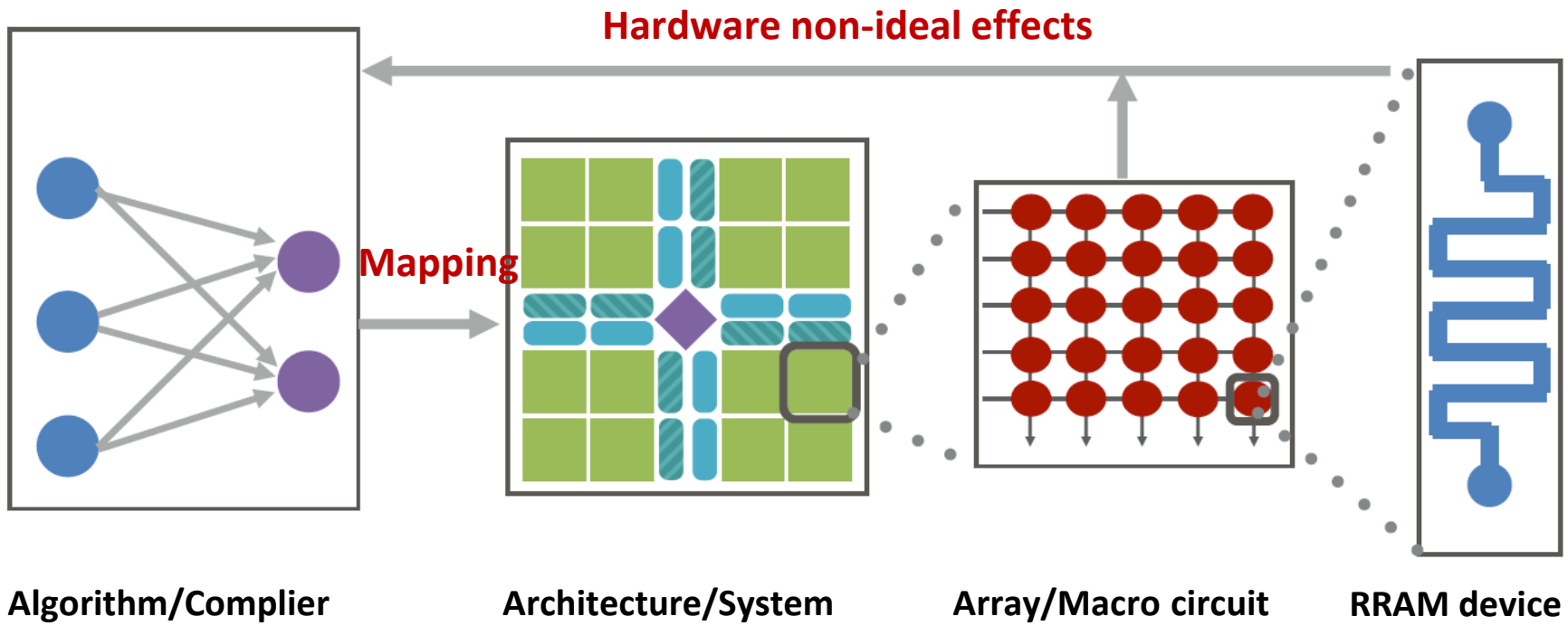




器件波动 → 逐器件累积 → **阵列误差** → 逐层累积 → **网络误差**

存算一体芯片的协同设计方法

- 存算一体芯片急需跨层次的协同优化方法：单一层面的优化难以达到高性能



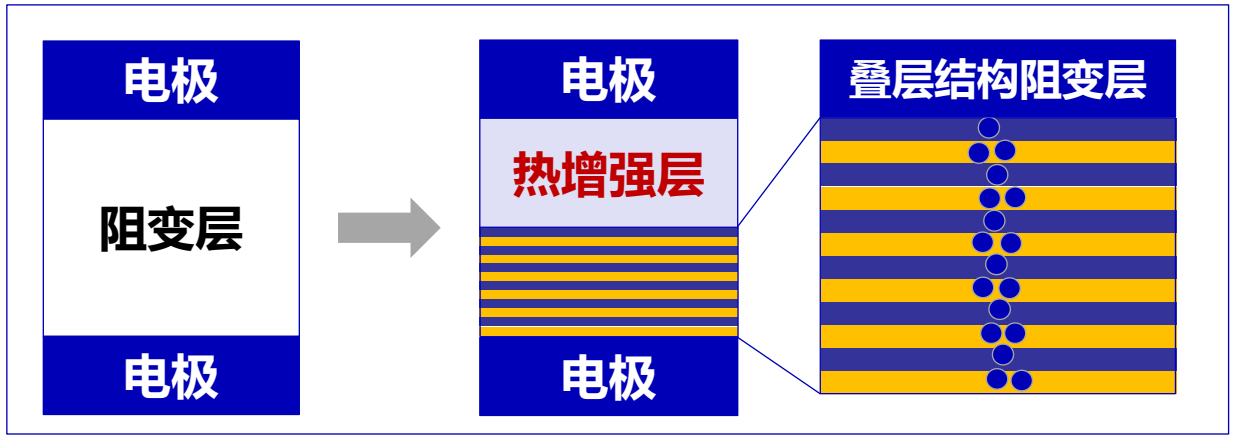
报告内容

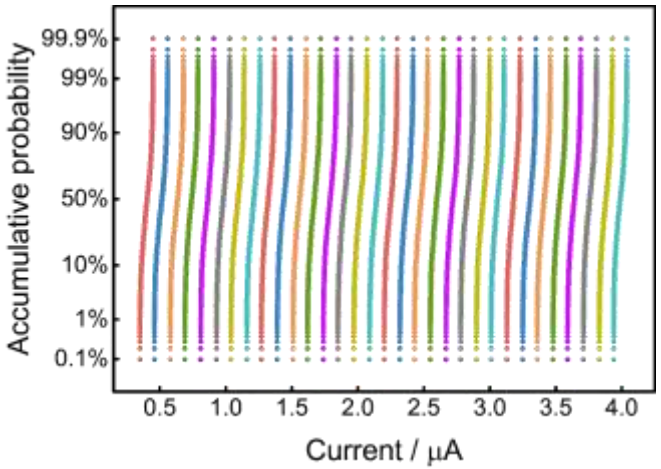
清华大学 Tsinghua University

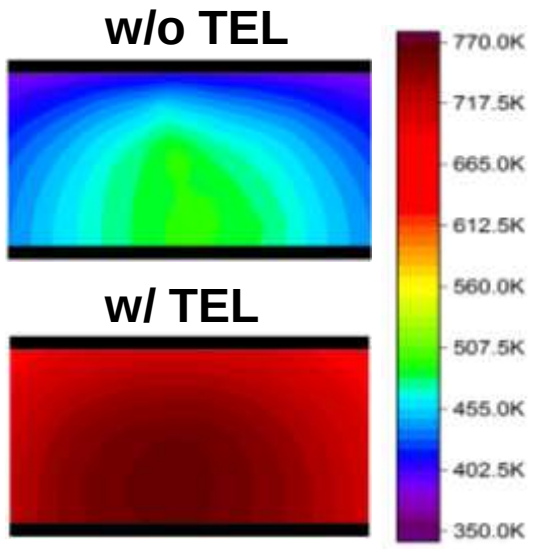
- 人工智能与芯片算力
- 存算一体技术
- **研究进展**
- **器件测试的新挑战**

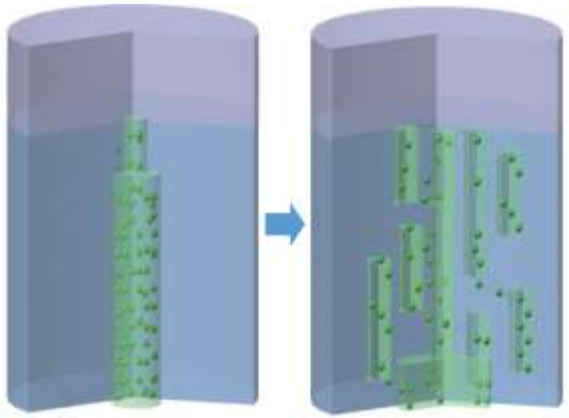
高性能忆阻器件

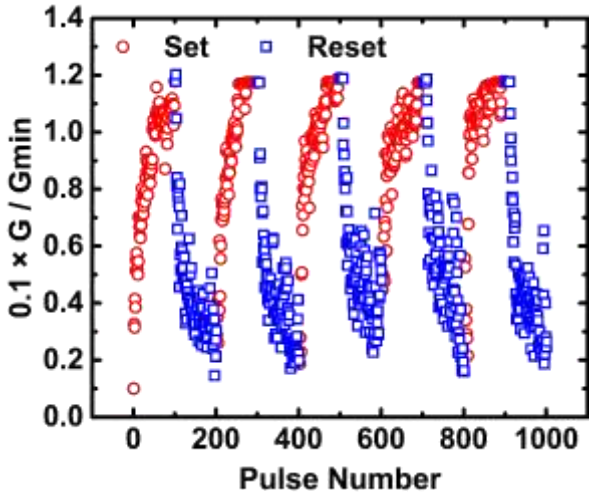
提出热增强层的新器件结构，在较小电流下实现了模拟阻变





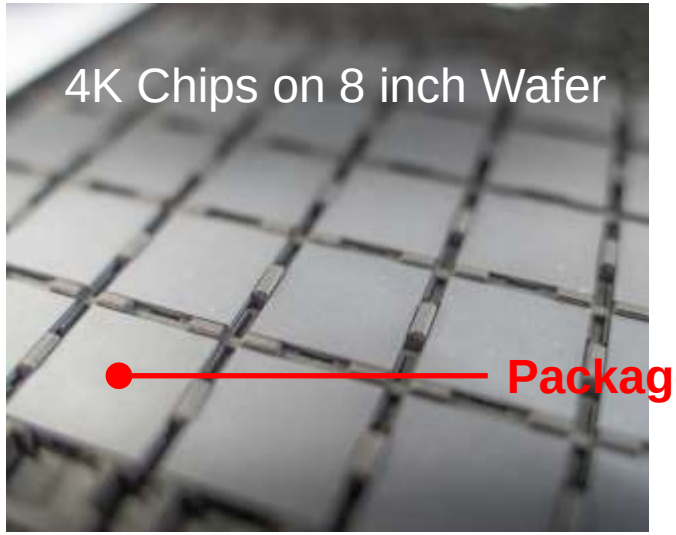


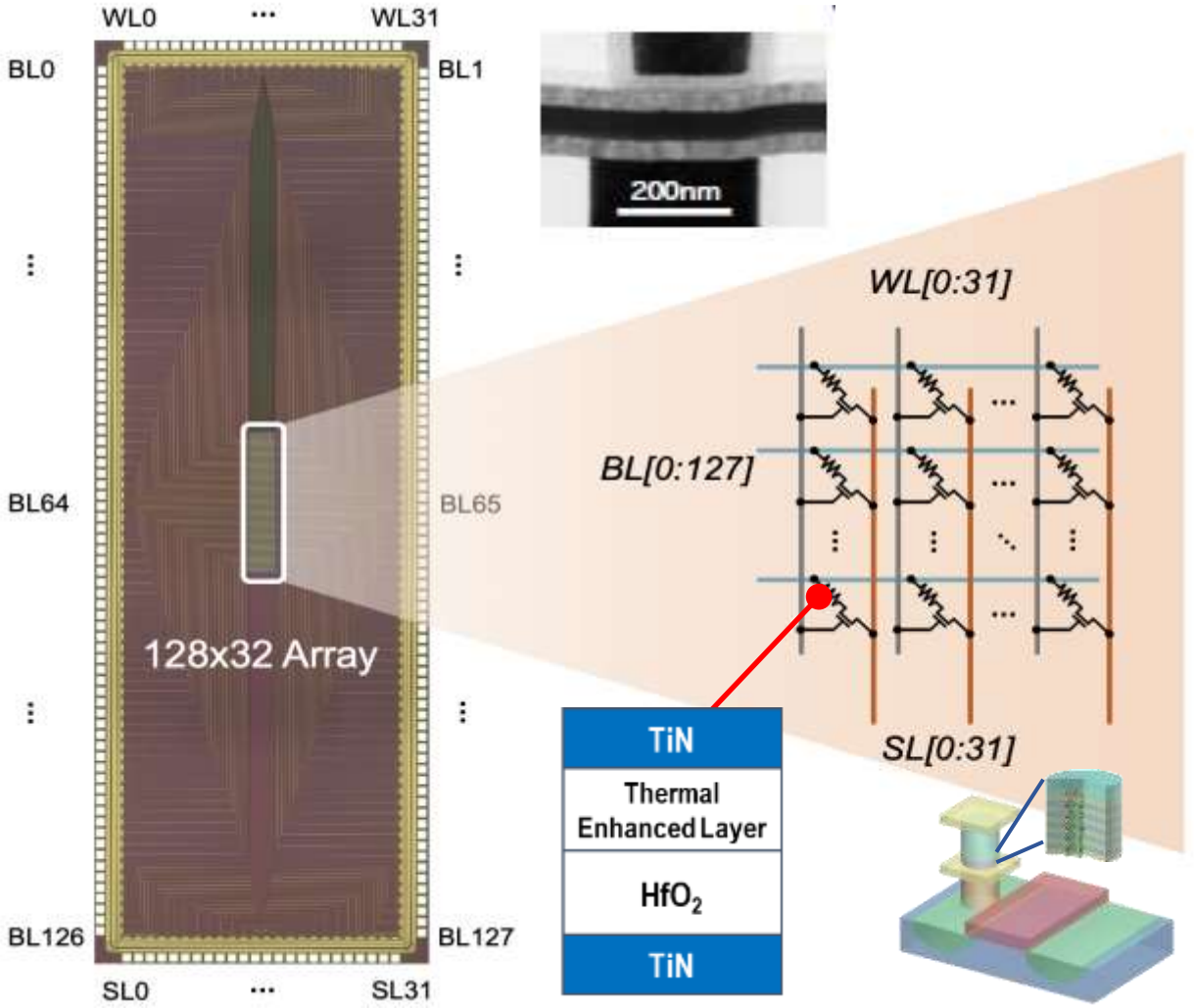




存算一体系统

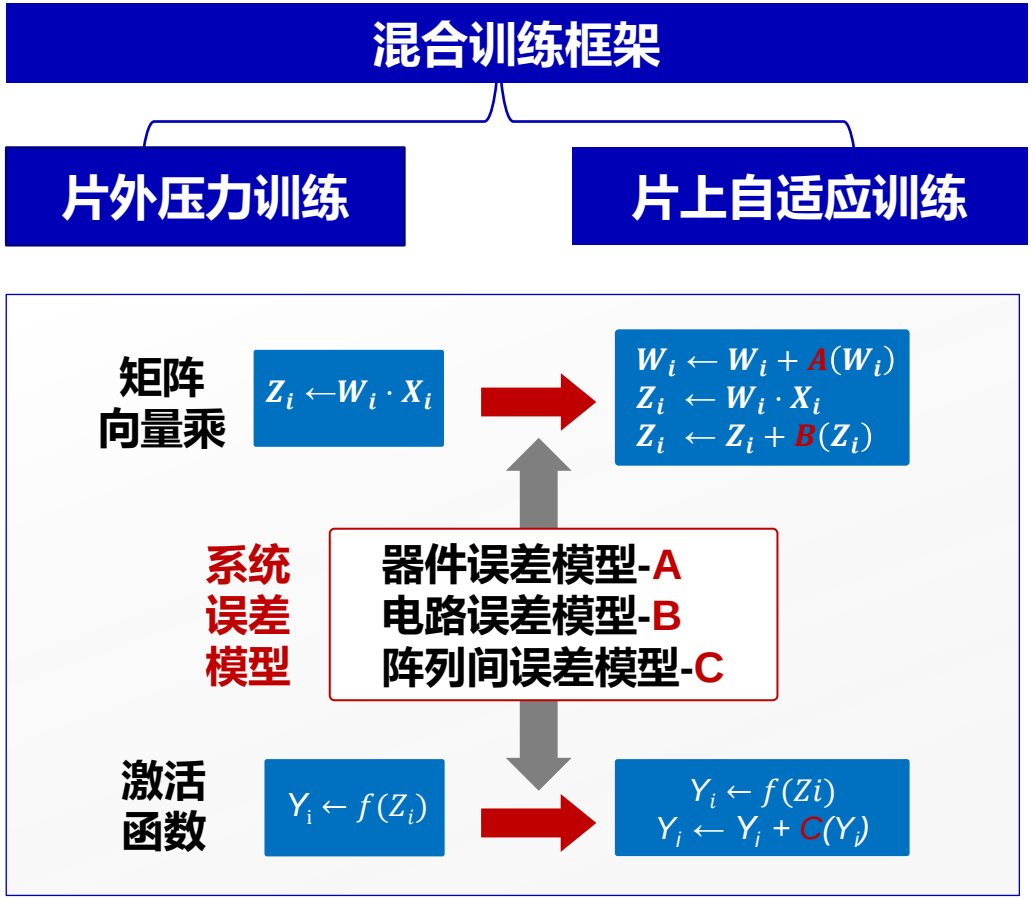
- Cell structure: 1T1R
- Array size: 128 × 32 or 1152 × 128
- Array operation: both cell by cell and fully parallel computing

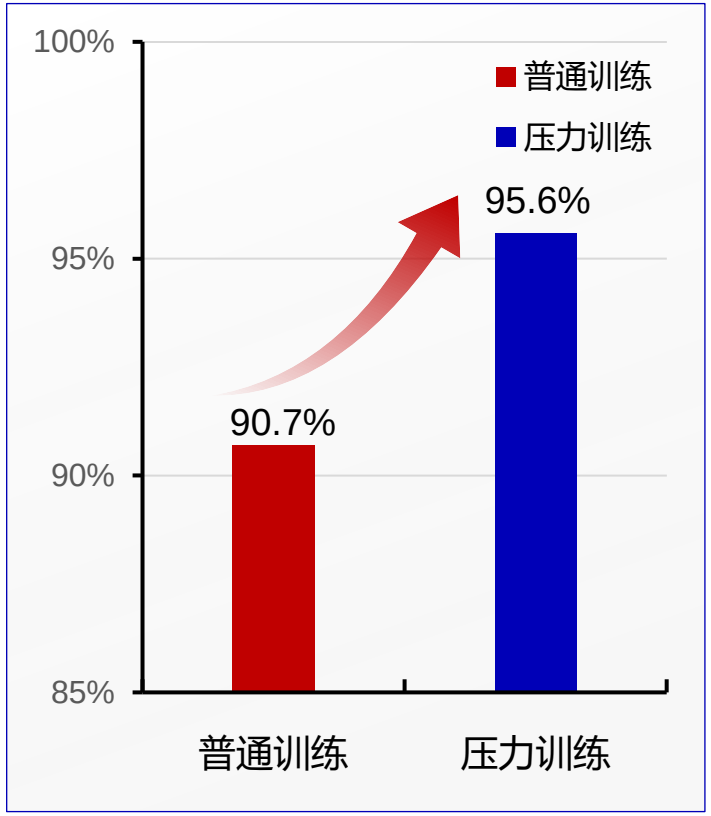




混合训练方法

提出由片外压力训练和片上自适应训练组成的混合训练框架。在片外压力训练中引入系统误差模型，构建具有误差耐受性的网络模型，提升实际硬件系统中的精度。

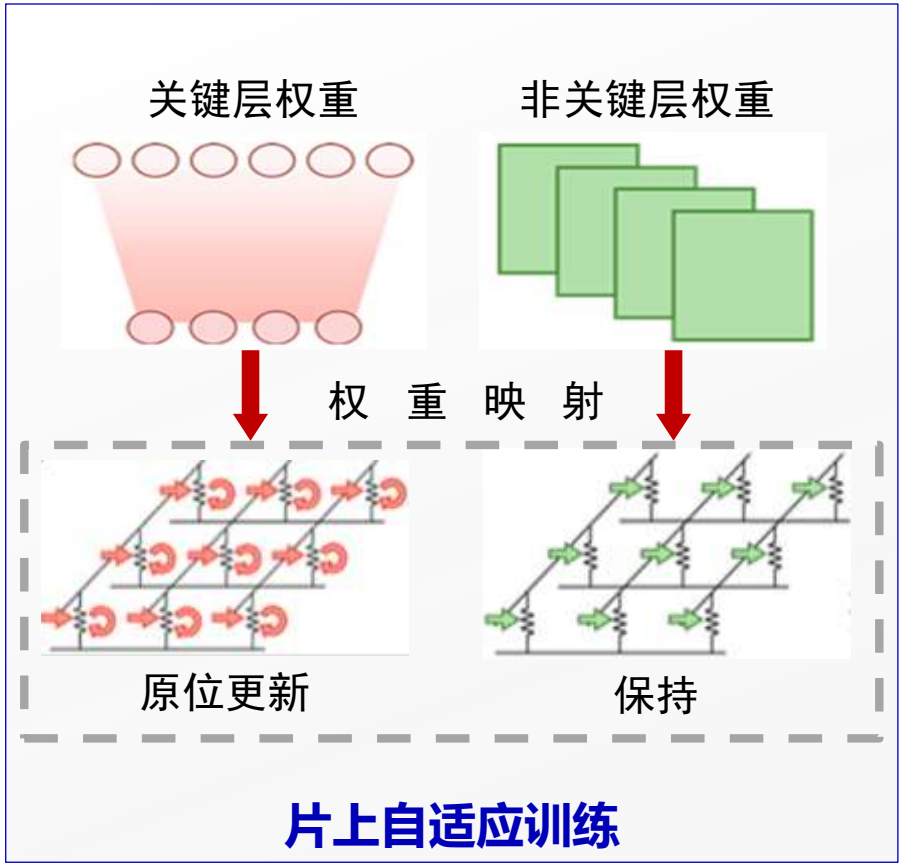


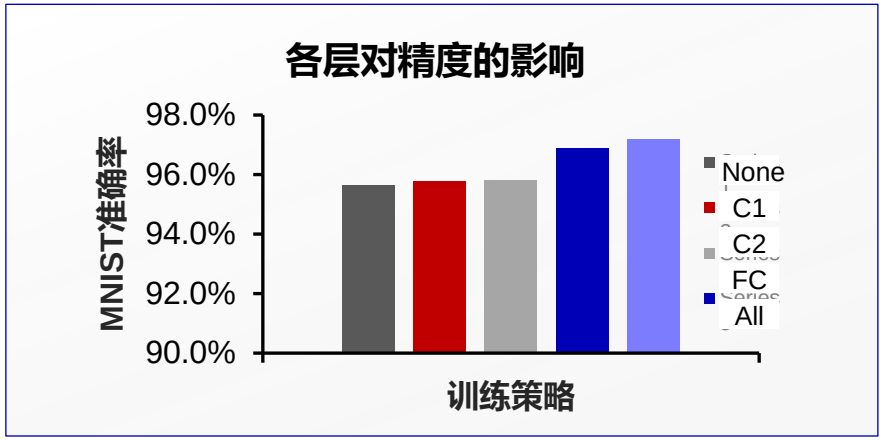


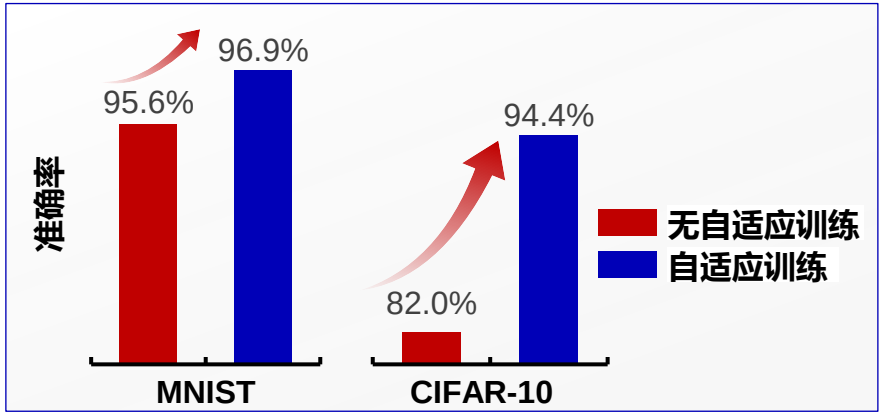
P. Yao, et al, Nature 2020

混合训练方法

在混合训练框架下，根据神经网络各层对系统精度的影响规律，在权重映射到芯片后，通过原位更新关键层权重的方法进行自适应训练，进一步提升系统精度。

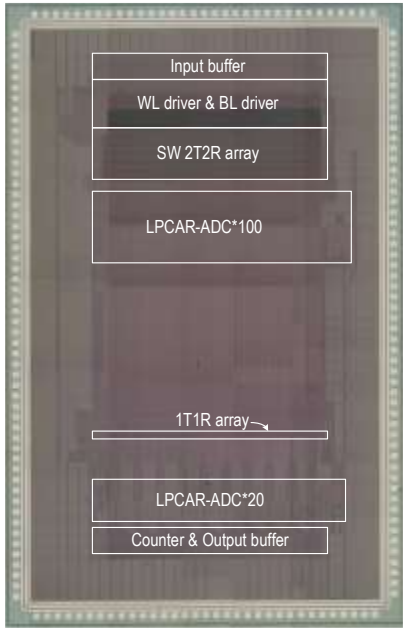


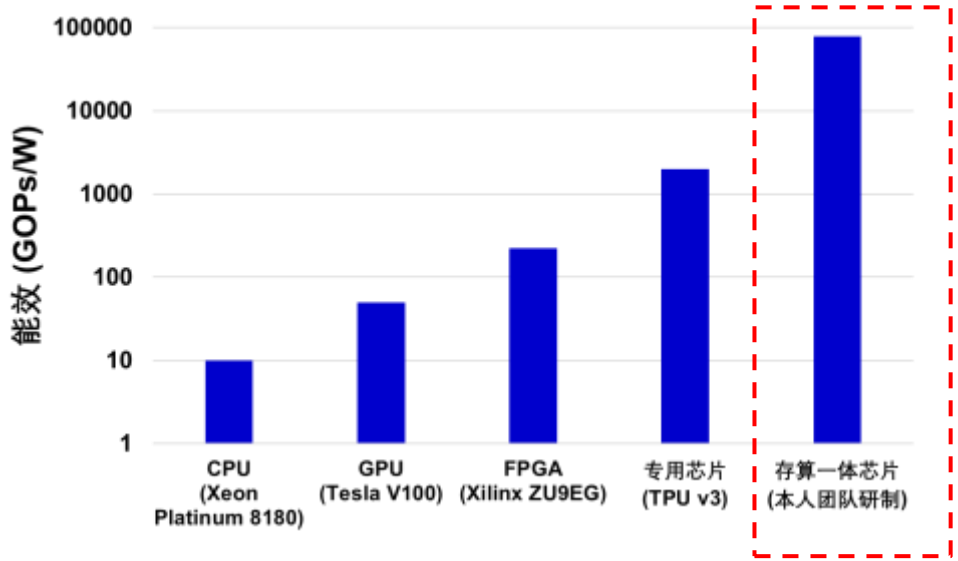




存算一体芯片

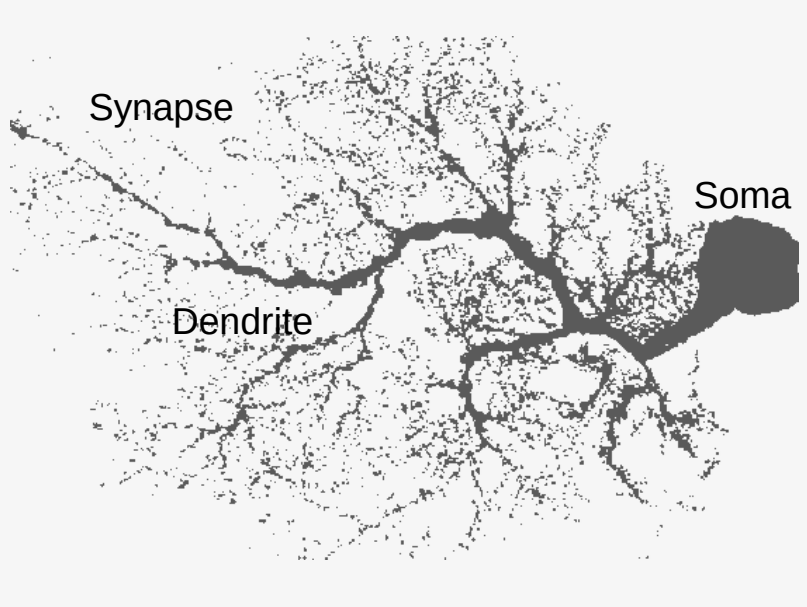
- **国际首款全系统集成的忆阻器存算一体芯片**
- **准确率与28nm CMOS 树莓派系统相当 ~95%**
- **推理速度比CMOS快 ~20倍 （3s vs. 59s)**
- **推理计算能效达到78.4TOPs/W**

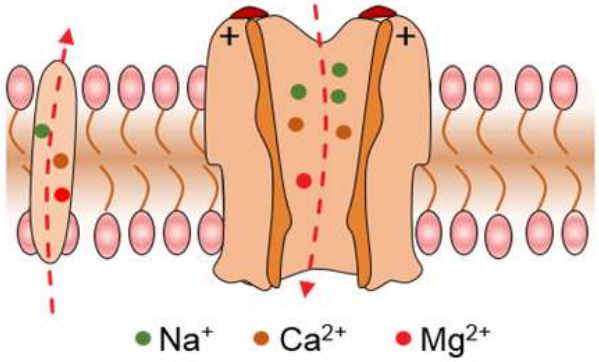


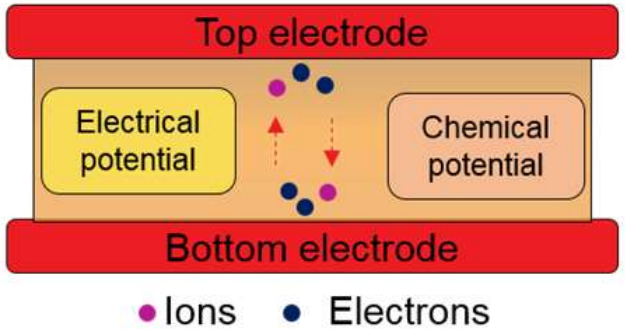


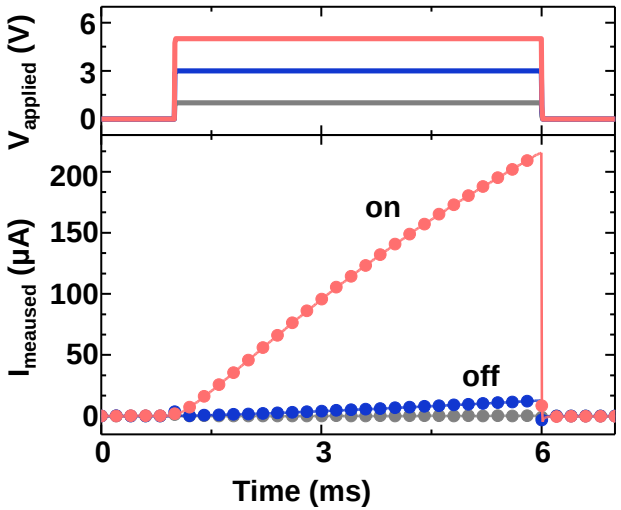
Q. Liu, et al, ISSCC 2020

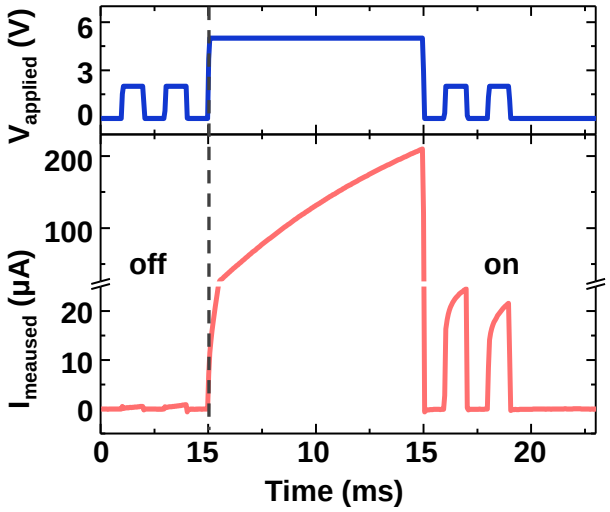
向人脑学习：树突的功能？











生物神经网络的3个基本单元:

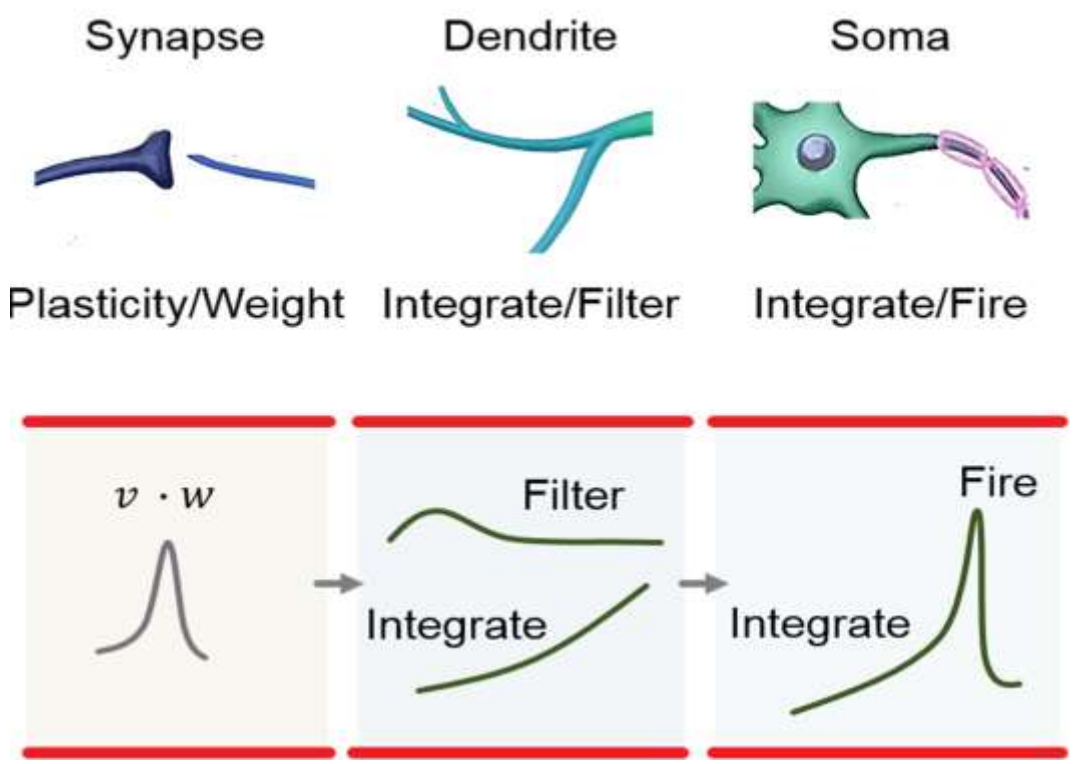
突触 (Synapse)：连接

树突 (Dendrite): 滤波与积分

胞体 (Soma): 积分与发放

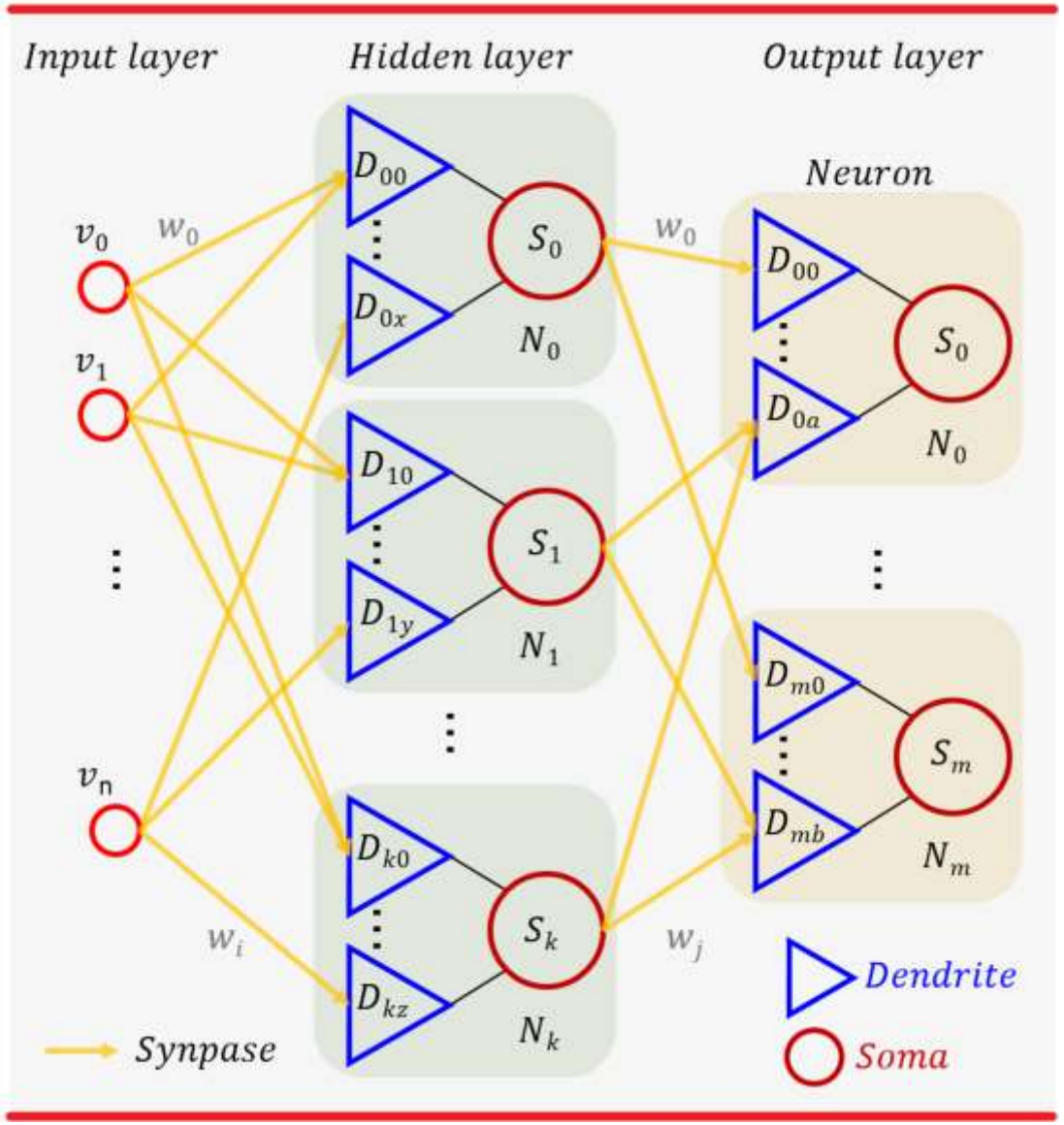
用单个忆阻器模拟电压门控NMDA通道，实现树突功能

树突器件与类脑系统探索



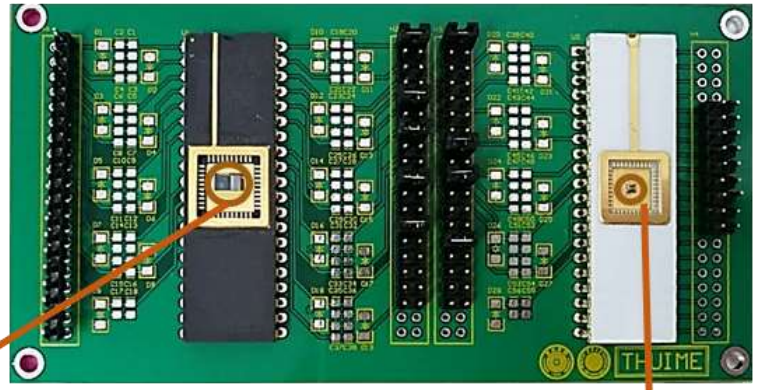
- 利用3种不同的忆阻器分别实现突触、树突、胞体的功能，完善人工神经网络

吴华强，Nature Nanotechnology 2020



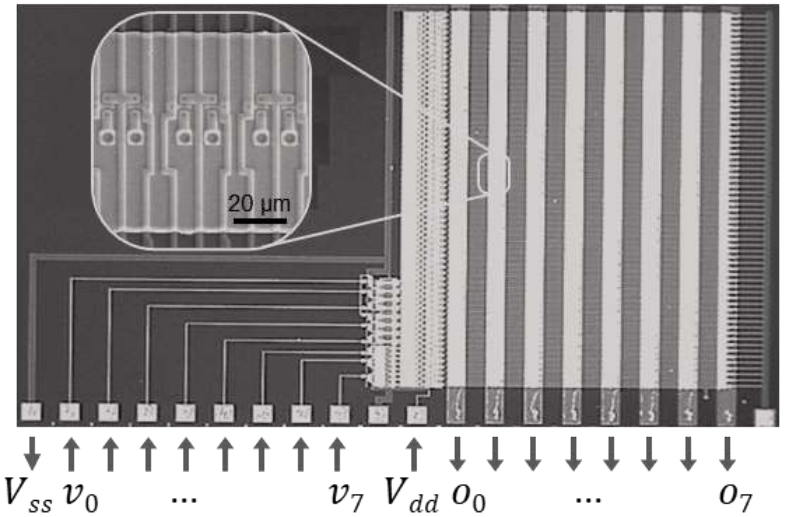
树突器件与类脑系统探索

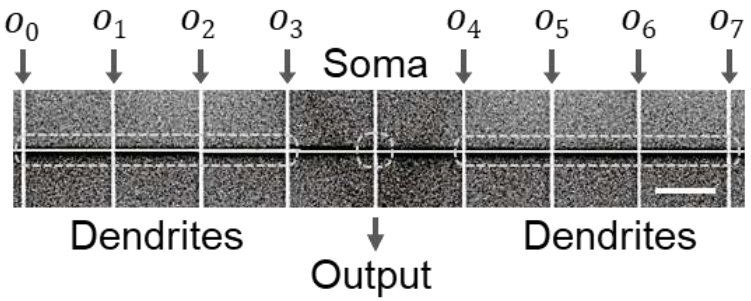
Input

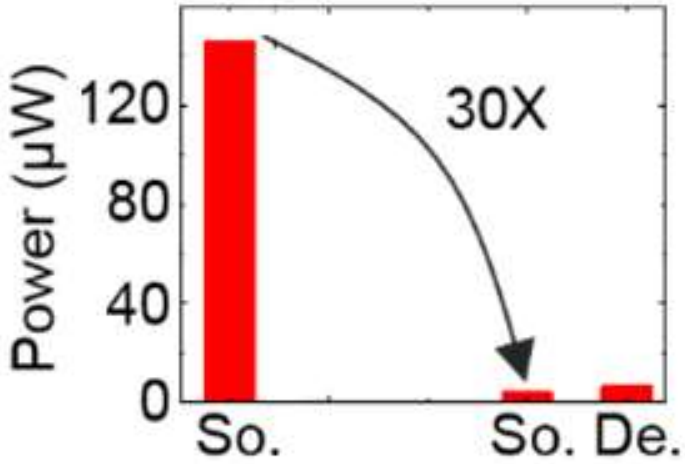


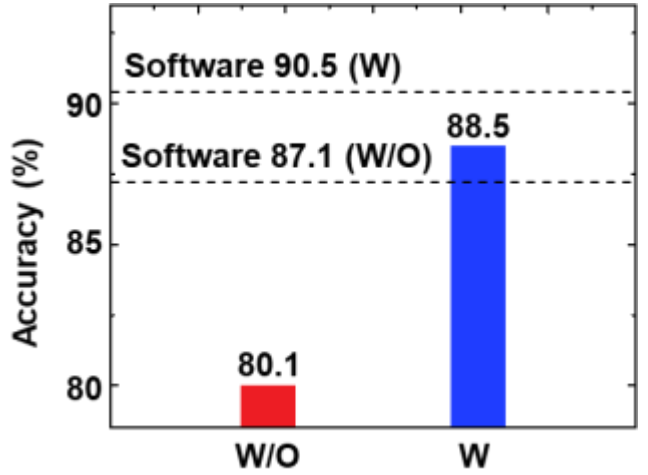
SVHN dataset

Synapses









集成突触、树突、胞体来构建完整类脑神经网络系统

- 引入树突后，动态功耗降低30倍！
- SVHN数据集识别准确率显著提高

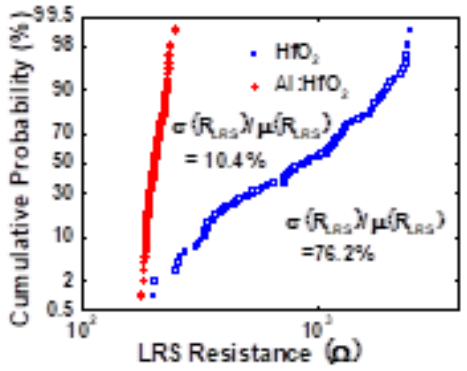
吴华强，Nature Nanotechnology 2020

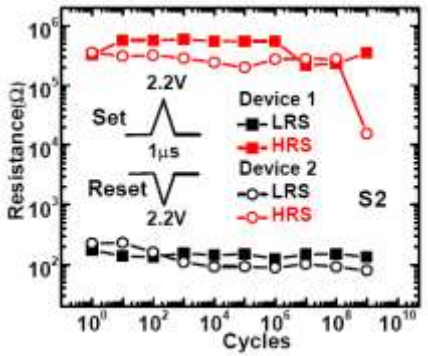
报告内容

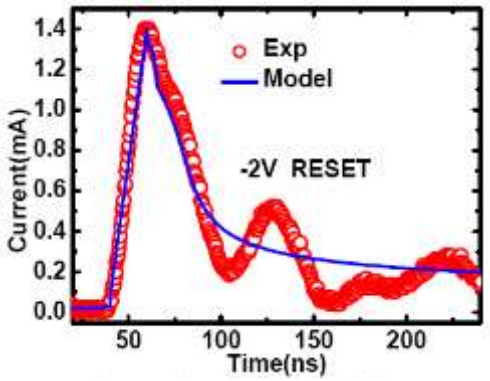
清華大學 Tsinghua University

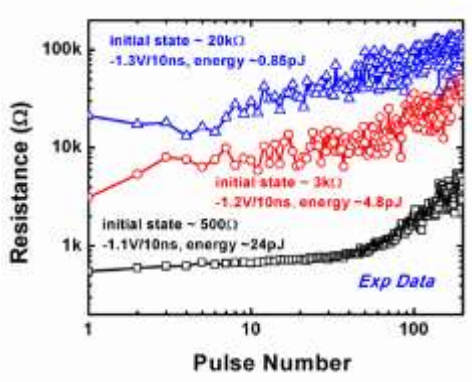
- 人工智能与芯片算力
- 存算一体技术
- 研究进展
- **器件测试的新挑战**

忆阻器电学测试的发展变化









2009 VLSI
忆阻器循环测试

2011 IEDM
快速测量二值循环次数

2012 IEDM
阻变瞬态过程的捕获

2013 AM
实现连续电阻调制

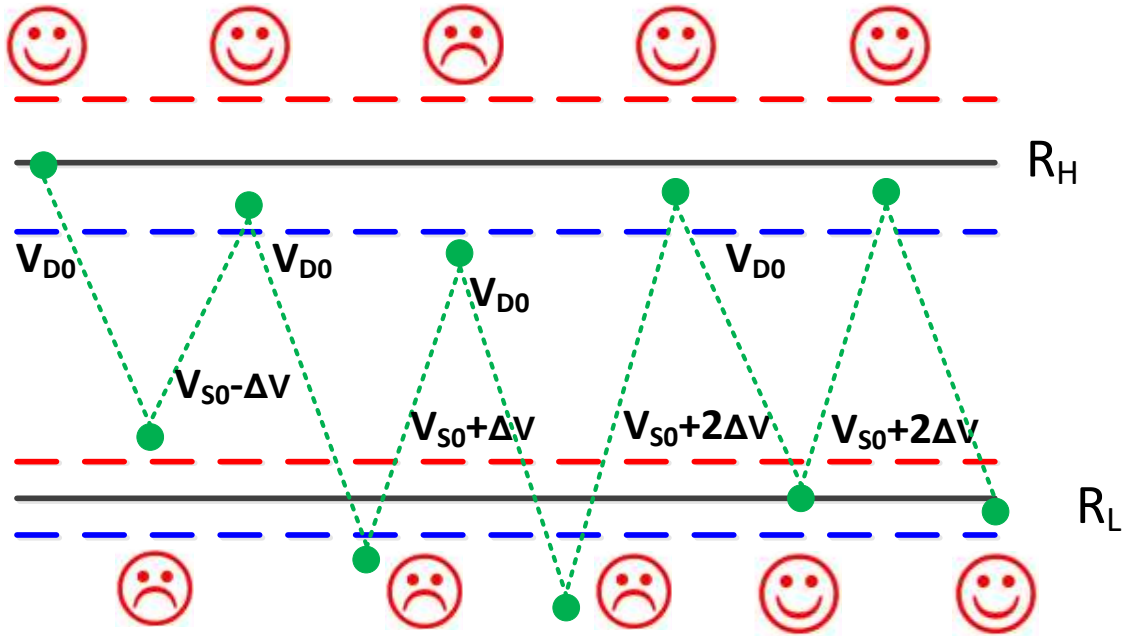
TEK设备一直在推动忆阻器研究手段的进步

模拟阻变的循环耐久性测试

存算一体技术对忆阻器特性要求更高，测试难度更大

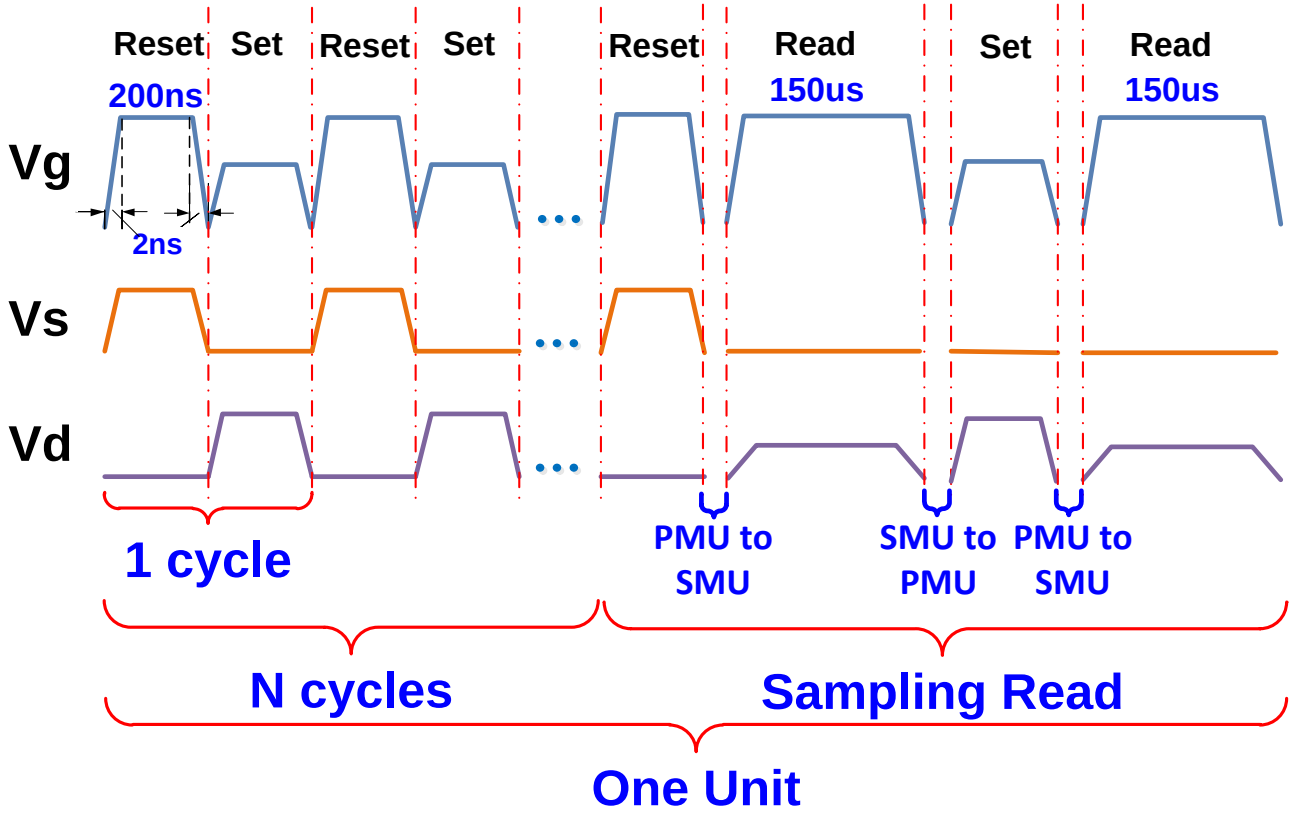
▪**关键技术点：**测试模拟型忆阻器的循环耐久性和二值存储器不同的地方在于阻变窗口不能大范围波动或偏离，需要控制循环窗口在设计的窗口内。

➜ **对阻变窗口的上下边界分别设定10%R_H或R_L的极限边界（多边界定位法）。**



M. Zhao, et al., JEDS, 2019；

模拟阻变的循环耐久性测试



Reset

Vg=4-5V;
Vs=1.6-2.5V;
Vd=0V;

Set

Vg=1.4-3.5V;
Vs=0V;
Vd=1.6-3V;

Read

Vg=4;
Vs=0V;
Vd=0.1V;

One Unit One Read

T_cycle=0.4us << t_read;
PUM to SUM to PUM~1s

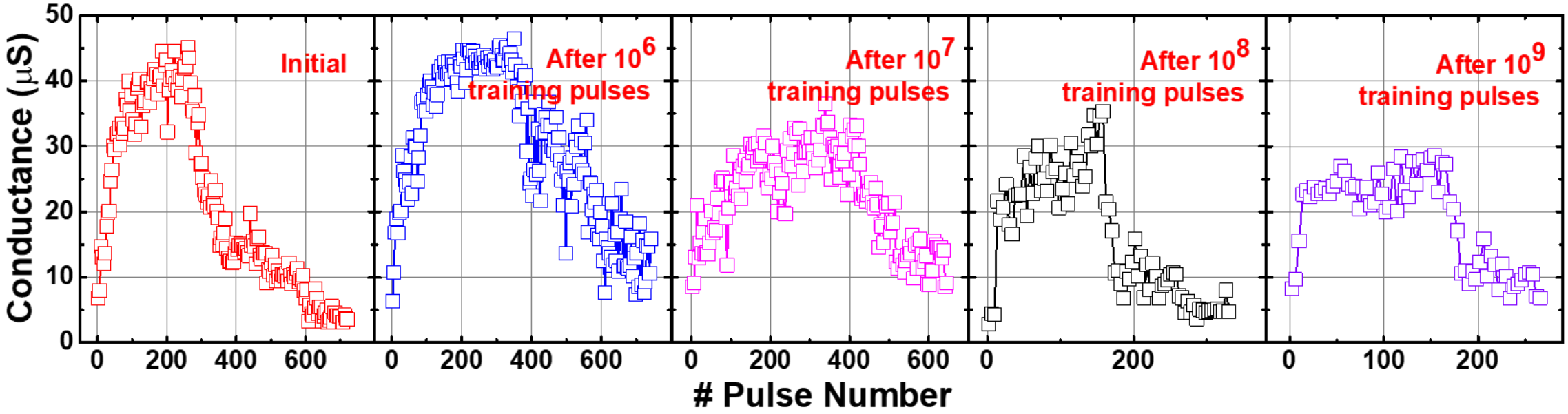
1e6 cycles~1s→**Best cycle unit**

1e7 cycles~10s
1e9 cycles~1000s
1e11 cycles~27.8h

测试效率相比传统设备提升1000倍，功能扩展性也更强

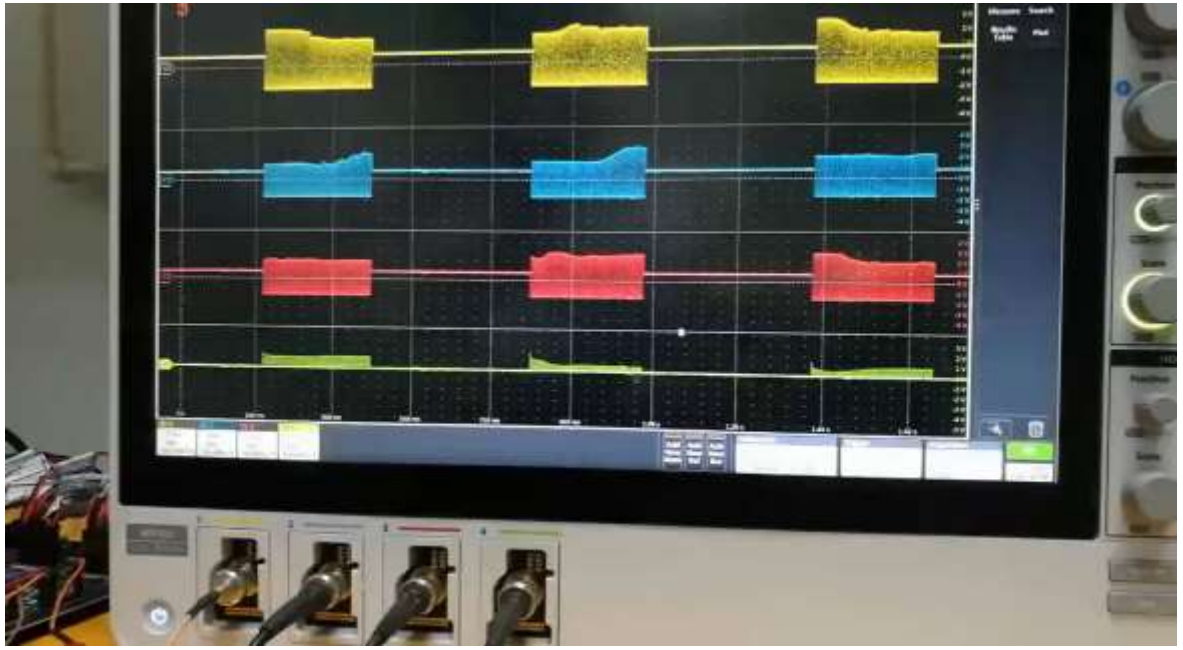
M. Zhao, et al., IEDM, 2018;

模拟阻变的循环耐久性测试

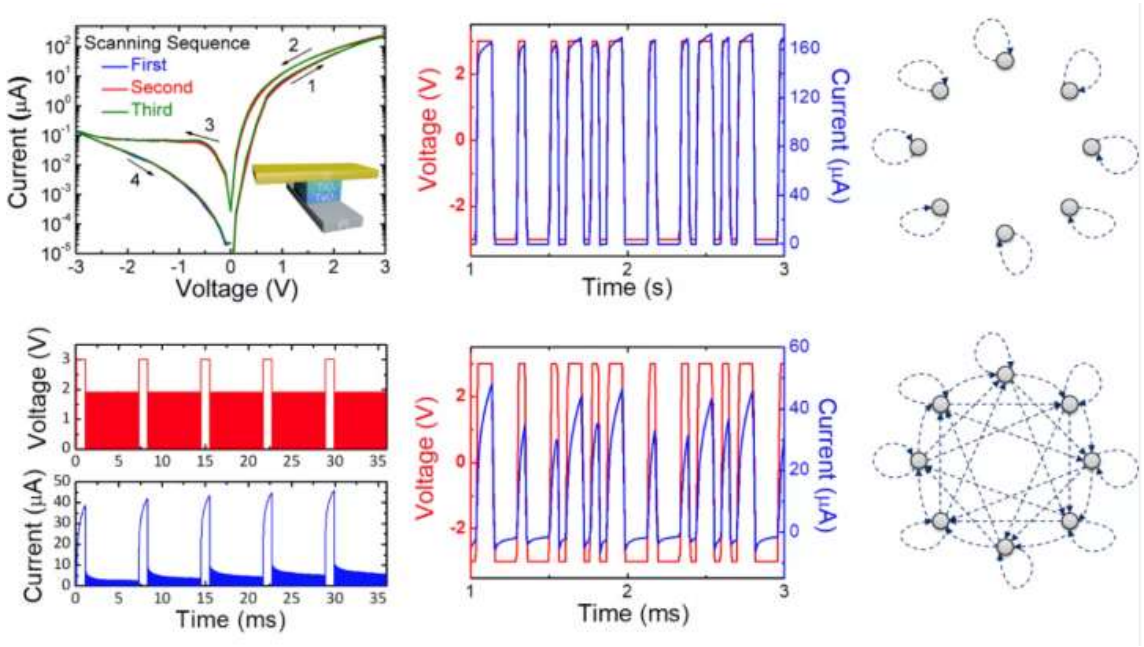


M. Zhao, et al., IEDM, 2018;

动态忆阻器测试



多通道高带宽高采样率示波器
分段内存采集模式
多通道分屏显示，输入输出对应关系一目了然



基于动态忆阻器的并行储备池计算系统
利用忆阻器的动态特性等效地实现复杂递归网络,当输入信号的单位时间步长小于器件的特征时间时不同历史时刻的状态之间开始相互耦合

Nature Nano 2020; Nature Comm. 2021

忆阻器测试的趋势与需求

- 单器件测试 → 小规模阵列测试 （3×9 ~ 12×12 规模）
- 低精度校验 → 高精度校验 （频繁的“写-读”循环，快速AC-DC切换）
- 静态测试 → 动态测试（在μA量级监测ns级的电流变化）
- 固化的测试流程 → 灵活多变的测试流程（与算法相关）
- 连探针 → 连探针卡或面包板
- ns级速度 → ps级速度

总结与展望

- 忆阻器在**数据存储**、**存算一体**、**类脑计算**等领域将发挥越来越重要的作用
- 忆阻器已具备在先进CMOS工艺平台集成的能力，**学术界**和**产业界**都广泛关注
- 存算一体技术可突破“**冯诺依曼瓶颈**”的限制，大幅提升AI计算的效率
- 存算一体等新应用对忆阻器的测试提出了更多的需求与挑战
- 测试设备的进步为忆阻器的研发做出了**重要的贡献**
- 未来仍期待在**多通道**、**快切换**、**高时间分辨率**等方面的进步

Tektronix®

清華大學 Tsinghua University

谢 谢 ！

Email: gaob1@tsinghua.edu.cn