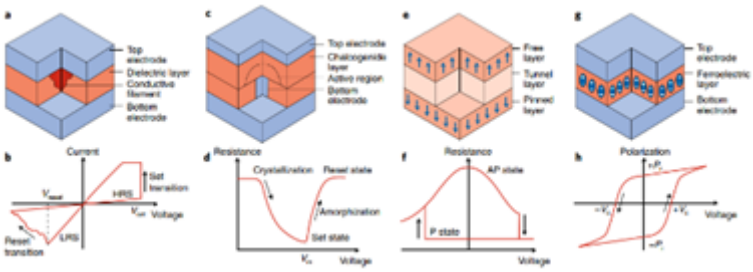
In-memory computing with resistive switching devices

> Ielmini D, Wong H S P. In-memory computing with resistive switching devices[J]. Nature electronics, 2018, 1(6): 333-343.

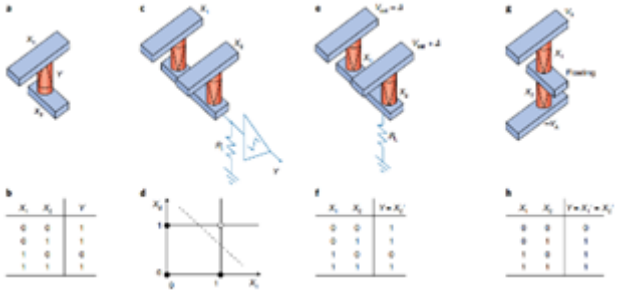
- **写作目的：**

  本文主要讨论电阻开关器件（resistive switching device）在内存计算方面的应用，主要包括二进制计算、多位计算、随机数生成器（RNG）、模拟计算。

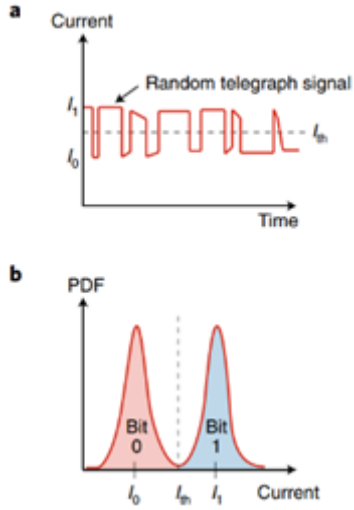
- **内容记录：**
  - 处理“内存墙”的若干方法：
    1. 增强处理器芯片的并行性。
    2. 增加带宽。
    3. 将新型非易失性存储器作为内存。
    4. CIM技术。
  - 可用于CIM技术的若干电阻开关器件（RRAM、PCM、MRAM、FeRAM）：



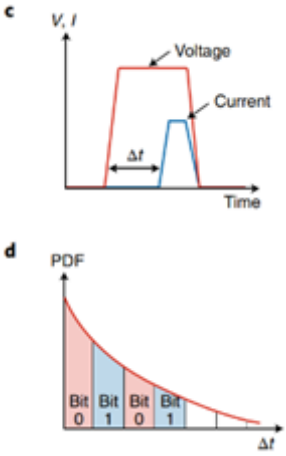
  - 几种二进制数字计算的器件实现：分为**V-R逻辑门**（无法实现有效输出）、**V-V逻辑门**（需外加比较器电路）、**R-R逻辑门**（可真正用于CIM技术）：



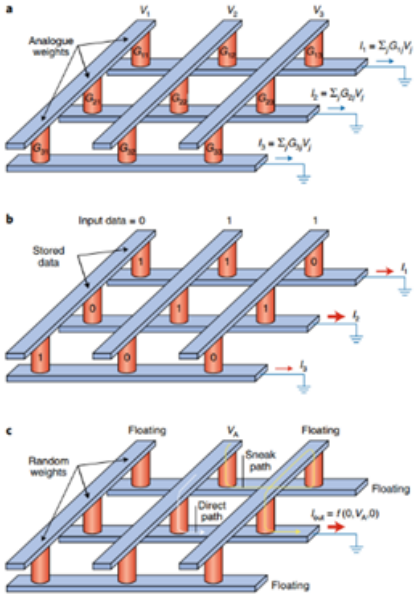
  - 利用**随机电报噪声（RTN）**生成二进制随机数（处于LHS与RHS均可）：



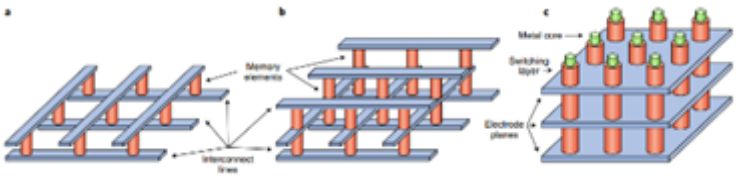
- 利用切换延迟（电阻开关在实现状态切换时有延迟）改良RNG的随机性，具体原理为将延迟时间进行等距划分，再根据完成切换的时刻落在奇/偶数个时间窗口内来给出随机数：



- 使用cross-point阵列实现模拟矩阵向量乘法（MVM）、可寻址存储器、物理不可克隆功能（PUF）：

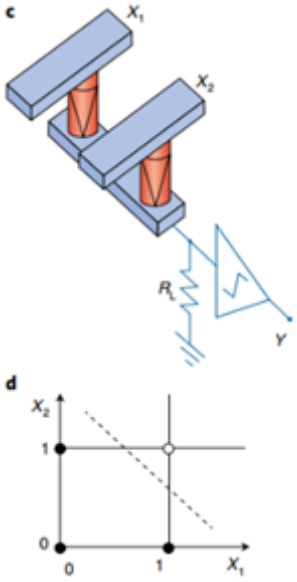


- 为实现尺寸缩放而设计的3D架构：



- **批注：**
  - **内存墙（memory wall）**：指基于冯诺依曼结构的计算机中，数据在运算器与存储器间的传输速率会限制计算机的性能。
  - FeRAM在实现电阻开关时并不改变MIM结构对应的电阻，而是改变MIM结构Metal电极上积累的电荷Q。
  - 对V-V逻辑门的解释（以下图为例）：2个电阻开关分别施加电压V_1、V_2，则据KCL、KVL有：

$$V_{\mathrm{com}}=\frac{V_1G_1+V_2G_2}{R_L^{-1}+G_1+G_2}$$



  - 可通过调制比较器的阈值电压V_T、电导G_1、G_2与电阻R_L来控制二进制0与1的“分界线”（图中虚线），进而实现不同的逻辑门。
  - 随机数生成器（RNG）严格来说不是CIM技术，但它在密码学和数据安全方面具有重要作用。

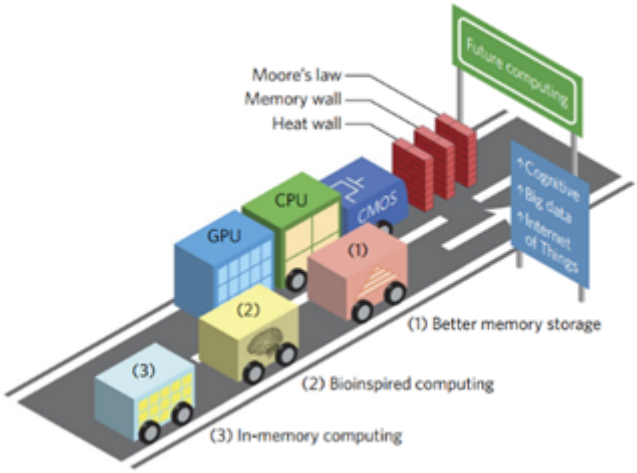
The future of electronics based on memristive systems

> Zidan M A, Strachan J P, Lu W D. The future of electronics based on memristive systems[J]. Nature electronics, 2018, 1(1): 22-29.

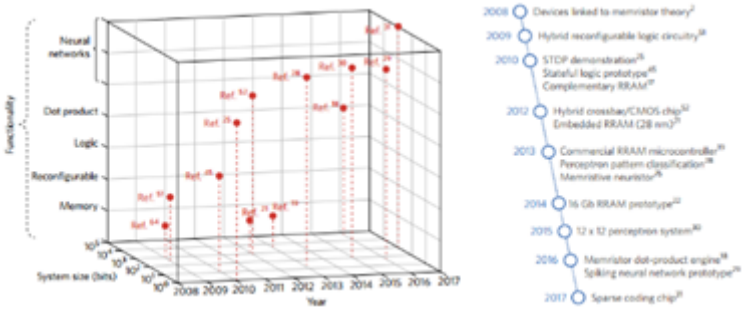
- **写作目的：**

  本文主要介绍忆阻器（memristor）在各领域的应用现状集未来，主要包括：片上存储器、通用内存计算与仿生计算。

- **内容记录：**
  - 忆阻器的本质特征是其电阻会随电流输入变化，且在停止输入时电阻具有“记忆性”，即保持不变。
  - 传统冯诺依曼结构计算机面临的挑战：摩尔定律、高能耗限制（the heat wall）、内存数据传输速度限制（the memory wall）。
  - 面向未来的计算解决方案：



- 忆阻器作为存储系统所面临的挑战：提高器件速度、开/关比、耐久性和数据保留时间、降低工作电压和电流、优化器件均匀性、减少漏电流和导线电阻。
- 忆阻器构建的存储系统在实际应用中需要扩大规模（scale up），主要包含以下几个方面：
  1. 增加忆阻器构建的存储网络的大小（即内部所含器件单元的数目）。
  2. 在单个存储系统内实现多重任务处理（multitasking）。
  3. 采用3D结构设计。



- **批注：**
  - 忆阻器在物理上用于表征磁通ϕ与电量Q之间的关系，即：

$$M = \frac{\mathrm{d}\phi}{\mathrm{d}Q} = \frac{\mathrm{d}\phi}{\mathrm{d}t} \cdot \frac{\mathrm{d}t}{\mathrm{d}Q} = \frac{V}{I}$$

  - 于是磁通ϕ随电量Q的变化率M仍具有电阻的量纲，但M是由过去流经器件的电荷总量Q决定的，因此具有记忆性。
  - 一般来说，任何忆阻器构建的存储系统仍需要一些 CMOS 电路来提供必要的接口和控制操作。

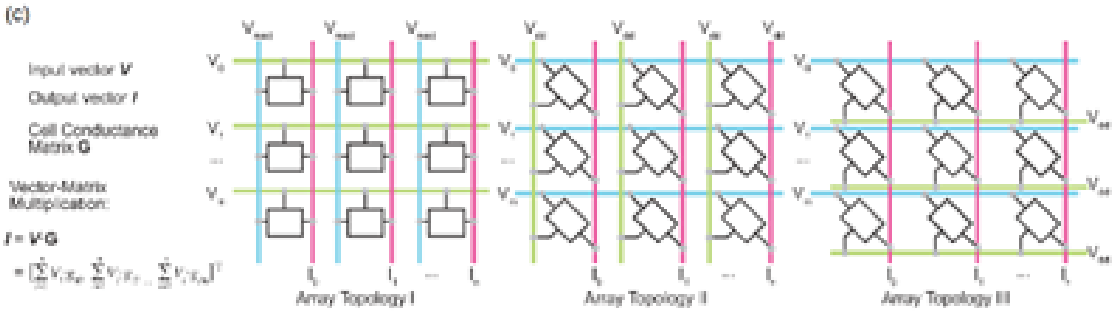
Resistive Memory-Based In-Memory Computing: From Device and Large-Scale Integration System Perspectives

> Yan B, Li B, Qiao X, et al. Resistive Memory-Based In-Memory Computing: From Device and Large-Scale Integration System Perspectives[J]. Advanced Intelligent Systems, 2019, 1(7): 1900068.

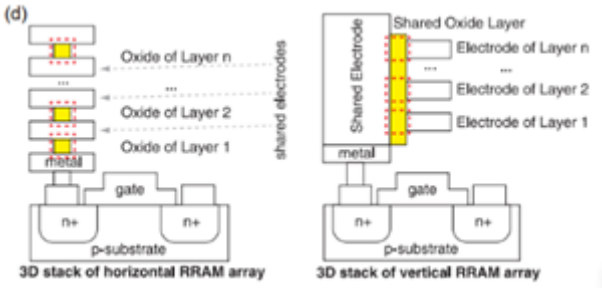
- **写作目的：**

本文主要从器件/架构层面讨论基于RRAM的CIM技术在神经网络、仿生计算等领域的应用，主要包括其器件选择、架构设计与性能评估。

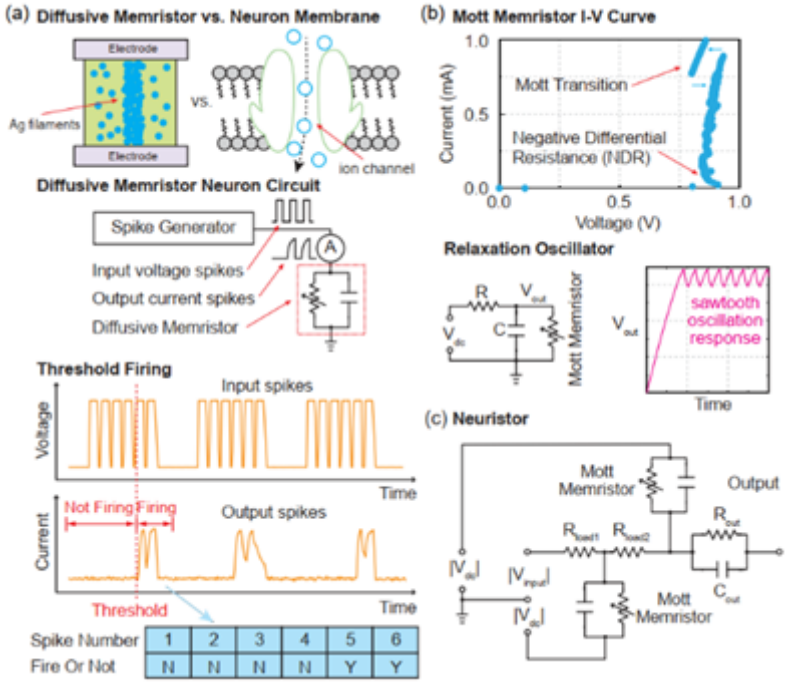
- **内容记录：**
  - 几种可用于CIM的存储设备各自的优势：
    1. **flash**：较强的稳定性与可靠性。
    2. **MRAM**：较快的写入速度、可实现随机编程。
    3. **racetrack存储器**：高密度、顺序读写。
    4. **PCM**：权重更新具有**线性**。
    5. **RRAM**：功能多样，包括高电阻、支持3D集成、可实现随机编程、可实现MLC。
  - 3种用于实现**矩阵向量乘法（VMM）**的RRAM拓扑架构：



  - 两种RRAM的3D堆叠架构：



  - 3D堆叠结构的一个突出问题是：每个器件单元工作时产生的焦耳热可能会影响相邻器件单元的工作（因为3D堆叠架构中的器件单元排列非常密集）。
  - 使用Mott忆阻器产生STDP中的尖峰信号：

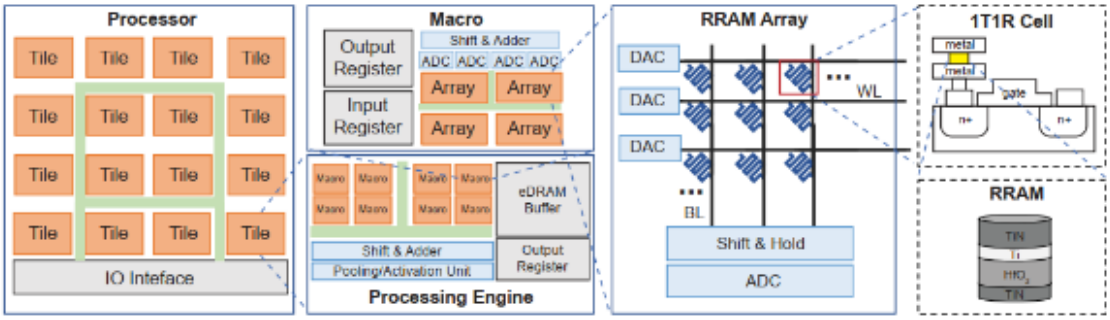


- 基于RRAM的**内存计算宏（in-memory computing macro）**：由于在CIM技术中，RRAM处理/计算的电压值为模拟值，因此需要对这种宏进行转换才能与外部数字系统进行交互，具体实现为：使用DAC将数字输入信号转化为模拟输入信号，使用ADC将模拟输出信号转化为模拟输出信号，以上使用ADC/DAC处理内存计算宏，而也可使用其他方法来消去ADC/DAC：

  1. 消去输入端的DAC：将二进制化的电压值分周期输入RRAM，并将结果进行加权累加，可消去输入端的DAC，但随之会增加运行周期数。
  2. 消去输出端的ADC：首先使用电平传感放大器（二进制）将输出转为数字信号，再将1bit的数字信号进行累加变为多位

  这种方案一般被称为基于**检测放大器（sense amplifier）**的处理方案。

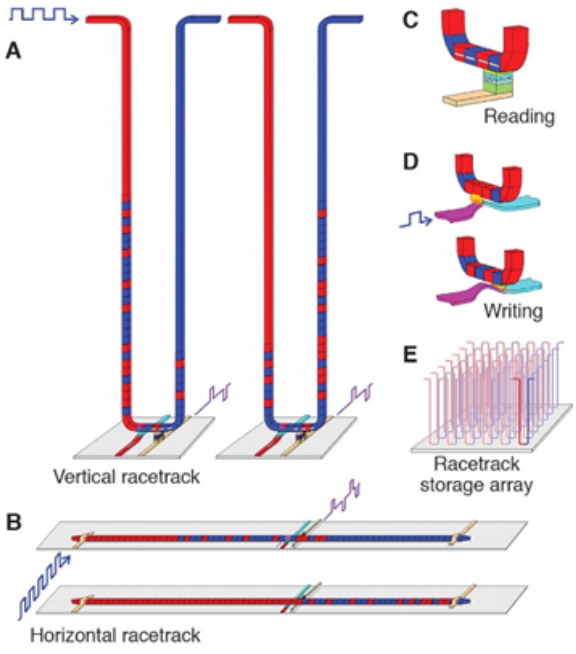
- 基于RRAM的CIM技术的层次结构（此处内存计算宏使用ADC/DAC处理）：



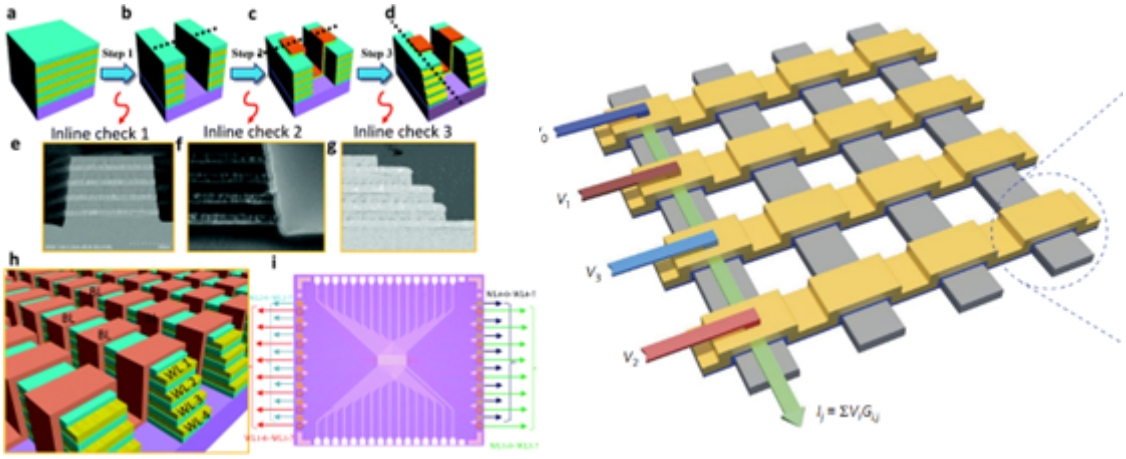
- RRAM的一些非理想的性能指标：

Nonideal behavior	Analog RRAM	Binary RRAM
Endurance[22,80]	500 k cycles 95.5% maintains resistance	$>10^6$ cycles
Variability[a)][126]	≈0.03	≈0.04 @ LRS, ≈0.4 @ HRS
Yield[61,80]	89.9%	>99%
Bit error rate before ECC[47]	N/A	$<10^{-5}$
Thermal-activated fluctuation variability[a)][94]	≈0.03	≈0.03
Read disturbance	Refer to Yan et al.[127]	Refer to Ho et al.[128]

- **批注：**
  - IBM设计的racetrack存储器的原理图：



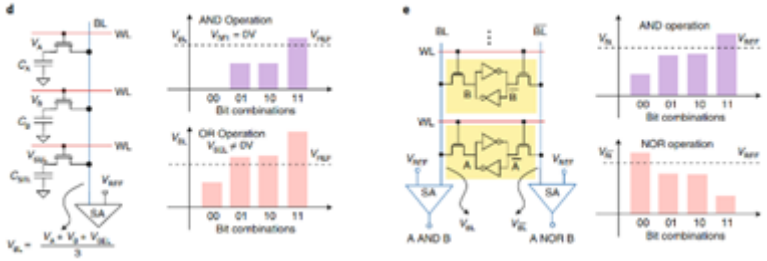
- 矩阵向量乘法（VMM）是内积运算，而一般仅使用RRAM无法实现如反向传播算法中涉及的偏导数或外积运算，因此有时需要额外的电路辅助。
- 一种RRAM的3D堆叠架构，顶层电极的个数决定输入矢量维数；堆叠层数决定输出矢量维数，可结合以下1层堆叠情形理解：



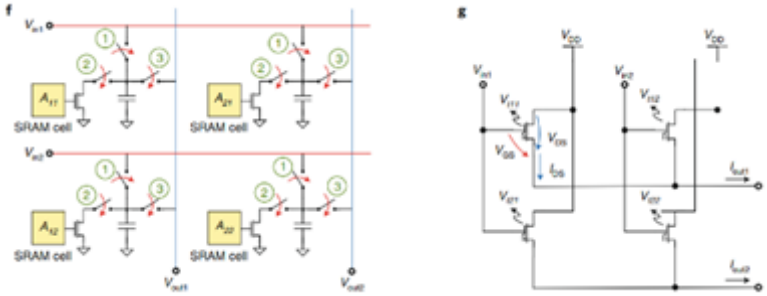
Memory devices and applications for in-memory computing

> Sebastian A, Le Gallo M, Khaddam-Aljameh R, et al. Memory devices and applications for in-memory computing[J]. Nature nanotechnology, 2020, 15(7): 529-544.

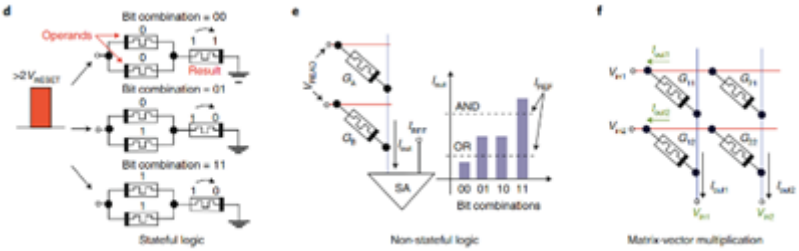
- **写作目的：**本文分别介绍了基于电荷（SRAM、DRAM、flash）以及基于电阻（RRAM、PCM、STT-MRAM）的存储设备的CIM技术，讨论其在科学计算、信号处理、优化、机器学习、深度学习和随机计算领域的应用。
- **内容记录：**
  - SRAM/DRAM实现通用逻辑门：



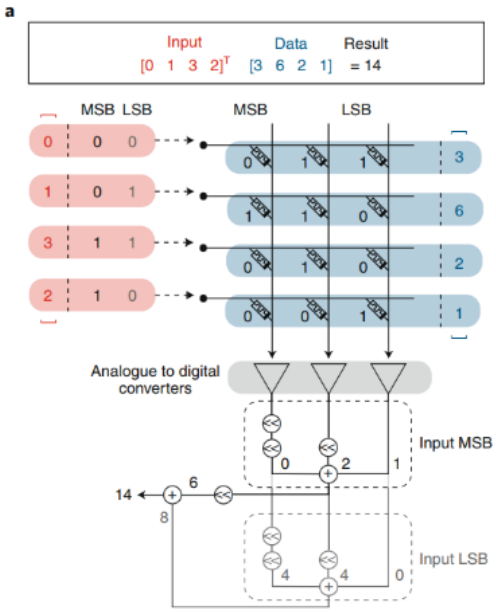
  - 使用SRAM/flash进行矩阵向量乘法（VMM）：



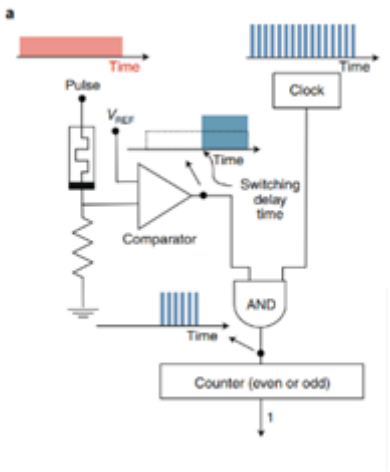
  - 使用RRAM实现VMM与通用逻辑门：



- 使用cross-point RRAM实现科学计算中的“位切片（bit slicing）”：



- 随机数生成器的具体构造：



结合文献：

> Ielmini D, Wong H S P. In-memory computing with resistive switching devices[J]. Nature electronics, 2018, 1(6): 333-343.

- **批注：**
  - flash作为非易失性存储器的缺点在于其：写入电压高，功耗大；存在着显著的延迟。
  - **位切片（bit slicing）**：假设需计算向量内积：

$$[0\ 1\ 3\ 2]^T[3\ 6\ 2\ 1]$$

则将前一向量化为二进制：

$$[00\ 01\ 11\ 10]^T$$

分2个计算周期输入，并将最终结果按位权展开式移位相加，同时将后一向量化为二进制：$[011\ 110\ 010\ 001]$存储在一个3×4的cross-point阵列中，并将各列计算结果按位权展开式移位相加。

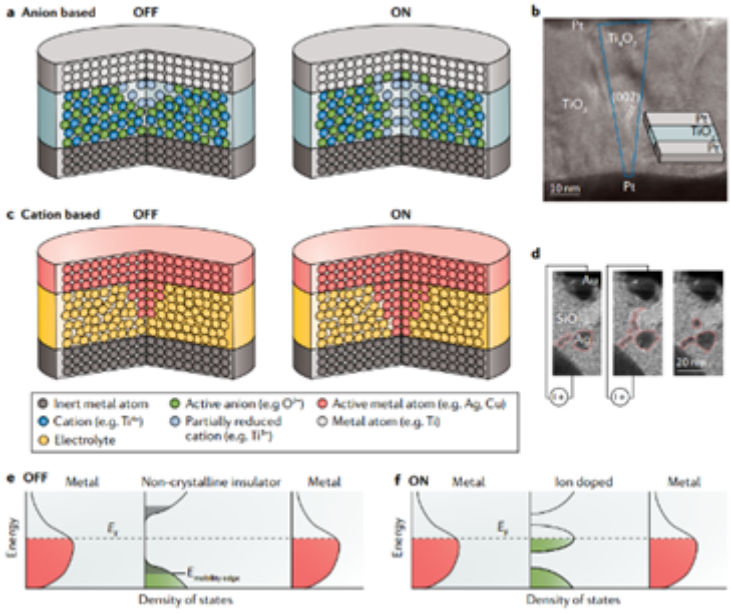
Resistive switching materials for information processing

> Wang Z, Wu H, Burr G W, et al. Resistive switching materials for information processing[J]. Nature Reviews Materials, 2020, 5(3): 173-195.

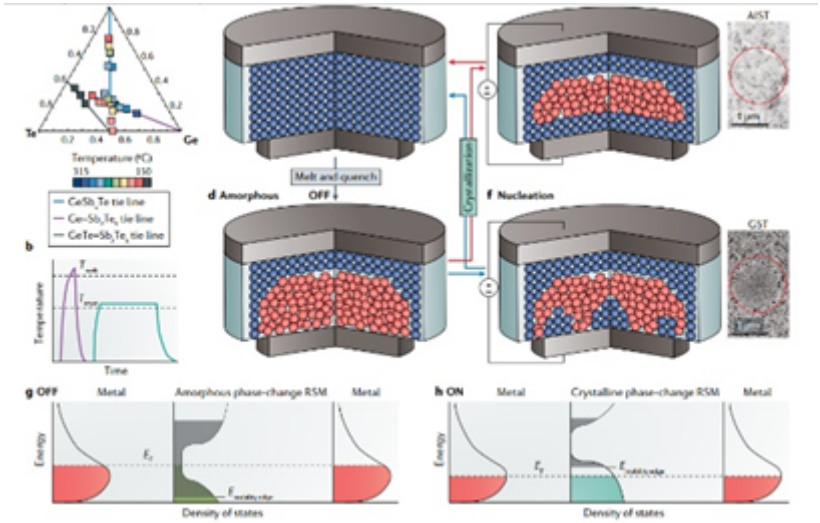
- **写作目的：**

  本文简介电阻开关材料（RSM）。对RSM实现电阻转换的物理原理、RSM的应用领域、RSM的性能指标进行了讨论。

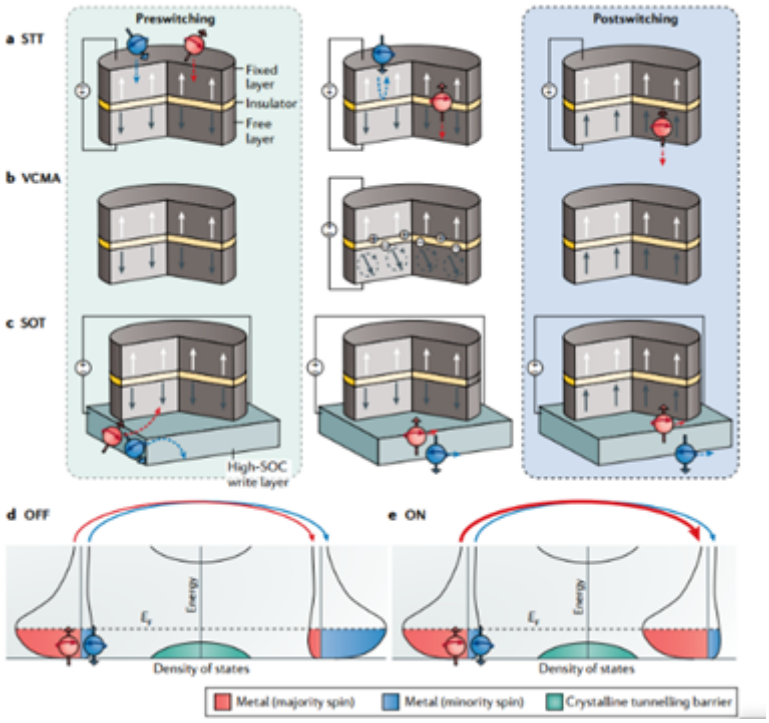
- **内容记录：**
  - RSM实现电阻切换的4种物理机制：氧化还原RSM、相变RSM、磁隧道RSM、铁电体RSM：
    1. **氧化还原RSM**：开关机制（阳离子型/阴离子型）：



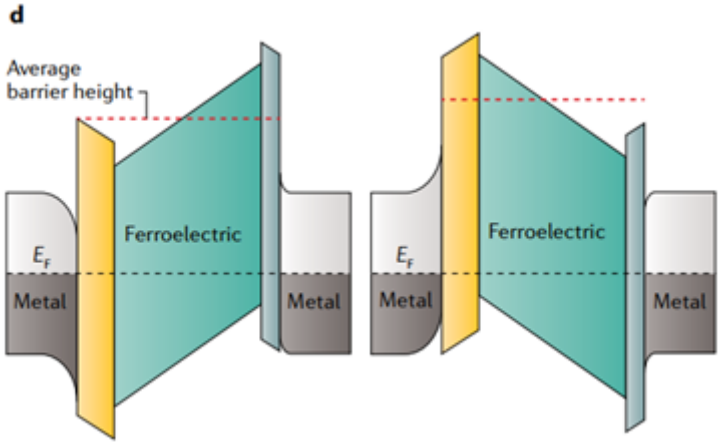
    2. **相变RSM**：长程无序的非晶相，低电导，OFF态；长程有序的晶相，高电导，ON态，开关机制：T_{melt}下晶相→非晶相；T_{cryst}下非晶相→晶相：



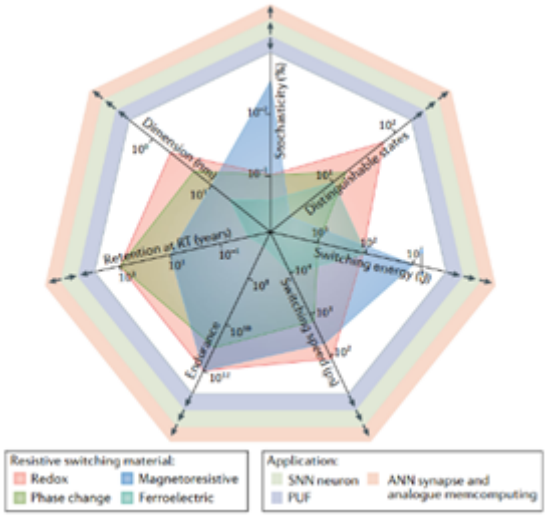
3. **磁隧道RSM**：按切换机制分类：STT-RSM（最成熟）、VCMA-RSM、SOT-RSM，开关机制：ON态两电极磁化方向相同；OFF态两电极磁化方向相反：



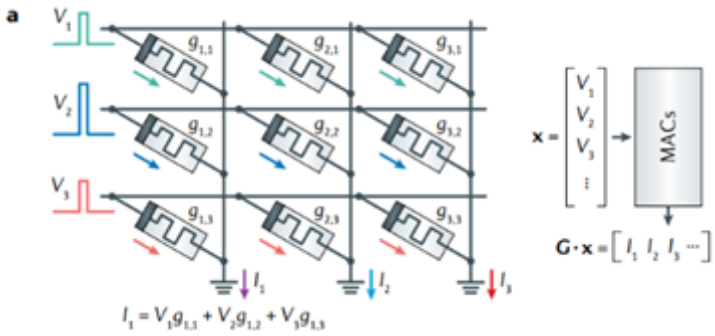
4. **铁电体RSM**：开关机制：



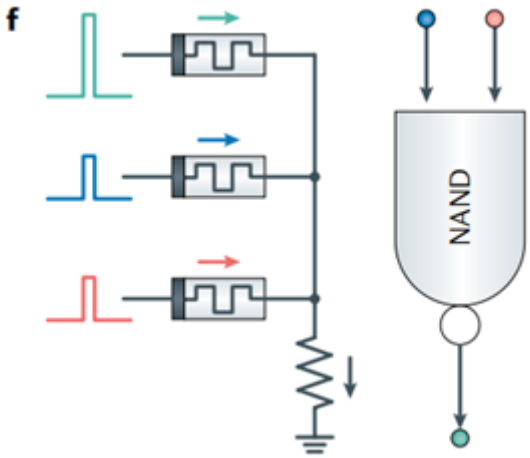
- 4类RSM的性能指标：



- RSM的应用领域：类脑计算、内存运算、硬件安全。
- RSM构建的权重矩阵：

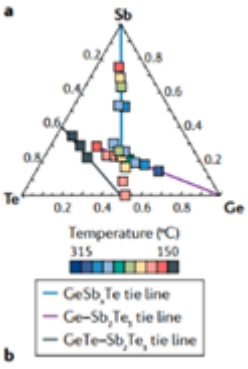


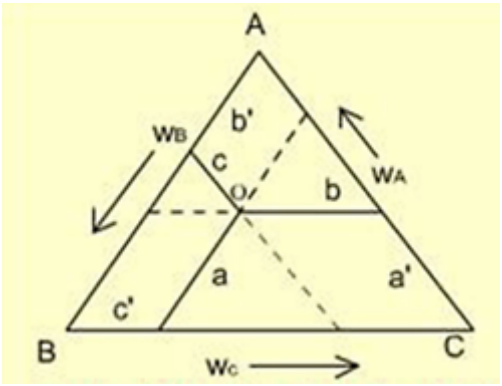
- RSM构建的通用逻辑门（NAND）：



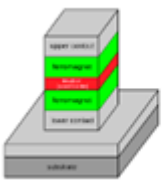
- RSM未来发展方向：材料工程、单元性能优化、算法优化、系统架构优化。

- **批注：**
  - **Dennard Scaling定律**：早期半导体工艺变化的规律，即将晶体管尺寸和电源电压一起变化，单位面积晶体管的总电容上升，但是电源电压在相应变小的单位面积能量消耗基本保持不变。
  - **三元相图（ternary phase diagram）**，用于表示一个三元体系各成分的含量：

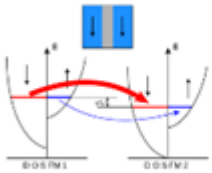


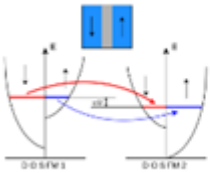


- **磁隧道结（MTJ）**：不同磁化方向会引起两种自旋方向的电子的态密度的改变：



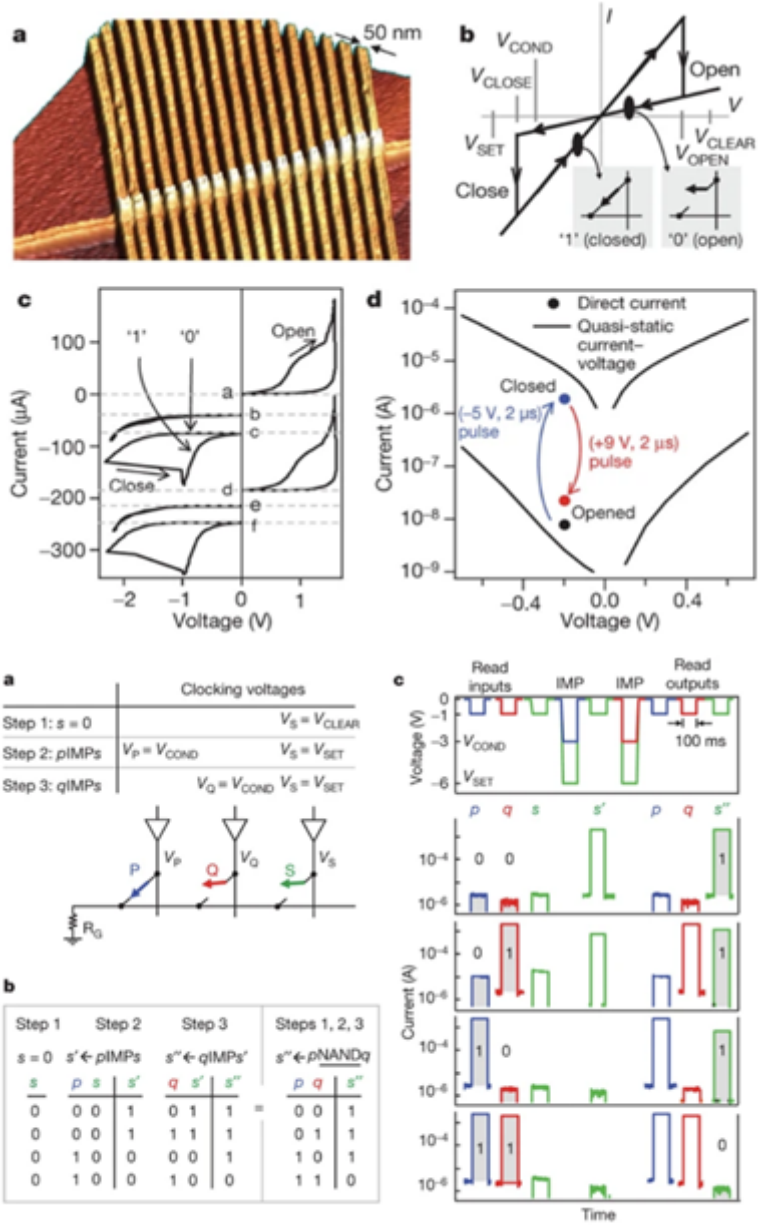
- **Magnetic tunnel junction** (MTJ): a component consisting of two ferromagnets separated by a thin insulator (typically a few nm). Electrons can tunnel from one ferromagnet into the other.
- MTJ is the basis of **MRAM** – Magnetoresistive random-access memory.





Parallel aligned magnetization assists the electron tunneling.

- RSM构建NAND逻辑门的具体原理：


a
50 nm
b
V_{COND}
V_{CLOSE}
V_{SET}
Close
Open
V_{CLEAR}
V_{OPEN}
'1' (closed)
'0' (open)
c
Current (μA)
'1'
'0'
Open
Close
a
b
c
d
e
f
Voltage (V)
d
Direct current
Quasi-static current–voltage
Closed
(–5 V, 2 μs) pulse
(+9 V, 2 μs) pulse
Opened
Current (A)
Voltage (V)
a
Clocking voltages
Step 1: s = 0
$V_S = V_{CLEAR}$
Step 2: pIMPs
$V_P = V_{COND}$
$V_S = V_{SET}$
Step 3: qIMPs
$V_Q = V_{COND}$ $V_S = V_{SET}$
P
V_P
Q
V_Q
S
V_S
R_G
b
Step 1
s = 0
Step 2
s′ ← pIMPs
Step 3
s″ ← qIMPs′
Steps 1, 2, 3
s″ ← pNANDq
c
Read inputs
IMP
IMP
Read outputs
100 ms
Voltage (V)
V_{COND}
V_{SET}
Current (A)
Time