

浅谈类脑计算





一场比赛
3000美元电费

引子

一场比赛
一块士力架



一场比赛
3000美元电费

引子

一场比赛
一块士力架



一场比赛
3000美元电费

以现在已有的技术想要模拟人脑, 直接外推的结果是需要有1-10 Exaflop/s的算力, 以及4 PB的内存

这一性能理论上可以在2022-2024年达到, 但需要一个能耗高达20-30MW的超级计算机

引子

一场比赛
一块士力架

人脑真实能耗：
20 W



一场比赛
3000美元电费

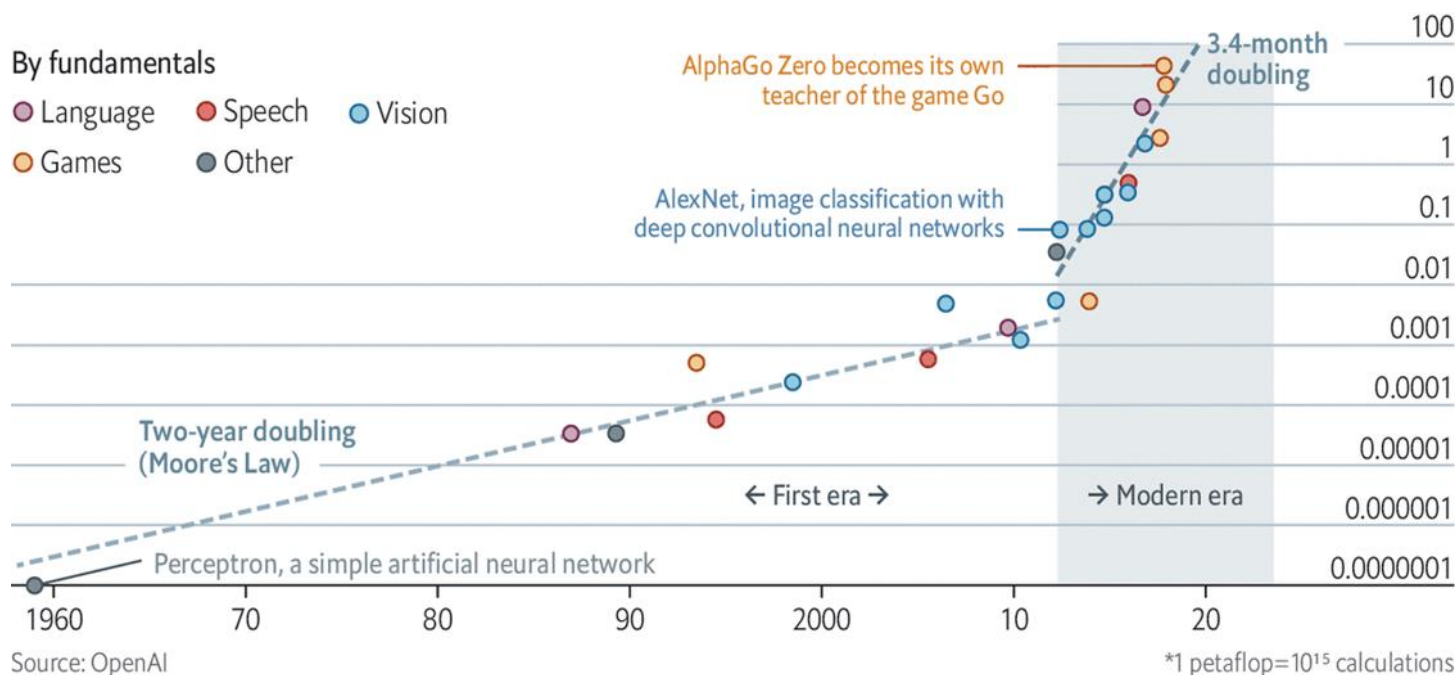
以现在已有的技术想要模拟人脑，直接外推的结果是需要有1-10 Exaflop/s的算力，以及4 PB的内存

这一性能理论上可以在2022-2024年达到，但需要一个能耗高达20-30MW的超级计算机

为什么要学习人脑，为什么要有类脑计算

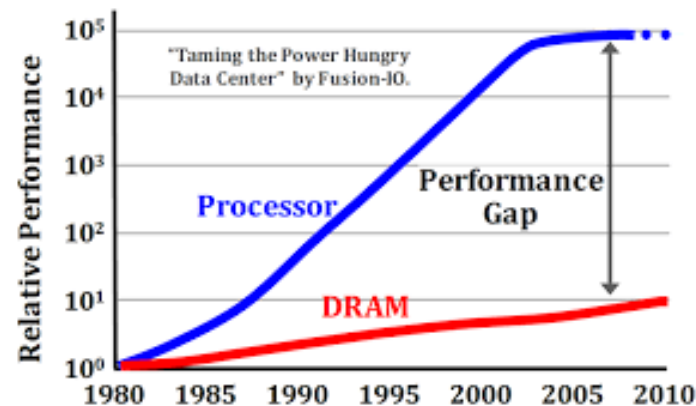
1. 现有架构能耗巨大，难以为继
2. 算力遭遇瓶颈（摩尔定律问题）

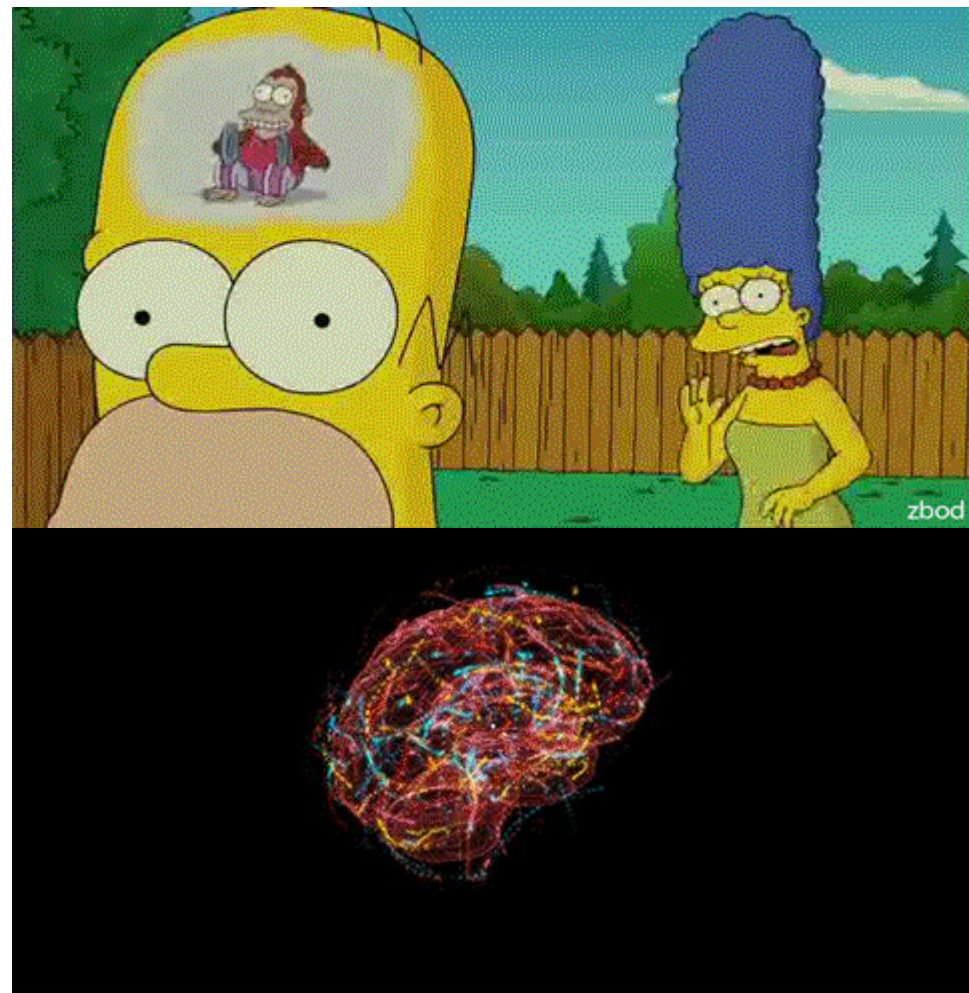
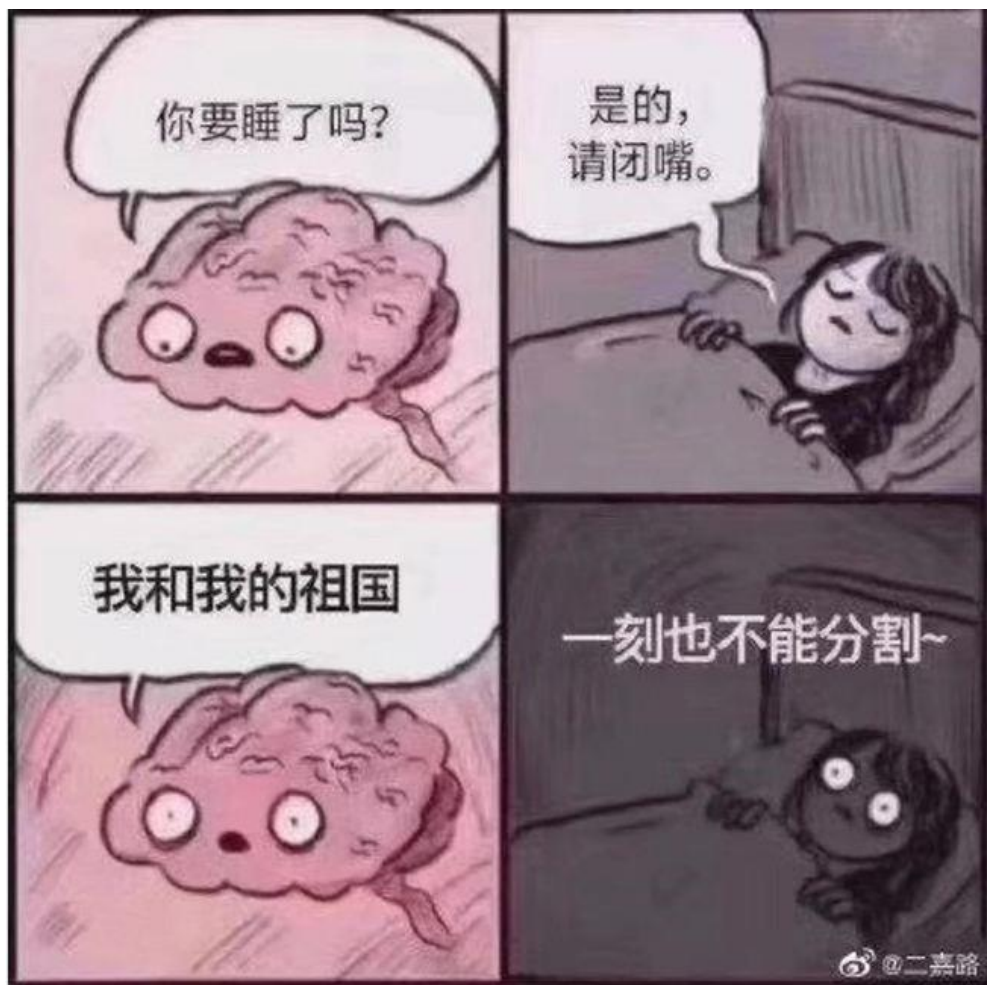
2.1 算力增长跑不过需求



一种可能的解决方案——**类脑计算**

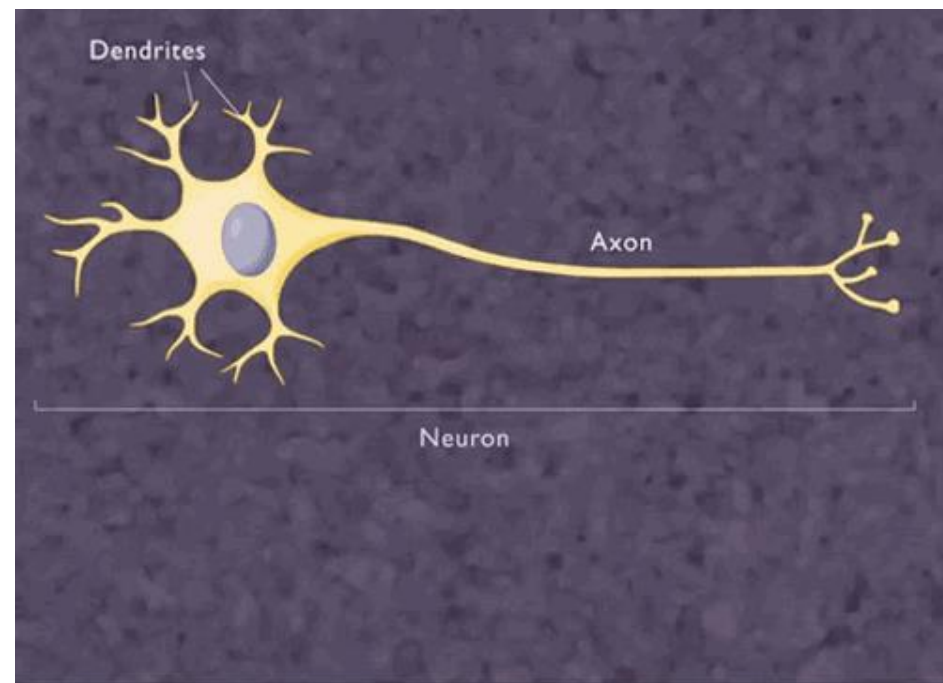
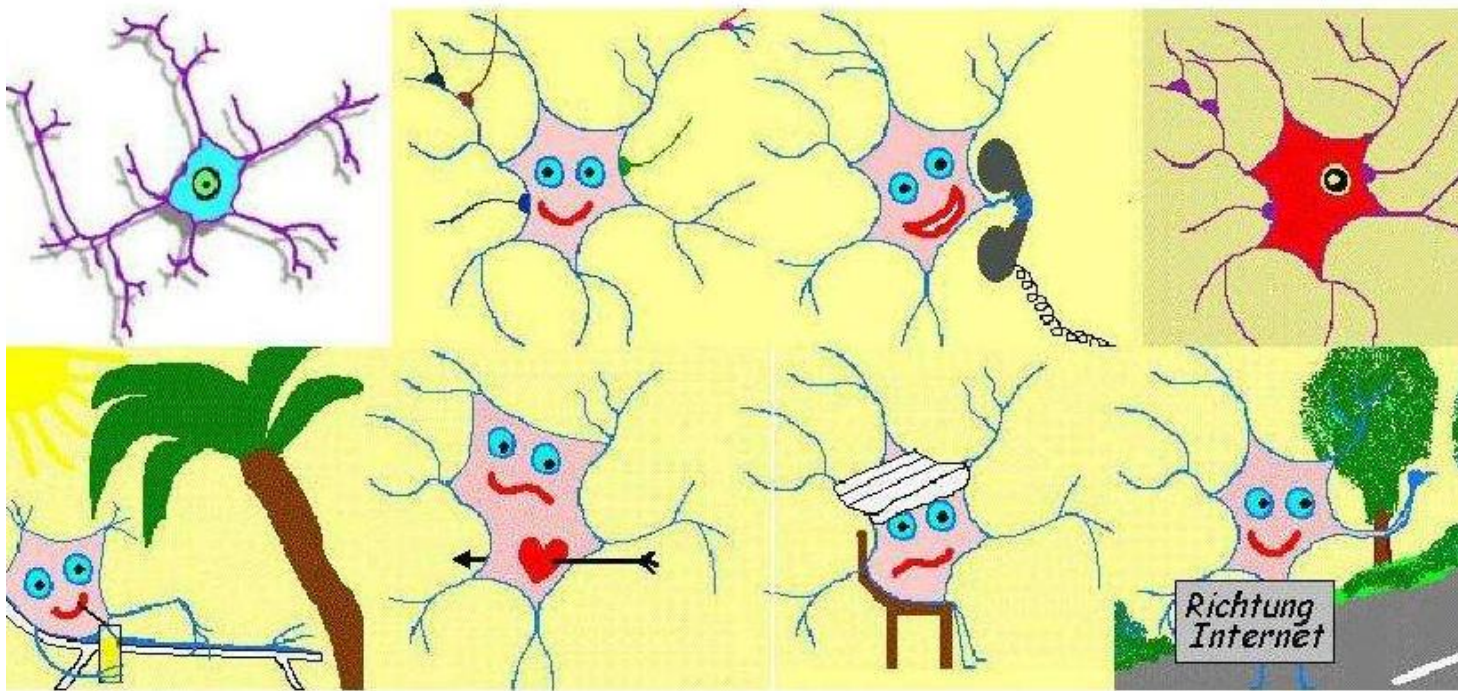
2.2 冯诺依曼架构造成通信瓶颈





- 底层基本原理逐步清晰, 但整个回路远未打通
- 自顶向下走不通, 自底向上再看看

人脑计算：神经元、突触



神经元的特点：

- ① 可塑性（有记忆）——可存储
 - ② 多环境响应（复杂神经网络）——可计算
 - ③ 尺寸小（人脑中有超过 10^{10} 个神经元）——可集成
- 存算一体
胜过经典计算机

怎么做类脑计算

IBM Blue Gene: 暴力模拟

• Human brain

parallel architecture	analog
10^{11} neurons	10 Hz
10^{15} synapses	20 W



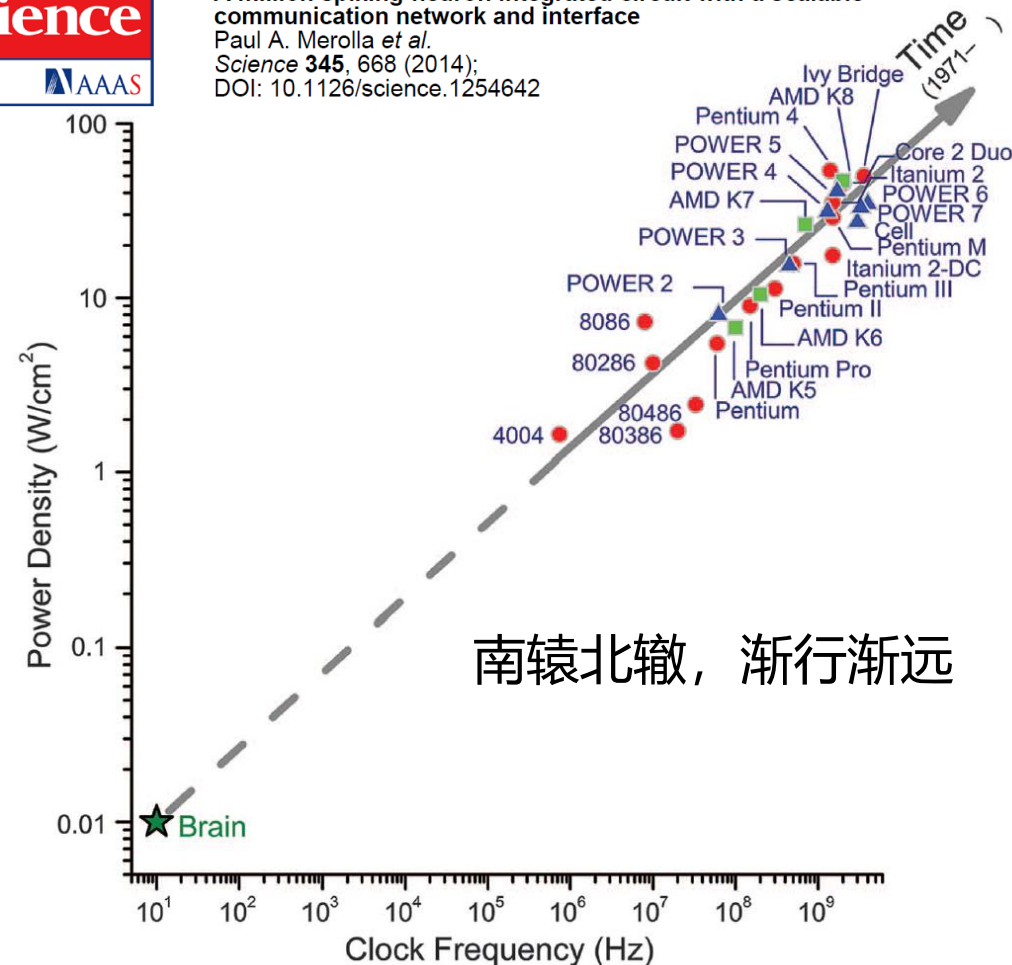
• Simulations of mouse cortex on Blue Gene L

Von-Neumann architecture	digital
$8 \cdot 10^4$ neurons	1 GHz
$5 \cdot 10^{10}$ synapses	40 kW
<i>super-computers slower than mouse ($\times 10$)</i>	



A million spiking-neuron integrated circuit with a scalable communication network and interface

Paul A. Merolla *et al.*
Science **345**, 668 (2014);
DOI: 10.1126/science.1254642



怎么做类脑计算

IBM Blue Gene: 暴力模拟

• Human brain

parallel architecture	analog
10^{11} neurons	10 Hz
10^{15} synapses	20 W



• Simulations of mouse cortex on Blue Gene L

Von-Neumann architecture	digital
$8 \cdot 10^4$ neurons	1 GHz
$5 \cdot 10^{10}$ synapses	40 kW
<i>super-computers slower than mouse ($\times 10$)</i>	

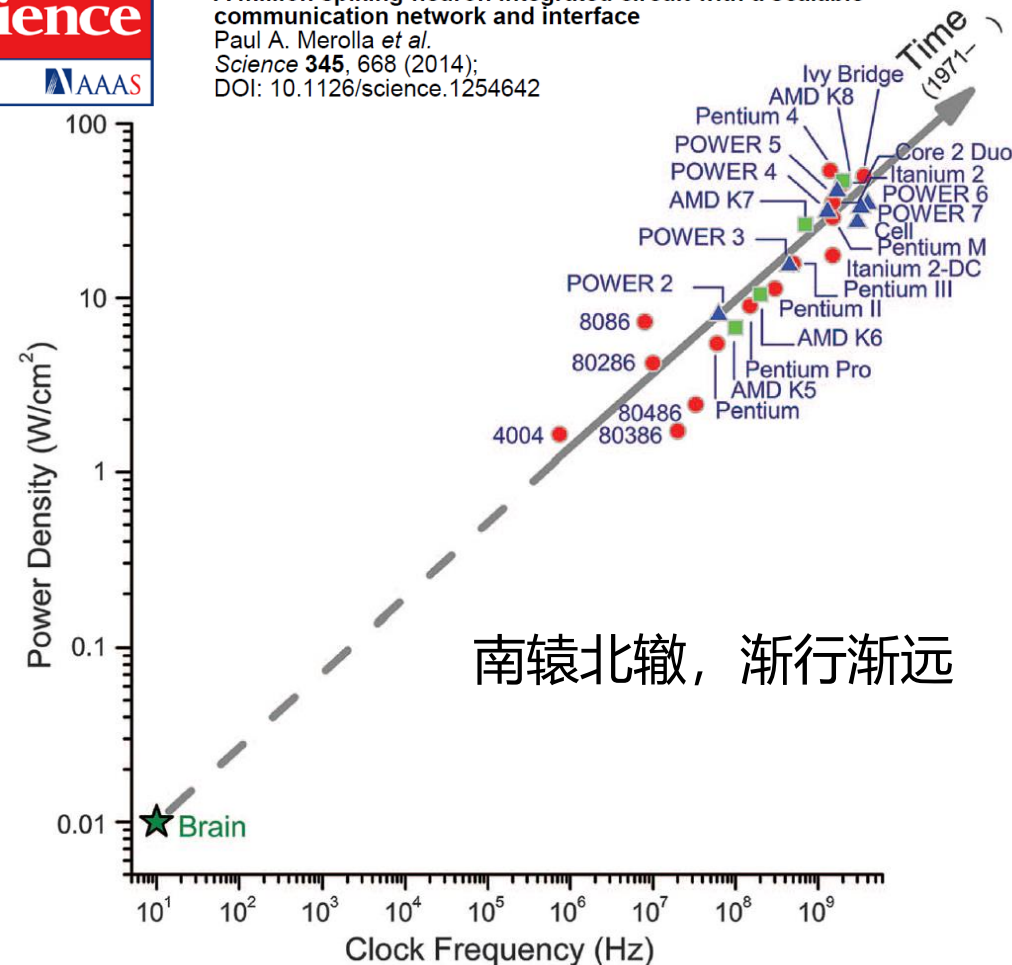


A million spiking-neuron integrated circuit with a scalable communication network and interface

Paul A. Merolla *et al.*

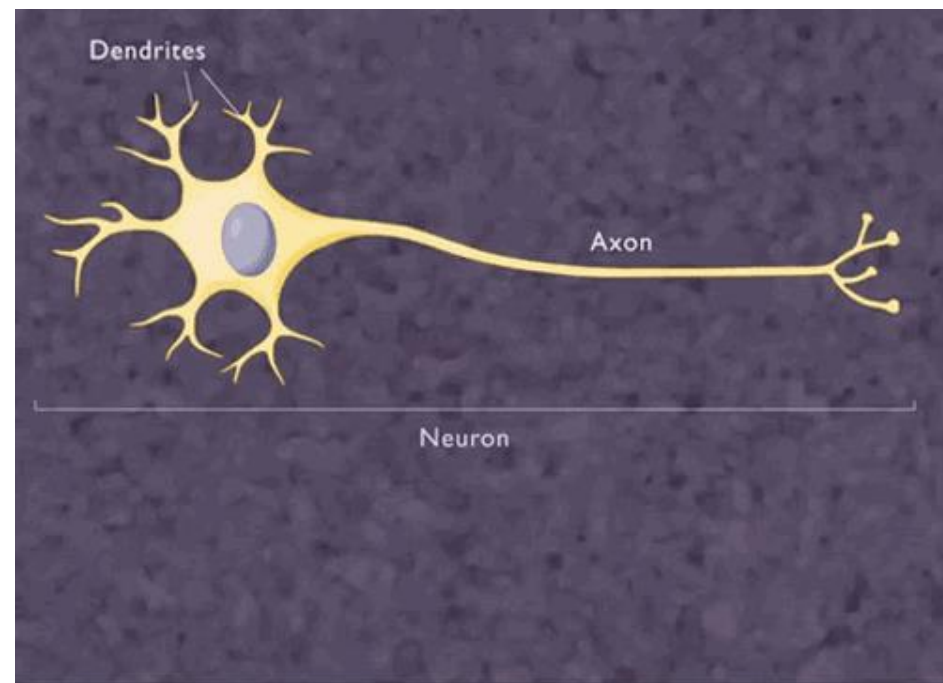
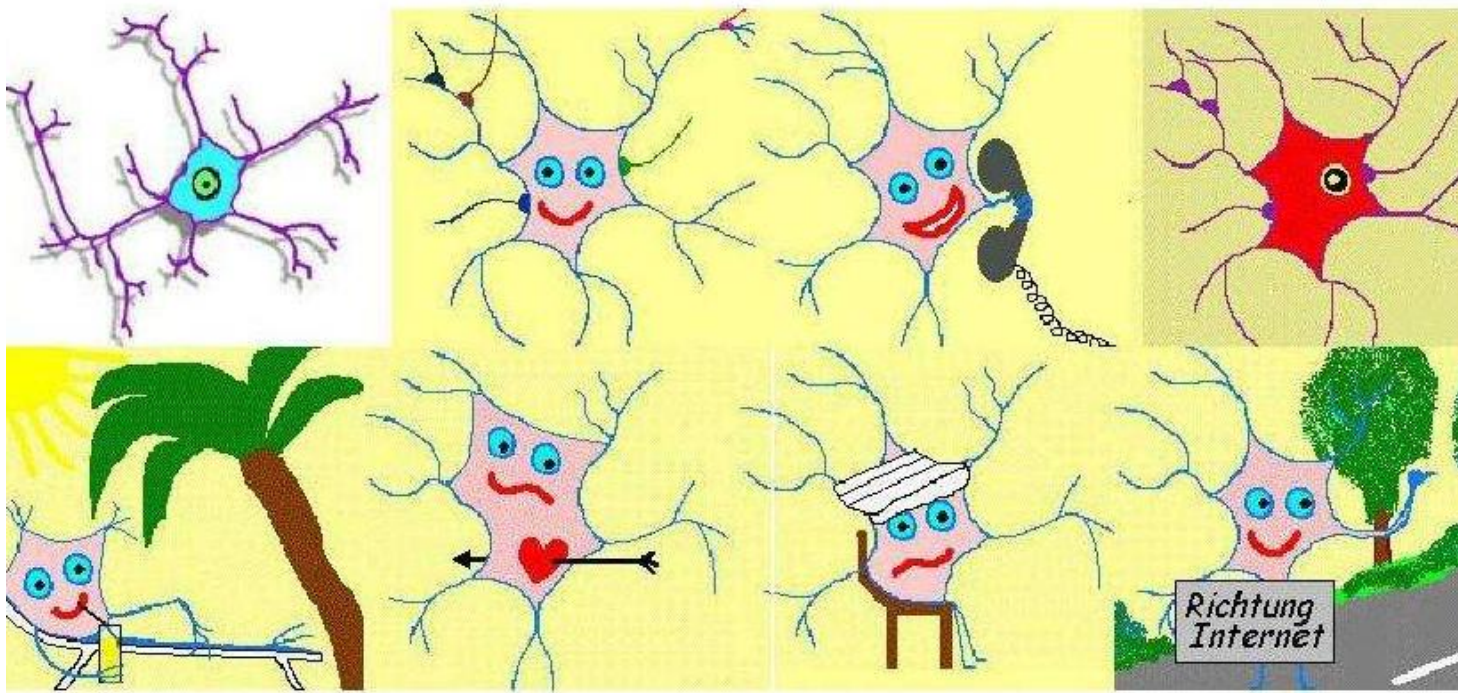
Science **345**, 668 (2014);

DOI: 10.1126/science.1254642



大清早亡了，技术要变革

怎么做类脑计算



神经元的特点:

① 可塑性 (有记忆) ——可存储

② 多环境响应 (复杂神经网络) ——可计算

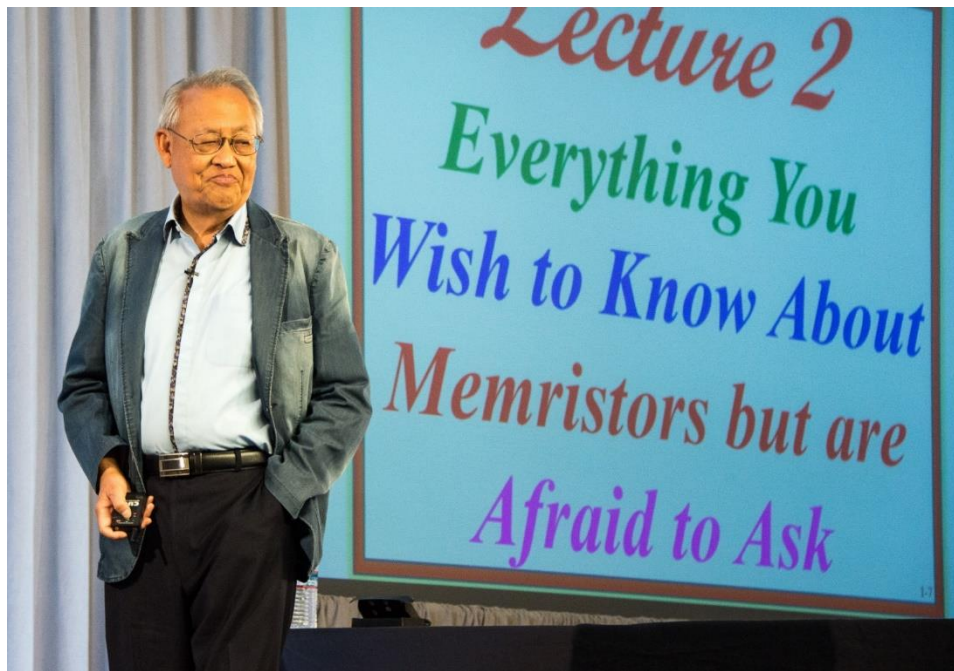
③ 尺寸小 (人脑中有超过 10^{10} 个神经元) ——可集成

存算一体

胜过经典计算机

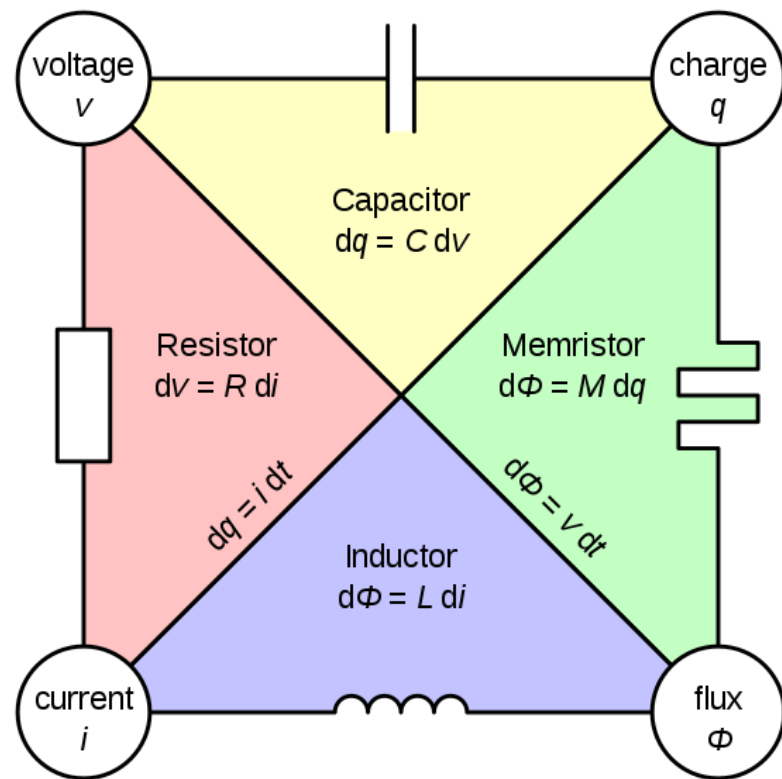
怎么做类脑计算：忆阻器

忆阻器登场



这种组件的效果，就是它的电阻会随着通过的电流而改变，而且就算电流停止了，它的电阻仍然会停留在之前的值，直到接受到反向的电流它才会被推回去

1971 年，任教于加州大学伯克利分校的蔡少棠推断在电阻、电容和电感器之外，应该还有一种组件，代表着电荷与磁通量之间的关系

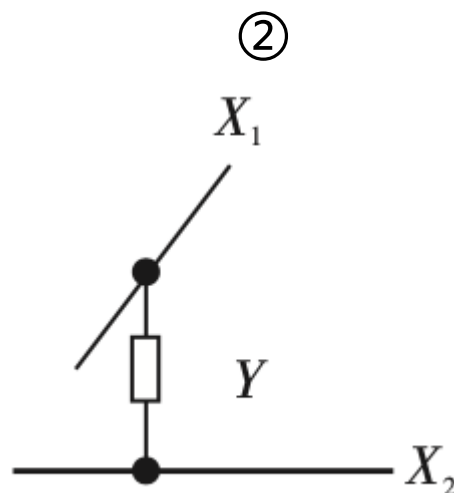
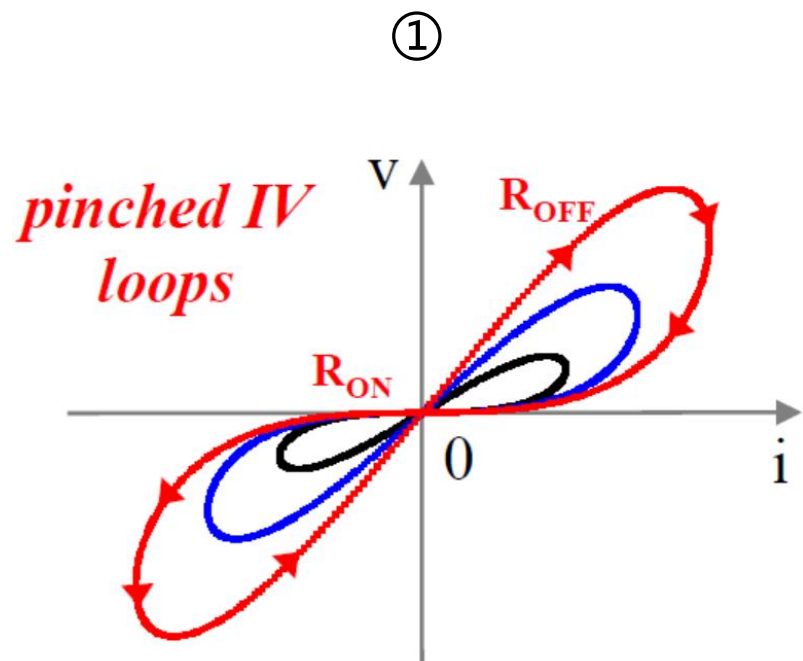


忆阻器能否对应神经元/突触

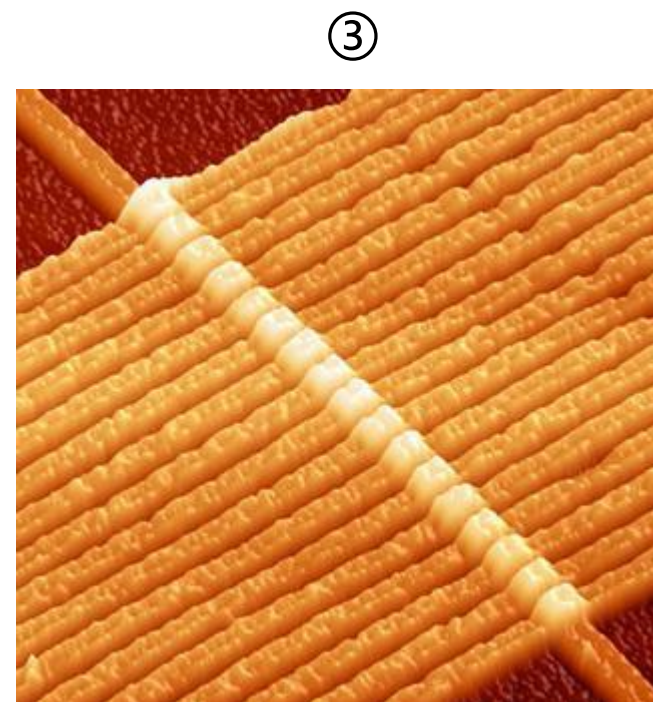
神经元的特点：

- ① 可塑性（有记忆）——可存储
- ② 多环境响应（复杂神经网络）——可计算
- ③ 尺寸小（人脑中有超过 10^{10} 个神经元）——可集成

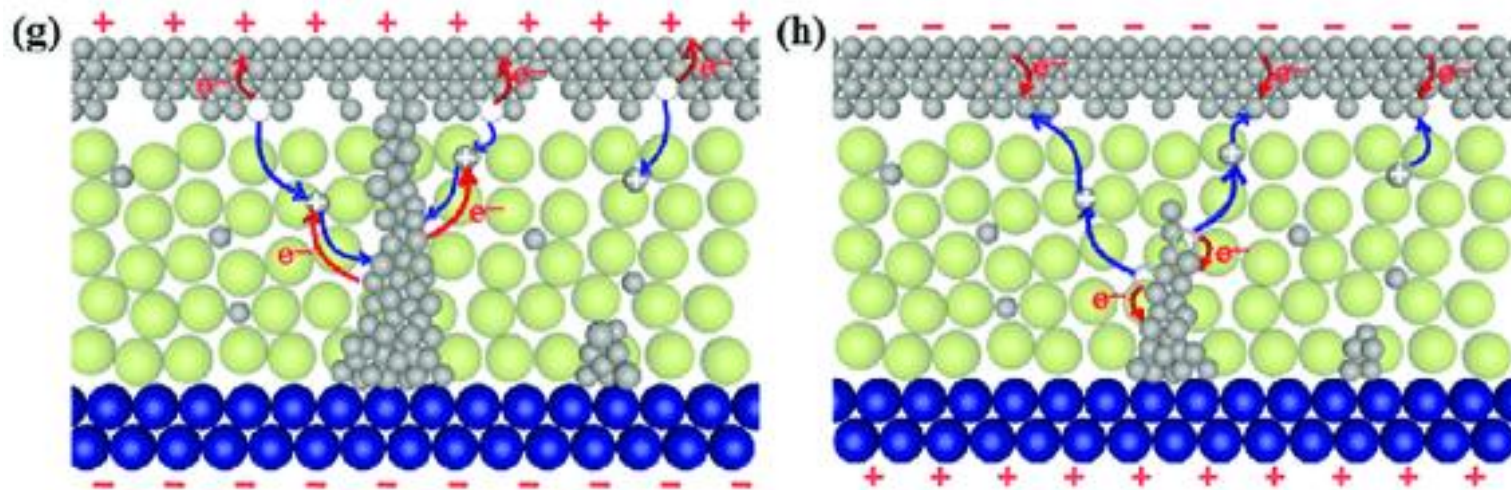
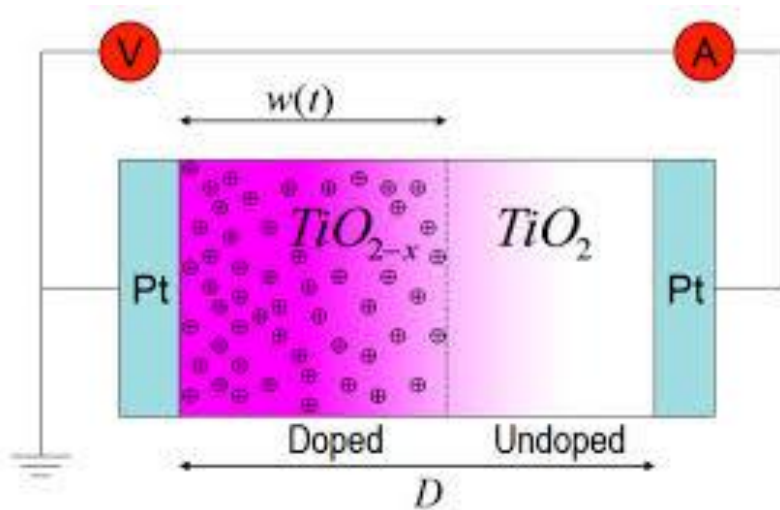
忆阻器：



$$I_i = \sum_{j=1}^N G_{ij} \cdot V_j$$

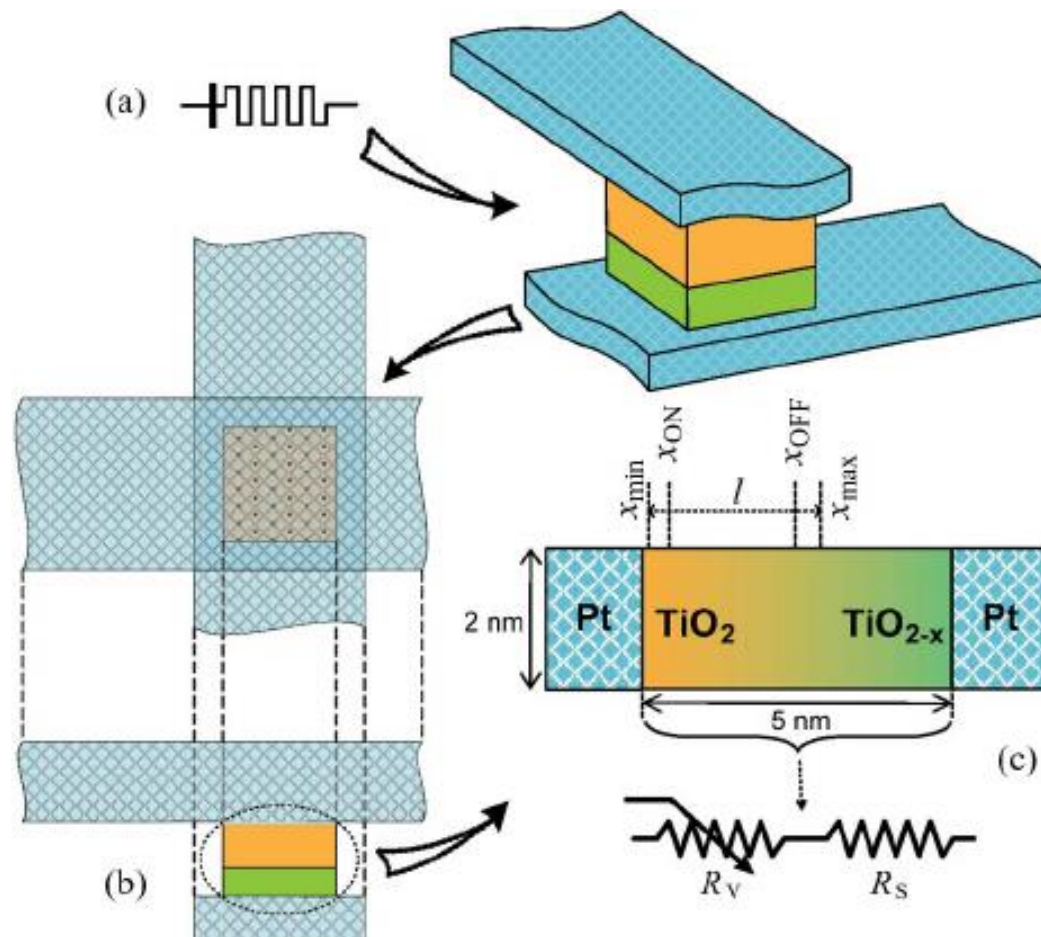
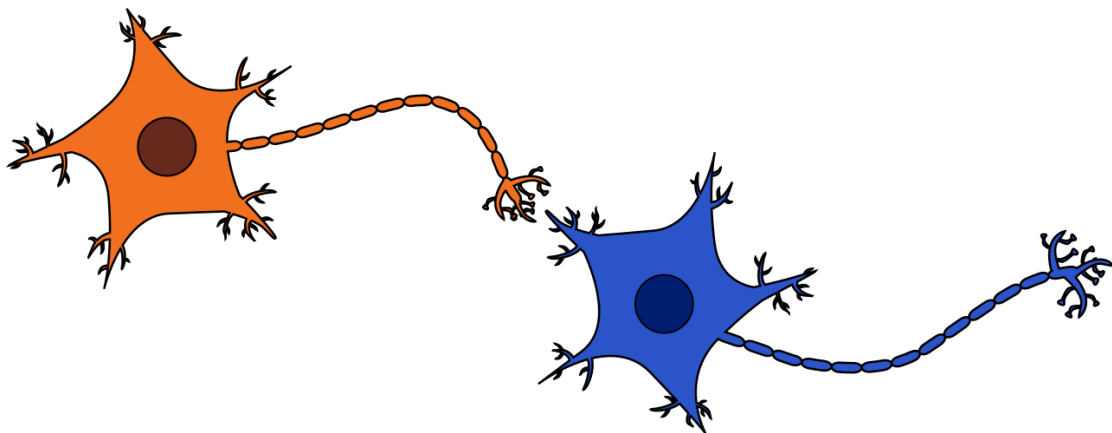


能不能找到忆阻器？



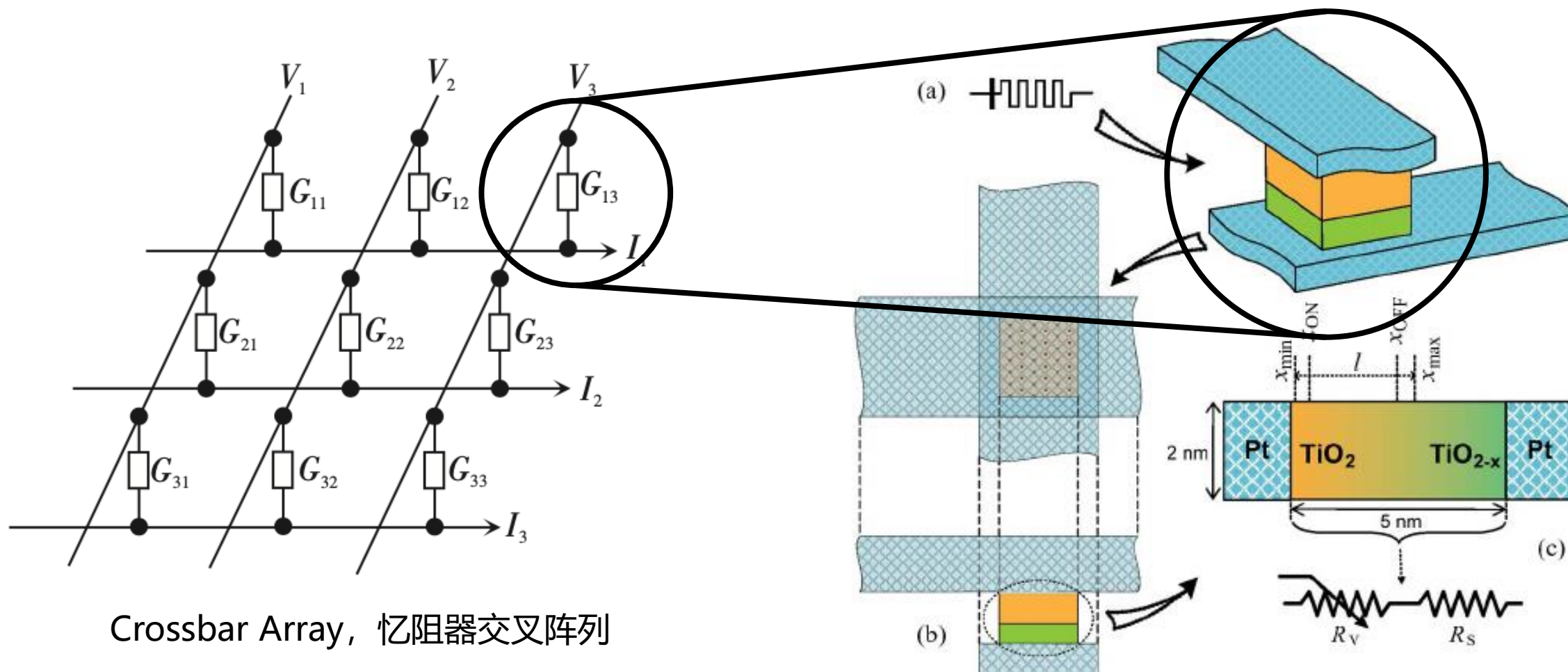
- 2008年，Stanley等人发现，一块极薄的二氧化钛被夹在两个电极中间，这些二氧化钛又被分成两个部分，一半是正常的二氧化钛，另一半进行了“掺杂”，少了几个氧原子
- “掺杂”的那一半带正电，电流通过时电阻比较小，而且当电流从“掺杂”的一边通向正常的一边时，在电场的影响之下缺氧的“掺杂物”会逐渐往正常的一侧游移
- 因此，整个器件就相当于一个滑动变阻器一样

有了忆阻器，怎么做计算？



耶和华把手伸向亚当，灌注灵魂
电流通过忆阻器，灌注参数

有了忆阻器，怎么做计算？

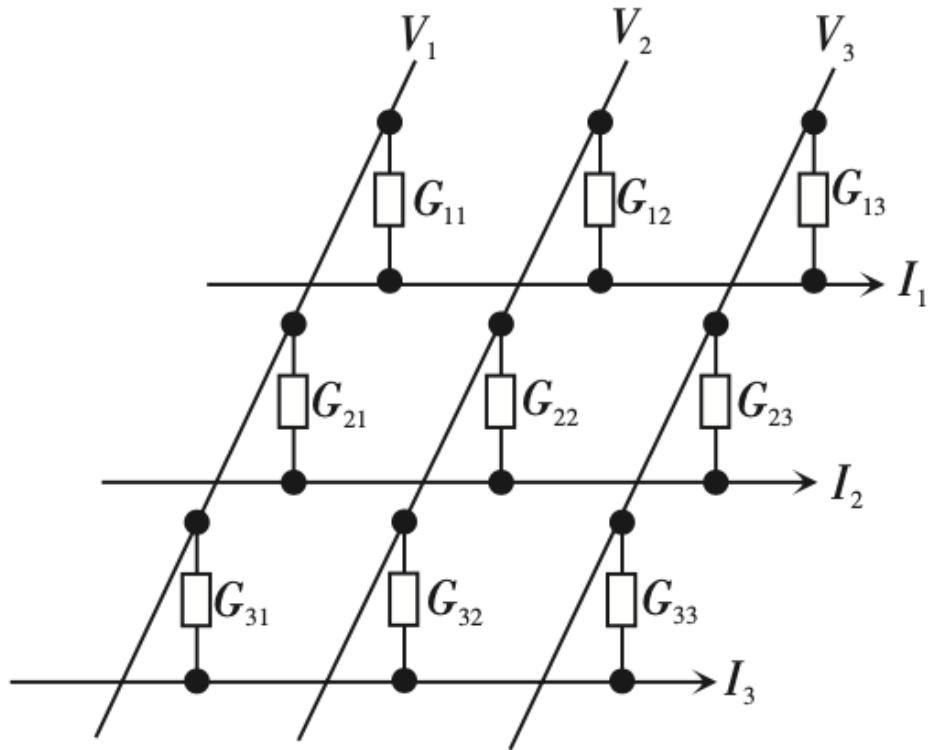


Crossbar Array, 忆阻器交叉阵列

$$I_i = \sum_{j=1}^N G_{ij} \cdot V_j$$

电压 V_j 是施加在第 j 列的电压值，根据欧姆定律和基尔霍夫定律，可以得到第 i 行的总电流值。其中 G_{ij} 为位于第 j 列第 i 行的忆阻器件的电导值

有了忆阻器，怎么做计算？



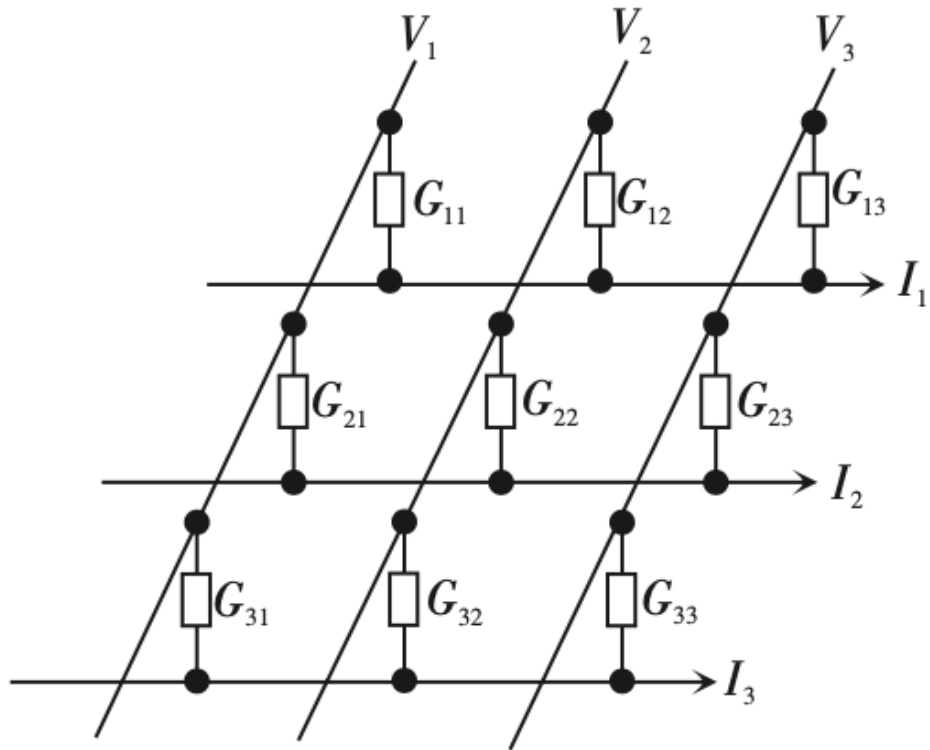
Crossbar Array, 忆阻器交叉阵列

$$I_i = \sum_{j=1}^N G_{ij} \cdot V_j$$

电压 V_j 是施加在第 j 列的电压值，根据欧姆定律和基尔霍夫定律，可以得到第 i 行的总电流值。其中 G_{ij} 为位于第 j 列第 i 行的忆阻器件的电导值

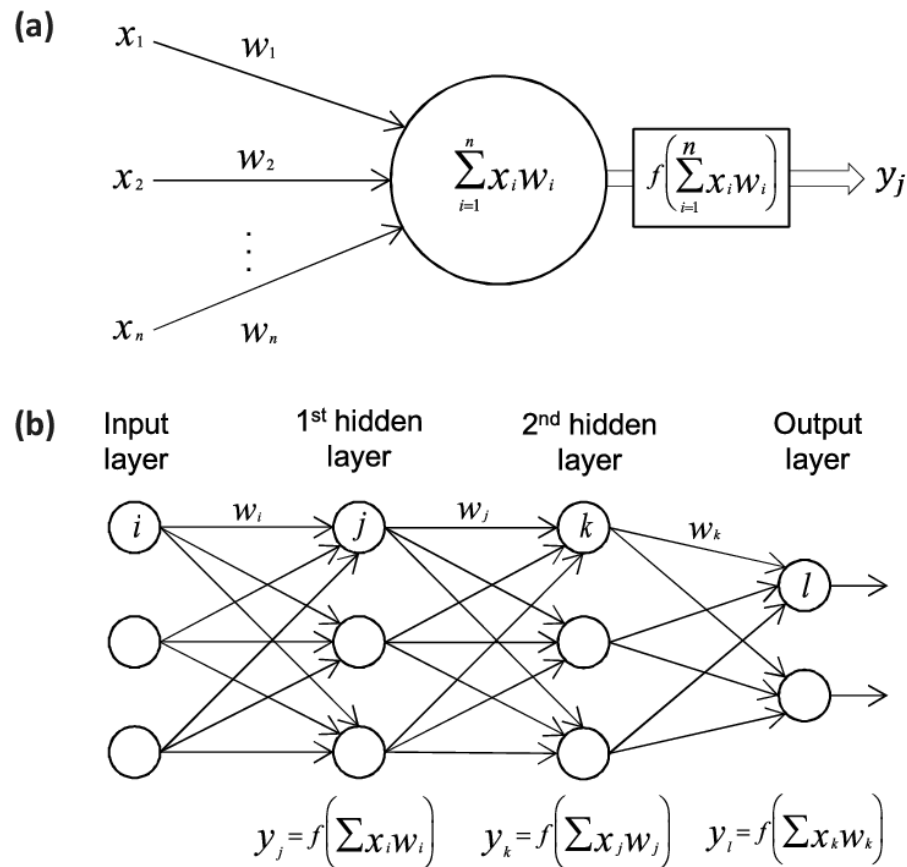
- 总电流值 I_i 是电导矩阵与电压向量的乘积结果
- 从存内计算角度来说，模拟型交叉阵列完成乘法-加法过程只需要一步，自然地可以实现矩阵向量乘的硬件加速
- 相比于传统的计算过程，这样的加速阵列更加节时、节能
- 模拟型交叉阵列可以在稀疏编码、图像压缩、神经网络等任务中担任加速器的角色

忆阻器-神经网络



Crossbar Array, 忆阻器交叉阵列

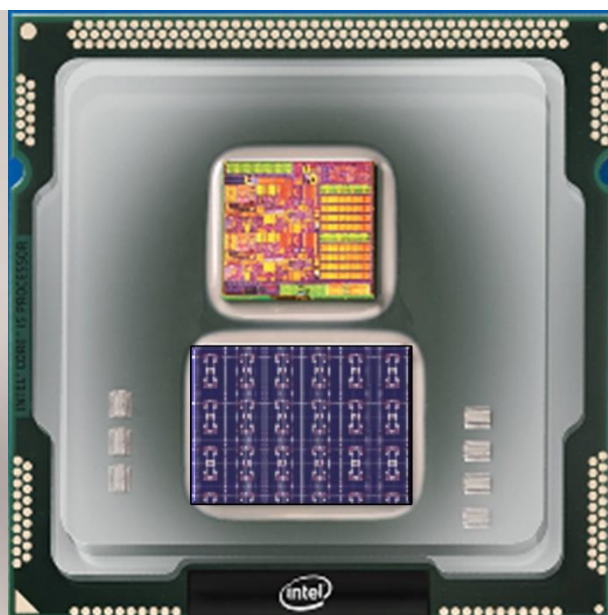
$$I_i = \sum_{j=1}^N G_{ij} \cdot V_j$$



- 在神经网络中, G_{ij} 代表突触权重的大小, V_j 是前神经元 j 的输出值, I_i 是第 i 个神经元的输入值
- 对于左图的交叉阵列, 列线与行线分别代表神经网络中的输入神经元和输出神经元, 忆阻器的电导值为神经元之间相互连接的突触权重值
- 利用反向传播等学习算法可以通过 SET/RESET 操作来原位更新网络权重

类脑计算进展

- 全球范围内主要在推进的类脑计算芯片，括 IBM 的 TrueNorth、Intel的Loihi、BrainScale、SpiNNaker、Neurogrid 等主要技术流派，中国则以清华的天机芯占据一席之地
- 脉冲神经网络芯片的代表是IBM2014年发布的 TrueNorth，其基本结构由硬件神经元和神经元之间的脉冲连接组成，硬件神经元接收输入脉冲，在累积到一定阈值后被激活产生输出脉冲。具有4096个处理核，每个内核包含256个硬件神经元，总共可以模拟100万个神经元和2.56亿个突触
- “天机芯”采用28纳米工艺制成，整个芯片尺寸为3.8 X 3.8mm²，由156个计算单元（Fcore）组成，包含约40000个神经元和1000万个突触。它把人工通用智能的两个方向，即基于计算机科学和基于神经科学这两种方法，集成到一个平台，可以同时支持机器学习算法和现有类脑计算算法

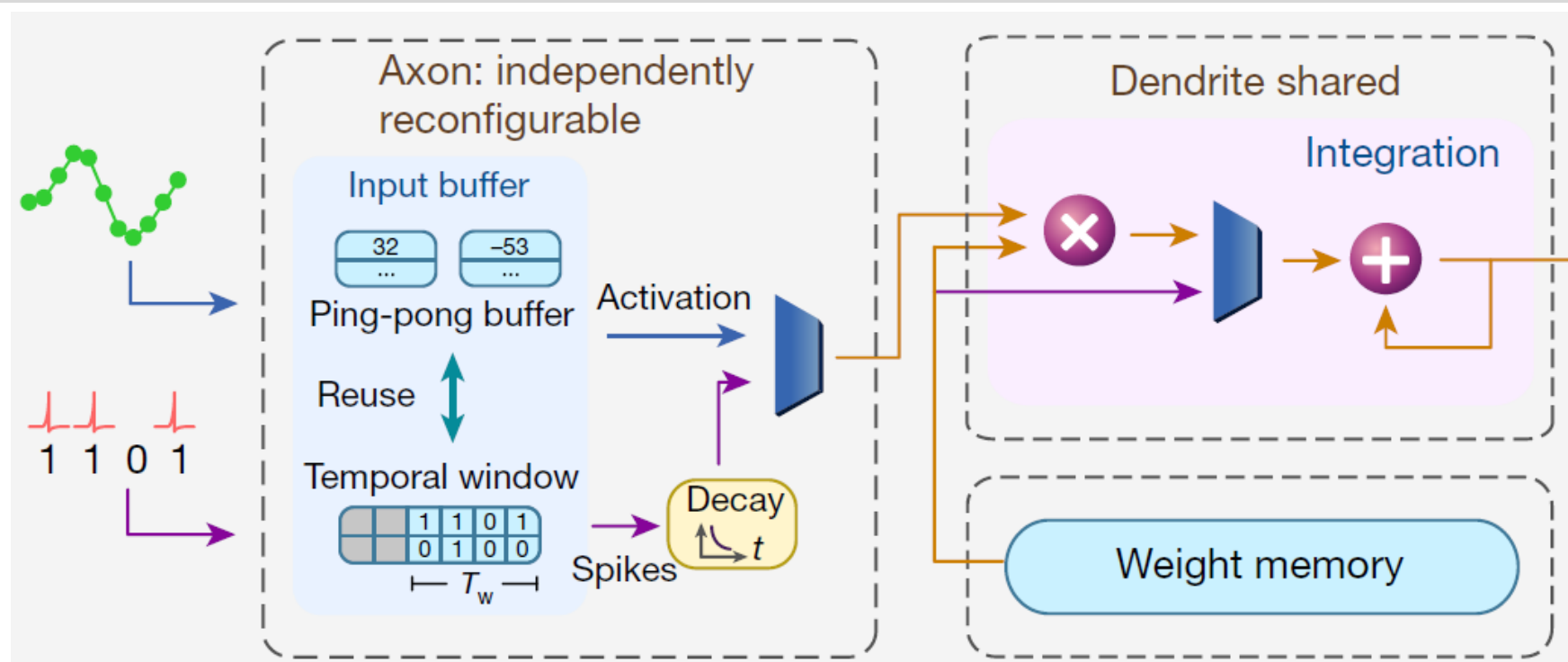
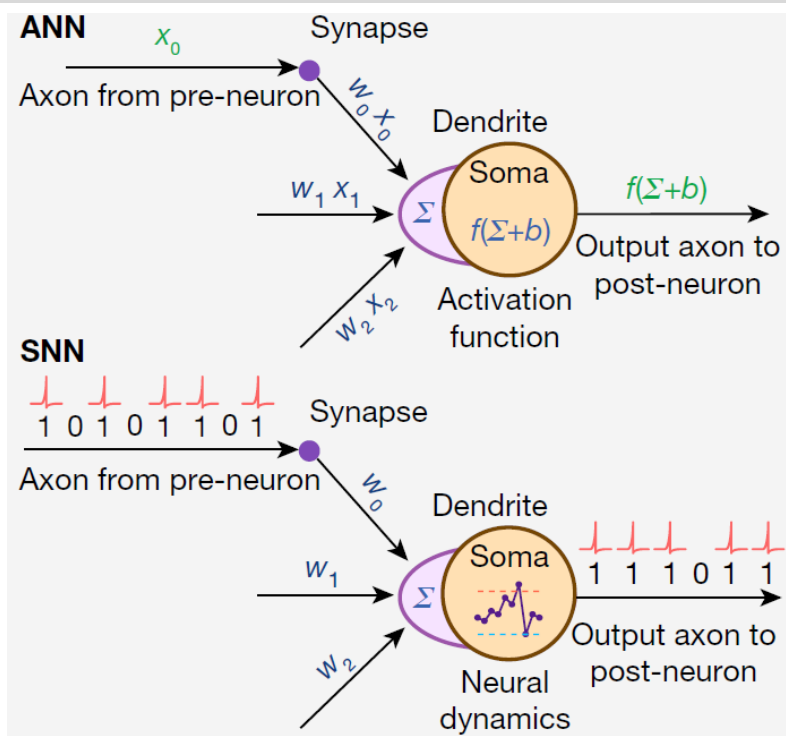


类脑计算进展：1 天机芯



- ❑ 目前人工智能的发展主要有2个方向，一是算法，一是仿生
- ❑ 算法方向主要是利用计算机运算的优势，通过大量数据的学习迭代形成有效的算法及参数，这是目前的主流方向，如我们熟知的深度学习框架CNN，RNN等，统称为ANN
- ❑ 另外一种是用电路器件模拟出生物神经元的连接和运行方式，搭建和人类大脑类似的模型来实现
- ❑ “天机芯”却能把这两种原本互不兼容的人工智能芯片融为一体，成为世界首款异构融合类脑芯片。这种融合技术有望实现人工通用智能（AGI）

类脑计算进展：1 天机芯



- ❑ 模拟神经元的代表性模型就是脉冲神经网络（SNN），主要通过计算神经电脉冲进行信息传递
- ❑ 这个和传统网络的权重连接+激活的方式有很大差别。Tianji的主要创新是将这两者统一在一个运算单元中，可以使一块芯片同时支持通用的深度学习模型和研究性的脉冲神经网络
- ❑ 语音识别、视觉追踪是受脑启发的模型；目标探测和运动控制是机器学习算法；而自主决策则是一个两者混合的模型
- ❑ 最大的挑战不是来自科学技术，而在于我们的学科分布过细，不利于解决这样的复杂问题。所以，多学科深度融合是解决问题的关键，可以把电脑思维和人脑思维的优势结合起来，帮助我们发展人工智能

类脑计算进展：2 图灵完备体系

nature

View all Nature Research journals

Search  Login 

Explore our content ▾



Journal information ▾

Subscribe

nature > articles > article

Article | Published: 14 October 2020

A system hierarchy for brain-inspired computing

Youhui Zhang , Peng Qu, Yu Ji, Weihao Zhang, Guangrong Gao, Guanrui Wang, Sen Song, Guoqi Li, Wenguang Chen, Weimin Zheng, Feng Chen, Jing Pei, Rong Zhao, Mingguo Zhao & Luping Shi 

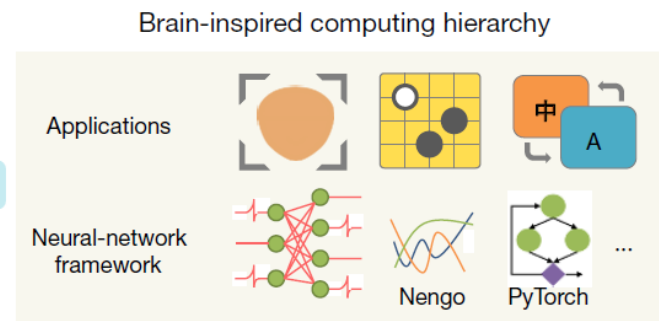
Nature 586, 378–384(2020) | Cite this article

Metrics

- ❑ 经典计算机是图灵完备的
- ❑ 张悠慧等人提出了"神经形态完备性"的概念
- ❑ 他们提出了一种全新的系统层次结构，在该系统层次结构下，各种程序可以用统一的表示来描述，在任何神经形态完备的硬件上都能转换为等效的可执行程序，从而确保编程语言的可移植性、硬件的完备性和编译的可行性

Software

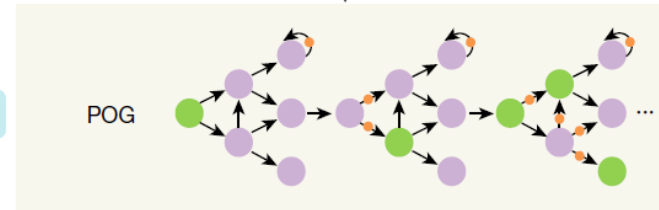
Turing-complete



Exactly equal

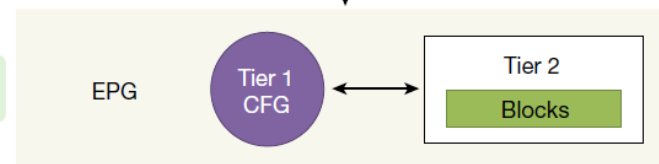
Compiler

Turing-complete



Approximately equal

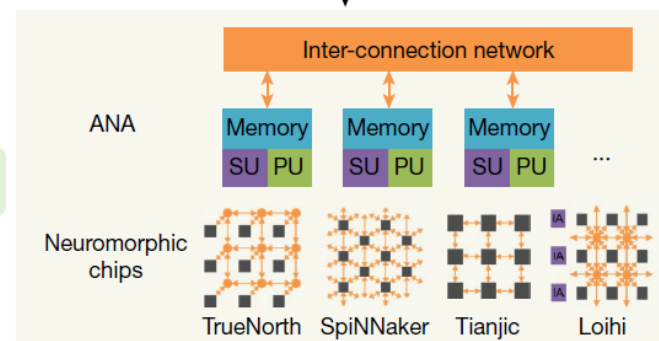
Neuromorphic-complete



Exactly equal

Hardware

Neuromorphic-complete



nature

Explore our content ▾


Journal information ▾

Subscribe

nature > articles > article

Article | Published: 29 January 2020

Fully hardware-implemented memristor convolutional neural network

Peng Yao, Huaqiang Wu , Bin Gao, Jianshi Tang, Qingtian Zhang, Wenqiang Zhang, J. Joshua Yang Qian

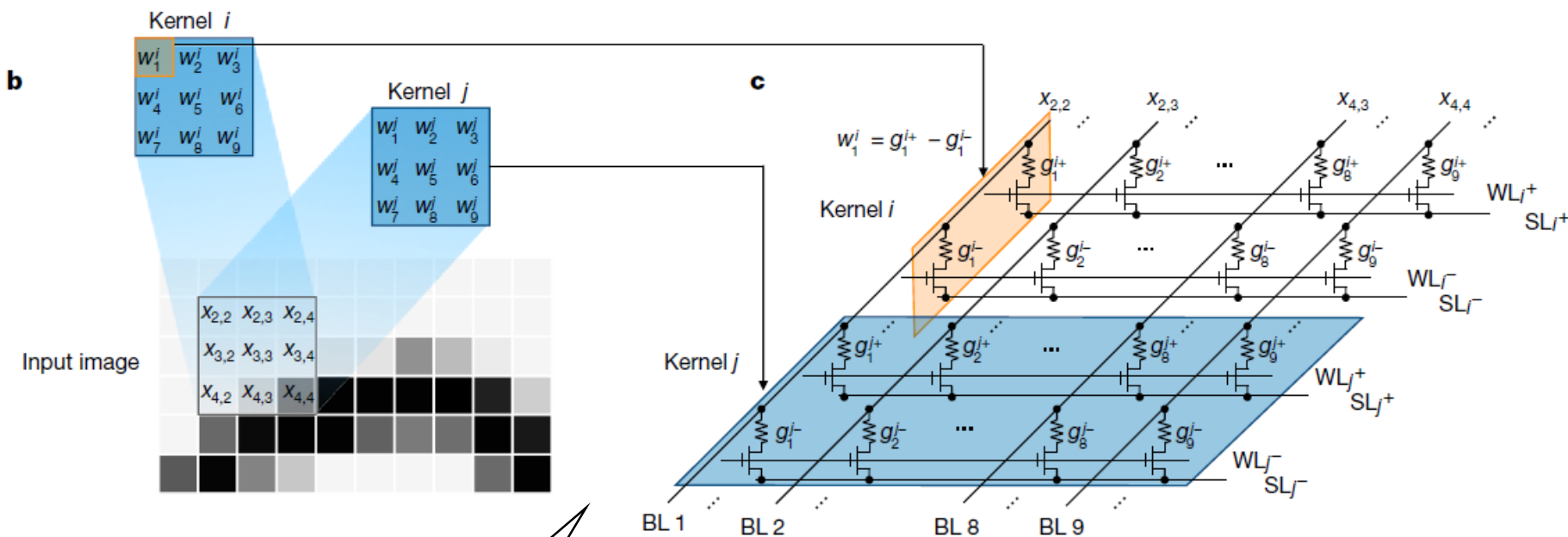
Nature **577**, 641–646(2020) | [Cite this article](#)

21k Accesses | **63** Citations | **97** Altmetric | [Metrics](#)



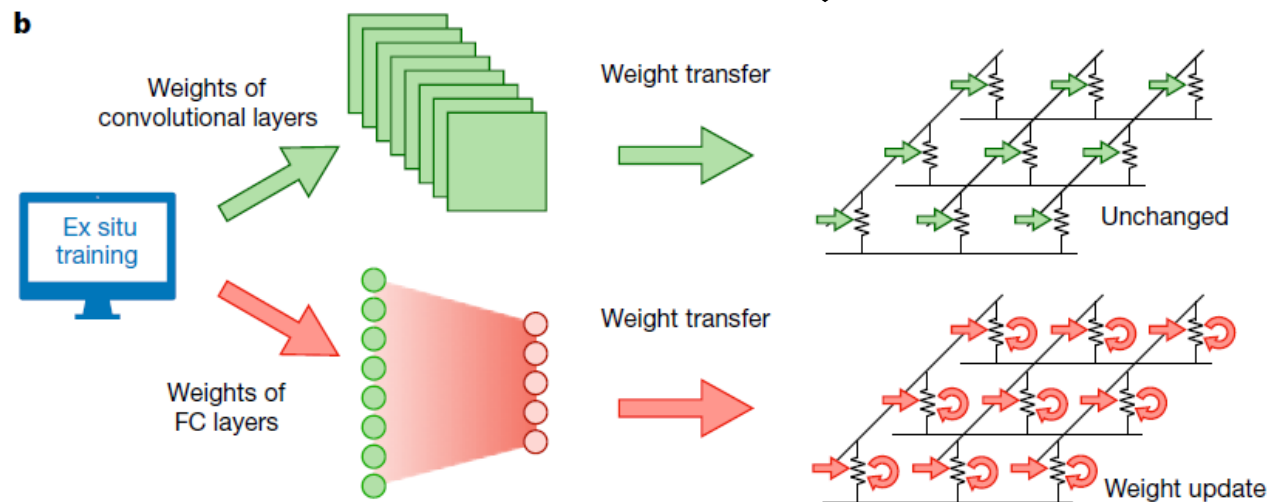
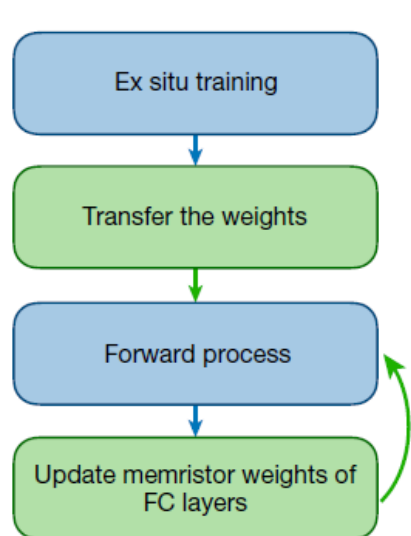
- ❑ 该成果所研发的基于多个忆阻器阵列的存算一体系统，在处理卷积神经网络（CNN）时比图形处理器芯片Tesla V100 GPU相比具有110倍的能效优势，大幅提升了计算设备的算力，成功实现了以更小的功耗和更低的硬件成本完成复杂的计算
- ❑ 混合训练方法在处理图像识别任务时，多值忆阻器硬件系统以达到和软件相当的识别准确率（96.19%）

类脑计算进展：3 忆阻器CNN



- 先将片外训练好的权重编程到各忆阻器阵列中
- 再通过片上训练重要层权重的方式，自适应的弥补器件非理想特性的影响

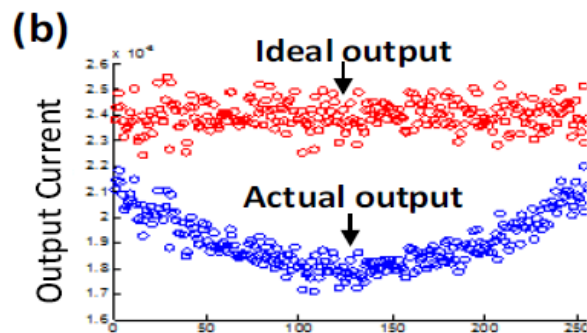
- 权重的硬件对应，
- 卷积的硬件对应



所以忆阻器无敌了吗，类脑计算解决了吗？

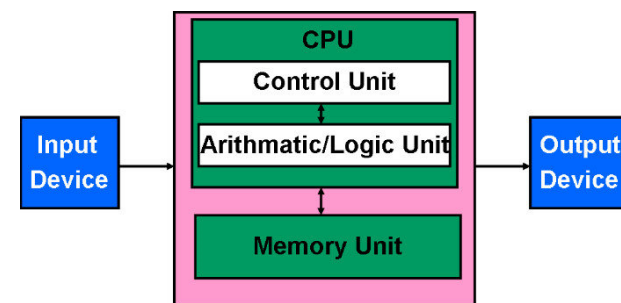
框架不完善，模块待开发

器件不稳定，性能不够好

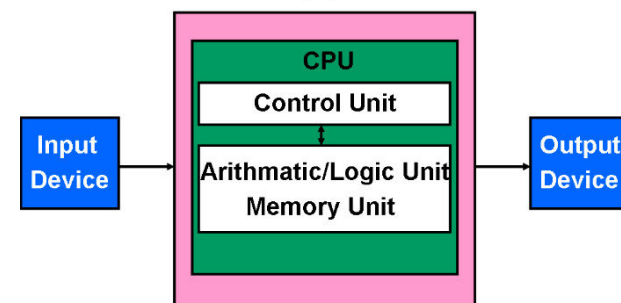


器件误差→逐器件累积
阵列误差→逐层累积
网络误差

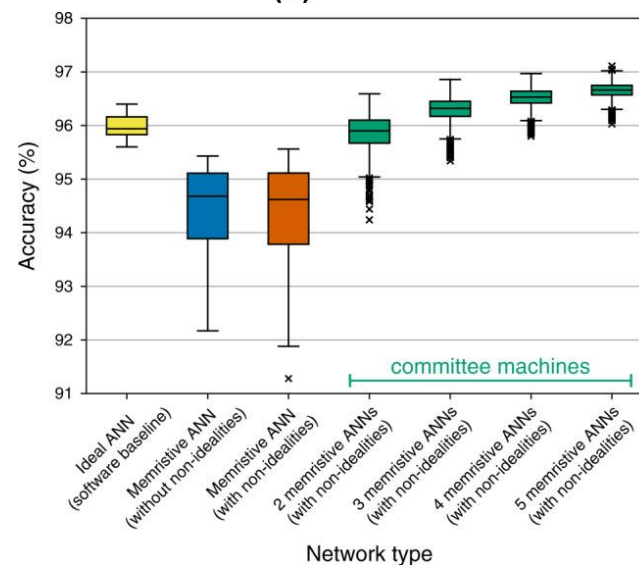
计算误差大，结果不精确



(a)



(b)



今天讲了点啥

类脑计算有望颠覆冯诺依曼架构，突破摩尔定律极限，算得快、能耗少、潜力大

Neurotransmitters

类脑计算可通过忆阻器等非易失性存储器件实现

类脑计算受到广泛关注，并正产生大量成果，逐步得到应用

类脑还存在诸多问题有待解决

谢谢！