北京大学
PEKING UNIVERSITY

# 人工智能的硬件基石

## 从物理器件到计算架构

### 第一讲：人工智能芯片与芯片发展简史

主讲：陶耀宇

2025年春季

# 课程简介

北京大学 PEKING UNIVERSITY

- **培养学生初步理解智能时代的硬件芯片的工作原理、设计原理与未来发展方向**
  - 分别从**物理器件**、**逻辑电路**、**计算架构**3方面进行全栈式介绍

| 指标 | 课程信息 |
| --- | --- |
| 课程号 | 04632043 |
| 学分 | 2 |
| 课程体系 | 专业任选 |
| 地址 | 理教417 |
| 优秀率 | 无强制限制 |
| 考核方式 | 出勤（5%）、6次课后作业（30%）<br>简单硬件编程实验（Lab 1 15% + Lab 2 30%）、期末汇报（20%） |



**扫描二维码**：加入【人工智能硬件基石】群
添加说明：年级-姓名

**前置知识要求**：无强制先修要求、建议具备最初步编程能力

**编程技能**：简单Python、Verilog（助教将通过习题课逐步进行教学）

**课程网站**：
https://aiarchpku.github.io/2025Spring/

**推荐教科书**：

- **物理器件/逻辑电路方面**：
  - Digital Integrated Circuits: A Design Perspective - Anantha Chandrakasan
  - CMOS数字集成电路：分析与设计 - 康松默
- **智能计算架构方面**：
  - Computer Architecture: A Quantitative Approach - John L. Hennessy
  - 智能计算系统 - 陈云霁
  - 人工智能芯片设计 - 尹首一

# 课程团队

- **培养学生初步理解智能时代的硬件芯片的工作原理、设计原理与未来发展方向**
  - 分别从**物理器件**、**逻辑电路**、**计算架构**3方面进行全栈式介绍

**主讲：**陶耀宇

**负责课程所有相关内容**

**助教：**王泊闻

**集成电路学院博士生1年级**

**主要负责课程网站、助教课、**

**作业批改、实验编程等**

**助教：**詹喆

**信息科学技术学院本科4年级**

**主要负责CLAB服务器平台、**

**实验编程等**

# 课程作业与实验

- 培养学生初步理解**智能时代的硬件芯片的工作原理、设计原理与未来发展方向**
  - 分别从**物理器件**、**逻辑电路**、**计算架构**3方面进行全栈式介绍

- **6次课后作业（总计30%，每次占总成绩5%）**
  1. 半导体物理基础与CMOS器件
  2. 逻辑门级电路与复杂计算单元
  3. 指令集、多级流水线及其控制
  4. 超标量与乱序执行
  5. 存储系统微架构
  6. AI加速器架构

- **2次编程实验（总计45%，分别为15%+30%）**
  1. （利用课程所学CMOS器件与逻辑电路知识，构建复杂计算单元（例如：如何在硬件上实现三角函数计算？如何在硬件上实现高效的大位宽乘法？），并且评估其硬件性能。Python/Matlab、Verilog，代码编写量：200-300行左右
     **（1个月时间完成）**
  2. 利用课程所学流水线、指令集、AI芯片架构知识，构建简单芯片架构与电路，能够实际运行某一类AI计算任务，并且评估其硬件性能。简单Linux脚本、Verilog，代码编写量：400-500行左右
     **（2个月时间完成）**

# 课程实验环境

- **培养学生初步理解智能时代的硬件芯片的工作原理、设计原理与未来发展方向**
  - 分别从**物理器件**、**逻辑电路**、**计算架构**3方面进行全栈式介绍
    - 项目采用CLAB平台：https://clab.pku.edu.cn/、具体使用方式请参考： CLAB使用手册
    - 实验采用Linux环境进行开发
      - 所需软件环境已为各位同学安装好，无需自己配置环境
      - Linux运行Lab的说明请参考：Linux使用参考信息
    - 助教正在为各位同学建立CLAB账号，具体事宜请同学们联系助教詹喆同学
    - **第2周开始，为大家提供Verilog入门习题课，支持硬件0基础的同学们**
    - 如有任何问题，欢迎联系授课老师或助教！

# 期末汇报

- **自行选择器件（IEDM等）、电路（ISSCC、VLSI等）或架构（MICRO、ISCA等）方面的论文，或Nature/Science系列相关论文，进行1-2篇文献阅读**
  - **占总成绩20%**，深入论文技术细节并做**12-15分钟汇报，以PPT+视频报告形式提交**

**AI加速器芯片**
- 传统AI加速器：Fused-layer cnn accelerators、Eyeriss，Google TPU等
- 新兴AI加速器（大模型Transformer、Neural ODE、MANN/DNC、PINN等）

**GPGPU芯片**
- 流式多处理器（Multithreaded Streaming Multiprocessors，CUDA的来源）

**FPGA芯片**等（可编程逻辑块、可编程路由等）

**安全与通信领域处理器芯片**
- 各类密钥编码（AES、RSA）、视频编码（MPEG等）、通信编码（LDPC、Polar等）

**传统CPU芯片**
- 优化Branch Predictor、Load-Store、缓存预读取、众核缓存一致性等

**新兴智能计算芯片**
- 存算一体/感存算一体、量子计算、生物信息处理、高维NoC、区块链
- 基于后摩尔非CMOS器件的架构（模拟计算架构、动力学计算架构等）

# 目 录

CONTENTS

# 人工智能产业蓬勃发展

- 人工智能产业是推动我国未来新质生产力发展和经济转型升级的核心驱动力，且持续增长

**人工智能市场规模（亿美元）**


数据来源：前瞻产业研究院

**国家发展战略**

- 2050年世界科技创新强国
- 2030年创新型国家行列
- 2025年制造业强国行列

**国家人工智能发展布局**

大模型为代表的新一代人工智能

## 国内外大模型相关产业现状

**OpenAI**

以OpenAI为代表的AI大模型公司持续发布**GPT系列模型，包括GPT-2/3/4、GPT-o1、Sora等**，复杂推理能力大幅提升，并整合进微软多款产品中，是全球领先的大模型科技公司

**Google DeepMind**

自2018年期发布**BERT、AlphaGo、LAMDA、PaLM、AlphaFold系列AI大模型**，领域覆盖科学计算、语言、视觉等，并最新推出**多模态大模型Gemini**

**Meta AI**

**Llama系列开源大模型**，在文本分析、视觉任务等多领域表现优异，Llama Lite等轻量化大模型用于边缘/端侧应用

**HUAWEI**

发布**盘古大模型系列**，面向视觉、自然语言、科学计算等多个场景；提出五大基础大模型，细分为多个行业大模型，深度渗透政务、金融、制造、气象、医学等多个行业，提升智能化水平

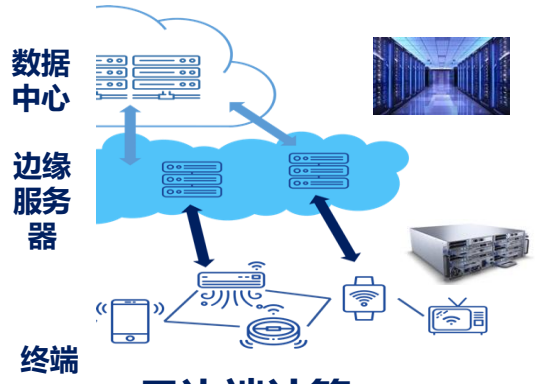
**文心一言** **通义千问** **腾讯混元** **DeepSeek**

# 人工智能产业蓬勃发展

- 提升全社会、全行业智能化水平，助力产业颠覆性发展，服务国家重大战略需求

## 以AI大模型为代表的先进智能技术

### 云边端计算




到 2025 年，全球 云计算系统产业 规模将突破 约 5.6 万亿元

到 2025 年，全球 边缘计算产业 规模将达到 约 1130 亿元

视频安防、智能手机、物联网、工业机器人等典型应用

### AR/VR


到 2025 年，AR/VR 产业规模将达到 约 1240 亿元

### 自动驾驶/无人机

到 2025 年，自动驾驶计算系统产业规模将达 3588 亿元

自动驾驶计算系统

| | |
|---|---|
| 环境感知 | CNN/RNN |
| 地图定位 | SLAM/GRU |
| 运动规划 | RL/LSTM |
| 控制决策 | RL/LSTM |

### 语音/图像


到 2025 年，智能语音/图像市场规模将达 约 3612 亿元

### 国家重大需求

航空航天、空间探索、国防工业等场景

### 海量工业制造场景

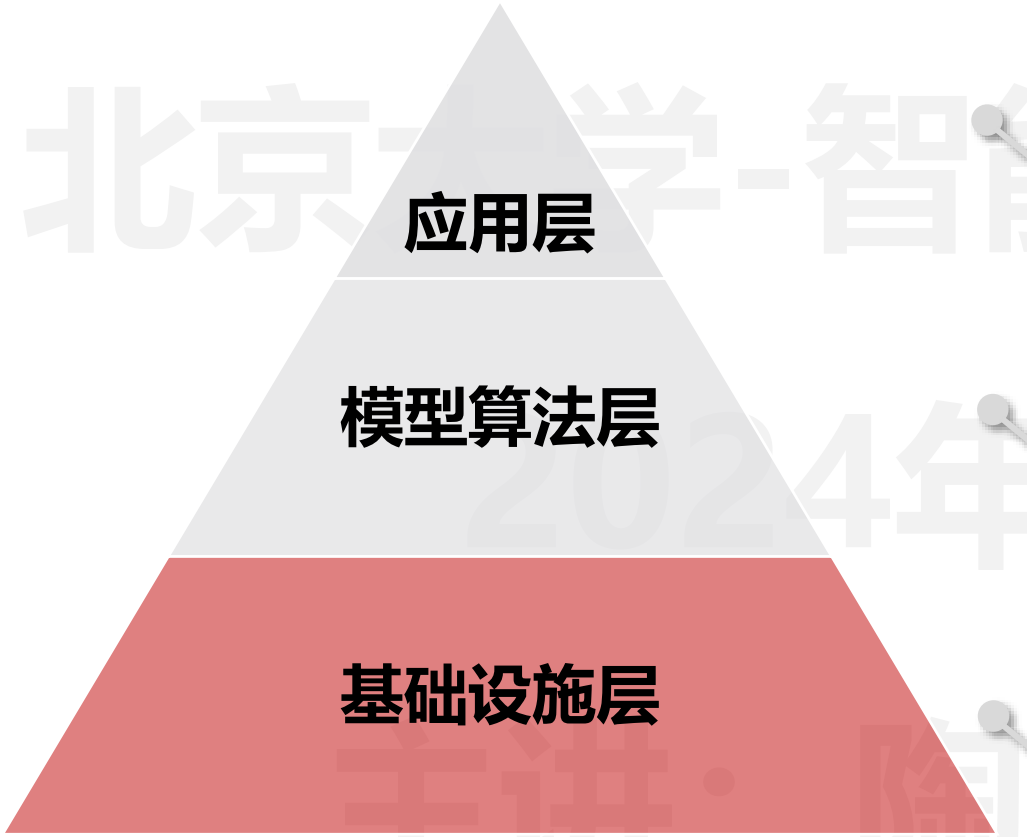
赋能电力、智能制造等工业场景，提升制造业自动化、智能化水平，产生数量级级别的生产效率提升

### 社会治理/经济管理

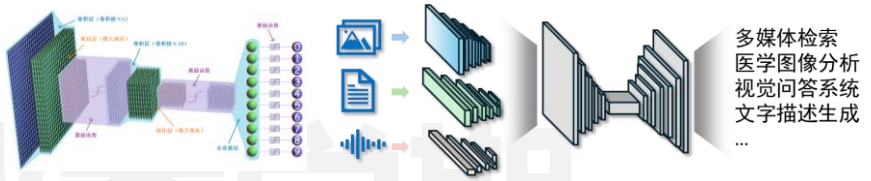
利用大模型提升社会治理水平和经济运行效率

# 人工智能硬件芯片

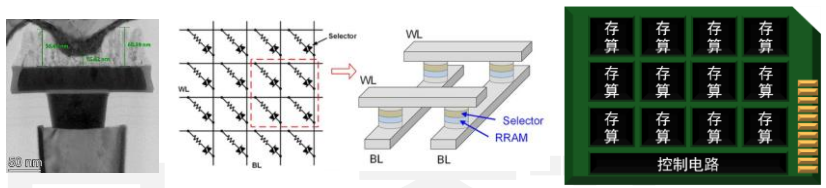
- 高性能芯片与硬件系统是不可或缺的算力基石




具身智能　自动驾驶　智慧医疗




机器学习模型　多模态大模型




大模型AI芯片与硬件系统(器件、电路、架构及编译器)

重要支撑

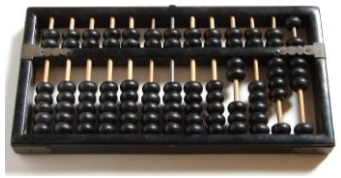
**人工智能竞争的一大核心：底层芯片与硬件的“军备竞赛”**

# 智能芯片的计算能力是未来新的生产力

- 数据是新的生产资料，计算能力是新的生产力，是支撑科技发展的源动力

古代

基于人力的算力



算盘

18世纪以前

机械化时代

基于机械的算力

机械式计算机

18世纪-20世纪初

电气化时代

基于电气元件的算力

电子管计算机

20世纪上半叶

电子信息时代

基于集成电路的算力

晶体管计算机

1960 – 2030

大数据智能时代

云边端多样化算力

新器件、新架构计算机

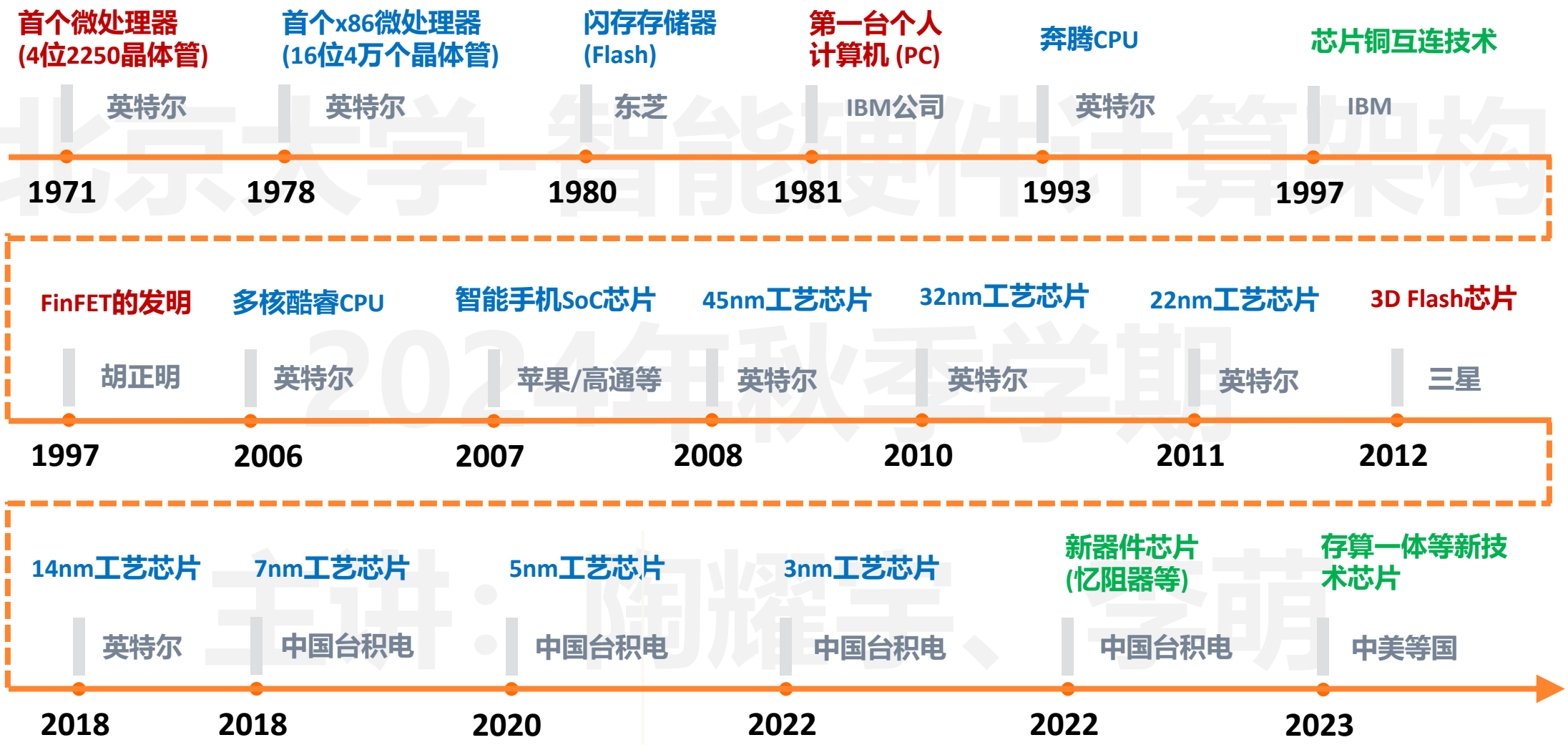
2030以后

# 波澜壮阔的智能芯片发展史

- 智能芯片的发展历史 (1833 - 1968)

# 波澜壮阔的智能芯片发展史

- 智能芯片的发展历史 (1968 - 2023)

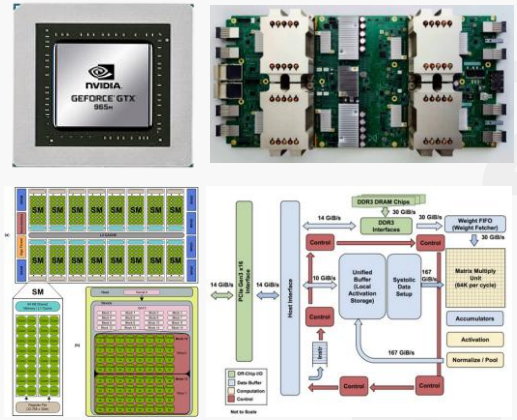


# 当前人工智能硬件芯片分类

- 高性能芯片与硬件系统是不可或缺的算力基石

## 当前AI芯片技术路线图及其发展现状

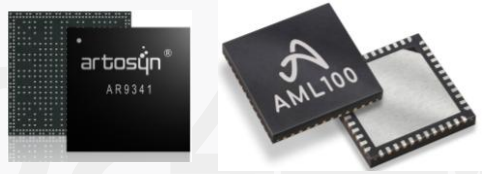
### 通用AI芯片

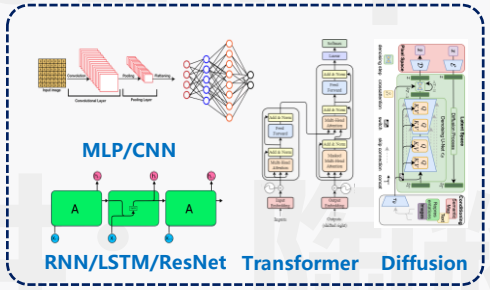



通用GPU　通用TPU/NPU等

- 通用计算指令集架构
- 高计算精度浮点数运算
- 规模算力大、编程性佳

### 定制AI芯片

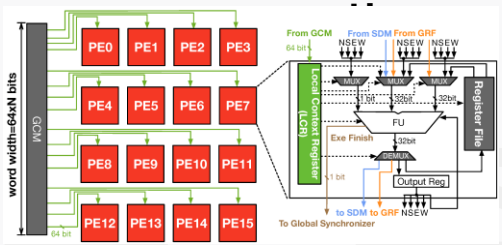



- 支持特定的AI模型
- 大部分定点数运算精度
- 中小规模、有限可编程性

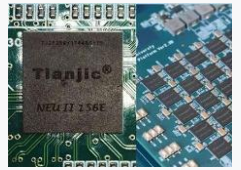
### 可重构AI加速芯片

FPGA加速卡　可重构阵列AI芯

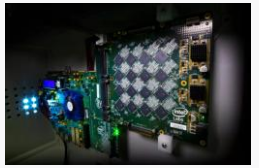



- 快速灵活、硬件可编程性
- 不受类型限制、吞吐高
- 即插即用、云边端均胜任

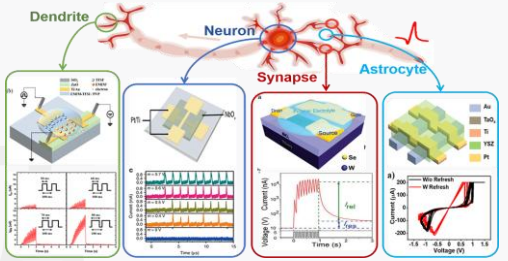
### 神经形态AI芯片



SNN芯片



神经形态芯片




- 电路器件模拟生物神经元
- 大量模拟神经元相连构成接近于人脑神经系统

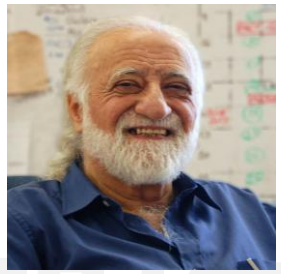
# 人工智能硬件芯片

- 学习人工智能硬件芯片是如何工作的至关重要


**硬件发展的历史关键人物**

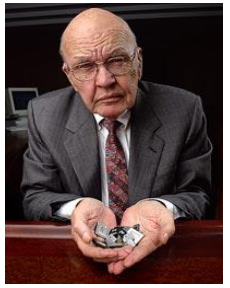
John L. Hennessy
Stony Brook/Stanford
美国科学院/工程院院士
图灵奖获得者

David Patterson
UCLA/UC Berkeley
美国科学院/工程院院士
图灵奖获得者

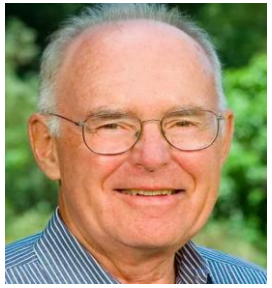


Yale Patt
UMich/UT Austin
美国工程院院士
富兰克林奖获得者

Robert Noyce
Intel创始人
美国国家技术奖章



Jack Kilby
集成电路发明人
诺贝奖获得者



Gordon Moore
美国工程院院士
美国总统自由勋章

胡正明 UC Berkeley
美国工程院院士、FinFet发明人
美国国家技术奖章

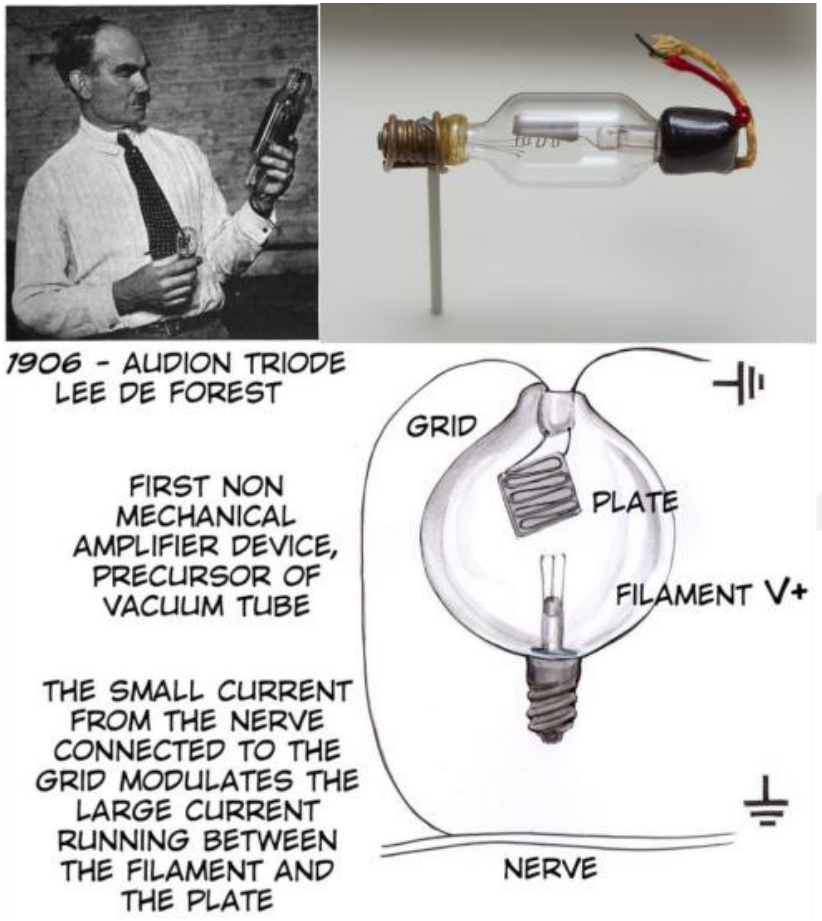
William Shockley、Walter Brattain、John Bardeen
半导体晶体管的3位发明人
均为美国科学院院士、诺贝奖获得者

# 人工智能硬件芯片 – 数模电路

- 学习人工智能硬件芯片是如何工作的至关重要

# 前半导体时代的器件霸主：真空三极管（电子管）- 1906年

- 佛雷斯特进一步再真空二极管中加入了栅极，提供额外电场调控阴极热电子向阳极运动的行为

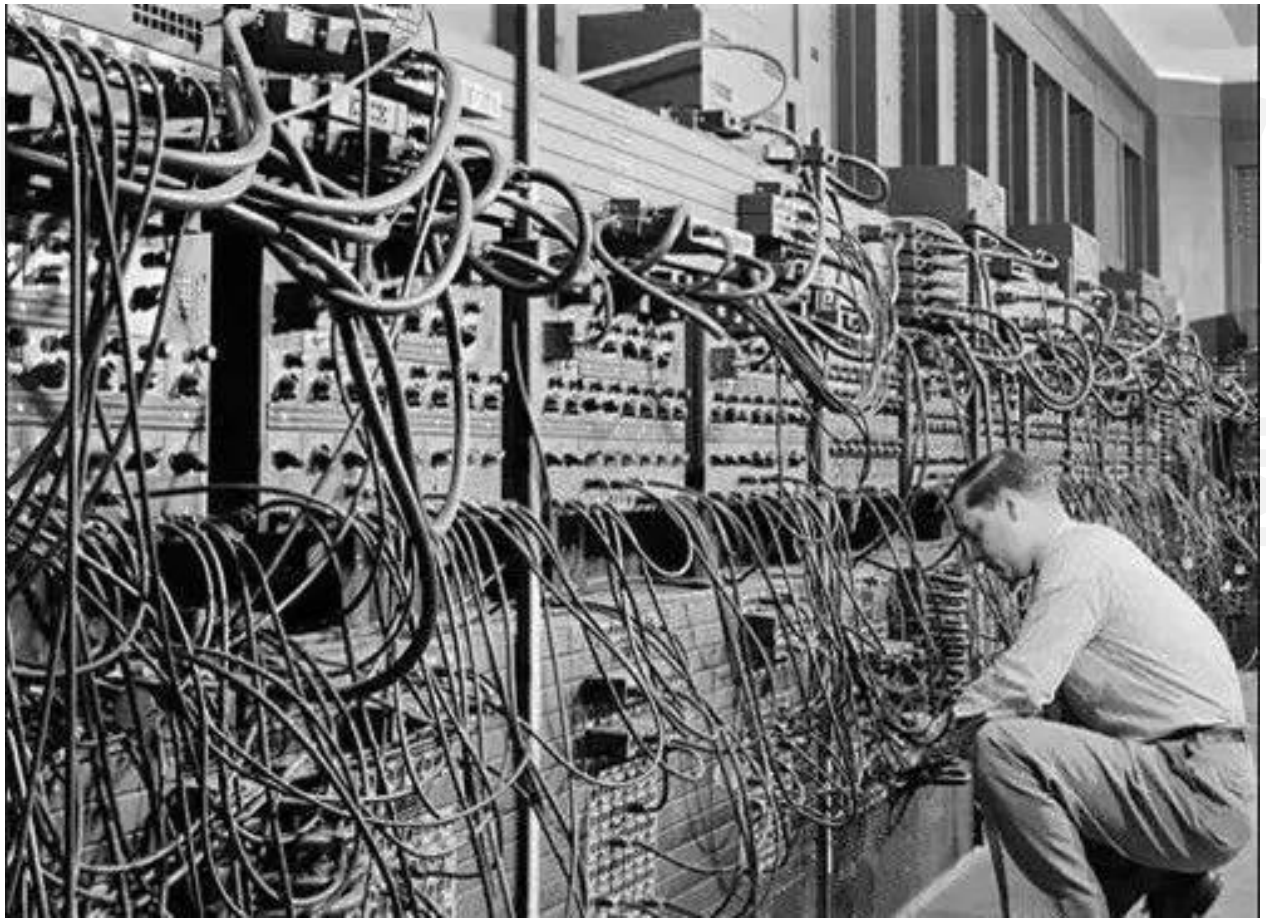



美国电子管计算机ENIAC

1906年，佛雷斯特进一步再真空二极管中加入了栅极，提供额外电场调控阴极热电子向阳极运动的行为，栅极电压就可以调控阴极的发射电流。这种新型真空管被称为三极管。三极管具备了检波、放大和振荡的功能，其应用场景被大大扩展，并促使了第一台现代意义的电子计算机埃尼阿克的诞生。

# 电子管的发展瓶颈 – 二十世纪中叶开始

- 消失的电子管 – 根本原因是体积过大无法大规模集成、寿命较低难以长时间工作



基于热电子发射的真空管寿命较短、功耗高、体积大、成本高。埃尼阿克有一半的机时都浪费在检修损坏的真空管上，这导致它难以长时间地处理复杂的计算任务。

# 电子管的发展瓶颈 – 前苏联/俄罗斯半导体芯片产业发展的教训

- 前苏联的计算机起步与美国几乎同时代，但在**电子管与晶体管的路线选择**上出现重大失误

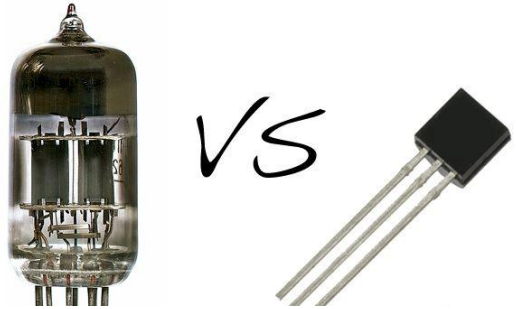
**美国电子管计算机ENIAC**

重达 30 吨，占地 170 平米

每分钟能执行 5000 次运算

**前苏联电子管计算机MESM**

6000 个电子管每分钟3000次运算，算力稍弱，但耐用和省电上有一些优势




**前苏联选择把主要精力放在了电子管的小型化上，在半导体晶体管时代逐渐落后**

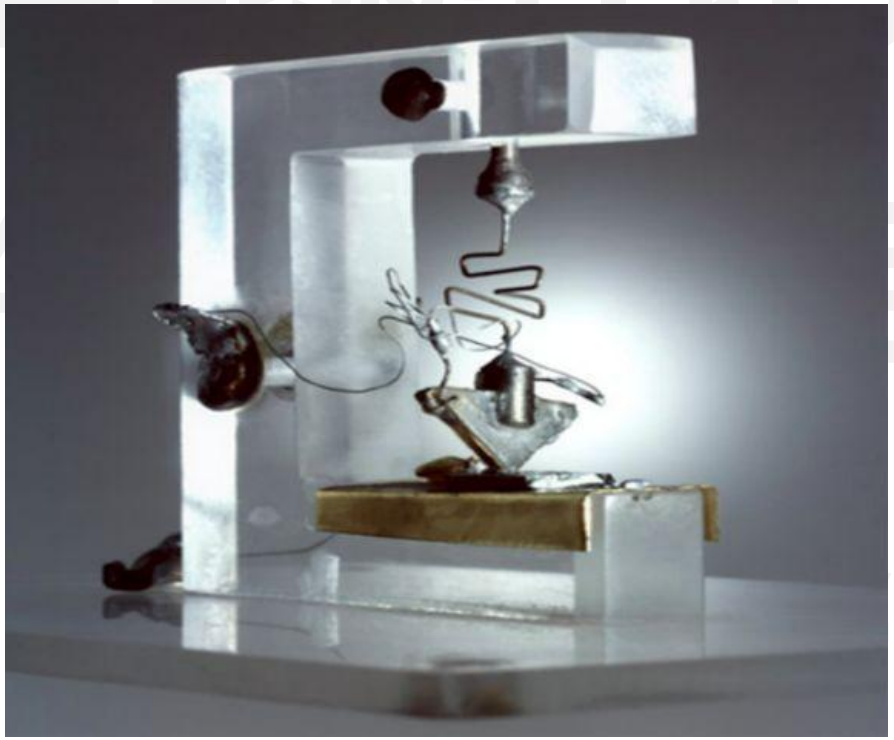
**电子管在特定的军事应用领域与国防工业中仍具有一定的作用**

**在电子管小型化方面，俄罗斯的实力在目前世界是最强的。俄罗斯S300/400等防空导弹系统极强的抗干扰能力，其实就来源于前苏联/俄罗斯的电子管小型化技术**

# 半导体锗晶体管的发明 – 1947年

- 半导体晶体管被誉为“21世纪最伟大的发明”，深刻的改变了人类历史发展进程

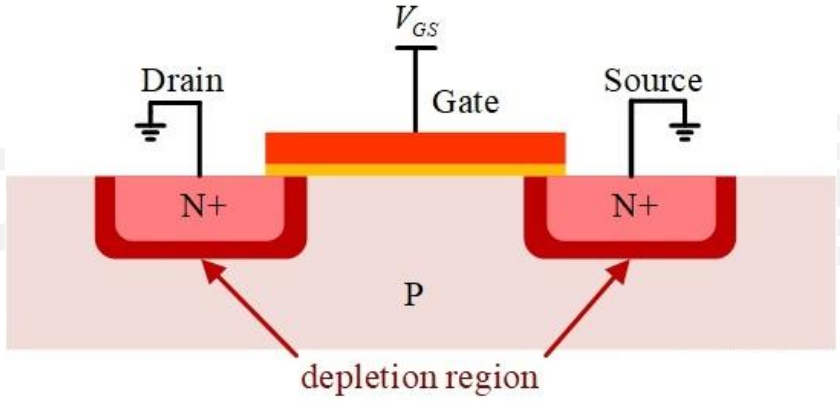
肖克利（前）、巴丁（后一）、布拉顿（后二），半导体锗晶体管的发明，共同获得了1956年的诺贝尔物理学奖



点接触式晶体管：把间距为50 μm的两个金电极压在锗半导体上，微小的电信号由一个金电极 (发射极) 进入锗半导体 (基极) 并被显著放大，然后通过另一个金电极 (集电极) 输出，这个器件在1kHz的增益为4.5

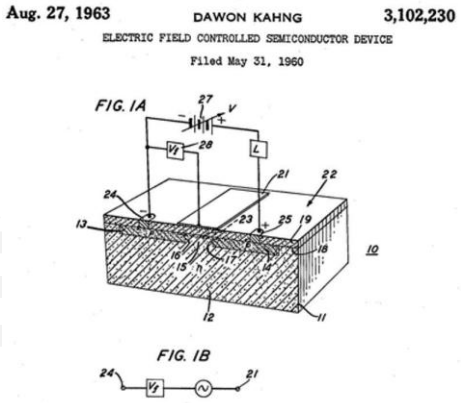
# 硅基MOSFET是支撑现代芯片的器件基石

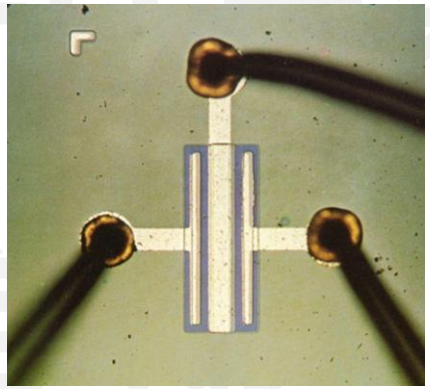
- 艾塔拉 (Martin Atalla) 和姜大元 (Dawon Kahng) 共同发明了硅基MOSFET场效应晶体管



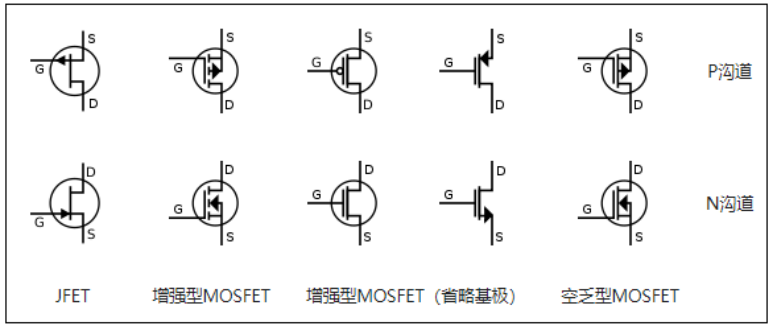

艾塔拉 (Martin Atalla) 和姜大元 (Dawon Kahng)

**Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET)**


Aug. 27, 1963 DAWON KAHNG 3,102,230
ELECTRIC FIELD CONTROLLED SEMICONDUCTOR DEVICE
Filed May 31, 1960
FIG. IA
FIG. IB



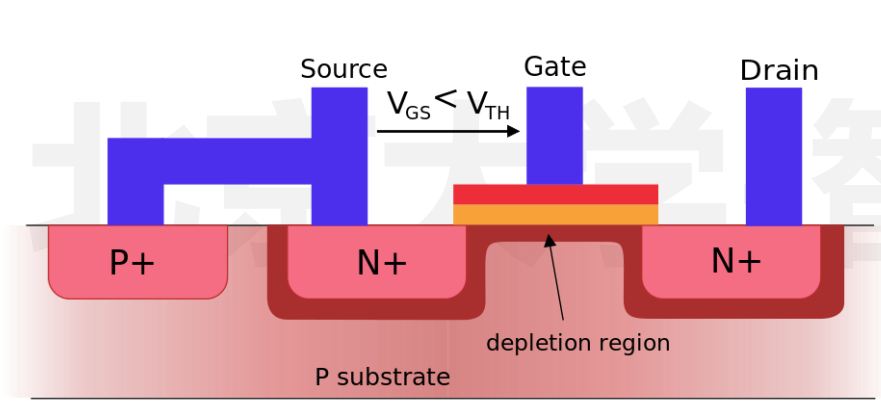
PMOS场效应晶体管实物图

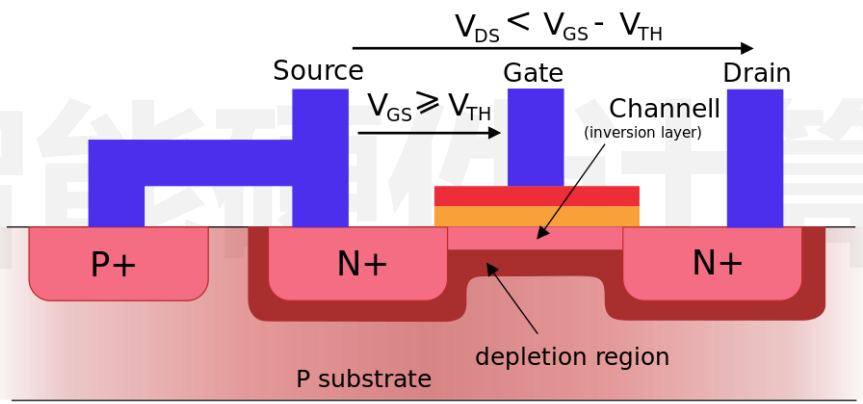



MOSFET已经成为集成电路的基本组成单元

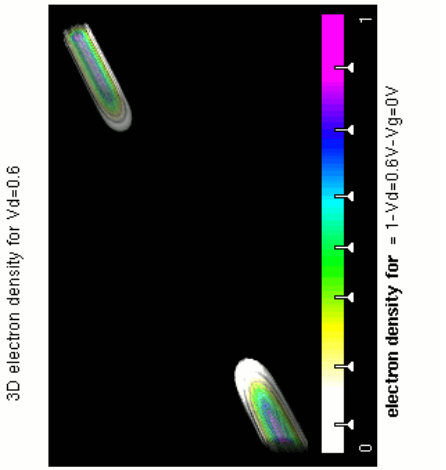
# MOSFET晶体管工作原理 – 可控开关

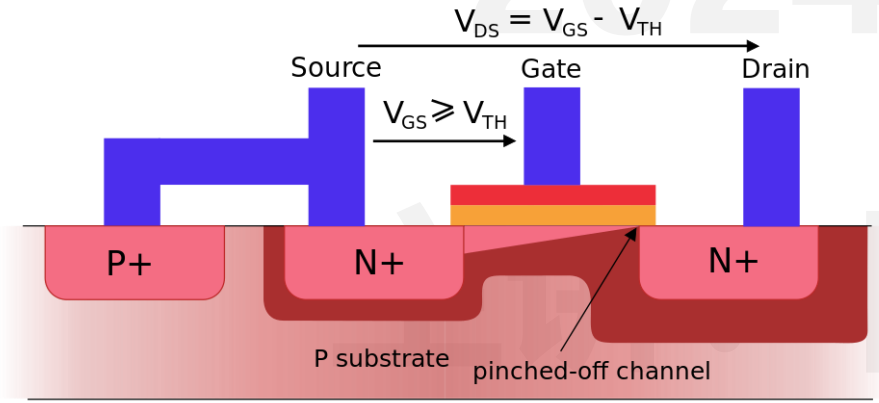
- MOSFET有三个工作区间：断开、线性（欧姆区间）、饱和（电压不随电流线性增加）



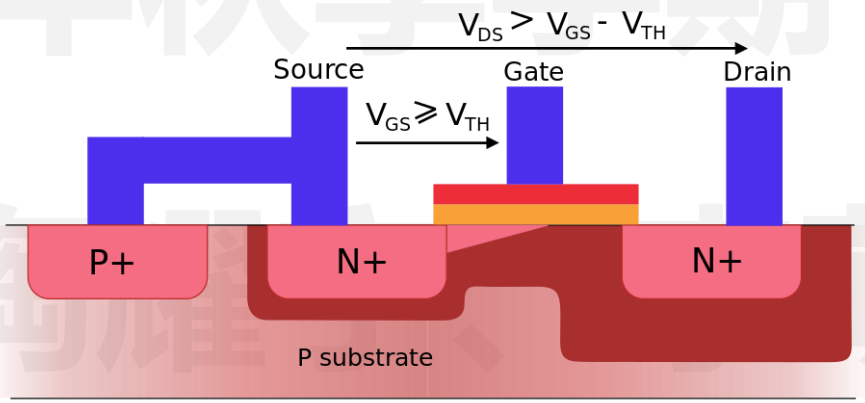




Linear operating region (ohmic mode)

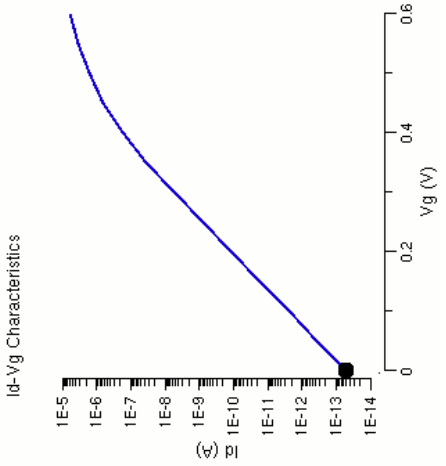



Saturation mode at point of pinch-off

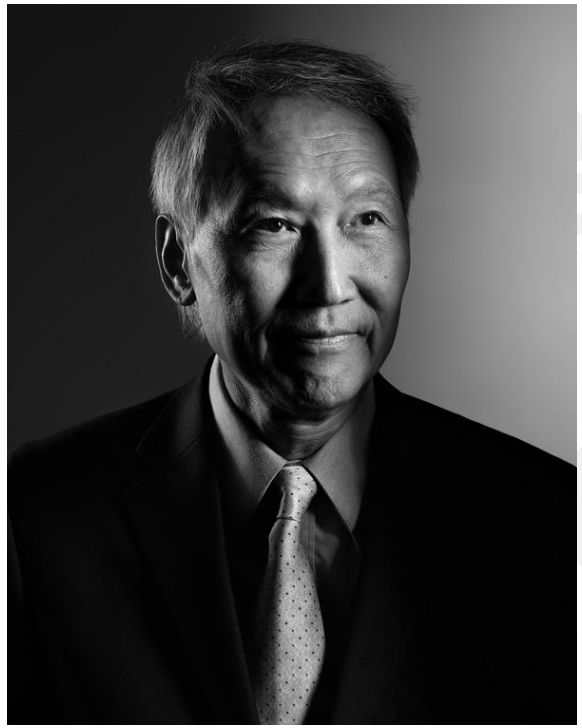



Saturation mode

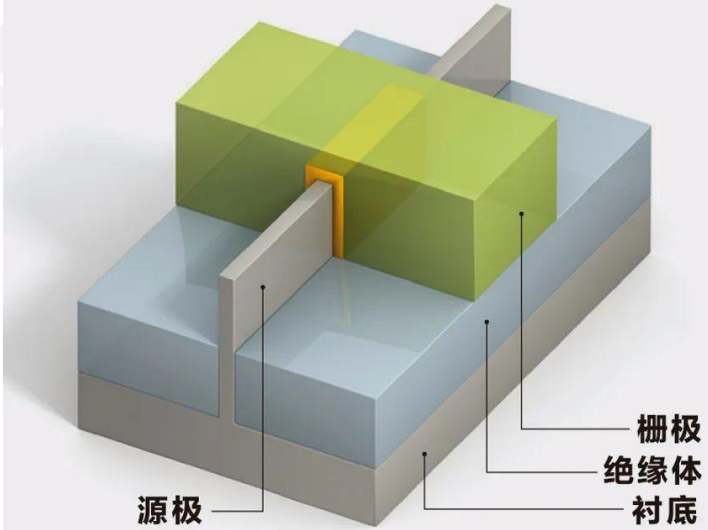


# 鳍式三维晶体管FinFET – 1999年

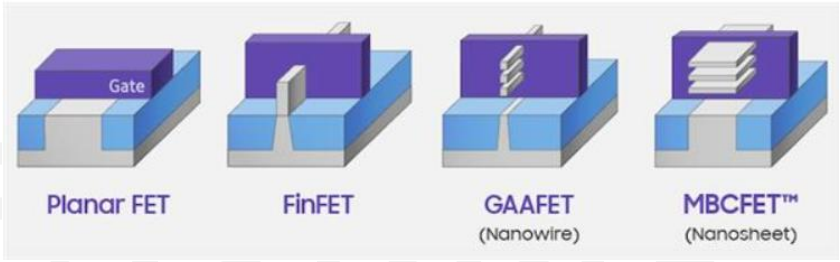
- 原本预计2010年后传统MOSFET在20nm走到尽头，胡正明的发明进一步推进制程缩小

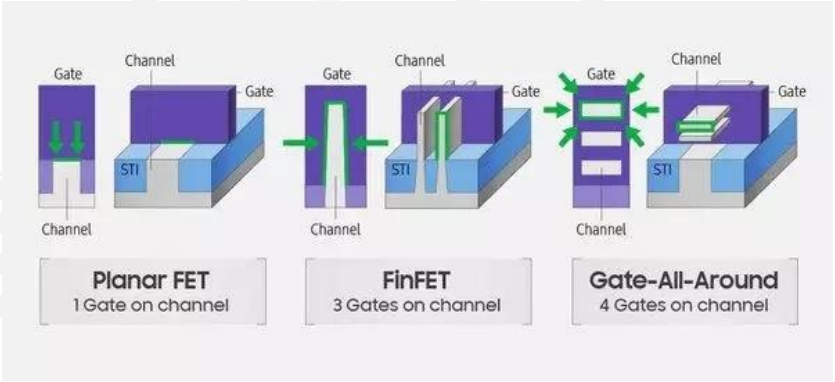


**加州大学伯克利分校的胡正明教授（IEEE Fellow，美国工程院院士，中国科学院外籍院士）**




**FinFET的特点：**每个晶体管包含一个源极、一个漏极、一个连接两者的导电通道和一个控制电流沿通道流动的栅极。在FinFET中，将通道抬高，如同鲨鱼鳍高出芯片表面，使得栅极三面环绕，从而给予栅极更大的控制力。



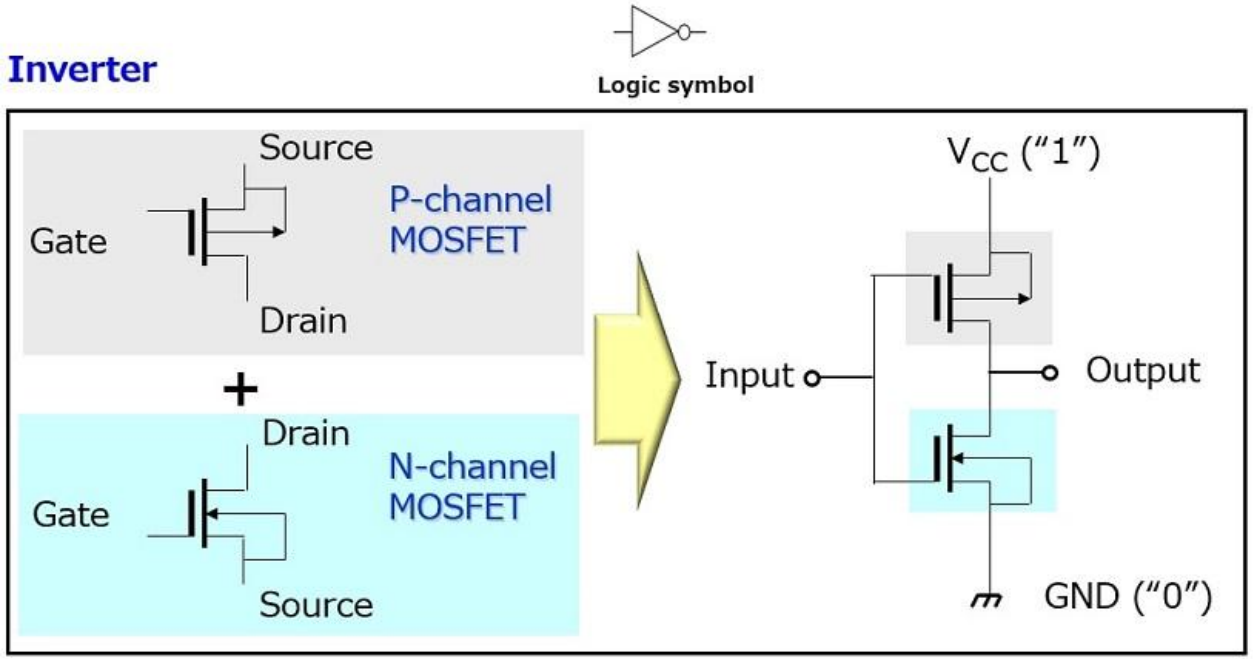

**由FinFET演化出多种三维晶体管构型，推动制程向3nm/1nm演进**

# 人工智能硬件芯片 – 逻辑电路

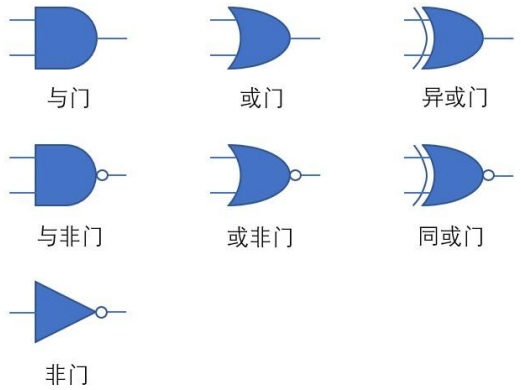
- 学习人工智能硬件芯片是如何工作的至关重要

# CMOS数字逻辑电路 – 当前芯片的主流电路技术方案

- 仙童半导体于1963年首次发明互补金属氧化物半导体 (Complementary Metal Oxide Sem.)




CMOS非门电路图




互补式金属氧化物半导体具有只有在晶体管需要切换启动与关闭时才需消耗能量的优点，因此非常节省电力且发热量少，且工艺上也是最基础而最常用的半导体器件

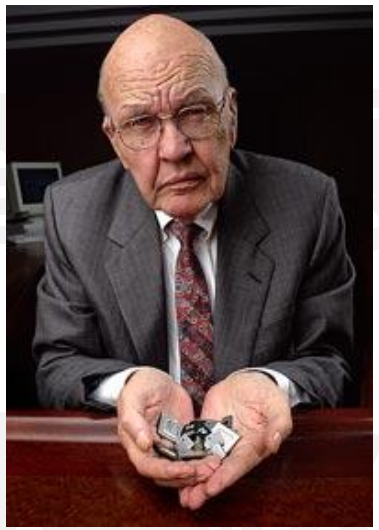
硅质晶圆模板上制出NMOS（n-type MOSFET）和PMOS（p-type MOSFET）的基本器件，由于NMOS与PMOS在物理特性上为互补性，因此被称为CMOS

# 重要历史节点：集成电路的发明 – 1958年/1959年

- 德州仪器公司的工程师基尔比 (Jack Kilby) 发明了第一块集成电路

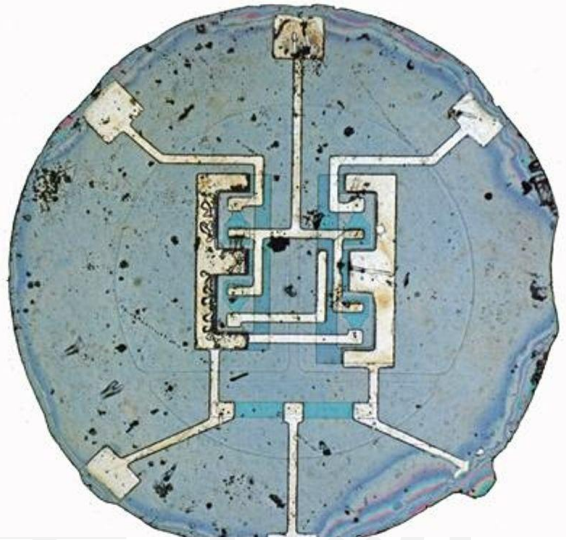
1958年8月28日世界第一块集成电路

尺寸7/16×1/16英寸



基尔比获2000年

诺贝尔奖



罗伯特-诺伊斯于1959年8月发明第一块

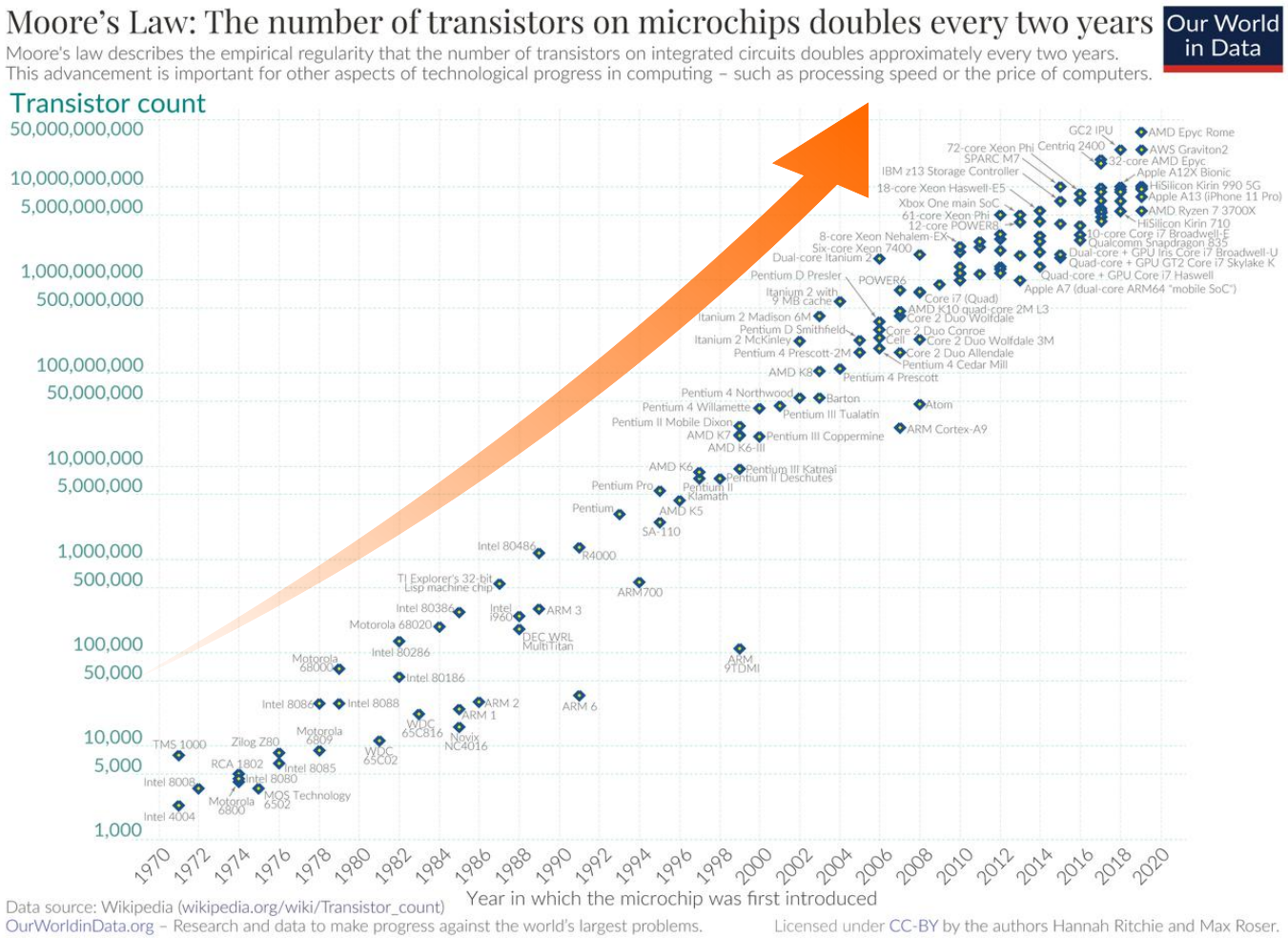
硅集成电路

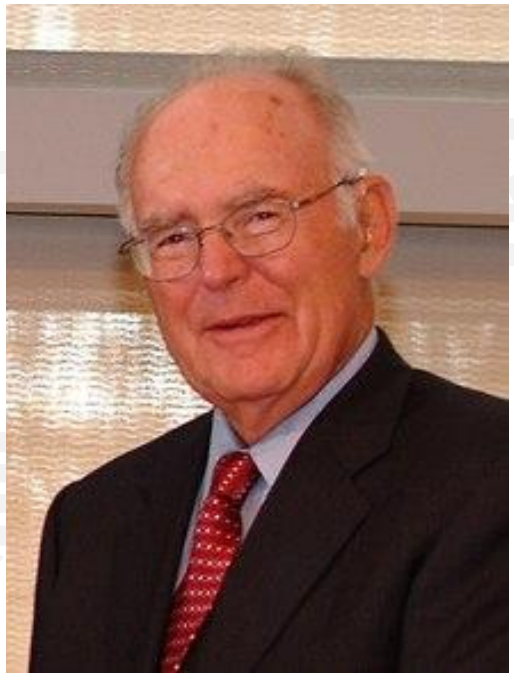
将包括锗晶体管在内的五个元器件集成在一起，基于锗材料制作了一个叫做相移振荡器的简易集成电路

参与创立仙童半导体 (Fairchild) 和英特尔 (Intel) 公司，奠定了硅谷的基石

# 重要历史节点：摩尔定律的提出 – 1964年

- 仙童半导体/英特尔的联合创始人戈登摩尔提出了著名的“摩尔定律”



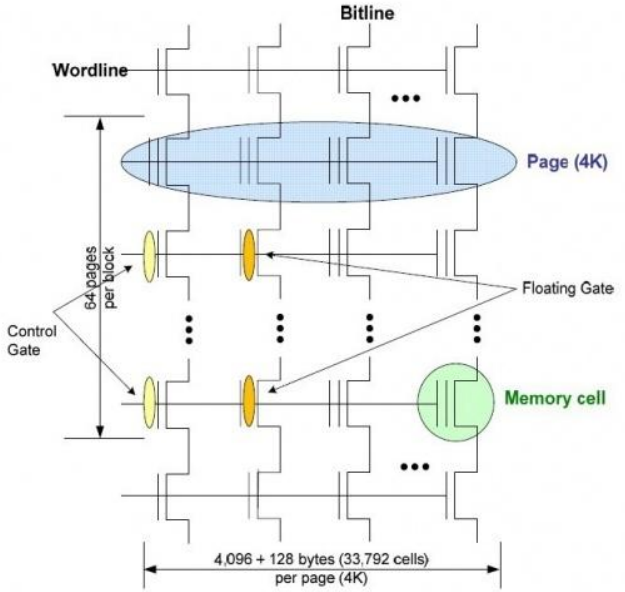
戈登·摩尔

**集成电路上可容纳的晶体管数目，**

**每隔两年便会增加一倍**

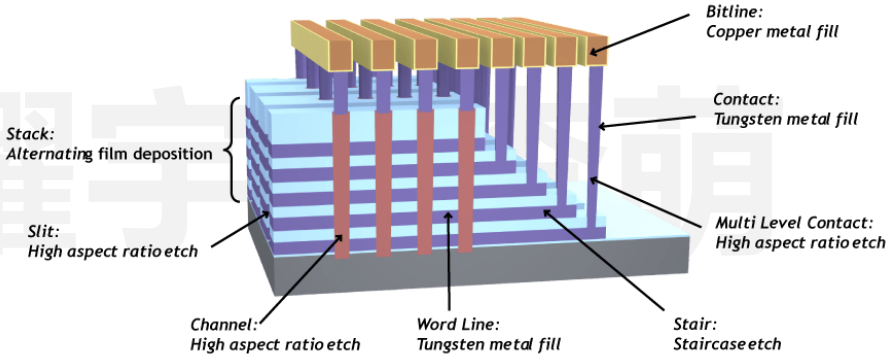
# 非易失性存储器Flash的发明 – 1967年

- 除了计算场景之外，存储也是占据智能芯片重要份额的典型应用场景


**传统平面型NAND Flash非易失性存储器**

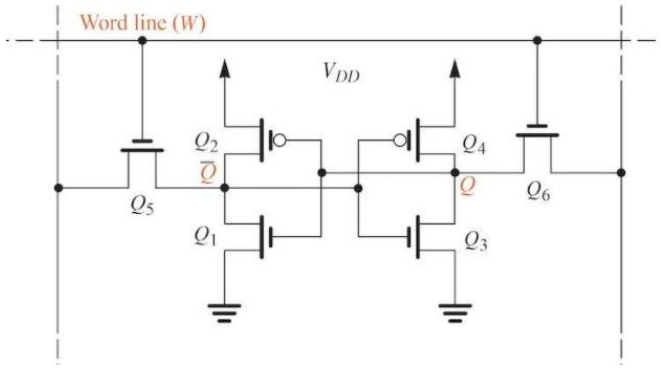
**Dawon Kahng (韩) 和 Simon Sze (华裔) 在贝尔实验室发明了非易失性存储器浮动门（Floating Gate ）本文发表为“A Floating Gate and Its Application to Memory Devices” （贝尔系统技术期刊）**

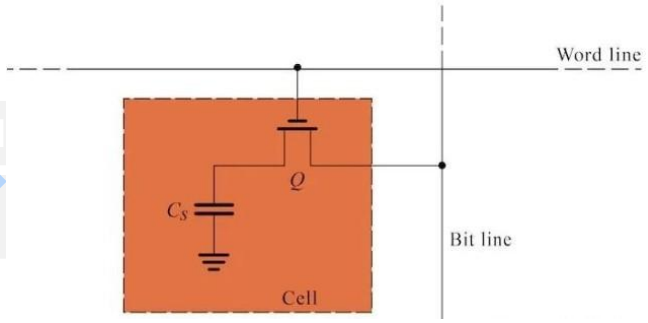


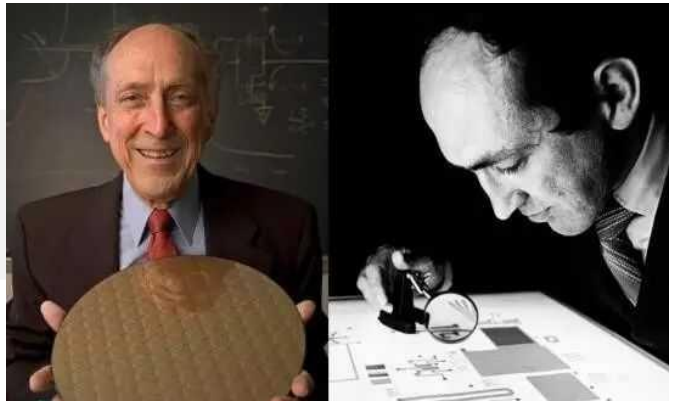

**现代三维堆叠 NAND Flash**

# 易失性存储器DRAM的发明 – 1968年

- SRAM/DRAM是两种最常用的易失性存储器件，广泛应用于现代智能芯片中



SRAM需要6个CMOS晶体管来存储数据

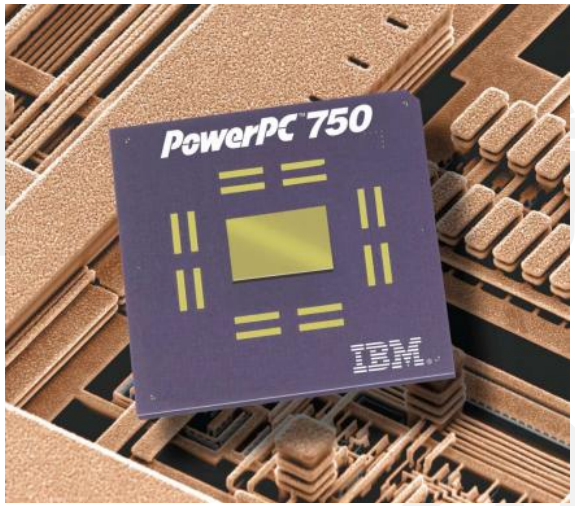
罗伯特·丹纳德发明了DRAM (**动态随机存取存储器**) 存储器

SRAM（**静态随机存取存储器**）的优点是它的速度快，它的存取速度比DRAM（动态随机存取存储器）快得多，因为它不需要每次访问数据都要重新刷新电容。

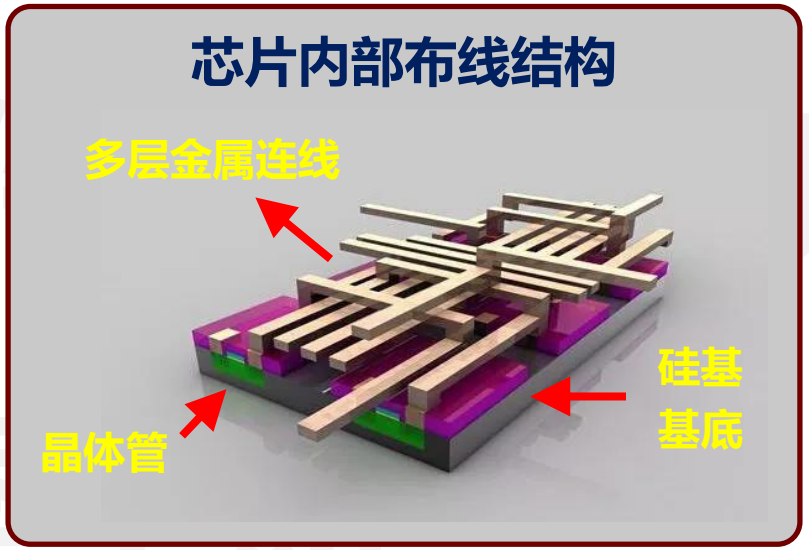
与SRAM相比，DRAM的优势在于结构简单—**每比特都只需一个电容跟一个晶体管来处理**，相比之下在SRAM上一个比特通常需要六个晶体管。 正因这缘故，**DRAM拥有非常高的密度，单位体积的容量较高因此成本较低**。 但相反的，DRAM也有访问速度较慢，耗电量较大的缺点。

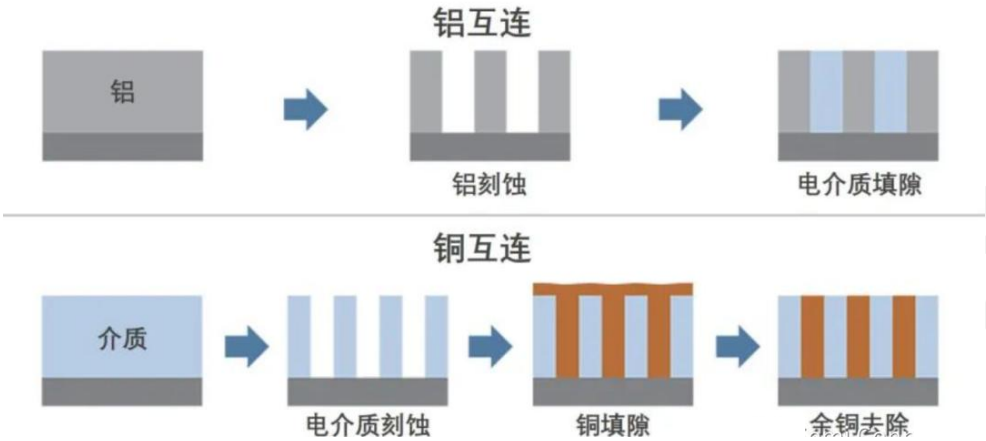
# 电路间的互连方式：智能芯片的铜互连技术 – 1997年

- IBM率先从铝互连转向铜互连，并推出了第一个铜基微处理器 IBM PowerPC 750




IBM PowerPC 750 最初是采用铝设计的，其工作频率高达300 MHz，采用铜互连之后，同一芯片的速度至少能达到400MHz，提高了33%




集成电路金属互连线制造工艺达到纳米级后，因为超高纯铜具有更佳的电阻率和抗电迁徙能力，很快高纯铜就替代超高纯铝合金成为金属互连线的主要材料

# 人工智能硬件芯片 – 计算架构

- 学习人工智能硬件芯片是如何工作的至关重要

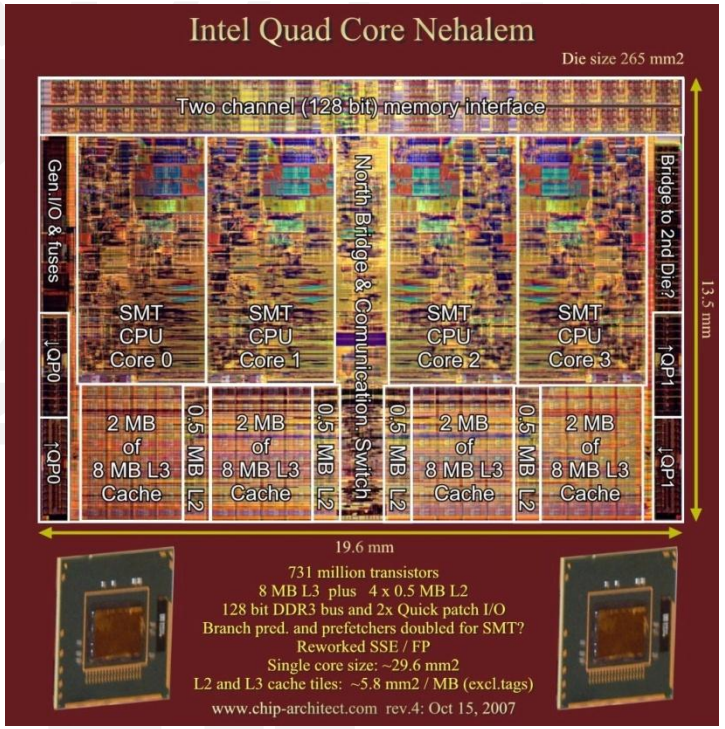
# 什么是硬件计算架构?

- 硬件计算架构这一概念的随着现代计算机的出现而出现，由Amdahl首次提出

"The term *architecture* is used here to *describe the attributes of a system as seen by the programmer*, i.e., the conceptual structure and functional behavior as distinct from the organization of the dataflow and controls, the logic design, and the physical implementation."

***Gene Amdahl*, IBM Journal of R&D, April 1964**

**吉恩•阿姆达尔：IBM大型机之父**




从1956年的达特茅斯会议开始，人工智能（Artificial Intelligence，AI）作为一个专门的研究领域出现

***硬件计算架构作为一个独立研究领域的出现，甚至晚于人工智能***

***随着人工智能的爆发式发展，硬件计算架构迎来大幅增长***

# 为什么要学习硬件计算架构?

- 硬件计算架构是为了解决上世纪60年代出现的实际工程问题 – 如何链接多种算法与单一硬件?

## IBM Compatibility Problem in Early 1960s

By early 1960's, *IBM had 4 incompatible lines of computers*

701 ➡ 7094
650 ➡ 7074
702 ➡ 7080
1401 ➡ 7010

Each system had its own:

- Instruction set architecture (ISA)
- I/O system and Secondary Storage: magnetic tapes, drums and disks
- Assemblers, compilers, libraries,...
- Market niche: business, scientific, real time, ...

Stanford

海量不同
计算任务

同一块
硬件

# 将计算任务分解为图灵机可运行的离散化操作

- 图灵计算理论催生出以图灵机为理论支撑的现代智能芯片体系结构


| 纸带符号 | 状态A | | | 状态B | | |
|---|---|---|---|---|---|---|
| | 写 | 移动 | 状态 | 写 | 移动 | 状态 |
| 0 | 1 | 右 | B | 1 | 左 | A |
| 1 | 1 | 左 | A | 0 | 右 | B |

***模拟人们用纸笔进行数学运算的过程***

- **纸带：**一条无限长的纸带（**TAPE**），被划分为一个接一个的小格子，每个格子上包含一个来自有限字母表的符号
- **笔：**一个读写头（**HEAD**），可以在纸带上左右移动，能读出当前所指的格子上的符号，并能通过写操作改变它
- **运算法则：**一套规则（**TABLE**），根据当前状态及当前读写头所指格子上的符号来确定读写头下一步的动作
- **状态：**一个状态寄存器堆栈（**STATE**），保存图灵机当前的状态

# 图灵计算理论

北京大学 PEKING UNIVERSITY

- **图灵机的数学理论框架由一个七元有序组定义**

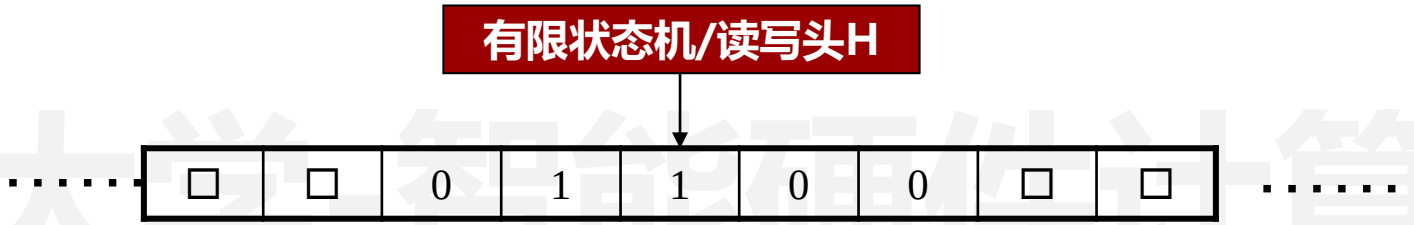
一台图灵机可被定义为 $T = \{Q, \Sigma, \Gamma, q_0, q_{accept}, q_{reject}, \delta(q,s)\}$

- **Q**：是非空有限状态集合
- **Σ**：非空有限输入符号表，其中特殊空白符 □ ∉ Σ
- **Γ**：非空有限带符号且 Σ⊂ Γ，空白符 □ ∈ Γ- Σ，也是唯一允许出现无限次的字符
- $q_0 \in Q$ 表示图灵机起始状态
- $q_{accept} \in Q$ 表示接受状态
- $q_{reject} \in Q$ 表示拒绝状态, 且$q_{reject} \neq q_{accept}$
- **$\delta(q,s)$:** $Q \times \Gamma \rightarrow Q \times \Gamma \times \{L, R\}$ 是转移函数， 根据当前读入符号s和当前状态q决定下一个状态、写入的符号、纸带移动方向和距离，L, R表示读写头是向左移还是向右移, - 表示不移动

# 图灵计算理论

- 图灵机的计算方式与工作流程

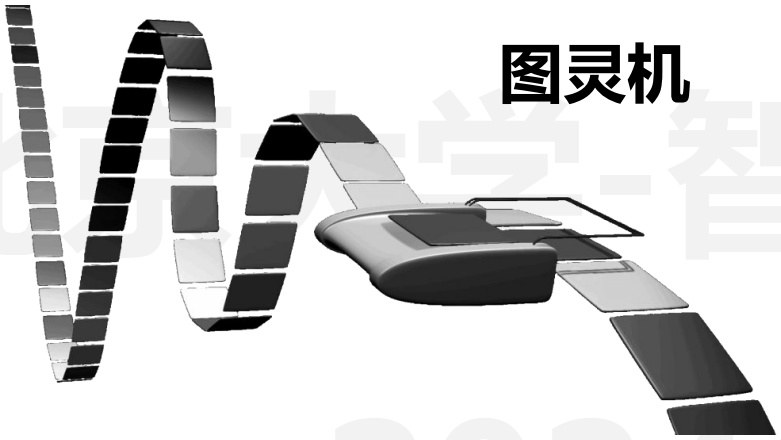
$T = \{Q, \Sigma, \Gamma, q_0, q_{accept}, q_{reject}, \delta(q,s)\}$




- **初始状态：**将输入符号串 $\omega = \omega_0\omega_1 \ldots \omega_{n-1} \in \Sigma^*$ ⟹ 纸带第0,1, ..., n-1 号格子
  - 读写头H指向0号格子，T @ $q_0$状态
- **运行方式：**T按照转移函数所描述的规则进行计算
  - T@$q$状态，H = $x$，设 $\delta(q, x) = (q', x', L)$
    - T→$q'$ ，H→$x'$，读写头左移一格
    - 若某时刻H指向0号格子，但根据 $\delta(q, x)$ 将继续左移，则T原地不动
- **停机情况：** 1）若某时刻T @ $q_{accept}$或$q_{reject}$，T停机，并接受或拒接$\omega$；
  2）$\delta$(q,s)对某些q和s可能无定义，T停机

# 由图灵计算理论衍生出的冯诺依曼体系结构

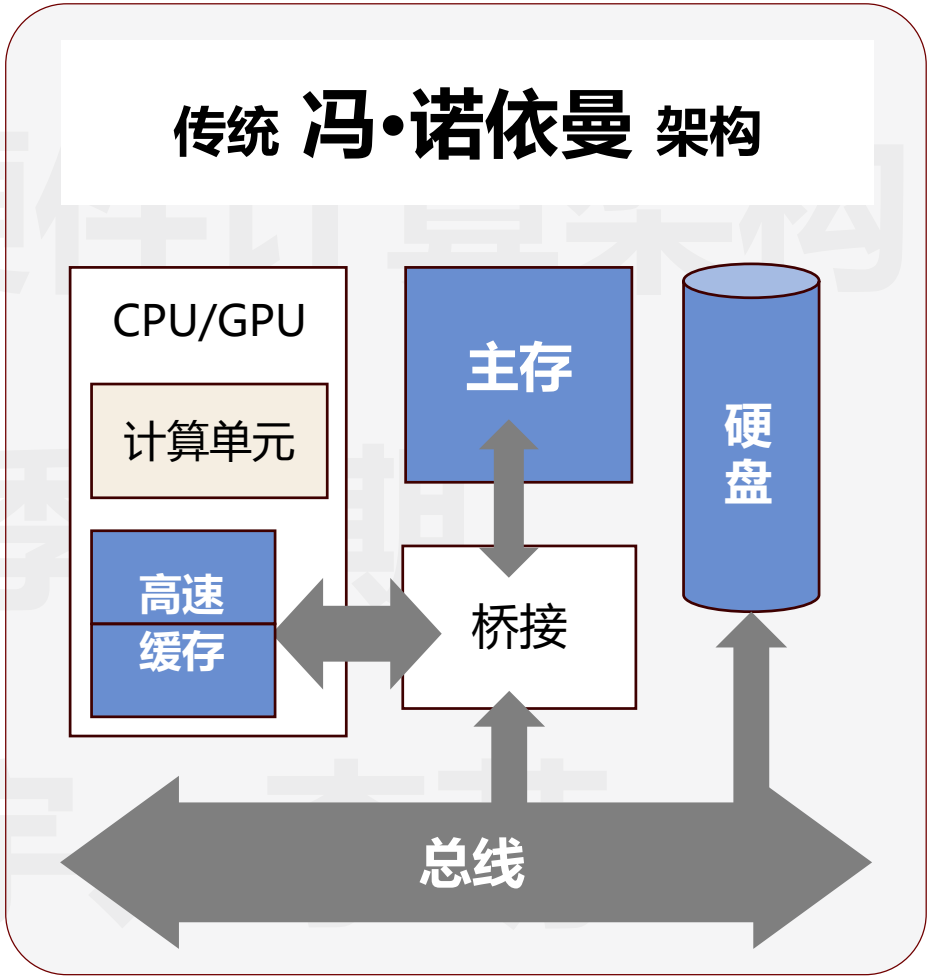
- 图灵机计算范式中的元素可在冯诺依曼架构中找到对应




笔 & 状态 → 数据存储器

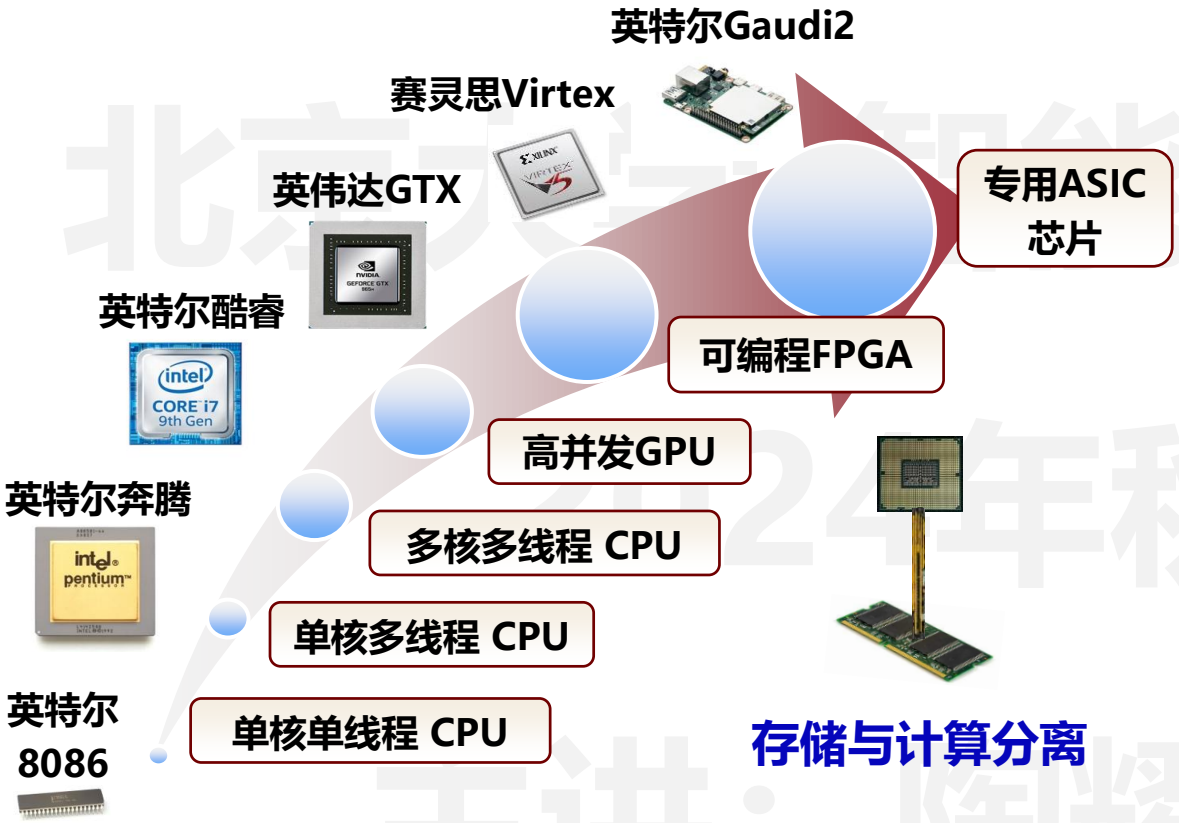
纸带 → 指令存储器

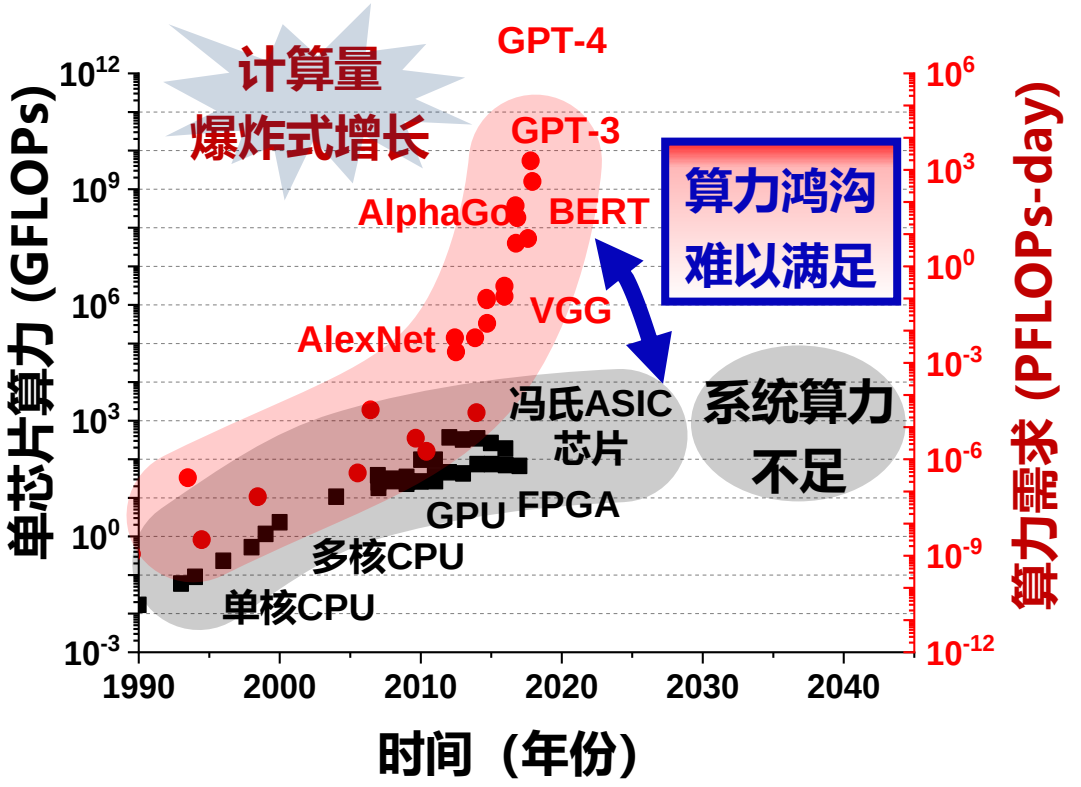
运算法则 → 控制逻辑&计算单元



# 冯诺依曼体系结构

- 从偏向通用计算任务的CPU、GPU到偏向定制化设计的FPGA、ASIC演进



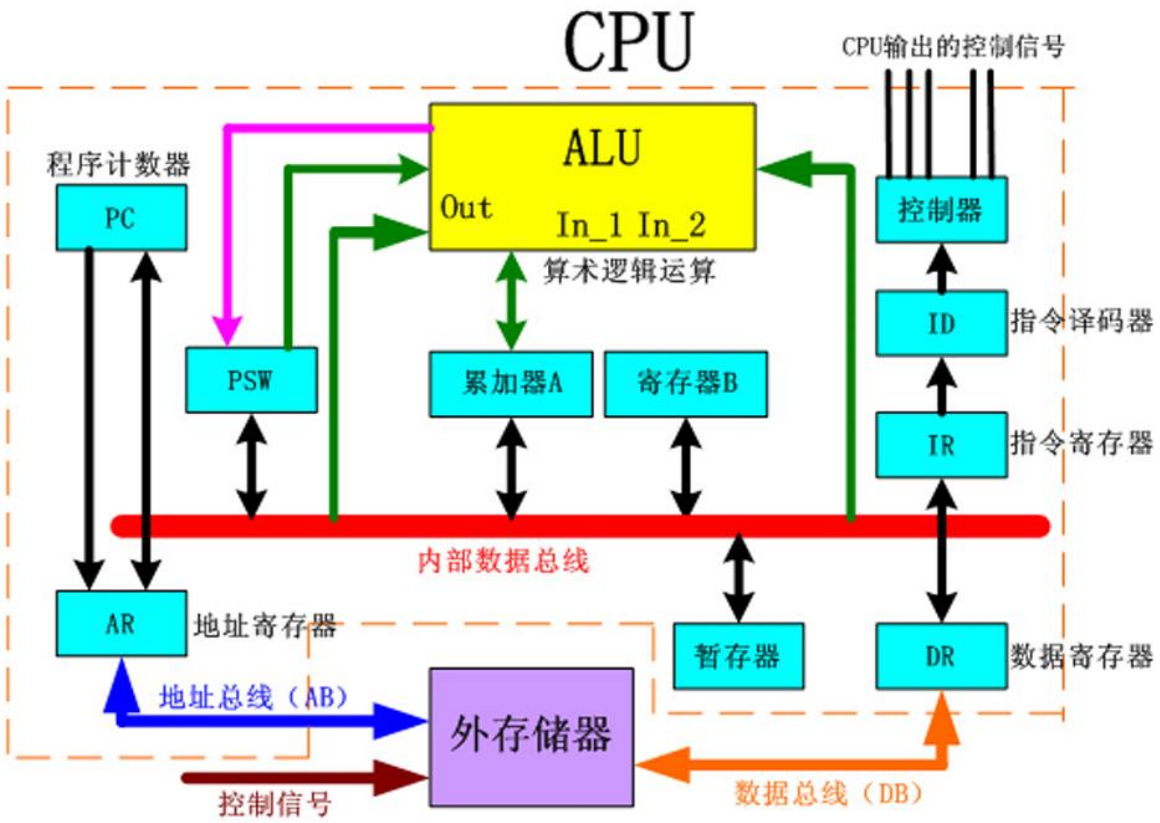



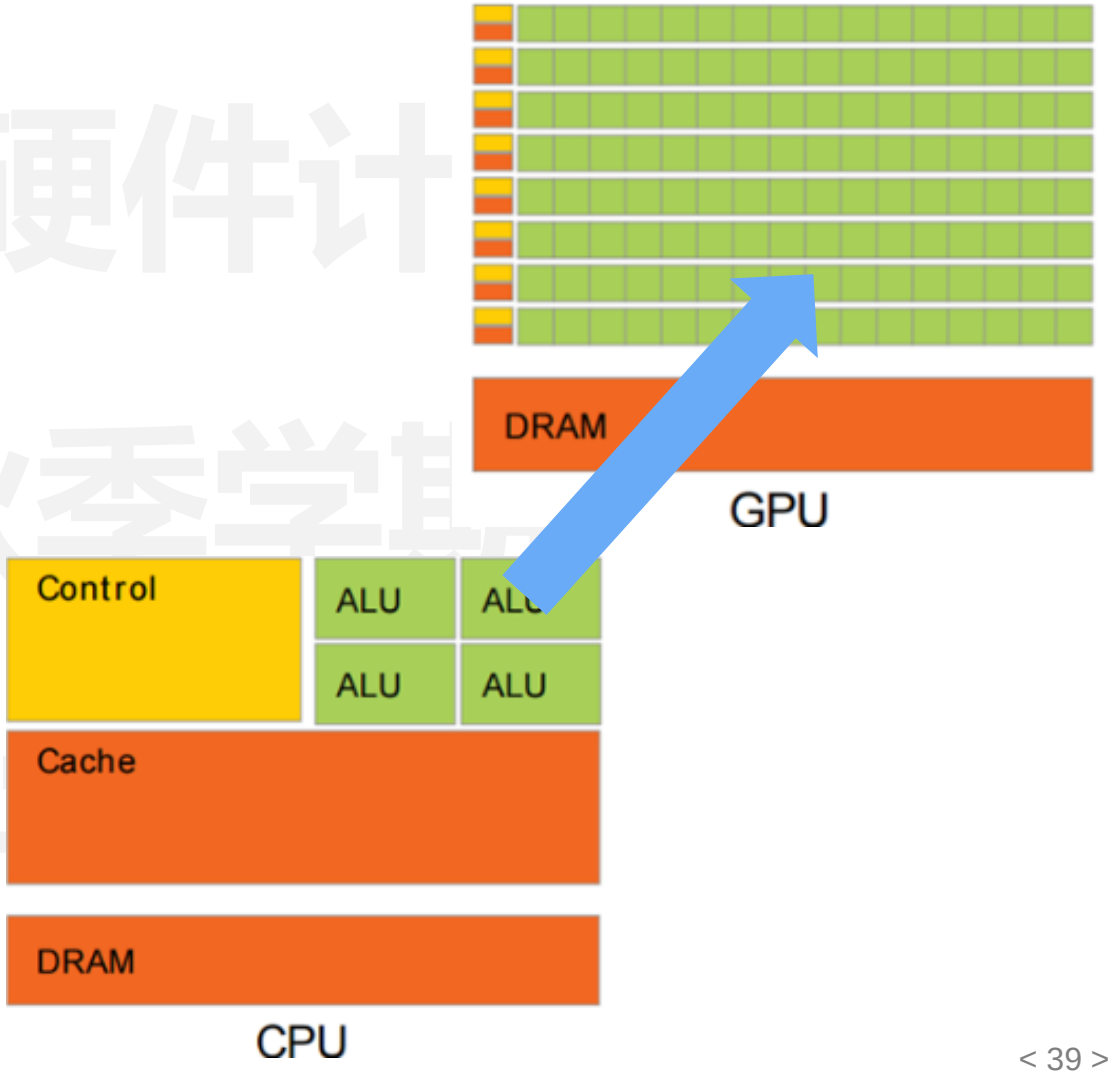

智能芯片体系结构演进图

新兴计算任务所需的计算量呈爆炸式增长

# 典型智能芯片体系结构 – CPU/GPU

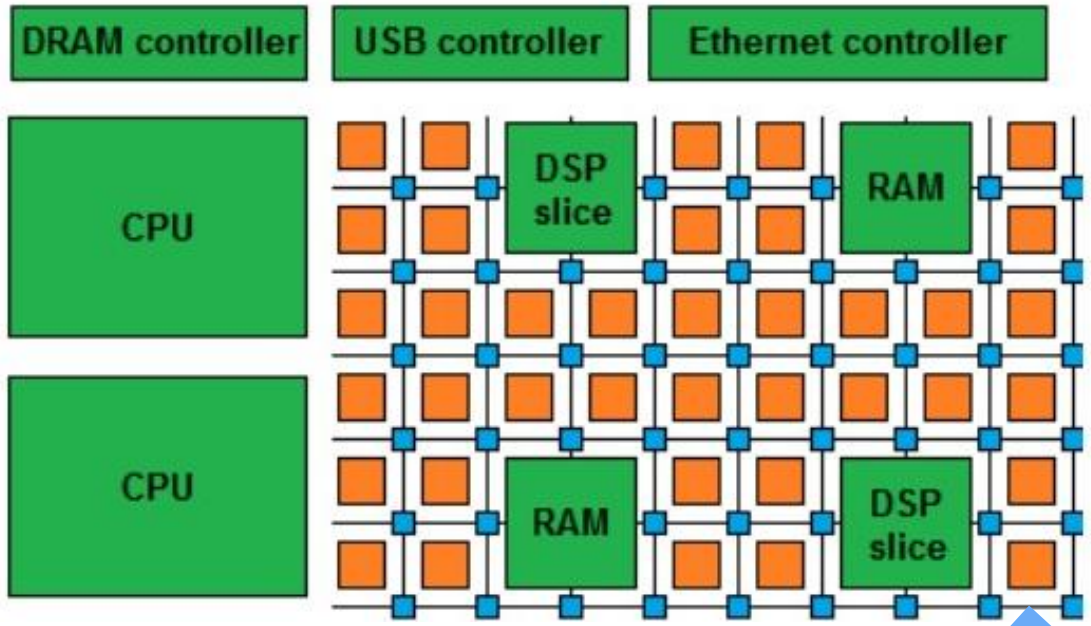
- 从偏向通用计算任务的CPU、GPU到偏向定制化设计的FPGA、ASIC



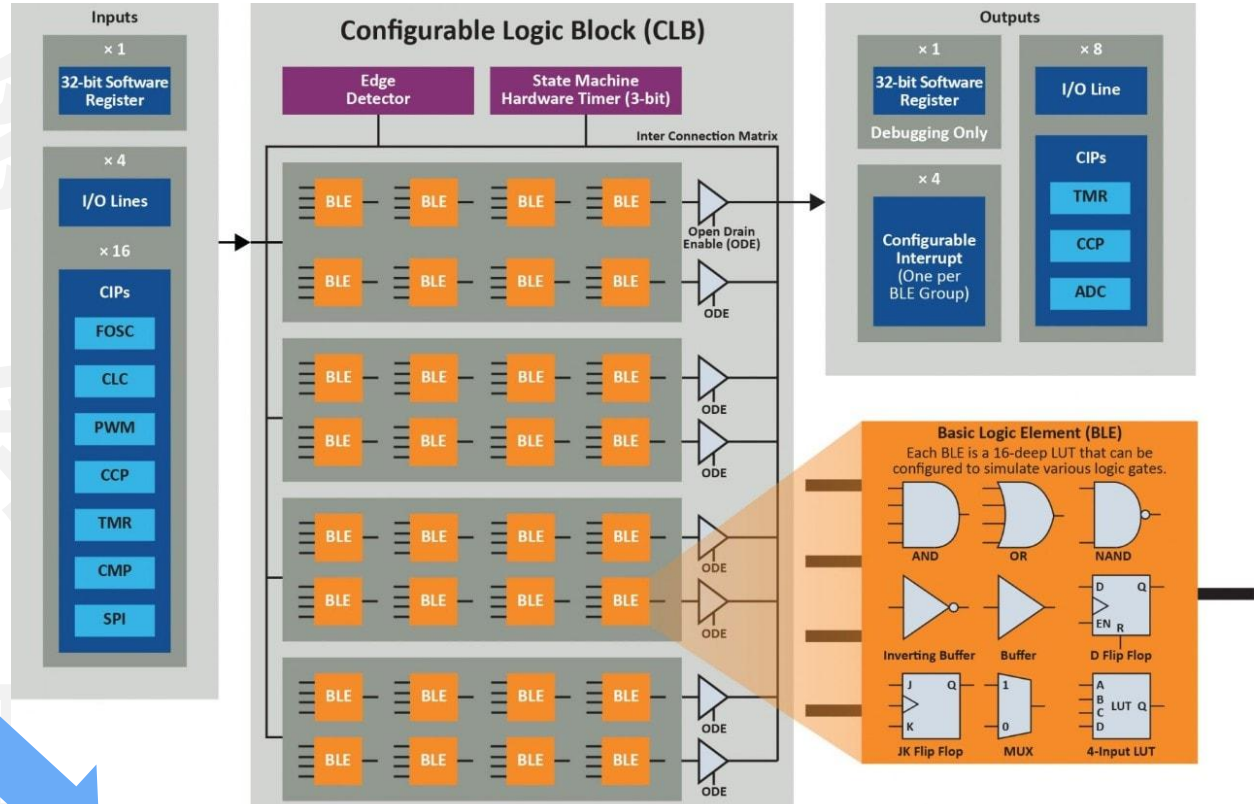
# 典型智能芯片体系结构 – FPGA

- 从偏向通用计算任务的CPU、GPU到偏向定制化设计的FPGA、ASIC




**可编程FPGA**

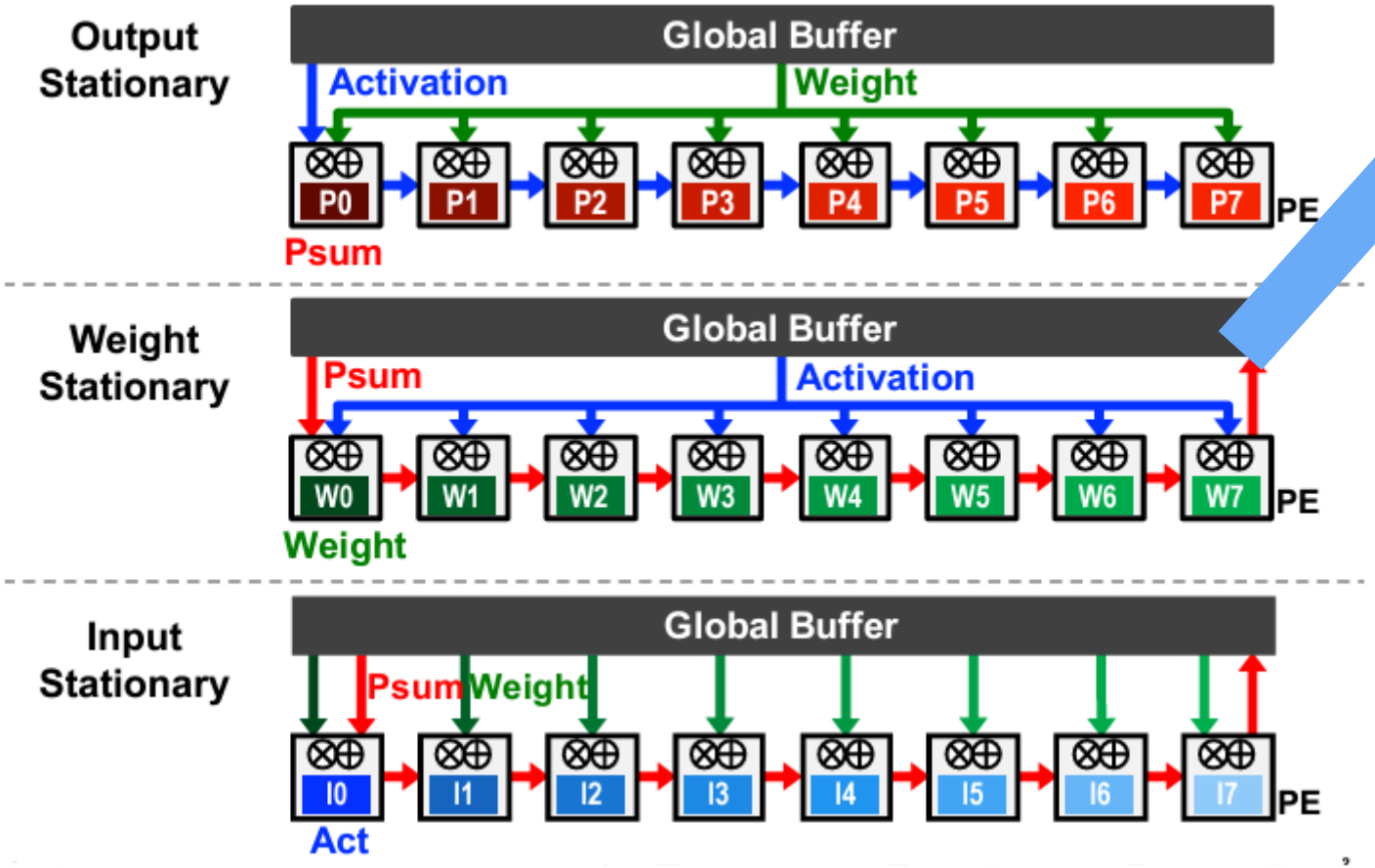
- 比CPU快
- 比GPU省功耗
- 比ASIC便宜流片周期短




**可编程逻辑模块CLB**

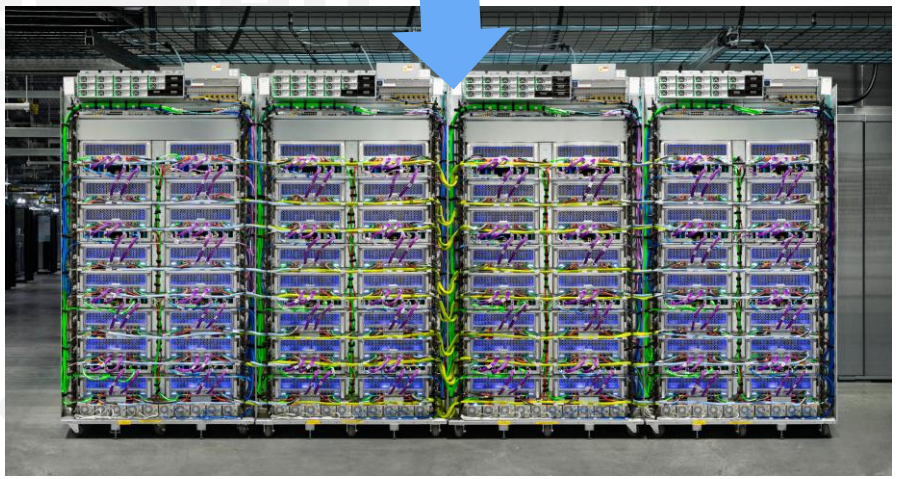
# 典型智能芯片体系结构 – ASIC

- 从偏向通用计算任务的CPU、GPU到偏向定制化设计的FPGA、ASIC




经典的神经网络加速器ASIC体系结构

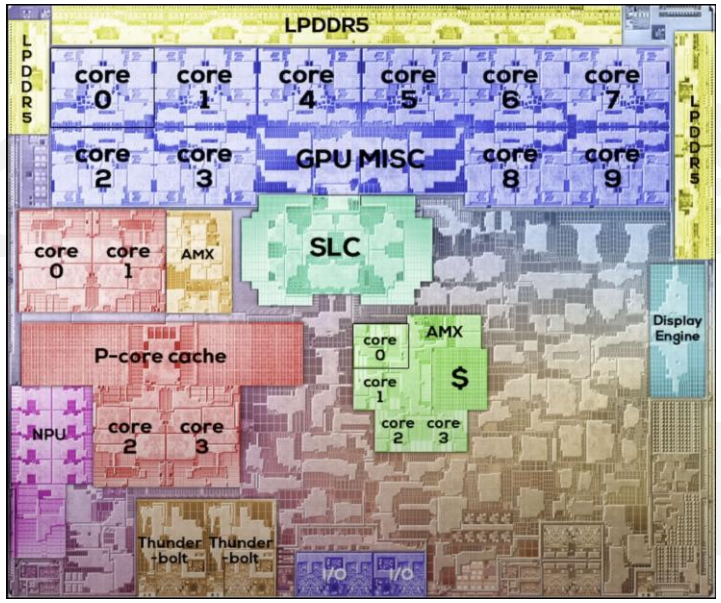
TPU
ASIC

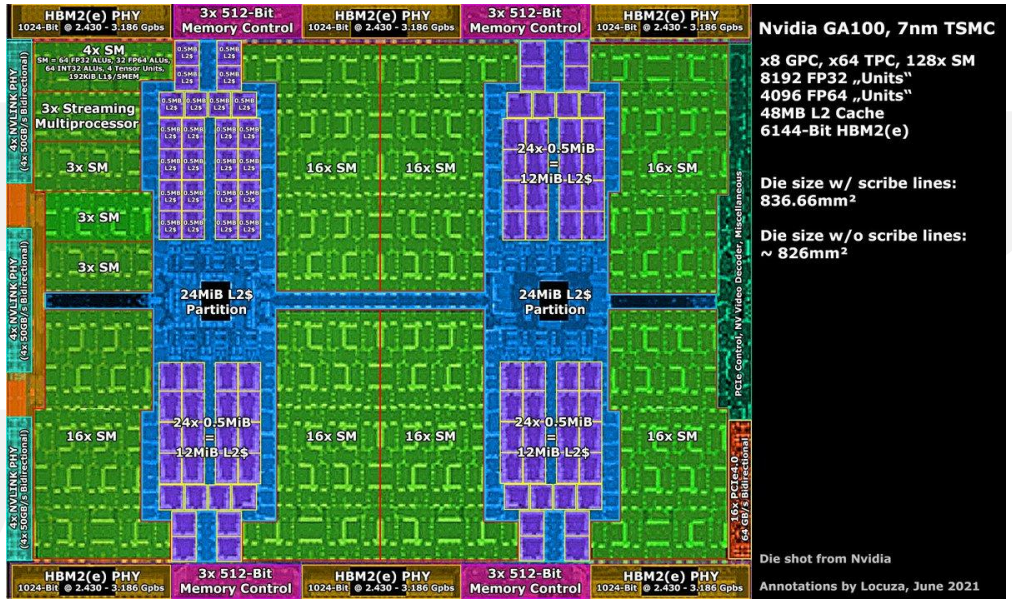


TPU组成的大型AI服务器

# 冯诺依曼体系结构

- 目前的成熟商用芯片基本均采用冯诺依曼体系结构




**Apple M3** **NVIDIA H100**

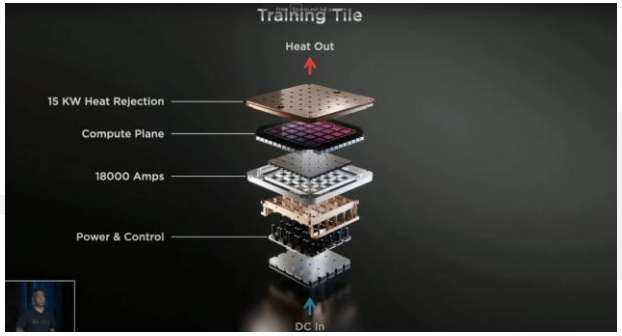
**特点：存储与计算单元分离，依靠总线进行连接，**

**执行程序时需要来回搬运数据（读出→计算→写入）**

# 推动智能时代飞速发展：AI芯片 – 2014/2015年至今

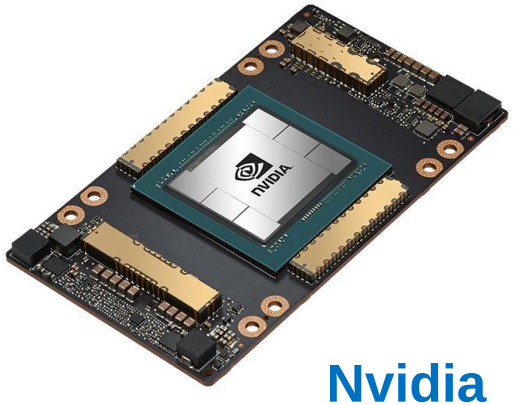
- 高性能AI芯片成为推动智能时代发展的算力基石，将引领未来十几年的技术革命

Google
TPU

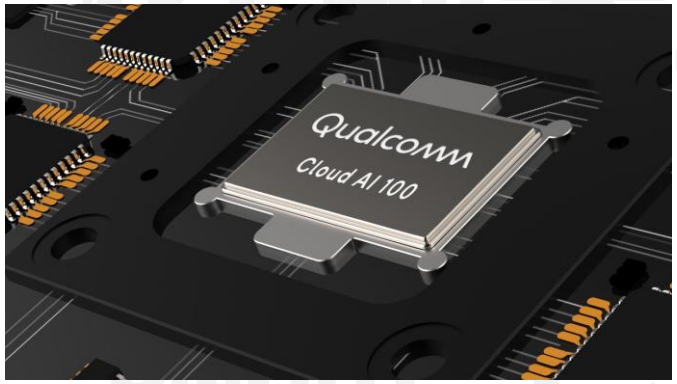



Tesla Dojo

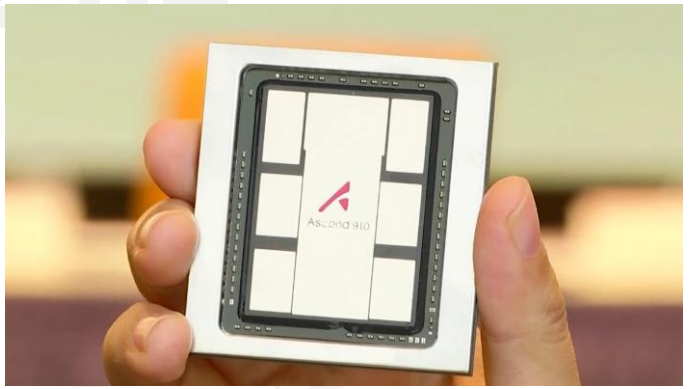

寒武纪




Nvidia
GPU



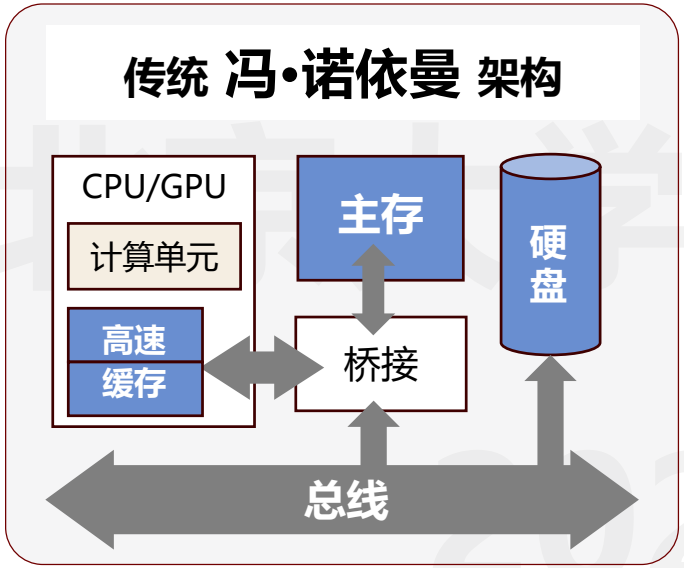

Qualcomm Cloud AI

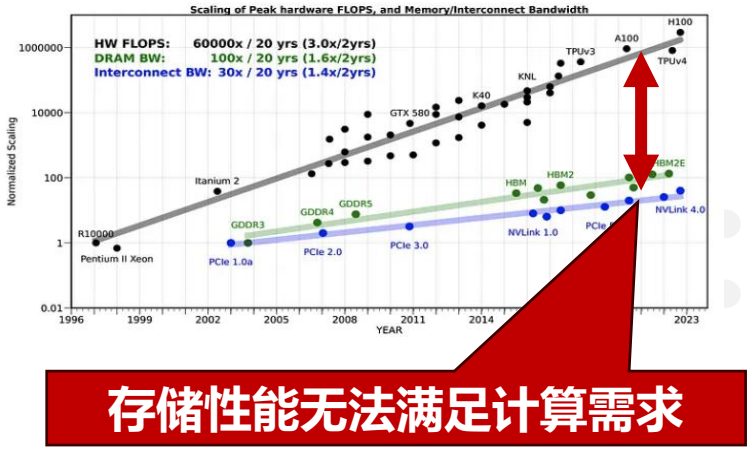



华为昇腾

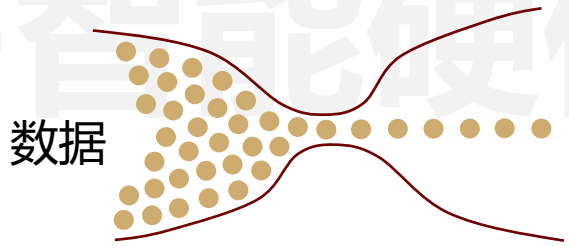
# 冯诺依曼体系结构

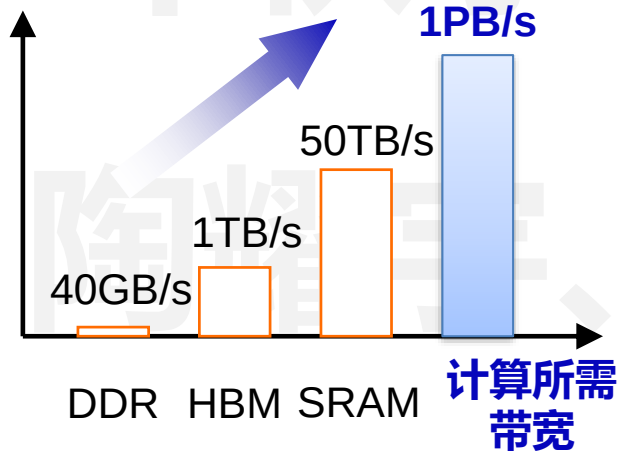
- 当前智能芯片体系结构的瓶颈



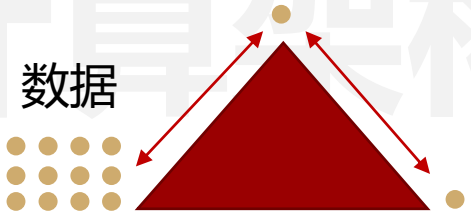



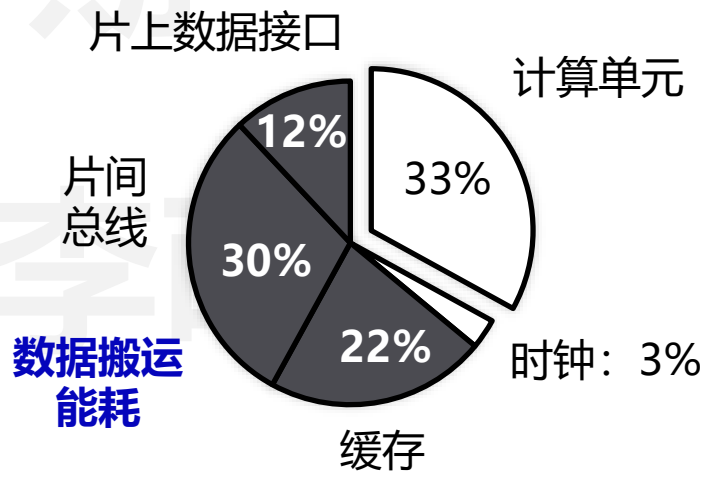

## 存储带宽限制算力




## 数据搬运限制能效



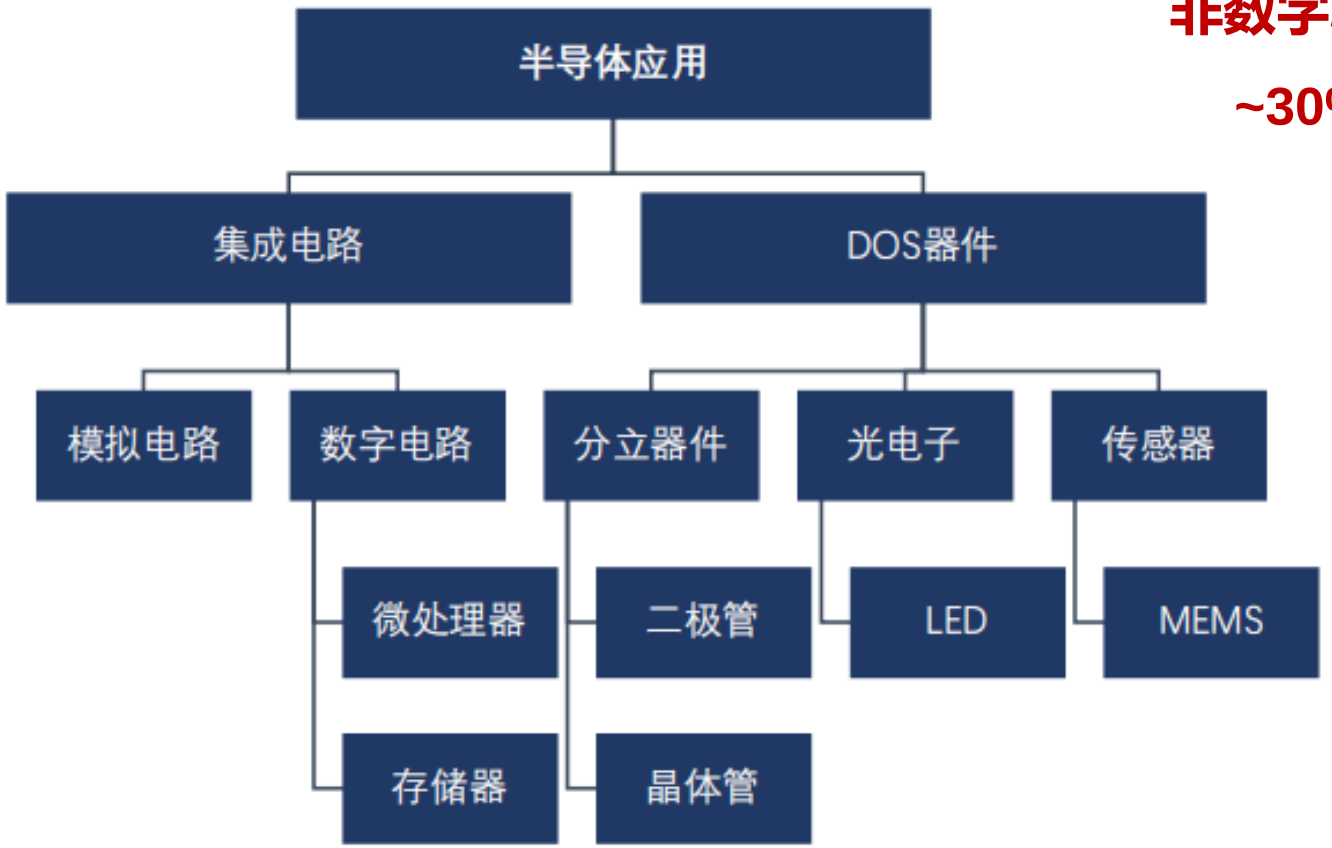



# 目 录

CONTENTS

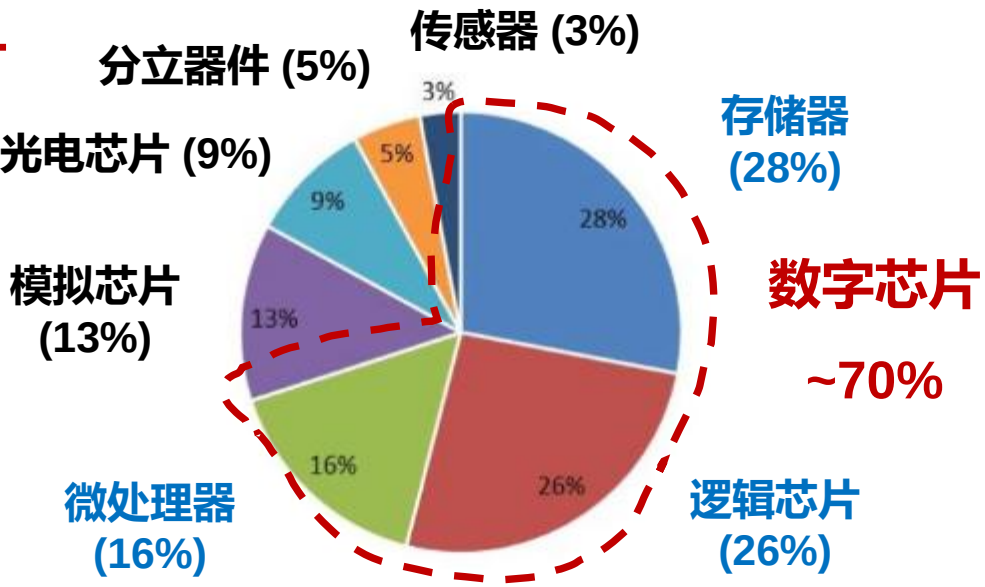
# 智能芯片产业按半导体应用分类

- 集成电路可分为模拟电路和数字电路，DOS器件分为光电子、传感器等

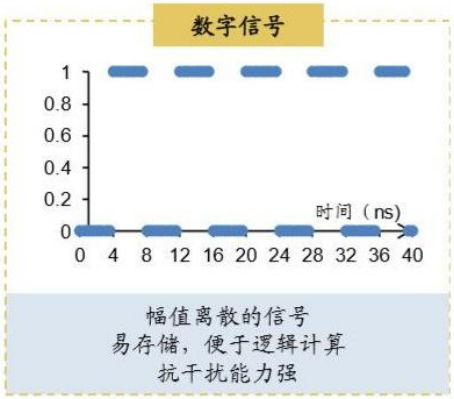


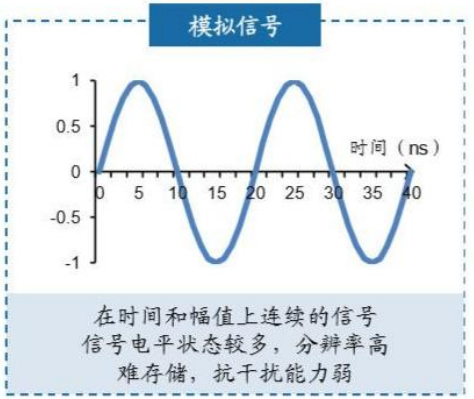

**非数字芯片**

**~30%**



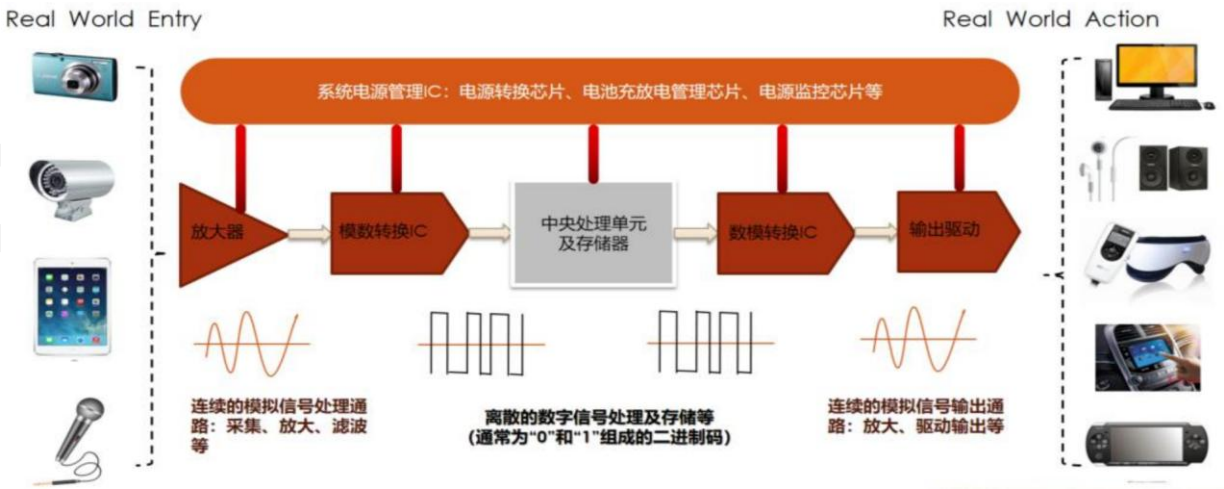

**数字电路的市场份额占70%**



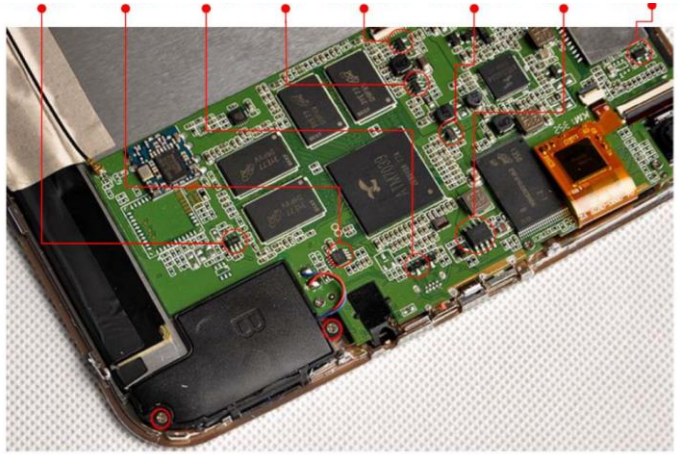



# 智能芯片产业按半导体应用分类

- 集成电路可分为模拟电路和数字电路，DOS器件分为光电子、传感器等

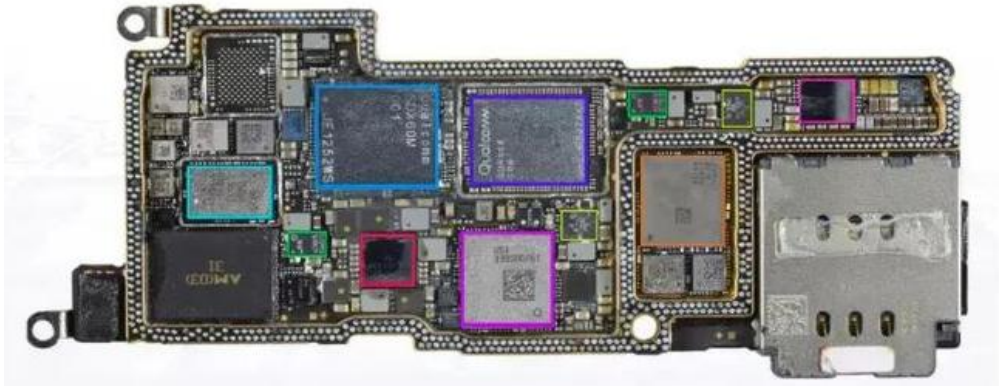



功率放大、电源管理、时钟生成、比较器、射频滤波、接口、数模转换、线性稳压等

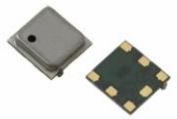


模拟芯片应用实例

CPU、GPU、存储器芯片、可编程逻辑芯片、MCU、DSP、NPU等



数字芯片应用实例



**传感芯片** 声、光、电、热、磁、压力、气体、震动、速度、湿度、惯性、流量、电磁波等

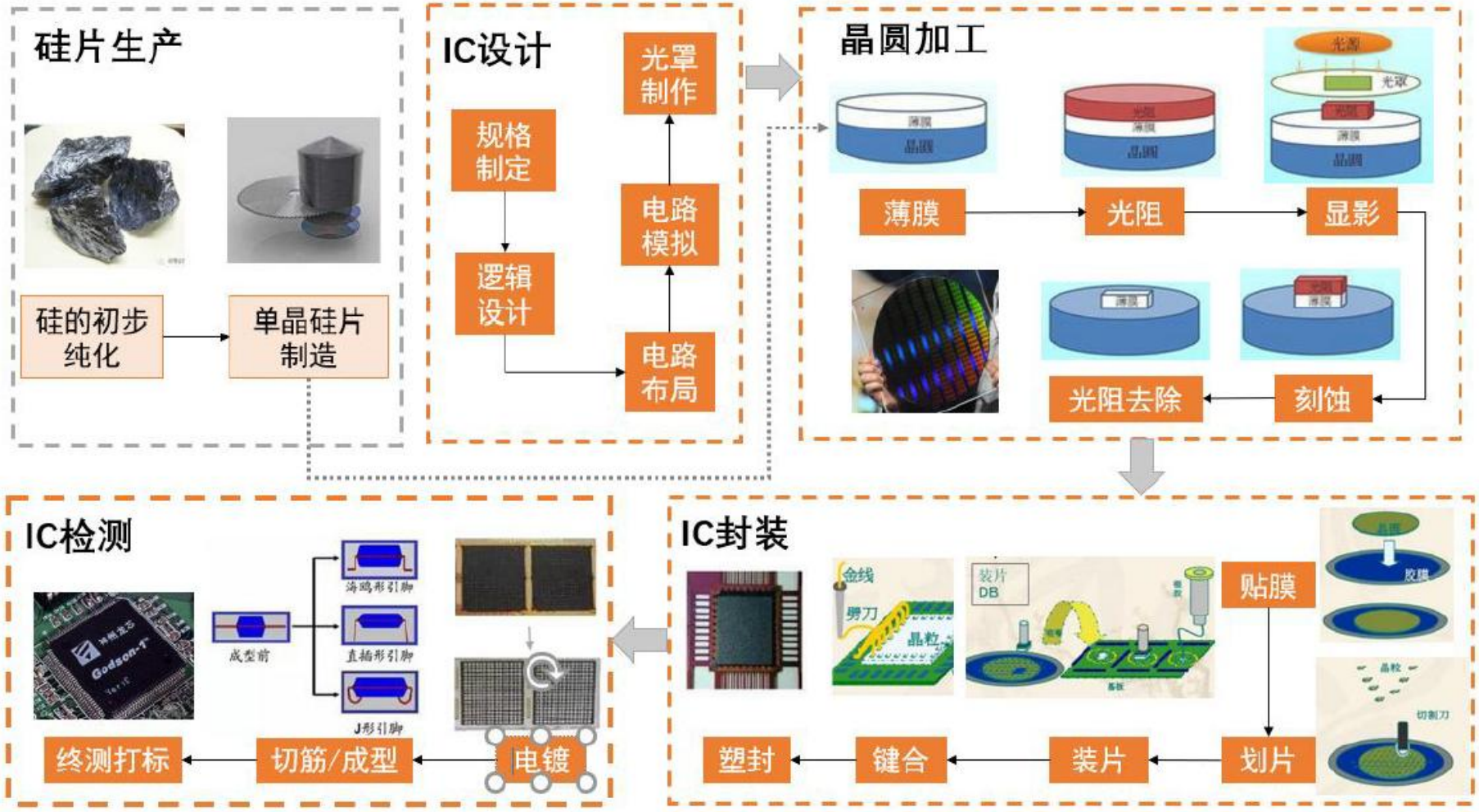
**光电芯片** 激光器芯片、半导体发光芯片等

**分立器件** 电阻、电容、电感、振荡器、晶体管、功率器件等

# 智能芯片产业现状 – 产业链极长、关联几乎所有工业门类

- 国际分工合作的庞大产业链生态

**中国与世界先进水平差距较大**




**硅片生产企业**
- 信越化学 (日本)
- 三菱住友 (日本)
- 环球晶圆 (台湾)

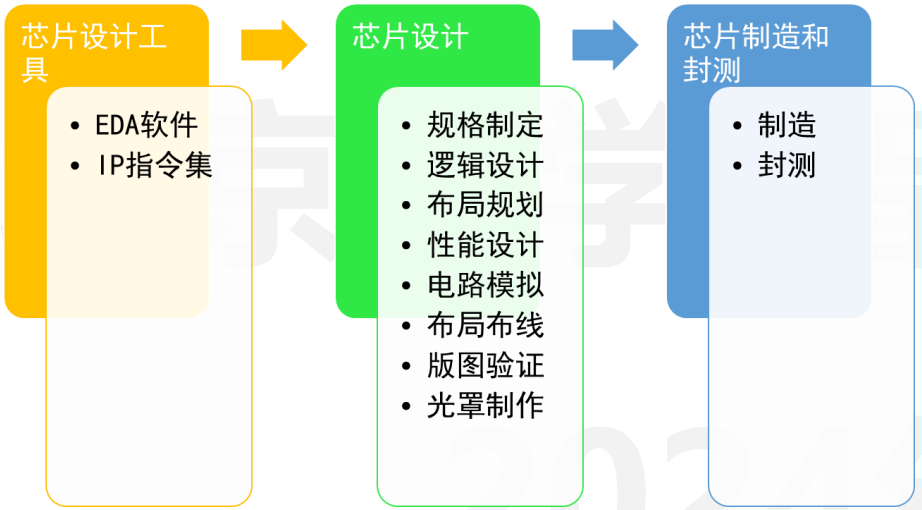
**晶圆加工企业**
- 台积电 (台湾)
- 三星 (韩国)
- 格芯 (美国)

**芯片设计企业**
- Intel (美国)
- Qualcomm (美国)
- 海思半导体 (中国)

**芯片封测企业**
- 日月光 (台湾)
- 安靠 (美国)
- 长电 (中国)

# 智能芯片产业现状 – 产业链极长、关联几乎所有工业门类

- 国际分工合作的庞大产业链生态




**芯片设计的EDA软件工具目前由美国公司所垄断**

| 芯片设计工具 | 芯片设计 | 芯片制造和封测 |
|---|---|---|
| • 美国英特尔 | • 美国苹果 | • 美国英特尔 |
| • 英国ARM | • 韩国三星 | • 韩国三星 |
| • 英国Imagination | • 英国Imagination | • 美国高通 |
| • 美国IBM | • 海思 | • 美国英伟达 |
| • 美国Mips | • 紫光展锐（展讯+锐迪科） | • 台积电 |
| • Synopsys | • 中兴微电子 | • 中芯国际 |
| • 美国Cadence | • 华大半导体 | • 华虹半导体 |
| • 美国Mentor Graphics | • 汇顶科技 | • 长江存储 |
| • 日本图研 | • 兆易创新 | • 合肥长鑫 |
| • 华大九天 | • 寒武纪 | • 福建晋华 |
| • 杭州中天微 | | • 紫光国微 |
| • 苏州国芯 | | • 长电科技 |
| • 寒武纪 | | • 华天科技 |
| | | • 通富微电 |

# 智能芯片产业的三种运作模式

- IDM (垂直整合)、Fabless (纯设计) 和 Foundry (晶圆加工)

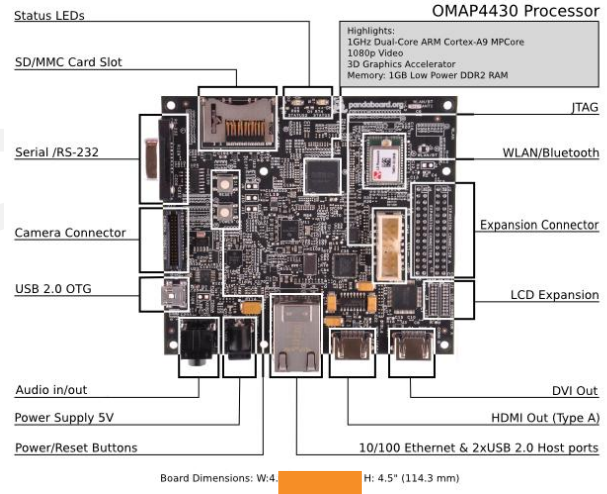
| | IDM模式 | Fabless模式 | Foundry模式 |
|---|---|---|---|
| 典型厂商 | SAMSUNG intel | QUALCOMM BROADCOM | tsmc SMIC |
| 基本特点 | 集芯片设计、制造、封测于一身。早期多数集成电路企业采用的模式，目前仅有极少数企业能够维持 | 只负责芯片的电路设计与销售；将生产、测试、封装等环节外包 | 不负责芯片设计，只负责制造或封测；可以同时为多家设计公司提供服务，但受制于公司间的竞争关系 |
| 主要优势 | 设计、制造等环节协同优化，有助于充分发掘技术潜力；能有条件率先实验并推行新的半导体技术 | 资产较轻，初始投资规模小，创业难度相对较小；企业运行费用较低，转型相对灵活 | 不承担由于市场调研不准、产品设计缺陷等决策风险。 |
| 主要劣势 | 公司规模庞大，管理成本较高；运营费用较高，资本回报率偏低 | 与IDM相比无法与工艺协同优化，因此难以完成指标严苛的设计；与Foundry相比需要承担各种市场风险 | 投资规模较大，维持生产线正常运作费用较高；需要持续投入维持工艺水平，一旦落后追赶难度较大。 |

早期企业都是IDM运营模式（垂直整合），这种模式涵盖设计、制造、封测等整个芯片生产流程，这类企业一般具有规模庞大、技术全面、积累深厚的特点，如Intel、三星等

随着专注于晶圆加工的台积电的出现，演化出Fabless和Foundry模式，专攻设计或者制造，各司其职

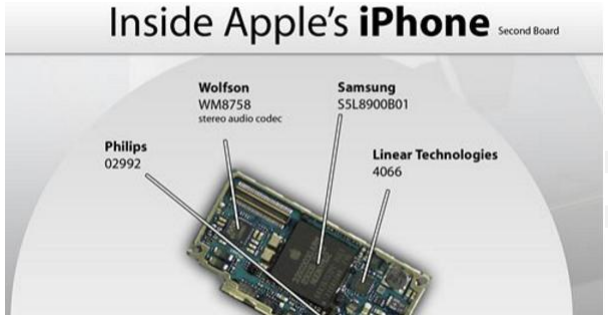
# 推动移动互联网飞速发展：移动SOC芯片 – 2007年至今

- 移动电话SOC芯片成为推动移动互联网飞速发展的算力基石，引领过去十几年的技术革命



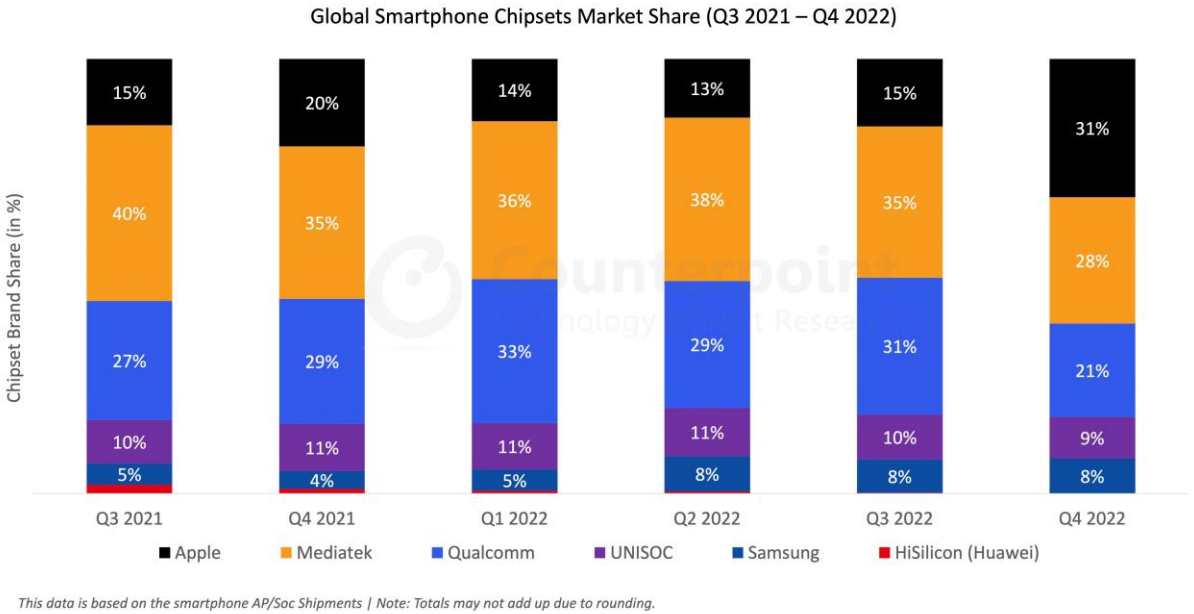

**德州仪器OMAP手机芯片**

**诺基亚 6630、 6680、6681、E50、E60、E61、E62、E65、E70、N70、N71、N72、N73、N80、N90、N91和N92 等**




**三星S5L8900 SOC芯片**

**2007年乔布斯发布了第一代iPhone采用90nm制程三星SOC芯片**




**苹果、高通、联发科、三星、紫光展锐、华为海思占据移动SOC市场的前列**

# 推动制程不断向前发展：中国台湾台积电/英特尔/三星 – 2008年至今

- 过去十几年，中国台湾积体电路公司、英特尔公司、三星公司是推动芯片制程发展的主要力量

**全球主要晶圆厂制程节点技术路线图**

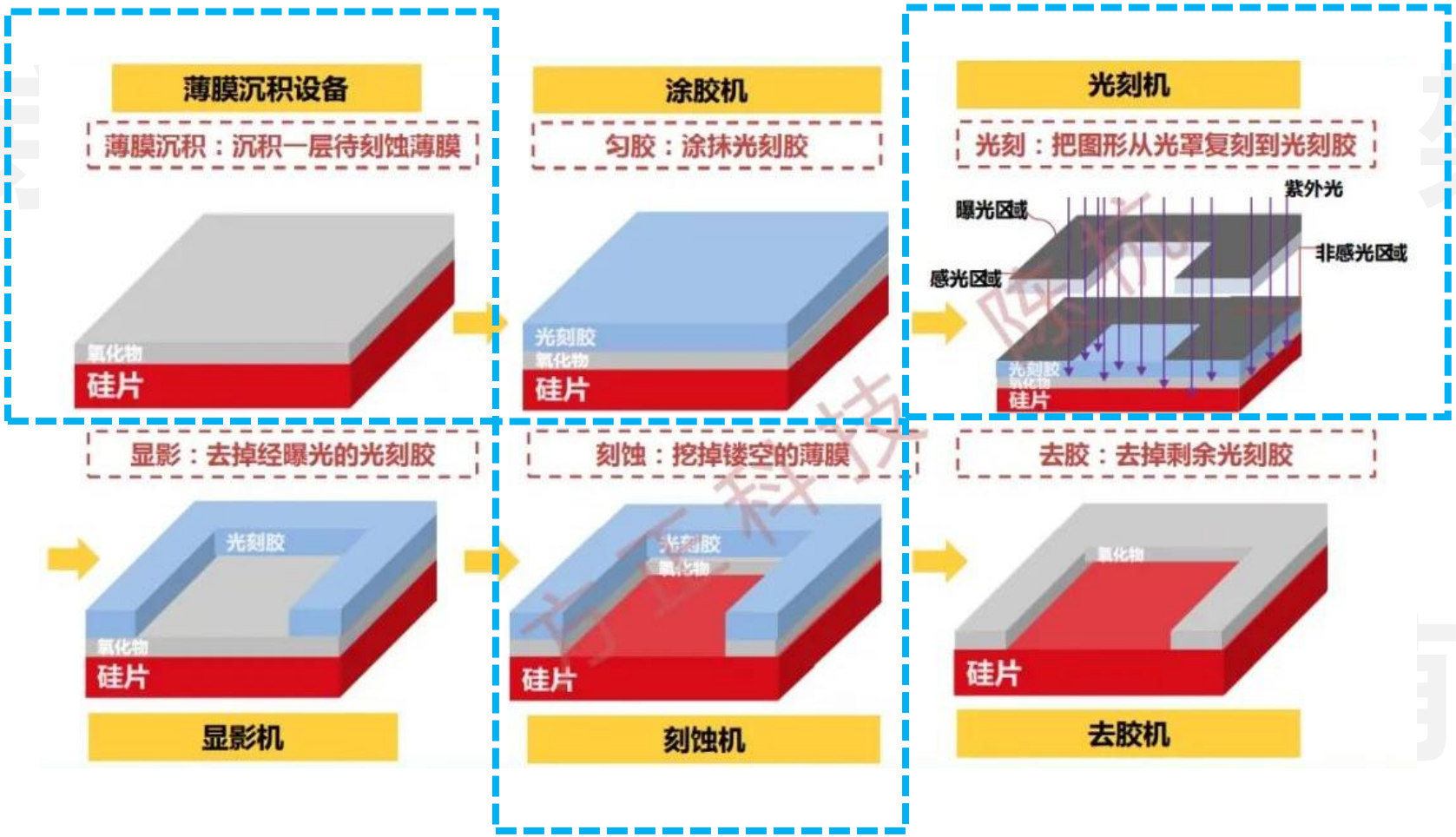
| 晶圆代工厂 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 台积电 | 28nm | | | 20nm | 16nm | | 10nm | 7nm | 7nm + | 5nm<br>6nm | | 3nm | | 2nm |
| 三星 | | 28nm | | 22nm | 14nm | | 10nm | 8nm | 7nm EUV<br>6nm | 5nm | 3nm | | | |
| 英特尔 | 22nm | | | 14nm | | 14nm + | 14nm ++ | | 10nm | 10nm + | 7nm<br>10nm ++ | 7nm + | 7nm ++ | |
| 格罗方德 | | | 28nm | | 14nm | | 12nm | | | | | | | |
| 联电 | | | | 28nm | | | 14nm | | | | | | | |
| 中芯国际 | 40nm | | | | 28nm | | | | 14nm | | | | | |

备注：以上信息整理自网络，如有错漏欢迎指正。

**中国台湾积体电路公司后来追上，超越英特尔与三星**

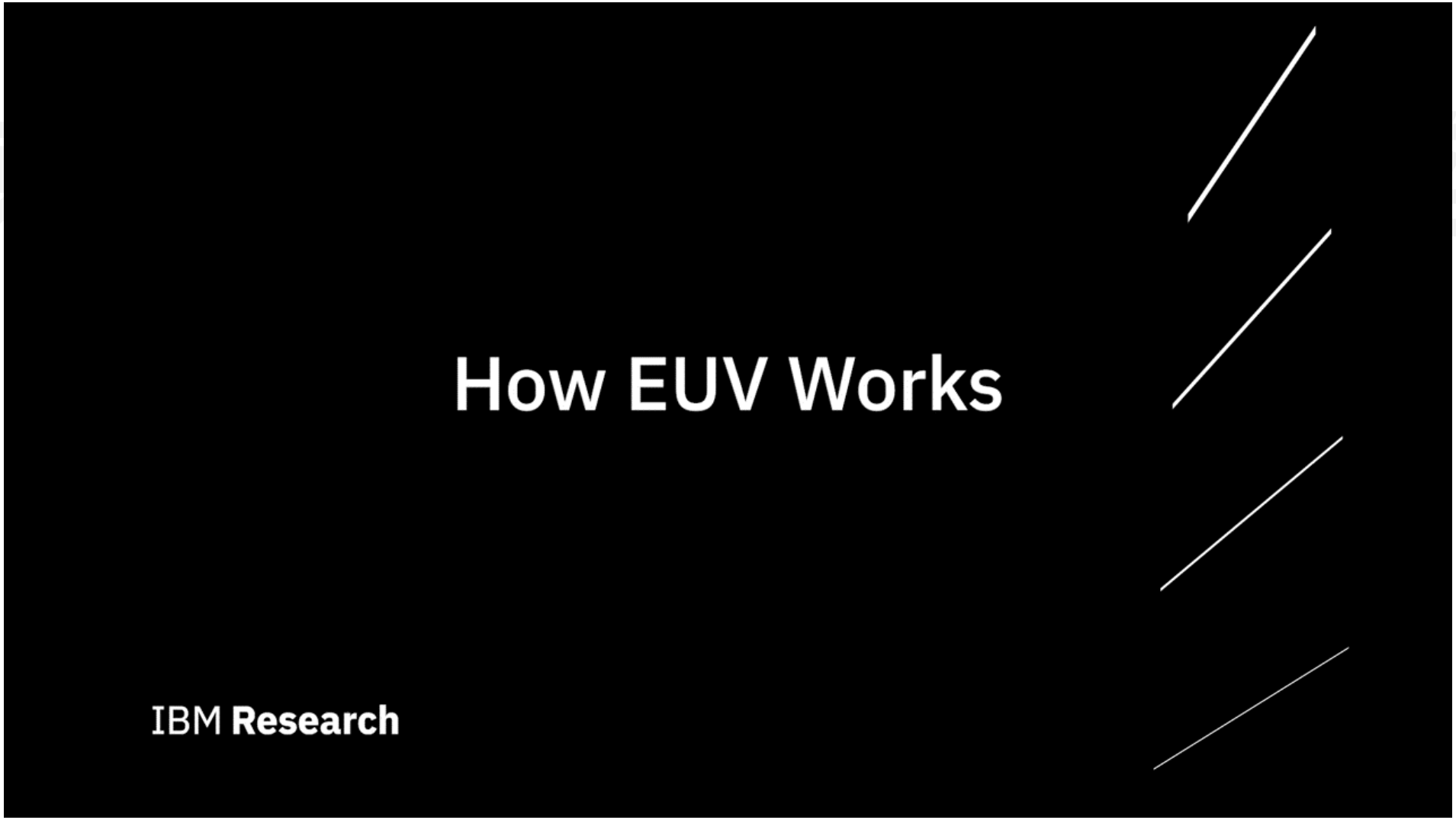
# 中国的“卡脖子”领域之器件制造：晶圆加工产业

- 更高精度、更高可靠性的光刻、刻蚀、薄膜沉积技术是亟待解决的三大瓶颈



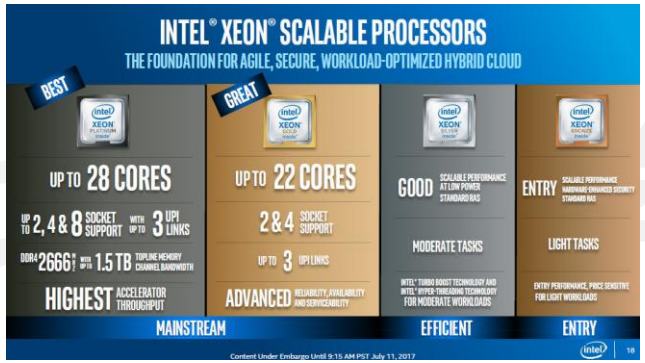
# 中国半导体芯片产业的关键瓶颈 – 光刻技术

- 高性能EUV光刻

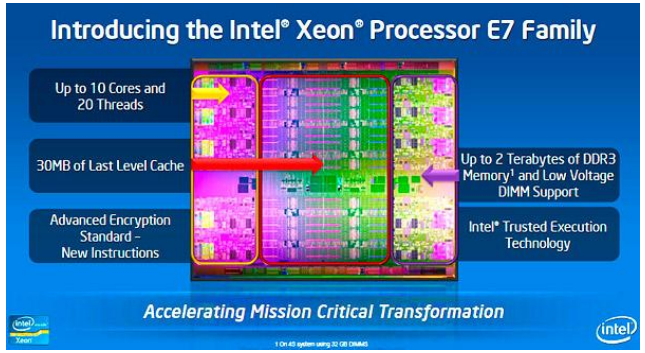


# 中国的“卡脖子”领域之计算架构：高性能处理器芯片

- 我国在高性能计算芯片CPU、GPU、FPGA的指令集与架构设计领域目前落后较多

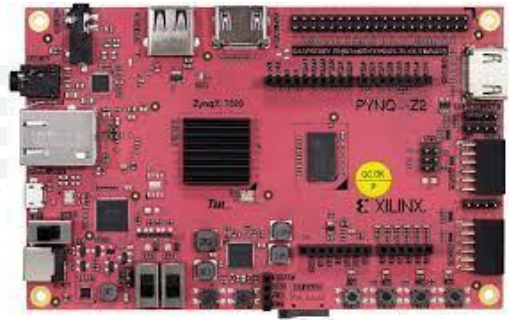







高性能CPU遭美国出口管制禁运


高性能GPU遭美国出口管制禁运




高性能可编程逻辑FPGA与美国主流厂商Altera、Xilinx差距明显


国产龙芯3C5000目前已可商用，但性能与至强系列仍有显著差异

国产GPU尚处于初级阶段

国产GPU包括摩尔线程、壁仞科技、燧原科技、天数智芯、景嘉微等，与英伟达差距很大

国产FPGA包括紫光同创、安路科技、复旦微等，在并行规模、功能灵活性上急需进步

# 中国的“卡脖子”领域之芯片设计软件：EDA产业

- 我国在高性能的电路辅助设计与仿真工业软件方面目前与发达国家差距明显

## 现状

**2019年**

EDA三巨头Cadence、Synopsys、Mentor等已经对华为断供EDA工具，不再出售新license

**2022年**

2022年8月美国将停止对中国出口GAA相关EDA软件，GAA主要用于3nm及以下晶体管


**国产EDA软件目前门类已补齐，但制程支持、设计仿真性能、与晶圆厂对接等多方面仍处于落后状态**

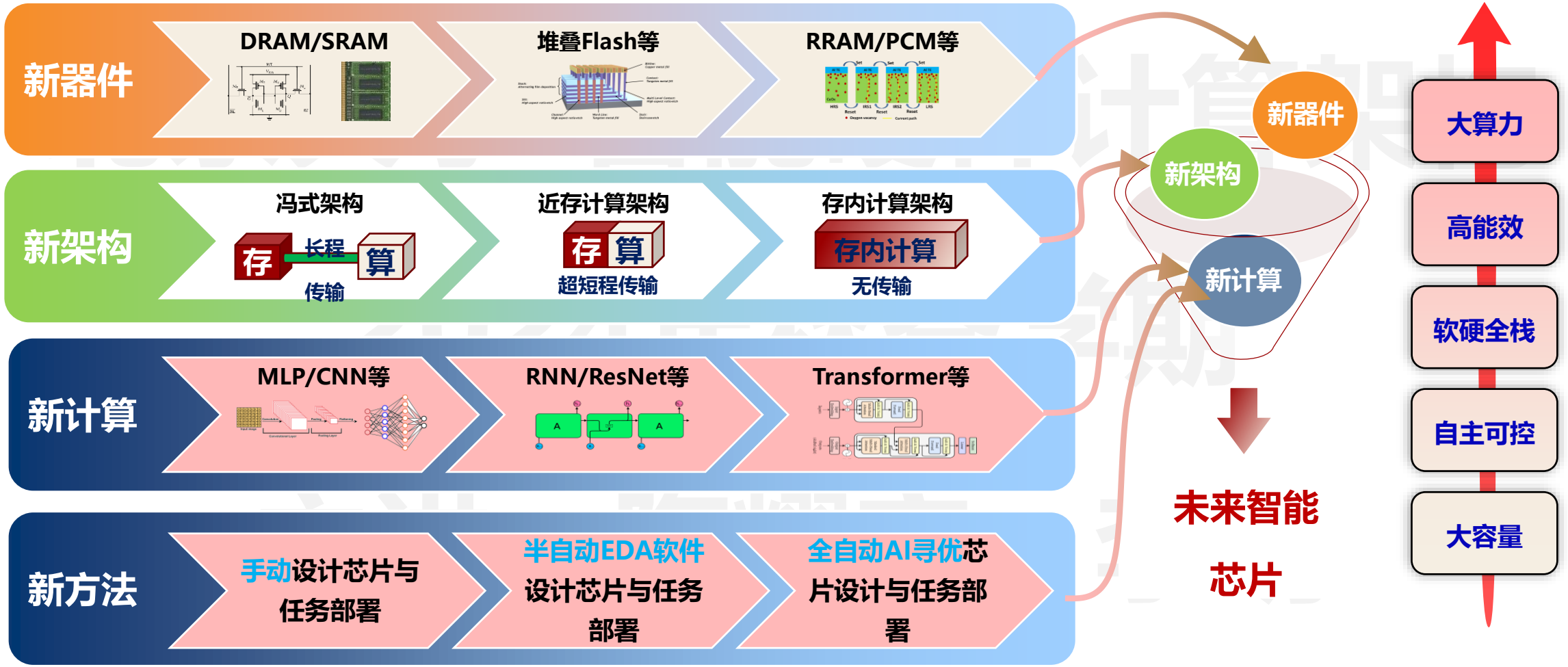
# 目 录

CONTENTS

# 融合新器件、新架构、新计算是后摩尔时代体系结构的发展趋势

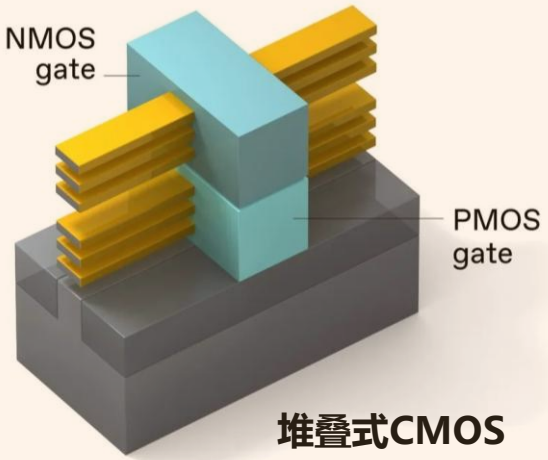
北京大学 PEKING UNIVERSITY

- 融合新器件、新架构、新计算是突破后摩尔时代大算力、高能效瓶颈的重大关键技术领域



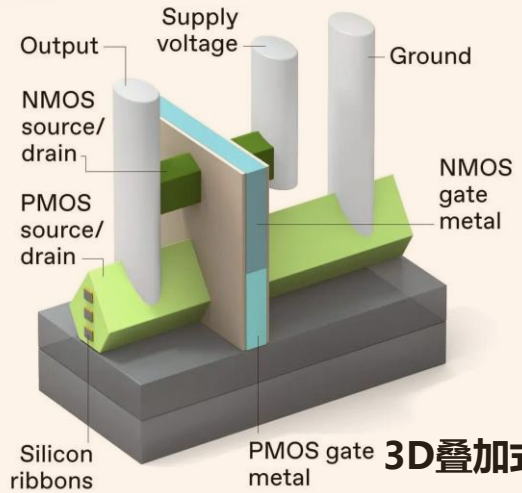
# 代表性智能芯片新兴技术– 新器件：高密度的逻辑器件

- 未来三维堆叠式晶体管与碳管器件

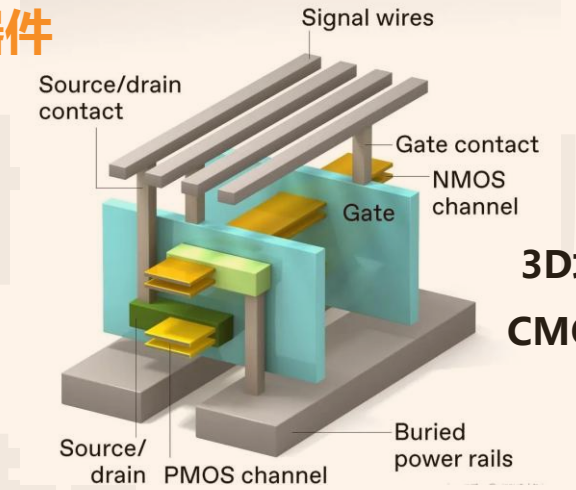



堆叠式CMOS

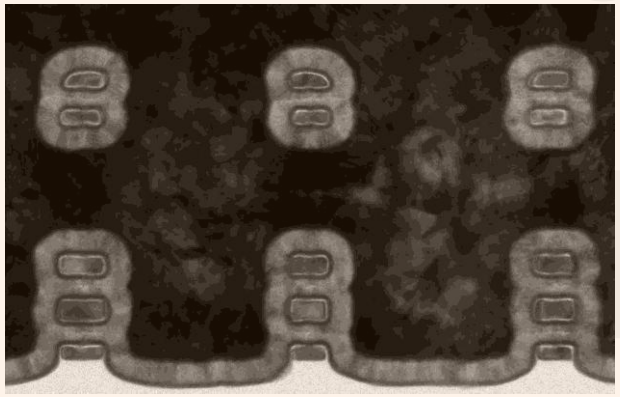
新兴硅基器件

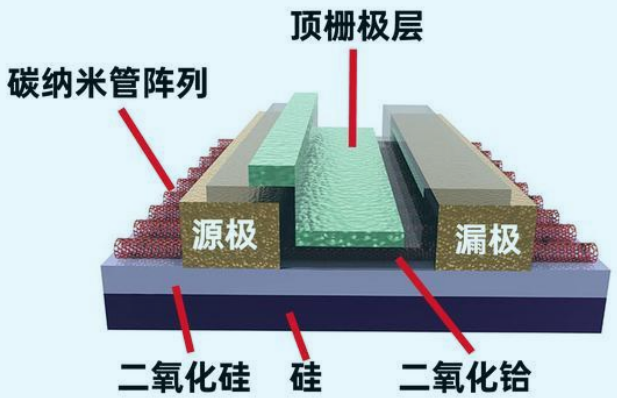


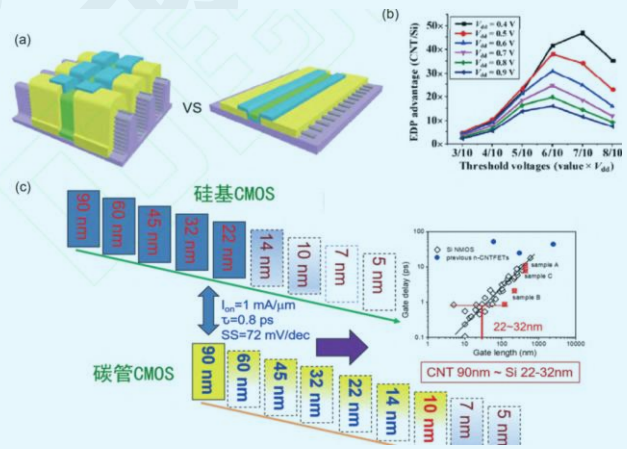

3D叠加式非门




3D堆叠的CMOS连接



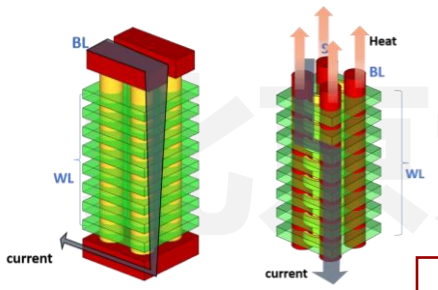



新兴碳基器件

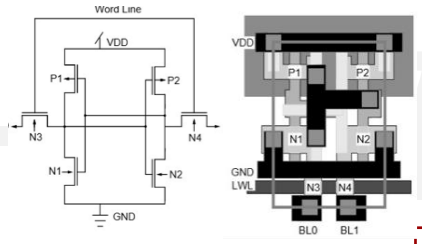
# 代表性智能芯片新兴技术– 新器件：存储-计算融合器件

- 未来存储器介质材料的创新



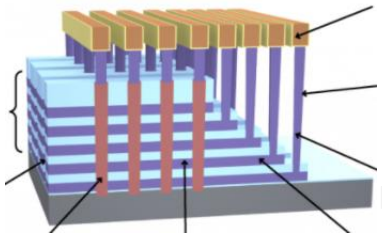

**DRAM**

优点：工艺成熟、密度高
缺点：速度低、刷新、只近存
非易失性：否
适合场景：冯氏架构过渡



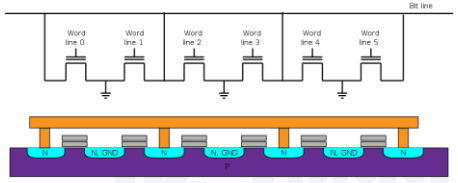

**SRAM**

优点：工艺成熟、IP化应用
缺点：能效低、密度低
非易失性：否
适合场景：端侧、边缘中小算力



**SSD/Nand Flash**

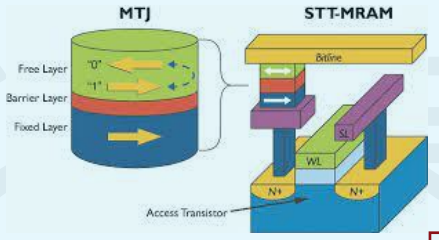
优点：工艺成熟、容量大、成本低
缺点：速度低、只能近存
非易失性：是
适合场景：云端大容量




**Nor Flash**

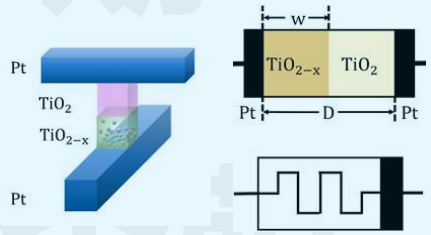
优点：工艺成熟、密度高、成本低
缺点：对PVT变化敏感、能效低
非易失性：是
适合场景：端侧、边缘低成本

## 新兴器件




**磁器件 (MRAM)**

优点：能效、速度、密度高
缺点：与CMOS大规模集成难
非易失性：是
适合场景：端侧、边缘中小算力

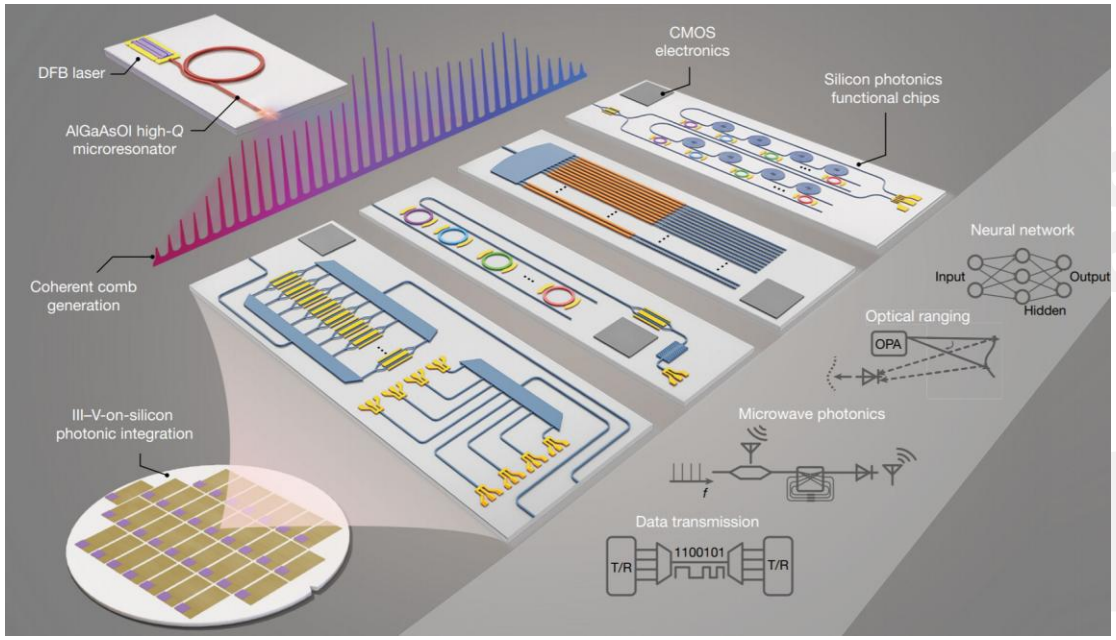


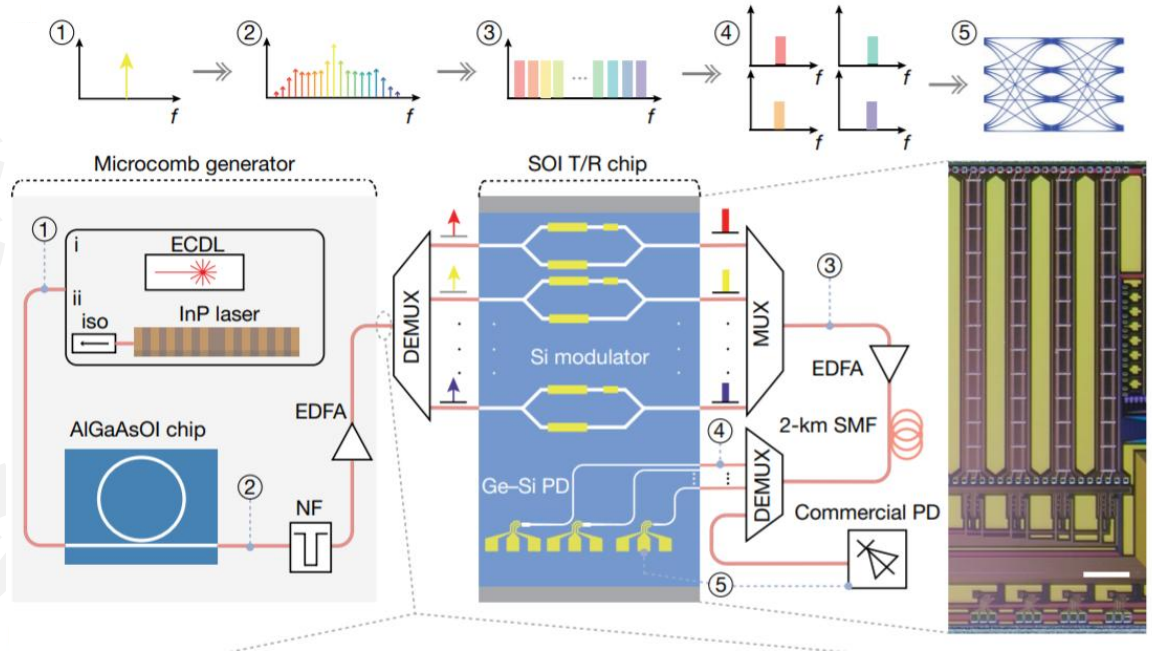

**忆阻器 (RRAM/PCM)**

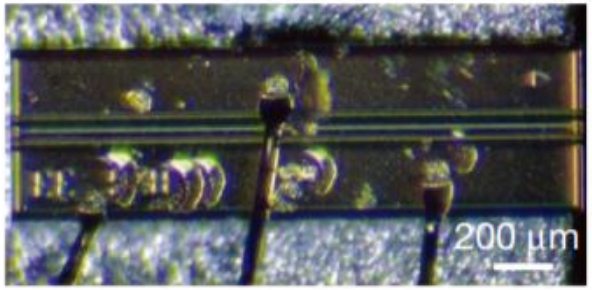
优点：算力、能效、密度高
缺点：工艺爬坡成熟中
非易失性：是
适合场景：云边端大算力

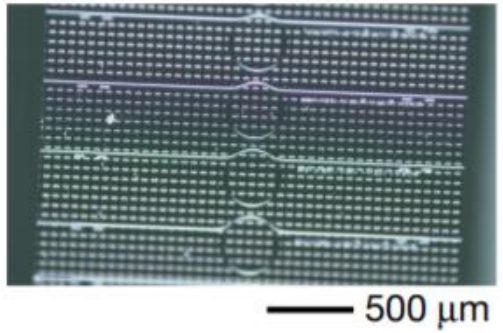
# 代表性新兴技术 – 新器件：光器件与片上光互连技术

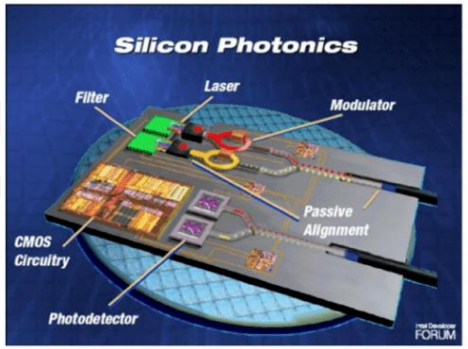
- 片上集成光电子通信系统有望突破信号传递延时的瓶颈，打破金属互连的物理上限

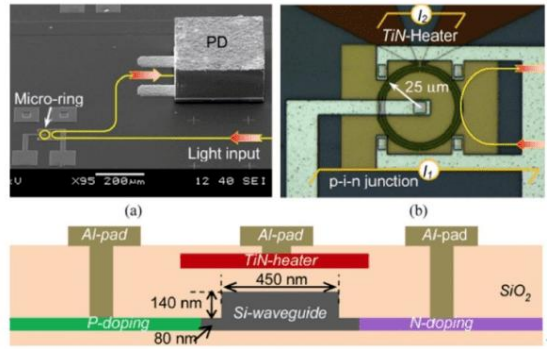


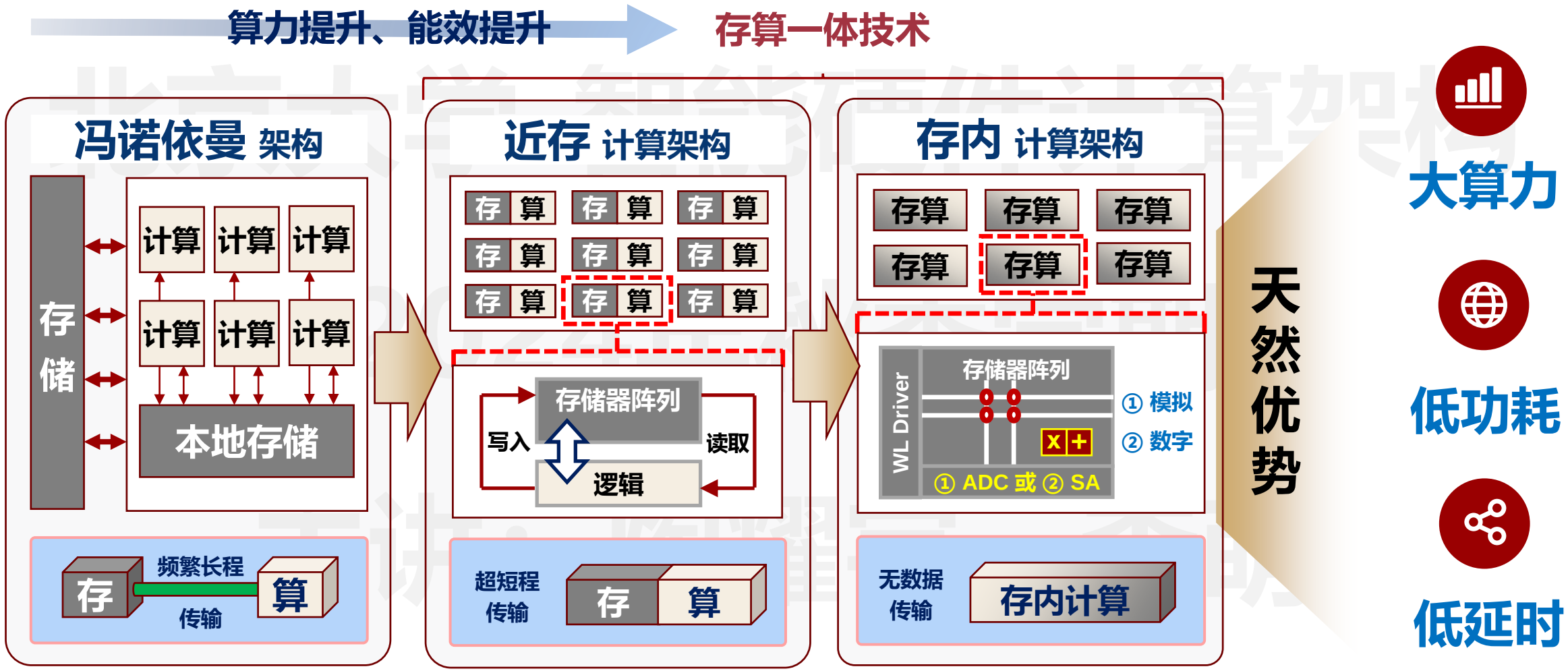



# 代表性智能芯片新兴技术 – 新架构：存算一体

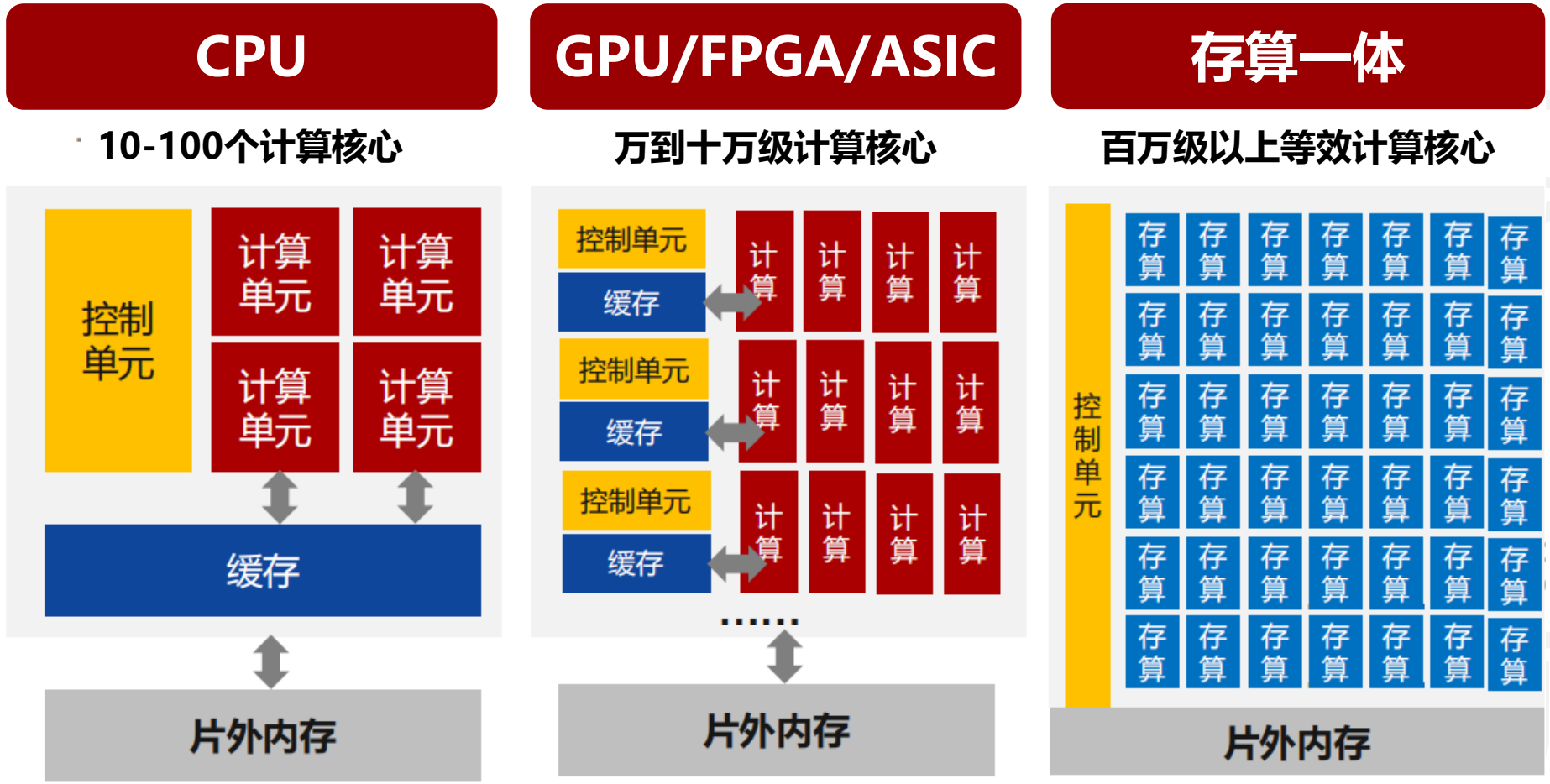
- 存算一体技术成为后摩尔时代打破算力瓶颈的重要路径



# 存算一体成为打破AI大模型推理算力极具潜力的技术路径

北京大学 PEKING UNIVERSITY

- 存算一体提供比GPU等冯氏芯片高多个数量级的并发度，有效支撑AI大模型推理

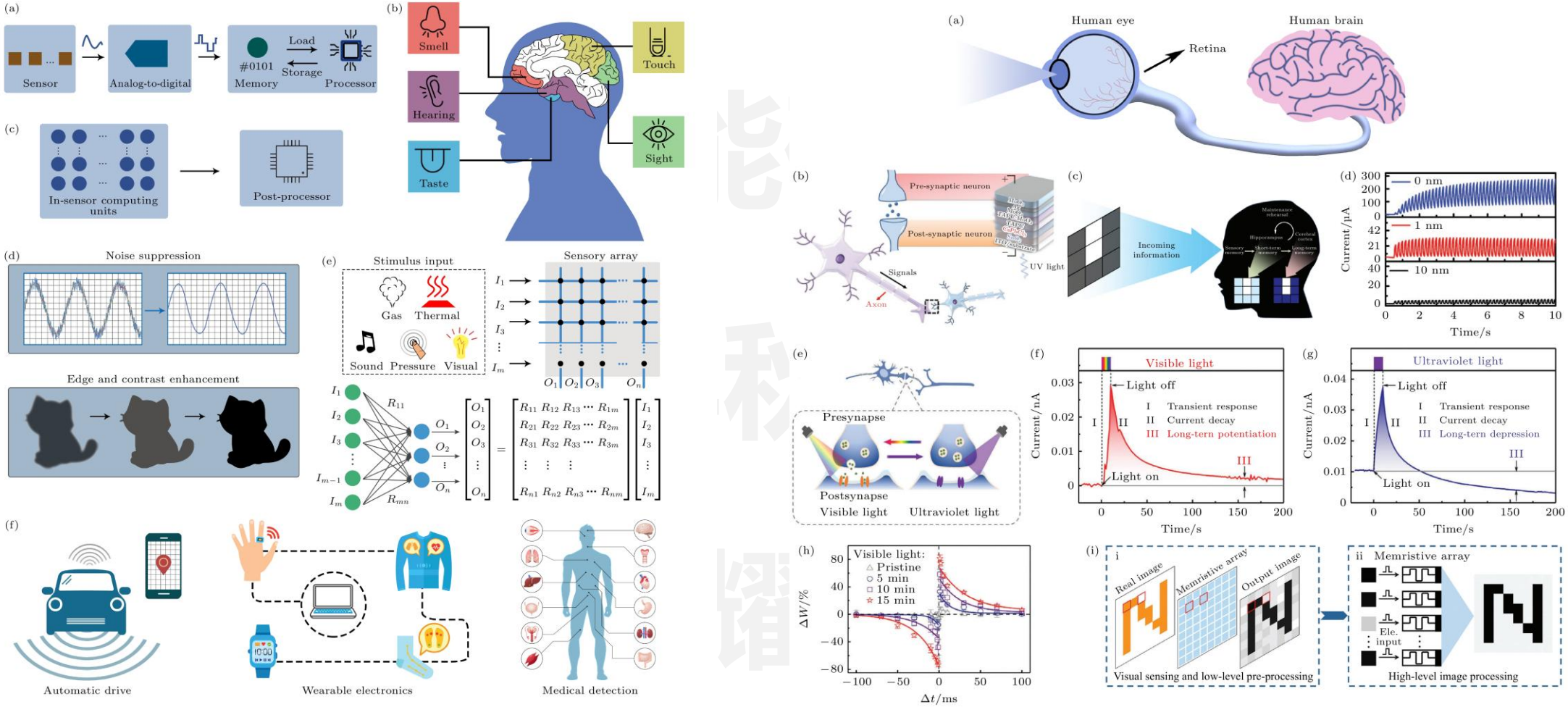



存算一体如何对AI大模型进行有效支持?

现有AI大模型推理基本上基于GPU/FPGA/ASIC等冯氏芯片

# 代表性智能芯片新兴技术 – 新架构：感存算一体

- 将传感、计算、存储融为一体，大幅降低系统功耗和计算延时，应用前景广阔

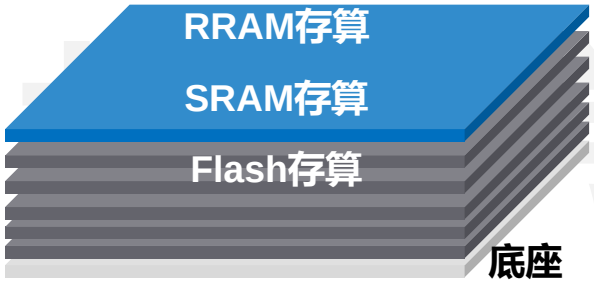



**视觉感存算一体芯片与硬件系统**

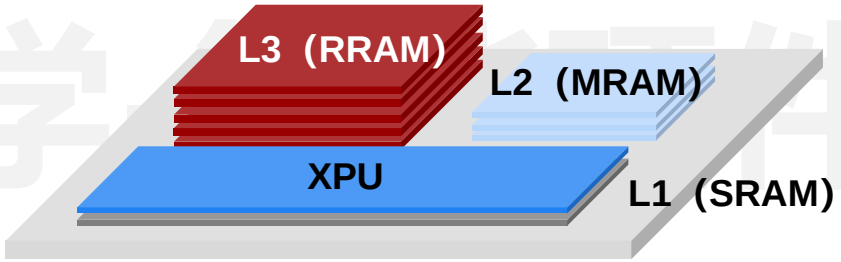
# 代表性智能芯片新兴技术 – 新架构：三维异质集成

- 协同先进封装技术，实现多种芯片方案相结合

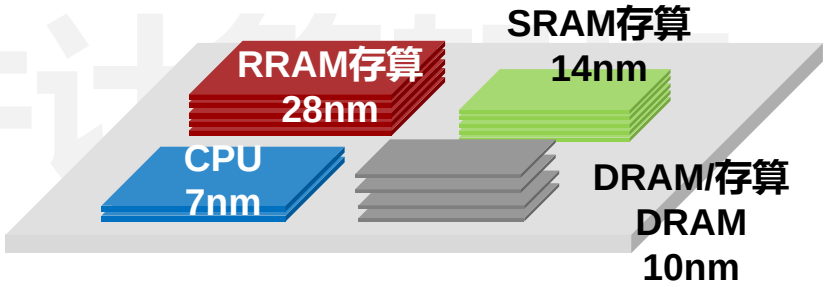
## 先进三维集成芯片示例图




三维集成

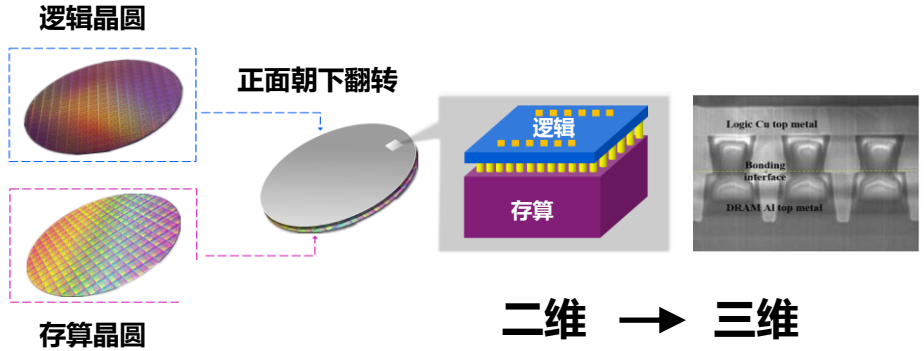



多级存储器堆叠SoC



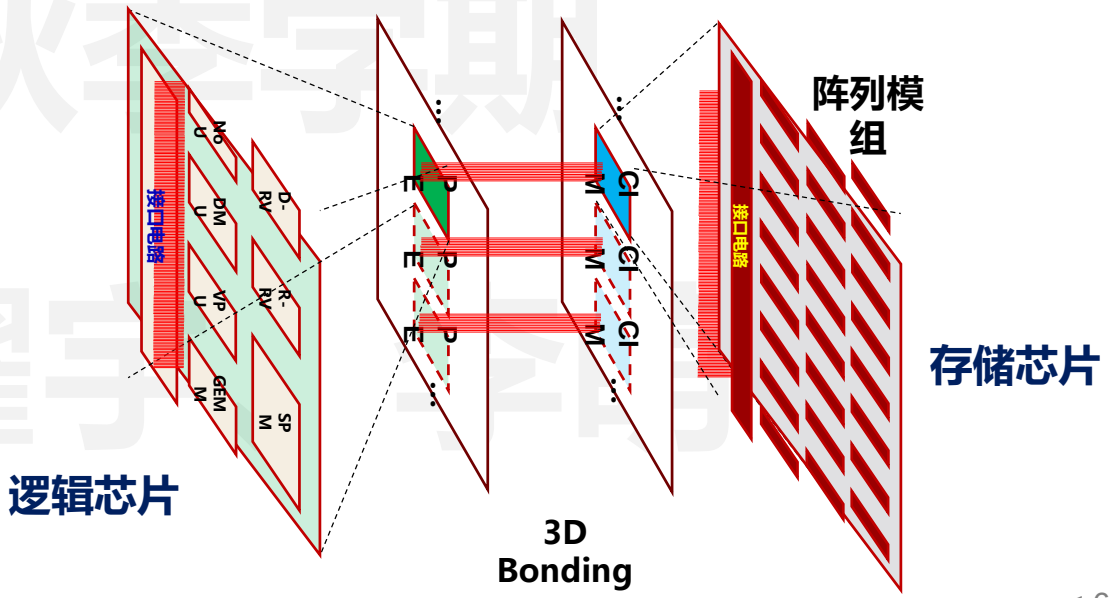

异构小芯粒封装

### 混合键合异质三维集成




二维 → 三维

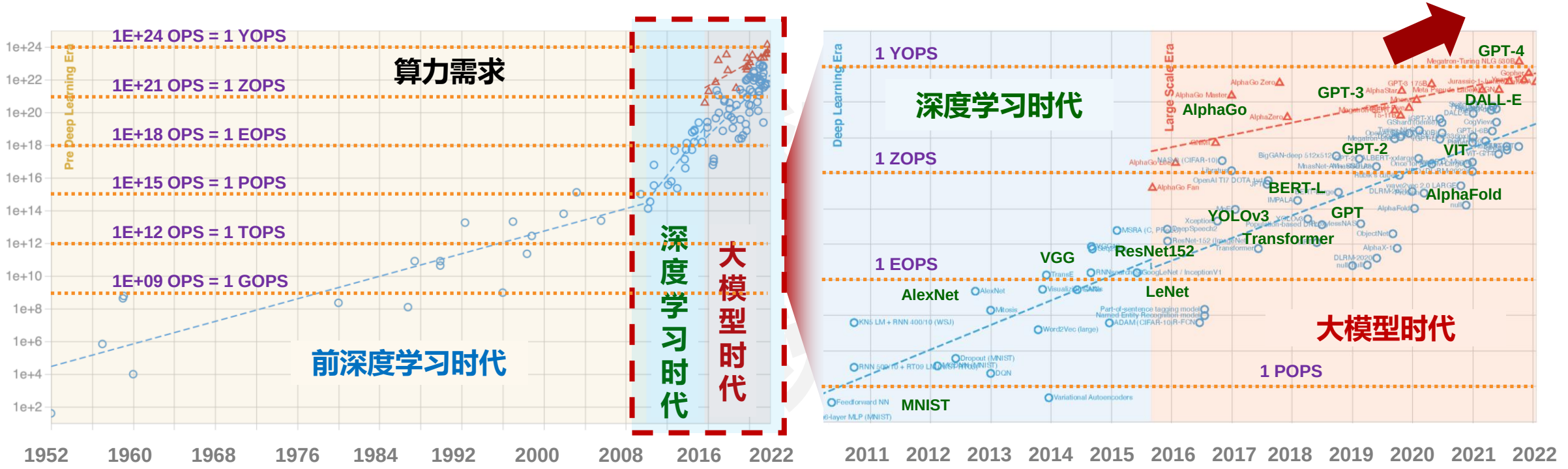
解决线路拥塞、突破面积约束、兼容不同制程、发挥各自优势



# 代表性智能芯片新兴技术 – 新计算：AI大模型

北京大学 PEKING UNIVERSITY

- 以AI大模型为代表的新一代人工智能系统对高性能AI芯片提出了新的要求




| 历史时期 | 算力需求 | 翻倍间隔 |
|---|---|---|
| 前深度学习时代 1952 – 2010 | 30 KOPS – 200 TOPS | 21.3月 |
| 深度学习时代 2010 – 2022 | 700 TOPS – 2 EOPS | 5.7月 |
| 大模型时代 2016 – 2022 | 1 ZOPS – 1 YOPS | 9.9月 |

| 代表性AI大模型 | 参数量 | 算力需求 |
|---|---|---|
| GPT-4 | ~1.5万亿个 | ~2.7 YOPS |
| GPT-3 | ~1746亿个 | ~314 ZOPS |
| GPT-3 Small | ~1.25亿个 | ~224 EOPS |

**芯片性能成为支撑智能系统从量变产生质变的基石**

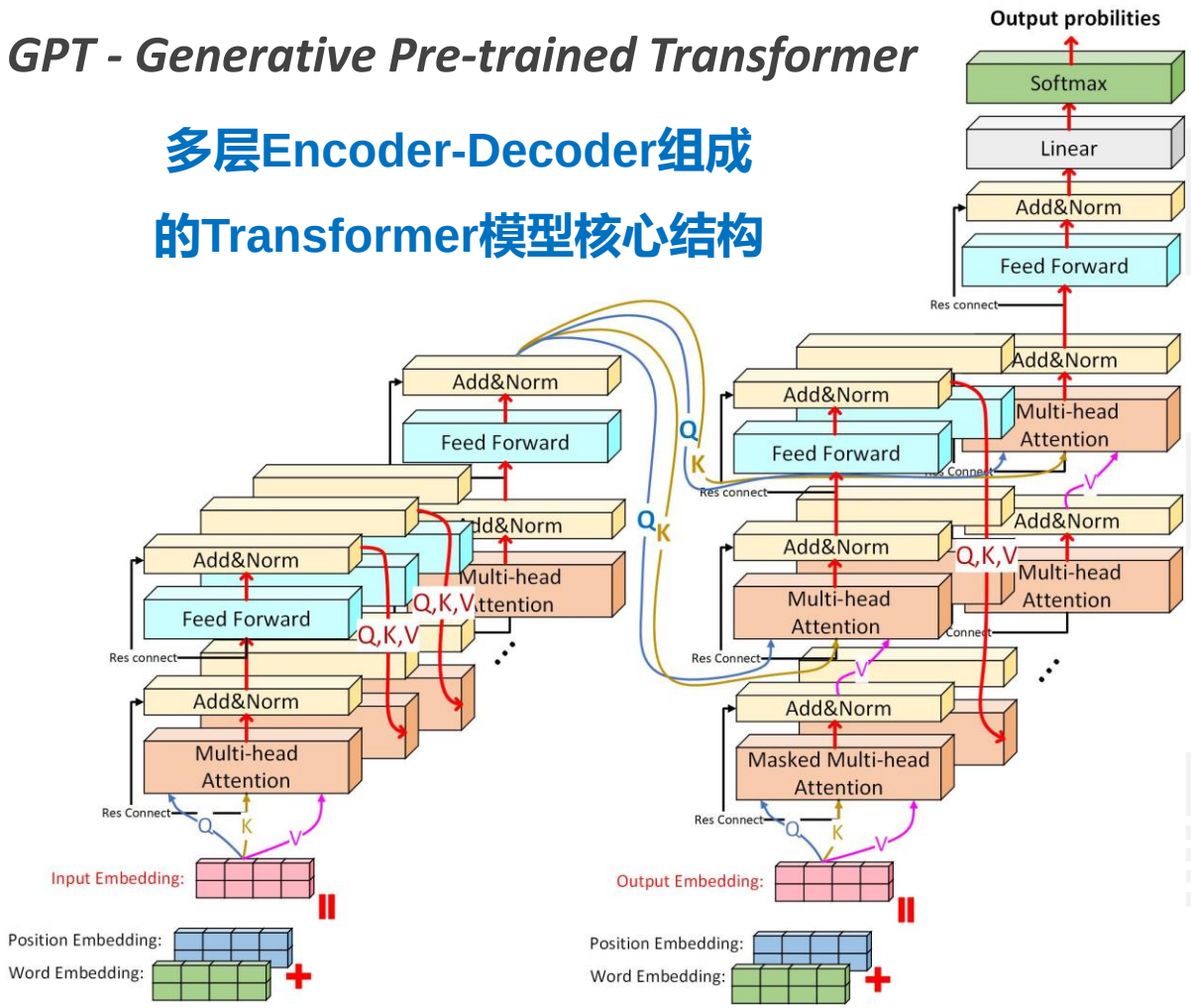
# 当前AI大模型以Transformer为基干网络 (以GPT为例)

- Decoder-Encoder层数、Token数量、掩码Mask尺寸、特征矩阵尺寸急剧增大

*GPT - Generative Pre-trained Transformer*

**多层Encoder-Decoder组成**

**的Transformer模型核心结构**




输入输出层

Input / Output Embedding

Positional Encoding

自注意力层

Multi-Head Attention

Masked Multi-Head Attention

归一化&前馈层

Add & Norm

Feed Forward

概率输出层

Linear Norm

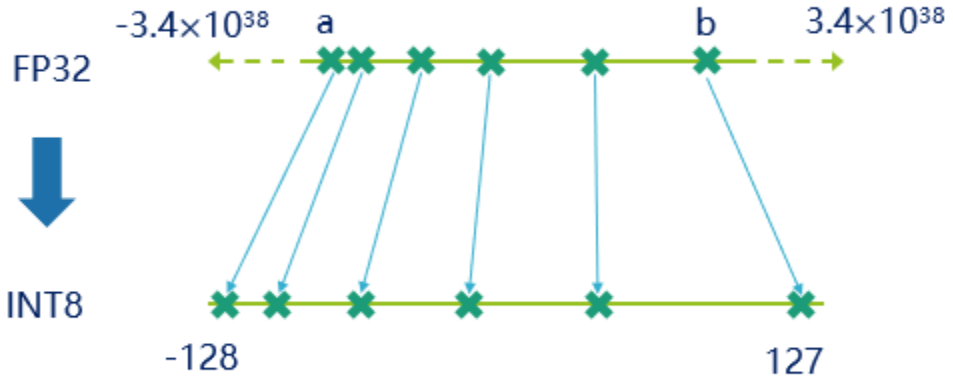
Softmax

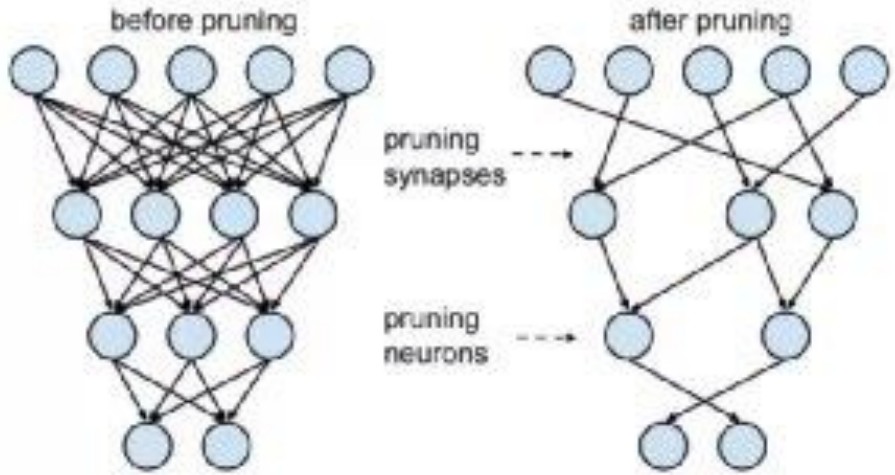
微软/OpenAI提出了**LongNet**，将Transformer的

**Token数提高到了10亿级别，并持续提升**

# 软硬件协同设计

- 面对复杂应用，单纯的硬件设计已经不足以支持性能需求，需要软硬件协同设计



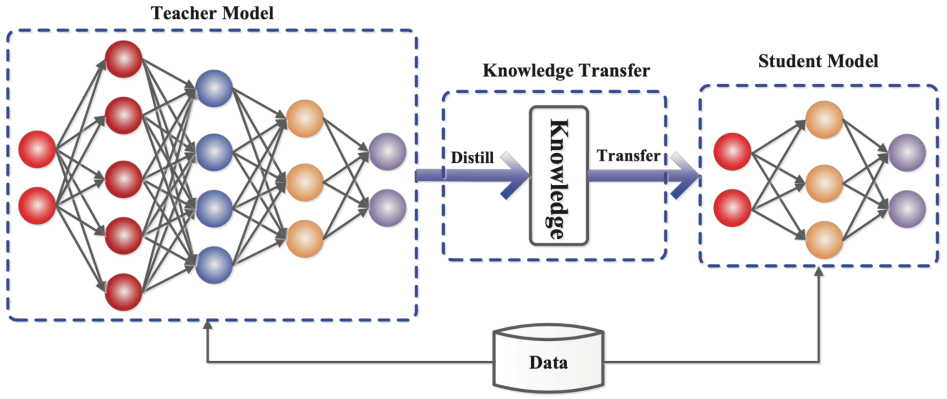




**模型量化**：将高精度的权重量化为低精度的权重，以一定的精度损失为代价换取更小的存储和计算开销

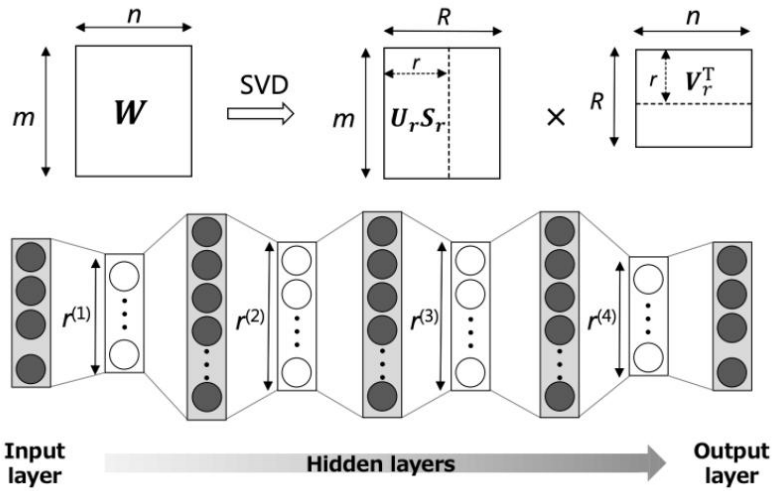
**模型剪枝**：将神经网络中重要性较小的神经元和权重删除，减少计算量，加速神经网络推理

# 软硬件协同设计

- 面对复杂应用，单纯的硬件设计已经不足以支持性能需求，需要软硬件协同设计



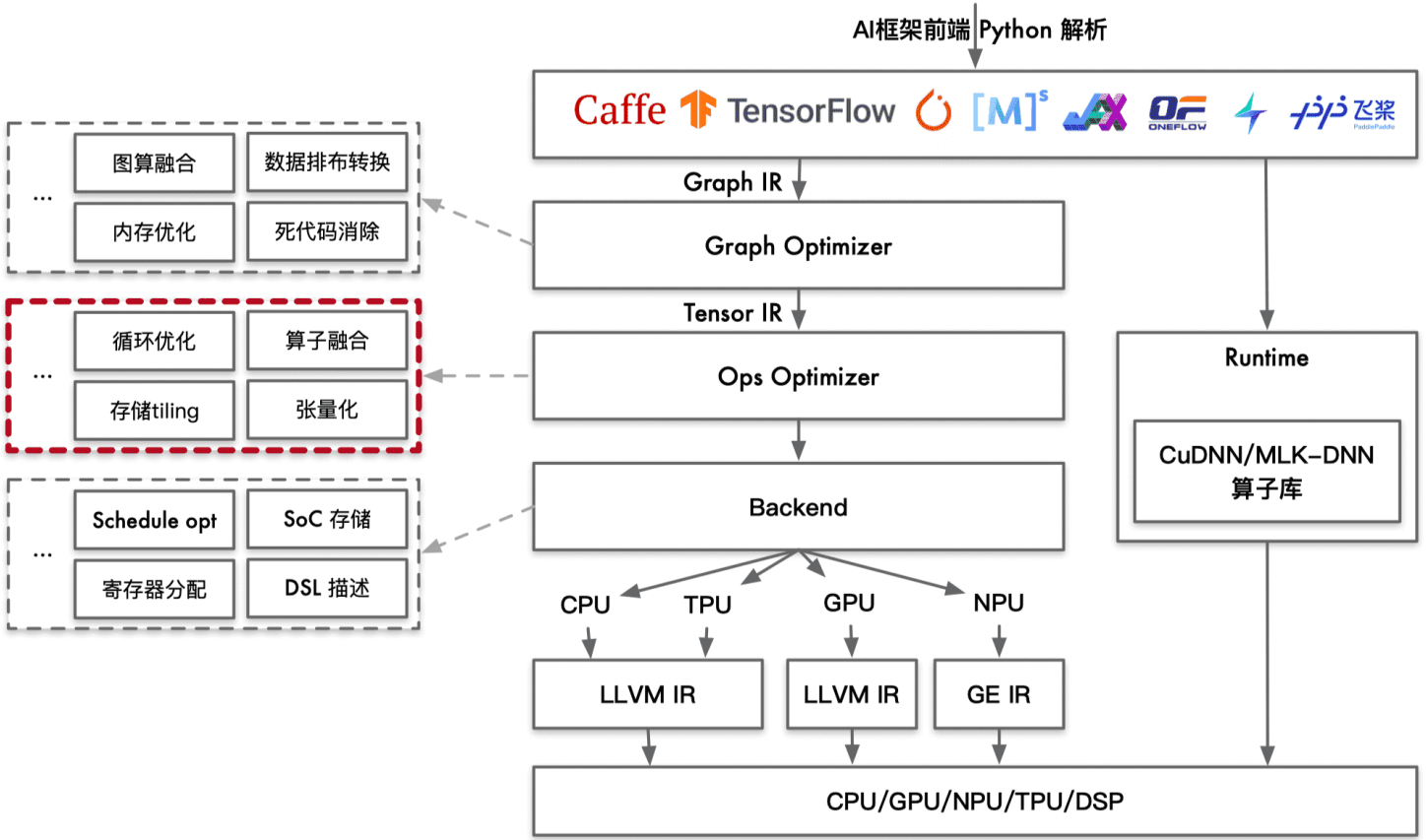

**知识蒸馏**：将规模较大的模型作为teacher model训练一个较小的student model，在尽可能保证性能的情况下减小模型规模




**低秩分解**：将大规模权重分解为两个小规模的权重矩阵相乘（SVD），减小矩阵向量乘的计算量

# 软硬件协同设计

- 编译层面优化




**在程序编译过程中**

**对算子、存储tiling**

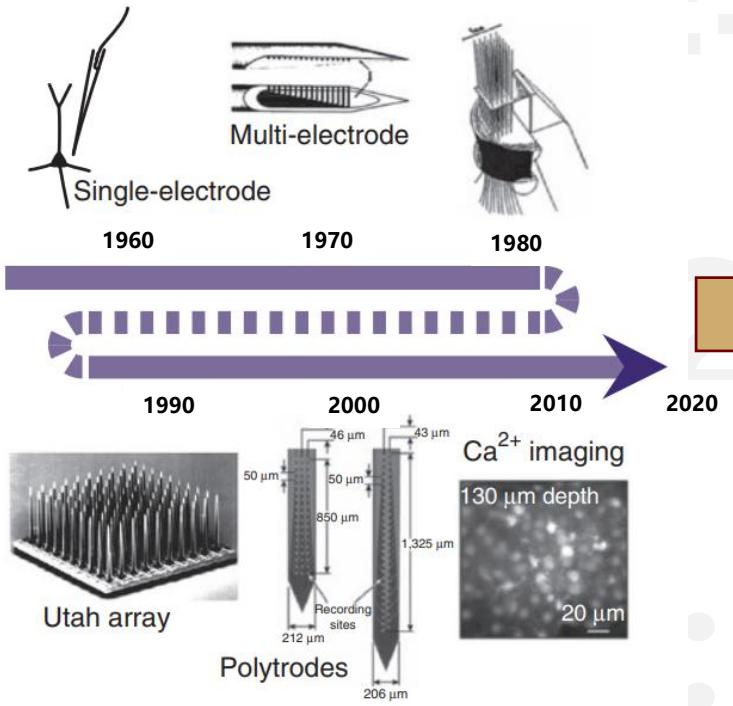
**和寄存器分配等等**

**方面进行优化**

# 代表性新兴技术 – 新计算：脑机接口芯片与系统

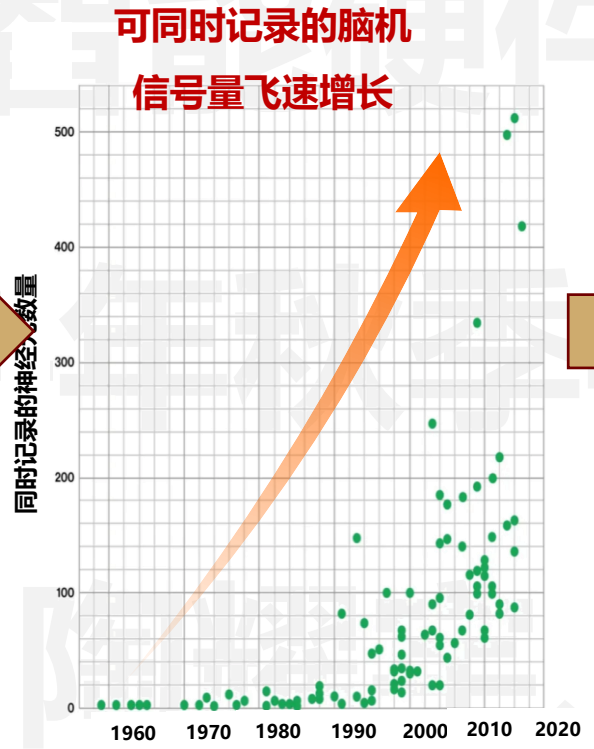
- 为脑机接口服务的芯片与系统将在未来数十年成为人类发展的方向之一

## 脑机数据采集工具的发展




Stevenson, I., et al. Nat Neuroscience

## 脑机数据量的“摩尔定律”

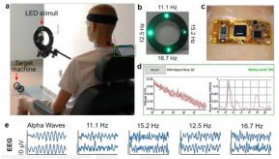



Stevenson, I., et al. Nat Neuroscience

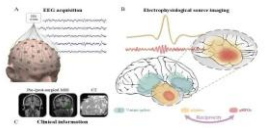
## 脑机打字

## 脑控无人机

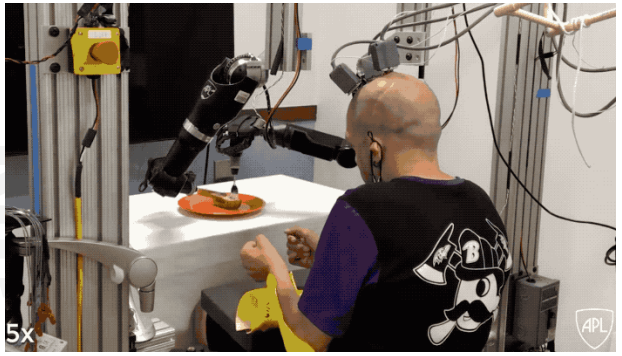
## 脑控座椅



## 脑机癫痫监测



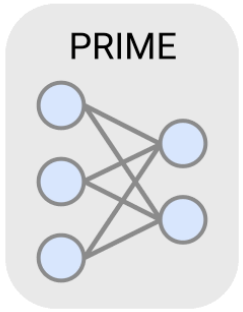
## 脑控机械臂



# 代表性新兴技术 – 新方法：AI设计AI芯片

- 设计AI芯片架构 -> 利用AI设计AI芯片架构

参数化硬件单元库　　针对某类任务的最优芯片设计




Optimizer

RL、大模型等方式

设计AI芯片的AI模型

AI芯片