北京大学
PEKING UNIVERSITY

智能硬件体系结构

第二讲：智能芯片发展史与未来趋势

主讲：陶耀宇、李萌

2025年秋季

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**

指标	课程信息
课程号	04632042
学分	2
课程体系	专业任选
地址	二教422
优秀率	无强制限制
考核方式	出勤（5%）、3次课后作业（10%+10%+10%） 简单硬件编程实验（Lab 1 15% + Lab 2 25% + Lab 3 25%）

扫描二维码：加入智能硬件体系结构（校内）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（将通过课程逐步进行教学）

课程网站：

https://aiarchpku.com

推荐教科书：

- **逻辑电路方面**：
  - Digital Integrated Circuits: A Design Perspective - Anantha P. Chandrakasan, Borivoje Nikolic, and Jan M. Rabaey
  - 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 - David A Patterson and John L. Hennessy
  - 智能计算系统 - 陈云霁；人工智能芯片设计 - 尹首一

目 录

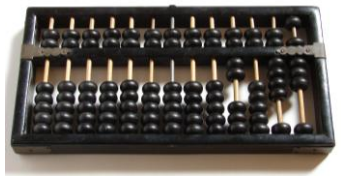
CONTENTS

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

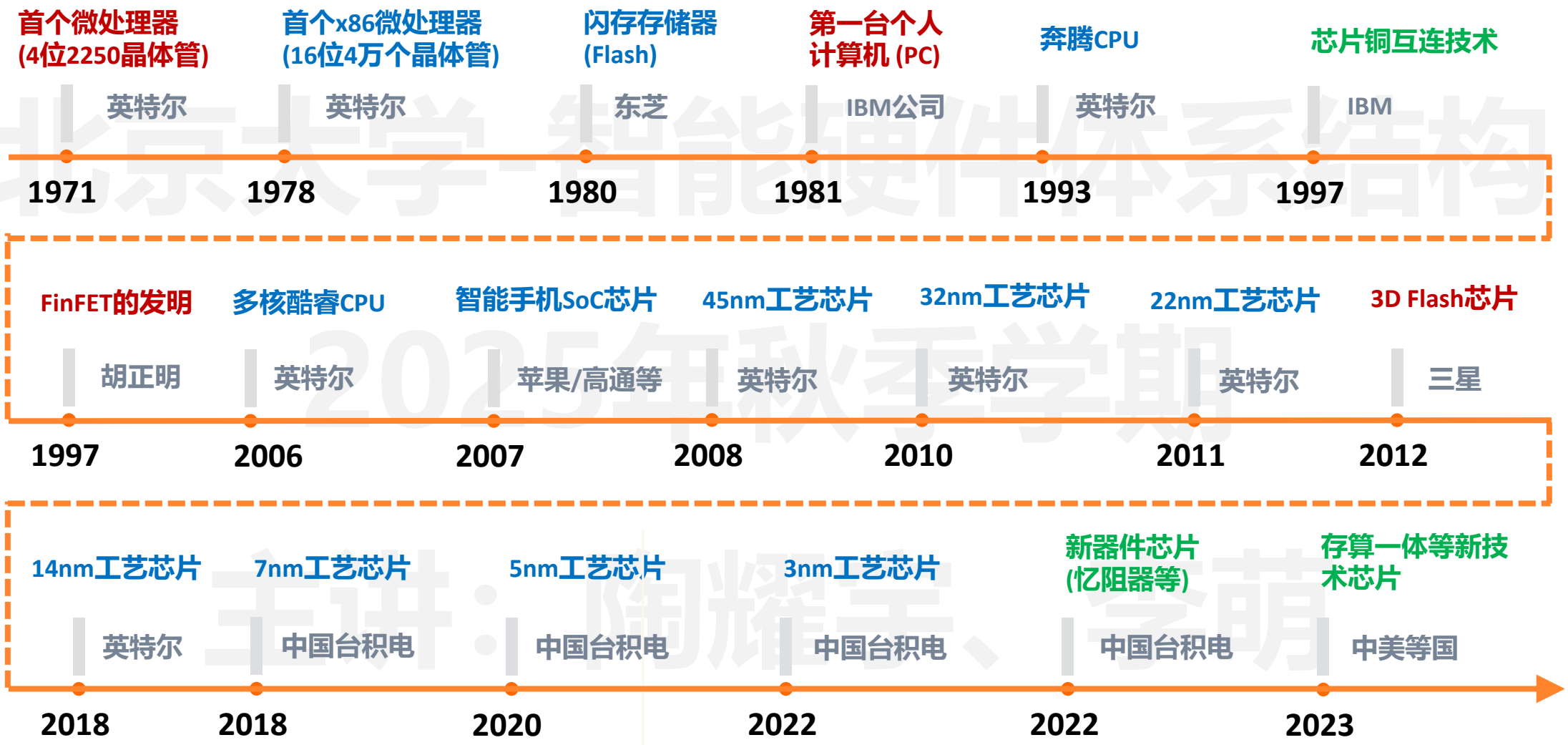
2030以后

波澜壮阔的智能芯片发展史

- 智能芯片的发展历史 (1833 - 1968)

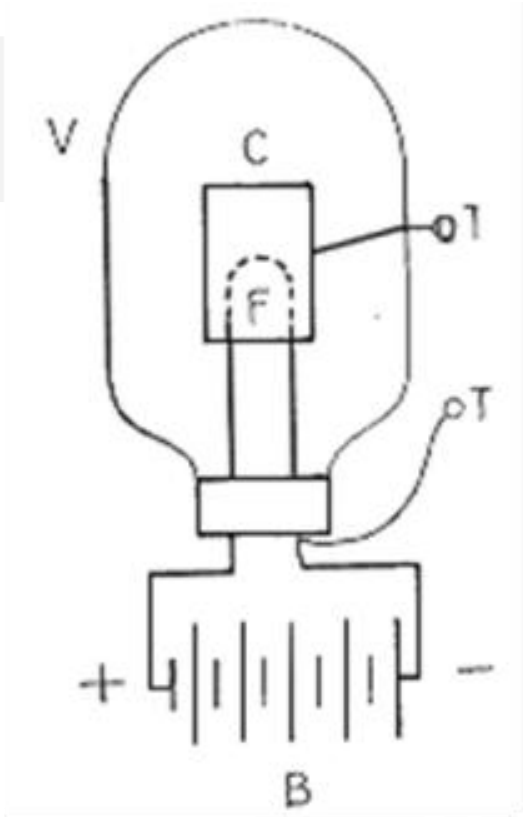
波澜壮阔的智能芯片发展史

- 智能芯片的发展历史 (1968 - 2023)



前半导体时代的霸主：真空二极管（电子管）- 1904年

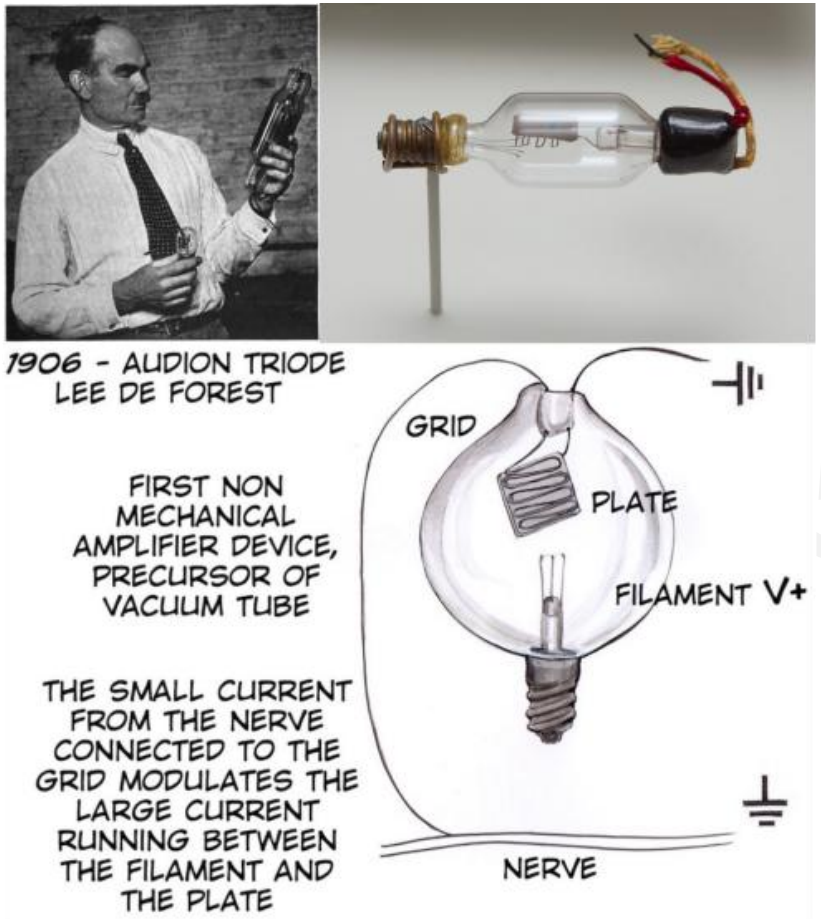
- 爱迪生的前雇员约翰弗莱明发明了可以真空二极管，也叫弗莱明管



- 1873年，弗雷德里克·格思里（Frederick Guthrie）发现：当加热一个接地的金属盘时，其旁边带正电的验电器会逐渐流失电荷；而当金属盘靠近带负电的验电器时，则不会有电荷流失。该现象表明加热的金属阴极可表现出单向导电特性；
- 1880年，爱迪生在未了解格思里的工作的情况下，也发现了类似的现象。这种现象是由于被加热物体的电子逸出功降低，更容易在外界电场的作用下逸散到外界导致的。

前半导体时代的霸主：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

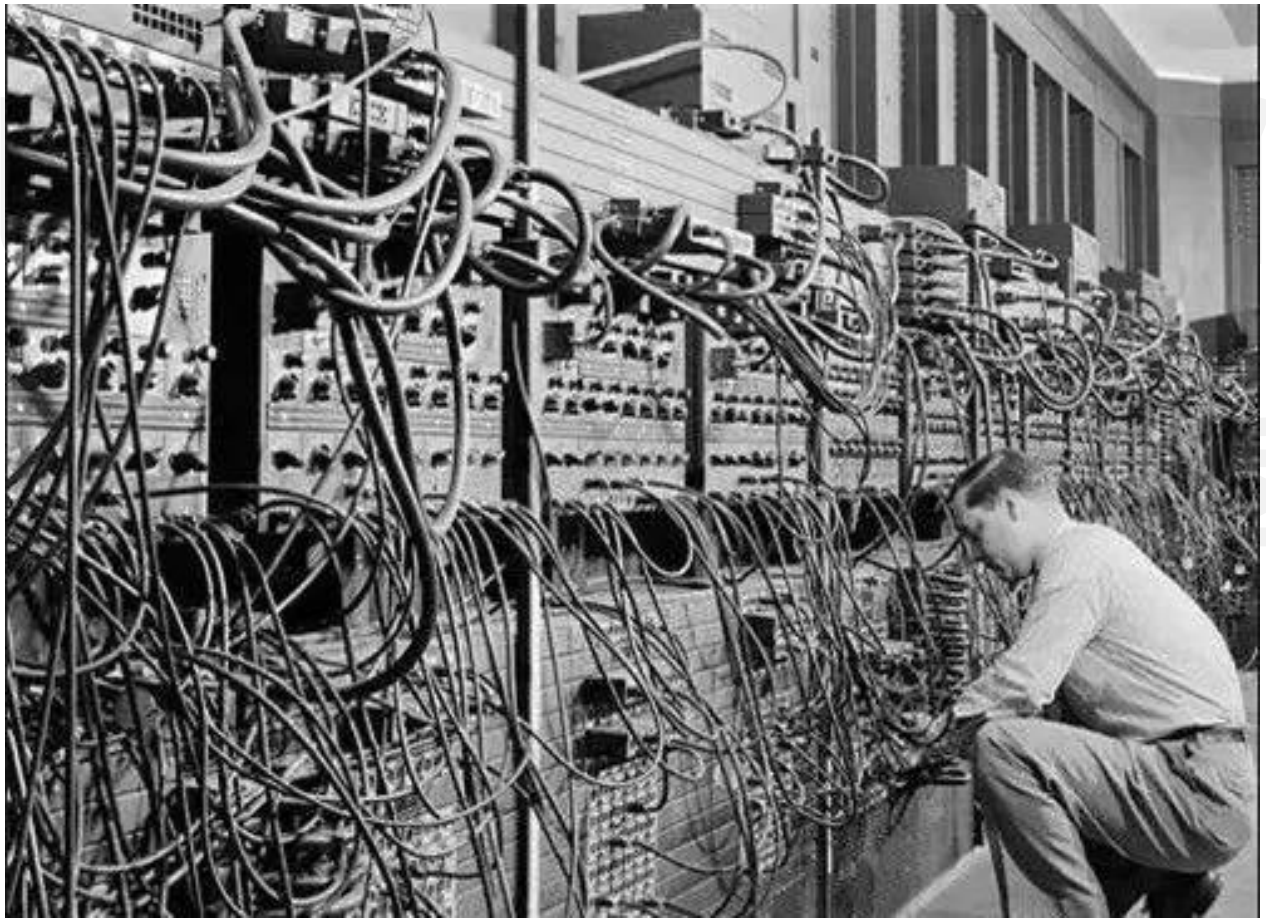


美国电子管计算机ENIAC

- 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

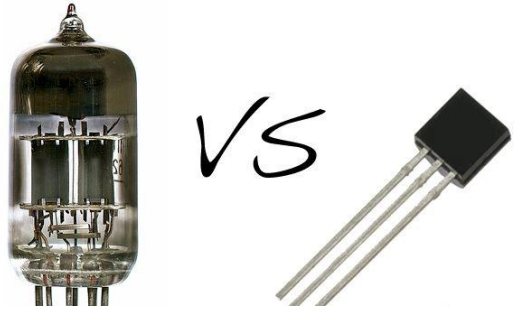
美国电子管计算机ENIAC

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

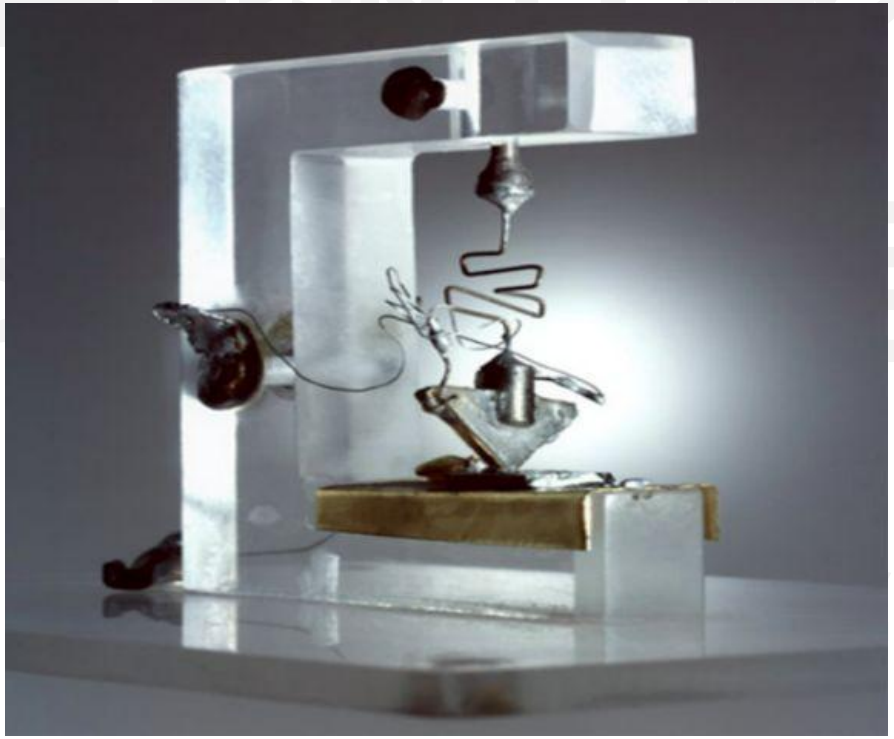
电子管在特定的军事应用领域与国防工业中仍具有一定的作用

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

重要历史节点：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

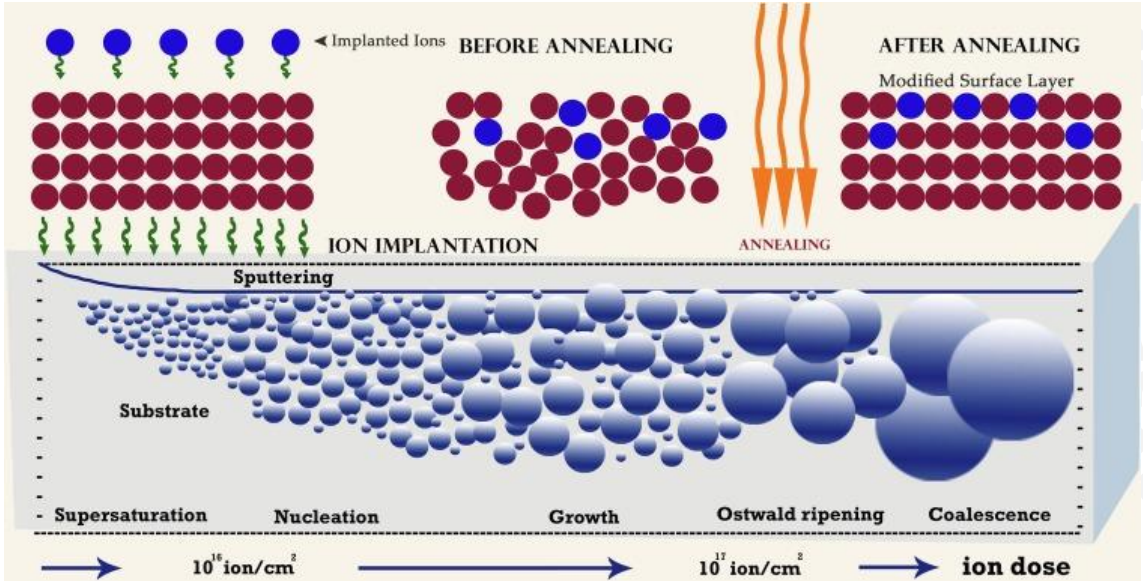
肖克利（前）、巴丁（后一）、布拉顿（后二），因为晶体管的发明，共同获得了1956年的诺贝尔物理学奖

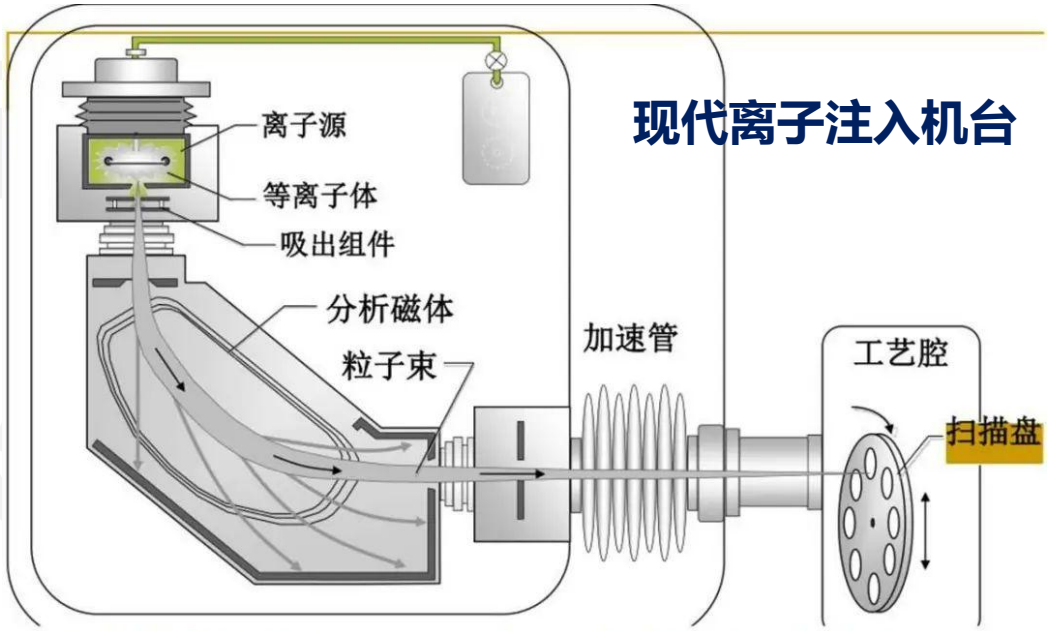


点接触式晶体管：把间距为50 μm的两个金电极压在锗半导体上，微小的电信号由一个金电极 (发射极) 进入锗半导体 (基极) 并被显著放大，然后通过另一个金电极 (集电极) 输出，这个器件在1kHz的增益为4.5

重要历史节点：离子注入工艺的发明 – 1950年

- 离子注入工艺是半导体掺杂技术的重要组成部分，也是控制晶体管阈值电压的重要手段





现代离子注入机台

肖克利首先发展了离子注入工艺

利用掺杂来调节半导体电学特性

离子注入工艺

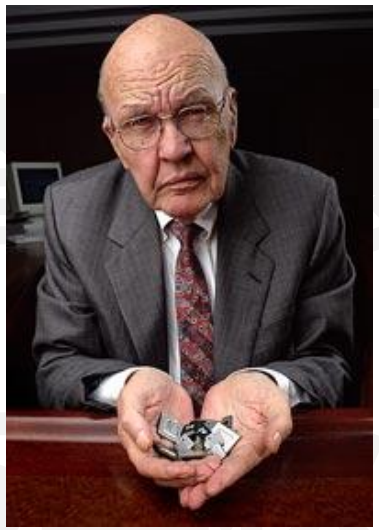
成为半导体芯片制造产业的核心技术之一

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

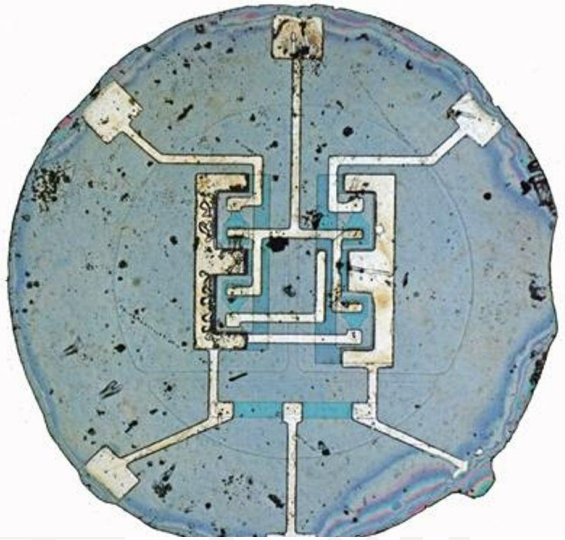
1958年8月28日世界第一块集成电路

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

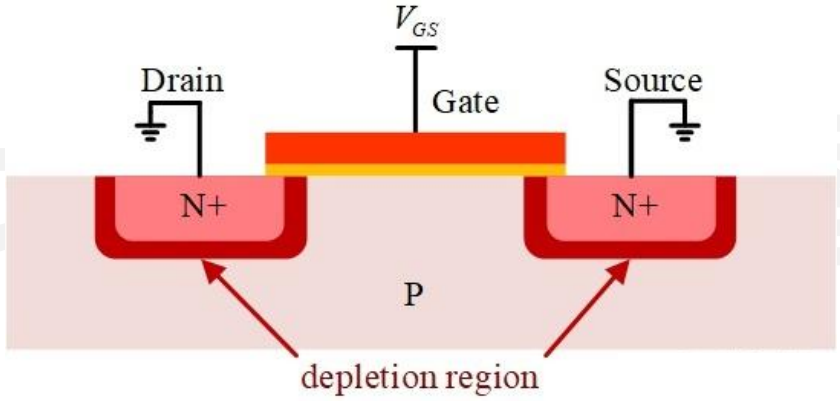
硅集成电路

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

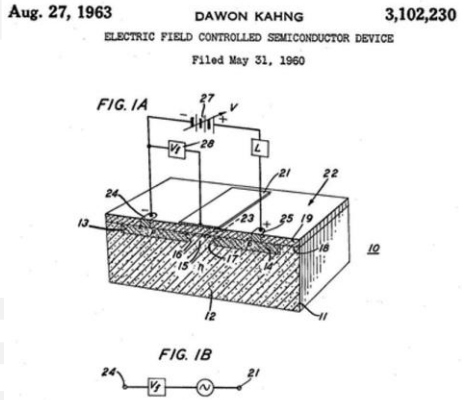
重要历史节点： MOSFET场效应晶体管的发明 – 1959年/1960年

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

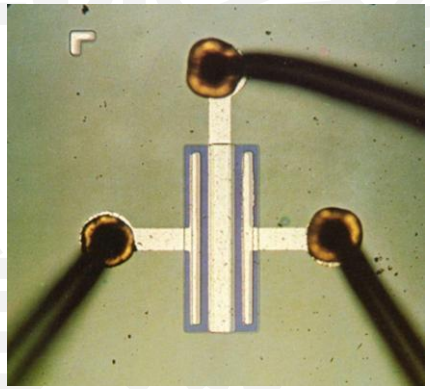


艾塔拉 (Martin Atalla) 和姜大元 (Dawon Kahng)

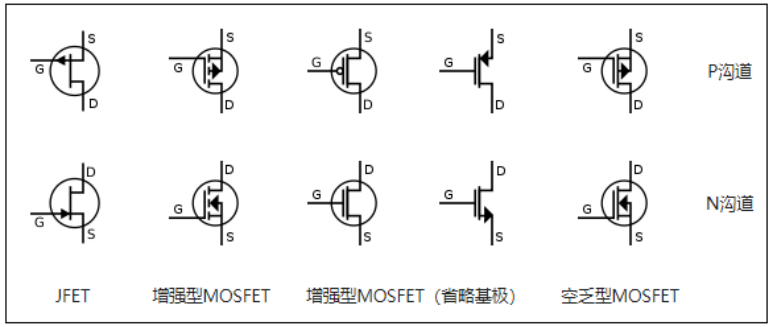
Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET)


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

1960年MOSFET专利



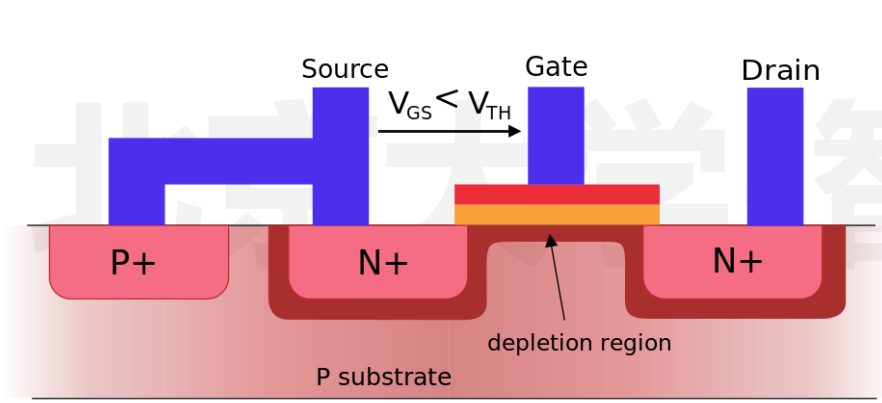
PMOS场效应晶体管实物图

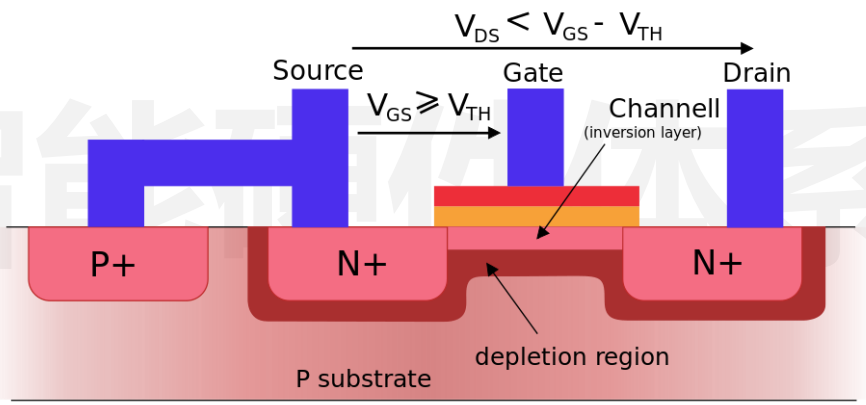


MOSFET已经成为集成电路的基本组成单元

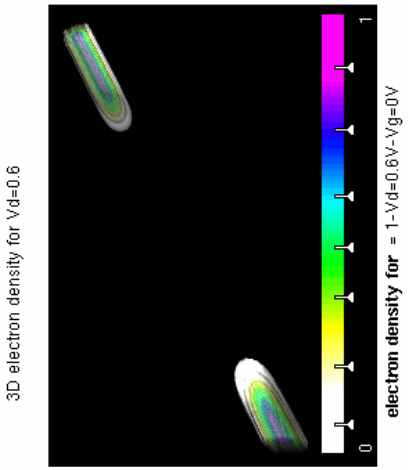
重要历史节点： MOSFET晶体管工作原理 – 1959年/1960年

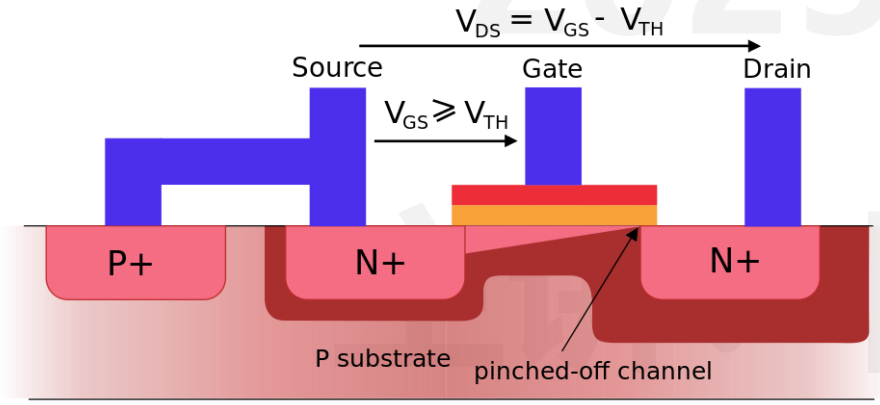
- MOSFET有三个工作区间：断开、线性（欧姆区间）、饱和（电压不随电流线性增加）



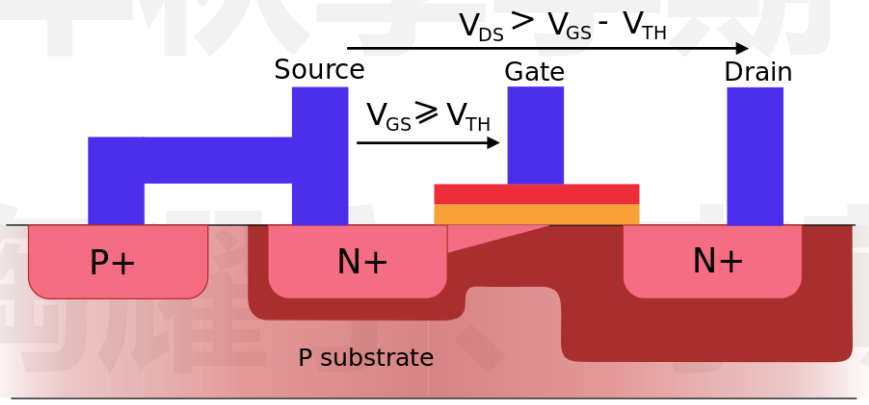


Linear operating region (ohmic mode)

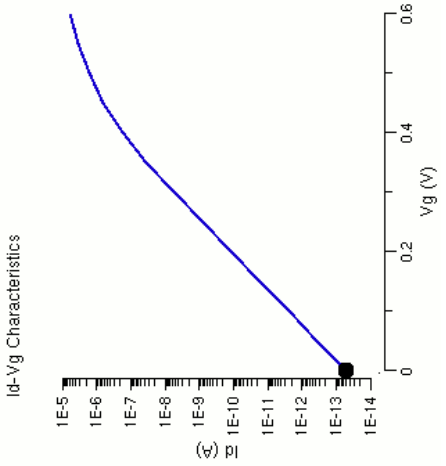




Saturation mode at point of pinch-off



Saturation mode

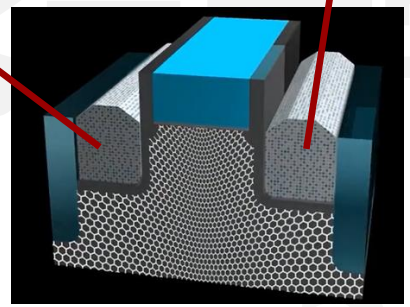


重要历史节点：外延/光刻等关键制造工艺 – 1960年

- 外延/光刻等关键制造工艺起源于1960年，已成为半导体芯片制造领域的王冠

外延生长

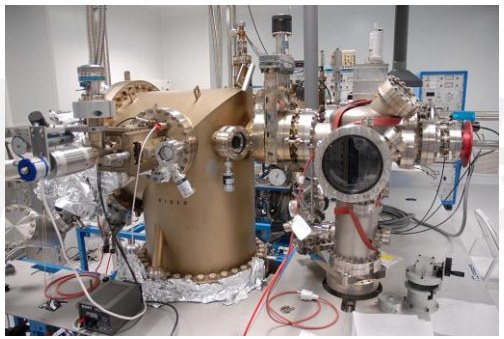
外延生长

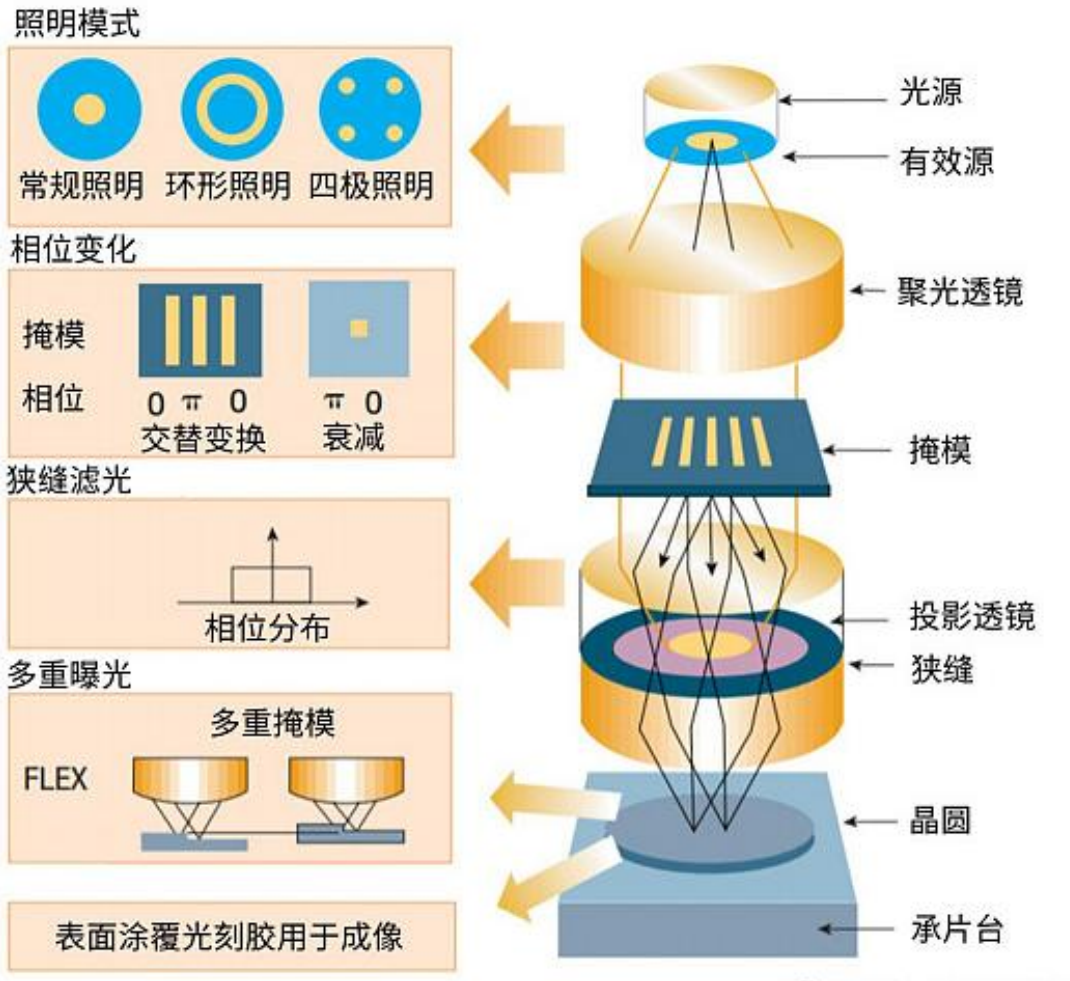


半导体器件制造过程中，在原有芯片上长出新结晶以制成新半导体层的技术

化学气相沉积（CVD）

分子束外延 (MBE)

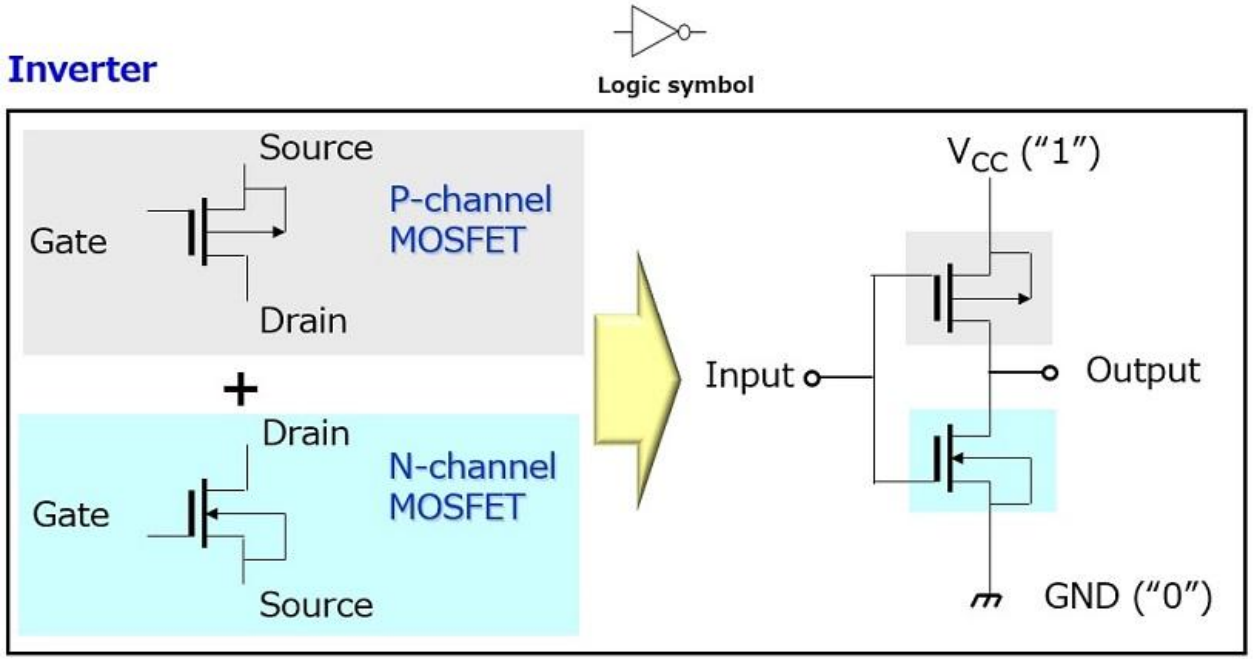




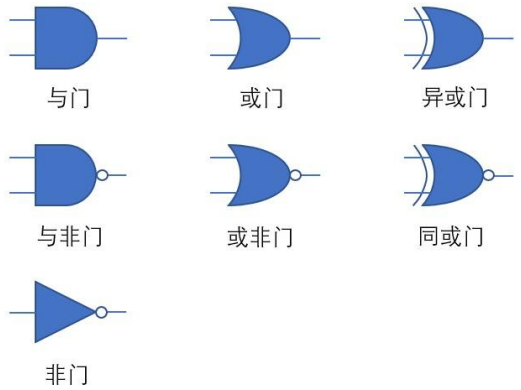
现代光刻技术示意图 (用光刻蚀山脊、走线等)

重要历史节点： CMOS电路的发明 – 1963年

- 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

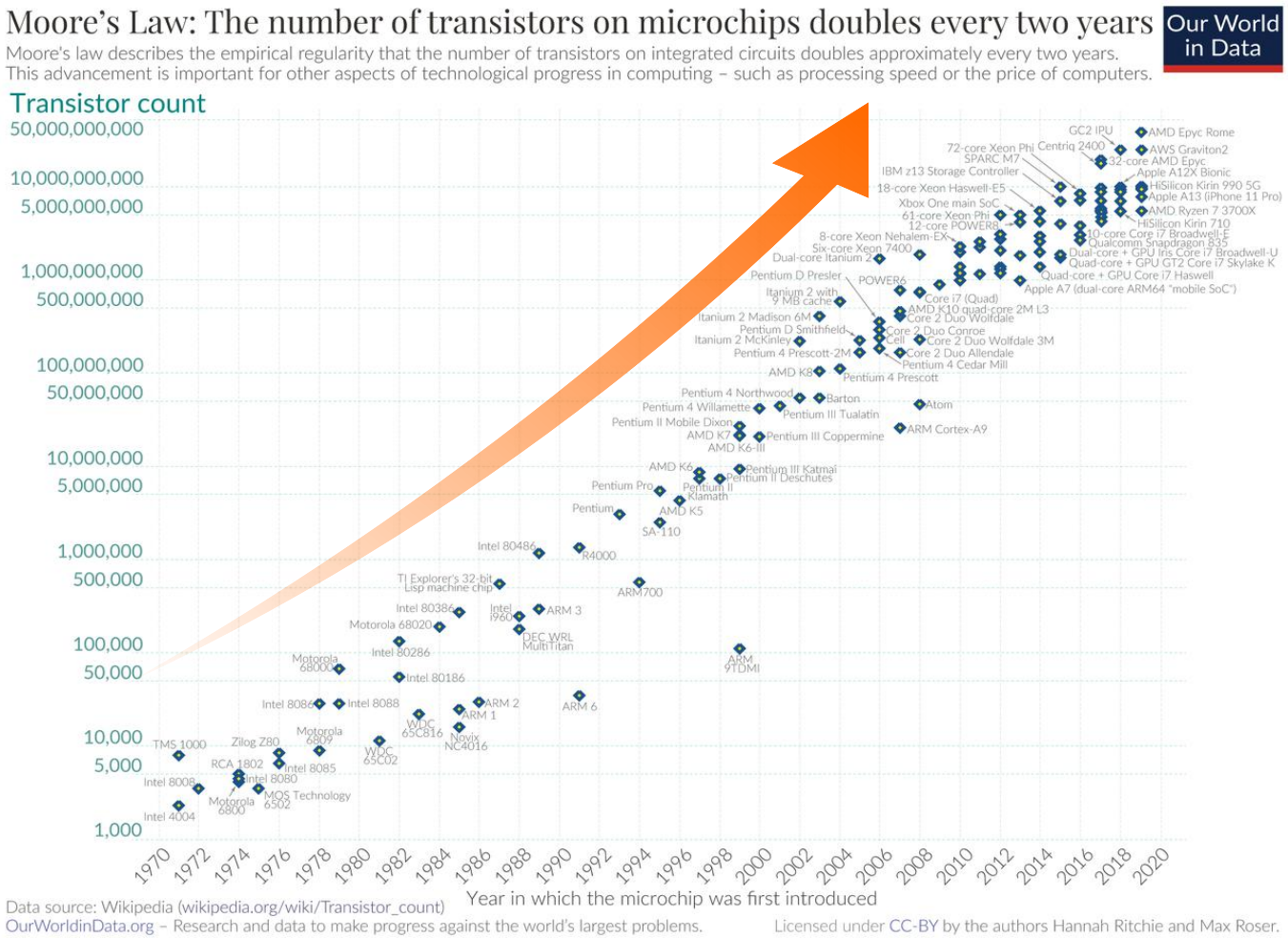


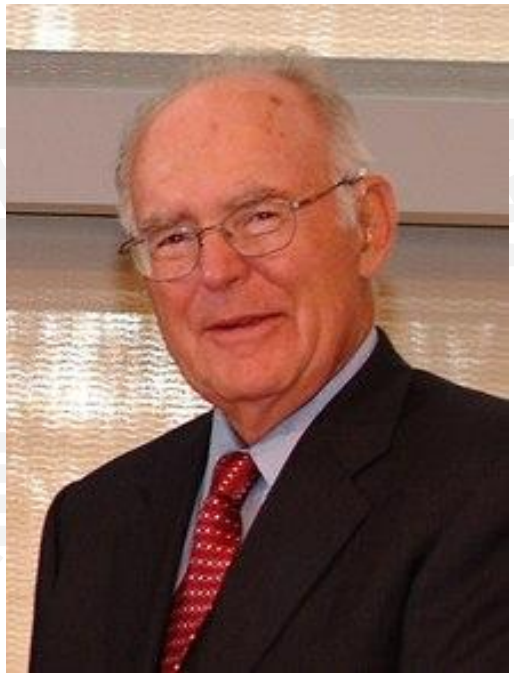
互补式金属氧化物半导体具有只有在晶体管需要切换启动与关闭时才需消耗能量的优点，因此非常节省电力且发热量少，且工艺上也是最基础而最常用的半导体器件

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：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，

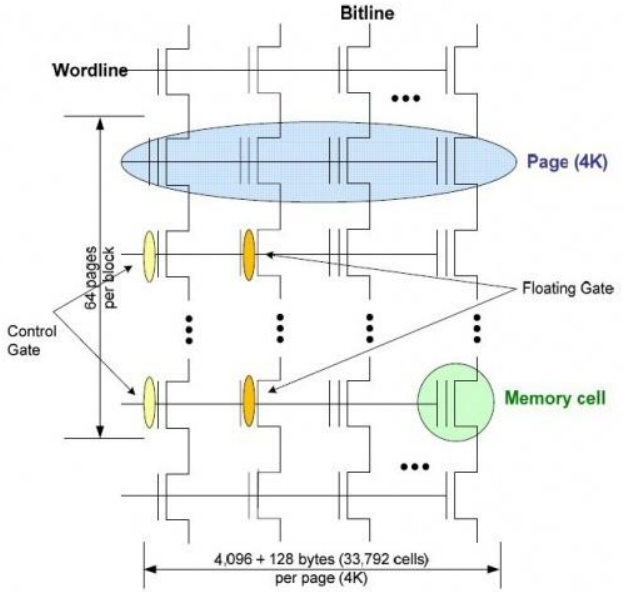
每隔两年便会增加一倍

重要历史节点：非易失性存储器Flash的发明 – 1967年

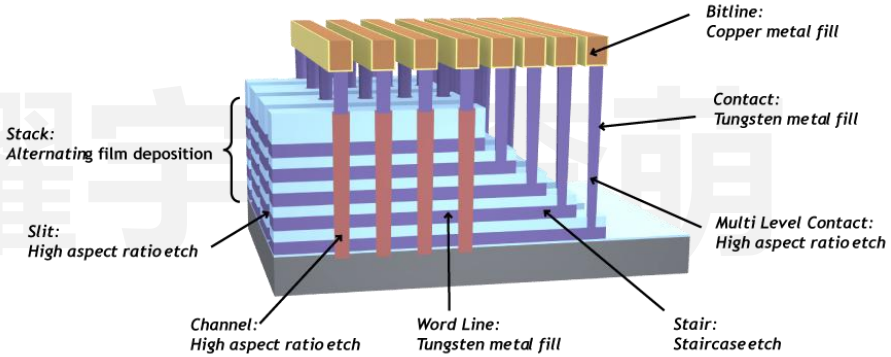
- 除了计算场景之外，存储也是占据半导体芯片重要份额的典型应用场景

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）



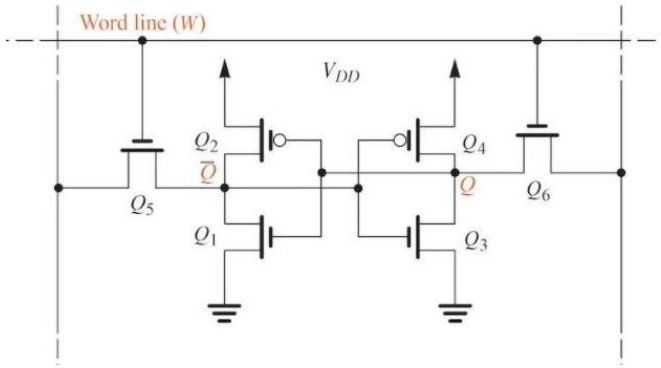
传统平面型NAND Flash非易失性存储器



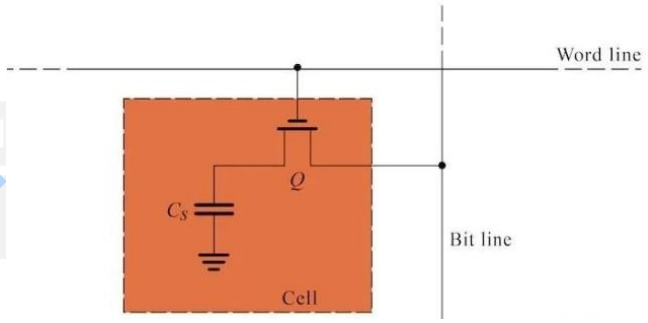
现代三维堆叠NAND Flash

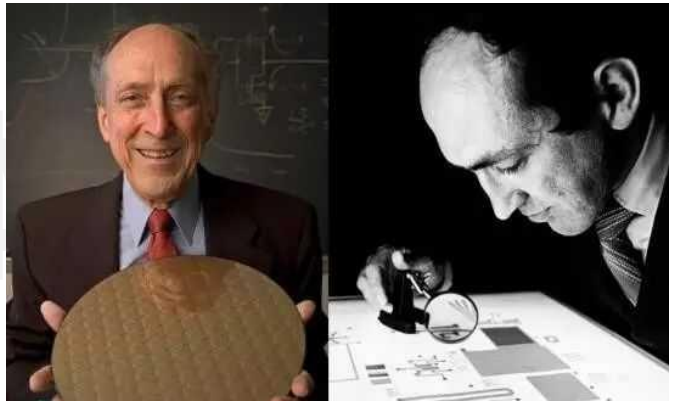
重要历史节点：易失性存储器DRAM的发明 – 1968年

- SRAM/DRAM是两种最常用的易失性存储器件，广泛应用于现代半导体芯片中



SRAM需要6个CMOS晶体管来存储数据





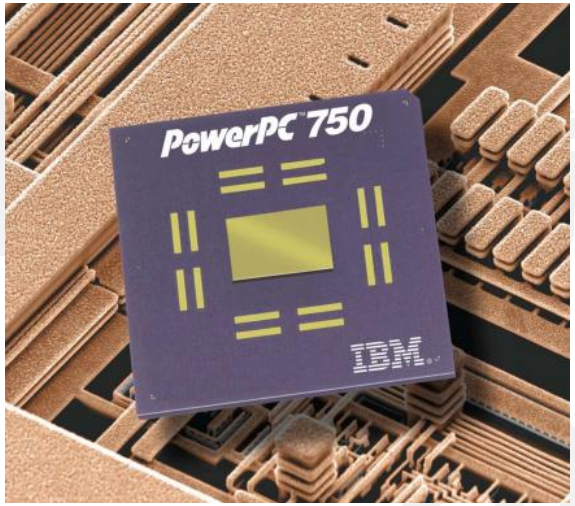
罗伯特·丹纳德发明了DRAM (**动态随机存取存储器**) 存储器

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。

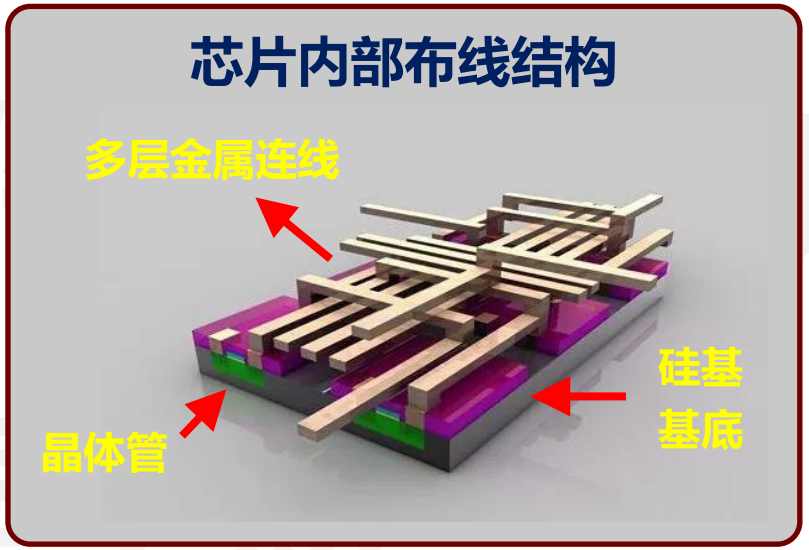
与SRAM相比，DRAM的优势在于结构简单—**每比特都只需一个电容跟一个晶体管来处理**，相比之下在SRAM上一个比特通常需要六个晶体管。 正因这缘故，**DRAM拥有非常高的密度，单位体积的容量较高因此成本较低**。 但相反的，DRAM也有访问速度较慢，耗电量较大的缺点。

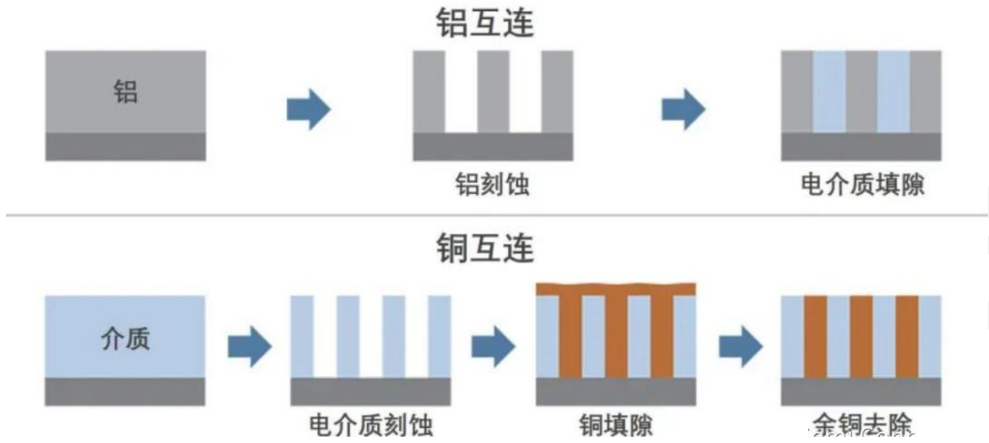
重要历史节点：半导体芯片的铜互连技术 – 1997年

- 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%

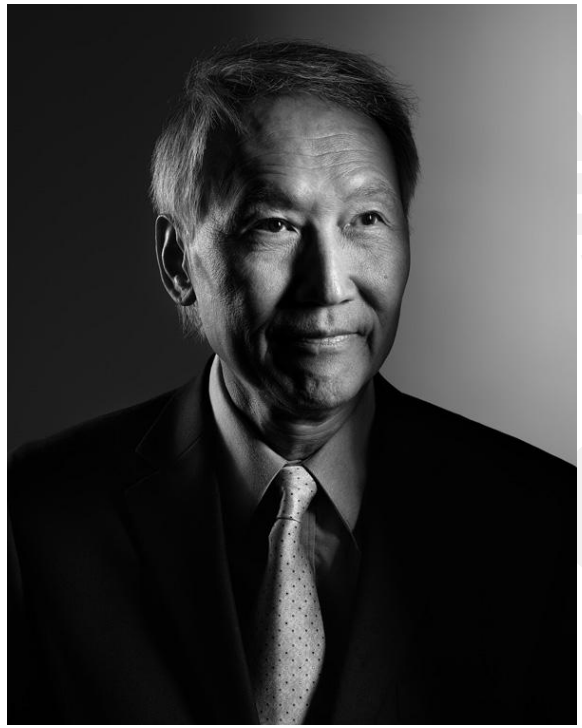




集成电路金属互连线制造工艺达到纳米级后，因为超高纯铜具有更佳的电阻率和抗电迁徙能力，很快高纯铜就替代超高纯铝合金成为金属互连线的主要材料

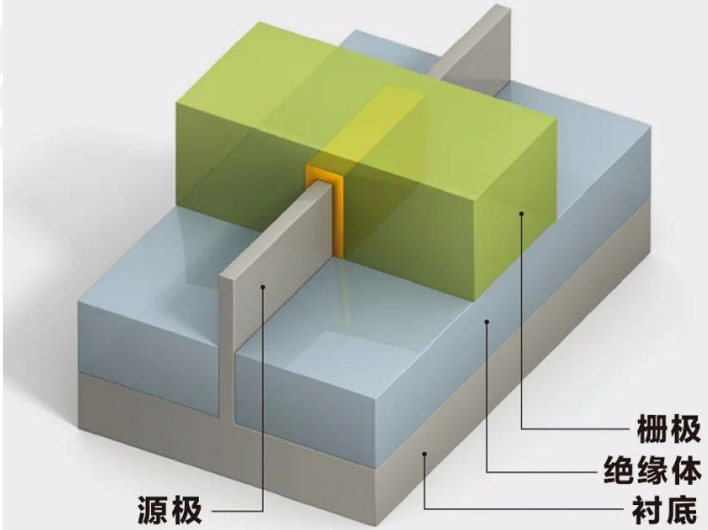
拯救摩尔定律的发明：鳍式三维晶体管FinFET – 1999年

- 原本预计2010年后，传统CMOS工艺技术在20nm走到尽头，胡正明的发明拯救了摩尔定律

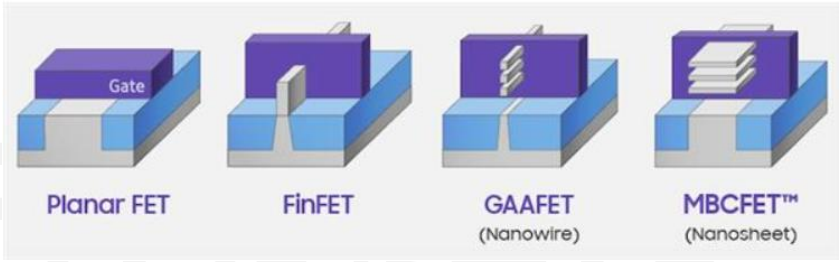


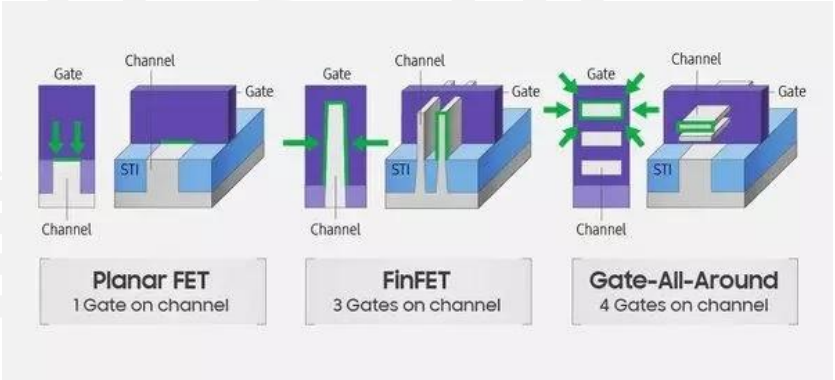
加州大学伯克利分校的胡正明教授

(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。

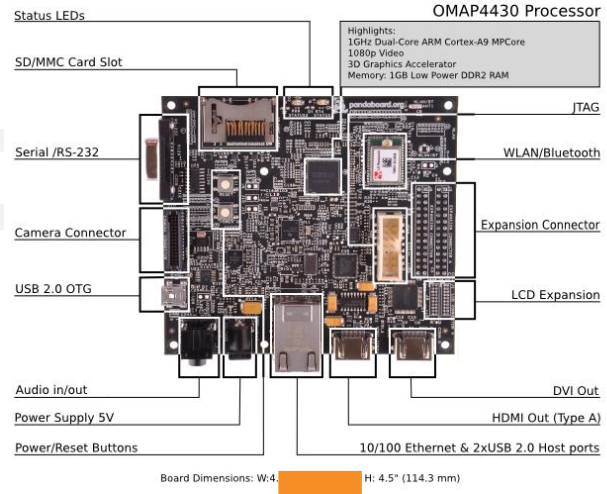




由FinFET演化出多种三维晶体管构型，推动制程向3nm/1nm演进

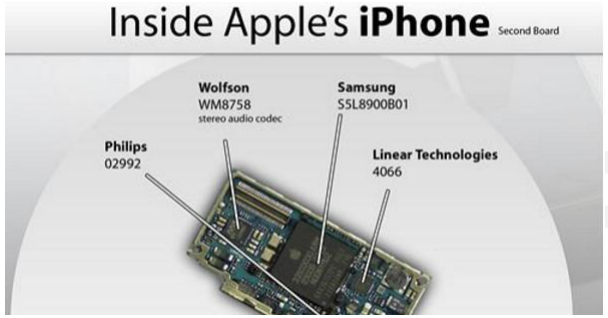
推动移动互联网飞速发展：移动SOC芯片 – 2007年至今

- 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



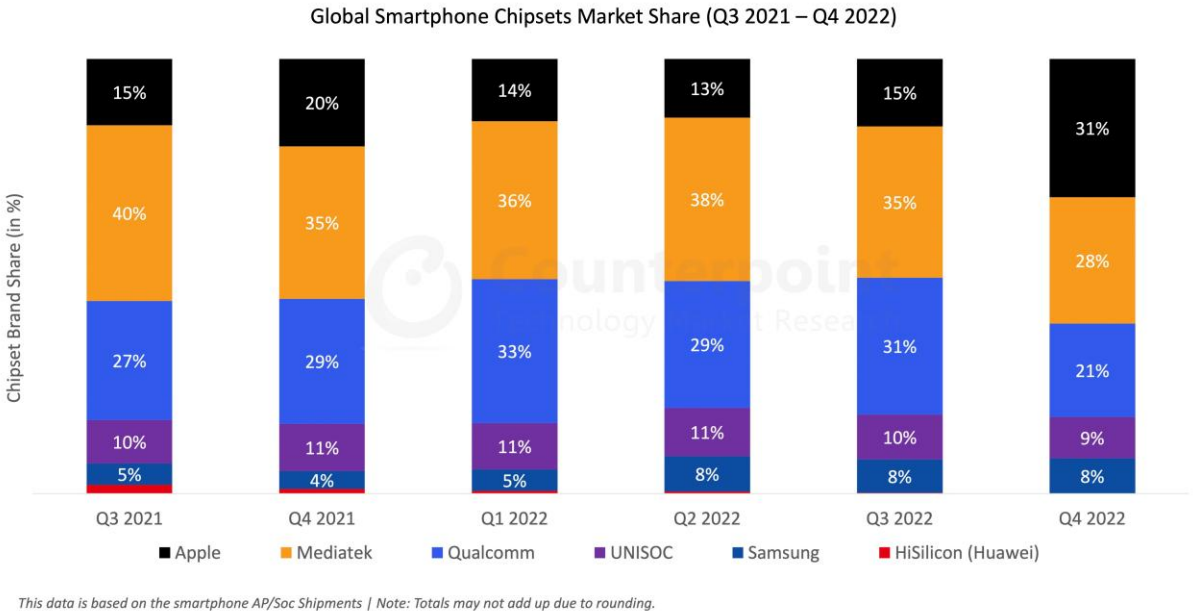
德州仪器OMAP手机芯片

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

北京大学 PEKING UNIVERSITY

- 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 6nm		3nm		2nm
三星		28nm		22nm	14nm		10nm	8nm	7nm EUV 6nm	5nm	3nm			
英特尔	22nm			14nm		14nm +	14nm ++		10nm	10nm +	7nm 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。

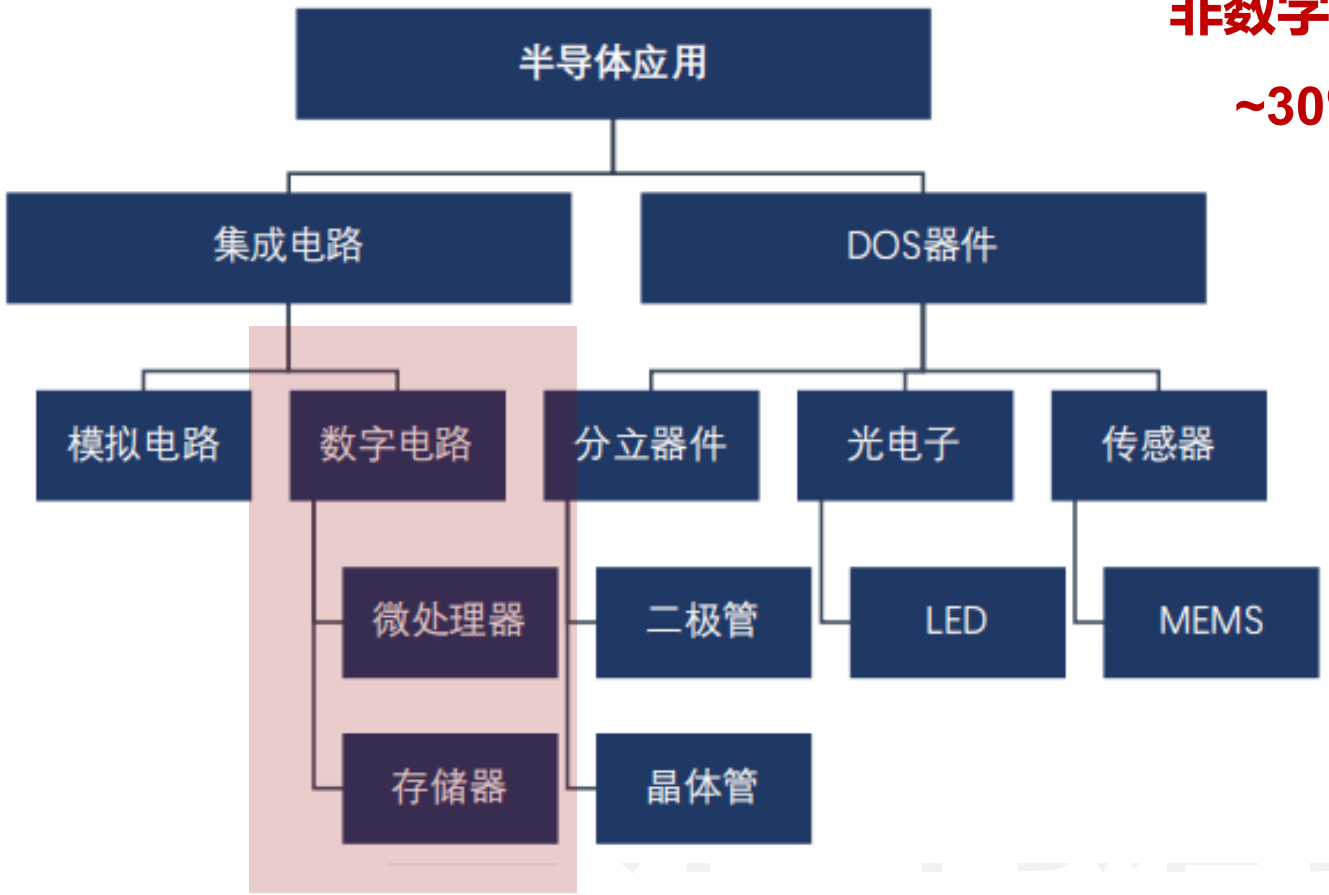
中国台湾积体电路公司后来追上，超越英特尔与三星

目 录

CONTENTS

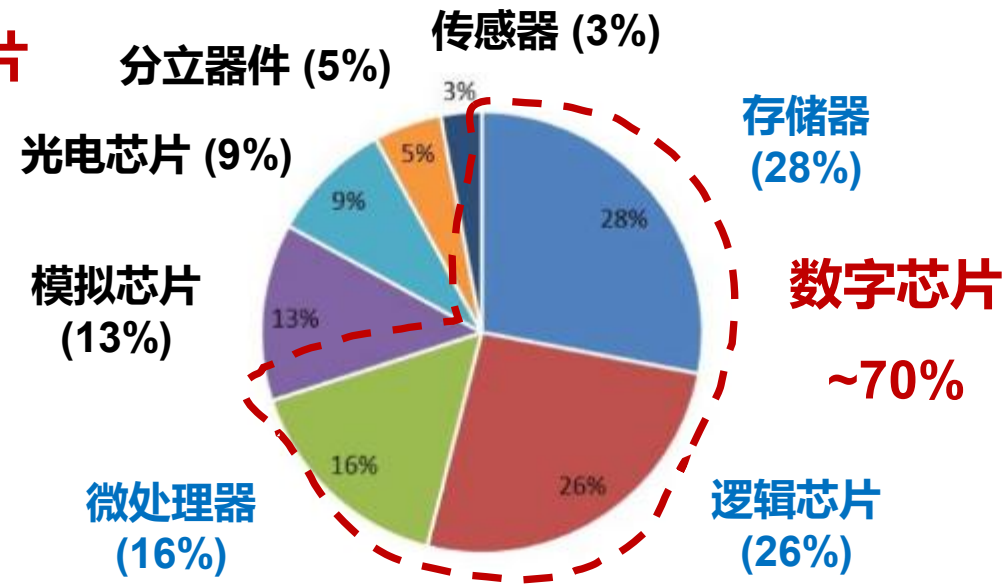
半导体芯片产业按半导体应用分类

- 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

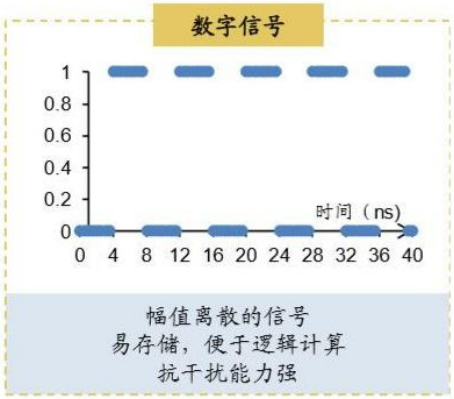


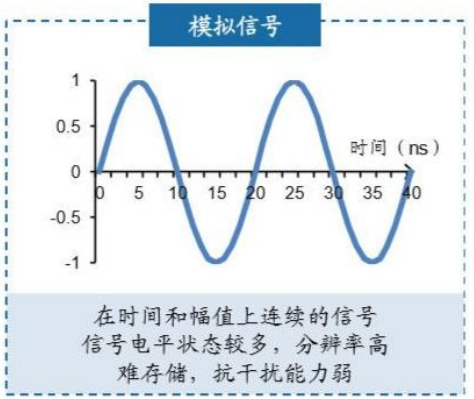
非数字芯片

~30%



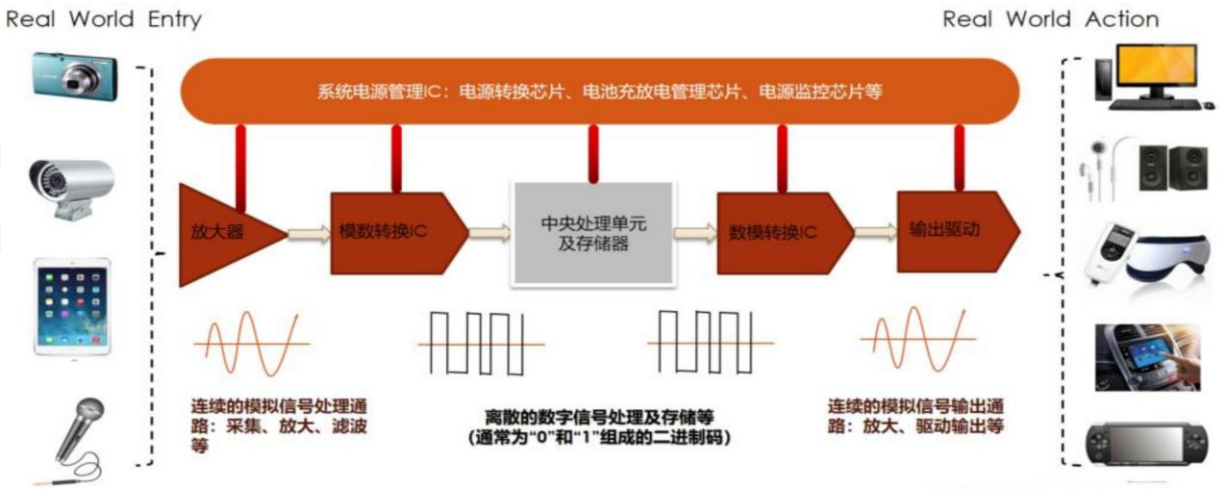
数字电路的市场份额占~70%



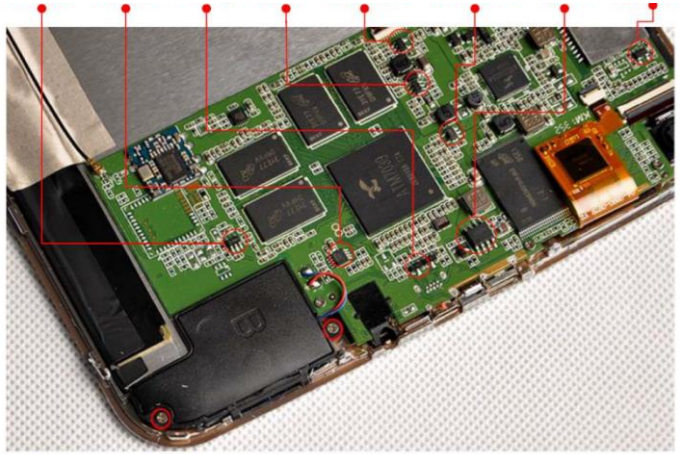


半导体芯片产业按半导体应用分类

- 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

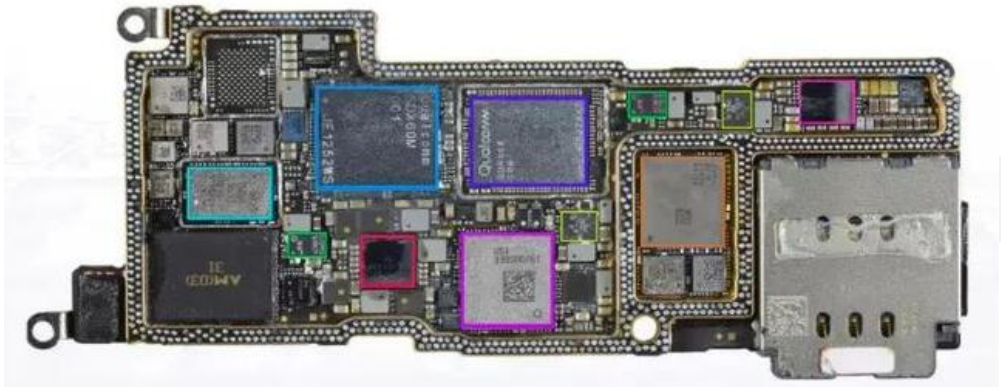


功率放大、电源管理、时钟生成、比较器、射频滤波、接口、数模转换、线性稳压等

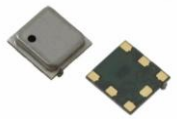


模拟芯片应用实例

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

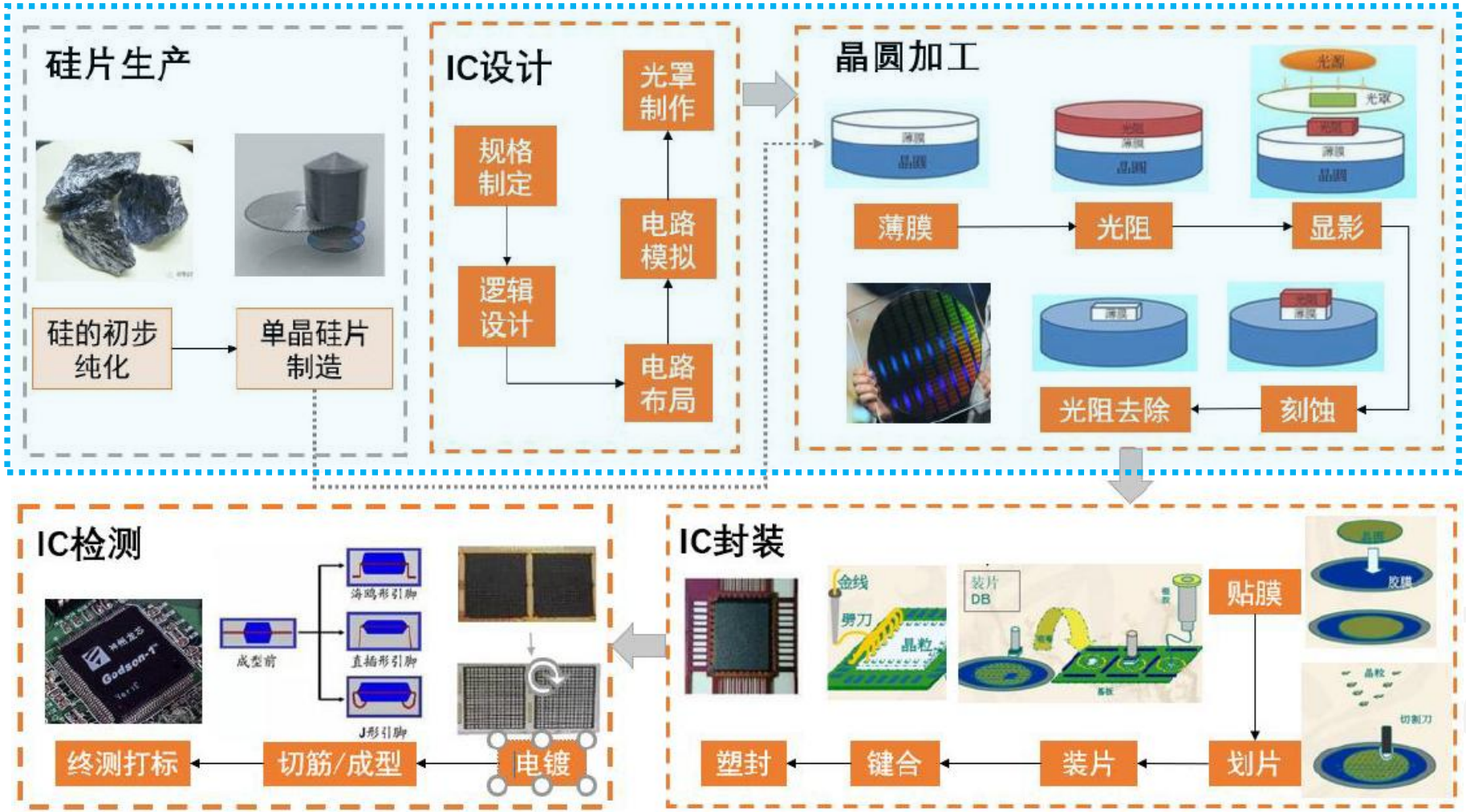
光电芯片 激光器芯片、半导体发光芯片等

分立器件 电阻、电容、电感、振荡器、晶体管、功率器件等

半导体芯片产业现状 – 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)

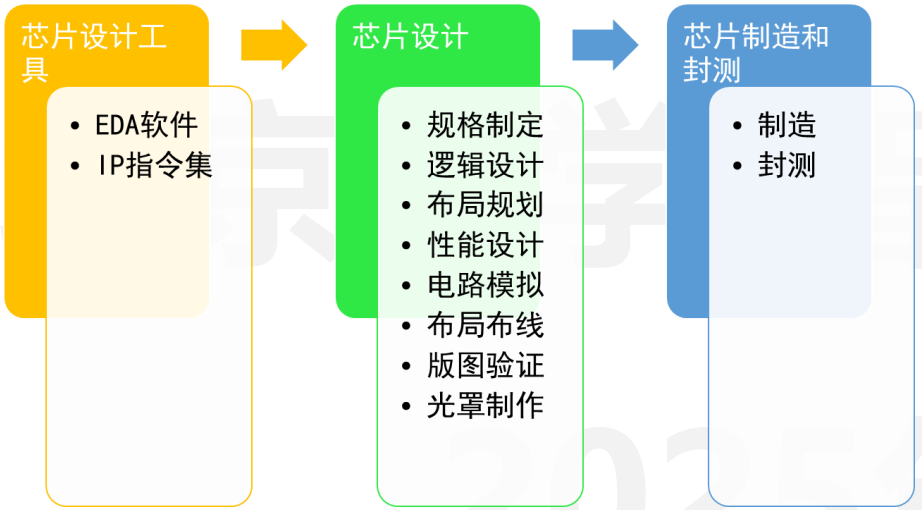
晶圆加工企业
- 台积电 (台湾)
- 三星 (韩国)
- 格芯 (美国)

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)

半导体芯片产业现状 – 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

半导体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)

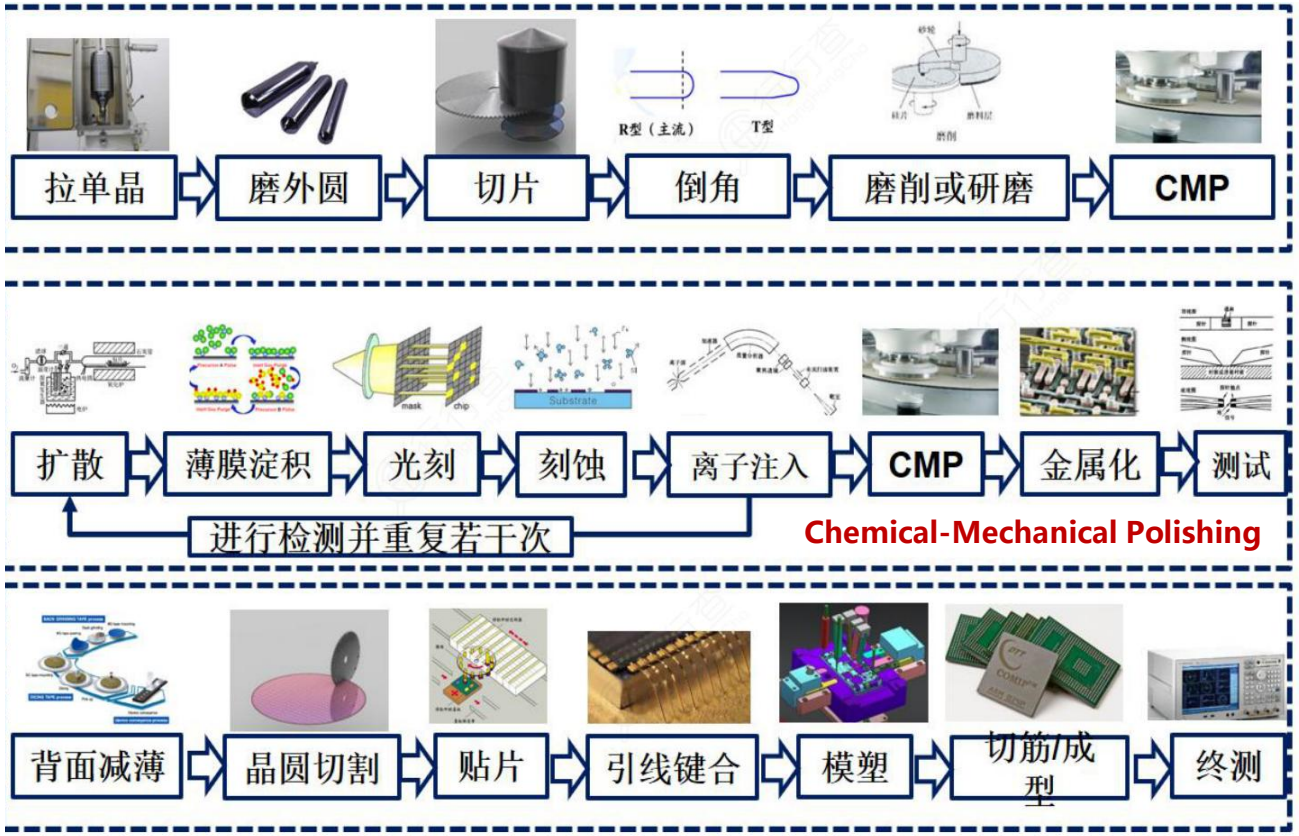
	IDM模式	Fabless模式	Foundry模式
典型厂商	SAMSUNG intel	QUALCOMM BROADCOM	tsmc SMIC
基本特点	集芯片设计、制造、封测于一身。早期多数集成电路企业采用的模式，目前仅有极少数企业能够维持	只负责芯片的电路设计与销售；将生产、测试、封装等环节外包	不负责芯片设计，只负责制造或封测；可以同时为多家设计公司提供服务，但受制于公司间的竞争关系
主要优势	设计、制造等环节协同优化，有助于充分发掘技术潜力；能有条件率先实验并推行新的半导体技术	资产较轻，初始投资规模小，创业难度相对较小；企业运行费用较低，转型相对灵活	不承担由于市场调研不准、产品设计缺陷等决策风险。
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

半导体芯片产业之晶圆加工

- 晶圆加工制造行业目前由台积电、三星、英特尔、格罗方德、中芯国际等公司占据主导地位



单晶硅片制造流程

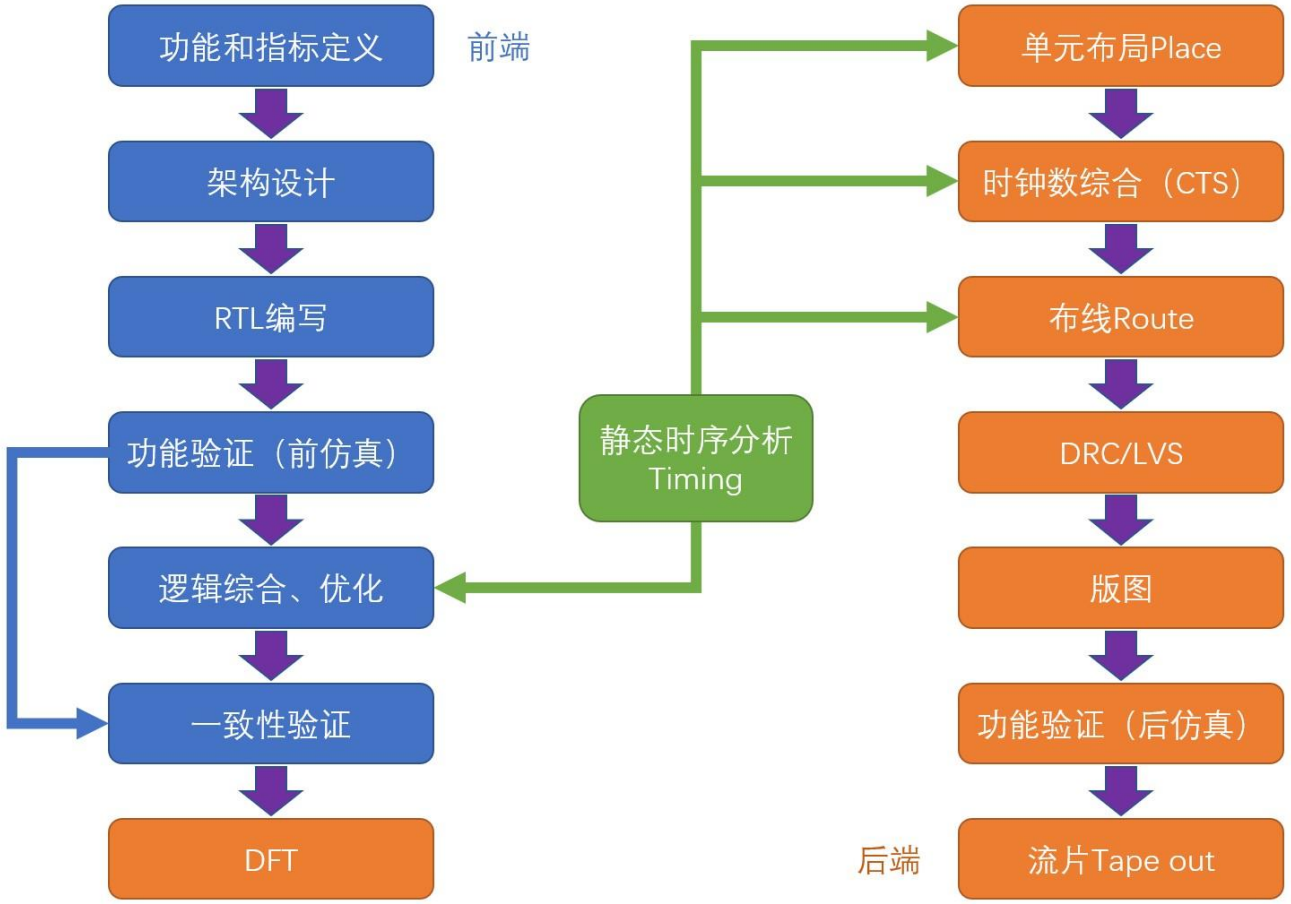
晶圆加工前道工艺

晶圆加工后道工艺

建设一条先进的晶圆线投资量级上百亿元，需要十几大类、几百种多种细分设备、数千多台各种高精尖设备，其中以光刻技术、刻蚀技术、薄膜沉积三大类为主要生产技术

半导体芯片产业之芯片设计 – 第一梯队均为美国公司

- 目前IC设计的重要产业公司 – 英特尔、英伟达、高通、华为



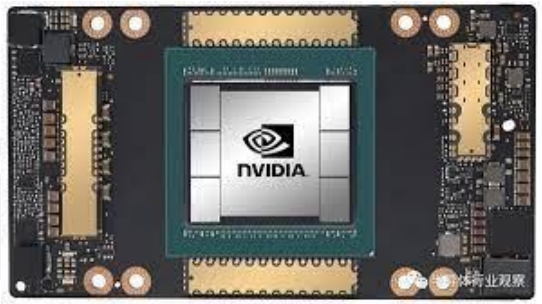
大型数字芯片设计基本流程

PC处理器CPU芯片

Qualcomm

手机SOC芯片

NVIDIA



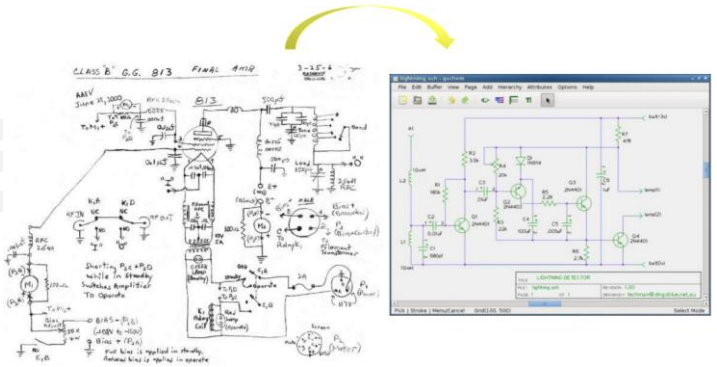
图形处理芯片GPU

半导体芯片产业之芯片设计 – 设计自动化软件产业

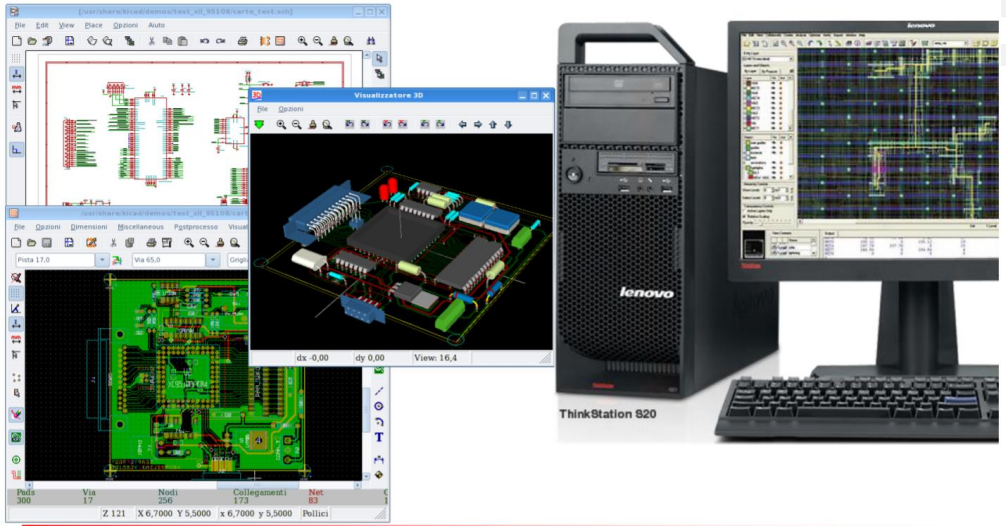
北京大学 PEKING UNIVERSITY

- 设计自动化软件 (EDA) 是提升芯片设计效率的关键因素，目前由**美国公司占主导地位**

电路设计软件与仿真工具



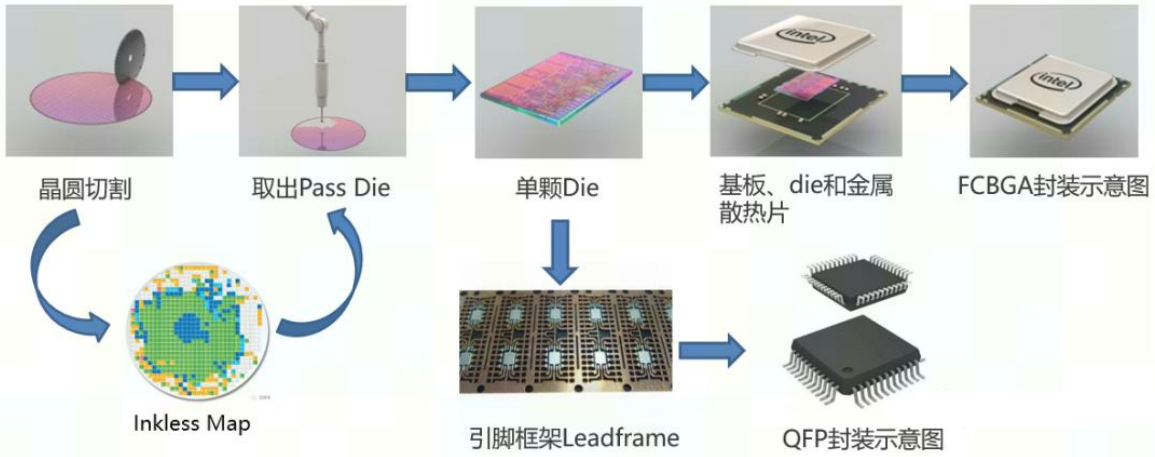
• Electronic Design Automation



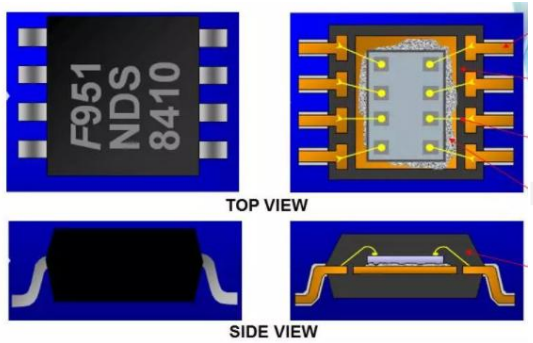
半导体芯片产业芯片封测产业

- IC封测直接关系到芯片的使用可靠性，目前由**日月光、安靠、长电**占据领先地位

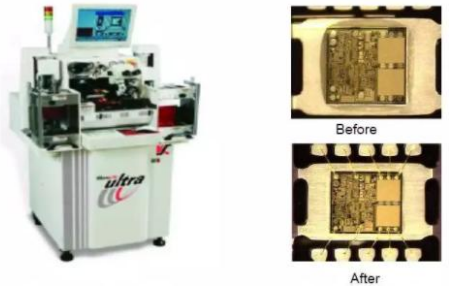
芯片封装



FT测试 Final Test

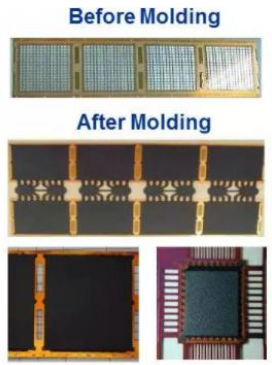


典型的芯片封装

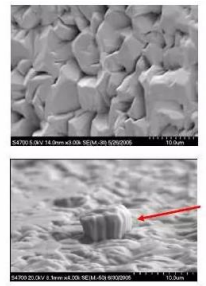


引线焊接

※为了防止外部环境的冲击，利用EMC把Wire Bonding完成后的产品封装起来的过程，并需要加热硬化。



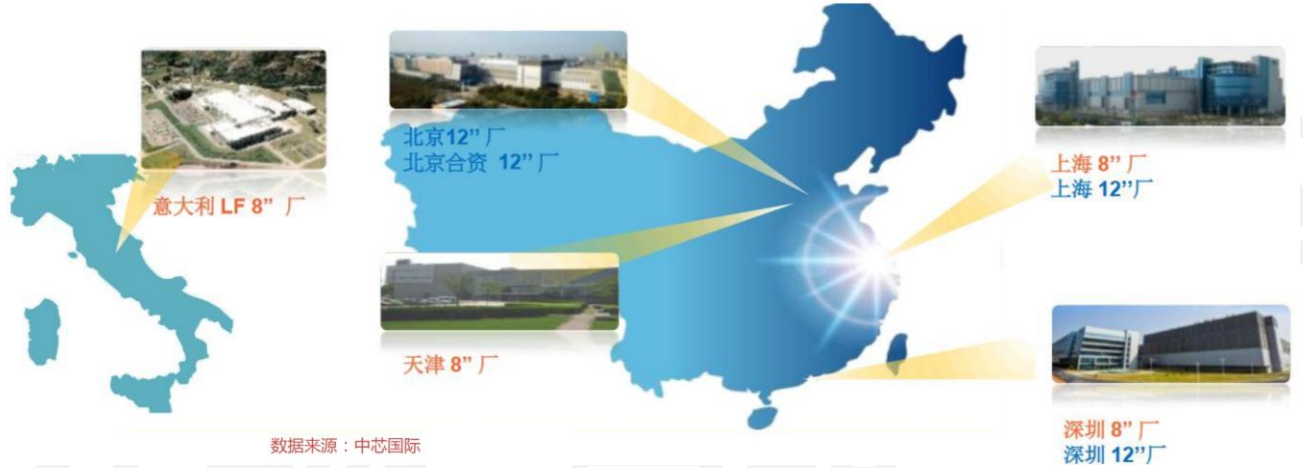
注塑



电镀退火

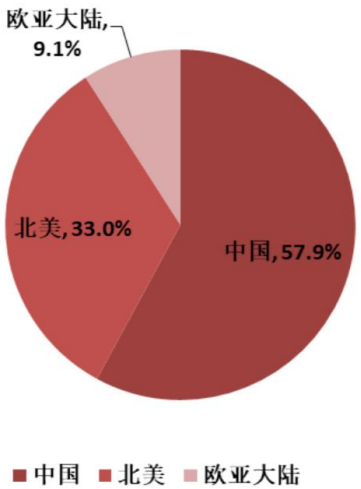
中国晶圆加工产业现状 – 中芯国际公司

- 中芯国际是大陆最大的晶圆加工和半导体芯片制造商，目前成熟制程为40nm/28nm

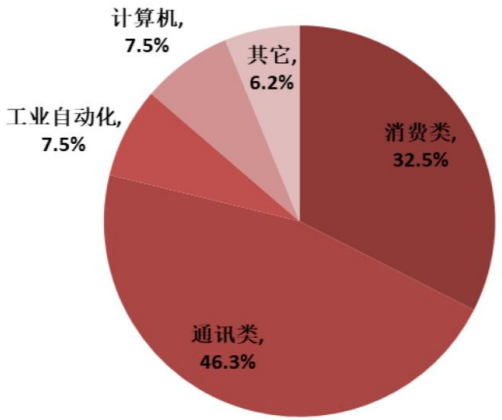


数据来源：中芯国际

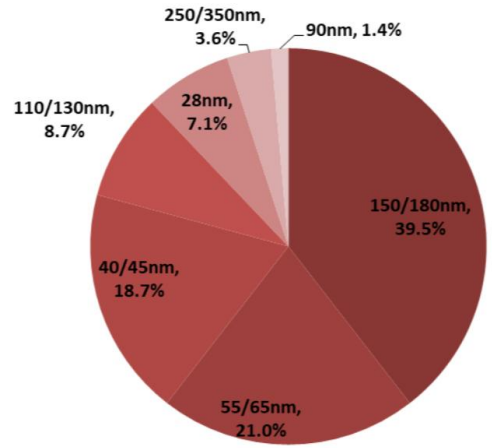
营收按地区划分



营收按下游应用划分



营收按工艺制程划分



中国晶圆加工产业现状 – 华虹半导体公司/长江存储

- 华虹半导体、长江存储是我国半导体芯片制造的重要企业，尤其是存储器领域

HGrace 華虹宏力

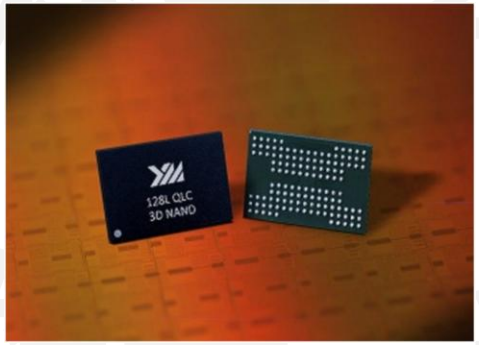
华虹半导体

华虹半导体晶圆厂房

上海一厂	上海二厂	上海三厂
工艺与产能	工艺与产能	工艺与产能
95nm 63千片晶圆/月 (200mm)	0.35um 57千片晶圆/月 (200mm)	90um 48千片晶圆/月 (200mm)

总产能168千片/月

长江存储科技有限责任公司
Yangtze Memory Technology Corp

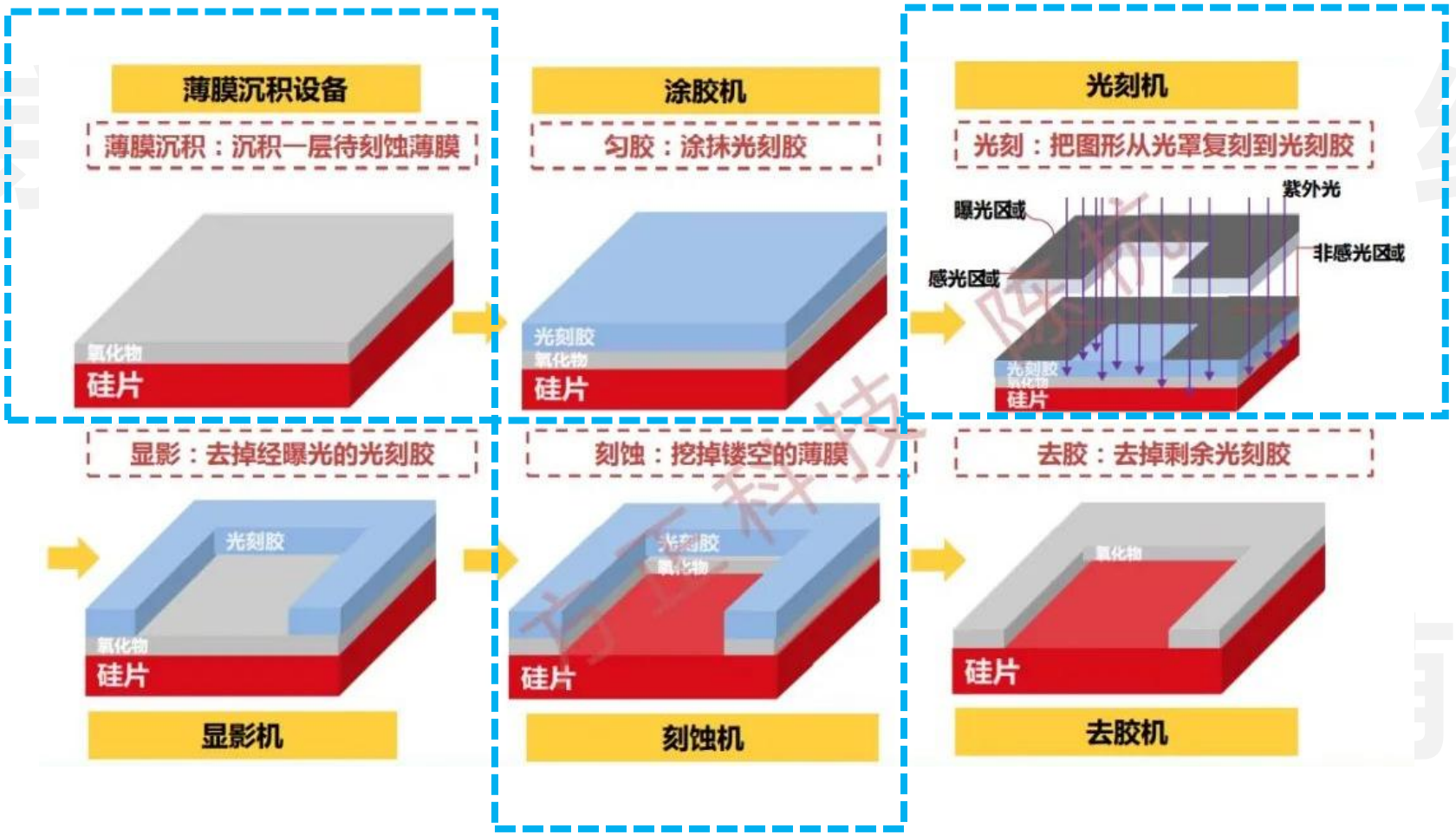


长存128层Nand Flash芯片

长存SSD固态硬盘芯片
(2TB PCIe4.0 ~535元)

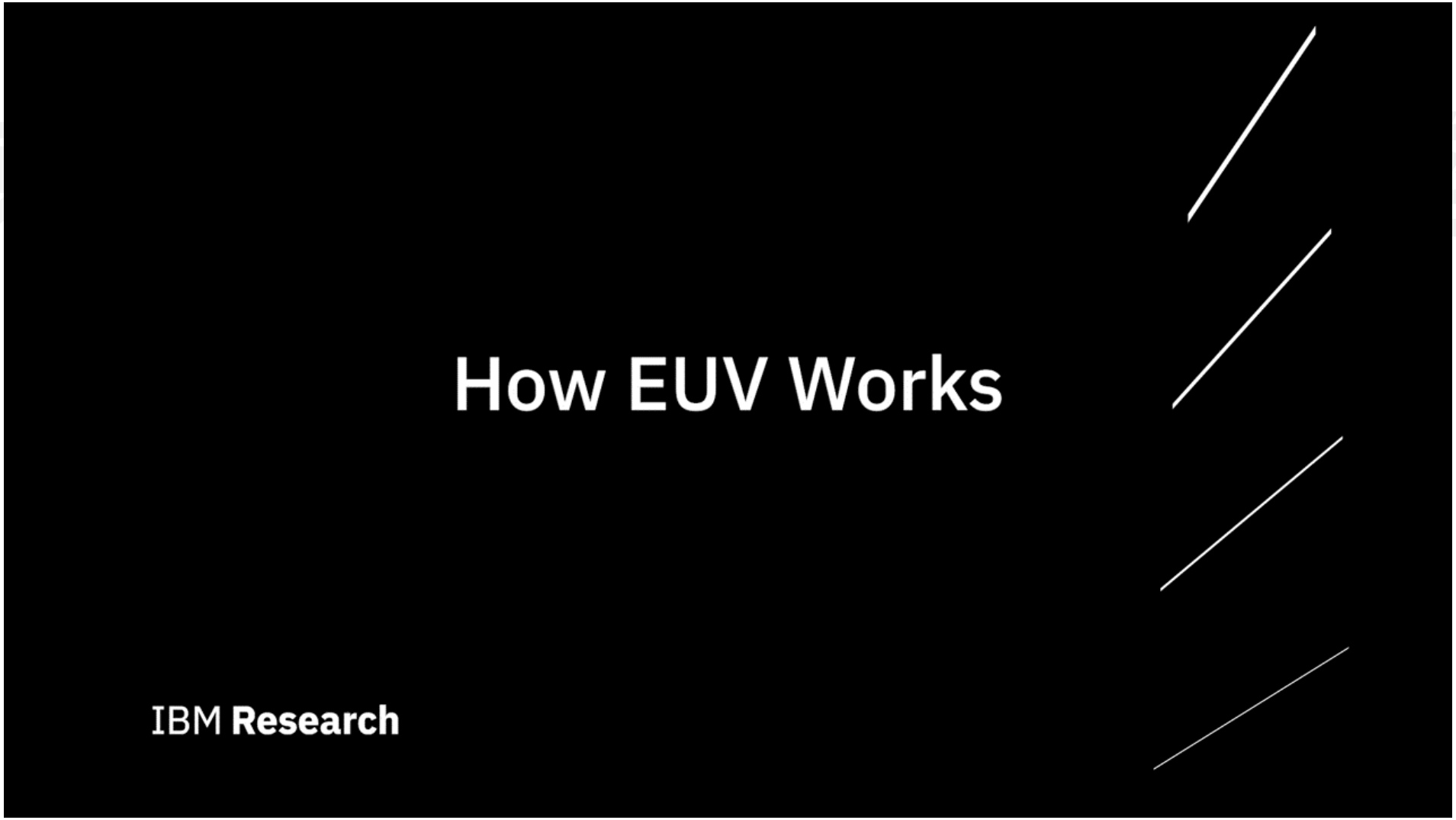
中国的“卡脖子”领域之一：晶圆加工产业

- 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



中国芯片设计产业现状 – 华为海思、紫光集团等

- 华为海思半导体、紫光集团是中国大陆最大的芯片设计公司

HUAWEI

HISILICON

国网信通产业集团
STATE GRID INFO & TELECOM GROUP
北京智芯微电子科技有限公司
BEIJING SMARTCHIP MICROELECTRONICS TECHNOLOGY CO., LTD.

电网相关芯片

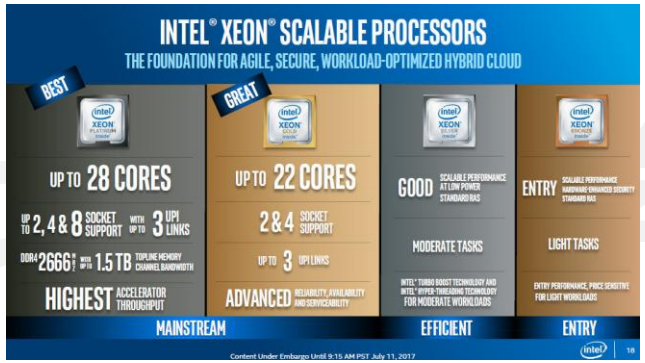
紫光集团
TSINGHUA UNIGROUP

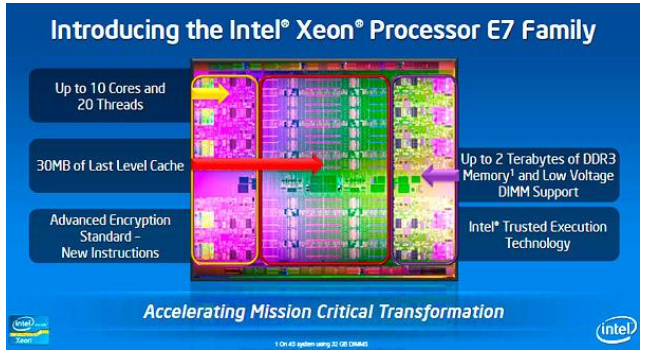
SANECHIPS™
中兴微电子

通信相关芯片

中国的“卡脖子”领域之二：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”领域之三：EDA软件产业

- 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

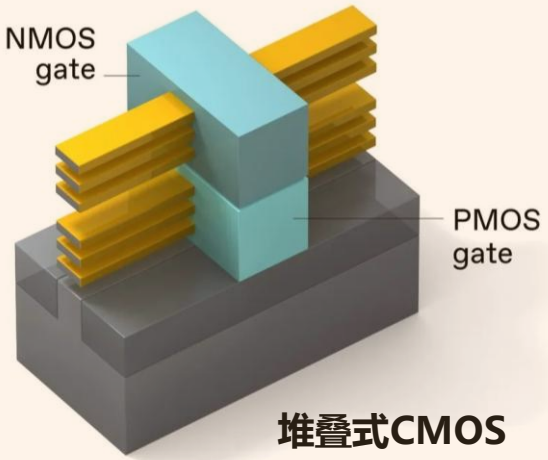
融合新器件、新架构、新计算是后摩尔时代体系结构的发展趋势

北京大学 PEKING UNIVERSITY

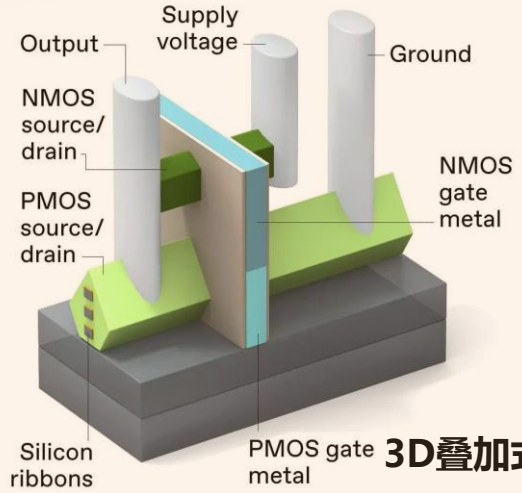
- 融合新器件、新架构、新计算是突破后摩尔时代大算力、高能效瓶颈的重大关键技术领域

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

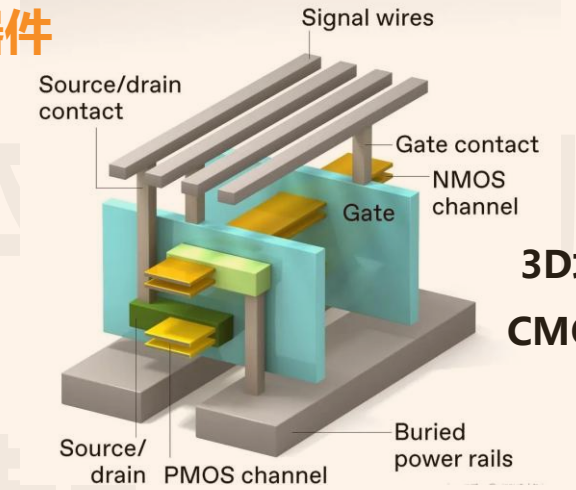


堆叠式CMOS

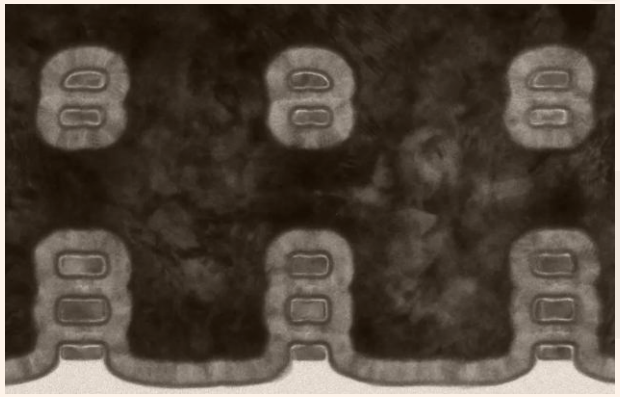


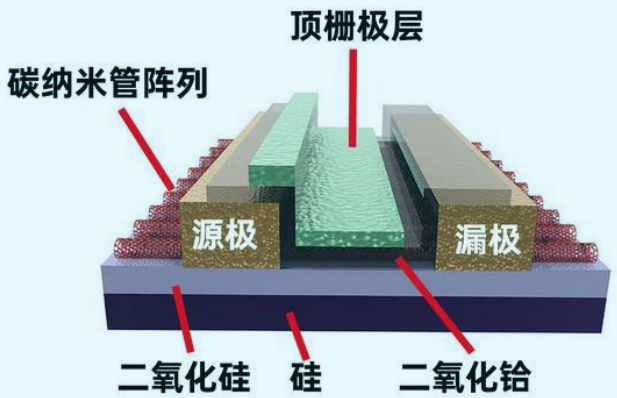
新兴硅基器件

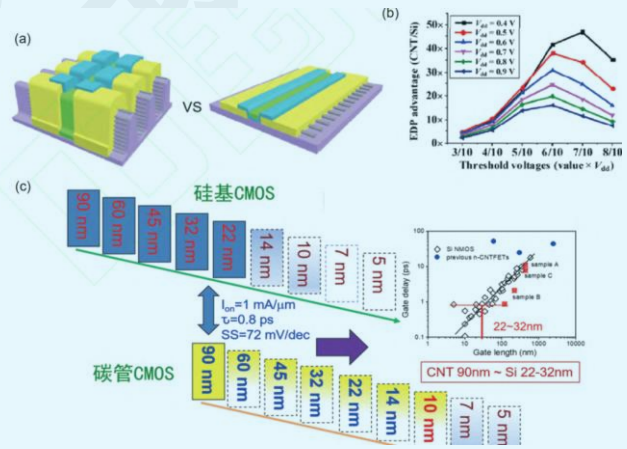
3D叠加式非门



3D堆叠的
CMOS连接



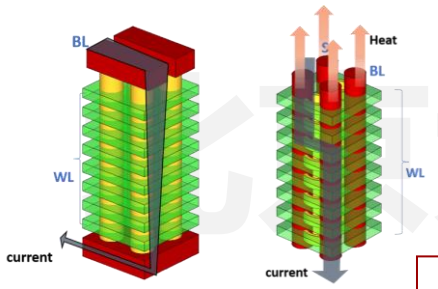




新兴碳基器件

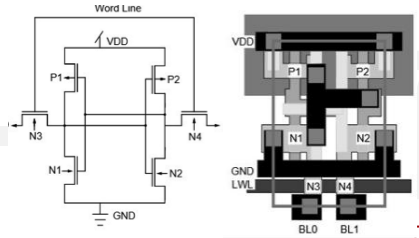
代表性智能芯片新兴技术– 新器件：存储-计算融合器件

- 未来存储器介质材料的创新



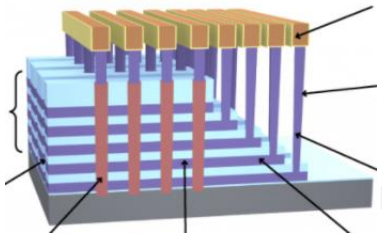
DRAM

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



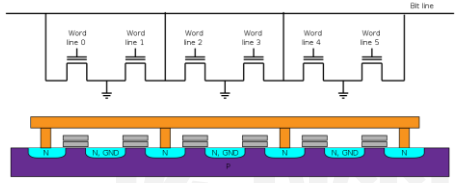
SRAM

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

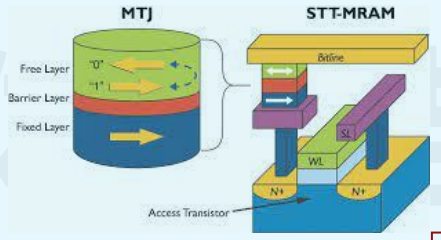
优点：工艺成熟、容量大、成本低
缺点：速度低、只能近存
非易失性：是
适合场景：云端大容量



Nor Flash

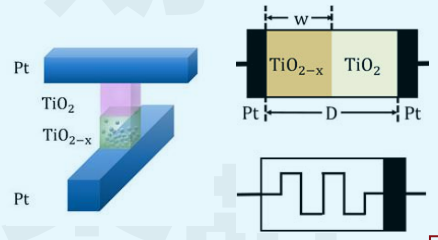
优点：工艺成熟、密度高、成本低
缺点：对PVT变化敏感、能效低
非易失性：是
适合场景：端侧、边缘低成本

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

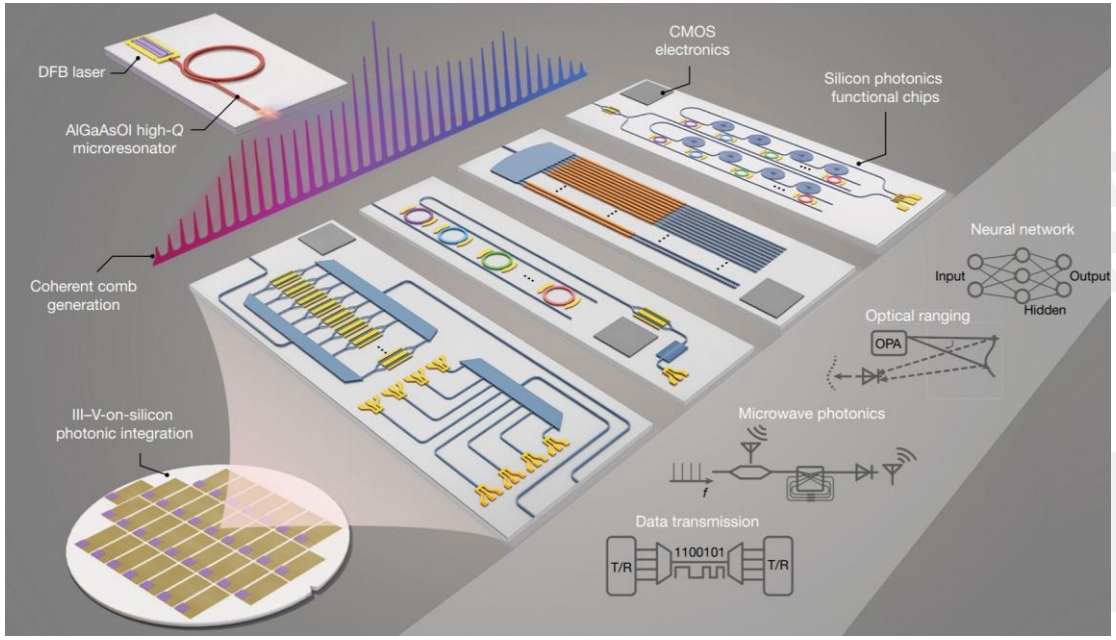


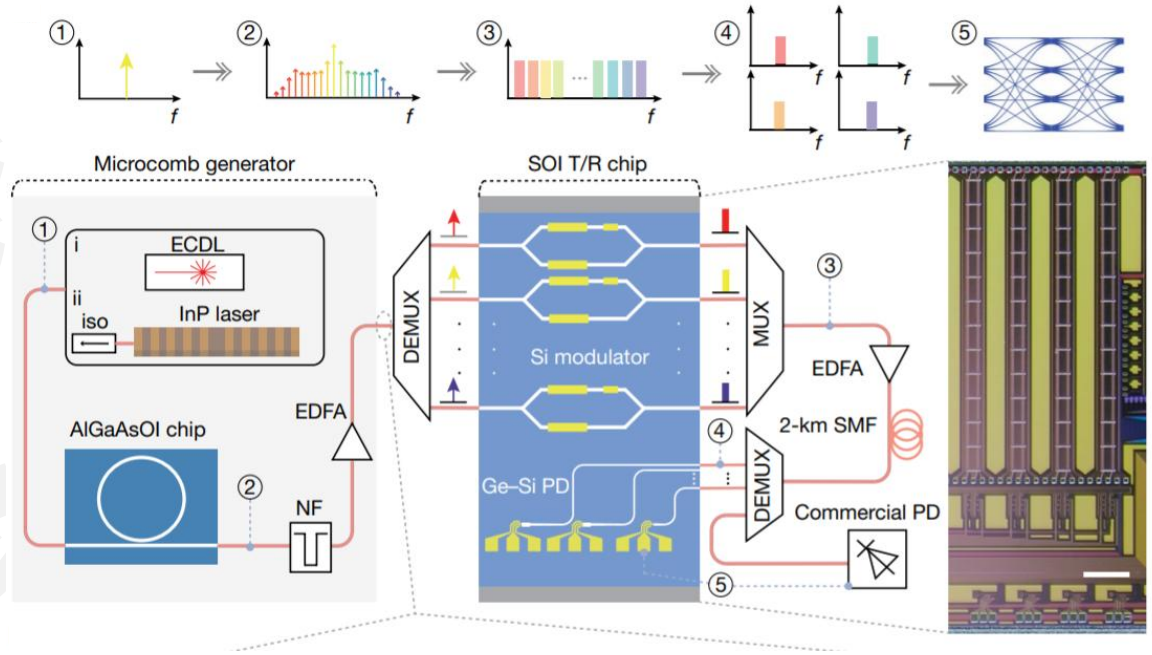
忆阻器 (RRAM/PCM)

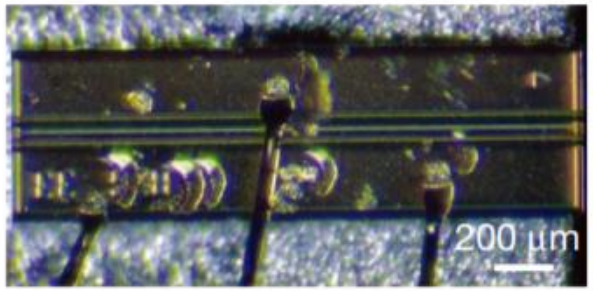
优点：算力、能效、密度高
缺点：工艺爬坡成熟中
非易失性：是
适合场景：云边端大算力

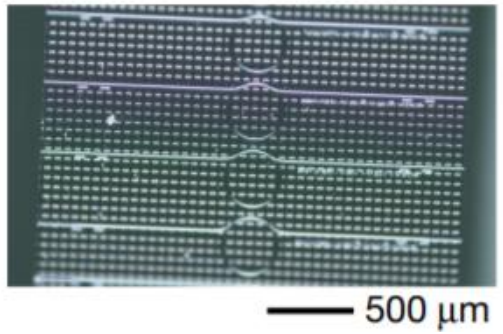
代表性新兴技术 – 新器件：光器件与光计算、光通信技术

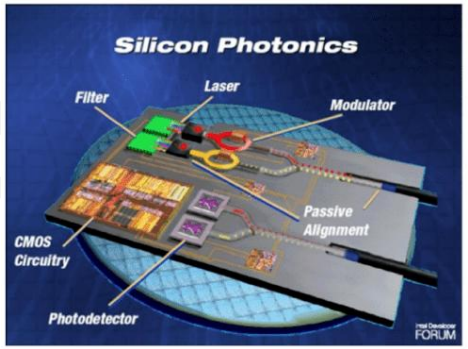
- 片上光计算、光通信有望突破信号传递延时的瓶颈，打破电芯片物理上限

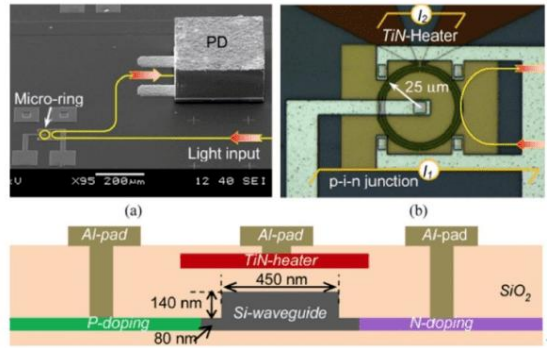






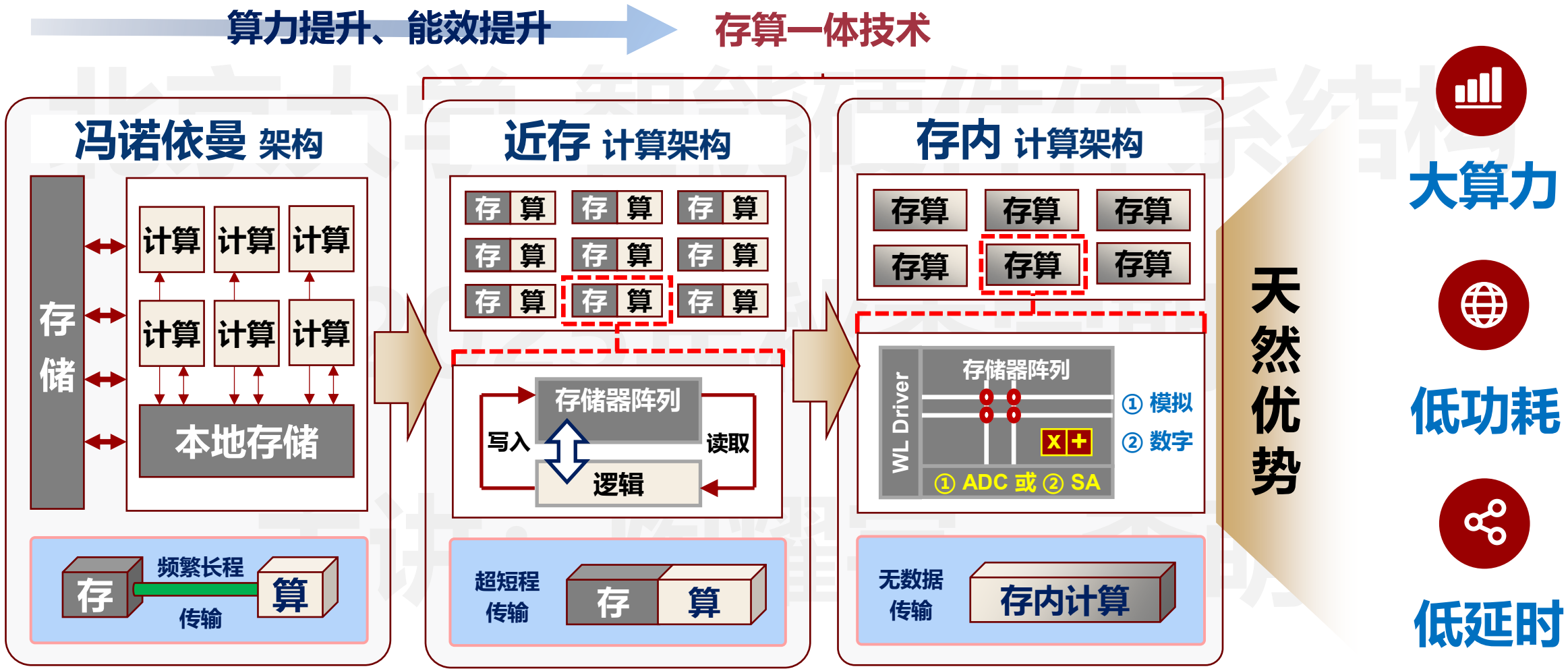






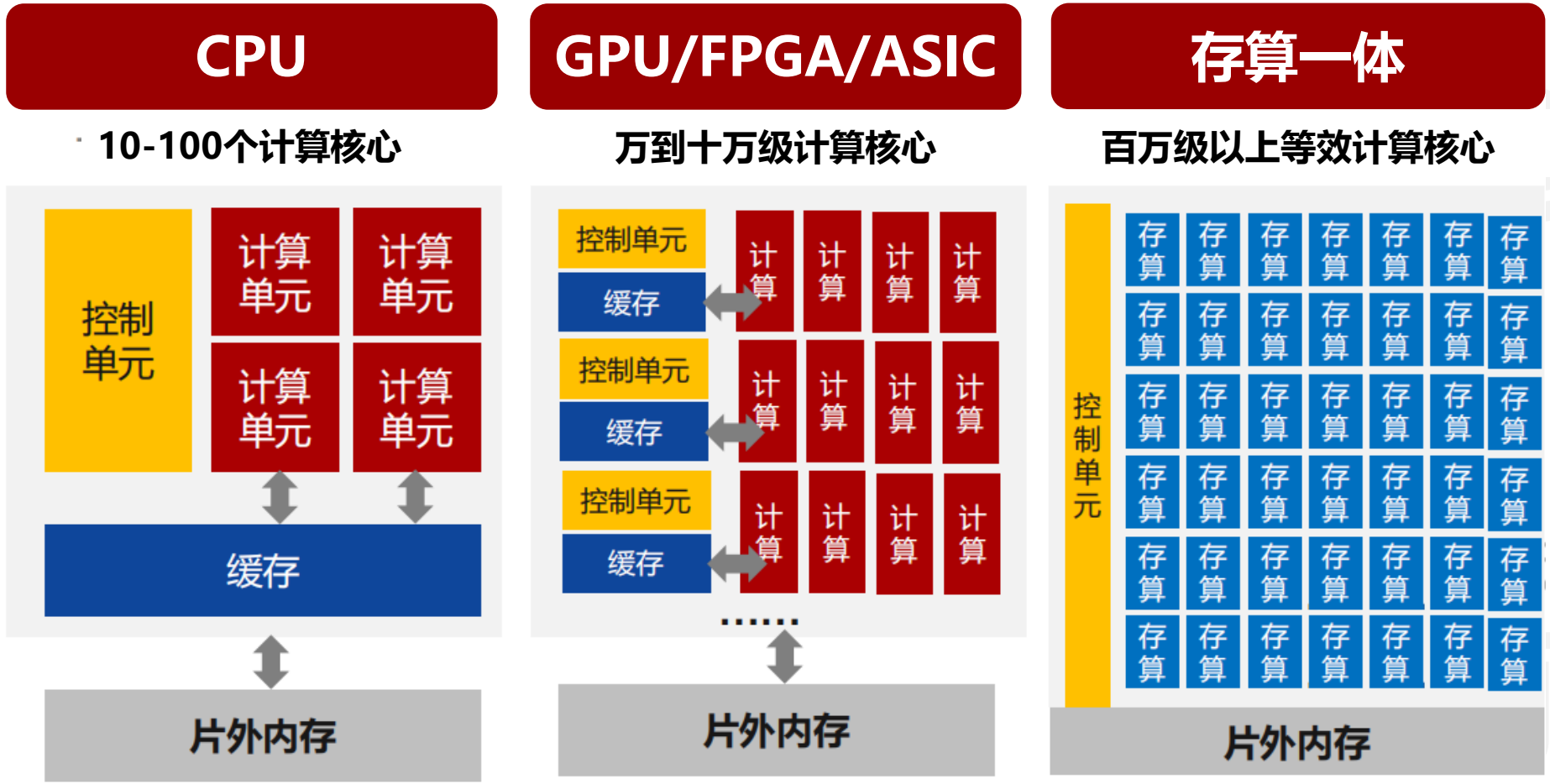
代表性智能芯片新兴技术 – 新架构：存算一体

- 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

- 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

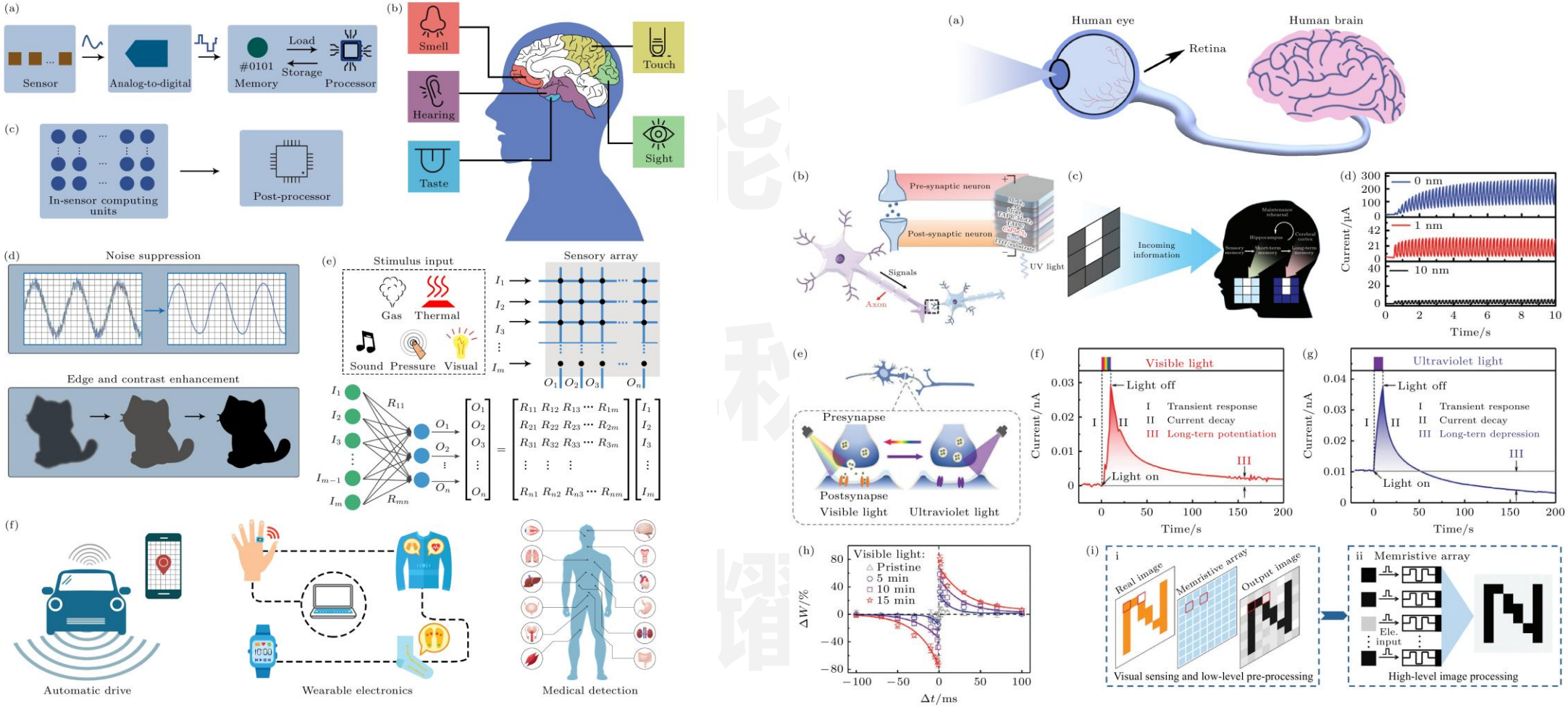


存算一体如何对AI大模型进行有效支持?

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

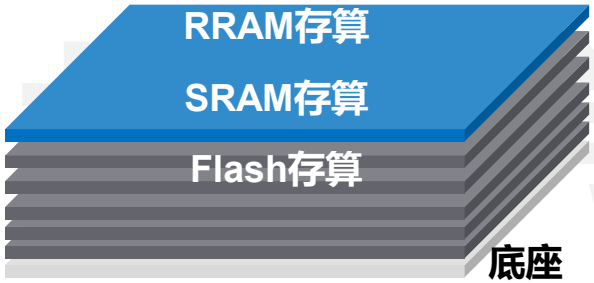


视觉感存算一体芯片与硬件系统

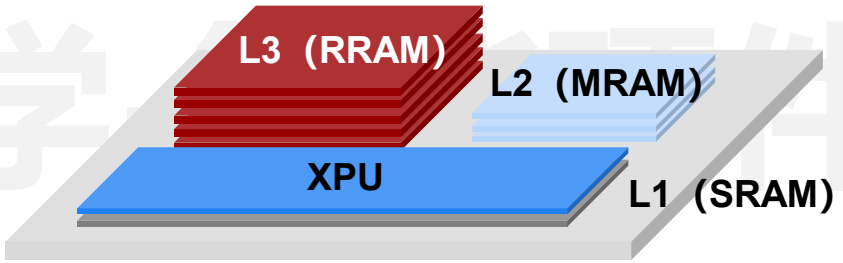
代表性智能芯片新兴技术 – 新架构：三维异质集成

- 协同先进封装技术，实现多种芯片方案相结合

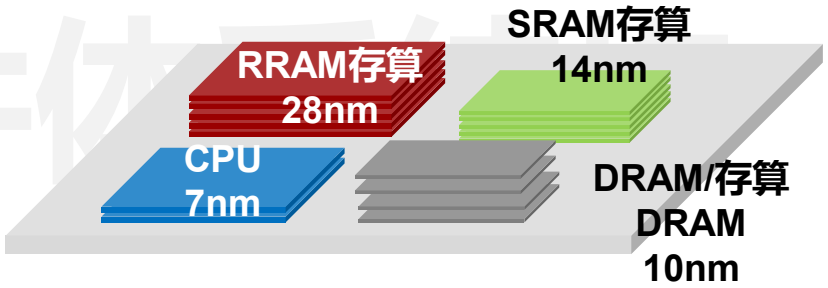
先进三维集成芯片示例图



三维集成

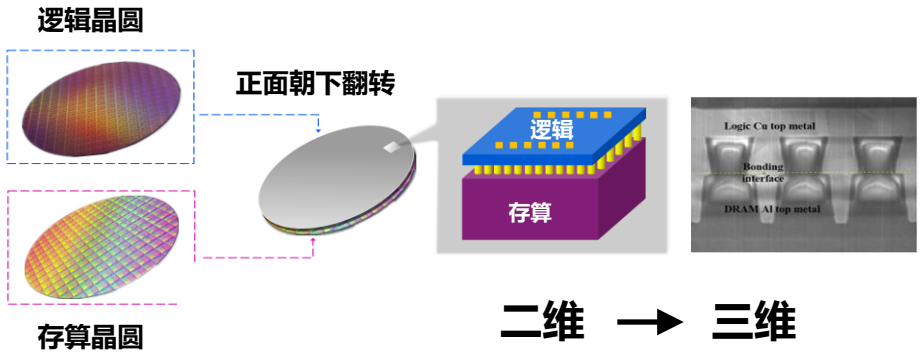


多级存储器堆叠SoC



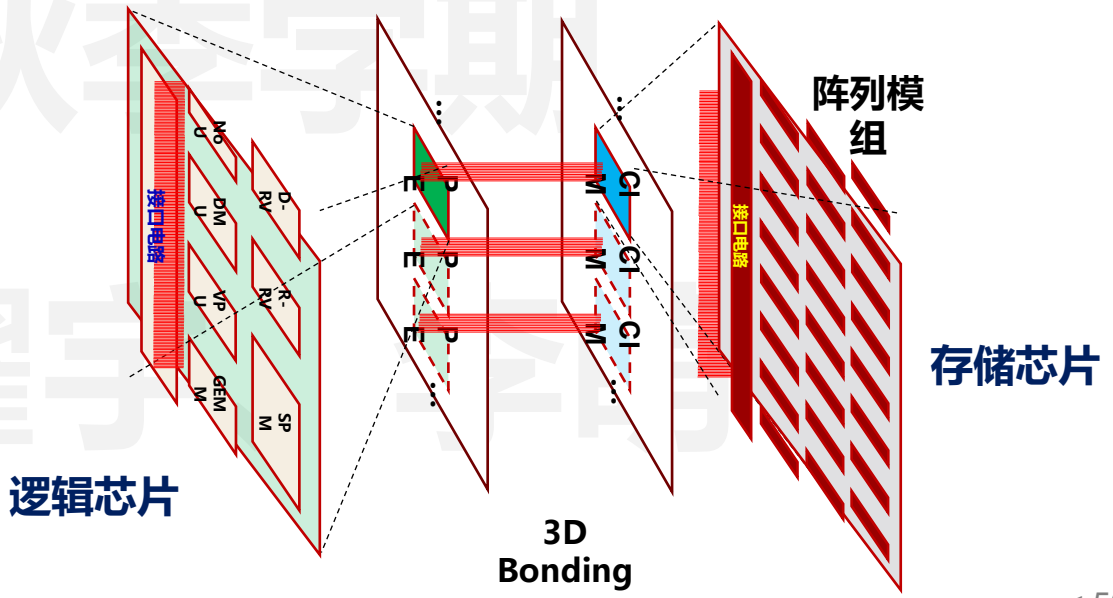
异构小芯粒封装

混合键合异质三维集成



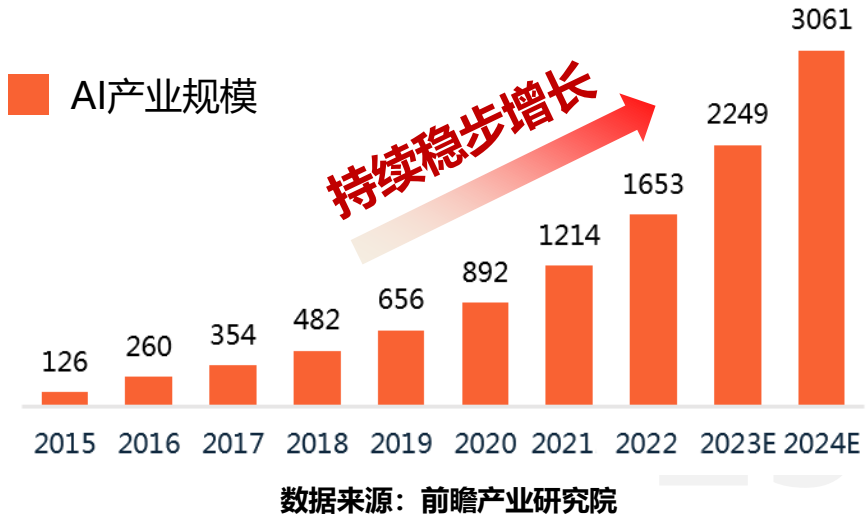
二维 → 三维

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

- 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



数据来源：前瞻产业研究院

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

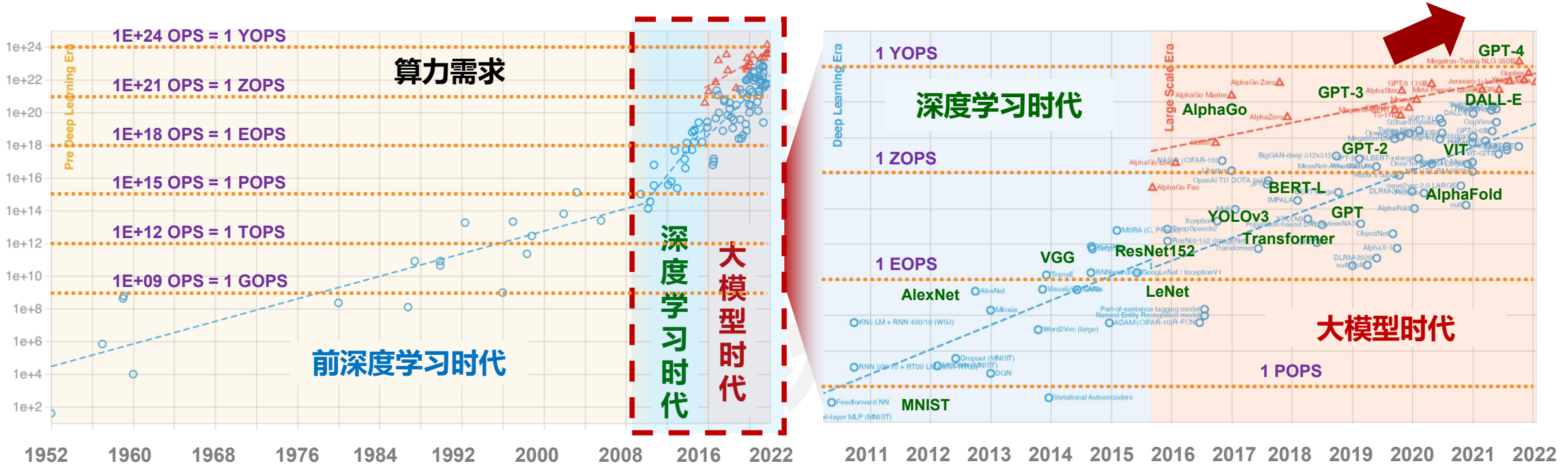
发布**盘古大模型系列**，面向视觉、自然语言、科学计算等多个场景；提出五大基础大模型，细分为多个行业大模型，深度渗透政务、金融、制造、气象、医学等多个行业，提升智能化水平

文心一言 **通义千问** **腾讯混元** **DeepSeek**

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

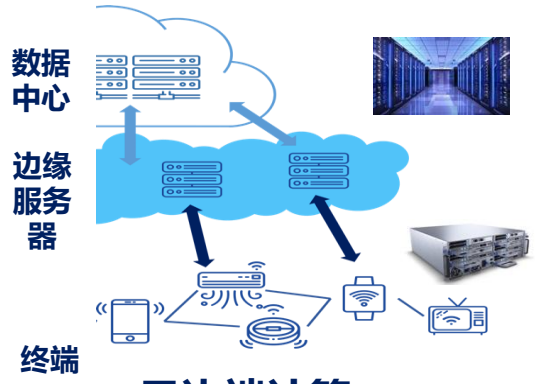
芯片性能成为支撑智能系统从量变产生质变的基石

代表性智能芯片新兴技术 – 新计算：AI大模型

- 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

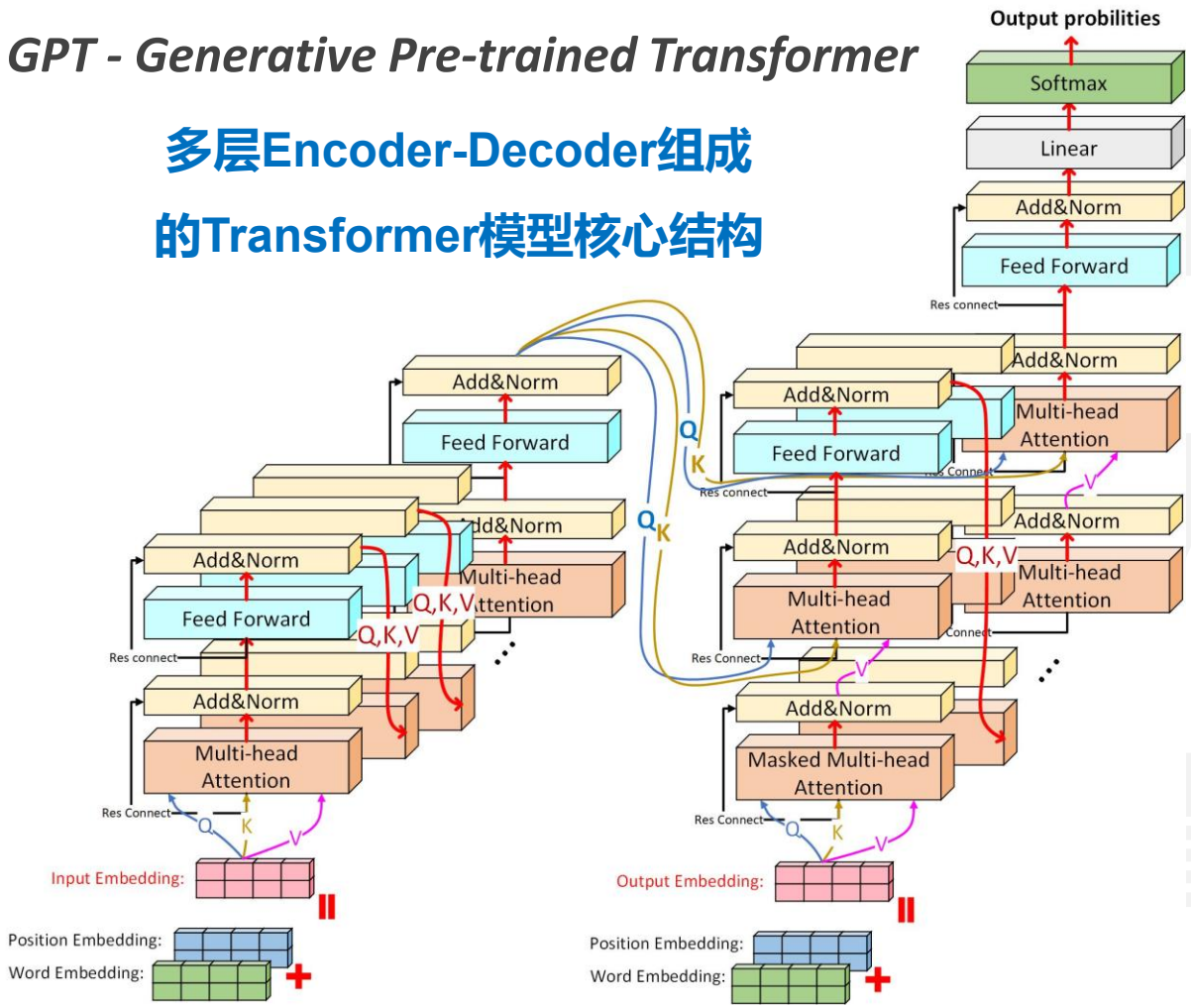
当前AI大模型以Transformer为基干网络 (以GPT为例)

- 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

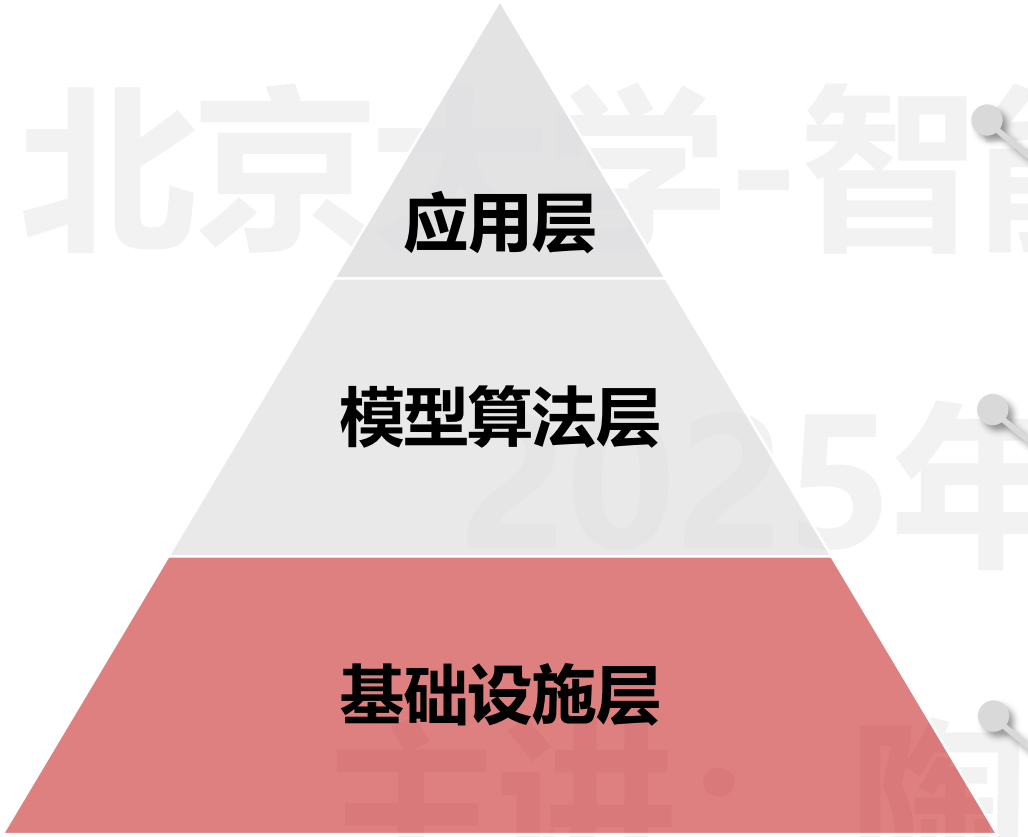
Softmax

微软/OpenAI提出了**LongNet**，将Transformer的

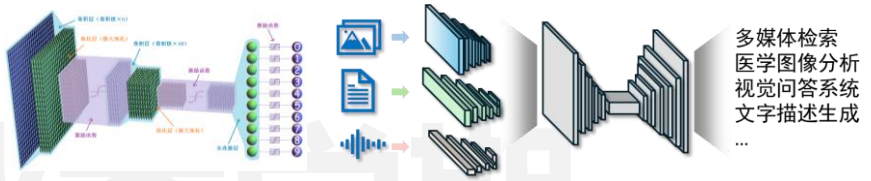
Token数提高到了10亿级别，并持续提升

代表性智能芯片新兴技术 – 新计算：AI大模型

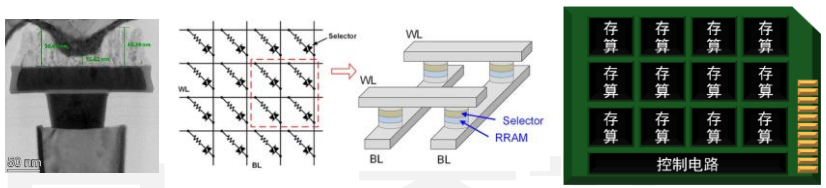
- 高性能芯片与硬件系统是不可或缺的算力基石



具身智能　自动驾驶　智慧医疗



机器学习模型　多模态大模型



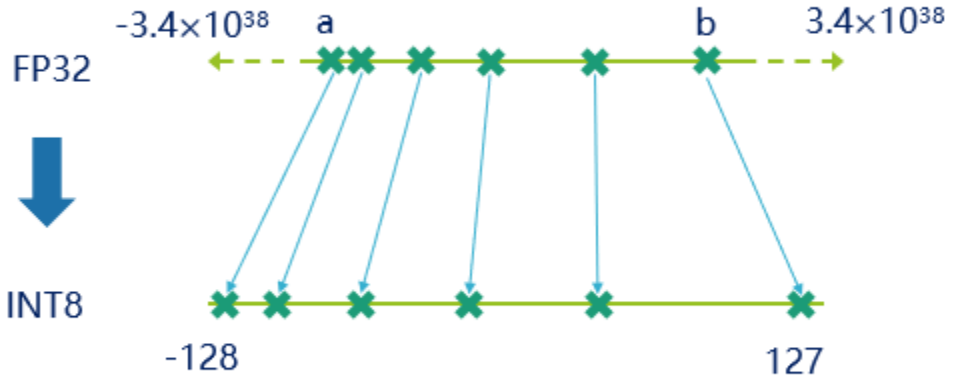
大模型AI芯片与硬件系统(器件、电路、架构及编译器)

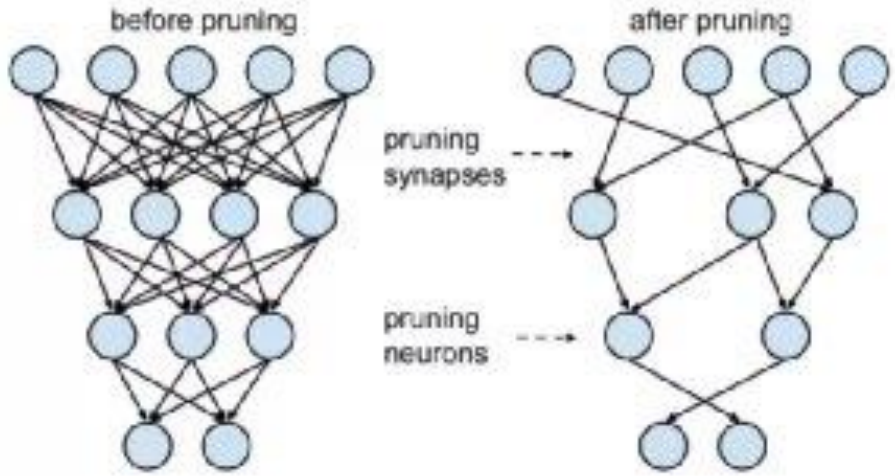
重要支撑

人工智能竞争的一大核心：底层芯片与硬件的“军备竞赛”

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



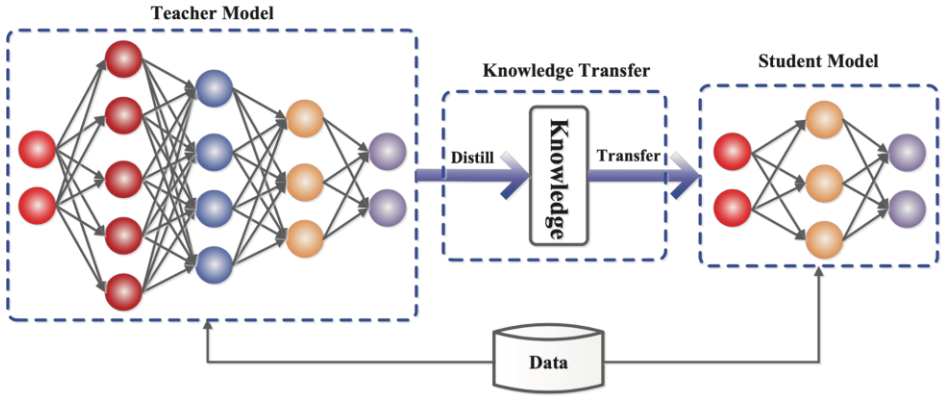


模型量化：将高精度的权重量化为低精度的权重，以一定的精度损失为代价换取更小的存储和计算开销

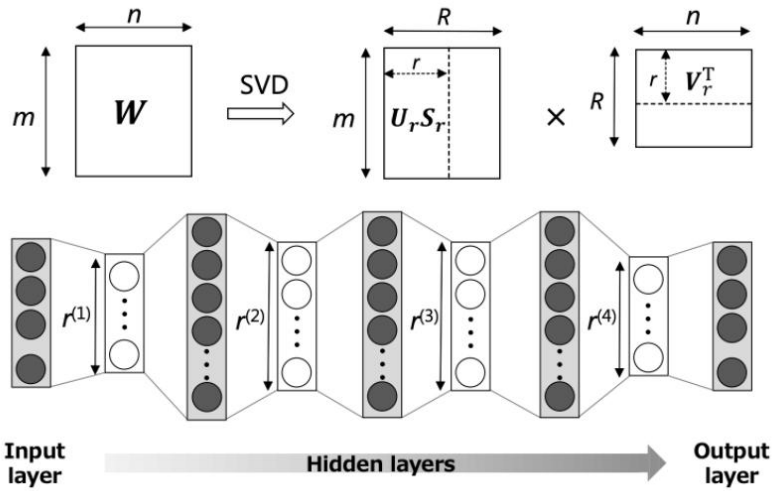
模型剪枝：将神经网络中重要性较小的神经元和权重删除，减少计算量，加速神经网络推理

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



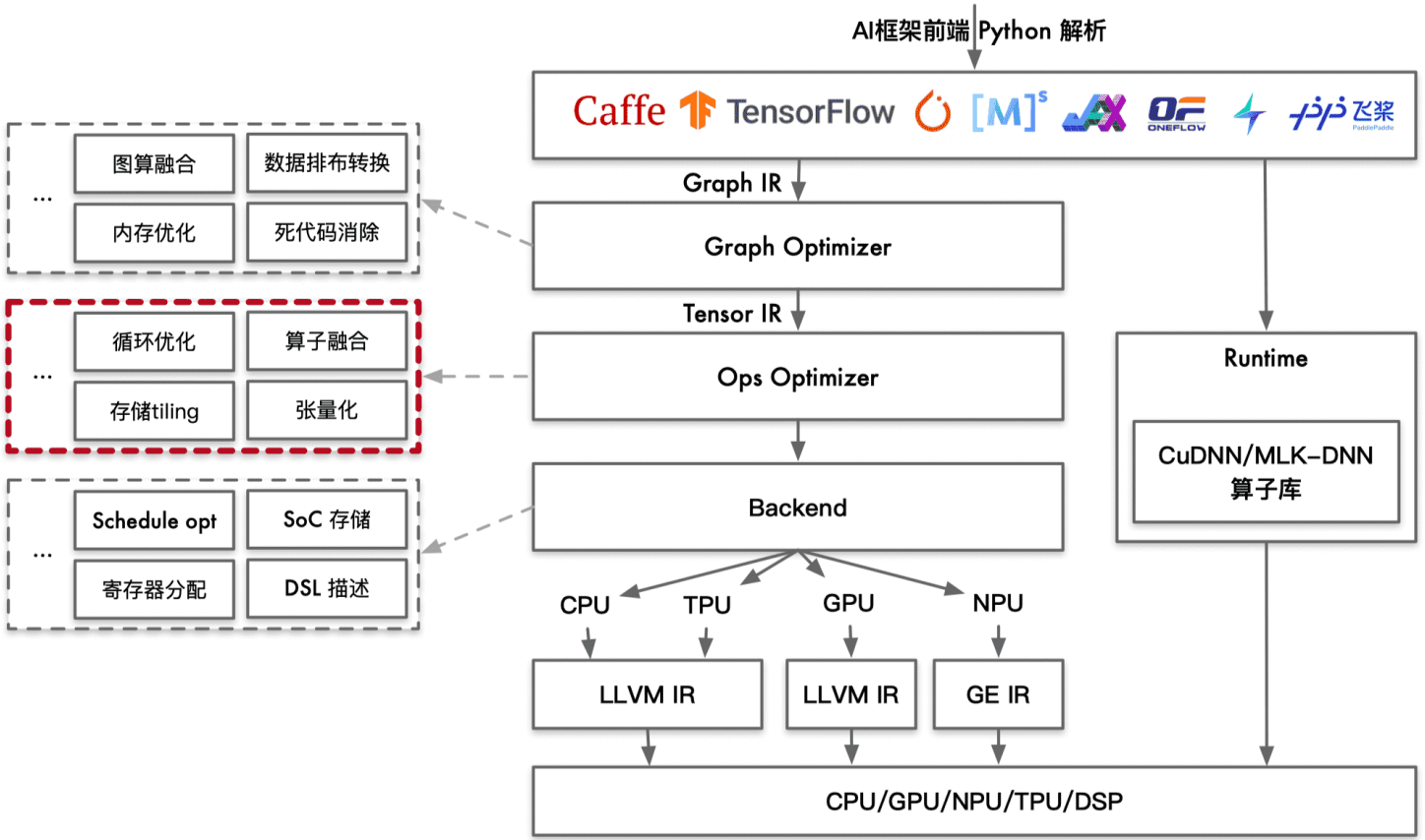
知识蒸馏：将规模较大的模型作为teacher model训练一个较小的student model，在尽可能保证性能的情况下减小模型规模



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

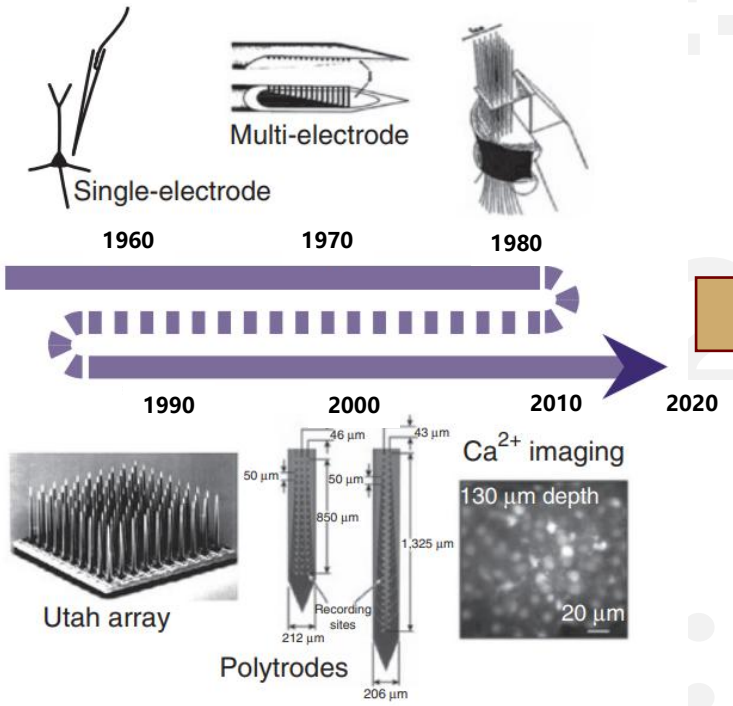
和寄存器分配等等

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

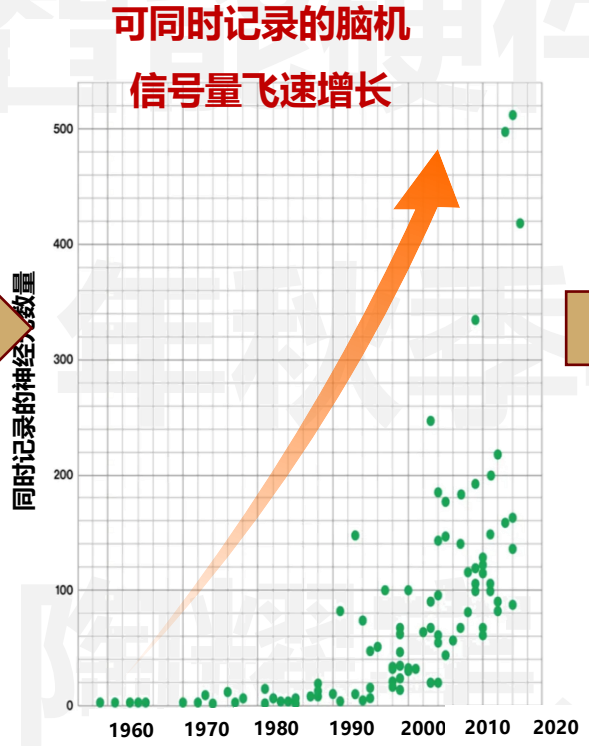
- 为脑机接口服务的芯片与系统将在未来数十年成为人类发展的方向之一

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”

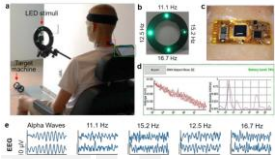


Stevenson, I., et al. Nat Neuroscience

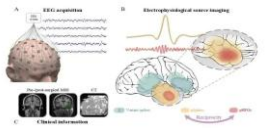
脑机打字

脑控无人机

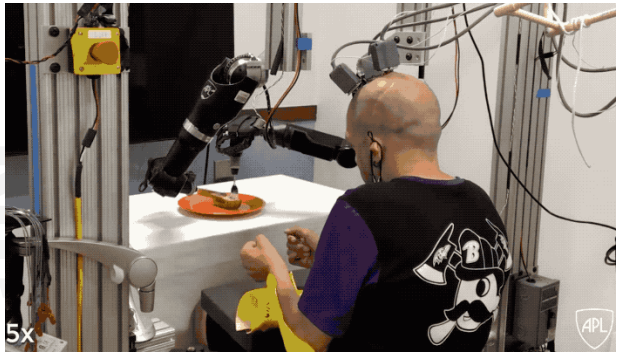
脑控座椅



脑机癫痫监测



脑控机械臂

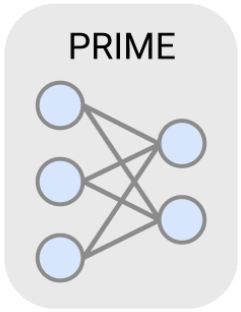


代表性新兴技术 – 新方法：AI设计AI芯片

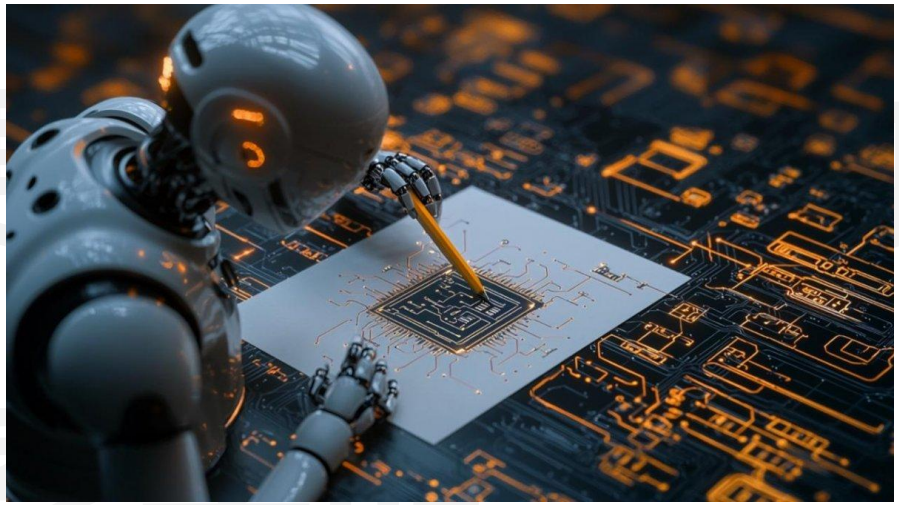
- 设计AI芯片架构 -> 利用AI设计AI芯片架构

参数化硬件单元库

针对某类任务的
最优芯片设计

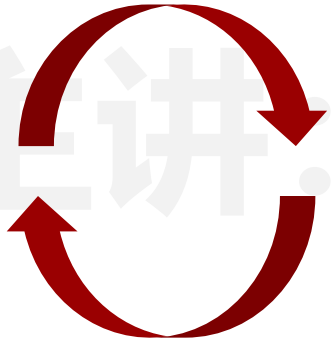


Optimizer



RL、大模型等方式

设计AI芯片的
AI模型



AI芯片