

华金证券电子团队一走进“芯”时代系列深度之七十四“算力芯”

# 以“芯”助先进算法，以“算”驱万物智能

分析师：孙远峰 S0910522120001

分析师：王臣复 S0910523020006

2024年2月18日

本报告仅供华金证券客户中的专业投资者参考  
请仔细阅读在本报告尾部的重要法律声明

- ◆ **半导体景气度有望迎来复苏：**从半导体行业景气度来看，在经历了2022、2023年的去库存周期后，半导体销售额有望在2024年迎来复苏。据WSTS数据显示，全球半导体产品销售总额从1999年的1494亿美元增长到2022年的5741亿美元，期间复合增速约6.03%。WSTS预计2023年全球年销售额同比下滑9.4%，但2024年将有望增长13.1%。
- ◆ **逻辑芯片国产替代空间巨大：**从半导体产品销售额来看，逻辑器件是半导体产品市场规模最大的细分品类，根据SIA的报告显示，2022年全球逻辑器件市场规模达1766亿美元，包括MPU、GPU、SOC等。根据集邦咨询统计的2023年第三季度前十大IC设计业者营收中（仅统计公开财报的），前五大厂商中，仅博通的主业不是逻辑芯片为主。国产厂商无论从料号的布局还是企业整体营收规模来看与国外厂商都有巨大的差距，目前具有较大的替代空间。
- ◆ **AI技术快速发展催生算力需求快速提升：**大模型技术所带来的突破以及商业化前景促使全球进入AI竞赛，其中，美国在AI领域依旧担当着创新者和引领者的身份。大模型的发展依托于强大的算法、庞大的数据量以及巨大的算力支撑，当下大模型逐步从单模态走向多模态，对于以英伟达GPU为代表的算力需求快速爆发，全球范围内算力芯片的供需缺口一定程度上降低了多模态大模型的发展速度，但依旧无法阻止AI技术向前演进以及商业化逐步落地，因此而对国内外算力产业链带来了巨大的发展机遇。国产算力同样面临着巨大的需求，但却要面对高端GPU芯片无法进口的局面，进而为国产厂商崛起带来了发展机遇。
- ◆ **走向推理，创新硬件不断涌现，边、端算力机遇有望启动：**以AI手机、AI PC、AI PIN等为代表的终端创新不断涌现，代表着AI从训练走向推理阶段，我们认为是AI技术走向商业化的必然之路，新技术的出现有望带来硬件应用场景的重新定义，并进而推动边、端侧算力的需求向前演进。
- ◆ **风险提示：技术创新风险、宏观经济和行业波动风险、国际贸易摩擦风险、复苏或需求不及预期相关风险**

- 01 逻辑芯片市场规模巨大，产业链具备高门槛
- 02 大模型走向多模态，技术持续演进，商业前景广阔
- 03 算力板块投资的几条主线
- 04 相关标的
- 05 风险提示

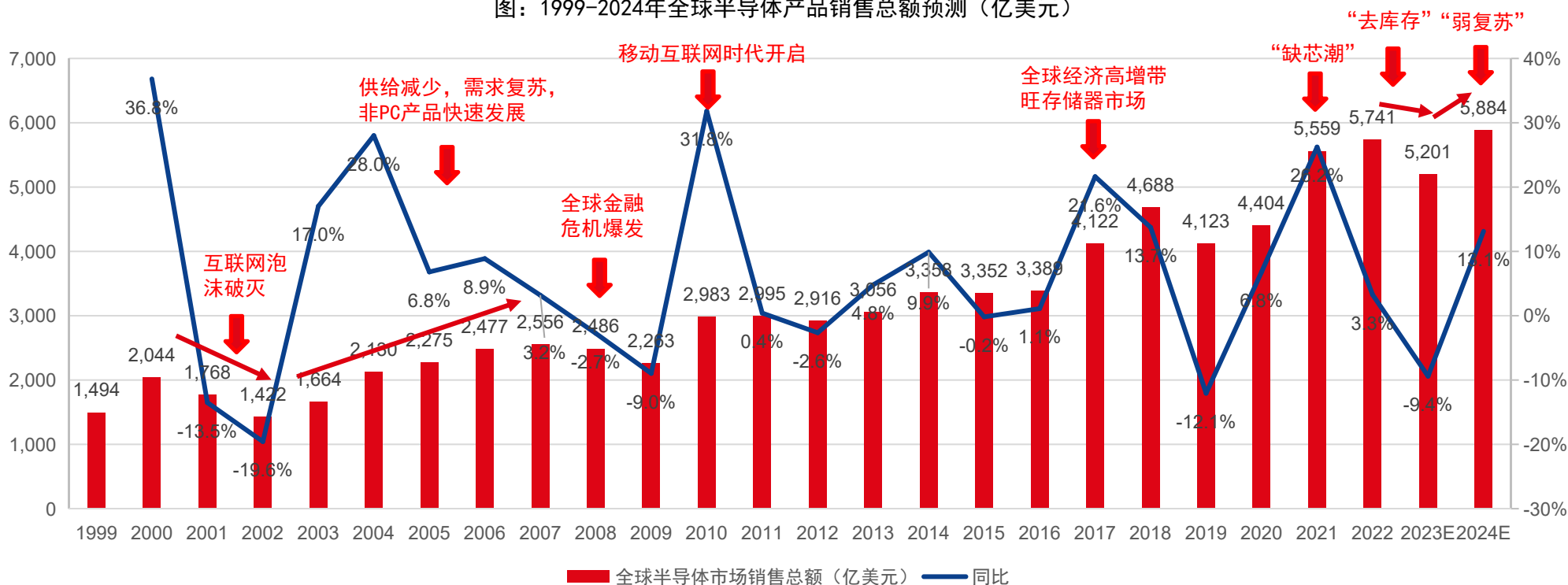
# 2024年全球半导体销售有望迎来复苏

- ◆ 据WSTS数据显示，全球半导体产品销售总额从1999年的1494亿美元增长到2022年的5741亿美元，期间复合增速约6.03%。
- ◆ WSTS预计2023年全球年销售额同比下滑9.4%，但2024年将有望增长13.1%。

图：2024年全球半导体产品销售总额预测（亿美元）

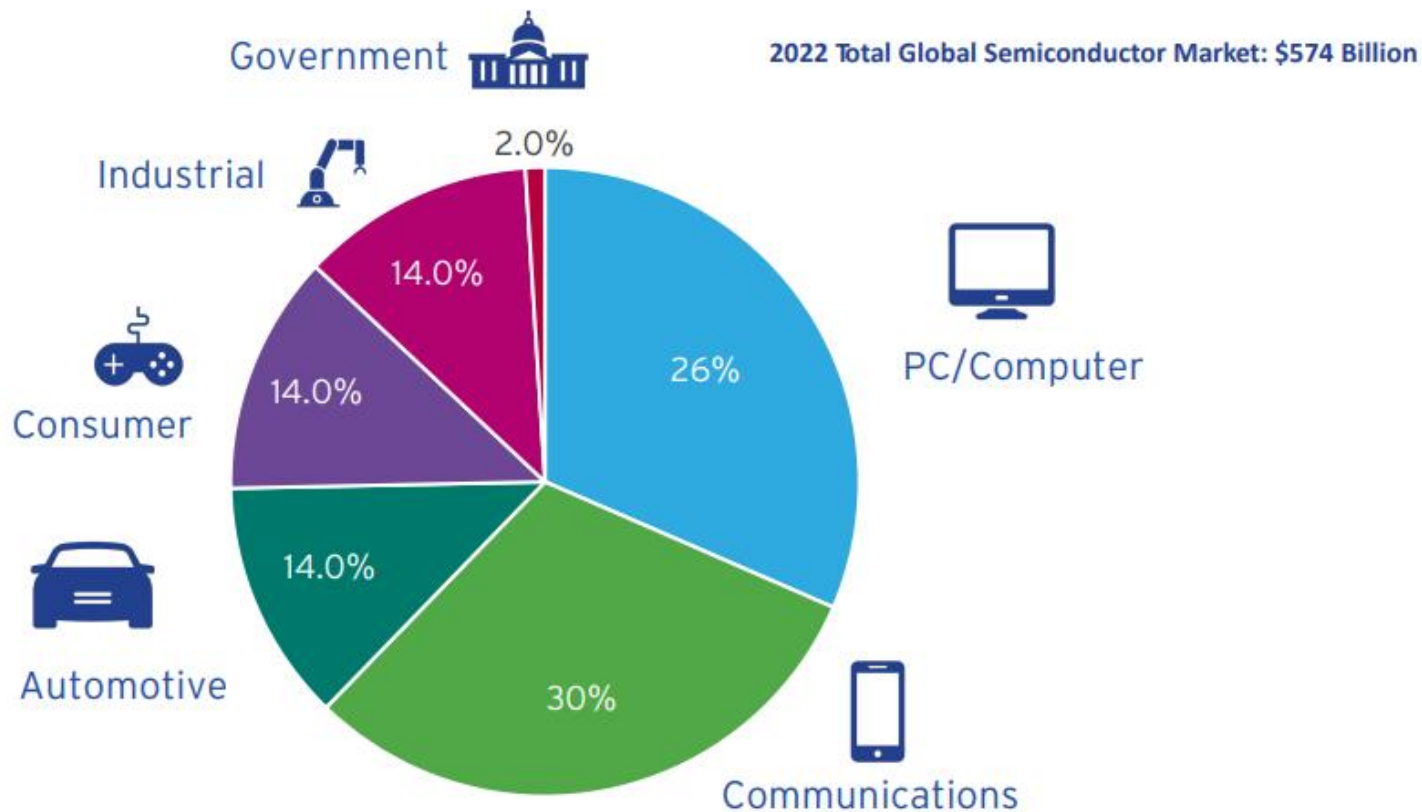
Fall 2023	Amounts in US\$M			Year on Year Growth in %		
	2022	2023	2024	2022	2023	2024
Americas	141,136	132,536	162,154	16.2	-6.1	22.3
Europe	53,853	57,048	59,480	12.8	5.9	4.3
Japan	48,158	47,209	49,275	10.2	-2.0	4.4
Asia Pacific	330,937	283,333	317,455	-3.5	-14.4	12.0
<b>Total World - \$M</b>	<b>574,084</b>	<b>520,126</b>	<b>588,364</b>	<b>3.3</b>	<b>-9.4</b>	<b>13.1</b>
Discrete Semiconductors	33,993	35,951	37,459	12.0	5.8	4.2
Optoelectronics	43,908	42,583	43,324	1.2	-3.0	1.7
Sensors	21,782	19,417	20,127	13.7	-10.9	3.7
Integrated Circuits	474,402	422,174	487,454	2.5	-11.0	15.5
Analog	88,983	81,051	84,056	20.1	-8.9	3.7
Micro	79,073	76,579	81,937	-1.4	-3.2	7.0
Logic	176,578	174,944	191,693	14.0	-0.9	9.6
Memory	129,767	89,601	129,768	-15.6	-31.0	44.8
<b>Total Products - \$M</b>	<b>574,084</b>	<b>520,126</b>	<b>588,364</b>	<b>3.3</b>	<b>-9.4</b>	<b>13.1</b>

图：1999-2024年全球半导体产品销售总额预测（亿美元）



- ◆ 根据美国SIA发布2023年Factbook白皮书，绝大多数半导体需求是由消费者最终购买的产品驱动的，如笔记本电脑和智能手机。新兴市场消费者的需求越来越多地发挥驱动作用，这些新兴市场包括亚洲、拉丁美洲、东欧和非洲等。

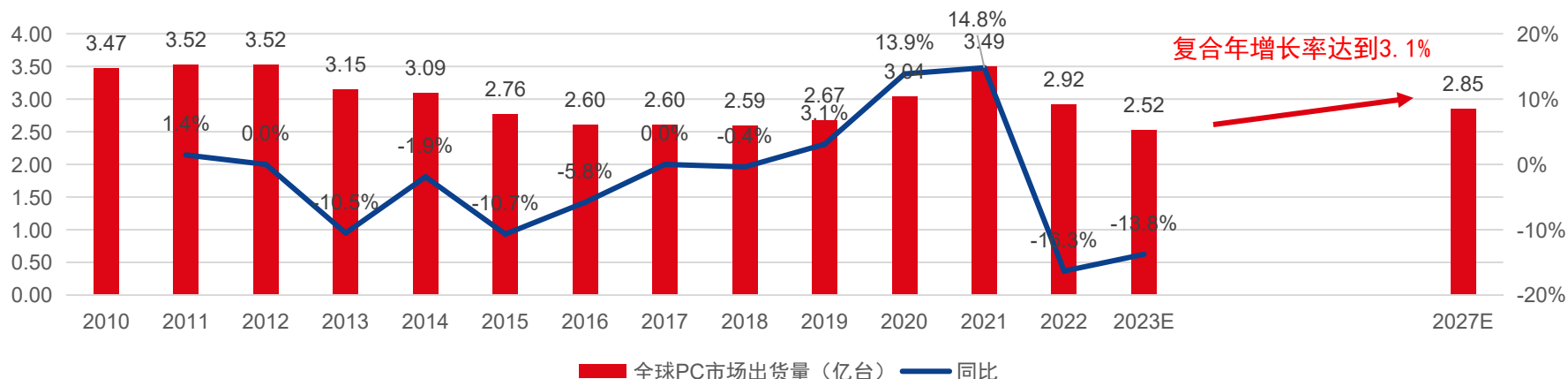
图：绝大多数半导体需求是由消费者最终购买的产品驱动



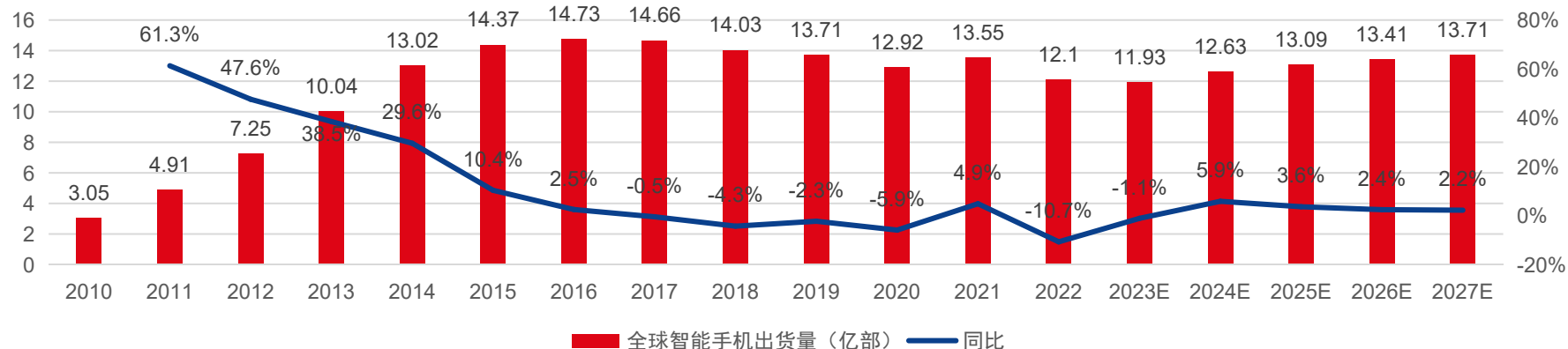
# 全球PC、智能手机等出货量预测

- ◆ 根据IDC预测，2023年至2027年全球PC出货量的复合年增长率达到3.1%，增长动因来自于老机换新、AI PC推动的产品升级等。智能手机方面，IDC预测，2023年，全球智能手机市场出货量将会低于12亿台，同比下降1.1%；而中国市场的出货量预计将仅有2.83亿台，同比也会下降1.1%。2024年全球智能手机市场出货量12.63亿，同比增长5.9%；中国智能手机市场出货重新回到3亿市场大盘，同比增长6.2%。

图：2010-2027年全球PC出货量预测（亿台）



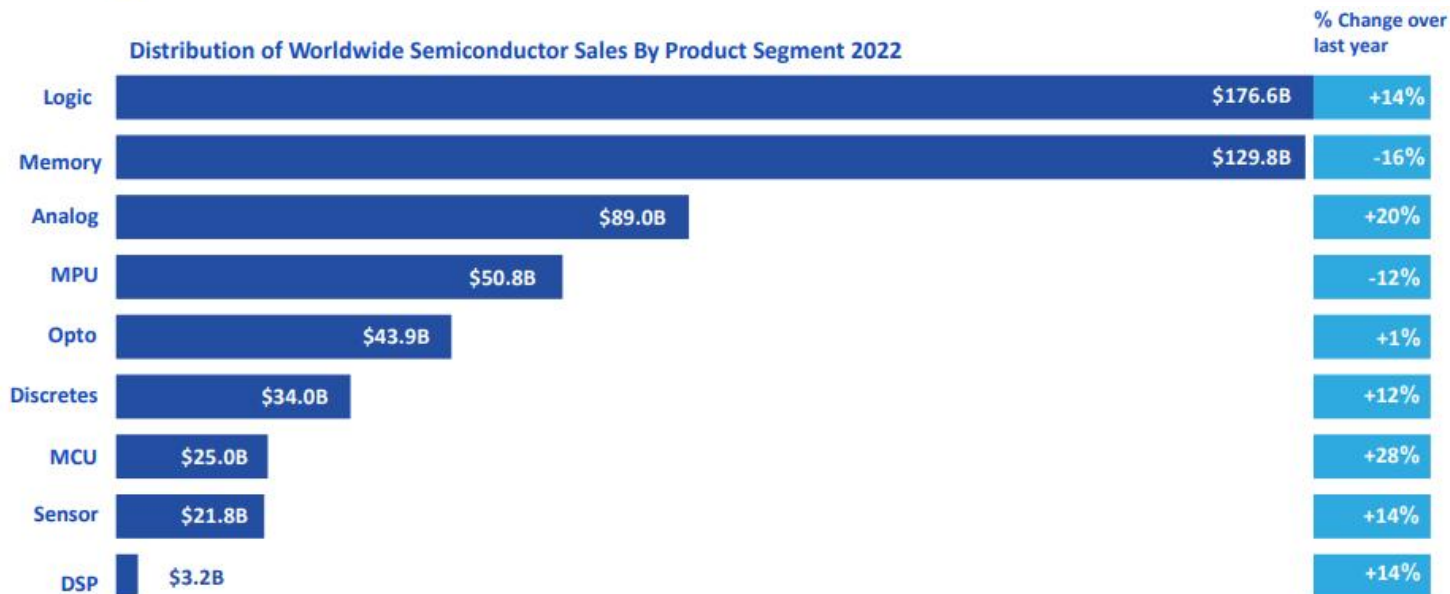
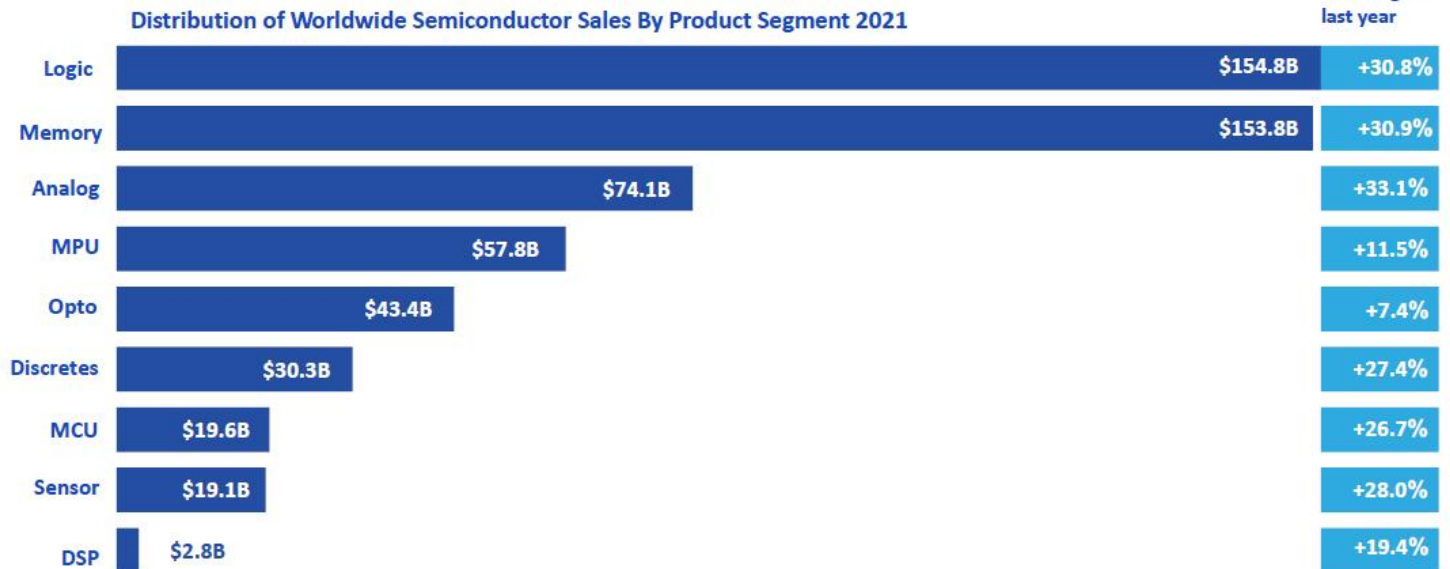
图：2010-2027年全球智能手机出货量预测（亿部）



# 逻辑器件是半导体产品市场规模最大的细分品类

◆ 根据SIA的报告  
显示，半导体产品细分市场中，逻辑、存储、模拟、MPU四类占据了大部分份额。其中逻辑器件规模最大。

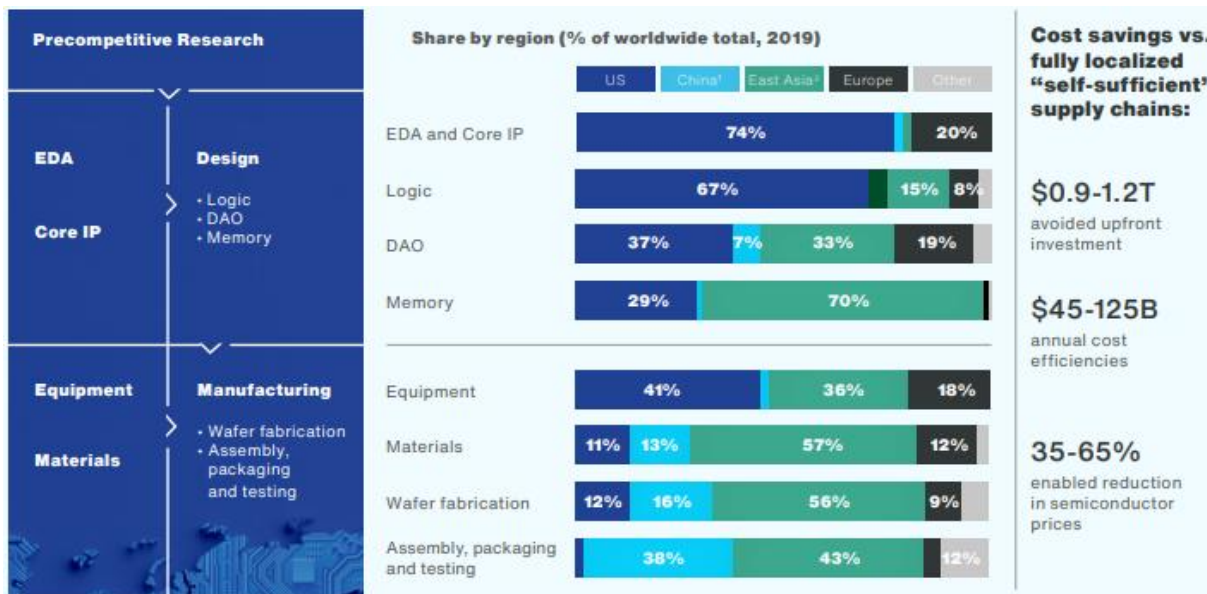
◆ 按照分类逻辑芯片主要包括CPU、GPU等通用处理器及FPGA、CPLD等专用性较强的逻辑芯片。



# 逻辑芯片市场主要是美国大厂主导

- ◆ 根据SIA发布的2023年Factbook白皮书，2022年美国半导体公司拥有最大的市场份额，占48%。其他国家半导体产业占全球市场份额的7%到20%不等。
- ◆ 根据BCG analysis的分析，2019年逻辑芯片市场美国半导体厂商占据了67%的市场份额，东亚（中国大陆除外）占了15%的市场份额，欧洲占了8%的市场份额，中国大陆占比仅个位数。
- ◆ 根据集邦咨询统计的2023年第三季度前十大IC设计业者营收中（仅统计公开财报的），前五大厂商中，仅博通的主业不是逻辑芯片为主。

图：不同区域在全球半导体核心环节占比

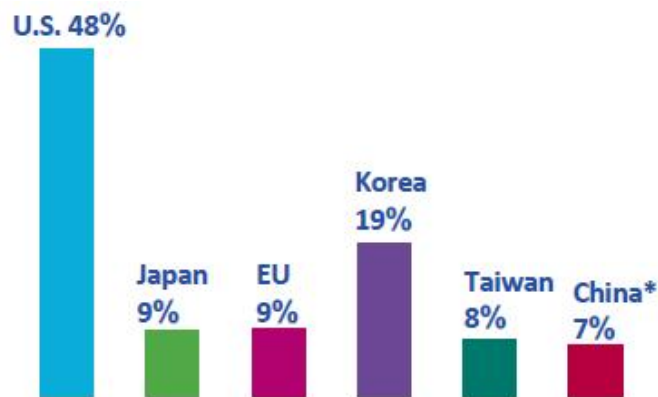


图：2023年第三季度全球前十大IC设计业者营收

Rank	Company	Revenue			Market Share	
		3Q23	2Q23	QoQ	3Q23	2Q23
1	NVIDIA	16,512	11,332	45.7%	36.9%	29.8%
2	Qualcomm	7,374	7,174	2.8%	16.5%	18.9%
3	Broadcom	7,198	6,897	4.4%	16.1%	18.2%
4	AMD	5,800	5,359	8.2%	13.0%	14.1%
5	MediaTek	3,474	3,195	8.7%	7.8%	8.4%
6	Marvell	1,393	1,335	4.4%	3.1%	3.5%
7	Novatek	913	987	-7.5%	2.0%	2.6%
8	Realtek	842	856	-1.7%	1.9%	2.3%
9	Will Semiconductor	752	528	42.3%	1.7%	1.4%
10	Cirrus Logic	481	317	51.7%	1.1%	0.8%
前十大业者合计		44,737	37,980	17.8%	100%	100%

注：  
 1. 此排名仅统计公开财报之前十大厂商。  
 2. Qualcomm仅计算QCT部门营收；NVIDIA扣除OEM/IP营收；Broadcom仅计算半导体部门营收；Will Semiconductor仅计算半导体设计及销售营收。  
 3. NTD:USD exchange rate= 3Q23 31.6957 ; 2Q23 30.7124  
 4. RMB: USD exchange rate=3Q23 7.2445; 2Q23 7.0130  
 Source: TrendForce, Dec., 2023

图：2022年不同区域在半导体全球市场占比

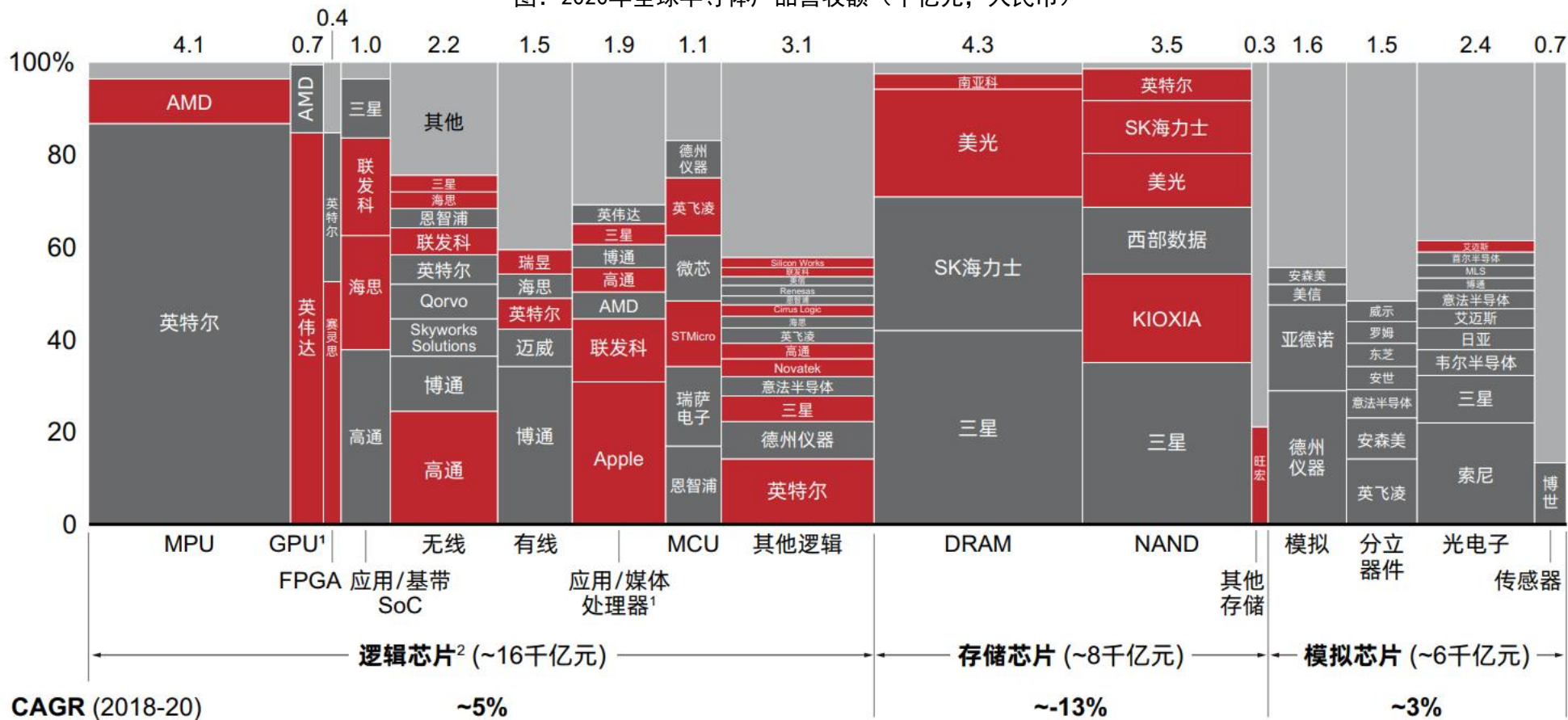


\*2022 data for the Chinese market is incomplete, so market share percentage based on 2021.

# 逻辑芯片有多种品类，市场相对分散

- 逻辑芯片具有多种品类，根据Gartner的定义，包括了MPU、GPU、FPGA、应用/基带SoC、无线、有线、应用/媒体处理器、MCU和其他逻辑等，市场相对分散。

图：2020年全球半导体产品营收额（千亿元，人民币）



注：1.GPU和应用/媒体处理器均为独立，而非集成到SoC里的；2.逻辑芯片包括模拟ASIC、ASSP；细分市场定义基于Gartner；其他包括细分市场内收入<9亿美元的公司

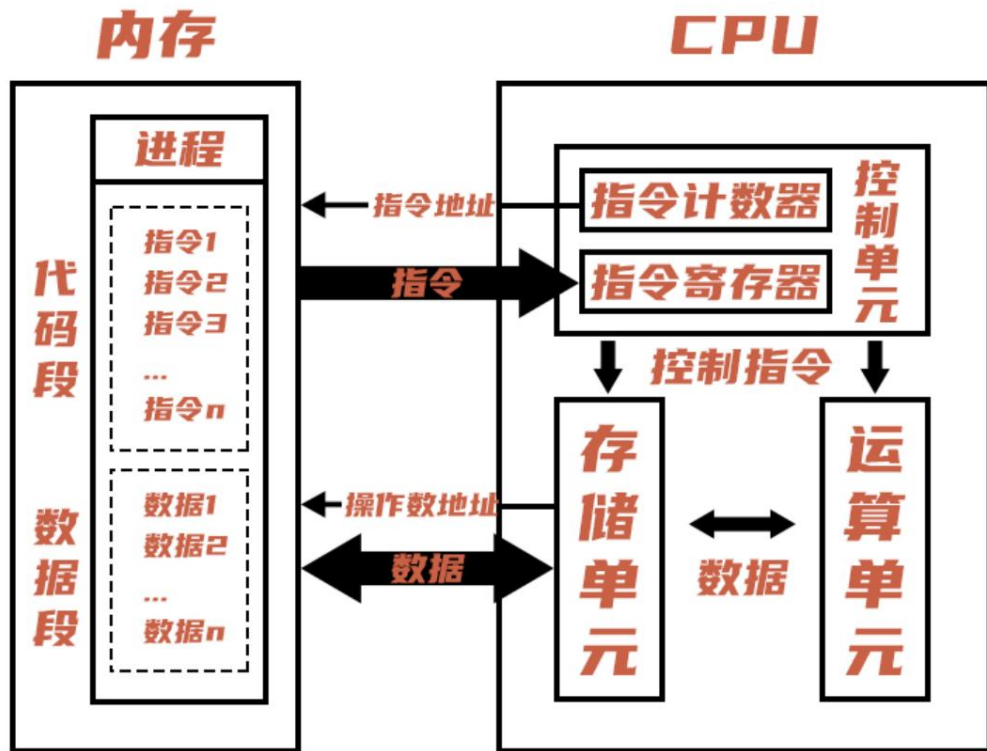
■ 增速跑赢细分市场平均水平的企业 (CAGR 2018 -20)

# 算力芯片的分类

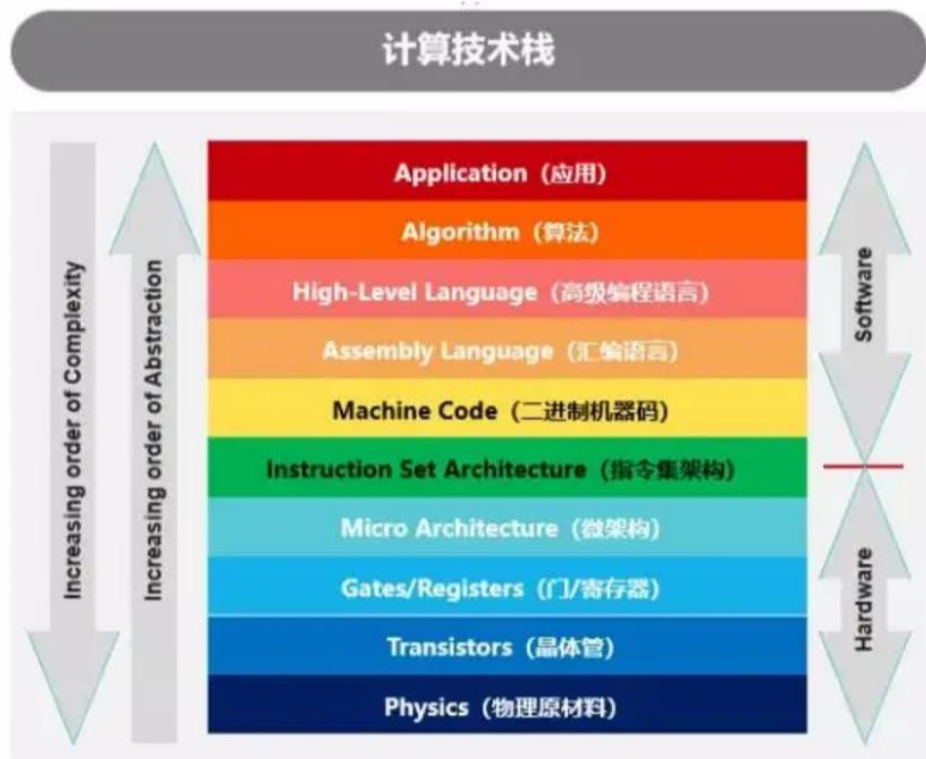
名称	CPU	GPU	FPGA	ASIC
定义	中央处理器	图像处理器/ 并行计算处理器	现场可编程逻辑门阵列	专用处理器
作用	逻辑运算和任务调度	图形显示/ 神经网络机器学习	定制化算法和计算 且可编程	定制化算法和计算
成本	高	高	高	低
计算类型	标量运算	矢量运算	稀疏运算	特定算法
能效比	低	中	优	优
优点	最通用、可计算复杂运算、产品成熟	峰值计算能力强、产品成熟	平均性能较高、功耗较低、灵活性强、时延小	AI算力强、体积小、重量轻、功耗低、保密性强
缺点	计算较为通用，算力相对小	能效比低、功耗高	量产单价高、峰值计算能力较低、编程难度大	前期投入成本高、不可编辑、研发时间长、技术风险大
代表厂商	英特尔、AMD	NVIDIA、AMD、英特尔	AMD（赛灵思）、英特尔	谷歌、寒武纪

- ◆ 中央处理器（central processing unit，简称CPU）作为计算机系统的运算和控制核心，是信息处理、程序运行的最终执行单元，电脑和智能设备的控制中枢。
- ◆ CPU内部主要由运算器、控制器和寄存器组成，运算器执行数值计算，寄存器储存数据。

图：CPU的工作原理

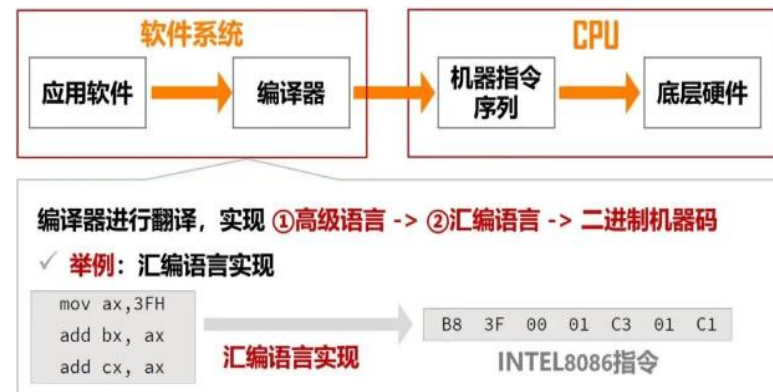


图：CPU的计算技术栈



- ◆ 指令就是指示计算机硬件执行某种运算、处理功能的命令。指令集又称指令系统架构，是CPU运行的软件的二进制编码格式，是一种指令编码的标准规范，存储于CPU内部，用来引导CPU进行加减运算和控制计算机操作系统的一系列指令集合。
- ◆ 指令集可以分为复杂指令集（CISC）和精简指令集（RISC），CISC的代表是X86；RISC的代表是ARM。CPU指令集的兼容性是生态规模的重要因素，构建高吸附性生态是竞争难点。RISC-V是一种新兴的开源精简指令集架构，因其具有开源、免费、开放、自由等特点而呈加速发展之势。

图：指令集可视为处理器支持所有操作的词汇表

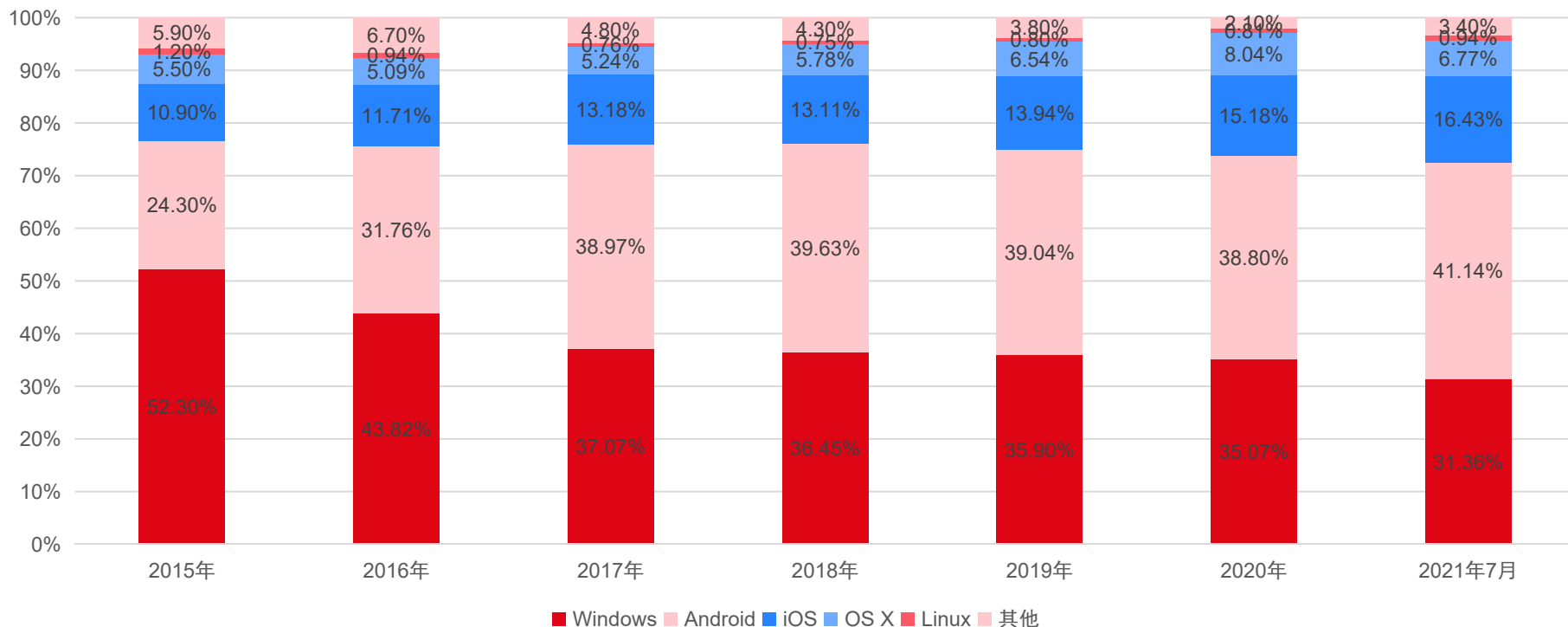


图：三大主流指令集架构对比

项目	X86架构	ARM架构	RISC-V架构
开发厂商	英特尔	ARM Holdings PLC	加州大学伯克利分校开发，目前由RISC-V基金会负责保护和推广
发布时间	1978年	1985年（ARM1 Sample）	2010年
指令集架构类型	CISC（复杂指令集架构）	RISC（精简指令集架构）	RISC（精简指令集架构）
架构文档篇幅	数千页	数千页	不足300页
指令集数目	指令集数目繁多	指令集数目繁多	基本指令集40多条，合计低于300条
模块化设计	不支持	不支持	支持
可扩展性	不支持	不支持	支持
版本兼容性	不同模块不兼容	不同的版本架构不兼容	兼容性较高
技术自主	否	否	是
技术可控	否	是	是
授权费收取情况	目前仅英特尔、AMD、VIA（台湾威盛）三家拥有授权，不对外授权	一次性授权费、与产品销量挂钩的版税提成Royalty、技术咨询费用	免费开源
主要应用领域	个人计算机、服务器、超级计算机	移动智能终端、物联网	物联网
主要短板	成本较高，应用弹性低	成本较高，应用弹性低	生态处于发展阶段，尚未成熟

- ◆ 操作系统是连接硬件和数据库、中间件、应用软件的纽带，是承载各种信息设备和软件应用运行基础平台的重要基础性软件。CPU行业目前由两大生态体系主导：一是基于X86指令系统和Windows操作系统的Wintel体系，垄断个人计算机和服务器处理器市场；二是基于ARM指令系统和Android操作系统的AA体系，主导了智能手机和物联网芯片处理器市场。前者生态相对封闭，后者生态相对开放，芯片厂商需要获得ARM的授权。
- ◆ 根据statcounter数据显示，在全球操作系统市场中，目前主要有Windows、Android、iOS、OS X、Linux五大操作系统竞争。

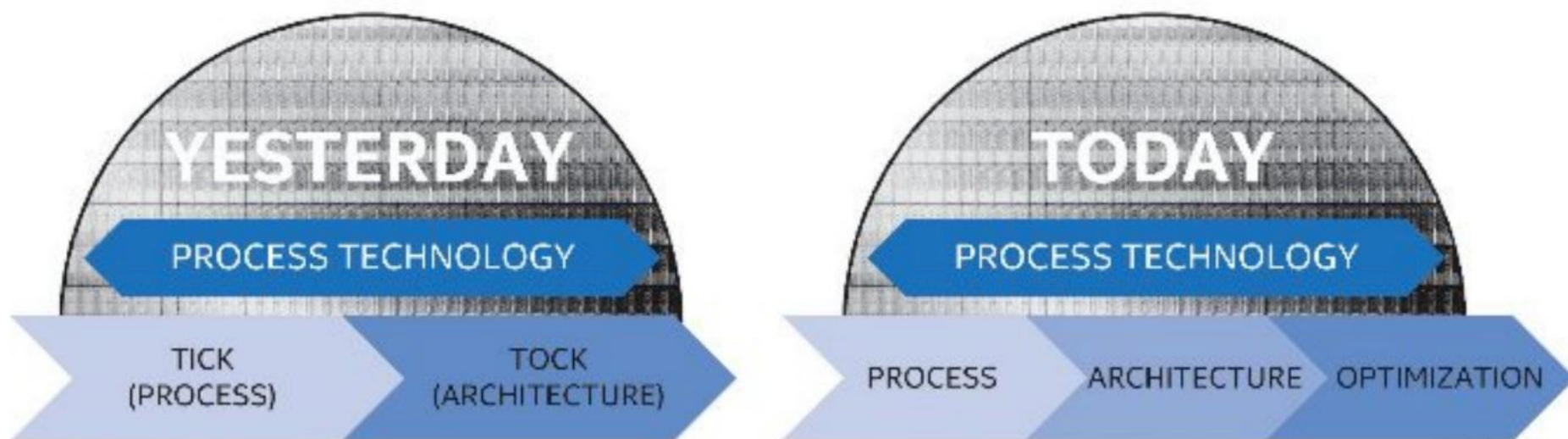
图：2015-2021年全球主要操作系统占比份额变化



# 从“Tick-Tock”模式到“PAO”模式

- ◆ Tick-Tock模型是Intel公司2006年提出的CPU发展路线，内含了提升CPU性能的两大主要驱动力，即是微结构改进和制程升级。“Tick”是提升CPU的制造工艺，“Tock”是带来更好的微结构设计。
- ◆ 2017年，Intel对Tick-Tock周期进行了修正，从10nm制程CPU开始改为“制程-架构-优化”的“三步走”战略，每次迭代周期拉升到3年。

图：从“Tick-Tock”模式到“PAO”模式



# 制程升级是提升CPU性能的重要手段之一

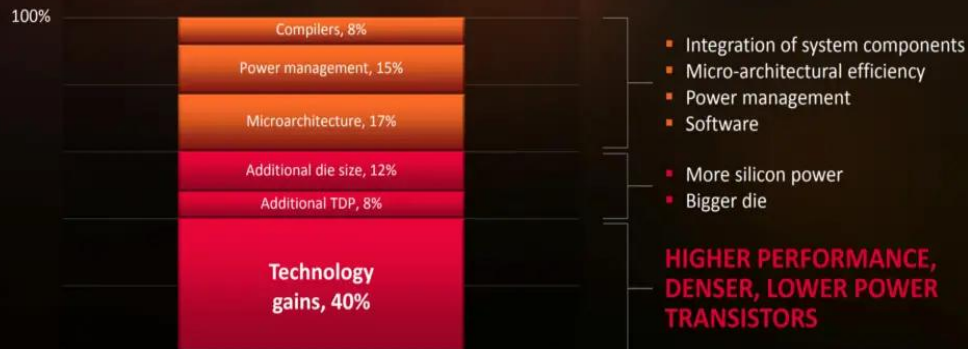
- ◆ 根据AMD的数据，过去十年（2017往前），制程升级对CPU性能提升的贡献度在40%，其他推动因素依次分别为软件优化、电源管理、增大芯片面积等。
- ◆ 随着制造技术越来越复杂，发展、研究和开发时间也越来越长。台积电在2022Q4宣布量产3纳米鳍式场效晶体管制程。根据台积电规划，台积电N2技术将采用纳米片晶体管架构，计划于2025年开始生产。

图：制程升级为半导体性能提升的主要贡献

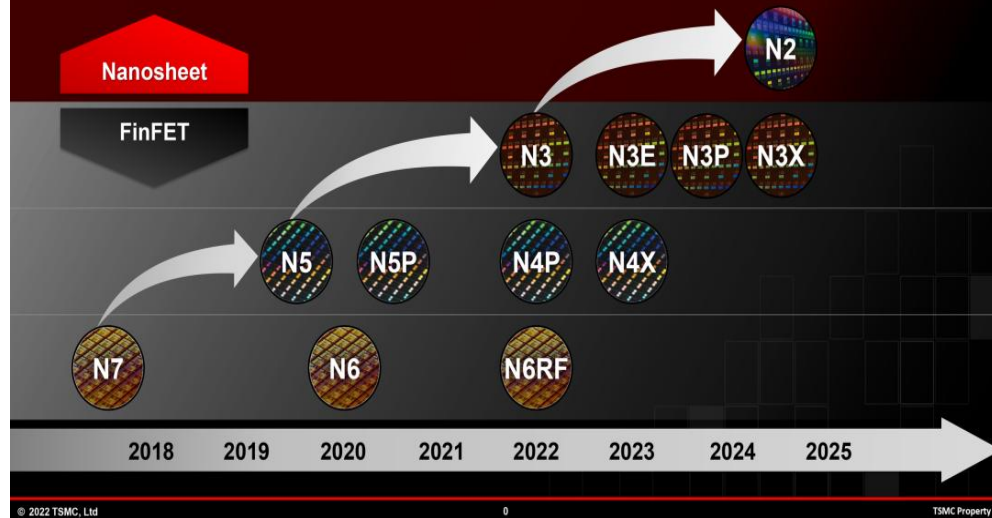
图：台积电制程规划图

## PERFORMANCE GAINS OVER THE PAST DECADE

Elements of 2X in 2.4 Year Performance Gain over the Past Decade



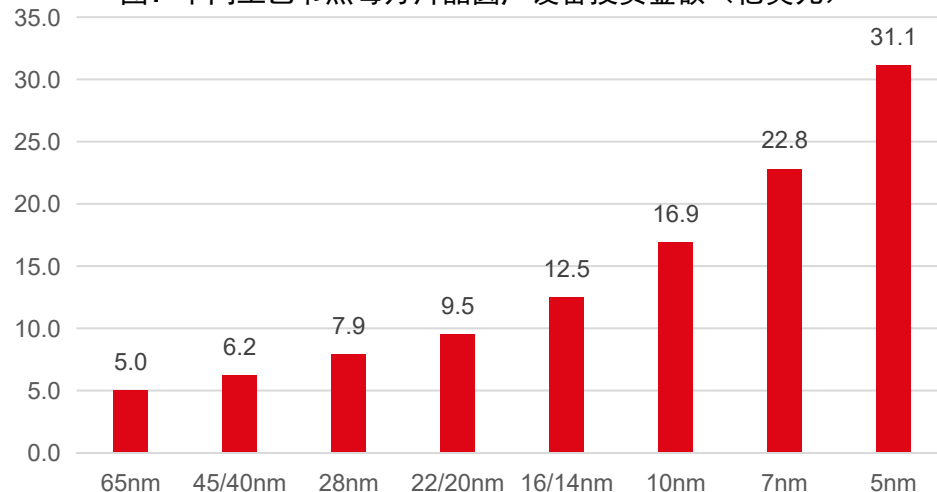
## Industry-leading Advanced Technology Portfolio



# 先进制程投入巨大，供应格局愈加集中

- ◆ 根据集微咨询的数据显示，晶圆代工厂的投资成本会随着工艺节点的先进程度提高而增长，随着先进制程的开发，芯片制程缩小对晶圆代工厂带来的建设成本急速上升，每万片5nm芯片的晶圆厂建设成本已达到28nm芯片晶圆厂建设成本的5-6倍。
- ◆ 根据集邦咨询报道，英特尔已经向台积电下达3nm工艺订单，用于生产即将推出的Lunar Lake芯片，这标志着台积电首次成为英特尔主流笔记本CPU的独家生产商。

图：不同工艺节点每万片晶圆厂设备投资金额（亿美元）



- ◆ GPU又称显示核心、视觉处理器，是一种专门在个人电脑、工作站、游戏机和一些移动设备（如平板电脑、智能手机等）上做图像和图形相关运算工作的微处理器。NVIDIA公司在1999年发布GeForce 256图形处理芯片时首先提出GPU的概念。从此NVIDIA显卡的芯片就用这个新名字GPU来称呼。GPU使显卡削减了对CPU的依赖，并执行部分原本CPU的工作，尤其是在3D图形处理时。
- ◆ 可编程性与浮点计算能力相结合，基于GPU的通用计算也开始出现，GPU朝着通用计算的方向持续演进。2006年，英伟达CUDA（Compute Unified Device Architecture，统一计算设备架构），及对应工业标准的OpenCL的出现，让GPU实现更广泛的通用计算功能，GPGPU的概念落地。

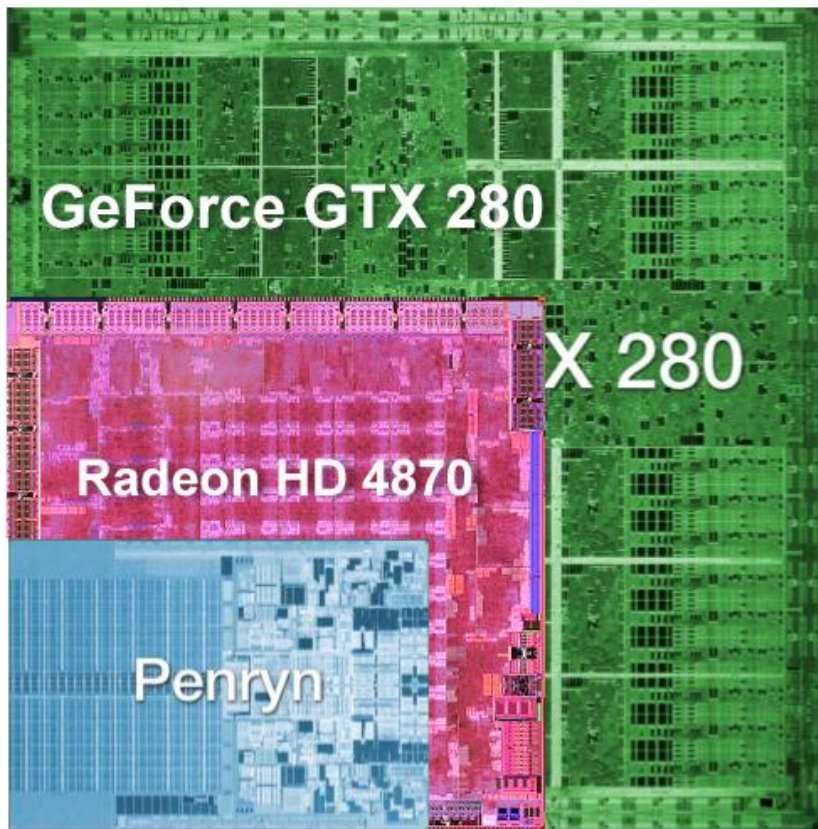
## 英伟达架构两年升级一次



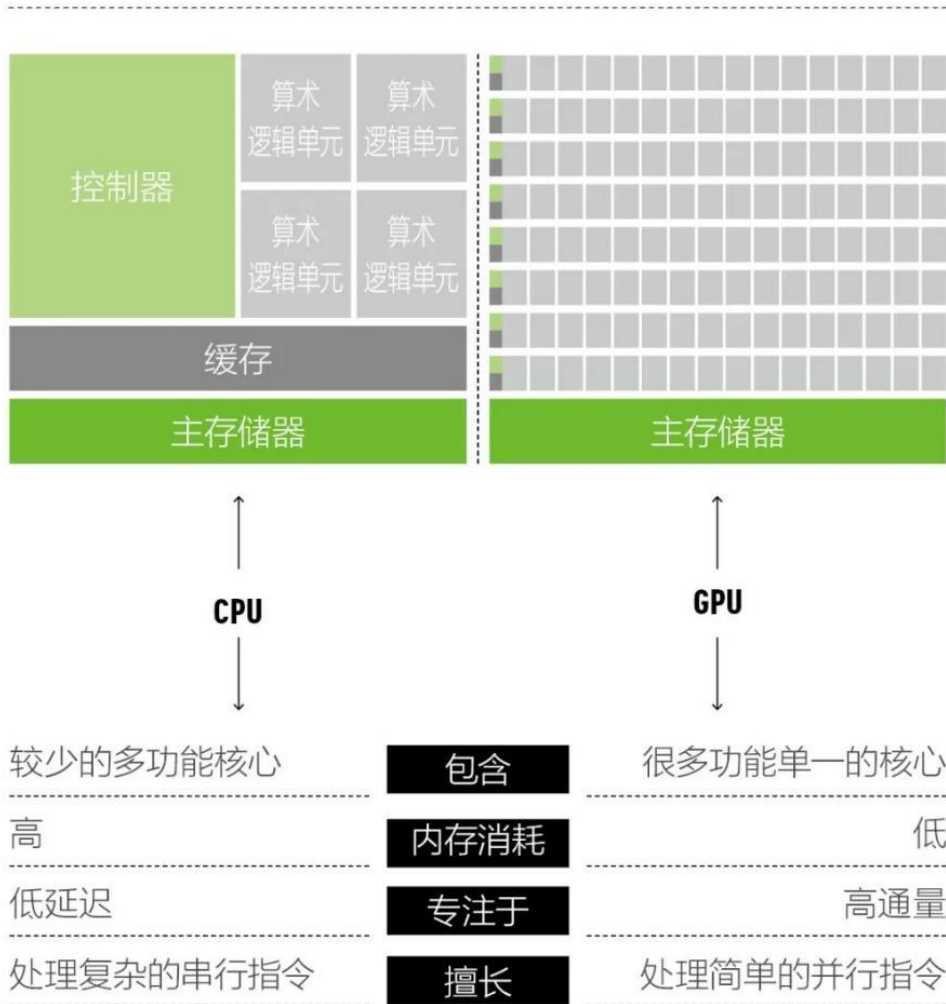
# 与CPU对比，GPU拥有更多的计算资源

- ◆ CPU和GPU都是计算机处理器，都能执行计算任务，其不同点在于，CPU更擅长线性计算，而GPU更擅长并行计算。CPU是基于低延时的设计；GPU是基于大的吞吐量设计。

图：CPU与GPU芯片面积对比

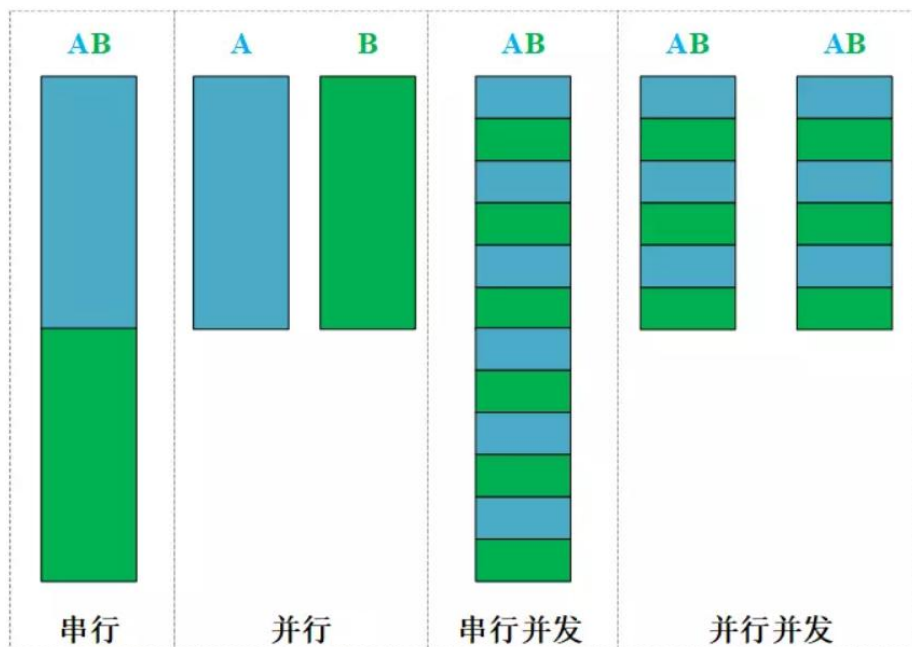


## CPU与GPU对比图

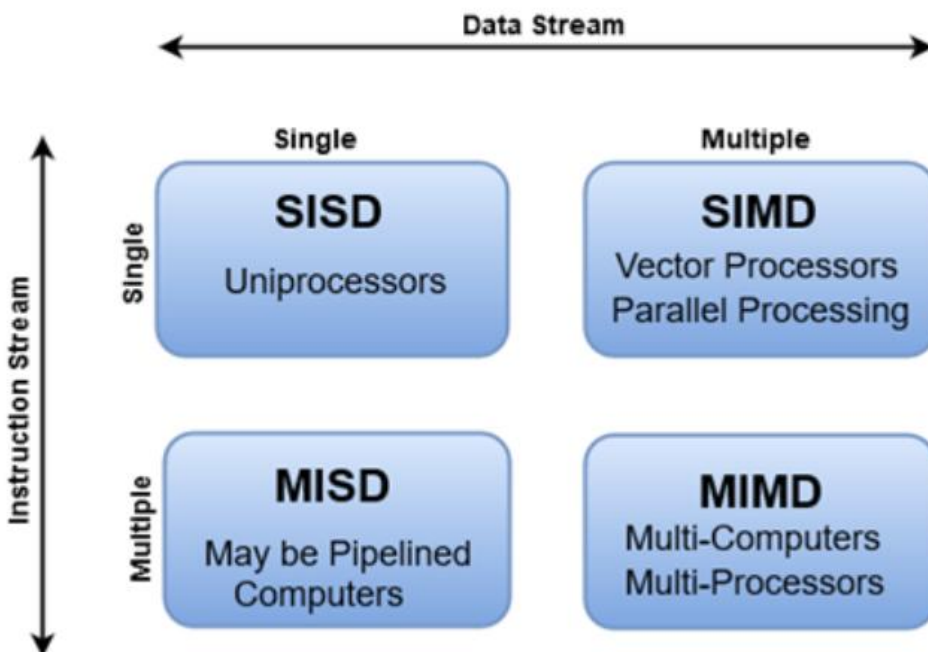


- ◆ 并行计算（英语：parallel computing）一般是指许多指令得以同时进行的计算模式。在同时进行的前提下，可以将计算的过程分解成小部分，之后以并发方式来加以解决。相对于串行计算，并行计算可以划分成时间并行和空间并行。时间并行即指令流水化，空间并行使用多个处理器执行并发计算。并行计算机体系结构的发展主要体现在计算节点性能的提高及节点间通信技术的改进两方面。
- ◆ 弗林分类法，根据指令流和数据流的不同组织方式把计算机体系的结构分为四类：单指令流单数据流（SISD）、单指令流多数据流（SIMD）、多指令流多单数据流（MISD）、多指令流多数据流（MIMD）。

图：串行、并行、串行并发与并行并发对比图



图：弗林分类法

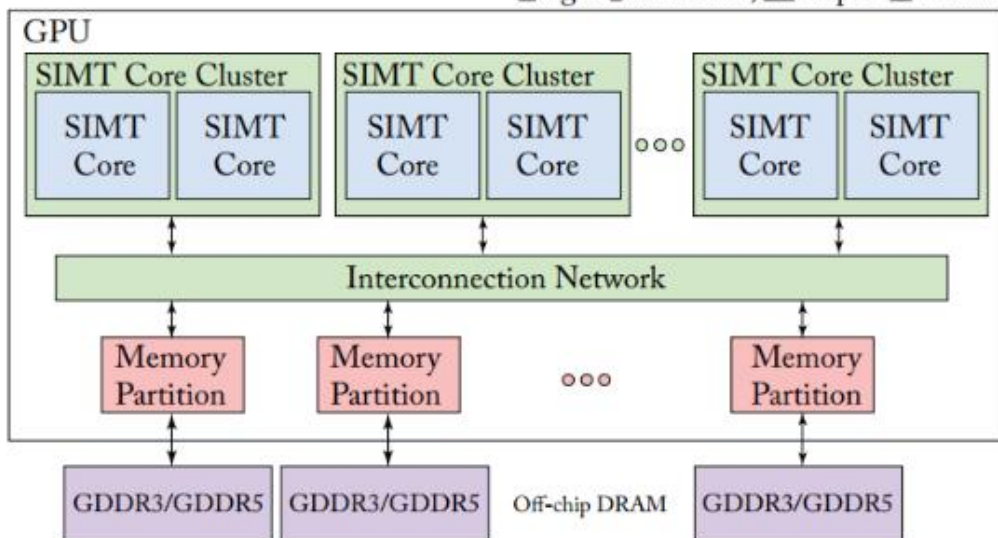


# SIMT, 主流GPU的系统架构核心

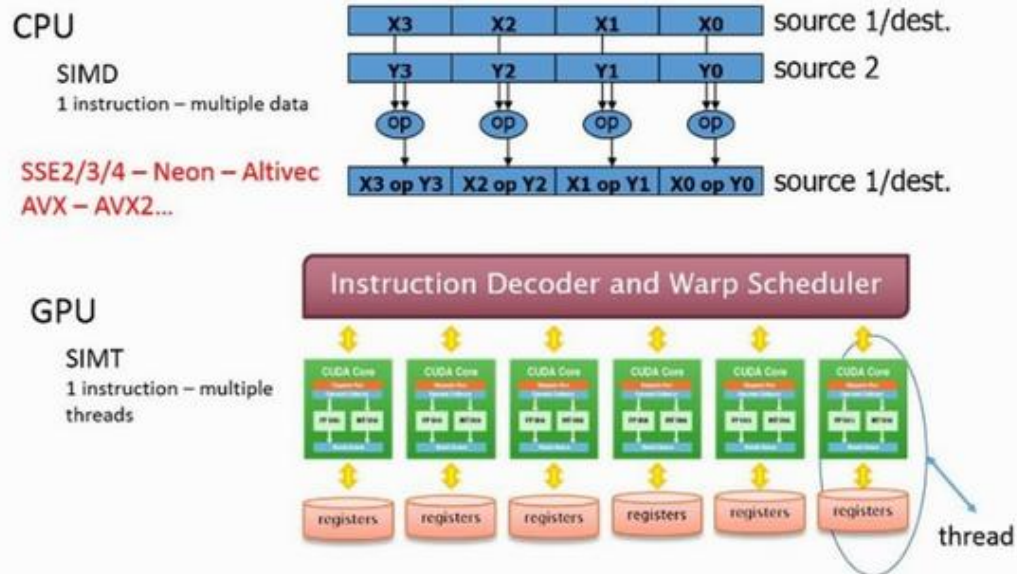
- ◆ SIMT与SIMD（同一条指令多个数据）的共同点是同一条指令。SIMT（单指令多线程）是SIMD的线程等价物，不同之处在于，SIMD使用执行单元或矢量单元，而SIMT将其扩展为利用线程。SIMT的好处是无需开发者费力把数据凑成合适的矢量长度，并且SIMT允许每个线程有不同的分支。SIMT的主要优点是它减少了指令预取带来的等待时间。
- ◆ SIMD是CPU上用的，SIMT是GPU上用的。SIMD更像是CPU的一个小扩展，SIMT是GPU上并发性的核心保证。

图：现代GPU简单架构示意图

Single-Instruction, Multiple-Threads



图：SIMD与SIMT对比



# Tensor Core, 面向AI 和 高性能计算 (HPC)

- ◆ Tensor Core (张量计算核心) 是由NVIDIA研发的新型处理核心, Tensor Core可实现混合精度计算, 并能根据精度的降低动态调整算力, 在保持准确性的同时提高吞吐量。英伟达目前已经发布了4代Tensor cores, 分别是2017年5月发布的NVIDIA Volta架构中的第一代 Tensor Core、2018年8月发布的NVIDIA Turing Tensor Core 技术、2020年5月发布的NVIDIA Ampere 架构和2022年9月20日发布的第四代Tensor Core——Ada Lovelace架构。

图: Tensor Core vs CUDA Core

图: TensorCore参数对比

Feature	Tensor Core	CUDA Core
Accuracy	Lower accuracy	High accuracy
Compute Speed	Faster compute speed	Lower compute speed
Machine Learning	Suited for machine learning	Can handle machine learning, but not ideal for it
Use Case	Low-end and high-end grade AI development	High throughput graphical processing
Use Cycle Reduction	Can reduce use cycles for mathematical operations	Cannot reduce use cycles
Operations Per Clock Cycle	Multiple operations per clock cycles	1
Graphical Processing	Not suited for graphical processing	Purpose-built for graphical processing, powers modern NVIDIA GPUs.

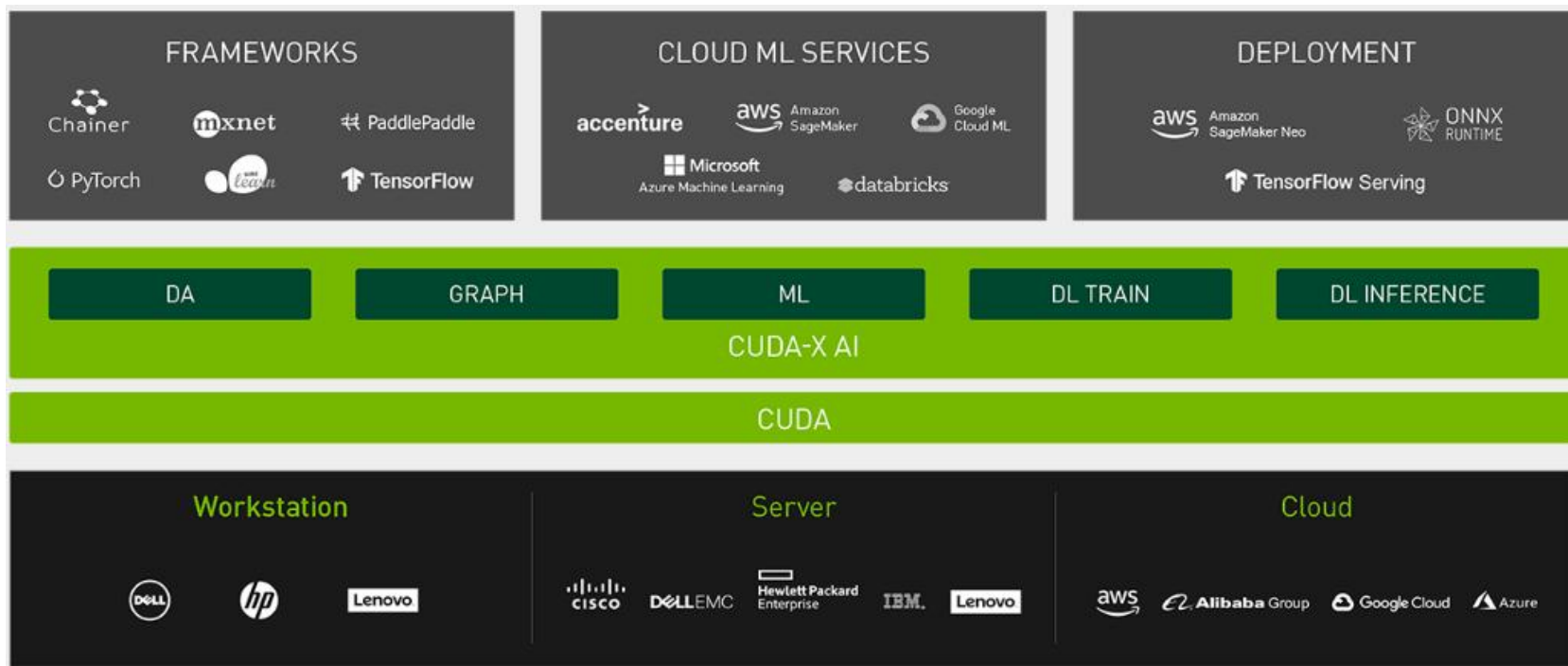
	Volta	Turing	Ampere	Hopper
Base Size*	64个F16(4x4x4)	64个F16(4x4x4)	256个F16(8x4x8)	512个F16(8x4x16)
tensorCore个数	672(8/SM)	576(8/SM)	512(4/SM)	576(4/SM)
SM内算力*	512个F16	512个F16	1024个F16	2048个F16
总算力 FOPS(F16)	125T	130.5T	312T	1000T
boost clock(MHz)	1530	1455	1410	≈A100x1.3
FP64	-	-	√	√
TF32	-	-	√	√
BF16	-	-	√	√
FP16	√	√	√	√
FP8	-	-	-	√
INT8/UINT8	-	√	√	√
INT4/UINT4	-	√	√	-
INT1	-	√	√	-

- ◆ 由于科学问题的求解需要大规模并行计算能力，很多科学家使用GPU来进行科学计算。这种场景中，经常用到高精度的数据类型。英伟达在GPU 新增了 DPU（FP64），负责64位浮点型的计算。浮点型计算是较为复杂的一个电路，随着计算位数的增加，电路面积呈指数级增长，相应的也会带来成本的大幅度增长。为了提升芯片的覆盖范围，英伟达需要在芯片面积和性能之间不断寻找平衡的方案。

架构名称	Volta-伏特	Turing-图灵	Ampere-安培	Hopper-赫柏
发布时间	2017	2018	2020	2022
核心参数	80个SM，每个SM包括32个FP64+64个Int32+64个FP32+8个Tensor Cores	102核心92个SM，SM重新设计每个SM包含64个Int32+64个FP32+8个Tensor Cores	108个SM，每个SM包含64个FP32+64个INT32+32个FP64+4个Tensor Cores	132个SM，每个SM包含128个FP32+64个INT32+64个FP64+4个Tensor Cores
特点	NVLink2.0, Tensor Cores第一代, 支持AI运算	Tensor Core2.0, RT Core第一代	Tensor Core3.0, RT Core2.0 NVLink3.0, 结构稀疏性矩阵MIG1.0	Tensor Core4.0, NVlink4.0结构稀疏性矩阵MIG2.0
纳米制程	12nm 211亿晶体管	12nm 186亿晶体管	7nm 283亿晶体管	4nm 800亿晶体管
代表产品	V100TiTan V	T4, 2080TI RTX 5000	A100、A800 A30系列	H100、H800

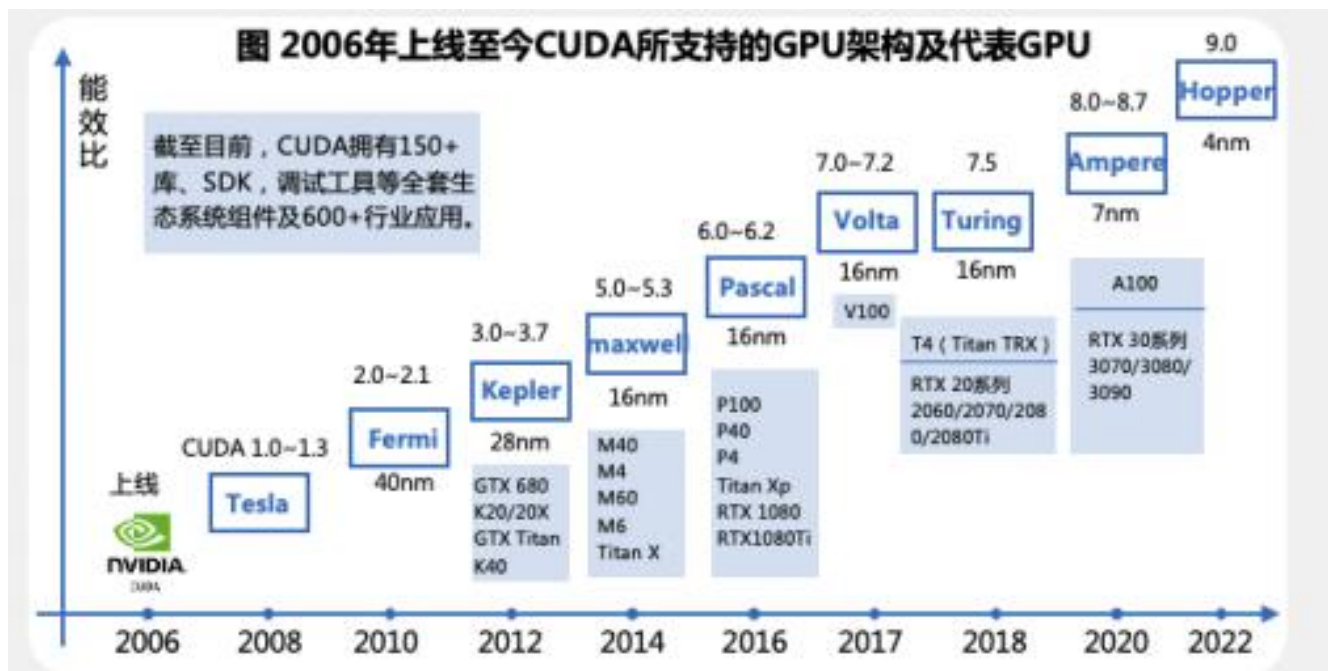
# 英伟达强大的重要推手之一：CUDA生态

- ◆ 2006年，NVIDIA推出CUDA，这是一种用于通用GPU计算的革命性架构。CUDA的存在使得开发者使用GPU进行通用计算的难度大幅降低，使得开发者可以相对简单有效地对英伟达GPU芯片进行编程，使科学家和研究人员能够利用GPU的并行处理能力来应对最复杂的计算挑战。
- ◆ 根据英伟达2023财年年报，目前有400万名开发者正在与CUDA合作。英伟达通过12年的时间达到200万名开发者，在过去的两年半里该数字翻了一番，目前CUDA的下载量已经超过了4000万次。



# 商业化竞争之“快”与生态构建之“慢”

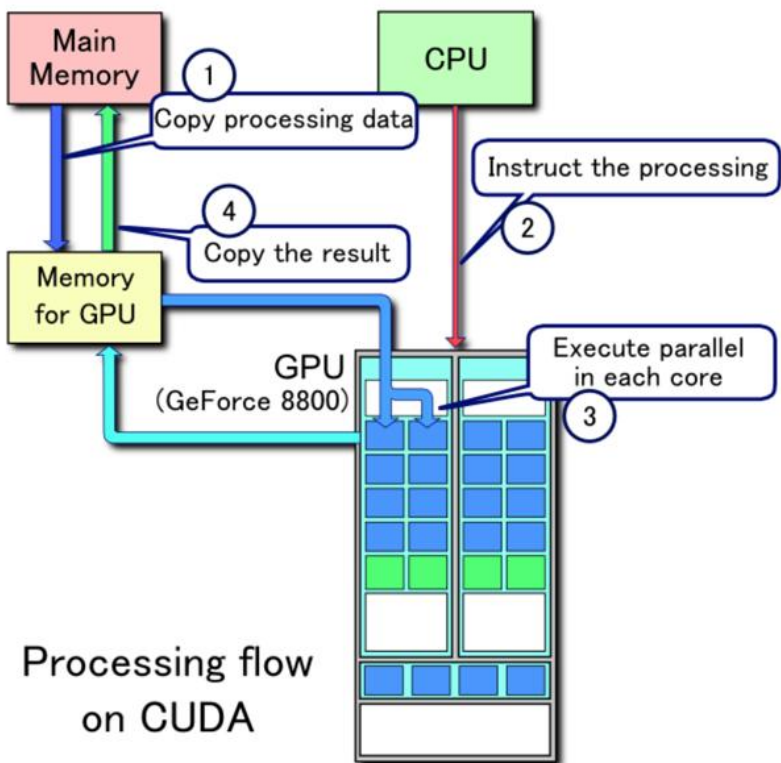
- ◆ 芯片是算力基础，但要充分发挥其性能，必须构建完备的系统软件底层库，英伟达构建了从底层系统软件、驱动软件、平台到上层的应用框架。此外，英伟达提供全面的算法库，几乎全部开源。
- ◆ CUDA包含了三个层面，编程语言、编译器和生态，是英伟达花费超15年，投入超100亿美元所构建。对于其他算力芯片厂商，软硬件的优化和生态的构建需要大量的时间，而当下AI发展迅猛，AI模型开发处于全球激烈竞争的时期，商业化竞争之“快”与生态构建之“慢”成为其他算力芯片厂商提升市场渗透率的核心阻碍之一。



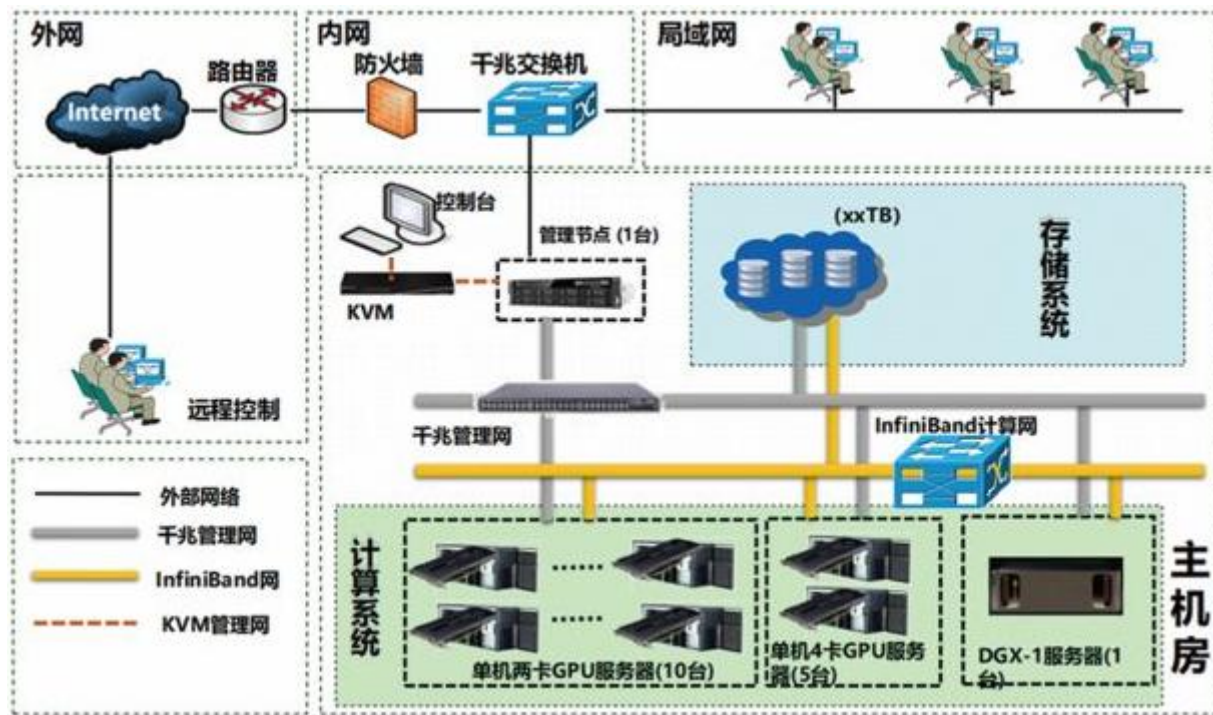
# GPU通信，影响GPU扩展集群性能的关键

- 在实际运用中，为缩减计算时间，应用程序往往同时使用多卡进行并行计算，这里的多卡包括单个服务器多张GPU卡(简称单机多卡)，和多个服务器，且每个服务器分别有多张卡的场景(简称多机多卡)。由于用于计算的数据量较大，GPU多卡之间的通信非常重要，成为影响数据吞吐量的核心要素。

图：分离式架构的CPU-GPU的数据流程图



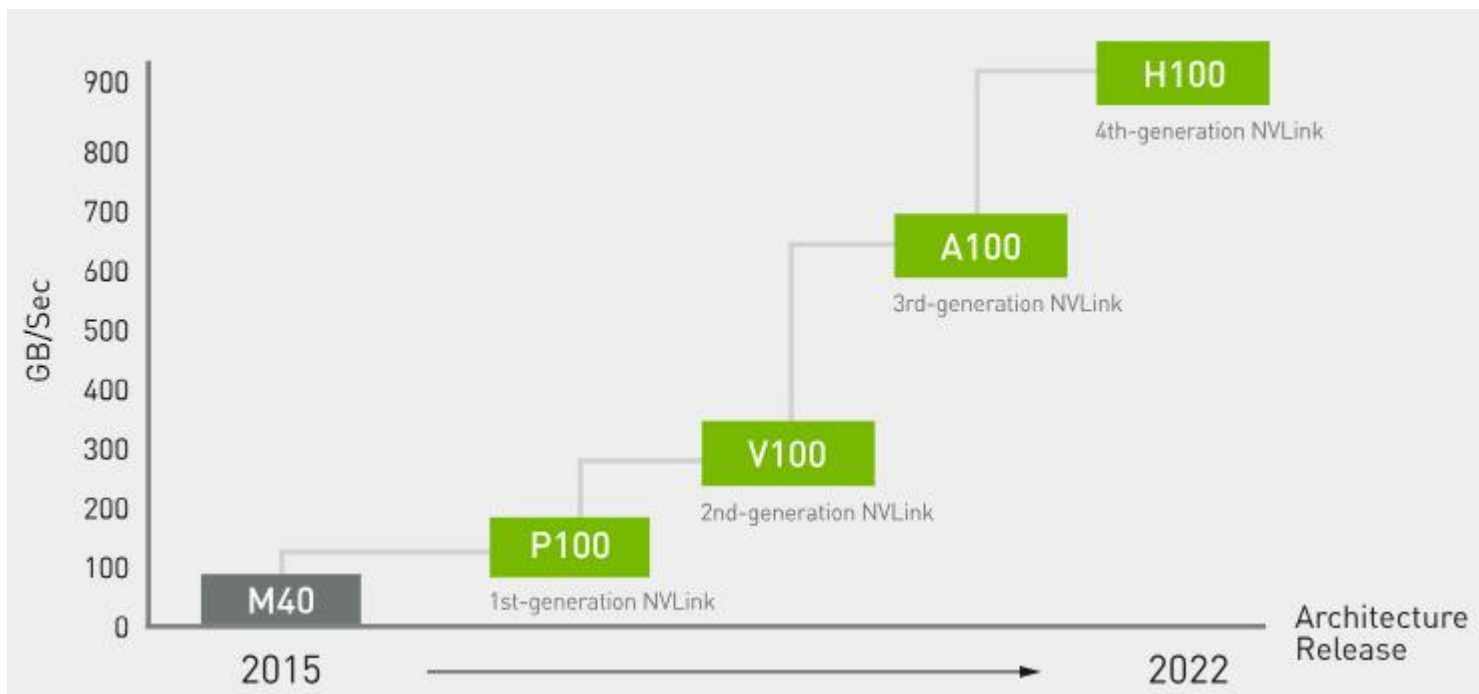
图：GPU集群架构图



# NVLink，最大化地提升系统吞吐量

- ◆ NVLink是英伟达（NVIDIA）开发并推出的一种总线及其通信协议。NVLink采用点对点结构、串行传输，用于中央处理器（CPU）与图形处理器（GPU）之间的连接，也可用于多个NVIDIA图形处理器之间的相互连接。借助该技术，可以最大化地提升系统吞吐量。
- ◆ 第四代NVIDIA® NVLink®技术可为多GPU系统配置提供高于以往1.5倍的带宽，以及增强的可扩展性。单个NVIDIA H100 Tensor Core GPU支持多达18个NVLink连接，总带宽为900GB/s，是PCIe 5.0带宽的7倍。

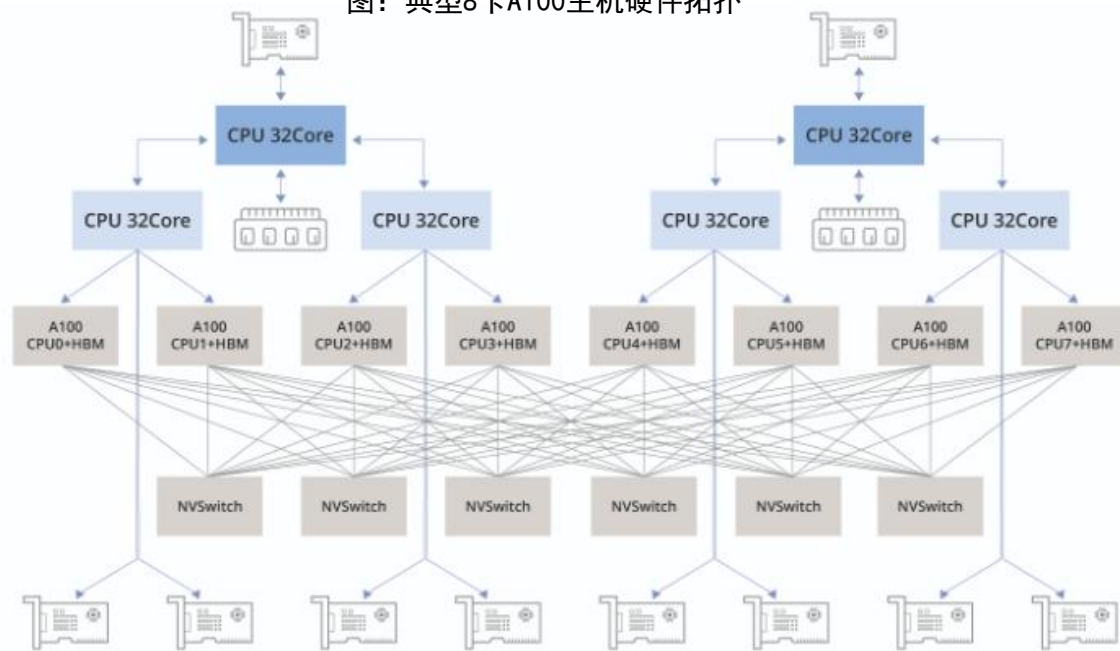
图：英伟达NVLink技术性能演进图



# NVSwitch, 助力多节点GPU集群构建

- ◆ NVSwitch是NVIDIA的一款交换芯片，封装在GPU module上，可连接多个NVLink，在单节点内和节点间实现以NVLink能够达到的最高速度进行多对多GPU通信。
- ◆ 借助NVSwitch，NVLink连接可在节点间扩展，以创建无缝、高带宽的多节点 GPU 集群，从而有效地形成数据中心大小的GPU。通过在服务器外部添加第二层NVSwitch，NVLink网络可以连接多达256个GPU，并提供57.6 TB/s的惊人多对多带宽，从而快速完成大型AI作业。

图：典型8卡A100主机硬件拓扑

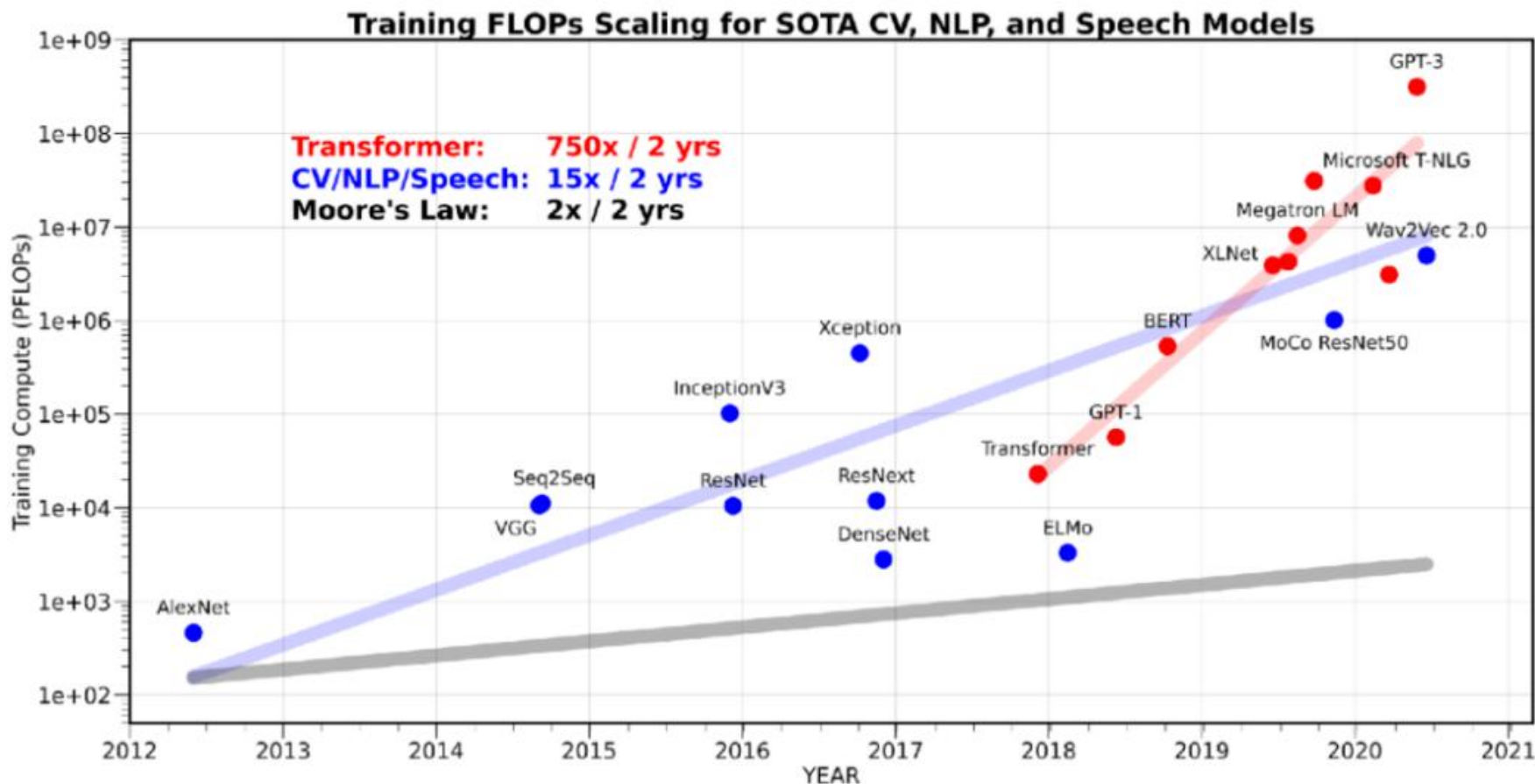


图：英伟达NVSwitch发展演进

	第一代	第二代	第三代
直连或节点中 GPU 数量	最多 8 个	最多 8 个	最多 8 个
NVSwitch GPU 之间带宽	300GB/s	600GB/s	900GB/s
聚合总带宽	2.4TB/s	4.8TB/s	7.2TB/s
NVIDIA 架构支持	NVIDIA Volta 架构	NVIDIA Ampere 架构	NVIDIA Hopper 架构

# 显存带宽，减小内存墙的关键

- ◆ 存储器数据访问速度跟不上处理器的数据处理速度，两者之间数据交换通路窄以及由此引发的高能耗两大难题，在存储与运算之间筑起了一道“内存墙”。
- ◆ 随着数据的爆炸式增长，内存墙对于计算速度的影响愈发显现。为了减小内存墙的影响，提升内存带宽一直是存储芯片聚焦的关键问题。



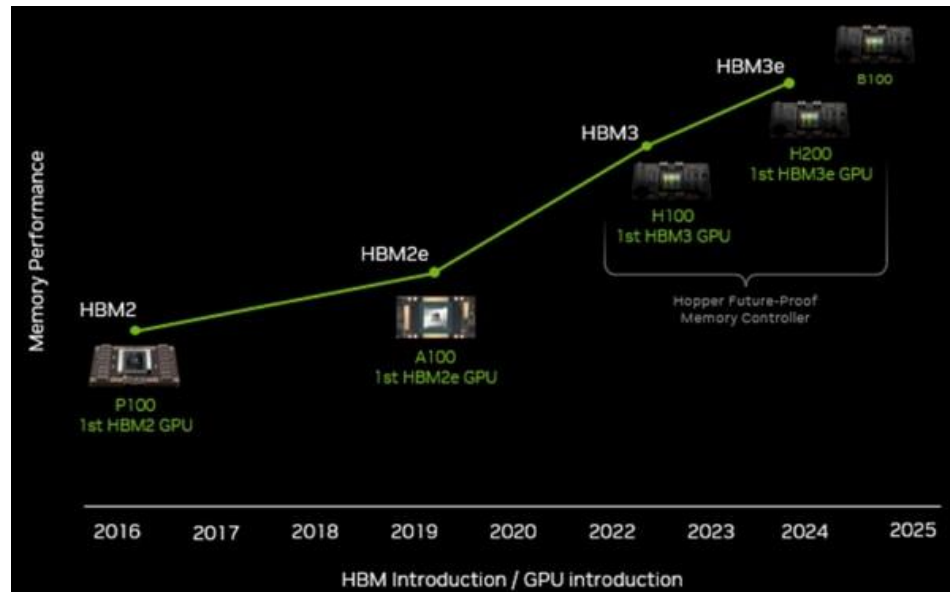
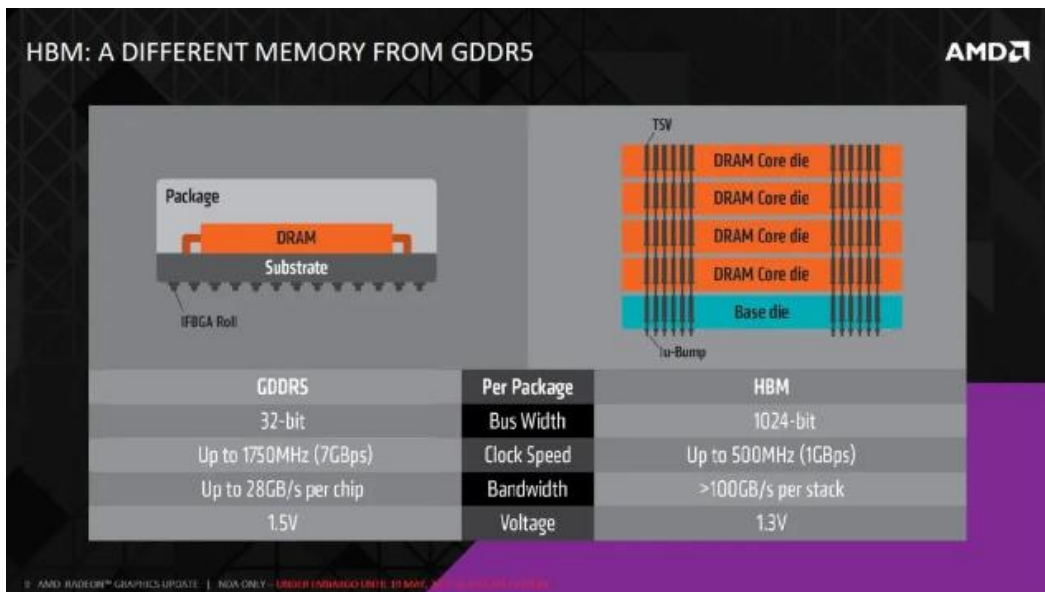
# HBM, 新一代DRAM解决方案

- ◆ HBM (High Bandwidth Memory, 高带宽内存) 将很多个DDR芯片堆叠在一起后和GPU封装在一起, 实现大容量, 高位宽的DDR组合阵列, 突破了内存容量与带宽瓶颈, 被视为新一代DRAM解决方案。



图: HBM与GDDR5对比图

图: 英伟达GPU芯片技术路线图



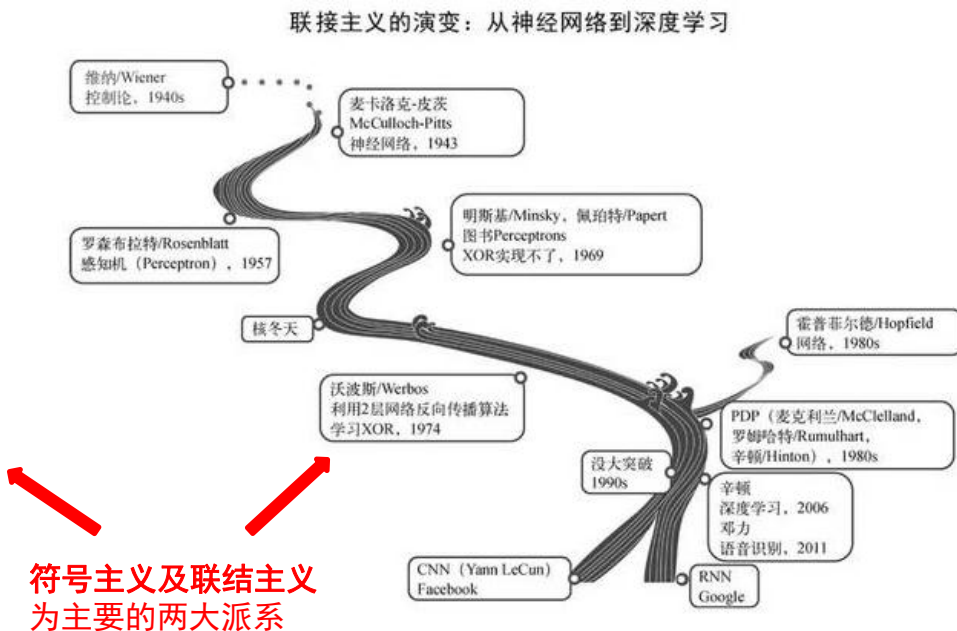
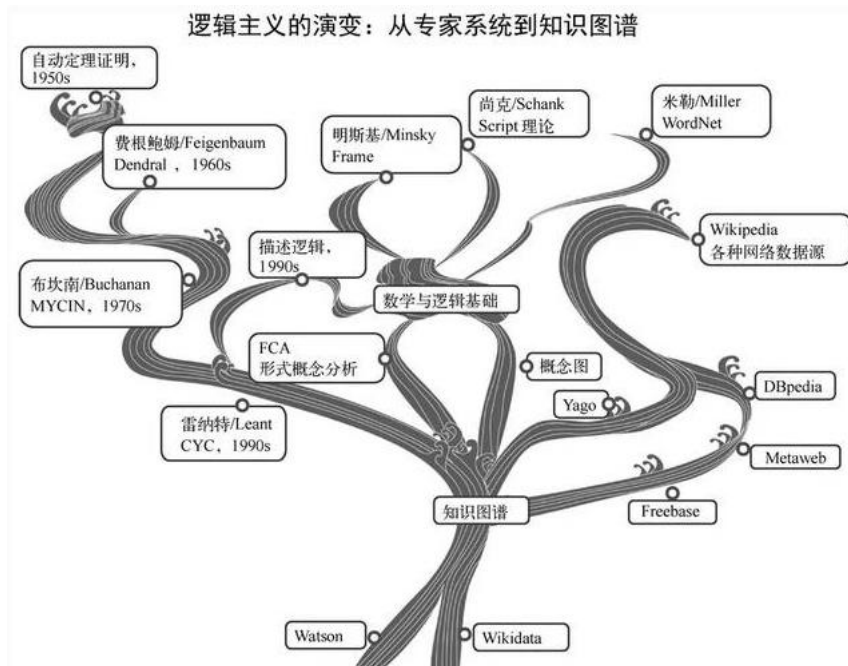
- 01 逻辑芯片市场规模巨大，产业链具备高门槛
- 02 大模型走向多模态，技术持续演进，商业前景广阔
- 03 算力板块投资的几条主线
- 04 相关标的
- 05 风险提示

- ◆ 1956年，达特茅斯会议召开，提出了“人工智能”这个名词，标志着AI作为一个独立研究领域的诞生，该会议确立了AI的目标，即创建一种机器，其智能行为可以与人类的智能相媲美。
- ◆ 由于“智能”没有定义，“人工智能”至今也尚无统一的定义。
- ◆ 现代人工智能技术是通过机器学习及由其驱动而发展起来的计算机视觉、自然语言处理和语音识别技术来实现多模态数据融合的现实交互。

序列	不同的人工智能定义
1	智能机器是一种能够呈现出人类智能行为的机器，而这种智能行为是人类用大脑考虑问题或创造思想。
2	智能机器是一种能够在不确定环境中执行各种拟人任务（anthropomorphic tasks）达到预期目标的机器。
3	人工智能是一种使计算机能够思维，使机器具有智力的激动人心的新尝试[Haugeland, 1985]
4	人工智能是那些与人的思维、决策、问题求解和学习等有关活动的自动化[Bellman, 1978]
5	人工智能是用计算模型研究智力行为[Charniak and McDermott, 1985]
6	人工智能是研究那些使理解、推理和行为成为可能的计算[Winston, 1992]
7	人工智能是一种能够执行需要人的智能的创造性机器的技术[Kurzweil, 1990]
8	人工智能研究如何使计算机做事让人过得更好[Rick and Knight, 1991]
9	人工智能是研究和设计具有智能行为的计算机程序，以执行人或动物所具有的智能任务 [Dean et al., 2003]
10	人工智能是一门通过计算过程力图理解和模仿智能行为的学科[Schalkoff, 1990]
11	人工智能是计算机科学中与智能行为的自动化有关的一个分支[Luger and Stubblefield, 1997]

# 人工智能（AI）的三大流派

- ◆ 符号主义，又称为逻辑主义、心理学派或计算机学派，其原理主要为物理符号系统(即符号操作系统)假设和有限合理性原理。知识工程是符号主义人工智能的典型代表，知识图谱就是新一代的知识工程技术。
- ◆ 联结主义，又称为仿生学派或生理学派，认为人工智能源于仿生学，神经网络，特别是对人脑模型的研究,主张模仿人类的神经元，用神经网络的连接机制实现人工智能。人工神经网络就是其典型代表性技术。
- ◆ 行为主义，又称进化主义或控制论学派，其原理为控制论及感知-动作型控制系统。波士顿动力公司的机器狗和机器人就是这个学派的代表作。

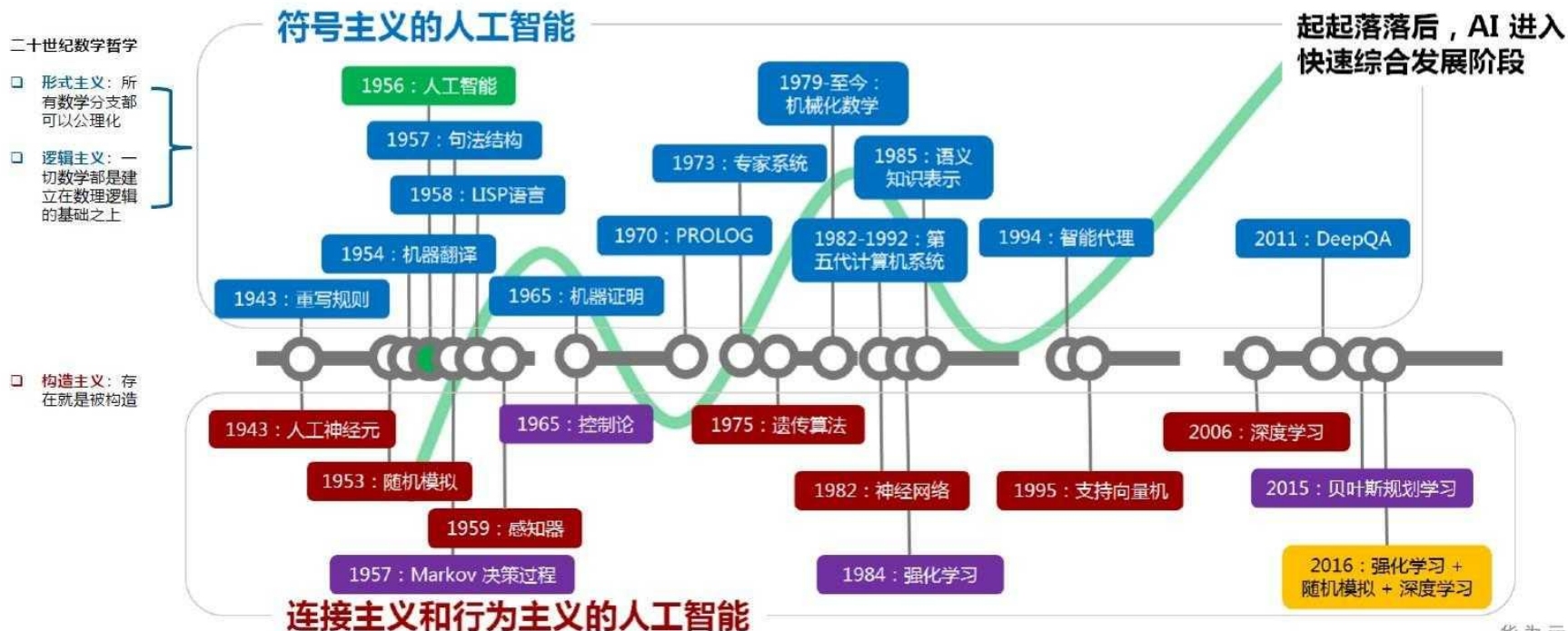


◆ 人工智能的演化形成了符号主义与联结主义人工智能两条代表性进路。符号主义人工智能的优点在于推理过程透明、可解释，但存在不完备、框架问题和知识接收瓶颈等问题。联结主义人工智能的泛化能力和可移植能力强，但缺陷是算法不可解释性、过拟合等问题。

图：符号主义与联结主义AI对比

	符号主义人工智能	联结主义人工智能
数据/知识	知识引导	数据驱动
推理/学习	知识表达、逻辑推理	机器学习、深度学习
优点	算法透明、可解释性	泛化能力、可移植性
缺点	不完备性、知识接收瓶颈、框架问题	不可解释、过拟合、语义理解问题

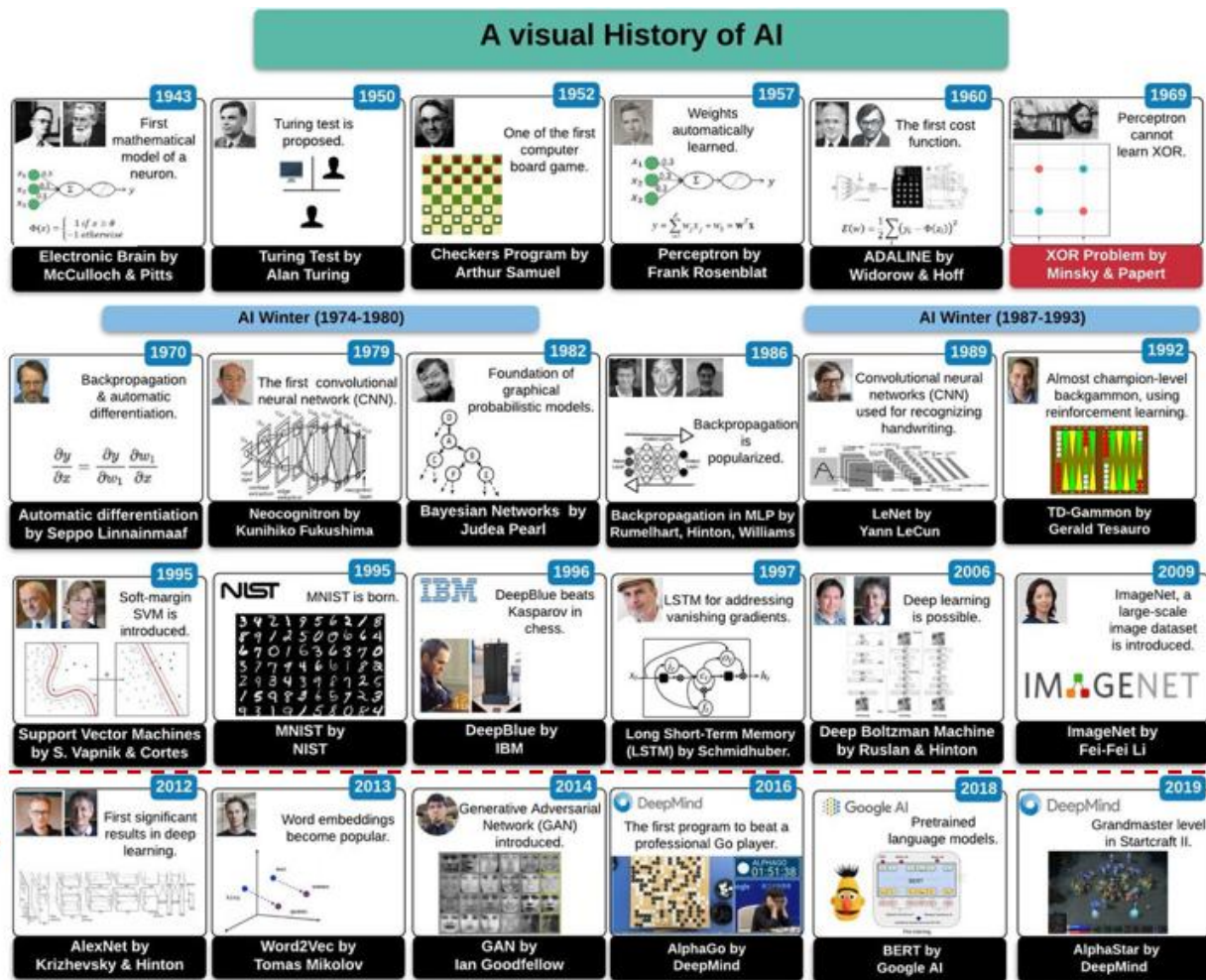
图：符号主义与联结主义各自发展



华为云社区

# 深度学习为代表的联结主义目前占主流

- ◆ 2012年，Hinton和他的学生Alex Krizhevsky设计的AlexNet神经网络模型在ImageNet竞赛大获全胜，这是史上第一次有模型在ImageNet数据集表现如此出色，并引爆了神经网络的研究热情。



成功解锁了计算机视觉乃至整体深度学习的全面进步，将深度学习推向主流！

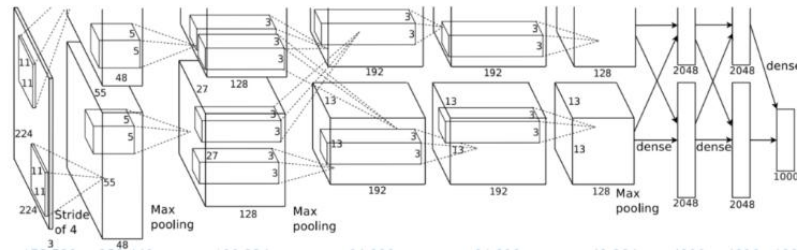
符号主义占主流，专家系统风靡一时。在2000年代初，由于专家系统的项目都需要编码太多的显式规则，这降低了效率并增加了成本，人工智能研究的重心从基于知识系统转向了机器学习方向

联结主义占主流，深度学习技术主导

# 两个突破奠定了深度学习的崛起

- ◆ 深度学习的崛起得益于两个突破：一是庞大数据量（ImageNet数据集）；另一个是以GPU为代表的算力的快速发展。
- ◆ 2012年的AlexNet神经网络模型是一个经典的CNN模型，采用8层的神经网络，5个卷积层和3个全连接层（3个卷积层后面加了最大池化层），包含6亿3000万个链接，6000万个参数和65万个神经元，使用CUDA加速神经网络的训练，利用了GPU强大的计算能力，在数据、算法及算力层面均有较大改进。

图： AlexNet神经网络模型层数



图：从1998年CMU的Vasc Faces 到2008年的TinyImage数据集




图： ImageNet数据集

## Hardly the First Image Dataset



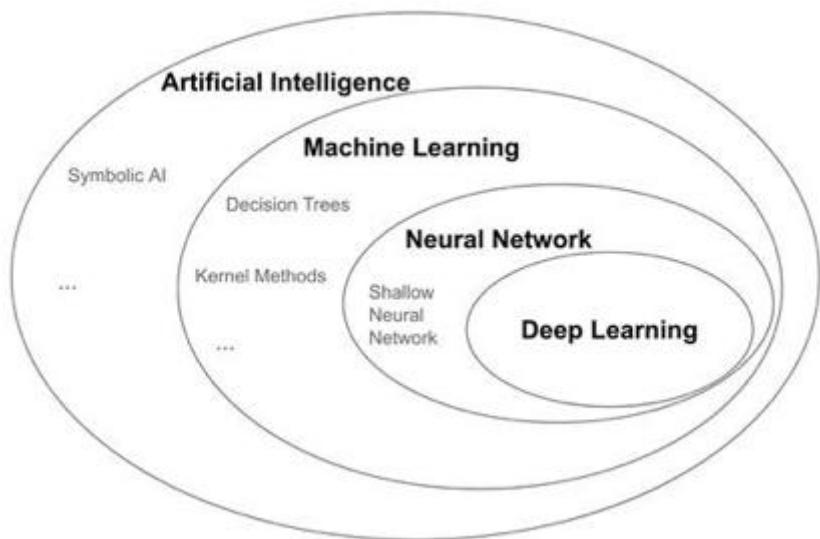
## Evaluation of ILSVRC Detection

Need to annotate the presence of all classes  
(to penalize false detections)

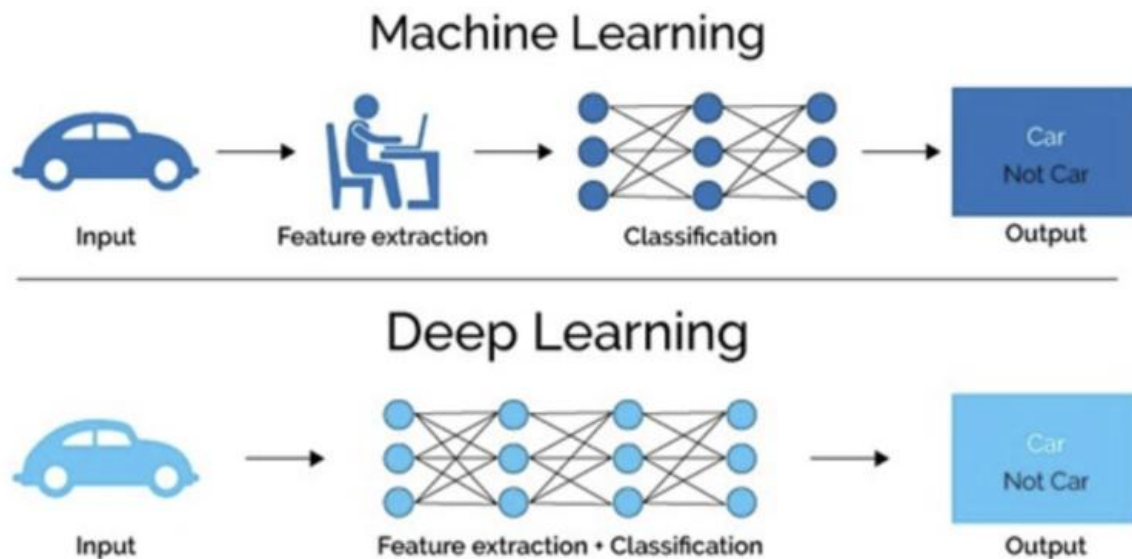
Table	Chair	Horse	Dog	Cat	Bird	# images: 400K # classes: 200 # annotations = 80M!
	+	-	-	-	-	
	+	-	-	+	-	
	+	+	-	-	-	

- ◆ 机器学习是人工智能的一个子领域。人工智能的某些领域使用了不同于机器学习的方法。神经网络是机器学习的一个分支。存在非神经网络机器学习技术，如决策树。深度学习是创造和应用“深度”神经网络的科学和艺术，即多“层”神经网络，这是相对于“浅层”神经网络，即层次少的神经网络而言的。
- ◆ 传统机器学习算术依赖人工设计特征，并进行特征提取，而深度学习方法不需要人工，而是依赖算法自动提取特征。

图：人工智能、机器学习、神经网络和深度学习之间的关系

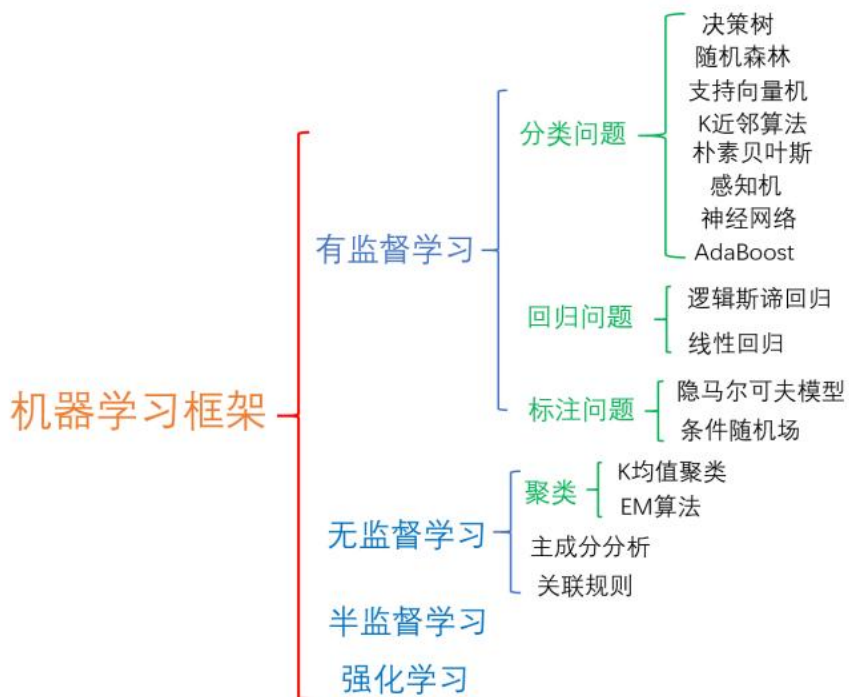


图：机器学习和深度学习对比



- ◆ 机器学习是通过一些让计算机可以自动“学习”的算法，从数据中分析获得规律，然后利用规律对新样本进行预测。从学科的角度看，机器学习是广义统计推理和决策领域的一部分，同时它也引入了古典统计学中不存在的计算机科学和优化的观点。
- ◆ 现阶段机器学习对数据有很大的依赖性，本质上是一种大数据统计分析技术。

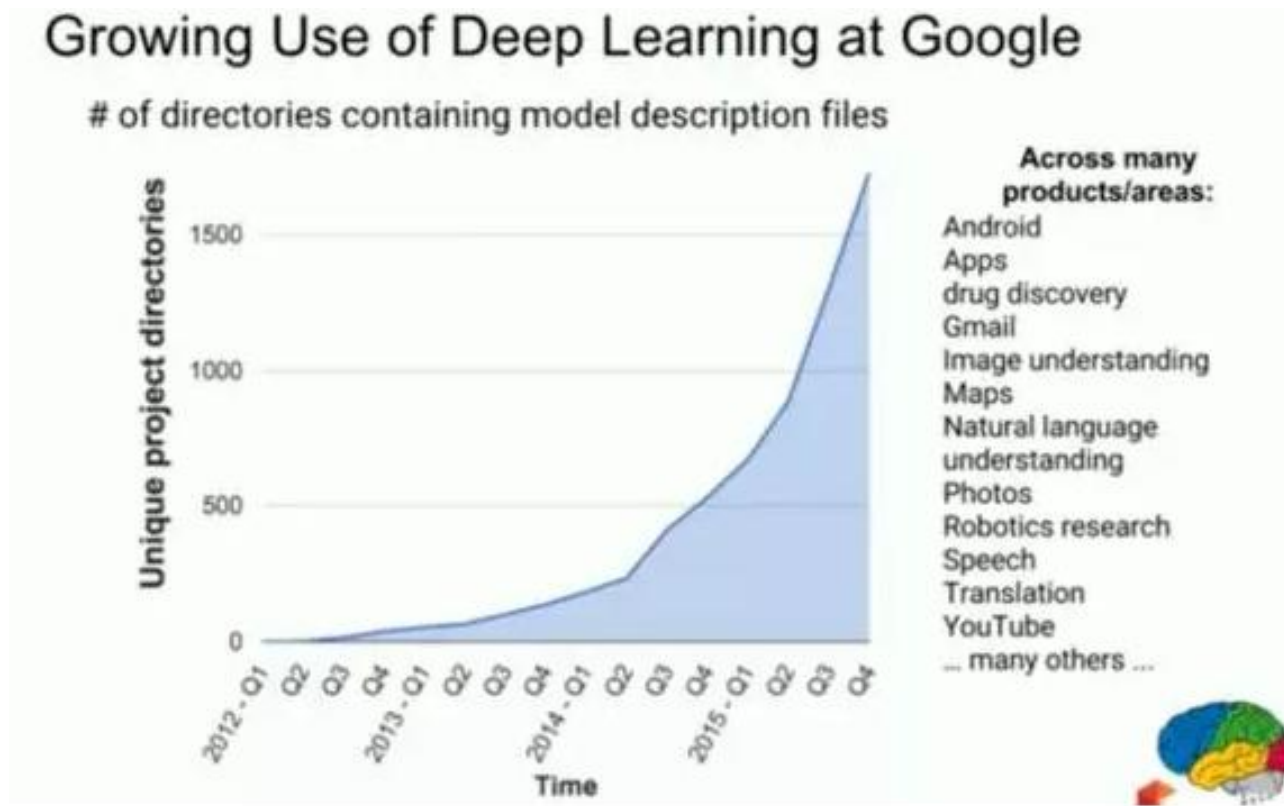
图：机器学习框架和算法



图：机器学习分类和特点

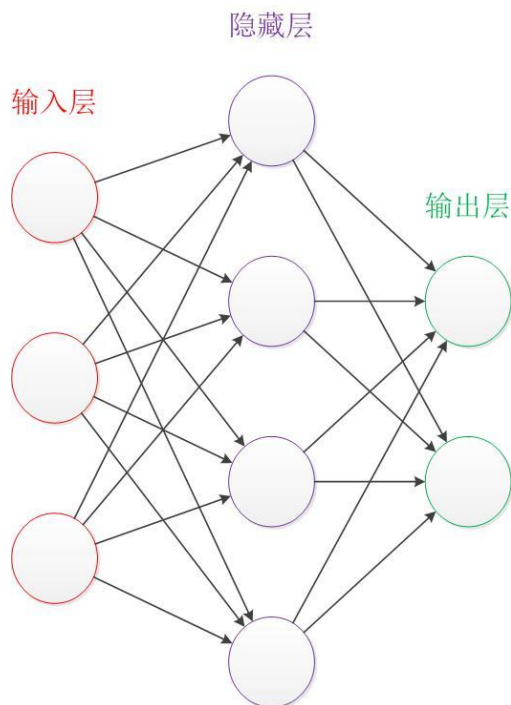
分类	特点
有监督学习	输入数据带有标签。监督学习建立一个学习过程，将预测结果与“训练数据”（即输入数据）的实际结果进行比较，不断的调整预测模型，直到模型的预测结果达到一个预期的准确率，比如分类和回归问题等。
无监督学习	输入数据没有标签，而是通过算法来推断数据的内在联系，比如聚类和关联规则学习等。
半监督学习	输入数据部分标签，是监督学习的延伸，常用于分类和回归。常用算法包括图论推理算法、拉普拉斯支持向量机等。
强化学习	输入数据作为对模型的反馈，强调如何基于环境而行动，以取得最大化的预期利益。与监督式学习之间的区别在于，它并不需要出现正确的输入/输出对，也不需要精确校正次优化的行为。强化学习更加专注于在线规划，需要在探索（在未知的领域）和遵从（现有知识）之间找到平衡。

- ◆ 神经网络又称人工神经网络(ANN)或模拟神经网络(SNN)，是机器学习的子集，同时也是深度学习算法的核心。神经网络其名称和结构均受到人脑的启发，可模仿生物神经元相互传递信号的方式。
- ◆ Google的搜索算法就是最著名的神经网络之一。ChatGPT也是一种基于神经网络的聊天AI模型。生成式预训练是一种在自然语言处理领域特别成功的技术，它涉及以无监督的方式在海量数据集上训练广泛的神经网络，以学习数据的通用表示。

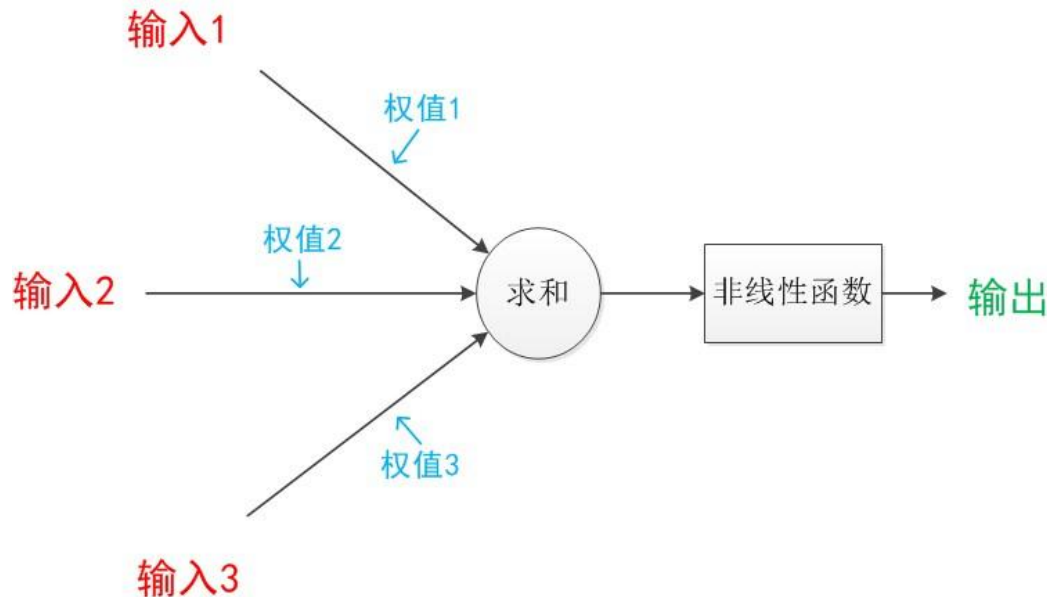


- ◆ 人工神经网络（ANN）由节点层组成，包含一个输入层、一个或多个隐藏层和一个输出层。每个节点也称为一个人工神经元，它们连接到另一个节点，具有相关的权重和阈值。如果任何单个节点的输出高于指定的阈值，那么会激活该节点，并将数据发送到网络的下一层。否则，不会将数据传递到网络的下一层。连接是神经元中最重要的东西。每一个连接上都有一个权重。
- ◆ 神经网络不直接告诉计算机如何解决问题，而让它从可观测数据中学习，让它自己找出解决问题的方法。

图：神经网络结构图



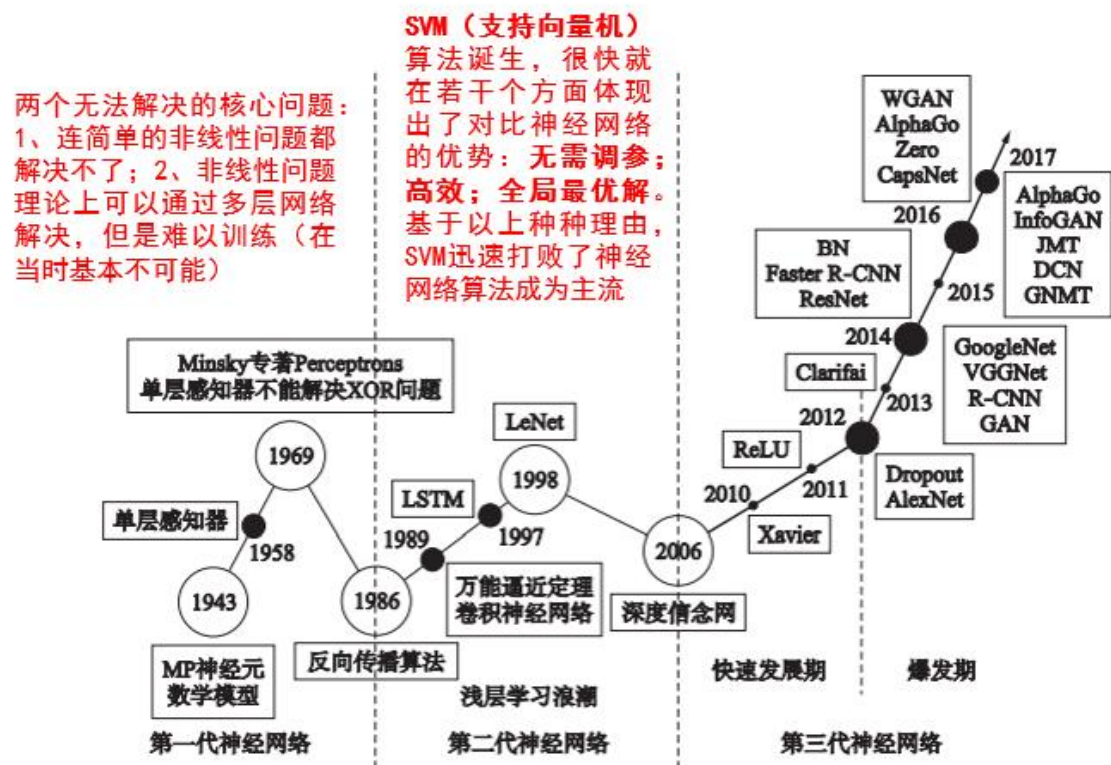
图：神经元模型



# 三代神经网络发展：由浅层学习走向深度学习

- ◆ 第一代神经网络又称为感知机：在1950年左右被提出来，算法分为输入层和输出层，输入和输出之间为线性关系，感知机无法处理非线性模型，即不能解决线性不可分的问题。
- ◆ 第二代神经网络为多层感知机（MLP）：在1980年提出，和第一代相比，在中间加了多个隐藏层，隐藏层可以引入非线性结构，能够处理非线性问题。
- ◆ 第三代神经网络为深度神经网络：2006年Hinton采取无监督预训练（Pre-Training）的方法解决了梯度消失的问题，使得深度神经网络变得可训练，自此逐步开启了第三代神经网络。

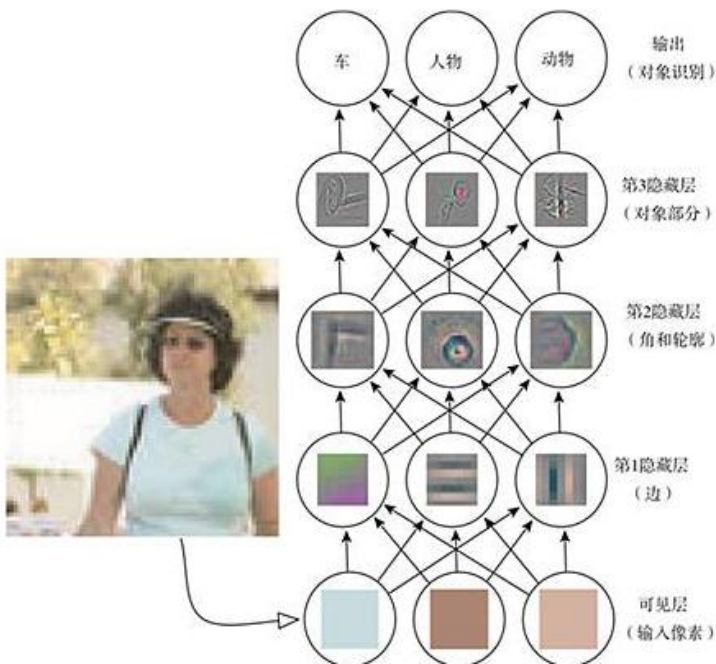
图：三代神经网络发展历程



深度学习的实质，是通过构建具有很多隐层的机器学习模型和海量的训练数据，来学习更有用的特征，从而最终提升分类或预测的准确性。因此，“深度模型”是手段，“特征学习”是目的。区别于传统的浅层学习，深度学习的不同在于：1) 强调了模型结构的深度，通常有5层、6层，甚至10多层的隐层节点；2) 明确突出了特征学习的重要性，也就是说，通过逐层特征变换，将样本在原空间的特征表示变换到一个新特征空间，从而使分类或预测更加容易。

- ◆ 深度学习是从层到层的函数的复杂组合，从而找到定义从输入到输出的映射的函数。深度学习可以通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。深度学习的实质，是通过构建具有很多隐层的机器学习模型和海量的训练数据，来学习更有用的特征，从而最终提升分类或预测的准确性。
- ◆ 深度学习的优点是：学习能力强；覆盖范围广，适应性好；数据驱动，上限高；可移植性好。
- ◆ 深度学习的缺点是：计算量大，便携性差；硬件需求高；模型设计复杂；可解释性不高，容易存在偏见。

图：深度学习模型的示意图

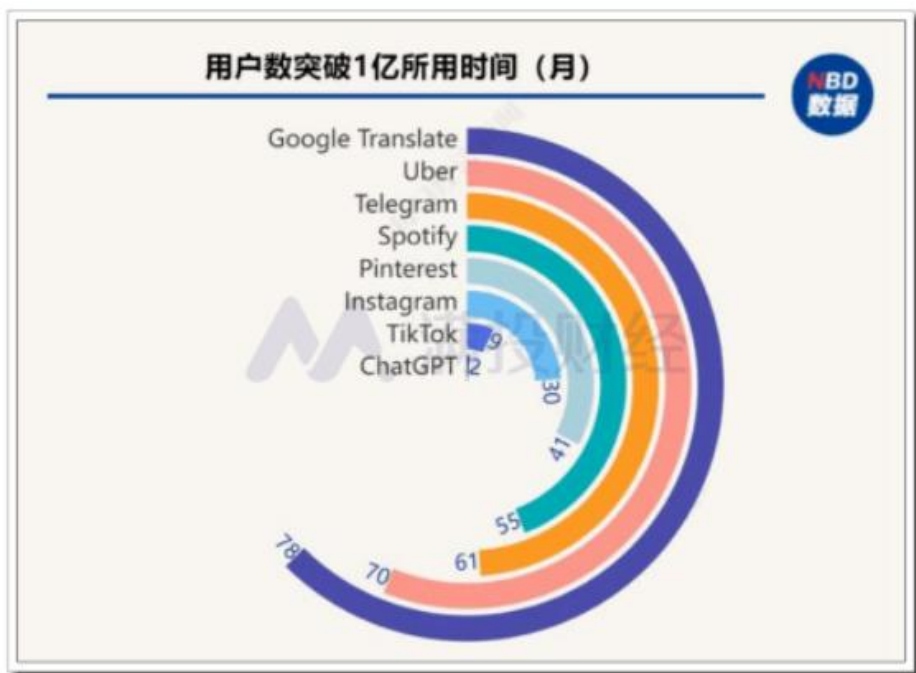


图：深度学习分类

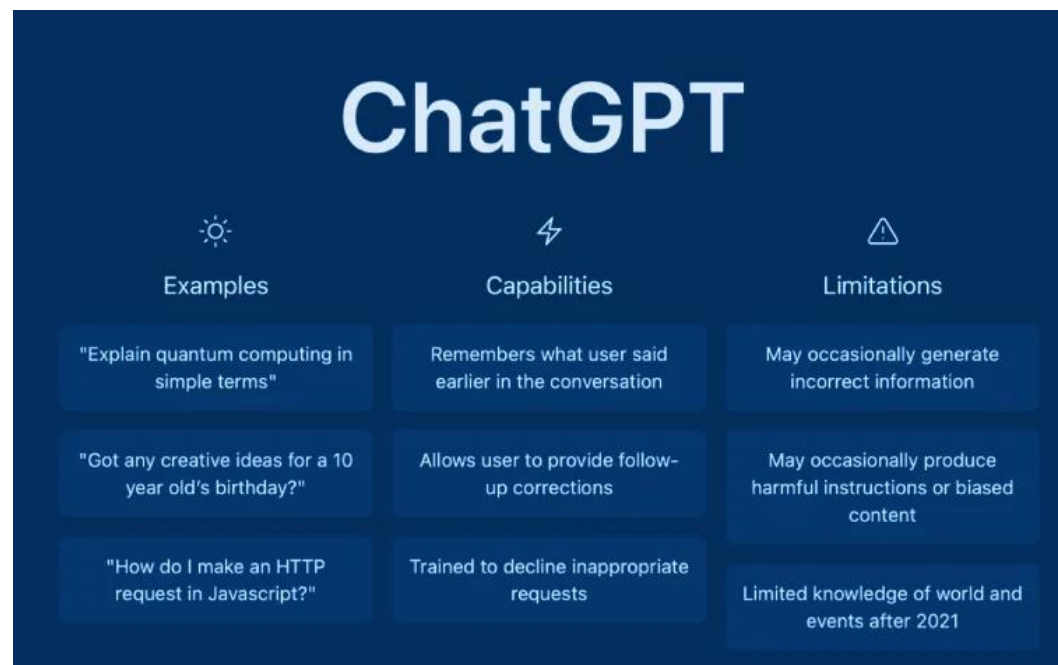


- ◆ 聊天生成型预训练变换模型（Chat Generative Pre-trained Transformer）简称ChatGPT，是OpenAI开发的人工智能聊天机器人程序，于2022年11月推出，上线两个月后已有上亿用户。
- ◆ ChatGPT是一种基于GPT（Generative pre-trained transformer）的大语言模型，是一款由人工智能驱动和自然语言处理工具实现的系统。

图：ChatGPT突破1亿用户数所需时间对比



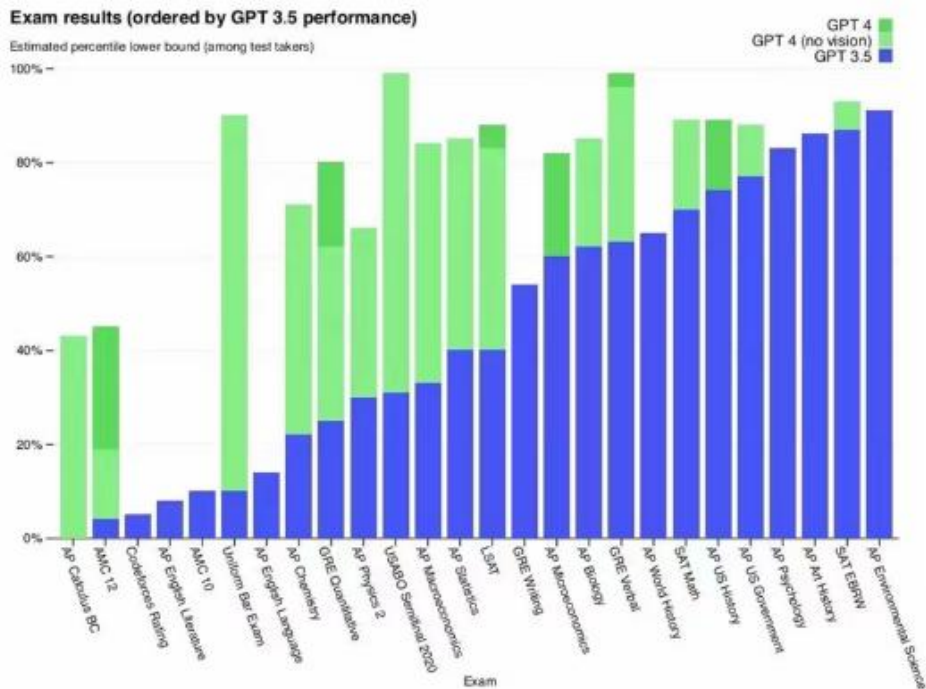
图：ChatGPT介绍



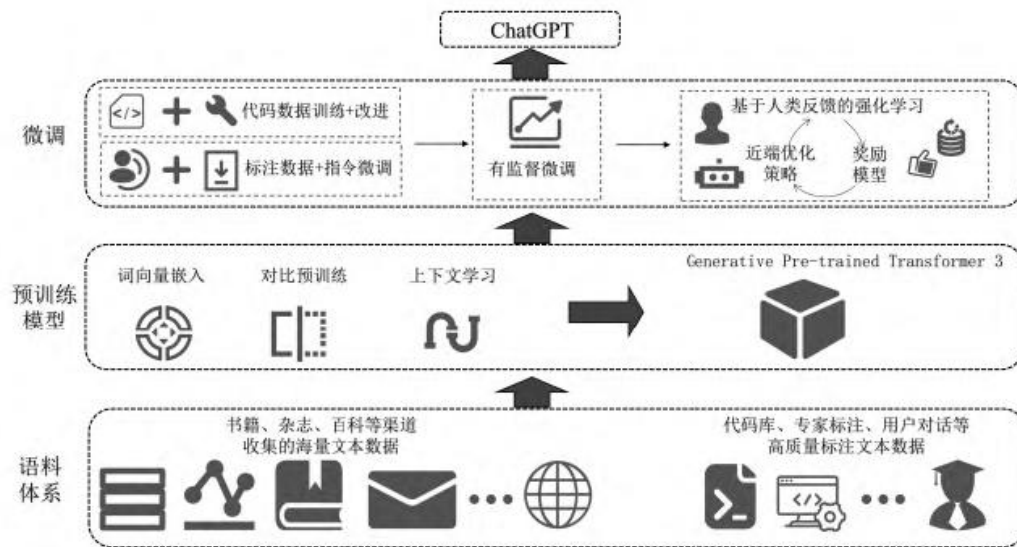
# ChatGPT是自然语言处理领域的重大技术突破

- ◆ ChatGPT是自然语言处理领域（NLP, Natural language processing）的一项重要技术突破，在多种任务中表现出卓越的性能。GPT是由OpenAI提出的采用transformer解码器的预训练模型，采用预训练加微调的范式。
- ◆ ChatGPT的主体架构遵从“基础语料+预训练+微调”的基本范式。

图：GPT4的卓越表现



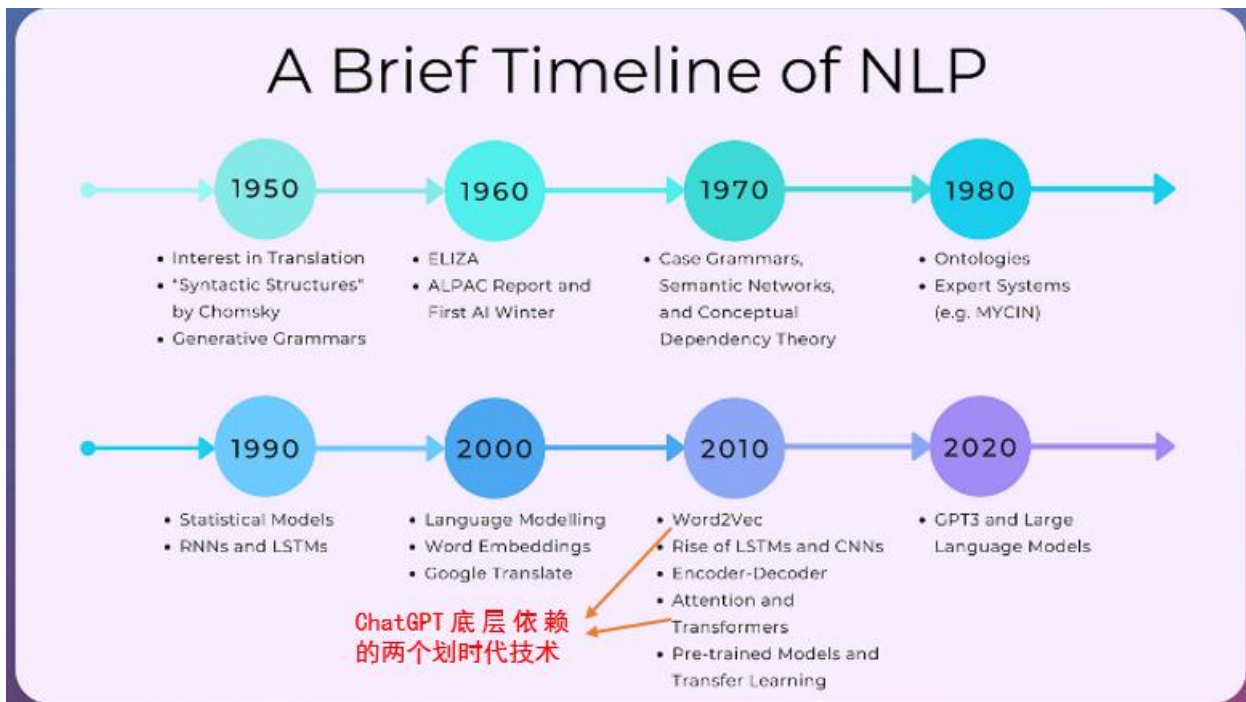
图：ChatGPT架构示意图



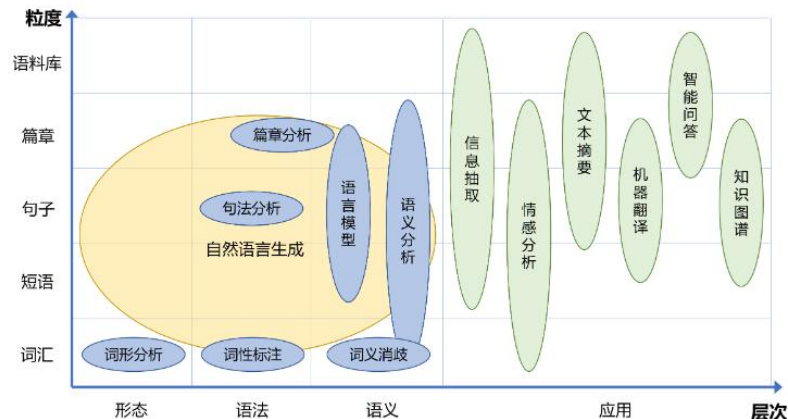
# NLP两大核心任务：NLU和NLG

◆ 自然语言主要指人们交流用的正式或非正式语言，NLP是AI领域的一大重要研究方向，目标是将人类语言转换成计算机能理解的符号或将机器语言转换成人可以理解的语言。

图：自然语言处理重大进展简史时间线



图：自然语言处理主要研究内容



◆ NLP有两大核心任务：自然语言理解（NLU）和自然语言生成（NLG）。让机器去理解语言是一件非常困难的事情，主要难点在于各种语言的鲁棒性，包括语言的多样性、歧义性、知识依赖、上下文关系等

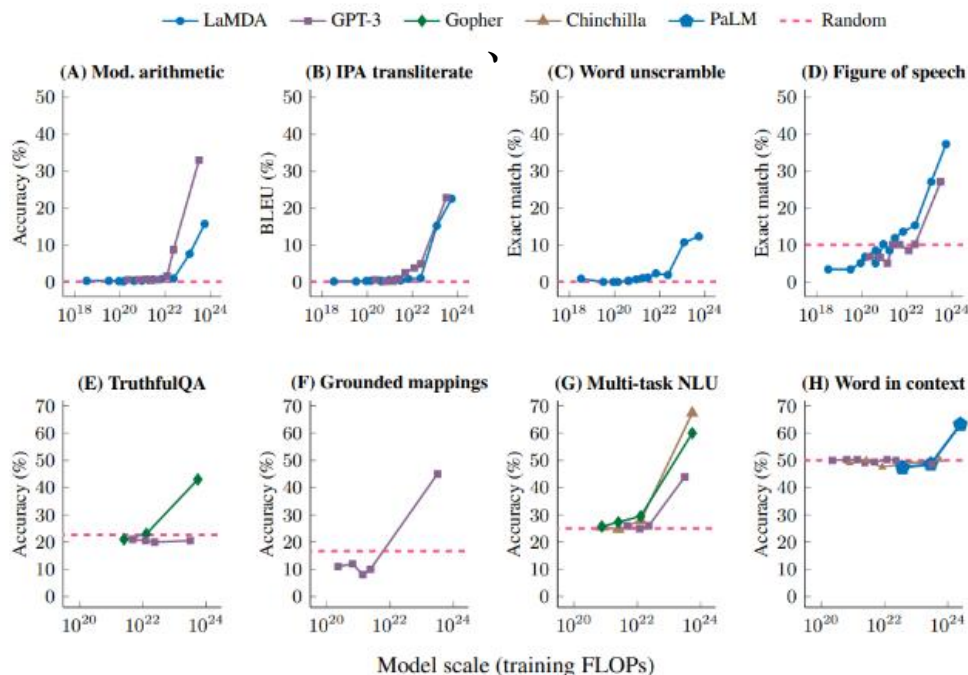
# 研究范式由规则到统计，“随机派”当下主导

- ◆ 从历史上来看，自然语言处理的研究范式变化是从规则到统计，从统计机器学习到基于神经网络的深度学习，这同时也是语言模型发展的历史。统计语言模型是当前自然语言处理处理的根本，当我们在判定一句话是否是人话的时候，可以通过计算概率的方式来判断该句子是否成立，如果一个句子成立的概率很大，那么这个句子是一个成立的句子概率就越大。
- ◆ 依据对自然语言处理的方法和侧重点不同，大致可划分为两个派别：符号派和随机派。符号派坚持对自然语言处理进行完整且全面的剖析，其过程具有较高的准确性和完整性。随机派坚持以概率统计的相关思想对自然语言处理的结果进行相关推测，并广泛应用计算假设概率的经典方法——贝叶斯方法。随机派的根本思想在于从大量数据中提取特征并进行“概率预测”，在逻辑性上远落后于“符合派”，但是当下自然语言处理领域的大规模性和真实可用性需求推动“随机派”成为主流。

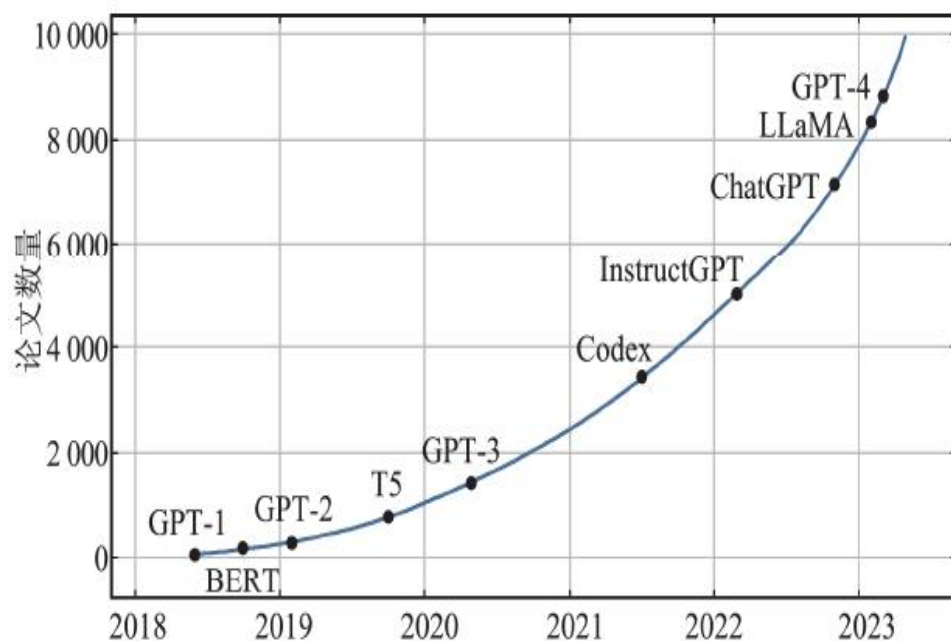
自然语言模型范式演进	主要内容	优缺点
<b>统计语言模型：</b> 计算条件概率的方法是简单的用词频做除法然后归一化不同	1975年，Frederick Jelinek等人在论文中提出并应用N-gram模型于语音识别任务，也即所谓的元文法模型。该模型基于这样一种假设，第N个词的出现只与前面N-1个词相关，而与其它任何词都不相关，整句的概率就是各个词出现概率的乘积。	优点：1) 采用极大似然估计，参数易训练；2) 完全包含了前 n-1 个词的全部信息；3) 可解释性强，直观易理解。 缺点：1) 缺乏长期依赖，只能建模到前 n-1 个词；2) 随着 n 的增大，参数空间呈指数增长；3) 数据稀疏，难免会出现OOV的问题；4) 单纯的基于统计频次，泛化能力差。
<b>神经语言模型：</b> 该模型与N-gram语言模型，该方法将所考虑的问题建模后先为其构造一个目标函数，然后对这个目标函数进行优化，从而求得一组最优的参数（即上述的参数量），最后再利用这组参数对应的模型来预测整个句子成立的概率。为了解决长信息依赖的问题。	<b>前馈神经语言模型（FFNN）：</b> 沿用了马尔可夫假设（当一个随机过程在给定现在状态及所有过去状态情况下，其未来状态的条件概率分布仅依赖于当前状态），即下一时刻的词只与过去n-1个词相关，通过将词映射到低维向量，从而解决了维度灾难问题。	优点：1) 长距离依赖，具有更强的约束性；2) 避免了数据稀疏所带来的OOV问题；3) 好的词表征能够提高模型泛化能力。 缺点：仅包含了有限的前文信息，并且无法解决长距离依赖。并且，模型训练时间长；神经网络黑盒子，可解释性较差。
	<b>循环神经语言模型（RNN）：</b> 不再基于马尔可夫假设，该模型是一类具有短期记忆能力的神经网络，在该模型中，神经元不但可以接受其他神经元的信息，也可以接受自身的信息，形成具有环路的网络结构。	<b>长短期记忆神经网络（LSTM）：</b> 是具有长期记忆能力的一种时间递归神经网络，其网络结构含有一个或多个具有可遗忘和记忆功能的单元组成。它在1997年被提出用于解决传统RNN的随时间反向传播中权重消失的问题
<b>预训练语言模型：</b> 探索一种更强的语言模型方法，即引入注意力机制，用来在长文本中刻画词语之间的相关性分配。	受到计算机视觉领域采用 ImageNet 对模型进行一次预选训练，使得模型可以通过海量图像充分学习如何提取特征，然后再根据任务目标进行模型精调的范式影响，自然语言处理领域基于预训练语言模型（PLM）的方法也逐渐成为主流。这种预训练模型是使用基于Transformer的编码器或解码器来实现的语言模型。	优点：引入注意力机制的预训练语言模型，能够比神经语言模型解决序列的长期依赖问题，并且能够捕捉出更丰富的上下文特征。 缺点：参数太多，模型复杂度高，模型太重，训练成本高以及黑盒程度加深

- ◆ 在对预训练模型的研究中，研究者很早就关注到了模型参数量对模型性能的影响。在2020年1月23日，OpenAI发表了论文《Scaling Laws for Neural Language Models》，研究了基于交叉熵损失的语言模型性能的经验尺度法则，并且发现：大模型使用样本的效率显著更高，因此最优的高效训练方式是在中等数据集上训练超大模型，并在显著收敛前提前停止。
- ◆ 当扩展大型语言模型时，偶尔会出现一些较小模型没有的新能力，这种类似于「创造力」的属性被称作「突现」能力。GPT-3的论文表明，语言模型执行多位数加法的能力对于从100M到13B参数的模型具有平坦的缩放曲线，近似随机，但会在一个节点造成性能的飞升。

图：大模型的「突现」能力



图：GPT-1发布后，与LLM相关的arXiv论文急剧增加

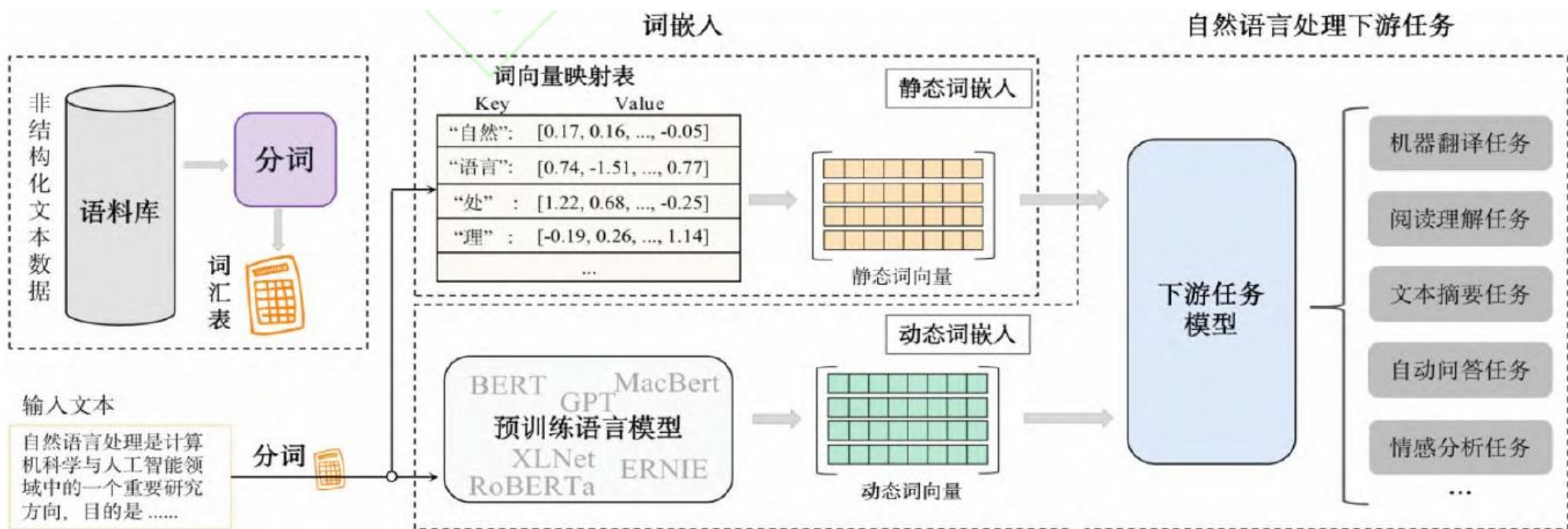


资料来源：《人工智能大模型综述及展望》罗锦钊等，华金证券研究所

# 始于word2vec，一种文本表示的方法

- ◆ 由于计算机无法读懂自然语言而只能处理数值，因此自然语言需要以一定的形式转化为数值方便计算机处理。
- ◆ 词嵌入（Word embedding）就是将自然语言中的词语映射为数值的一种方式，是一种分布式单词表示方式，也是自然语言处理任务的第一步。词嵌入能够将语料库中所蕴含的丰富内隐知识信息嵌入到低维稠密的词向量中，进而为每个字符格式的单词生成一个数值格式的词向量。Word2Vec模型是Google公司在2013年开源的一种将词语转化为向量表示的模型。

图：词嵌入方法在自然语言处理任务中的作用

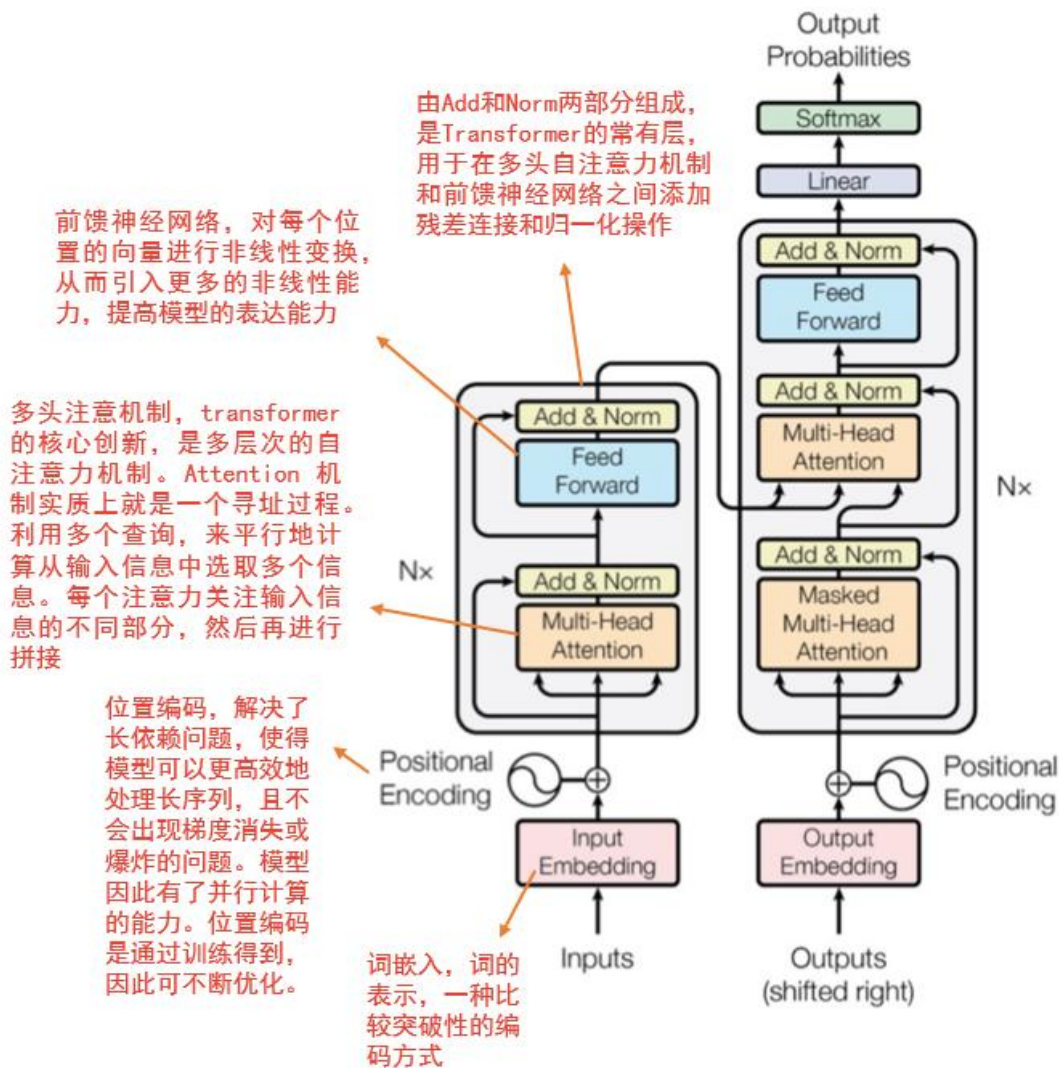


- ◆ Transformer架构是一种深度学习架构，由2017年12月6日Google发布的论文《Attention is all you need》提出。
- ◆ 注意力机制是人工神经网络中一种模仿认知注意力的技术。这种机制可以增强神经网络输入数据中某些部分的权重，同时减弱其他部分的权重，以此将网络的关注点聚焦于数据中最重要的一小部分。数据中哪些部分比其他部分更重要取决于上下文。可以通过梯度下降法对注意力机制进行训练。
- ◆ Transformer最初是作为机器翻译的序列到序列模型提出的，而后来的研究表明，基于Transformer的预训练模型（PTM）在各项任务中都有最优的表现。因此，Transformer已成为NLP领域的首选架构。

图：Transformer架模型的整体流程

步骤	描述
1. 输入嵌入层	将输入序列转换为向量表示
2. 位置编码	对输入序列中的每个元素添加位置信息
3. 多头自注意力机制	学习输入序列中元素之间的依赖关系
4. 前馈神经网络	对每个位置的向量进行非线性变换
5. 残差连接和层归一化	提高梯度流动和模型训练效果
6. 编码器层堆叠	多层编码器用于提取更高级的语义表示
7. 解码器层堆叠	多层编码器用于生成输出序列

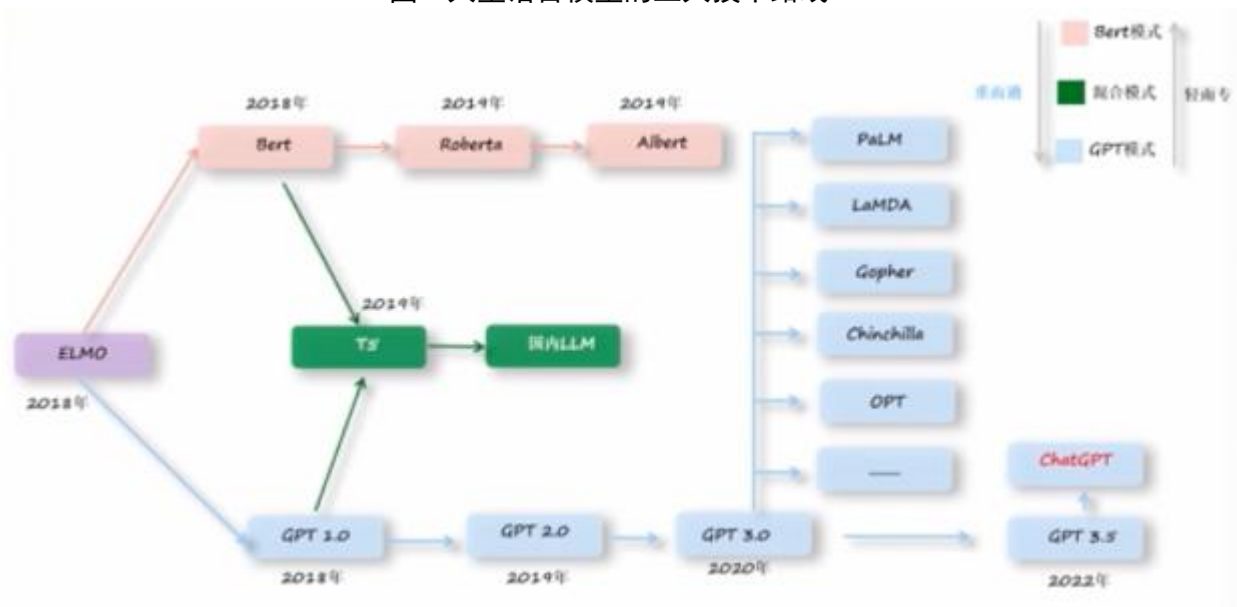
图：Transformer架构示意图



# 大型语言模型三大路线： Bert、 GPT、 混合模式

◆ 从大预言模型出现开始，技术上大致有三大技术路线： Bert、 GPT、 混合模式。BERT全称为Bidirectional Encoder Representations from Transformers，是由 Google AI Language团队在2018年提出的预训练语言模型。GPT是2018年OpenAI公司发布典型的生成式预训练模型，发布时间晚于BERT。Bert和GPT虽然都是基于Transformer的语言模型，但是其结构完全不一样。混合模式是将前两者的方法结合，代表性模型是2019年 Facebook提出了BART模型和Google同年提出的T5模型，国内早期大多采用混合模式。

图：大型语言模型的三大技术路线



图：大型语言模型三大技术路线对比

LLM技术发展路线： Bert模式（Encoder-Only）

Bert模式：两阶段（双向语言模型预训练+任务Fine-tuning）

适用场景：比较适合自然语言理解类，某个场景的具体任务，专而精；



LLM技术发展路线： GPT模式（Decoder-Only）

GPT模式：由两阶段到一阶段（单向语言模型预训练+zero shot/few shot prompt / Instruct）

适用场景：比较适合自然语言生成类任务，目前规模最大的LLM，都是这种模式； GPT 系列， PaLM, LaMDA.....,重而通；生成类任务 / 通用模型 建议GPT模式；



LLM技术发展路线： Google T5模式（Encoder-Decoder）

T5模式：形似GPT，神似Bert

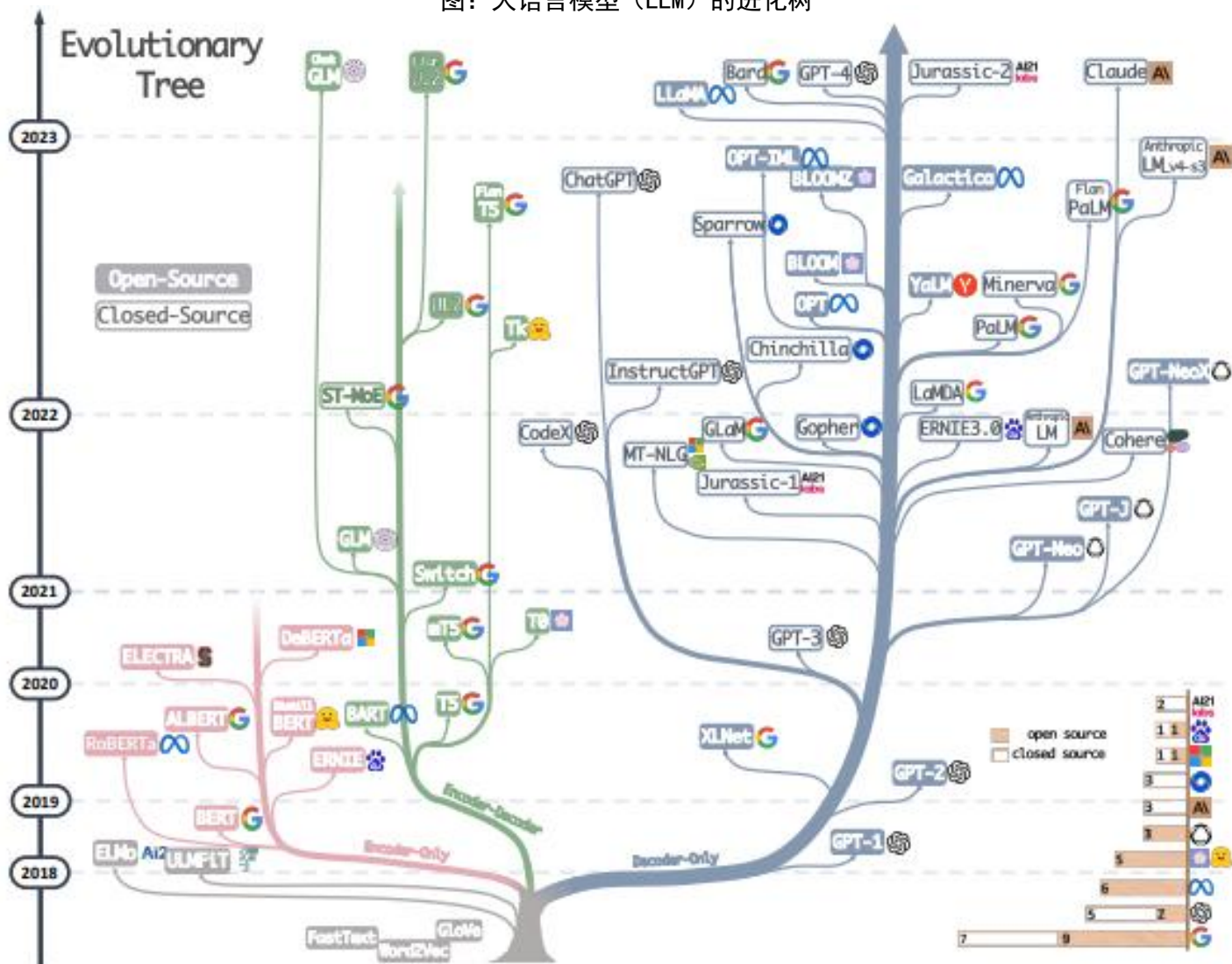
适用场景：生成和理解都行，从效果上看比较适合自然语言理解类任务，国内很多大型LLM采取这种模式；如果是单一领域的自然语言理解类任务，建议使用T5模式；



# “GPT派”一枝独秀

- ◆ 从JINGFENG YANG等学者2023年发布的《*Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*》中的关于大语言模型进化树的信息显示，以GPT为代表的Decoder-only模型目前是主流技术路径，非Transformer架构的大语言模型最近几年基本没有更新的模型出现，以BERT为代表的Encoder-only模型技术路线在2021年后也没有更新的模型出现。
- ◆ 此外，GPT为代表的技术路线中闭源模型的占比较高。

图：大语言模型（LLM）的进化树



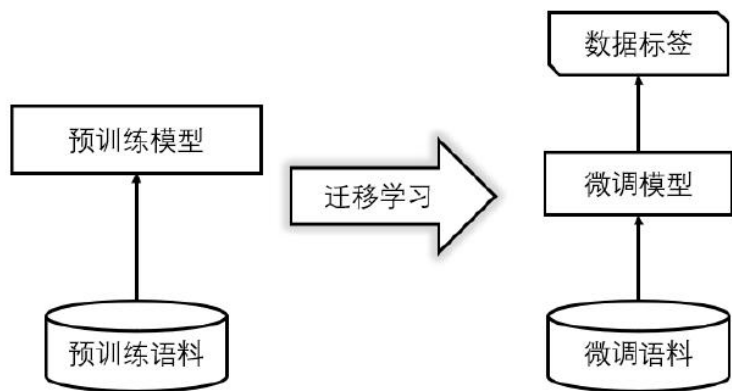
注：基于Transformer的模型显示为非灰色颜色；Decoder-only模型显示为蓝色分支，Encoder-only模型显示为粉红色分支，而Encoder-decoder模型显示为绿色分支。时间轴上模型的垂直位置代表其发布日期。开源模型由实心方块表示，而闭源模型由空心方块表示。右下角的堆叠条形图显示来自各公司和机构的模型数量。

# “预训练+微调” 范式

- ◆ 在发现Word2Vec通过神经网络学习词向量的优势后，研究者们开始考虑能否应用神经网络模型学习更全面的语言表示，于是使用语言模型进行预训练的思路被提出。
- ◆ 预训练（pre-training）是指在一个较小的、特定任务的数据集上进行微调（fine-tuning）之前，在一个大数据集上训练一个模型的过程。这个初始训练阶段允许模型从数据中学习一般的特征和表征，然后可针对具体任务进行微调。
- ◆ 预训练模型极大推进了AI的落地，让AI模型的开发从手工作坊模式走向工厂模式。当下，“预训练+微调”也成为了AI模型开发领域的标准范式。

图：基础模型的“预训练-微调”范式与传统深度学习的对比

图：“预训练+微调”流程图



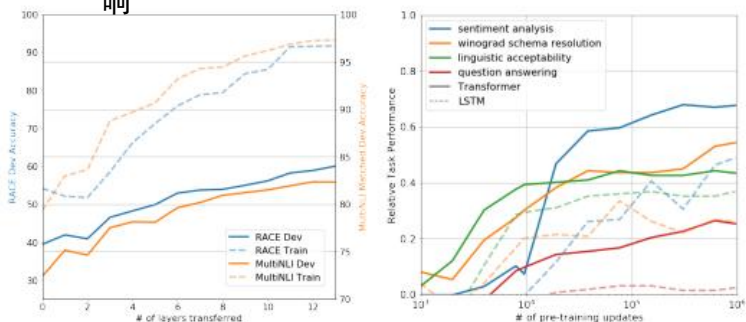
(a) 传统深度学习范式

(b) 基础模型的“预训练-微调”范式

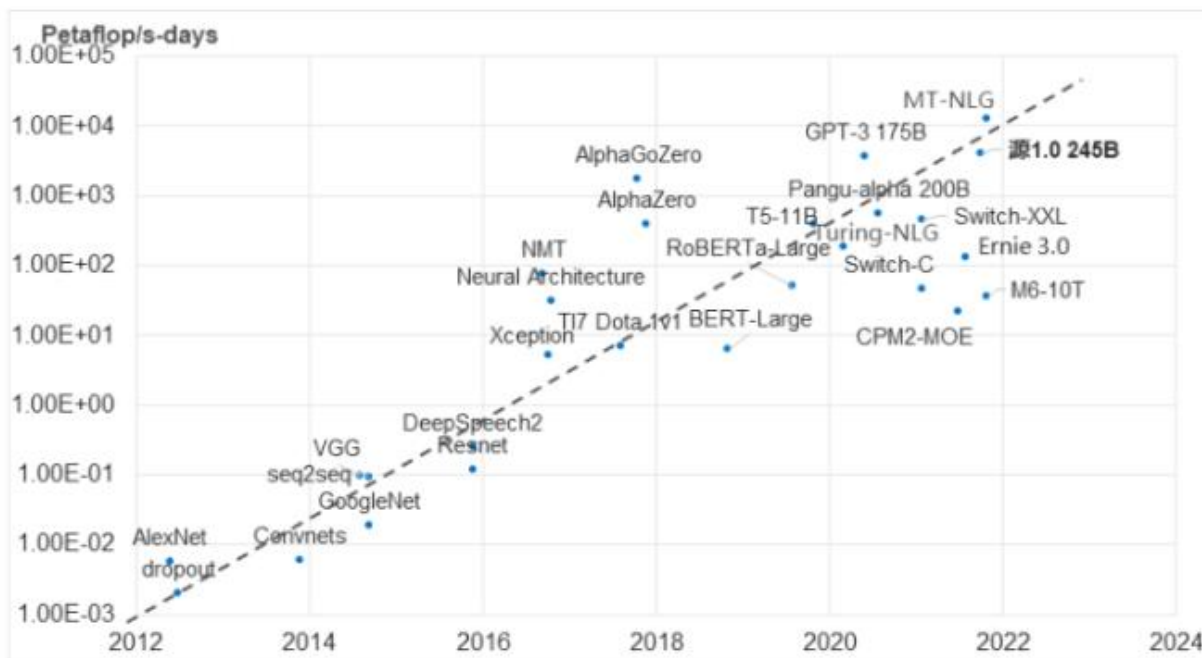
# 不断扩大的语言模型的规模

- ◆ 2018年，OpenAI发布了论文《Improving Language Understanding by Generative Pre-Training》，即使用通用的预训练模型来提升语言理解能力。论文中，GPT的实验证明了模型的精度和泛化能力会随着解码器层数增加而不断提升，而且目前还有提升空间。另外，预训练模型具有zero-shot的能力，并且能随着预训练的进行不断增强。
- ◆ 2019年2月，OpenAI官宣GPT-2模型，发布了1.24亿参数版本的GPT-2预训练结果，其公布的论文《Language Models are Unsupervised Multitask Learners》里提到，当模型的容量非常大且数据量足够丰富时，仅仅靠语言模型的学习便可以完成其他有监督学习的任务，不需要在下游任务微调。

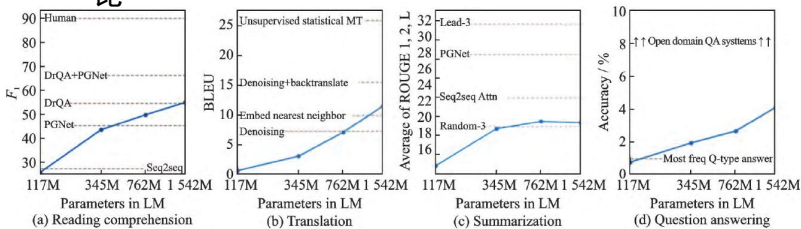
图：GPT- 1 解码器层数增加带来的影响



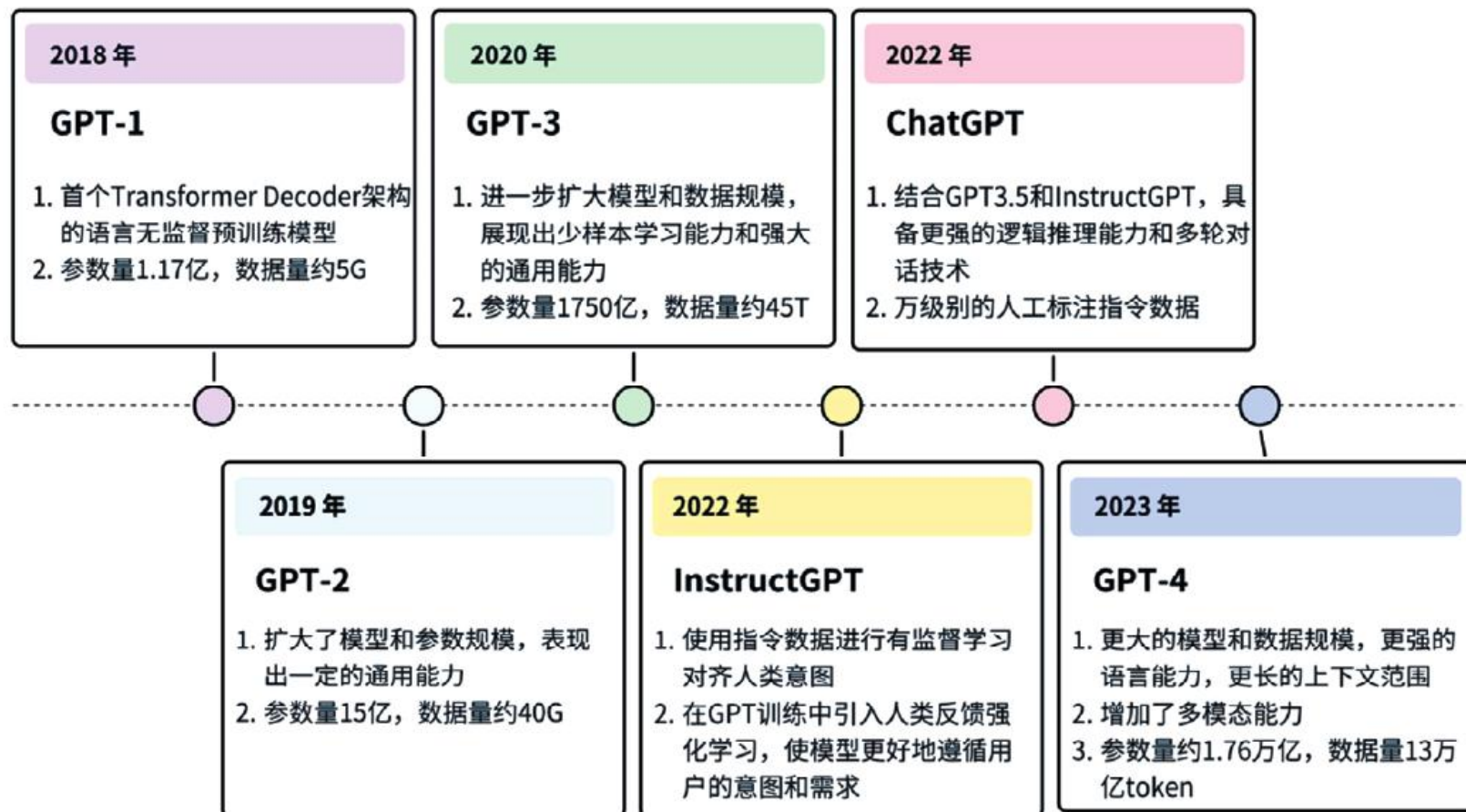
图：AI大模型参数量和计算量越来越大



图：GPT- 2 在不同任务上的不同规模模型对比



- ◆ GPT-1利用Transformer的解码器并通过大量无标注数据进行自监督预训练，然后通过有监督微调来解决不同的下游任务；GPT-2进一步扩大了模型与参数规模；GPT-3使用1750亿参数规模的模型和45TB的数据量；InstructGPT模型通过结合有监督的指令微调和人类反馈的强化学习方式，实现模型的自我优化和更新；ChatGPT基于GPT-3.5，结合InstructGPT的训练方式，加入更丰富的数据类型进行训练；GPT-4步入多模态大模型时代。



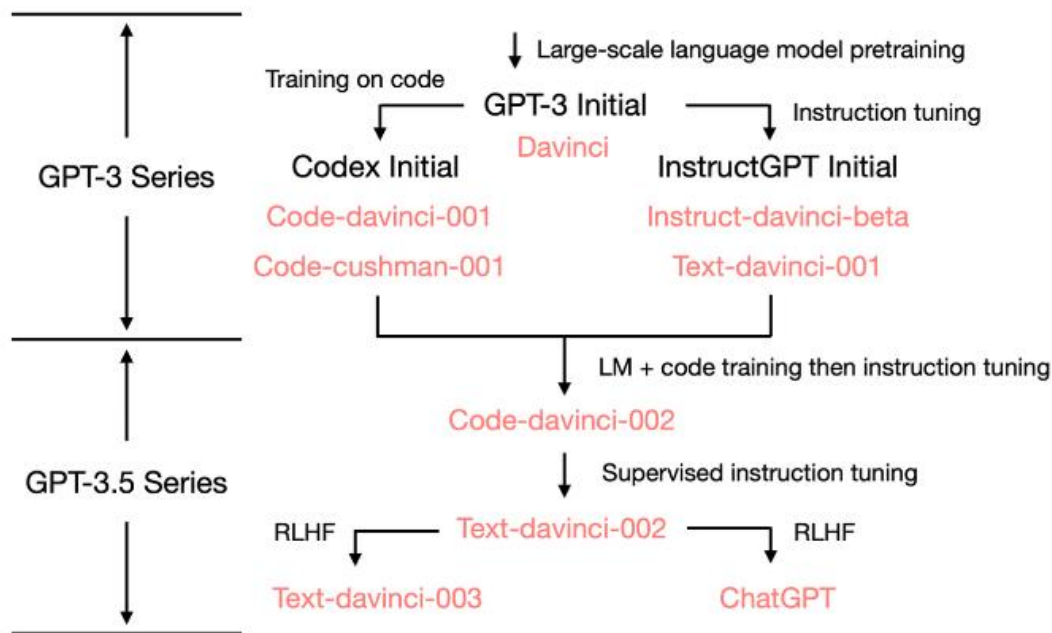
# 从GPT-3模型到ChatGPT, OpenAI用了两年

- ◆ GPT-3相比GPT-2最大的一个变化是大幅度提高了模型的参数量, OpenAI通过自监督训练获得基础大模型GPT-3实现了流畅性、知识性。在GPT-3的模型基础上, OpenAI通过引入代码训练和指令微调对模型进行增强。
- ◆ GPT-3解决了知识存储的问题, ChatGPT解决了知识如何调用的问题。

图: GPT-3和GPT-3.5的能力对比

能力	OpenAI 模型	训练方法	OpenAI API	OpenAI 论文	近似的开源模型
GPT-3系列					
语言生成 + 世界知识 + 上下文学习	GPT-3初始版本 **大部分的能力已经存在于模型中, 尽管表面上看起来很弱。	语言建模	Davinci	GPT-3论文	Meta OPT
+ 遵循人类的指令 + 泛化到没有见过的任务	Instruct-GPT初始版本	指令微调	Davinci-Instruct-Beta	Instruct-GPT论文	T0论文 Google FLAN论文
+ 代码理解 + 代码生成	Codex初始版本	在代码上进行训练	Code-Cushman-001	Codex论文	Salesforce CodeGen
GPT-3.5系列					
++ 代码理解 ++ 代码生成 ++ 复杂推理/思维链 + 长距离的依赖 (很可能)	现在的Codex **GPT3.5系列中最强大的模型	在代码+文本上进行训练 在指令上进行微调	Code-Davinci-002 (目前免费的版本 = 2022年12月)	Codex 论文	
++ 遵循人类指令 - 上下文学习 - 推理能力 ++ 零样本生成	有监督的Instruct-GPT ** 通过牺牲上下文学习换取零样本生成的能力	监督学习版的指令微调	Text-Davinci-002	Instruct-GPT论文, 有监督的部分	T0论文 Google FLAN论文
+ 遵循人类价值观 + 包含更多细节的生成 + 上下文学习 + 零样本生成	经过RLHF训练的Instruct-GPT ** 和002模型相比, 和人类更加对齐, 并且更少的性能损失	强化学习版的指令微调	Text-Davinci-003	Instruct-GPT论文, RLHF部分, 从人类反馈中的学习摘要。	DeepMind Sparrow 论文 A12 RL4LMs
++ 遵循人类价值观 ++ 包含更多细节的生成 ++ 拒绝知识范围外的问题 ++ 建模对话历史的能力 -- 上下文学习	ChatGPT ** 通过牺牲上下文学习的能力换取建模对话历史的能力	使用对话数据进行强化学习指令微调			DeepMind Sparrow论文 A12 RL4LMs

图: GPT-3.5的进化历程




# GPT-4, 迈入多模态时代

◆ OpenAI 在2023年3月14日公开了第4代生成式预训练变换模型（Generative Pre-trained Transformer 4），这是一种多模态模型，是对ChatGPT的多模态升级，能够将图像输入视为文本提示。

图：GPT-4的图片识别能力展示

User Can you explain this meme?

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.




GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

图：各代GPT能力演进

GPT-1 (2018)	GPT-2 (2019)	GPT-3 (2020)	ChatGPT (2022)	GPT-4 (2023)
<ul style="list-style-type: none"> <li>12层 Transformer, 每层12个注意力头</li> </ul>	<ul style="list-style-type: none"> <li>48层, embedding 1600维;</li> <li>词表扩大到 50257;</li> <li>层归一化移动;</li> <li>残差初始化权重改变</li> </ul>	<ul style="list-style-type: none"> <li>96层, 每层96个注意力头, embedding 12888维;</li> <li>上下文窗口 1024-&gt;2048;</li> <li>采用稀疏注意力模式</li> </ul>	<ul style="list-style-type: none"> <li>基于人类反馈对话数据训练;</li> <li>使用PPO强化学习策略;</li> </ul>	<ul style="list-style-type: none"> <li>能力升级: 更可靠、更有创意、处理更细微的指令;</li> <li>强大的识图能力;</li> <li>幻觉、安全问题的改善;</li> <li>训练后处理</li> <li>模型训练行为预测;</li> </ul>

图：GPT-4的图片识别能力展示

User What is unusual about this image?



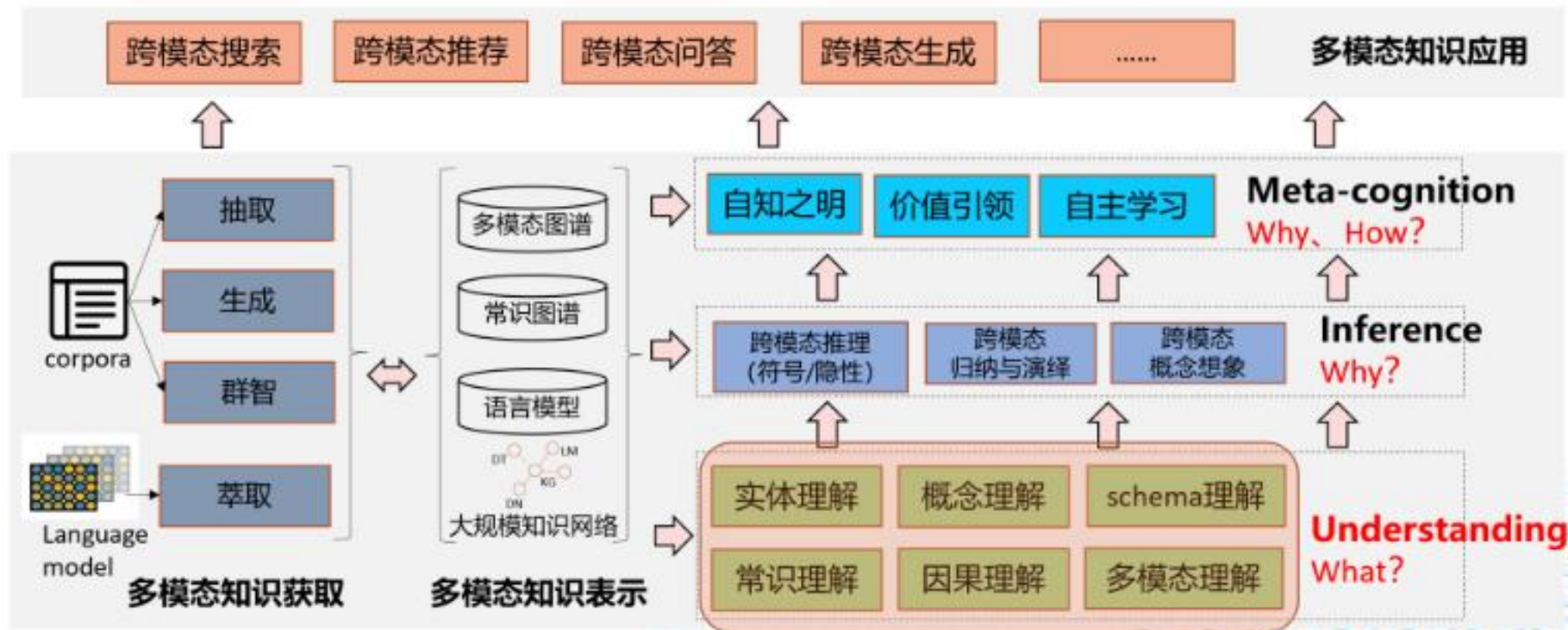
Source: Barnorama

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

# 什么是多模态模型？

- ◆ 多模态认知模型旨在处理和分析来自多个来源并且不同种类的数据，例如文本、语音、图像、视频、结构化知识源和三维数据等更复杂的数据种类。为通用认知大模型引入多种模态的信息处理能力，是通用人工智能技术发展的必然趋势。
- ◆ 常见的多模态任务大致可以分为两类：（1）多模态理解任务，包括视频分类、视觉问答、跨模态检索、指代表达等；（2）多模态生成任务，包括以文生图和视频、基于对话的图片编辑等。

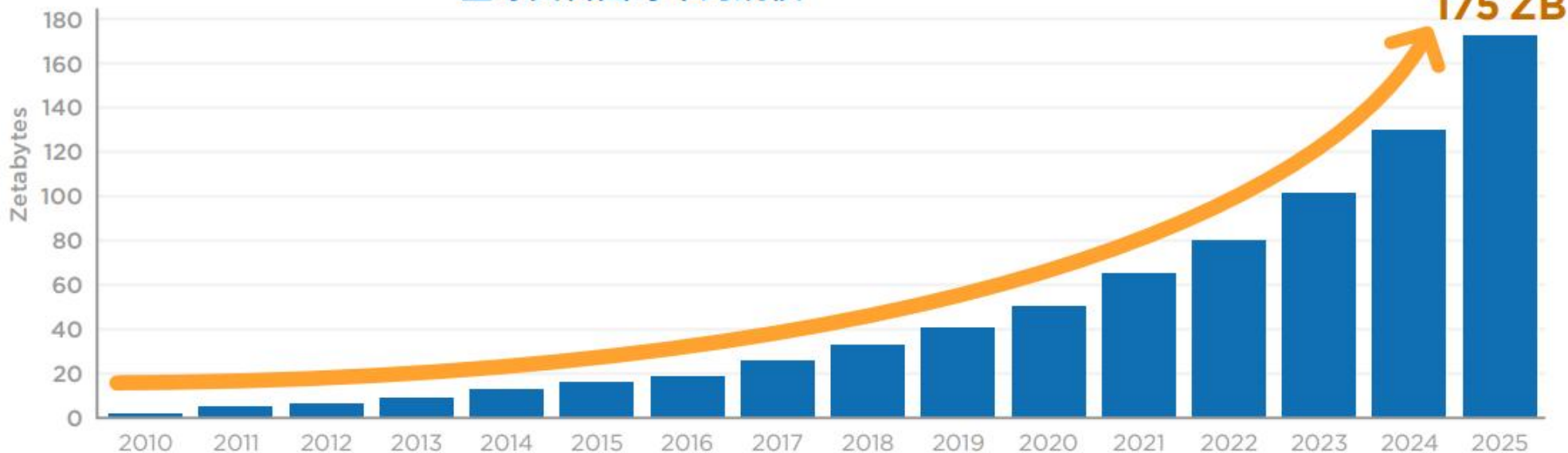
图：多模态认知智能研究框架



# 多模态，迈向通用人工智能（AGI）的重要路径

- ◆ 数据是人类知识的重要载体。据IDC预测，2018年到2025年之间，全球产生的数据量将会从33ZB增长到175ZB，复合增长率达到27%，其中超过80%的数据都会是处理难度较大的非结构化数据。非结构化数据包括全文文本、图象、声音、影视、超媒体等。
- ◆ 通用人工智能AGI是“智能代理理解或学习人类可以完成的任何智力任务的能力”，是人工智能领域的长期目标之一。多模态之间的互补性可以增强人工智能的学习能力，为了探索实现通用人工智能（AGI）的路径，人工智能（AI）从单模态走向多模态已是大势所趋。

全球資料圈每年的規模

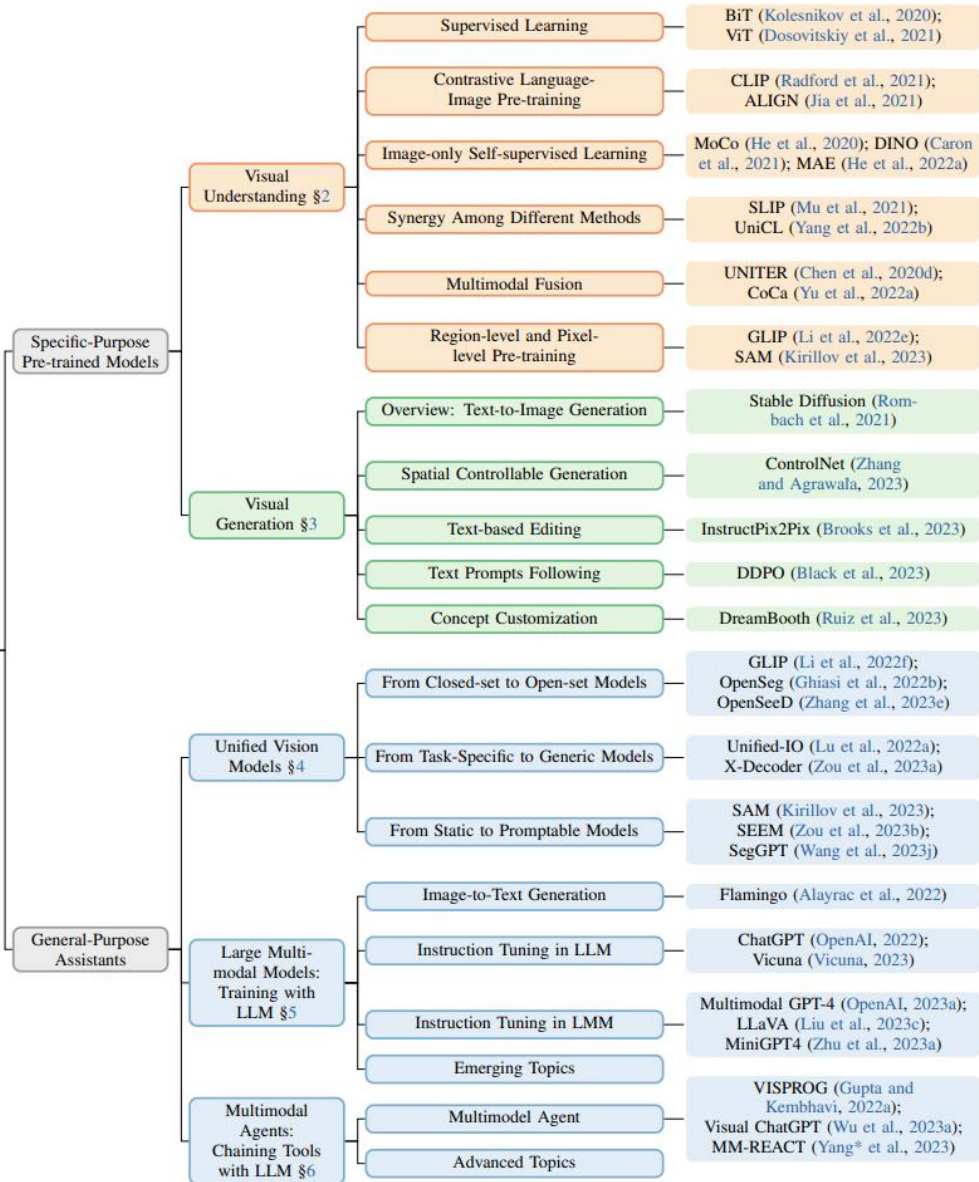
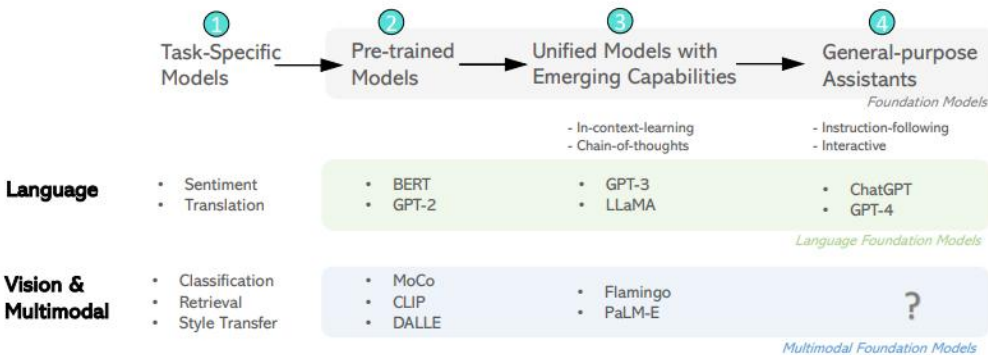


五大挑战	主要目标	实现方法	难点
多模态表示学习	指通过利用多模态之间的互补性，剔除模态间的冗余性，从而学习到更好的特征表示。	主要包括两大研究方向：联合表示（将多个模态的信息一起映射到一个统一的多模态向量空间）和协同表示（将多模态中的每个模态分别映射到各自的表示空间，但映射后的向量之间满足一定的相关性约束）	如何结合异质性的来源的数据，比如文字是符号性的，图片是一个RGB矩阵，视频是时序的RGB矩阵，声音需要采样成一个一位数组；如何处理不同级别的噪音，原因是不同模态产生的噪音是不同的
模态转化	转化也称为映射，负责将一个模态的信息转换为另一个模态的信息。	机器翻译（将输入的语言A（即时）翻译为另一种语言B）、图片描述或者视频描述（对给定的图片/视频形成一段文字描述，以表达图片/视频的内容）、语音合成（根据输入的文本信息，自动合成一段语音信号）	模态间的转换主要有两个难点，一个是open-ended，即未知结束位；另一个是subjective，即主观评判性，是指很多模态转换问题的效果没有一个比较客观的评判标准，也就是说目标函数的确定是非常主观的。
对齐	负责对来自同一个实例的不同模态信息的子分支/元素寻找对应关系。这个对应关系可以是时间维度的，又可以是空间维度的	(1) 多模态转换器。转换器促进将多模态信息转化为LLM能理解或学习的对象，利用LLM的能力学习这些转换后的对象。(2) 多模态感知器。这些方法专注于设计与LLM接口的多模态感知器，主要是为了增强对多模态信息的感知能力。(3) 工具辅助。鼓励使用工具将不同模态转换为统一的模态，主要是文本，最终完成多模态任务。(4) 数据驱动。数据驱动方法旨在通过让LLM在特定数据集上学习，赋予它与特定数据集相关的能力。	模态对齐面临着许多困难：少有显示标注模态对齐的数据集；很难设计模态间的相似性度量；存在多种可能的模态对齐，而且一个模态中的elements可能在另一个模态中没有对应。
多模态融合	负责联合多个模态的信息，进行目标预测（分类或者回归）。	按照融合的层次，可以将多模态融合分为 pixel level, feature level 和 decision level 三类，分别对应对原始数据进行融合、对抽象的特征进行融合和对决策结果进行融合。而 feature level 又可以分为 early 和 late 两个大类，代表了融合发生在特征抽取的早期和晚期。当然还有将多种融合层次混合的 hybrid 方法。	多模态融合任务中有如下挑战（1）signal 可能不是时序对齐的，例如密集连续信号vs稀疏的事件；（2）很难建立一个模型来发掘补充信息而非辅助信息；（3）每个模态在不同时间点可能展现出不同类型、不同级别的噪音。
协同学习	使用一个资源丰富的模态信息来辅助另一个资源相对贫瘠的模态进行学习，比如迁移学习	三种co-learning的方法：并行，非并行，混合；第一种方法需要一种模态的observation和另一种模态的observation直接连接，；非并行数据方法不需要两种observation的直接连接，通常利用类别间的交集；混合数据的方法通过一种共享模态或者数据连接起来。	多模态协同学习通过寻找模态之间的互补信息，使一种模态影响另一种模态的训练过程。多模态协同学习是与任务无关的，可以用于更好的多模态特征融合、转换和对齐

# 多模态大模型当下仍然处于技术发展期

◆ 微软7位华人研究员2023年发布的论文《Multimodal Foundation Models: From Specialists to General-Purpose Assistants》从目前已经完善的和还处于最前沿的两类多模态大模型研究方向出发，全面总结了五个具体研究主题：视觉理解、视觉生成、统一视觉模型、LLM加持的多模态大模型、多模态agent。其中前两个属于已经完善的主题，后三个属于还处于最前沿的主题。

图：语言和视觉/多模态的基础模型发展历程



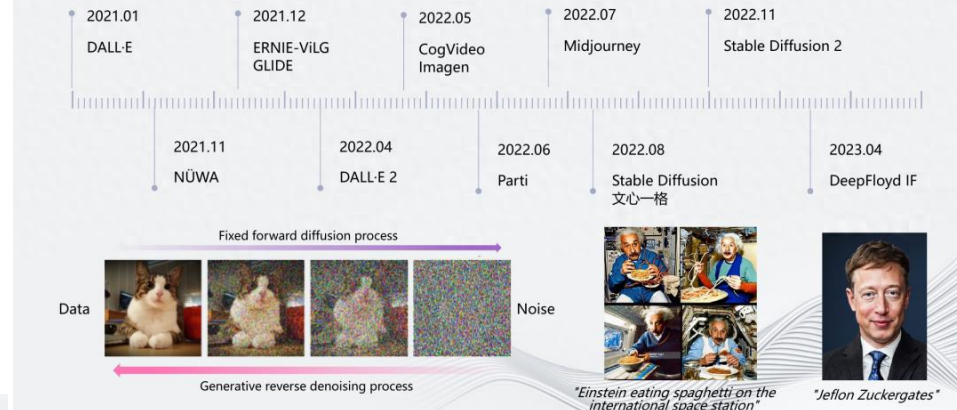
# 多模态具备更广泛的应用场景

◆ 多模态大模型由于能够处理多种类型数据（如图像、视频、文本、声音等），与传统的深度学习模型相比，可以提供模型的泛化能力与鲁棒性。这使得多模态大模型在自然语言处理、计算机视觉、语音识别等领域具有广泛的应用前景。

## AIGC大模型一览，单模态&多模态



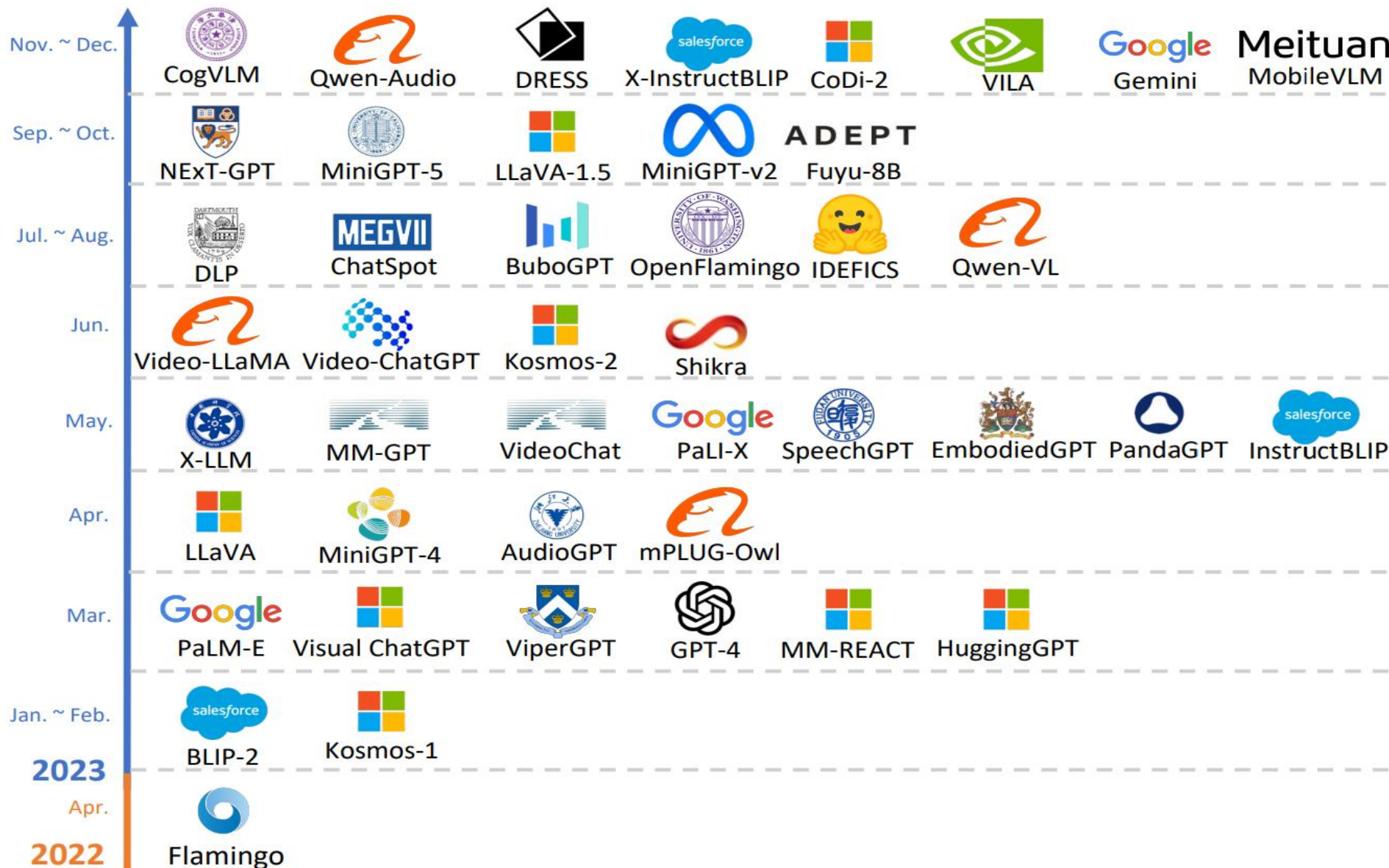
## 多模态生成之图像创作



## 多模态生成之视频创作



# 主流多模态大模型发布演进



# 中国AIGC产业链图谱



# AI框架是AI算法开发的必备工具

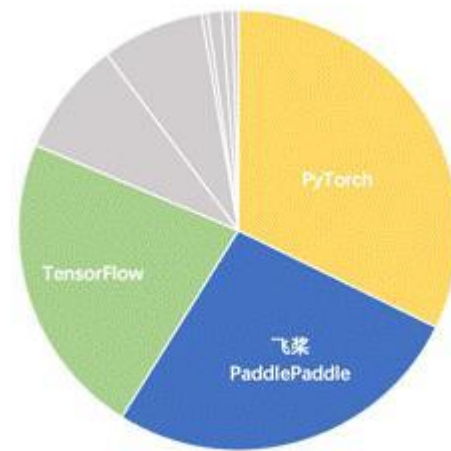
- ◆ AI框架是AI算法模型设计、训练和验证的一套标准接口、特性库和工具包，集成了算法的封装、数据的调用以及计算资源的使用，同时面向开发者提供了开发界面和高效的执行平台，是现阶段AI算法开发的必备工具。深度学习平台下接芯片、上承应用，相当于“智能时代的操作系统”。
- ◆ 根据国际权威数据调研机构IDC发布《中国深度学习框架和平台市场份额，2022H2》报告显示，百度稳居中国深度学习平台市场综合份额第一，领先优势进一步扩大。中国深度学习开源框架市场形成三强格局，飞桨PaddlePaddle超过TensorFlow，框架市场前三份额超过80%。

图：AI框架发展历程



图：中国深度学习开源框架市场份额（2022H2）

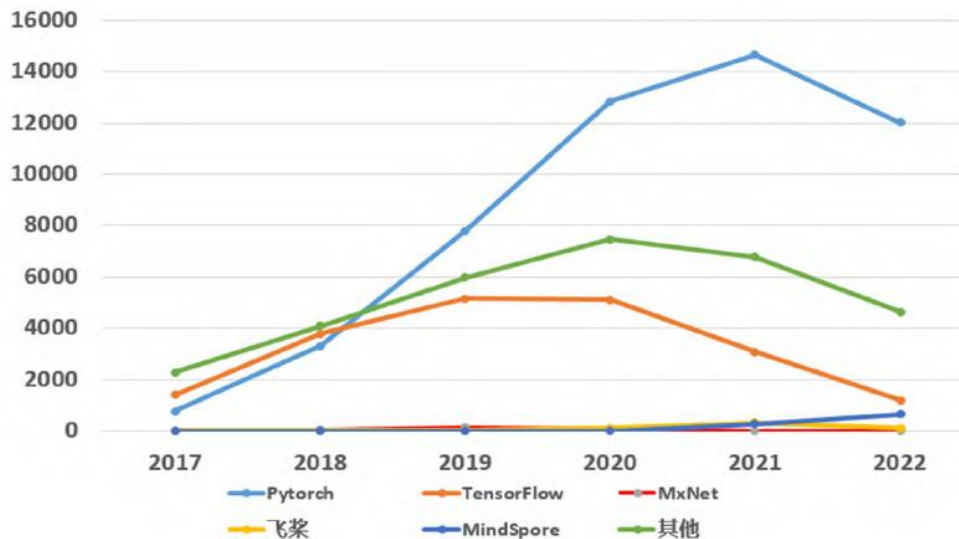
中国深度学习开源框架市场份额，2022H2



数据来源：IDC中国，《中国深度学习框架和平台市场份额，2022H2》

- ◆ 深度学习框架是一种抽象软件，当安装配置好某深度学习框架后，可通过该框架的接口语言直接调用框架及相关库文件。深度学习框架可以有效降低使用门槛，使得开发者能够有更多精力专注于产品设计。
- ◆ 从深度学习框架相关论文年度发表数量对比来看，飞桨从2019年呈现迅猛发展的势头，MindSpore从2020年开源后也是快速发展。

图：深度学习框架相关论文年度发表数量对比图



图：部分 GitHub 开源深度学习框架对比

框架名称	所属机构	贡献者数量	license	适用场景	主要开发语言	接口语言	星数 (K)
TensorFlow	Google	3045	Apache2.0	生产、科研	C++、Python	C++,Python , JavaScript, Java	161
PyTorch	Facebook	2105	自定义	生产、科研	C++、Python	C++,Python	53
CNTK	微软	201	自定义	生产	C++	C++,Python ,C#,Java	17.1
MXNet	DMLC	868	Apache2.0	生产	C++	Python, C++,R,Go,JavaScript	19.8
飞桨	百度	513	Apache2.0	生产、科研	C++、Python	C++,Python ,Go,Java,R	17.3
MindSpore	华为	262	Apache2.0	生产	C++、Python	C++,Python , Java	2.6
计图	清华	28	Apache2.0	科研	C++、Python	Python	2.3

- ◆ 飞桨（PaddlePaddle）是由百度开发的国内首个集深度学习训练和测试、模型库、工具组建等为一体的开源深度学习框架。在2016年8月，百度正式将飞桨开源在GitHub网站，框架以Python为前端，底层架构依赖C++。目前，飞桨已经部署到计算机视觉、语音、自然语言处理等领域。在计算视觉方面，飞桨提供了度量学习、目标检测、图像生成、图像分割、关键点检测、视频分类等算法模型。同时，飞桨还提供了各种优秀的工具组件。目前飞桨已广泛应用于工业、农业、服务业等，服务535万开发者，创造67万个模型，服务20万企事业单位，产学研用紧密协同培养超过200万AI人才，与合作伙伴一起帮助越来越多的行业完成AI赋能。

图：飞桨四大领先技术

## 开发便捷的深度学习框架

飞桨深度学习框架基于编程一致的深度学习计算抽象以及对应的前后端设计，拥有易学易用的前端编程界面和统一高效的内部核心架构，对普通开发者而言更容易上手并具备领先的训练性能。飞桨自然完备兼容命令式和声明式两种编程范式，默认采用命令式编程范式，并完美地实现了动静统一，开发者使飞桨可以实现动态图编程调试，一行代码转静态图训练部署。飞桨框架还提供了低代码开发的高层API，并且高层API和基础API采用了一体化设计，两者可以互相配合使用，做到高低融合，确保用户可以同时享受开发的便捷性和灵活性。

## 超大规模深度学习模型训练技术

大规模分布式训练历来是飞桨非常有特色的一个功能。飞桨突破了超大规模深度学习模型训练技术，领先其它框架实现了千亿稀疏特征、万亿参数、数百节点并行训练的能力，解决了超大规模深度学习模型的在线学习和部署难题。此外，飞桨还覆盖支持包括模型并行、流水线并行在内的广泛并行模式和加速策略，率先推出业内首个通用异构参数服务器模式和4D混合并行策略，引领大规模分布式训练技术的发展趋势。

## 多端多平台部署的高性能推理引擎

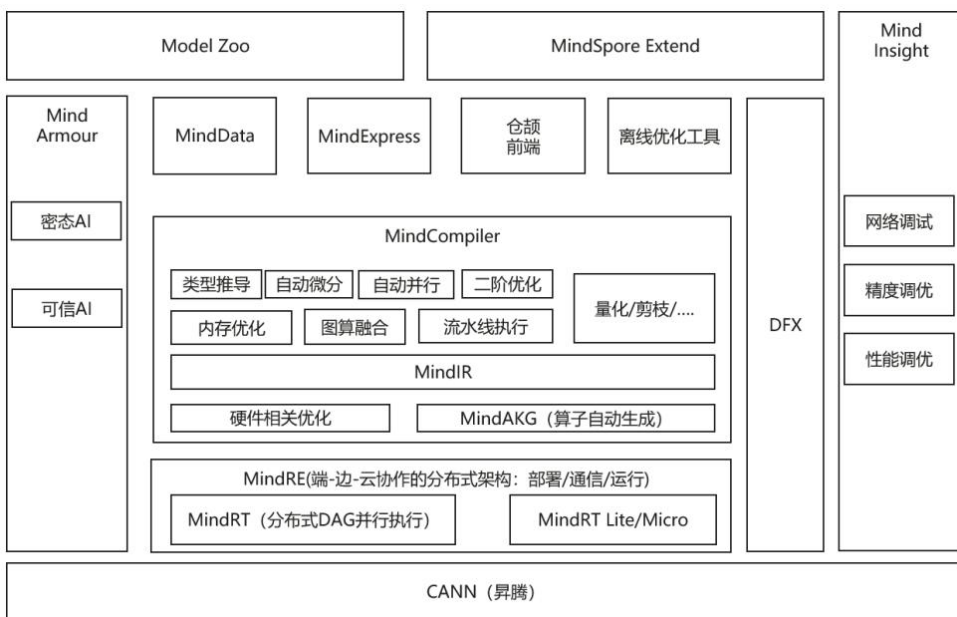
飞桨对推理部署提供全方位支持，可以将模型便捷地部署到云端服务器、移动端以及边缘端等不同平台设备上，并拥有全面领先的推理速度，同时兼容其它开源框架训练的模型。飞桨推理引擎支持广泛的AI芯片，特别是对国产硬件做到了全面的优化适配。

## 产业级开源模型库

基于飞桨框架2.0，官方建设的算法数量达到270+，并且绝大部分模型已升级为动态图模型，包含经过产业实践长期打磨的主流模型以及在国际竞赛中的夺冠模型；提供面向语义理解、图像分类、目标检测、语义分割、文字识别、语音合成等场景的多个端到端开发套件，满足企业低成本开发和快速集成的需求，助力快速的产业应用。

- ◆ MindSpore是华为开发的一款深度学习框架。框架在2020年3月首次开源在GitHub网站上，支持C++、Python等语言开发，与华为昇腾AI处理器协同开发。该框架旨在实现三个目标：易开发、高效执行和全场景覆盖。MindSpore 由几个主要组件组成：MindExpression (ME)、MindCompiler (MC)、MindData (MD)、MindRE 和 MindArmour (MA)。
- ◆ 截至2023年12月，昇思MindSpore下载量突破657W+，2023年增长了301W+；昇思大模型平台下载量突破57W+，作为首个基于自主创新的AI算力和框架、服务全球开发者的一站式大模型平台，已有28.7K+名用户注册使用，其中2.2K+核心贡献者。

图： MindSpore 架构图



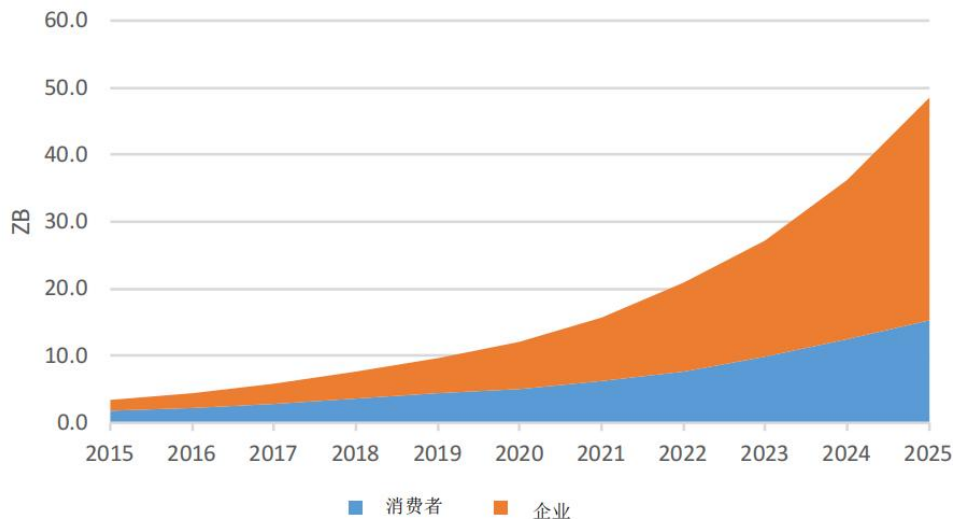
图： MindSpore 2023年年度总结



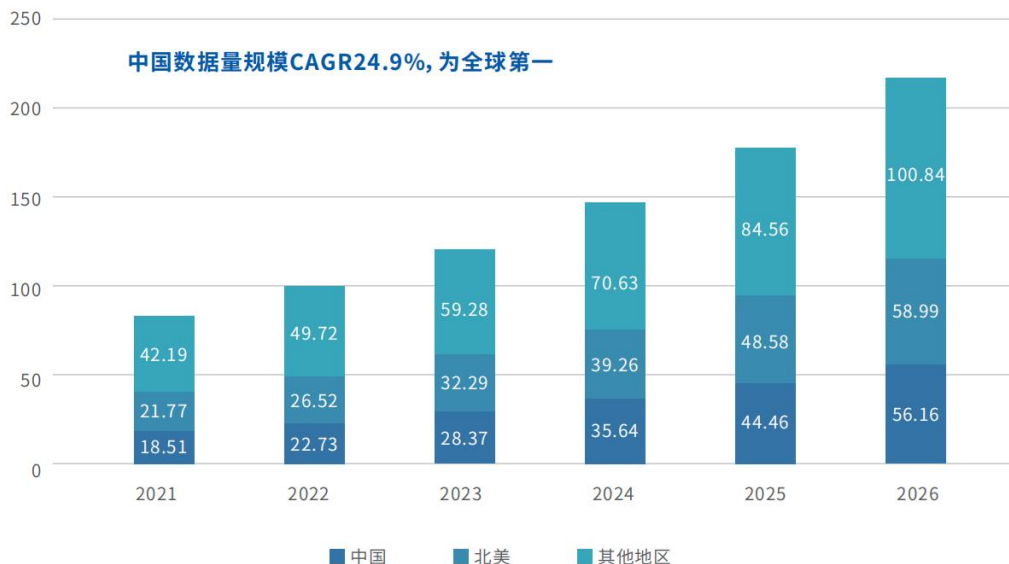
# 中国具备庞大的数据量

- ◆ 数据是AI发展的基石，据IDC统计，中国数据规模将从2021年的18.51ZB增长至2026年的56.16ZB，年均增长速度CAGR为24.9%，增速位居全球第一。
- ◆ 根据IDC的预测，中国产生的非娱乐数据占比逐步提升，生产力数据、物联网数据占比快速增长。

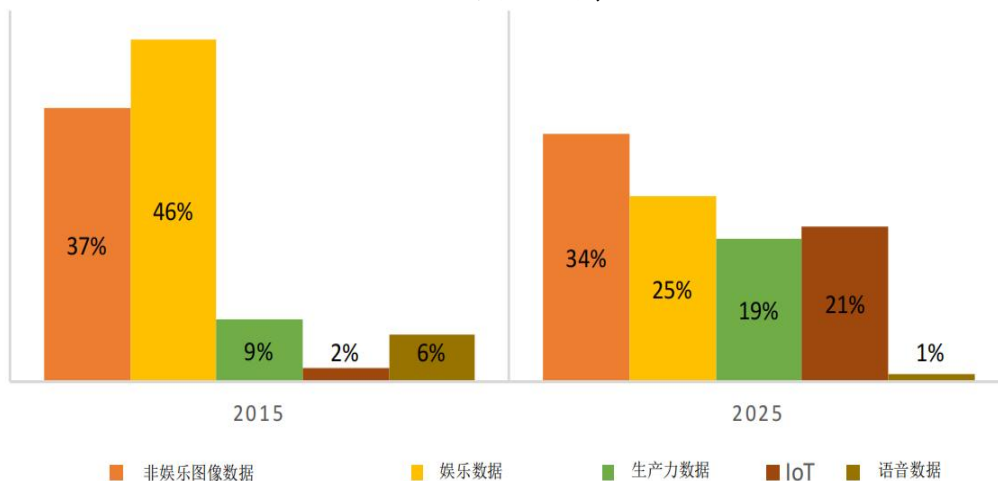
图：中国数据圈：消费者与企业级的数据圈份额，2015—2025



图：全球不同区域数据量规模预测

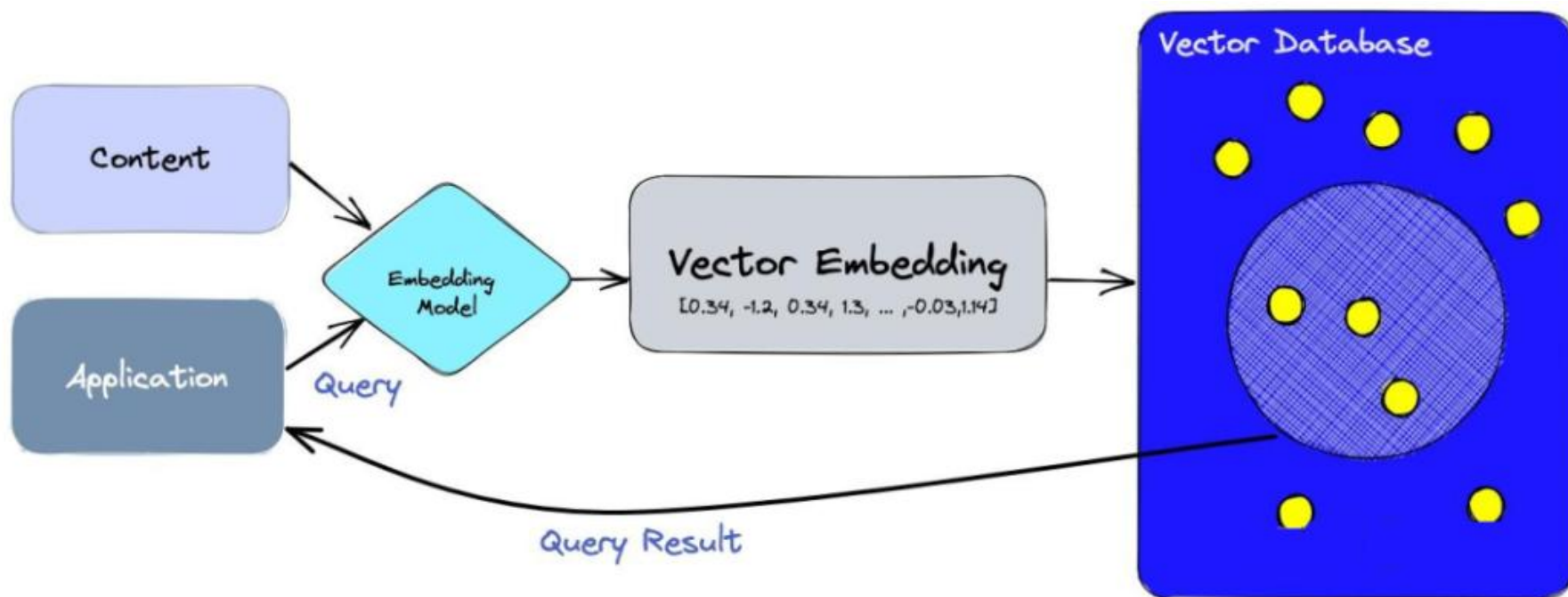


图：中国数据圈：数据类型份额，2015年和2025年



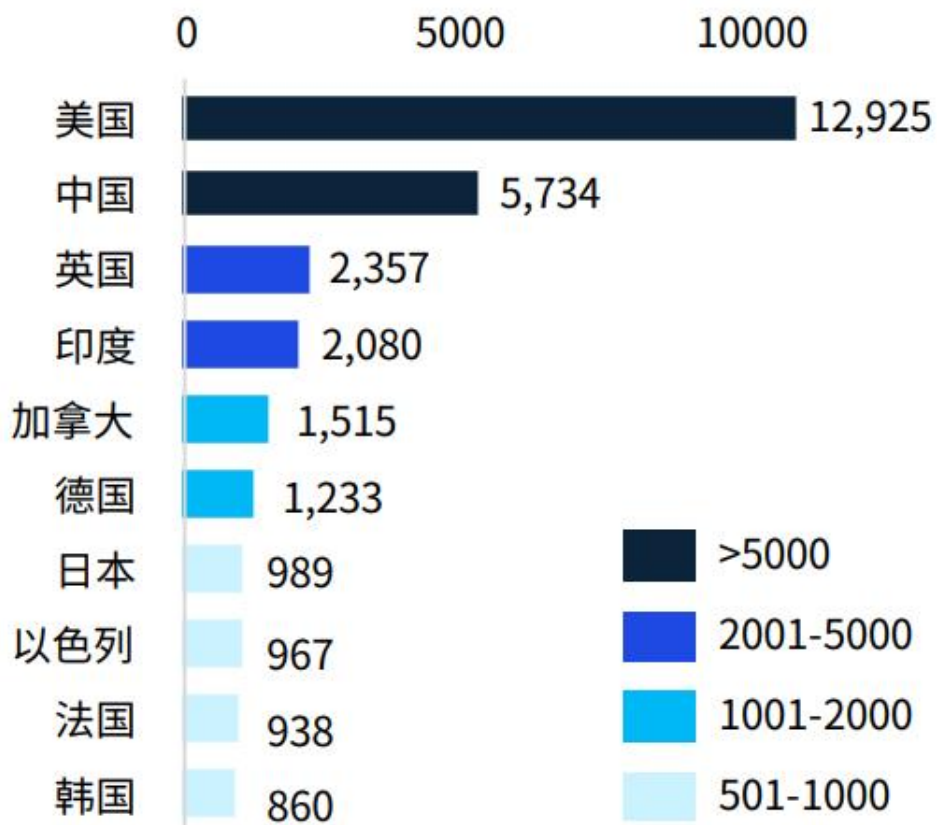
- ◆ 向量数据库是一种以向量嵌入（高维向量）方式存储和管理非结构化数据（如文本、图像或音频）的数据库，以便于快速查找和检索类似对象。最核心是相似度搜索，通过计算一个向量与其他所有向量之间的距离来找到最相似的向量（最相似的知识或内容）。
- ◆ 文本、图像、音视频等海量的非结构化数据占数据总量不断上升，这些数据需要通过机器学习算法从中提取出以向量为表示形式的“特征”。向量数据库便是为了解决对这些向量进行存储与计算的问题而兴起。

图：向量数据库结构图

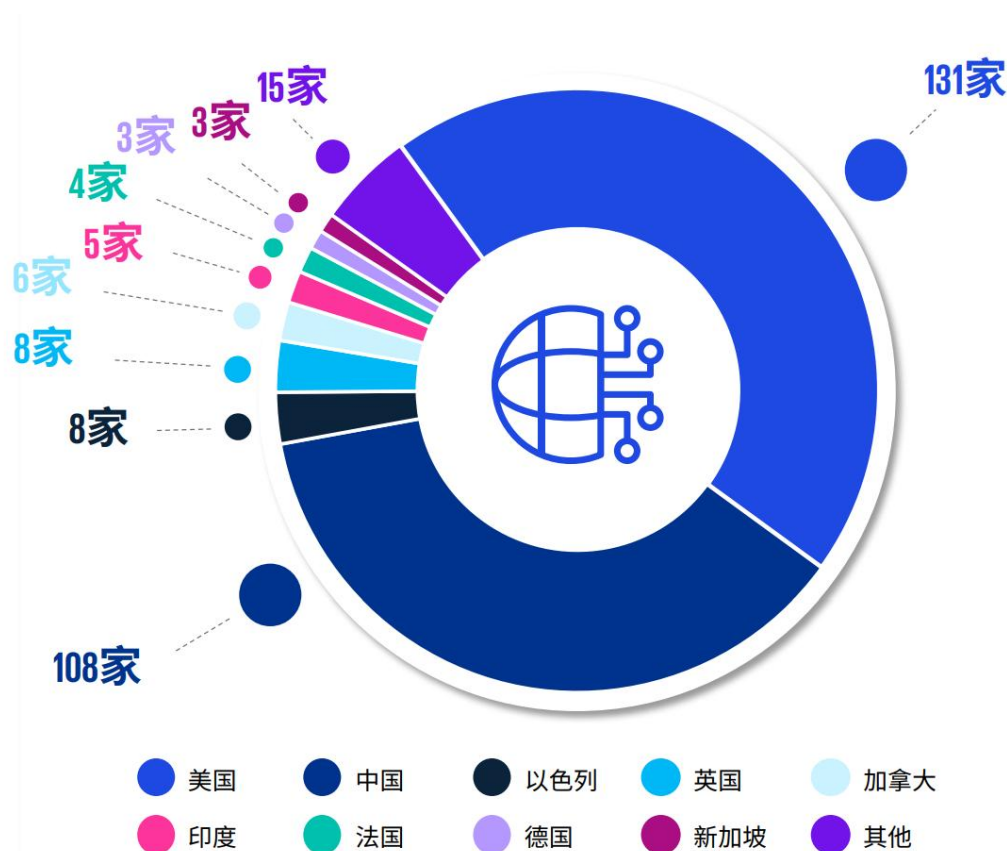


◆ 美国AI产业布局全面领先，在基础层、技术层和应用层，尤其是在算法、芯片和数据等产业核心领域，积累了强大的技术创新优势，各层级企业数量全面领先中国。

图：全球AI企业主要分布国家

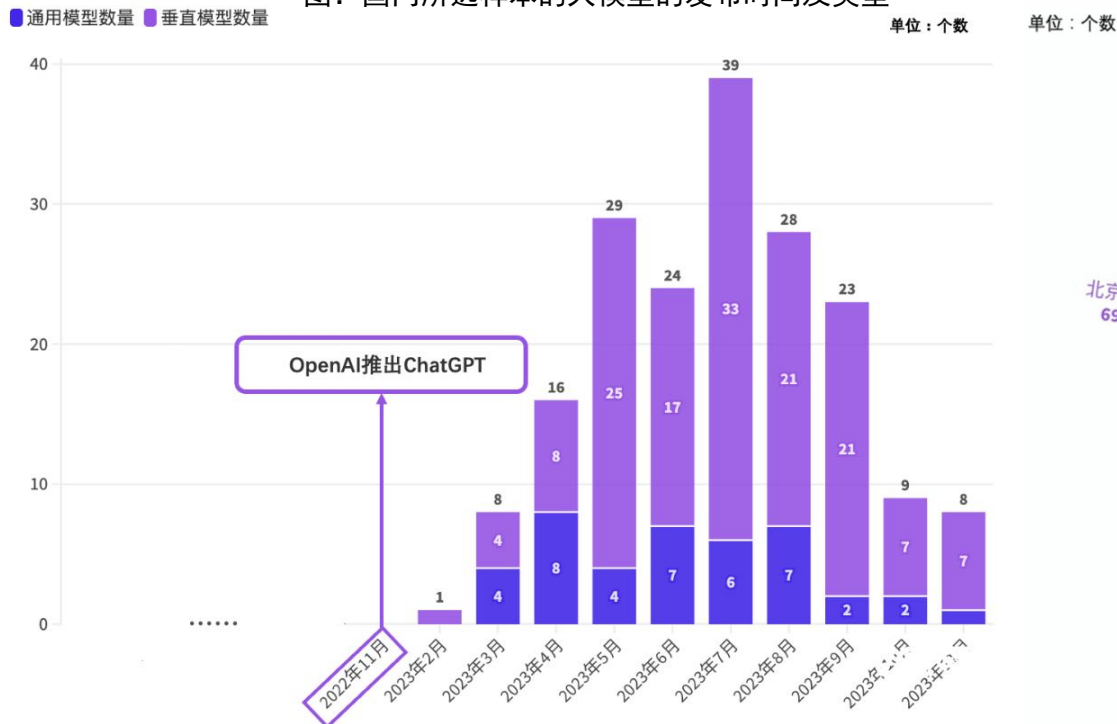


图：全球AI独角兽按国家分布

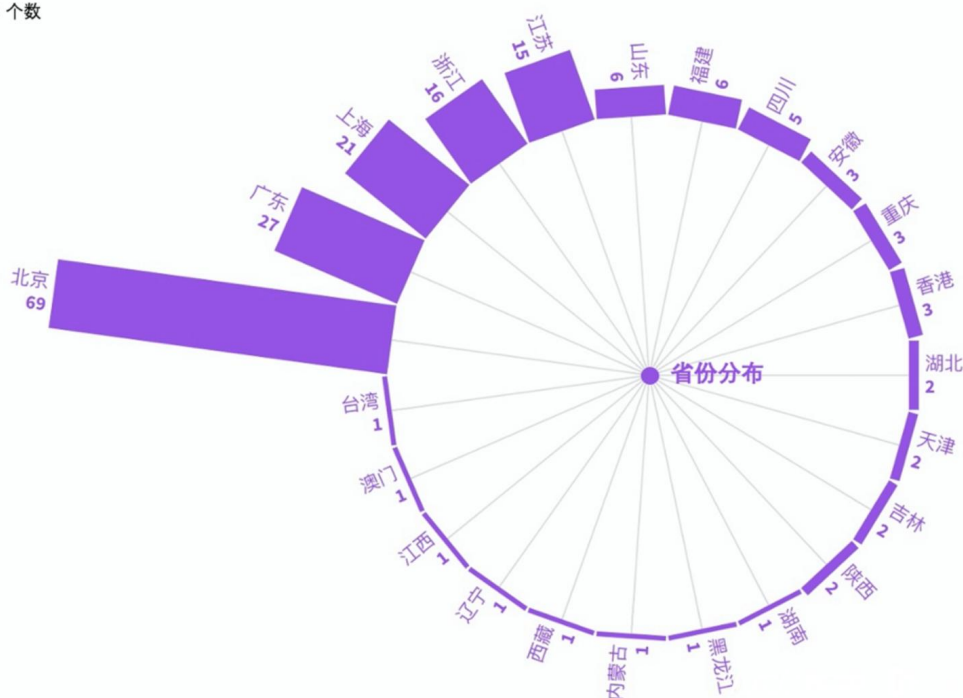


- ◆ 根据公开资料显示，截至今年10月份，国内已经发布了238个大模型。从大模型数量来看，中美两国已经是全球大模型领域的领跑者。从大模型应用类型分析，国外研发的大模型主要以通用大模型为主，而中国国产大模型呈现出行业大模型占比较高的发展形态。
- ◆ 根据2023年12月南都数字经济治理研究中心发布了《生成式AI发展与治理观察报告（2023）》显示，其统计的185个大模型样本中，通用领域的大模型共有41个，垂直领域的大模型则有144个。

图：国内所选样本的大模型的发布时间及类型



图：国内所选样本的大模型的发布地域

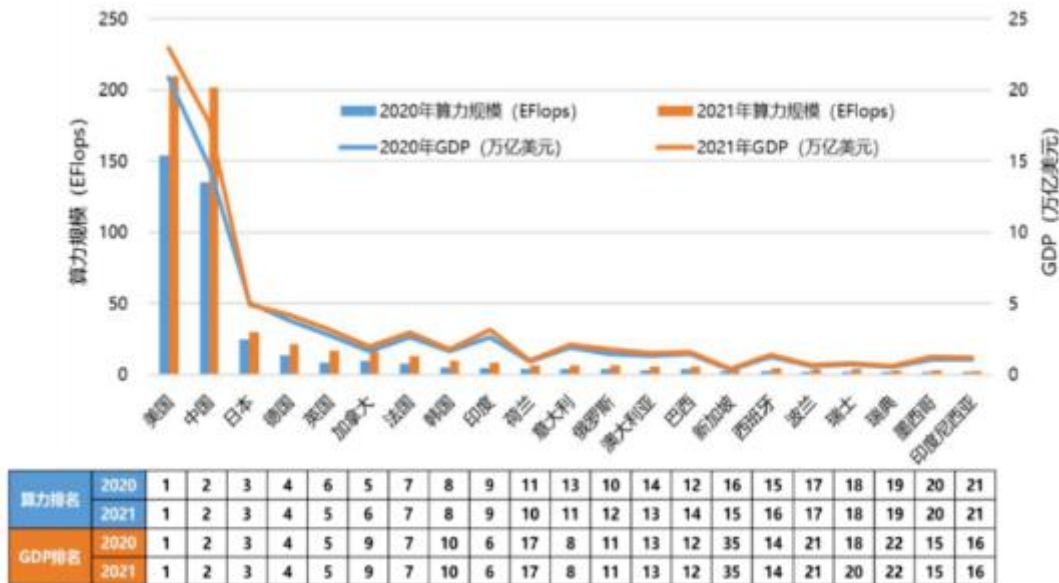


- 01 逻辑芯片市场规模巨大，产业链具备高门槛
- 02 大模型走向多模态，技术持续演进，商业前景广阔
- 03 算力板块投资的几条主线
- 04 相关标的
- 05 风险提示

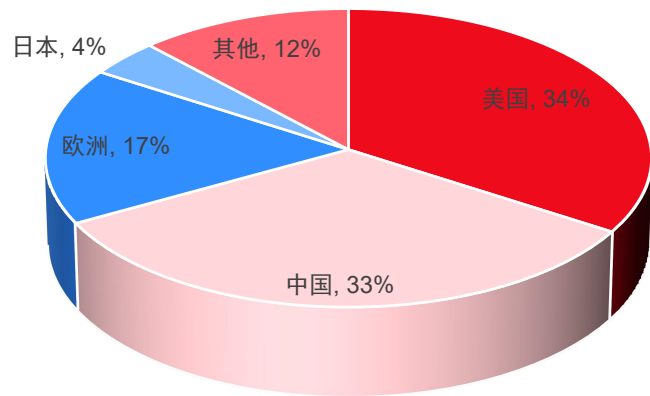
# 中国算力资源位居世界第二

- ◆ 根据中国信通院发布的《中国算力发展指数白皮书（2023）》显示，从基础设施侧看，我国通用数据中心、智能计算中心持续加快部署，2022年基础设施算力规模达到180EFlops，位居全球第二，在用数据中心机架规模超过650万标准机架，已投运智能计算中心达到25个，在家智能计算中心超过20个。
- ◆ 从计算设备侧看，我国近六年累计出货超过2091万台通用服务器，82万台AI服务器，算力总规模达到302EFlops，全球占比33%，增速达50%，其中智能算力保持稳定高速增长，增速达72%。

图：算力与经济发展关系示意图



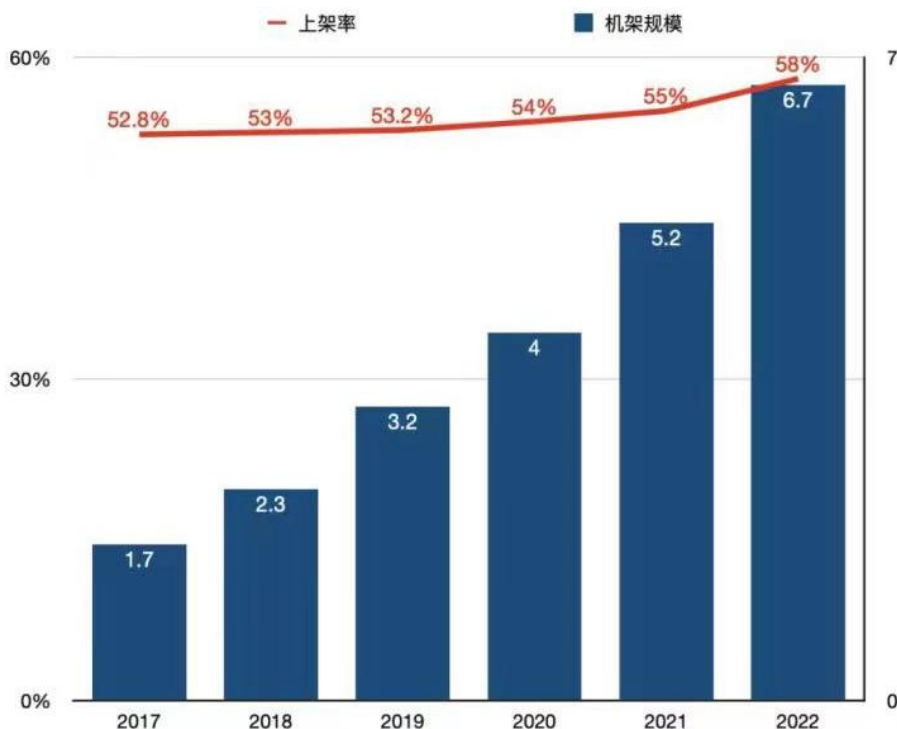
图：2022年全球算力规模分布情况



# 区域失衡、云计算占比较低

- ◆ 根据《财经》的数据，中国算力资源位居世界第二，仅次于美国，但算力利用效率不高。其中有两个重要表现，一是数据中心上架率不够高，二是算力形态中云计算的占比也低。
- ◆ 市场调研机构沙利文2023年数据显示，2022年中国数据中心整体平均上架率为58%。中国算力地区供需也相对失衡。华东、华北、华南地区上架率均在65%以上，西北、西南地区仅分别约为30%和40%。
- ◆ 根据《财经》掌握的数据，美国服务器总规模约为2200万台，公有云服务比例超过65%。中国服务器总规模约2000万，公有云输出比例仅有28%，大部分服务器以私有云形式存在。

图：2017-2022年中国数据中心整体平均上架率

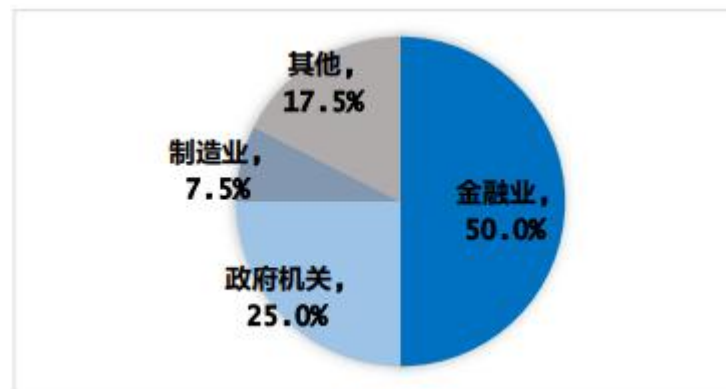


图：2023年中国算力规模现状（不完全统计）

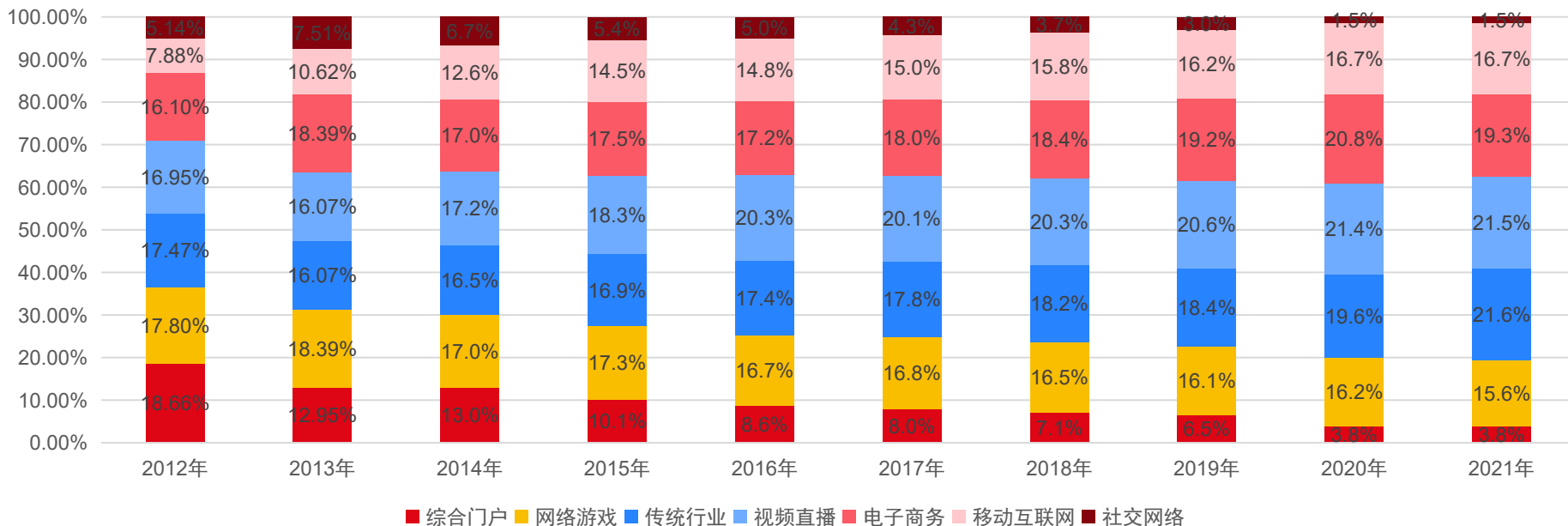
算力占比	算力提供方	服务器数量 (万台)	机架数量 (万个)
云计算与互联网公司 占比40%	阿里云	200-300	
	腾讯云	200-250	
	字节跳动	约100	
	百度	60-80	
电信运营商 占比15%	华为云	约100	
	中国移动	80	47.8
	中国电信	100-200	53.4
数据中心提供商 占比10%	中国联通	约50	38
	数据中心服务商	合计约200	
	地方大数据产业园		
政府与企业自用 占比35%	其他IT服务公司		
	金融行业	合计约200	
合计100%	其他用户自用	合计约500	
		超过2000万	650

# 传统行业等需求占比持续提升

◆ 从下游细分来看,《中国数据中心产业发展白皮书(2023年)》显示,头部互联网仍是需求基本面,腰部互联网、传统行业与外资企业需求成为增长新动力。随着产业互联网崛起,传统行业客户需求逐步被激发,其中金融、政务和制造业为需求主体。据统计,2021年数据中心下游应用在传统行业占比近20%,金融、政府机关、制造业居前三,占比超八成(82.5%)。



图：2012-2021年数据中心下游客户需求占比



◆ 2022年2月，国家发展改革委、中央网信办、工业和信息化部、国家能源局联合印发通知，同意在京津冀、长三角、粤港澳大湾区、成渝、内蒙古、贵州、甘肃、宁夏等8地启动建设国家算力枢纽节点，并规划了10个国家数据中心集群。这标志我国“东数西算”项目拉开序幕。全国一体化大数据中心体系完成总体布局设计。

图：“东数西算”工程规划的八大枢纽和十大集群



# 各地方数据中心集群建设规划汇总

◆ 根据工业和信息化部副部长张云明在2023年7月的发言，我国算力基础设施建设扎实推进。2018年以来，我国数据中心机架数量年复合增长率超过30%。截至2022年底，在用标准机架超过650万架。根据中国信息通信研究院测算，2022年我国算力核心产业规模达1.8万亿元。

地区	枢纽节点	集群	规划
东部	京津冀枢纽	张家口数据中心集群	到2023年，新增标准机柜18万架，总量达到33万架，投运数据中心平均上架率达到55%以上。到2025年，新增标准机柜55万架，总量达到70万架，数据中心平均上架率达到65%以上
	长三角枢纽	长三角生态绿色一体化发展示范区数据中心集群	预计到2025年，上海市数据中心标准机架总规模将达到28万架，江苏省数据中心标准机架总规模将达到70万架，浙江省将达到45万以上标准机架服务能力
		芜湖数据中心集群	按照“三步走”战略设定发展目标，到2023年底，在建机架规模达到10万架；2025年达到约60万架；2030年达到100万架。
	粤港澳大湾区枢纽	韶关数据中心集群	到2025年，韶关数据中心集群建成50万架标准机架、500万台服务器规模，投资超500亿元。
	成渝枢纽	天府数据中心集群	到2025年，天府数据中心集群起步区全面建成，数据中心规模将达50万机架（折合单机架功率2.5kW，下同），上架率将达到60%
西部		重庆数据中心集群	到2025年，全市标准机架规模达50万，总体上架率>75%
	贵州枢纽	贵安数据中心集群	力争到2025年，贵安集群数据中心平均上架率不低于65%，全省数据中心标准机架达到80万架、服务器达到400万台。
	内蒙古枢纽	和林格尔数据中心集群	截至2023年7月初，自治区已落地数据中心项目50个，服务器承载能力突破240万台，2028年规划潜力超过800万台，至2025年数据中心PUE（能效指标）将达到1.2，实现绿色算力赋能产业绿色转型。当前，自治区现有智算规模760P，2025年预计达到8000P
	甘肃枢纽	庆阳数据中心集群	到2023年底，庆阳国家数据中心集群要新增2.5千瓦标准机架15万个，到2025年底，累计新增2.5千瓦标准机架30万个
	宁夏枢纽	中卫数据中心集群	到2023年完成起步区建设，标准机架达到30万架。到2025年，宁夏枢纽全面建成，标准机架达到72万架，中卫数据中心集群机架达到69万架，集群PUE平均值降低到1.2，可再生能源利用率达到65%

- ◆ 2023年12月29日，由国家发展改革委、国家数据局、中央网信办、工业和信息化部、国家能源局五部门联合印发的《关于深入实施“东数西算”工程 加快构建全国一体化算力网的实施意见》正式公布。
- ◆ 《关于深入实施“东数西算”工程 加快构建全国一体化算力网的实施意见》提出，到2025年底，普惠易用、绿色安全的综合算力基础设施体系初步成型，东西部算力协同调度机制逐步完善，通用算力、智能算力、超级算力等多元算力加速集聚，国家枢纽节点地区各类新增算力占全国新增算力的60%以上，国家枢纽节点算力资源使用率显著超过全国平均水平。1ms时延城市算力网、5ms时延区域算力网、20ms时延跨国家枢纽节点算力网在示范区域内初步实现。算力电力双向协同机制初步形成，国家枢纽节点新建数据中心绿电占比超过80%。
- ◆ 《关于深入实施“东数西算”工程 加快构建全国一体化算力网的实施意见》提出，坚持需求牵引、应用导向，对数据中心整体上架率低的地区加强规划指导，显著提高通用算力资源利用率，加快实现智能算力资源供需平衡，切实推动超级算力资源便捷易用。加强新型算力基础设施系统设计，建设涵盖通用计算、智能计算、超级计算的融合算力中心，促进不同计算精度算力资源服务有机协同。引导算力基础设施建设主体以更加灵活的建设运营方式响应快速迭代的算力市场需求，促进智能计算和高性能计算等算力资源综合应用。提升智能算力在人工智能等领域适配水平，增强计算密集型、数据密集型等业务的算力支撑能力。

- ◆ 根据中国通服数字基建产业研究院发布的《中国数据中心产业发展白皮书（2023年）》显示，国内数据中心产业总体处于平稳增长期，且受数字经济深化、“东数西算”战略、AI新业态等新驱动力影响，预计未来10年仍有空间。在“双碳”与“东数西算”政策复合作用下，产业进入高质量转型期，数据中心向新一代智算中心演进。
- ◆ 2022年ChatGPT开启AIGC这一全新业态，推动AI发展进入以多模态和大模型为特色的AI2.0时代，推动智算设施建设进入新阶段。

图：2008-2025 年国内数据中心产业规模发展情况

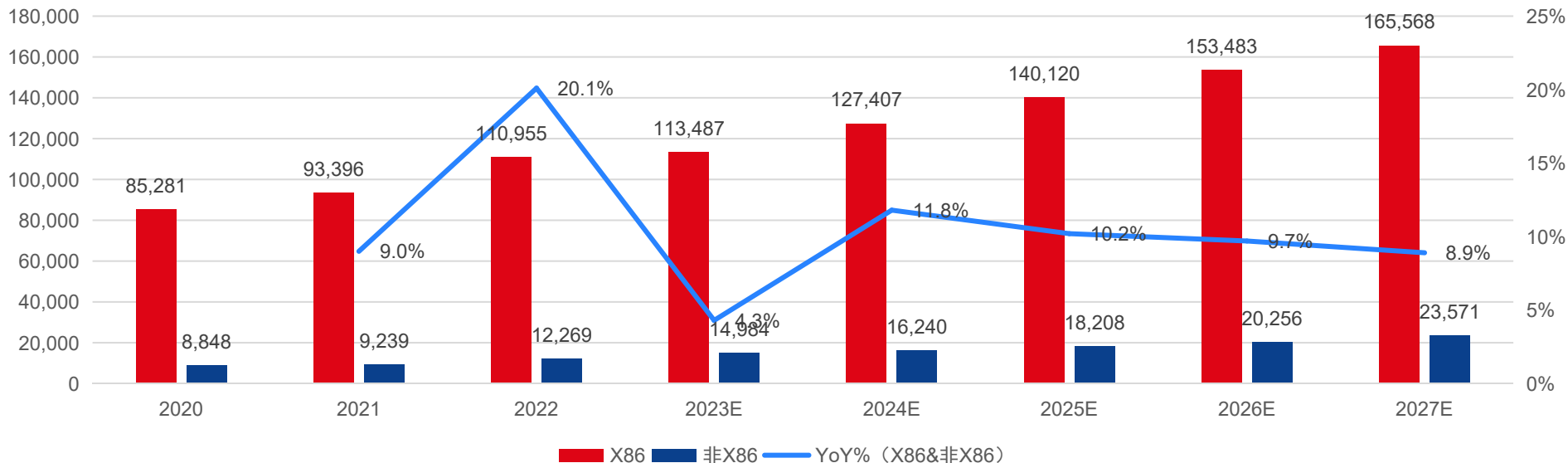


◆ IDC数据显示，2022年全球服务器市场规模同比增长20.1%达到了1232.24亿美元，预计到2027年市场规模将达1891.39亿美元。从整体的趋势来看，未来数年内，虽然非x86服务器保持了持续的高速增长，但是x86服务器仍然是整个服务器市场的主流。

图：2014-2020年全球服务器出货量

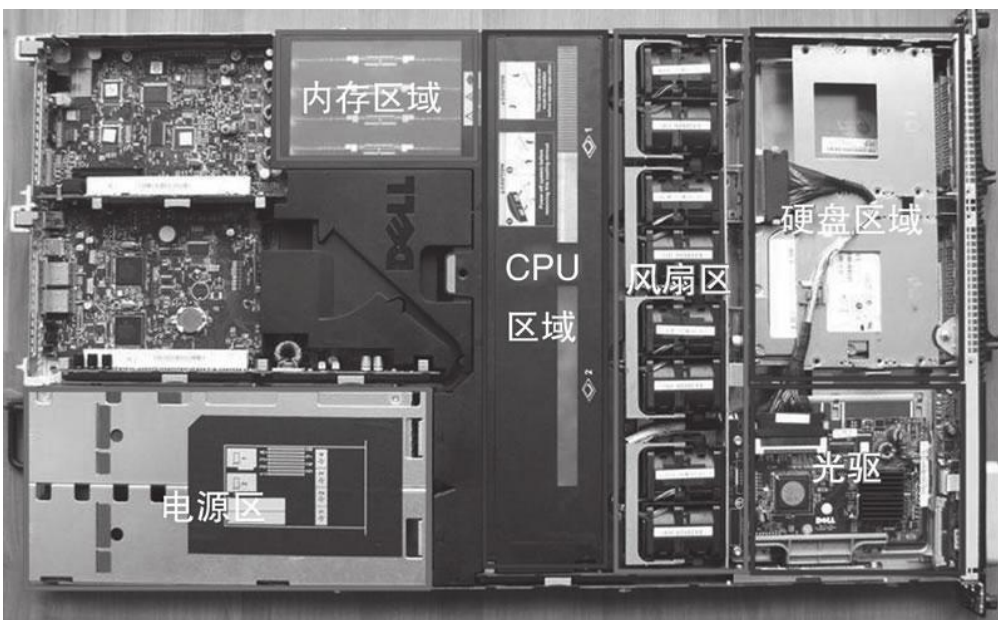


图：2020-2027年全球服务器销售额预测（百万美元）

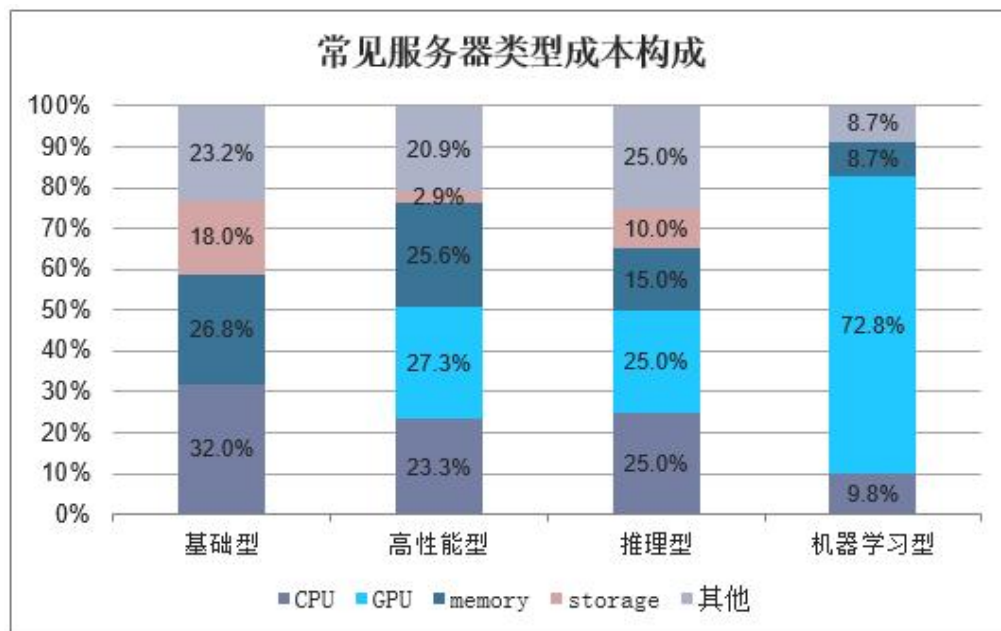


- ◆ 据Counterpoint统计显示，2022年全球数据中心CPU市场收入同比下降4.4%。AMD的数据中心CPU收入同比增长62%，占据20%的市场份额。英特尔的数据中心CPU收入在2022年同比下降16%，而其市场份额下降至71%。基于ARM的CPU受到Ampere、Graviton（亚马逊）和 Yitian（阿里巴巴）的青睐，收入首次超过10亿美元。
- ◆ 根据英特尔公布的财报数据，2022年英特尔DCG（数据中心事业部）营收为192亿美元，同比下滑26%。结合Counterpoint关于英特尔的市场份额来看，全球数据中心CPU市场规模大约在250~300亿美元。

图：服务器内部构成示意图



图：常见服务器类型成本构成



- ◆ 2021年，中国服务器市场出货量达到412万台，同比增长9.6%；厂商收入达到264.5亿美元，同比增长15.4%。2021年，按处理器类型划分，X86服务器厂商收入达到254.6亿美元，同比增长15.2%；RISC/EPIC服务器厂商收入达到2.5亿美元，同比下滑36.5%；CISC服务器厂商收入达到1.6亿美元，同比增长31.7%；ARM服务器厂商收入达到5.7亿美元，同比增长85.4%。
- ◆ 中国x86服务器以双路服务器为主，从2016年到2020年，双路服务器总占比均在80%以上；其次分别是单路、4路服务器，合计占比在10%至20%之间；8路以上的服务器较少，占比未超过0.3%。

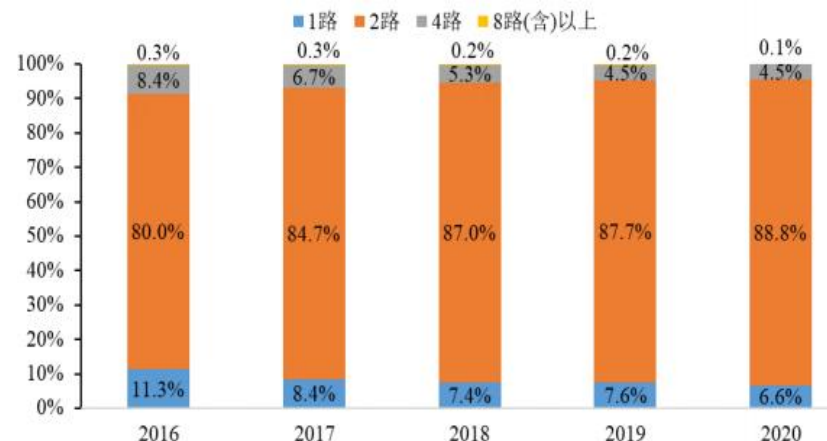
图：中国X86服务器出货量（万台）

中国x86服务器出货量和市场规模预测

■ 中国x86服务器出货量（万台）    — 同比增长



图：中国x86服务器以双路服务器为主（万台）



# 中国数据中心服务市场未来五年增速预测

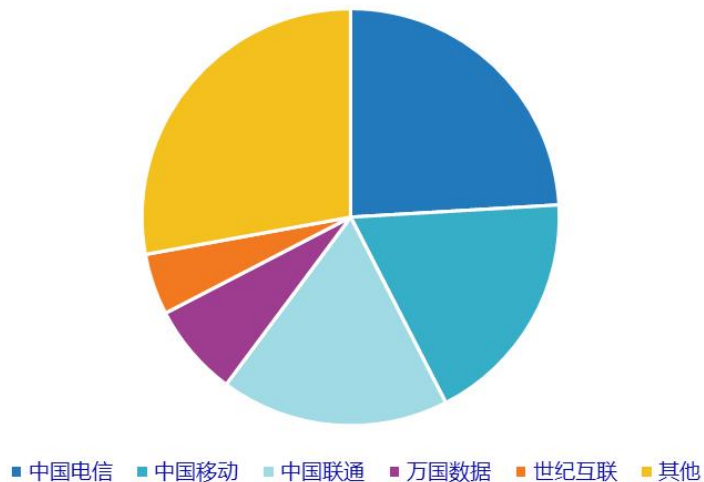
- ◆ IDC发布的《中国数据中心服务市场(2022年)跟踪》报告显示, 2022年, 中国数据中心服务市场同比增长12.7%, 市场规模达1293.5亿元人民币。IDC预计, 未来五年, 中国数据中心服务市场将以18.9%的复合增速持续增长, 预计2027年市场规模达3075亿元人民币。

图: 2022年中国Top5数据中心服务商市场份额

图: 2022年中国Top5第三方数据中心服务商市场份额



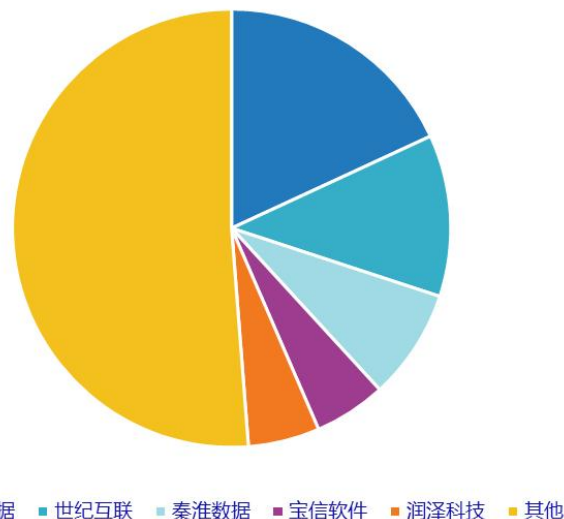
中国 Top 5 数据中心服务商市场份额, 2022



来源: IDC中国, 2023



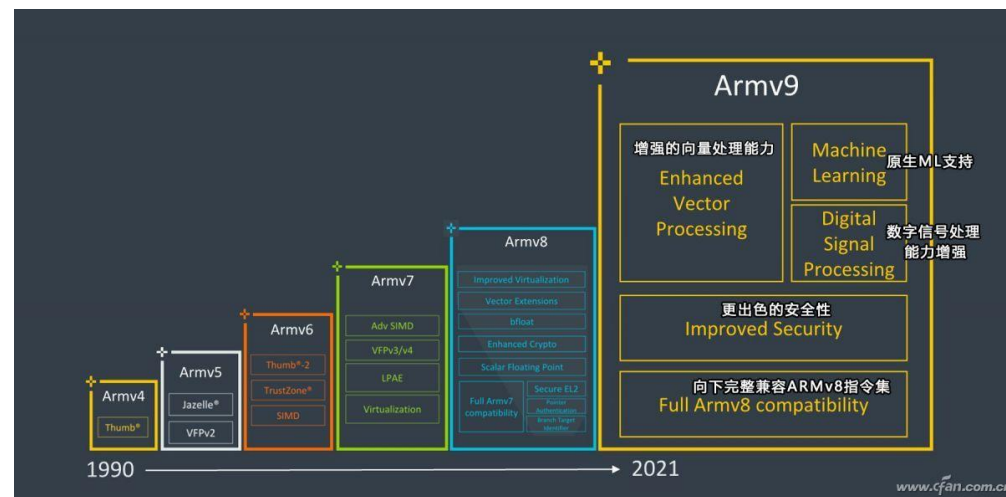
中国 Top 5 第三方数据中心服务商市场份额, 2022



来源: IDC中国, 2023

◆ 国产CPU企业目前主要有6家，分别是龙芯中科、电科申泰、华为海思、飞腾信息、海光信息、上海兆芯。按采用的指令系统类型可大致分为三类：第一类，是龙芯中科和电科申泰，早期曾分别采用MIPS兼容的指令系统和类Alpha指令系统，现已分别自主研发指令系统；第二类，是华为海思和飞腾信息，采用ARM指令系统；第三类，是海光信息和上海兆芯，采用X86指令系统。

图：ARM指令集架构发展演进



图：国产六大CPU厂商概况

	龙芯	申威	鲲鹏	飞腾	海光	兆芯
指令集体系	LoongArch+MIPS	Alpha, SW-64	ARMv8	ARMv8	x86	x86
来源	MIPS	DEC	ARM v8	ARM v8	AMD	威盛
授权模式	指令集授权	指令集授权	指令集授权	指令集授权	指令集授权	指令集授权
代表产品	龙芯1号	SW1600	鲲鹏920	腾云S系列	海光1号	KX-5000
	龙芯2号	SW1610		腾锐D系列	海光2号	KX-6000
	龙芯3号	SW26010		腾珑E系列	海光3号	KH-20000
自主迭代能力	强	强	未获得v9授权 迭代受限	未获得v9授权 迭代受限	强	强
优势	起步最早	底层应用、超算为主 力方向	移动端占比高 产品线丰富	产品线丰富	性能强	应用生态丰富
	自主化程度高				应用生态丰富	

# 大模型发展需要强大的算力支撑

- ◆ 大模型本质上是大算力、强算法、大数据共同作用下的产物，随着模型走向多模态，对于算力的需求更加巨大。根据媒体 *semianalysis* 的报道，GPT-4在120层中总共包含了1.8万亿参数，而GPT-3只有约1750亿个参数。也就是说，GPT-4的规模是GPT-3的10倍以上。OpenAI训练GPT-4的FLOPS约为 $2.15e^{25}$ ，在大约25000个A100上训练了90到100天。
- ◆ 本质上，大模型对于算力的需求存在于两个维度：第一是大模型持续训练升级所需要的算力快速增长；第二是面对大模型的商业化，算力的需求更是巨大。

图：不同模型所需算力

Model	Evolved Transformer NAS	T5	Meena	Gshard -600B	Switch Transformer	GPT-3
Number of Parameters (B)	0.064 per model	11	2.6	619	1500	175
Percent of model activated on every token	100%	100%	100%	0.25%	0.10%	100%
Developer	Google					OpenAI
Datacenter of original experiment	Google Georgia	Google Taiwan	Google Georgia	Google North Carolina	Google Georgia	Microsoft
When model ran	Dec 2018	Sep 2019	Dec 2019	Apr 2020	Oct 2020	2020
Datacenter Gross CO <sub>2</sub> e/KWh (kg/KWh when it was run)	0.431	0.545	0.415	0.201	0.403	0.429
Datacenter Net CO <sub>2</sub> e/KWh (kg/KWh when it was run)	0.431	0.545	0.415	0.177	0.330	0.429
Datacenter PUE (when it was run)	1.10	1.12	1.09	1.09	1.10	1.10
Processor	TPU v2		TPU v3			V100
Chip Thermal Design Power (TDP in Watts)	280		450			300
Measured System Average Power per Accelerator, including memory, network interface, fans, host CPU (W)	208	310	289	288	245	330
Measured Performance (TFLOPS/s) <sup>12</sup>	24.8	45.6	42.3	48.0	34.4	24.6
Number of Chips	200	512	1024	1024	1024	10,000
Training time (days)	6.8	20	30	3.1	27	14.8
Total Computation (floating point operations)	2.91E+21	4.05E+22	1.12E+23	1.33E+22	8.22E+22	3.14E+23
Energy Consumption (MWh)	7.5	85.7	232	24.1	179	1,287
% of Google 2019 total energy consumption (12.2 TWh = 12,200,000 MWh) [Goo20]	0.00006%	0.00070%	0.00190%	0.00020%	0.00147%	0.01055%
Gross tCO <sub>2</sub> e for Model Training	3.2	46.7	96.4	4.8	72.2	552.1
Net tCO <sub>2</sub> e for Model Training	3.2	46.7	96.4	4.3	59.1	552.1
Fraction of NAS Estimate in [Str19] (284 tCO <sub>2</sub> e)	0.011	0.164	0.340	0.015	0.208	1.944
Fraction of equivalent jet plane CO <sub>2</sub> e round trip San Francisco ↔ New York (~180 t; see Ap. A)	0.018	0.258	0.533	0.024	0.327	3.054
tCO <sub>2</sub> e savings by Meena using Evolved Transformer	--	--	48.5	--	--	--
% 24/x7 carbon free energy (when run)	31%	19%	30%	73%	43%	N/A

- ◆ 英伟达H100由于具备更好的拓展性以及更快的训练时间而成为训练大模型公司的首选，此类型海外客户既有OpenAI、Anthropic等训练大模型的初创公司，也有Azure、GCP等各种云服务提供商。
- ◆ 外媒曝出Sam Altman正计划筹集数十亿美元，为OpenAI建起全球性的半导体晶圆厂网络，GPT-5训练严重缺芯。
- ◆ H100的物料核心主要包括核心逻辑芯片、HBM内存芯片、CoWoS封装，逻辑芯片采用台积电4N工艺制造（5nm+），并使用台积电的CoWoS先进封装技术，HBM内存芯片来自SK海力士的HBM3内存芯片。据科技新报报道，台积电已释出明年CoWoS将翻倍量产的消息，比原本目标的计划再增20%，计划明年月产35000片晶圆。

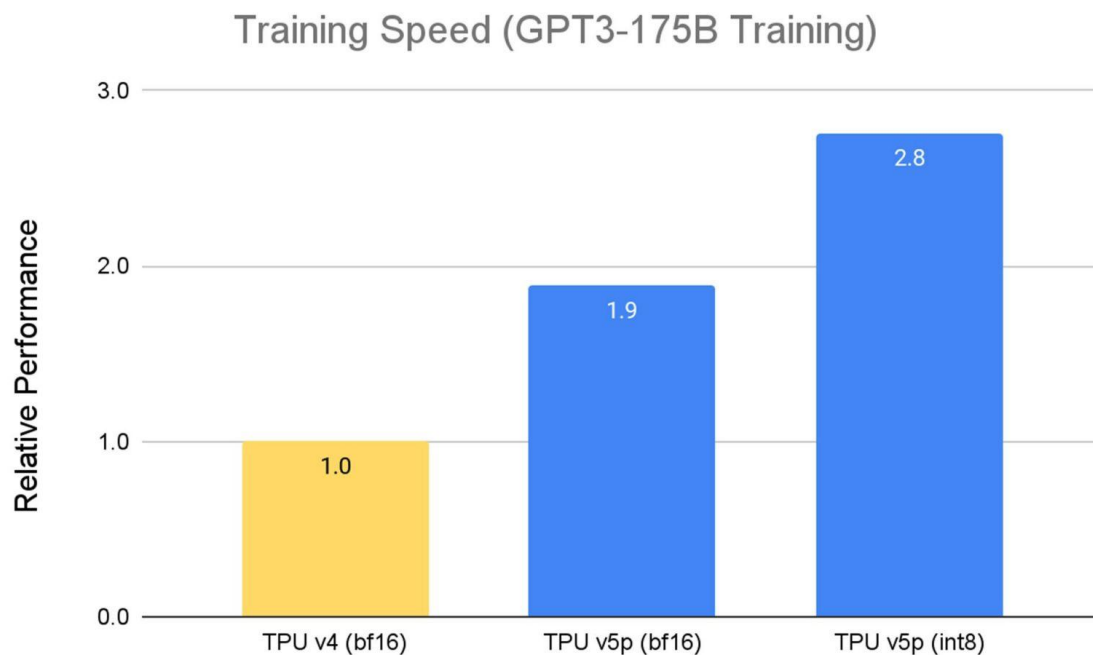
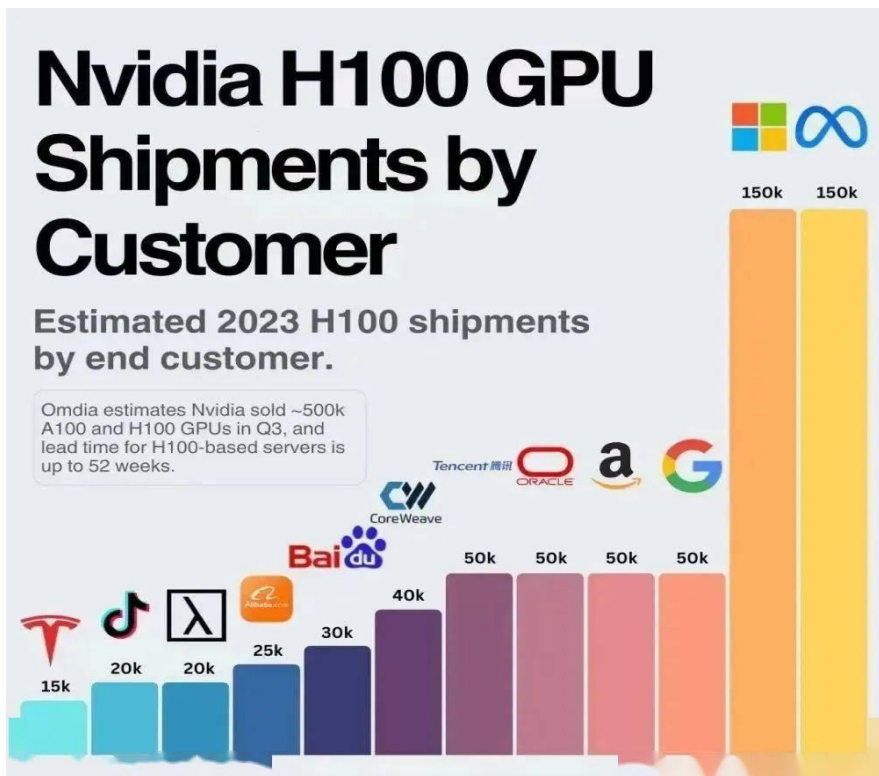
图：英伟达H100核心物料



- ◆ 为了减少对英伟达的依赖，AI芯片使用者开始寻求备选方案，主要有两个方向：第一，对于大厂来说，自研AI芯片当下成为一个趋势；第二，AMD等芯片厂商也持续推出更高端的AI芯片，预计使用者对于相关厂商的态度也会变得更加积极。
- ◆ 据报道，OpenAI 首席执行官山姆·奥尔特曼（Sam Altman）正在筹集资金建设半导体生产设施，生产用于人工智能（AI）应用的处理器。

图：英伟达2023年Q3季度Top12的H100显卡购买客户

图：谷歌Cloud TPU v5p的训练速度



# 发布MI300X GPU芯片加速卡，AMD挑战英伟达


- ◆ 2023年12月7日，AMD正式宣布了旗舰 AI GPU 加速器 MI300X，该芯片完全基于 CDNA 3 架构设计，混合使用 5nm 和 6nm IP，晶体管数量达到 1530 亿个，内存方面，MI300X 采用HBM3内存，容量最高192GB。

## AMD Instinct™ MI300X GPU vs. Competition

		MI300X (Up to)	H100 SXM	AMD Instinct™ Advantage (Up to)
Hardware Specifications	TBP	750W	700W	-
	Memory Capacity	192 GB HBM3	80GB HBM3	2.4x
	Memory Bandwidth (Peak Theoretical)	~5.3 TB/s	3.3TB/s	1.6x
HPC Performance (Peak Theoretical)	FP64 Matrix / DGEMM (TFLOPS)	163.4	66.9 (Tensor)	2.4x
	FP32 Matrix / SGEMM (TFLOPS)*	163.4	N/A	N/A
	FP64 Vector / FMA64 (TFLOPS)	81.7	33.5	2.4x
	FP32 Vector / FMA32 (TFLOPS)	163.4	66.9	2.4x
AI Performance (Peak Theoretical)	TF32 (Matrix)	653.7	494.7	1.3x
	TF32 w// Sparsity (Matrix)	1307.4	989.4	1.3x
	FP16 (TFLOPS)	1307.4	133.8   989.4 (Tensor)	9.8x   1.3x
	FP16 w/Sparsity (TFLOPS)	2614.9	1978.9 (Tensor)	1.3x
	BFLOAT16 (TFLOPS)	1307.4	133.8 / 989.4 (Tensor)	9.8x   1.3x
	BFLOAT16 w/Sparsity (TFLOPS)	2614.9	1978.9 (Tensor)	1.3x
	FP8 (TFLOPS)	2614.9	1978.9	1.3x
	FP8 w/Sparsity (TFLOPS)	5229.8	3957.8 (Tensor)	1.3x
	INT8 (TOPS)	2614.9	1978.9	1.3x
	INT8 w/Sparsity (TOPS)	5229.8	3957.8 (Tensor)	1.3x

52 | AMD INSTINCT™ MI300 PRESS AND ANALYST PRE-BRIEF DECK | UNDER EMBARGO UNTIL DECEMBER 6, 2023

See endnotes: MI300-05A, MI300-17 MI300-18  
\* Nvidia H100 GPUs don't support FP32 Tensor. Nvidia H100 source: <https://resources.nvidia.com/en-us-tensor-rt>

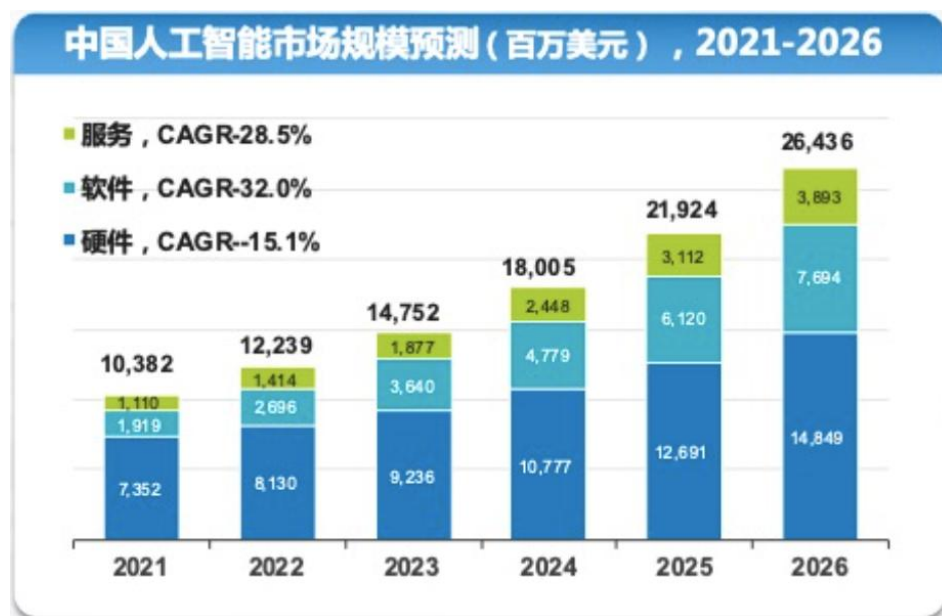
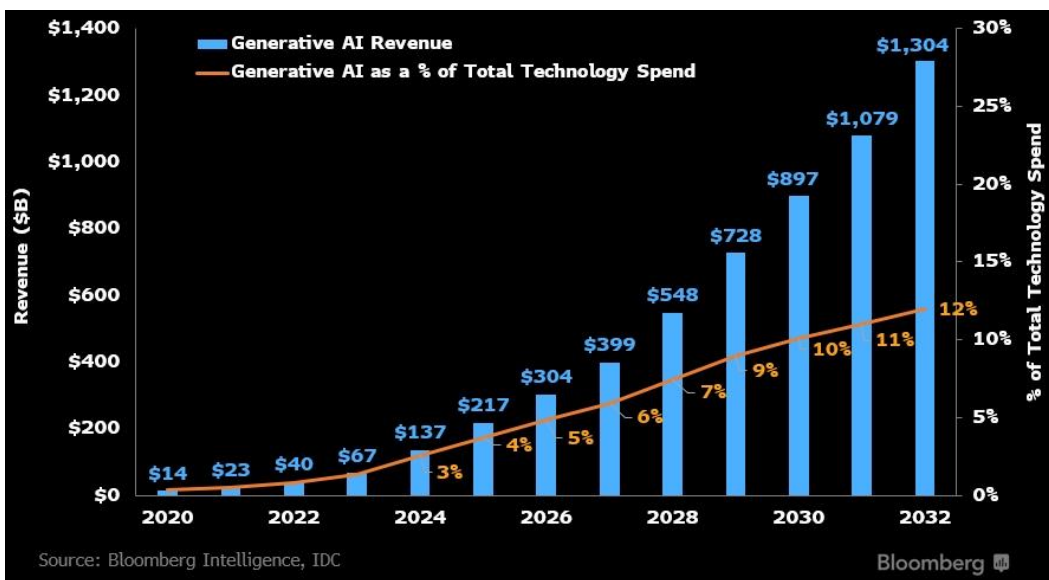


together we advance.

- ◆ 根据国际数据公司IDC预测，全球AI计算市场规模将从2022年的195.0亿美元增长到2026年的346.6亿美元。其中，生成式AI计算市场规模将从2022年的8.2亿美元增长到2026年的109.9亿美元，复合增长率高达91%。
- ◆ 华为昇腾计算业务总裁张迪焯在2023世界人工智能大会上表示，“这两年内，大模型带来了750倍算力需求的增长，而硬件的算力供给（增长）仅有3倍。”而这个算力缺口还在进一步扩大当中，张迪焯预计，到2030年AI所需的算力，相对2020年还会增长500倍。

图：生成式AI市场规模预测

图：中国人工智能市场规模预测



# 高端GPU进口受限，国产崛起号角吹响

- ◆ BIS（美国商务部工业与安全局）于2023年10月17日发布了1017新规，针对高算力芯片3A090的管控指标，改为“总处理性能”及“性能密度”两个指标。
- ◆ 英伟达的H20等特供中国版合规AI芯片预计将在2024年2季度量产，但英伟达降级版芯片领先中国本土产品的性能优势正在缩小。科大讯飞董秘、副总裁江涛在2023年三季度业绩说明会上指出，当前华为昇腾910B的能力已经基本达到英伟达A100的水平。

图：英伟达的H20等特供中国版合规AI芯片参数

Product Line Up Specifications			
	HGX H20	L20 PCIe	L2 PCIe
GPU Architecture	NVIDIA Hopper	NVIDIA Ada Lovelace	NVIDIA Ada Lovelace
GPU Memory	96 GB HBM3	48 GB GDDR6 w/ ECC	24 GB GDDR6 w/ ECC
GPU Memory Bandwidth	4.0 TB/s	864 GB/s	300 GB/s
INT8   FP8 Tensor Core*	296   296 TFLOPS	239   239 TFLOPS	193   193 TFLOPS
BF16   FP16 Tensor Core*	148   148 TFLOPS	119.5   119.5 TFLOPS	96.5   96.5 TFLOPS
TF32 Tensor Core*	74 TFLOPS	59.8 TFLOPS	48.3 TFLOPS
FP32	44 TFLOPS	59.8 TFLOPS	24.1 TFLOPS
FP64	1 TFLOPS	N/A	N/A
RT Core	N/A	Yes	Yes
MIG	Up to 7 MIG	N/A	N/A
L2 Cache	60 MB	96 MB	36 MB
Media Engine	7 NVDEC 7 NVJPEG	3 NVENC (+AV1) 3 NVDEC 4 NVJPEG	2 NVENC (+AV1) 4 NVDEC 4 NVJPEG
Power	400 W	275W	TBD
Form Factor	8-way HGX	2-slot FHFL	1-slot LP
Interconnect	Pcie Gen5 x16: 128 GB/s NVLink: 900GB/s	PCIe Gen4 x16: 64 GB/s	PCIe Gen4 x16: 64 GB/s
Availability	PS: Nov 2023 MP: Dec 2023	PS: Nov 2023 MP: Dec 2023	PS: Dec 2023 MP: Jan 2024

图：部分国产算力公司芯片参数

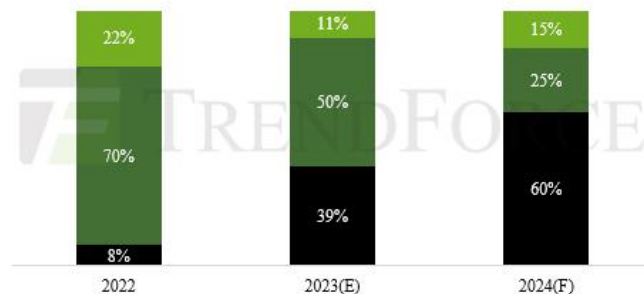
公司	产品	使用架构	算力 (INT8)	功耗	应用	制程
华为	昇腾910	ASIC	640TOPS	310W	训练	7nm
寒武纪	思元370	ASIC	256TOPS	150W	推理/训练	7nm
昆仑芯	R200	ASIC	256TOPS	150W	推理	7nm
燧原科技	T20	GPGPU	256TOPS	300W	训练	12nm
天数智芯	智铠100	GPGPU	384TOPS	150W	推理/训练	7nm
平头哥	含光800	ASIC	820TOPS	-	推理	12nm
沐曦科技	MXN100	GPGPU	160TOPS	-	推理	7nm
壁刃科技	BR100	GPGPU	-	300W	推理/训练	7nm

# HBM助力算力大发展，前景广阔

◆ TrendForce预估，2024年全球HBM的位元供给有望增长105%；HBM市场规模也有望于2024年达89亿美元，同比增长127%；预计至2026年市场规模将达127.4亿美元，对应CAGR约37%。

图：HBM代际份额占比

Projected Shift in Proportions of HBM2e and HBM3 from 2022 to 2024



Source: TrendForce, Aug., 2023 ■ HBM3 ■ HBM2e ■ Others

图：HBM技术演进图



# 供需不平衡，大厂积极扩产

◆ HBM市场目前被SK海力士、三星及美光三大DRAM原厂牢牢占据，集邦咨询（TrendForce）调查显示，2022年三大原厂HBM市占率分别为SK海力士50%、三星约40%、美光约10%。SK海力士在HBM3生产方面处于领先地位，是英伟达服务器GPU的主要供应商。另一方面，三星则专注于满足其他云服务提供商（CSP）的订单。

图：三大原厂HBM市场份额

Projected HBM Market Share Distribution from 2022 to 2024

Company	2022	2023(E)	2024(F)
SK hynix	50%	46~49%	47~49%
Samsung	40%	46~49%	47~49%
Micron	10%	4~6%	3~5%
<b>Total</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>

Source: TrendForce, Aug., 2023

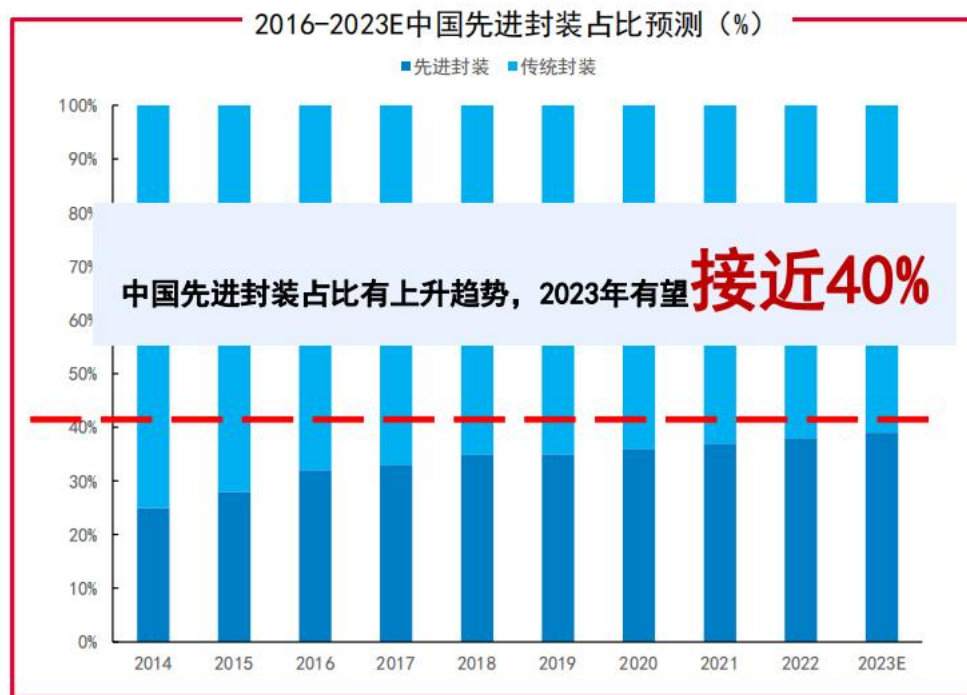
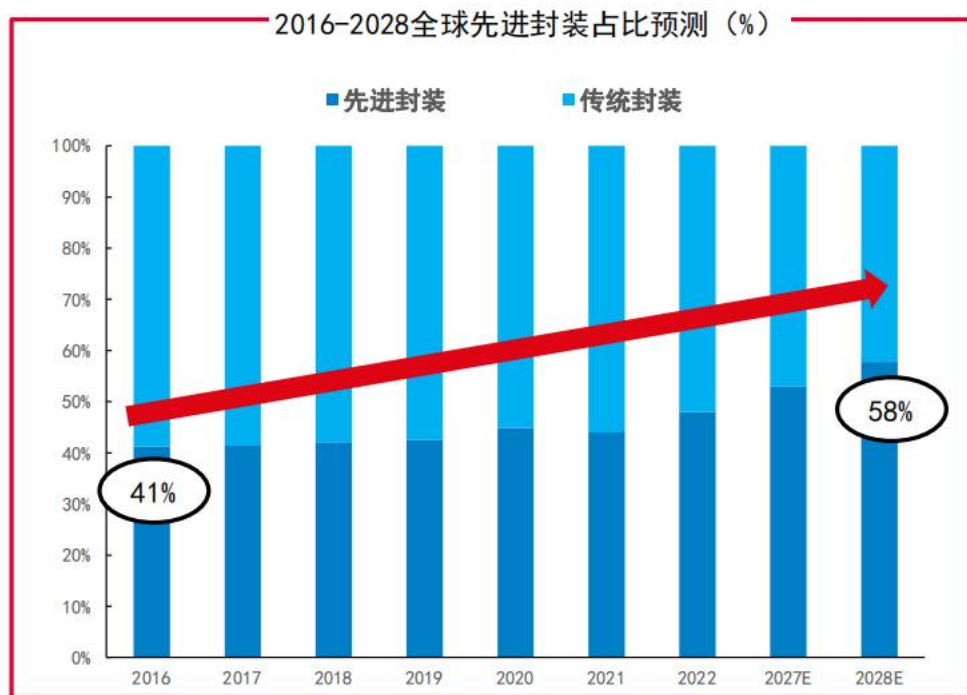
图：三大原厂HBM解决方案开发进度



Source: TrendForce, Nov., 2023

# 先进封装是封装市场主要发展趋势

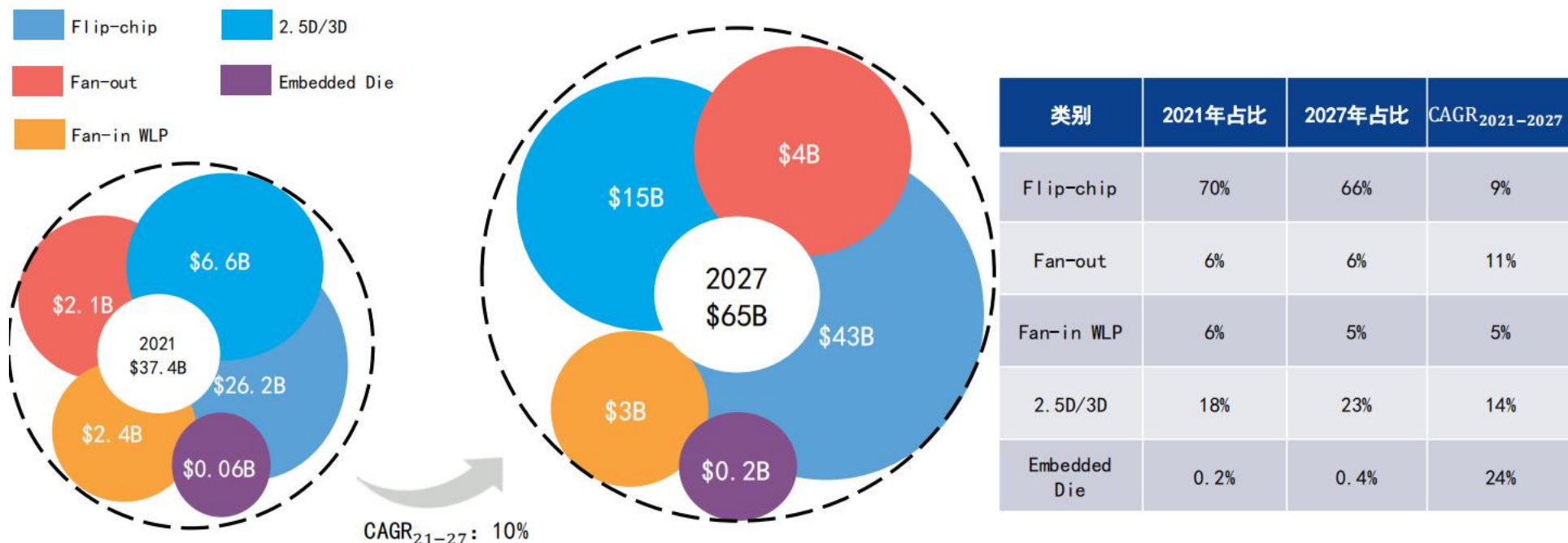
- ◆ 先进封装技术通过采用更紧凑、更高级设计和制程技术，可提供更高集成度，更小尺寸，更高性能及更低能耗芯片。
- ◆ 2028年封装市场规模有望达1360亿美元，先进封装占比约为58%。根据Yole数据，预计2022年至2028年封装市场预计将以6.9%复合年增长率增长，2028年将达到1,360亿美元，其中传统封装市场年均复合增长率将放缓至3.2%，达到575亿美元，先进封装为786亿美元，占比为57.79%。



# 2027年先进封装市场规模有望达650亿美元

- ◆ 先进封装市场有望达650亿美元，芯片倒装占比最大，芯片嵌入式封装增速最快。根据Yole数据，2021年全球先进封装市场规模为374亿美元，其中芯片倒装占比最大为70%，2.5D/3D封装次之；2027年全球先进封装市场规模预计为650亿美元，其中芯片倒装占比为66%（较2021年下滑4pcts），2.5D/3D占23%，约150亿美元，芯片嵌入式封装增速最快，21-27年CAGR为24%。

图：2021-2027先进封装市场规模预测



- ◆ 从长期来看，先进封装技术必将随着终端应用的升级和对芯片封装性能的提升而蓬勃发展。随着新技术演进，以2.5D/3D为代表的先进封装工艺已深入大多数高端芯片生产。

图：不同芯片对于封装技术的需求

应用领域	CPU/GPU	APU	DPU	MCU	ASIC	FPGA	存储	传感器	模拟	光电子		
人工智能	FC、2.5D/3D、FO、SiP	FC、FO、ED		FC、WB、QFN、WLCSP	FC、FO	FC、2.5D/3D、FO		FC、FO、WB、QFN、WLCSP、SiP	FC、FO、WB、QFN、ED、SiP			
智能驾驶												
AR/VR												
HPC			FC、FO、ED			FC、2.5D/3D、FO				FC、2.5D/3D、WB、SiP		
IoT				FC、WB、QFN、WLCSP			FC、3D、WB、QFN、WLCSP、SiP	FC、FO、WB、QFN、WLCSP、SiP				
5G	FC、2.5D/3D、FO、SiP	FC、FO、ED									FC、FO、WB、QFN、ED、SiP	FC、2.5D/3D、WB、SiP
手机通信										FC、FO、WB、QFN、WLCSP、SiP		
区块链	FC、2.5D/3D、FO				FC、2.5D/3D、FO							

- ◆ 在AI领域，模型轻量化是指将复杂的深度学习模型转换为更小、更快、更低功耗的模型，以便在资源有限的设备上推理。模型轻量化的主要目标是在保持模型性能的前提下，将模型的大小和计算复杂度最小化。这样可以降低模型的存储和计算开销，从而提高模型的实时性和可扩展性。
- ◆ 模型轻量化可以有两种方式：第一是可以使用模型压缩和优化技术来达到轻量化；第二是让大模型只维持“基本功能”，外挂模块实现能力扩展，也可以实现大模型的轻量化。

图：AI大模型轻量化的驱动技术



- ◆ 2023年10月24日的高通骁龙峰会上，高通推出最新一代旗舰移动平台骁龙8 Gen3，采用台积电4nm制程工艺打造，相较于骁龙8 Gen2，CPU整体性能提升30%，能效提升20%。NPU方面，AI算力超过73TOPS，高通AI引擎实现了终端设备上世界首次支持运行100亿参数的模型，并且针对70亿参数LLM每秒能够生成20个token。
- ◆ 2023年11月6日晚，联发科正式发布了年度旗舰SoC天玑9300，天玑9300集成了MediaTek第七代AI处理器APU 790，专为生成式AI而设计，其性能和能效相较上一代得到显著提升，整数运算和浮点运算的性能是前一代的两倍，功耗降低了45%。早前在2023年10月18日，联发科宣布与vivo在AI领域深度合作和联调，率先实现了10亿和70亿AI大语言模型以及10亿AI视觉大模型在手机端侧的落地。



**终端侧智能 集大成者**

**AI**  
高通®Hexagon™ NPU  
98% 性能提升  
40% 能效提升  
高速运行 Stable Diffusion 和 ControlNet  
<1秒 目前全球最快速度  
77GB/s, 4.8GHz 内存带宽

**Qualcomm AI Stack**  
PyTorch ExecuTorch 和优化的模型  
首个支持多模态生成式AI模型  
终端侧个性化 利用高通传感器中核  
全球首个 支持推测解码技术

**4纳米 制程工艺**  
高通®Kryo™ CPU  
提升的性能和能效  
3.3GHz CPU最高主频达3.3GHz  
30% 性能提升  
20% 能效提升

**安全**  
Truepic 符合C2PA Android strongbox保护  
始终感知的双ISP

**游戏**  
25% 性能提升  
25% 能效提升  
虚幻引擎5 支持Lumen光照系统  
240FPS 全局光照 下一代光照效果

**影像**  
照片扩展 视频对象擦除  
Night vision 视频拍摄  
Dolby HDR 照片拍摄

**音频**  
24bit 96kHz 无损音质  
高通®扩展个人局域网 (XPAN)



**MEDIATEK | VIVO**

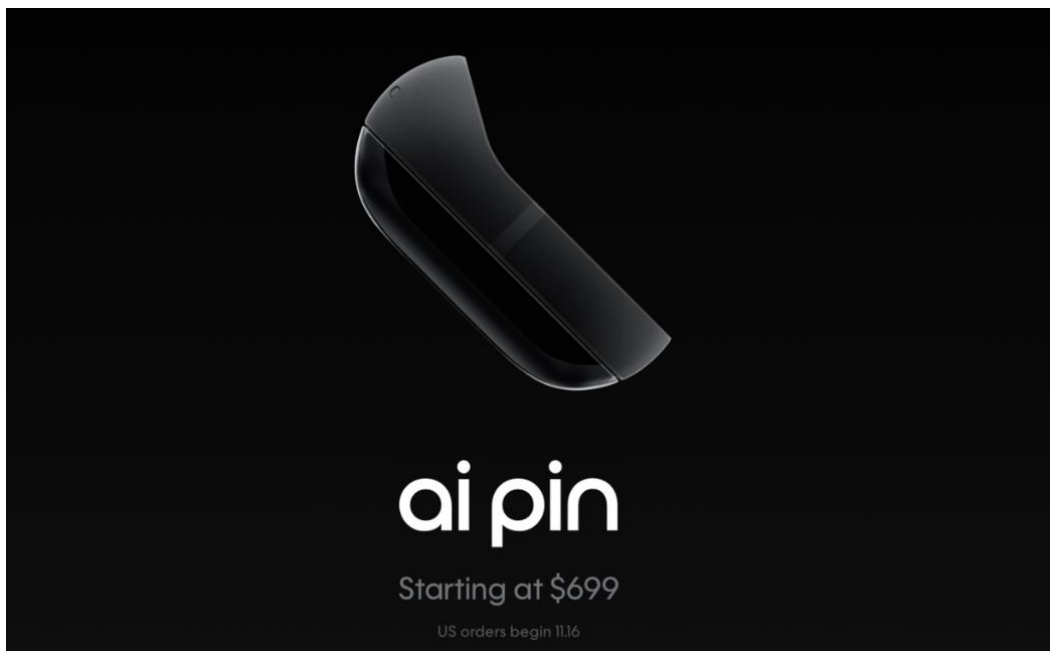
10亿和70亿生成式AI大语言模型  
10亿视觉大模型  
率先落地终端侧

5G

**MediaTek 天玑**

- ◆ 2023年11月，Humane 公司推出了首款无屏幕可穿戴设备Ai Pin，这是一种安置在服装上的微型投影，可以投屏在手掌上进行交互，目的是“替代手机”，售价为699美元。Ai Pin无需与智能手机或其他辅助设备配对，是一款无屏幕的独立设备和软件平台，并通过专有软件和OpenAI的GPT-4的组合来支持设备功能。
- ◆ 2024年初的CES盛会上，初创公司Rabbit推出的手持式AI硬件“Rabbit R1”（R1）。其内嵌的LAM模型可以使用户通过语音方式与R1进行对话交流。

图：AIPIN



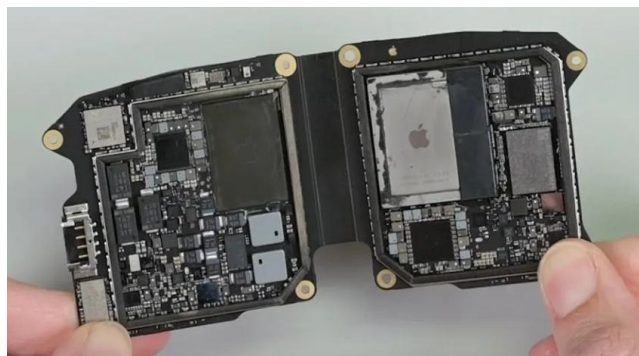
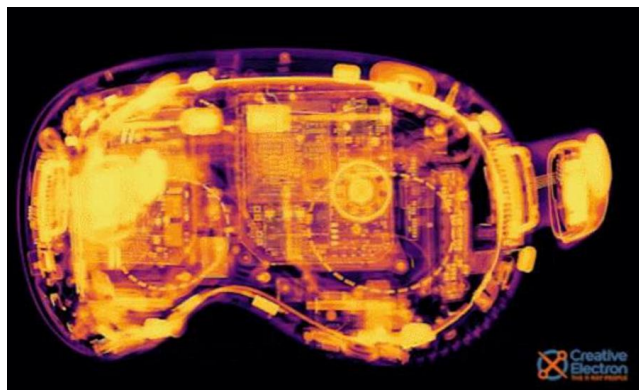
图：Rabbit R1



# Vision Pro发售，空间计算时代开启

- ◆ 2024年2月2日，苹果混合现实（MR）头戴式设备Vision Pro正式发售，售价为3499美元（约合人民币2.5万元），搭载M2芯片和R1芯片，整机包括12个摄像头、5个传感器和6个麦克风，若插电，即可全天使用；若用外接电池，则有2个小时的续航时间；Apple Vision Pro采用Micro OLED（硅基OLED）屏幕，两块屏幕像素2300万，支持4K广色域视频和HDR渲染。

图：VisionPro拆机



01

逻辑芯片市场规模巨大，产业链具备高门槛

02

大模型走向多模态，技术持续演进，商业前景广阔

03

算力板块投资的几条主线

04

相关标的

05

风险提示

- ◆ 公司的主营业务是研发、设计和销售应用于服务器、工作站等计算、存储设备中的高端处理器。公司的产品包括海光通用处理器（CPU）和海光协处理器（DCU）。海光CPU系列产品兼容x86指令集以及国际上主流操作系统和应用软件，软硬件生态丰富，性能优异，安全可靠，已经广泛应用于电信、金融、互联网、教育、交通等重要行业或领域。公司研发出的CPU产品包括海光一号、海光二号、海光三号。
- ◆ 海光DCU属于GPGPU的一种，能够较好地适配、适应国际主流商业计算软件和人工智能软件，软硬件生态丰富，可广泛应用于大数据处理、人工智能、商业计算等计算密集类应用领域，主要部署在服务器集群或数据中心，为应用程序提供高性能、高能效比的算力，支撑高复杂度和高吞吐量的数据处理任务。

产品类型	主要产品	指令集	产品特征	典型应用场景
高端处理器	通用处理器-海光CPU	兼容x86指令集	内置多个处理器核心，集成通用的高性能外设接口，拥有完善的软硬件生态环境和完备的系统安全机制。针对不同应用场景对高端处理器计算性能、功能、功耗等技术指标的要求，分别提供海光7000系列产品、5000系列产品、3000系列产品	云计算、物联网、信息服务等
	协处理器-海光DCU	兼容“类CUDA”环境	内置大量运算核心，具有较强的并行计算能力和较高的能效比，适用于向量计算和矩阵计算等计算密集型应用	大数据处理、人工智能、商业计算等

- ◆ 根据公司2023年半年报显示，海光CPU系列产品海光三号为主力销售产品，海光四号、海光五号处于研发阶段。公司新产品加速迭代，性能持续提升，研发团队在高端处理器设计、SoC架构设计、处理器安全、处理器验证、高主频与低功耗处理器实现、高端芯片IP设计、工艺物理设计、先进封装设计、基础软件等关键技术上不断实现突破。
- ◆ DCU方面，根据公司2023年09月19日发布的投资者调研纪要显示，公司深算二号已经发布并实现商用，深算二号实现了在大数据处理、人工智能、商业计算等领域的商业化应用，具有全精度浮点数据和各种常见整型数据计算能力，性能相对于深算一号实现了翻倍的增长。深算三号研发进展顺利。

图：海光信息在研项目一览（2023年半年报）

序号	项目名称	预计总投资规模	本期投入金额	累计投入金额	进展或阶段性成果	拟达到目标	技术水平	具体应用前景
1	海光处理器产品实现技术	不适用	77,006,539.31	261,187,613.30	开发阶段	海光处理器设计	国内领先	海光通用处理器
2	新一代海光协处理器芯片设计	不适用	184,945,792.08	413,233,723.23	开发阶段	新一代海光协处理器芯片	国内领先	海光协处理器
3	新一代海光协处理器产品工程技术	不适用	101,829,098.73	206,437,128.13	开发阶段	新一代海光协处理器芯片	国内领先	海光协处理器
4	海光处理器工艺实现技术	不适用	29,563,109.25	40,129,918.96	开发阶段	海光处理器设计	国内领先	海光通用处理器
5	新一代海光通用处理器芯片设计	不适用	431,807,338.62	1,428,353,693.18	开发阶段	新一代海光通用处理器芯片	国内领先	海光通用处理器
6	海光处理器新工艺设计	不适用	109,129,385.09	190,105,953.58	开发阶段	海光处理器设计	国内领先	海光通用处理器
7	2023年处理器关键技术研发项目	不适用	203,166,159.04	220,549,016.88	开发阶段	海光处理器设计	国内领先	海光通用处理器
合计	/		1,137,447,422.12	2,759,997,047.26	/	/	/	/

# 龙芯中科：采用自主指令系统LoongArch

- ◆ 公司主营业务为处理器及配套芯片的研制、销售及服务，主要产品与服务包括处理器及配套芯片产品与基础软硬件解决方案业务。目前，龙芯中科基于信息系统和工控系统两条主线开展产业生态建设，面向网络安全、办公与业务信息化、工控及物联网等领域与合作伙伴保持全面的市场合作，系列产品在电子政务、能源、交通、金融、电信、教育等行业领域已获得广泛应用。龙芯中科研制的芯片包括龙芯 1 号、龙芯 2 号、龙芯 3 号三大系列处理器芯片及桥片等配套芯片。
- ◆ 2020年，通过十余年的自主研发和市场化运作，在处理器研发基础软件研发、生态体系建设等方面已具备充足的技术和经验积累的条件下，龙芯中科推出了自主指令系统LoongArch（龙芯架构）。

图：龙芯中科部分芯片产品一览

型号	简介	主要应用场景
龙芯 1C101	面向智能门锁等应用的 MCU 芯片，集成 Flash、TSENSOR、VPWM、ADC 等功能模块	门锁应用等
龙芯 1C102	面向智能家居以及其他物联网设备，采用龙芯 LA132 处理器核心，集成 Flash、SPI、UART、I2C、RTC、TSENSOR、VPWM、ADC、GPIO 等功能模块	智能门锁类产品、电动助力车、跑步机等
龙芯 1C103	集成 Flash、ATIM、GTIM、ADC、SPI、I2C、UART、RTC 等功能模块，可输出带有死区的互补 PWM 信号，具备驱动舵机、有刷电机、无刷电机的原生支持，同时具备常见的通讯模块	高性价比的常见电机应用，如筋膜枪、修枝机、电锯等
龙芯 2K0500	64 位单核 SoC 芯片，主频 500MHz，集成 DDR3、2D GPU、DVO、PCIE2.0、SATA2.0、USB2.0、USB3.0、GMAC、PCI、彩色黑白打印接口、HDA 及其他常用接口	工控互联网应用、打印终端、BMC 等
龙芯 2K1000LA	64 位双核 SoC 芯片，主频 1.0GHz，基于 LA264 处理器核，集成 DDR2/3、PCIE2.0、SATA2.0、USB2.0、DVO 等接口	交换机、边缘网关、工业防火墙、工业平板、智能变电站、挂号自助机等
龙芯 2K1500	64 位双核 SoC 芯片，主频 1.0GHz，基于 LA264 处理器核，集成 DDR3、PCIE3.0、SATA3.0、USB2.0 接口，提供数量丰富的 SPI、CAN、I2C、PWM 等小接口，支持 eMMC 功能	低功耗场景下的工控需求

型号	简介	主要应用场景
龙芯 3A5000	64 位四核处理器，主频 2.3-2.5GHz，片上集成 4 个 LA464 处理器核，集成双通道 DDR4-3200 和 HT3.0 接口	桌面与终端类应用
龙芯 3C5000L	64 位十六核处理器，主频 2.0-2.2GHz，集成四个 3A5000 硅片，集成四通道 DDR4-3200 和 HT3.0 接口，最高支持四路互联	服务器类应用
龙芯 3C5000	64 位十六核处理器，主频 2.0-2.2GHz，片上集成 16 个高性能 LA464 处理器核，集成四通道 DDR4-3200 和 HT3.0 接口，最高支持十六路互联	服务器类应用
龙芯 3D5000	64 位三十二核处理器，主频 2.0GHz，集成两个 3C5000 硅片，集成八通道 DDR4-3200 和 HT3.0 接口，最高支持四路互联	服务器类应用
龙芯 7A1000	龙芯 3 号系列处理器的配套桥片，通过 HT3.0 接口与处理器相连，外围接口包括 PCIE2.0、GMAC、SATA2.0、USB2.0 和其他低速接口	与龙芯 3 号系列配套使用
龙芯 7A2000	第二代龙芯 3 号系列处理器配套桥片，通过 HT3.0 接口与处理器相连。外围接口包括 PCIE3.0、USB3.0、SATA3.0；显示接口为 2 路 HDMI 和 1 路 VGA，可直连显示器；内置一个网络 PHY，直接提供网络端口输出；片内集成自研 3D GPU，采用统一渲染架构，搭配 32 位 DDR4 显存接口，最大支持 16GB 显存容量	与龙芯 3 号系列配套使用

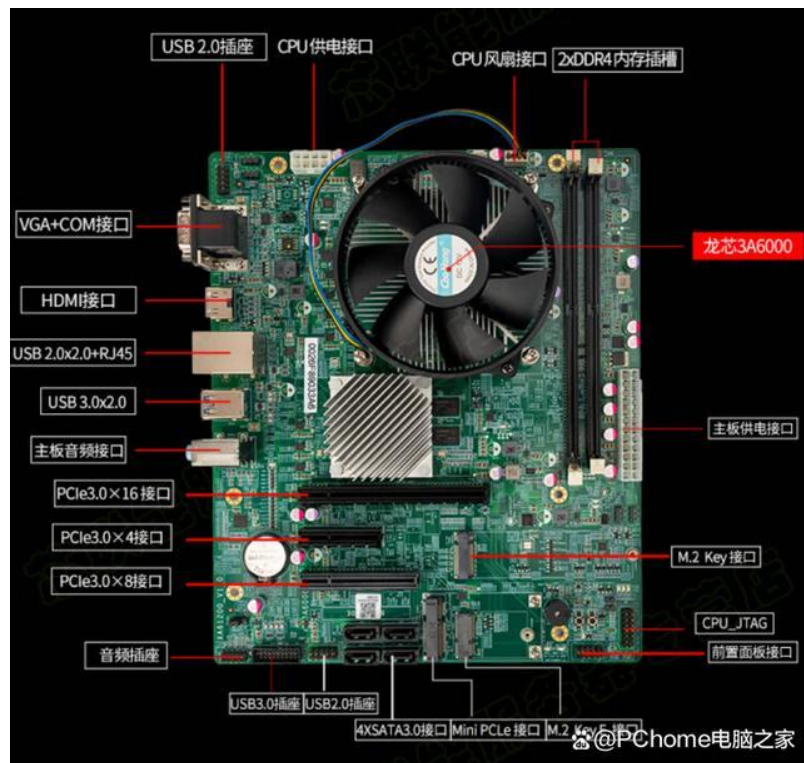
# 龙芯中科：发布3A6000和龙芯2P0500

- ◆ 2023年11月28日，公司发布新一代通用处理器龙芯3A6000、打印机主控芯片龙芯2P0500重磅成果并对外公布龙芯处理器核IP及龙芯自主指令系统架构授权计划。龙芯3A6000处理器采用龙芯自主指令系统龙架构（LoongArch），是龙芯第四代微架构的首款产品，主频达到2.5GHz，集成4个最新研发的高性能LA664处理器核，支持同时多线程技术（SMT2），全芯片共8个逻辑核。根据中国电子技术标准化研究院赛西实验室测试结果，龙芯3A6000处理器总体性能与Intel公司2020年上市的第10代酷睿四核处理器相当。龙芯2P0500是一款适用于单/多功能打印机的主控SOC芯片，单芯片即可满足打印、扫描、复印等多种典型应用需求。

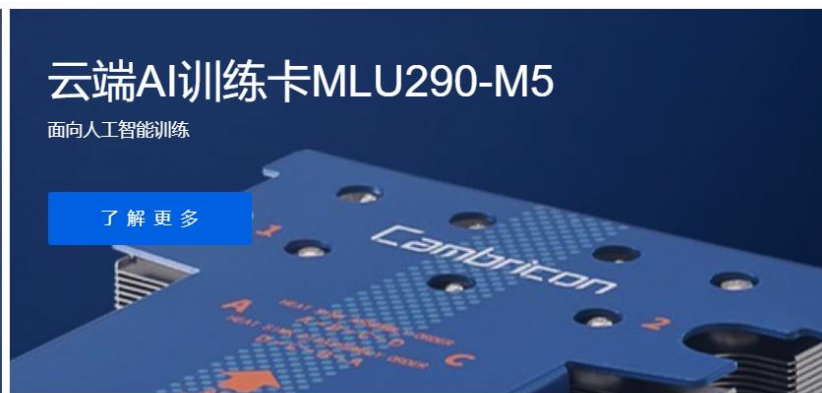
图：龙芯中科3A6000性能参数

主频	2.0GHz - 2.5GHz
峰值运算速度	240GFlops@2.5GHz
物理核个数	4
逻辑核个数	8
处理器核	64位超标量处理器核 LA664 支持 LoongArch®指令集 支持 128/256 位向量指令 六发射乱序执行 4个定点单元、4个向量单元和4个访存单元
高速缓存	每个核包含 64KB 私有二级指令缓存和 64KB 私有二级数据缓存 每个核包含 256KB 私有三级缓存 共 16MB 三级缓存
内存接口	2个 72位 DDR4-3200 支持 ECC 校验
高速 I/O	1个 HyperTransport 3.0 IO 接口 (HT0)
其它 I/O	1个 SPI、1个 UART、2个 I2C、1个 AVS、16个 GPIO 接口
封装方式	FCBGA1190
功耗管理	支持主要模块时钟动态关闭 支持主要时钟域动态变频 支持主电压域动态调压
典型功耗	38W@2.5GHz

图：龙芯中科3A6000板卡一览



- ◆ 公司自成立以来一直专注于人工智能芯片产品的研发与技术创新，致力于打造人工智能领域的核心处理器芯片，让机器更好地理解和服务人类。公司的主营业务是各类云服务器、边缘计算设备、终端设备中人工智能核心芯片的研发、设计和销售，主要产品为云端智能芯片及加速卡、训练整机、边缘智能芯片及加速卡、终端智能处理器IP以及上述产品的配套软件开发平台。目前，公司的主要产品线包括云端产品线、边缘产品线、IP授权及软件。



- ◆ 公司是专业的集成电路设计企业，是国内领先的AIoT SoC芯片供应商，经过二十余年发展，公司拥有丰富的产品矩阵和市场布局，在推动芯进百行百业的过程中积累了强大的技术能力和庞大的客户资源，形成了独特的竞争优势，赋能汽车电子、边缘计算、机器视觉、智能家居、教育办公、消费电子、商业金融、工业应用等众多领域的数字化、智能化升级。
- ◆ 公司高性能的应用处理器，如RK3588系列、RK3399系列、RK3288系列、RK3568系列，拥有高性能CPU和GPU内核、强大的多媒体处理能力、丰富的外设接口，充分满足众多复杂场景应用的需求。

## RK3588

### 典型应用方向

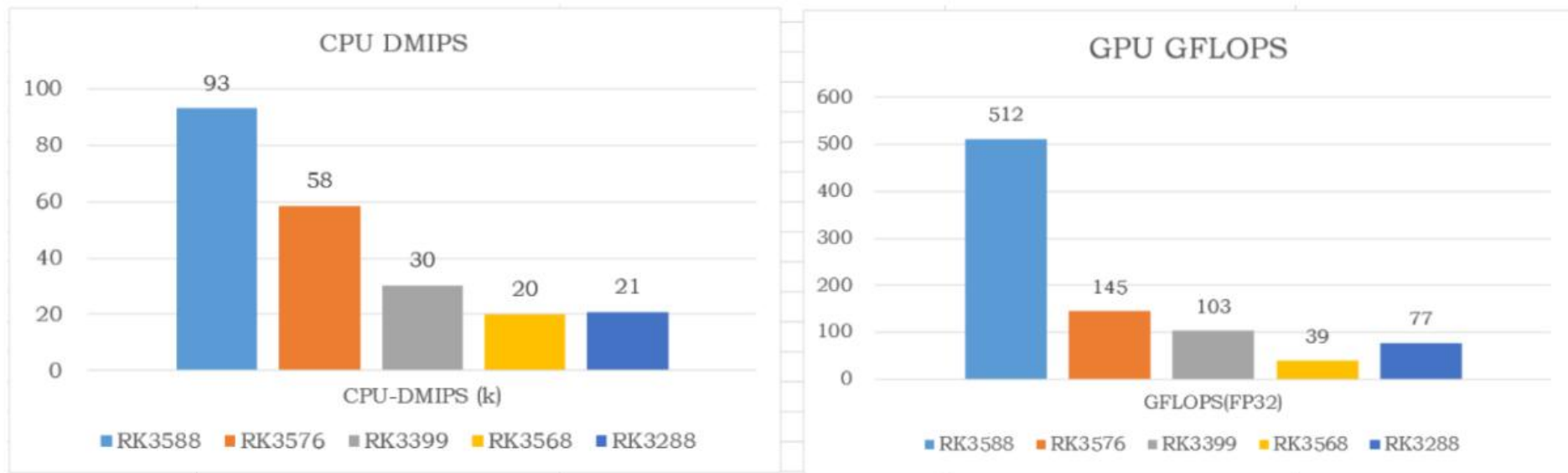
### 主要特性

- 8nm先进制程，8核64位架构，高性能，低功耗
- ARM Mali-G610 MC4 GPU, 专用2D图形加速模块
- 6TOPs NPU，赋能各类AI场景
- 8K 视频编解码，8K显示输出
- 内置多种显示接口，支持多屏异显
- 超强影像处理能力，48MP ISP, 支持多摄像头输入
- 丰富的高速接口(PCIe, TYPE-C, SATA, 千兆以太网)，易于扩展
- Android 和Linux OS



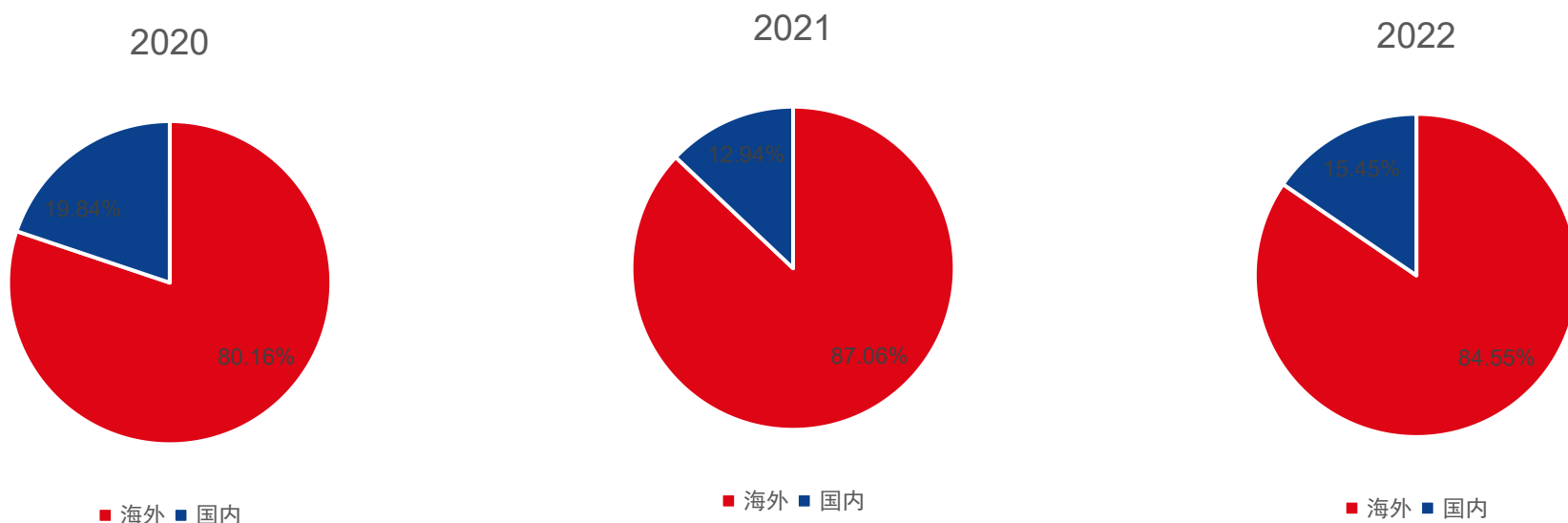
- ◆ 根据scensmart的文章显示，公司即将发布RK3576八核芯片，专为AIoT设计，可用于平板电脑、AIoT应用程序、电子墨水显示器、ARM PC和汽车电子中，集成独立的6TOPS NPU，支持4K视频编解码，性能定位于RK3588和RK3399之间。
- ◆ 对于算力需求的演进，我们认为泛在算力发展大势所趋，从云端发展到边端侧趋势明确。2024年将是电子业以人工智能为主的大变动的一年。公司作为国内边端侧算力的优秀厂商，伴随着AIoT SoC芯片、AI专用芯片等产品线的突破，有望迎来新一轮业绩成长周期。

图：瑞芯微几大芯片性能对比



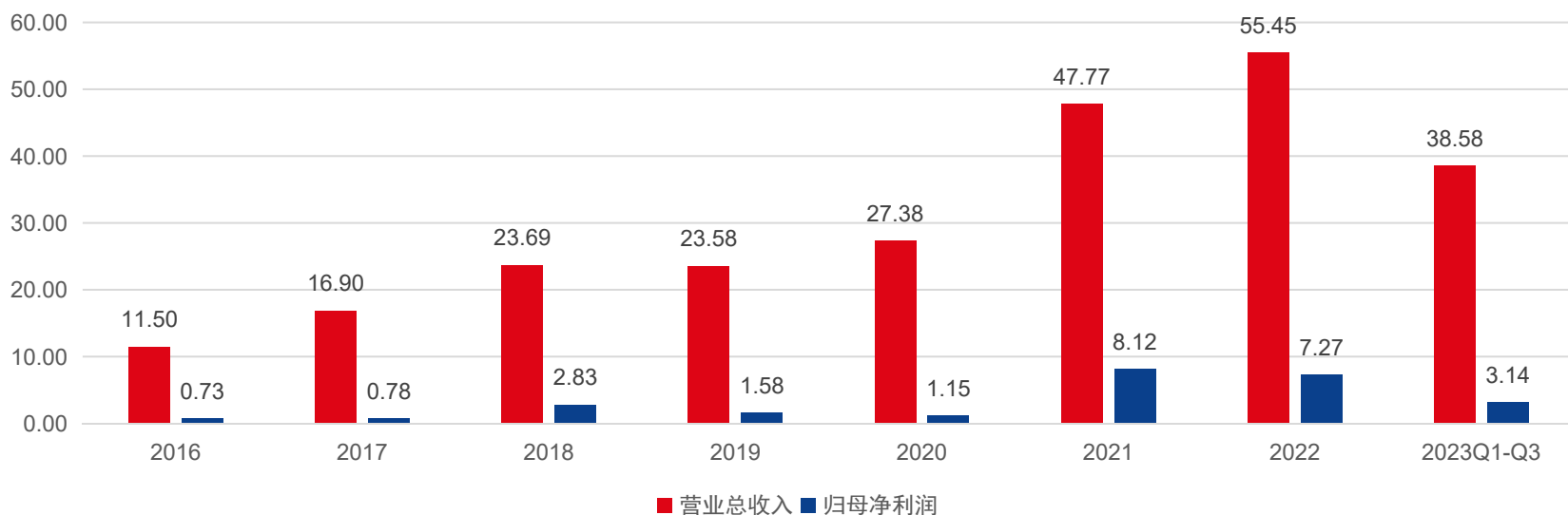
- ◆ 公司是全球布局、国内领先的无晶圆半导体系统设计厂商，主营业务为系统级SoC芯片及周边芯片的研发、设计与销售，目前主要产品有多媒体智能终端SoC芯片、无线连接芯片、汽车电子芯片等，为众多消费类电子领域提供SoC主控芯片和系统级解决方案。公司产品已广泛应用于家庭、汽车、办公、教育、体育健身、工业、商业、农业、娱乐、仓储等领域。
- ◆ 公司业务已覆盖中国大陆、香港、北美、欧洲、拉丁美洲、亚太、非洲等全球主要经济区域，2020-2022年，公司国外销售收入占比分别为80.16%、87.06%、84.55%。

图：晶晨股份海外销售占比



- ◆ 公司有五大产品线：S系列SoC芯片主要有全高清系列芯片和超高清系列芯片，已广泛应用于IPTV机顶盒、OTT机顶盒、混合模式机顶盒及其他智能终端领域；T系列SoC芯片是智能显示终端的核心关键部件，公司T系列芯片已广泛应用于智能电视、智能投影仪、智慧商显、智能会议系统等领域；A系列SoC芯片广泛应用于众多消费类电子领域；W系列芯片为自主研发的高速数传Wi-Fi蓝牙二合一集成芯片，可应用于高吞吐视频传输；汽车电子芯片目前有车载信息娱乐系统芯片和智能座舱芯片。
- ◆ 公司的T系列SoC芯片解决方案已广泛应用于小米、海尔、TCL、创维、海信、长虹、联想、腾讯、Maxhub、Seewo（希沃）、百思买、亚马逊、Epson、Sky等境内外知名企业及运营商的智能终端产品。Wi-Fi产品线的销售收入逐季提升，新产品Wi-Fi62×2也于2023年8月中旬正式商用出货，后续随着Wi-Fi产品进一步与公司主控SoC平台适配并配套销售，Wi-Fi产品线销量将进一步提升。

图：晶晨股份过往几年营收及归母净利润（亿元）



◆ 公司目前的主营业务为智能应用处理器SoC、高性能模拟器件和无线互联芯片的研发与设计。主要产品为智能应用处理器SoC、高性能模拟器件和无线互联芯片。公司产品满足消费、工业、车载领域的应用需求，产品广泛适用于智能硬件、智能家电、智能物联网、智能汽车电子、平板电脑、网络机顶盒以及电源模拟器件、无线通信模组等多个产品市场。

◆ 全志耕耘多年，积累诸多设计、验证、制造、测试、应用经验，以40/28/22nm制程工艺为主，可以在该工艺节点下做到较高的性价比。同时，为满足日渐提升的应用需求，需要有更高算力、更低功耗的系统性能、采用更先进更低成本的FinFET制程将是一个很好的选择。

图：全志科技在研项目（2023年半年报）

主要研发项目名称	项目目的	项目进展	拟达到的目标	预计对公司未来发展的影响
8K视频解码AI SoC	在先进工艺下研发支持AI加速、8K视频解码的芯片	研发阶段	实现通用算力、AI算力的多核异构；支持8K级视频解码；支持多路大屏显示；工业车规级质量设计。	支撑公司在工业控制、流媒体播放器、大屏显示、智能汽车电子领域的发展。
先进工艺高可靠性车规设计	可通过车规认证	研发阶段	研发先进工艺下的高可靠性车规设计。	支持公司智能汽车电子、工业控制领域产品研发需求。
超低功耗无线语音AIOT SoC	基于语音AI加速的智能无线AIOT芯片	试量产阶段	实现异构架构；内置低功耗语音AI处理器，支持低功耗语音唤醒；高集成设计，内置电源器件和WiFi无线通讯。	支撑公司在智能家电语音识别、智能音箱领域的发展。
高性能通用处理SoC平台	带AI加速的通用处理器	量产阶段	采用先进工艺，集成高性能AI模块，支持高性能ISP图像，支持多种通用接口，支持视频编解码和显示输出。	支撑公司在智能家居、智能扫地机、智能汽车电子、智能教育领域的发展。
高级安全SoC系统设计	可防护物理攻击的安全架构	研发阶段	基于通用处理器架构，设计相关IP和SoC系统架构，实现安全启动、固件防修改、安全TEE环境、防物理旁路攻击等安全防护功能。	支撑公司在工业控制、智能车载系统、流媒体播放器领域的发展。
先进工艺高性能车规SoC	面向车载领域的高性能AI应用处理器芯片	研发阶段	采用先进工艺、高可靠性芯片设计，内置高性能处理器和高性能AI模块，支持多种通用接口，支持视频编解码和显示输出。	支撑公司在智能汽车电子的高性能AI领域、智能驾驶领域的发展。
先进工艺高性能通用SoC平台	研发先进工艺的通用智能处理平台芯片	研发阶段	采用先进工艺，内置高性能处理器和高性能AI模块，支持多种显示接口和通用接口，支持视频编解码和显示输出。	支撑公司在智能家居、智能教育领域的发展。
智慧屏幕芯片	研发用于智慧显示领域的芯片	研发阶段	基于ARM处理器平台，集成高画质显示引擎，支持主流显示接口，支持多种视频解码格式和视频标准。	支撑公司在智能电视、智能投影、智能显示领域的发展。
智能机器人芯片	研发用于智能机器人领域的芯片	研发阶段	基于多核异构架构，内置通用处理器、高性能AI模块和实时处理器，支持丰富的控制接口。	支撑公司在智能机器人、工业机器人领域的发展。

- ◆ 公司自设立以来，始终以蜂窝基带技术为核心，专注于各类无线通信芯片的研发设计和技术创新。作为国内知名的蜂窝基带芯片提供商，公司产品已经广泛应用于各类消费电子和工业控制领域。此外，公司也是国内少数兼具提供大型芯片定制业务和半导体IP授权业务的企业。
- ◆ 公司已经实现从2G到5G的蜂窝基带技术累积，构建起高效、完整的基带芯片研发、技术体系，在信号处理、高性能模拟/射频电路、通信协议栈、低功耗电路设计等多个方面拥有了大量的自研IP，且已经具备了WiFi、蓝牙、LoRa、全球导航定位等多协议的无线通讯芯片设计能力。

类别	系列	功能	应用场景
蜂窝基带芯片	基带通信芯片	支持2G、3G、4G、5G通信标准下多种网络制式的通信	适用于车联网、智能支付、工业物联网、智慧安防、智能电网等各种场景
	移动智能终端芯片	支持2G、3G、4G通信标准下多种网络制式的通信。集成了语音通话、视频、拍照等多媒体功能	适用于手机、智能可穿戴设备、智能支付、智能家居等场景
非蜂窝物联网芯片	低功耗LoRa系统芯片	支持LoRa网络制式下的通信，拥有较长的通信距离及低功耗的优点	适用于智能表计、工业物联网、智慧安防等场景
	高集成度WiFi芯片	可作为智能物联网设备的主控芯片或仅提供数据网络连接的功能芯片 单芯片同时实现WiFi及蓝牙芯片通信功能，实现了更高的集成度	适用于智慧安防、智能家居、智能可穿戴设备等场景
	高集成低功耗蓝牙芯片	高度集成射频收发器、蓝牙信号处理、MCU、电源管理一体化	适用于智能可穿戴设备、智能家居等场景
	全球导航定位芯片	可与北斗导航、GPS、Glonass、Galileo四种卫星定位系统进行通信定位，覆盖了目前世界上所有的卫星定位系统	适用于智能可穿戴设备、车联网、工业物联网、手机等场景

◆ 公司是业内较早采用RISC-V指令集架构作为技术开发路线的芯片设计企业，基于开源的RISC-V指令集架构，配合开源实时操作系统RT-Thread，自主开发出高性能CPU内核和DSP指令，实现了各种音频算法。采用RISC-V指令集架构一方面不需要高昂的授权费，另一方面是指令集相对精简，而公司基于RISC-V指令集架构自研CPU内核更是可以使得公司不断优化芯片设计进而持续提升产品的性价比优势，在中低端市场构建起了较高的护城河。公司持续升级现有芯片产品，通过技术的迭代和制程工艺的提升，不断提升芯片性能、综合性价比优势和市场竞争力。

图：公司部分产品线芯片型号一览

图：“蓝讯讯龙”三代产品

## 蓝讯低延时音频芯片型号



### 无线麦克风、无线话筒

 BTS-3, 40nm Flash: 4Mbit 封装: QFN32 延时 < 20ms 功能: 一拖一, 一拖二无线麦克风	 BTS-3, 40nm Flash: 4Mbit 封装: QFN20 延时 < 20ms 功能: 一拖一, 一拖二无线话筒
--	--

 SOP16封装 MCU, MP3音箱, 扩音器, 解码板, usb mic, spi flash音频播放器, iis转usb芯片, 录音笔, 儿童玩具	 QFN32封装 MCU, MP3音箱, 扩音器, 解码板, usb mic, spi flash音频播放器, iis转usb芯片, 录音笔, 儿童玩具
---	---

 SSOP24封装 A: 无软件开机功能 B: 有软件开机功能
---

## MCU

 SSOP24封装 通用MCU应用	 QFN32封装 通用MCU应用
---	--

## 蓝讯IoT芯片型号

### BLE

 SOP16封装 BLE Only 灯带市场	 QFN20封装 BLE Only 灯带市场/通用无线遥控
 BLE Only Flash: 2Mbit 封装: QFN20	 BLE Only Flash: 2Mbit 封装: SOP16

## 蓝讯数字音频芯片型号

### 标准方案

 QFN20封装 Type-C耳机 lightning耳机 转接头 96K高采样率	 QFN20封装 Type-C耳机 lightning耳机 转接头
--	--



- ◆ 公司主营业务为智能音视频SoC芯片的研发、设计与销售，为客户提供AIoT场景下具有语音交互能力的边缘智能主控平台芯片。公司主要产品为蓝牙音频芯片、智能手表芯片和智能家居芯片，并基于公司在无线连接领域的技术积累，逐步延伸至Wi-Fi/BT连接芯片。公司智能音视频SoC芯片能够集成多核CPU、Wi-Fi/BT基带和射频、声学 and 音频系统、电源管理、存储、嵌入式AI处理器和2.5D GPU等多个功能模块，是智能音视频设备的主控平台芯片。
- ◆ 公司产品已经进入三星、OPPO、小米、荣耀、华为、vivo等全球主流安卓手机品牌，同时也进入包括安克创新、哈曼、漫步者、万魔等专业音频厂商，并在阿里、百度、谷歌等互联网公司的智能音频产品中得到应用。品牌客户的深度及广度是公司重要的竞争优势和商业壁垒。

图：公司核心技术



超低功耗射频技术



主动降噪和通话降噪



具有IBRT专利的TWS

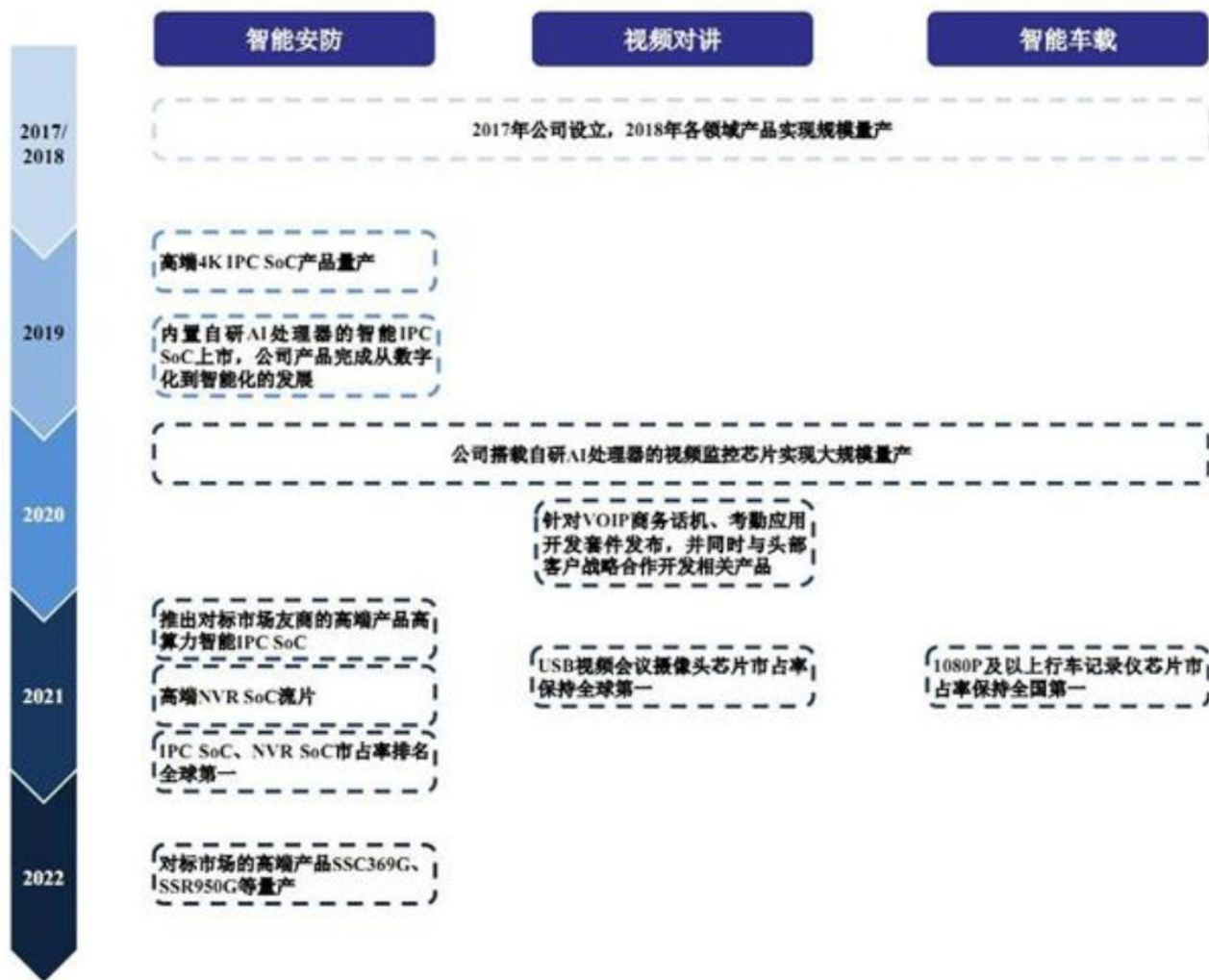


语音AI技术

◆ 公司为全球领先的视频监控芯片企业，主营业务为视频监控芯片的研发及销售，产品主要应用于智能安防、视频对讲、智能车载等领域。公司在芯片设计全流程具有丰富经验，可支撑大型先进工艺下的SoC设计。公司自研全套AI技术，包含AI处理器指令集、AI处理器IP及其编译器、仿真器等全套AI处理器工具链。公司拥有大量核心IP资源，包含图像IP、视频IP、高速模拟IP和音频IP等。

◆ 公司在视频监控领域持续研发创新，在图像信号处理、音视频编解码、显示处理等领域具有领先优势，并积极投入AI等新领域的芯片研发。公司拥有ISP技术、AI处理器技术、多模视频编码技术、高速高精度模拟电路技术、先进制程SoC芯片设计技术等多项核心技术

图：公司产品线核心技术的发展及新产品推出情况



- ◆ 公司的主要产品为蓝牙音频SoC芯片系列、便携式音视频SoC芯片系列、端侧AI处理器芯片系列等，广泛应用于蓝牙音箱、无线家庭影院、智能手表、无线麦克风、无线收发dongle、蓝牙耳机、无线电竞耳机、蓝牙语音遥控器及低功耗端侧AI处理器等领域。
- ◆ 顺应人工智能的发展大势，从高端音频芯片入手，公司将在产品线整合低功耗AI加速引擎，逐步全面升级为CPU、DSP加NPU（神经网络处理器）的三核异构AI计算架构，以打造低功耗端侧AI算力。

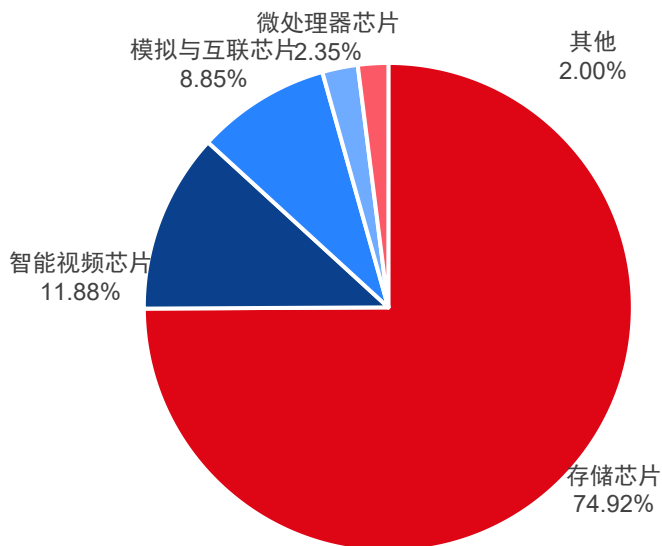
图：公司部分产品线一览

应用示例	主要产品系列	主要应用领域	部分终端品牌
	ATS281X 系列、ATS282X 系列、ATS283X 系列(除ATS2837)、ATS285X 系列、ATS283XP 系列、ATS303X 系列	普通蓝牙音箱(含TWS 音箱)、智能蓝牙音箱、家庭影院的无线音响系统、蓝牙车载产品、K歌麦克风、无线麦克风、无线收发dongle	华为、哈曼、SONY、OPPO、罗技、安克创新、沃尔玛、小米、荣耀、天猫精灵、漫步者、不见不散、唱吧、现代、绿联、Vizio、LG、海信、TCL、Razer、Polk、RODE、科大讯飞、猛犸、枫笛等
	ATS301X 系列、ATS302X 系列、ATS303X 系列	TWS 耳机、颈挂式耳机、头戴式耳机、开放式耳机、无线电竞耳机	荣耀、realme、传音、JBL、倍思、TOZO、QCY、西伯利亚等
	ATS308X 系列	智能手环/手表	Noise、realme、boAt、Fire-Boltt、小米等

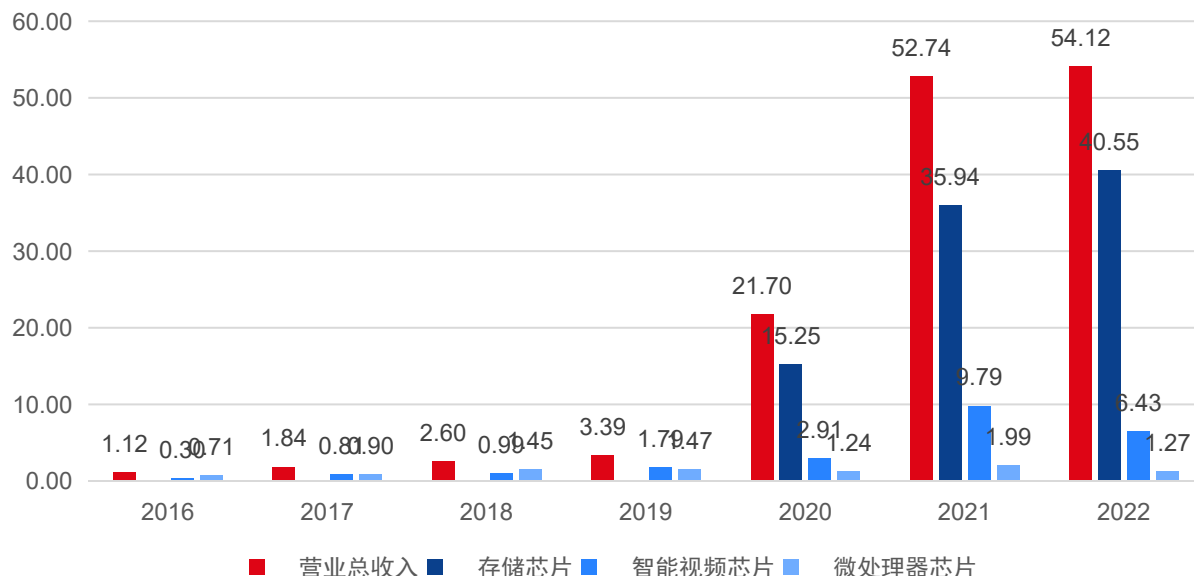
应用示例	主要产品系列	主要应用领域	部分终端品牌
	ATS360X 系列	智能办公类产品(如会议音箱)、智能家居和家电话音交互模组	eMeet、音络、向往、飞利浦等
	ATS2837	智能录音笔	科大讯飞、飞利浦、汉王等
	ATB110X 系列、ATB111X 系列	蓝牙语音遥控器、语音鼠标、语音键盘、翻译棒及其它数据传输类产品	罗技、小米、FetchTV、创维、长虹、Caixun、Vestel、当贝、极米、峰米等

- ◆ 公司芯片产品所属领域涵盖了处理器、存储器和模拟电路等，具体产品类别分为微处理器芯片、智能视频芯片、存储芯片、模拟芯片和互联芯片，其中微处理器芯片和智能视频芯片的现有产品均采用了MIPS架构，同时，随着RISC-V架构的发展，公司也在积极布局RISC-V相关技术的研发，公司自研的RISC-V CPU核已应用于公司部分芯片产品中。
- ◆ 从市场应用上，公司的芯片产品主要面向汽车、工业与医疗、通讯和消费等几大市场领域，其中微处理器产品线主要应用于生物识别、二维码识别、商业设备、智能家居、智能穿戴、教育电子及其他物联网相关领域，智能视频产品线主要应用于安防监控、智能门铃、人脸识别设备等智能视觉相关领域。

图：2022年公司营业收入构成



图：2016-2022年公司营业收入及几块主要业务收入（亿元）



01

逻辑芯片市场规模巨大，产业链具备高门槛

02

大模型走向多模态，技术持续演进，商业前景广阔

03

算力板块投资的几条主线

04

相关标的

05

风险提示

**技术创新风险：**随着下游市场对产品的性能需求不断提升，集成电路设计行业技术升级和产品更新换代速度较快，企业需紧跟市场发展步伐，及时对现有产品及技术进行升级换代，以维持其市场地位。同时，集成电路产品的发展方向具有一定不确定性，因此企业需要对主流技术迭代趋势保持较高的敏感度，根据市场需求变动和工艺水平发展制定动态的技术发展战略。未来若公司技术研发水平落后于行业升级换代水平，或公司技术研发方向与市场发展趋势偏离，将导致公司研发资源浪费并错失市场发展机会，对公司产生不利影响。

**宏观经济和行业波动风险：**集成电路行业是面临全球化的竞争与合作并得到国家政策大力支持的行业，受到国内外宏观经济、行业法规和贸易政策等宏观环境因素的影响。近年来，全球宏观经济表现平稳，国内经济稳中有升。未来，如果国内外宏观环境因素发生不利变化，可能会对公司经营带来不利影响。

**国际贸易摩擦风险：**伴随全球产业格局的深度调整，国际贸易摩擦不断，集成电路产业成为贸易冲突的重点领域，也对中国相关产业的发展造成了客观不利影响。2022年8月以来，美国推出了多项贸易管制政策通过限制产品、设备以及技术等项目的出口以限制中国半导体行业的发展。

**复苏或需求不及预期相关风险：**当下全球局势比较复杂多变，如果国内外宏观环境等诸多因素导致复苏或需求不及预期，进而带来产业链需求不及预期的风险。

- 1、芯时代之一\_半导体重磅深度《新兴技术共振进口替代，迎来全产业链投资机会》
- 2、芯时代之二\_深度纪要《国产芯投资机会暨权威专家电话会》
- 3、芯时代之三\_深度纪要《半导体分析和投资策略电话会》
- 4、芯时代之四\_市场首篇模拟IC深度《下游应用增量不断，模拟 IC加速发展》
- 5、芯时代之五\_存储器深度《存储产业链战略升级，开启国产替代“芯”篇章》
- 6、芯时代之六\_功率半导体深度《功率半导体处黄金赛道，迎进口替代良机》
- 7、芯时代之七\_半导体材料深度《铸行业发展基石，迎进口替代契机》
- 8、芯时代之八\_深度纪要《功率半导体重磅专家交流电话会》
- 9、芯时代之九\_半导体设备深度《进口替代促景气度提升，设备长期发展明朗》
- 10、芯时代之十\_3D/新器件《先进封装和新器件，续写集成电路新篇章》
- 11、芯时代之十一\_IC载板和SLP《IC载板及SLP，集成提升的板级贡献》
- 12、芯时代之十二\_智能处理器《人工智能助力，国产芯有望“换”道超车》
- 13、芯时代之十三\_封测《先进封装大势所趋，国家战略助推成长》
- 14、芯时代之十四\_大硅片《供需缺口持续，国产化蓄势待发》
- 15、芯时代之十五\_化合物《下一代半导体材料，5G助力市场成长》
- 16、芯时代之十六\_制造《国产替代加速，拉动全产业链发展》
- 17、芯时代之十七\_北方华创《双结构化持建机遇，由大做强倍显张力》
- 18、芯时代之十八\_斯达半导《铸IGBT功率基石，创多领域市场契机》
- 19、芯时代之十九\_功率半导体深度②《产业链逐步成熟，功率器件迎黄金发展期》
- 20、芯时代之二十\_汇顶科技《光电传感创新领跑，多维布局引领未来》
- 21、芯时代之二十一\_华润微《功率半导专芯致志，特色工艺术业专攻》
- 22、芯时代之二十二\_大硅片\*重磅深度《半导体材料第一蓝海，硅片融合工艺创新》
- 23、芯时代之二十三\_卓胜微《5G赛道射频芯片龙头，国产替代正当时》
- 24、芯时代之二十四\_沪硅产业《硅片“芯”材蓄势待发，商用量产空间广阔》
- 25、芯时代之二十五\_韦尔股份《光电传感稳创领先，系统方案展创宏图》

- 26、芯时代之二十六\_中环股份《半导硅片厚积薄发，特有赛道独树一帜》
- 27、芯时代之二十七\_射频芯片《射频芯片千亿空间，国产替代曙光乍现》
- 28、芯时代之二十八\_中芯国际《代工龙头创领升级，产业联动芯火燎原》
- 29、芯时代之二十九\_寒武纪《AI芯片国内龙头，高研发投入前景可期》
- 30、芯时代之三十\_芯朋微《国产电源IC十年磨一剑，铸就国内升级替代》
- 31、芯时代之三十一\_射频PA《射频PA革新不止，万物互联广袤无限》
- 32、芯时代之三十二\_中微公司《国内半导刻蚀巨头，迈内生&外延平台化》
- 33、芯时代之三十三\_芯原股份《国内IP龙头厂商，推动SiPaaS模式发展》
- 34、芯时代之三十四\_模拟IC深度PPT《模拟IC黄金赛道，本土配套渐入佳境》
- 35、芯时代之三十五\_芯海科技《高精度测量ADC+MCU+AI，切入蓝海赛道超芯星》
- 36、芯时代之三十六\_功率&化合物深度《扩容&替代提速，化合物布局长远》
- 37、芯时代之三十七\_恒玄科技《专注智能音频SoC芯片，迎行业风口快速发展》
- 38、芯时代之三十八\_和而泰《从高端到更高端，芯平台创新格局》
- 39、芯时代之三十九\_家电芯深度PPT《家电芯配套渐完善，增存量机遇筑蓝海》
- 40、芯时代之四十\_前道设备PPT深度《2021年国产前道设备，再迎新黄金时代》
- 41、芯时代之四十一\_力芯微《专注电源管理芯片，内生外延拓展产品线》
- 42、芯时代之四十二\_复旦微电《国产FPGA领先企业，高技术壁垒铸就护城河》
- 43、芯时代之四十三\_显示驱动深度PPT《显示驱动芯一面板国产化最后1公里》
- 44、芯时代之四十四\_艾为电子《数模混合设计专家，持续迭代拓展产品线》
- 45、芯时代之四十五\_紫光国微《特种与安全两翼齐飞，公司步入快速发展阶段》
- 46、芯时代之四十六\_新能源芯\*PPT深度《乘碳中和之风，基础元件腾飞》
- 47、芯时代之四十七\_AIoT \*PPT深度《AIoT大时代，SoC厂商加速发展》
- 48、芯时代之四十八\_铂科新材《双碳助力发展，GPU新应用构建二次成长曲线》
- 49、芯时代之四十九\_AI芯片《AI领强算力时代，GPU启新场景落地》
- 50、芯时代之五十\_江海股份《乘“碳中和”之风，老牌企业三大电容全面发力》

- 51、芯时代之五十一\_智能电动车1000页PPT（多行业协同）《智能电动车★投研大全》
- 52、芯时代之五十二\_瑞芯微PPT深度《迈入全球准一线梯队，新硬件十年前景可期》
- 53、芯时代之五十三\_峰昭科技《专注BLDC电机驱动控制芯片，三大核心技术引领成长》
- 54、芯时代之五十四\_纳芯微《专注高端模拟IC，致力国内领先车规级半导体供应商》
- 55、芯时代之五十五\_晶晨股份《核心技术为躯，全球开拓为翼》
- 56、芯时代之五十六\_国微&复微《紫光国微与复旦微的全面对比分析》
- 57、芯时代之五十七\_国产算力SoC《算力大时代，处理器SoC厂商综合对比》
- 58、芯时代之五十八\_高能模拟芯《高性能模拟替代渐入深水区，工业汽车重点突破》
- 59、芯时代之五十九\_南芯科技《电荷泵翘楚拓矩阵蓝图，通用产品力屡复制成功》
- 60、芯时代之六十\_AI算力GPU《AI产业化再加速，智能大时代已开启》
- 61、芯时代之六十一\_瑞芯微②深度《人工智能再加速，AIoT SoC龙头多点开花》
- 62、芯时代之六十二\_华峰测控《技术/产品为基石，SoC/模数/功率测试机助拓全球市场》
- 63、芯时代之六十三\_裕太微《以太网PHY芯片稀缺标的，国产化渗透初期前景广阔》
- 64、芯时代之六十四\_华虹公司《立足成熟制程，“特色IC+功率器件”代工龙头底部加码12寸》
- 65、芯时代之六十五\_汇顶科技《指纹&触控保持市场领先，新品营收逐步起量》
- 66、芯时代之六十六\_中科蓝讯《产品结构升级&品牌客户突破，八大产品线拓未来》
- 67、芯时代之六十七\_2.5D/3D封装PPT《技术发展引领产业变革，向高密度封装时代迈进》
- 68、芯时代之六十八\_显示驱动芯片PPT《显示驱动芯片——面板国产化最后一公里》
- 69、芯时代之六十九\_菱电电控《双转战略促量价齐升逻辑凸显，T-BOX塑造第二增长极》
- 70、芯时代之七十\_华海清科《国产CMP设备龙头，持续走向高端化、平台化》
- 71、芯时代之七十一\_东芯股份《利基型存储国内领先，强周期属性2024年迎拐点》
- 72、芯时代之七十二\_通富微电《VISIONS技术护城河&AMD深度合作，在AI浪潮中更上层楼》
- 73、芯时代之七十三\_长电科技《深度分析-深度分析-XDFOI™平台为支撑，吹响算力存力汽车三重奏》
- 74、芯时代之七十四\_算力芯《以“芯”助先进算法，以“算”驱万物智能》

孙远峰：华金证券总裁助理&研究所所长&电子行业首席分析师，哈尔滨工业大学工学学士，清华大学工学博士，近3年电子实业工作经验；2018年新财富上榜分析师（第3名），2017年新财富入围/水晶球上榜分析师，2016年新财富上榜分析师（第5名），2013~2015年新财富上榜分析师团队核心成员；多次获得保险资管IAMAC、水晶球、金牛奖等奖项最佳分析师；2019年开始未参加任何个人评比，其骨干团队专注于创新&创业型研究所的一线具体创收&创誉工作，以“产业资源赋能深度研究”为导向，构建研究&销售合伙人队伍，积累了健全的成熟团队自驱机制和年轻团队培养机制，充分获得市场验证；2023年带领崭新团队获得《证券时报》评选的中国证券业最具特色研究君鼎奖和2023年Wind第11届金牌分析师进步最快研究机构奖；清华校友总会电子工程系分会副秘书长；

王臣复：电子行业高级分析师，北京航空航天大学工学学士和管理学硕士，2年半导体产业一级股权投资经历，曾就职于华西证券研究所、欧菲光集团投资部、融通资本、平安基金等，2023年2月加入华金证券研究所；

王海维：电子行业联席首席分析师，华东师范大学硕士，电子&金融复合背景，主要覆盖半导体板块，善于个股深度研究，2018年新财富上榜分析师（第3名）核心成员，先后任职于安信证券/华西证券研究所，2023年2月入职华金证券研究所；

## 公司评级体系

### 收益评级：

- 买入 — 未来6个月的投资收益率领先沪深300指数15%以上；
- 增持 — 未来6个月的投资收益率领先沪深300指数5%至15%；
- 中性 — 未来6个月的投资收益率与沪深300指数的变动幅度相差-5%至5%；
- 减持 — 未来6个月的投资收益率落后沪深300指数5%至15%；
- 卖出 — 未来6个月的投资收益率落后沪深300指数15%以上。

### 风险评级：

- A — 正常风险，未来6个月投资收益率的波动小于等于沪深300指数波动；
- B — 较高风险，未来6个月投资收益率的波动大于沪深300指数波动。

## 行业评级体系

### 收益评级：

领先大市 — 未来6个月的投资收益率领先沪深300指数10%以上；

同步大市 — 未来6个月的投资收益率与沪深300指数的变动幅度相差-10%至10%；

落后大市 — 未来6个月的投资收益率落后沪深300指数10%以上；

### 风险评级：

A — 正常风险，未来6个月投资收益率的波动小于等于沪深300指数波动；

B — 较高风险，未来6个月投资收益率的波动大于沪深300指数波动。

## 分析师声明

孙远峰/王臣复声明，本人具有中国证券业协会授予的证券投资咨询执业资格，勤勉尽责、诚实守信。本人对本报告的内容和观点负责，保证信息来源合法合规、研究方法专业审慎、研究观点独立公正、分析结论具有合理依据，特此声明。

## 本公司具备证券投资咨询业务资格的说明

华金证券股份有限公司（以下简称“本公司”）经中国证券监督管理委员会核准，取得证券投资咨询业务许可。本公司及其投资咨询人员可以为证券投资人或客户提供证券投资分析、预测或者建议等直接或间接的有偿咨询服务。发布证券研究报告，是证券投资咨询业务的一种基本形式，本公司可以对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向本公司的客户发布。

## 免责声明：

本报告仅供华金证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因为任何机构或个人接收到本报告而视其为本公司的当然客户。

本报告基于已公开的资料或信息撰写，但本公司不保证该等信息及资料的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映本公司于本报告发布当日的判断，本报告中的证券或投资标的价格、价值及投资带来的收入可能会波动。在不同时期，本公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，本公司将随时补充、更新和修订有关信息及资料，但不保证及时公开发布。同时，本公司有权对本报告所含信息在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以本公司向客户发布的本报告完整版本为准。

在法律许可的情况下，本公司及所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务，提请客户充分注意。客户不应将本报告为作出其投资决策的惟一参考因素，亦不应认为本报告可以取代客户自身的投资判断与决策。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，无论是否已经明示或暗示，本报告不能作为道义的、责任的和法律的依据或者凭证。在任何情况下，本公司亦不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告版权仅为本公司所有，未经事先书面许可，任何机构和个人不得以任何形式翻版、复制、发表、转发、篡改或引用本报告的任何部分。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“华金证券股份有限公司研究所”，且不得对本报告进行任何有悖原意的引用、删节和修改。

华金证券股份有限公司对本声明条款具有惟一修改权和最终解释权。

## 风险提示:

报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价或询价。投资者对其投资行为负完全责任，我公司及其雇员对使用本报告及其内容所引发的任何直接或间接损失概不负责。

华金证券股份有限公司

办公地址:

上海市浦东新区杨高南路759号陆家嘴世纪金融广场30层

北京市朝阳区建国路108号横琴人寿大厦17层

深圳市福田区益田路6001号太平金融大厦10楼05单元

电话: 021-20655588

网址: [www.huajinsec.com](http://www.huajinsec.com)