

AI算力研究框架

——AI系列专题研究报告

刘熹(证券分析师)
S0350523040001
liux10@ghzq.com.cn

◆ GPT4：AI技术和工程的伟大创新，开启科技“十年新周期”

GPT-4 是世界首个最接近AGI的先进AI系统，展现出强大的“涌现能力”。GPT的成功，得益于其参数规模扩大，RLHF、Transformer、Prompt、插件、系统工程等方面的伟大创新。我们预计，ChatGPT将对科技产业产生深远的影响，类似于操作系统，ChatGPT将接入现有的全部软硬件系统。GPT-4的诞生将加速AGI时代的到来，开启科技“十年新周期”。

◆ AI算力：科技企业大模型竞赛的核心“装备”，AIGC应用的关键基建

Transformer架构大模型训练对算力的消耗呈指数级增长。2023年1月，ChatGPT计划再向微软融资100亿美金，该融资将是新一代大模型算力基建的主要资金来源。ChatGPT激发“鲶鱼效应”，全球科技巨头将AI战略提升到空前高度，算力作为新一轮科技竞赛的核心“装备”，迎来需求的脉冲式增长。未来，ChatGPT应用的全面落地还将释放更为广阔的算力需求。

◆ 计算是AI算力的核心引擎，存储、网络、软件是AI算力的主要发展方向

1) **计算**：GPU是ChatGPT训练和推理的核心支柱，其更新速度远超过“摩尔定律”，受益于AI和高性能市场需求增长，GPU行业景气度显著提升。AI服务器作为GPU的重要载体，预计其市场规模、渗透率将随着GPU放量迎来同步高增。

2) **网络**：已成为限制AI算力提升的主要瓶颈，英伟达推出InfiniBand架构下的NVLink、NVSwitch等方案，将GPU之间的通信能力上升到新高度。而800G、1.6T高端光模块作为AI训练的上游核心器件，将受益于大模型训练需求的增长。

3) **存储**：“内存墙”是制约算力提升的重要因素。NAND、DRAM等核心存储器在制程方面临近极限，不断探索“3D”等多维解决方案。HBM基于其高宽带特性，成为了高性能GPU的核心组件，市场前景广阔。

◆ 投资建议

ChatGPT对算力的影响远不止当前可见的基建投入，未来Transformer大模型的迭代推动模型训练相关需求的算力增长，以及AIGC大模型应用的算力需求，将是算力市场不断超预期的源泉。相关公司：

1、计算

- 1) 服务器：浪潮信息、中科曙光、紫光股份、工业富联、纬创、广达、英业达、戴尔、联想集团、超威电脑、中国长城、神州数码、拓维信息、四川长虹；
- 2) GPU：英伟达、AMD、Intel、海光信息、寒武纪、龙芯中科、景嘉微；

2、网络

- 1) 网络设备：紫光股份、中兴通讯、星网锐捷、深信服、迪普科技、普天科技、映翰通；
- 2) 光模块：中际旭创、新易盛、光迅科技、华工科技、联特科技、剑桥科技、天孚通信；

3、存储

- 1) 存储器：紫光国微、江波龙、北京君正、兆易创新、澜起科技、东芯股份、聚辰股份、普冉股份、朗科科技。

◆ 风险提示：宏观经济影响下游需求，大模型迭代不及预期，大模型应用不及预期，市场竞争加剧，中美博弈加剧，相关公司业绩不及预期等。

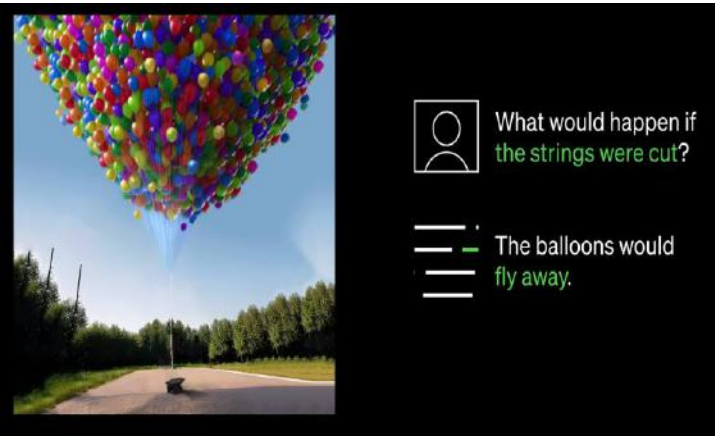
1. GPT4：AI技术和工程的伟大创新，迈向AGI时代
2. AI算力：GPT基座，率先受益于AI“十年新周期”
3. 算力—计算：GPU为算力核心，服务器为重要载体
4. 算力—网络：数据中心算力瓶颈，光模块需求放量
5. 算力—存储：AI训练“内存墙”，3D工艺持续突破
6. 投资建议和风险提示

1. GPT4：AI技术和工程的伟大创新，迈向AGI时代

1.1 GPT4：全球领先的“智能涌现” AI大模型

- GPT-4 是世界第一个最接近AGI的先进AI系统。Generative Pre-trained Transformer 4 (GPT-4) 是OpenAI创建的多模态大型语言模型，于 2023 年 3 月 14 日发布，并已通过ChatGPT Plus和商业API形式对外提供服务。
- ChatGPT是OpenAI在2022年11月推出的基于 GPT-3.5的新型 AI聊天机器人，只需向ChatGPT提出需求，即可实现文章创作、代码创作、回答问题等功能。ChatGPT从推出到月用户过亿仅用了2个月时间，是世界上增速最快消费级应用。

GPT4：全球最为领先的AI应用之一



1.1 GPT4：全球领先的“智能涌现” AI大模型

- 2018年以来，GPT系列模型的参数规模、训练数量规模曾指数级增长，算法和框架持续升级。
- 预计GPT4的参数规模接近1万亿，数据量20万亿 Tokens。

	GPT-1	BERT(开源)	GPT-2 (开源)	GPT-3	ChatGPT/GPT-3.5	GPT4 (多模态)
发布时间	2018	2018	2019	2020	2022年11月30日	2023年3月14日
模型技术结构特点与安全性措施	- 基于Transformer架构 - 12层，768维词嵌入向量 - 无监督预训练和有监督微调	- 基于Transformer架构，12层 - 双向编码器(全文本) - 词嵌入向量维度为768或1024使用掩码语言模型(MLM)进行预训练 - 下游任务微调	- 从GPT-1基础上增加到48层，使用1600维向量进行词嵌入 - 修改初始化的残差层权重，缩放为原来的1/N (N为残差层的数量) - 交替密度和局部带状稀疏注意模式	- 96层，参数数量增加到约1750亿 (175B) - 上下文窗口宽度增加到2048个tokens - 训练数据更大 - 采用交替密度和局部带状稀疏注意模式	- 基于GPT-3的架构，进行了针对性的优化和调整 - 参数数量、层数和词表与GPT-3相近，但具体参数可能有所不同 - RLHF训练	- 预期参数数量远大于1750亿 - 视觉语言模型组件(VLM) - 对抗性测试和红队测试 - RLHF训练 - 基于规则的奖励模型(RBRM)
模型参数	117M	340M	1.5B	1758B	大概175B-6B-1.3B	未知
上下文窗口	512 token	512 token	1024 token	2048 token	4096 token	32,000 token
训练方法	无监督预训练和有监督微调	掩码语言模型预训练	无监督预训练	无监督预训练	1) 有监督微调 2) RLHF训练奖励模型 3) PPO强化学习	1) 有监督微调 2) RLHF训练奖励模型 3) 构造基于规则的奖励模型(RBRM) 4) PPO强化学习
数据集	BooksCorpus	BooksCorpus与English Wikipedia	WebText	Common Crawl, WebText2, Books1, Books2和Wikipedia	类似GPT-3,但可能有更新	更大规模和多样化
数据量规模	5G		40GB	45TB	45T+X	20万亿Tokens

1.1 GPT4：全球领先的“智能涌现” AI大模型

- GPT4的显著特征“涌现能力”， LLM的涌现能力被正式定义为“在小型模型中不存在，但在大型模型中出现的能力”。
- 涌现能力出现时的一个显著特点：当模型规模达到一定程度时，性能显著提升。这种涌现模式与物理学中的相变现象有着密切的联系。原则上，涌现能力也可以根据一些复杂任务来定义。
- 涌现是非线性深度网络的基本特征，也是群体智能行为与复杂思维，感知与认知的基本特质。

GPT4：“智能涌现” 特征描述	
涌现特征	特征 描述
In-context learning (情景学习)	上下文学习能力由GPT-3正式引入：假设语言模型已经提供了一个自然语言指令和/或几个任务演示，它可以通过完成输入文本的单词序列为测试实例生成预期的输出，而不需要额外的训练或梯度更新
Instruction following (指令跟随)	通过对经自然语言描述(即指令)格式化的多任务数据集的混合进行微调，LLM在同样以指令形式描述的未见任务上表现良好。有了这种能力，指令调优使LLM能够通过理解任务指令来执行新任务，而无需使用明确示例，这可以在很大程度上提高泛化能力
Step-by-step reasoning (逻辑推理)	对于小型语言模型，通常很难解决涉及多个推理步骤的复杂任务，例如数学应用题。而在思维链推理策略下，LLMs可以利用包含中间推理步骤以获得最终答案的提示机制来解决此类任务。据推测，这种能力可能是通过对代码的训练获得。

1.1 GPT4：全球领先的“智能涌现”AI大模型

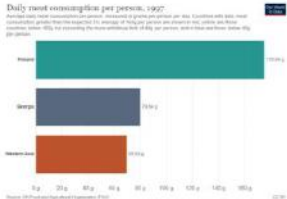
GPT-4 在以下六个方面实现飞跃式提升。

- **创造能力更强：** GPT-4 比以往任何时候都更具创造性和协作性。它可以生成、编辑并与用户一起迭代创意和技术写作任务，例如创作歌曲、编写剧本或学习用户的写作风格。
- **推理能力更强：** GPT-4在高级推理能力方面超过ChatGPT。
- **更长的上下文：** GPT-4能够处理超过25,000个单词的文本，允许使用诸如长表单内容创建、扩展对话、文档搜索和分析等用例。
- **图片识别能力：** GPT-4可以接受图像作为输入，并生成标题、分类和分析。
- **专业知识更强：** 大多数专业和学术考试中，GPT-4表现出了人类水平的表现。
- **更安全、更一致：** 与GPT-3.5相比，GPT-4对禁用内容的请求做出响应的可能性要低82%，产生事实性响应的可能性要高40%。

GPT4：专业考试、逻辑推理、多模态能力突出

GPT-4 visual input example, Chart Reasoning:

User: What is the sum of average daily meat consumption for Georgia and Western Asia? Provide a step-by-step reasoning before providing your answer.



GPT-4: To find the sum of average daily meat consumption for Georgia and Western Asia, we will follow these steps:

1. Identify the average daily meat consumption for Georgia.
2. Identify the average daily meat consumption for Western Asia.
3. Add the values from steps 1 and 2.

Step 1: From the provided chart, we see that the average daily meat consumption for Georgia is 79.84 grams per person per day.

Step 2: From the chart, we see that the average daily meat consumption for Western Asia is 69.62 grams per person per day.

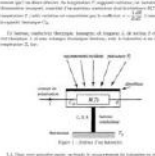
Step 3: Add the values from steps 1 and 2.

79.84 g (Georgia) + 69.62 g (Western Asia) = 149.46 g

So, the sum of average daily meat consumption for Georgia and Western Asia is 149.46 grams per person per day.

GPT-4 visual input example, Ecote Polytechnique Exam Question:

User: Answer question 1.1a. Think step-by-step.



GPT-4: To answer question 1.1a, we need to express the temperature $T(x)$ in every point of the conductive bar, which is characterized by the abscissa x . We are given that $T(0) = T_0$ and $T(L) = T_b$.

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:

$$\frac{d^2 T}{dx^2} = 0$$

Integrating once, we get:

$$\frac{dT}{dx} = C_1$$

Integrating a second time:

$$T(x) = C_1 x + C_2$$

To determine the constants C_1 and C_2 , we use the boundary conditions:

$$T(0) = T_0 \Rightarrow T_0 = C_1 \cdot 0 + C_2 \Rightarrow C_2 = T_0$$

$$T(L) = T_b \Rightarrow T_b = C_1 \cdot L + T_0 \Rightarrow C_1 = \frac{T_b - T_0}{L}$$

Finally, the expression for $T(x)$ is:

$$T(x) = \left(\frac{T_b - T_0}{L} \right) x + T_0$$

This represents a linear temperature profile along the conductive bar, with a slope of $(T_b - T_0)/L$.

Example of GPT-4 visual input:

User: What is funny about this image? Describe it panel by panel.



GPT-4: The image shows a package for a "Lightning Cable" adapter with three panels.

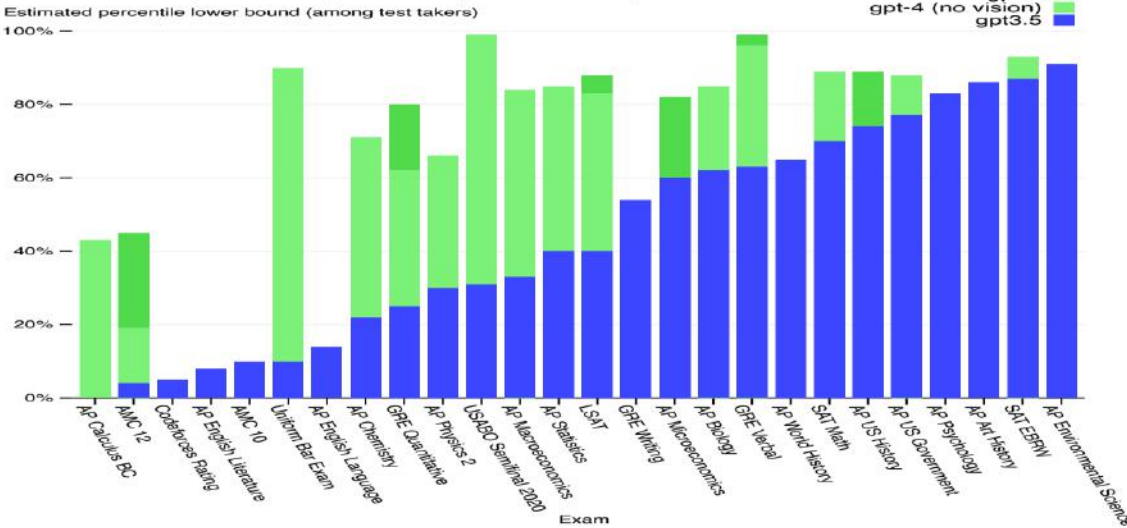
Panel 1: A smartphone with a VGA connector to large, blue, 15-pin connector typically used for computer monitors plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

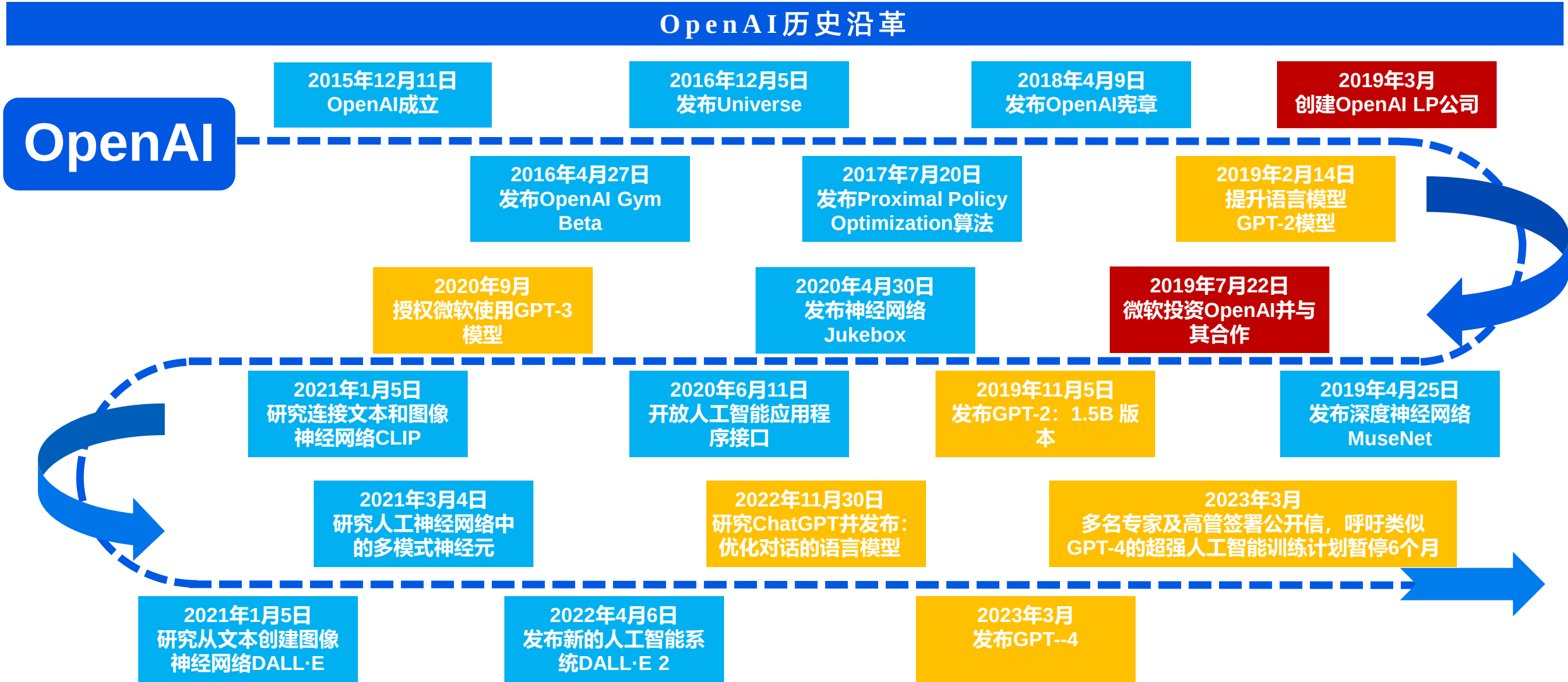
The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

Exam results (ordered by GPT-3.5 performance)

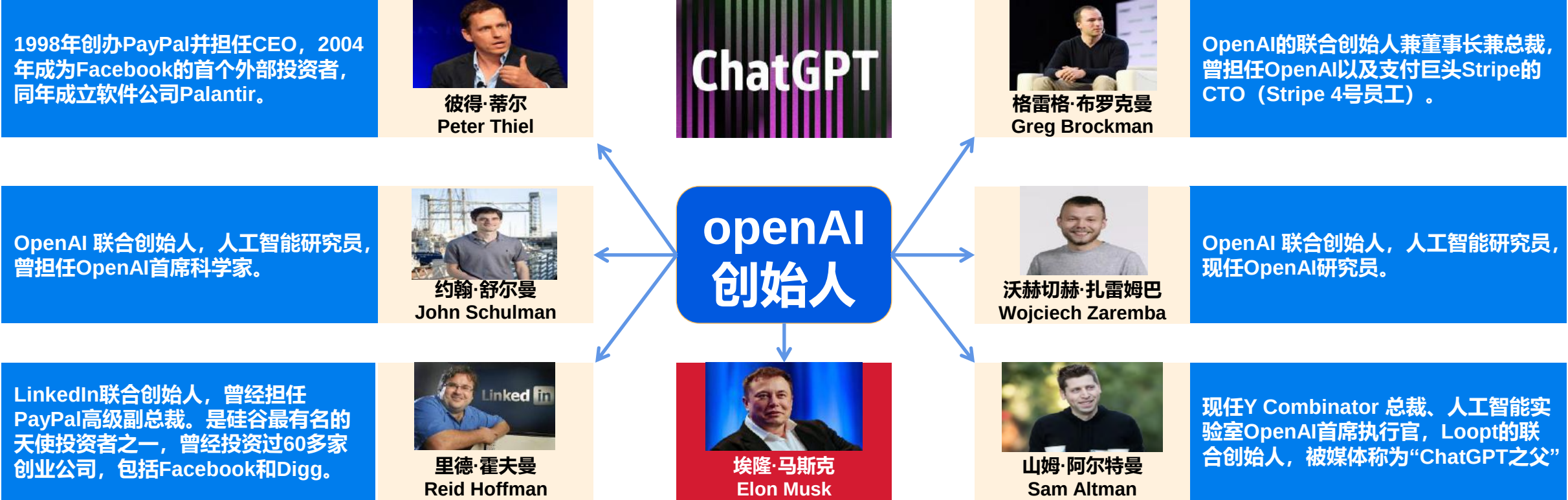


1.2 OpenAI：AGI的重要推手

OpenAI，在美国成立的人工智能研究公司，核心宗旨在于“实现安全的通用人工智能(AGI)”，使其有益于人类。



1.2 OpenAI: AGI的重要推手



1.2 OpenAI： AGI的重要推手

- GPT-4 幕后的研发团队大致可分为七个部分：预训练（Pretraining）、长上下文（Long context）、视觉（Vision）、强化学习 & 对齐（RL & alignment）、评估 & 分析（Evaluation & analysis）、部署（Deployment），以及其他。

OpenAI团队主要分工			
主要团队	主要团队任务		
预训练	●计算机集群扩展 ●硬件正确性	●数据 ●优化 & 架构	●分布式训练基础设施 ●Training run babysitting
长上下文	●长上下文研究	●长上下文内核	
视觉	●架构研究 ●硬件正确性 ●Training run babysitting	●计算机集群扩展 ●数据 ●部署 & 后训练	●分布式训练基础设施 ●对齐数据
强化学习 & 对齐	●数据集贡献 ●模型安全 ●Flagship training runs	●数据基础设施 ●Refusals ●代码功能	●ChatML 格式 ●基础 RLHF 和 InstructGPT 工作
评估 & 分析	●OpenAI Evals 库 ●ChatGPT评估 ●真实世界用例评估 ●新功能评估	●加速预测 ●能力评估 ●污染调查	●模型等级评估基础设施 ●编码评估 ●指令遵循和 API 评估
部署	●核心贡献者 ●GPT4网络体验 ●信托与安全工程	●推理研究 ●推理基础设施 ●产品管理	●GPT-4 API和ChatML部署 ●可靠性工程



右起三个人分别是：
1）首席科学家伊尔亚·苏茨克维 (Ilya Sutskever)
2）总裁兼董事长格雷格·布罗克曼 (Greg Brockman)
3）CEO萨姆·奥特曼(Sam Altman)
三人皆是2015年OpenAI成立时的创始元老。

1.3 GPT-4六大颠覆式技术创新

1、大参数+大数据+算法创新

2、Transformer：自注意力机制

3、对齐调优：RLHF

4、Prompt：情境学习、思维链

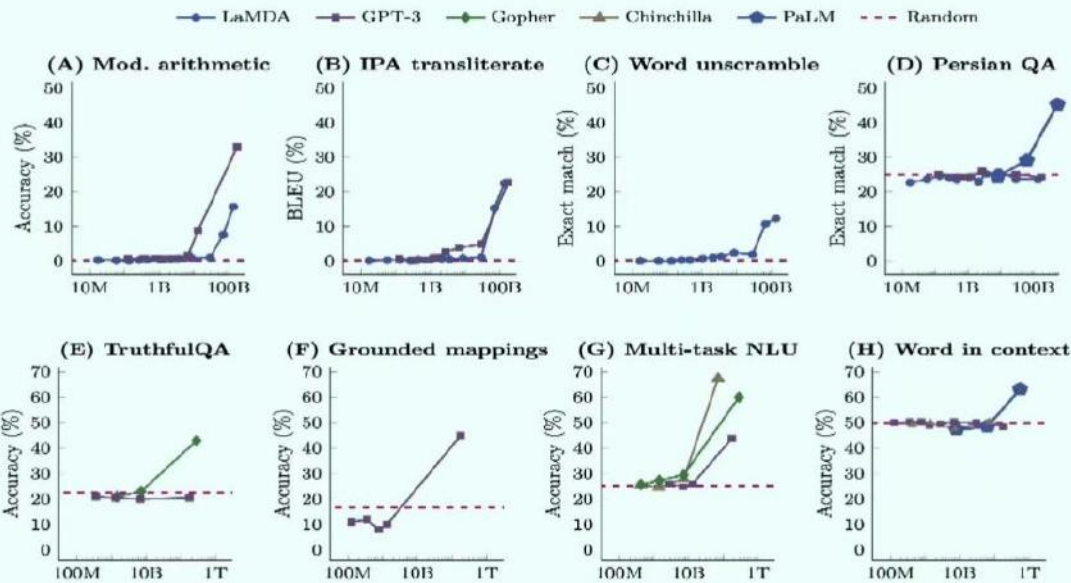
5、工具引入：卡片、互联网

6、工程创新

1.3 GPT4六大颠覆式技术创新（1）：大参数+大数据+算法创新

- 参数扩大是提高LLM模型能力的关键因素。GPT-3首先将模型大小增加到175B参数的极大规模。语言模型前期的性能和模型规模大致呈线性关系，当模型规模大到一定程度时，任务性能有了明显的突变。
- 大规模语言模型基座的可扩展性很强，实现反复自我迭代。因此，LLM也被看作是实现通用人工智能AGI的希望。

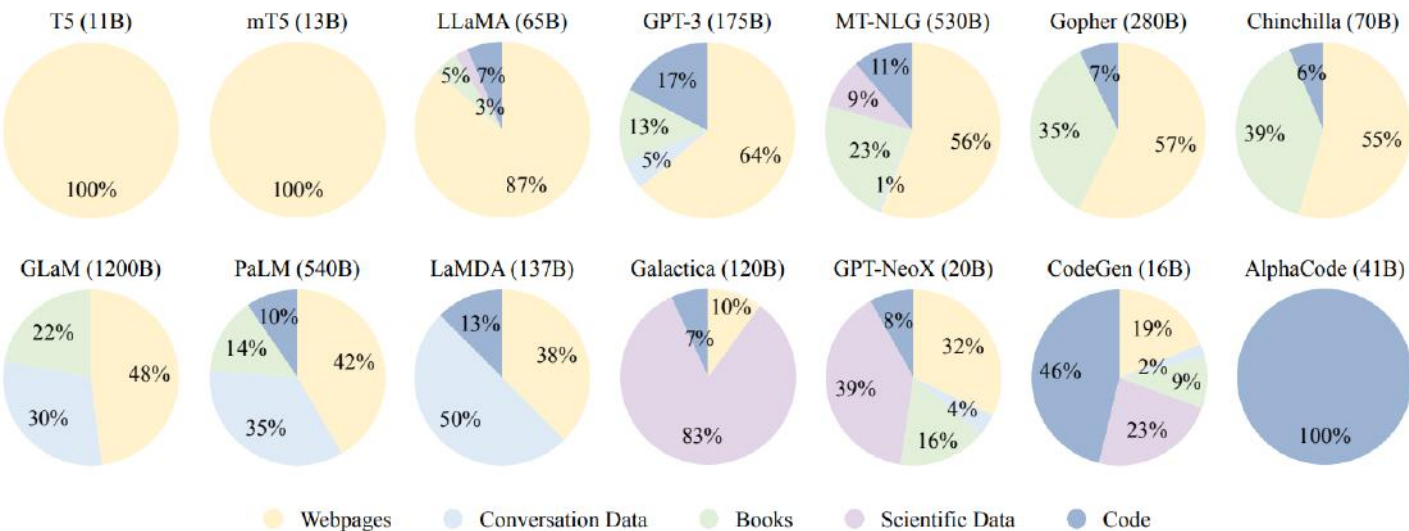
参数对大模型性能起到明显作用



参数对大模型性能起到明显作用

Model	Release Time	Size (B)	Base Model	Adaptation IT	RLHF	Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation ICL	Cot
T5 [71]	Oct-2019	11	-	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
mT5 [72]	Mar-2021	13	-	-	-	1T tokens	Apr-2019	-	-	✓	-
PanGu-α [73]	May-2021	13*	-	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
CPM-2 [74]	May-2021	198	-	-	-	2.6TB	-	-	-	✓	-
T0 [28]	Oct-2021	11	T5	✓	-	-	-	512 TPU v3	27 h	✓	-
GPT-NeoX-20B [75]	Feb-2022	20	-	-	-	825GB	Dec-2022	96 40G A100	-	✓	-
CodeGen [76]	Mar-2022	16	-	-	-	577B tokens	-	-	-	✓	-
Tk-Instruct [77]	Apr-2022	11	T5	✓	-	-	-	256 TPU v3	4 h	✓	-
UL2 [78]	Apr-2022	20	-	✓	-	1T tokens	Apr-2019	512 TPU v4	-	✓	✓
OPT [79]	May-2022	175	-	-	-	180B tokens	-	992 80G A100	-	✓	-
BLOOM [66]	Jul-2022	176	-	-	-	366B	-	384 80G A100	105 d	✓	-
GLM [80]	Aug-2022	130	-	-	-	400B tokens	-	768 40G A100	60 d	✓	-
Flan-T5 [81]	Oct-2022	11	T5	✓	-	-	-	-	-	✓	✓
mT0 [82]	Nov-2022	13	mT5	✓	-	-	-	-	-	✓	-
Galactica [35]	Nov-2022	120	-	-	-	106B tokens	-	-	-	✓	✓
BLOOMZ [82]	Nov-2022	176	BLOOM	✓	-	-	-	-	-	✓	-
OPT-IML [83]	Dec-2022	175	OPT	✓	-	-	-	128 40G A100	-	✓	✓
LLaMA [57]	Feb-2023	65	-	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-
GShard [84]	Jan-2020	600	-	-	-	1T tokens	-	2048 TPU v3	4 d	-	-
GPT-3 [55]	May-2020	175	-	-	-	300B tokens	-	-	-	✓	-
LaMDA [85]	May-2021	137	-	-	-	2.81T tokens	-	1024 TPU v3	57.7 d	-	-
HyperCLOVA [86]	Jun-2021	82	-	-	-	300B tokens	-	1024 A100	13.4 d	✓	-
Codex [87]	Jul-2021	12	GPT-3	-	-	100B tokens	May-2020	-	-	✓	-
ERNIE 3.0 [88]	Jul-2021	10	-	-	-	375B tokens	-	384 V100	-	✓	-
Jurassic-1 [89]	Aug-2021	178	-	-	-	300B tokens	-	800 GPU	-	✓	-
FLAN [62]	Oct-2021	137	LaMDA	✓	-	-	-	128 TPU v3	60 h	✓	-
MT-NLG [90]	Oct-2021	530	-	-	-	270B tokens	-	4480 80G A100	-	✓	-
Yuan 1.0 [91]	Oct-2021	245	-	-	-	180B tokens	-	2128 GPU	-	✓	-
WebGPT [70]	Dec-2021	175	GPT-3	-	✓	-	-	-	-	✓	-
Gopher [59]	Dec-2021	280	-	-	-	300B tokens	-	4096 TPU v3	920 h	✓	-
ERNIE 3.0 Titan [92]	Dec-2021	260	-	-	-	300B tokens	-	2048 V100	28 d	✓	-
GLaM [93]	Dec-2021	1200	-	-	-	280B tokens	-	1024 TPU v4	574 h	✓	-
InstructGPT [61]	Jan-2022	175	GPT-3	✓	✓	-	-	-	-	✓	-
AlphaCode [94]	Feb-2022	41	-	-	-	967B tokens	Jul-2021	-	-	-	-
Chinchilla [34]	Mar-2022	70	-	-	-	1.4T tokens	-	-	-	✓	-
PaLM [56]	Apr-2022	540	-	-	-	780B tokens	-	6144 TPU v4	-	✓	✓
AlexaTM [95]	Aug-2022	20	-	-	-	1.3T tokens	-	128 A100	120 d	✓	✓
Sparrow [96]	Sep-2022	70	-	-	✓	-	-	64 TPU v3	-	✓	-
U-PaLM [97]	Oct-2022	540	PaLM	-	-	-	-	512 TPU v4	5 d	✓	✓
Flan-PaLM [81]	Oct-2022	540	PaLM	✓	-	-	-	512 TPU v4	37 h	✓	✓
Flan-U-PaLM [81]	Oct-2022	540	U-PaLM	✓	-	-	-	-	-	✓	✓
GPT-4 [46]	Mar-2023	-	-	✓	✓	-	-	-	-	✓	✓
PanGu-Σ [98]	Mar-2023	1085	PanGu-α	-	-	329B tokens	-	512 Ascend 910	100 d	✓	-

大模型主要利用各种公共文本数据集做预训练

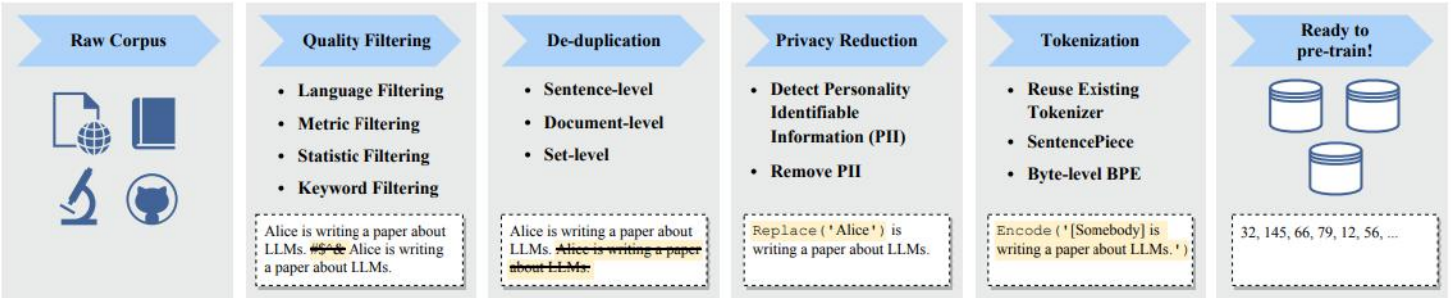


模型能力不仅与模型大小有关，还与数据大小和总计算量有关。同时，预训练数据的质量对取得良好的性能起着关键作用，因此在扩展预训练语料库时，数据收集和清洗策略是非常重要的考虑。

预训练语料库的来源大致可以分为两类：

- **通用数据**：如网页、书籍和对话文本，由于其庞大、多样化和可访问性，被大多数LLM使用，可以增强LLM的语言建模和泛化能力。
- **专业数据**：如多语言数据、科学数据和代码，使LLM具有特定的任务解决能力。

预训练大语言模型典型的数据处理过程



1.3 GPT4六大颠覆式技术创新（1）：大参数+大数据+算法创新



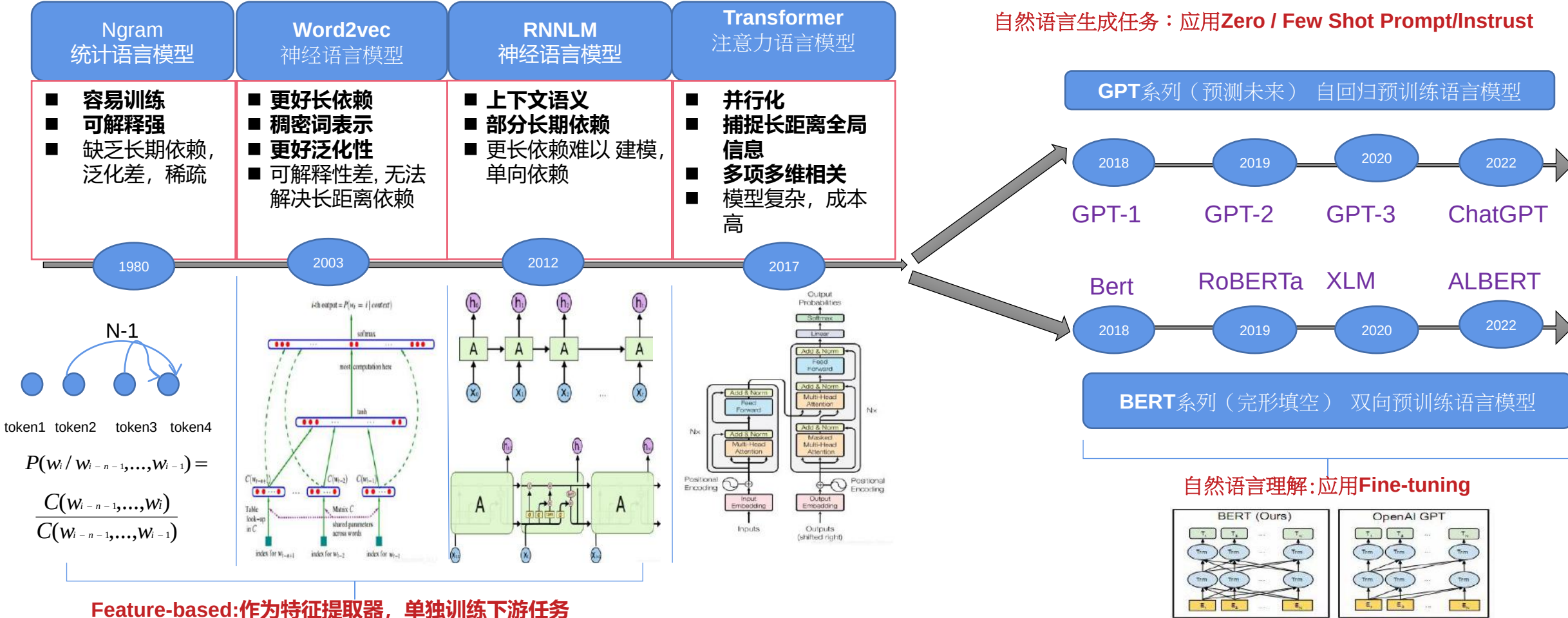
- **并行训练**。由于模型规模巨大，成功训练一个强大的LLM是非常具有挑战性的。LLM的网络参数学习通常需要联合使用多种并行策略，一些优化框架已经发布，以促进并行算法的实现和部署，如Transformer、DeepSpeed和Megatron-LM。此外，优化技巧对训练稳定性和模型性能也很重要。最近，GPT-4提出开发特殊的基础设施和优化方法，用小得多的模型的达到大型模型的性能。
- 目前，常用的训练LLM的库包括Transformers、DeepSpeed、Megatron-LM、JAX、Colossal-AI、BMTrain、FastMoe等。此外，现有的深度学习框架(如PyTorch、TensorFlow、MXNet、PaddlePaddle、MindSpore和OneFlow)也提供了对并行算法的支持。

目前开发大模型主要使用的算法库

名称	具体介绍
Transformers	开源的Python库，用于使用Transformer架构构建模型，由Hugging Face开发和维护。它具有简单和友好的API，使其易于使用和定制各种预训练模型，以及用于数据集处理和评估的工具。它是一个强大的库，拥有一个庞大而活跃的用户和开发人员社区，他们定期更新和改进模型和算法。
DeepSpeed	由微软开发的基于PyTorch的深度学习优化库，已用于训练许多LLM，如GPT-Neo和BLOOM。它为分布式训练提供了各种优化技术，如内存优化(零技术)、梯度检查点和流水线并行。此外，它还提供了微调和评估这些模型的API。
Megatron-LM	基于PyTorch的深度学习库，由NVIDIA开发，用于训练大规模语言模型。它还为分布式训练提供了丰富的优化技术，包括模型和数据并行、混合精度训练、闪光注意力和梯度检查点。这些优化技术可以显著提高训练效率和速度，实现跨gpu和机器的高效分布式训练。
JAX	谷歌Brain开发的用于高性能机器学习的Python库，允许用户在支持硬件加速(GPU或TPU)的数组上轻松执行计算。它支持各种设备上的计算，并提供了一些方便的功能，如即时编译加速和自动批处理。
Colossal-AI	由EleutherAI开发的深度学习库，用于训练大规模语言模型。它建立在JAX之上，并支持训练的优化策略，如混合精度训练和并行。最近，一个名为ColossalChat的类似聊天gpt的模型已经公开发布了两个版本(7B和13B)，它们是使用基于LLaMA的colossalai开发的。
BMTrain	OpenBMB开发的一个高效的分布式训练大规模参数模型的库，它强调代码简单、低资源和高可用性。BMTrain已经将几个常见的LLM(例如Flan-T5和GLM)合并到其ModelCenter中，开发人员可以直接使用这些模型。
FastMoE	专门用于MoE(即，专家混合)模型的训练库。它是在PyTorch之上开发的，在设计中优先考虑效率和用户友好性。FastMoE简化了将Transformer模型转换为MoE模型的过程，并在训练期间同时支持数据并行和模型并行。

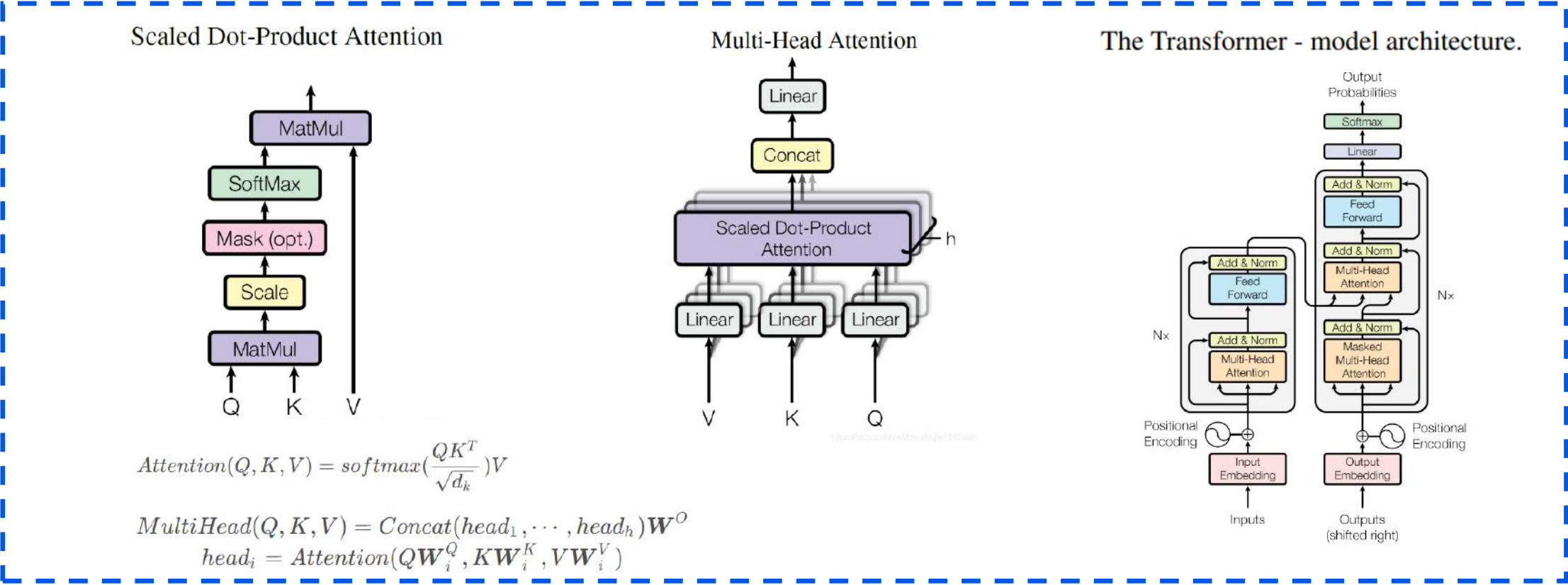
1.3 GPT4六大颠覆式技术创新（2）：Transformer

- Transformer由Google 在2017年的论文 Attention is All you need 中提出，**GPT与BERT均采用了Transformer模型。**
- Transformer基于显著性的注意力机制为输入序列中的任何位置提供上下文信息，使得Transformer具有**全局表征能力强，高度并行性，位置关联操作不受限，通用性强，可扩展性强**等优势，从而使得GPT模型具有优异的表现。



1.3 GPT4六大颠覆式技术创新（2）：Transformer

- **Self-Attention自注意力机制**：当模型处理每个词（输入序列中的每个位置）时，Self-Attention 机制使得模型不仅能够关注当前位置的词，而且能够关注句子中其他位置的词，从而可以更好地编码这个词。即单词自己记住我和哪些单词在同一句话里面。
- Transformer基于自注意力机制，学会单词和单词之间共同出现的概率，在语料输入后，可以输出单词和单词共同出现的概率，同时，Transformer能够挖掘长距离上下文的词之间的双向关系。

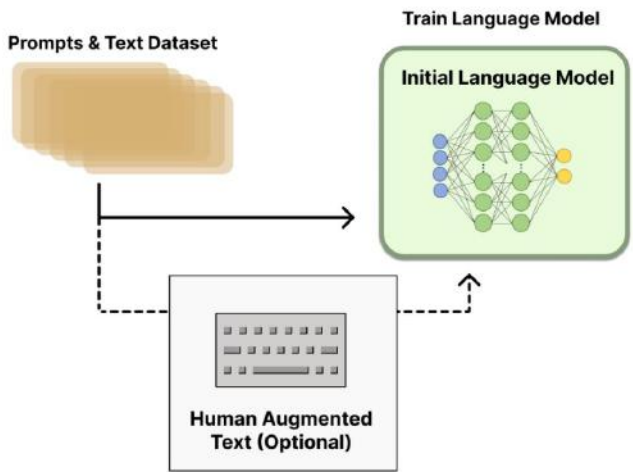


1.3 GPT4六大颠覆式技术创新（3）：RLHF

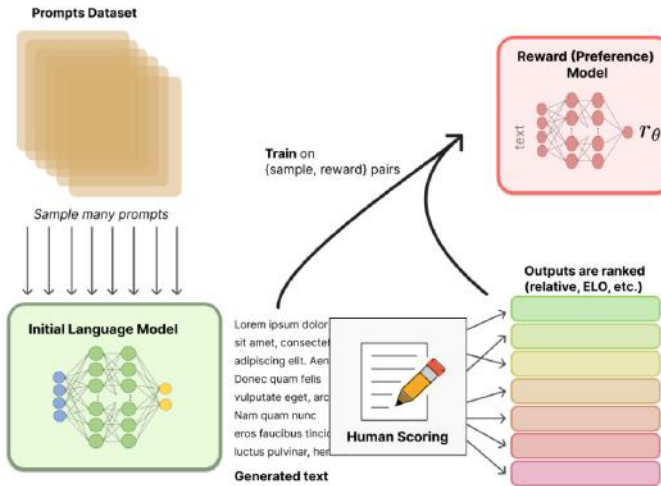
- **RLHF是ChatGPT所采用的关键技术之一。Reinforcement Learning with Human Feedback (RLHF) 是强化学习 (RL) 的一个扩展分支，RLHF 将人类的反馈信息纳入到训练过程，通过使用这些反馈信息构建一个奖励模型神经网络，以此提供奖励信号来帮助 RL 智能体学习，从而更加自然地将人类的需求，偏好，观念等信息以一种交互式的学习方式传达给智能体，对齐 (align) 人类和人工智能之间的优化目标，产生行为方式和人类价值观一致的系统。**

RLHF 训练步骤

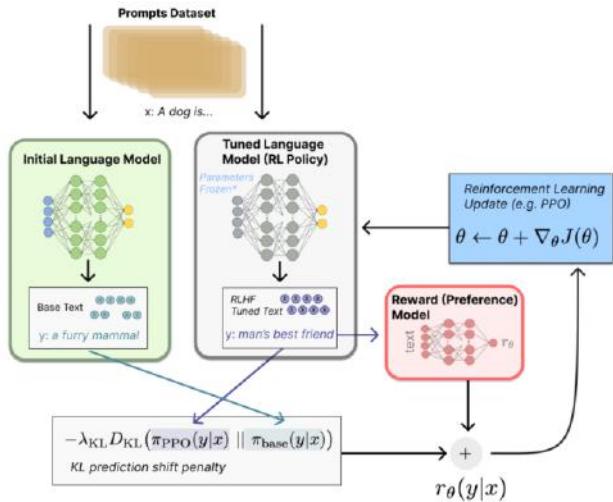
Step 1. 预训练语言模型 (LM)



Step 2. 训练奖励模型 (RM)



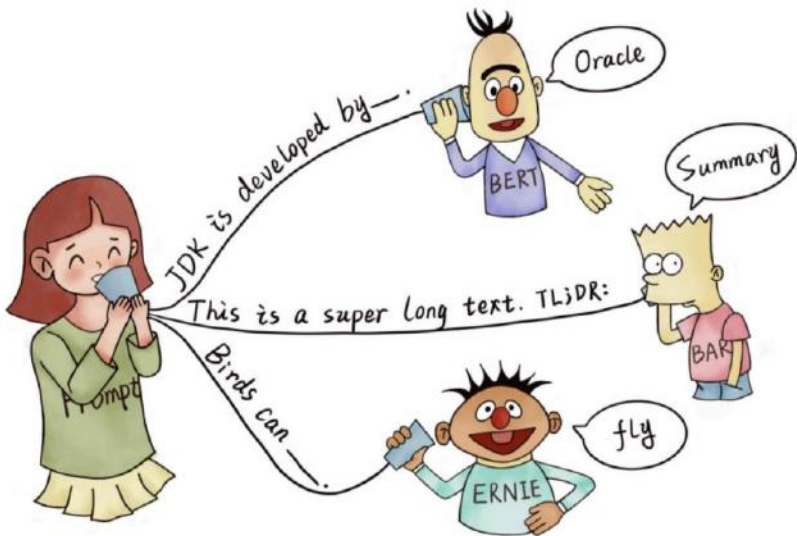
Step 3. 用强化学习 (RL) 微调



1.3 GPT4六大颠覆式技术创新（4）：Prompt

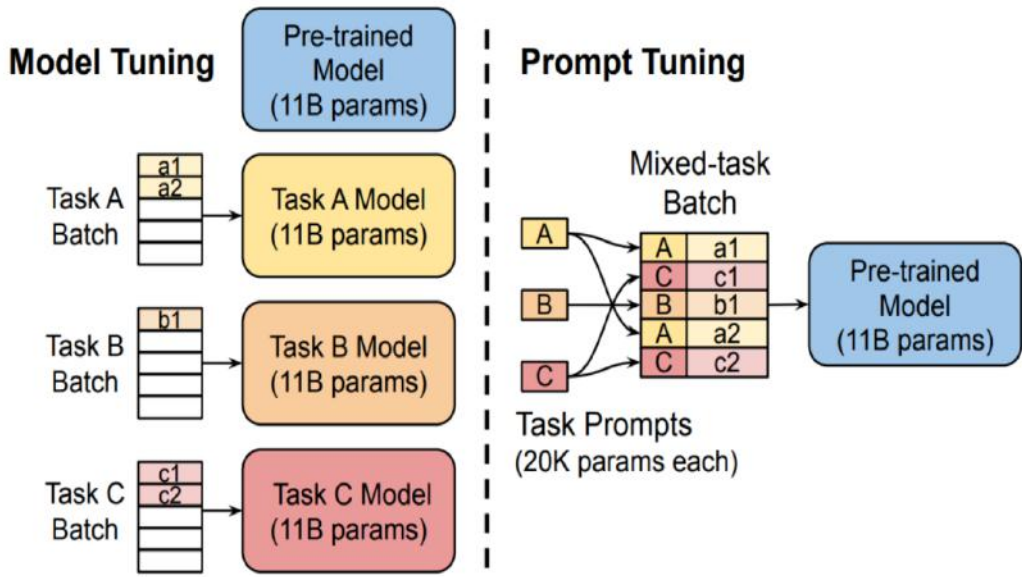
- “提示” 是一种提供给预训练语言模型的线索，让预训练语言模型能更好的理解人类的问题。通过在输入中增加额外的文本（clue/prompt），以更好地利用预训练模型中的知识。
- 提示学习的基本流程主要包括以下四个步骤：提示构造（Prompt Construction），答案构造（Answer Construction），答案预测（Answer Prediction），以及答案-标签映射（Answer-Label Mapping）。
- 提示学习的优势：1）对预训练模LM的利用率高；2）小样本场景训练效果提升；3）fine-tune成本大幅度下降等。

Prompt的案例演示



根据提示，BERT 能回答/补全出“JDK是由 Oracle 研发的”，BART 能对长文本进行总结，ERNIE 能说出鸟类的能力。

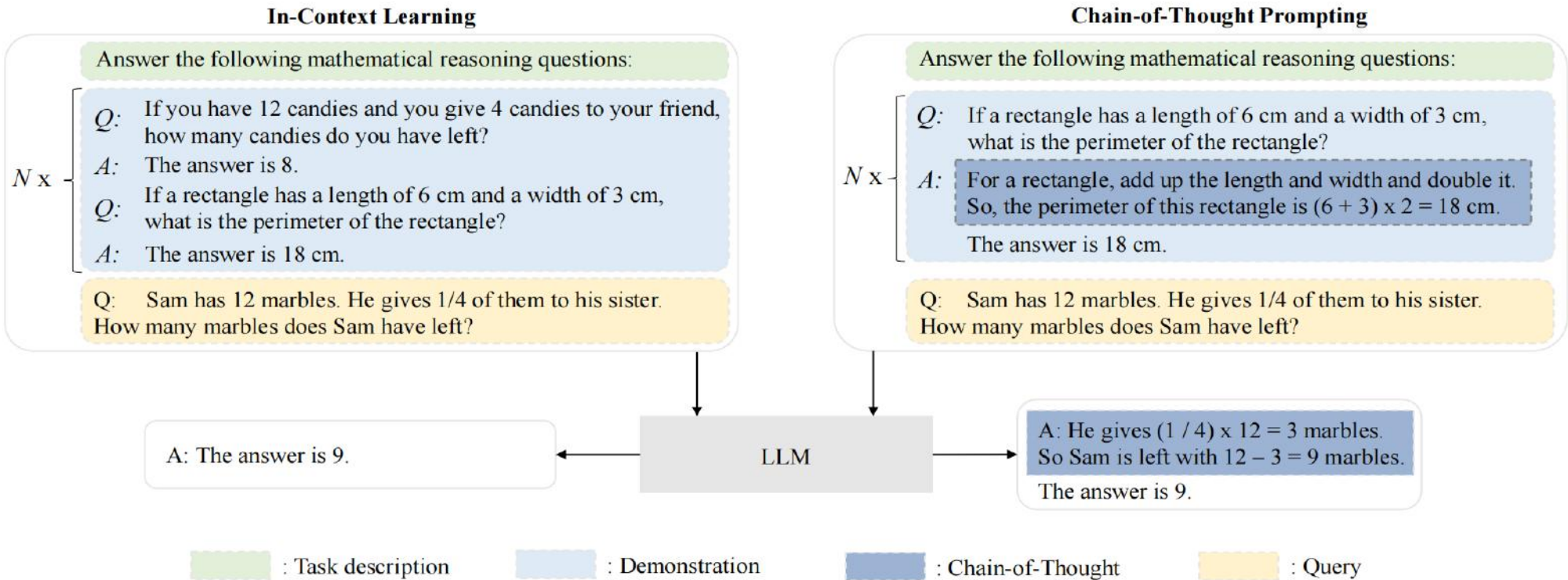
Prompt tuning与pre-train and fine-tune对比



1.3 GPT4六大颠覆式技术创新（4）：Prompt

- 语境学习(in-context learning, ICL)作为一种特殊的提示形式与GPT-3一起被首次提出，并已成为一种典型的利用LL的方法。首先，从任务描述开始，从任务数据集中选择一些示例作为演示。然后，将它们按照特定的顺序组合起来，形成具有特殊设计模板的自然语言提示。最后，测试实例被附加到演示中，作为LLM生成输出的输入。基于任务演示，LLM可以在不显式梯度更新的情况下识别并执行新任务。

情境学习(ICL)与思维链(CoT)提示的比较研究



注释：ICL用自然语言描述、几个演示和一个测试查询提示LLM，而CoT提示则涉及提示中的一系列中间推理步骤。

1.3 GPT4六大颠覆式技术创新（5）：插件

- 由于LLM是在大量纯文本语料库上训练成文本生成器的，因此在非文本生成方面(例如，数值计算)表现不佳。此外，LLM的能力局限于预训练数据，例如无法捕捉最新信息。
- 为解决这些问题，ChatGPT启用了外部插件的机制，帮助ChatGPT访问最新信息，运行计算或使用第三方服务，类似于LLM的“眼睛和耳朵”，可以广泛扩展LLM的能力范围。
- 2023年5月，ChatGPT更新了包括网络浏览功能和70 个测试版插件。此更新有望彻底改变ChatGPT 的使用方式，从娱乐和购物到求职和天气预报。
- ChatGPT建立一个社区，插件开发者来构建ChatGPT插件，然后在语言模型显示的提示符中列出启用的插件，以及指导模型如何使用每个插件的文档。

ChatGPT可以启用网络浏览和插件功能

General

Beta features

Data controls

As a Plus user, enjoy early access to experimental new features, which may change during development.

Web browsing

Try a version of ChatGPT that knows when and how to browse the internet to answer questions about recent topics and events.

Plugins

Try a version of ChatGPT that knows when and how to use third-party plugins that you enable.

ChatGPT插件部分展示

**Expedia**
Bring your trip plans to life—get there, stay there, find things to see and do.

**FiscalNote**
Provides and enables access to select market-leading, real-time data sets for legal, political, and regulatory data and information.

**Instacart**
Order from your favorite local grocery stores.

**KAYAK**
Search for flights, stays and rental cars. Get recommendations for all the places you can go within your budget.

**Klarna Shopping**
Search and compare prices from thousands of online shops.

**Milo Family AI**
Giving parents superpowers to turn the manic to magic, 20 minutes each day. Ask: Hey Milo, what's magic today?

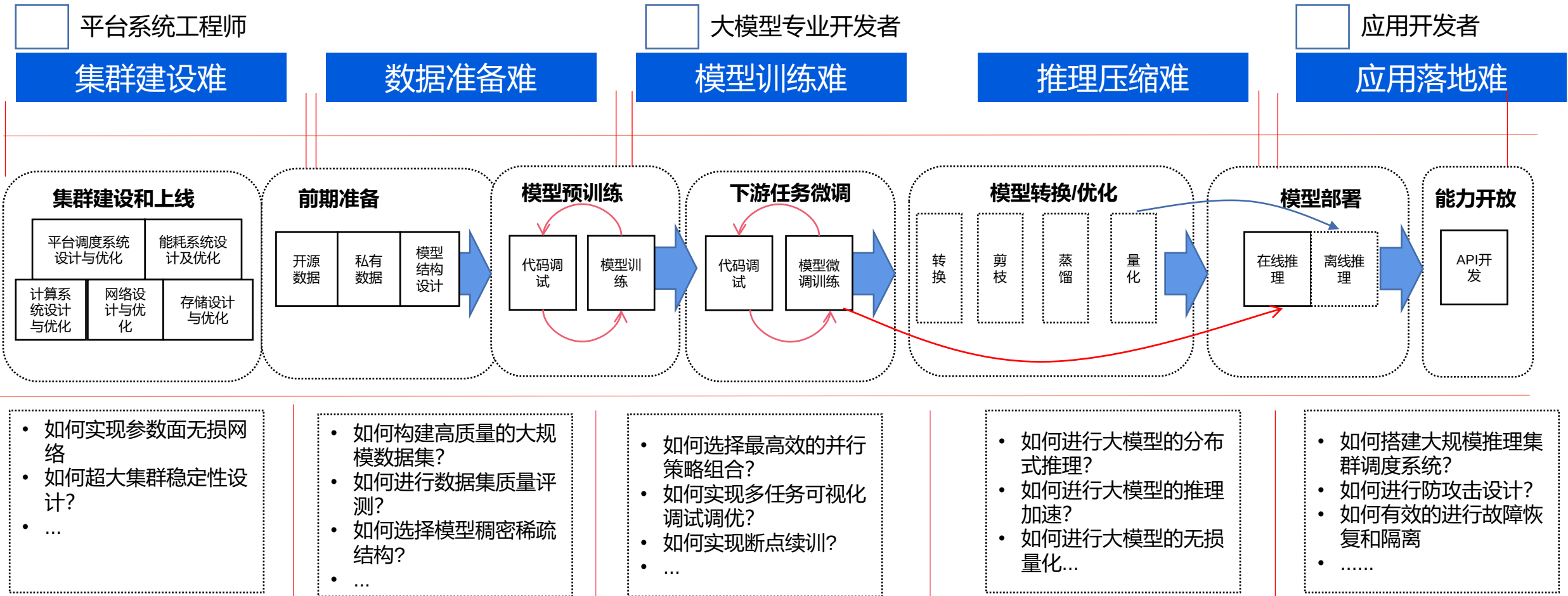
**OpenTable**
Provides restaurant recommendations, with a direct link to book.

**Shop**
Search for millions of products from the world's greatest brands.

**Speak**
Learn how to say anything in another language with Speak, your AI-powered language tutor.

1.3 GPT4六大颠覆式技术创新（6）：系统工程

- OpenAI 联合创始人& CEO Sam Altman表示：GPT-4是人类迄今为止最复杂的软件系统。
- LLM的发展使得研发和工程的界限不再清晰。LLM的训练需要在大规模数据处理和分布式并行训练方面的广泛实践经验。
开发LLM，研究人员必须解决复杂的工程问题，与工程师一起工作或成为工程师。



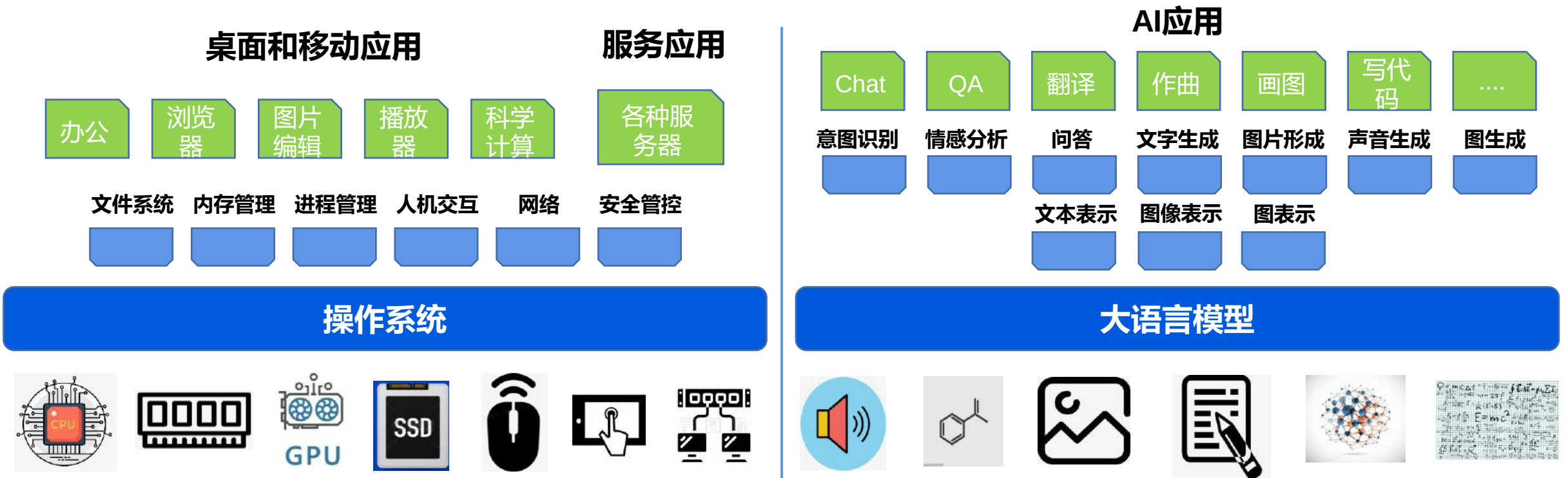
2. AI算力：GPT的基座，显著受益于新一轮科技革命

2.1 GPT开启AI新纪元：对标 ows的生态价值

ChatGPT的发布类似于ows的诞生。

- ChatGTP作为大语言模型，将会起到信息系统入口的作用，同时，ChatGPT或将重塑目前的软件生态。
- 2022年，ows在全球PC操作系统市占率约75%，应用数量3000万以上，是世界上生态规模最庞大的商业操作系统。
- 围绕ows所创造的桌面软件生态，诞生了现有的全球互联网巨头，亚马逊、谷歌、META、阿里巴巴、腾讯、百度等。

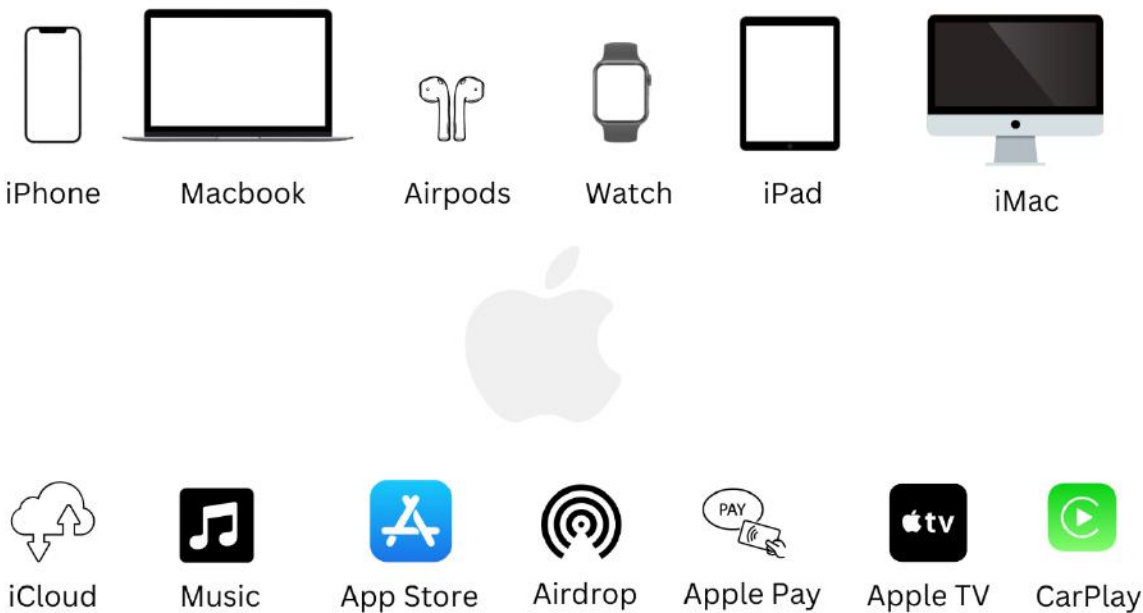
ChatGPT成为人工智能时代新的信息系统入口



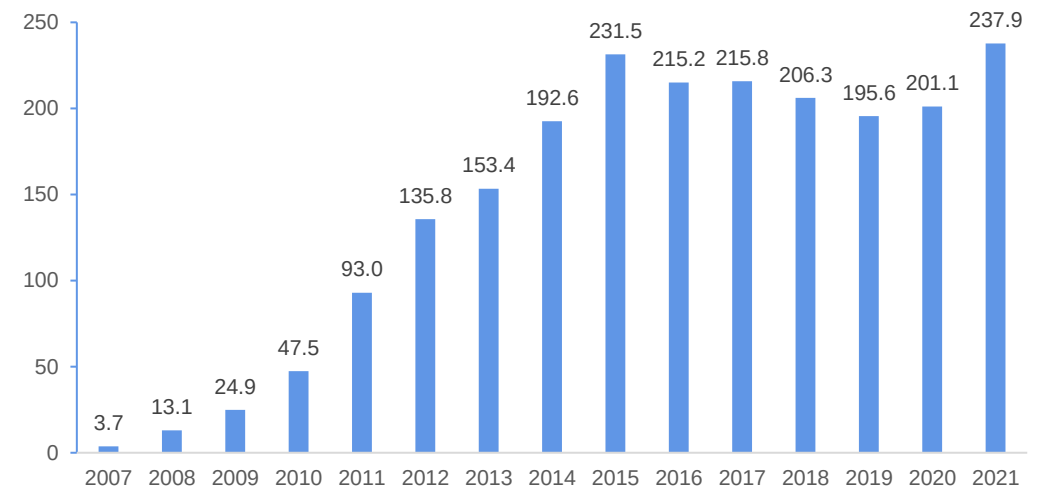
2.1 GPT开启AI新纪元：对标Iphone的“十年周期”

- 英伟达CEO黄仁勋：ChatGPT是AI的“Iphone时刻”
- iPhone是苹果公司（Apple Inc.）于2007年1月9日开始发售的搭载iOS操作系统的系列手机产品。
- 同时，围绕苹果创造的智能手机生态，诞生了抖音、微信等应用公司，以及苹果供应链环节的广阔市场机遇。

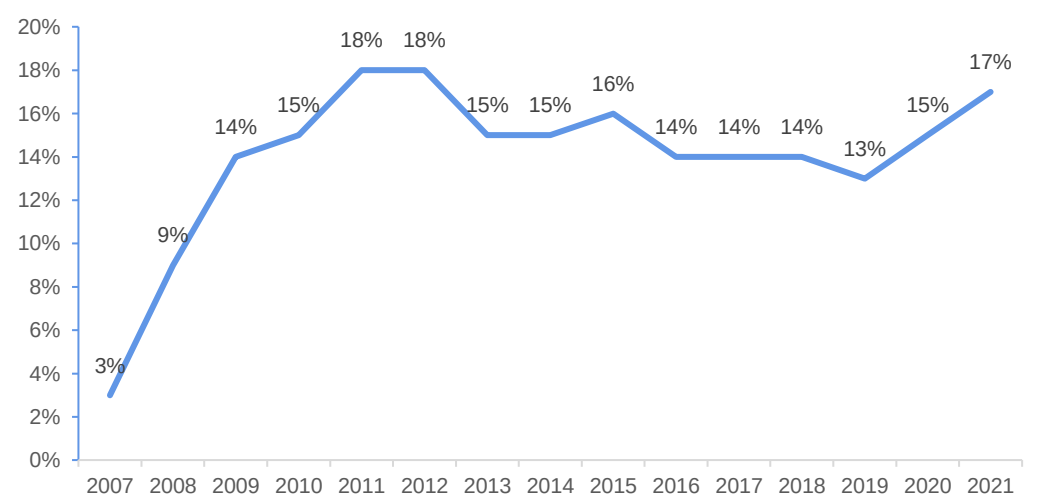
苹果的生态



2007-2021年苹果iPhone出货量（百万台）



2007-2021年苹果iPhone市占率

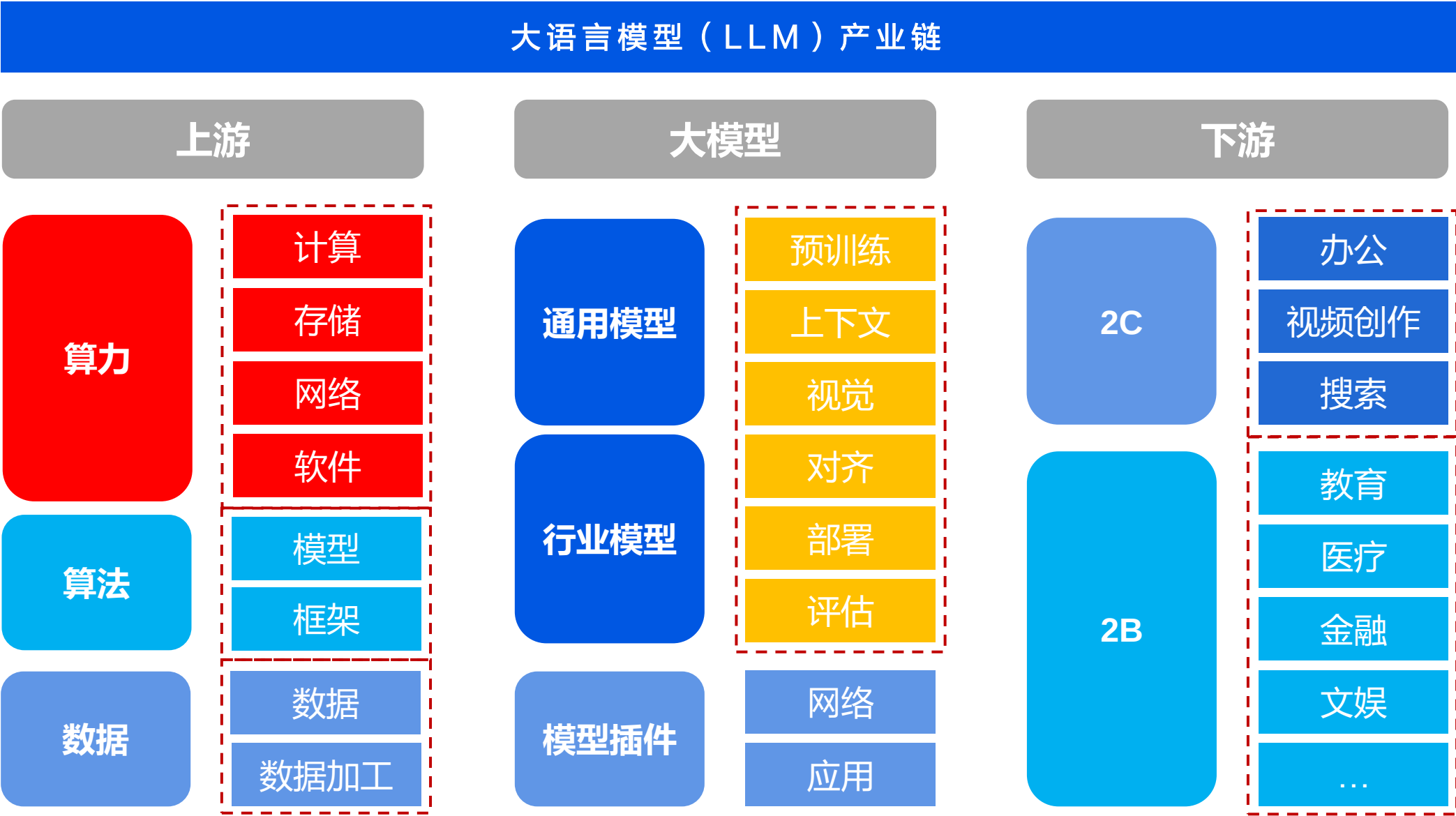


2.1 GPT开启AI新纪元：科技产业的“十年新周期”

时间	1970s	1980s	1990s	2000s	2010s	2020s
类型	大型机	小型机	个人电脑	桌面互联网	移动互联网	AI
代表公司	IBM Control Data Sperry Burroughs	DEC HP Prime Data General	Microsoft Cisco Intel Dell	Google Ebay Sina 百度 阿里巴巴 腾讯	Apple Meta Qualcomm 腾讯 字节跳动	Nvidia Tesla Microsoft OpenAI ...
图示						

全球科技产业“十年一周期”，新的科技革命诞生新的巨头！

2.2 算力是大模型的根基，GPT的率先受益赛道



2.2 算力是大模型的根基，GPT的率先受益赛道

- ◆ 算力是对信息数据进行处理输出目标结果的计算能力。随着社会数字化转型的持续深入，算力已成为支撑和推动数字经济发展的核心力量，并对推动科技进步、社会治理等发挥着重要的作用。根据中国算力发展指数白皮书测算，算力每投入1元，将带动3-4元的经济产出。
- ◆ 2021年全球计算设备算力总规模达到615EFlops，同比增长44%，其中智能算力规模为232EFlops，超级算力规模为14EFlops。智算中心、边缘数据中心将保持高速增长。

算力需求场景

超算



- 科学实验
- 气象研究
- 航空航天
- 石油勘探

通用、智算



- 互联网
- 通信
- 金融
- 政府

边缘算力

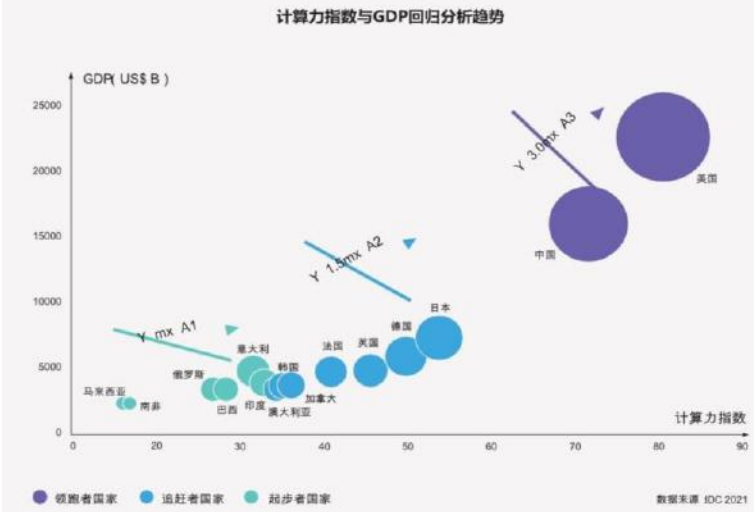


- 智能汽车
- 视频监控
- 移动设备
- 传感设备

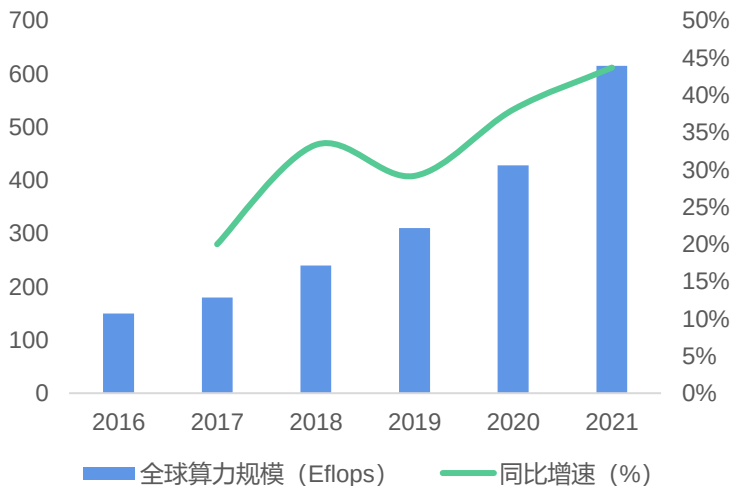
先进计算技术产业体系框架



算力指数与GDP走势显著正相关



2016-2021年全球算力规模及增速



2.2 算力是大模型的根基，GPT的率先受益赛道

- **微软投资10亿美金打造OpenAI超算平台。**2020年5月，微软投资10亿美金与OpenAI独家合作打造了Azure AI超算平台亮相，性能位居全球前五，拥有超过28.5万个CPU核心、1万个GPU、每GPU拥有400Gbps网络带宽的超级计算机，主要用于大规模分布式AI模型训练。
- 据OpenAI报告，训练一次1746亿参数的 GPT-3模型需要的算力约为3640 PFlop/s-day。即假如每秒计算一千万亿次，也需要计算3640天。

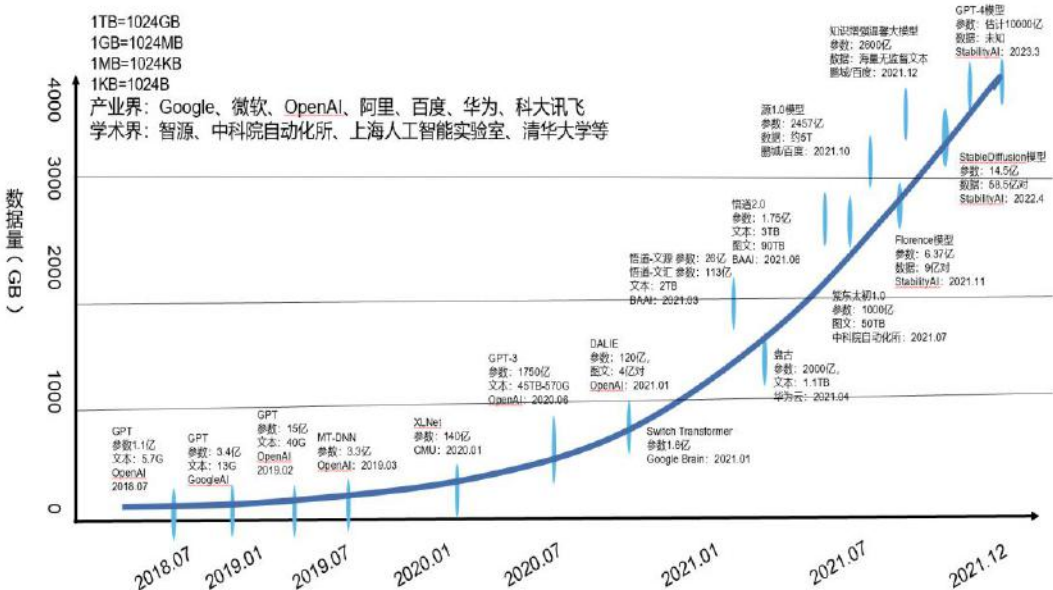


OpenAI融资历程			
融资时间	融资轮次	融资金额	投资方
2023/1	B轮	100亿美元	微软
2022/1	A轮	2.5亿美元	微软、Google Ventures等
2021/1	Pre-A轮	未披露	Sequoia Capital Andreessen Horowitz Tiger Global Managernernt Bedrock Capital
2021	--	未披露	微软
2019/7	天使+轮	10亿美元	微软
2019/4	天使轮	1000万美元	Matthew Brown
2019/3	种子轮	未披露	Khosla Ventures Reid Hoffman Foundation
2016/8	Pre-种子轮	12万美元	Y Combinator

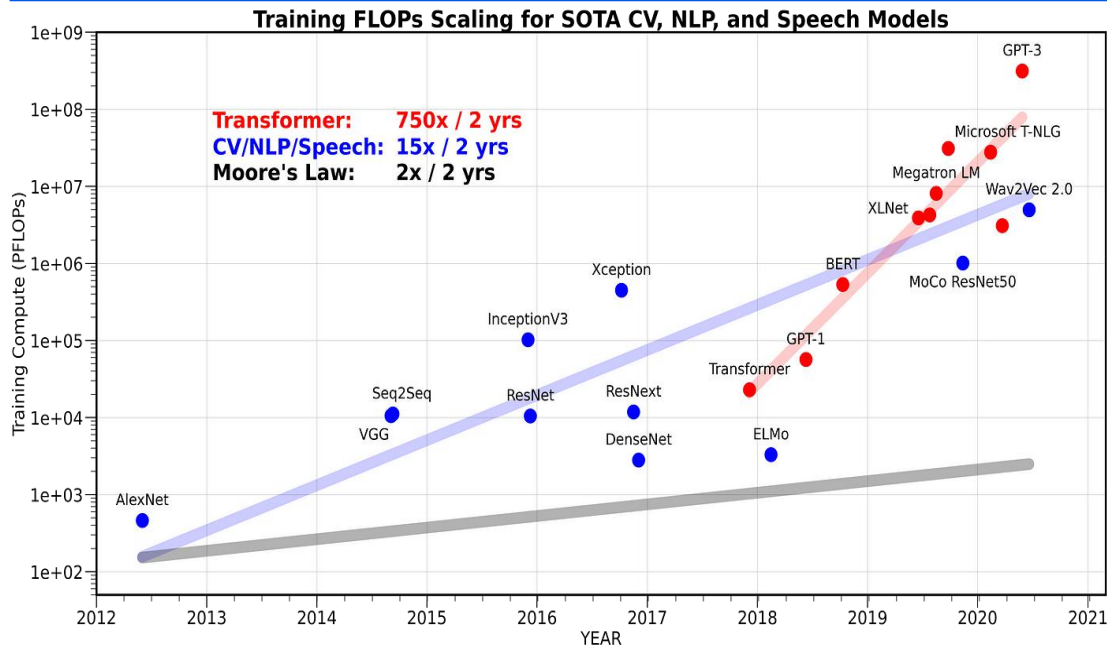
2.2 算力是大模型的根基，GPT的率先受益赛道

- Transformer相关模型对算力的需求提升数倍，远超过算力产业的摩尔定律增速。基于Transformer体系结构的大型语言模型(Large Language Models) 涉及高达数万亿从文本中学习的参数。开发它们是一个昂贵、耗时的过程，需要深入研究技术专长、分布式数据中心规模的基础设施和完整的堆栈加速计算方法。
- “大模型+大数据”成为预训练模型的“新范式”。近年新推出的大语言模型所使用的数据量和参数规模呈现“指数级”增长，参数和数据规模的增长带来的是算力消耗增加。我们预计未来的大模型所消耗的算力将远超过目前已有的大模型。

“大模型+大数据”成为AI预训练模型的“新范式”



Transformer模型消耗的算力远超过摩尔定律



2.2 算力是大模型的根基，GPT的率先受益赛道

- ChatGPT产品运营需要更大的算力：据SimilarWeb数据，2023年1月ChatGPT官网总访问量为6.16亿次。据Fortune杂志，每次用户与ChatGPT互动，产生的算力云服务成本约0.01美元。
- 根据科技云报道如果使用总投资30.2亿元、算力500P的数据中心来支撑ChatGPT的运行，至少需要7-8个这样的数据中心，基础设施的投入都是以百亿计的。

GPT的应用场景

应用场景

金融

决策指挥

业务调度
实时分析
风险预警

客户服务

对话机器人
分析报告

风险合规

报告审核
信用风险

营销

金融营销

仿真演练
售前资料准备
客户外呼

电商营销

脚本设计
文案生成
产品素材
海报生成

法律

咨询

法律咨询
法条检索
司法解释

案件

类案推荐
法条查询
信息研判
辅助写作

教育

在线答疑
课程规划
智能辅导

医疗

医学影像处理
病历录入
心理辅导

其他

工业设计
数字人主播
信息采集

1 大模型+金融:

文档知识信息智能抽取审核
辅助金融从业人员智能化写作

2 大模型+营销:

营销文案智能生成
营销资料、素材、底稿快速生成
基于客服对话、直播反馈的快速话术指导

3 大模型+法律:

帮助律师智能化完成客户咨询、信息检索
帮助司法人员快速查询类似案件、信息研判

4 大模型+教育:

帮助客户智能规划课程，针对性辅导

5 大模型+医疗:

提升病历录入效率
帮助医生对患者进行心理辅导

6 大模型+其他:

赋能垂直行业，降本增效

GPT-4的API价格

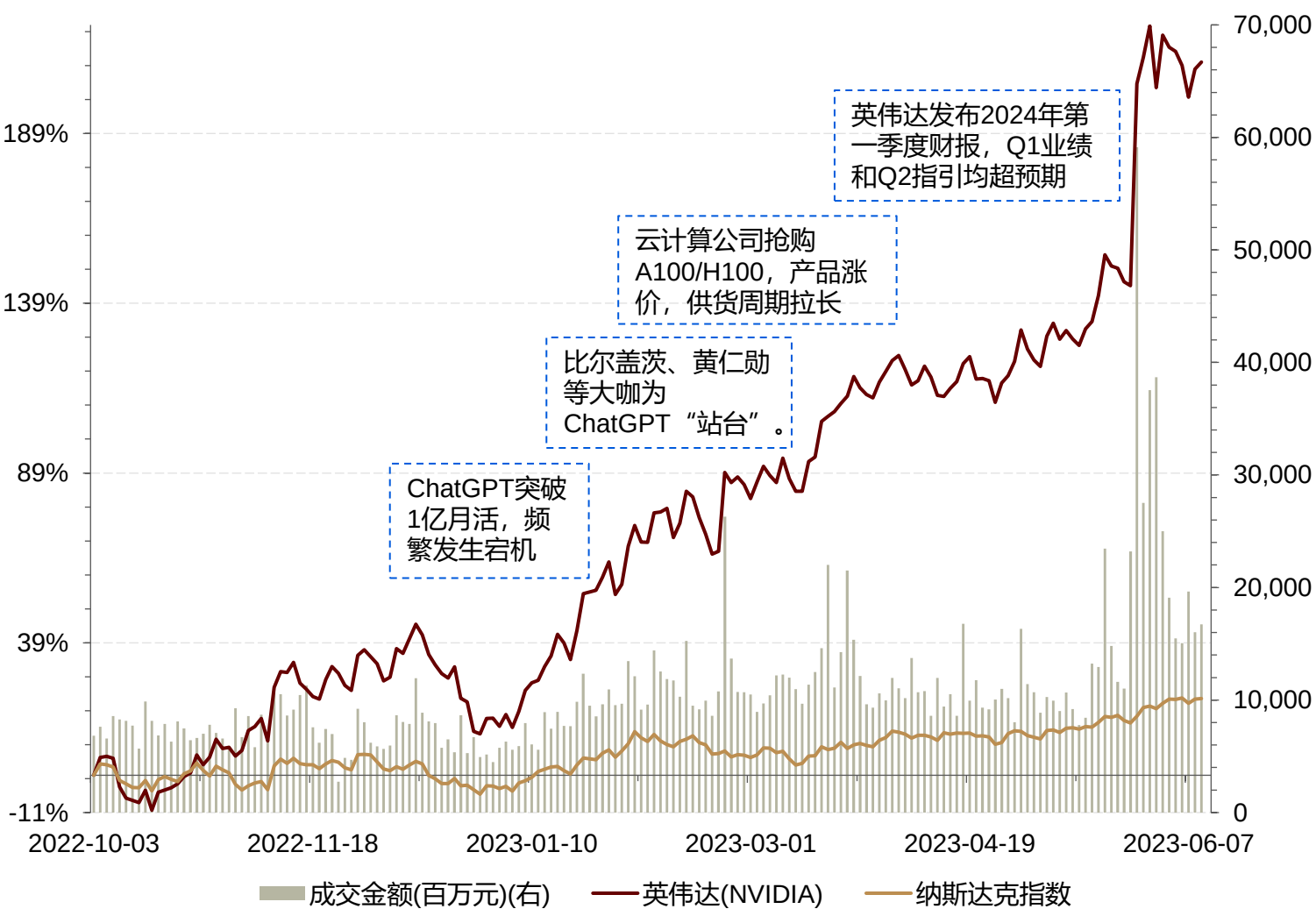
Model	Prompt	Completion
8K context	\$0.03 / 1K tokens	\$0.06 / 1K tokens
32K context	\$0.06 / 1K tokens	\$0.12 / 1K tokens

2.3 2023年，全球算力上市公司市值涨幅可观

2022年10月-至今 全球算力相关公司市值涨幅

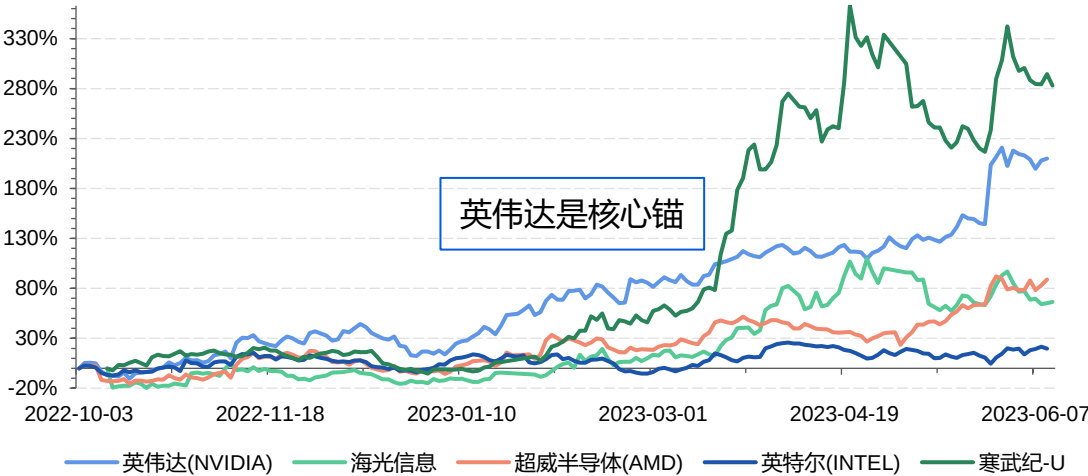
行业	证券代码	证券简称	涨幅排名
光模块	603083.SH	剑桥科技	681.24%
光模块	301205.SZ	联特科技	496.05%
服务器	SMCI.O	超微电脑	375.14%
光模块	300308.SZ	中际旭创	366.71%
光模块	300502.SZ	新易盛	303.38%
存储	300042.SZ	朗科科技	270.54%
GPU	688256.SH	寒武纪	251.32%
光模块	300394.SZ	天孚通信	226.95%
GPU	NVDA.O	英伟达	219.55%
服务器	3231.TW	纬创	166.29%
服务器	601138.SH	工业富联	150.59%
服务器	000977.SZ	浪潮信息	129.01%
服务器	603019.SH	中科曙光	116.81%
服务器	000938.SZ	紫光股份	106.24%
GPU	688047.SH	龙芯中科	97.65%
GPU	AMD.O	AMD	97.16%
光模块	000988.SZ	华工科技	95.48%
GPU	300474.SZ	景嘉微	93.95%
光模块	002281.SZ	光迅科技	89.64%
服务器	2382.TW	广达	67.53%
服务器	000066.SZ	中国长城	63.11%
服务器	2356.TW	英业达	61.35%
服务器	DELL.N	戴尔	43.05%
服务器	0992.HK	联想集团	42.22%
存储	000660.KS	海力士	38.15%
存储	005930.KS	三星	33.71%
GPU	688041.SH	海光信息	33.27%
存储	MU.O	美光科技	31.44%
GPU	INTC.O	INTEL	25.27%
存储	300223.SZ	北京君正	19.43%
存储	688008.SH	澜起科技	11.22%
存储	603986.SH	兆易创新	9.76%

2022年10月-至今 英伟达和全球部分证券市场走势图

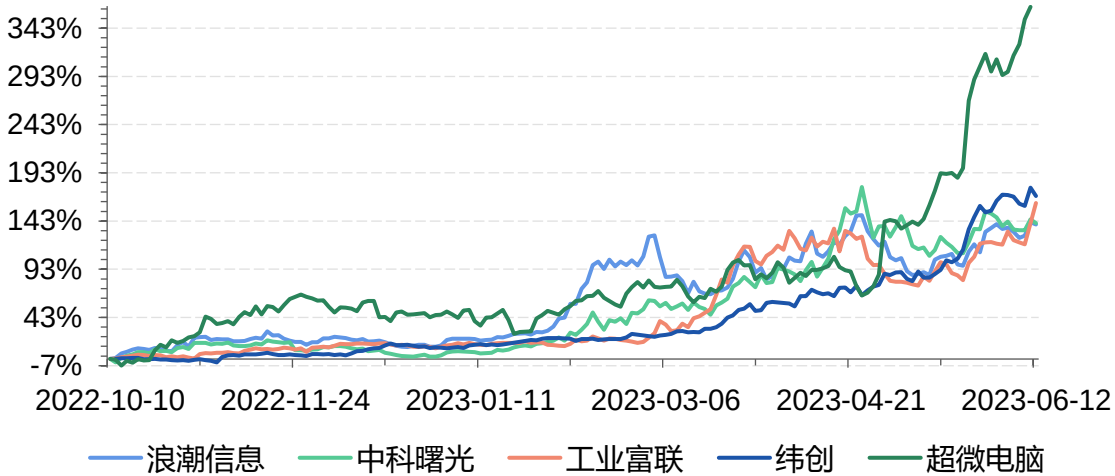


2.3 2023年，全球算力上市公司市值涨幅可观

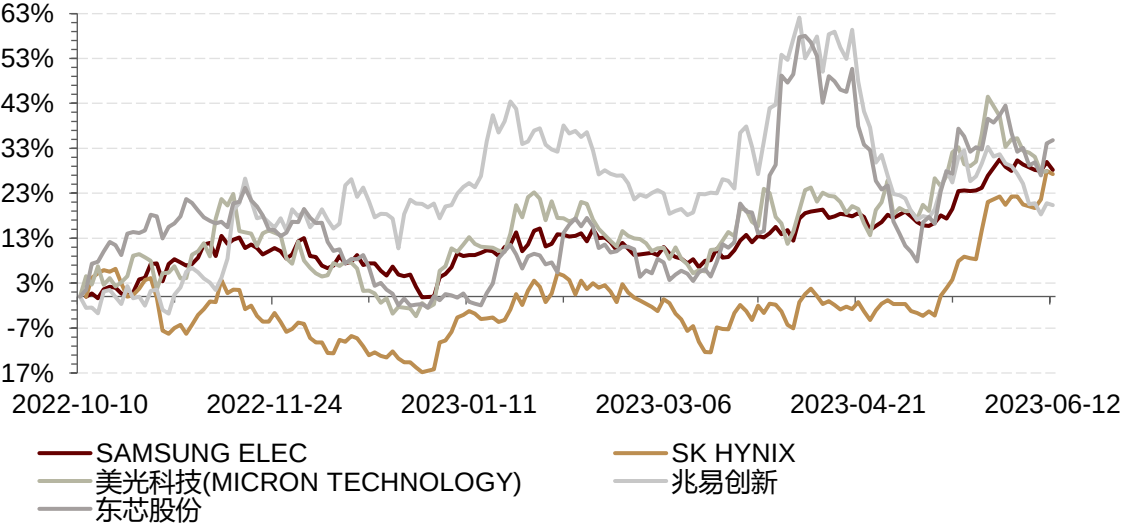
2022.10-至今 GPU公司股价走势



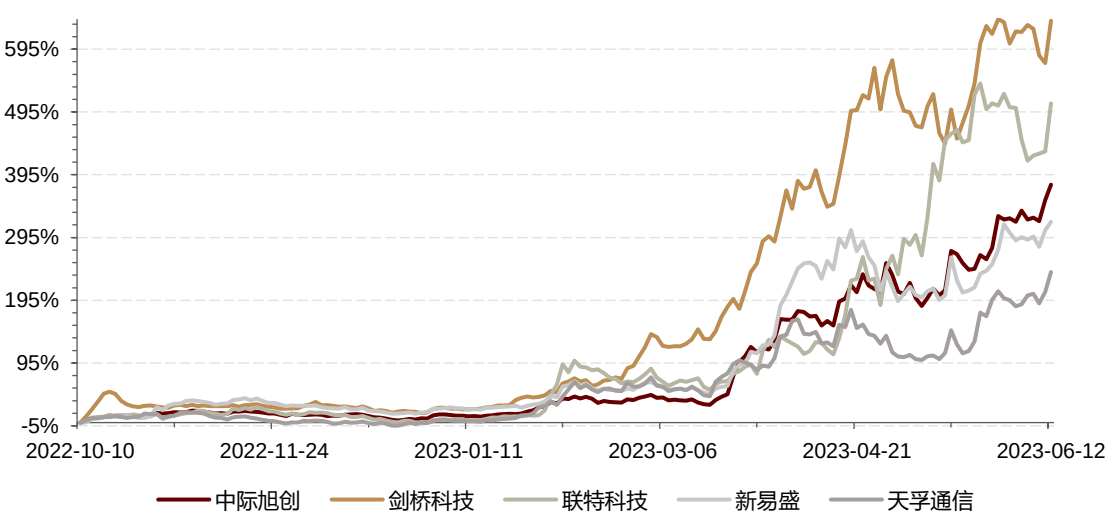
2022.10-至今 服务器公司股价走势



2022.10-至今 存储公司股价走势



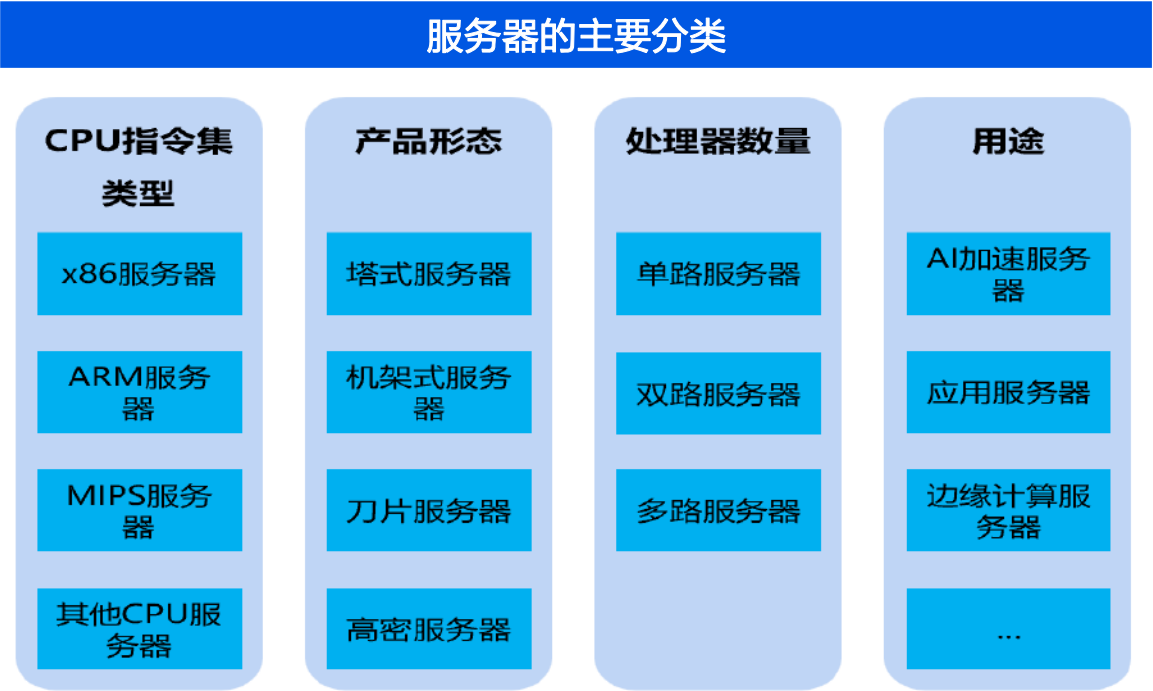
2022.10-2023.6 光模块公司股价走势



3. 计算：GPU为算力核心，服务器为重要载体

3.1 服务器：AI算力的重要载体

- 服务器通常是指那些具有较高计算能力，能够提供给多个用户使用的计算机。服务器与PC机的不同点很多，例如PC机在一个时刻通常只为一个用户服务。服务器与主机不同，主机是通过终端给用户使用的，服务器是通过网络给客户端用户使用的，所以除了要拥有终端设备，还要利用网络才能使用服务器电脑，但用户连上线后就能使用服务器上的特定服务了。
- AI服务器是一种能够提供人工智能（AI）计算的服务器。它既可以用来支持本地应用程序和网页，也可以为云和本地服务器提供复杂的AI模型和服务。AI服务器有助于为各种实时AI应用提供实时计算服务。AI服务器按应用场景可分为训练和推理两种，其中训练对芯片算力要求更高，推理对算力的要求偏低。



NVIDIA A100服务器

GPUs	8x NVIDIA A100
GPU Memory	320 GB total
Peak performance	5 petaFLOPS AI 10 petaOPS INT8
NVSwitches	6
System Power Usage	6.5kW max
CPU	Dual AMD Rome 7742 128 cores total, 2.25 GHz(base), 3.4GHz (max boost)
System Memory	1TB
Networking	8x Single-Port Mellanox ConnectX-6 200Gb/s HDR Infiniband (Compute Network) 1x (or 2x*) Dual-Port Mellanox ConnectX-6 200GB/s HDR Infiniband (Storage Network also used for Eth*)
Storage	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 15TB (4x 3.84TB) U.2 NVME drives
Software	Ubuntu Linux OS (5.3+ kernel)
System Weight	271 lbs (123 kgs)
Packaged System Weight	315 lbs (143 kgs)
Height	6U
Operating temp range	5°C to 30°C (41°F to 86°F)

* Optional upgrades

3.1 服务器：AI算力的重要载体

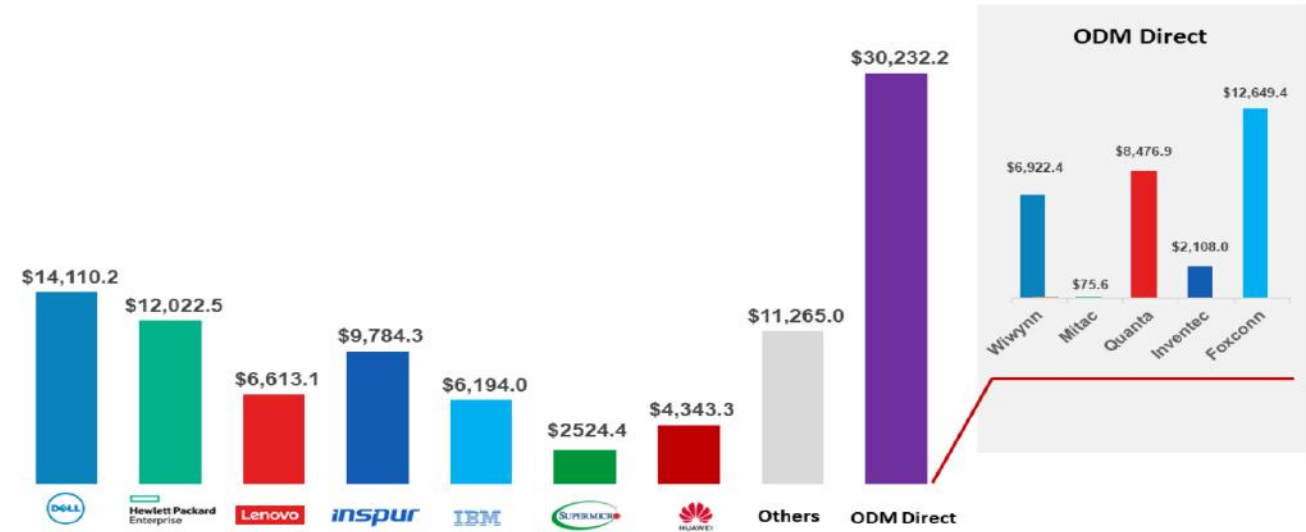
1) 全球服务器市场：

- 根据Counterpoint报告，2022年，全球服务器市场的收入将同比增长17%，达到1117亿美元。
- **主要企业：**戴尔、惠普、联想、浪潮和超微等服务器公司，以及富士康、广达、纬创、英业达等ODM厂商。ODM Direct的增长速度比整体市场高3个百分点，ODM Direct将成为大规模数据中心部署的硬件选择。

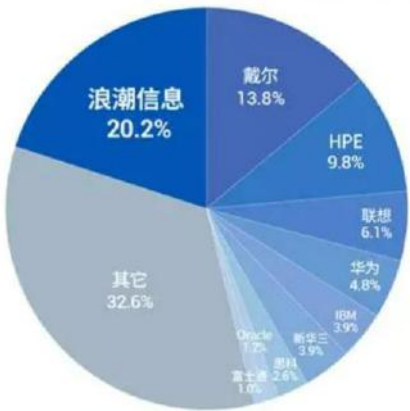
2) 全球AI服务器市场：

- 根据IDC数据，预计2022年市场规模约为183亿美元，2023年市场规模211亿美元。
- **市场份额：**浪潮信息市场占有率达20.2%。其次，戴尔、HPE、联想、华为占比分别为13.8%、9.8%、6.1%、4.8%。

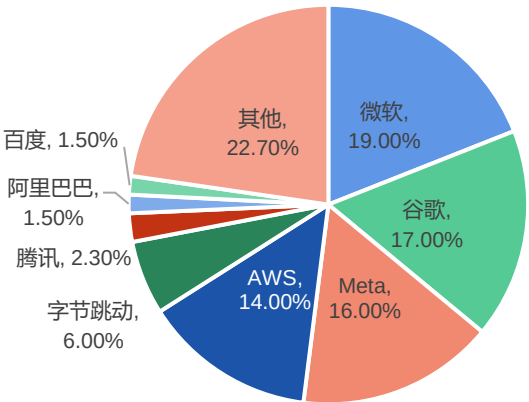
2021年全球各服务器公司收入（单位：百万美元）



2021全球AI服务器市场份额

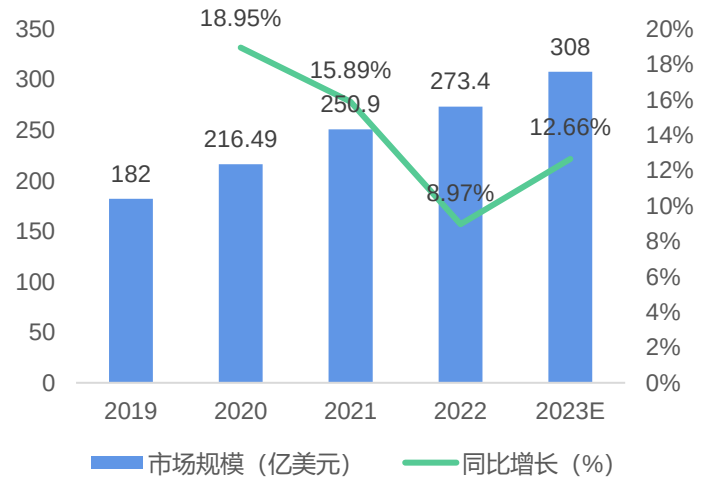


2022年全球AI服务器主要客户

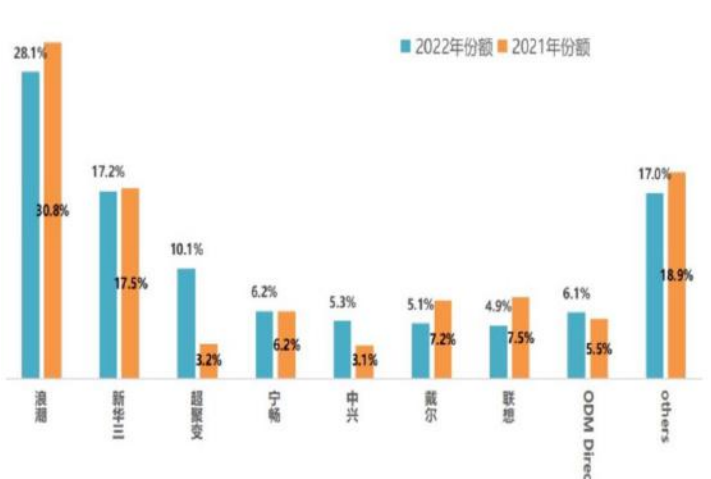


3.1 服务器：AI算力的重要载体

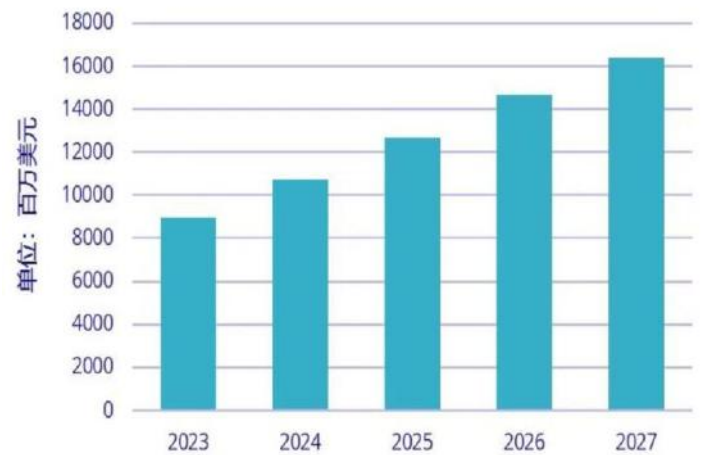
2019-2023年中国服务器市场规模



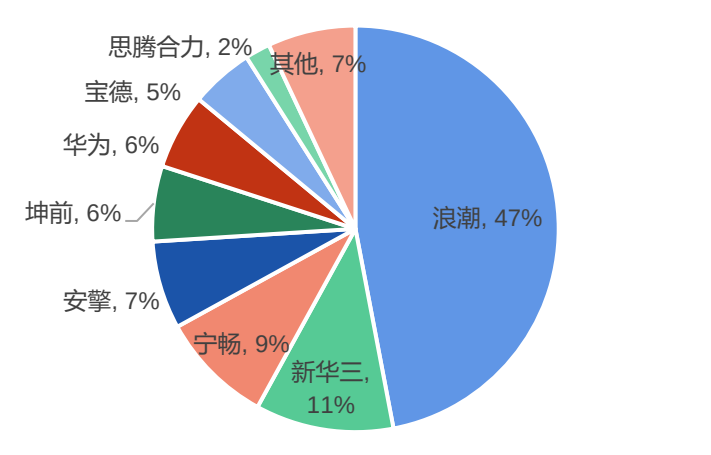
2021-2022年中国服务器市场份额



2023-2027E中国加速服务器市场规模



2022年中国AI服务器市场份额



1) 中国服务器市场：

- 2022年，中国服务器市场规模为273.4亿美元。
- 2022年，浪潮以28.1%的市场份额位列第一，收入达530.63亿。

2) 中国加速服务器市场：

- 根据IDC数据，2022年，中国加速服务器市场规模达到67亿美元，同比增长24%。
- 2022年，浪潮、新华三、宁畅位居前三，占据了60%以上的市场份额
- 互联网依然是最大的采购行业，占整体加速服务器市场接近一半的份额。

3.2 GPU：AI算力的核心

- AI芯片是算力的核心。AI芯片也被称为AI加速器或计算卡，即专门用于处理人工智能应用中的大量计算任务的模块（其他非计算任务仍由CPU负责）。伴随数据海量增长，算法模型趋向复杂，处理对象异构，计算性能要求高，AI芯片在人工智能的算法和应用上做针对性设计，可高效处理人工智能应用中日渐多样繁杂的计算任务。
- GPU是目前最广泛应用的AI芯片。AI芯片主要包括图形处理器（GPU）、现场可编程门阵列（FPGA）、专用集成电路（ASIC）、神经拟态芯片（NPU）等。GPU属于通用型芯片，ASIC属于专用型芯片，而FPGA则是介于两者之间的半定制化芯片。2022年，我国GPU服务器占AI服务器的89%。



不同AI芯片之间对比			
	GPU	FPGA	ASIC
特点	通用型	半定制化	专用型
芯片架构	叠加大量计算单元和高速内存，逻辑控制单元简单	具备可重构数字门电路和存储器,根据应用定制	电路结构可根据特定领域应用和特定算法定制
擅长领域	3D图像处理，密集型并行运算	算法更新频繁或者市场规模较小的专用领域	市场需求量大的专用领域
优点	计算能力强，通用性强，开发周期短，难度小，风险低	功能可修改，高性能、功耗远低于GPU，一次性成本低	专业性强、性能高于FPGA、功耗低、量产成本低
缺点	价格贵、功耗高	编程门槛高、量产成本高	开发周期长、难度太、风险高、一次性成本高

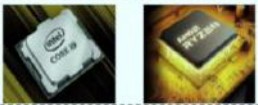
3.2 GPU：AI算力的核心

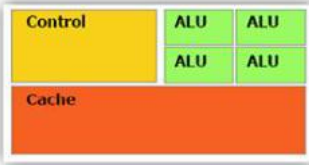
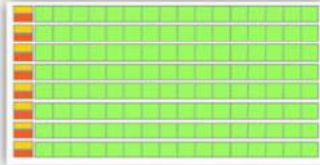
- GPU是NVIDIA公司在1999年8月发表NVIDIA GeForce 256绘图处理芯片时首先提出的概念。在此之前，电脑中处理影像输出的显示芯片，通常很少被视为是一个独立的运算单元。

年份	主要进展	代表产品	晶体管数量	总线	API	渲染模型
1980-1990	图形工作站系统	SGI Iris Geometry engine	<1M		Iris GL	
1995-1998	GPU,硬件光栅化	3dfx Voodoo NVIDIA TNT2 ATI Rage	10M	PCI,AGP 2X	OpenGL 1.12,DirectX6	
1999-2000	硬件几何处理	NVIDA Geforce2 ATI Radeon 7000 S3 Savage 3D	25M	AGP4X	OpenGL 1.12,DirectX7	
2001	可编程顶点程序	Geforce 3/4 Radeon8000	60M		OpenGL 1.13,DirectX8	1.0
2003	可编程程序像素	Geforce FX Radeon9000	100M	AGP8X	OpenGL 1.13,DirectX9	2.0
2004-2005	64位颜色，视频出来，增强可编程性	Geforce 6/7 Radeon X	200M	PCI-E	OpenGL 2.10, DirectX 9.10c	3.0
2006-至今	可编程同意渲染器GPGPU	Geforce 8/9 Radeon 2000/3000/4000	700M	PCIe2.0	OpenGL 2.11,Directx10, CUDA , OpenCL	4.0

3.2 GPU：AI算力的核心

- 图形处理器（GPU），又称显示核心（display core）、视觉处理器（video processor）、显示芯片（display chip）或图形芯片（graphics chip），是一种专门执行绘图运算工作的微处理器。
- GPU具有数百或数千个内核，经过优化，可并行运行大量计算。虽然GPU在游戏中以3D渲染而闻名，但它们对运行分析、深度学习和机器学习算法尤其有用。GPU允许某些计算比传统CPU上运行相同的计算速度快10倍至100倍。GPGPU，即是将GPU的图形处理能力用于通用计算领域的处理器。

GPU主要产品形态及玩家			
	游戏（桌面+云）	计算（AI、DC、HPC）	移动产品（手机、Pad）
独立显卡 (产品模式) nVidia和AMD独大			
集成显卡 (产品模式) Intel和AMD独大			
移动GPU (IP授权模式) IMG和ARM的IP授权 厂商众多			

GPU和CPU架构对比	
CPU	GPU
 ★ Low compute density ★ Complex control logic ★ Large caches (L1\$/L2\$, etc.) ★ Optimized for serial operations ★ Fewer execution units (ALUs) ★ Higher clock speeds ★ Shallow pipelines (<30 stages) ★ Low Latency Tolerance ★ Newer CPUs have more parallelism	 ★ High compute density ★ High Computations per Memory Access ★ Built for parallel operations ★ Many parallel execution units (ALUs) ★ Graphics is the best known case of parallelism ★ Deep pipelines (hundreds of stages) ★ High Throughput ★ High Latency Tolerance ★ Newer GPUs: ★ Better flow control logic (becoming more CPU-like) ★ Scatter/Gather Memory Access ★ Don't have one-way pipelines anymore

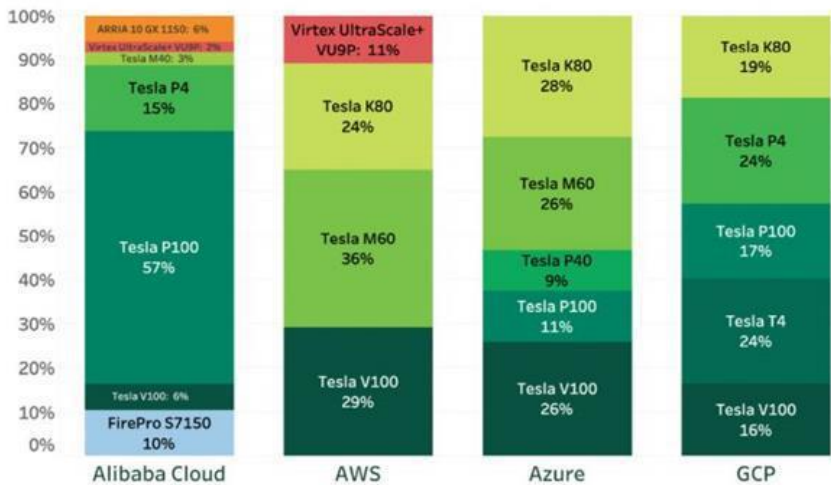
3.2 GPU：AI算力的核心

- 根据LC数据，2019年，NVIDIA在全球AI加速云市场份额高达97%。此外，AMD GPU、赛灵思FPGA、 Intel FPGA的云计算份额分别为1.0%、1.0%、0.6%的份额。
- 2021年，中国加速卡出货量超80万片，Nvidia市占率80%+。

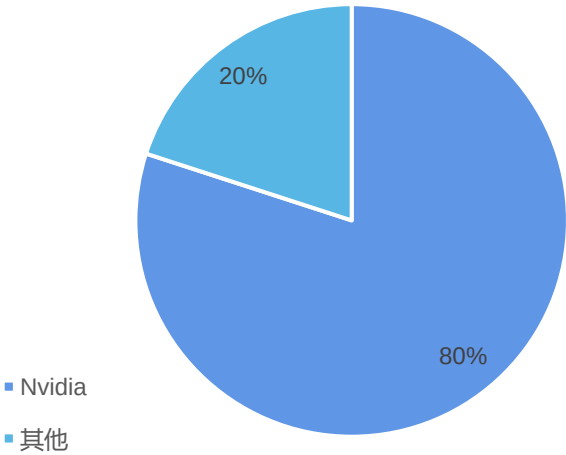
英伟达数据中心GPU类别

VideoCardz.com	NVIDIA H100	NVIDIA A100	NVIDIA Tesla V100	NVIDIA Tesla P100
Picture				
GPU	GH100	GA100	GV100	GP100
Transistors	80B	54.2B	21.1B	15.3B
Die Size	814 mm²	828 mm²	815 mm²	610 mm²
Architecture	Hopper	Ampere	Volta	Pascal
Fabrication Node	TSMC N4	TSMC N7	12nm FFN	16nm FinFET+
GPU Clusters	132/114*	108	80	56
CUDA Cores	16896/14592*	6912	5120	3584
L2 Cache	50MB	40MB	6MB	4MB
Tensor Cores	528/456*	432	320	-
Memory Bus	5120-bit	5120-bit	4096-bit	4096-bit
Memory Size	80 GB HBM3/HBM2e*	40/80GB HBM2e	16/32 HBM2	16GB HBM2
TDP	700W/350W*	250W/300W/400W	250W/300W/450W	250W/300W
Interface	SXM5*/PCIe Gen5	SXM4/PCIe Gen4	SXM2/PCIe Gen3	SXM/PCIe Gen3
Launch Year	2022	2020	2017	2016

2019年全球四大云平台加速卡占比

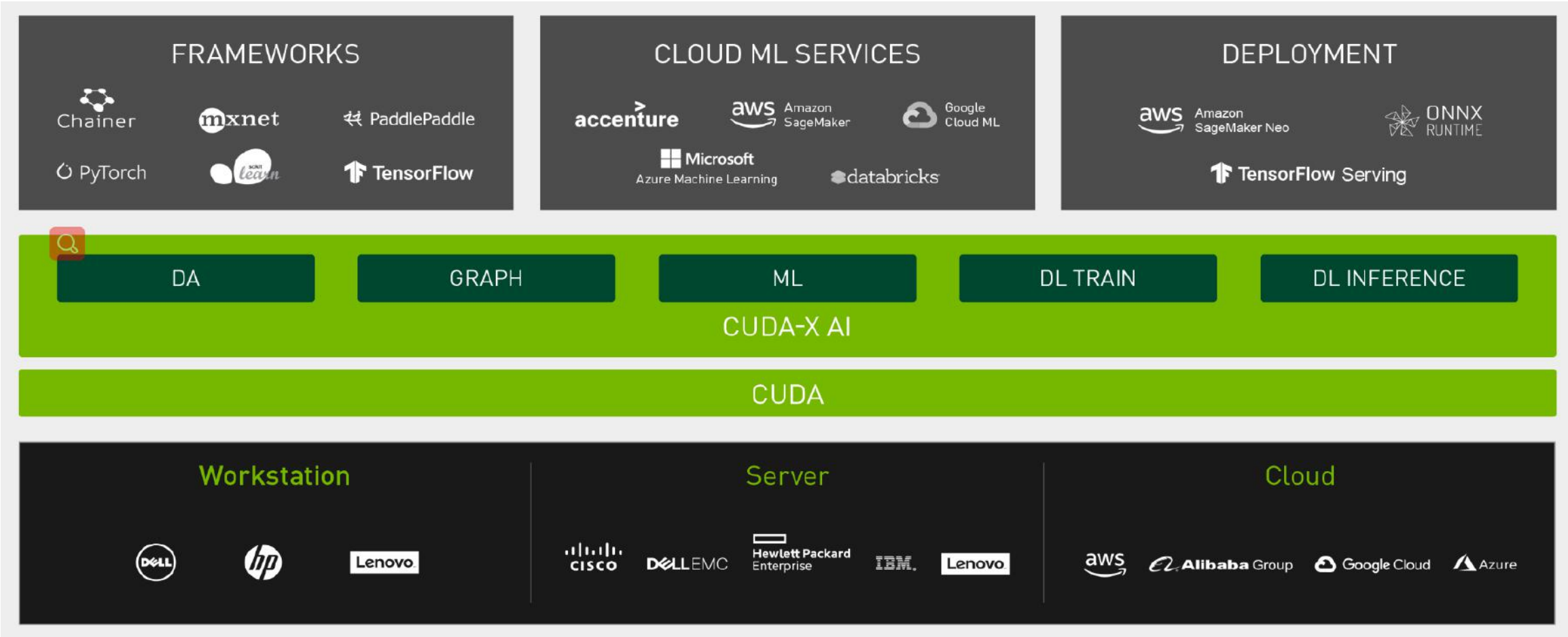


2021年我国服务器AI加速卡市场份额



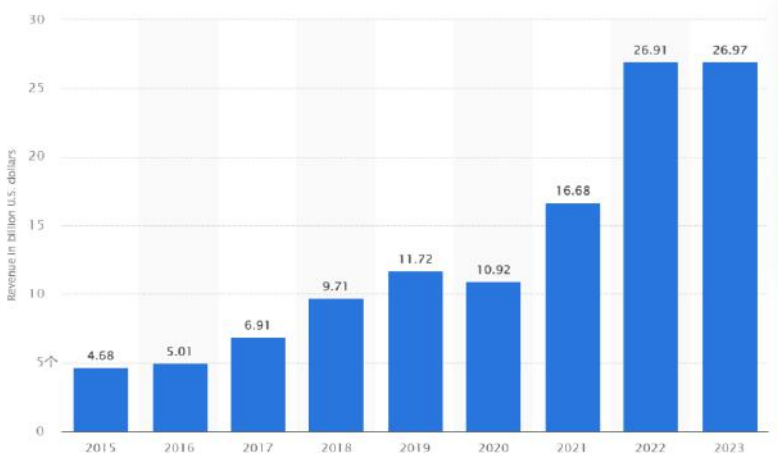
3.2 GPU：AI算力的核心

- CUDA是英伟达2007年推出的一种并行计算平台和应用程序编程接口(API)，允许软件使用某些类型的GPU进行通用计算机处理。CUDA与 NVIDIA GPU 无缝协作，加速跨多个领域的应用程序开发和部署。
- 目前，超过一百万的开发人员正在使用 CUDA-X，它提供了提高生产力的能力，同时受益于持续的应用程序性能。

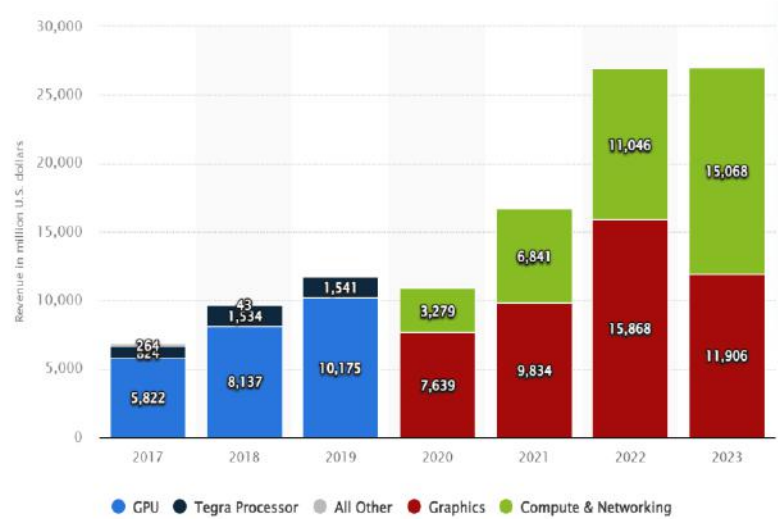


3.2 GPU：AI算力的核心

2015-2023年英伟达全球收入（十亿美元）

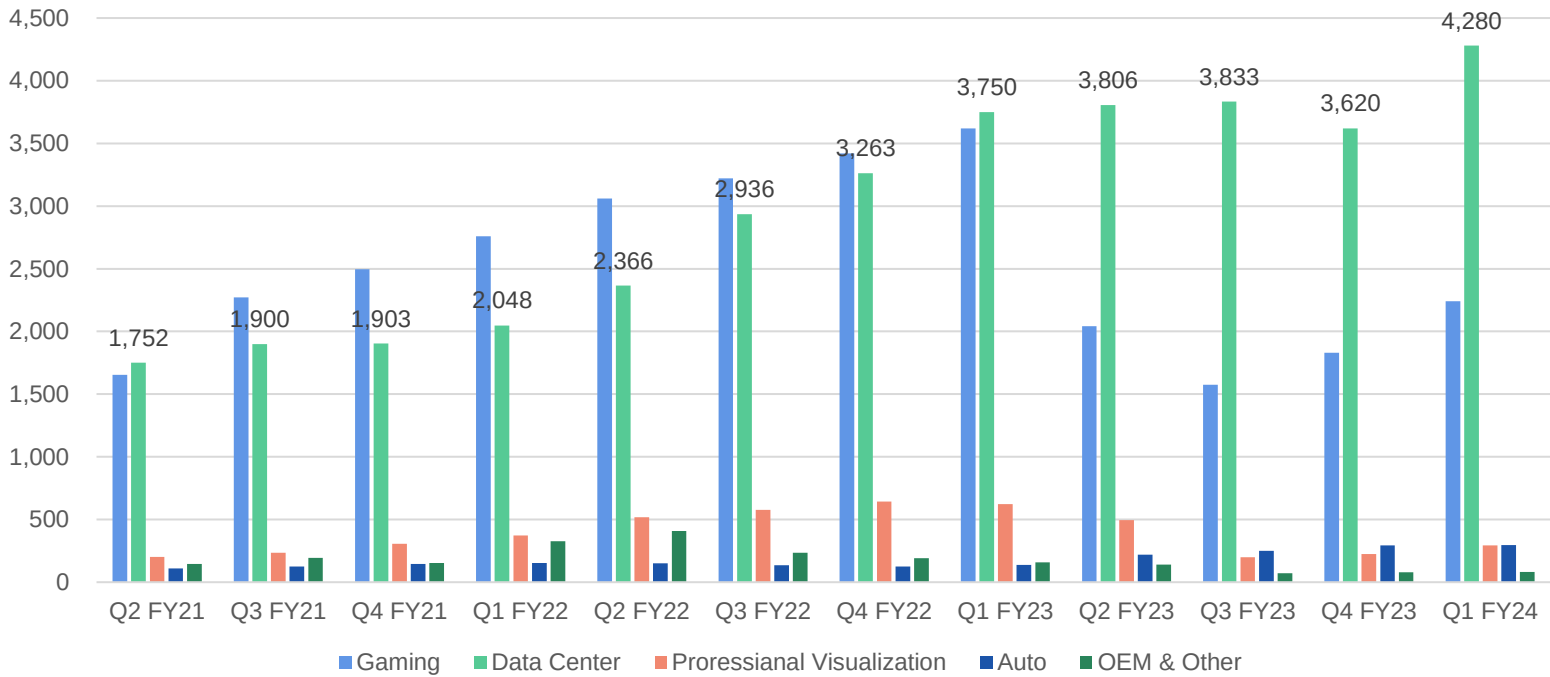


2015-2023年英伟达按部门收入（百万美元）



- 2023 财年，英伟达全球收入269.7亿美元，其中，图形业务部门的收入约为 119 亿美元，计算与网络部门的收入为 151 亿美元。
- 2023Q1（Q1 FY24）财季，英伟达数据中心业务营收42.8亿美元，历史新高，同比增长14%，环比增长18%。

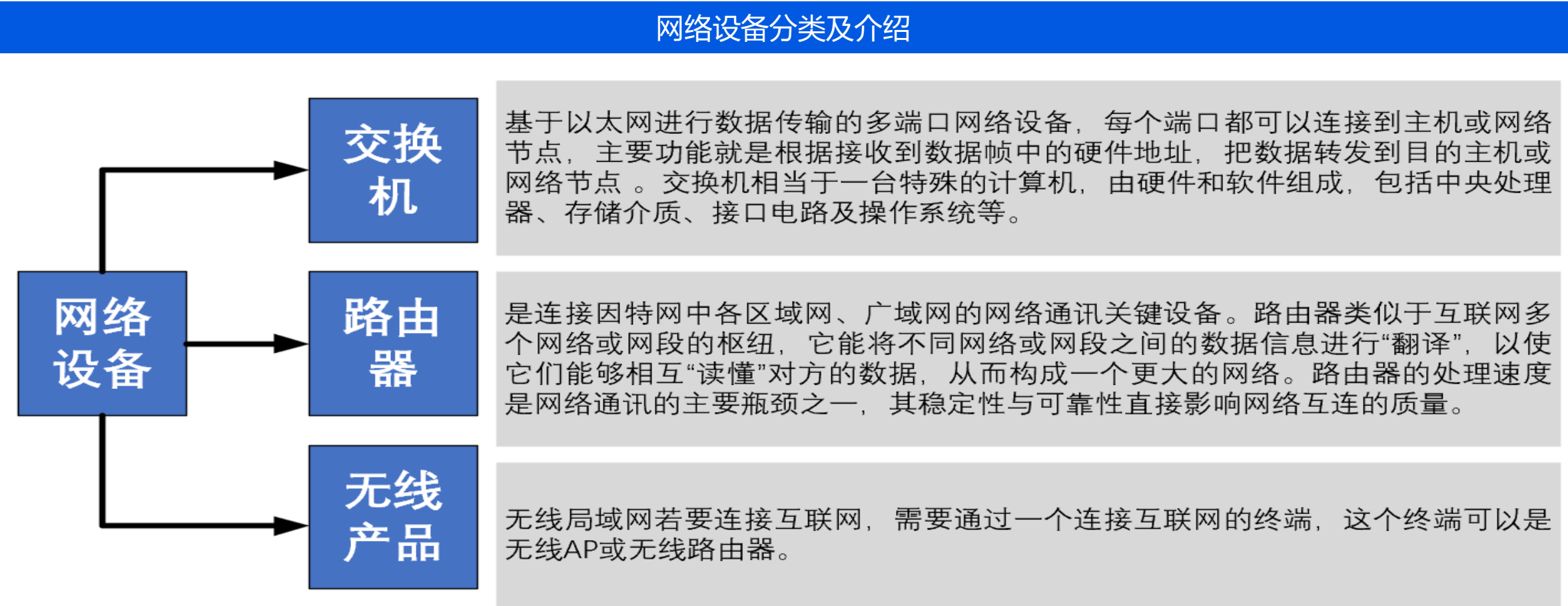
2021Q2-2024Q1 英伟达各季度收入-按市场



4. 网络：数据中心算力瓶颈，光模块需求放量

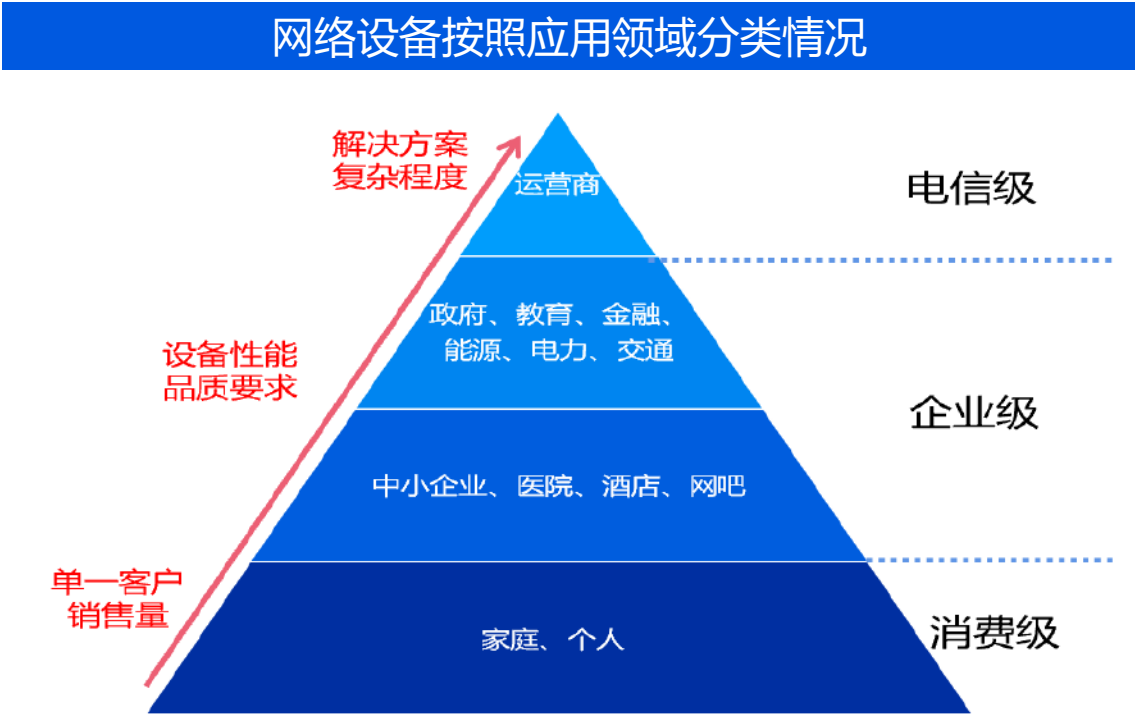
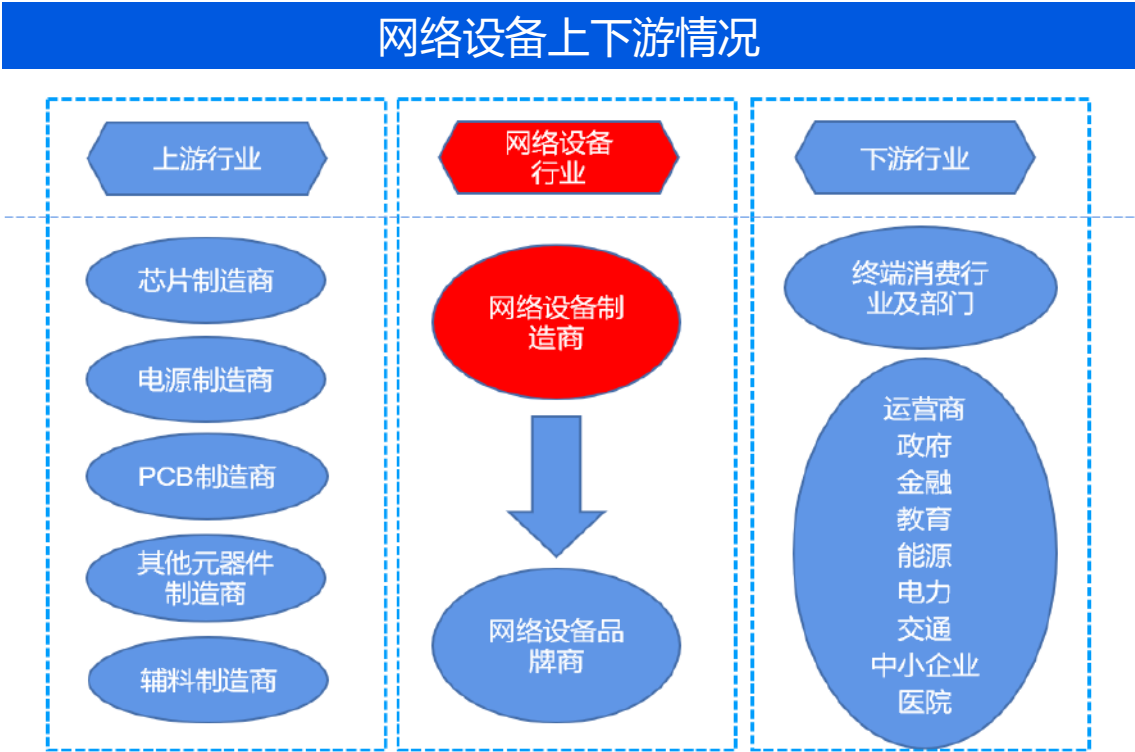
4.1 网络：算力的瓶颈之一，英伟达布局InfiniBand

- 数据通信设备（网络设备、ICT设备）泛指实现IP网络接入终端、局域网、广域网间连接、数据交换及相关安全防护等功能的通信设备，主要大类包括交换机、路由器、WLAN。其中主要的是交换机和路由器。
- 网络设备是互联网基本的物理设施层，属于信息化建设所需的基础架构产品。



4.1 网络：算力的瓶颈之一，英伟达布局InfiniBand

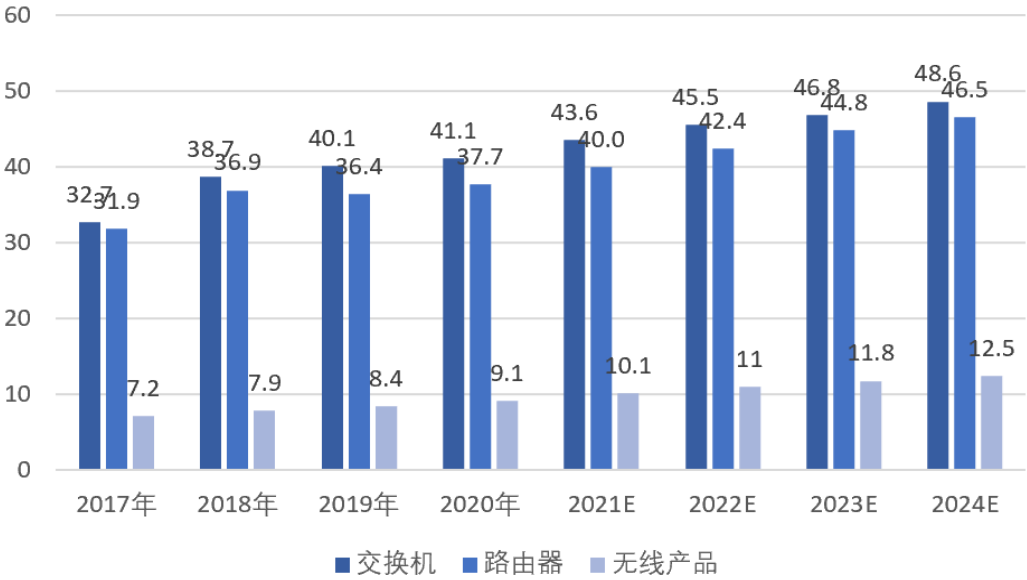
- 网络设备制造服务行业，上游主要为芯片、PCB、电源、各类电子元器件等生产商，直接下游为各网络设备品牌商，终下游包括运营商、政府、金融、教育、能源、电力、交通、中小企业、医院等各个行业。
- 网络设备根据应用领域分为电信级、企业级和消费级。电信级网络设备主要应用于电信运营商市场，用于搭建核心骨干互联网；企业级网络设备主要应用于非运营商的各种企业级应用市场，包括政府、金融、电力、医疗、教育、制造业、中小企业等市场；消费级网络设备主要针对家庭及个人消费市场。



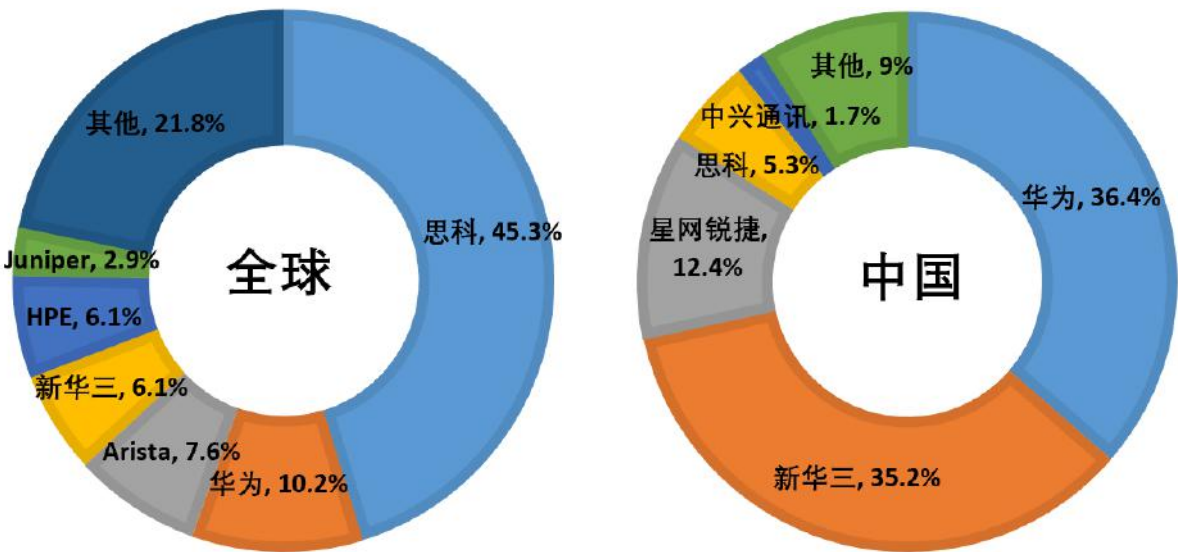
4.1 网络：算力的瓶颈之一，英伟达布局InfiniBand

- **市场规模**，1) **全球**：根据IDC报告，2021年网络市场规模为542.4亿美元，同比增长10.1%。其中交换机、路由器和WLAN增速分别为9.7%、6.5%和20.4%。2) **中国**：2021年网络市场规模为102.4亿美元（约677.89亿元人民币），同比增长12.1%。数字经济、5G、云计算、网络设备升级、大型数据中心建设等驱动网络设备行业需求持续提升。
- **竞争格局**，行业集中度较高，思科、华为、新华三等少数几家企业占据着绝大部分的市场份额，呈现寡头竞争的市场格局。
- **人工智能和高性能计算(HPC)日益增长的计算需求推动了多节点、多GPU系统的无缝、高速通信的需求，为了构建最强大的、能够满足业务速度的端到端计算平台，需要一个快速、可扩展的互连网络。**

2017-2024年中国网络设备市场规模统计（亿美元）

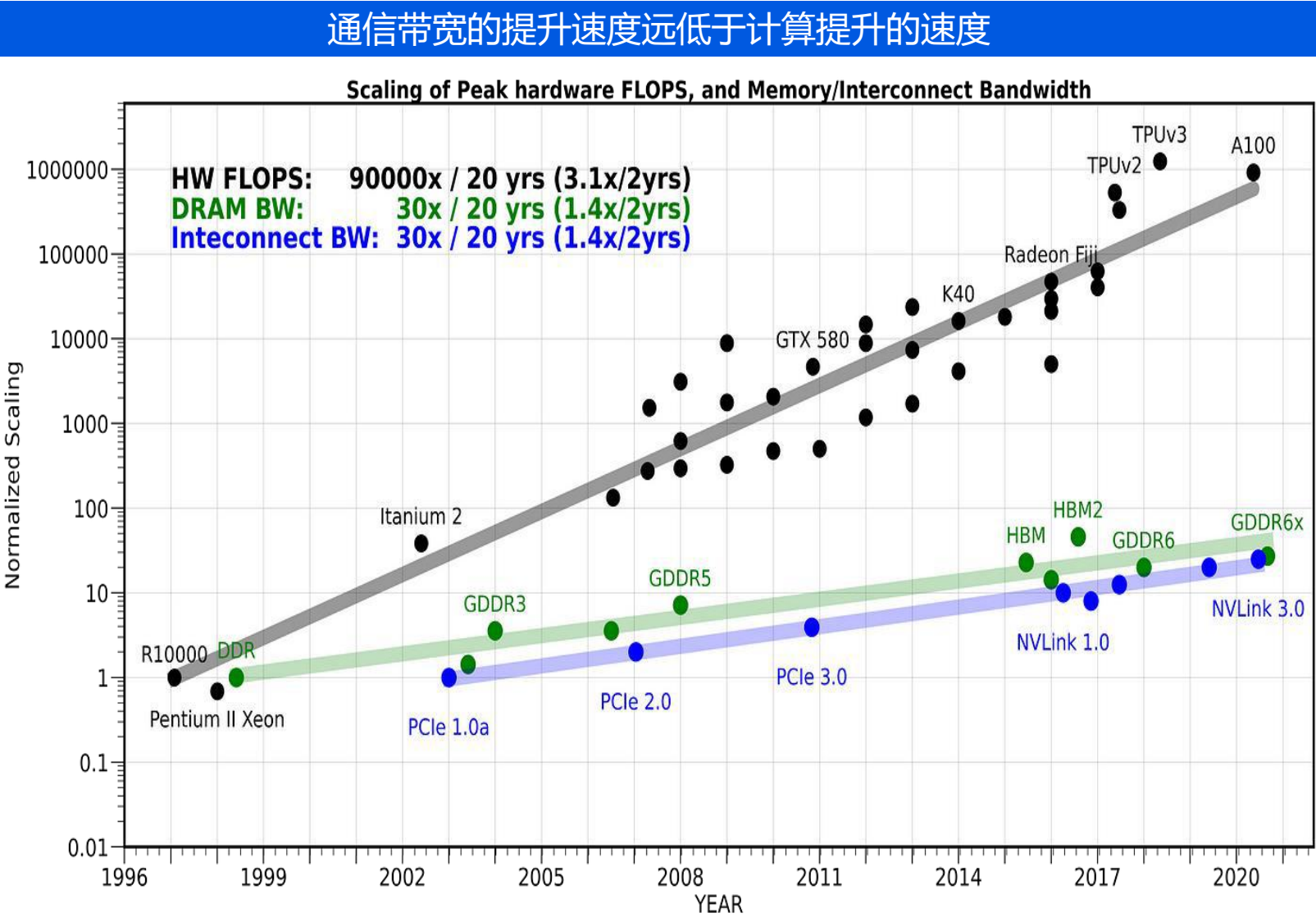


2021年全球及中国交换机行业市场份额情况



4.1 网络：算力的瓶颈之一，英伟达布局InfiniBand

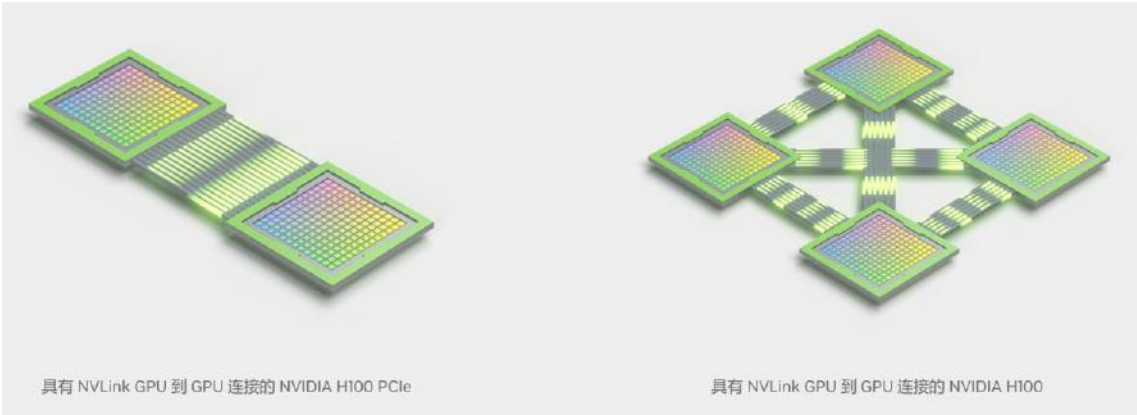
- **通信成为算力的瓶颈。** AI加速器通常会简化或删除其他部分，以提高硬件的峰值计算能力，但是却难以解决在内存和通信上的难题。无论是芯片内部、芯片间，还是AI加速器之间的通信，都已成为AI训练的瓶颈。
- **扩展带宽的技术难题还尚未被攻克。** 过去20年间，运算设备的算力提高了90,000倍，虽然存储器从DDR发展到GDDR6x，接口标准从PCIe1.0a升级到NVLink3.0，但是通讯带宽的增长只有30倍。



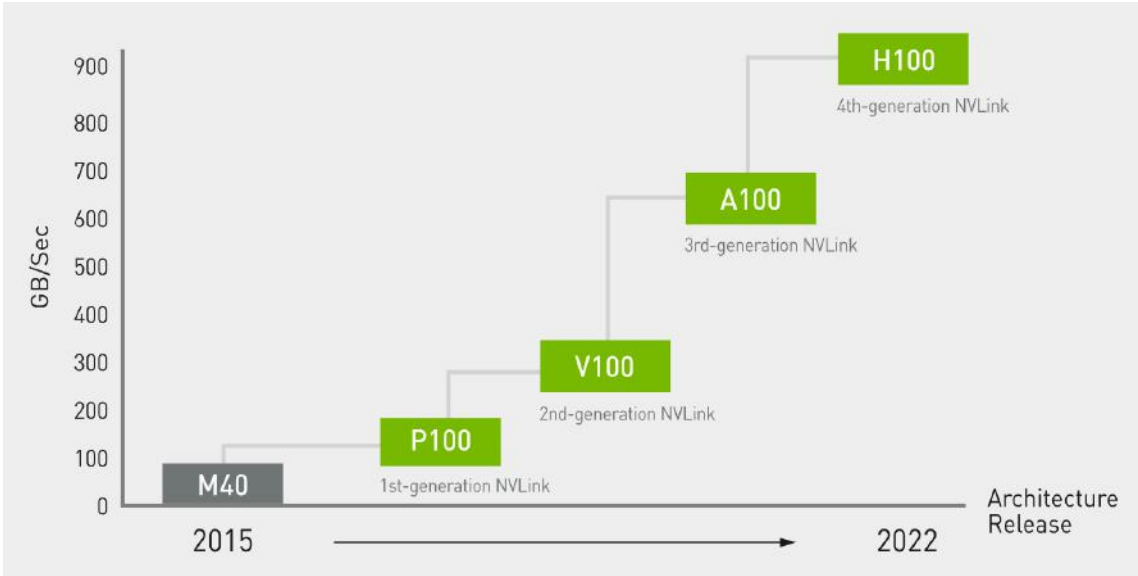
4.1 网络：算力的瓶颈之一，英伟达布局InfiniBand

- 英伟达NVLink。NVLink是 NVIDIA 的高带宽、高效能、低延迟、无损的 GPU 到 GPU 互连技术，其中包含诸如链路级错误检测和数据包回放机制等弹性特性，可保证数据的成功传输。
- 对比上一代，第四代NVLink可将全局归约操作的带宽提升 3 倍，通用带宽提升 50%，单个NVIDIA H100 Tensor Core GPU最多支持18个NVLink连接，多 GPU IO 的总带宽为 900GB/s，是 PCIe 5.0 的 7 倍。

NVLink 链接图



NVLink 性能表现



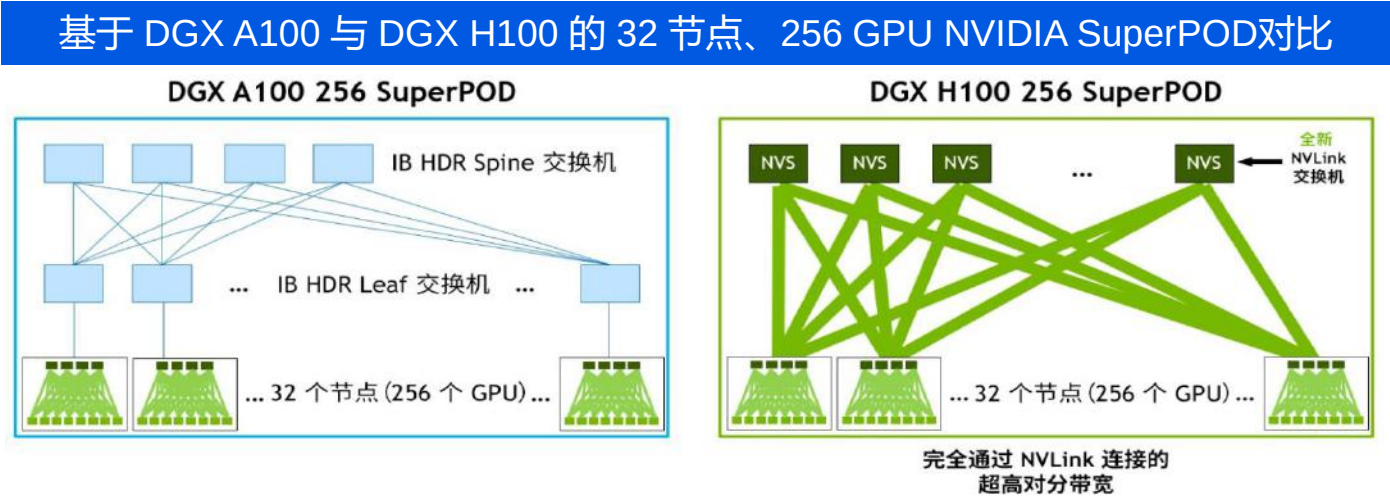
NVLink 规格

	第二代	第三代	第四代
每个 GPU 的 NVLink 带宽	300GB/秒	600GB/秒	900GB/秒
每个 GPU 的最大链接数	6个	12	18
支持的 NVIDIA 架构	NVIDIA Volta™ 架构	NVIDIA 安培架构	NVIDIA Hopper™ 架构

4.1 网络：算力的瓶颈之一，英伟达布局InfiniBand

- 英伟达NVSwitch。第三代 NVSwitch 技术包括位于节点内部和外部的交换机，用于连接服务器、集群和数据中心环境中的多个 GPU。节点内的每个 NVSwitch 具有 64 个第四代 NVLink 链路端口，可加速多 GPU 连接。交换机总吞吐量从上一代的 7.2Tb/s 提升到 13.6Tb/s。新的第三代NVSwitch 技术还通过组播和 NVIDIA SHARP 在网计算，为集合运算提供硬件加速。
- 英伟达结合全新 NVLINK 和 NVSwitch 技术，构建大型NVLink Switch 系统网络，实现前所未有的通信带宽水平。NVLink Switch 系统最多可支持 256 个 GPU。互连节点能够提供 57.6 TB 的多对多带宽，可提供高达 1 exaFLOP 级别的 FP8 稀疏计算算力。

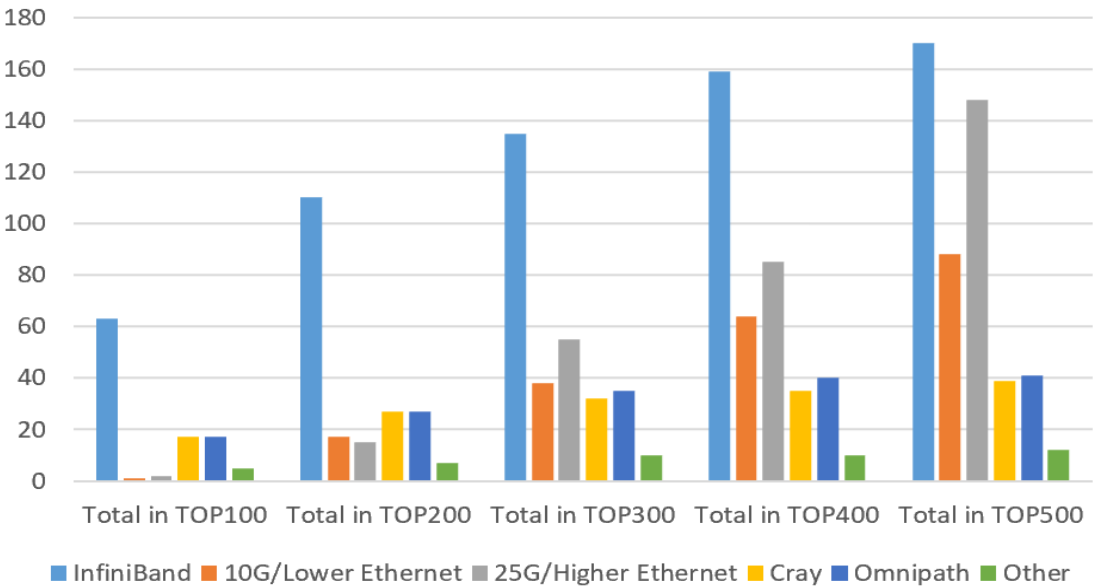
NVSwitch规格参数			
	第一代	第二代	第三代
直连GPU数量/节点	最多 8 个	最多 8 个	最多 8 个
NVSwitch GPU 到 GPU 带宽	300GB/秒	600GB/秒	900GB/秒
总聚合带宽	2.4TB/秒	4.8TB/秒	7.2TB/秒
支持的 NVIDIA 架构	NVIDIA Volta 架构	NVIDIA安培架构	NVIDIA Hopper 架构



4.1 网络：算力的瓶颈之一，英伟达布局InfiniBand

- InfiniBand是一个用于高性能计算的计算机网络通信标准，特点是高带宽、低延迟，主要用于高性能计算（HPC）、高性能集群应用服务器和高性能存储。
- 2019年3月，英伟达以69亿美金收购Mellanox，加大InfiniBand投入。2021年，发布NVIDIA Quantum-2，采用第七代 NVIDIA InfiniBand 架构，为AI 开发者和科学研究人员提供超强网络性能和丰富功能。

InfiniBand广泛应用于全球超算中心



NVIDIA Quantum-2 InfiniBand 交换机系列

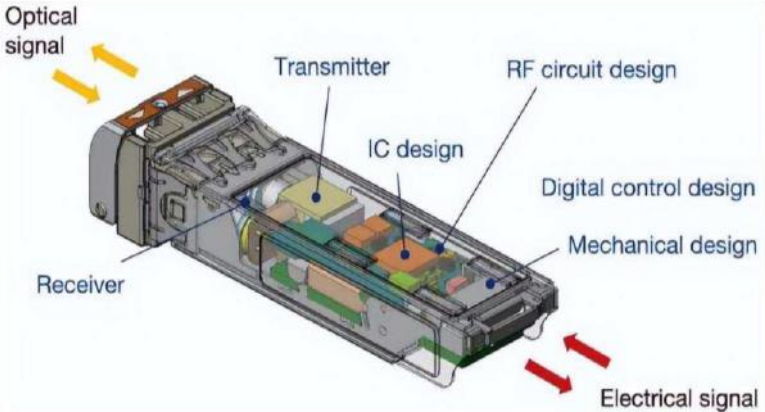


NVIDIA Quantum-2 InfiniBand 交换机可提供海量吞吐、出色的网络计算能力、智能加速引擎、杰出的灵活性和健壮架构，在高性能计算 (HPC)、AI 和超大规模云基础设施中发挥出色性能，并为用户降低成本和系统复杂性。

InfiniBand与万兆以太网的比较

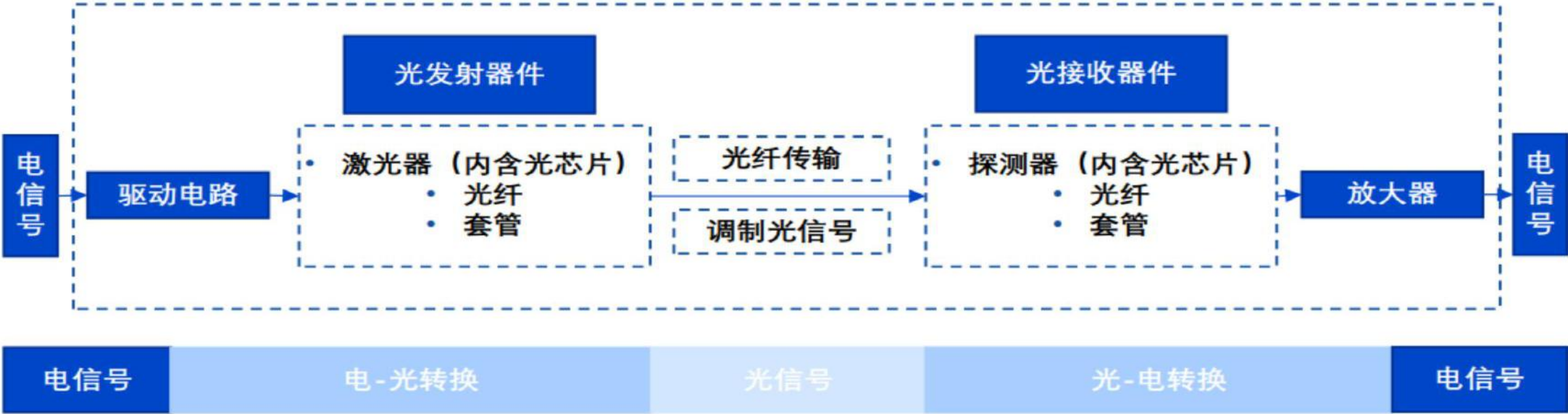
	InfiniBand(12X)	万兆以太网
带宽	30G,60G,120G,168G,312G	10G
延迟	小于等于1微秒	接近10微秒
应用领域	超级计算企业存储领域	互联网、域域网、数据中心骨干等
优点	极低的延迟和高吞吐量	应用范围广、已成为业界普遍认可的标准互联技术
缺点	再服务器硬件上需要昂贵的专有互联设备	延迟难以进一步降低

4.2 光模块：网络核心器件，AI训练提振800G需求



- 光模块是光纤通信系统的核心器件之一，其为多种模块类别的统称，包括：光接收模块，光发送模块，光收发一体模块和光转发模块等。
- 通常情况下，光模块由光发射器件（TOSA，含激光器）、光接收器件（ROSA，含光探测器）、驱动电路、放大器和光（电）接口等部分组成。
- 光模块主要用于实现电-光和光-电信号的转换。

光模块：光电转换示意图



4.2 光模块：网络核心器件，AI训练提振800G需求

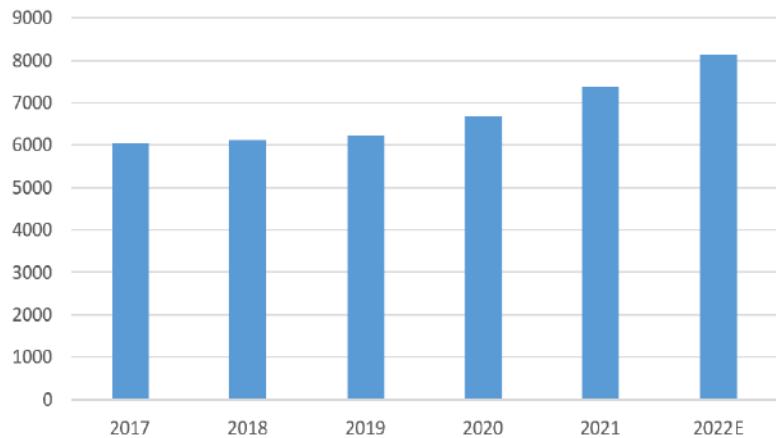
- 光模块行业的上游主要包括光芯片、电芯片、光组件企业。光组件行业的供应商较多，但高端光芯片和电芯片技术壁垒高，研发成本高昂，主要由境外企业垄断。光模块行业位于产业链的中游，属于封装环节。光模块行业下游包括互联网及云计算企业、电信运营商、数据通信和光通信设备商等。



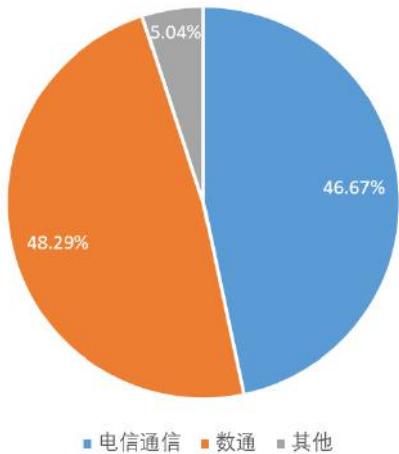
4.2 光模块：网络核心器件，AI训练提振800G需求

- 作为信息化和互连通信系统中必需的核心器件，光通信模块的发展对 5G 通信、电子、大数据、互联网行业的影响至关重要。全球数据流量的增长，光通信模块速率的提升，光通信技术的创新等推动光模块产业规模持续增长。全球光模块市场 Lightcounting 预测，全球光模块的市场规模在未来 5 年将以 CAGR11%保持增长，2027 年将突破 200 亿美元。另外，高算力、低功耗是未来市场的重要发展方向，CPO、硅光技术或将成为高算力场景下“降本增效”的解决方案。
- 光模块应用场景主要可以分为数据通信和网络通信两大领域。数据通信领域主要是指互联网数据中心以及企业数据中心。网络通信主要包括光纤接入网、城域网/骨干网以及5G接入、承载网为代表的移动网络应用。

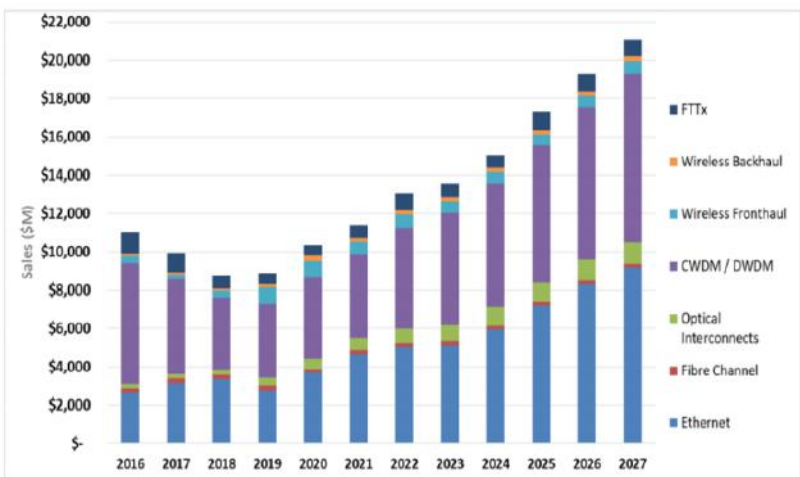
2017-2022全球光模块市场规模统计预测（百万美元）



2020年我国光模块应用市场结构



全球光模块细分市场规模及预测



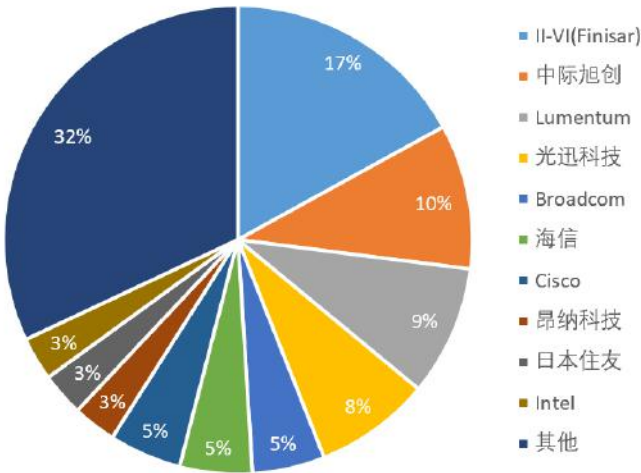
4.2 光模块：网络核心器件，AI训练提振800G需求

- **2010年至2021年，全球前十家光模块厂商中，中国企业增长至5家。**Omdia发布的全球前十大光模块厂商名单及其2021年市场份额变动情况显示，前十名分别为：高意集团、中际旭创、朗美通、光迅科技、博通、海信宽带多媒体、**Acacia**、昂纳集团、住友电工、英特尔。国内入围的厂商有中际旭创、光迅科技、海信宽带多媒体、昂纳集团，前四大国内光模块厂商占据全球的26%市场份额。
- **海关数据显示，2017-2021年中国光模块行业贸易顺差额逐年增长。**2017年我国光模块行业贸易顺差额为14.85亿美元，其中进口额为10.80亿美元，出口额为25.65亿美元;2021年光模块行业贸易顺差额增长至33.23亿美元，其中进口额为8.77亿美元，出口额为42.10亿美元。

全球前十大光模块厂商

Ranking of Top 10 Transceiver Suppliers				
2010	2016	2018	2021	
Finisar	Finisar	1	Finisar	II-VI & Innolight (tie)
Opnext	Hisense	2	Innolight	
Sumitomo	Accelink	3	Hisense	Huawei (HiSilicon)
Avago	Acacia	4	Accelink	Cisco (Acacia)
Source Photonics	FOIT (Avago)	5	FOIT (Avago)	Hisense
Fujitsu	Oclaro	6	Lumentum/Oclaro	Broadcom (Avago)
JDSU	Innolight	7	Acacia	Eoptolink
Emcore	Sumitomo	8	Intel	Accelink
WTD	Lumentum	9	AOi	Molex
NeoPhotonics	Source Photonics	10	Sumitomo	Intel

2021年全球光模块市场份额



2017-2021年中国光模块行业进出口状况表 (亿美元)

项目	2017年	2018年	2019年	2020年	2021年
进口额	10.80	8.93	9.94	9.22	8.77
出口额	25.65	28.99	30.51	36.90	42.10
进出口额	36.45	37.92	40.45	46.12	50.87
贸易顺差	14.85	20.06	20.57	27.67	33.23

4.2 光模块：网络核心器件，AI训练提振800G需求

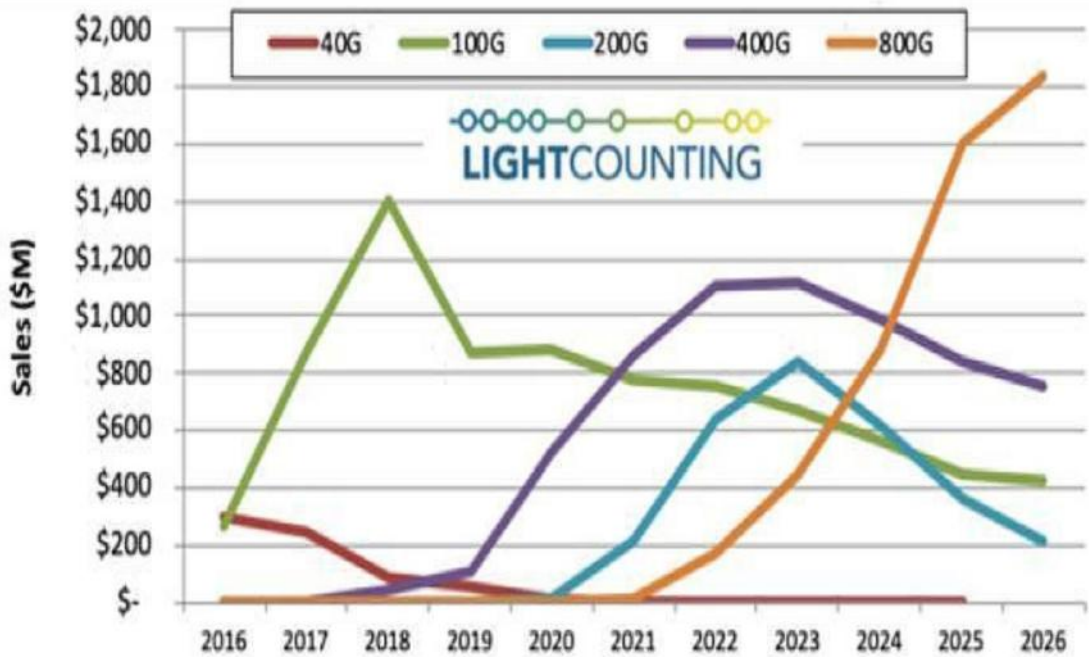
光模块：光模块行业上市公司基本信息对比

企业简称	业务/产品类型	光模块业务概况	2022 年 光 模 块 销 量 (万 块)	同比增速
光迅科技	光电子器件、模块和子系统	数据中心 800G 高速产品、固网接入 combo PON 系列产品、400G DCO 系列产品等均取得较好进展。	21075.27	-8.57%
中际旭创	高端光通信收发模块、光器件	800G 和相干系列产品等已实现批量出货，1.6T 光模块和 800G 硅光模块已开发成功并进入送测阶段，CPO（光电共封）技术和 3D 封装技术也在持续研发进程中。	946	-9.13%
新易盛	光模块	已成功研发出涵盖 5G 前传、中传、回传的25G、50G、100G、200G 系列光模块产品并实现批量交付。前已成功推出800G 的系列高速光模块产品，基于硅光解决方案的 800G、400G 光模块产品及 400G ZR/ZR+相干光模块产品、以及基于LPO 方案的 800G 光模块产品。	759	-2.57%
联特科技	光模块	研发生产的不同型号光模块产品累计1000 余种，公司兼具产品研发和生产制造能力，拥有光芯片集成、光器件以及光模块的设计、生产能力，是国内少数可以批量交付涵盖 10G、25G、40G、50G、100G、200G、400G、800G 全系列光模块的厂商。	298	18.73%
博创科技	PLC光分路器、光收发模块等	坚持走光电结合和器件模块化、集成化、小型化的道路，专注于集成光电子器件的规模化应用，为电信传输网和接入网以及数据通信提供关键的光电子器件。产品包括PLC 光分路器PON光收发模块、AWG和VMUX、用于无线承载网的前传、中回传光收发模块，25G-400G bps 速率的光收发模块、AOC、DAC、ACC、CPO 等。	532.64	42.12%
德科立	光收发模块、光放大器、光传输子系统	在非相干光模块领域，25G 单波速率下，100G（4×25G）80km 产品传输距离业界领先；50G单波速率下，200G（4×50G）40km 产品传输距离业界领先；100G 单波速率下，400G（4×100G）10km/40km 产品传输距离业界领先。在相干光模块领域，公司最新研发的 100G/200G 相干光模块产品在国内市场仍处于技术领先水平。公司是行业内少数能够同时提供高速率相干与高速率非相干光模块的公司。	133.19	11.23%
剑桥科技	高速光模块	进一步优化 800G 8×FR1 产品和 800G 2×FR4 的工艺和良率并在上海工厂逐步量产，推动新一代低功耗、低成本 400G DR4/DR4+产品的生产导入和量产，对 100G CWDM4、100G LR4、100G DR/FR、400G FR4 进行持续的降本改进，加快 800G 产品大规模量产的步伐。	89.03	2.77%

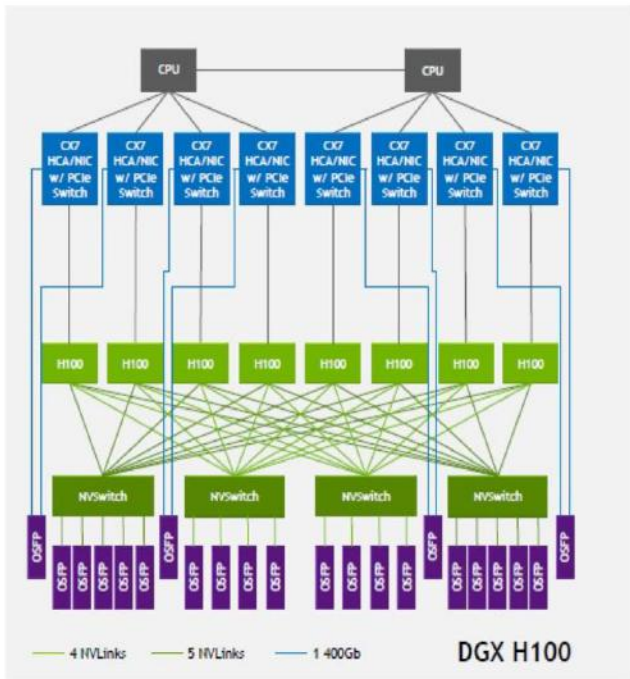
4.2 光模块：网络核心器件，AI训练提振800G需求

- AIGC 的高速发展将进一步促进数据流量的持续增长和包括光模块在内的 ICT 行业的发展，加速光模块向 800G 及以上产品迭代。LightCounting 数据显示，2023 年开始，800G 有望拉动新一轮增长，预计在 2026 年突破 30 亿美元大关。
- **预计英伟达 H100 GPU：800G光模块为1:2-1:4。** 英伟达采用的InfiniBand无阻塞、胖树架构。我们预计算力网络中，服务器层GPU与800G光模块比例为1:1；交换机层GPU与800G光模块比例为1:2；此外，考虑核心层交换机，以及管理网络、存储网络等，以及安装率相关因素，预计**英伟达 H100 GPU：800G光模块的比例约为1:2-1:4。**

数据中心光模块需求演进（2021-2026年数据为预测值）



DGX H100 数据网络配置图



DGX H100: DATA-NETWORK CONFIGURATION

Full-BW Intra-Server NVLink

- All 8 GPUs can simultaneously saturate 18 NVLinks to other GPUs within server
- Limited only by over-subscription from multiple other GPUs

Half-BW NVLink Network

- All 8 GPUs can half-subscribe 18 NVLinks to GPUs in other servers
- 4 GPUs can saturate 18 NVLinks to GPUs in other servers
- Equivalent of full-BW on AllReduce with SHARP
- Reduction in All2All BW is a balance with server complexity and costs

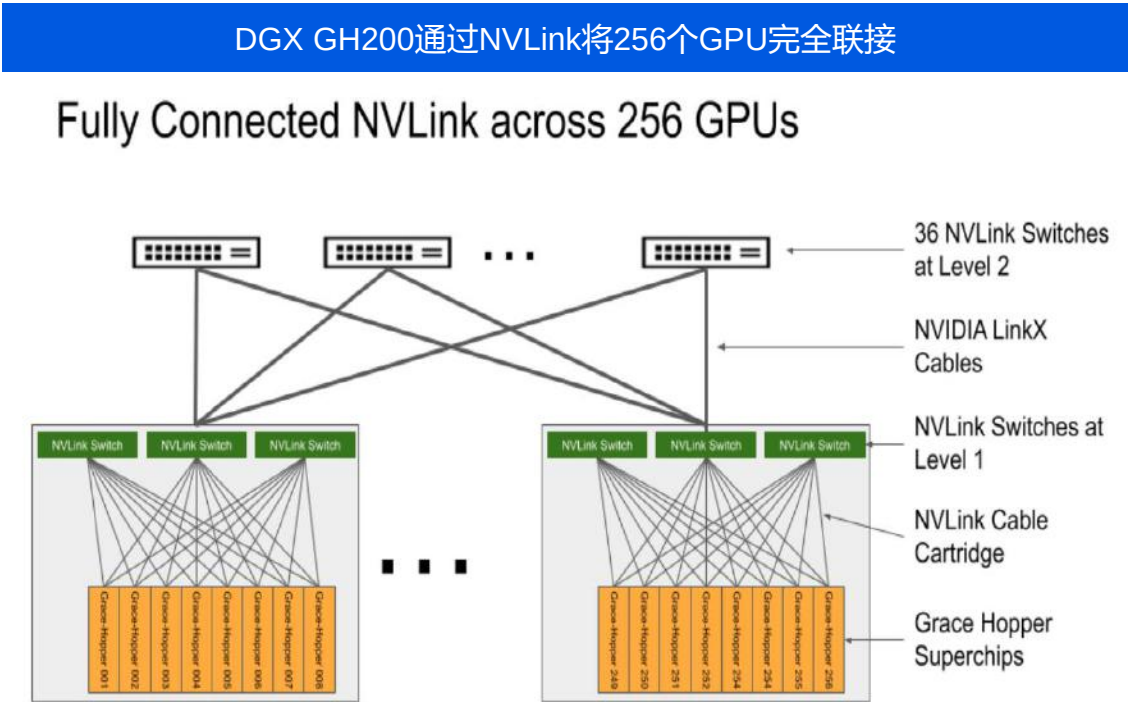
Multi-Rail InfiniBand/Ethernet

- All 8 GPUs can independently RDMA data over its own dedicated 400 Gb/s HCA/NIC
- 800 Gbps of aggregate full-duplex to non-NVLink Network devices

4.2 光模块：网络核心器件，AI训练提振800G需求

- 2023年5月，英伟达推出DGX GH200，GH200是将 256 个NVIDIA Grace Hopper超级芯片完全连接,旨在处理用于大规模推荐系统、生成式人工智能和图形分析的太字节级模型。
- NVLink交换系统采用两级、无阻塞、胖树结构。如下图：L1和L2层分为96和32台交换机，承载Grace Hopper超级芯片的计算底板使用NVLink fabric第一层的定制线缆连接到NVLink交换机系统。LinkX电缆扩展了NVLink fabric的第二层连接。我们预计GH200的推出将进一步促进800G光模块的需求增长。

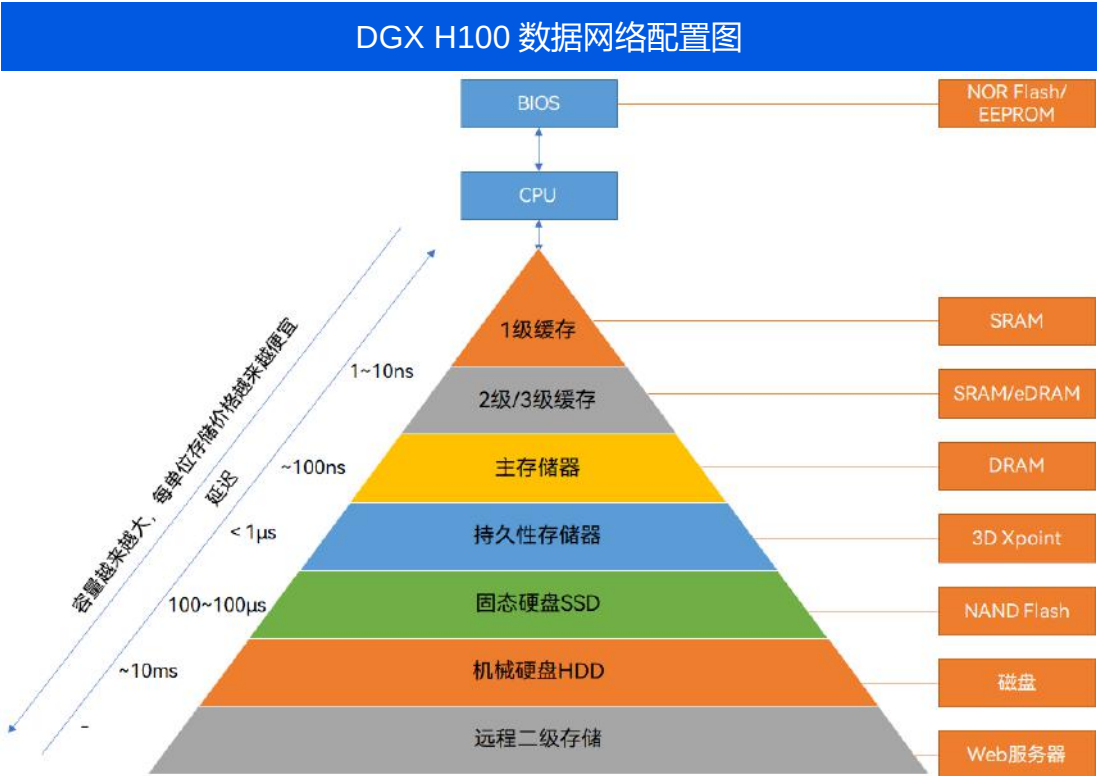
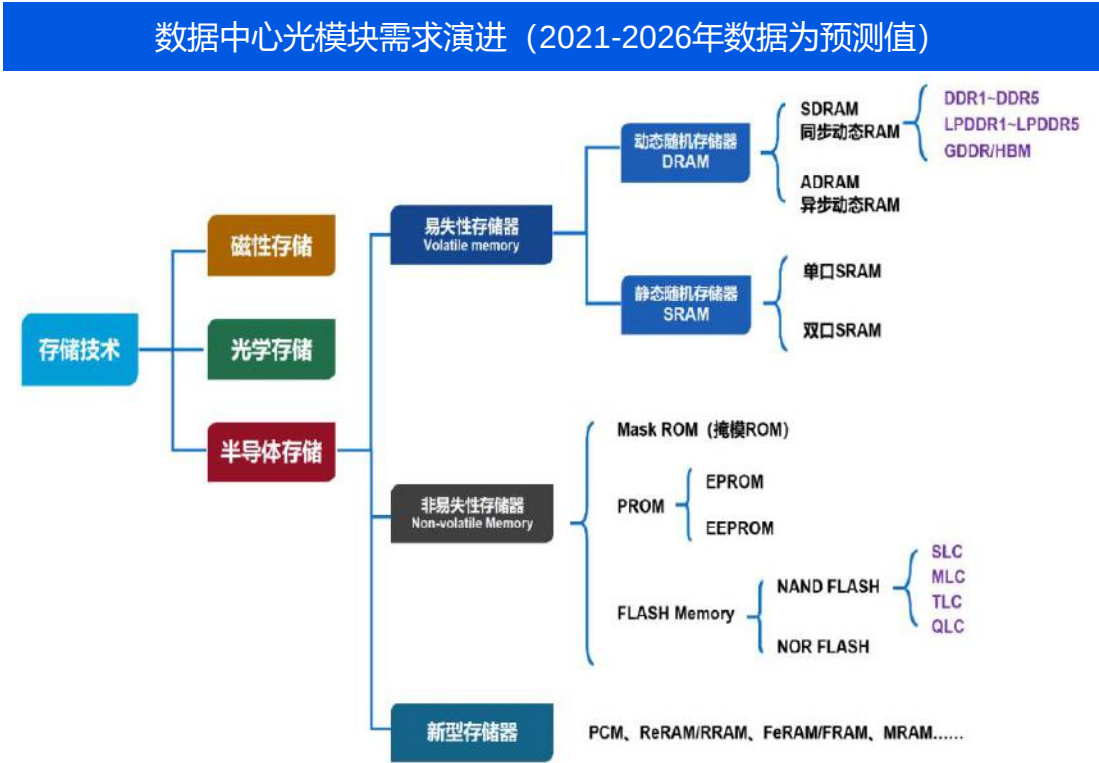
DGX GH200技术参数	
CPU and GPU	256x NVIDIA Grace Hopper Superchips
CPU Cores	18,432 Arm® Neoverse V2 Cores with SVE2 4X 128b
GPU Memory	144TB
Performance	1 exaFLOPS
Networking	256x OSFP single-port NVIDIA ConnectX®-7 VPI with 400Gb/s InfiniBand
	256x dual-port NVIDIA BlueField®-3 VPI with 200Gb/s InfiniBand and Ethernet
	24x NVIDIA Quantum-2 QM9700 InfiniBand Switches
	20x NVIDIA Spectrum™ SN2201 Ethernet Switches 22x NVIDIA Spectrum SN3700 Ethernet Switches
NVIDIA NVLink Switch System	96x L1 NVIDIA NVLink Switches
	36x L2 NVIDIA NVLink Switches
Management Network	Host baseboard management controller (BMC) with RJ45
Software	NVIDIA AI Enterprise (optimized AI software)
	NVIDIA Base Command (orchestration, scheduling, and cluster management)
	DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky (operating system)
Support	Comes with three-year business-standard hardware and software support



5. 存储：人工智能“内存墙”，3D工艺持续突破

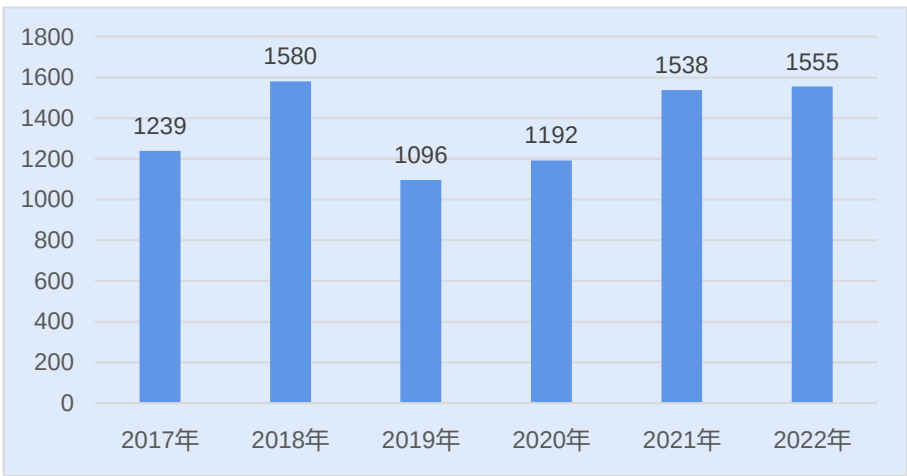
5.1 存储：半导体产业核心支柱，AI算力的“内存墙”

- 计算机存储器是一种利用半导体、磁性介质等技术制成的存储资料的电子设备。其电子电路中的资料以二进制方式存储，不同存储器产品中基本单元的名称也不一样。
- 存储芯片可分为掉电易失和掉电非易失两种，其中易失存储芯片主要包含静态随机存取存储器（SRAM）和动态随机存取存储器（DRAM）；非易失性存储器主要包括可编程只读存储器（PROM），闪存存储器（Flash）和可擦除可编程只读寄存器（EPROM/EEPROM）等。NAND Flash和DRAM存储器领域合计占半导体存储器市场比例达到95%以上。

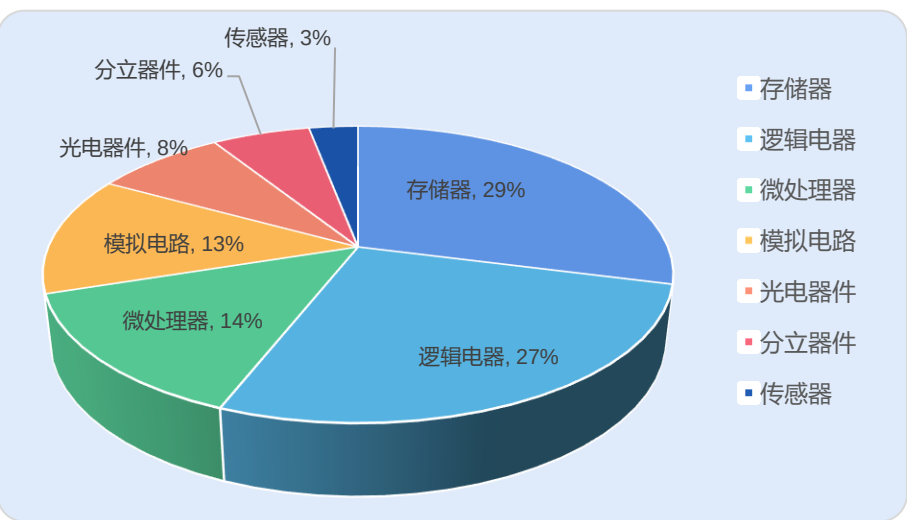


5.1 存储：半导体产业核心支柱，AI算力的“内存墙”

2017-2022年全球存储器行业市场规模（单位：亿美元）



2021年全球半导体细分行业结构



- 半导体存储器也是整个半导体产业的核心支柱之一。半导体行业分为集成电路、光电器件、分立器件、传感器等子行业，根据功能的不同，集成电路又可以分为存储器、逻辑电路、模拟电路、微处理器等细分领域。根据世界半导体贸易统计（WSTS）对世界半导体贸易规模的最新报告，2021年全球半导体行业的整体规模达到5529.61亿美元，同比增长25.6%。其中存储器的市场规模接近1600亿美元，是半导体中规模最大的子行业，占比超过1/4。

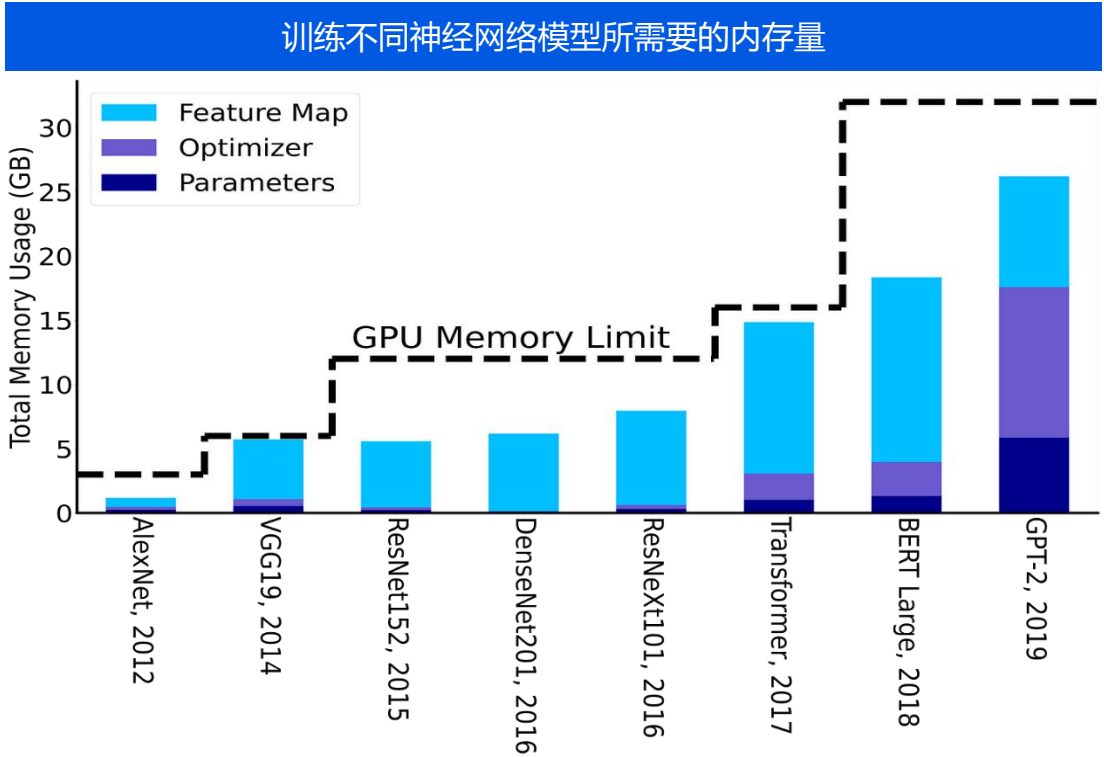
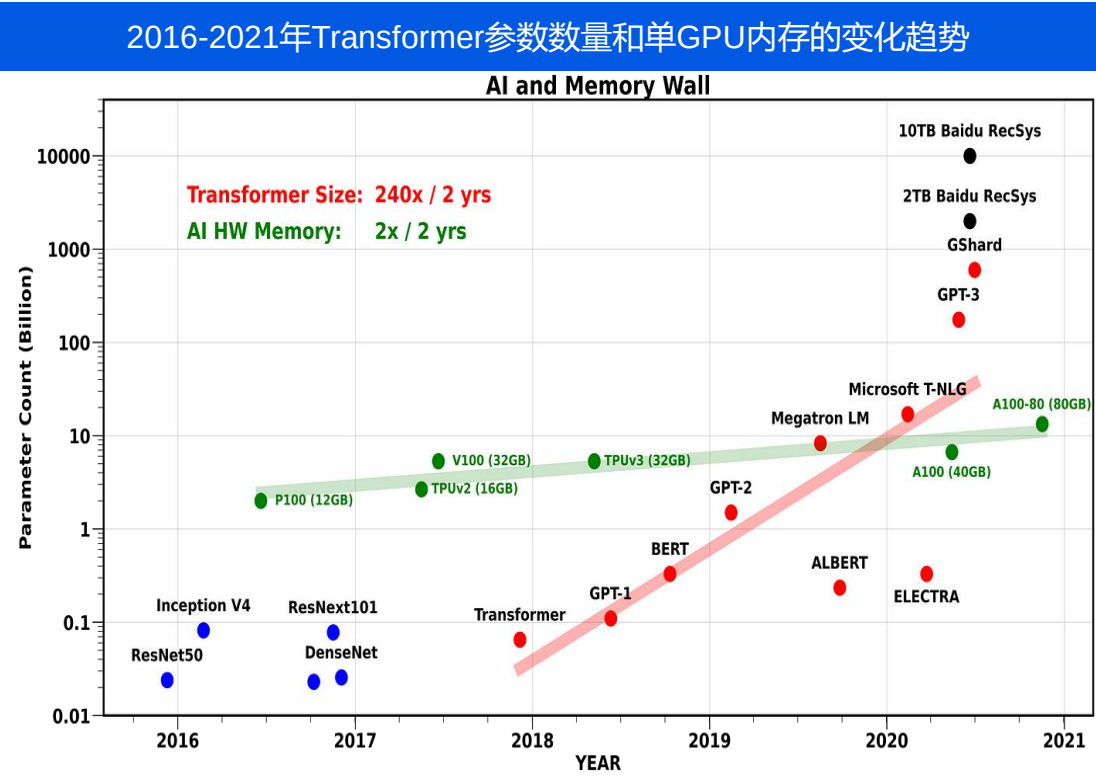
存储芯片产业链



5.1 存储：半导体产业核心支柱，AI算力的“内存墙”

“内存墙” (memory wall)是制约AI算力提升的重要因素。

- Transformer模型中的参数数量呈现出2年240倍的超指数增长，而单个GPU内存仅以每2年2倍的速度扩大。
- 训练AI模型的**内存需求，通常是参数数量的几倍**。因为训练需要存储中间激活，通常会比参数（不含嵌入）数量增加3-4倍的内存。自谷歌团队在2017年提出Transformer，模型所需的内存容量开始大幅增长。



5.1 存储：半导体产业核心支柱，AI算力的“内存墙”

存算一体的原理、优势

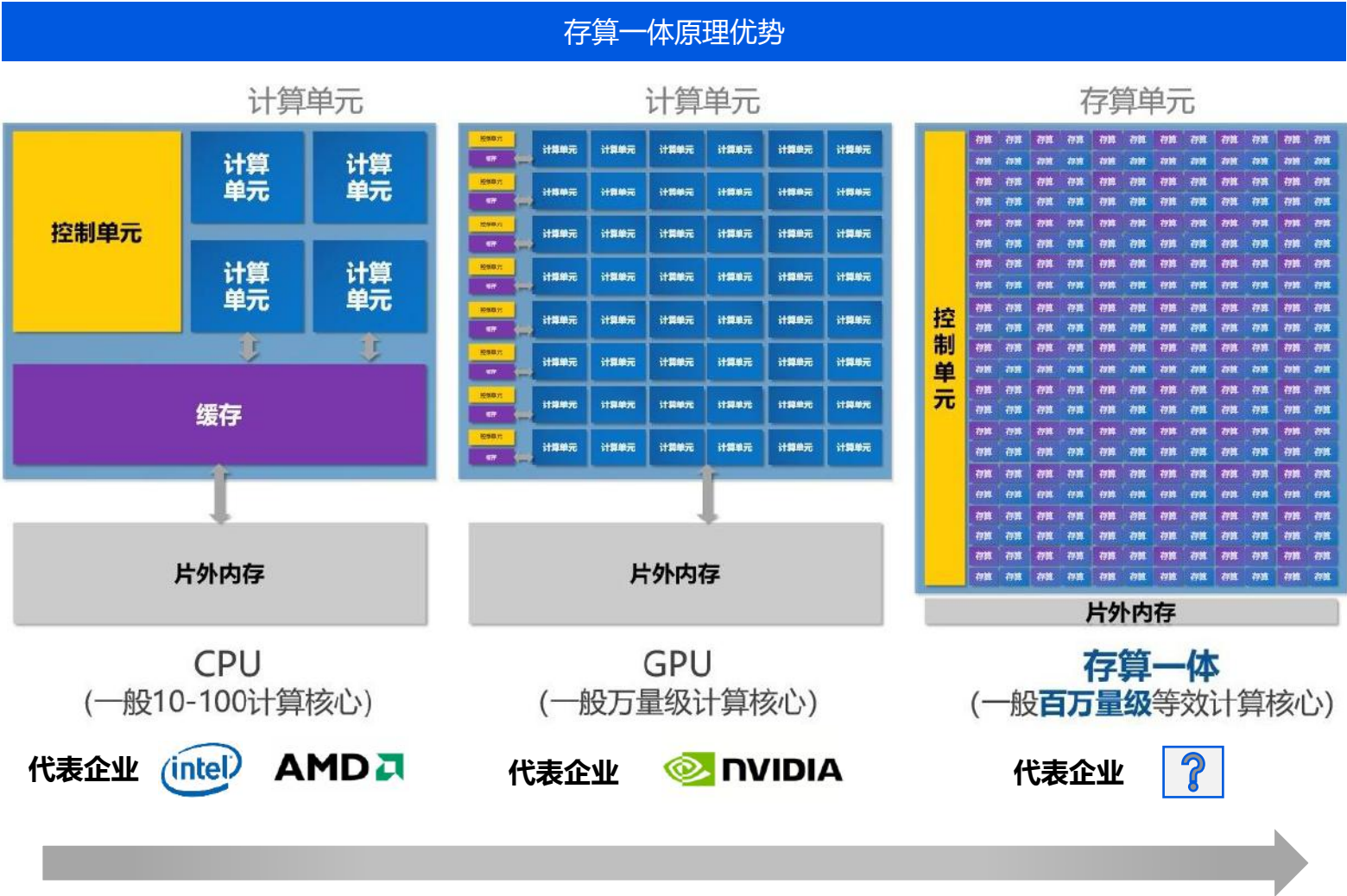
- 存算一体就是存储器中叠加计算能力，以新的高效运算架构进行二维和三维矩阵计算。

存算一体的优势包括：

- 具有更大算力（1000TOPS以上）
- 具有更高能效（超过10-100TOPS/W），超越传统ASIC算力芯片
- 降本增效（可超过一个数量级）

存算一体技术的技术底层特征包括：

- 减少数据搬运（降低能耗至1/10~1/100）
- 存储单元具备计算能力（等效于在面积不变的情况下规模化增加计算核心数，或者等效于提升工艺代）
- 单个存算单元替代“计算逻辑+寄存器”更小更快

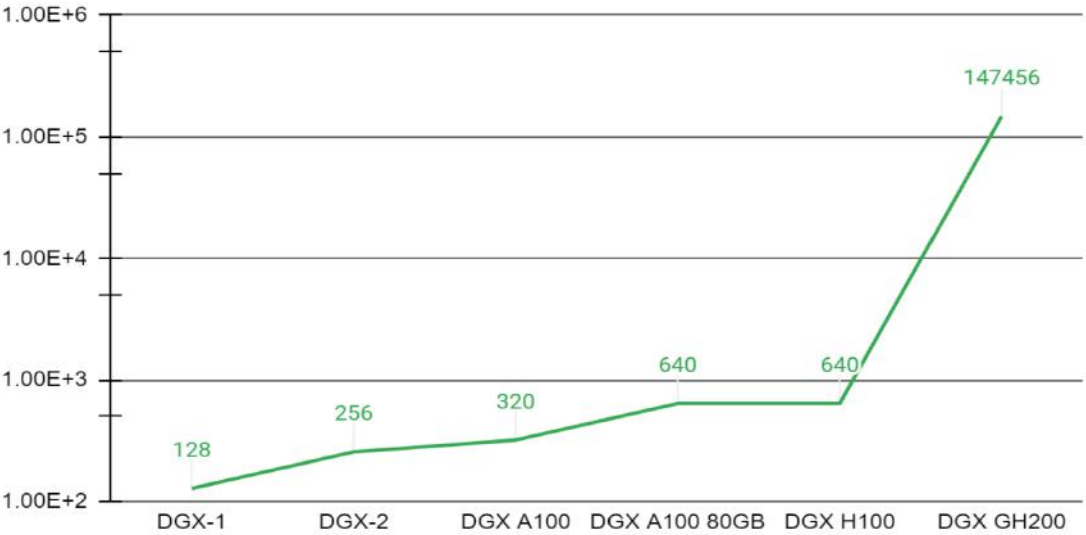


5.1 存储：半导体产业核心支柱，AI算力的“内存墙”

- NVIDIA DGX GH200 是第一台通过GPU的NVLink 联接实现144TB 内存的超级计算机。NVIDIA DGX GH200 通过 NVLink 为 GPU 共享内存编程模型提供了近 500 倍的内存，形成了一个巨大的数据中心大小的 GPU。NVIDIA DGX GH200 是第一台通过在 256 个NVIDIA Grace Hopper 超级芯片上提供 144TB 海量共享内存空间的 AI 超级计算机。
- NVIDIA DGX GH200中的每个NVIDIA Grace Hopper超级芯片具有480 GB LPDDR5 CPU内存，每GB的功率是 DDR5和96 GB fast HBM3的八分之一。NVIDIA Grace CPU和Hopper GPU通过NVLink互连，每个GPU都可以以 900GBps的速度访问其他GPU的内存和NVIDIA Grace CPU的扩展GPU内存。

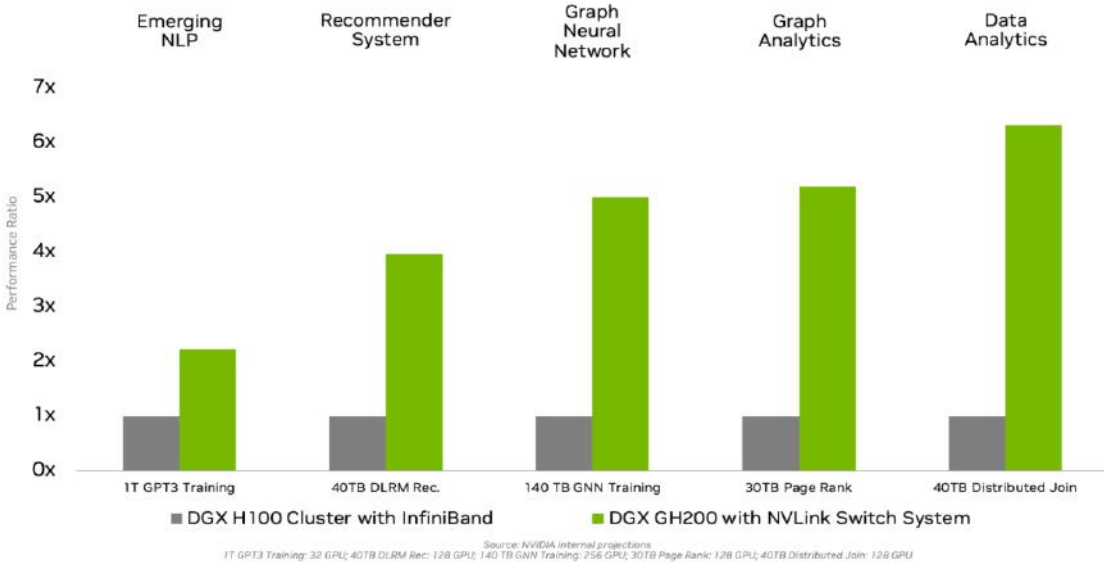
DGX每一代产品的 存储能力

GPU memory (in GB) over DGX generations



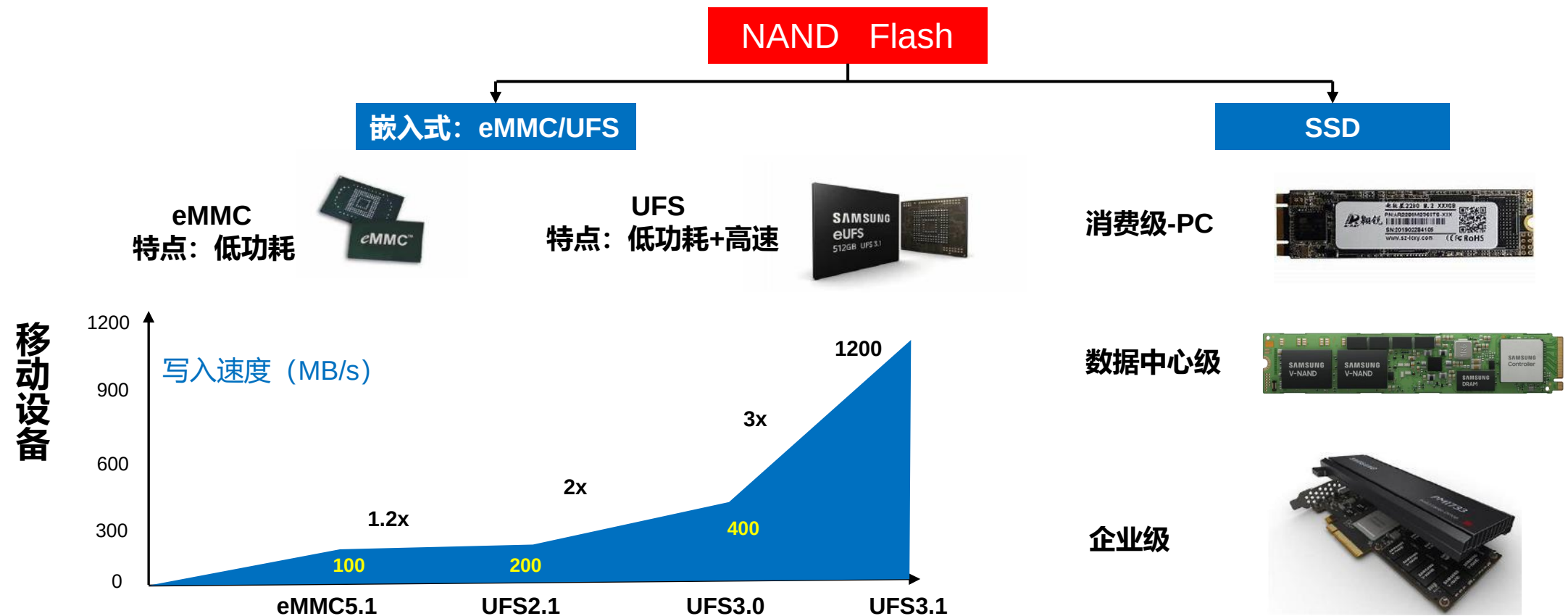
DGX每一代产品的 存储能力

DGX GH200 Fastest for Giant Memory Models



5.2 NAND：大容量存储的最佳方案，3D NAND技术持续突破

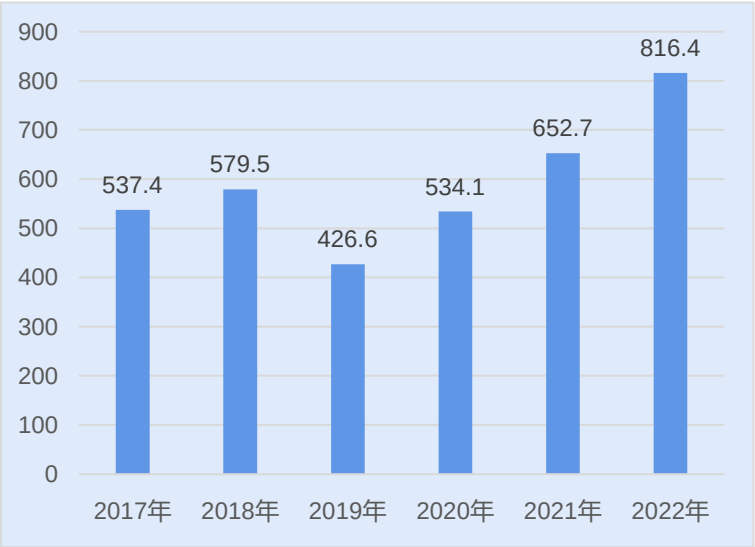
- NAND基本原理：非易失性存储，通过外部施加电压控制存储单元中的电荷量，实现电荷在内存单元中的存储。
- NAND优势：非易失、读写速度快、抗震、低功耗、体积小、价格较低等；NAND挑战：耐久性，高密度、高容量。
- NAND Flash出货形态以 eMMC/UFS（主要应用在移动设备、智能手机、平板电脑等）、SSD（主要应用在服务器和PC）产品为主。主要应用在数码相机、MP3随身听记忆卡、体积小巧的U盘等。



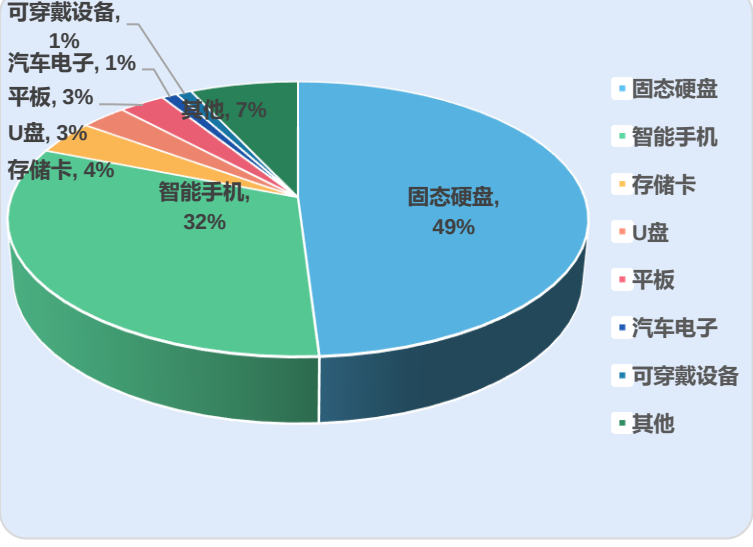
5.2 NAND：大容量存储的最佳方案，3D NAND技术持续突破

- NAND Flash是大容量存储器当前应用最广和最有效的解决方案。据Gartner统计,NAND Flash2020年市场规模为534.1亿美元。随着人工智能、物联网、大数据、5G等新兴应用场景不断落地,电子设备需要存储的数据也越来越庞大,NAND Flash需求量巨大,市场前景广阔。
- 目前全球具备NAND Flash晶圆生产能力的主要有三星、铠侠、西部数据、美光、SK海力士、英特尔等企业,国产厂商长江存储处于起步状态,正在市场份额与技术上奋起直追。根据Omdia的数据统计,2020年六大NAND Flash晶圆厂占据了98%的市场份额。

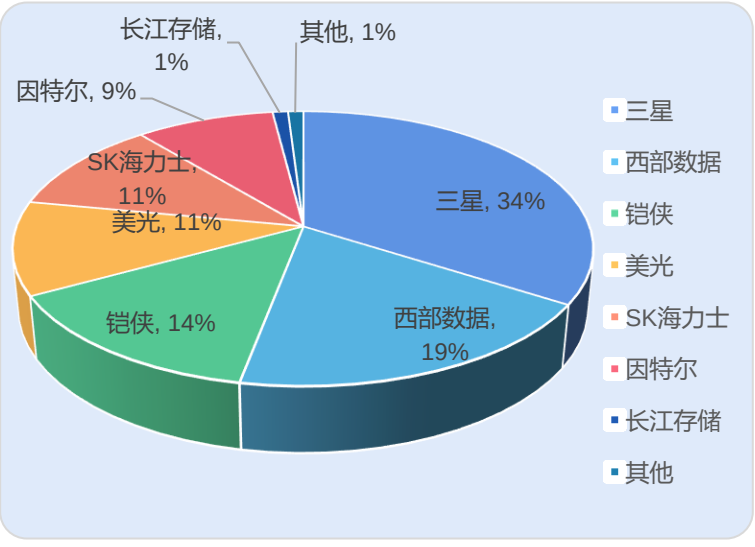
2017-2022年NAND Flash市场规模 (单位: 亿美元)



全球DAND Flash行业需求应用统计



2020年全球NAND Flash晶圆场市场份额



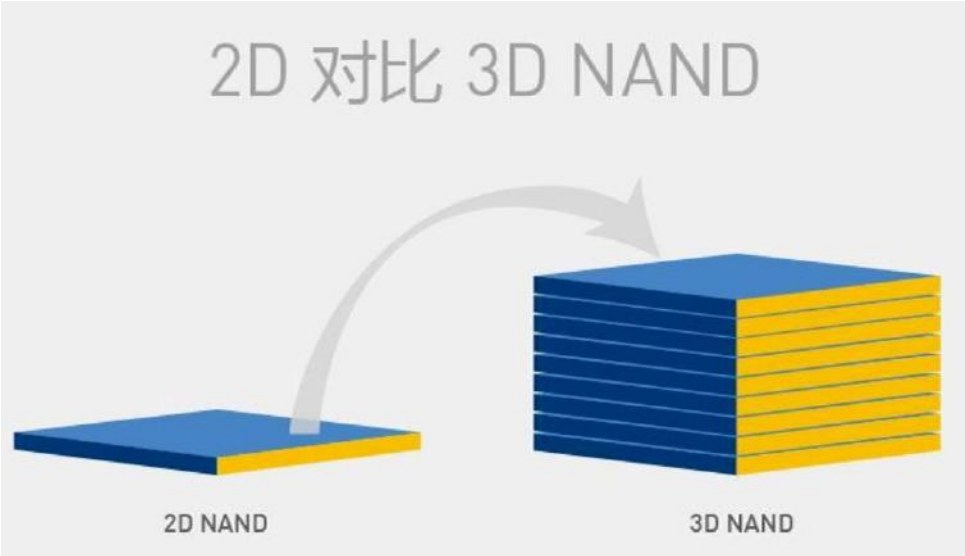
5.2 NAND：大容量存储的最佳方案，3D NAND技术持续突破

全球主要NAND厂商的堆叠层数



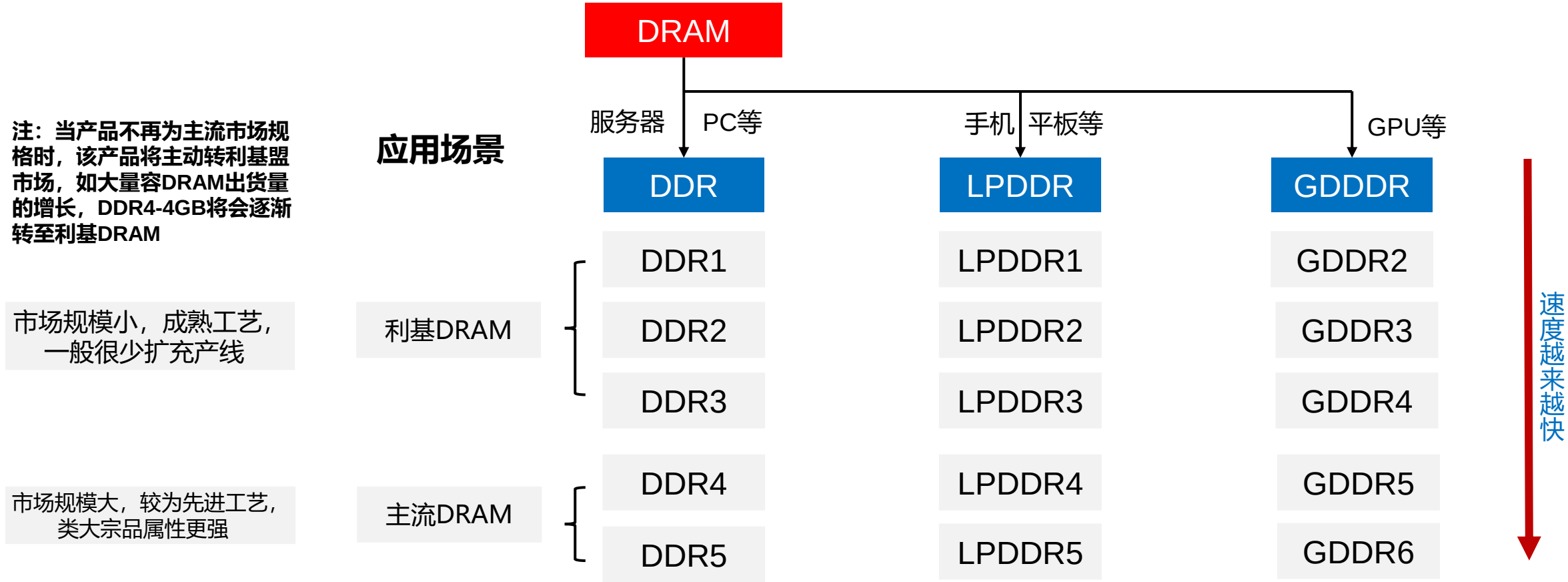
	2020	2021		2022		2023
	2H	1H	2H	1H	2H	1H
SAMSUNG	128L V6			176L V7 2021H2 MP		224L 2023
KIOXIA Western Digital	112L BiCS5 2020 MP			162L BiCS6 2022 MP		212L 2023
Micron	128L TLC 2020 MP	176L TLC 2021 MP		232L 2022Q3 MP		
SK hynix	128L TLC 2020Q3 MP	176L TLC 2021Q4 MP		238L 2022~2023		
长江存储	64L TLC		128L TLC/QLC 2021 MP		232L 2022~2023	

- NAND最核心的是数据的稳定性。根据厂家测试，在做到15、16nm 工艺之后，NAND 闪存的可靠性就会断崖性下跌。不能一味地通过先进制程，缩小晶体管体积、提高密度的方式来提升的性能。
- 全球存储巨头都在投入3D NAND。长江存储发布独家 3D 堆叠技术Xtacking。



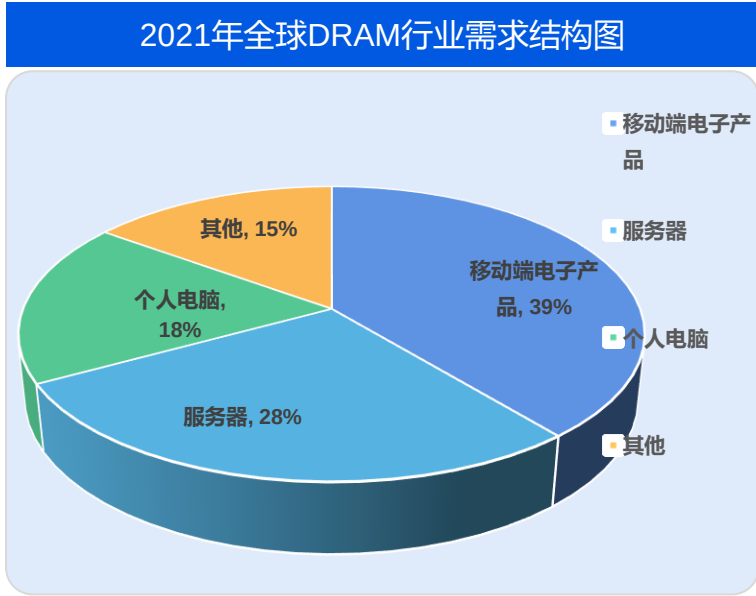
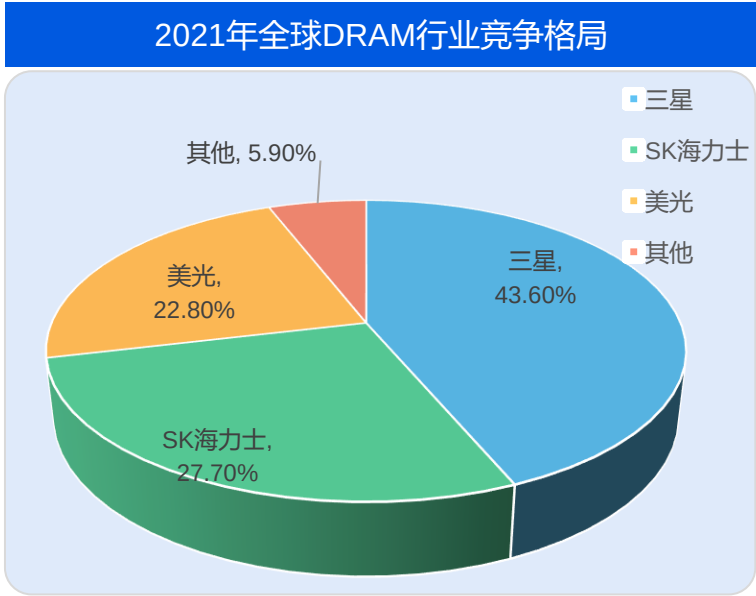
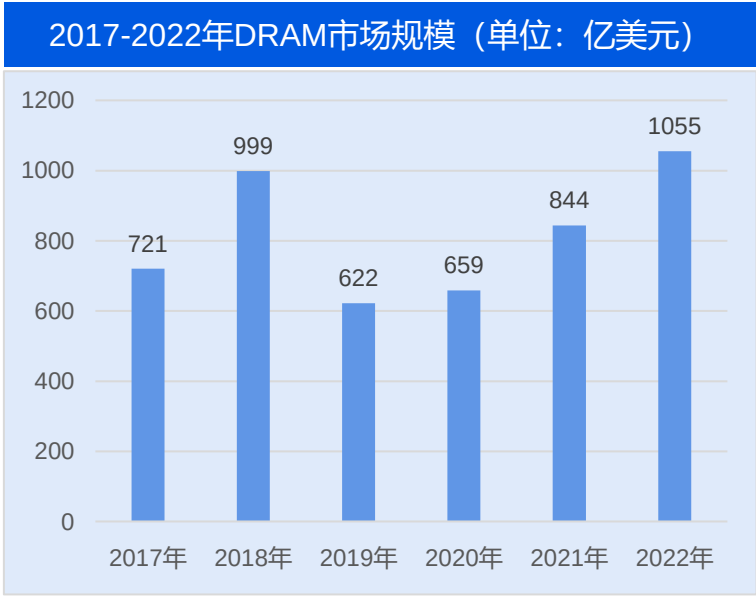
5.3 DRAM：存储器最大细分市场，3D成为重要方向

- DRAM（Dynamic Random Access Memory，动态随机存取存储器）是一种半导体存储器，主要的作用原理是利用电容内存储电荷的多寡来代表一个二进制比特（bit）是1还是0。DRAM根据应用设备可分为计算机（DDR）、移动（LPDDR）、图形存储器DRAM（GDDR），DDR和LPDDR合计占DRAM应用比例约90%。
- DRAM优势：体积容量高、成本低、高密度、结构简单； DRAM挑战：访问速度慢、耗电量大。



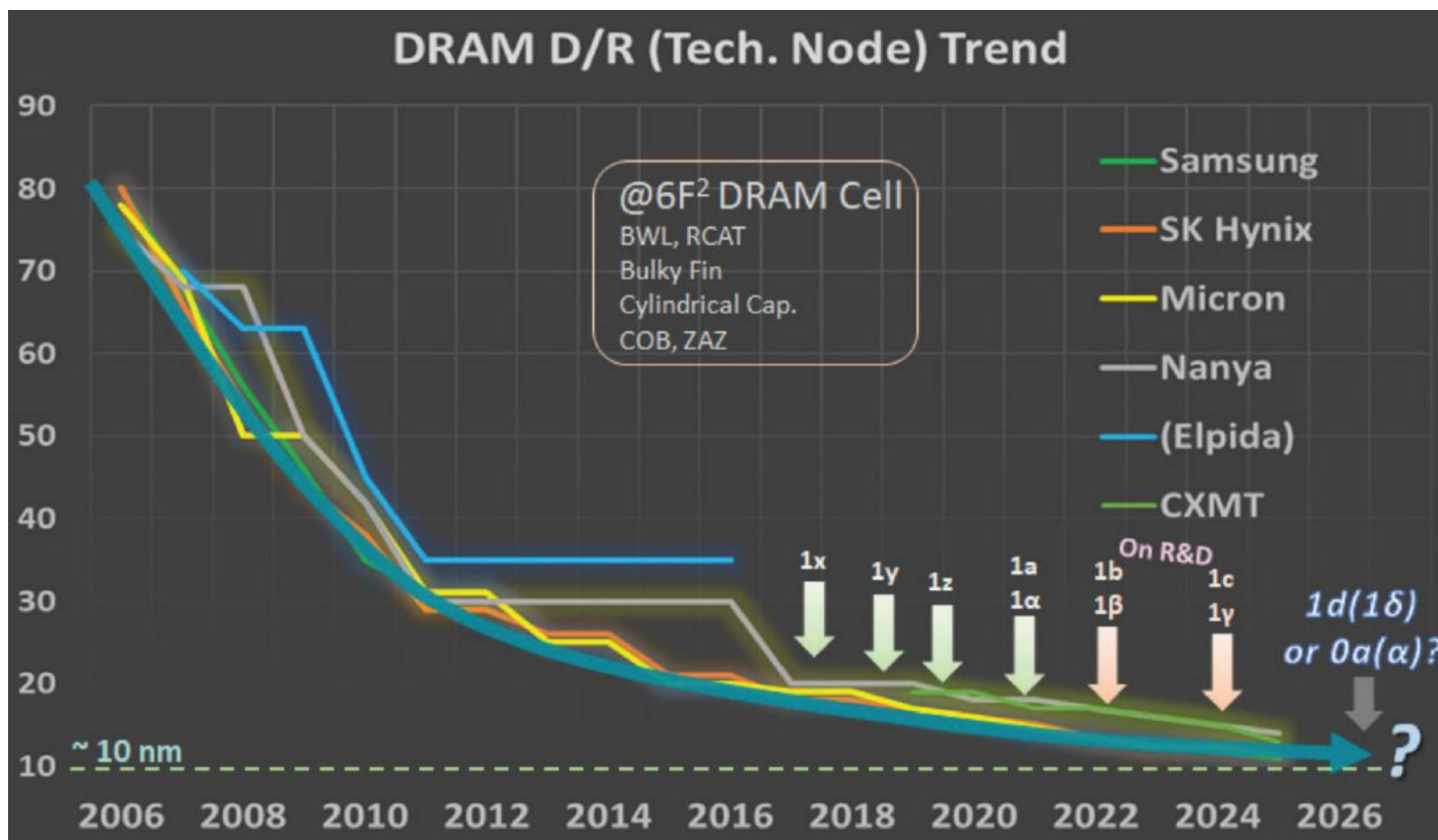
5.3 DRAM：存储器最大细分市场，3D成为重要方向

- DRAM是存储器市场规模最大的芯片，根据Trend Force数据统计，2022年DRAM市场规模预计达到1055亿美元。
- 市场格局：目前DRAM晶圆的市场供应主要集中在三星、SK海力士和美光，三大厂商2021年市场占有率合计已达到94.1%，其中三星市场占有率43.6%，SK海力士与美光分别占比27.7%与22.8%。国内DRAM晶元厂商主要为合肥长鑫，目前尚处于起步阶段。
- 下游行业：根据Gartner统计及预测，DRAM下游需求市场格局较为稳定，移动端电子产品、服务器、个人电脑占为主，个人电脑占比近年来呈现缓慢下降的趋势。



5.3 DRAM：存储器最大细分市场，3D成为重要方向

厂商的DRAM路线图：D1z-D1z DRAM在2020-2022年上市

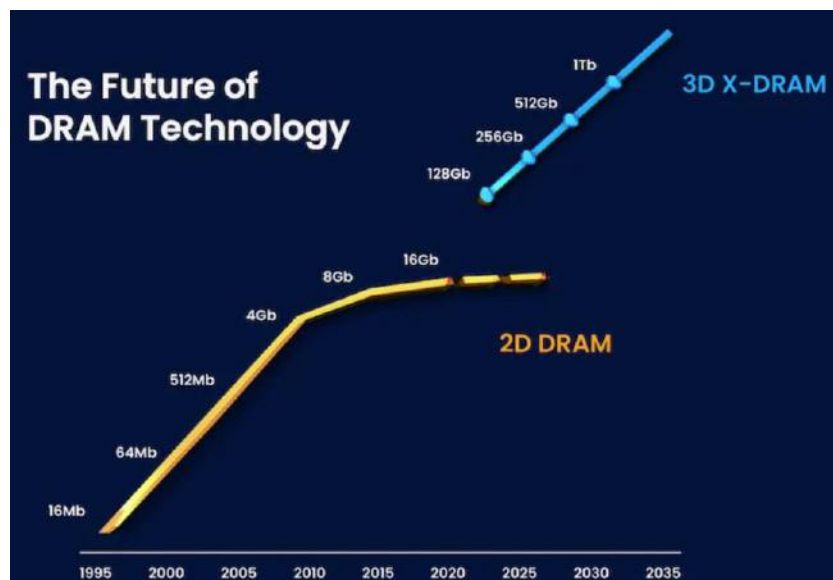


- DRAM停滞在10nm。DRAM 容量取决于 DRAM 的物理尺寸和其中的存储单元数量。存储单元密度受单元尺寸、用于构建芯片的层的厚度以及运行芯片所需的功率量的限制。由于多项技术挑战，DRAM小型化的速度已经放缓。自 2016 年以来，由于电容器尺寸和低于 10 纳米水平的其他电气限制， DRAM 尺寸停滞在 10 纳米。
- 现有的DRAM技术将可能在2026年逼近终点。现有6F2结构DRAM单元设计下，10nm的D/R可能是2027年或2028年成为DRAM的最后一代。

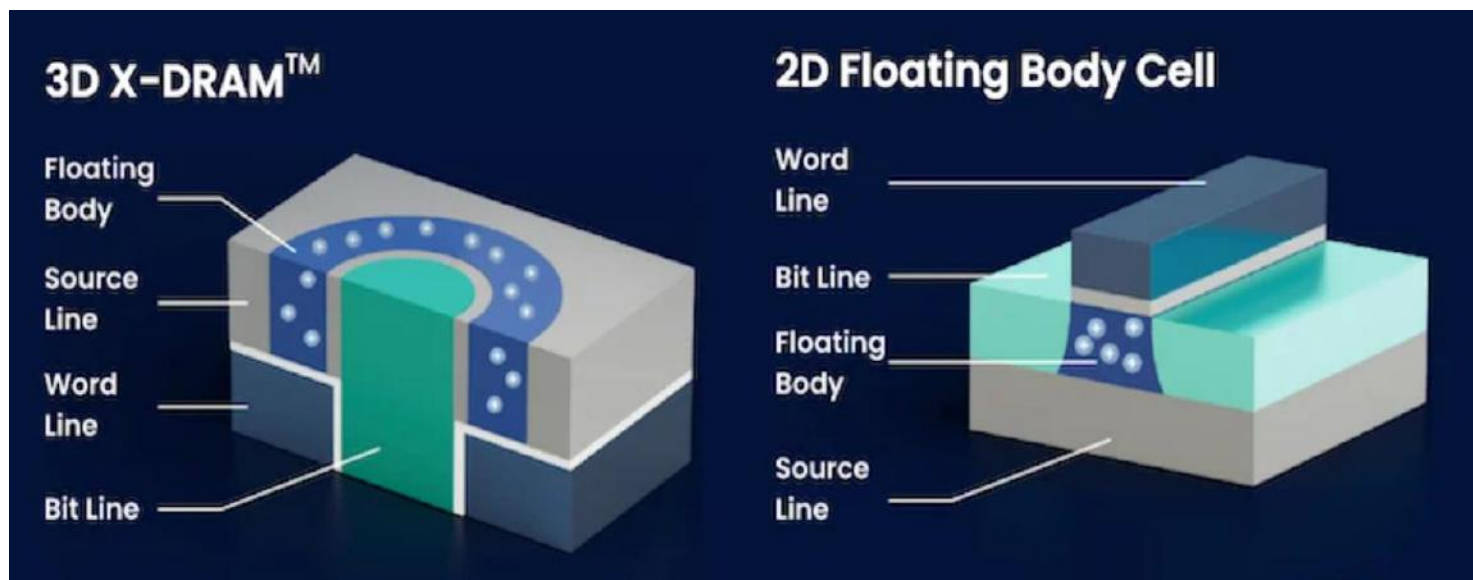
5.3 DRAM：存储器最大细分市场，3D成为重要方向

- 3D DRAM 视为存储器半导体市场的游戏规则改变者，它或将在未来 3 到 4 年内克服超精细工艺的局限性，发展3D DRAM 架构对聊天 GPT 和人工智能 (AI) 的发展是必要的。
- Neo Semiconductor 推出3D X-DRAM，通过垂直堆叠存储单元以增加存储容量，而不会增加存储芯片的物理占用空间。根据公司估计，这项技术可以通过 230 层实现 128 Gb 的密度——是当今 DRAM 密度的八倍。
- 三星电子、SK海力士和美光将 3D DRAM 视为未来将改变内存市场游戏规则的关键技术，正在根据各种路线图加速研究。

Neo Semiconductor 预测的 DRAM 技术的未来

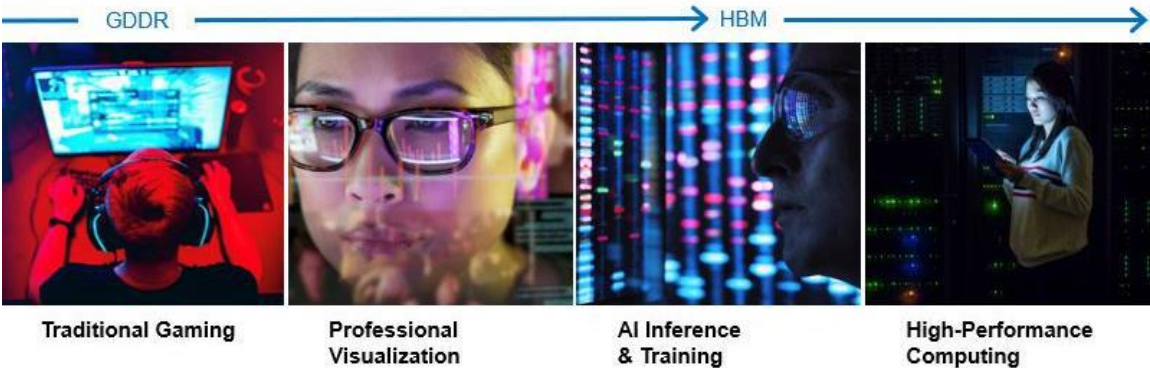


3D X-DRAM概念图



5.3 DRAM：存储器最大细分市场，3D成为重要方向

HBM更适用于AI、高性能计算等场景



- HBM（High Bandwidth Memory）高带宽存储器，是一种面向需要极高吞吐量的数据密集型应用程序的DRAM。
- HBM特点：更高带宽、更多I/O数量、更低功耗、更小尺寸。
HBM挑战：灵活性不足、容量小、访问延迟高。
- 超高的带宽让HBM成为了高性能GPU的核心组件。根据TrendForce报告，目前市场上主要的HBM制造商为SK 海力士、三星、美光，市占率分别为50%、40%、10%。

HBM性能演进

- 专为图形应用开发
- 现使用于多个应用领域，包括：
 - AI/ML, HPC, 网络



目前推出的搭载HBM和GDDR的GPU产品

Graphics Card Name	Memory Technology	Memory Speed	Memory Bus	Memory Bandwidth	Release
AMD Radeon R9 Fury X	HBM1	1.0 Gbps	4096-bit	512 GB/s	2015
NVIDIA GTX 1080	GDDR5X	10.0 Gbps	256-bit	320 GB/s	2016
NVIDIA Tesla P100	HBM2	1.4 Gbps	4096-bit	720 GB/s	2016
NVIDIA Titan XP	GDDR5X	11.4 Gbps	384-bit	547 GB/s	2017
AMD RX Vega 64	HBM2	1.9 Gbps	2048-bit	483 GB/s	2017
NVIDIA Titan V	HBM2	1.7 Gbps	3072-bit	652 GB/s	2017
NVIDIA Tesla V100	HBM2	1.7 Gbps	4096-bit	901 GB/s	2017
NVIDIA RTX 2080 Ti	GDDR6	14.0 Gbps	384-bit	672 GB/s	2018
AMD Instinct MI100	HBM2	2.4 Gbps	4096-bit	1229 GB/s	2020
NVIDIA A100 80GB	HBM2e	3.2 Gbps	5120-bit	2039 GB/s	2020
NVIDIA RTX 3090	GDDR6X	19.5 Gbps	384-bit	936 GB/s	2020
AMD Instinct MI200	HBM2e	3.2 Gbps	8192-bit	3200 GB/s	2021
NVIDIA RTX 3090 Ti	GDDR6X	21.0 Gbps	384-bit	1008 GB/s	2022
AMD Instinct MI300	HBM3	满血6.4 Gbps	8192-bit	>5000 GB/s	E2023
NVIDIA H100	HBM3	满血6.4 Gbps	8192-bit	>5000 GB/s	E2023

6. 投资建议和风险提示

◆ 投资建议

ChatGPT对算力的影响远不止当前可见的基建投入，未来Transformer大模型的迭代推动模型训练相关需求的算力增长，以及AIGC大模型应用的算力需求，将是算力市场不断超预期的源泉。相关公司：

1、计算

1) 服务器：浪潮信息、中科曙光、紫光股份、工业富联、纬创、广达、英业达、戴尔、联想集团、超威电脑、中国长城、神州数码、拓维信息、四川长虹；

2) GPU：英伟达、AMD、Intel、海光信息、寒武纪、龙芯中科、景嘉微；

2、网络

1) 网络设备：紫光股份、中兴通讯、星网锐捷、深信服、迪普科技、普天科技、映翰通；

2) 光模块：中际旭创、新易盛、光迅科技、华工科技、联特科技、剑桥科技、天孚通信；

3、存储

1) 存储器：紫光国微、江波龙、北京君正、兆易创新、澜起科技、东芯股份、聚辰股份、普冉股份、朗科科技。

6.1 相关公司

业务	股票名称	归母净利润（亿元）				PE			归母净利润 CAGR
		2022A	2023E	2024E	2025E	2023E	2024E	2025E	2022-2025E
601138.SH	工业富联	200.73	237.26	266.70	299.99	17.87	15.90	14.14	14.33%
000977.SZ	浪潮信息	20.80	26.48	32.91	38.40	25.14	20.23	17.34	22.67%
603019.SH	中科曙光	15.44	20.16	26.10	32.85	37.19	28.71	22.82	28.61%
000938.SZ	紫光股份	21.58	26.99	33.13	39.99	34.66	28.24	23.39	22.83%
000066.SZ	中国长城	1.20	5.50	7.72	10.34	81.65	58.21	43.47	104.83%
688041.SH	海光信息	8.04	13.02	18.94	25.03	139.95	96.19	72.78	46.05%
688256.SH	寒武纪-U	-12.57	-8.13	-5.59	-3.38	-112.84	-164.25	-271.04	-35.42%
688047.SH	龙芯中科	0.52	2.07	3.94	4.50	270.40	142.10	124.29	105.65%
300474.SZ	景嘉微	2.89	4.09	5.75	7.59	103.28	73.44	55.65	37.96%
300223.SZ	北京君正	7.89	8.99	12.06	14.83	47.25	35.24	28.66	23.40%
603986.SH	兆易创新	20.53	13.36	19.25	25.50	51.36	35.66	26.91	7.51%
688008.SH	澜起科技	12.99	12.07	19.89	24.71	54.80	33.24	26.76	23.89%
300308.SZ	中际旭创	12.24	15.05	19.86	24.91	64.81	49.13	39.16	26.72%
300502.SZ	新易盛	9.04	9.06	12.59	16.58	47.63	34.30	26.05	22.41%
002281.SZ	光迅科技	6.08	6.87	8.02	9.23	32.77	28.08	24.38	14.91%
300394.SZ	天孚通信	4.03	5.06	6.53	8.53	67.32	52.14	39.89	28.41%
000988.SZ	华工科技	9.06	12.14	15.58	19.38	29.83	23.23	18.68	28.83%
301205.SZ	联特科技	1.13	1.29	1.98	2.94	128.59	83.66	56.30	37.36%

- 1、**宏观经济影响下游需求：**宏观经济环境下行将影响政府、行业客户对数字基础设施的采购需求。
- 2、**大模型迭代不及预期：**未来大模型迭代速度不及预期将导致大模型训练对算力的需求增长放缓。
- 3、**大模型应用不及预期：**受监管等方面原因，如果大模型不能广泛向商业市场应用，其推理端算力需求的释放或放缓。
- 4、**市场竞争加剧：**新兴企业投入的GPU、网络、存储、等算力产业，将加剧市场竞争，影响企业的盈利能力。
- 5、**中美博弈加剧：**国际形势持续不明朗，美国不断通过“实体清单”等方式对中国企业实施打压，若中美紧张形势进一步升级，将可能导致中国半导体供应链供应受到影响。
- 6、**相关公司业绩不及预期：**市场环境变化、公司治理情况变化、其他非主营业务经营不及预期等原因或将导致相关公司的整体业绩不及预期。

计算机小组介绍

刘熹，计算机行业首席分析师，上海交通大学硕士，多年计算机行业研究经验，2021年新浪财经金麒麟新锐分析师（计算机行业）第三名主要成员。主要研究AI、数据要素、信创、网络安全、工业软件、智能网联、云计算、地理信息化等赛道。

分析师承诺

刘熹，本报告中的分析师均具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观的出具本报告。本报告清晰准确的反映了分析师本人的研究观点。分析师本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收取到任何形式的补偿。

国海证券投资评级标准

行业投资评级

推荐：行业基本面向好，行业指数领先沪深300指数；
中性：行业基本面稳定，行业指数跟随沪深300指数；
回避：行业基本面向淡，行业指数落后沪深300指数。

股票投资评级

买入：相对沪深300 指数涨幅20%以上；
增持：相对沪深300 指数涨幅介于10%~20%之间；
中性：相对沪深300 指数涨幅介于-10%~10%之间；
卖出：相对沪深300 指数跌幅10%以上。

免责声明

本报告的风险等级定级为R3，仅供符合国海证券股份有限公司（简称“本公司”）投资者适当性管理要求的客户（简称“客户”）使用。本公司不会因接收人收到本报告而视其为客户。客户及/或投资者应当认识到有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司的完整报告为准，本公司接受客户的后续问询。

本公司具有中国证监会许可的证券投资咨询业务资格。本报告中的信息均来源于公开资料及合法获得的相关内部外部报告资料，本公司对这些信息的准确性及完整性不作任何保证，不保证其中的信息已做最新变更，也不保证相关的建议不会发生任何变更。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。报告中的内容和意见仅供参考，在任何情况下，本报告中所表达的意见并不构成对所述证券买卖的出价和征价。本公司及其本公司员工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。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或者金融产品等服务。本公司在知晓范围内依法合规地履行披露义务。

风险提示

市场有风险，投资需谨慎。投资者不应将本报告为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向本公司或其他专业人士咨询并谨慎决策。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。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息。本报告不构成本公司向该机构之客户提供的投资建议。

任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

郑重声明

本报告版权归国海证券所有。未经本公司的明确书面特别授权或协议约定，除法律规定的情况外，任何人不得对本报告的任何内容进行发布、复制、编辑、改编、转载、播放、展示或以其他方式非法使用本报告的部分或者全部内容，否则均构成对本公司版权的侵害，本公司有权依法追究其法律责任。

国海证券 · 研究所 · 计算机研究团队

心怀家国，洞悉四海



国海研究上海

上海市黄浦区绿地外滩中心C1栋
国海证券大厦

邮编：200023

电话：021-61981300

国海研究深圳

深圳市福田区竹子林四路光大银行大厦28F

邮编：518041

电话：0755-83706353

国海研究北京

北京市海淀区西直门外大街168号腾达大厦25F

邮编：100044

电话：010-88576597