安信证券 ESSENCE SECURITIES

2023 年 03 月 21 日
半导体

行业分析

证券研究报告

AI 算力产业链梳理——技术迭代推动瓶颈突破，AIGC 场景增多驱动算力需求提升

投资评级 领先大市-A
首次评级

首选股票	目标价（元）	评级

行业表现

资料来源：Wind 资讯

升幅%	1M	3M	12M
相对收益	5.9	4.2	-10.2
绝对收益	4.0	4.3	-16.8

马良 分析师
SAC 执业证书编号：S1450518060001
maliang2@essence.com.cn

程宇婷 分析师
SAC 执业证书编号：S1450522030002
chengyt@essence.com.cn

相关报告

设备国产化关键环节，半导体零部件蓝海起航	2022-09-23
中芯国际拟再建新厂，持续推荐上游设备及材料	2022-08-30
国产替代渗透率提升+国产芯片开发需求增多，掩膜版行业进入高速增长通道	2022-08-26
晶圆平坦化的关键工艺，CMP 设备材料国产替代快速推进	2022-06-10
市场空间广阔，电池管理（BMS/BMIC）芯片国产替代进程加速	2022-05-19

AI 大模型引领应用层百花齐放，算力层长期受益：

ChatGPT、GPT4.0、Microsoft 365 Copilot、文心一言等相继发布，以 ChatGPT 为代表的 AI 大模型及其初步应用“一石激起千层浪”，其相关技术变革预计将对个体的工作、生活及社会组织方式带来的广泛影响。以海内外 IT 龙头为代表的企业界也开始深入挖掘此次技术变革对公司经营方式、商业模式的潜在颠覆性变化，并重新评估未来的发展战略。我们认为，AI 大模型在参数规模、计算量简化、安全性及多模态融合等方向虽然仍有迭代进步空间，但其迄今展示出的“思维能力”可作为先进生产力工具已是不争事实。随着多模态大模型 GPT-4 的发布，基于文字、图片等垂直场景的应用步伐有望“从 1 到 10”加速，类似于移动互联网时代各类型 APP 的百花齐放，其竞争格局也会逐步加剧。而类比 19 世纪末的美国西部“淘金热”对铲子、牛仔裤的大量需求，我们认为以 GPGPU 为代表的算力基础设施作为 AI 大模型底座将长期稳定受益。

ChatGPT 算力需求加速增长，基于大算力、先进制程领域的技术创新企业有望受益：

我们根据 GPT-4 对使用次数的限制推论，目前 AI 大模型的算力水平显著供不应求。以 Open AI 的算力基础设施为例，芯片层面 GPGPU 的需求最为直接受益，其次是 CPU、AI 推理芯片、FPGA 等。AI 服务器市场的扩容，同步带动高速网卡、HBM、DRAM、NAND、PCB 等需求提升。同时，围绕解决大算力场景下 GPU“功耗墙、内存墙”问题的相关技术不断升级，如存算一体、硅光/CPO 产业化进程有望提速；先进制程芯片演进中已有的 Chiplet 等技术路径也将受益；Risk-V 由于开源免费、开发者自由度高、自主可控度高、更适应 AIoT 处理器架构需求等优势，带动围绕 AI 场景的参与企业数量提升。

本报告的创新点：

1）以 GPT-3 模型为例的 GPGPU 市场测算：预计用于高端 GPGPU 显卡的训练及推理部分市场空间合计约 145 亿元，其中训练市场规模约 28 亿元，推理市场规模约 117 亿元。分别对应约 3200 张和 135031 张英伟达 A100 GPU 芯片。

2）对 GPT-4 算力需求及未来趋势的推论：GPT-4 由于复杂度提升、图片识别功能加入，推算算力需求翻倍增长。长期看来，伴随编

译器等软件端技术迭代，新产品推出有望提速。我们认为 AI 大模型有望向小型化、高效化方向发展，对算力需求趋势从单模型所需高性能芯片价值转变为应用端规模增长带来的用量提升。

3）重点技术梳理：存算一体技术、HBM 技术、Chiplet 技术、CPO 技术等技术。

4）系统梳理潜在受益的产业链环节及标的。

投资建议：我们建议关注国产大算力芯片、英伟达/AMD 产业链、上游硬件供应商、下游多模态应用落地等。1）GPU/AI 芯片：寒武纪、海光信息、景嘉微、澜起科技；2）英伟达产业链配套：胜宏科技、和林微纳；3）CPU：海光信息、龙芯中科、澜起科技；4）FPGA：紫光国微、复旦微电、安路科技；5）芯片 IP：芯原股份、华大九天;6）服务器：浪潮信息、工业富联、中科曙光;7）Chiplet 等先进封装相关：通富微电、长电科技、兴森科技、深南电路、生益科技、华正新材;8）光模块：天孚通信、新易盛、中际旭创;9）AIoT：乐鑫科技、恒玄股份、炬芯科技;10）SoC：晶晨股份、瑞芯微、全志科技、恒玄科技、富瀚微;11）Risk-V：兆易创新、芯原股份、国芯科技、北京君正;12）存算一体：兆易创新、恒烁股份;13）存储芯片/模组：兆易创新、佰维存储、江波龙、北京君正、聚辰股份;14）CPU/GPU 等供电芯片：杰华特、晶丰明源；15）多模态下游应用：海康威视、大华股份、萤石网络、漫步者等

风险提示：技术研发不及预期的风险；应用落地不及预期的风险；中美贸易摩擦的风险。

内容目录

图表目录

1. ChatGPT 浪潮带动算力需求提升，以 GPU 为核心的硬件市场扩容

1.1. ChatGPT:基于生成式 AI 技术的大型语言模型，商业化迅速开启

ChatGPT（Chat Generative Pre-trained Transformer）是由 OpenAI 开发的聊天机器人程序，于 2022 年 11 月推出。ChatGPT 是目前为止最先进的语言生成模型之一，使用基于 GPT-3.5 架构的大型语言模型（LLM），并通过强化学习进行训练，拥有语言理解和文本生成能力，适用于问答、对话、生成文本等多种场景。ChatGPT 用户规模扩增迅速，根据 Similar Web **数据，**2023 年 1 月期间，ChatGPT 平均每天大约有 1300 万独立访客，数量是 2022 年 12 月的两倍多（Similarweb 数据）；根据瑞银公开报告数据，2023 年 1 月（即上线两个月后）实现全球 1 亿月活跃用户，是历史上增长最快的消费者应用程序。

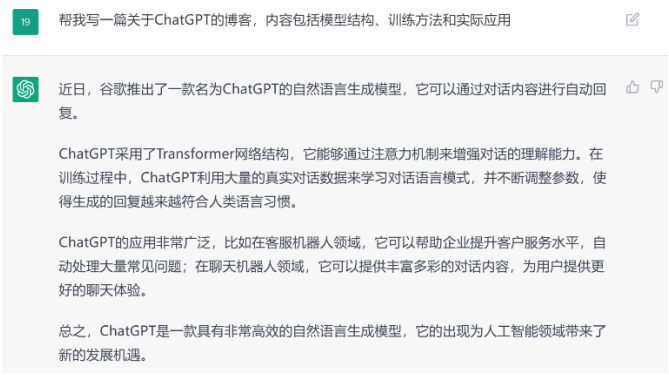
ChatGPT 是生成式 AI 在文字生成领域的应用，创造能力是其核心优势。传统 AI 依靠逻辑进行分类和判断，而生成式 AI 的壁垒在于能够创造新内容，可以是多模态的、模仿人类情感的、具有互动和审美性质的。传统的聊天机器人（Chatbot），例如客服机器人，只能根据用户输入的内容在数据库中查询到对应答案，再机械地以模板的形式反馈给客户；而 ChatGPT 采取生成式 AI 技术，并且有工作人员每日优化模型，在应对用户提问时会根据上下文内容调整回答内容，增强互动式、对话式的情感体验，更加智能。

OpenAI 开放 API，降本 90%扩大覆盖用户面。2023 年 3 月 1 日，OpenAI 官网宣布 ChatGPT 和 Whisper（OpenAI 去年发行的语音识别生成模型）的 API 开放使用，开发者可将模型集成到 APP 和其他产品中。ChatGPT API 接入的模型为 GPT-3.5-turbo，与 GPT-3.5 相比更加快捷、准确，成本也更低，定价为每 1000 个 tokens（约 750 个单词） 0.002 美元，用户则需要按照输入和输出的 tokens 总数来付费。OpenAI 官方表示自 2022 年 12 月以来 ChatGPT 降低了 90%的成本，开放 API 旨在使更多人受益于生成式 AI 技术。

图1. 不同程序实现 1 亿月活跃用户所花费的时间

资料来源：Yahoo Finance，安信证券研究中心

图2. 使用 ChatGPT 撰写博客内容



资料来源：CSDN，ChatGPT，安信证券研究中心

1.2. 采用 GPT-3.5 预训练模型，参数量随模型换代呈指数型增长

GPT3.5 是一种大型语言模型（LLM），参数量大，精准度高。GPT-3.5 采用深度学习中的 Transformer 架构，并通过大规模预训练（pre-training）的方式来学习自然语言处理任务，可以进行文本生成、对话生成、文本分类、命名实体识别、关键词提取等自然语言处理任务。

- ✓ **语言模型（LM）**是指对语句概率分布的建模。具体是判断语句的语序是否正常，是否可以被人类理解。它根据句子中先前出现的单词，利用正确的语序预测句子中下一个单词，以达到正确的语义。例如，模型比较“我是人类”和“是人类我”出现的概率，前者是

正确语序，后者是错误语序，因此前者出现的概率比后者高，则生成的语句为“我是人类”。

- ✓ **大型语言模型（LLM）**是基于海量数据集进行内容识别、总结、翻译、预测或生成文本等的语言模型。相比于一般的语言模型，LLM 识别和生成的精准度会随参数量的提升大幅提高。

ChatGPT 需要通过预训练来形成 GPT3.5 的模型，从而可以在用户端的网页或 APP 进行推理。

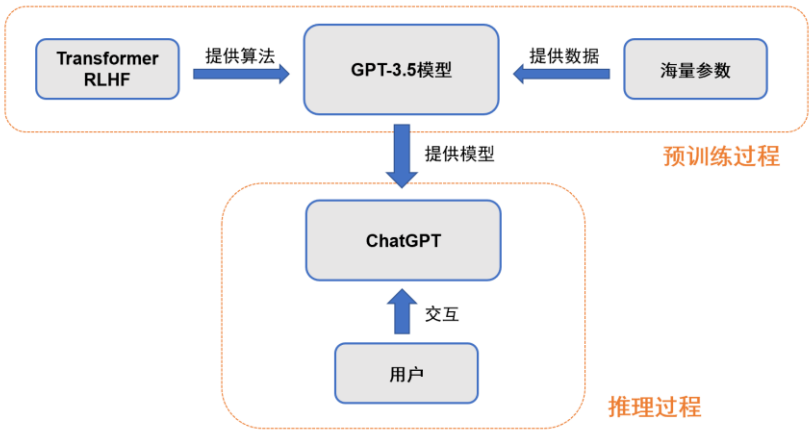
- ✓ **预训练**指先通过一部分数据进行初步训练，再在这个初步训练好的模型基础上进行重复训练，或者说是“微调”；
- ✓ **推理**指将预训练学习到的内容作为参考，对新的内容进行生成或判断。
- ✓ 预训练是模型运作的主要部分，所需要的精度较高，算力需求也较高；推理则相反。

ChatGPT 通过 Transformer 和 RLHF 两种语言模型进行预训练，可并行训练并大量优化反馈。采用深度学习中的 Transformer 架构，并通过大规模预训练（pre-training）的方式来学习自然语言处理任务，可以进行文本生成、对话生成、文本分类、命名实体识别、关键词提取等自然语言处理任务。

- ✓ **长短期记忆网络算法（LSTM）**是一种时间循环神经网络。传统的循环神经网络（RNN）拥有链式形式，就像人脑会忘记很久以前发生的事件，RNN 也会忘记它在较长序列中学习的内容，因此具有短时记忆。LSTM 是一种特殊的 RNN，它解决了传统 RNN 的短时记忆问题，在 Transformer 问世前曾主导 NLP 领域，但也拥有无法并行训练、建模长度有限的缺点。
- ✓ **Transformer** 是一个完全依赖于自注意力机制来计算其输入和输出的表示的转换模型，所以与 LSTM 的顺序处理不同，它可以并行同时处理所有的输入数据，模仿人类联系上下文的习惯，从而更好地为 LLM 注入意义并支持处理更大的数据集。
- ✓ **人类反馈信号强化学习（RLHF）**指使用强化学习的方式直接优化带有人类反馈的语言模型，使得语言模型能够与复杂的人类价值观“对齐”。它负责 ChatGPT 预训练中微调的部分，首先在人类的帮助下训练一个奖赏网络（RM），RM 对多个聊天回复的质量进行排序，从而增加 ChatGPT 对话信息量，使其回答具有人类偏好。

ChatGPT 的预训练需要处理海量参数，从而实现超高文本识别率。OpenAI 目前没有公布 ChatGPT 所使用的 GPT-3.5 的相关数据，由表 2 可知，随着新模型推出，新的参数量需求呈翻倍式增长。OpenAI 首席执行官 Sam Altman 接受公开采访表示，GTP-4 参数量为 GTP-3 的 20 倍，需要的计算量为 GTP-3 的 10 倍；GTP-5 在 2024 年底至 2025 年发布，它的参数量为 GTP-3 的 100 倍，需要的计算量为 GTP-3 的 200-400 倍。

图3. ChatGPT 预训练和推理过程



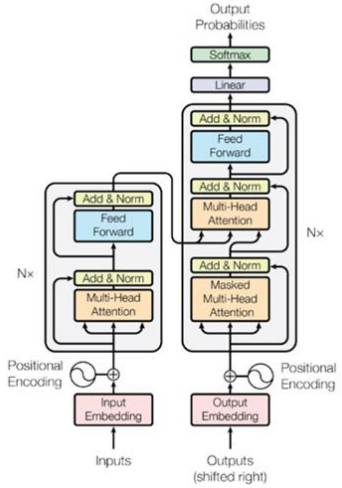
资料来源：OpenAI 官网，安信证券研究中心

表1：ChatGPT 预训练相关概念

中文名称	英文缩写/名称	特性/作用	ChatGPT 是否使用
语言模型	LM	根据语句概率进行文字预测	是
大型语言模型	LLM	需要海量数据集的 LM	是
循环神经网络	RNN	顺序处理；短时记忆	否
长短期记忆网络算法	LSTM	顺序处理；建模长度有限	否
/	Transformer	并行处理；注意力机制	是
人类反馈信号强化学习	RLHF	使模型与人类价值观对齐	是
奖赏网络	RM	RLHF 的重要步骤	是

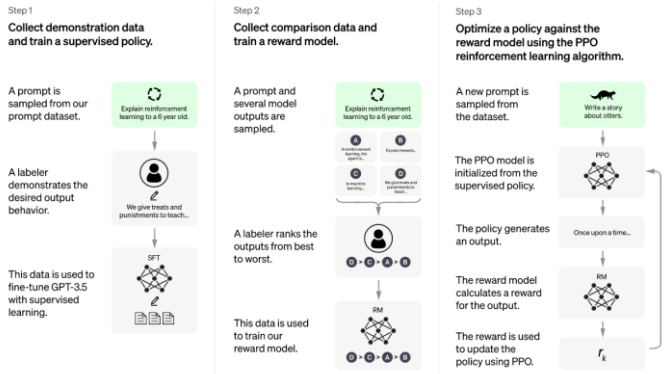
资料来源：CSDN，电子发烧友，澎湃新闻，安信证券研究中心

图4. Transformer 架构示意图



资料来源：Attention is all you need，安信证券研究中心

图5. RLHF 原理示意图



资料来源：OpenAI 官网，安信证券研究中心

表2：各代 GPT 系列所需要参数量

模型	发布时间	参数量
GPT-1	2018 年 6 月	1.17 亿
GPT-2	2019 年 2 月	15 亿
GPT-3	2020 年 5 月	1750 亿
GPT-4	2023 年 3 月	暂未公布
GPT-5（预期）	2021 年底至 2025 年	175000 亿

资料来源：OpenAI 官网，安信证券研究中心

GPT-4 功能升级，多模态拓展应用场景。2023 年 3 月 14 日，OpenAI 正式发布 GPT-4 模型，早于此前 23 年下半年发布的时间规划。根据 OpenAI 官方，GPT-4 模型于 2022 年 8 月完成训练，之后通过 6 个月时间对模型进行了安全性研究、风险评估和迭代。GPT-4 作为大型多模态模型，在多方面提升显著：

1）多模态大模型——新增接受图片和文本输入并产生文本输出能力，能分析图片的符号意义，如理解图片中的“笑梗”;文字方面，GPT-4 的输入限制由 3000 字提升至 2.5 万字，对于英语以外的语种支持有更多优化。

2）提升各种专业和学术水准并有较好表现。能处理更长更复杂的文本，在没有针对考试内容进行特别训练的基础上，GPT-4 在各项测试中均取得较高成绩，如 GPT-4 在 GRE 考试中取得 332+4 分，GPT-4（no vision）取得 322+4 分，而 GPT-3.5 分数为 301+4 分。

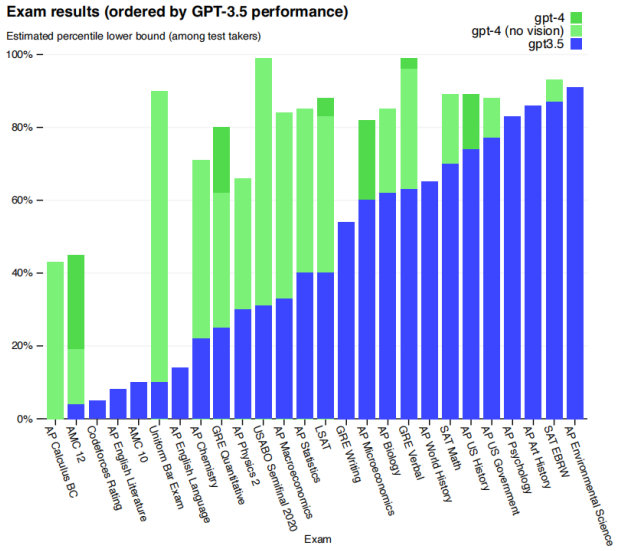
3）在安全、一致性上有较为明显的提升。根据 OpenAI 的对抗性测试和红队测试结果，相比 GPT-3.5，GPT-4 产生客观事实回答的可能性提升 40%，响应违禁内容请求的可能性降低 82%。

根据公开新闻整理，目前接入 GPT-4 支持的应用端已有微软的必应浏览器 new Bing、嵌入于办公软件的 Microsoft 365 Copilot 人工智能服务，外语培训教育机构多邻国的付费产品 DuolingoMax、摩根士丹利等。**我们认为，随着 GPT-4 等模型复杂度升级，并逐步支持图片视频识别等多模态，对应的算力及基础设施需求有望持续增长；下游则有望拓展更多图片视频内容端的商业化应用场景。**

图6. GPT-4 对图片输入的理解

资料来源：《GPT-4 Technical Report》，安信证券研究中心

图7. GPT-4 考试表现相较 GPT-3.5 的提升



资料来源：《GPT-4 Technical Report》，安信证券研究中心

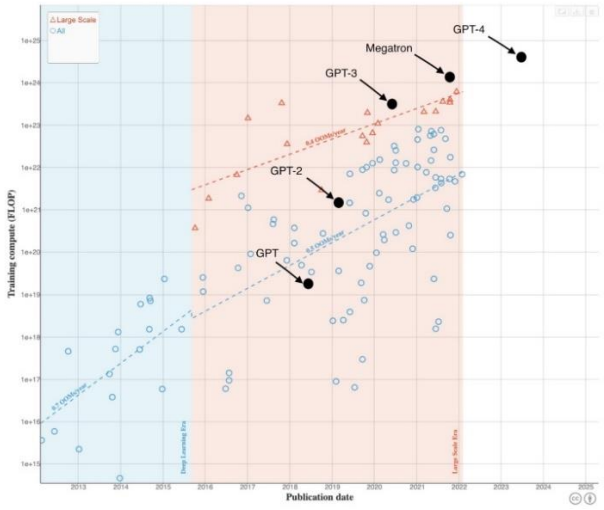
1.3. 海量参数产生大算力需求，GPGPU 等高壁垒 AI 芯片受益

ChatGPT 算力需求与参数量呈正相关，对硬件的内存容量和带宽提出高要求。算力即计算能力，具体指硬件对数据收集、传输、计算和存储的能力，算力的大小表明了对数字化信息处理能力的强弱，常用计量单位是 FLOPS（Floating-point operations per second），表示每秒浮点的运算次数。硬件方面，运算量取决于 GPU 运算执行时间的长短，而参数量取决于占用显存的量。运算量（FLOPS）的数值通常与参数量（parameter count）成比例，不同模型架构的换算关系不同。模型越复杂、参数量越大，所需计算量越大。

GPGPU 拥有硬件技术的核心壁垒：大显存带宽，进行超高能效比的并行运算，可同时用于 GPT 模型的训练和推理过程。GPGPU（通用图像处理器）是一种由 GPU 去除图形处理和输出，仅保留科学计算、AI 训练和推理功能的 GPU（图形处理器）。GPU 芯片最初用于计算机系统图像显示的运算，但因其相比于擅长横向计算的 CPU 更擅长于并行计算，在涉及到大量的矩阵或向量计算的 AI 计算中很有优势，GPGPU 应运而生。目前，GPGPU 的制造工艺在英伟达等企业的领导下已趋向成熟，成本在 AI 芯片中也较低，成为市场主流选择， ChatGPT 引起的 AI 浪潮有望提升其应用规模。

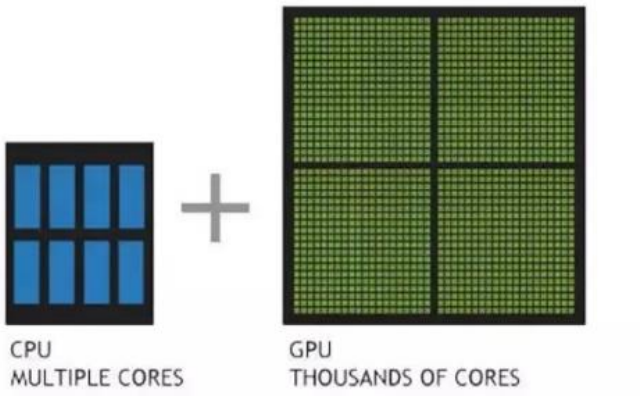
FPGA 具有可编程的灵活性，ASIC 性能佳、具有定制化特点，但成本方面与 GPU 相比稍显劣势，在 GPT 等 AI 模型的运用占比较 GPU 低。FPGA 指现场可编程逻辑门阵列，具有静态可重复编程和动态在系统重构的特性，但其开发难度大、只适合定点运算，同时价格也比较昂贵，性能方面也不及 GPU 与 ASIC，只在精度较低的推理过程有所应用。ASIC 指专用集成电路，是一种应不同用户需求和不同系统需要而设计、制造的集成电路。ASIC 芯片的性能较 GPU 佳，能耗也较低，但因其定制性价格昂贵，在人工智能平台和推理过程中有部分应用。

图8. 近年主流生成型 AI 对算力的需求



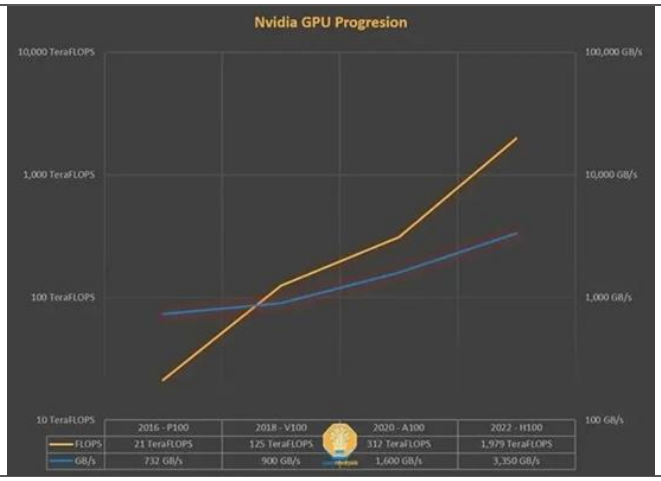
资料来源：NextBigFuture，安信证券研究中心

图9. GPU 与 CPU 并行运算能力对比



资料来源：维基百科，安信证券研究中心

图10. 近年英伟达 GPU 的 FLOPS 与带宽速率增长



资料来源：Semianalysis，安信证券研究中心

表3：各 AI 芯片性能对比

类别	GPU	FPGA	ASIC
优点	性能高 通用性好	可编程性 灵活	定制化设计 性能稳定 功耗控制优秀
缺点	功耗高	开发难度大 价格昂贵	灵活性不足 价格昂贵
代表公司	英伟达 AMD	Altera（Intel 收购） Xilinx（AMD 收购）	寒武纪 地平线 谷歌（TPU）

资料来源：CSDN，安信证券研究中心

1.4. 类 ChatGPT 成本高昂产品涌现，国产大模型方兴未艾

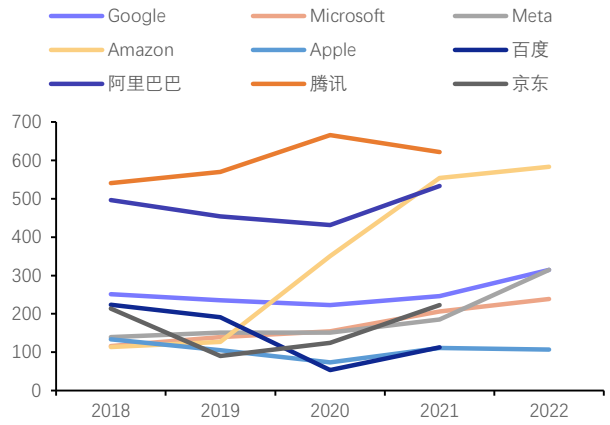
大模型运行成本高昂，准入壁垒较高。大模型对于训练时间和参数量都有高要求，以 OpenAI CEO Altman 在推特上回复马斯克的留言可知，ChatGPT 平均一次聊天成本为几美分。根据 Similar Web 数据，2023 年 1 月 27 日至 2 月 3 日 ChatGPT 日活跃用户达 2500 万人。中性假设下，以平均单人单日对话 7 次，每次 3 美分成本进行测算，对应一年支出对话成本约为 19.2 亿美元。根据英伟达官网，A100 作为 DGX A100 系统的一部分进行销售，该系统搭载 8

个 A100 GPU，一个由 5 台 DGX A100 系统组成的机架可替代一个包括 AI 训练和推理基础设施的数据中心，且功耗仅为其 1/20，成本为其 1/10，系统售价 19.9 万美元。因此，在中性假设条件下，考虑到服务器约占数据中心成本的 70%（中商产业研究院），则 ChatGPT 运营一年将需要 6741 个 DGX A100 系统用于支撑访问量。因此我们推断，在高昂成本及大数据量需求的限制下，仅有限数量的科技巨头具备参与 AI 竞赛的实力。

图11. Musk 和 Altman 关于 ChatGPT 对话成本聊天截图

资料来源：Twitter，安信证券研究中心

图12. 2018-2022 年科技厂商资本支出（亿美元）



资料来源：Wind，安信证券研究中心

表4：ChatGPT 对话成本测算

对话成本			
	保守	中性	乐观
日活跃用户（万）	2500	2500	2500
对话次数（次）	5	7	10
每次对话成本（美元）	0.02	0.03	0.05
每年总成本（亿美元）	9.1	19.2	45.6
服务器成本占比	70%	70%	70%
服务器总成本（亿美元）	6.4	13.4	31.9
DGX A100 系统价格（美元）	199000	199000	199000
DGX A100 系统需求量（个）	3210	6741	16049

资料来源：Similar Web，英伟达官网，Twitter，安信证券研究中心

ChatGPT 带动大模型竞品发布，海内外科技巨头先后加码 AI 布局。1）谷歌向 AI 公司 Anthropic 投资近 4 亿美元，后者正在测试生成式 AI 工具 Claude，且谷歌也推出对标 ChatGPT 的聊天机器人 Bard。2）微软以 100 亿美元投资 ChatGPT 的开发商 OpenAI，并获得其 49%股权。2023 年 2 月，微软发布基于 ChatGPT 的 new Bing。3）亚马逊云服务 AWS 宣布与 AI 公司 Hugging Face 开展合作，Hugging Face 将在 AWS 上开发针对 ChatGPT 的开源竞品，构建开源语言模型的下个版本 Bloom。4）阿里达摩院正研发类 ChatGPT 的对话机器人，目前已处于内测阶段。5）百度开发类 ChatGPT 项目“文心一言”（ERINE Bot）。6）京东推出产业版 ChatJD。

表5：各科技公司关于类 ChatGPT 的技术布局概览（统计截止日期：2023.03.19）

公司	AI 模型	参数规模	领域	应用场景
Google	BERT	4810 亿	NLP	语言理解与生成
	LaMDA	1370 亿	NLP	对话系统
	PaLM	5620 亿	多模态	语言理解与图像生成
	Imagen	110 亿	多模态	语言理解与图像生成
	Parti	200 亿	多模态	语言理解与图像生成
Microsoft	Florence	6.4 亿	CV	视觉识别
	Turing-NLG	170 亿	NLP	语言理解、生成
Meta	OPT-175B	1750 亿	NLP	语言模型
	M2M-100	154 亿	NLP	100 种语言互译
Deep Mind	Gato	12 亿	多模态	通才智能体
	Gopher	2800 亿	NLP	语言理解与生成
	AlphaCode	414 亿	NLP	代码生成
OpenAI	CLIP&DALL-E	120 亿	NLP	图像生成、跨模态检索
	Codex	120 亿	多模态	代码生成
	ChatGPT	-	NLP	语言理解与生成、推理等
百度	NLP 大模型	千亿	NLP	语言理解、生成
	跨模态大模型	240 亿	多模态	语言理解与图像生成
	CV 大模型	170 亿	多模态	语言理解与图像生成
	生物计算大模型	-	CV	化合物表征学习、分子结构预测
阿里巴巴	M6	十万亿	多模态	语言理解与图像生成
腾讯	混元大模型	万亿	NLP	语言理解与生成
京东	K-PLUG	10 亿	NLP	语言理解与生成、推理、代码生成
华为	盘古大模型	2000 亿	NLP、CV、多模态	内容生成与理解、分类分割检测、跨模态检索
复旦大学	MOSS	175 亿	NLP	语言理解与生成
360	-	-	NLP	智能搜索
字节跳动	DA	-	NLP	语言理解

资料来源：IT 资讯、虎嗅网、华为云官网、腾讯云官网、百度云官网、量子位、《超大规模多模态预训练模型 M6 的关键技术及产业应用》，公开信息整理，安信证券研究中心

基于昆仑芯+飞桨+文心大模型 AI 底座，百度推出“文心一言”拉开国产生成式 AI 序幕。2023 年 3 月 16 日，百度正式推出国内首款生成式 AI 产品“文心一言”，可支持文学创作、文案创作、数理推算、多模态生成等功能，目前已有多家厂商宣布接入。“文心一言”基于全栈自研的 AI 基础设施进行学习和训练：

- 昆仑芯 2 代 AI 芯片：“文心一言”的芯片层核心能力，采用自研 XPU-R 架构，通用性和性能显著提升；256 TOPS@INT8 和 128 TFLOPS@FP16 的算力水平，较一代提升 2-3 倍，保障“文心一言”算力需求；采用 7nm 先进工艺，GDDR6 高速显存，支持虚拟化，芯片间互联和视频编解码等功能。
- 飞桨深度学习平台：“文心一言”的框架层核心能力，系业内首个动静统一的框架、首个通用异构参数服务器架构，支持端边云多硬件和多操作系统，为文心大模型提供有效、快捷、完整的训练框架。

- 文心知识增强大模型:“文心一言”的模型层核心能力，该产品主要采用 ERNIE 系列文心 NLP 模型，拥有千亿参数级别的 ERNIE 3.0 Zeus 为该系列最新模型，进一步提升了模型对于不同下游任务的建模能力，大大拓宽了“文心一言”的应用场景。

我们认为，随着国产 AI 大模型应用的不断拓展，算力基础设施加速升级，伴随产业链自主研发需求及地缘政治不确定性，关于进口高端 AI 芯片及服务器中美博弈升级，国产高算力 GPU 芯片、服务器及数据中心等厂商有望加速迭代，长期充分受益。

图13. 百度 AI 大底座示意图

资料来源：百度智能云官网，安信证券研究中心

1.5. 以 GPT-3 为例测算：大算力需求驱动 AI 硬件市场空间提升

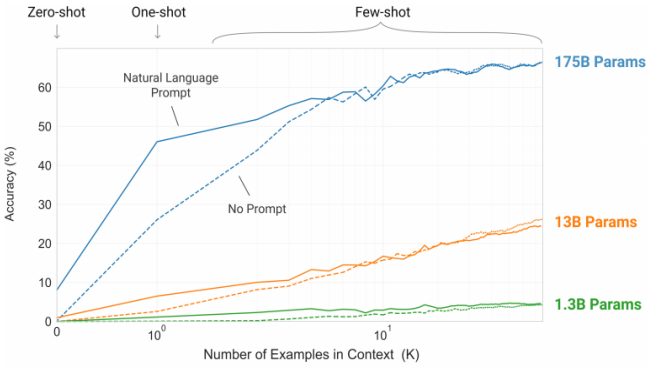
GPT-3（Generative Pre-trained Transformer 是 GPT-3.5 的上一代语言模型，目前一般所说的 GPT-3 即为拥有 1750 亿参数的最大 GPT-3 模型，OpenAI 在公开发表的论文《Language Models are Few-Shot Learners》中对 GPT-3 模型进行了详细分析。对于以 ChatGPT 为例的大模型算力需求，根据测算，我们预计用于**高端 GPGPU 显卡的训练及推理部分市场空间合计约 145.32 亿元，其中训练市场规模为 27.84 亿元，推理市场规模为 117.48 亿元。**

图14. GPT-3 模型大小、架构及参数

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

资料来源：《Language Models are Few-Shot Learners》，安信证券研究中心

图15. 不同参数量模型的上下文学习曲线



资料来源：《Language Models are Few-Shot Learners》，安信证券研究中心

图16. 用于训练语言模型所需要的算力情况

Model	Total train compute (PF-days)	Total train compute (flops)	Params (M)	Training tokens (billions)	Flops per param per token	Mult for bwd pass	Fwd-pass flops per active param per token	Frac of params active for each token
T5-Small	2.08E+00	1.80E+20	60	1,000	3	3	1	0.5
T5-Base	7.64E+00	6.60E+20	220	1,000	3	3	1	0.5
T5-Large	2.67E+01	2.31E+21	770	1,000	3	3	1	0.5
T5-3B	1.04E+02	9.00E+21	3,000	1,000	3	3	1	0.5
T5-11B	3.82E+02	3.30E+22	11,000	1,000	3	3	1	0.5
BERT-Base	1.89E+00	1.64E+20	109	250	6	3	2	1.0
BERT-Large	6.16E+00	5.33E+20	355	250	6	3	2	1.0
RoBERTa-Base	1.74E+01	1.50E+21	125	2,000	6	3	2	1.0
RoBERTa-Large	4.93E+01	4.26E+21	355	2,000	6	3	2	1.0
GPT-3 Small	2.60E+00	2.25E+20	125	300	6	3	2	1.0
GPT-3 Medium	7.42E+00	6.41E+20	356	300	6	3	2	1.0
GPT-3 Large	1.58E+01	1.37E+21	760	300	6	3	2	1.0
GPT-3 XL	2.75E+01	2.38E+21	1,320	300	6	3	2	1.0
GPT-3 2.7B	5.52E+01	4.77E+21	2,650	300	6	3	2	1.0
GPT-3 6.7B	1.39E+02	1.20E+22	6,660	300	6	3	2	1.0
GPT-3 13B	2.68E+02	2.31E+22	12,850	300	6	3	2	1.0
GPT-3 175B	3.64E+03	3.14E+23	174,600	300	6	3	2	1.0

资料来源：《Language Models are Few-Shot Learners》，安信证券研究中心

具体分为训练及推理两部分进行分别测算：

- **训练部分：**以 Nvidia A100 GPU 为例，其理论峰值算力为 312 TFLOPS，Nvidia 联合发表的论文《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》中，通过使用流水线并行（pipeline parallelism）、张量并行（tensor parallelism）和数据并行（data parallelism）等并行技术将 GPU 的算力利用率提升到 52%。我们参考 OpenAI 论文公开数据，标准 GPT-3 模型的 175B 模型参数(parameter)，完整训练需要 3.14E+23 FLOPs。单个模型训练时间越短，所需 GPU 越多，反之亦然。我们假设 GPT-3 模型训练时长为一周，以此作为参考，则该训练过程所需 A100 GPU 数量约为 3200 张。根据中关村在线数据，单张 A100 80G 售价约 87000 元，我们假设将有 10 家科技厂商采购 A100 卡参与 AI 大模型训练，则 A100 GPU 对应市场规模预计为 27.84 亿元。计算过程如下：

  1. 单张 A100 GPU 实际使用过程中的算力
     = 312 TFLOPS * 52%
     = 162 *10^12 FLOPS
  2. 训练一周所需时间
     = 7days*24h/day*60min/h*60s/min
     = 604800s
  3. A100 GPU 所需数量
     = 总算力需求 / (单张 GPU 实际算力 * 训练一周时间)
     = 3.14 * 10^23 / (162 * 10^12 * 604800)
     = 3200 张
  4. A100 GPU 市场规模
     = A100 数量 * 单价 * 厂商数
     = 3200 * 87000 * 10 = 27.84 亿元

- **推理互动部分：**推理端需求较训练端占比逐渐提升。根据 Similar Web 数据，每人每天平均 1000 词左右问题回答，目前 ChatGPT 日活跃用户为 2500 万人，即合计每日产生 250 亿单词，相当于 333 亿 tokens（根据 OpenAI 官网，token 是一种非结构化文本单位，英文语境下 1 个 token 相当于 4 个字母，0.75 个词，中文语境下 1 个中文字被视为 1 个 token）。根据马里兰大学副教授 Tom Goldstein 推文表示，30 亿参数模型使用单张 A100 GPU（使用半精度、TensorRT 和激活缓存）生成 1 个 token 需要 6ms，扩大至 1750 亿参数模型则需要 350ms (=1750/30*6)。以单日时长计算，推理过程需要 135031 张 A100 GPU，对应市场规模 117.48 亿元。计算过程如下：

  1. 用户每日产生总 token 数
     =日活跃人数 * 平均问题字数 / 0.75

=2500*10^4*1000/0.75
=333.33 亿个

2. 模型生成总 token 数所需时间
=总 token 数 * 单 A100 GPU 输出单 token 所需时间
=333.33*10^8*350ms
=116.67*10^8s

3. A100 GPU 所需数量
=模型所需总时间 / 一天时间
=116.67*10^8/(24*60*60)
=135031 张

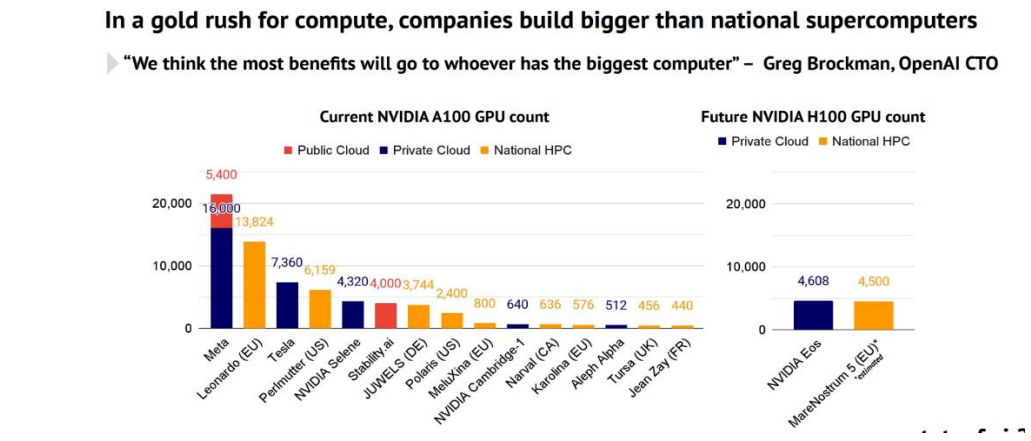
4. A100 GPU 市场规模
=A100 数量 * 单价
=135031 * 87000 = 117.48 亿元

表6：ChatGPT 对应 A100 GPU 市场规模

训练部分	算力总需求量	3.14E+23 FLOPs
	A100 GPU 算力	312 TFLOPS
	算力利用率	52%
	A100 GPU 实际算力	162 TFLOPS
	训练时长（s）	7days*24h*3600s
	A100 GPU 所需数量	3200
	A100 GPU 单价（元）	87000
	参与厂商数量（个）	10
	A100 GPU 市场空间（亿元）	27.84
推理互动部分	日活跃用户（万）	2500
	每人问题单词数（words）	1000
	单 token 对应单词数（个）	0.75
	24h 生成总 token 数（亿个）	333.33
	单 GPU 输出单 token 所需时间	350ms
	单日 token 输出所需总时间（s）	116.67*10^8
	A100 GPU 所需数量	135031
	A100 GPU 市场空间（亿元）	117.48

资料来源：英伟达官网，《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》，Twitter，Similar Web，中关村在线，安信证券研究中心

图17. 下游企业拥有英伟达 A100 GPU 数量（截止至 2022）



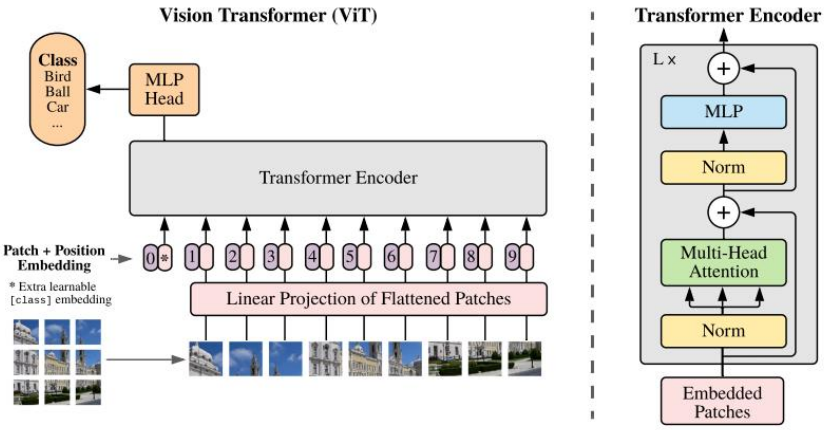
资料来源：tateof.AI，安信证券研究中心

1.6. GPT-4 模型算力需求扩增，架构升级降本增效未来可期

根据 OpenAI 官网显示，目前 GPT-4 每 4 小时只能处理 100 条消息，且并没有开放图片识别功能。大模型升级带来的运算需求逐渐加码，且可推测目前算力已处于供不应求状态。

多模态拓展，图片识别算力需求升级十倍以上。关于从图片到 token 的转换方式，OpenAI 未公布 GPT-4 的模型参数，假设 GPT-4 处理图片视觉任务使用 VisionTransformer 模型(ViT)，则输入图片尺寸必须为 224×224（ViT-B/16 版本）。根据 2021 年 ICLR 论文，模型原理大致为把一张图片分成 nxn 个 Patch，每一个 Patch 作为一个 Token。即把一张 224×224×3 的图片，切分为 16×16 大小的 Patch，每个 Patch 是三通道小图片，得到 16×16×3=768 个维向量作为一个 token 输入，则一张大图片能转成 14*14=196 个 token。相较之下，根据前文 GPT-3 部分假设，假设每个文字问题 50-100 词，即 67-133token。**我们可以粗略推论，图像识别的所需算力是文字推理部分所需算力的 2-4 倍。**

图18. Vision Transformer 模型对图片进行切割输入



资料来源：ICLR，安信证券研究中心

编译器性能升级，带动大模型产品加速迭代。随着 2023 年 3 月 15 日 Pytorch2.0 正式版的发布，编译器的性能有大幅提升。Pytorch 作为主流深度学习框架，用于构建及训练深度学习模型。Pytorch2.0 正式版包含的新高性能 TransformAPI 能使 GPT-3 等使用的先进 transformer 模型的训练和部署更加容易、快速。根据 PyTorch 基金会数据，在 Nvidia A100 GPU 上使用 PyTorch 2.0 对 163 个开源模型进行的基准测试，其中包括图像分类、目标检测、图像生成，以及各种 NLP 任务，2.0 版本的编译时间比 1.0 提高 43%。**我们认为，编译器性能的提升带动 AI 大模型编译时间缩短，新产品推出进展或将超预期。**

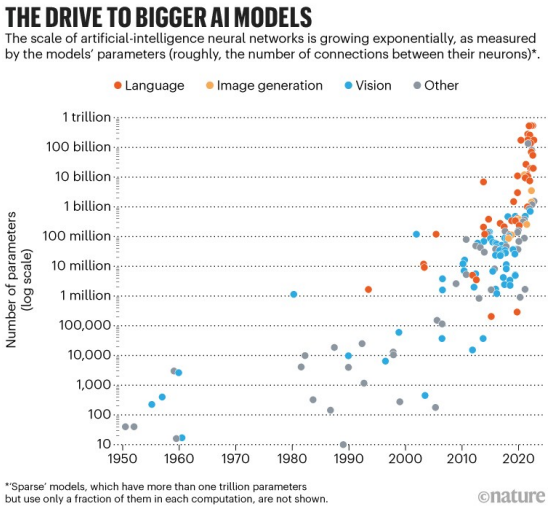
同时我们认为，目前模型的计算成本高，参数量大，长期看模型架构的升级将缩小训练成本，并拓宽边缘设备等部署场景，对算力的需求有望从单模型所需芯片价值量高的推演转变为应用场景快速拓展的量的增长。

（1）根据 Nature2023 年 3 月 8 日文章<In AI，is bigger always better?>，有观点认为，更大参数量的模型只是在回答训练数据相关范围的查询上表现更好，并不具备获得回答新问题的更优能力。过往几年，AI 大模型的训练使用更高的算力和参数量，但一些小型性能好的模型涌现，在训练中用了更高数据。具体而言，2023 年 2 月 Meta 发布 LLaMA 小参数模型，130 亿参数但训练量多达 1.4 万亿个，表现优于 GPT-3。而同年 3 月 14 日，斯坦福发布基于 LLaMA 的 Alpaca7B 微调模型，其 52000 个指令的 OpenAI API 总成本不到 500 美元；微调过程在云计算平台使用 8 个 A100 80GB GPU，用时 3 小时，成本约 100 美元。测试结果表明 Alpaca7B 性能和其指令资料来源的 GPT-3 模型相近。长期来看，大模型有望向规模更小、更智能高效的方向演进。

（2）多模态方面，举例说明，根据清华大学 2021 年论文< DynamicViT: Effificient Vision Transformers with Dynamic Token Sparsification>，ViT 的最终预测仅基于信息最丰富的 token 的一个子集，该子集足以进行图像准确识别，论文提出的动态 token 稀疏化框架可

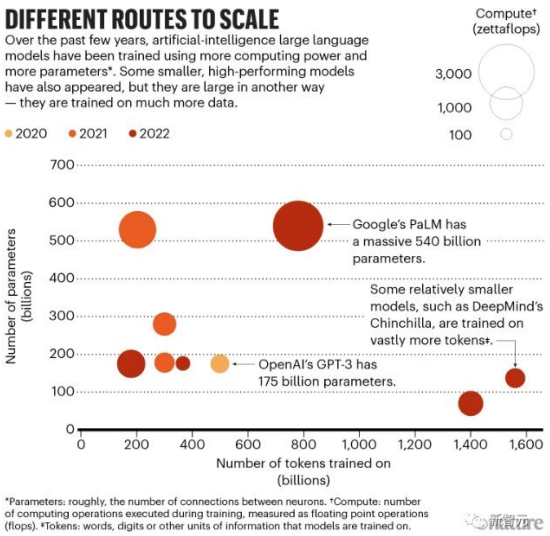
以理解为轻量化预测模块，估计每个 token 的重要性，从而动态删除冗余 token，其框架的结论减少了 31-37%FLOPS，提升 40%以上吞吐量，同时精度下降小于 5%。

图19. AI 大模型的参数规模持续加速攀升



资料来源：Nature，安信证券研究中心

图20. 小参数模型逐渐有出色表现

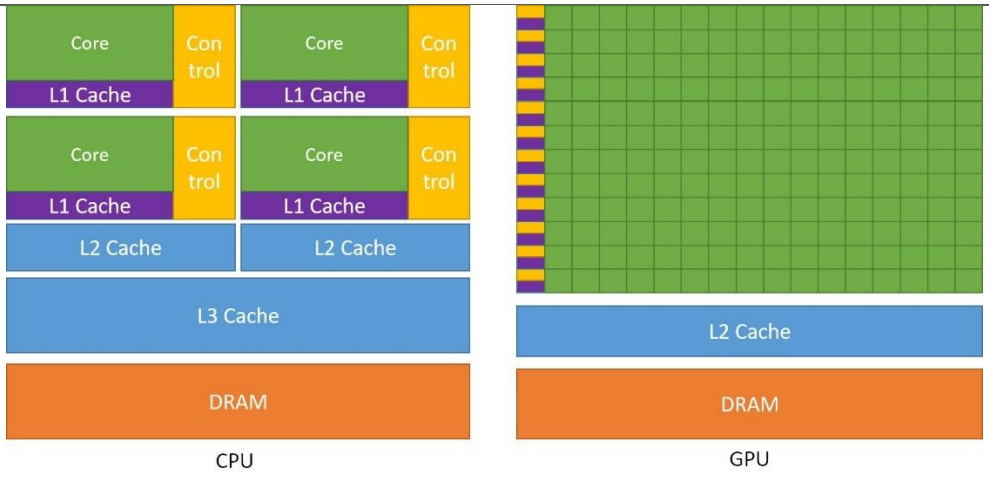


资料来源：Nature，安信证券研究中心

1.7. 英伟达引领硬件端产品升级，国产 GPU 静待花开

大 GPU 优势在于通过并行计算实现大量重复性计算。GPGPU（General Purpose GPU）即通用 GPU，能够帮助 CPU 进行非图形相关程序的运算。在类似的价格和功率范围内，GPU 能提供比 CPU 高得多的指令吞吐量和内存带宽。GPGPU 架构设计时去掉了 GPU 为了图形处理而设计的加速硬件单元，保留了 GPU 的 SIMT(Single Instruction Multiple Threads)架构和通用计算单元，通过 GPU 多条流水线的并行计算来实现大量计算。所以基于 GPU 的图形任务无法直接运行在 GPGPU 上，但对于科学计算，AI 训练、推理任务（主要是矩阵运算）等通用计算类型的任务仍然保留了 GPU 的优势，即高效的搬运和运算有海量数据的重复性任务。目前主要用于例如物理计算、加密解密、科学计算以及比特币等加密货币的生成。

图21. CPU 和 GPU 架构对比



资料来源：腾讯技术工程，安信证券研究中心

表7：GPU 发展历程

时间	类型	相关标准	代表产品	基本特征	意义
80 年代	图形显示	CGA. VGA	IBM 5150	光栅生成器	最早图形显示控制器
80 年代末	2D 加速	GDI，DirectFB	S3 86C911	2D 图元加速	开启 2D 图形硬件加速时
90 年代初	部分 3D 加速	OpenGL（1.1~4.1），DirectX（6.0-11）	3DLabs Glint300SX	硬件 T&L	第一颗用于 PC 的 3D 图形加速芯片
90 年代后期	固定管线		NVIDIA GeForce256	shader 功能固定	首次提出 GPU 概念
2004~2010	统一渲染		NVIDIA G80	多功能 shader	CUDA 与 G80 一同发布
2011~至今	通用计算	CUDA，Open CL1.2~2.0	NVIDIA TESLA	完成与图形处理无关的科学计算	NVIDIA 正式将用于计算的 GPU 产品线独立出来

资料来源：半导体行业观察，安信证券研究中心

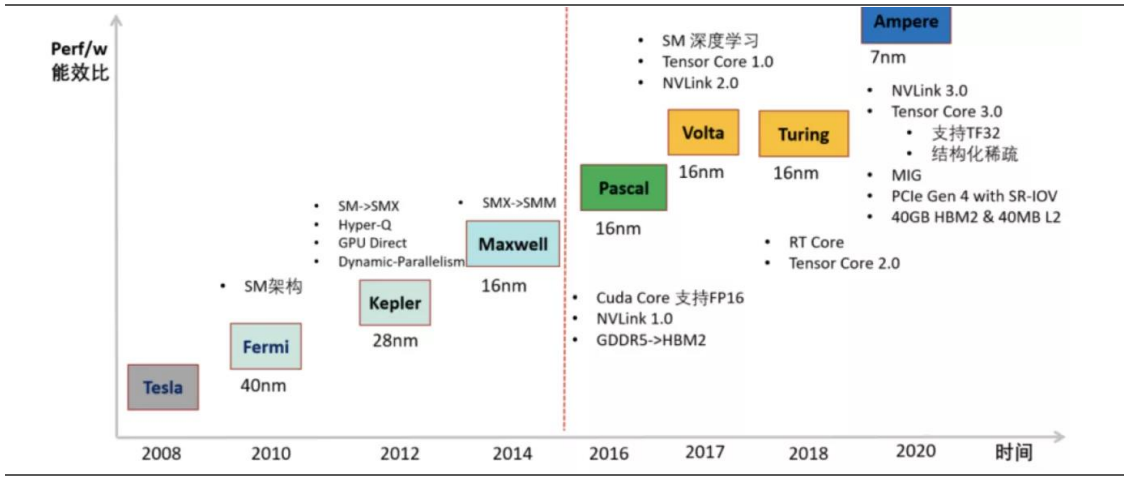
英伟达 CUDA 架构引领 GPGPU 开发市场，算力底座筑造核心护城河。随着超算等高并发性计算的需求不断提升，英伟达以推动 GPU 从专用计算芯片走向通用计算处理器为目标推出了 GPGPU，并于 2006 年前瞻性发布并行编程模型 CUDA，以及对应工业标准的 OpenCL。CUDA 是英伟达的一种通用并行计算平台和编程模型，它通过利用图形处理器（GPU）的处理能力，可大幅提升计算性能。CUDA 使英伟达的 GPU 能够执行使用 C、C++、Fortran、OpenCL、DirectCompute 和其他语言编写的程序。在 CUDA 问世之前，对 GPU 编程必须要编写大量的底层语言代码；CUDA 可以让普通程序员可以利用 C 语言、C++等为 CUDA 架构编写程序在 GPU 平台上进行大规模并行计算，在全球 GPGPU 开发市场占比已超过 80%。**GPGPU 与 CUDA 组成的软硬件底座，构成了英伟达引领 AI 计算及数据中心领域的根基。**

通过与云计算平台的集成，CUDA 可在未购买 GPU 硬件的基础上提供强大计算能力。例如，假设客户需要训练一个深度学习模型需要大量的计算资源和时间，通过在 AWS 上租用一个带有 NVIDIA GPU 的实例，并在该实例上安装 CUDA，客户可以使用 CUDA API 和库来利用 GPU 的计算能力运行计算密集型工作负载，从而可以无需购买 GPU 硬件并快速完成训练任务。除了 AWS，其他云计算提供商如 Microsoft Azure、Google Cloud Platform 等也提供了与 CUDA 集成的服务。这些服务可以为客户提供强大的 GPU 计算能力，从而加速计算密集型工作负载的处理速度。

GPU 架构升级过程计算能力不断强化，Hopper 架构适用于高性能计算（HPC）和 AI 工作负载。英伟达在架构设计上，不断加强 GPU 的计算能力和能源效率。在英伟达 GPU 架构的演变中，从最先 Tesla 架构，分别经过 Fermi、Kepler、Maxwell、Pascal、Volta、Turing、Ampere 至发展为今天的 Hopper 架构。以 Pascal 架构为分界点，自 2016 年后英伟达逐步开始向深度学习方向演进。根据英伟达官网，**Pascal 架构，**与上一代 Maxwell 相比，神经网络训练速

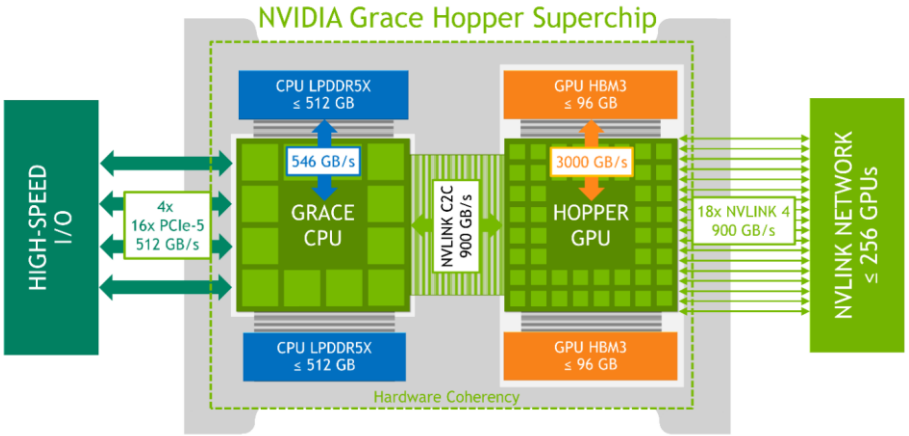
度提高 12 倍多，并将深度学习推理吞吐量提升了 7 倍。**Volta 架构，**配备 640 个 Tensor 内核增强性能，可提供每秒超过 100 万亿次（TFLOPS）的深度学习性能，是上一代 Pascal 架构的 5 倍以上。**Turing 架构，**配备全新 Tensor Core，每秒可提供高达 500 万亿次的张量运算。**Ampere 架构，**采用全新精度标准 Tensor Float 32（TF32），无需更改任何程序代码即可将 AI 训练速度提升至 20 倍。**最新 Hopper 架构是第一个真正异构加速平台，**采用台积电 4nm 工艺，拥有超 800 亿晶体管，主要由 Hopper GPU、Grace CPU、NVLINK C2C 互联和 NVSwitch 交换芯片组成，根据英伟达官网介绍，其性能相较于上一代 Megatron 530B 拥有 30 倍 AI 推理速度的提升。

图22. GPU 架构演变历程



资料来源：汽车人参考，安信证券研究中心

图23. Grace Hopper 超级芯片示意图



资料来源：英伟达官网，安信证券研究中心

AMD 数据中心领域布局全面，形成 CPU+GPU+FPGA+DPU 产品矩阵。与英伟达相比，AMD 在服务器端 CPU 业务表现较好，根据 Passmark 数据显示，2021 年 Q4 AMD EPYC 霄龙系列在英特尔垄断下有所增长，占全球服务器 CPU 市场的 6%。依据 CPU 业务的优势，AMD 在研发 GPGPU 产品时推出 Infinity Fabric 技术，将 EPYC 霄龙系列 CPU 与 Instinct MI 系列 GPU 直接相连，实现一致的高速缓存，形成协同效应。此外，AMD 分别于 2022 年 2 月、4 月收购 Xilinx 和 Pensando，补齐 FPGA 与 DPU 短板，全面进军数据中心领域。

软件方面，AMD 推出 ROCm 平台打造 CDNA 架构，但无法替代英伟达 CUDA 生态。AMD 最新的面向 GPGPU 架构为 CDNA 系列架构，CDNA 架构使用 ROCm 自主生态进行编写。AMD 的 ROCm 生态采取 HIP 编程模型，但 HIP 与 CUDA 的编程语法极为相似，开发者可以模仿 CUDA 的编程方式

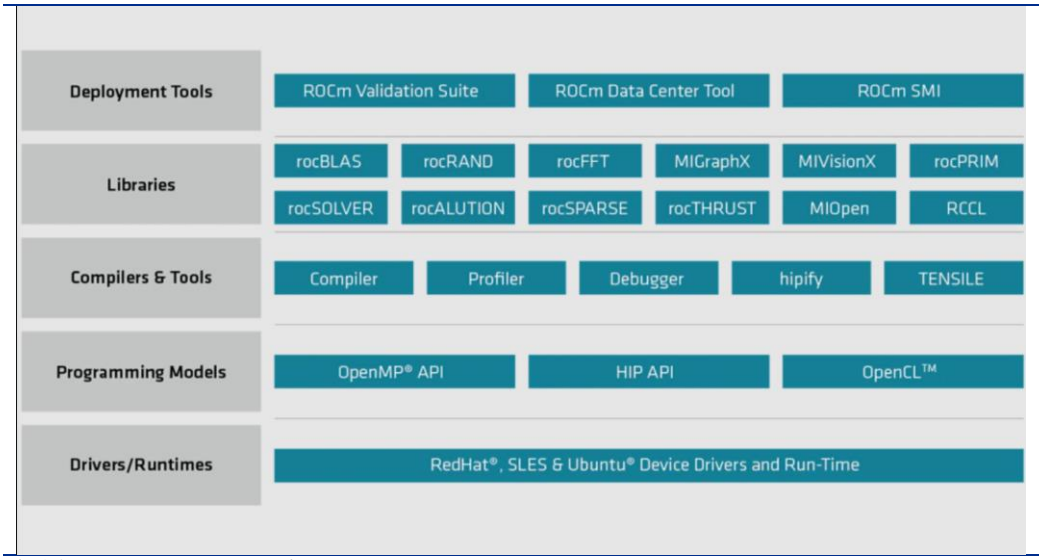
为 AMD 的 GPU 产品编程，从而在源代码层面上兼容 CUDA。所以从本质上来看，ROCm 生态只是借用了 CUDA 的技术，无法真正替代 CUDA 产生壁垒。

表8：AMD GPGPU 相关产品一览

系列	产品	主要参数
Instinct MI 系列 GPU	MI50	7nm Vega20 架构，3840 个流处理器，32GB 显存，1024GB/s 带宽，单精度 13.3T
	MI60	7nm Vega20 架构，4096 个流处理器，32GB 显存，1024GB/s 带宽，单精度 14.75T
	MI100	7nm CDNA 架构，7680 个流处理器，32GB 显存，1228.8GB/s 带宽，单精度 23.1T
	MI210	6nm CDNA2 架构，6656 个流处理器，64GB 显存，1638.4GB/s 带宽，单精度 22.6T
	MI250	6nm CDNA2 架构，13312 个流处理器，128GB 显存，3276.8GB/s 带宽，单精度 45.3T
	MI250X	6nm CDNA2 架构，14080 个流处理器，128GB 显存，3276.8GB/s 带宽，单精度 47.9T

资料来源：AMD 官网，安信证券研究中心

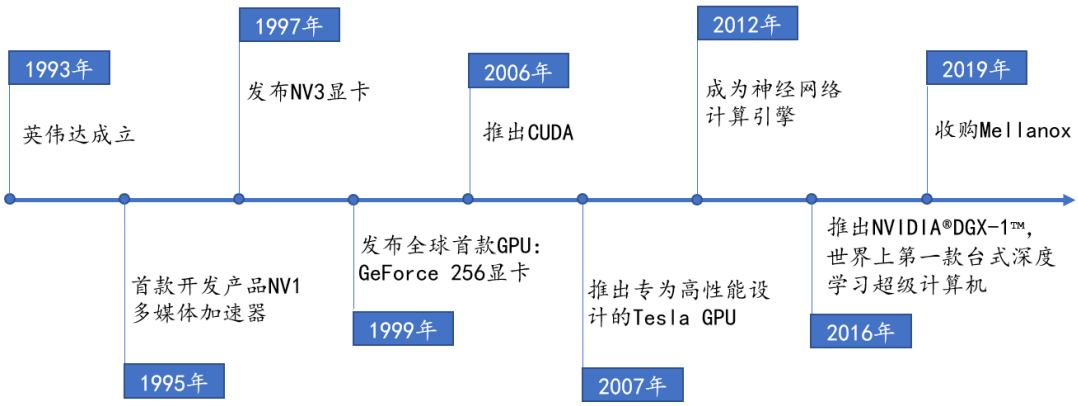
图24. ROCm 5.0 生态技术



资料来源：AMD 官网，安信证券研究中心

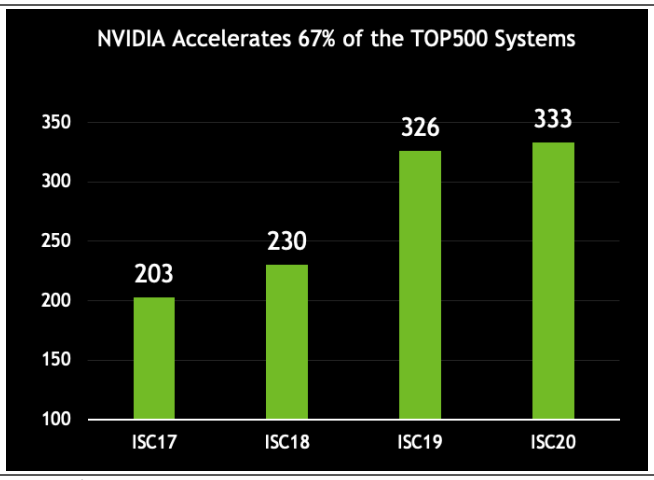
前瞻性布局 AI 和云计算领域，英伟达独占鳌头。回顾英伟达发展历程，在传统游戏业务外，公司始终关注数据中心业务布局：英伟达早在 2006 年便推出 CUDA 架构，提高 GPU 解决复杂计算的能力；2007 年推出专为高性能计算设计的 Tesla 系列 GPU 产品，此后开始快速迭代，性能不断提升，至今已发展出 8 个架构；2016 年推出世上首款台式超级计算机 DGX-1，主要应用于 AI 领域；2019 年收购 Mellanox，降低云数据中心的运营成本。与 AMD、英特尔相比，英伟达在 AI 计算领域独占鳌头：在 2020 年全球 TOP500 超级计算机榜单中，有 333 台超级计算机采用了英伟达的技术，占总数的 66.6%，英伟达的统治地位可见一斑。

图25. 英伟达发展历程



资料来源：英伟达官网，安信证券研究中心

图26. 2017-2020 年英伟达技术在 TOP500 超算的占比



资料来源：英伟达官网，安信证券研究中心

软硬件共同布局形成生态系统，造就英伟达核心技术壁垒。

➢ **硬件端：基于 GPU、DPU 和 CPU 构建英伟达加速计算平台生态：**

（1）主要产品 Tesla GPU 系列迭代速度快，从 2008 年至 2022 年，先后推出 8 种 GPU 架构，平均两年多推出新架构，半年推出新产品。超快的迭代速度使英伟达的 GPU 性能走在 AI 芯片行业前沿，引领人工智能计算领域发生变革。

（2）DPU 方面，英伟达于 2019 年战略性收购以色列超算以太网公司 Mellanox，利用其 InfiniBand（无限带宽）技术设计出 Bluefield 系列 DPU 芯片，弥补其生态在数据交互方面的不足。InfiniBand 与以太网相同，是一种计算机网络通信标准，但它具有极高的吞吐量和极低的延迟，通常用于超级计算机的互联。英伟达的 Bluefield DPU 芯片可用于分担 CPU 的网络连接算力需求，从而提高云数据中心的效率，降低运营成本。

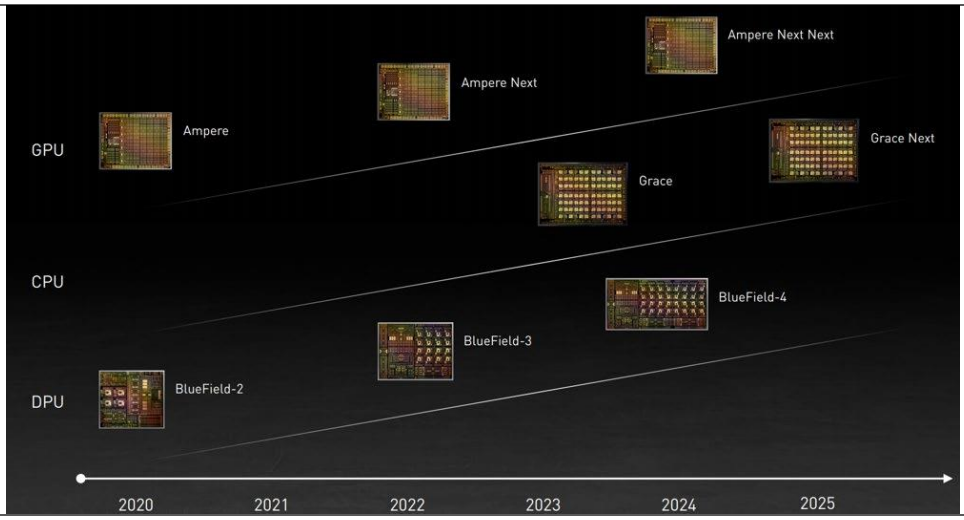
（3）CPU 方面，自主设计 Grace CPU 并推出 Grace Hopper 超级芯片，解决内存带宽瓶颈问题。采用 x86 CPU 的传统数据中心会受到 PCIe 总线规格的限制，CPU 到 GPU 的带宽较小，计算效率受到影响；而 Grace Hopper 超级芯片提供自研 Grace CPU+GPU 相结合的一致内存模型，从而可以使用英伟达 NVLink-C2C 技术快速传输，其带宽是第 5 代 PCIe 带宽的 7 倍，极大提高了数据中心的运行性能。

表9：英伟达 AI 相关产品一览

系列	产品	功能	主要参数
Tesla GPU	P100	进行 AI 领域高性能计算	Pascal 架构，3584 个 CUDA 核心，16GB 显存，732GB/s 带宽，单精度 10.6T
	P4		Pascal 架构，3584 个 CUDA 核心，8GB 显存，192GB/s 带宽，单精度 5.5T
	V100		Volta 架构，5120 个 CUDA 核心，32GB 显存，1134GB/s 带宽，单精度 16.4T
	T4		Turing 架构，2560 个 CUDA 核心，16GB 显存，300GB/s 带宽，单精度 8.1T
	A100		Ampere 架构，6912 个 CUDA 核心，80GB 显存，2039GB/s 带宽，单精度 19.5T
	H100		Hopper 架构，7296 个 CUDA 核心，80GB 显存，3.35TB/s 带宽，单精度 60T
DGX	DGX-1	针对企业和组织的大规模 AI 基础架构，一站式超级计算机	8 块 V100 GPU，双路英特尔至强 E5-2698 CPU，512GB DDR4
	DGX-2		16 块 V100 GPU，双路英特尔至强 8168 CPU，1.5TB 内存
	DGX-A100		8 块 A100 GPU，双路 AMD Rome 7742 CPU，2TB 内存
	DGX-H100		8 块 H100 GPU，双路 x86 CPU，2TB 内存
Grace CPU	Grace CPU	超高内存带宽，专为搭载英伟达 GPU 的数据中心而设计	144 个 Arm Neoverse V2 核心，1TB 带宽，封装密度是 DIMM 的两倍
	Grace Hopper	适用于大规模 AI 和 HPC 应用，提供 CPU+GPU 相结合的一致内存模型	900 GB/s 一致性接口，比 PCIe 5.0 快 7 倍
Bluefield DPU	Bluefield-2 DPU	为云端、数据中心或边缘计算等环境中的各种工作负载提供安全加速	速度 100GB/s 双端口或 200GB/s 单端口，8GB/16GB DDR4
	Bluefield-3 DPU		速度最高 400GB/s，1/2/4 个端口，16GB DDR5

资料来源：英伟达官网，安信证券研究中心

图27. 英伟达产品规划图



资料来源：英伟达官网，安信证券研究中心

相较于 A100 GPU，H100 性能再次大幅提升。在 H100 配备第四代 Tensor Core 和 Transformer 引擎（FP8 精度），同上一代 A100 相比，AI 推理能力提升 30 倍。其核心采用的是 TSMC 目前最先进的 4nm 工艺，H100 使用双精度 Tensor Core 的 FLOPS 提升 3 倍。

表10：Nvidia A100 GPU 和 H100 GPU 规格对比

规格	A100 SXM	A100 PCIe	H100 SXM	H100 PCIe
FP64	9.7 TFLOPS		34 TFLOPS	26 TFLOPS
FP64 Tensor Core	19.5 TFLOPS		67 TFLOPS	51 TFLOPS
FP32	19.5 TFLOPS		67 TFLOPS	51 TFLOPS
TF32	156 TFLOPS | 312 TFLOPS*		989 TFLOPS*	756 TFLOPS*
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*		1979 TFLOPS*	1513 TFLOPS*
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*		1979 TFLOPS*	1513 TFLOPS*
FP8 Tensor Core	-		3958 TFLOPS*	3026 TFLOPS*
INT8 Tensor Core	624 TOPS | 1248 TOPS		3958 TOPS*	3026 TOPS*
GPU 显存	80GB HBM2e	80GB HBM	80GB	80GB
GPU 显存带宽	2039 GB/s	1935 GB/s	3.35 TB/S	2 TB/s
解码器	-	-	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
最大热设计功率（TDP）	400W	300W	700W（可配置）	300-350W（可配置）
多实例 GPU	最大为 7 MIG @ 10GB	最大为 7 MIG @ 5G8	最多 7 个 MIG @每个 10GB	
互连	NVLink：600 GB/s PCIe4.0：64 GB/s	NVIDIA® NVLink®桥接器 2 块 GPU：600GB/s PCIe 4.0：64 GB/s	NVLink：900GB/S PCIe 5.0：128GB/S	NVLink：600GB/S PCIe 5.0：128GB/S

资料来源：英伟达官网，安信证券研究中心

（注：* 采用稀疏技术显示。在不采用稀疏技术的情况下，规格降一半）

表11：Nvidia 计算卡进化历程

训练	K40	M40	P100	V100	A100
发布时间	2013.11	2015.11	2016.4	2017.05	2020.05
架构	Kepler	Maxwell	Pascal	Volta	Ampere
制程	28nm	28nm	16nm	12nm	7nm
晶体管数量	71 亿	80 亿	153 亿	211 亿	510 亿
Die Size	551mm²	601mm²	610mm²	815mm²	826mm²
最大功耗	235W	250W	300W	300W	400W
Streaming Multiprocessors	15	24	56	80	108
Tensor Cores	NA	NA	NA	640	432
FP64 CUDA Cores	960	960	1792	2560	3456
FP32 CUDA Cores	2880	3072	3584	5120	6912
FP32 峰值算力	5.04 TFLOPS	6.08 TFLOPS	10.6 TFLOPS	15.7 TFLOPS	19.5 TFLOPS
稀疏 Tensor Core FP32 峰值算力	NA	NA	NA	NA	312TFLOPS

资料来源：英伟达官网，安信证券研究中心

国内 GPGPU 生态起步较晚，国产 GPU 亟待补位。根据华为 2021 年 9 月发布的《智能世界 2030》报告，人类将于 2030 年进入 YB 数据时代，通用算力相较 2020 年增长 10 倍、人工智能算力

增长 500 倍。在算力需求快速增长的进程中，国产 GPU 正面临机遇与挑战并存的局面。目前，国产 GPU 厂商的核心架构多为自研，难度极高，需投入海量资金以及高昂的人力和时间成本。由于我国 GPU 行业起步较晚，缺乏相应生态，目前同国际一流厂商仍存在较大差距。在中美摩擦加剧、经济全球化逆行的背景下，以海光信息、天数智芯、壁仞科技和摩尔线程等为代表的国内 GPU 厂商进展迅速，国产 GPU 自主可控未来可期。

表12：国产 GPU 厂商情况

公司名称	成立时间	核心产品
海光信息	2014 年	DCU 8000 系列
天数智芯	2015 年	天垓 100、智铠 100
壁仞科技	2019 年	壁砺™100P、壁砺™104P
摩尔线程	2020 年	MTT S80、MTT S50、MTT S3000
景嘉微	2006 年	JM5、JM7、JM9 系列
沐曦集成电路	2020 年	MXN、MXC、MXG
芯瞳半导体	2019 年	GenBu01
龙芯中科	2010 年	龙芯 7A1000、7A2000
芯动科技	2007 年	风华系列 GPU

资料来源：各公司官网，安信证券研究中心

表13：国产 GPU 与国际 GPU 参数对比

厂商	海光	NVIDIA	AMD
品牌	深算一号	A100	MI100
生产工艺	7nm	7nm	7nm
核心数量	4096(64CUs)	2560 CUDA processors 640 Tensor processors	120CUs
内核频率	Up to 1.5GHz(FP64) Up to 1.7GHz(FP32)	Up to 1.53Ghz	Up to 1.5GHz (FP64) Up to 1.7Ghz (FP32)
显存容量	32GB HBM2	80GB HBM2e	32GB HBM2
显存位宽	4096 bit	5120 bit	4096bit
显存频率	2.0 GHz	3.2 GHz	2.4 GHz
显存带宽	1024 GB/s	2039 GB/s	1228 GB/s
TDP	350W	400W	300W
CPU to GPU 互联	PCIe Gen4 x16	PCIe Gen4 x16	PCIe GEN4 x16
GPU to GPU 互联	xGMI x 2 Up to 184 GB/s	NVLink up to 600 GB/s	Infinity Fabric x3 up to 276 GB/s

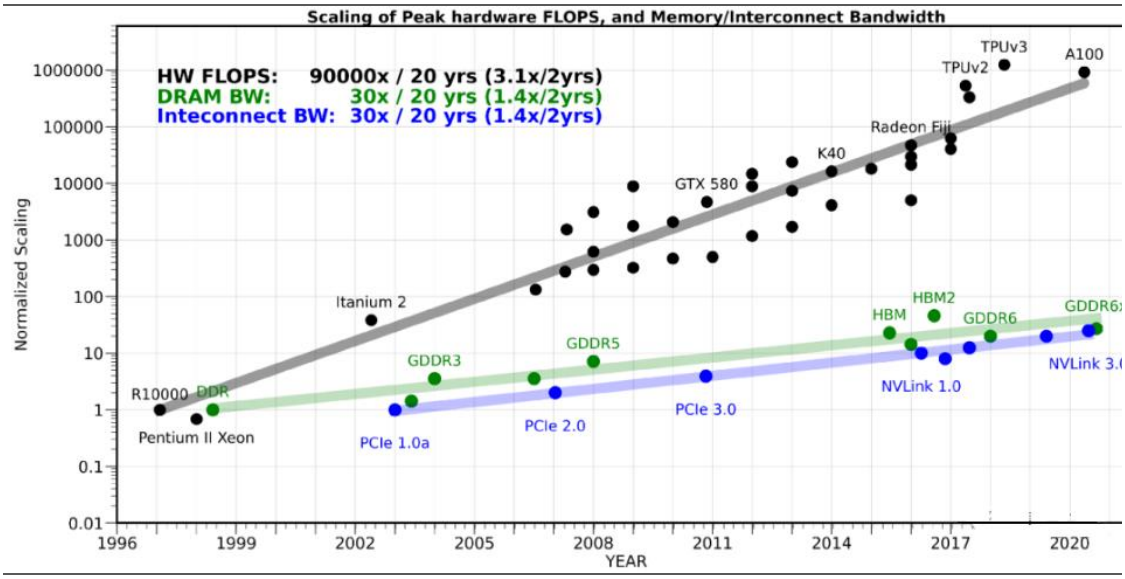
资料来源：中国计量科学研究院，海光信息招股书，安信证券研究中心

2. 大算力场景遇到的问题及解决途径

2.1. “内存墙”、“功耗墙”等掣肘 AI 的算力发展

“存”“算”性能失配，内存墙导致访存时延高，效率低。内存墙，指内存的容量或传输带宽有限而严重限制 CPU 性能发挥的现象。内存的性能指标主要有“带宽”(Bandwidth)和“等待时间”(Latency)。近 20 年间，运算设备的算力提高了 90000 倍，提升非常快。虽然存储器从 DDR 发展到 GDDR6x，能够用于显卡、游戏终端和高性能运算，接口标准也从 PCIe1.0a 升级到 NVLink3.0，但是通讯带宽的增长只有 30 倍，和算力相比提高幅度非常缓慢。

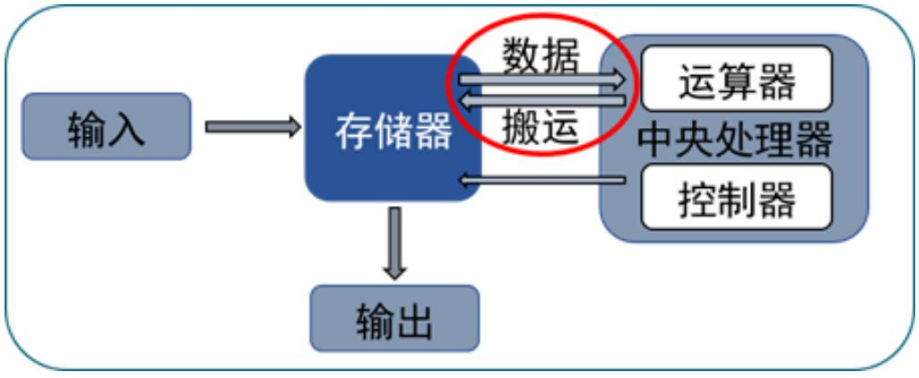
图28. 存储计算“剪刀差”



资料来源：OneFlow 公司公众号，安信证券研究中心

冯诺依曼架构下，数据传输导致严重的功耗损失。冯·诺依曼架构要求数据在存储器单元和处理单元之间不断地“读写”，这样数据在两者之间来回传输就会消耗很多的传输功耗。根据英特尔的研究表明，当半导体工艺达到 7nm 时，数据搬运功耗高达 35pJ/bit，占总功耗的 63.7%。数据传输造成的功耗损失越来越严重，限制了芯片发展的速度和效率，形成了“功耗墙”问题。

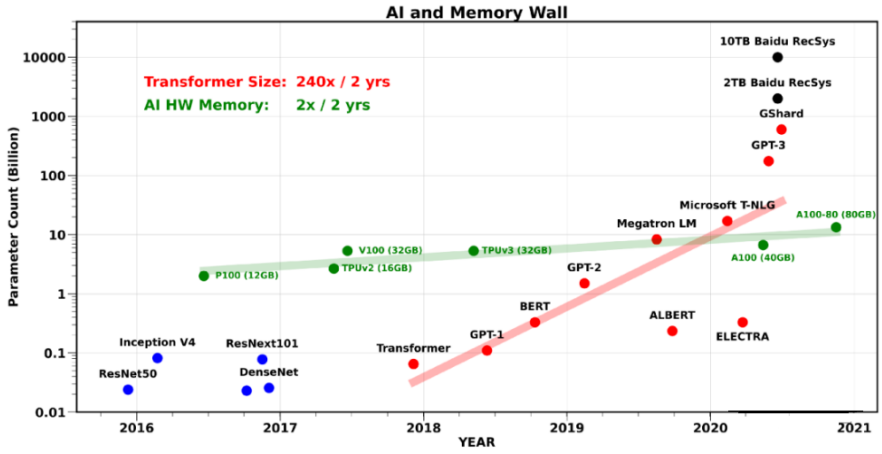
图29. 冯诺依曼架构下的数据传输



资料来源：中国科学信息科学，安信证券研究中心

AI 模型参数量极速扩大，GPU 内存增长速度捉襟见肘。在 GPT-2 之前的模型时代，GPU 内存还能满足 AI 大模型的需求。近年来，随着 Transformer 模型的大规模发展和应用，模型大小每两年平均增长了 240 倍。GPT-3 等大模型的参数增长已经超过了 GPU 内存的增长。传统的设计趋势已经不能适应当前的需求，芯片内部、芯片之间或 AI 加速器之间的通信成为了 AI 训练的瓶颈。AI 训练不可避免地遇到了“内存墙”问题。

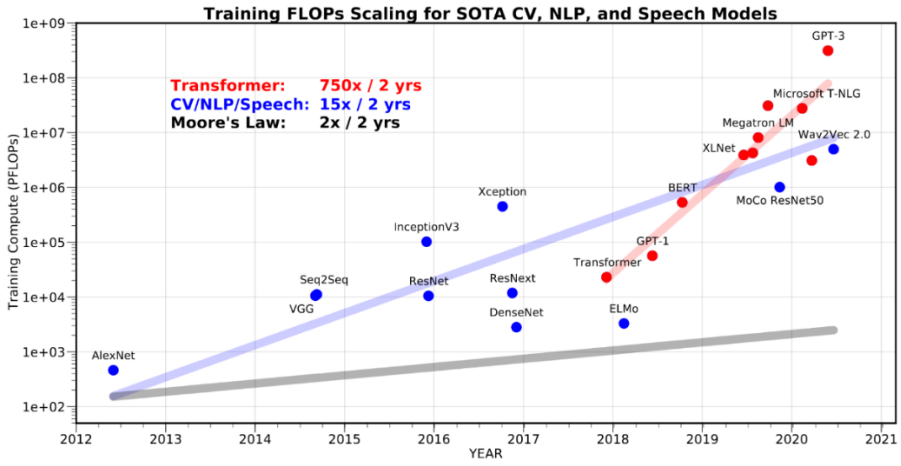
图30. AI 模型大小增长与 GPU 内存增长



资料来源：OneFlow 公司公众号，安信证券研究中心

AI 模型运算量增长速度不断加快，推动硬件算力增长。预训练技术的进步导致了各领域模型计算量的快速增长，大约每两年就要增加 15 倍。而 Transformer 类模型的运算量更是每两年就要增加 750 倍。这种近乎指数的增长趋势促使 AI 硬件的研发方向发生变化，需要更高的峰值算力。当前的研究为了实现更高的算力，甚至不惜简化或者优化其他部分组件，例如内存的分层架构，将 DRAM 容量用于需要高性能访问的热数据，将容量层用于处理需要大容量但性能要求不那么高的任务，以适应不同的数据类型、用例、技术需求和预算限制，适用于 AI、ML 和 HPC 等众多应用场景，能帮助企业以经济高效的方式满足内存需求。

图31. AI 模型计算量增长速度



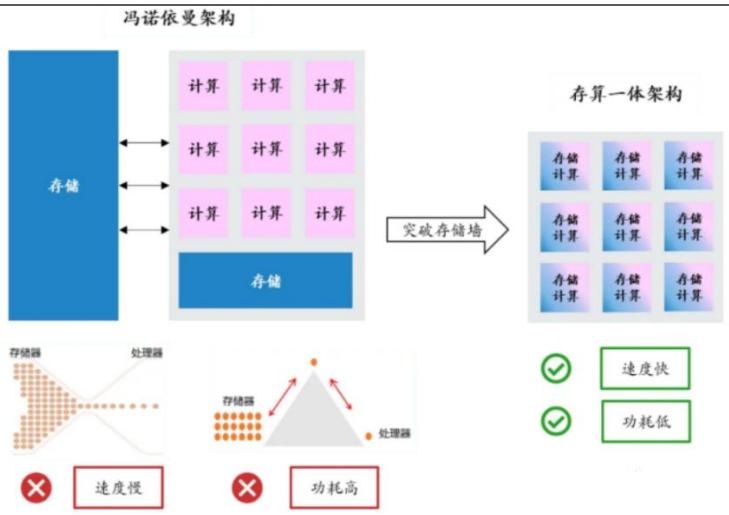
资料来源：OneFlow 公司公众号，安信证券研究中心

2.2. “内存墙”、“功耗墙”等问题解决路径

2.2.1. 存算一体技术：以 SRAM、RRAM 为主的新架构，大算力领域优势大

存算一体在存储器中嵌入计算能力，以新的运算架构进行乘加运算。存算一体是一种以数据为中心的非冯诺依曼架构，它将存储功能和计算功能有机结合起来，直接在存储单元中处理数据。存算一体通过改造“读”电路的存内计算架构，可以直接从“读”电路中得到运算结果，并将结果“写”回存储器的目标地址，避免了在存储单元和计算单元之间频繁地转移数据。存算一体减少了不必要的数据搬移造成的开销，不仅大幅降低了功耗（降至 1/10~1/100），还可以利用存储单元进行逻辑计算提高算力，显著提升计算效率。它不仅适用于 AI 计算，也适用于感存算一体芯片和类脑芯片，是未来大数据计算芯片架构的主流方向。

图32. 冯诺依曼架构 vs 存算一体架构

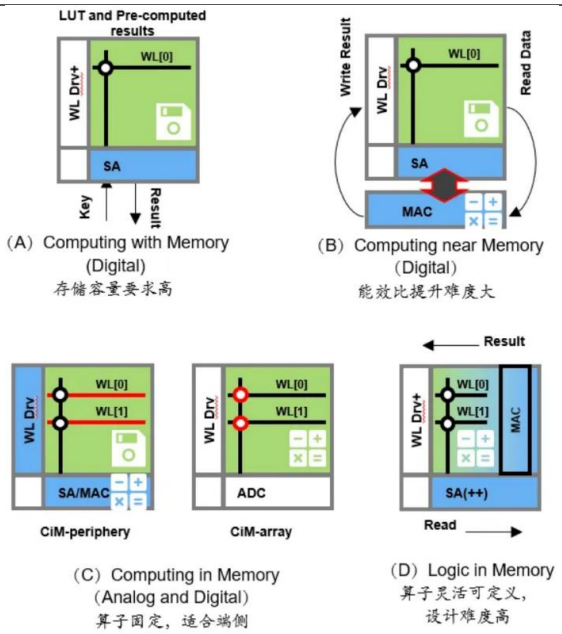


资料来源：九章智驾，安信证券研究中心

存算一体技术可分为查存计算、近存计算、存内计算和存内逻辑，提供多种方式解决内存墙问题。

- **查存计算：**早期技术，在存储芯片内部查表来完成计算操作。
- **近存计算：**早已成熟，计算操作由位于存储区域外部的独立计算芯片/模块完成。典型代表是 AMD 的 Zen 系列 CPU，以及封装 HBM 内存（包括三星的 HBM-PIM）与计算模组（裸 Die）的芯片。
- **存内计算：**计算操作由位于存储芯片/区域内部的独立计算单元完成，存储和计算可以是模拟或数字的。典型代表是 Mythic、千芯科技、闪亿、知存、九天睿芯等。
- **存内逻辑：**通过在内部存储中添加计算逻辑，直接在内部存储执行数据计算。典型代表包括 TSMC（在 2021 ISSCC 发表论文）和千芯科技。

图33. 四种存算一体架构对比



资料来源：清华大学微电子所，安信证券研究中心

SRAM、RRAM 是存算一体介质的主流研究方向。存算一体的成熟存储器有几种，比如 NOR FLASH、SRAM、DRAM、RRAM、MRAM 等 NVRAM。

- FLASH 是非易失性存储，成本低，可靠性高，但制程有瓶颈。

- SRAM 速度快，能效比高，在存内逻辑技术发展后有高能效和高精度的特点。
- DRAM 容量大，成本低，但速度慢，需要不断刷新电力。
- 新型存储器 PCAM、MRAM、RRAM 和 FRAM 也适用于存算一体。其中 RRAM 在神经网络计算中有优势，是下一代存算一体介质的主流方向之一。除了 SRAM 之外，RRAM 也是未来发展最快的新型存储器之一，它结构简单，速度高，但材料不稳定，工艺还需 2-5 年才能成熟。

表14：不同存储器介质对比

存储器类型	优势	不足	适合场景
SRAM(数字模式)	能效比高，高速高精度，对噪声不敏感，工艺成熟先进，适合 IP 化	存储密度略低	大算力、云计算、边缘计算
SRAM(模拟模式)	能效比高，工艺成熟先进	对 PVT 变化敏感，对信噪比敏感，存储密度略低	小算力、端侧、不要求待机功耗
各类 NVRAM（包括 RRAM/MRAM 等）	能效比高，高密度，非易失，低漏电	对 PVT 变化敏感，有限写次数，相对低速，工艺良率尚在爬坡中	小算力、端侧/边缘 Inference、待机时间长的场景力
Flash	高密度低成本，非易失，低漏电	对 PVT 变化敏感，精度不高，工艺迭代时间长	小算力、端侧、低成本、待机时间长的场景
DRAM	高存储密度，整合方案成熟	只能做近存计算，速度略低，工艺迭代慢	适合现有冯氏架构向存算过渡

资料来源：陈巍谈芯，安信证券研究中心

存算一体有着广泛的应用场景，在不同大小设备上均有需求。

- 从技术领域来看，存算一体可以应用于：

（1）AI 和大数据计算：将 AI 计算中大量乘加计算的权重部分存在存储单元中，从而在读取的同时进行数据输入和计算处理，在存储阵列中完成卷积运算。

（2）感存算一体：集传感、储存和运算为一体构建感存算一体架构，在传感器自身包含的 AI 存算一体芯片上运算，来实现零延时和超低功耗的智能视觉处理能力。

（3）类脑计算：使计算机像人脑一样将存储和计算合二为一，从而高速处理信息。存算一体天然是将存储和计算结合在一起的技术，是未来类脑计算的首选和产品快速落地的关键。

- 从应用场景来分，存算一体可以适用于各类人工智能场景和元宇宙计算，如可穿戴设备、移动终端、智能驾驶、数据中心等。

（1）针对端侧的可穿戴等小设备，对成本、功耗、时延难度很敏感。端侧竞品众多，应用场景碎片化，面临成本与功效的难题。存算一体技术在端侧的竞争力影响约占 30%。（例如 arm 占 30%，降噪或 ISP 占 40%，AI 加速能力只占 30%）

（2）针对云计算和边缘计算的大算力设备，是存算一体芯片的优势领域。存算一体在大算力领域的竞争力影响约占 90%。

传统存储大厂纷纷入局，新兴公司不断涌现。

（1）国外方面，三星电子在多个技术路线进行尝试，发布新型 HBM-PIM（存内计算）芯片、全球首个基于 MRAM（磁性随机存储器）的存内计算研究等。台积电在 ISSCC 2021 上提出基于数字改良的 SRAM 设计存内计算方案。英特尔也早早提出近内存计算战略，将数据在存储层级向上移动，使其更接近处理单元进行计算。

（2）国内方面，阿里达摩院成功研发全球首款基于 DRAM 的 3D 键合堆叠存算一体芯片，可突破冯·诺依曼架构的性能瓶颈。千芯科技是可重构存算一体 AI 芯片的领导者和先驱，核心产品包括高算力低功耗的存算一体 AI 芯片/IP 核（支持多领域多模态人工智能算法）。后摩智能致力于突破智能计算芯片性能及功耗瓶颈，其提供的大算力、低功耗的高能效比芯片及解决方案，可应用于无人车、泛机器人等边缘端，以及云端推荐、图像分析等云端推理场景。

表15：云和边缘大算力企业对比

企业名称	标识	场景	架构类型	存储器类型	主力产品	算力(TOPS)	其他
亿铸科技	亿铸科技	边缘为主大算力(ADAS)	全数字存算一体	RRAM	未公布	未公布	上海
千芯科技	千芯科技 TensorChip	云和边缘大算力	存内计算/存内逻辑	RRAM/SRAM	云计算卡 G40710E、G41210E、F11610E、F12010	>1000-4000@INT8	北京，最早支持多实例(虚拟化）计算的存算一体架构
后摩智能	后摩智能 HOUMO.AI	边缘为主大算力(ADAS)	模拟存内计算	SRAM/MRAM/RRAM	智能驾驶芯片	20TOPS	上海
中科声龙	SUNLUNE 中科声龙	云为主大算力	近存计算	SRAM	矿机		北京
d-Matrix	d-Matrix	Transformer 加速	近存 or 存内	SRAM	计算卡	未公布	结合 Chiplet

资料来源：陈巍谈芯，安信证券研究中心

表16：端和边缘小算力企业对比

企业名称	标识	场景	架构类型	存储器类型	主力产品	算力(TOPS)	其他
闪易半导体/闪亿	FLASH BILLION 闪易半导体	端侧小算力	模拟存内计算	闪存/自主核心工艺	语音/图像 HEXA01	未公布能效比明显优于某家	陈大同投资产品量产
Mythic	MYTHIC	边缘小算力	模拟存内计算	闪存	MP10304	100	
SST/Cypress	SST A Microchip Technology Company	端侧小算力	模拟存内计算	闪存/SF	memBrainIP 核	未公布	2017 年设计 memBrain 技术
知存科技	WITINMEM 知存科技	端侧小算力	模拟存内计算	闪存/SF	语音 WTM2101	0.05@INT8(50GOPS)	使用了 SST 工艺单元，产品量产
每刻深思	MokeSens 每刻深思	端侧小算力	模拟存内计算	SRAM	未公布		感存算一体
九天睿芯	REXEN 九天睿芯	端侧小算力	模拟存内计算	SRAM	图像 ADA20X	0.3-200@INT8	感存算一体
恒烁半导体	恒烁半导体 Zbit Semi.Inc.	端侧小算力	模拟存内计算	闪存/ETOX	CiNOR	未公布	已 IPO 上市
新忆科技	INNOVATION MEMORY	端侧小算力	模拟存内计算	RRAM	xuanwu	未公布	
智芯科	AISTARTEK	端侧小算力	模拟存内计算	SRAM	语音 AT660x	未公布	
苹芯科技	苹芯科技	端侧小算力	存内计算	SRAM	图像 PIMCHIP-S200、语音 PIMCHIP-S100	未公布	

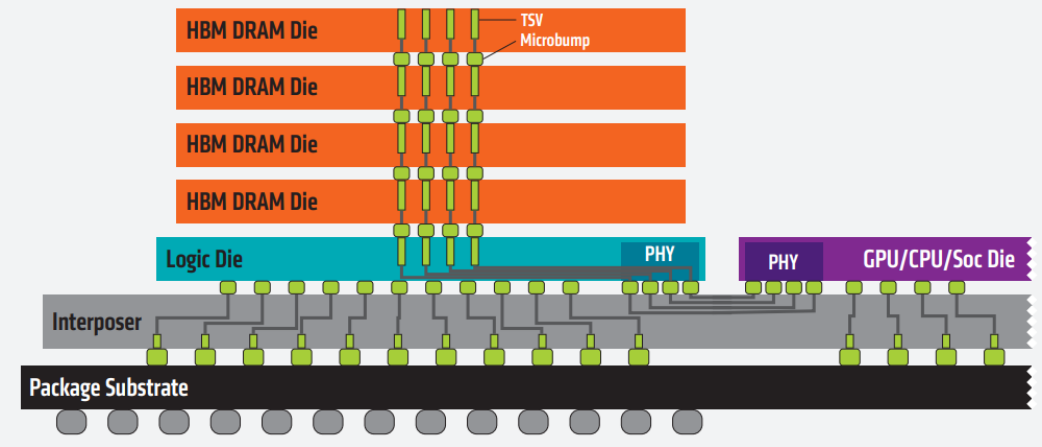
资料来源：陈巍谈芯，安信证券研究中心

2.2.2. HBM 技术：高吞吐高带宽，AI 带动需求激增

HBM（High Bandwidth Memory）意为高带宽存储器，是一种硬件存储介质，是高性能 GPU 的核心组件。HBM 具有高吞吐高带宽的特性，受到工业界和学术界的关注。它单颗粒的带宽可以达到 256 GB/s，远超过 DDR4 和 GDDR6。DDR4 是 CPU 和硬件处理单元的常用外挂存储设备，但是它的吞吐能力不足以满足当今计算需求，特别是在 AI 计算、区块链和数字货币挖矿等大数据处理访存需求极高的领域。GDDR6 也比不上 HBM，它单颗粒的带宽只有 64 GB/s，是 HBM 的 1/4。而 DDR4 3200 需要至少 8 颗粒才能提供 25.6 GB/s 的带宽，是 HBM 的 1/10。

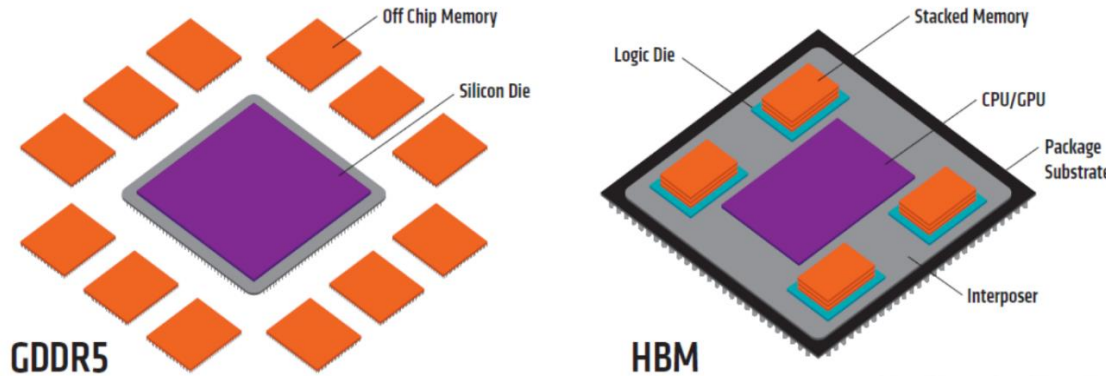
HBM 使用多根数据线实现高带宽，完美解决传统存储效率低的问题。HBM 的核心原理和普通的 DDR、GDDR 完全一样，但是 HBM 使用多根数据线实现了高带宽。HBM/HBM2 使用 1024 根数据线传输数据，作为对比，GDDR 是 32 根，DDR 是 64 根。HBM 需要使用额外的硅联通层，通过晶片堆叠技术与处理器连接。这么多的连接线保持高传输频率会带来高功耗。因此 HBM 的数据传输频率相对很低，HBM2 也只有 2 Gbps，作为对比，GDDR6 是 16 Gbps，DDR4 3200 是 3.2 Gbps。这些特点导致了 HBM 技术高成本，容量不可扩，高延迟等缺点。

图34. HBM 设计结构



资料来源：AMD 官网，安信证券研究中心

图35. GDDR5 vs HBM



资料来源：AMD 官网，安信证券研究中心

HBM 可以被广泛的应用到汽车高带宽存储器，GPU 显存芯片，部分 CPU 的内存芯片，边缘 AI 加速卡，Chiplets 等硬件中。在高端 GPU 芯片产品中，比如 NVDIA 面向数据中心的 A100 等加速卡中就使用了 HBM；部分 CPU 的内存芯片，如目前富岳中的 A64FX 等 HPC 芯片中也有应用到。车辆在快速移动时，摄像头、传感器会捕获大量的数据，为了更快速的处理数据，HBM 是最合适的选择。Chiplets 在设计过程中没有降低对内存的需求，随着异构计算（尤其是小芯片）的发展，芯片会加速对高带宽内存的需求，无论是 HBM、GDDR6 还是 LPDDR6。

HBM 缓解带宽瓶颈，是 AI 时代不可或缺的关键技术。AI 处理器架构的探讨从学术界开始，当时的模型简单，算力低，后来模型加深，算力需求增加，带宽瓶颈出现，也就是 IO 问题。这个问题可以通过增大片内缓存、优化调度模型等方法解决。但是随着 AI 大模型和云端 AI 处理的发展，计算单元剧增，IO 问题更严重了。要解决这个问题需要付出很高的代价（比如增加 DDR 接口通道数量、片内缓存容量、多芯片互联），这便是 HBM 出现的意义。HBM 用晶片堆叠技术和硅联通层把处理器和存储器连接起来，把 AI/深度学习完全放到片上，提高集成度，降低功耗，不受芯片引脚数量的限制。HBM 在一定程度上解决了 IO 瓶颈。未来人工智能的数据量、计算量会越来越大，超过现有的 DDR/GDDR 带宽瓶颈，HBM 可能会是唯一的解决方案。

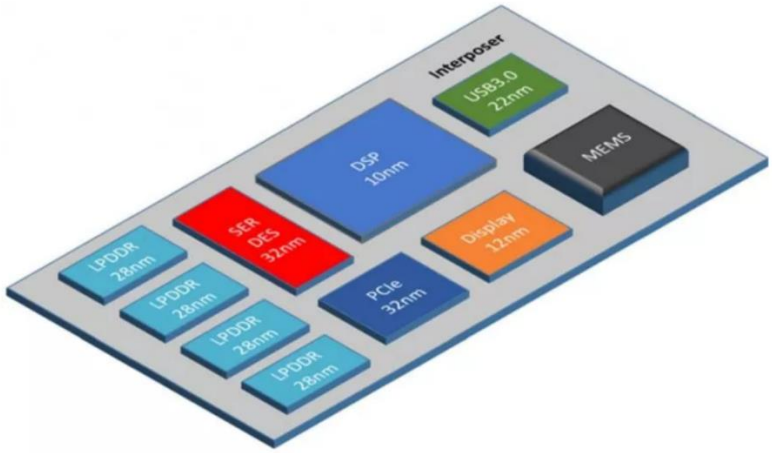
巨头领跑，各大存储公司都已在 HBM 领域参与角逐。SK 海力士、三星、美光等存储巨头在 HBM 领域展开了升级竞赛，国内佰维存储等公司持续关注 HBM 领域。SK 海力士早在 2021 年 10 月就开发出全球首款 HBM3，2022 年 6 月量产了 HBM3 DRAM 芯片，并将供货英伟达，持续

巩固其市场领先地位。三星也在积极跟进，在 2022 年技术发布会上发布的内存技术发展路线图中，HBM3 技术已经量产。伴随着 ChatGPT 的火热，整个市场对于高性能计算卡等硬件产品的需求水涨船高，上游大厂如三星和海力士目前的 DRAM 业务相关订单激增。GPU 公司英伟达一直在要求 SK 海力士提供最新的 HBM3 内存颗粒。服务器 CPU 公司英特尔在全新的第四代至强可扩展处理器当中也推出了配备 SK 海力士 HBM 的产品。

2.2.3. Chiplet 技术：全产业链升级降本增效，国内外大厂前瞻布局

Chiplet 即根据计算单元或功能单元将 SOC 进行分解，分别选择合适制程工艺制造。随着处理器的核越来越多，芯片复杂度增加、设计周期越来越长，SoC 芯片验证的时间、成本也急剧增加，特别是高端处理芯片、大芯片。当前集成电路工艺在物理、化学很多方面都达到了极限，大芯片快要接近制造瓶颈，传统的 SoC 已经很难继续被采纳。Chiplet，俗称小芯片、芯粒，是将一块原本复杂的 SoC 芯片，从设计的时候就按照不同的计算单元或功能单元进行分解，然后每个单元分别选择最合适的半导体制程工艺进行制造，再通过先进封装技术将各自单元彼此互联。Chiplet 是一种类似搭乐高积木的方法，能将采用不同制造商、不同制程工艺的各种功能芯片进行组装，从而实现更高良率、更低成本。

图36. Chiplet 设计结构



资料来源：eefocus，安信证券研究中心

Chiplet 可以从多个维度降低成本，延续摩尔定律的“经济效益”。随着半导体工艺制程推进，晶体管尺寸越来越逼近物理极限，所耗费的时间及成本越来越高，同时所能够带来的“经济效益”的也越来越有限。Chiplet 技术可从三个不同的维度来降低成本：

（1）可大幅度提高大型芯片的良率：芯片的良率与芯片面积有关，Chiplet 设计将大芯片分成小模块可以有效改善良率，降低因不良率导致的成本增加。

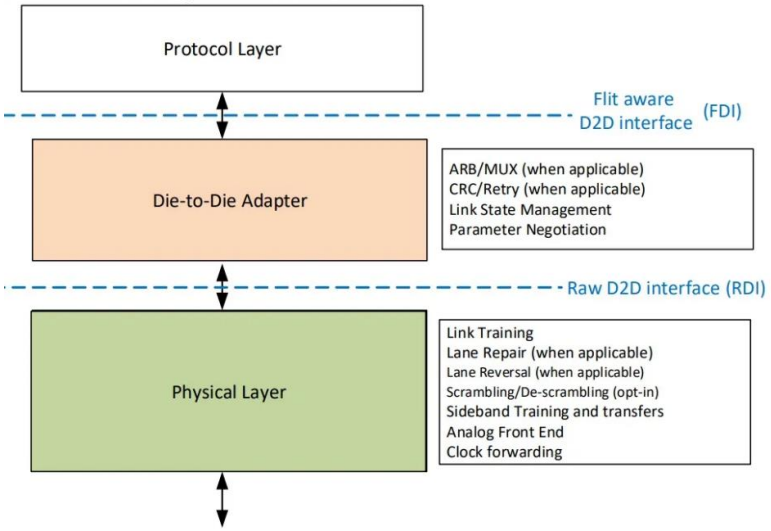
（2）可降低设计的复杂度和设计成本：Chiplet 通过在芯片设计阶段就将 Soc 按照不同功能模块分解成可重复云涌的小芯粒，是一种新形式的 IP 复用，可大幅度降低设计复杂度和成本累次增加。

（3）可降低芯片制造的成本：在 Soc 中的一些主要逻辑计算单元是依赖于先进工艺制程来提升性能，但其他部分对制程的要求并不高，一些成熟制程即可满足需求。将 Soc 进行 Chiplet 化后对于不同的芯粒可选择对应合适的工艺制程进行分开制造，极大降低芯片的制造成本。

Chiplet 为全产业链提供了升级机会。在后摩尔时代，Chiplet 可以开启一个新的芯片生态。2022 年 3 月，Chiplet 的高速互联标准——UCIe（UniversalChiplet Interconnect Express，通用芯粒互联技术）正式推出，旨在芯片封装层面确立互联互通的统一标准，打造一个开放性的 Chiplet 生态系统。巨头们合力搭建起了统一的 Chiplet 互联标准，将加速推动开放的 Chiplet 平台发展，并横跨 x86、Arm、RISC-V 等架构和指令集。Chiplet 的影响力也从设计端走到芯片制造与封装环节。在芯片小型化的设计过程中，需要添加更多 I/O 与其他芯片芯

片接口，裸片尺寸必须要保持较大的空白空间。而且，要想保证 Chiplet 的信号传输质量就需要发展高密度、大宽带布线的先进封装技术。另外，Chiplet 也影响到从 EDA 厂商、晶圆制造和封装公司、芯粒 IP 供应商、Chiplet 产品及系统设计公司到 Fabless 设计厂商的产业链各个环节的参与者。

图37. UCIe 标准



资料来源：UCIe，安信证券研究中心

乾坤未定，Chiplet 是国内芯片相关公司的重要发展机遇。（1）最先受到影响的是芯片 IP 设计企业，Chiplet 本质就是不同的 IP 芯片化，国内类似 IP 商均有望参与其中，比如华为海思有 IP 甚至指令集开发实力的公司，推出基于 RISC-V 内核的处理器（玄铁 910）阿里平头哥半导体公司，独立的第三方 IP 厂商，如芯动科技、芯原股份、芯耀辉、锐成芯微、芯来等众多 IP 公司等。（2）Chiplet 需要 EDA 工具从架构探索、芯片设计、物理及封装实现等提供全面支持，为国内 EDA 企业发展带来了突破口。芯和半导体已全面支持 2.5D Interposer、3DIC 和 Chiplet 设计。（3）Chiplet 也推动了先进封装技术的发展。根据长电科技公告，在封测技术领域取得新的突破。4nm 芯片作为先进硅节点技术，是导入 Chiplet 封装的一部分通富微电提供晶圆级及基板级封装两种解决方案，其中晶圆级 TSV 技术是 Chiplet 技术路径的一个重要部分。

表17：Chiplet 相关公司产品

产业链	公司	相关研发
芯片 IP	华为海思	有自己的 IP 甚至指令集开发实力，但不对外
	阿里平头哥	在 2019 年推出基于 RISC-V 内核的处理器（玄铁 910）
	芯原股份	基于 Chiplet 架构所设计了高端应用处理器平台
	芯动科技	发布了自研的首套跨工艺、跨封装物理层兼容 UCIe 国际标准的 Innolink Chiplet 解决方案。提供从 0.18um 到 5nm 全套高速混合电路 IP 核
	芯来科技	RISC-V 生态引领者
	华夏芯	拥有完全自主知识产权的 CPU、DSP、GPU 和 AI 处理器 IP
	华大九天	高速接口 IP
EDA 工具	芯和半导体	首发了“3DIC 先进封装设计分析全流程”EDA 平台，是业界首个用于 3DIC 多芯片系统设计分析的统一平台，全面支持 2.5D Interposer、3DIC 和 Chiplet 设计
	瞬曜	只面 Chiplet 设计方法学对数字验证的新挑战，为解决系统级高速验证和仿真方面的需求做出了攻关
先进封装	大和股份	掌握了 2.5D、3D 先进封装的关键掌握了 TSV 技术
	长电科技	在封测技术领域取得新的突破，实现 4nm 工艺制程手机芯片的封装，以及 CPU、GPU 和射频芯片的集成封装。4nm 芯片作为先进硅节点技术，也是导入 Chiplet 封装的一部分
	通富微电	公司技术布局进展顺利，已开始大规模生产 Chiplet 产品，工艺节点方面 7nm 产品实现量产，5nm 产品完成研发。针对 Chiplet，通富微电提供晶圆级及基板级封装两种解决方案，其中晶圆级 TSV 技术是 Chiplet 技术路径的一个重要部分

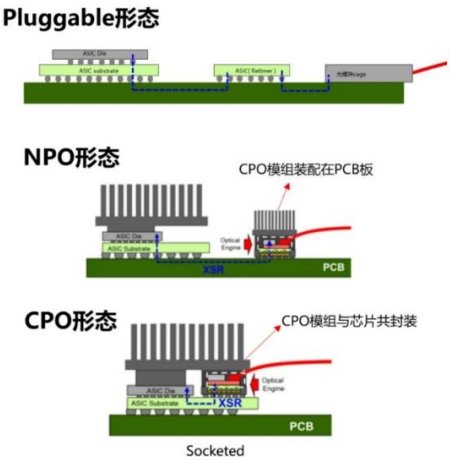
资料来源：公司官网，安信证券研究中心

国外芯片厂率先发力，通过 Chiplet 实现收益。AMD 的 EPYC 率先采用了 Chiplet 结构，实现了在服务器 CPU 市场上的翻身。随后，Ryzen 产品上重用了 EYPC Rome 的 CCD，这样的 chiplet 设计极好的降低了总研发费用。2023 年 1 月，Intel 发布了采用了 Chiplet 技术的第四代至强可扩展处理器 Sapphire Rapids 以及英特尔数据中心 GPU Max 系列等。Sapphire Rapids 是 Intel 首个基于 Chiplet 设计的处理器，被称为“算力神器”。Xilinx 的 2011 Virtex-7 2000T 是 4 个裸片的 Chiplet 设计。Xilinx 也是业界唯一的同构和异构的 3D IC。

2.2.4. CPO 技术：提升数据中心及云计算效率，应用领域广泛

CPO（Co-packaged，共封装光学技术）是高速电信号能够高质量的在交换芯片和光引擎之间传输。在 5G 时代，计算、传输、存储的带宽要求越来越高，同时硅光技术也越来越成熟，因此板上和板间的光互连成为了一种必要的方式。随着通道数大幅增加，需要专用集成电路（ASIC）来控制多个光收发模块。传统的连接方式是 Pluggable（可插拔），即光引擎是可插拔的光模块，通过光纤和 SerDes 通道与网络交换芯片（AISC）连接。之后发展出了 NPO（Near-packaged，近封装光学），一种将光引擎和交换芯片分别装配在同一块 PCB 基板上的方式。而 CPO 是一种将交换芯片和光引擎共同装配在同一个 Socketed（插槽）上的方式，形成芯片和模组的共封装，从而降低网络设备的功耗和散热问题。NPO 是 CPO 的过渡阶段，相对容易实现，而 CPO 是最终解决方案。

图38. 共封装光学技术



资料来源：CSDN，安信证券研究中心

随着大数据及 AI 的发展，数据中心的需求激增，CPO 有着广泛的应用前景。在数据中心领域，CPO 技术可以实现更高的数据密度和更快的数据传输速度，还可以减少系统的功耗和空间占用，降低数据中心的能源消耗和维护成本，能够应用于高速网络交换、服务器互联和分布式存储等领域，例如，Facebook 在其自研的数据中心网络 Fabric Aggregator 中采用了 CPO 技术，提高了网络的速度和质量。在云计算领域，CPO 技术可以实现高速云计算和大规模数据处理。例如微软在其云计算平台 Azure 中采用了 CPO 技术，实现更高的数据密度和更快的数据传输速度，提高云计算的效率和性能。在 5G 通信领域，CPO 技术可以实现更快的无线数据传输和更稳定的网络连接。例如华为在其 5G 通信系统中采用了 CPO 技术，将收发器和芯片封装在同一个封装体中，从而实现了高速、高密度、低功耗的通信。除此之外，5G/6G 用户的增加，人工智能、机器学习（ML）、物联网（IoT）和虚拟现实流量的延迟敏感型流量激增，对光收发器的数据速率要求将快速增长；AI、ML、VR 和 AR 对数据中心的带宽要求巨大，并且对低延迟有极高的要求，未来 CPO 的市场规模将持续高速扩大。

CPO 技术壁垒高，通信公司成为主要参与者，发展迅速。锐捷网络于 2022 年正式推出了首款应用 CPO 技术的数据中心交换机，截至目前正式发布了多款同时应用硅光技术和液冷技术的交换机，散热成本对比同性能的可插拔光模块设备降低了 35%。联特科技专注研发基于 EML（电吸收调制激器）、SIP（硅光）、TFLN（薄膜铌酸锂）调制技术的 800G 光模块，以及用于下一代产品 NPO（近封装光学）/CPO（共封装光学）所需的高速光连接技术、激光器技术和芯片级光电混合封装技术等。新易盛的光膜块 400G 已广泛应用在各大数据中心，更高端的 800G 已实现产业化出货走在行业引领前端，且光模块已突破低功耗极限，同时布局了光电共同封装（CPO）技术，双重受益，行业需求增量大。中际旭创 400G 系列相干产品已逐步在国内主流设备商和互联网云厂商中得到了应用，同时也发布了 800G 的解决方案，部分光模块使用自家研制的硅光芯片。

3. 投资建议

通过探讨对 AI 大模型的算力需求及相关衍生问题，我们梳理了硬件及应用端的产业链和核心技术路径，建议关注如下国内标的：

（1） GPU/AI 芯片：寒武纪、海光信息、景嘉微、澜起科技
（2） 英伟达产业链配套：胜宏科技、和林微纳
（3） CPU：海光信息、龙芯中科、澜起科技
（4） FPGA：紫光国微、复旦微电、安路科技
（5） 芯片 IP：芯原股份、华大九天
（6） 服务器：浪潮信息、工业富联、中科曙光

（7） Chiplet 等先进封装相关：通富微电、长电科技、兴森科技、深南电路、生益科技、华正新材
（8） 光模块：天孚通信、新易盛、中际旭创
（9） AIoT:乐鑫科技、恒玄科技、炬芯科技
（10） SoC：富瀚微、晶晨股份、瑞芯微、全志科技、恒玄股份
（11） Risk-V：兆易创新、芯原股份、国芯科技、北京君正
（12） 存算一体：兆易创新、恒烁股份
（13） 存储芯片/模组：兆易创新、佰维存储、江波龙、北京君正、聚辰股份
（14） CPU/GPU 等供电芯片：杰华特、晶丰明源
（15） 多模态下游应用：海康威视、大华股份、萤石网络、漫步者等

4. 风险提示

4.1. 技术研发不及预期的风险

大语言模型涉及对高性能硬件（如 GPU、TPU）、大规模高质量数据集的需求以及软件算法的提高等多方面要求。如果公司不能紧跟 AI 模型的技术发展趋势，及时进行技术升级迭代，公司将面临市场竞争力下降的风险，公司产品和技术存在被替代的风险。

4.2. 应用落地不及预期的风险

AI 大模型发展时间较短，尚处于商业化探索的早期阶段。如果公司无法研发出具有商业应用价值的 AI 产品，或者相关 AI 产品不符合市场需求、性价比超出市场承受能力，公司将面临研发投入无法获得收入回报的风险。

4.3. 中美贸易摩擦的风险

我国 AI 进度和 GPU 芯片与国际巨头存在差距，如果未来中美贸易摩擦进一步加剧，可能会对公司供应稳定性、及时性和价格产生不利影响，进而影响公司技术迭代升级和业务发展，从而可能对公司生产经营和盈利能力带来潜在的不利影响。

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%及以上；
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%及以上；
风险评级：
A —— 正常风险，未来 6 个月的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 个月的投资收益率的波动大于沪深 300 指数波动；

分析师声明

本报告署名分析师声明，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

安信证券股份有限公司（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

本报告仅供安信证券股份有限公司（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，如有需要，客户可以向本公司投资顾问进一步咨询。

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。

本报告的估值结果和分析结论是基于所预定的假设，并采用适当的估值方法和模型得出的，由于假设、估值方法和模型均存在一定的局限性，估值结果和分析结论也存在局限性，请谨慎使用。

安信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。

安信证券研究中心

深圳市

地　址：　深圳市福田区福田街道福华一路 19 号安信金融大厦 33 楼

邮　编：　518026

上海市

地　址：　上海市虹口区东大名路 638 号国投大厦 3 层

邮　编：　200080

北京市

地　址：　北京市西城区阜成门北大街 2 号楼国投金融大厦 15 层

邮　编：　100034