华泰证券 HUATAI SECURITIES

AI 芯片风继续吹：群贤毕至，花落谁家？

华泰研究 - 海外科技

2023 年 9 月 22 日 | 美国　　首次覆盖

电子	增持（首评）
半导体	增持（首评）

研究员　何翩翩
SAC No. S0570523020002　purdyho@htsc.com
SFC No. ASI353　+(852) 3658 6000

华泰证券研究所分析师名录

重点推荐

股票名称	股票代码	目标价（当地币种）	投资评级
英伟达	NVDA US	650.00	买入
超威半导体	AMD US	150.00	买入

资料来源：华泰研究预测

人工智能风继续吹，AI 芯片乘风而起，但 B 端应用落地才是制胜关键

本轮 AI 浪潮由 ChatGPT 掀起，并引发各中外科技企业展开对大语言模型及生成式 AI 的追逐和对算力的军备竞赛。GPT 背后的核心算法是谷歌在 2017 年提出的 Transformer，相对于深度学习，其创新在于采用了接近无监督的自我监督预训练，因此需要大量训练数据，加上少量有监督的微调和强化学习相结合。随着更复杂和多元模型不断涌现，高算力的 AI 芯片将充分受惠。然而，若以上技术只停留在 C 端应用意义却并不大，因此我们更认为，本轮 AI 热潮能否持续将取决于 B 端的大规模应用落地。AI 浪潮方兴未艾，我们首次覆盖给予 AI 芯片行业增持评级，重点推荐龙头英伟达及突围者 AMD。

首选软硬件一体双护城河的英伟达，同时看好突围二战有望凯旋的 AMD

作为全球 AI 芯片的龙头厂商，英伟达在高算力硬件和高粘性 CUDA 生态的双护城河下优势明显，将充分受益于 AI 需求高涨。我们认为，短期内英伟达将主要由数据中心业务带动，长期成长则取决于 AI 商业应用落地及芯片竞争格局的演变。随着 CoWoS 产能瓶颈的改善，我们认为英伟达 GPU 放量节奏将迎来加速。AMD 曾凭台积电的领先制程颠覆了一家独大的英特尔，如今在 AI 领域面对英伟达的突围战似曾相识。AI 已是 AMD 战略首位，MI300 也蓄势待发，我们认为 AI 新赛道乃 AMD 重估之钥。对比英伟达完善的软件生态 CUDA，AMD 的 ROCm 会否成为其阿克琉斯之踵？

AI 芯片竞争趋白热化：训练端“一超多强”，推理端百花齐放

英伟达 GPU 一直为 AI 训练端首选。我们认为只有少数芯片能与其匹敌，如谷歌 TPU 和 AMD MI300 系列。当算法开始稳定和成熟，ASIC 定制芯片凭着专用性和低功耗，能承接部分算力。因此，头部云计算及互联网大厂出于削减 TCO、提升研发可控性及集成生态等考量，均陆续发力自研芯片，我们认为或将成为英伟达最大的竞争对手。初创企业如 Cerebras、Graphcore 等，以晶圆级芯片拼内存和传输速度，也有望异军突起。AI 推理市场规模大，但对算力要求比训练较低，因此百花齐放，在大模型和多模态趋势下 GPU 或能夺份额。但目前推理端还是以 CPU 主导，多方涌入下竞争愈发激烈。

台积电 CoWoS 封装产能乃 AI 芯片厂商“必争之地”

英伟达 H100 采用台积电 CoWoS 先进封装技术，而 AMD MI300 采用台积电 CoWoS 和 SoIC 技术，二者都需依赖台积电先进封装产能。目前，AI 芯片需求旺盛，台积电 CoWoS 封装乃限制出货量的瓶颈之一。但据 Digitimes 在 7 月 14/21 日报道，公司正积极扩产，到本年底至少达 12 万片，24 年将达 24 万片，而英伟达将取得约 15 万片；当前三大客户为英伟达、博通和赛灵思，而 MI300 在四季度推出后，AMD 或将一举跻身前五大客户。英伟达的订单或也将外溢到联电和 Amkor。另外，CoWoS 的瓶颈也许是来自日本的 Tazmo、Shibaura 等的封装设备厂商，交货周期往往需要 6-8 个月。

配置建议：英伟达“买入”，TP 650 美元；AMD“买入”，TP 150 美元

英伟达作为全球数据中心 GPU 龙头，该业务已占总营收逾 75%，为主要盈利和营收贡献，将充分受惠于台积电先进封装产能明年翻倍，以及美国加息步入尾声。游戏显卡逐渐从挖矿和疫情间 PC 高基数影响中恢复，叠加高期待新游戏带动。首次覆盖英伟达（NVDA US）给予“买入”，FY24-25 年动态 20 倍 PS，目标价 650 美元。AMD 数据中心业务 CPU 制程仍领先英特尔，MI300 系列有力冲击英伟达，看好 CPU 和 GPU 均能抢夺份额；游戏业务将受益主机“半代升级”；客户端业务 PC 市场下滑收窄渐入佳境，首次覆盖 AMD（AMD US）给予“买入”，目标价 150 美元，对应 24 PS 8.5x。

风险提示：AI 技术落地和推进不及预期、行业竞争激烈、中美竞争加剧。

正文目录

人工智能风继续吹，AI 芯片乘风而起.....4
人工智能风再起，产业发展空间广阔.....4
看好整体 AI 芯片需求将伴随着大模型和生成式 AI 的 B 端应用落地而增加.....4
人脑神经网络的运作模式始终是人工智能追求的终极形态.....7
我们处于 AI 的“iPhone”时刻吗？.....8
生成式 AI 将带动云大厂加码硬件基础设施.....9
“一超多强”的 AI 芯片市场竞争格局.....11
传统芯片巨头：英伟达在多方入局下能否继续笑傲江湖？.....12
英伟达在推理端能否复制其在训练端的成功？.....12
英伟达的两大护城河：高算力芯片和高粘性 CUDA 软件生态.....12
先进的网络技术 NVLink+NVSwitch，为吞吐量和可扩展性带来突破.....13
CUDA 软件生态的先发优势构筑英伟达第二护城河.....16
生成式 AI 下较复杂的推理需求或为英伟达 GPU 赋新机.....17
英伟达 SWOT 分析.....19
传统芯片巨头：备受期待的突围者 AMD.....20
以 AI 为战略首位，AMD 突围二战能否凯旋？.....20
MI300A 和 GH200：CPU+GPU AI 芯片架构仿生人脑结构.....20
ROCm 生态圈会否成为 AMD 的“阿克琉斯之踵”？分而治之或可解困.....24
传统芯片巨头：多元布局的追赶者英特尔.....27
云计算和互联网大厂：或许是传统芯片厂商的最大竞争对手.....32
谷歌 TPU：少数能与英伟达高算力 GPU 匹敌的 AI 芯片.....32
亚马逊 AWS：Trainium & Inferentia，训练推理双管齐下.....38
微软：“闭门造芯”Athena.....40
Meta：首个自研推理端芯片 MTIA 将于 2025 年问世.....41
异军突起者：晶圆级芯片持续突破性能极限，内存和传输成破局关键.....43
Cerebras：向晶圆级大尺寸芯片迈出第一步，但良率和有效运行占比暂成疑.....43
Graphcore：Bow IPU 实现精细数据多指令并行.....48
特斯拉：Dojo 超算为自动驾驶而生，为公司四大全栈自研科技支柱之一.....52
晶圆级芯片跟传统芯片的各项对比.....55
AI 芯片产业链：聚焦兵家必争之地 CoWoS 封装.....58
台积电大扩 CoWoS 产能，供给紧张有望得解.....58
硅晶圆供应商：台积电的 6 家硅晶圆供应商占全球总产能 90%以上.....59
衬底/基板（Substrate）：揖斐电、景硕、欣兴电子等.....60
HBM 内存：SK 海力士、三星、美光，三足鼎立.....61
服务器相关供应商：惠与、戴尔、联想、美超微、广达、纬创等.....61
AI 不只是大模型，自动驾驶芯片群雄逐鹿，谁能突围？.....63
Mobileye：ADAS 技术奠基者，“黑箱子模式”优势不再，转型将面临挑战.....65

地平线：基于 BPU 架构布局自动驾驶生态追击.....67
黑芝麻：第一家递交港股 18C 上市文件的车载芯片股，华山对标英伟达 Orin，武当实现跨域融合.....70
高通：可扩展体系开展差异化竞争，对标英伟达 Thor 打造跨域融合.....71
华为：边缘端 AI 芯片赋能 MDC 计算平台.....72
特斯拉：车企破局者，FSD 和 DOJO 软硬件全栈自研.....73
重点推荐：英伟达为 AI 芯片行业龙头，AMD 突围有望迎来重估.....75
英伟达：AI 龙头软硬一体双护城河（NVDA US，买入，目标价：650.00 美元）.....75
超威半导体：AI 新赛道为重估之钥（AMD US，买入，目标价：150.00 美元）.....76
风险提示.....78
首次推荐公司.....79
英伟达（NVDA US，买入，目标价：650.00 美元）.....79
超威半导体（AMD US，买入，目标价：150.00 美元）.....125

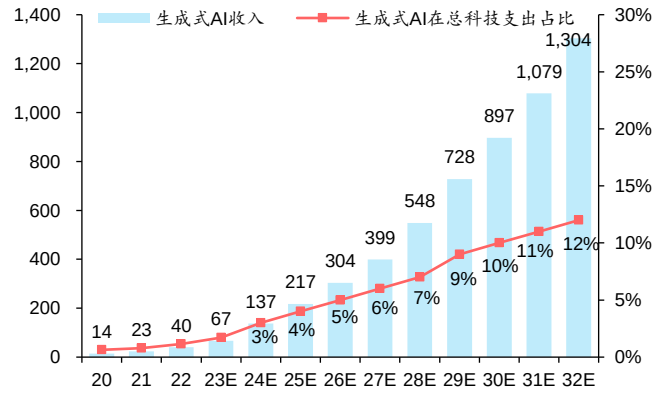
人工智能风继续吹，AI 芯片乘风而起

人工智能风再起，产业发展空间广阔

本轮人工智能浪潮由 ChatGPT 掀起，并以语言大模型（Large Language Model, LLM）和生成式 AI（Generative AI）应用作为切入点。自谷歌在 2017 年发表至今，Transformer 除了带来像 ChatGPT 这样的 C 端爆款产品外，其早已在自然语言处理、计算机视觉、自动驾驶等领域里广泛应用。各中外科技企业持续加大对相关的投入，包括谷歌（GOOGL US）、Meta（META US）、微软（MSFT US）、字节跳动（未上市）、百度（BIDU US）等海内外一众科技巨头和初创企业均希望分一杯羹，其他非技术公司也不断在人才、技术和资源方面进行布局。根据 Bloomberg Intelligence 的预测数据，到 2032 年，生成式 AI 在总体信息技术硬件、软件、服务、广告和游戏等支出中的占比或将从目前不到 1%的水平扩大至 12%。

ChatGPT（Chat Generative Pre-trained Transformer）自 22 年 11 月发布后就引起了全球关注，5 天内注册用户超 100 万，仅两个月后，月活跃用户已达 1 亿。ChatGPT 将生成式 AI 在文本、图像、视频等领域的多模态应用正式带入 C 端大众用户的视野。然而，我们认为，语言大模型若只是停留在 C 端应用，提供给一些网民娱乐，其实意义并不大。我们更加认为，生成式 AI 的发展必须要配合 B 端应用的落地，才能成为真正可以改变世界的高端科技。目前，微软已发布 Microsoft 365 Copilot 等生成式 AI 产品，作为率先发布的重磅商业化应用。Copilot 依托微软庞大的用户群体、产品生态及使用场景，有望开启 AI 的 B 端应用发展新里程，并带动微软打开新的 AI 商业化空间。Bloomberg Intelligence 预测，全球生成式 AI 下游软件市场规模将在 2032 年扩大至 2799 亿美元，2022-2023 年十年复合增速达到 69%。

图表1：生成式 AI 占科技投入支出不断增加（单位：十亿美元）



资料来源：Bloomberg Intelligence、IDC、华泰研究

图表2：全球生成式 AI 的市场机遇（单位：百万美元）

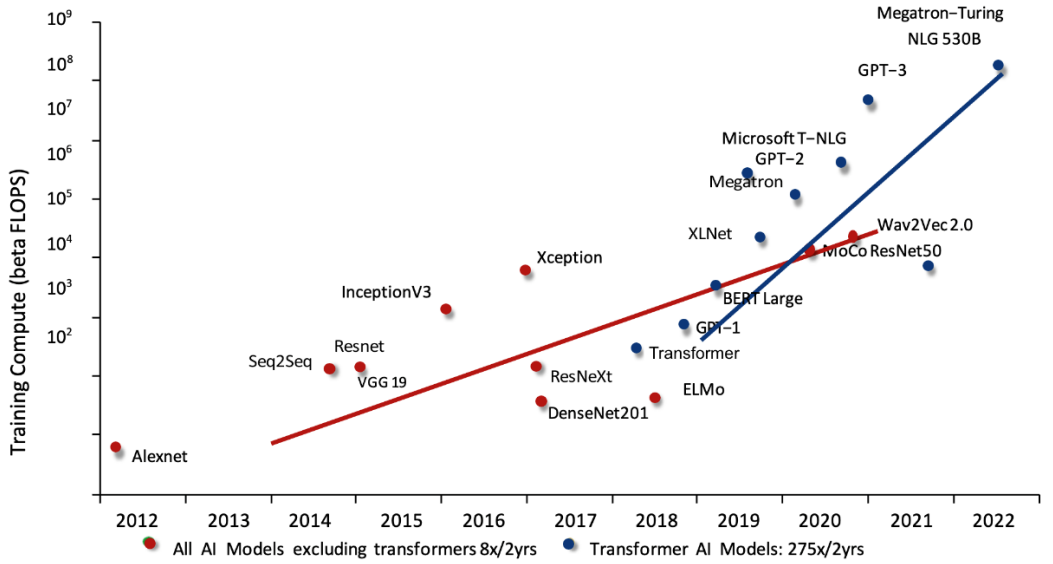
生成式 AI 项目	2022	2032E	CAGR
专业智能助手	$447	$89,035	70%
代码编写、DevOps 等	$213	$50,430	73%
工作负载基础设施软件	$439	$71,645	66%
药物发现软件	$14	$28,343	113%
网络安全支出	$9	$13,946	109%
教育支出	$370	$26,500	53%
软件	**$1,493**	**$279,899**	**69%**

资料来源：Bloomberg Intelligence、IDC、华泰研究

看好整体 AI 芯片需求将伴随着大模型和生成式 AI 的 B 端应用落地而增加

2022 年开始，大模型的数量及参数量均呈指数级增长。总体来说，我们认为模型的数量和所需的训练数据才是对于算力要求的关键，因此，我们看好整体 AI 芯片的需求，将伴随着大模型和生成式 AI 所支持的 B 端商业应用落地而增加。自 2018 年 OpenAI（未上市）发布了包含 1.17 亿参数的第一代 GPT（Generative Pre-trained Transformer）模型以来，每一代 GPT 模型的迭代都伴随着参数量的飞跃。一众中外的科技巨头们也不甘示弱，包括 Google、Meta、百度等纷纷发布了 PaLM 、LaMDA 、Llama、文心一言等为代表的大语言模型。2020 年 1 月，OpenAI 团队论文《Scaling Laws for Neural Language Models》提出“缩放定律”（Scaling Laws），即大模型表现伴随模型参数量、数据集大小和计算量增长而增长，他们于 2023 年 5 月也再次强调，目前缩放定律仍未出现瓶颈。但我们也看到，谷歌在今年 5 月的 I/O 大会里发布的新一代 PaLM 大模型，PaLM2，就是通过算法上的改进达到训练数据增加为上一代 PaLM（7800 亿 tokens）的约 5 倍，达到 3.6 万亿个 tokens，但参数量为 3400 亿，小于 PaLM 的 5400 亿。

图表3：AI 训练对算力的需求成倍上涨，尤其是 Transformer 相关模型

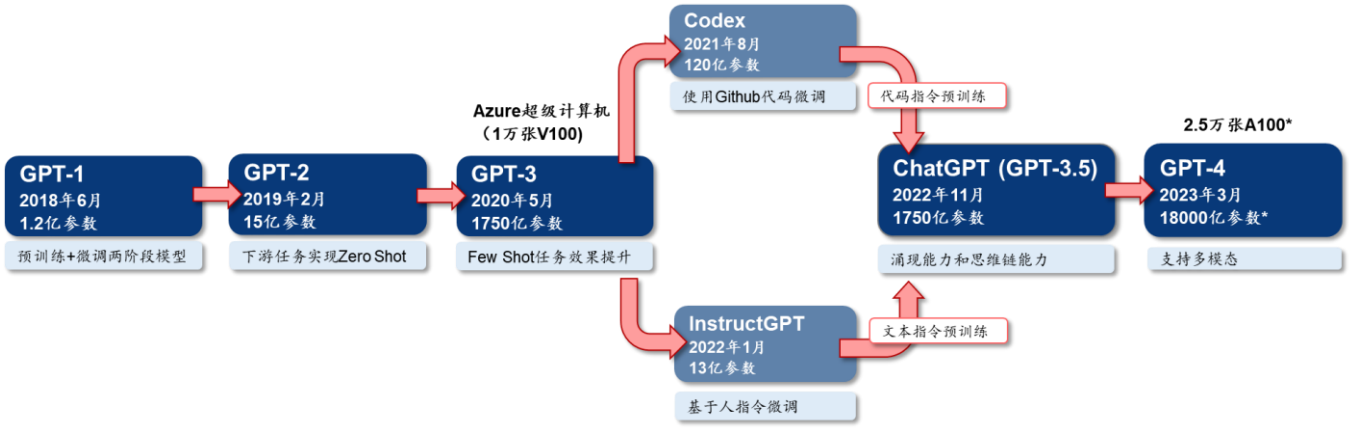


注：不同颜色代表不同模型种类

资料来源：英伟达官网、华泰研究

“大模型”通常指的是有大量参数的自我监督和预训练模型，其背后的核心技术是 Transformer 架构，目前比较广泛应用于文本生成等自然语言处理领域。Transformer 在 2017 年由谷歌大脑团队在论文《Attention Is All You Need》中提出。该架构主要用于处理序列数据，主要采用自注意力机制（self-attention mechanism），为序列中的每个元素赋予不同的权重，从而捕获序列内部的长距离依赖关系。在 Transformer 之前，深度学习模型更多是采用监督学习的方法进行训练，因此需要大量标注的数据。相对来说，GPT 模型的创新之处在于采用了接近无监督学习（具体叫“自我监督学习”，self-supervised learning）的预训练，加上少量有监督的微调相结合。

图表4：GPT 各代模型迭代情况



注：GPT-4 参数及训练基础设施数据来自 semianalysis

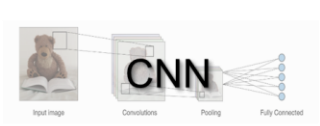
资料来源：OpenAI 官网、福布斯官网、InfoQ 官网、semianalysis 官网、微软官网、华泰研究

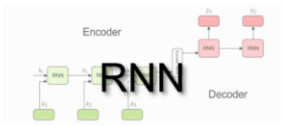
图表5：大模型训练过程示意图

资料来源：CSDN 官网、华泰研究

在文本生成、上下文语义理解、文章修订及摘要总结等需要泛化能力的大语言模型中，Transformer 架构相对以往的 CNN 和 RNN 网络结构取得了较大的进展。Transformer 架构突破了 RNN（Recurrent Neural Network）模型固定顺序属性所带来的计算限制，其通过自注意力机制，可以同时处理整个序列的所有元素，从而实现了高效的并行化，提高了计算速度。同时，相比 CNN（Convolutional Neural Network）模型中伴随距离增大，计算位置关联所需操作将不断增多，Transformer 通过自注意力机制，可直接计算序列中任何两个元素之间的关联，且通过权重显示序列元素间的关系，从而为模型提供更为丰富的全局上下文信息，有效提高对复杂结构及语义的理解。故 Transformer 被认为与大部分白领工作相契合，在当前人力成本较高及生产力急需提升的背景下，或将开始下沉至办公、会计、法律、编程和医疗等各领域进行结合。我们可将 Transformer 模型比作人类的右脑，在浅层关联性上表现优异，适用于需要创造性的生成式领域，但其仍然需要加强左脑的逻辑判断能力。

图表6：Transformer 架构与 CNN 和 RNN 对比情况

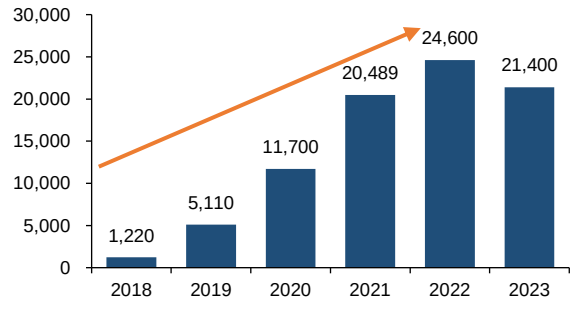




	CNN架构	RNN架构	Transformer架构
	• 局部连接、权值共享的卷积结构 • 通常包括：输入层（input layer）、卷积层（convolutional layer）、池化层（pooling layer）和输出层（全连接层+softmax layer）	• 每个时间片t的计算依赖t-1时刻的计算结果： • Encoder:$h_t = \mathrm{RNN}(x_{t-1}, h_{t-1})$ • Decoder:$y_t = \mathrm{RNN}(y_{t-1}, h'_{t-1})$ • 具有循环连接结构，如LSTM、GRU等	• 采用attention机制建模序列中位置关系 • 将序列中的任意两个位置之间的距离缩小为一个常量 • 使用位置编码代表位置信息
特征表示能力	对局部特征提取能力强，适用于图像识别、视觉任务中	适用于处理序列数据，能够捕捉序列信息的演化过程，适用于语音识别、自然语言处理任务	具有较强的建模能力，适用于处理声音序列、图像等多种数据形式
训练效率	每个卷积操作都是独立的，可并行化处理	每个时间步的计算都依赖于前一个时间步的输出，故与其他二者相比难以并行化	使用attention机制绘制输入和输出之间的全局依赖关系，可并行化处理
模型复杂度	模型结构相对简单，参数较少，不适用于处理序列数据	模型结构相对复杂，参数较多	模型结构较为复杂，参数较多

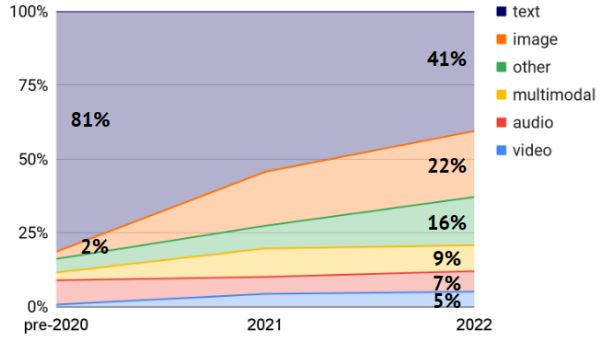
资料来源：Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).、OpenAI、KDnuggets、斯担福大学官网、华泰研究

图表7：2018 年开始，Transformer 在谷歌学术的引用数量（次）



注：2023 年的数据截止至 7 月
资料来源：谷歌学术、华泰研究

图表8：Transformer 相关的论文领域占比



资料来源：State of AI report、华泰研究

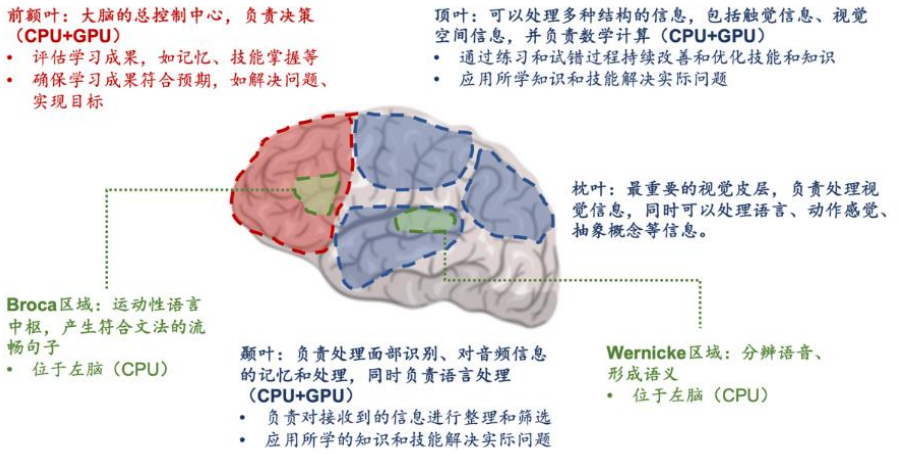
人脑神经网络的运作模式始终是人工智能追求的终极形态

类比人类大脑，左脑主要负责对信息逻辑的处理，如串行运算、数字和算术、分析思维、理解、分类、整理等，而右脑负责并行计算、多模态、创造性思维和想象等。因此，左右脑从功能上分别代表 CPU 和 GPU，对比人类可实现左右脑协同工作，整体调动神经网络，将为 AI 的最终愿景。

早在 2011 年，AMD 产品构想中就以 CPU 和 GPU 分别类比人类左右脑，并基于此提出了 CPU+GPU 的异构产品策略。（详见 AMD 部分）

目前 AMD 的 MI300A 和英伟达的 Grace Hopper（GH200）均为 CPU+GPU 的异构集成。GPU 的算力高并针对并行计算，但须由 CPU 进行控制调用，发布指令。在 AI 训练端，CPU 可负责控制及发出指令，指示 GPU 处理数据和完成复杂的浮点运算（如矩阵运算）。在面对不同模态数据的推理时，我们认为，CPU 与 GPU 的分工也各有不同，因此，同时部署 CPU 和 GPU 能提供更大的运算支撑。例如，在处理语音、语言和文本数据的推理时，AI 模型需逐个识别目标文字，计算有序，因此或更适合使用擅长串行运算的 CPU 进行运算支持；但在处理图像、视频等数据的推理时（对比人类的操作，每一个像素是同时进入眼睛），需要大规模并行运算，或更适宜由 GPU 负责，例如英伟达 L4 GPU 可将 AI 视频性能提高 120 倍，据英伟达测试，L4 与基于 CPU 的传统基础设施相比能源效率提高 99%。

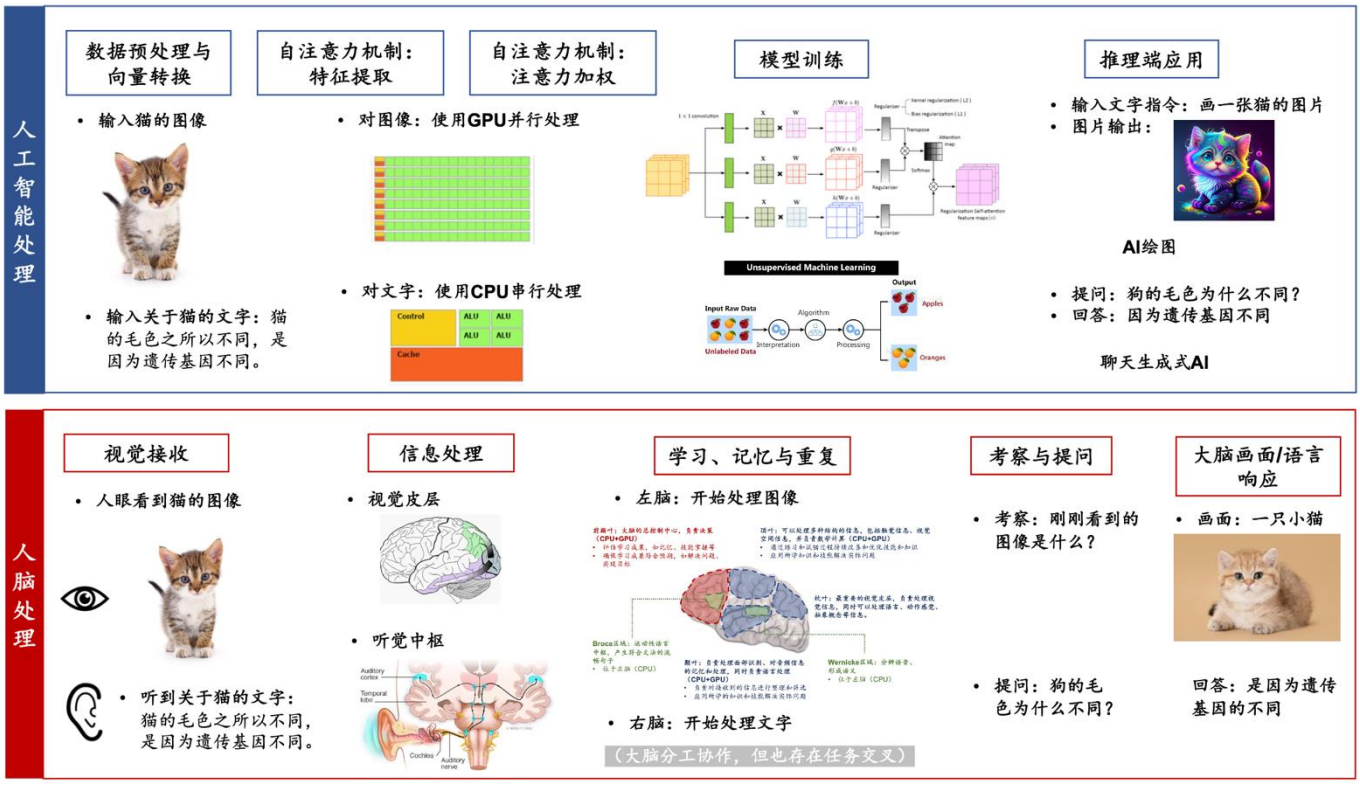
图表9：人类大脑不同部分功能及对应芯片种类



资料来源：Hari R. From brain–environment connections to temporal dynamics and social interaction: principles of human brain function[J]. Neuron, 2017, 94(5): 1033-1039.、BrainFacts、华泰研究

AI 推理市场规模大，但对算力要求比训练较低，因此我们认为各类芯片的使用将百花齐放，在大模型和多模态趋势下 GPU 或能夺份额。但目前推理端还是以 CPU 主导，多方涌入下竞争愈发激烈。值得一提的是，数据中心里拥有各类不同的芯片，而不同的 AI 工作负载应该在哪一片芯片上运行，将取决于以上提到的适配度以及性价比。因此，各类芯片也有其不同的优势。

图表10：人脑处理信息与人工智能训练和推理的流程对比



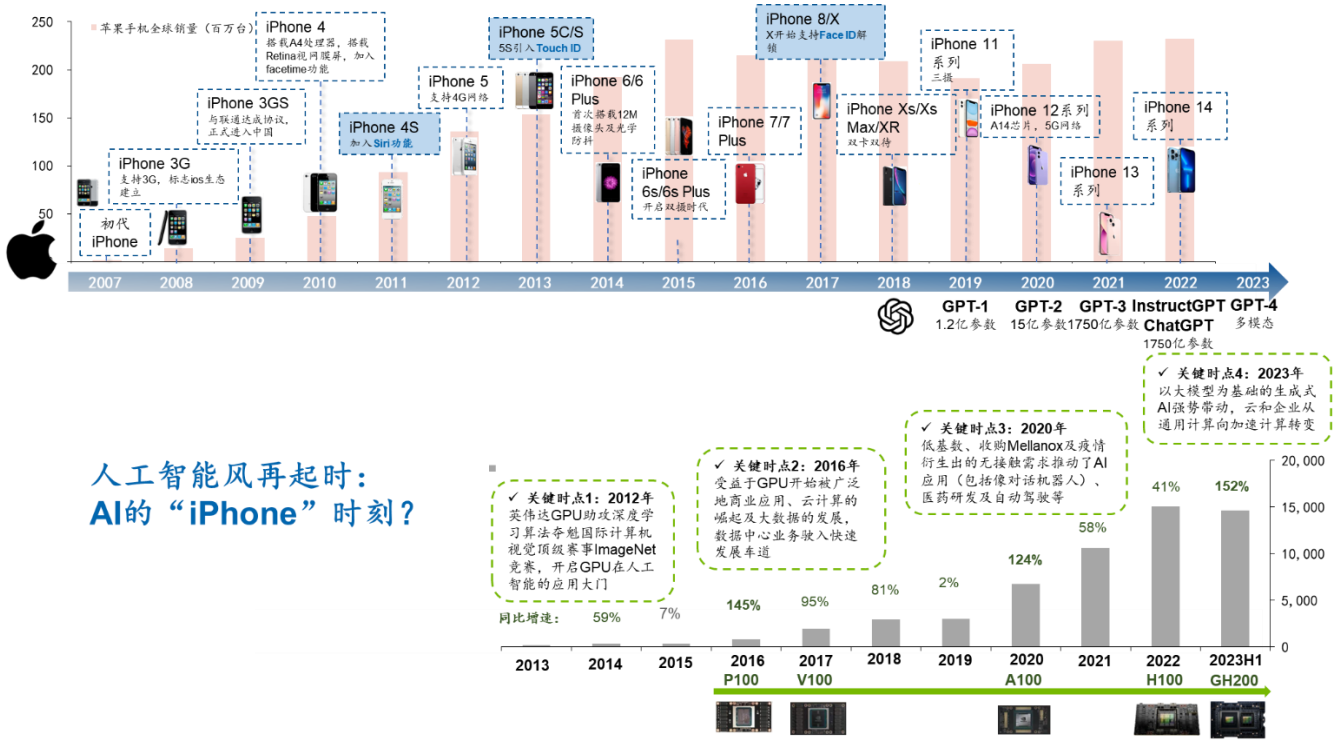
资料来源：CSDN、谷歌官网、Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.、华泰研究

我们处于 AI 的“iPhone”时刻吗？

人工智能的概念可以追溯到上世纪五六十年代。很多我们现在熟悉的算法，如神经网络，在 20-30 年前已经存在，但由于缺乏算力和数据，因此无法有效地运行。随着 GPU 被应用到 AI、云计算的普及和海量数据的产生和存储，AI 技术才得以快速发展和应用。

对于“现在是 AI 的 iPhone 时刻”的观点，我们更倾向于认为，这是描述跟 GPT 相关的生成式 AI 开始应用于 To B 端及解放生产力的重要突破。至于在 To C 端，AI 技术其实早已有众多应用融入了我们的生活，如智能手机中的语音助手 Siri 和人脸识别等功能。

图表11：AI 的“iPhone”时刻

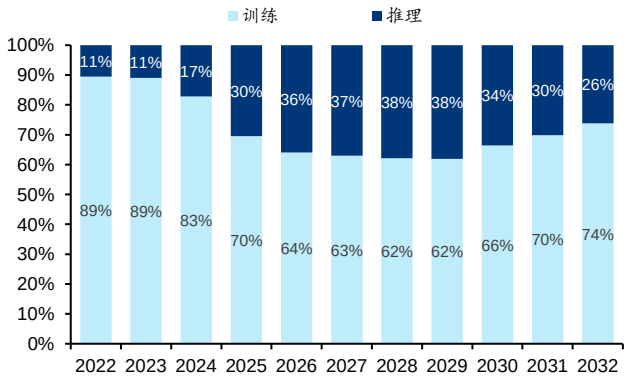


资料来源：太平洋电脑网、苹果官网、英伟达官网、苹果派、OpenAI 官网、各公司公告、华泰研究

生成式 AI 将带动云大厂加码硬件基础设施

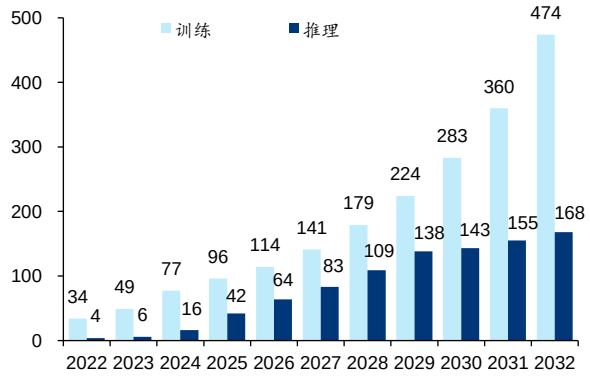
我们认为，硬件设备的规模和性能是 AI 大模型时代的必然要求。鉴于目前生成式 AI 主要以大参数模型路径实行，随着模型数量和所需要处理的数据量增长，其训练与推理均需要大量的计算能力与存储资源，**故生成式 AI 应用的蓬勃发展将带动高算力 AI 芯片以及云计算的需求增长。**Bloomberg Intelligence 及 IDC 数据显示，到 2024 年，AI 训练和推理硬件市场规模将达 930 亿美元，而到 2032 年将逾 6000 亿美元。

图表12：AI 训练和推理硬件市场占比



资料来源：Bloomberg Intelligence、IDC、华泰研究

图表13：AI 训练和推理硬件市场规模（单位：十亿美元）



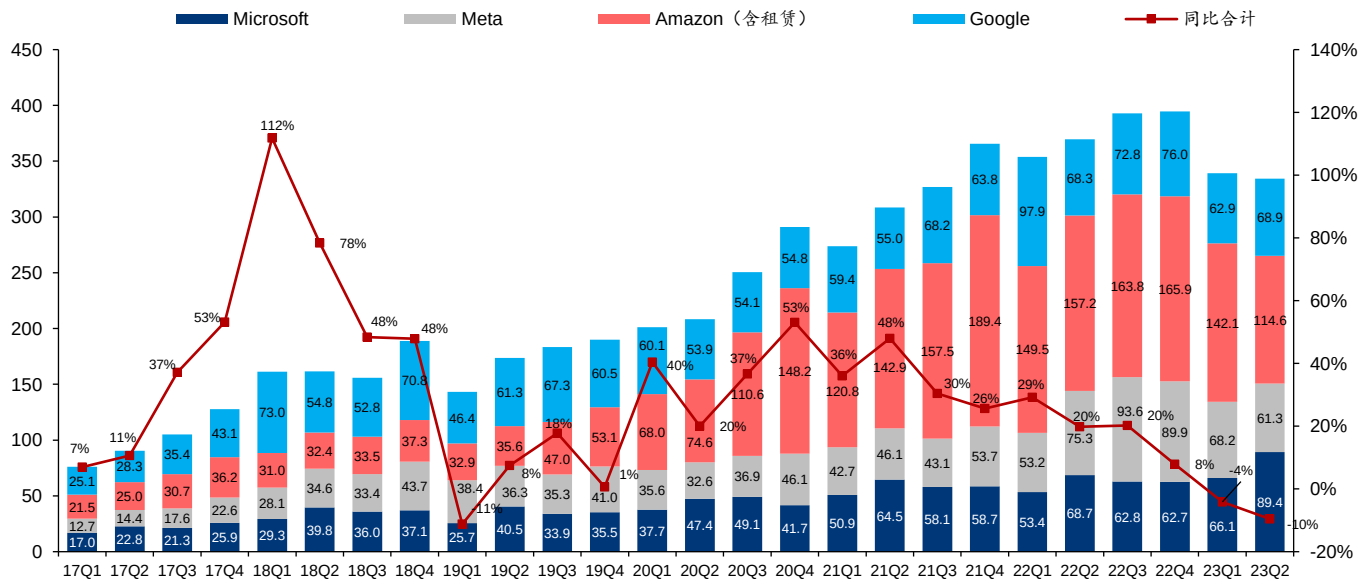
资料来源：Bloomberg Intelligence、IDC、华泰研究

云大厂和互联网巨头预计将继续加大资本开支，AI 硬件为重点领域。谷歌、微软、亚马逊以及 Meta 在二季度业绩说明会中提到：

- **微软 FY23Q4：**资本开支（不含融资租赁）89.43 亿美元，同比增加 30.16%，计划持续加大数据中心、CPU 和 GPU 等投入；
- **谷歌 232Q：**资本开支环比增长 10%至 69 亿美元，主要投放在服务器以及 AI 大模型计算领域，增幅低于彭博一致预期主因数据中心建设项目延迟，但公司预计对技术基础设施的投资将在 2023 下半年增加；
- **亚马逊 23Q2：**资本开支（含融资租赁）为 114.55 亿美元，同比下跌 27%，虽受逐渐下滑的运输投入影响，公司预计 2023 全年资本开支同比下滑至略高于 500 亿美元的水平，但仍将继续增加对 AI 和大语言模型的投入以满足客户需求；
- **Meta 23Q2：**资本开支（不含融资租赁）为 61.34 亿美元，同比下跌 19%，主要鉴于非 AI 服务器费用的减少，以及部分项目和设备交付的延误将转入 2024 年，公司预计 2024 年资本开支将伴随数据中心、服务器及人工智能方面投资的推进而增加。

总体来看，2023 上半年，以谷歌、微软及亚马逊及 Meta 为代表的互联网巨头在资本开支方面，虽受到项目延期或宏观与其他业务规划等因素扰动，但在 AI 相关的资本开支方面正逐渐加大。展望 2024 年，AI 基础设施将是重点投入领域。**故我们认为头部云厂商和互联网巨头加大 AI 领域资本开支将进一步支撑 AI 的产业趋势。**我们认为，2022 年以来，美联储稳步提高利率导致企业削减数据中心支出，后续美联储或将停止加息，叠加 AI 需求增长，有望提振科技巨头的资本开支，将持续带动 AI 芯片等基础硬件设施放量。

图表14：17Q1-23Q2 四大互联网巨头季度资本开支情况（单位：亿美元）



资料来源：各公司官网、华泰研究

“一超多强”的 AI 芯片市场竞争格局

在人工智能的训练端（training），我们认为英伟达的 GPU 凭着高算力为门槛，一直以来都是训练端的首选。但随着人工智能应用市场的不断扩大，AMD MI300 系列的 GPU、云计算厂商自研专用芯片、以及大尺寸晶圆级芯片也开始异军突起。推理端对算力要求较训练端低，目前推理芯片市场百花齐放，各类芯片均据有一席之地。我们认为，推理端和训练端一样，传统芯片厂商也将面临云计算厂商和 AI 芯片初创企业的挑战。

AMD 在 23Q1 财报会议上表示，AMD 已将人工智能列为战略重点的首位，同时推出新产品 MI300 系列，在制程、架构及算力等多方面向英伟达同类产品看齐。另外，作为英伟达客户的云计算大厂也陆续发力自研专用芯片。谷歌的 TPU（Tensor Processing Unit）专为神经网络和 TensorFlow 框架量身打造，性能不断提升，目前已发展到第五代 TPU v5e，而于 2020 年推出的 TPU v4，在多种机器学习框架模型上已与英伟达的 A100 可比。亚马逊 AWS 自己造芯早已有迹可循，2018 年开发了基于 ARM 的服务器 CPU Graviton，也为 AI 推理端自研专用芯片 Inferentia (2018 年末推出第一代，目前已发展到第二代)和 AI 训练端定制芯片 Trainium（2020 年末推出）。对比基于 GPU 的实例，Trainium 速度能提升 140%，成本能降低 70%。据 2023 年 4 月 18 日美国科技媒体 The Information 报道，微软也正在闭门造芯，研发支撑 AIGC 训练和运行的专用芯片 Athena（雅典娜），希望进一步降低开发 AI 的成本。

除了传统芯片龙头和云计算大厂之外，我们也发现一些新兴初创 AI 芯片企业（如 Cerebras、Graphcore 等），以及芯片行业以外的企业，包括特斯拉等，正在异军突起，试图在芯片设计上另辟蹊径，通过大尺寸晶圆级芯片的技术路线，在持续上升的算力、传输和内存需求市场中抢占份额。短期来看，我们认为，使用先进封装技术的 GPU 相较晶圆级芯片或是更优选择，但长期来看，晶圆级芯片的瓶颈若能突破，也将成为传统技术路径的有力挑战者。

图表15：主流 AI 芯片对比

训练端	通用性强，但较难适应于人工智能时代大数据并行计算工作。	通用性强，多维计算及大规模并行计算架构，适合深度学习需要；在训练端是第一选择	-	针对特定框架进行深度优化定制，能耗较低，但开发周期较长，固定成本也较高
推理端	需要大量空间去放置存储单元（Cache）和控制单元（Control），用于逻辑控制。	英伟达从 18 年开始通过 T4 芯片等布局推理端到边缘计算；对算力要求较训练端要低	多以加速器形式跟 CPU 一起搭载；依靠可编程性，适用于开发周期较短的产品，以及开发试错阶段等；较成熟的量产设备多采用 ASIC	若特定领域产生大规模需求，在大批量生产下固定成本可有效给摊分；能耗也较低
代表厂商	Intel/AMD	NVIDIA/AMD	Altera （Intel） / Xilinx （AMD）	Google 的 TPU AWS Tranium / Inferentia

资料来源：nextplatform 官网、HUAWEI、华泰研究

传统芯片巨头：英伟达在多方入局下能否继续笑傲江湖？

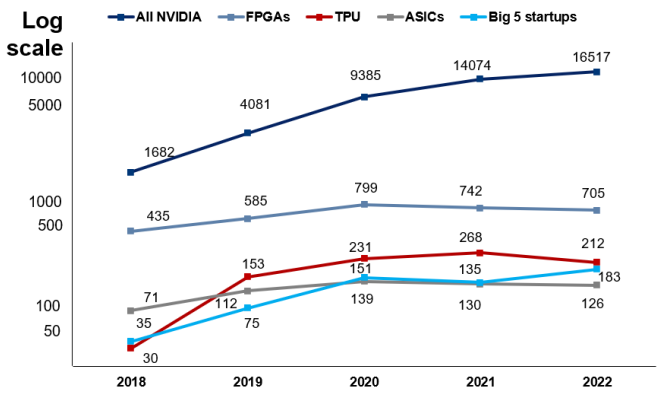
英伟达在推理端能否复制其在训练端的成功？

英伟达的 GPU 虽在 AI 训练端属不二之选，对比市场同类训练产品也具有显著优势。然而，推理端的门槛相对较低，且应用场景和需求更为多元，目前各类芯片都在此领域获得一席之地，因此 AI 推理方面英伟达仍面临着激烈竞争。从发展历程来看，传统推理端主要依赖 CPU 去处理相对简单和对算力要求较低的推理任务。不过，目前 AI 模型的规模和复杂度跟过去相比已提升了不少，随着更多工作负载将逐渐纳入到推理领域，因此对于推理算力的要求也会越来越高，或将在一定程度上带动更多 GPU 在推理领域的应用。但我们需强调，推理所需要的算力本身比训练所需要的算力低，因此英伟达的高算力在推理端不一定像其在训练端般拥有明显优势。另外，数据中心里拥有各类不同的芯片，而不同的 AI 工作负载应该在哪一片芯片上运行，将取决于适配度以及性价比。因此，各类芯片也有其不同的优势。在这领域里英伟达的高性能推理芯片正面对着各种竞争对手，除了 CPU 之外，也包括 AMD 的 GPU、英伟达的 GPU（包括老款）、FPGA 和 ASIC 等。

英伟达的两大护城河：高算力芯片和高粘性 CUDA 软件生态

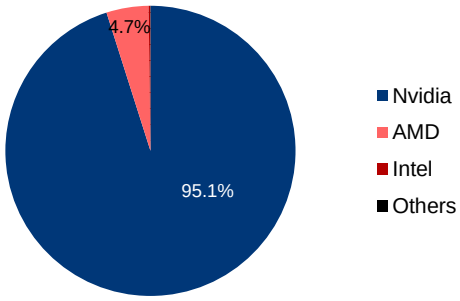
我们认为，英伟达凭着高算力硬件和高粘性软件生态两大护城河，是人工智能训练端的不二之选。根据不同芯片在 AI 论文中的引用数量可知，英伟达的芯片在 AI 研究论文中最受欢迎，其产品的使用率是 ASIC 的 131 倍，是 Graphcore（未上市）、Habana（给 Intel 收购了）、Cerebras（未上市）、SambaNova（未上市）和寒武纪（688256 CH）五家总和的 90 倍，是谷歌 TPU 的 78 倍，是 FPGA 的 23 倍。一般来说，在人工智能领域，新模型的推出都会发表相关论文便于信息交流与学术合作，英伟达在人工智能相关的论文中遥遥领先的引用数量，也反映了新算法需采用英伟达 GPU 的必要性，以及其在学术界长期以来的重要地位和影响力。

图表16：英伟达芯片在 AI 论文中的引用数量遥遥领先



注：Big 5 startups 是指 Graphcore, Habana, Cerebras, SambaNova 和寒武纪五家初创芯片企业，其中寒武纪为中国企业；Habana 在 2019 年被 Intel 收购。
资料来源：Zeta Alpha analysis、华泰研究

图表17：2022 年服务器 GPU 单元份额情况



资料来源：IDC、华泰研究

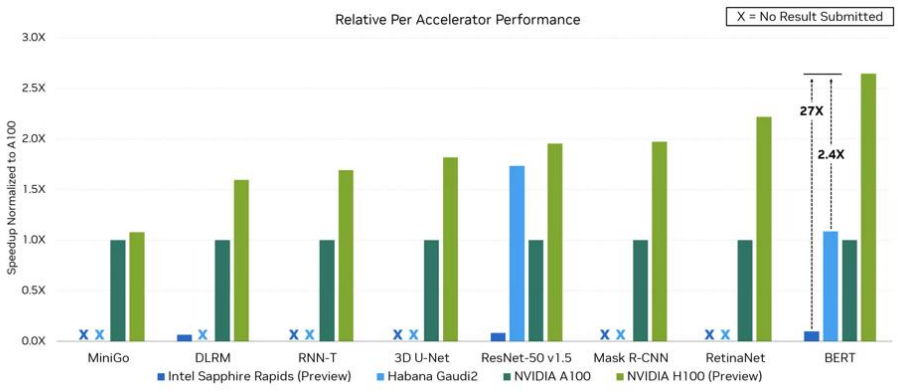
英伟达持续迭代高性能计算芯片，在产品工艺、计算能力和存储带宽等不断创新。面向高性能计算和深度学习场景，英伟达基于其芯片架构，打造了一系列支持提升张量核心和稀疏矩阵计算等能力的 GPU 产品。2023 年，英伟达已不满足于单 GPU 的更新换代，重磅推出结合 Grace CPU 与 Hopper GPU 的 GH200 超级芯片，实现了高达 900GB/s 的总带宽，加速大规模 AI 和 HPC 应用计算。在一年后的 SIGGRAPH 上，英伟达的 AI 芯片再迎升级，推出了全球首次采用 HBM3e 内存的 GH200 超级芯片。该芯片的带宽高达每秒 5TB 并能提供 141GB 的内存容量，适用于复杂的生成式人工智能工作负载，如大型语言模型、推荐系统和矢量数据库等。

图表18：主要人工智能芯片的参数对比

	英伟达				
产品名称	A100 PCIe丨SXM	H100 PCIe丨SXM丨NVL	L40S	GH200 （HBM3）	GH200 （HBM3e）
发布时间	2020.6	2022.3	2023.8	2023.5	2023.8
峰值算力（TFLOPS）	FP16: 312丨624 FP32: 19.5 FP64: 19.5	FP8: 3026丨3958丨7916 FP16: 1513丨1979丨3958 FP32: 51丨67丨134 FP64: 51丨64丨134	FP8:1466 FP16: 733 FP32: 91.6	-	-
工艺制程	TSMC 7nm	TSMC 4nm	TSMC	TSMC 4nm	TSMC 4nm
芯片面积	826mm2	814 mm2	-	-	-
晶体管数 （B）	54	80	76.3	200	-
内存容量 （GB）	80 GB （HBM2e）	80丨80丨188 （HBM3）	48GB（GDDR6）	96 （HBM3）	141 （HBM3e）
内存带宽	1935丨2039 GB/s	2丨3.35 TB/s丨7.8TB/s	864GB/s	<=4TB/s	5TB/s
Interconnect	NVLink 600 GB/s	NVLink 600丨900丨600 GB/s	16 Links PCIe Gen4 64GB/s	NVLink 900GB/s	NVLink 900GB/s
TDP（W）	300丨400	300-350丨700丨2x350-400	350	450 - 1000	450 - 1000
	AMD			谷歌	英特尔
产品名称	MI250X	MI300A	MI300X	TPUv4	Habana Gaudi 2
发布时间	2021.11	2023.1	2023.6	2021.5	2022.5
峰值算力（TFLOPS）	FP16: 383 FP32/64: 47.9 FP32/64 Matrix: 95.7	-	-	Bf16: 275	-
工艺制程	TSMC 6nm	TSMC 5nm	TSMC 5nm	TSMC 7nm	TSMC 7nm
芯片面积	724 mm2	1017 mm2	1017 mm2	780mm2	-
晶体管数 （B）	58	154	146	31*	-
内存容量 （GB）	128 （HBM2e）	-	128 （HBM3）	32 （HBM2）	96 （HBM2E）
内存带宽	3.2 TB/s	-	5.2TB/s	1200GB/s	2.45TB/s
Interconnect	Infinity Fabric >=500GB/s	Infinity Fabric 800GB/s	Infinity Fabric 896GB/s	3D torus	RDMA （RoCE v2） 100GB/s
TDP（W）	500	600	-	192	600

资料来源：英伟达官网、AMD 官网、谷歌官网、habana 官网、ANANDTECH、semianalysis、tom's Hardware、TechPowerUp、THENEXTPLATFORM、华泰研究

图表19：英伟达 H100 与部分同业产品在运行不同 AI 负载时表现

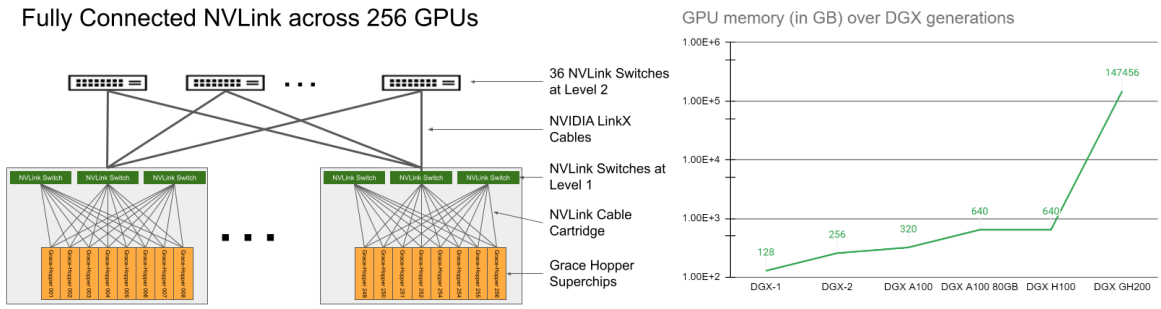


资料来源：福布斯官网、华泰研究

先进的网络技术 NVLink+NVSwitch，为吞吐量和可扩展性带来突破

NVLink+NVSwitch，英伟达先进的网络技术为吞吐量和可扩展性带来突破。大规模的计算负载需要实现多节点连接。在 2023 年 5 月 29 日举办的 COMPUTEX 大会上，英伟达正式发布 NVIDIA DGX GH200 超级计算系统。该系统通过 NVLink 互连技术及 NVLink Switch 串联 32 台由 8 块 GH200 超级芯片（总计 256 块），组成了一个 144 TB 内存及 1 exaFLOPS = 1000 petaFLOPS 算力的超级计算系统。大规模的共享内存解决了 AI 大模型训练的关键瓶颈，Google Cloud、Meta 与微软将是其首批用户。NVLink 与 NVSwitch 相结合实现了服务器节点间通信拓展和高速互联，使大规模并行处理成为可能，是支撑英伟达 GPU 系统实现高速通信的基石。

图表20：NVIDIA DGX GH200 通过 NVLink + NVSwitch 整合了 256 个 GPU，实现高达 144 TB 内存的容量



资料来源：英伟达官网、华泰研究

图表21：英伟达 DGX H100 VS DGX GH200

	DGX H100	DGX GH200
GPU and CPU	8x NVIDIA H100 Tensor Core GPUs + Dual Intel® Xeon® Platinum 8480C Processors	256x NVIDIA Grace Hopper Superchips （each Grace Hopper Superchip includes Grace Arm® CPU+ H100 Tensor Core GPU）
CPU Cores	112 Cores total, 2.00 GHz （Base）， 3.80 GHz （Max Boost）	18,432 Arm® Neoverse V2 Cores with SVE2 4X 128b
GPU memory	640GB	144TB
Performance（FP8）	32 petaFLOPS	1 exaFLOPS
NVIDIA® NVSwitch	4x	96x L1 NVIDIA NVLink Switches 36x L2 NVIDIA NVLink Switches
Networking	4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet	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
Management network	10Gb/s onboard NIC with RJ45 100Gb/s Ethernet NIC Host baseboard management controller （BMC） with RJ45	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 3-year business-standard hardware and software support	

资料来源：英伟达官网、华泰研究

英伟达独家的 NVLink 网络连接技术，解决了数据传输带宽瓶颈。NVLink 是英伟达针对 GPU 加速计算而开发的高速互连技术，自 2016 年推出后已发展至第四代。NVLink 能提供比 PCIe 更大的带宽，满足 AI 工作负载大规模的传输需求。传统的 PCIe 技术下，每个设备拥有专用的点对点连接，对于运行大规模并行功能和移动大量数据的 CPU 和 GPU 往往存在性能瓶颈。为了构建满足人工智能等发展需求的端到端计算平台，英伟达推出的 NVLink 技术，为 CPU、GPU 与系统其他部分之间提供高带宽的连接路径，也允许多个 GPU 通过高速互连直接通信，从而实现 GPU 之间可用内存的组合和访问。

图表22：历代 NVLink 参数

	NVLink 2	NVLink 3	NVLink 4
DSR 数据信号速率（Gbps）	25	50	100
单链路的通道数	8	4	2
单链路的单向带宽（GB/s）	25	25	25
单链路总带宽（GB/s）	50	50	50
总链路数	6	12	18
总带宽（GB/s）	300	600	900
推出年份	2016	2017	2020

资料来源：英伟达官网、ICspec 官网、华泰研究

NVLink 4 的总带宽可达每秒 900 GB/s，为 PCIe 5.0 总线带宽的 7 倍，也远高于下一代 PCIe 6.0 的 256 GB/s 速率，能满足 AI 和 HPC 持续增长的对多节点、多 GPU 系统的计算需求，为深度学习训练提供了更大的延展空间。此外，使用 NVLink 技术的设备有多个路径可供选择，和共享中央集线器的 PCIe 相比，加快了系统的运行速度，提升了数据流和总系统吞吐量。

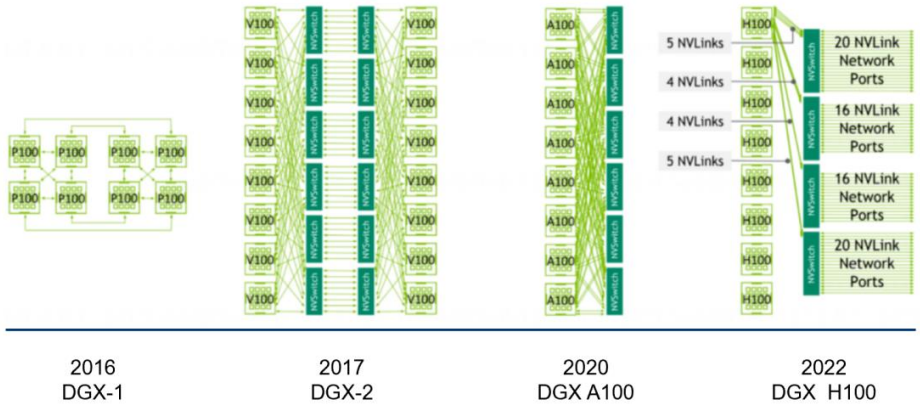
图表23：历代 PCIe 总线标准

PCIe 标准	单通道数据传输速率（GT/s）	编码	x16 带宽（GB/s）	标准批准年份
1.x	2.5	8b/10b	8	2003
2.x	5	8b/10b	16	2007
3.x	8	128b/130b	32	2010
4.0	16	128b/130b	63	2017
5.0	32	128b/130b	128	2019
6.0	64	PAM4/FLIT	256	2022

资料来源：WCCFtech 官网、Rambus 官网、华泰研究

NVSwitch 是英伟达的节点交换架构，通过连接多个 NVLink，在单节点内和节点间实现多 GPU 的拓展。NVSwitch 在 2018 年随第二代 NVLink 一起推出，将多个 NVLink 加以整合，实现多对多的 GPU 通信，进一步提高系统的可拓展性。最新的第三代 NVSwitch 采用台积电 4N 工艺打造，能在单个服务器节点中支持 8 到 16 个完全链接的 GPU，支持以 900GB/s 的速度互连每个 GPU，保障它们之间的完整点对点通信。

图表24：NVLink 和 NVSwitch 协同工作



资料来源：英伟达官网、ICspec 官网、华泰研究

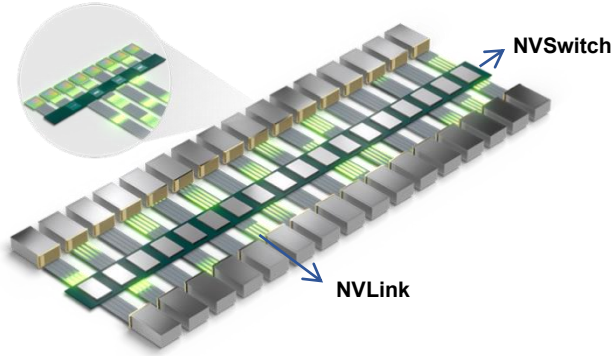
图表25：各代 NVSwitch 性能演变

	第一代	第二代	第三代
直接连接的 GPU 数量/节点数量	最多 8 个	最多 8 个	最多 8 个
NVSwitch GPU-to-GPU 带宽	300GB/s	600GB/s	900GB/s
总聚合带宽	2.4TB/s	4.8TB/s	7.2TB/s
支持的 NVIDIA 架构	NVIDIA Volta architecture	NVIDIA Ampere architecture	NVIDIA Hopper architecture

资料来源：英伟达官网、华泰研究

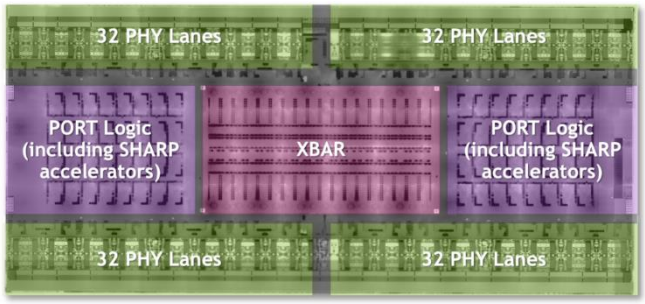
PCI Express→NVLink→NVLink+NVSwitch 的演变历程体现了英伟达对通信效率和扩展性的不断追求。随着深度学习算法的复杂化和规模的扩大，权重、梯度和偏差的同步与交换对通信延迟和高带宽提出了更高的要求。PCIe 在原始带宽、延迟以及缓存一致性等方面限制了 GPU 之间大规模的组合和连接，于高性能计算和数据中心里适用性较低。NVLink 和 NVSwitch 的推出解决了多个 GPU 大规模集群的传输，实现更快和更可扩展的计算系统。

图表26：NVLink 与 NVSwitch 可以纵向扩展，训练大参数模型



资料来源：英伟达官网、华泰研究

图表27：NVIDIA NVLink 4 NVSwitch 示意图

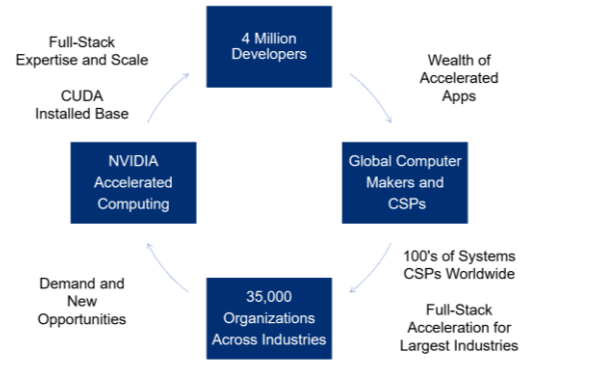


资料来源：英伟达官网、华泰研究

CUDA 软件生态的先发优势构筑英伟达第二护城河

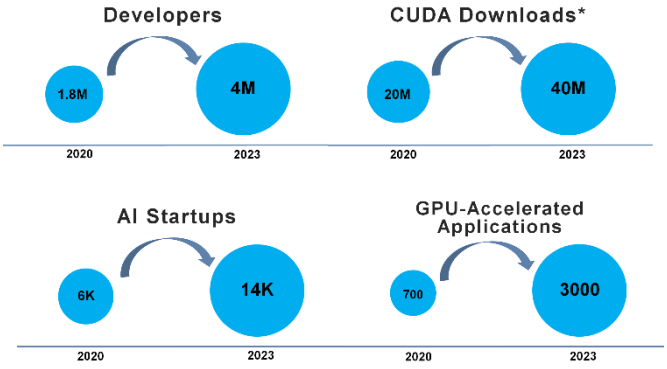
CUDA 工具包包括一系列的编程工具、加速库和框架，可用于协助开发人员更为便捷地进行 GPU 编程和并行计算，其核心竞争力主要在于其硬件集成、完善的社区资源和广泛的应用支持形成的正循环。2023 COMPUTEX 大会上，英伟达 CEO 黄仁勋表示，CUDA 拥有超过 400 万开发人员和超过 3000 个应用程序。受益于英伟达 CUDA 的先发优势与长期耕耘，搭配其新手友好的安装与编程体验，CUDA 庞大的用户群体致使其搭建起由大量专业开发者与领域专家组成的开发者社区。CUDA 也涵盖各类计算应用的代码库资源，这无疑为 CUDA 的学习和应用提供进一步支持。截止 2023 年 5 月，CUDA 下载量已累计超过 4000 万次，且仅去年一年便达到 2500 万次。

图表28：CUDA 软件生态圈效应



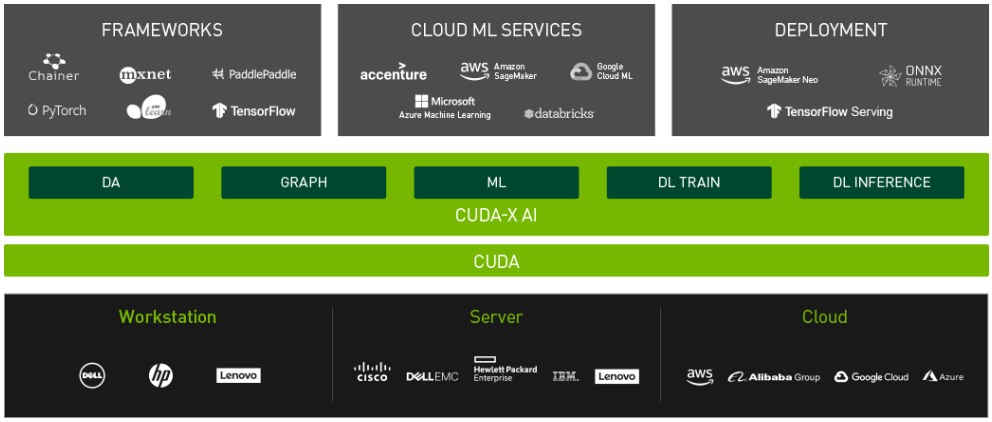
资料来源：英伟达官网、华泰研究

图表29：英伟达不断扩展加速计算生态圈



备注：*表示 Cumulative
资料来源：CES 2023、华泰研究

图表30：英伟达 CUDA-X AI 生态圈及相关客户矩阵



资料来源：英伟达官网、华泰研究

同类对比下，CUDA 软件生态丰富成熟，在应用广泛性、市场份额和开发者支持方面均较为突出。由英伟达在 2007 年推出的编程平台 CUDA 先发优势较明显，历年来大量机器学习算法工程师均使用。与 CUDA 生态系统对标的平台方面，OpenCL 虽在通用性上更胜一筹，但其缺少针对并行计算的相关优化及深度学习相关功能库较少；而 AMD ROCm 发布时间较晚，加上此前长期只支持 Linux 导致跨平台能力不足，且对比 CUDA 其在科学计算与深度学习领域的功能库、开发工具和应用支持的完善方面仍有改进空间。目前，虽有众多算力芯片厂商选择兼容 CUDA 的路径打开市场，如 ROCm 可以通过 HIP（Heterogeneous-Computing Interface for Portability）兼容，但 CUDA 并未开源，因此 100%兼容 CUDA 较为被动。我们认为，CUDA 生态凭借稳定的先发优势与用户粘性，将持续为英伟达的软件生态圈壁垒。

图表31：CUDA、ROCm 以及 OpenCL 三者对比

	CUDA	ROCm	OpenCL
发布时间	2007 年	2016 年	2008 年
硬件支持	支持 2006 年以来所有的英伟达 GPU	较多支持 AMD 高端 GPU 系列，自 2023 年 6 月 29 日更新的 ROCm 5.6.0 版本开始逐步向消费级 GPU 拓展	可移植到 NVIDIA、AMD、Intel 等等各种其它硬件设备，包括 FPGA 和 ASIC 除了供应商特定的扩展
操作系统	Linux 和 Windows	支持 Linux，自 2023 年 6 月 29 日更新的 ROCm 5.6.0 版本开始在 Windows 支持部分 AMD 消费级 GPU	支持包括 Linux、Windows 以及 macOS 在内的几乎所有的操作系统
功能库	拥有广泛的高性能库，覆盖广泛的终端应用场景，包括资源受限的物联网设备、自动驾驶及超级计算机等领域：九大部分组成的合作伙伴库、八大部分组成的数学库、五大部分组成的深度学习库、四大部分组成的图像和视频库、两大部分组成的通讯库、并行算法库以及计算光刻库	仅包括 CUDA 库中的一部分：由 Linear Algebra Libraries 线性代数库、Fast Fourier Transforms 快速傅里叶变换和 Random Numbers 随机数三部分组成的数学库；rocPRIM、rocThrust 以及 hipCUB 三部分组成的 C++ 原始库；MIOpen、Composable Kernel 和 MIGraphX 三大部分组成的 AI 库和以 RCCL 构成的通讯库	OpenCL 作为通用平台，并行计算能力远不如 CUDA（并行编程模型），并行计算能力只达到后者的 1/10 到 1/200 不等，目前很少用于机器学习和深度学习，因此可用的相关库很少
技术细节	一种使用 CUDA 关键字实现并行化的平台和编程模型；CUDA 提供 C/C++ 语言扩展和 API，用于编程和管理 GPU。	类似于 NVIDIA 的 CUDA，ROCm 支持多种编程语言、编译器、库和工具，目前主要通过异构计算可移植接口（HIP）这种 C++方言来简化 CUDA 应用程序到可移植 C++代码的转换。HIP 提供了 C 风格的 API 和 C++的内核语言。	不支持使用 C++ 编写代码，而是提供了类 C 语言编程工作环境

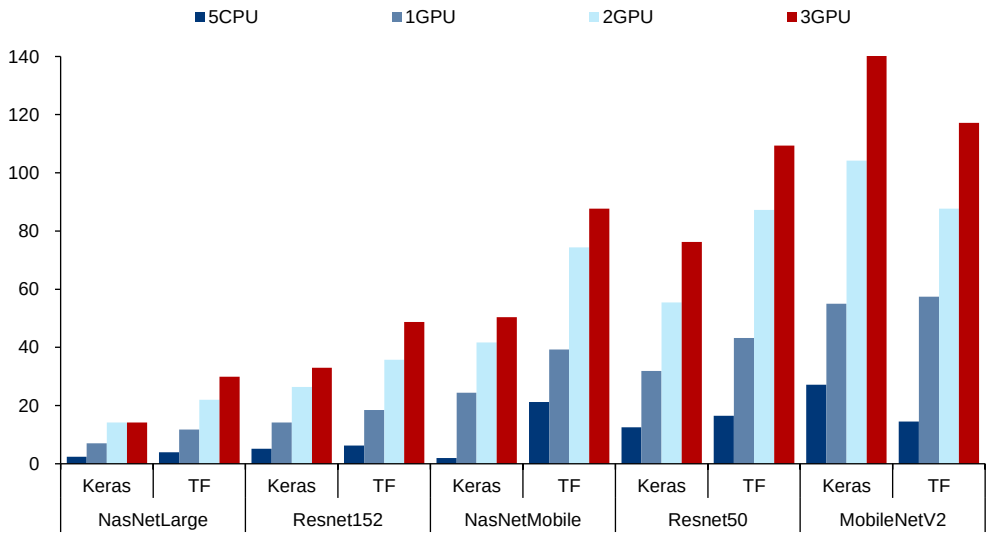
资料来源：CUDA 官网、run:ai 官网、phoronix、incredibuild、华泰研究

生成式 AI 下较复杂的推理需求或为英伟达 GPU 赋新机

长期以来，AI 推理市场一直由 CPU 主导。根据 The nextplatform 在 2023 年 4 月 5 日的报道，数据中心 70%的推理在 Intel Xeon CPU 上运行。主要鉴于过去的 AI 推理任务较为简单，以通用 AI 计算为主，如个性化广告、搜索推荐、中小型模型等应用，这些基础的推理任务 CPU 便可胜任。此外，CPU 在 AI 领域的应用较早，云厂商已建立了大量的基础设施和工具来支持，更换及重新配置的成本或也是考虑因素之一。

面对不同的 AI 负载任务，GPU、CPU 和 ASIC 或分别具有性能和成本优势。GPU 擅长并行计算，特别适用于深度学习模型中的大规模矩阵计算。相比之下，CPU 更擅长串行计算。因此，面对不同的 AI 负载，不同种类的芯片或能发挥不同的性能和成本优势。虽然 GPU 单个成本高于 CPU，但在处理深度学习推理任务时，特别面向较复杂和多模态的 AI 负载，GPU 或能提供更高的性能和效率，从而实现更快的推理速度和更高的吞吐量，节省了部署和运行成本。在 2018 年 9 月，微软发表了一篇 GPU 与 CPU 在深度学习模型推理部署对比的博客：测试中所采用的 3 节点 GPU 集群与 5 节点 CPU 集群的成本大致相同，在该测试所用的模型和框架中，3 节点 GPU 集群的吞吐量优于 5 节点 CPU 集群。

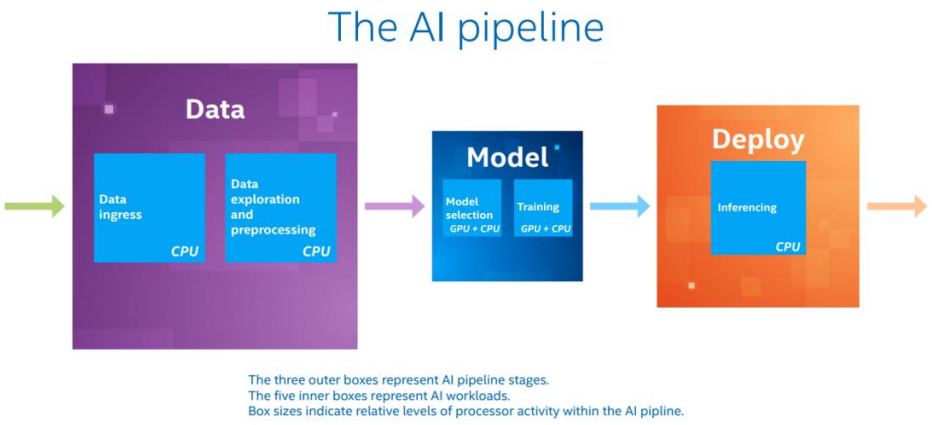
图表32：微软的深度学习推理测试：GPU 与 CPU 吞吐量对比（张图片/秒）



资料来源：微软官网，华泰研究

我们认为随着生成式 AI 和大模型的兴起，高复杂度的推理任务变得更加普遍，或将推动 GPU 在推理端的需求。对于较大的模型和较复杂的计算任务，CPU 单独来说或不足以满足，未来这部分的推理应用或将从 CPU 转移至 CPU+加速器（ASIC 或 FPGA）或者 CPU 转移至 GPU。英伟达 CEO 黄仁勋也在 8 月 8 日的 SIGGRAPH 2023 表示，在 1 亿美元的预算下，数据中心可以采购 2500 块 GH200 进行推理，功耗为 3MW，能实现同等预算下 x86 CPU 方案 12 倍的 AI 推理性能和 20 倍的能效。

图表33：在 AI 的通用计算中，推理以 CPU 为主



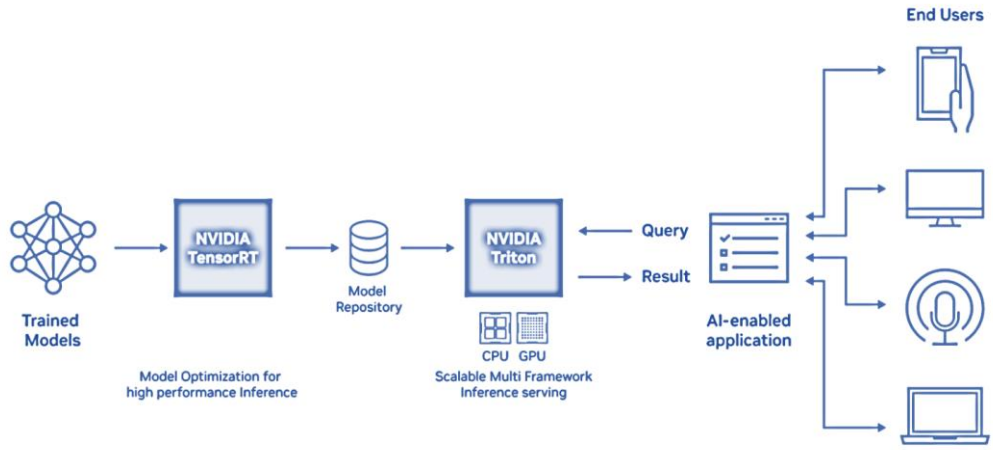
资料来源：Intel 官网，华泰研究

针对推理市场，英伟达推出了一系列的新品。英伟达的推理平台 GPU 产品组合包括用于 AI 视频的 NVIDIA L4、用于图像生成的 NVIDIA L40、用于大型语言模型推理部署的 NVIDIA H100 NVL 和用于推荐模型的 NVIDIA Grace Hopper。这些平台的软件层采用 NVIDIA AI Enterprise 软件套件，包括用于高性能深度学习推理的软件开发套件 NVIDIA TensorRT，以及帮助实现模型部署的开源推理服务软件 NVIDIA Triton Inference Server。

图表34： NVIDIA 在 2023 GTC 推出的四款推理产品

资料来源：NVIDIA，华泰研究

图表35： NVIDIA AI 推理流程



资料来源：NVIDIA 官网、华泰研究

图表36： 英伟达新推理产品与旧版 GPU、其他厂商 GPU 对比

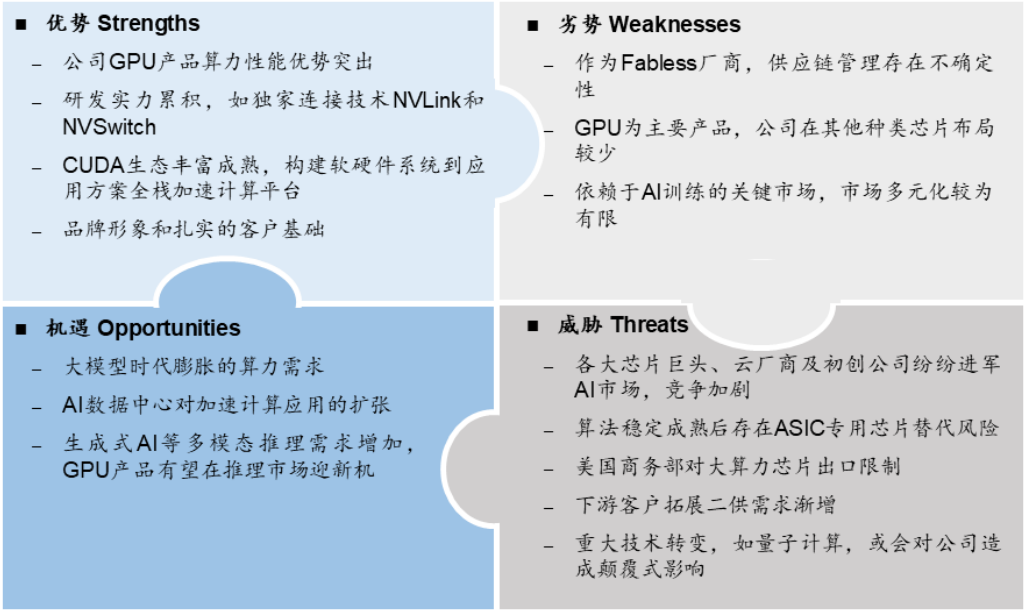
	NVIDIA L4	NVIDIA L40	NVIDIA T4	NVIDIA A2	NVIDIA A10	Intel Data Center GPU Flex 170	Intel Data Center GPU Flex 140
Release Year	2023	2022	2018	2021	2021	2022	2022
Memory Size （GB）	24	48	16	16	24	16	12
Memory Type	GDDR6	GDDR6	GDDR6	GDDR6	GDDR6	GDDR6	GDDR6
Memory Bus （bit）	192	384	256	128	384	256	192
Bandwidth （GB/s）	300.1	864.0	320.0	200.0	600.0	512.0	372.0
Pixel Rate （Gpixel/s）	163.20	478.10	101.80	56.64	162.70	262.40	124.80
Texture Rate （Gtexel/s）	489.60	1,414.00	254.40	70.80	488.20	524.80	249.60
FP16 half （TFLOPS）	31.33	90.52	65.13	9.00	62.50	33.59	15.97
FP32 float （TFLOPS）	31.33	90.52	8.14	9.00	31.20	16.79	7.99
FP64 double （GFLOPS）	489.60	1,414.00	254.40	70.80	976.30	-	-
Price	Unknown	Around $9000	Around $1500	Around $1400	Around $9000	Probably around $6000	Probably around $4000

资料来源：英伟达、英特尔、SHI 官网，华泰研究

英伟达 SWOT 分析

我们认为英伟达的 AI 芯片，凭借高算力，以及 NVLink 等独家技术，叠加高粘性的 CUDA 软件生态圈加持，优势显著。

图表37：英伟达 SWOT 分析



资料来源：华泰研究

传统芯片巨头：备受期待的突围者 AMD

以 AI 为战略首位，AMD 突围二战能否凯旋？

AMD 的 AI 战略主要包括三个方向：1）广泛的 CPU 和 GPU 产品组合；2）开放稳定且已证明（open, steady and proven）的软件平台；3）ROCm 生态系统。CEO Lisa Su 于 5 月 31 日的《福布斯》采访时强调“放眼 5 年，将在 AMD 每一个产品中看到 AI”，AI 是公司目前的战略首位。

目前英伟达领军 AI 训练端，但随着 AMD 在 AI 芯片上逐步发力，或能开始撼动英伟达在行业里独占鳌头的地位。我们也认为云厂商应不希望 AI 芯片呈现一家独大的局面，MI300 恰逢其时地出现，为市场提供了英伟达以外的选择。MI300 虽备受瞩目，但截至 2023 年 8 月在客户方面几乎未有正式披露，因此，我们认为，一旦有大型云客户正式宣布部署，或将有效提振市场信心。目前云厂商应还在对 MI300 进行测试和下单阶段，我们将对后续公布的订单情况保持关注。公司在 23Q2 业绩电话会中提到 AI 业务进展势头强劲，截至 23Q2 客户对 AI 产品的“参与度”环比增长超过七倍，主要来自 MI300 的新订单和 MI250 的增量订购，MI300 将在本年四季度开始出货。

MI300A 和 GH200：CPU+GPU AI 芯片架构仿生人脑结构

MI300 系列目前包括两款产品：1）MI300X：纯 GPU，由 12 个 chiplets（8 个 GPU+4 个 IO+Cache）构成；**2）MI300A：**CPU+GPU，由 13 个 chiplets（6 个 GPU+3 个 CPU+4 个 IO+Cache）构成。MI300X 作为纯 GPU 产品或对标英伟达 GPU H100，而 MI300A 为 APU 架构（Zen 4 CPU + CNDA 3 GPU），与英伟达的异构 CPU+GPU 芯片 GH200 正面交锋。

我们认为，MI300A 和 X 是客户在英伟达 GPU 之外的有力选择，或也可对 AI 芯片定价造成一定影响。MI300 系列在参数上值得关注的亮点包括：1）MI300X 的 192GB HBM3 内存，领先英伟达 H100 双卡 NVL 的 188GB HBM3，更远超 H100 PCIe 和 SMX 的 80GB HBM3，而 MI300A 的 128GB HBM3 内存也不遑多让；2）MI300X 晶体管数量 1530 亿，MI300A 晶体管数量 1460 亿，对比 H100 的 800 亿；3）内存带宽 5.2TB/s 与英伟达 H100 的 2-7.2TB/s 相近；4）Infinity Fabric 互联带宽的 896GB/s 与 NVLink 的 900GB/s 也相差无几；5）比 H100 高 2.4X 的 HBM 密度，以及 1.6X HBM 带宽。

AMD 在 2023 年 CES 大会和 2023 年数据中心和人工智能首映式上，分别展示了 MI300 的 AI 训练和推理能力：1）MI300A 可将 GPT 级别的模型训练时间从“几个月”降低至“几周”；2）单片 MI300X 可推理 400 亿参数的 Falcon 模型。AMD 称这是这种参数规模的大语言模型第一次在单 GPU 上运行（需要注意的是，400 亿参数在如今千亿参数大模型时代并不大，公司主要强调的是单 GPU），公司进而称单片 MI300X 还可推理规模更大的模型，比如 Meta 的 OPT 模型（660 亿参数版本）和 LLaMA（650 亿参数）。若使用 FP16 精度，单片 MI300X 最高可推理 800 亿参数的模型。

在 AMD 的 MI 系列 GPU 中，除 MI300A 以外的所有产品都是纯 GPU。AMD MI 系列 GPU 始自发布于 2018 年 11 月的 MI50 和 MI60，二者制程都是台积电 7nm（早于英伟达 2020 年发布 7nm 的 A100），晶体管数量均为 132 亿，分别有 16GB 和 32GB HBM2。2020 年 11 月，AMD 发布 MI100，晶体管数量 256 亿，依然是 7nm 制程和 32GB HBM2。2021 年 11 月，MI200 系列（MI250 和 MI250X）发布，对标英伟达 A100，据 AMD 称其可达到 A100 AI 大模型训练性能的 80%；6nm 工艺，582 亿晶体管，128GB HBM2e，其中 MI250 比 MI250X 的算力略低。2022 年 5 月 Build 大会上，微软宣布成为部署 MI200 系列的首个云厂商。2023 年 3 月 Morgan Stanley TMT 大会上微软再次宣布已在云上部署了大量 MI250。2022 年 3 月，AMD 正式发布了 MI200 系列的第三个产品 MI210，仅 64GB HBM2e，且算力也比另外两款 MI200 产品约低 50%，是 MI200 系列的基础版。

图表38： AMD Instinct MI 系列 GPU 发展历程

产品名称	MI50	MI60	MI100	MI250	MI250X	MI210	MI300A	MI300X
发布时间	2018.11	2018.11	2020.11	2021.11	2021.11	2022.3	2023.1	2023.6
峰值算力（TFLOPS）	FP16: 26.82 FP32: 13.41	FP16: 29.49 FP32: 14.75	FP16: 184 FP32: 23.1	FP16: 362.1 FP32: 45.3	FP16: 383 FP32: 47.87	FP16: 181 FP32: 22.6	相比 MI250 约有 8 倍提升	暂无
工艺制程及芯片面积	7nm, 331mm^2	7nm, 331 mm^2	7nm, 750 mm^2	6nm, 724 mm^2	6nm, 724 mm^2	6nm, 724 mm^2	5nm, 1017 mm^2	5nm, 1017 mm^2
晶体管数量（亿）	132	132	256	582	582	582	1460	1530
内存容量	16 GB HBM2	32 GB HBM2	32 GB HBM2	128 GB HBM2e	128 GB HBM2e	64 GB HBM2e	128 GB HBM3	192 GB HBM3
内存带宽	1024 GB/s	1024GB/s	1.2TB/s	3.2TB/s	3.2TB/s	1.6TB/s	3.2TB/s	5.2TB/s
热设计功耗 TDP（W）	300	300	300	500	500	300	600	700/350（air-cooled）

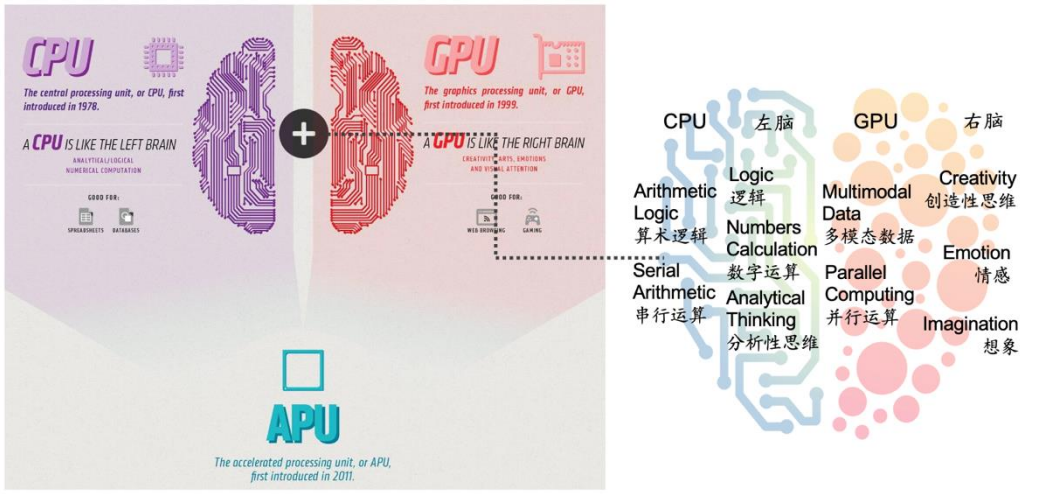
资料来源：AMD 官网、华泰研究

2023 年 1 月，AMD 在 2023 CES 大会上首次推出了 CPU+GPU 的 MI300，后改称 MI300A。MI300A 是 MI 系列的第一款 CPU+GPU 异构产品，我们认为 CPU+GPU 架构已成为 AI 芯片的趋势，鉴于 AI 的最终目标是模仿人类大脑的操作，因此 AI 芯片也应仿生人脑结构，并顺应多模态模型的发展需求。如前文所述，英伟达的 Grace Hopper 也是 CPU+GPU 架构。

在 AI 应用里，GPU 算力高，针对并行计算，在视频处理、图像渲染等方面的优势毋庸置疑，但并非所有工作负载都是单纯的 GPU-bound，也须由 CPU 进行控制调用，发布指令。因此，在 CPU+GPU 架构里的 CPU 可负责控制及发出指令，指示 GPU 处理数据和完成运算（如矩阵运算）。值得一提的是，MI300A 里的 CPU，AMD 选用了 x86 架构，而 GH200 里的 CPU，英伟达则采用了 ARM 架构。我们认为，两者的选择各有优势。一般来说，ARM 架构主要应用于移动端，因此相比 x86 能耗较低，这点不管在 AI 或是数据中心的应用也会受到青睐。我们认为英伟达也是看准这点（公司也曾对 ARM 提出收购），加上在这类 CPU+GPU 架构中，CPU 或仅需发挥其部分性能，如向 GPU 发出指令等，其他性能如 AI 训练和推理可交由 GPU 负责，因此 ARM 架构已能胜任。反过来，x86 架构则追求高性能和拥有较丰富的指令集，在 AI 里也可分担推理负载，与 GPU 在功能上互补。另外，AMD 的 x86 CPU 也主打较高能效 (Performance per Watt)，因此也符合 AI 和数据中心的条件。

在面对不同模态数据的推理时，我们认为 CPU 与 GPU 的分工各有不同，因此同时部署 CPU 和 GPU 能提供更大的运算支撑。例如，在处理语音、语言和文本时，AI 模型需计算有序，因此或更适合使用擅长串行运算的 CPU；但在处理图像、视频等推理时（对比人类在看到一幅图片时，每一个像素同时进入眼睛），需要大规模并行运算，更适宜由 GPU 负责。

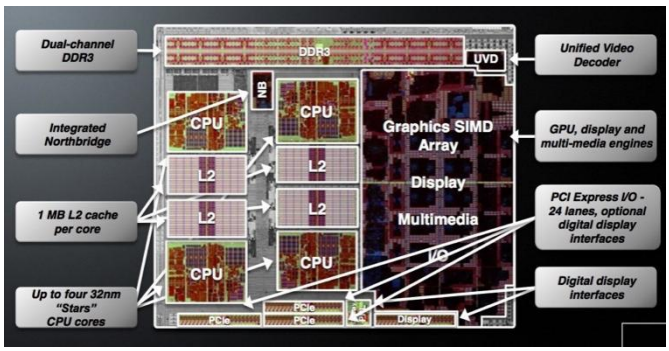
图表39：2011 年 AMD 提出 APU 概念，将结合 CPU 与 GPU 在左右脑层面的分工区别和组合构想



资料来源：AMD 推特官方、华泰研究

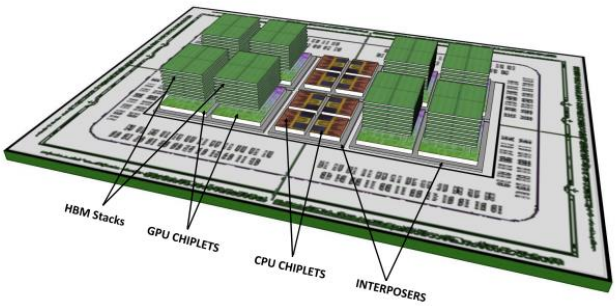
我们认为 AMD 在 CPU+GPU 架构具备深厚的 know-how。MI300A 虽是 AMD 的首个 CPU+GPU 架构的数据中心产品，但其实 AMD 早在 2011 年发布的 APU（Accelerated Processing Unit）已是 CPU+GPU 架构，当时只用于 PC 端。再向前追溯，我们发现 AMD 的 CPU+GPU 架构理念较早有迹可循。在 2006 年，当时 AMD 通过收购 ATI 获得 GPU 以及芯片组技术，并在同年开展 AMD Fusion 项目（即后来的 APU），提出将 CPU 和 GPU 集成到一颗芯片上的理念，但当时 AMD 的 CPU 和 GPU 采用 45nm 制程，将两者放在同一芯片上的难度较大，直到 2011 年，AMD 发布首款 APU 产品 Llano，真正把异构的理念落地。2017 年，AMD 发布的论文《Design and Analysis of an APU for Exascale Computing》中讨论了包含 CPU、GPU 和 HBM 内存堆栈的 APU 芯片设计。

图表40：2011 年 AMD 的首款 APU 产品 Llano：CPU+GPU 结合的设计第一次落地



资料来源：AnandTech、华泰研究

图表41：2017 年 AMD 发布的论文中讨论了包含 CPU、GPU 和 HBM 内存堆栈的 APU 芯片设计



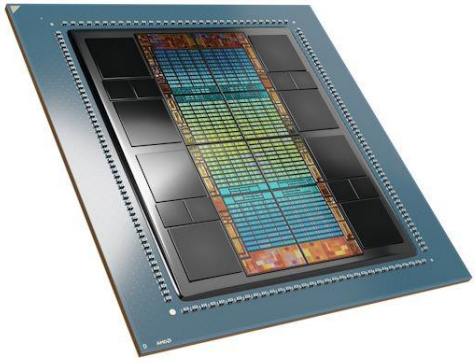
资料来源：T. Vijayaraghavan et al., "Design and Analysis of an APU for Exascale Computing," 2017 IEEE International Symposium on High Performance Computer Architecture （HPCA）, Austin, TX, USA, 2017, pp. 85-96, doi: 10.1109/HPCA.2017.42.、华泰研究

图表42：英伟达及 AMD 主要 GPU 相关产品参数对比

	英伟达		AMD		
产品名称	A100 PCIe | SXM	H100 PCIe | SXM | NVL	MI250X	MI300A	MI300X
发布时间	2020.6	2022.3	2021.11	2023.1	2023.6
峰值算力（TFLOPS）	FP16: 312 | 624 FP32: 19.5 FP64: 19.5	FP8: 3,026 | 3,958 | 7,916 FP16: 1,513 | 1,979 | 3,958 FP32: 51 | 67 | 134 FP64: 51 | 64 | 134	FP16: 383 FP32/64: 47.9 FP32/64 Matrix: 95.7	暂无	暂无
工艺制程及芯片面积	7nm, 826mm^2	4nm, 814 mm^2	6nm, 724 mm^2	5nm, 1017 mm^2	5nm, 1017 mm^2
晶体管数量（亿）	540	800	582	1460	1530
内存容量	80 GB HBM2e	80 | 80 | 188 GB HBM3	128 GB HBM2e	128 GB HBM3	192 GB HBM3
内存带宽	1.9 | 2TB/s	2 | 3.35 TB/s | 7.8TB/s	3.2 TB/s	3.2TB/s	5.2TB/s
Interconnect	600 GB/s NVLink for 2 GPUs 64 GB/s PCIe Gen4	600 | 900 | 600 GB/s NVLink 125 GB/s PCIe Gen5	100GB/s	约 800GB/s	896GB/s
热设计功耗 TDP（W）	300 | 400	300-350 | 700 | 2x350-400	500	600	暂无

资料来源：AMD 官网、英伟达官网、华泰研究

图表43：AMD MI300X 产品实物图（共 8 个 GPU chiplets）



资料来源：AMD 数据中心&AI 首映式、华泰研究

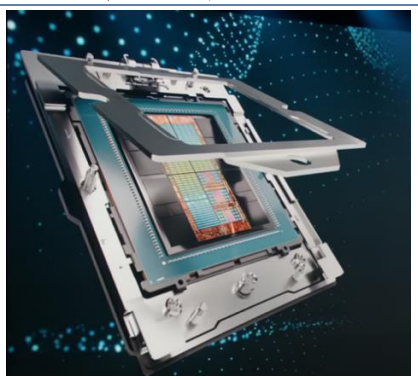
图表44：搭载 8 个 MI300X 的 Instinct Platform

资料来源：AMD 数据中心&AI 首映式、华泰研究

图表45：单片 MI300X 运行 400 亿参数的 Falcon 模型

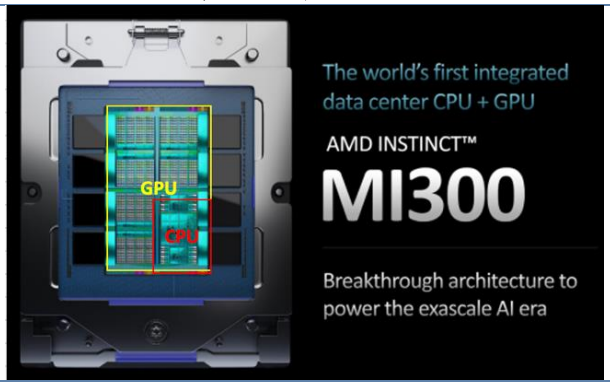
资料来源：AMD 官网、华泰研究

图表46：AMD MI300A 产品实物图



资料来源：AMD 数据中心&AI 首映式、华泰研究

图表47：AMD MI300A 产品示意图



资料来源：CES 2023、华泰研究

ROCm 生态圈会否成为 AMD 的“阿克琉斯之踵”？分而治之或可解困

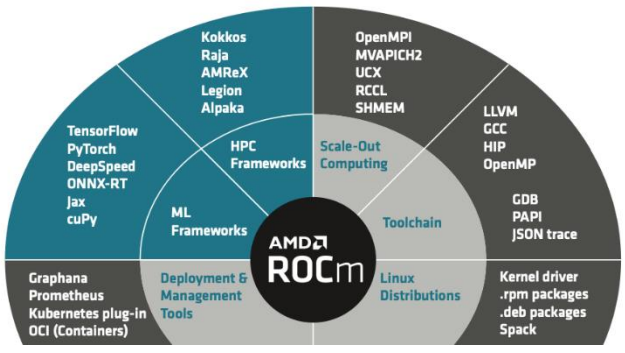
AMD 的软件生态圈 ROCm（Radeon Open Compute Ecosystem）于 2016 年 4 月发布，相比 2007 年发布的英伟达 CUDA 起步较晚。目前，ROCm 具备完全兼容 CUDA 的能力，为 AMD 提供了说服客户迁移的条件和理由，然而一味兼容只会导致 ROCm 受 CUDA 的掣肘，加上需应对 CUDA 的每一次更新迭代，或会导致 ROCm 陷入长期被动的局面，我们认为，这或已成为 AMD 的“阿克琉斯之踵”。为了更有效地破解此困境，AMD 进行了三类努力：1）如上所述持续兼容 CUDA；2）继续完善 ROCm 生态圈；3）与大型云和互联网厂商分别进行直接合作，分而治之（divide and conquer）与 CUDA 脱钩。

目前，ROCm 有以下三点不足：1）操作系统：长期只支持 Linux，在 2023 年 4 月才宣布登录 Windows；反观，CUDA 从 1.0 版就同时支持 Linux 和 Windows；**2）产品支持：**ROCm 长期只支持 AMD 的较高端 GPU，例如 Radeon Pro 系列等，直到 2023 年 4 月才宣布扩展至一些消费级 GPU 如 Radeon RX 6900 XT、Radeon RX 6600、以及 Radeon R9 Fury；反观 CUDA，2006 年发布的 G80 系列及以后所有的英伟达 GPU 都能支持；**3）开发者数量和生态扩展：**CUDA 在 2023 年达 400 万以上的开发者，这是 ROCm 暂时无法相比，我们发现，截至 2023 年 8 月 9 日，全球最大的开发者社区之一 StackOverflow 中“CUDA”的标签下已有超过 14000（14259）个问题讨论，而 ROCm 在“amd-rocm”的标签下只有 120 个问题讨论；在 Github 上，CUDA 已有超过 33400 个开发者贡献的软件包仓库，而 ROCm 只有不到 600（559）个。更多的开发者意味着不断迭代的工具和更广泛的多行业应用，因此 ROCm 需要更多的开发者以形成生态的良性循环。

为了破解“阿克琉斯之踵”，AMD 进行了三类努力：

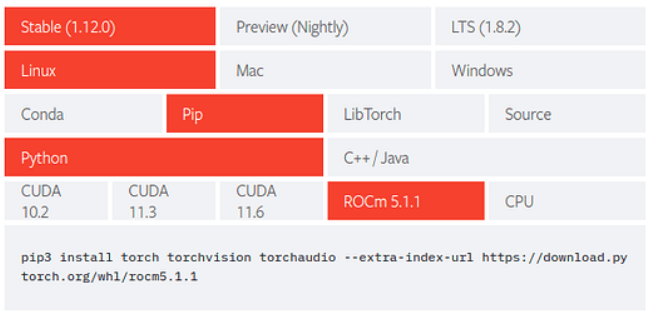
1）AMD 正积极拓展 ROCm 的生态圈：ROCm 从 2023 年 4 月 14 日开始支持 Windows 操作系统，终于迈出了来迟的一步。另外，ROCm 在 AI 领域进行了更广泛的框架和软件兼容，已支持 TensorFlow 和 PyTorch 等主流机器学习框架，而且与英伟达和英特尔同属 PyTorch 的 Founding Member（PyTorch 在 2022 年 12 月成立的联盟，包括四大云厂商、三大芯片巨头等）一员。软件库方面，优化深度学习库 MIOpen 和计算机视觉和机器智能库 MIVisionX，PyTorch 2.0 也在 2023 年一季度开始提供对 ROCm 的支持。

图表48：ROCm 支持主流机器学习框架



资料来源：AMD ROCm 手册 2022 版、华泰研究

图表49：PyTorch 中可以选择 ROCm



资料来源：AMD ROCm 手册 2022 版、华泰研究

丰富 ROCm 软件栈并非 AMD 的一厢情愿，AI 初创企业和 AI 开发者社区都愿为了更低的算力成本和更多的芯片选择而助 AMD 一臂之力。我们认为，AI 初创企业对于获取英伟达之外的其他可选算力意愿也不低。2023 年 6 月 30 日 **Mosaic ML**（初创生成式 AI 公司，MIT 背景）发布了基于 ROCm 使用 AMD 的 MI250 GPU 进行大语言模型的训练日志，称其希望在“这个全由英伟达供应的世界里”提高选择性。日志中，MosaicML“无需转码（no code changes were needed）”基于 AMD 的 MI250 和 ROCm 实现了模型训练。更多类似的尝试将推动 ROCm 的边界向外拓展。2023 年 6 月 14 日 AMD 数据中心与人工智能发布会上，**HuggingFace**（人工智能开发者社区，开源共享模型和数据集，可认为是 AI 的 Github）宣布与 AMD 建立合作，这项合作的重点正是把 Hugging Face 的 Transformer 库集成进 ROCm 中，目的是让用户在 AMD 的芯片上训练和推理在库中的模型时无需其他操作，正如 Hugging Face CEO Clement Delangue 在会上直言“我们希望所有人都能在 AMD 的芯片上运行模型（we want everyone to be able to run their models on AMD hardware）”。在 9 月 3 日 OpenAI 也宣布，其 Python 类的开源编程语言 Triton 也开始将 ROCm 并入。

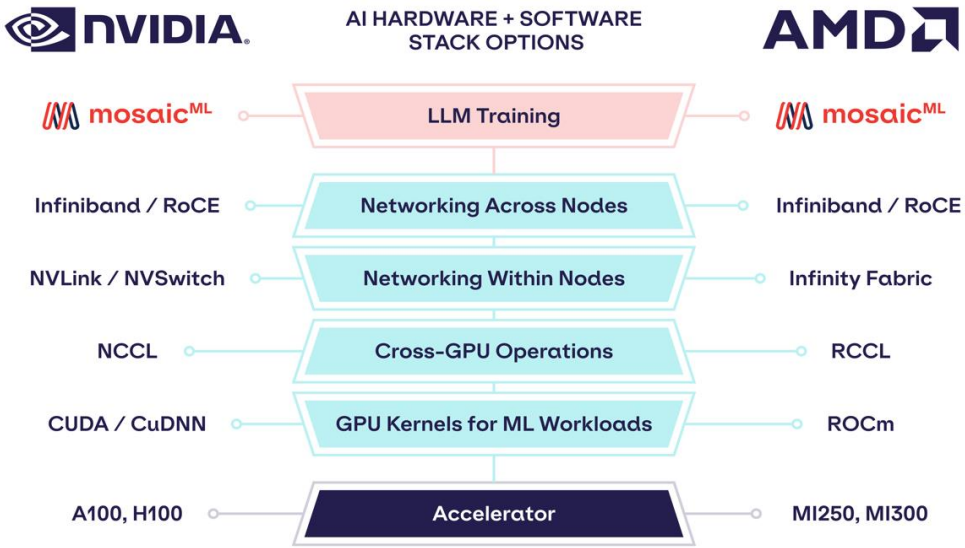
图表50：AMD 与 Hugging Face 的合作伙伴关系

资料来源：AMD 数据中心与人工智能首映式、华泰研究

图表51：AMD 与 Mosaic ML 的合作伙伴关系

资料来源：MosaicML 官网、华泰研究

图表52：MosaicML 希望能同时使用英伟达和 AMD 两套硬件+软件

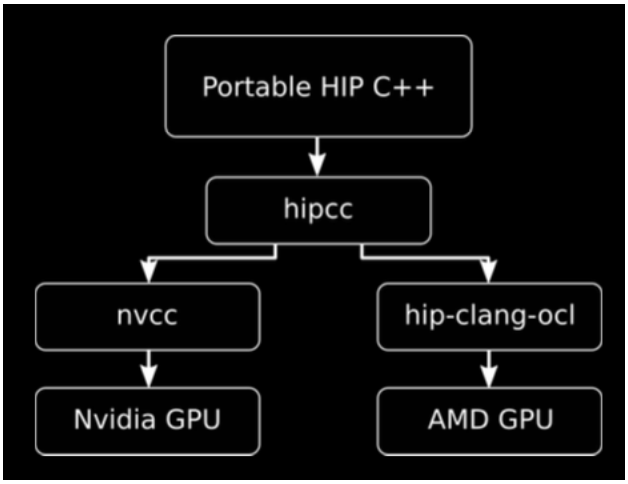


资料来源：MosaicML 官网、华泰研究

2）进一步兼容 CUDA：ROCm 可通过 HIP（Heterogeneous-Computing Interface for Portability）完全兼容 CUDA。HIP 是 AMD 的 GPU 软件开发框架，它提供的 HIPify-perl 和 HIPify-clang 工具，可用于 CUDA 到 HIP 的代码转换，转码后可在 AMD GPU 上编译运行，而基于 HIP 开发的应用也可同时用于 AMD 和英伟达的 GPU 上。虽然这种兼容方式依然需要开发者进行一定的转码工作，不过已可大量节省开发者针对 AMD 产品进行代码重新开发的时间。**这为 AMD 提供了说服客户进行迁移的条件和理由。但兼容 CUDA 属权宜之计，能让 AMD 在短期内争取客户和抢占市场。**

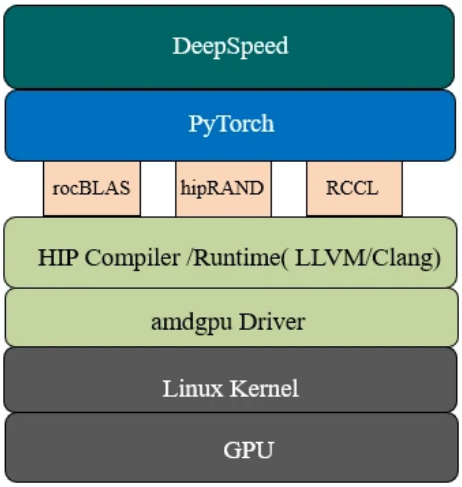
3）与云和互联网厂商等重要客户分别合作，重构自己的库，分而治之，以此与 CUDA 脱钩：长期一味兼容 CUDA 只会导致 ROCm 受 CUDA 的掣肘，加上需应对 CUDA 的每一次更新迭代，会导致 ROCm 陷入长期被动的局面。对云厂商来说，培育 AMD 与英伟达分庭抗礼，能有效影响芯片的定价权力，对重成本的云和互联网厂商来说也符合利益。因此，AMD 和这些厂商有一拍即合的基础，可通过分别与厂商客户合作构建兼容度更高的生态，分而治之与 CUDA 脱钩。例如 2022 年 3 月，微软在其开发的深度学习最优化函数库 DeepSpeed 中支持 ROCm，使开发者无需修改代码，就可以直接在 AMD 的 GPU 上运行。

图表53： AMD HIP 使 ROCm 可以部署英伟达和 AMD 的 GPU



资料来源：AMD 官网、华泰研究

图表54： 微软 DeepSpeed 函数库支持 ROCm 和 AMD GPU



资料来源：微软官网、华泰研究

图表55： 英伟达 CUDA 软件库及应用场景

种类	软件包组成	功能/应用
数学库	cuBLAS 、 cuFFT 、 cuRAND 、 cuSOLVER 、 cuSPARSE、cuTENSOR、AmgX	为分子动力学、计算流体学、计算化学、医学成像和地震勘探等领域的计算密集型应用奠定基础
并行算法库	Thrust	用于 C++中的多项运算，并在研究自然科学、物流、旅行规划等领域的关系时与图形一起使用
图像和视频库	nvJPEG、NVIDIA 性能基元、NVIDIA 视频编解码器 SDK、NVIDIA 光流 SDK	用于通过 CUDA 和 GPU 的专用硬件组件来进行图像和视频的解码、编码和处理
通信库	NVSHMEM、NCCL	性能经过优化的多 GPU 和多节点通信基元
深度学习库	NVIDIA cuDNN、NVIDIA TensorRT、NVIDIA Riva、NVIDIA DeepStream SDK、NVIDIA DALI	用于利用 CUDA 和专用硬件组件的深度学习应用
合作伙伴库	OpenCV、FFmpeg、ArrayFire、MAGMA、IMSL Fortran 数值库、Gunrock、CHOLMOD、Triton Ocean SDK、CUVIlib	包含 GPU 加速的开放源代码库等，覆盖矩阵、信号、图像、音频、视频等多种数据类型处理

资料来源：英伟达官网、华泰研究

图表56： ROCm 系统对应英伟达 CUDA 部分名称

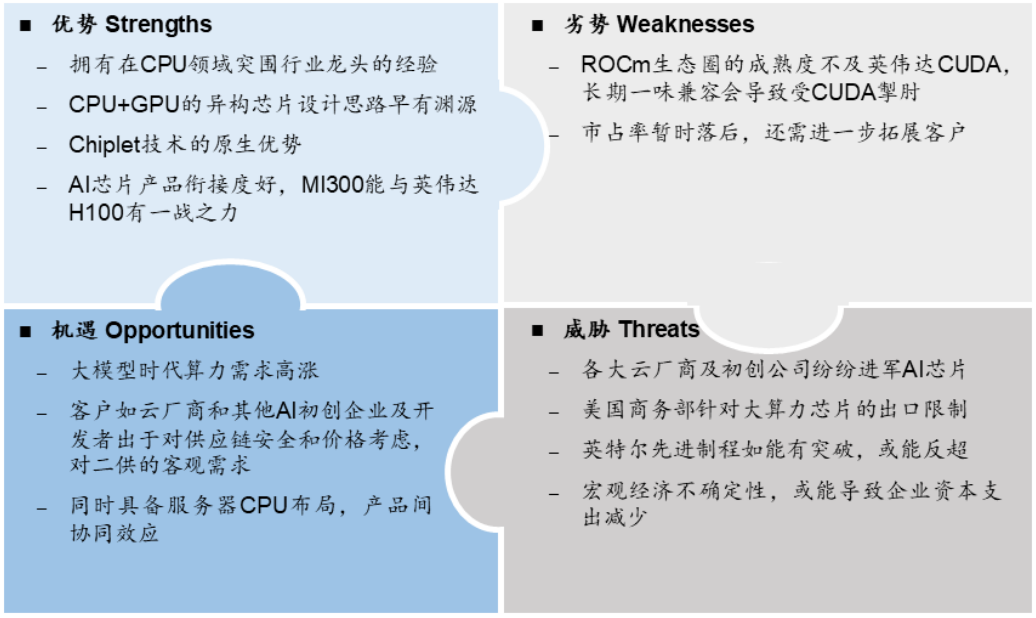
CUDA	ROCm
CUDA API	HIP
NVCC	HCC
CUDA 函数库	ROC 库、HC 库
Thrust	Parallel STL
Profiler	ROCm Profiler
CUDA-GDB	ROCm-GDB
NVIDIA-smi	ROCm-SMI
Direct GPU RDMA	ROCm RDMA
TensorRT	Tensile
CUDA-Docker	ROCm-Docker
cuDNN	MIOpen

资料来源：CSDN、华泰研究

AMD SWOT 分析

我们认为，AMD 目前最大的短板就在于 ROCm 软件生态圈的成熟度难敌英伟达的 CUDA，但 AMD 拥有突围行业龙头的经验，且产品稳步布局 AI，MI300 系列竞争力凸显。因此，我们认为 AMD 能在群贤毕至的 AI 芯片领域抓住客户拓展二供的需求，再次突围。

图表57： AMD SWOT 分析



资料来源：华泰研究

传统芯片巨头：多元布局的追赶者英特尔

英特尔在 AI 芯片追赶上以 ASIC 挂帅另辟蹊径

英特尔的 AI 芯片布局包括 GPU 产品 Ponte Vecchio（今年一季度已推出）和下一代 Falcon（准备在 2025 年推出）。在 2024 年的真空期内，公司准备以 ASIC 芯片 Gaudi 3 来填补。

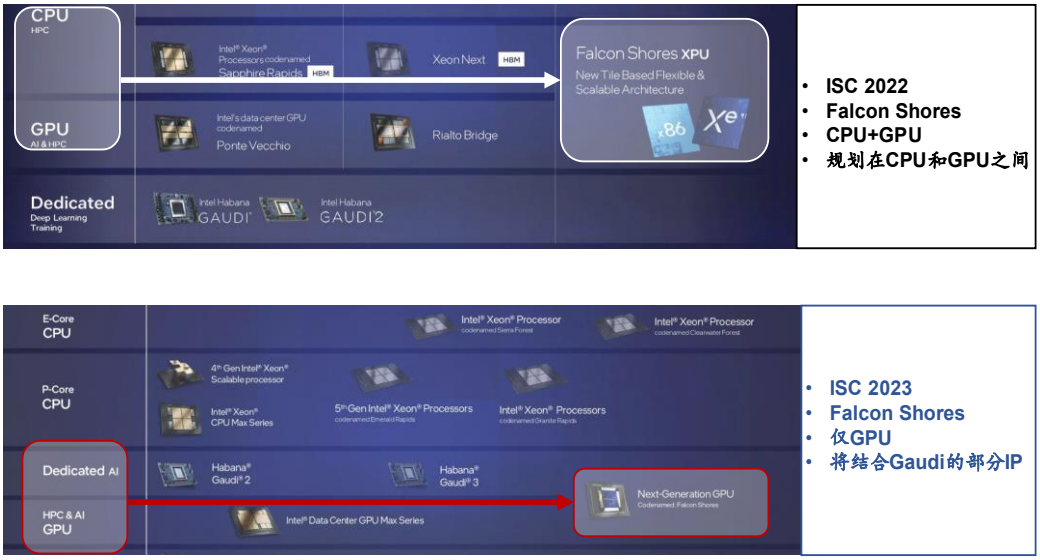
2023 年 3 月，英特尔称 Falcon Shores 为其下一个 GPU 产品，取消原定的 Rialto Bridge GPU，直接接棒 Ponte Vecchio GPU，且将 Falcon Shores 从原定的 2024 年推出延后至 2025 年推出。这意味着英特尔在 Ponte Vecchio 之后和 Falcon 之前将出现 GPU 真空期。当时，英特尔还发布了加速计算系统和图像组暂代总经理 Jeff McVeigh 的一封信，信中提到公司 GPU 产品的未来发展路径，包括取消一些原定在未来 1 至 1.5 年准备推出的服务器 GPU 产品，如 Rialto Bridge，但会将更多资源投放到 Falcon 上。

2023 年 5 月，英特尔在 ISC 2023 会议上再次确认 Falcon Shores 将推出仅 GPU 版，而不是之前说的 XPU（CPU+GPU 异构版本）。英特尔解释称，此前确实曾认为 AI 工作负载已经进入成熟阶段（Mature Workload），但目前发现 AI 工作负载依然处在动态阶段（Dynamic Workload），因此 CPU 与 GPU 的比例（CPU-GPU Ratio）难以确定下来故放弃异构设计。2023 年 3 月，负责 Falcon Shores 异构设计的架构师 Raja Koduri 宣布从英特尔辞职，创办自己的生成式 AI 公司。Raja Koduri 在 2017 年 11 月离开 AMD 后加入英特尔，曾是英特尔在 2021 年成立的 AXG 事业群（Accelerated Computing Systems and Graphics，加速计算与图像处理）中加速计算部门的主负责人。2022 年 12 月，AXG 重组分别并入 CCG 和 DCAI 后，Koduri 成为英特尔的首席架构师，公司当时称 Koduri 会聚焦高性能计算技术，肩负 AI 架构责任，例如“集成不同类型的芯片”。Koduri 离开后，Jeff McVeigh 担任拆分重组后的 AXG 临时总负责人，同月，McVeigh 在日志中宣布了 Falcon 为下一代 GPU（而非 XPU）的消息。

仅 GPU 的 Falcon Shores 将结合 Gaudi 系列 ASIC 产品的部分 IP(例如网络接口的设计)。因此,在最新的路线图上,Falcon Shores 的位置从 CPU 与 GPU 之间转移到 GPU 与 Gaudi 之间。其实，英特尔对 AI 芯片的多元布局由来已久，可以通过梳理其收购历史观察发展脉络：2015 年，英特尔以 167 亿美元收购 FPGA 制造商 Altera，试图以 FPGA 作为加速器，跟 CPU 搭配做 AI 推理；2016 年，公司又以 4.08 亿美元收购 AI 芯片初创公司 Nervana。英特尔当时想通过 Nervana 的 ASIC 芯片 Spring Crest NNP-T 实现 AI 产品，但在 2020 年 1 月，英特尔宣布结束四年来的 Nervana ASIC 项目，转而对其在 2019 年 12 月以 20 亿美元收购的 Habana Labs 全力押注 AI。2019 年 4 月，英特尔又收购了又一家 FPGA 初创公司 Omnitek。

目前来看，在这些以 AI 芯片为目标的收购中，Habana Labs 应该更能为英特尔提供有效的 AI 收益。英特尔目前在 GPU 产品方面暂时掉队，但 ASIC 产品方面的 Gaudi 2 和 Gaudi 3 或能有效填补了 2025 年 Falcon Shores 推出前的空白时间。ASIC 芯片 Habana Gaudi 系列：Gaudi 2 目前正在出货，而公司计划将在 2024 年推出 Gaudi 3，英特尔希望以 CPU+Gaudi 作为加速器的配合主攻 AI 训练和推理。值得一提的是，英特尔也拥有自己的软件生态 oneAPI，但其同样也难敌 CUDA 的根深蒂固。

图表58：ISC2022 英特尔将 Falcon Shores 规划在 CPU 与 GPU 之间，ISC2023 移动至 Gaudi 系列和 GPU 之间



资料来源：英特尔官网，华泰研究

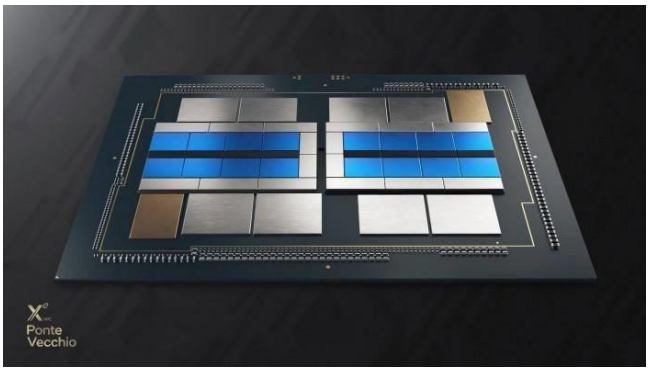
英特尔的 AI 相关 GPU 和 ASIC 芯片

1）Ponte Vecchio GPU：在 2021 年发布，但 2023 年一季度才推出。Ponte Vecchio GPU 结构复杂，一共有 47 个功能片，分 5 个制程。英特尔在 2021 年 8 月 19 日的 Architecture Day 宣布 Ponte Vecchio GPU（Xe HPC）的计算层采用了台积电 N5 工艺，基底采用了 Intel 7（对标台积电 7nm），Xe Link I/O 是台积电 N7，另外 Rambo Cache 采用的是 Intel 7，对比 H100 和 MI300 的台积电 N5 制程。晶体管数量超 1000 亿个，高于 H100 的 800 亿，但低于 MI300X 的 1530 亿。内存属 HBM2e，落后于 H100 和 MI300X 的 HBM3。

2）Falcon Shores GPU：Falcon Shores 原定为 XPU（即 CPU+GPU 架构）产品，并准备于 2024 年推出，但目前改为纯 GPU 架构，推迟到 2025 年推出。反观，英伟达的 GH200 和 AMD 的 MI300A 也属 CPU+GPU 架构的 AI 芯片。目前 Falcon Shores 的产品参数细节还未完全公布，已知道有 288GB 的 HBM3 和 9.8TB/s 的内存带宽，并能支持较低的数据精度，如 BF16 和 FP8。

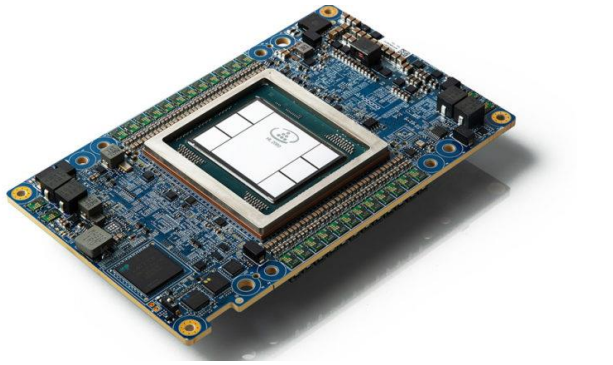
3）Habana Gaudi ASIC：Habana Gaudi 是由英特尔在 2019 年 12 月，以 20 亿美元收购的 Habana Labs 设计的 ASIC 芯片。第一款 Gaudi（16nm）于 2019 年 6 月推出，目前已迭代至 Habana Gaudi 2（7nm），在 2022 年末已推出。**Gaudi 2 的架构特点是异构**，包含 2 个 MME（Matrix Multiplication Engine，矩阵乘法引擎）和 24 个 TPC（Tensor Processor Core，张量处理核），前者负责处理所有可以转换成矩阵运算的任务，例如卷积、GEMM（General Matrix Multiplication）等，后者处理其他类型的运算。这两种计算引擎可以并行使用，因此两种类型的运算可以重叠进行，这也是 Gaudi2 可有效提高运行 AI 模型速度的原因。Gaudi 主要用于 AI 训练端，并适用于配合公司的 CPU 一同使用，提升 CPU 在 AI 的处理效果。下一代的 5nm Gaudi 3，计划在 2024 年推出。

图表59： 英特尔 Ponte Vecchio GPU



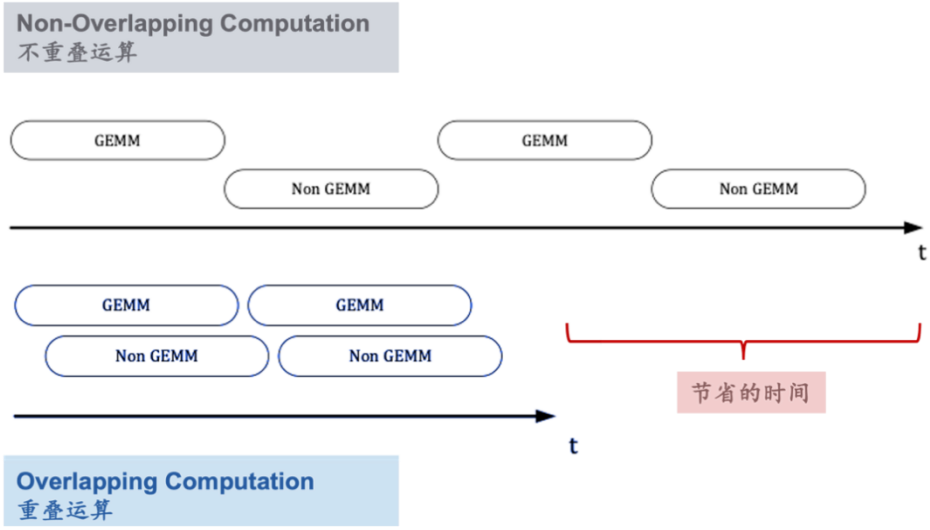
资料来源：英特尔官网、华泰研究

图表60： 英特尔 Habana Gaudi 2



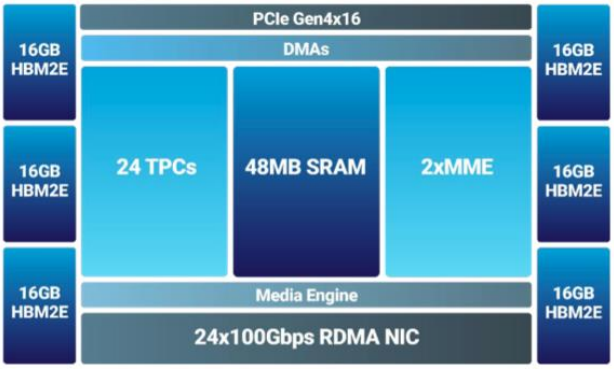
资料来源：英特尔官网，华泰研究

图表61： MME 与 TPC 异构可以使运算重叠，显著加速工作



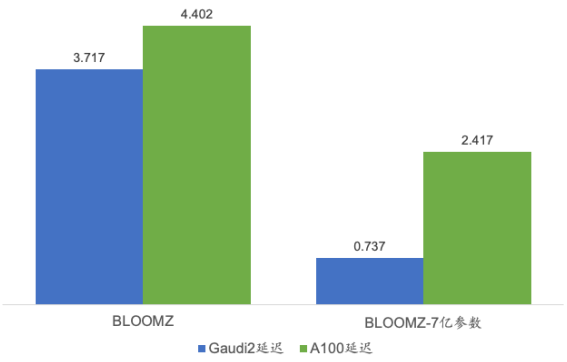
资料来源：Habana Labs，华泰研究

图表62： Habana Gaudi2 产品架构图



资料来源：Habana Labs，华泰研究

图表63： Habana Gaudi2 与 A100 的推理延迟基准测试（单位：秒）

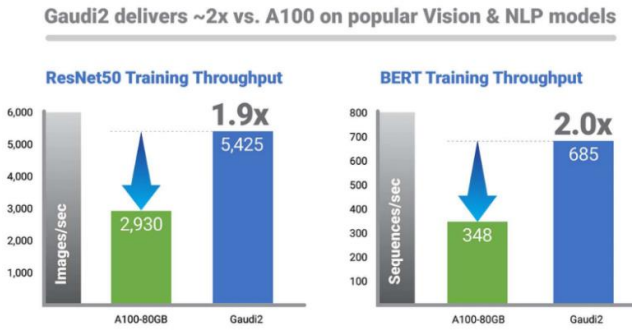


资料来源：Hugging Face，华泰研究

ASIC 在特定领域性能具备优势已有先例

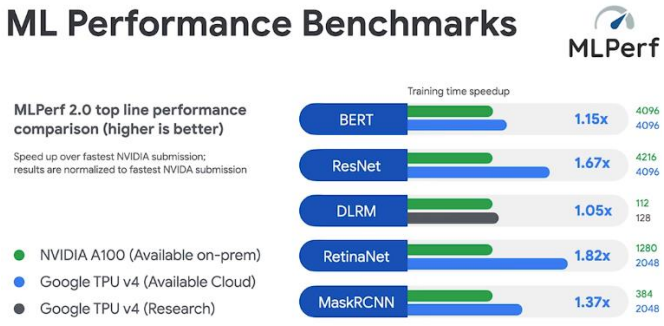
受益于架构特点，Gaudi 2 在一些基准测试里表现较 A100 优秀：1）推理端：HuggingFace 在 2023 年 3 月对 Habana Gaudi 2 与 A100 进行了大模型（BLOOMZ）推理的基准测试。BLOOMZ 是一个 1760 亿参数的文本生成模型，推理延迟测试的结果显示，Gaudi 2 比 A100 快 1.2 倍；还针对小参数版本的 BLOOMZ-7 进行了测试，在 70 亿参数的 BLOOMZ-7 模型推理中，Gaudi 2 比 A100 快 3 倍；**2）训练端：**Habana Labs 对 Habana Gaudi 2 与 A100 进行了基准测试，运行了两款芯片在 RestNet50 和 BERT 模型训练的测试，称其训练吞吐量可达到英伟达 A100 GPU 的两倍。

图表64：Habana Gaudi 2 比英伟达 A100 快 2 倍



资料来源：Habana Labs，华泰研究

图表65：谷歌 TPU 同样能与英伟达 A100 一战



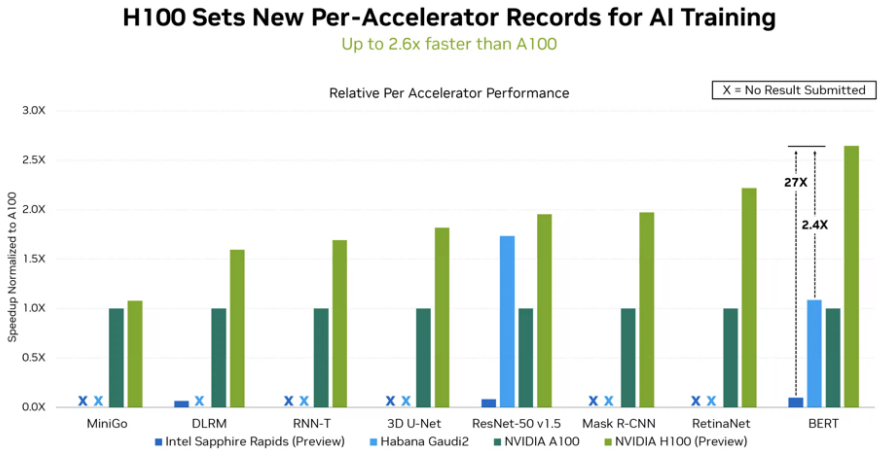
资料来源：谷歌，华泰研究

另外，同为 ASIC 的谷歌的 TPU（已发展到第五代），在架构和性能上也不断迭代。根据谷歌对 TPUv4 和 A100 的对比，其在 BERT 和 ResNet 模型上较 A100 的性能表现分别高 1.15 倍和 1.67 倍。**然而，想要与英伟达在训练端匹敌，Gaudi 还要面对 H100。**英伟达在 2022 年 11 月对 H100 和 Gaudi 2 进行了对比，在 ResNet 模型上，H100 比 Gaudi 2 快 1.1 倍左右，而在 BERT 模型上，H100 比 Gaudi 2 快 2.4 倍。

我们需强调，目前 ASIC 仍然面临通用性和灵活性问题，以及开发成本较高等局限，目前还未能大规模抢占 GPU 在 AI 训练市场的份额。ASIC 芯片的固定成本较高，因此使用它的公司或机构必需拥有大规模计算需求部署量才能分摊成本。同时，ASIC 开发周期较长，在目前变化多端的人工智能应用里，或会出现硬件开发跟不上算法迭代的情况。

但我们也重申，ASIC 的优点在于当算法开始稳定和成熟，其在一定程度上能承接部分算力。ASIC 也具备专用性、低能耗等优势，在计算量足够的情况下是一个能有效降本增效的合理选择。在具体应用中，可考虑 CPU 与 ASIC 同时部署的方案。如今年 5 月英特尔宣布和 BCG（波士顿咨询）的一项合作，就使用了 CPU+ASIC 的组合解决方案。BCG（未上市）将结合使用 Xeon 以及 Habana Gaudi 系列产品，为 BCG 以自己在咨询行业多年来的海量文件数据训练出的人工智能模型提供算力支持。类似的行业应用场景广阔，尤其是在行业内积累了大量数据，希望使用 AI 来赋能这些数据的企业。

图表66：英伟达 H100 比 Gaudi2 在 BERT 模型上快 2.4 倍

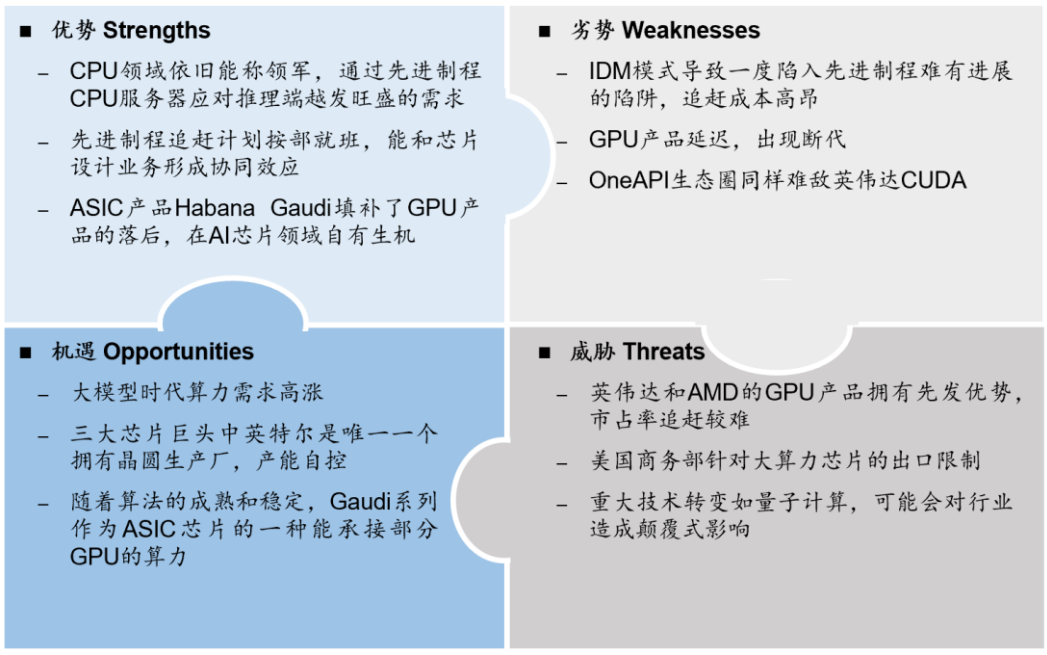


资料来源：英伟达，华泰研究

英特尔 AI 芯片 SWOT 分析

英特尔目前在制程上仍落后于竞争对手。在 AI 芯片方面，GPU 产品将出现真空期，但收购 Habana Labs 后会为英特尔带来 Gaudi 系列的 ASIC 芯片，从而在 AI 芯片领域中再获生机。叠加自身作为老牌 CPU 龙头厂商，可通过服务器 CPU+ASIC 加速器来满足大量的 AI 训练和推理端算力需求。

图表67：英特尔 AI 芯片业务 SWOT 分析



资料来源：华泰研究

云计算和互联网大厂：或许是传统芯片厂商的最大竞争对手

我们认为，大型云计算供应商及互联网巨头拥有财力物力，面对较高的外购成本和内部研发 IP 保密等因素，选择自己设计 AI 专用芯片并非意外，也许是 AMD 和英伟达的高算力 GPU 共同的竞争对手。云厂商的目标是降低 TCO（total cost of ownership，总拥有成本），因此我们认为他们具备较高意愿去自研芯片。不过，云厂商自研芯片除了需具备较强的自身研发能力外，也需符合其他条件，包括拥有较多自有的软件生态和应用，鉴于自研和设计定制芯片的固定成本也不低，需要足够的计算需求去摊分成本，而大型云厂商和互联网巨头内部就能产生足够规模的芯片需求。另外，算法也需较为成熟和稳定。一个典型的例子就是挖掘虚拟货币的芯片，鉴于挖掘虚拟货币需要大量芯片和算力，且能耗高，算法稳定并不断重复做同样的计算，因此设计专用的 ASIC 芯片是非常合适。各大云厂商和互联网巨头的 AI 工作负载不仅来自训练大模型和后续推理，还包括信息流推荐、广告排名等 AI 深度学习和 Transformer 算法。

目前，云厂商和互联网巨头们的自研 AI 芯片进度不同：1）**谷歌**在人工智能领域有多年布局，其 TPU 是市场上少数能与英伟达 GPU 匹敌的 AI 芯片；2）**亚马逊**已在训练端和推理端双管齐下，Trainium 和 Inferentia 已上 AWS 云供客户使用；3）**微软和 Meta** 则依然在较大程度上依赖英伟达、AMD 和英特尔的芯片，但二者同样不甘落后，微软“闭门造芯”Athena 已是公开的秘密，而 Meta 的 MTIAv1 则时间较迟，公司预计将于 2025 年问世。

图表68：海外云大厂和互联网巨头的自研芯片

	云厂商	芯片	发布时间	代际	工艺制程	设计	类型
AI芯片	amazon	Trainium	2022	1	5 nm	In-house	Training
	amazon	Inferentia	2019	2	5 nm	In-house	Inference
	Google	Tensor(TPU)	2015(internal) 2017(public)	4	7 nm	Co-designed with Broadcom	Training and Inference
	Meta	MTIA v1	As soon as 2025	1	7 nm	In-house†	Inference
	Microsoft	Athena	As soon as 2024	1	5 nm	In-house†	Training and Inference
CPU	amazon	Graviton	2018	3	5 nm	In-house†	CPU
	Google	Maple	As soon as 2025	1	5 nm	Marvell Technology†	CPU
	Google	Cypress	As soon as 2025	1	5 nm	In-house†	CPU
	Microsoft	Cascade	As soon as 2024	1	5 nm	In-house	CPU

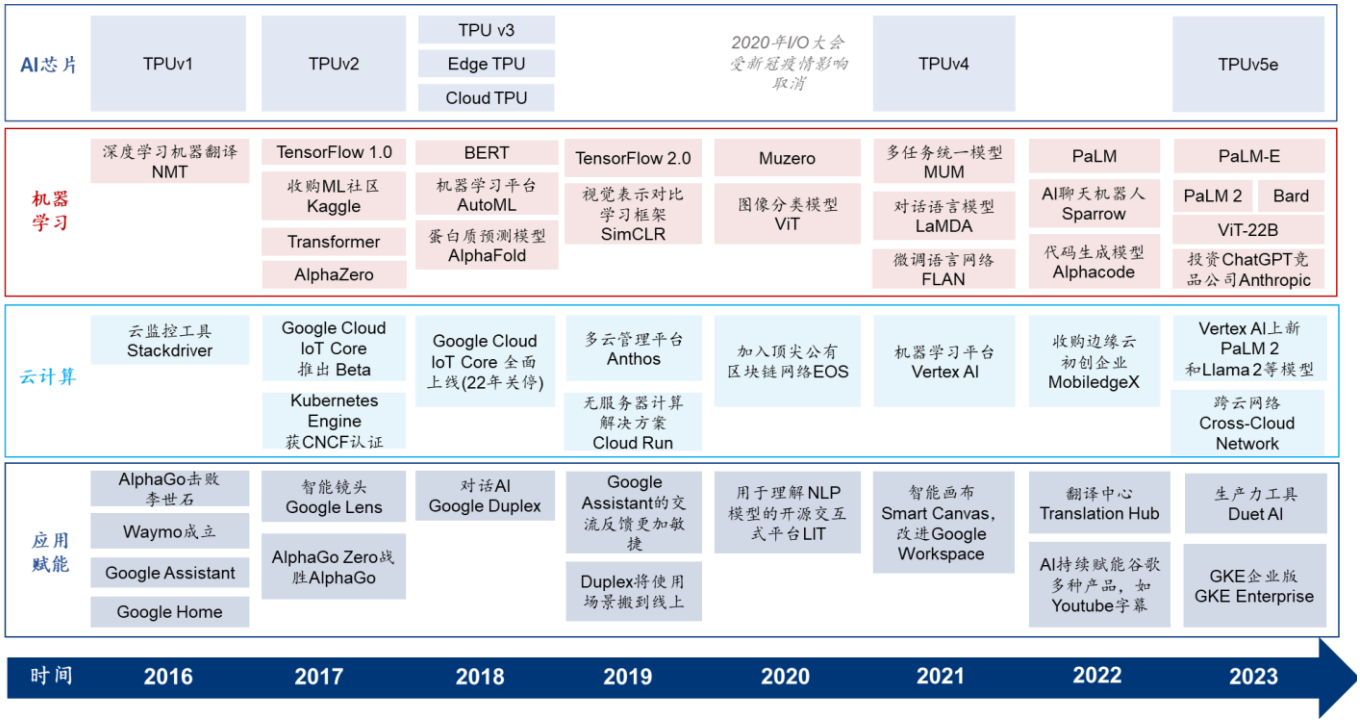
注：†指 Manufactured by TSMC
资料来源：The Information 官网、各公司官网、华泰研究

谷歌 TPU：少数能与英伟达高算力 GPU 匹敌的 AI 芯片

云厂商自研 AI 芯片的典型例子是谷歌的 TPU（Tensor Processing Unit，从 2017 年开始已具备训练和推理能力）。**谷歌 TPU 是少数能与英伟达高算力 GPU 匹敌的 AI 芯片。**

谷歌 TPU 在架构与性能参数上不断迭代。第一代 TPU 从 2015 年开始被使用于谷歌云计算数据中心的机器学习应用中，彼时仅面向推理端，但从 2017 年推出第二代开始，TPU 已同时拥有训练和推理能力。第三代 TPU 于 2018 年发布，旨在提高性能和能效以满足不断增长的机器学习任务需求。第四代 TPU 于 2021 年发布，而专为中大规模训练和推理而构建的 TPUv5e 于 2023 年发布。与 TPU v4 相比，TPU v5e 可为大语言模型提供高达 2 倍的训练性能和 2.5 倍的推理性能，并能节约一半以上的成本。谷歌目前仅通过谷歌云服务平台向外部客户提供 TPU 的算力租赁服务，而未有将其作为硬件产品出售。

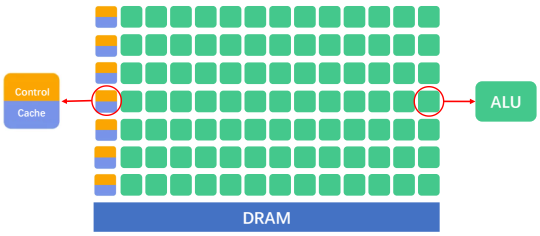
图表69：2016 年至今谷歌云计算、AI 芯片、机器学习及 AI 应用赋能进程梳理

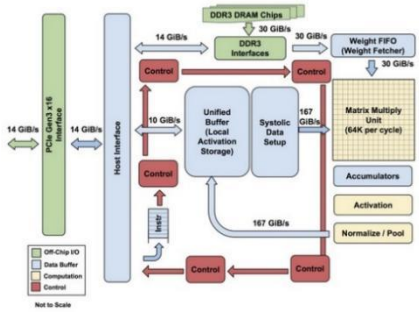


资料来源：谷歌官网、华泰研究

值得强调的是 TPU 属于定制化 ASIC 芯片，是专门针对谷歌自身的开源深度学习框架，TensorFlow 和 Jax 量身打造并全面优化的集成芯片，在此二框架下能发挥出较高运行效率，因此其他学习框架在 TPU 上运行的效率或不及 TensorFlow。

图表70：GPU 与 TPU 优缺点对比





	GPU	TPU
优点	**1. 灵活性和通用性：**GPU 可处理多种任务，包括图形渲染、模拟和科学计算等，以及机器学习工作负载。 **2.生态圈成熟：**GPU 已被广泛应用于深度学习，围绕它构建了丰富的软件和工具生态系统，如 CUDA。 **3. 精度：**GPU 提供从低精度 FP16 到高精度 FP64 等各种选项，使它适用于具有不同精度要求的工作负载。	**1. 性能：**TPU 专为张量运算而设计，因此在特定情况下，神经网络的训练和推理效率更高。 **2. 能效：**从单元上看，TPU 或比 GPU 更节能。 **3. 集成性：**TPU 与一些的机器学习框架（如 TensorFlow）集成在一起，因此在一同使用下效率或更高。
缺点	**1. 能耗：**从 per unit 角度看，GPU 或会比 TPU 高。 **2. 成本：**从 per unit 角度看，高性能 GPU 价格或较高，但从 TCO 来看或更节省。	1. **生态系统：**英伟达 GPU CUDA 优势较为明显，其他芯片较难与之相比 **2. 必须要在谷歌云上使用：**TPU 主要通过 Google Cloud Platform 提供，芯片没有对外售卖。

资料来源：谷歌、CSDN、Openmetal、华泰研究

谷歌 TPU 的应用可以分为两类，分别是对内承担 AI 工作负载和其他应用，以及对外在云服务上提供给客户。对内：根据路透社 4 月 5 日的报道，目前谷歌内部 90%的 AI 工作负载都使用 TPU，例如在拥有 5400 亿参数的 PaLM 模型训练中，就使用了 6144 个 TPU v4；今年的 PaLM2 也是采用 TPU v4 来训练。除 AI 工作负载之外，TPU 也已广泛应用于翻译、相册、搜索、Google 助理和 Gmail 等众多 Google 产品。**对外：**谷歌云服务同时提供 GPU 和 TPU 给客户选择。谷歌云上 TPU 相比 GPU 的价格：A100 80GB 价格为 3.93 美元/芯片/小时，TPU v4 价格为 3.22 美元/芯片/小时；但 TPU 的应用也在一定程度上受到英伟达 CUDA 生态圈一家独大的影响。

谷歌云上的 TPU 相比 GPU 的价格虽较有优势，但在一定程度上受到英伟达 CUDA 生态圈一家独大的影响。谷歌云作为 AI 云服务商，需积极满足有 AI 训练和推理需求的客户，而英伟达 GPU 拥有生态圈成熟和开发者众多的 CUDA，是目前大部分 AI 训练所必需的工具。**总的来说:1)谷歌的 TPU 或其他云厂商的自研芯片不会在一夜之间取代所有英伟达的 GPU；2）若算法已相对成熟，可使用 TensorFlow 框架编程并在 TPU 上运行，可有效利用其优化和加速，节省成本，是性价比较高的选择；3）面对英伟达 CUDA 的成熟生态圈，云厂商自研芯片无需以完全取代作为目标，而仅需为客户提供更多算力选择即可有效打开市场。**

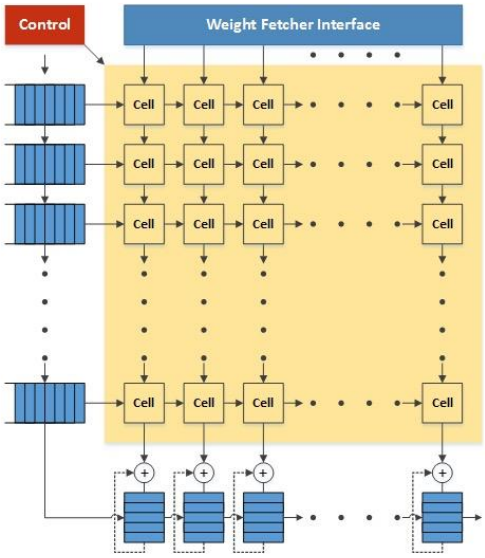
TensorFlow 在深度学习里的生态虽成熟，使用者众，但其他机器学习库如 PyTorch 的开发者或也有自己的偏好：深度学习框架的主流为 Meta 开发的 PyTorch 和谷歌开发的 TensorFlow。2015 年，谷歌宣布开源 TensorFlow。2016 年 11 月发布了 TensorFlow 的第一个正式版本，主要基于 Python 和 C++，推出后很快超过 Torch、Theano 和 Caffe 等一众框架。2016 年，Meta 发布 PyTorch，也是基于 Python 和 C++等机器学习热门编程语言，由于上手简单，很快受到开发者青睐。**另外，谷歌大脑（Google Brain）团队于 2017 年开发了 JAX，提供 TPU 和 GPU 都能使用的深度学习软件库。**基于 Jax 构建的软件库包括谷歌大脑的 Trax、Flax、Jax-md 及 DeepMind 的神经网络库 Haiku 和强化学习库 RLax 等。

TPU 针对矩阵乘法进行专门的优化

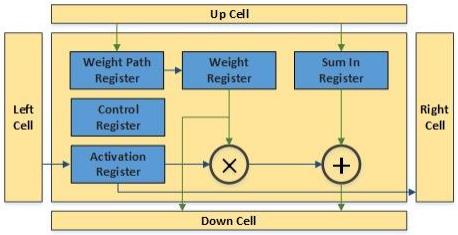
矩阵乘法是机器学习中非常关键的计算步骤，目前各类大模型基于的 Transformer 是注意力机制，涉及大量的矩阵运算，特别是矩阵乘法，而 GPU 非常擅长于矩阵运算。**与通用计算 GPU 相比，TPU 也是针对矩阵乘法进行了专门的优化**，采用脉动阵列（Systolic Array），对大规模矩阵乘法可以最大化数据复用，减少访存次数，大幅提升机器学习的训练速度，同时节约训练成本。在 TPUv4 中有 2 个 TensorCore，每个 TensorCore 中有 4 个 MXU（矩阵乘法单元），MXU 采用的是脉动阵列。典型的脉动阵列将数据转为向量形式输入 MXU，并进行矩阵乘法得到结果。TPU v5e，相对于 TPU v5（未发布），是以 e 为后缀的高效版本，因此从架构复杂度看跟 TPUv4 相差不大。TPU v5e 的算力为 393 TOPS（int8），包含 1 个 TensorCore，而每个 TensorCore 中有 4 个 MXU（矩阵乘法单元），这点跟 TPUv4 一样。每个 TPU 可通过芯片间互连 (ICI) 以 400Gbps 的速度连接到另外四个 TPU（最多支持 256 个芯片互连），因此，单片 TPU v5e 的累积带宽为 1.6T。TPU v5e Pod 是由 256 个 TPU v5e 芯片组成，总带宽超过 400 Tb/s，每秒可提供高达 100 PetaOps（int8）。

图表71：脉动阵列卷积加速示意图

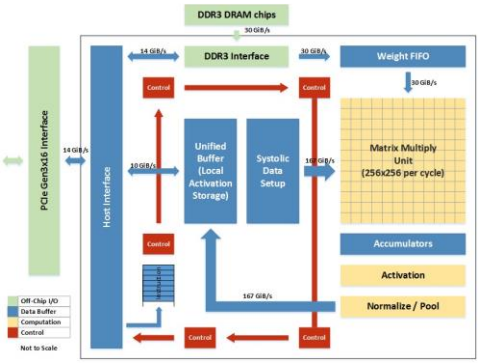
脉动阵列卷积加速示意图



MXU：典型的脉动阵列



TPU结构示意图



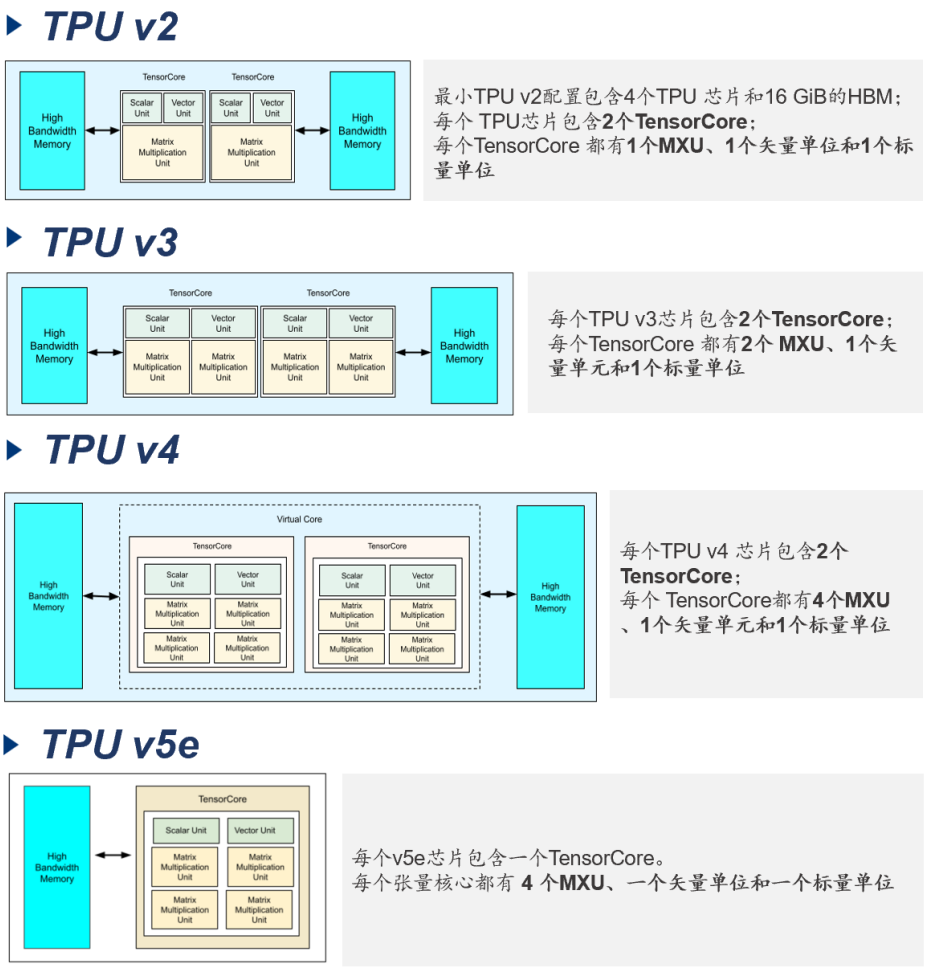
资料来源：谷歌官网、CSDN（《脉动阵列：因 Google TPU 获得新生》）、华泰研究

图表72：五代 TPU 性能对比

	TPUv1	TPUv2	TPUv3	TPUv4	TPUv5e
发布年份	2016	2017	2018	2021	2023
每颗芯片的峰值计算能力（TFLOPS）	92 （int8）	46 （bf16）	123 （bf16）	275 （bf16 or int8）	197 TFLOP（bf16）
HBM2 容量与带宽	28 GiB, 34 GB/s	32 GiB, 700 GB/s	32 GiB, 900 GB/s	32 GiB, 1200 GB/s	16 GB， 819 GBps
最小/平均/最大测量功耗 （W）	-	-	123/220/262	90/170/192	-
TPU Pod 规模（芯片数量）	-	256	1024	4096	256
互连拓扑结构	-	2D torus	2D torus	3D torus	2D torus
每个 Pod 的峰值计算能力（PFLOPS）	-	12 （bf16）	126 （bf16）	1100 （bf16 or int8）	-
每个 Pod 的 All-reduce 带宽（TB/s）	-	120	340	1126.4	-
每个 Pod 的切分带宽 （TB/s）	-	2	6.4	24	-
目标应用场景	仅推理端	训练&推理端	训练&推理端	训练&推理端	训练&推理端

资料来源：谷歌官网、nextplatform 官网、华泰研究

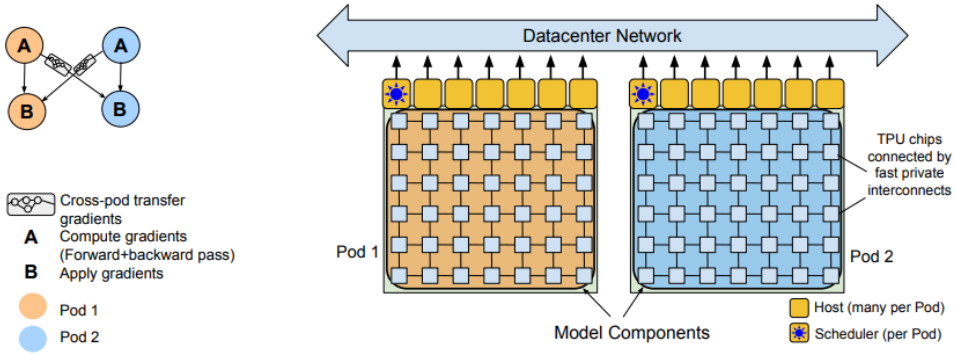
图表73：TPU v2/v3/v4/v5e 对比图



资料来源：谷歌官网、华泰研究

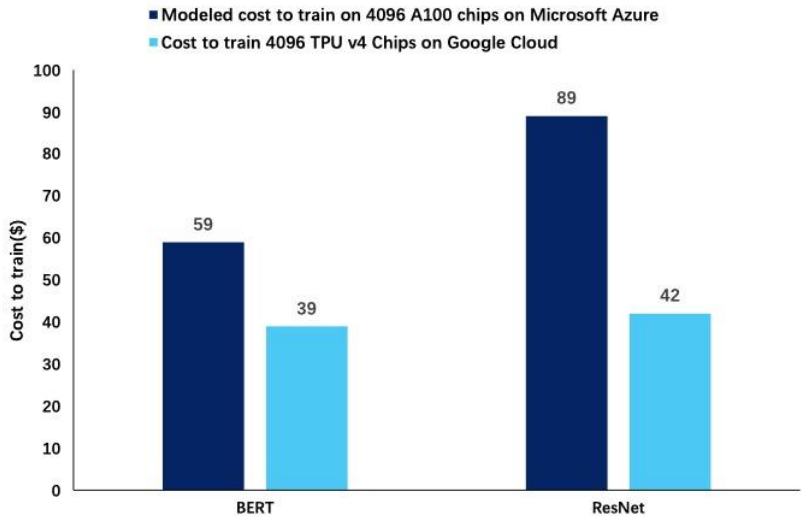
谷歌在内部使用 GPU 的主要场景之一是为 TPU 进行基准测试。2023 年 4 月 4 日谷歌发布关于 TPU v4 的论文《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》中就提到，以 A100 为基准做了对 TPUv4 的测试。根据论文，在相似芯片规模的系统中，TPU v4 在 BERT 模型上比 A100 快 1.15 倍，而在 ResNet 模型上比 A100 快 1.67 倍；在 MLPerf 3.0 基准测试上的功耗使用情况，A100 能耗为 TPUv4 的 1.3-1.9 倍。与微软 Azure 中布置的 A100 相比，谷歌云的 TPU v4 在 BERT 上最多可节省 35%的训练成本，在 ResNet 上最多可节省 50%。

图表74：谷歌使用 TPUv4 训练 5400 亿参数的 PaLM 模型



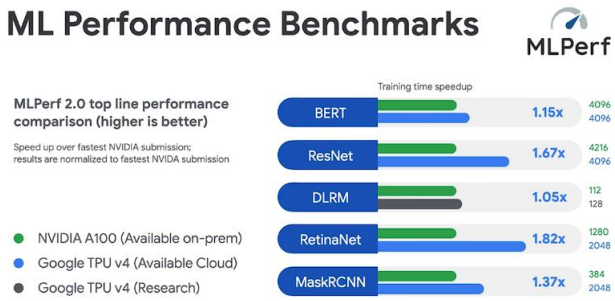
资料来源：Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint arXiv:2204.02311, 2022、华泰研究

图表75：BERT 模型中使用 TPU v4 相比 A100 能节省成本



注：具体计算方法及假设见谷歌云 Blog
https://cloud.google.com/blog/products/ai-machine-learning/cloud-tpu-v4-mlperf-2-0-results?hl=en
资料来源：谷歌官网、华泰研究

图表76：论文展示-TPU v4 在不同机器学习任务上性能优于 A100



资料来源：谷歌官网、华泰研究

图表77：论文展示-TPU 对 FLOPS 利用率高达 46.2%

Chip feature	# of Parameters (in billions)	Accelerator chips	Model FLOPS utilization
GPT-3	175B	V100	21.3%
Gopher	280B	4096 TPU v3	30.8%
Megatron-Turing NLG	530B	2240 A100	30.2%
PaLM	540B	6144 TPU v4	46.2%

资料来源：谷歌官网、华泰研究

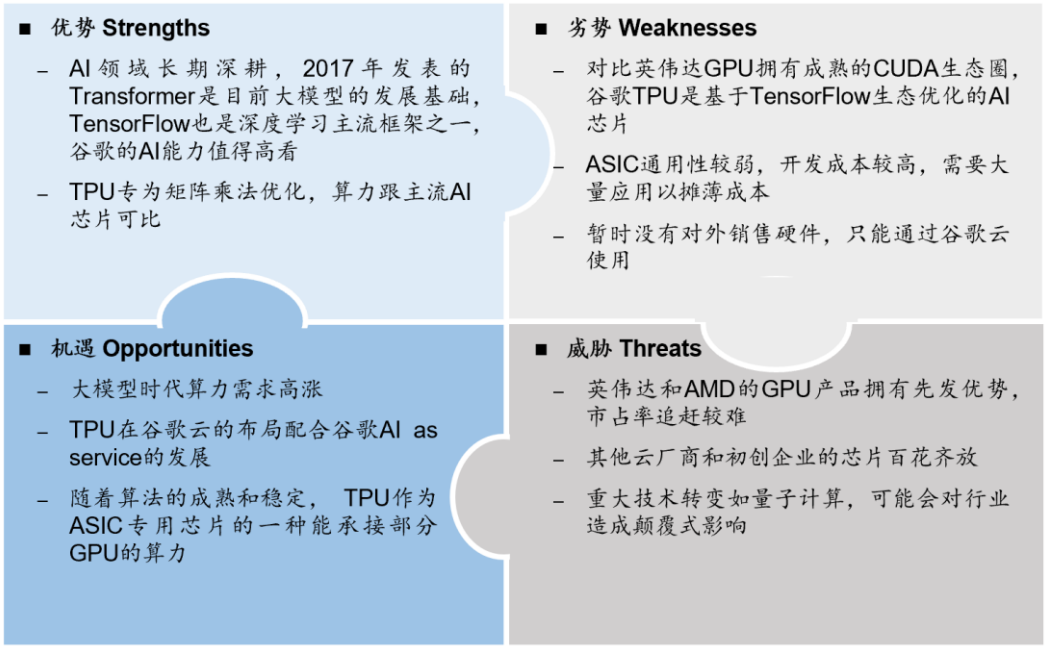
谷歌云上提供英伟达的 GPU 包括 A100、V100、P100、L4、P4、T4 等，多种训练和推理芯片供客户选择，进行机器学习、科学计算和 3D 可视化等应用。最近的使用案例包括在 2023 年 I/O 大会上发布的 AI 超级计算机 A3，每台配备 8 个英伟达 H100，面向需要训练大语言模型的谷歌云客户，最高可提供 26000 片 H100 的算力。A3 虚拟机的训练速度是上一代 A2 的 3 倍，网络带宽达 A2 的 10 倍，并支持大规模拓展。在 2023 年 8 月的 Google Cloud Next'23 大会上，谷歌正式宣布 A3 将于 2023 年 9 月全面上市。推理端方面，谷歌发布了 G2 VM，是业界首款由英伟达推理芯片 L4 Tensor Core GPU 提供支持的云 VM。

谷歌 TPU 的云上客户包括许多人工智能初创企业，如：1）著名的 AI 文本到图像生成软件 Midjourney(未上市)，2023 年 3 月宣布正在使用 TPUv4 训练 Midjourney 的第四代大模型；2）AI 生命科技公司 InstaDeep（未上市），2022 年 11 月宣布使用 512 片 TPUv4 成功训练了基因组学大模型；3）微软创始人 Paul Allen 的 AI 研究所 Allen Institute for AI（AI2，未上市）；4）为企业提供 NLP（自然语言处理）模型的 Cohere（创始人来自谷歌大脑）。

谷歌 TPU SWOT 分析

谷歌作为云大厂中自研芯片的先行者和 AI 领域的奠基者之一，其自研 TPU 针对矩阵乘法优化，适合训练大模型，且具备价格优势，但 ASIC 存在通用性较弱等问题。因此，我们认为 TPU 或其他云大厂自研芯片不会取代所有英伟达的 GPU，但若算法已相对成熟，设计 ASIC 去取代部分算力则挺合适。

图表78： 谷歌 TPU SWOT 分析

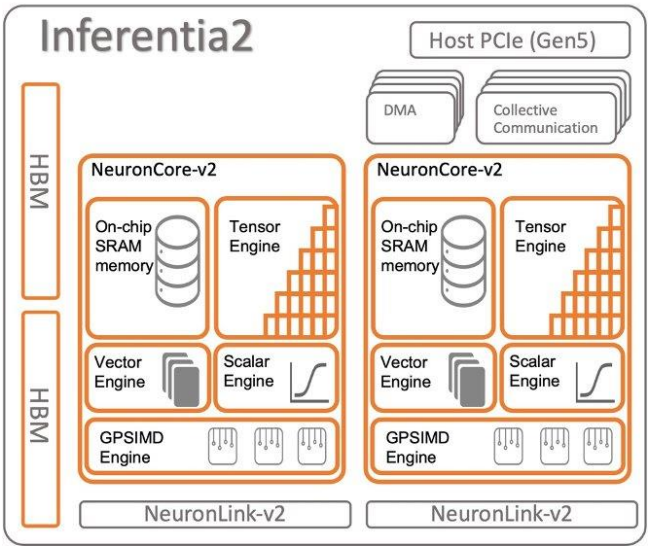


资料来源：华泰研究

亚马逊 AWS：Trainium & Inferentia，训练推理双管齐下

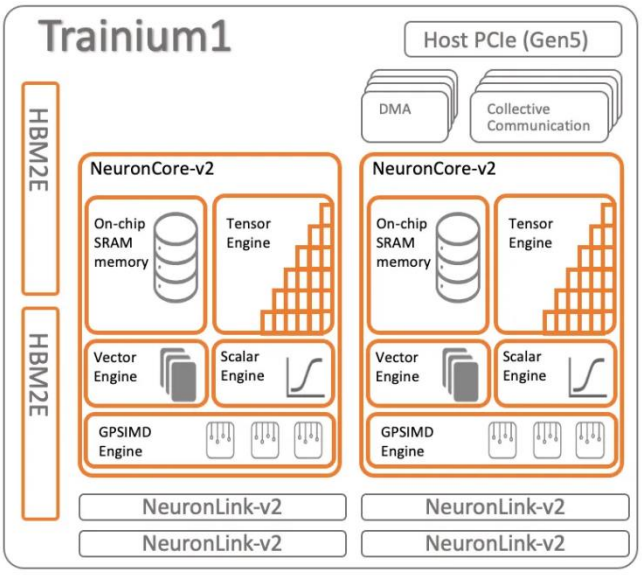
AWS 分别在 2018 和 2020 年发布 AI 推理芯片 Inferentia 以及训练芯片 Trainium，随后在 2023 年推出 Inferentia 2，并在 AWS 云上提供给客户使用。2015 年，亚马逊收购了以色列芯片设计公司 Annapuma Labs，开启了其自研 AI 芯片的脚步，2018 年的第一代 Inferentia 就出自它手。AWS 的 AI 芯片搭配有 AWS Neuron 开发软件包，其中包含里可用于兼容 TensorFlow 和 PyTorch 的编译器。2023 年 5 月，亚马逊表示计划将其自研大语言模型“Alexa Teacher Model”（AlexaTM）接入智能语音助手 Alexa。Alexa 此前已接入亚马逊 Echo 智能音箱等智能硬件设备，且使用 Inferentia 进行推理。

图表79： AWS Inferentia2 芯片架构



资料来源：AWS 官网、华泰研究

图表80： AWS Trainium 芯片架构



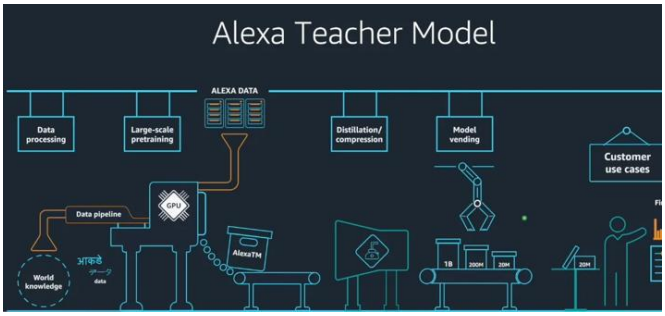
资料来源：AWS 官网、华泰研究

Trainium 在云端训练表现或较 A100 好，性价比也较高。Trainium 是 AWS 为超过 1000 亿参数规模的大模型打造的 AI 芯片，2020 年发布，目前仍处于第一代。每一个 Trainium 配备容量 32GB 带宽 820GB/s 的 HBM2e，FP16 算力 190 TFLOPS（英伟达 A100 为 624 TFLOPS），FP32 算力 47.5 TFLOPS，支持包括可配置的 FP8 在内的多种数据精度。其使用的互联技术为 AWS 的 NeuronLink（超高速非阻塞互连技术，v2 代），互联速度达到 768 GB/s，对比 NVLink 4.0 互联速度为 900GB/s。据 AWS 官网，Trainium 实例内存比英伟达 A100 实例内存容量高 60%，互联带宽高 2 倍，在 130 个 Trainium 实例上训练 GPT-3 只需要 2 周，而据英伟达与微软论文《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》使用 1024 颗 A100 进行训练需要 34 天。2022 年，AWS 推出的 Trn1 AI 平台可以部署最多 16 个 Trainium，在 AWS 云上进行 AI 模型训练。相较于同类型的 Amazon EC2 实例，以 Trainium 为支撑的 Trn1 实例节约 50%训练成本，在亚马逊广告模型训练中更是将节约成本高至 70%。

图表81：130 片 Trainium 需要两周就可以完成 GPT-3 训练

资料来源：AWS 官网、华泰研究

图表82：AWS 的 in-house 大语言模型 Alexa Teacher Model

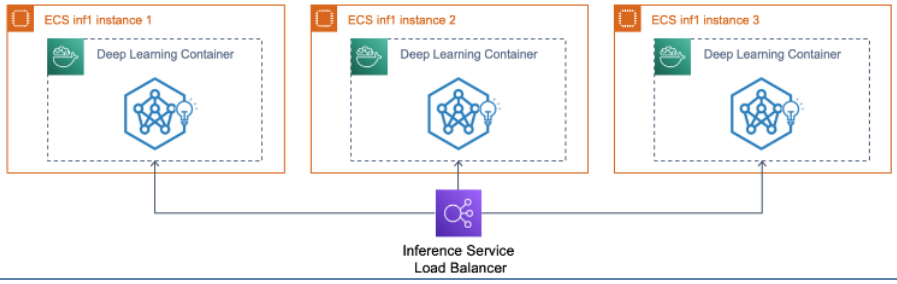


资料来源：AWS 官网、华泰研究

推理卡 Inferentia 已迭代两代，用于亚马逊硬件终端 AI 服务。2018 年推出的第一代 Inferentia 配备 8 GB 带宽为 50GB/s 的 DDR4 内存，于 2023 年 4 月正式推出的第二代 Inferentia 2 配备 32 GB 带宽为 820GB/s 的 HBM2e 内存，FP16 算力达到 190 TFLOPS，相比一代 Inferentia（64 TFLOPS）提高 2 倍，主要为高性能深度学习推理应用程序而设计。据亚马逊官网，相比第一代延迟降低至前者的 1/10，吞吐量提高了 4 倍。由于大规模终端设备 AI 模型对云端推理能力要求较高，而自研 AI 芯片等信息基础设施和自身应用可针对性地进行相互适配与优化，Amazon 人工智能助手 Alexa 使用以 Inferentia 为支撑的 Inf 实例进行推理负载。

与 AWS 生态捆绑，借助云服务进行推理卡客群逐步渗透。除此之外，客户可通过开发工具包 AWS Neuron，通过 Amazon Sagemaker（AWS 机器学习平台）、Amazon Elastic Container Service （ECS，AWS 容器托管方案）、Amazon Elastic Kubernetes Service （EKS）等服务快速开始使用 Inf 和 Trn 实例，分别使用底层 Inferentia 和 Trainium 芯片能力。目前 AWS 上使用 Inferentia 承担推理工作负载的客户包括 Airbnb（爱彼迎，房屋租赁平台，ABNB US）、Snap（图片类社交媒体平台，SNAP US）、Sprinklr（SCRM 社交媒体营销公司，客户包括麦当劳、耐克、微软等，CXM US）、Money Forward（金融科技公司，3994 JT）、Finch Computing（AI 初创公司，主要为政府机构等设计 AI 应用程序，未上市）等；使用 Inferentia2 的客户包括 Hugging Face（机器学习公司，主要项目为 BLOOM 模型库）、Qualtrics（自动化管理软件公司，客户包括法拉利、阿迪达斯等，未上市）、Finch Computing（亦为 Inf1 客户）等。

图表83：使用 inf1 实例将推理服务部署到 AWS ECS 容器托管集群



资料来源：AWS 官网、华泰研究

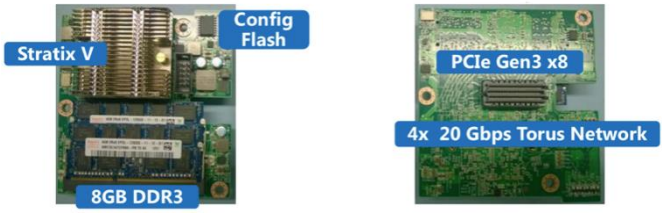
微软：“闭门造芯”Athena

微软早在 2010 年开始已希望自研 AI 硬件，当时以外采 FPGA 然后自己搭建运算平台为主。FPGA 的方案最早由微软的计算机架构研究员、FPGA 专家 Doug Burger 提出。据半导体行业观察梳理，微软自研 FPGA 的第一阶段采用了单板多 FPGA 的方案，即每块加速卡上集成 6 片 Xilinx Virtix-6 FPGA，各 FPGA 之间通过自身的通用 I/O 端口相连和通信，但整体稳定性存在隐患，若一块 FPGA 出问题，整个板卡都会出问题。第二阶段从单板多 FPGA，变成了单板单 FPGA。但是，这种方案为实现 FPGA 之间的低延时通信，FPGA 之间的互联需要通过两类特殊定制的线缆，成本高昂且实现难度较高。第三阶段为解决互联问题，取消了 FPGA 互连的网络，直接将 FPGA 与数据中心网络互连，提出了 HaaS（硬件即服务）的概念。2017 年起，微软宣布其已在 Azure 和 Office 365 中扩展了 FPGA 的使用，且采取外部采购 Intel 和 Xilinx 等厂商的 FPGA 进行每一台新 Azure 服务器的搭建。

图表84：第一阶段：单板多 FPGA 方案

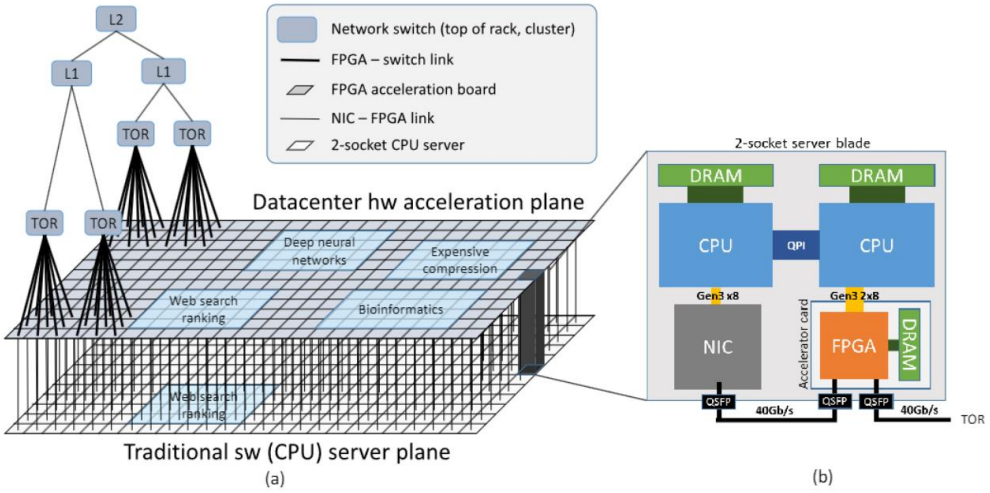
资料来源：微软官网、半导体行业观察（Shilicon）、华泰研究

图表85：第二阶段：单板单 FPGA 方案



资料来源：微软官网、半导体行业观察（Shilicon）、华泰研究

图表86：第三阶段：Catapult FPGA 与数据中心网络紧密连接

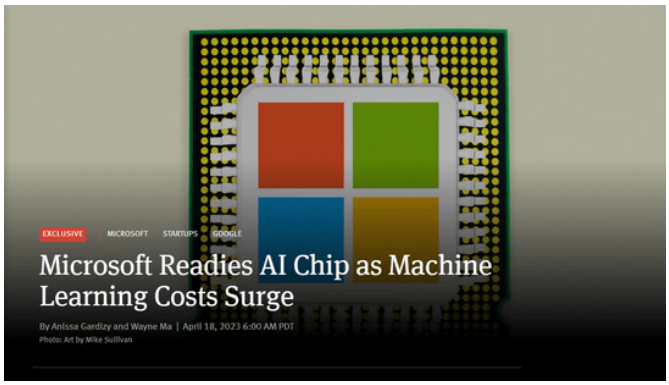


资料来源：《A Cloud-Scale Acceleration Architecture》Microsoft Corporation、华泰研究

目前，微软云 Azure 已开始自研 AI 芯片，芯片代号为 Athena。The Information 2023 年 4 月 18 日的报道称，微软至少从 2019 年开始研发代号为“Athena”的新芯片组。据 Tom’s Hardware 2023 年 4 月 18 日消息，Athena 使用台积电 5nm 制程，专门为大语言模型训练设计。据彭博社 5 月 4 日报道，微软将注资 AMD 并开展合作，目前合作研发的微软 AI 芯片即为 Athena，希望为 ChatGPT 等大语言模型的训练及推理提供英伟达芯片以外的替代方案。随后 5 月 5 日，微软发言人 Frank Shaw 表示 AMD 参与“Athena”项目的报道不实，但却并未明确微软与 AMD 的合作关系。事实上，微软和 AMD MI 系列早有合作。2022 年 5 月，微软宣布 Azure 将成为首个部署 AMD MI200 系列 GPU 进行大规模 AI 加速的公有云。AMD Instinct MI200 加速器将协同其他 AMD 产品共同被用于微软 Azure 上，包括全新的使用 AMD 3D V-Cache 技术第三代 AMD EPYC 处理器的 Azure HBv3 虚拟机。此外，微软还宣布正在与 AMD 数据中心软件团队以及 PyTorch Core 团队密切合作，以确保 PyTorch 项目开发者可充分利用 AMD Instinct 加速器的性能与功能。目前，官方仍未透露关于 Athena 的具体架构及参数信息。

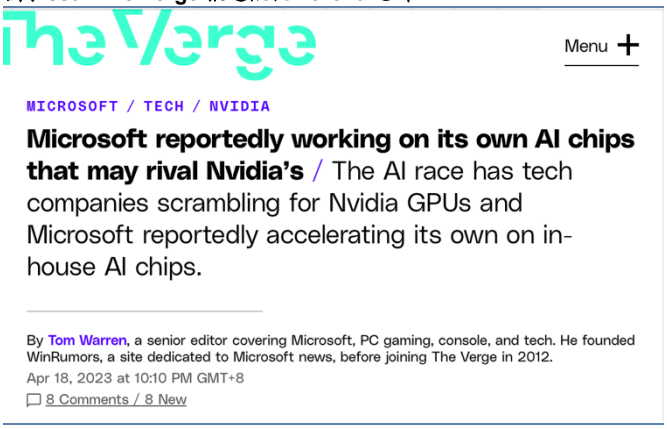
我们认为，大型云计算供应商拥有财力物力，面对在 AI 应用需求激增下较高的外购成本和较有限的供应，选择自己研发 AI 芯片也是无可厚非，且微软与 OpenAI 的合作中用到大量的英伟达芯片。我们认为云计算大厂也希望出现一个二供选择。

图表87： The Information 报道微软 Athena 芯片



资料来源：The Information、华泰研究

图表88： The Verge 报道微软 Athena 芯片



资料来源：The Verge、华泰研究

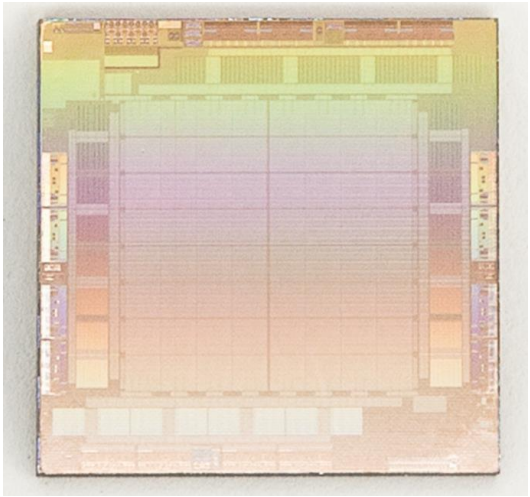
Meta：首个自研推理端芯片 MTIA 将于 2025 年问世

Meta 在 2023 年 5 月发布了主要针对推理工作的自研 AI 芯片 MTIA（Meta Training and Inference Accelerator）。此芯片自 2020 年开始设计，公司预计于 2025 年正式推出，采用台积电 7nm 制程。MTIAv1 是针对推理端的产品，使用最高 128GB 的 LPDDR5 内存，RISC-V 架构，配合基于 PyTorch 的软件包。如其他云厂商自研 AI 芯片一样，MTIA 也是针对公司内部应用和模型量身定做的 ASIC，尤其是针对 Meta 旗下产品所需要的 feed（信息流页面，如 Instagram 的用户浏览界面）贴文推荐算法进行了优化，相比通用芯片能实现降本增效。

图表89：MTIA 产品实物图

资料来源：Meta 官网、华泰研究

图表90：Meta MTIA v1（主要针对推理端）芯片产品图



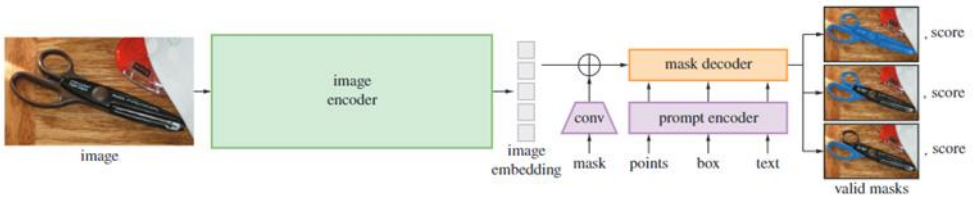
资料来源：Meta 官网、华泰研究

Meta 的超级计算机由约 16,000 片英伟达 A100 构成，已用于 LLaMA 模型训练。目前，Meta 还没有推出专用于训练阶段的芯片。Meta 的 AI 超级计算机 RSC（Research Super Cluster，研究超级集群）由约 16,000 片英伟达 A100 构成（2000 台英伟达 DGX A100），通过 NVIDIA Quantum InfiniBand 16 Tb/s 结构网络连接。Meta 表示其使用 RSC（除此外还包括由 A100 组成的内部生产集群）来训练其在 2023 年 2 月发布的 70-650 亿参数的开源大模型 LLaMA，其中 650 亿参数的 LLaMA 模型在 2048 片英伟达 A100 上花费 21 天预训练完成。2023 年 7 月，Meta 发布免费可商用版本 LLaMA2，与一代相比，LLaMA2 作为其升级版本包括 70 亿、130 亿和 700 亿三个参数版本，使用 1.4 倍容量的数据集，并采用了分组查询注意力机制，同样使用 RSC 工作负载进行预训练。据 Meta 评估，多项测评结果显示 LLaMA 2 在推理、精通性、编码和知识测试等诸多外部基准测试中均优于其他开源语言模型。

和谷歌 TPU 相比，Meta MITA 不具先发优势，RSC 超算性能或被谷歌压制。对比来说，谷歌 TPU 和 Meta MTIA 同为互联网厂商自研的 ASIC 芯片。目前 TPU 已用于训练 PaLM-2 和 Imagen 等大型生成式 AI 模型，而 MTIA 落地情况仍然未披露。谷歌 2022 年发表的论文《PaLM: Scaling Language Modeling with Pathways》显示，初代 PALM 模型（5400 亿参数）在 6144 颗 TPU v4 芯片上进行了 1200 小时（50 天）的训练，并分散到 3072 颗 TPU v4 芯片上进行了 336 小时的训练。另外，谷歌 AI 超级计算机 A3 目前拥有 26,000 个 Nvidia H100 GPU，在算力（Meta RSC 可高至 5 exaFLOPS，谷歌 A3 可高至 26 exaFLOPS）及规模上均领先于 Meta RSC。

Meta AI 团队：重振旗鼓角逐生成式 AI 赛场。Meta AI 团队前身为 2013 年成立的 Facebook AI 研究院，由 Yann LeCun 带领（2018 年图灵奖得主），多年来持续投入资金支撑包括社交媒体人脸识别与自动标注功能在内的 AI 工作研发。继 2022 年 8 月 Meta 对话机器人 BlenderBot 3 受迫于主流媒体“政治不正确”舆论压力而关停，11 月 Meta 大语言模型 Galactica 上线 48 小时后因散布错误信息被迫下线。屡屡受挫之后，根据路透社 5 月 9 日消息，Meta 于 2023 年初收购了一支位于挪威奥斯陆的团队，该团队在英国芯片独角兽 Graphcore 开发 AI 网络技术（AI networking technology），自此 Meta 开始其在生成式 AI 领域奋起直追的步伐。2023 年 2 月，Meta 开源其大模型 LLaMA，首席执行官马克·扎克伯格宣布将成立一个由 Ahmad Al-Dahle（Meta 生成式 AI 事业群副总裁，曾于苹果工作 16 年）引领的新生成式人工智能团队，将专注于开发能够以包括文本（WhatsApp 和 Messenger 中聊天数据）、图像（创意 Instagram 滤镜和广告格式）、视频和多模式等在内的多种方式协助用户的 AI 角色。此后，Meta 4 月发布 SAM 图像分割模型，5 月发布其自研芯片 MTIA，并于 7 月开放 LLaMA 商业版本。Meta 发布的 MTIA 及其开源贡献的历史或表明该公司致力于推进 AI 软件及硬件研究的决心。

图表91：SAM 模型总览示意图



资料来源：《Segment Anything》Meta AI Research、华泰研究

异军突起者：晶圆级芯片持续突破性能极限，内存和传输成破局关键

AI 大模型对训练端的算力提出了更高要求，新兴初创 AI 芯片企业（如 Cerebras、Graphcore 等），以及芯片行业以外的企业，包括特斯拉（TSLA US）等，正在异军突起，试图在芯片设计上另辟蹊径，**通过大尺寸晶圆级芯片内存共享和低延时的技术路线突破 AI 芯片瓶颈**，试图在持续上升的算力需求中抢占份额。

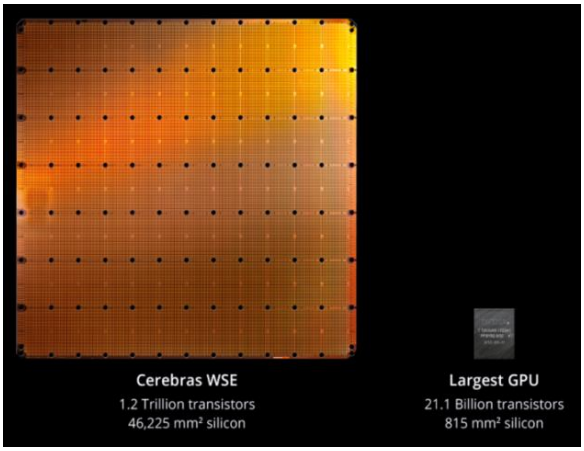
我们认为晶圆级芯片的优势在于其提出了有效应对 AI 应用中内存和传输速度的瓶颈的措施。具体来说：1）大尺寸晶圆内核比传统芯片上的互连更加紧密，提升内核之间以及跟内存之间的通信速度，降低延迟并提供更快传输速度；2）晶圆级芯片所有内存都保留在芯片上，而不必考虑片外内存速度缓慢，内核可直接访问整个晶圆级芯片的全局共享内存，突破内存带宽瓶颈。**但晶圆级芯片由于尺寸大，因此或存在良率低、功耗高、框架定制化需求高等问题。**低良率问题直接增加制造成本，同时影响芯片后续量产和质量。虽然目前基本是通过增加冗余面积及内核数量的方式以绕过制造缺陷，但仍较难完全解决良率与性能之间的平衡。大尺寸芯片的设计也会导致功率增加，出现高耗能和散热问题，影响稳定性，目前解决方案主要通过增加液冷、风冷和水冷散热模块。我们认为，晶圆级芯片在内核、SRAM、内存带宽、晶体管几方面的表现，或能吸引到 B 端的高性能计算行业成为其主要客户。**但短期来看，晶圆级芯片的有效运行面积与晶圆面积之比例仍未可知。长期来看，我们认为晶圆级芯片的瓶颈若能给突破，将能对传统技术路径发起重大挑战。**

Cerebras：向晶圆级大尺寸芯片迈出第一步，但良率和有效运行占比暂成疑

Cerebras Systems（未上市）成立于 2015 年，是一家美国人工智能芯片初创企业，由 Andrew Feldman 等五位前 SeaMicro 员工创立，目前在硅谷、圣地亚哥、多伦多、东京、和班加罗尔设有办公室。SeaMicro（未上市）成立于 2007 年，是一家小型低功率计算机服务器厂商，2012 年被 AMD 以 3.34 亿美元收购，2015 年开始停止运营。

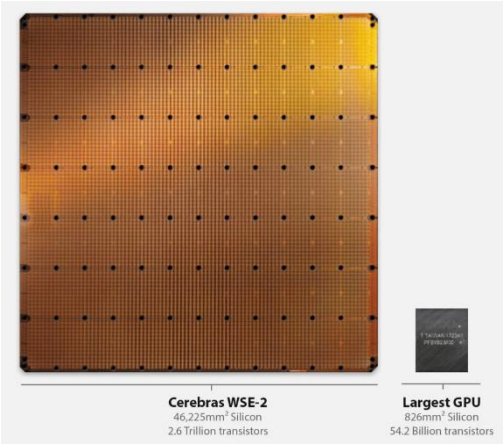
2019 年 8 月，Cerebras Systems 发布第一代 WSE（Wafer-Scale Engine，晶圆级引擎）芯片。到了 2021 年 4 月，Cerebras Systems 推出为超级计算机而打造的 WSE-2，在第一代的基础上进行了优化。WSE 与传统芯片最大的不同在于整片晶圆不进行切割。而在芯片结构上，WSE 对内存和横向扩展也采用了独特的设计。单块 WSE-1 拥有 1.2 万亿个晶体管，采用台积电 16nm 制程工艺，面积 46,225 平方毫米，包含 40 万个内核，片上存储 18G，功耗 1.5 万瓦，内存带宽 9PB/秒，通信结构带宽 100PB/秒。WSE-1 的内核为专为 AI 训练设计的可编程计算内核，即稀疏线性代数（Sparse Linear Algebra, SLA™）内核。对比当时英伟达的旗舰 GPU Titan V，其采用 12nm 制程，晶体管数为 211 亿，12GB HBM2 内存，内存带宽 652.8GB/秒，最大功耗 250W，包含 5120 个 CUDA 内核以及 640 个 Tensor 内核，而 WSE-1 晶体管数量是其 57 倍，内核数是其 70 倍，功耗是其 60 倍。换言之，WSE-1 单个晶圆芯片或能达到约 60 个 GPU 的性能水平。

图表92： WSE-1 对比当时英伟达的旗舰 GPU Titan V



资料来源：Cerebras 官网、华泰研究

图表93： WSE-2 对比当时英伟达的旗舰 A100



资料来源：Cerebras 官网、华泰研究

单块 WSE-2 在面积上没有明显变化，但内核、晶体管、内存带宽、通信结构带宽，对比第一代都有 2 倍以上的提升，主要鉴于 WSE-2 为台积电 7nm 制程工艺，并在一定程度上缩小了 SRAM，可以容纳更多内核。WSE-2 拥有 2.6 万亿个晶体管，内核数达 85 万个。WSE-2 采用 40GB SRAM 内存，可平均分配到整个晶圆芯片的表面。对比当时英伟达旗舰 GPU A100 也为台积电 7nm 制程，面积 826mm²，晶体管数量为 542 亿，内存带宽 2.039 TB/s，片上内存 80GB HBM，功耗 400W，包括 6,912 个 CUDA 内核以及 432 个 Tensor 内核。WSE-2 具备 27.5PB/s Fabric 带宽，对比 A100 采用 GPU 到 GPU 的互连所达到的 600GB/s Fabric 带宽。WSE-2 以约 200 倍的价格，达到 A100 晶体管数的 48 倍，内核数的 123 倍，内存带宽的 12733 倍，功耗的 50 倍。

图表94： WSE-1 和 WSE-2 性能指标对比

性能指标	Wafer Scale Engine 1	Wafer Scale Engine 2	英伟达 A100
发布时间	2019	2021	2020
内核	400,000	850,000	6,912 CUDA + 432 Tensor
制程	台积电 16nm	台积电 7nm	台积电 7nm
面积	46,225mm²	46,225mm²	826mm²
晶体管	1.2 万亿	2.6 万亿	542 亿
SRAM	18 GB	40 GB	40 MB
内存带宽	9 PB/s	20 PB/s	2.039 TB/s
通信带宽	100 Pb/s	220Pb/s	600 GB/s
能耗（系统/芯片）	20 kW/15 kW	20 kW/15 kW	400W
售价	约 200 万美元	约 300 万美元	约 1.5 万美元

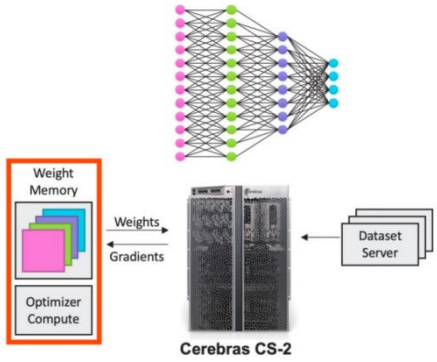
资料来源：Cerebras 官网、英伟达官网、华泰研究

Cerebras WSE-1 定价为 200-300 万美元，相当于约 200 片英伟达 A100，WSE-2 则更高。高昂的成本可能对大规模商业化造成一定阻碍，客户群的扩充势必对成本的降低提出更高的要求，而目前我们仍未看到 Cerebras 在成本方面有降低的提示。当然就目前来看，WSE 主要面向高校、研究机构以及政府等预算充足的客户，包括美国国家能源技术实验室（NETL），葛兰素史克（GlaxoSmithKline），日本东京电子器件株式会社（Tokyo Electron Devices），匹兹堡超级计算中心（the Pittsburgh Supercomputing Center），以及爱丁堡大学（the University of Edinburgh）等。

图表95： WSE-2 介绍

资料来源：Cerebras 官网、华泰研究

图表96： WSE-2 SwarmX 示意图

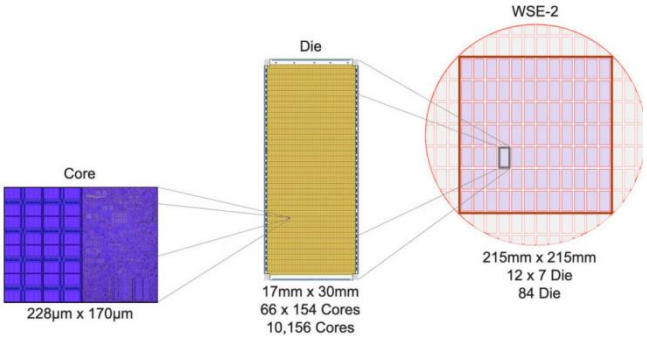


资料来源：Cerebras 官网、华泰研究

从流程上来说，WSE-2 首先在整片直径约 300 毫米（12 英寸）的晶圆上做出一个个传统晶粒（Die），每个晶粒含有约 10,000 个核心。与传统芯片不同的是，WSE-2 不将单个晶粒切割出来，而是在整片晶圆内切割出一个边长 215 毫米的方块，方块包含 84 个晶粒，共有 85 万个计算核心。我们认为，内存离计算内核越近，计算速度就越快，延迟越短，功率也越少。

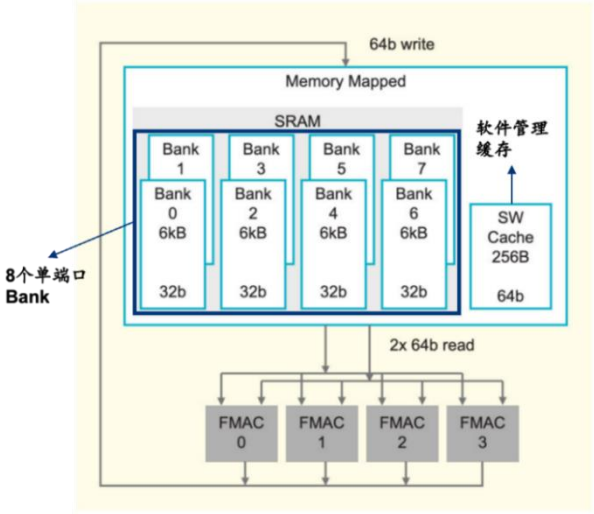
对比传统芯片架构使用共享中央 DRAM，其存取速度较慢且距离较远。目前主流技术使用中介层（interposer）和 HBM 等技术来解决这问题，但人工智能深度学习要求每个内核都在最高水平运行，使得内核和内存之间距离须尽量缩短。Cerebras 的方案试图通过在每个核心配置 48 KB 的本地 SRAM，并以 8 个 32 位宽的单端口 bank，以及 256B 的软件管理缓存（供频繁访问的数据结构使用）使其具备高密度，解决内存瓶颈问题，实现全速访问。另外，大模型的高算力对横向扩展也提出了更高的要求。WSE 架构能够以数据并行扩展的方式在单个晶圆芯片运行所有模型，内核通过为数据并行专门设计的 SwarmX 通信方式（允许模型以线性方式扩展）以网格结构连接，从而实现模块化和低成本扩展。

图表97： Cerebras WSE-2 结构示意图



资料来源：Cerebras 官网、华泰研究

图表98： Cerebras WSE-2 内存设计示意图

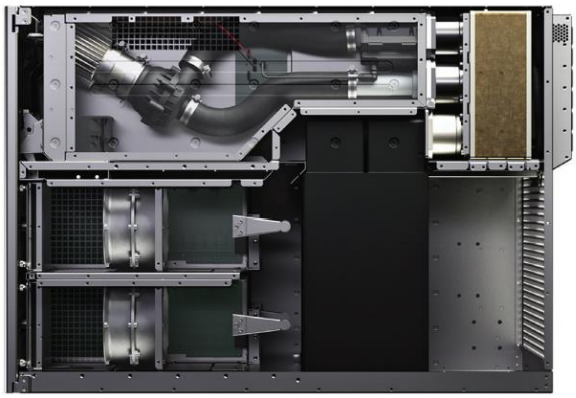


资料来源：Cerebras 官网、华泰研究

此外，公司针对性地设计了 Cerebras CS-2 系统，用单块晶圆芯片实现集群级计算。CS-2 是一个系统级解决方案，由以下三个部分组成：1）一个 WSE-2 晶圆级芯片、2）Cerebras 系统、以及 3）Cerebras 软件平台。CS-2 高 26 英寸（15 个机架单位），可容纳标准数据中心机架的三分之一。Cerebras 指出，GPU 数据中心可能需要几周或几个月的时间来设置，并需要大量超参数调优以及数十个数据中心机架，而 CS-2 仅需要几分钟即可设置完成，只需将基于标准 100 千兆以太网连接插入交换机，就可以训练模型。

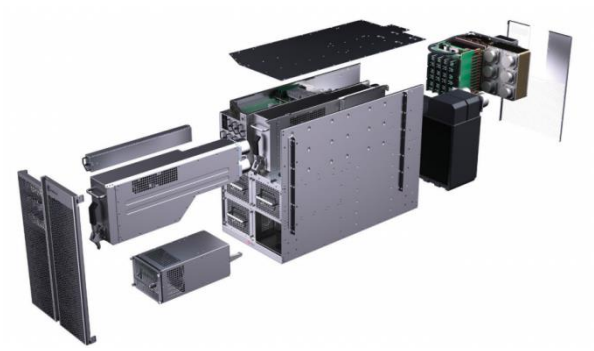
大芯片设计也无法避免高功耗带来的散热挑战。Cerebras WSE 采用液冷和风冷两套散热系统，其功耗高达 15KW，需在仅有 46,225 平方毫米的芯片上散发出来。CS-2 内置定制电源传输和冷却技术，使芯片能够在低于传统芯片工作温度下保持全速运行。CS-2 内部的一套水冷散热系统，用水冷来对 WSE-2 散热，再由风冷来降低水温。

图表99：CS-2 散热系统



资料来源：Cerebras 官网、华泰研究

图表100：CS-2 内部构成示意图



资料来源：Cerebras 官网、华泰研究

在良率提升上，晶圆级芯片尝试通过设计额外面积以及冗余内核来绕过缺陷。一般来说，在相同工艺制程下，芯片面积越小良率越能有保证。芯片制造的缺陷会随着芯片面积的增长而同时大幅增加。当蚀刻电路时，晶圆会产生一些无法修复的缺陷区域，在同样的缺陷分布下，晶圆分割的数量越少，裸片越大，缺陷的影响就越大。据 AnandTech 于 2021 年 4 月的报道，Cerebras 必须建立额外面积，并设计 1.5%的内核冗余，以提升良率。公司表示，通过台积电的高级金属层（high-level metal layers）工艺，加上 Cerebras 的训练和自动校正能力，实现高效、源同步、数据并行的晶粒对晶粒（die-to-die interface）接口，即便在制造过程中存在瑕疵，整个晶圆的结构也能做到完全均质（uniform fabric）。然而，即便 Cerebras 可能使用了 1.5%的额外内核，绕过制造缺陷来实现更高良率，但由于芯片的整体面积过大，导致晶圆级芯片最终能够运行的部分，占总面积比例或不会特别理想，但公司则未公布这一比例数字。

Cerebras 认为，使大模型普及既需要解决基础算力，也需要向开发者社区提供更多的开源。为此，他们设计了 Cerebras-GPT 开源大语言模型，并与云计算服务提供商 Cirrascale（未上市）合作建立 Cerebras AI Model Studio，在其专为深度学习所设计的 CS-2 上提供大模型训练服务。Cerebras AI Model Studio 云服务分为固定参数价格和微调模型价格，按固定参数的定价最低为在 10 小时内训练一个 13 亿参数的 GPT-3 模型，需要 2500 美元，最高为在 85 天内训练 700 亿参数的 GPT 模型，需要 250 万美元。根据 Cerebras，平均来说，用户为完成同等工作而支付的租用成本是其他竞争者的一半，基于 CS-2 的云实例训练速度是英伟达 A100 云实例的 8 倍。

图表101：Cerebras AI Model Studio 固定价格表

Model	Parameters	Tokens to Train to Chinchilla Point (B)	Cerebras Model Studio CS-2 Day to Train	Cerebras Model Studio Price to Train
GPT3-XL	1.3	26	0.4	$2,500
GPT-J	6	120	8	$45,000
GPT-3 6.7B	6.7	134	11	$40,000
T-5 11B	11	34*	9	$60,000
GPT-3 13B	13	260	39	$150,000
GPT NeoX	20	400	47	$525,000
GPT 70B	70	1,400	85	$2,500,000
GPT 175B	175	3,500	Contact For Quote	Contact For Quote

* - T5 tokens to train from the original T5 paper. Chinchilla scaling laws not applicable.

资料来源：Cirrascale、华泰研究

Cerebras 的技术背景多来自于过去 SeaMicro 的成员，在 Fabric 服务器领域积累深厚。SeaMicro 为一家从事 Fabric 服务器的公司，其早期产品 SM15000 在 2 个机架中高达 512 核的 10RU 和 5PB 的存储，也是唯一一家拥有超过 1.28 Tbps 带宽的第二代 Fabric 服务器的公司，于 2015 年被 AMD 收购。

图表102：Cerebras 技术团队背景

职位	姓名	履历
首席技术官	Gary Lauterbach	AMD 服务器 CPU 部门首席技术官； SeaMicro 联合创始人和 CTO
首席硬件架构师	Sean Lie	AMD 数据中心首席架构师； SeaMicro 首席硬件架构师（ IO virtualization fabric ASIC）
首席系统架构师	Jean-Philippe Fricker	DSSD 高级硬件架构师； SeaMicro 首席系统架构师
首席先进技术架构师	Michael James	SeaMicro 首席软件架构师（分布式系统软件）
硬件工程高级副总裁	Dhirai Mallick	英特尔数据中心业务首席技术官和架构副总裁； SeaMicro 工程副总裁
软件工程高级副总裁	Vinay Srinivas	Synopsys 仿真工程副总裁； Archpro Design Automation 研发副总裁；Sequence Design 研发副总裁

资料来源：Cerebras 官网、华泰研究

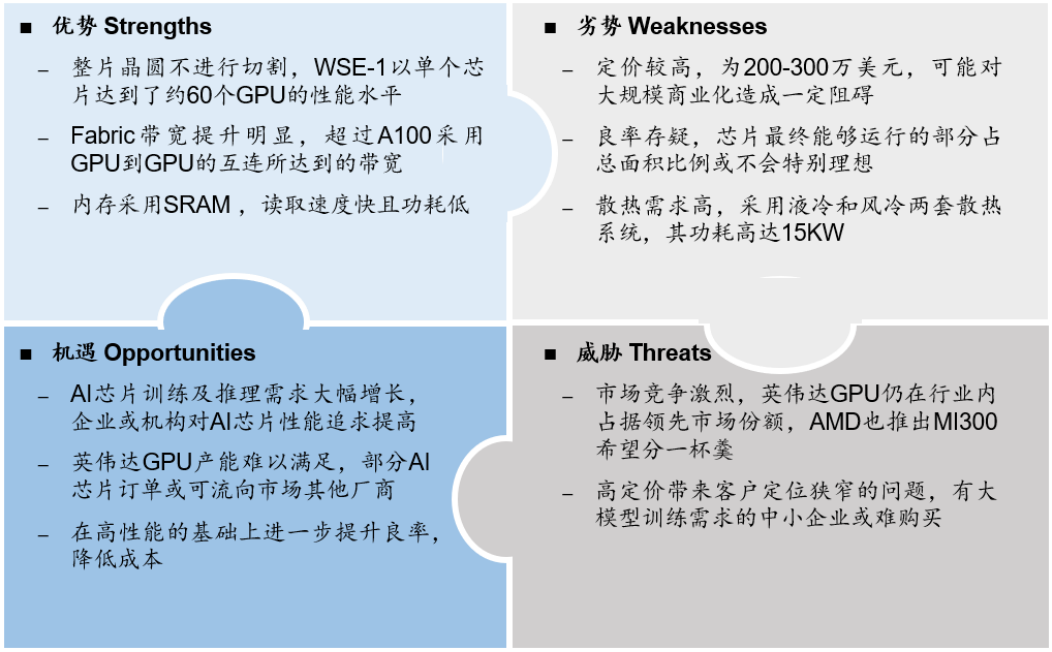
Cerebras 目前已进行六轮融资，总额达到约 7.2 亿美元。最近一轮为 2021 年 11 月进行的 F 轮融资，金额为 2.5 亿美元，由 Alpha Wave Ventures 和 Abu Dhabi Growth Fund 领投，估值超过 40 亿美元。

图表103：Cerebras 融资历程

融资时间	轮次	金额	投资者	估值
2021 年 11 月	F 轮	2.50 亿美元	Alpha Wave Ventures, Abu Dhabi Growth Fund, & G42	40+亿美元
2019 年 11 月	E 轮	2.72 亿美元	-	24 亿美元
2018 年 11 月	D 轮	8,800 万美元	Altimeter, VY Capital, Coatue, Foundation Capital, Benchmark, & Eclipse	-
2017 年 1 月	C 轮	6,000 万美元	VY Capital	-
2016 年 12 月	B 轮	2,500 万美元	Coatue Management	-
2016 年 5 月	A 轮	2,700 万美元	Benchmark, Foundation Capital & Eclipse Ventures	-

资料来源：Wikipedia、Crunchbase、华泰研究

图表104： Cerebras SWOT 分析

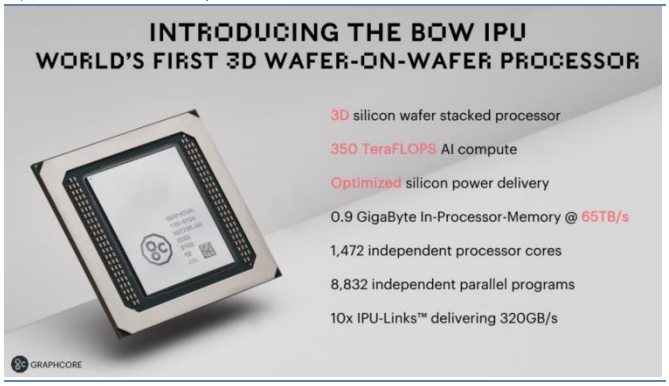


资料来源：华泰研究

Graphcore：Bow IPU 实现精细数据多指令并行

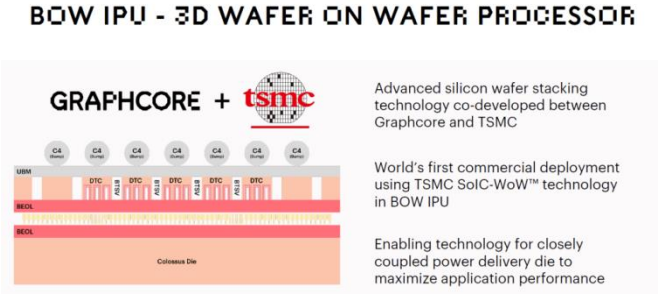
Graphcore（未上市）是一家专注于研发人工智能芯片及打造计算机系统的初创公司，2016年在英国布里斯托成立。Graphcore通过智能处理器IPU（Intelligence Processing Unit）提供满足人工智能计算的存储要求，包括低时延访问、使用非结构化数据以及管理随机与非时序数据模式。

图表105： IPU 性能介绍



资料来源：Graphcore 官网、华泰研究

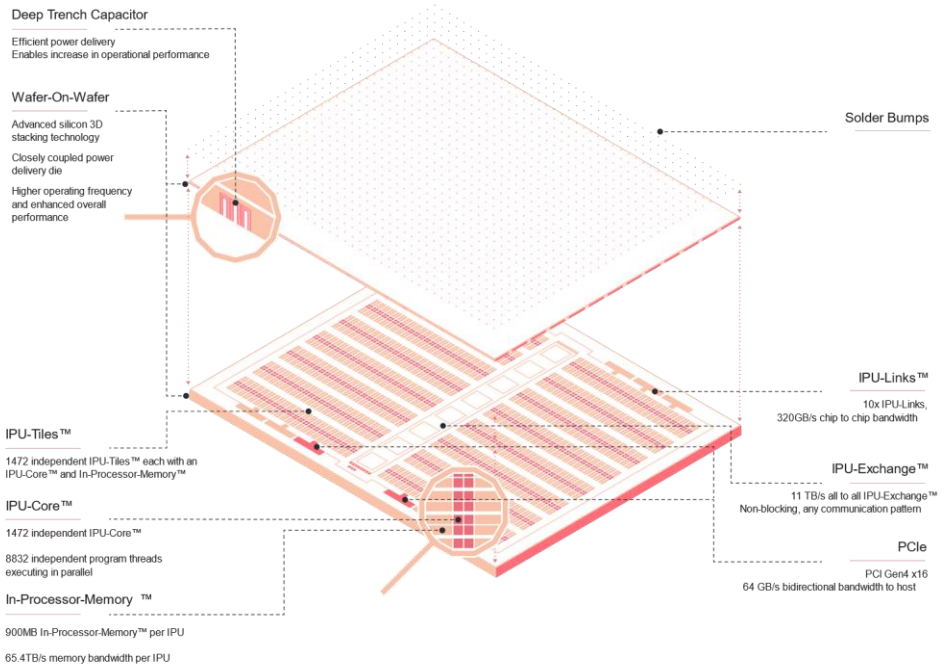
图表106： 基于台积电 SoIC-WoW 多晶圆堆叠 3D 封装技术



资料来源：Graphcore 官网、华泰研究

Graphcore 的计算系统应用在中外众多行业的人工智能中，包括科学研究、气象预测等领域。Graphcore 与阿里云（BABA US）、百度（BIDU US）、金山云（KS US）、腾讯云（00700 HK）、戴尔（DELL US）、神州数码（000034 CH）、鑫联大（大联大控股旗下，3702TT）等展开了合作。目前，Graphcore 正在努力切入中国市场，为中国大型互联网公司提供硬件服务，如百度深度学习平台飞桨正式发布在 Graphcore IPU 上实现训练和推理全流程支持的开源代码库，飞桨模型库如 ERNIE 等也逐步在 IPU 上实现支持；与腾讯共同推出腾讯云公有云上的 IPU 产品；以及支持阿里云深度学习开放接口标准。Graphcore 也发布了中国首款 IPU 开发者云，部署在金山云的数据中心。此外，微软曾与 Graphcore 达成合作，并发布 Azure 上 Graphcore IPU 的预览版。

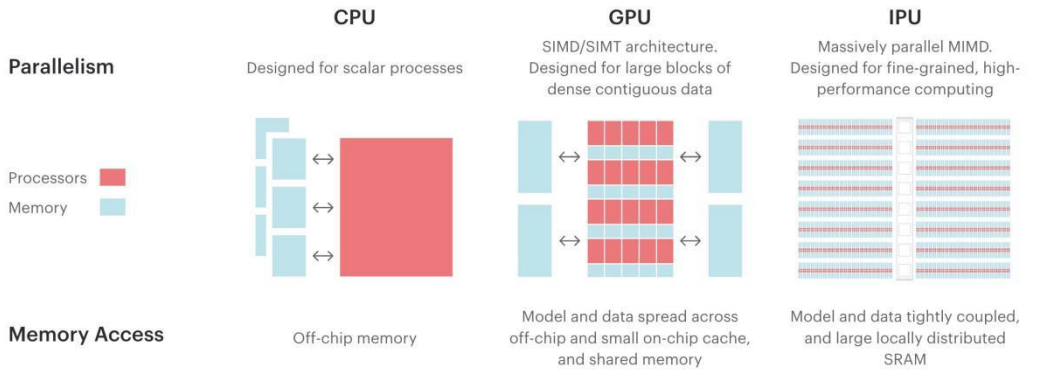
图表107：IPU 产品示意图



资料来源：Graphcore 官网、华泰研究

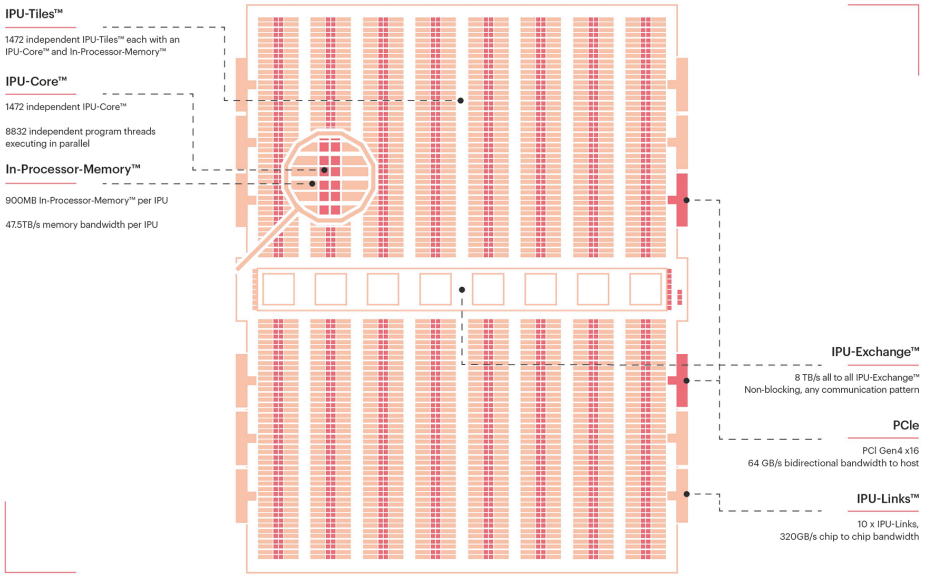
2022 年 3 月，Graphcore 推出 Bow IPU，是一种全新的大规模并行处理器，主要用于加速 AI 计算。Bow IPU 采用台积电 7nm 制程工艺，并基于台积电 SoIC-WoW（System on IC Wafer on Wafer）多晶圆堆叠 3D 封装技术。通过采用背面硅通孔 TVS（Through-silicon via）技术，在现有底部的计算晶圆上堆叠一个拥有供电裸片的晶圆，产生一个新的 3D 裸片，分别用于人工智能处理和实现降低电压的功效，使得设备的性能和能效均得到提升。Bow IPU 具备高运算能力以处理高性能计算和负载。单个封装中包含超过 600 亿个晶体管、1,472 个独立处理器内核和 8,832 个可并行执行的独立线程。FP16 算力达 350 TeraFLOPS，0.9GB 处理器内存储，带宽达 65TB/s，10x IPU-Links 达 320GB/s 芯片到芯片的传输带宽。

图表108：Graphcore IPU 产品与 CPU 和 GPU 的区别



资料来源：Graphcore 官网、华泰研究

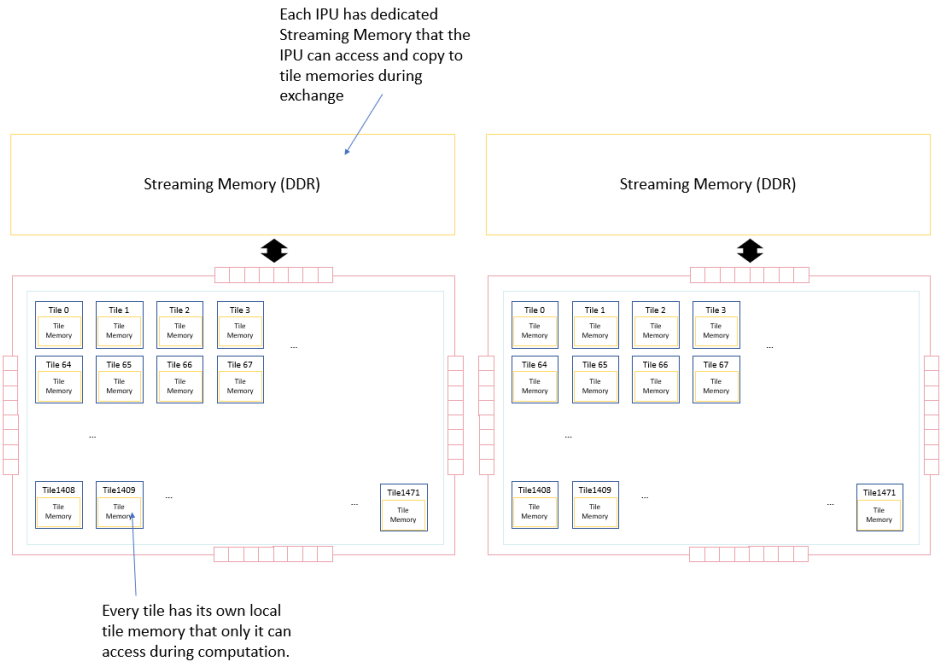
图表109： IPU 内部架构



资料来源：Graphcore 官网、华泰研究

IPU 采用 MIMD（Multiple Instruction & Multiple Data）架构，具有多指令和多数据并行的特点，适用于精细化和高性能计算，与英伟达 GPU 采用的常规 SIMT（Single Instruction Multiple Thread）架构不同。IPU 强调细粒度并行性，可以在较小的数据块上运行单独的处理线程，而每个线程有不同代码和执行流程，因此不会损失性能。**为了减少内存延迟时间，IPU 摒弃了共享内存并在架构上采用大规模分布式的片上 SRAM，存储器与每个内核紧密耦合。**SRAM 提供了比 DRAM 更高带宽（45 TB/s）和更低延迟（6 个时钟周期），主要适用于自然语言处理、计算机视觉、图神经网络等领域的人工智能及图计算。

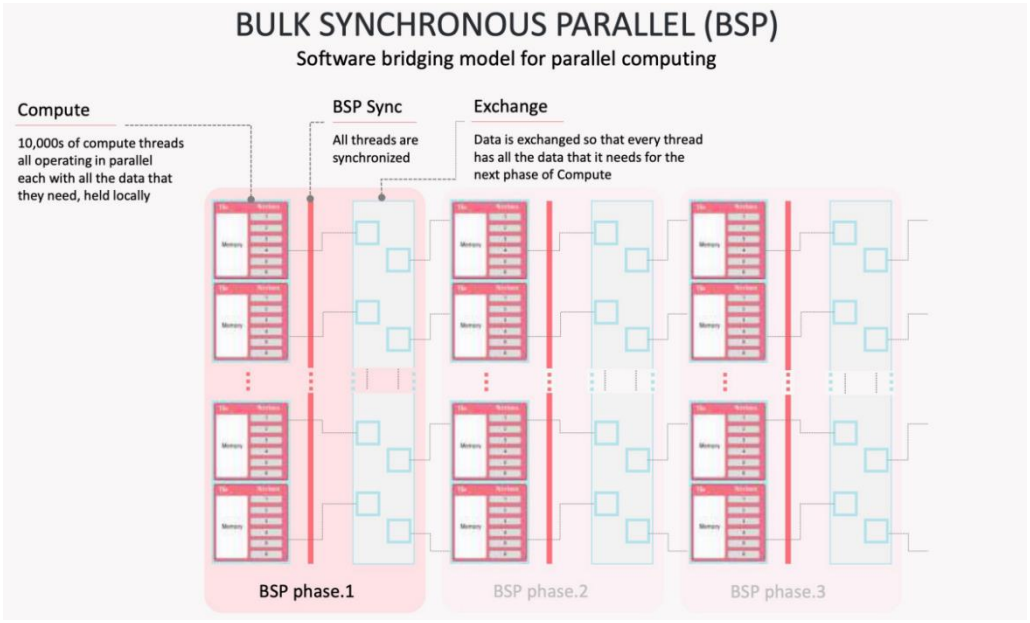
图表110： Graphcore IPU 内存架构



资料来源：Graphcore 官网、华泰研究

IPU 采用多核通信技术 BSP（Bulk-Synchronous Parallel），芯片内包含 1,000 多个内核，同时多芯片实现跨 IPU 连接。通过硬件支持 BSP 协议，将 IPU 计算逻辑分成了计算、同步、交换三个阶段。每个阶段处理器仅需本地内存、交换数据和等待同步，从而使并行算法在 BSP 模型中清晰呈现。由于无需处理 Java 平台的线程同步机制锁（Locks），IPU 能降低了编程的复杂性。

图表111：Graphcore IPU BSP 技术



资料来源：Graphcore 官网、华泰研究

在良率提升方面，Graphcore 也是通过增加冗余核心来实现。Graphcore 表示，IPU 使用分布式的存储技术，每块芯片上集成了 300MB 的内存和海量异步核，通过增加冗余内核和芯片面积，减少易受缺陷影响的“有效面积”，从而提升大芯片良率以及控制成本。在散热改善方面，Graphcore 通过在表面区域使用各种形状的散热孔来改善散热问题。

图表112：Graphcore 管理团队简介

管理层	职位	简介
Nigel Toon	联合创始人兼 CEO	曾担任两家由风投支持的芯片公司 CEO，领导并建立了多项半导体业务
Simon Knowles	联合创始人、CTO 兼工程执行副总裁	IPU 的创始设计师，拥有近 30 年的经验，一直从事针对新兴工作负载的新型处理器设计工作
卢涛	总裁、大中华区总经理和执行委员会成员	拥有近 20 年芯片领域经验，全面负责公司在大中华区的业务

资料来源：Graphcore 官网、华泰研究

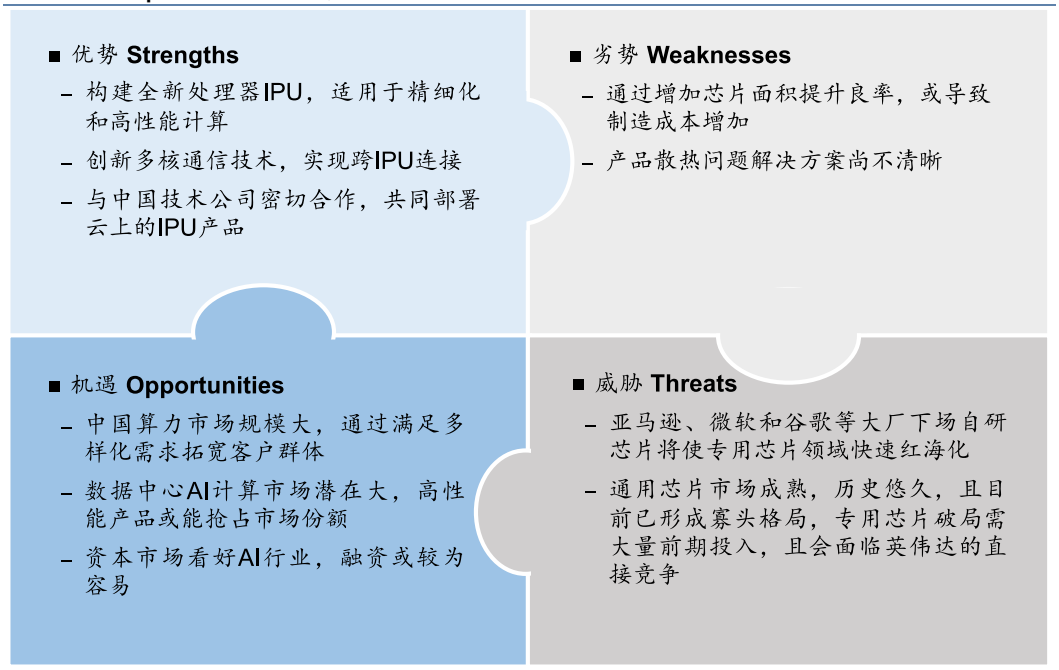
Graphcore 目前已进行了六轮融资，总额达到约 6.82 亿美元。最近一轮为 2020 年 12 月进行的 E 轮融资，金额为 2.22 亿美元，由 Ontario Teachers' Pension Plan 领投，估值达 25 亿美元。

图表113：Graphcore 融资

日期	轮次	金额	主要投资者	估值
2020 年 12 月	E 轮	2.22 亿美元	Ontario Teachers' Pension Plan	25 亿美元
2020 年 2 月	D 轮	1.50 亿美元	Mayfair Equity Partners	15 亿美元
2018 年 12 月	D 轮	2 亿美元	BMW i Ventures，微软	15 亿美元
2017 年 11 月	C 轮	5,000 万美元	红杉资本	-
2017 年 7 月	B 轮	3,000 万美元	Atomico	-
2016 年 10 月	A 轮	3,000 万美元	Amadeus Capital Partners, Bosch Ventures, 三星战略与创新中心	-

资料来源：Crunchbase、华泰研究

图表114： Graphcore SWOT 分析

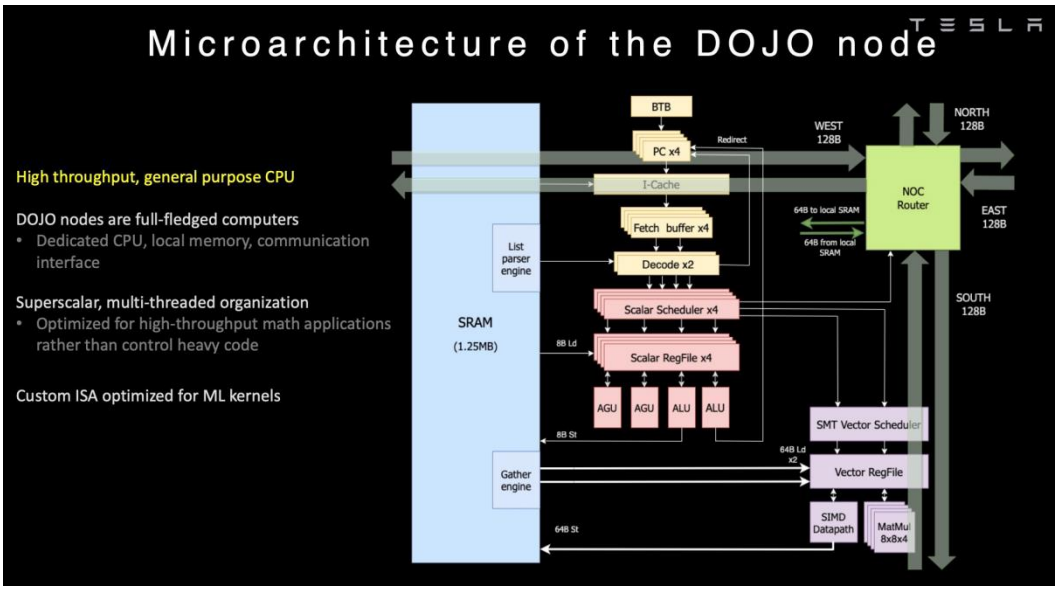


资料来源：Graphcore 官网、华泰研究

特斯拉：Dojo 超算为自动驾驶而生，为公司四大全栈自研科技支柱之一

2021 年 8 月，特斯拉（TSLA US）在 AI Day 里首次介绍 Dojo 超级计算机，一个基于 D1 Dojo 芯片的晶圆上（System on Wafer）系统级方案。特斯拉计划利用 Dojo 对海量的视频数据进行无监督学习，加速特斯拉的 Autopilot 和完全自动驾驶（FSD）系统的迭代，同时为特斯拉的人形机器人 Optimus 提供算力支持。根据特斯拉 23Q2 财报发布会和科技媒体 The Verge 于 7 月 20 日报道，特斯拉列出四大全栈自研自动驾驶的科技支柱：超大现实数据、神经网络训练、车载硬件和车载软件。当中，Dojo 超级计算机将提供更快和性价比更高的神经网络训练方案，已在 7 月开始量产。

图表115： Dojo 分布式架构示意图

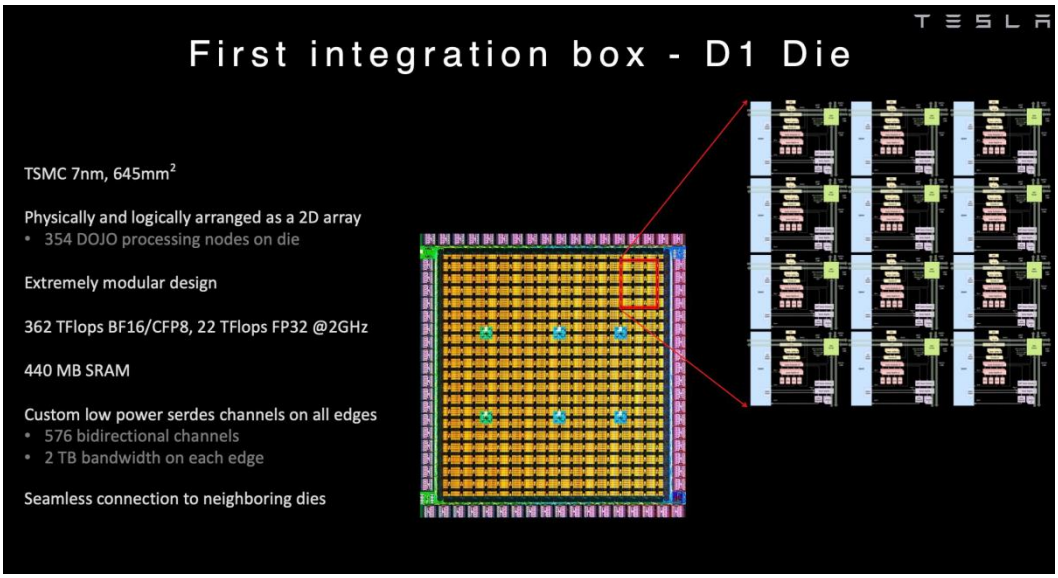


资料来源：Hot Chips 34、华泰研究

Dojo 架构能提供高宽带和低延迟的性能，其采用台积电 InFO_SoW（晶圆上集成扇出，Integrated Fan-Out System on Wafer）技术。InFO_SoW 技术整合了 InFO 技术、动力和散热模块。**通过将晶圆作为载体本身，不使用衬底和 PCB，从而获得低延迟的芯片间通信、高带宽密度和低配电网络阻抗，以提升计算性能和功耗效率。**Dojo 的数据传输方向与芯片平面平行，供电及水冷却方向与芯片平面垂直。

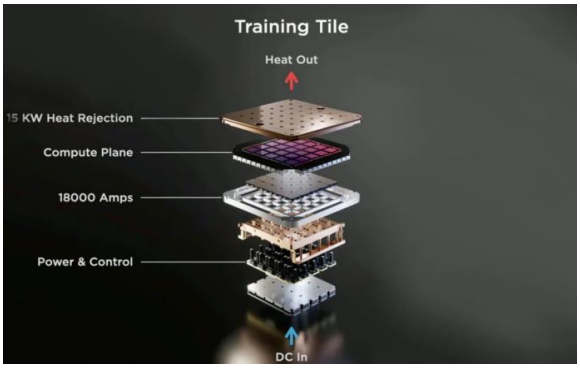
从 D1 Dojo 芯片到 Dojo ExaPOD：D1 Dojo 芯片采用台积电 7nm 制程工艺，芯片面积为 645 平方毫米，晶体管数量达 500 亿。**每个 Dojo D1 芯片**虽由 360（18 x 20）个 Dojo 核心拼接构成，但为了提升良率和处理器核心稳定，其中**只有 354 个 Dojo 内核**可用。**25 颗 D1 Dojo 芯片**（5x5 排列）集成到**一枚 Dojo Training Tile** 上，以二维 Mesh 结构无缝互连。**6 枚 Dojo Training Tiles**（整块 12 英寸重构晶圆）安装在铜质散热**盘上(tray)**，可为芯片提供刚度并将散热能力从 7 KW 提升至 15 KW。**2 个盘子**安装在一起成为 **1 个柜子(cabinets)**，包含 4,248 个内核(354 x 6 x 2)。最后，**10 个柜子**一起成为机柜集群 **Dojo ExaPOD**，而每个 Dojo ExaPOD **算力达到 1.1 ExaFLOPs**（10^{18}浮点运算），等于 1,100 PFLOPS 或内置 3,000 个 D1 芯片(25 x 6 x 2 x 10)。Dojo ExaPOD 集成 120 个训练模块(6 x 2 x 10)，拥有 1.3TB 的高速 SRAM 和 13TB 的高宽带 DRAM。

图表116： D1 Dojo 芯片



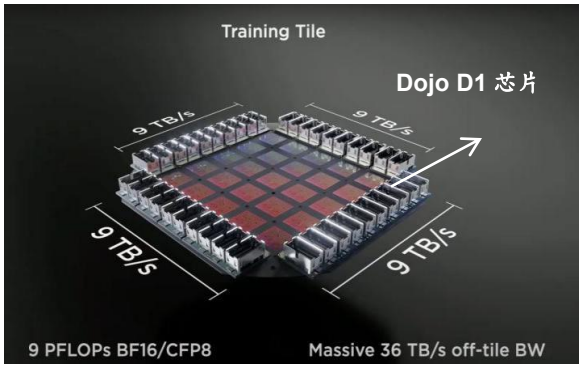
资料来源：Hot Chips 34、华泰研究

图表117： Dojo 结构示意图



资料来源：Tesla AI Day（2021）、华泰研究

图表118： Dojo Training Tile 性能



资料来源：Tesla AI Day（2021）、华泰研究

D1 Dojo 芯片的算力为：BF16/CFP8 达 362 TFLOPS，FP32 达 22.6 TFLOPS，热设计功耗达 400W。每个接口处理器（DIP）包括 32GB 的 HBM（800GB/s 存储带宽）、900GB/s 对外传输带宽（特斯拉自定义的 TTP 协议）、32GB/s PCIe Gen4 接口，以及 50GB/s 以太网带宽（特斯拉自定义的 TTPoE 协议）。每个 Dojo 节点都有一个内核，具有 CPU 专用内存和 I/O 接口，而每个内核则拥有一个 1.25MB SRAM 作为主存储器。Dojo 架构通过矩阵计算单元增强算力，SRAM 能以 400GB/s 速度加载，并以 270GB/s 速度存储。
Dojo Training Tile 包含了一整层液冷模块和铜质结构两种散热设计。每个边缘带宽为 4.5TB/s，每个模组功率 15kW，通过低延迟和高带宽实现大量的计算集成。Dojo Training Tile 采用直流电直接输入模式，单枚模组的总电流高达 18000A。

Dojo ExaPOD 集群体积约为 1.5 立方英尺，在 15 kW 液冷封装中能实现每秒 556 万亿次 FP32 浮点运算。特斯拉在 2021 年 AI Day 表示，准备于 Palo Alto 数据中心里放置 7 台 Dojo ExaPOD；公司也预计，2024 年 10 月算力的总规模将达 100 Exa-Flops。内存方面，特斯拉主要利用计算网格中的分布式 SRAM，通过大量更快更近的片上存储和片上存储之间的流转减少对内存的访问频度，来提升整个系统的性能。

图表119：特斯拉 Dojo 系列产品

名称	层级	片上 SRAM	算力	说明
Dojo Core	内核	1.25MB	1.024 TFLOPS	单个计算核心，64 位处理器，运行频率为 2GHz
Dojo D1	芯片	442.5MB	362 TFLOPS	单芯片，由 354 个内核组成一颗芯片，面积为 645mm2
Dojo Training Tile	模组	11GB	9050 TFLOPS	单枚训练模组，由 25 颗芯片组成一个训练模组
ExaPOD	集群	1320GB	1.1 EFLOPS	单个训练集群，由 10 个机柜组成，包含 3000 个 D1 芯片（120 个训练模块）

资料来源：Tesla AI Day （2021 & 2022）、华泰研究

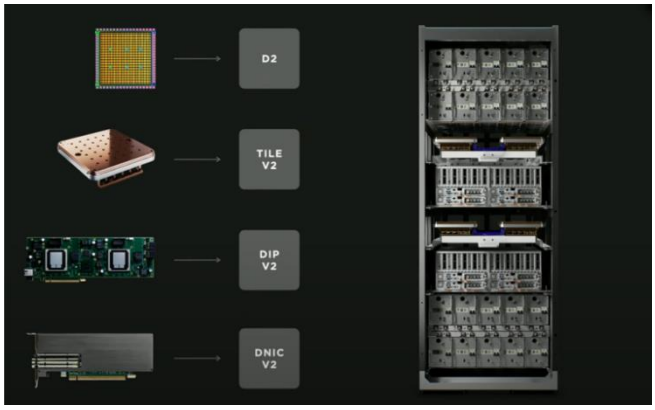
在良率问题上，特斯拉 2022 年 8 月在 Hot Chips 大会上表示，也可通过增加冗余核心数量，保证芯片即使存在个别杂质，也能正常运行，从而提升芯片良率。每个 Dojo D1 芯片虽由 360（18 x 20）个 Dojo 核心拼接构成，但为了提升良率和处理器核心稳定，其中仅只有 354 个核心可用。**在散热问题上，特斯拉表示通过全自研的 VRM（Voltage Regulator Module，电压调节模组）解决 Dojo 超算平台的散热控制。**通过 MEMS 振荡器（Oscillator）技术来感知电源调节模组的热形变，从而主动调节电源功率，满足芯片运行对热膨胀系数（CTE，Coefficient of Thermal Expansion）指标的要求。特斯拉的自研 VRM 在过去 2 年内迭代了 14 个版本，目前单个 VRM 可以在不足 25 美分硬币面积的电路上，提供 52V 电压和超过 1000A 的电流。

在 2022 年 9 月 30 日的 AI Day 上，特斯拉公布了 Dojo 的未来路线图，同时也表示，AI 团队正在研发新版本的 Dojo 超级计算机组件，其中包括 Dojo D2 芯片、Dojo Training Tile V2、Dip V2 以及 DNIC V2。特斯拉计划通过持续改进 Dojo 的硬件以突破现有算力瓶颈，同时预计相较于原有版本，新版本 Dojo 超级计算机性能将提升近 10 倍。在软件层面，FSD 或已实现落地层面的重要突破。6 月 26 日，马斯克在 X（推特）宣布，特斯拉将在 2023 年推出 L4-L5 级全自动驾驶汽车，从 V12 版本开始，FSD 将去掉 Beta 后缀，或将意味着 FSD 将成为正式版。当地时间 8 月 26 日，马斯克在硅谷帕洛阿尔托（Palo Alto）的街道上直播测试特斯拉全自动驾驶系统 FSD 12，直播一周后，特斯拉在加拿大和美国推出“城市自动驾驶”功能。

图表120： Dojo ExaPOD 示意图

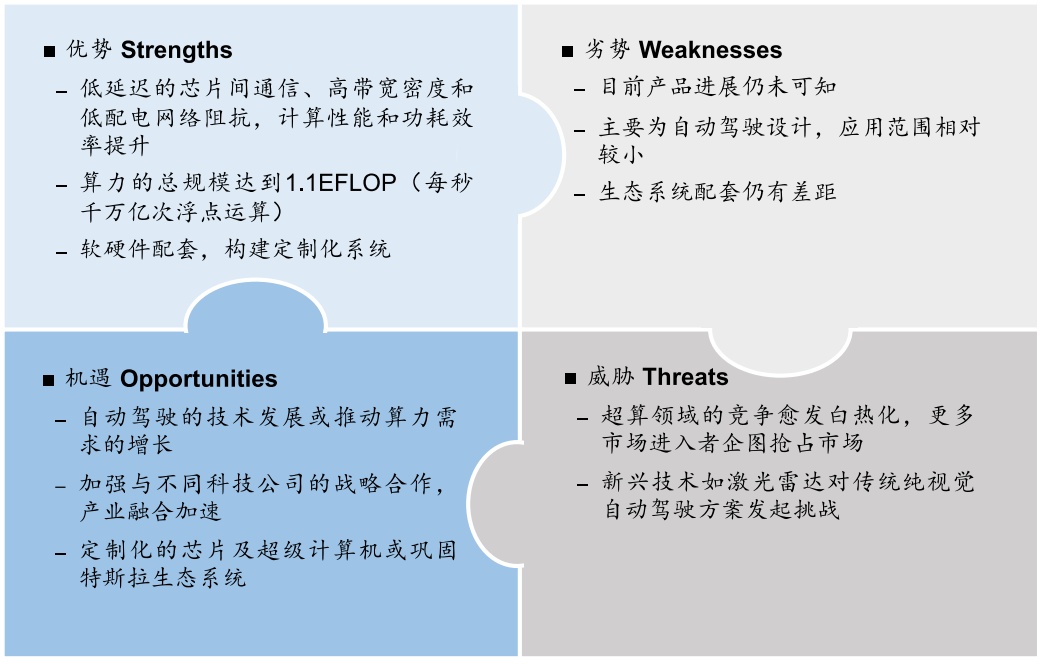
资料来源：Tesla AI Day（2022）、华泰研究

图表121： Dojo 未来路线图



资料来源：Tesla AI Day（2022）、华泰研究

图表122： 特斯拉 DOJO SWOT 分析



资料来源：Tesla AI Day （2021 & 2022）、华泰研究

晶圆级芯片跟传统芯片的各项对比

我们将 Cerebras WSE、Dojo D1、Dojo Training Tile（由 25 块 D1 组成）、Graphcore IPU、英伟达 H100 和 AMD MI300X 的性能指标进行对比（但我们需强调，理论上他们也并非同一基准的比较，因此以下的图表仅作参考）。综合考虑下，A100 和 H100 仍然是大部分企业开展 AI 训练的芯片首选。Cerebras WSE 在内核、SRAM、内存带宽、通信带宽和晶体管都有突出的表现，但值得注意的是，Cerebras 没有公布浮点运算算力；而 Dojo 只提供 FP32 和 BF16/CFP8 算力；IPU 只公布 FP32 和 FP16；而英伟达 A100 和 H100 作为训练芯片，提供最全的精度模式。另外，几款晶圆级芯片如 Cerebras WSE、Dojo D1、Graphcore 都没有使用 DRAM 而是采 SRAM，但 SRAM 在具备更高存储速度的同时，成本也较高。

图表123：AI 芯片性能对比（但并非同一基准的比较）

名称	超算级别		晶圆级芯片			传统GPU		
	NVIDIA DGX A100	Dojo Training Tile	Cerebras WSE-2	Dojo D1	Graphcore Colossus MK2 GC200 IPU	NVIDIA A100 80GB SXM	NVIDIA H100 SXM	AMD MI300X
面积	6,608mm²(8*826)	<92,903mm²	46,225mm²	645mm²	823mm²	826mm²	814mm²	
内核	55,296 CUDA+3,456 Tensor	53,400	850,000	1,416	1,472	6,912 CUDA + 432Tensor	8,448/16,896 FP64/FP32+538 Tensor	8 CDNA 3 GPU chiplet + 4 IO chiplet
制程	7nm	7nm	7nm	7nm	7nm	7nm	4nm（台积电5nm）	5nm/6nm
晶体管	4320亿（8*540）	12,500亿	26,000亿	500亿	594亿	540亿	800亿	1530亿
静态随机存储器(SRAM)		11 GB	40 GB	442.5 MB	900 MB	40 MB		
动态随机存储器(DRAM)	640GB HBM					80 GB HBM	80 GB HBM	192GB
内存带宽		10 TB/s	20 PB/s	10 TB/s	47.5 TB/s	2.039 TB/s	3.35TB/s	5.2 TB/s
Fabric带宽		36 TB/s	27.5 PB/s	4 TBs/s	320 GB/s	600 GB/s		896 GB/s
BF16/CFP8		9 PFLOPS		362 TFLOPS		312 TFLOPS		
FP64						9.7 TFLOPS	34TFLOPS	
FP32		565 TFLOPS		22.6 TFLOPS	64 TFLOPS	19.5 TFLOPS	67 TFLOPS	
FP16					250 TFLOPS	312 TFLOPS	1979 TFLOPS	
INT8	10PFLOPS					624 TOPS	3958 TOPS	
功耗	6.5kW	5kW	20kW/15kW	400W		400W	700W	750W
售价	20万美元		300万美元			2万美元	4万美元	
用途	通用	训练	通用	训练	通用	通用	通用	通用

资料来源：各公司官网、华泰研究

DRAM vs SRAM：片上 VS 片外内存

对比内存容量路线，Cerebras、Graphcore 和特斯拉的 Dojo 技术路线是大尺寸芯片上 SRAM +大容量 DDR，而英伟达和 AMD 的技术路线则是片上小容量 SRAM +堆叠式 HBM。大尺寸芯片上 SRAM 读取速度快且功耗低，但由于 SRAM 占用面积大以及内部结构复杂，因此成本也较高，为 DRAM 成本的约 100 倍。SRAM 主要用于 CPU 高速缓存。1 个 SRAM 单元通常由 4-6 只晶体管组成，存储器只要保持通电，里面储存的数据就可以恒常保持。通过将存储器分散地集成在运算单元旁，尽可能减少数据搬移，可减少负载突破内存瓶颈，也能降低功耗。因此大尺寸芯片使用 SRAM 可避免多芯片之间通信带宽的限制，同时获得带宽提升。

DRAM 和 SRAM 两种内存容量路径主要的差异在于：1）SRAM 传输速度比较快。SRAM 可以一次接收所有的地址位，使用行列独立技术，而 DRAM 则使用行列地址复用技术，因此相较 SRAM 的接口更加复杂。此外，SRAM 主要用于二级高速缓存（Level 2 Cache），利用晶体管来存储数据，因此与 DRAM 主要用于内存相比，SRAM 的访问和读取速度更快；2）**SRAM 的功耗较小。**SRAM 不需要刷新电路即能保存内部存储的数据，而 DRAM 每隔一段时间，需要刷新充电一次，否则内部的数据即会消失，因此 SRAM 功耗较小；3）**SRAM 的缺点在于集成容量较低，体积较大。**由于存储单元结构不同导致 SRAM 和 DRAM 的体积和集成容量的不同，一个 DRAM 存储单元大约需要一个晶体管和一个电容（不包括行读出放大器等），而一个 SRAM 存储单元大约需要六个晶体管。此外，SRAM 可能需要增加冗余面积保证芯片安全性和提升良率，因此相同容量的 DRAM 内存体积或较小。

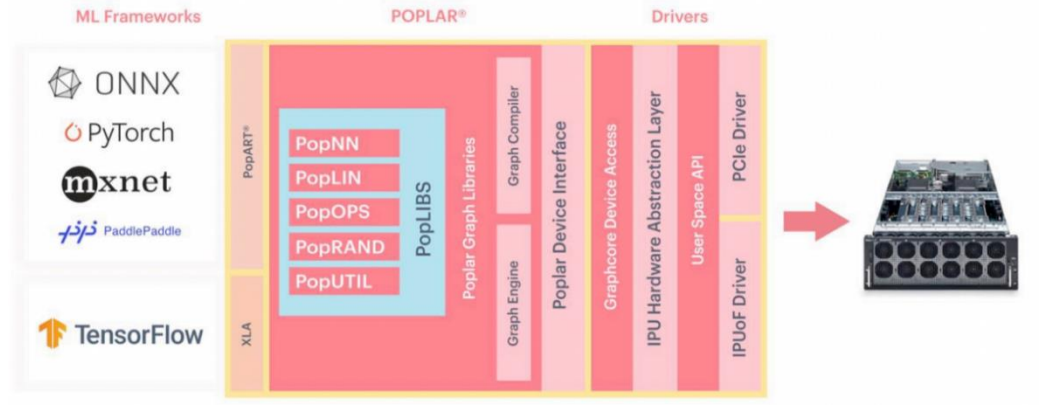
片上内存 VS 片外内存：晶圆级芯片上共享内存相较传统 GPU 访存速度更快，对比英伟达利用 NVLink 和 NVSwitch 技术突破片外内存访存效率瓶颈。片上内存的访存调度属于**共享内存**，按着线程块分配，因此块中的所有线程都可访问同一共享内存，进行读写操作，提升片上共享内存的访存速度。相较而言，一般的 GPU 是利用**片外内存**，需要通过频繁读取片外的 DRAM 进行内存调度，因此访存速度较低和功耗较高。为了解决片外 DRAM 传输问题，英伟达通过 NVLink 和 NVSwitch 技术实现了服务器中所有 GPU 之间的高带宽连接，并提高可扩展性。

HBM 是一种基于 3D 堆栈工艺的高性能 DRAM，适用于高储存器带宽需求的应用场合，如图像处理器、网络交换及转发设备（交换机、路由器）等。堆叠式 HBM 技术具有高密度、低延迟、高性能、耐用和低功耗等特性。HBM 利用 TSV（硅通孔）技术打造立体堆栈式的显存颗粒，通过硅中介层，让显存连接至 GPU 核心，并封装在一起，完成显存位宽和传输速度的提升。因此，DRAM 颗粒可相互堆叠，使得芯片在垂直面上能实现小面积和高容量。英伟达的 A100 搭载 80GB 的 HBM2e，而 H100 搭载 80GB 的 HBM3，其中 HBM3 的最大容量高达 24GB，是 HBM2e 最大容量 16GB 的 1.5 倍。GH200 超级芯片的 HBM3 搭载容量较 H100 再提升 20%，达 96GB；新一代 GH200 则搭载了最新的 HBM3e，容量达 141GB，对比之前的 GH200 大幅提升。AMD 的 MI300 系列也搭载了 HBM3，其中，MI300A 容量与上一代 MI250X 相同为 128GB，而 MI300X 则达 192GB，容量提升了 50%。

软件生态：英伟达凭借 CUDA 形成稳固的护城河。对任何一种计算平台和编程模型来说，开发人员与其学习、磨合和建立生态圈都需要时间，更多的开发者意味着不断迭代的工具和更广泛的多行业应用。CUDA 是英伟达于 2006 年推出的一种基于 C 语言的编程框架，包含 CUDA 指令集架构（ISA, Instruction Set Architecture）和 GPU 内部的并行计算引擎。通过先发优势和长期耕耘，CUDA 生态圈已较为成熟，为英伟达 GPU 开发、优化和部署多种行业应用提供了独特的护城河。根据英伟达 2022 年报，全球有 380 万开发人员使用 CUDA。这进一步为选择英伟达 GPU 提供了更为充分的理由，正向循环、不断完善的生态也将进一步提高其用户粘性。

Cerebras 软件栈名为 CSoft，由 Cerebras ML 软件与机器学习框架 TensorFlow 和 PyTorch 集成。Cerebras 图形编译器（Graph Compiler）能够自动将神经网络转换成 WSE 计算资源可用的可执行程序。特斯拉推出 **Dojo 专属的全栈软件系统**，通过自研软件和开源软件的结合，其中包括底层驱动软件、编译器引擎、PyTorch 插件和上层的神经网络模型等。通过 Dojo 编译器实时生成代码。用户通过软件栈可为硬件创建和优化代码，无需重新手写的内核代码，提升研发效率。据科技分析机构 Moor Insights & Strategy 报告，**Graphcore** 将约一半研发人员投入开发 **Poplar 开源软件栈**，其由计算图、元素编译器、优化库以及用于运行时调度的计算图引擎组成，能够基于 TensorFlow 和 PyTorch 进行编译。Graphcore 在中国积极组建创新社区，已在微信、知乎、微博和 GitHub 开通了官方频道，旨在与开发者更有效交流和互动。总体来看，Cerebras、特斯拉和 Graphcore 均通过已广泛使用的机器学习框架来开发插件，以减少软件开发人员的学习成本，并通过自研的编译器来实现软件的执行。但与英伟达 CUDA 相比，开发者生态未形成规模，目前暂未实现生态的正向循环和客户粘性增长。

图表124：Graphcore Poplar 软件栈示意图



资料来源：Moor Insights & Strategy、华泰研究

AI 芯片产业链：聚焦兵家必争之地 CoWoS 封装

台积电大扩 CoWoS 产能，供给紧张有望得解

在市场对 AI 芯片的旺盛需求下，导致先进封装产能供不应求。我们认为，CoWoS 是限制 AI 芯片出货量的主要瓶颈。英伟达 A100 和 H100 采用台积电（TSM US）CoWoS（2.5D）封装技术，MI300 系列需采用台积电 CoWoS（2.5D）和 SoIC（3D）技术。据 Digitimes 在 7 月 14 和 21 日的报道中提到，台积电正积极扩大 CoWoS 产能，包括竹南、龙潭和台中三地；2023 年产能至少 12 万片，2024 年将达 24 万片，而英伟达将取得约 15 万片。报道中还提到，为了实现扩产，台积电或将把部分 oS（on Substrate）释放给其他封装厂商，订单或会外溢到包括中国台湾封测龙头日月光、矽品精密（2018 年被日月光控投收购）、台湾的联华电子、美国的 Amkor Technology、中国大陆的通富微电（6 月 27 日披露）等。

此外，英伟达也在积极寻找台积电以外的选择。在台积电产能供应日益紧张之下，英伟达也正在考虑将其部分 GPU 外包给三星电子（009150 KS）进行制造。据 Digitimes 7 月 5 日报道，如果三星的 3nm 试验产品通过性能验证，且其 2.5D 先进封装技术符合美国芯片制造商的要求，英伟达可能会外包一些订单至三星电子。目前，台积电 CoWoS 的三大客户包括：英伟达、博通和赛灵思；AMD 的 MI300 系列在 4 季度推出后，或将跻身前五大客户；亚马逊在 2024 年也或将成为第三大客户；而我们认为鉴于博通与谷歌共同设计 TPU，据路透社 5 月 31 日新闻，Meta 也已成为博通的 ASIC 客户，因此我们认为这里博通的量或是代表谷歌和 Meta。

上文提到台积电已在改装一些厂房来应对供不应求的 CoWoS 产能。我们认为，改装厂房其实还是相对简单，因此 CoWoS 的真正瓶颈之一，也许是封装用的机械设备所需的交货周期较长。据 DigiTimes 在 7 月 25 日报道，CoWoS 产能扩充缓慢的原因在于其使用的设备交货周期，关键设备如研磨液供给设备、半导体清洁装置（湿制程设备）等，主要供应厂商包括日本的 Tazmo（6266 JT）和 Shibaura（6590 JT），设备的完整交货周期往往在 6-8 个月（6 个月交货+2 个月调试参数）。然而，同样可以提供这类设备的还有如台湾本地供应商亚泰半导体（未上市），台积电或能通过其他供应商缓解一定压力。

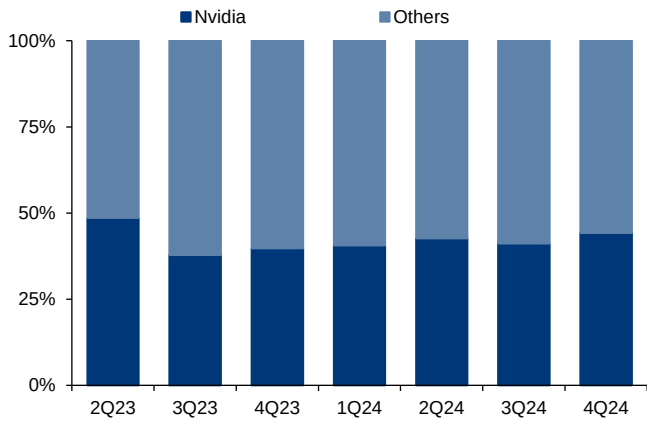
图表125：Tazmo 半导体制造设备产品

资料来源：Tazmo 官网、华泰研究

图表126：Tazmo 研磨液供给系统设备

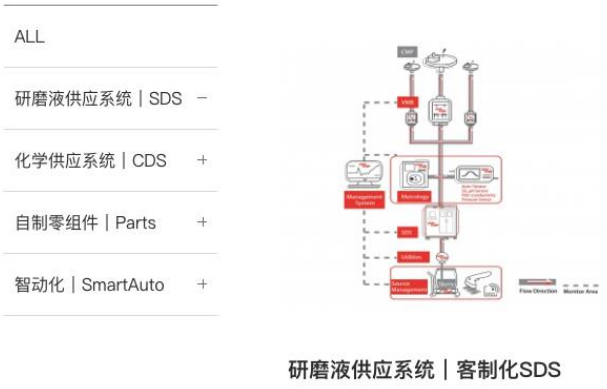
资料来源：Tazmo 官网、华泰研究

图表127：英伟达在 CoWoS 季度 Output 的占比



资料来源：semianalysis 官网、华泰研究

图表128：亚泰半导体的研磨液供应系统



资料来源：Tazmo 官网、华泰研究

硅晶圆供应商：台积电的 6 家硅晶圆供应商占全球总产能 90%以上

台积电制作硅中介层和 SoC 的原材料是硅晶圆。根据台积电 2022 年报中披露的原材料供应商信息，总共列有六家，两台、两日、一德、一韩，分别是：FST（Formosa Sumco Technology Corporation，台塑胜高，中国台湾和日本合资，股票代码：3532 TT）、Global Wafers（环球晶圆，中国台湾，股票代码：6488 TT）、SEH（Shin-Etsu Handotai，日本信越半导体，日本，股票代码：4063 JT）、Siltronic（世创电子，德国，股票代码：WAF DE）、SK Siltron（SK 矽得荣，韩国，未上市）和 SUMCO（胜高，日本，股票代码：3436 JT）。台积电通过向这六家供应商分别采购硅晶圆材料，以分摊供应风险。这六家供应商的总硅晶圆产能约占全球供应量的 90%以上。

图表129：台积电年报中列举的硅晶圆供应商

主要原物料名称	供应商	市场状况	台积电的采购策略
硅晶圆	FST、Global Wafers、SEH、Siltronic、SK siltron、SUMCO	6 家供应商硅晶圆产能合计约占全球供应量的 90%以上。	• 硅晶圆供应商必须通过台积电最严格的制程认证程序。 • 台积电向多个不同供应商购买矽晶圆，以确保量产无虞，并分散采购风险。 • 推动矽晶圆的品质改善以维持台积电的技术领先地位。 • 台积电定期检讨供应商的产品品质、交货状况、成本、永续管理及服务绩效各方面表现，并将结果列为未来采购决策参考。 • 定期稽核供应商的品管系统，以确保台积电能持续提供高品质的产品。 • 为了优化成本面与供给面的管理，台积电采取多种方式与供应商合作。

资料来源：台积电 2022 年报，华泰研究

除作为主要原材料的硅晶圆之外，晶圆制造流程中还需要其他原材料如制程用化学原料、黄光制程材料、特殊气体、研磨液、研磨垫、钻石碟等。可以看到台积电年报中披露的供应商分布在中国台湾、日本、德国、韩国、法国等，显示了台积电原材料全球采购的布局和分摊供应链风险的能力。

图表130：台积电晶圆制造的其他供应商（包括原材料）

主要原物料名称	供应商	中文名称	所在地区
制程用化学原料	BASF	巴斯夫	德国
	Merck	默克集团	德国
	Air Liquide	液化空气集团	法国
	DuPont	杜邦公司	美国
	Entegris	英特格公司	美国
	Fujifilm Electronic Materials	富士电子材料集团	日本
	Kanto PPC	关东鑫林科技股份有限公司	日本
	RASA	RASA 工业	日本
	Tokuyama	德山株式會社	日本
	Kuang Ming	广明实业股份有限公司	中国台湾
	Shiny	胜一化工股份有限公司	中国台湾
	Wah Lee	华立企业股份有限公司	中国台湾
黄光制程材料	3M	明尼苏达矿业及机器制造公司	美国
	Fujifilm Electronic Materials	富士电子材料集团	日本
	JSR	JSR 株式会社	日本
	Nissan	日产化学株式会社	日本
	Shin-Etsu Chemical	信越化学工业株式会社	日本
	Sumitomo Chemical	住友化学株式会社	日本
	T.O.K	东京应化工业株式会社	日本
特殊气体	Air Liquide	液化空气集团	法国
	SK Materials	SK 材料公司	韩国
	Air Products	空气化工公司	美国
	Entegris	英特格公司	美国
	Praxair	普莱克斯工业气体公司	美国
	Central Glass	中央玻璃株式会社	日本
	Nippon Sanso Taiwan	大阳日酸株式会社	日本
	Linde LienHwa	联华林德气体公司	中国台湾
	Taiwan Material Technology	茂泰利科技股份有限公司	中国台湾
研磨液、研磨垫、钻石碟	3M	明尼苏达矿业及机器制造公司	美国
	Cabot Microelectronics	嘉柏微电子材料公司	美国
	DuPont	杜邦公司	美国
	AGC	AGC 株式会社	日本
	Fujibo	爱媛株式会社	日本
	Fujifilm Electronic Materials	富士电子材料	日本
	Fujimi	Fujimi 株式会社	日本
先进封装设备	Shibaura Mechatronics	芝浦机电	日本
研磨液供给设备	Tazmo	TAZMO 株式会社	日本

资料来源：台积电 2022 年报，asiatimes 官网，华泰研究

衬底/基板（Substrate）：揖斐电、景硕、欣兴电子等

根据英伟达 2022 年报，衬底供应商包括揖斐电（Ibiden，日本）、景硕（Kinsus，中国台湾）、欣兴电子（Unimicron，中国台湾）等；而 AMD 曾公开披露过的衬底供应商有新光电气（Shinko Electric，日本）和三星电机（Samsung Electro-Mechanics，韩国）。除了这些被公开披露信息明确与英伟达和 AMD 有供应关系的衬底供应商之外，奥特斯（AT&S，奥地利）和南亚科技（Nanya，中国台湾，股票代码：2408 TT）等，也常出现在其他芯片巨头如英特尔的衬底供应商名录中。中国大陆企业，例如景旺电子（603228 CH）、沪电股份（002463 CH）、生益科技（600183 CH）等，更多专攻服务器用大型 PCB 板，虽然不直接供应台积电或芯片巨头，但产品最终或会应用到英伟达或 AMD 芯片组成的服务器中。

图表131：衬底/基板供应商

硬件结构与制造流程	供应商	地区	目前与英伟达的明确合作关系	目前与 AMD 的明确合作关系
衬底/基板（Substrate）	Ibiden Co. Ltd.（揖斐电株式会社）	日本	✓	X
	Kinsus Interconnect Technology Corporation（景硕科技）	中国台湾	✓	X
	Unimicron Technology Corporation（欣兴电子）	中国台湾	✓	X
	Shinko Electric Industries Co. LTD.（新光电气）	日本	X	✓
	Samsung Electro-Mechanics Co., Ltd.（三星电机）	韩国	X	✓

资料来源：英伟达年报，AMD 官网，三星财报，华泰研究

HBM 内存：SK 海力士、三星、美光，三足鼎立

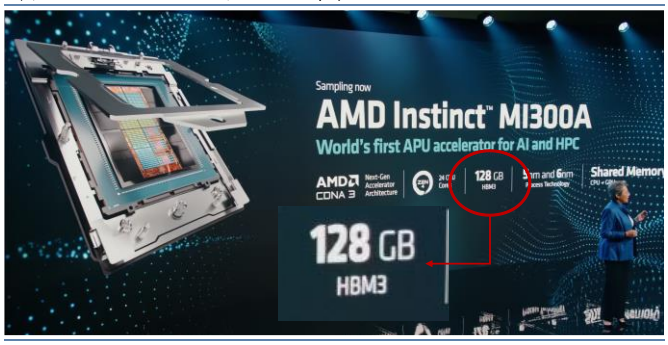
HBM 内存供应商目前的市场格局为 SK 海力士（000660 KS）、三星、美光（MU US）三分天下。据中国台湾《科技新报》2023 年 4 月 18 日报道，H100 和 MI300 系列使用的 HBM3 内存，唯一生产商是 SK 海力士，但三星和美光预计也将在今年底至明年初开始量产。虽然目前 H100 的 HBM3 来自 SK 海力士，但为了满足英伟达拓展二供的需求，我们认为三星或也将提速进入 HBM3 的供应商列表。另外，英伟达在 2023 年 8 月 8 日发布了使用 HBM3e 的 GH200 新版本，这是海力士于 2023 年 6 月发布的增强版 HBM3。

2023 年 4 月 20 日，SK 海力士发布了大内存（24GB）的 HBM3 产品，同时宣布已给客户送样，因此考虑到 AMD MI300X 的 192GB HBM3（24*8=192），我们认为，MI300X 搭载的 HBM3 或即是 24GB 的海力士 HBM3 产品。而 MI300A 的 128GB HBM3（16*8=128）则或对应的是海力士于 2022 年 6 月发布的 16GB 的 HBM3 产品。Hopper GPU 的 96GB HBM3 也同样对应 16GB 的 HBM3（16*6=96）。据中国台湾《科技新报》2023 年 4 月 18 日报道，SK 海力士在 2022 年 6 月发布 16GB HBM3 的同时就立即供给英伟达。同样据上文报道，除海力士外，另外两家内存巨头三星和美光预计在今年底到明年年初开始量产 HBM3，或可满足英伟达和 AMD 拓展其他供应商的需求。

图表132：MI300X 使用 HBM3 内存

资料来源：AMD 官网、华泰研究

图表133：MI300A 使用 HBM3 内存



资料来源：ADM 官网、华泰研究

服务器相关供应商：惠与、戴尔、联想、美超微、广达、纬创等

英伟达 GPU 和 AMD MI300 系列使用的散热方案供应商也值得关注。由于英伟达 Grace Hopper 单颗芯片最高功耗可达 1000W，对散热系统要求更高，英伟达在 2023 年 Computex 上发布了液冷版本 H100 HPC 服务器。风冷系统方面，全球最大散热模组厂商 AVC（3017 TT）是前代 DGX H100 的风冷散热系统供应商。AMD 的散热供应商为中国台湾散热解决方案厂商双鸿科技（3324 TT）。此外，中国台湾散热材料厂商健策精密（3653 TT）为 AMD 提供均热片、LED 导线架、电子周边零组件、通讯周边零组件等元器件。

服务器 PCB 板供应商包括胜宏科技（300476 CH）、景旺电子、沪电股份、生益科技等，他们不一定直接与英伟达或 AMD 进行合作，但他们的产品可能会用在搭载英伟达和 AMD 芯片的服务器上。另外，奥士康（002913 CH）作为国内高密度印制电路板的领先企业，较早布局 AI 服务器和数据中心业务，据公司 2023 年 5 月 4 日投资者关系活动记录，其在服务器 PCB 市场占比超 20%，已进入英特尔和 AMD 新一代服务器供应商目录。

最后，使用 AI 芯片组装服务器的厂商可分为白牌（指不贴服务器品牌名）厂商和品牌商（指贴品牌名）两种，近年白牌厂商份额逐渐增长。品牌商包括浪潮（600756 CH）、戴尔、联想（00992 HK）、IBM（IBM US）等；白牌厂商包括广达（2382 TT）、纬创（3231 TT）、英业达（2356 TT）、美超微（SMCI US）等。互联网云厂商或服务器品牌商将硬件组装委托给白牌厂商，合作开发定制化“白牌”服务器。近年来，部分服务器上游硬件厂商通过直接委托白牌 ODM 厂商进行组装，以自身品牌进行服务器销售，绕过服务器品牌商，推动白牌厂商份额上涨。据 AMD 2022 年报，服务器合作厂商包括戴尔、HPE（惠与）（HPE US）、联想、美超微、思科（CSCO US）等；据数位时代 2023 年 3 月 30 日报道，英伟达的服务器供应商包括纬创和美超微。

图表134： AMD 与英伟达 AI 芯片产品各硬件结构、原材料等供应商一览

硬件结构与制造流程	供应商	地区	目前有无与英伟达的明确合作关系	目前有无与 AMD 的明确合作关系
衬底/基板（Substrate）	Ibiden Co. Ltd.（揖斐电株式会社）	日本	✓	X
	Kinsus Interconnect Technology Corporation（景硕科技）	中国台湾	✓	X
	Unimicron Technology Corporation（欣兴电子）	中国台湾	✓	X
	Shinko Electric Industries Co. LTD.（新光电气）	日本	X	✓
	Samsung Electro-Mechanics Co., Ltd.（三星电机）	韩国	X	✓
PCB	景旺电子*	中国大陆	✓	✓
	沪电股份*	中国大陆	✓	X
	生益电子*	中国大陆		
内存（Memory）	Micron Technology（美光科技）	美国	✓	✓
	SK Hynix（SK 海力士）	韩国	✓	✓
	Samsung Semiconductor, Inc（三星电子）	韩国	✓	✓
晶圆制造	Taiwan Semiconductor Manufacturing Company Limited （台湾积体电路制造）	中国台湾	✓	✓ （<=7nm）
	Samsung Electronics Co. Ltd（三星电子）	韩国	✓	X
	GLOBALFOUNDRIES Inc. （格罗方德）	美国	X	✓ （>7nm）
封测（Testing and Packaging）	ASE Technology Holding Co.,Ltd（日月光控投）	中国台湾	✓	✓
	Amkor Technology （安靠科技）	美国	✓	X
	SILICONWARE PRECISION INDUSTRIES CO., LTD.（矽品精密工业）	中国台湾	✓	X
	通富微电	中国大陆	X	✓
	联华电子	中国台湾		
服务器相关	Dell（戴尔）	美国	✓	✓
	HPE（惠与科技）	美国	✓	✓
	Lenovo（联想）	中国大陆	✓	✓
	Supermicro（美超微）	美国	✓	✓
	Wistron（纬创）	中国台湾	✓	✓
散热	AVC（奇鋐科技）	中国台湾	✓	✓
	双鸿科技	中国台湾	✓	✓
	健策精密	中国台湾	×	✓

注：*不一定直接供应英伟达或 AMD，产品或应用在搭载英伟达和 AMD 芯片的服务器上

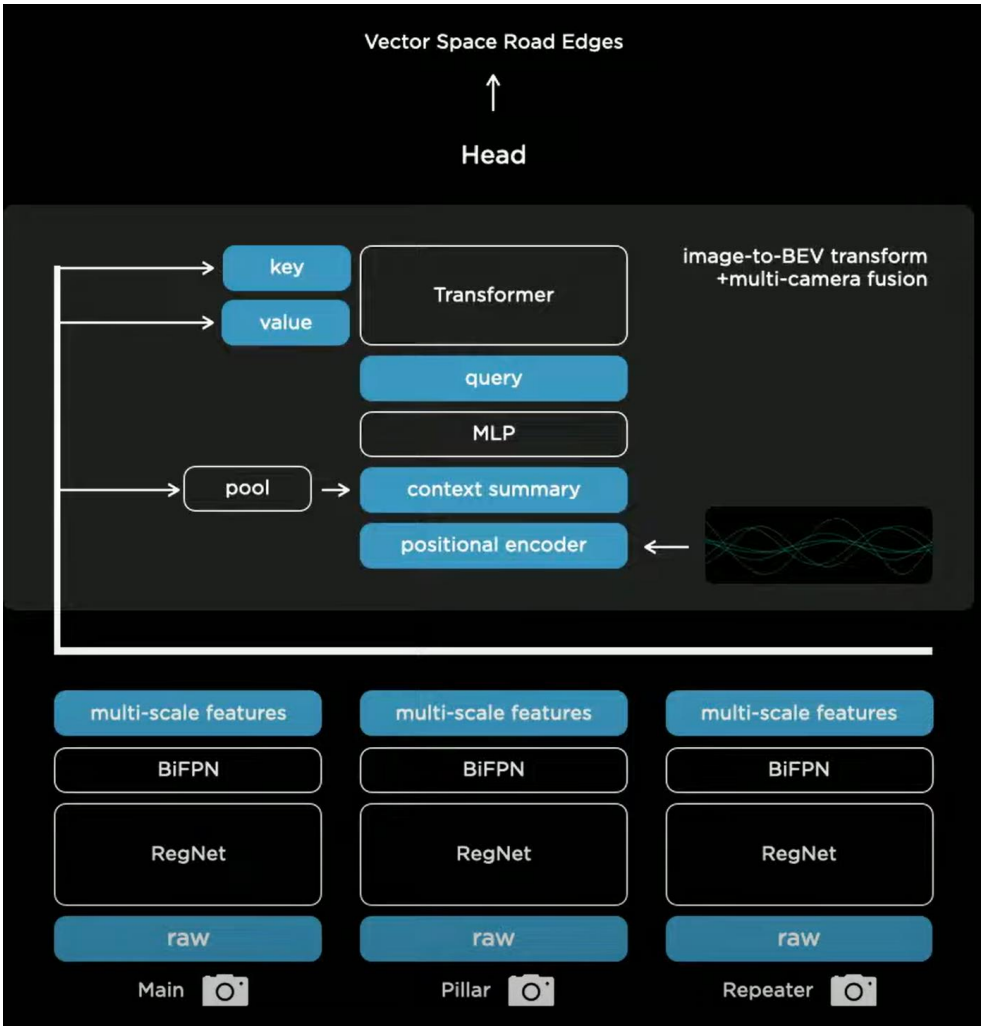
资料来源：英伟达年报，AMD 年报，AMD 官网，三星财报，招股说明书，华泰研究

AI 不只是大模型，自动驾驶芯片群雄逐鹿，谁能突围？

自动驾驶是目前 AI 应用落地较全面的领域。在自动驾驶解决方案中，特征提取、场景模拟与生成，感知和决策以及路径规划等任务，均需应用深度学习算法，包括机器学习、计算机视觉等，而近年也加入了不少 Transformer 算法，因此对于自动驾驶芯片的算力与性能也提出了更高的要求。Transformer 的基本原理是在预训练的前提下，根据一系列信息的全局关联性去预测下一步，从而输出结果。在处理自动驾驶长尾问题时，谷歌 Waymo 将 AIGC 应用于仿真场景生成去模拟罕见的驾驶场景，以优化模型训练；特斯拉基于 Transformer 进行 BEV（Bird's Eye View，即俯视角度）特征提取。

2021 年，特斯拉通过将感知任务内置于 Transformer 中，使神经网络能够自动完成从 2D 图像到 BEV 特征信息的转换，具体步骤为：1）通过 RegNet（Regularization Network，用于图像分类的神经网络结构）对摄像头获取的 2D 图像信息依据分辨率进行分层；2）利用 BiFPNs（Bi-directional Feature Pyramid Networks，用于图像语义分割的神经网络结构）使分类信息发生交互，形成可被神经网络理解的输入信息；3）基于预训练获取的信息，根据相关性提取观测对象的特征（如类别和位置等），并整合至 3D 空间。这一架构已成为目前在自动驾驶领域中广泛应用的范式，地平线、百度 Apollo 和毫末智行等均已完成 Transformer 的自研，并将其纳入智能驾驶解决方案中。

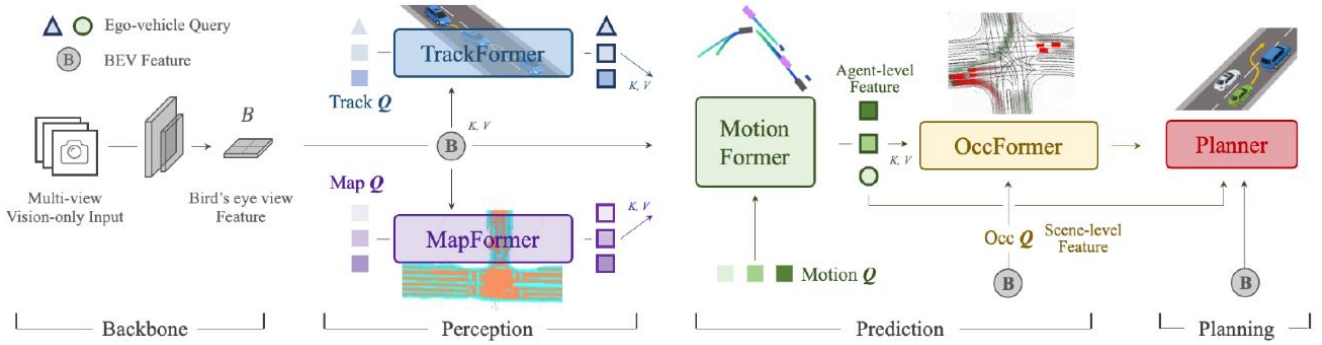
图表135： 特斯拉 Transformer 架构



资料来源：特斯拉 2022 AI Day、华泰研究

Transformer 在自动驾驶中的应用正不断衍生，并开始覆盖全栈任务的发展。2023 CVPR（Conference on Computer Vision and Pattern Recognition，计算机视觉与模式识别会议，是以计算机视觉和机器学习等为导向的人工智能顶级会议）的最佳论文《以路径规划为导向的自动驾驶（Planning-oriented Autonomous Driving）》（发表机构包括上海人工智能研究所、武汉大学、商汤科技等）提出感知决策一体化的端到端自动驾驶解决方案 Unified Autonomous Driving（UniAD）。UniAD 通过 Transformer 多层架构覆盖感知、预测和规划任务，不同层并行处理图像信息并进行交互，最终根据车辆对周围环境的感知及预测信息以实现路径规划。

图表136：UniAD 运作流程图



资料来源：《Planning-oriented Autonomous Driving》（2022，Shanghai AI Laboratory, Wuhan University, SenseTime Research 等）、华泰研究

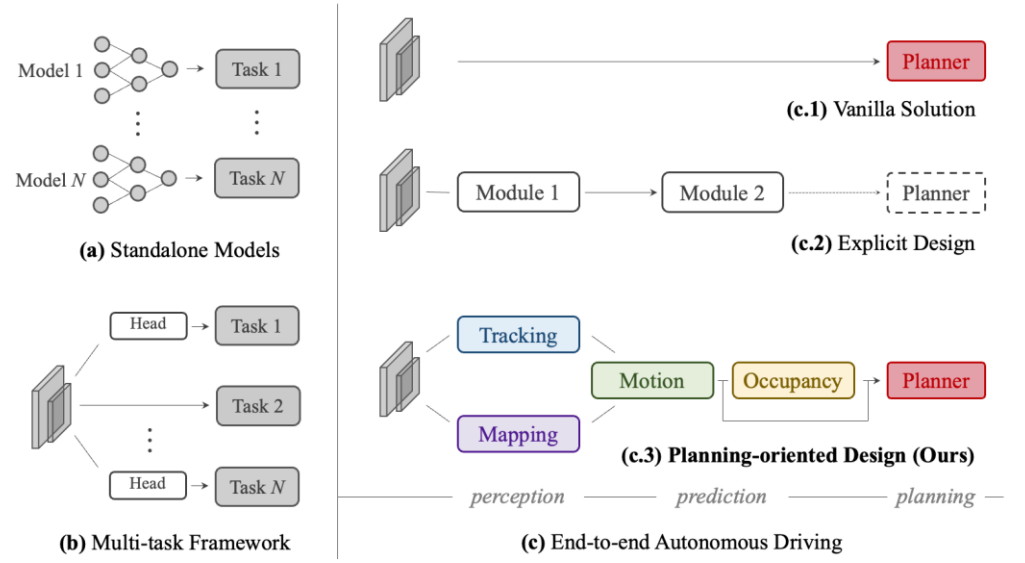
UniAD 的运作流程共包括四个基于 Transformer 的感知和预测模块，以及最后的规划模块（Planner）。流程的第一步是特征提取，由多个摄像头获取图像，经过 BEVFormer 转换为 BEV 特征信息，这一步骤和特斯拉的过程类似。这些信息分别输入感知、预测及规划层：感知层包括 TrackFormer 和 MapFormer，分别提取代理的轨迹信息和整体道路信息，将其输入至预测层的 MotionFormer 与 OccFormer，MotionFormer 捕捉代理与地图、代理与代理和代理与行动目标之间的相互作用，基于此预测每个代理最可能采用的 k 种轨迹，并对轨迹进行平滑以减少预测不确定性，OccFormer 再进一步将 BEV 特征信息与轨迹信息结合，预测是否存在可能被占用的区域以避免出现碰撞；最后，规划层 Planner 根据以上信息决定路线。

论文发表机构之一上海人工智能实验室表示，UniAD 这种端到端多模块融合流程的亮点在于：1）模型通过 Transformer 的多层架构覆盖关键任务，不同 Transformer 层间信息的输入输出融合了每个环节，并行处理多个不同的任务；2）感知与预测环节均包含本车信息（Ego-vehicle），最后 Planner 结合本车信息与 BEV 特征信息进行决策，从而使整个网络均以规划为目标，有效提升解决方案的整体性能。

端到端方案相比模块化方案更利于提高路径规划的效率。行业内自动驾驶解决方案包括模块化系统、多任务模块架构系统和端到端自动驾驶系统三类。据论文所述，模块化系统将一个方案分解成不同模块处理，是目前最常见的解决方案，有助于通过跨团队合作提升研发效率，但存在跨模块信息丢失和误差累积等风险。而多任务模块架构系统虽然将大量任务合并到共享特征提取器中，能够节省芯片的计算成本，但可能会导致负迁移（negative transfer，指一个任务中获得的知识会对下一个任务的表现产生干扰）。端到端系统则是将各模块融合，使前置任务也以规划为目标，能够避免模块化的累计误差及多任务的负迁移问题。

论文中称 UniAD 为首个实现感知、预测和规划三个模块融合的端到端方案。据论文对比，现有的端到端解决方案包括两类：1）基础的端到端方案基于预设的环境条件直接规划路线，在闭环仿真中效果较好，但由于未将代理与环境的交互作用考虑入内，方案难以应对复杂真实场景中可能存在的突发情况；2）按照任务划分网络的显式设计将感知及预测作为关键前置任务纳入模型，以放宽基础方案对环境固定的假设，但目前对模块下子任务（目标检测、目标跟踪、场景建图、轨迹预测、栅格预测和路径规划）的覆盖并不全面。而据论文，UniAD 使输入的传感器数据经过感知、预测和规划模块下六小类子任务后进行输出，相比另两种方案更注重交互和融合，可有效优化以决策为目标的各项任务的表现。

图表137：自动驾驶框架的对比



资料来源：《Planning-oriented Autonomous Driving》（2022，Shanghai AI Laboratory, Wuhan University, SenseTime Research 等）、华泰研究

目前全球自动驾驶芯片供应商可依据其主营业务分为三类：

1）**专注于自动驾驶领域的芯片供应商：**提供软硬件自动驾驶解决方案，如 Mobileye、地平线和黑芝麻等；
2）**通用芯片供应商：**除自动驾驶芯片外，其主营业务往往包括传统芯片或其他消费电子芯片，如英伟达、高通和华为等；
3）**自研芯片车企：**为其自有品牌车辆研发专用芯片，目前主要包括特斯拉。

我们将分别选择自动驾驶专用芯片供应商代表（Mobileye、地平线和黑芝麻）、通用芯片供应商代表（高通和华为）和自研芯片车企代表厂商（特斯拉）进行对英伟达自动驾驶芯片的竞争格局分析，并讨论各厂商的核心竞争力。

Mobileye：ADAS 技术奠基者，“黑箱子模式”优势不再，转型将面临挑战

ADAS 市场领头羊，REM 高精地图打造数据优势。Mobileye 成立于 1999 年，自进入 ADAS 领域起，以视觉感知技术为基础持续迭代其核心产品 EyeQ 系列芯片。凭借以摄像头为主的图像识别技术壁垒，在 ADAS 技术发展早期，大量主流汽车厂商均选择 Mobileye 作为辅助驾驶方案，帮助其迅速拿下了初期的自动驾驶芯片市场份额。Mobileye 在 ADAS 市场占有率在 2015 年一度高达 90%（截至 2023 年 6 月，其占有率约为 70%左右）。截止 2023 年 7 月 1 日，Mobileye SoC 及其解决方案已累计搭载 1.5 亿辆汽车，2023 年上半年出货量为 0.16 亿套。以出货量累计超过 1.5 亿颗，超过 800 种车型搭载，Mobileye 占据 ADAS 市场当之无愧的龙头位置。

2016 年，Mobileye 推出 REM 高精地图服务。REM 以众包方式与车企进行数据采集合作，通过车辆摄像头进行道路网络信息采集，再加密上传到云端进行调整与整合后，生成高精度的地图数据。这标志着 Mobileye 从硬件供应商的角色基础上拥有了数据合作商和服务供应商的双重身份，帮助其在一定程度上占据了无人驾驶的数据入口进行数据储备。2017 年，英特尔以 153 亿美元价格收购 Mobileye，随后于 2022 年分拆再上市，IPO 当日市值达 230 亿美元。

图表138：历代 EyeQ 芯片参数表

型号	上市时间	AI 算力	功率	制程	适用于自动驾驶级别
EyeQ1	2008 年	0.0044 TOPS	2.5W	180nm CMOS	L1
EyeQ2	2010 年	0.026 TOPS	2.5W	90nm CMOS	L1
EyeQ3	2014 年	0.256 TOPS	2.5W	40nm CMOS	L2
EyeQ4	2018 年	2 TOPS	3W	28nm FD-SOI	L2+
EyeQ5	2021 年	15 TOPS	10W	7nm FinFET	L4
EyeQ6 Light	2021 年	5 TOPS	3W	7nm FinFET	L1/L2
EyeQ6 High	2024 年	34 TOPS	-	7nm FinFET	L4
EyeQ ULTRA	2025 年	176 TOPS	<100W	5nm FinFET	L4

资料来源：Mobileye 官网、华泰研究

作为机器视觉龙头，Mobileye 主要占据 L3 以下的自动驾驶芯片市场，但在“软件定义汽车”广泛发展的情况下，“黑箱子模式”优势不再。在面对 ADAS 市场中地平线等厂商的追击，以及英伟达在高阶芯片市场中壁垒的深厚，Mobileye 转型将面临挑战。

Mobileye 的技术积累主要基于以摄像头附带传感器为主的自动驾驶方案。以往 Mobileye 主要采用较为封闭的模式——软硬件结合的解决方案。虽使得产品交付效率提高，但车企无法对“黑箱子”内的软件与算法进行调整与修改，这在自动驾驶渗透率逐渐提升，全栈开发成为车企主流的趋势下，导致其近年原有客户出现流失，包括蔚来 ES8、ES6、EC6 和 2020 款理想 ONE 等车型均曾搭载 EyeQ4，但均已表示与英伟达建立合作伙伴关系，使用 Orin 芯片。另外，与 Mobileye 合作长达十余年的宝马也自 2025 年开始选用高通 Snapdragon Ride 自动驾驶平台芯片作为解决方案，相比目前 Mobileye 已官宣的大算力芯片 EyeQ Ultra（2024 年量产），该平台搭载高通 Snapdragon Ride Flex 芯片，算力达到 2000+TOPS，将在 2024 年开始量产。在 2022 年 CES 上，Mobileye 推出高算力 EyeQ ULTRA SoC（达到 L4 自动驾驶级别），并放弃黑盒模式，与极氪进行 L4 级别车辆的合作开发。不过，Mobileye 与极氪共同开发的 L4 级别车辆将最快于 2024 年上市，而芯片从设计到流片生产至少需要 2-3 年。

图表139：EyeQ 解决方案客户正转向其他厂商

车企	过去		目前		未来	
蔚来	ES8、ES6、EC6	EyeQ4	ES8、ES6、EC6	EyeQ4	ET7	英伟达 Orin
理想	2020 款理想 ONE	EyeQ4	2021 款理想 ONE	地平线 征程 3	X01	英伟达 Orin
长城	第三代哈弗 H6	EyeQ4	WEY 摩卡	EyeQ4	WEY 沙龙机甲龙	高通 Ride 异腾 310
特斯拉	2014 款 ModelS	EyeQ3	全系车型	自研 FSD 芯片	全系车型	自研 FSD 芯片
宝马	1 系、X1 等	EyeQ1-4	iX	EyeQ5	-	高通 Ride
奥迪	第四代 A8	EyeQ3	-	英伟达 Xavier	-	华为
沃尔沃	上一代 XC60	EyeQ1-2	-	-	下一代 XC90	英伟达 Orin
极氪	极氪 001	EyeQ5	极氪 001	EyeQ5	-	英伟达 Thor

资料来源：车东西、华泰研究

图表140：头部自动驾驶芯片参数表

厂商	产品	制程	适用等级	AI 算力
Mobileye	EyeQ4	28nm	L2	2 TOPS
	EyeQ5	7nm	L4	15 TOPS
	EyeQ ULTRA	5nm	L4	176 TOPS
英伟达	Xavier	12nm	L2/L3	30 TOPS
	Orin	8nm	L2-L5	254 TOPS
	Thor	-	L2-L5	2000 TOPS
高通	Ride Flex	-	L4/L5	2000 TOPS
特斯拉	FSD 芯片	14nm	L3	72 TOPS
华为	昇腾 310	12nm	L4	16 TOPS
	昇腾 610	-	L4	512 TOPS
地平线	征程 3	16nm	L1-L2	5 TOPS
	征程 5	16nm	L3-L4	128 TOPS
黑芝麻	A1000	16nm	L0-L2	58 TOPS
	A1000L	16nm	L2+/L3	16 TOPS
	A1000L Pro	16nm	L3-L4	>106 TOPS

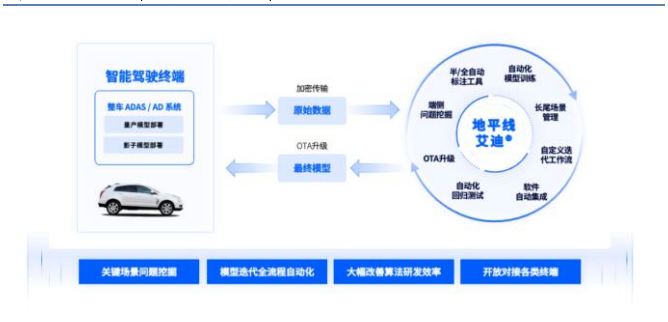
资料来源：各公司官网、华泰研究

地平线：基于 BPU 架构布局自动驾驶生态追击

基于软硬协同构建 BPU（Brain Processing Unit）架构芯片，打造“芯片+工具链+参考算法”的开放技术方案。地平线自动驾驶主要产品线为征程系列芯片，自 2017 年底征程 1 发布以来，地平线始终针对最新的神经网络架构与自动驾驶应用场景，从 AI 模型与开发工具全体系的角度出发，遵循软硬结合的技术路径，对 BPU 计算架构与征程系列芯片进行迭代。同时，地平线基于汽车芯片与工具链，为客户提供用于加速模型训练、模型编译、优化转换和应用部署的天工开物工具链，和用于自动化数据标注与回传、软件自动集成和自动化回归测试与 OTA 升级的艾迪开发工具平台，逐渐形成“芯片+工具链+参考算法”的开放链条方案，助力合作伙伴进行高效开发，实现基于底层硬件能力模型快速迭代。

地平线聚焦 BPU 架构，专用芯片与通用芯片各有所长。云端 AI 芯片更侧重数据支持多种 AI 工作负载和大规模数据吞吐的高拓展能力，自动驾驶芯片则更为注重端侧推理，在低功耗、低延迟性和高计算能效等方面有更多要求。以贝叶斯 BPU 架构为例，其采用脉动张量计算核、大规模异构近存计算和高灵活大并发数据桥等方式进行端侧矩阵运算优化，配合其向量加速单元助力 MAC 阵列利用效率提升，且支持各种自动驾驶场景下 AI 算法，可有效契合自动驾驶场景的高能效比需求。

图表141：地平线艾迪开发平台



资料来源：地平线官微、华泰研究

图表142：地平线天工开物工具链

资料来源：地平线官网、华泰研究

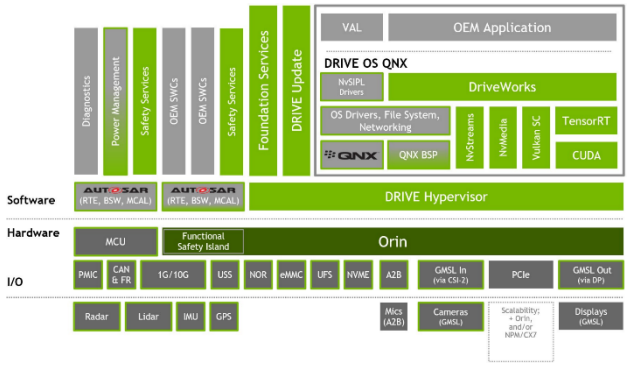
图表143：部分搭载地平线征程系列芯片车型

产品代际	第一代	第二代	第二代	第三代	第四代
芯片	征程 1	征程 2	征程 3	征程 5	征程 6
发布时间	2018 年 4 月	2019 年 8 月	2020 年 9 月	2021 年 7 月	尚未发布
应用架构	BPU 1.0 高斯架构	BPU 2.0 伯努利架构	BPU 2.0 伯努利架构	BPU 3.0 贝叶斯架构	BPU4.0
AI 算力	-	>4 TOPS	5 TOPS	128 TOPS	纳什架构
典型功耗	1.5W	2W	2.5W	30W	>400 TOPS
每帧延时	小于 30ms	小于 100ms	-	60ms	-
功能安全	-	-	-	ASIL-B (D)	ASIL C/D
可靠性	-	AEC-Q100 Grade2	AEC-Q100 Grade2	AEC-Q100 Grade2	-
应用场景	-	座舱交互	L2+辅助驾驶	行泊一体	智能驾驶

资料来源：佐思汽研、华泰研究

软件方面，虽地平线天工开物平台和艾迪开发工具平台可在一定程度上对标英伟达的 CUDA 和 NVIDIA Drive，但生态培养需要大量的技术积累与实际场景数据驱动迭代，因此英伟达的 CUDA 先发优势较明显。英伟达的 GPU 计算核心模块直接对标云端数据场景，且 Orin 集成 CUDA Tensor GPU 开发，与 GPU 底层开发包 CUDA 以及针对深度学习网络优化的软件开发包 TensorRT 绑定，客户可直接在其之上开发相关的自动驾驶软硬件系统。

图表144：NVIDIA Drive OS 提供 CUDA 与 TensorRT 开发包

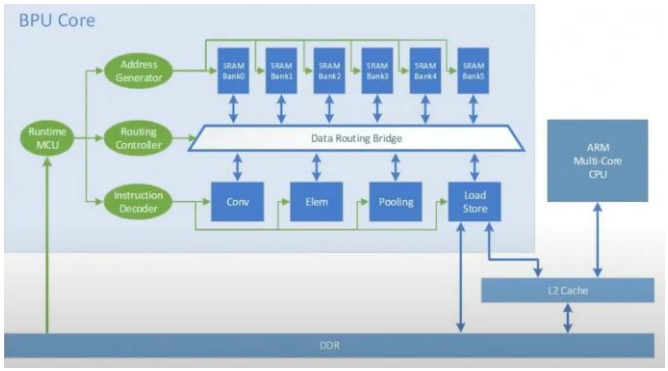


资料来源：英伟达官网、华泰研究

图表145：地平线产品服务生态体系

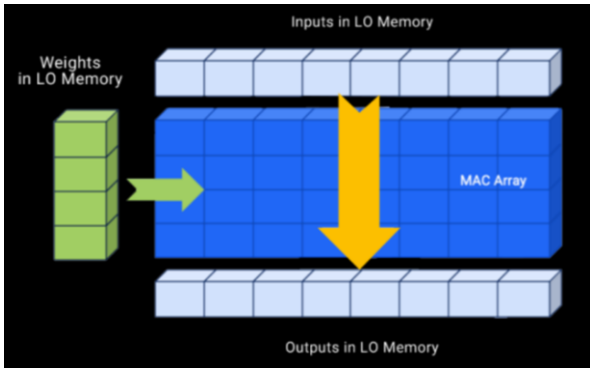
资料来源：亿欧汽车、华泰研究

图表146：地平线 BPU 架构



资料来源：地平线官网、华泰研究

图表147：脉动张量计算核可有效降低计算功耗、延迟和所需数据带宽



资料来源：地平线官网、华泰研究

地平线征程系列定点数量可观。2020 年是搭载地平线征程系列芯片车型的量产元年，截止 2023 年 4 月已与 20 家车企共计 120 款车型达成量产定点合作，**其中，征程 5 从首次量产上车至今的近半年时间里，合作车型增长至 20 款。**

图表148：部分搭载地平线征程系列芯片乘用车车型

品牌	代表车型	搭载产品	量产年份
比亚迪	-	地平线 J5	2023
理想 Li	2021 款 ONE/L7 Pro/L8 Pro/L7 Air/L8 Air	地平线 J3/J5	2021/2023
长安	UNI-T/UNI-V/UNI-K/深蓝 S7/深蓝 SL03	地平线 J2/J3	2020/2021/2021/2023/2023
埃安	埃安 Y	地平线 J2	2021
广汽传祺	GS4 PLUS/GS8/M6/影豹	地平线 J2	2021/2023/2023/2023
江淮	思皓 QX/X8 PLUS	地平线 J2	2021/2023
一汽	红旗	地平线 J5	2023
哪吒汽车	U-Ⅱ/GT	地平线 J3	2023
上汽荣威	第三代 RX5	地平线 J3	2022
自游家	NV	征程 J2/J5	2023
比亚迪	腾势 N7	地平线 J3	2023
上汽大通	MAXUS MIFA 9	地平线 J3	2021
上汽通用五菱	凯捷 HEV	地平线 J2	2023
岚图	FREE	地平线 J2	2021
长城	哈弗 H9	地平线 J2	2021
奇瑞	瑞虎 8PRO、欧萌达 OMODA5	地平线 J3	2022
吉利	博越 L	地平线 J3	2022

资料来源：各公司官网、IT 之家、汽车之家、中国日报、新浪汽车、凤凰新闻、车东西、华泰研究

在高算力市场方面，地平线目前还未进一步印证其在高端车型市场的实际应用能力。2022 年地平线发布的征程 5 芯片算力达到 128 TOPS，单芯片功耗 30W，带领地平线跻身全覆盖 L1-L5 的高算力芯片厂商行列。值得注意的是，搭载征程 5 的首发量产车型为 2023 年 2 月正式发售的理想 L7 Pro/L8 Pro，但在同系列发布的理想高端车型 L7 Max/L8 Max 中，理想则选择搭载两颗英伟达 Orin 作为高级智能辅助驾驶系统 AD Max 的标配。在 2023 年 8 月正式发售的 L9 Pro 与 L9 Max 中，理想同样分别为其搭配了标配智能驾驶 AD Pro（地平线征程 5）和高级智能辅助驾驶系统 AD Max（英伟达 Orin）。

图表149：地平线征程 5 芯片架构

资料来源：地平线官网、华泰研究

图表150：理想 L7 车型自动驾驶配置情况

车型	L7 Pro	L7 Max	L7 Air
全国统一零售价	31.98 万元	33.98 万元	37.98 万元
智能驾驶系统	AD Pro	AD Pro	ADMax
智能驾驶处理器	**地平线征程 5**	**地平线征程 5**	**Orin-X ×2**
激光雷达装备	否	否	是
800 万像素摄像头	1 个	1 个	6 个
200 万像素摄像头	9 个	9 个	5 个
前向毫米波雷达装备	是	是	是

资料来源：理想汽车官网、华泰研究

而在低算力市场中，地平线凭借中国市场的体量规模建立生态，与 Mobileye 抗衡。在产品性能方面，地平线 J2 和 J3 定位为低等级自动驾驶车辆芯片，与全球 ADAS 市场龙头 Mobileye 的 EyeQ4 芯片相比，J2 和 J3 提供的能效比（2 TOPS/W）高于 EyeQ4（0.8TOPS/W），在同等算力所需要配备的散热设备较少，有利于集成上车的应用。但值得注意的是，对比已经实现量产的 7nm 制程 Mobileye EyeQ5，地平线目前所量产芯片没有使用 5nm 和 7nm 工艺，最新征程 5 采用 16nm 工艺。

图表151： 地平线、Mobileye、黑芝麻地算力 SoC 参数对比

	Mobileye		地平线		黑芝麻		
	EyeQ4	EyeQ6 Light	J2	J3	A500	A1000L	A1000
AI 算力（TOPS）	2	5	>4	5	5.8	16	58
功耗（W）	3	3	2	2.5	<2	<5	<8
能效比（TOPS/W）	0.7	1.7	>2.0	2.0	>2.4	>3.0	>7.3
量产时间	2018	2023	2019	2020	2020	2021	2021

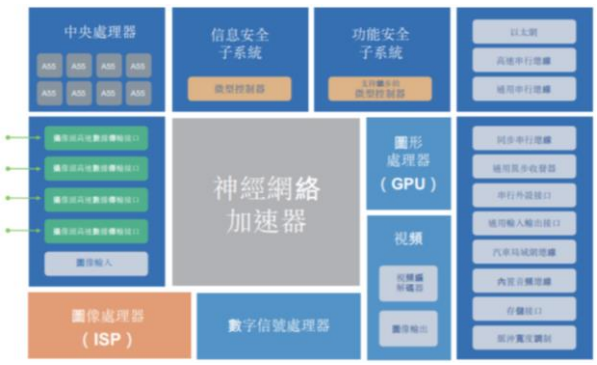
资料来源：各公司官网、华泰研究

生态建设方面，麦肯锡预测，中国未来或成为全球最大的自动驾驶市场，伴随中国市场自动驾驶渗透率逐步提升，相较 Mobileye“黑箱子”模式，地平线秉持其开放的开发生态，利用本土厂商开展服务的优势，已在中国市场获得相当数量的定点客户。**另外，在地图数据方面，**数据作为自动驾驶渐进式发展路径中的决胜点，帮助厂商进行算法迭代，实现产品适用等级的突破。地平线从成立之初便已渐进式发展路径为主，其作为本土企业在国内具有数据采集的天然优势，而 Mobileye 作为外企在国内收集数据较为受限，需与国内车企（如极氪等）进行合作采集高精地图数据。

黑芝麻：第一家递交港股 18C 上市文件的车载芯片股，华山对标英伟达 Orin，武当实现跨域融合

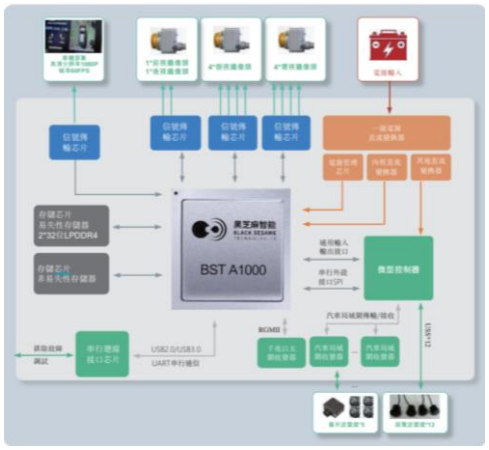
打造“华山”+“武当”两大产品线，为第一家按照港交所 18C 规则递交上市文件的企业。黑芝麻自行研发的 IP、算法和支持软件，打造自动驾驶全栈式解决方案，以满足客户的感知系统需求。目前，黑芝麻产品线已覆盖自动驾驶芯片（华山系列）与跨域计算（武当系列）两大领域。其中，已流片的华山系列芯片目前包括华山一号 A500、华山二号 A1000、A1000L 和 A1000 Pro，所适用的自动驾驶等级分别 L1-L2、L3 和 ADAS/L2+和 L3/L4。2023 年 6 月 30 日，黑芝麻智能向港交所递交的 IPO 申请获受理，计划在港交所挂牌上市，这使得黑芝麻成为 2023 年 3 月 31 日港交所 18C 规则（18C 章的最大特点是容许满足相关市值要求的未商业化特专科技公司上市）生效以来，第一家按此规则正式递交 A-1 上市文件的企业。目前，黑芝麻 C+轮融资的交易后隐含估值折合约 173 亿元。

图表152： 华山 A1000 SoC 内部架构



资料来源：黑芝麻招股书、华泰研究

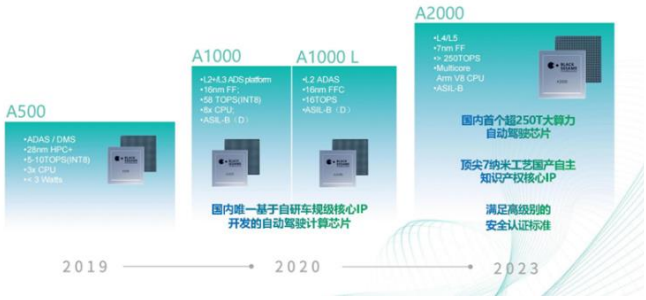
图表153： 华山 A1000 SoC 外部系统架构



资料来源：黑芝麻招股书、华泰研究

华山系列持续探索高算力，武当 C1200 对标英伟达定位跨域融合。黑芝下一代华山 A2000 正在开发中，公司预计将于 2024 年发布。据黑芝麻官方介绍，A2000 采用 7nm 制程，250+ TOPS（INT8）算力，将直接对标英伟达 Orin（7nm 制程，254 TOPS）。另外 2023 年 4 月，黑芝麻推出基于 7nm 制程的武当 C1200，面向自动驾驶、智能座舱和车身控制等计算功能，计划于下半年提供样片，2024 年实现量产，而此前只有英伟达 Thor（2000 TOPS、2025 年量产）定位为自动驾驶和智能座舱的跨域能力。

图表154：黑芝麻华山系列迭代进程



资料来源：21 世纪经济网、华泰研究

图表155：黑芝麻智能武当 C1200

资料来源：黑芝麻官网、华泰研究

定点方面，截至 2023 年 6 月 28 日，黑芝麻已获得 10 家车企及一级供应商的 15 款车型的定点订单，并与超过 30 家车企及一级供应商达成合作关系，如一汽集团（红旗 E001/E202）、东风集团（首款纯电轿车和首款纯电 SUV）和合创汽车（V09）等。

图表156：已公开搭载黑芝麻智能芯片车型情况

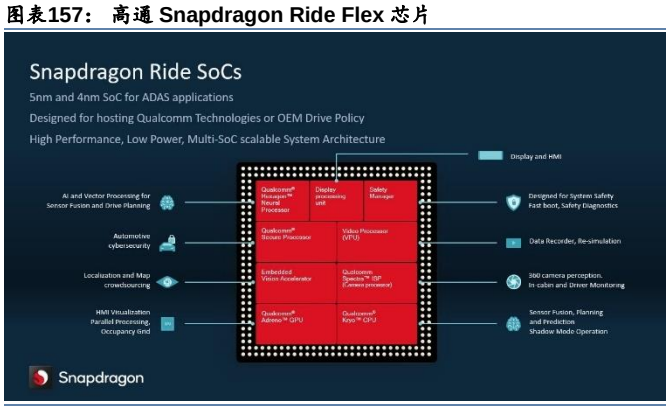
品牌	代表车型	搭载产品	量产年份
江汽	思皓系列多款车型	A1000	2022
东风	首款纯电轿车和 SUV 车型	A1000	2022
吉利	领克 08	A1000	2022
合创	V09	A1000	2023
一汽红旗	E001 和 E202	A1000 L	2024
三一专汽	-	A1000	2023

资料来源：各公司官网、汽车之家、华泰研究

高通：可扩展体系开展差异化竞争，对标英伟达 Thor 打造跨域融合

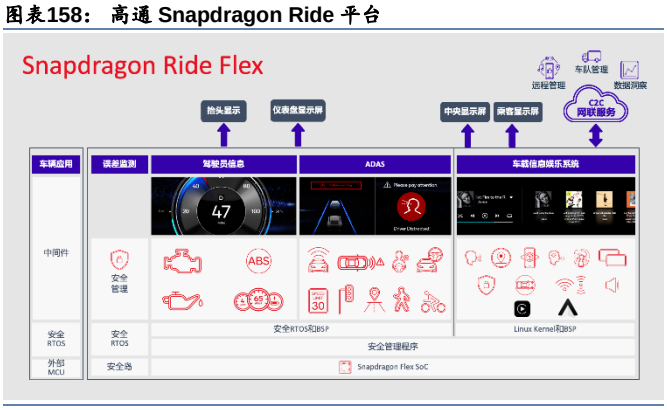
移动通信芯片龙头，降维打击智能座舱市场，逐步扩展至舱驾一体的 SoC。高通汽车产品及解决方案包括数字底盘、座舱平台、自动联网平台、车道云服务平台等。作为移动通信芯片市场的领头羊，高通自 2014 年推出第一代基于 Snapdragon（骁龙）28nm 芯片 620A 的智能座舱平台开始，高通智能座舱产品已迭代至基于 5nm 芯片 SA8295P 的第四代智能座舱平台。凭借其在安卓生态的优势，高通 2019 年发布的 Snapdragon SA8155P 智能座舱芯片基于手机芯片 Snapdragon 855 改进推出，作为全球首个 7nm 制程以下的座舱芯片，SA8155P 至今为应用最广泛的座舱芯片产品之一。2020 年，高通推出算力可达 700 TOPS 的自动驾驶平台 Snapdragon Ride 进行其汽车产品组合扩展。2023 年，高通带来二代 Ride 平台芯片 Snapdragon Ride Flex，其作为舱驾一体智能芯片综合算力可达 2000 TOPS，公司预计将于 2024 年量产上车。

图表157：高通 Snapdragon Ride Flex 芯片



资料来源：高通官网、华泰研究

图表158：高通 Snapdragon Ride 平台



资料来源：高通官网、华泰研究

打造可扩展体系开展差异化竞争，对标英伟达 Thor 打造跨域融合。2021 年 10 月，高通联合投资机构 SSW Partner 以 45 亿美元价格收购瑞典汽车零部件制造商 Weininger，收购完成后，高通将 Weininger 旗下软件部门 Arriver 的辅助驾驶和计算机视觉等业务能力与 Snapdragon 进行逐步整合。高通 2023 年 1 月宣布的新一代 Snapdragon Ride 平台为一个可扩展的自动驾驶加速计算平台，包括 SoC、加速器、视觉系统和自动驾驶软件栈等部分。其中最为主要的产品 Snapdragon Ride Flex SoC 包括 Mid、High 和 Premium 三个级别，采用 4nm 制程，Ride Flex Premium SoC 单芯片 AI 算力可达 600 TOPS 以上，通过双 Flex SoC+双 AI 加速芯片，可达到 2000 TOPS 算力水平，主要定位为舱驾一体芯片，直接对标 2022 年 9 月英伟达宣布的 Thor Soc（2000+ TOPS，未公布制程）。Ride Flex SoC 可用于车内数字座舱的同时，也可实现智能驾驶端的可扩展功能，兼容高通数字底盘平台所包含的 SoC 组合，公司预计 2024 年开始大规模生产。**截至 2023 年 8 月，高通智能驾驶芯片已与长城、通用和奇瑞等国内车企达成合作。**

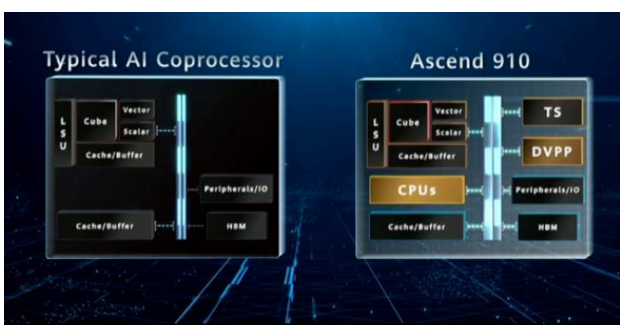
图表159：基于 Snapdragon Flex 芯片打造算力高至 2000TOPS 的可扩展自动驾驶体系

资料来源：高通官网、华泰研究

华为：边缘端 AI 芯片赋能 MDC 计算平台

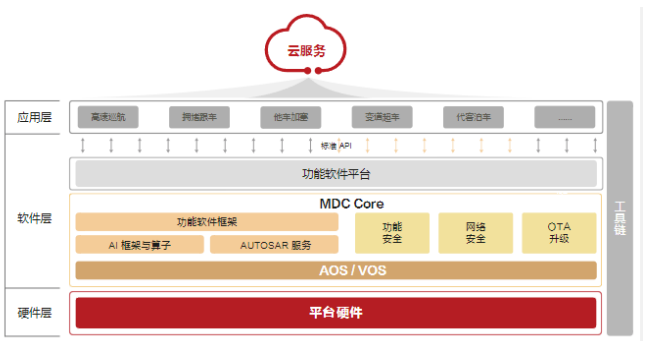
华为海思半导体的边缘端 AI 芯片赋能 MDC 计算平台。华为旗下半导体设计公司海思半导体专注于为华为及其他客户提供高性能、低功耗的芯片解决方案，其产品涵盖通信、消费电子、自动驾驶和物联网等多个领域。2018 年，华为发布首款自动驾驶芯片升腾 310（Ascend 310），其采用 12nm 制程与自研的华为达芬奇芯片结构，主要面向自动驾驶、安防和智制造等边缘计算场景，单芯片算力达 16 TOPS，功耗 8W。2019 年发布的华为升腾 910（7nm）则定位于人工智能全场景，半精度(FP16)算力达到 256 TFLOPS，整数精度(INT8) 算力达到 512 TOPS，可支持包括智能驾驶在内的全栈云边端应用负载。另外，基于 8 颗升腾 310，华为 2018 年发布的 MDC 600 自动驾驶计算平台算力达到 352 TOPS，可处理 16 个摄像头、6 个毫米波雷达、16 个超声波雷达和 8 个 LiDAR 的数据，支持 L3-L4 级别自动驾驶，在当时国内市场中性能力压其他国产厂商产品。

图表160：华为 Ascend 910 与普通 AI 协处理器对比图



资料来源：digitimes 官网、华泰研究

图表161：华为 MDC 计算平台整体架构图



资料来源：华为官网、华泰研究

垂直一体 Tier1 模式，构建自动驾驶生态圈。背靠华为在 ICT 领域 30 余年的研发与生产制造经验，华为 MDC 已作为中央计算平台逐渐被定位为汽车“大脑”，帮助华为构建其 Tier1 厂商角色。2020 年，华为发布智能汽车解决方案 Huawei Inside，重新将企业定位为 Tier1 厂商与车企合作，共同打造智能汽车，MDC 平台则开始致力于构建涵盖产业链上下游的传感器、执行器和应用算法的三大类合作伙伴生态圈，推动产业快速成熟与商用落地。2022 年搭载华为 HI 版的极狐阿尔法 S 的发售标志着华为自动驾驶能力的首番落地，整车搭载华为 MDC 810 计算方案，采用华为鸿蒙 OS 驾驶系统，算力达到 400 TOPS。除北汽极狐外，阿维塔（2022 年发售的阿维塔 11 采用 MDC 810 平台）和广汽埃安（AION LX 采用 MDC 610 平台）等亦为华为 HI 合作伙伴。

图表162：华为以 MDC 智能驾驶计算平台为核心搭建的“平台+生态”全景图

资料来源：华为官网、华泰研究

特斯拉：车企破局者，FSD 和 DOJO 软硬件全栈自研

软硬件全栈自研，历经多轮升级。特斯拉自 2014 年起已与 Mobileye 合作，发布第一代自动驾驶硬件，从早期使用 Mobileye EyeQ3，到英伟达 Drive PX2，再到自研的 FSD 芯片。FSD 芯片的上车标志着特斯拉实现车企从软件到硬件的自主研发。

图表163：各代 Autopilot 硬件配置

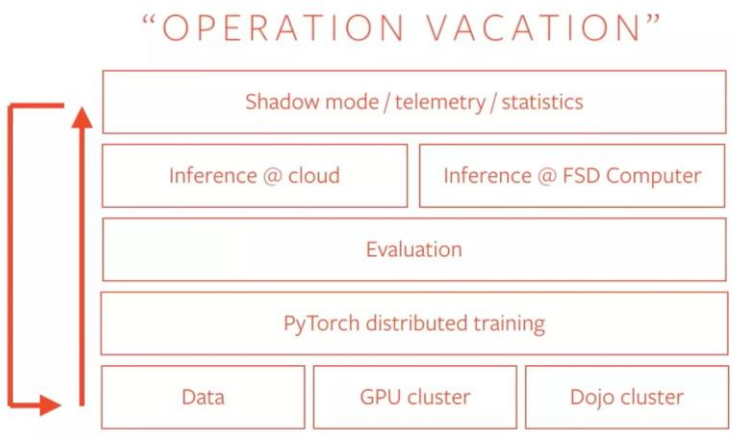
Autopilot Hardware 版本	HW 1.0	HW 2.0	HW 2.5	HW 3.0
日期	2014	2016	2017	2019
摄像头	1 个前视摄像头	8 个摄像头，360° 环视	8 个摄像头，360° 环视	8 个摄像头，360° 环视
毫米波雷达	前向毫米波雷达(博世)	前向毫米波雷达(博世)	前向毫米波雷达(大陆)	前向毫米波雷达(大陆)
超声波雷达	12 个	12 个	12 个	12 个
GPS	GPS&地图	GPS&地图	GPS&地图	GPS&地图
处理器	Mobileye EyeQ3	1-英伟达 Parker SoC	2-英伟达 Parker SoC	2-特斯拉自研芯片
		1-英伟达 Pascal GPU	1-英伟达 Pascal GPU	2-神经网络处理器
		1-英飞凌三核 CPU	1-英飞凌三核 CPU	1-CPU 容错设计
FPS	36	110	110	2300
TOPS	0.256	12	12	144

资料来源：汽车之心、华泰研究

在硬件芯片端，自研芯片有助于特斯拉FSD系统面向其算法与场景进行针对性优化。目前，HW3.0 架构已伴随 FSD 芯片迭代，据 IT 之家 7 月 11 日报道，特斯拉已开始交付其加州弗里蒙特工厂生产的部分 Model Y 车型，这些车型安装有最新版本的自动驾驶计算机系统 HW4.0，实现图像处理和其他 AI 计算。在数据方面，车企身份还有利于特斯拉向车主借力收集数据。截止 2023 年 5 月，特斯拉 FSD beta 累计积累里程已接近 2 亿英里，将帮助其充分训练其神经网络进行软件迭代。不过，我们认为数据多既是好事，但数据的多样化却更为重要。对比谷歌的 Waymo，虽然在数据收集量方面不及特斯拉，但上文提到其对罕见场景的模拟并以其为训练数据，能有效提升处理长尾问题的能力。特斯拉在 2021 年 AI Day 中也曾介绍，特斯拉 Autopilot 也通过大量仿真测试模拟来训练车辆极限交通场景及路况下自动驾驶能力。

另外，在前面章节提到，特斯拉基于 D1 芯片打造的自研超级计算机 DOJO，目标也是为了提高基于车端海量视频与数据的神经网络训练速度与效率。公司预计将在 2024 年投入使用，并预计将进一步搭配 HW 4.0，形成以车主数据为驱动的算法+芯片+软件的三端开发协同，实现针对其算法的芯片架构高效与快速迭代。特斯拉在 2023 年上海人工智慧大会开幕式称在年底前能实现全自动驾驶，不过，CEO Elon Musk 其实早在 2016 年也做出该承诺（根据 The Verge 2016 年 1 月 11 日的报道），因此，我们对于 DOJO 以及全自动驾驶将拭目以待。

图表164：特斯拉自动驾驶自动改进概念图



资料来源：特斯拉官网、华泰研究

重点推荐：英伟达为 AI 芯片行业龙头，AMD 突围有望迎来重估

我们对英伟达和 AMD 采用 PS 的估值逻辑。PS 估值适用于处于高速发展阶段且和技术进步高度相关的企业，这也是我们选择 PS 对两家公司进行估值的原因。我们列出 AI 芯片产业链的相关公司，包括上游和芯片相关的英特尔、博通、高通（QCOM US）、Mobileye（MBLY US）、台积电和阿斯麦（ASML US）；以及下游云计算客户，微软、谷歌、亚马逊和甲骨文（ORCL US）。我们综合考虑英伟达和 AMD 二者业务前景、技术壁垒和市场地位等，对英伟达和 AMD 重点推荐，判断逻辑如下所述：

图表165：AI 芯片行业产业链的相关公司估值（数据截至 2023 年 9 月 20 日）

		PS			PE		
		2023E	2024E	2025E	2023E	2024E	2025E
芯片设计/晶圆代工/半导体设备							
NVIDIA CORPORATION	NVDA US	19.9	13.1	10.8	39.9	25.1	21.0
ADVANCED MICRO DEVICES	AMD	7.2	6.1	5.4	37.1	24.8	19.7
INTEL CORPORATION	INTC	2.9	2.6	2.4	56.7	20.6	15.6
BROADCOM	AVGO	10.1	9.4	8.8	25.3	22.5	19.4
QUALCOMM	QCOM	3.5	3.3	3.0	13.0	11.8	10.6
MOBILEYE	MBLY	14.8	11.6	8.4	54.4	44.5	29.8
TSMC	2330 TT	6.6	5.5	4.6	18.5	14.6	12.1
ASML Holding N.V.	ASML	8.0	7.4	6.2	28.8	25.5	19.7
均值		9.1	7.4	6.2	34.2	23.7	18.5
大型云厂商							
MICROSOFT	MSFT	11.0	9.7	8.5	31.7	28.0	24.1
ALPHABET	GOOGL	5.8	5.1	4.5	24.3	20.5	17.7
AMAZON	AMZN	2.5	2.3	2.0	61.6	41.8	29.8
ORACLE	ORCL	6.1	5.7	5.3	21.2	19.2	17.0
均值		6.3	5.7	5.1	34.7	27.4	22.2

注：预测值参考 Visible Alpha 一致预期
资料来源：Visible Alpha 官网、华泰研究

图表166：重点推荐公司一览表（数据截至 2023 年 9 月 20 日）

股票代码	股票名称	投资评级	收盘价	目标价	预测营收(M USD)			目前市值对应P/S		
			(USD)	(USD)	FY2024	FY2025	FY2026	FY2024	FY2025	FY2026
NVDA US	英伟达	买入	422.39	650	51,725	82,823	114,207	20.17	12.60	9.14
股票代码	股票名称	投资评级	收盘价	目标价	预测营收(M USD)			目前市值对应P/S		
			(USD)	(USD)	2023	2024	2025	2023	2024	2025
AMD US	AMD（超威半导体）	买入	100.34	150	24,188	28,510	31,887	6.70	5.69	5.08

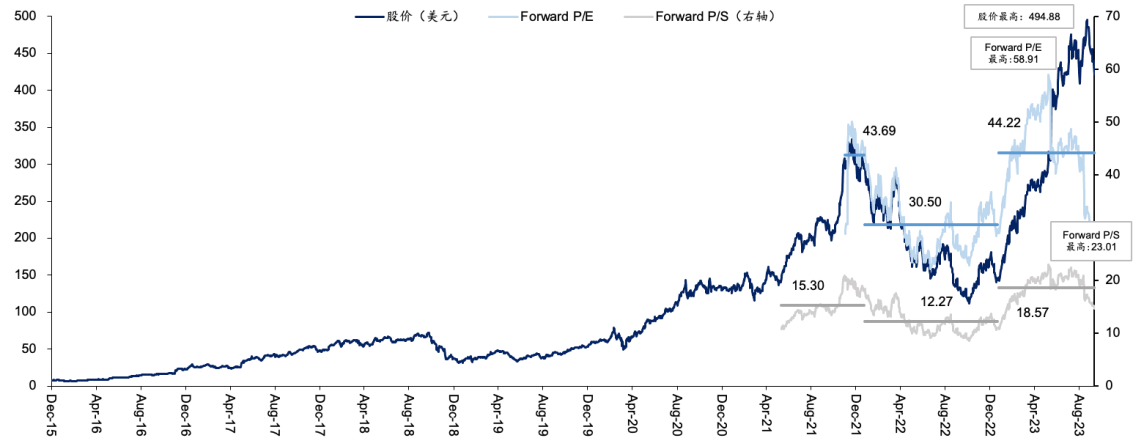
资料来源：Bloomberg、华泰研究

英伟达：AI 龙头软硬一体双护城河（NVDA US，买入，目标价：650.00 美元）

英伟达作为全球数据中心 GPU 的龙头厂商，凭借高算力的硬件及高粘性的软件生态 CUDA，软硬一体平台布局充分受益于 AI 的喷发需求。我们认为，公司中短期内将主要由数据中心业务带动，长期成长性则取决于 AI 商业化应用落地及 AI 芯片竞争格局的发展。我们认为英伟达在 AI 芯片的龙头优势将持续，并带动毛利率和净利率上升，高估值得以支撑。我们预计 FY24/25/26 年营收为 517.3/828.2/1142.1 亿美元，同比 91.8%/60.1%/37.9%。我们给予公司 FY24-25 年动态营收 801 亿美元和 20 倍 PS，目标价 650 美元，首次覆盖给予买入评级。

英伟达的主营业务已从游戏显卡转移到数据中心。英伟达两大业务：1）数据中心营收占比已逾 75%，为公司主要营收和盈利贡献。台积电对未来五年 AI 服务器销售额 CAGR ~50%的预测，加上 2024 年先进封装 CoWoS 产能将迎来翻倍，以及美国加息步入尾声，均表明需求旺盛。我们预计 23/24/25 年营收 378/654/916 亿美元，同比 152%/73%/40%。2）游戏显卡逐渐从挖矿影响中恢复，叠加 PC 市场下滑收窄、疫情间高基数开始消化和高期待新游戏带动，我们预计 23/24/25 年营收 109/136/177 亿美元，同比 20%/25%/30%。公司 23-25 年毛利率和净利率受益于业务转移，将分别从 70%提升到 73%及 47%提升到 50%。

图表167：2016 年至今英伟达历史股价、Forward PE 和 Forward PS（数据截至 2023 年 9 月 20 日）



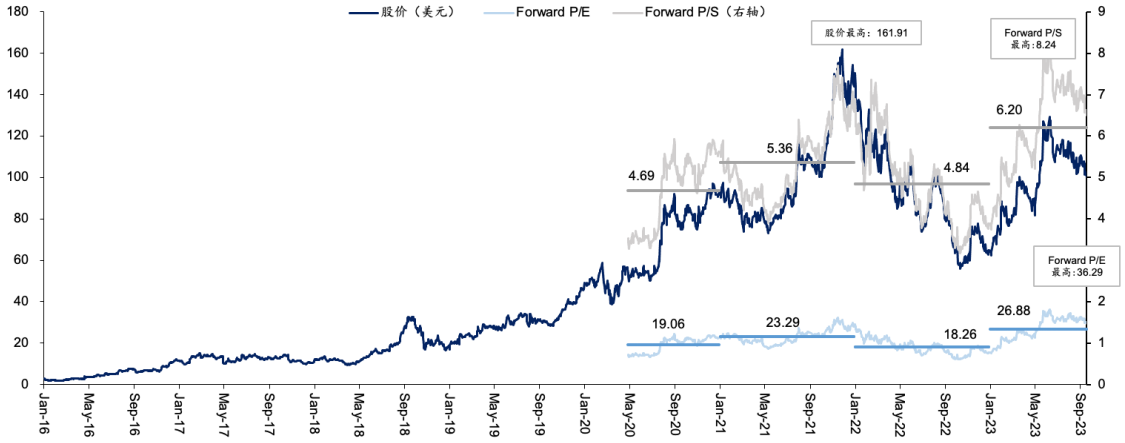
资料来源：Bloomberg、华泰研究

超威半导体：AI 新赛道为重估之钥（AMD US，买入，目标价：150.00 美元）

2016 年开始 AMD 突围英特尔，在抢占份额中估值不断攀升，从 3x PS 到 21 年 5-6x。如今突围二战一触即发，面对 AI 新机遇 AMD 有望再次来到重估分水岭。我们预测 AMD 23/24/25 年营收为 241.9/285.1/318.9 亿美元，同比 2.5%/17.9%/11.8%。对比竞争对手英特尔和英伟达 2024 年的 PS 分别为 2.6 倍和 14.6 倍，AMD 在 CPU 制程上仍领先英特尔，但在 GPU 却奋力追赶英伟达，因此我们认为估值应在两家之间且略低于两家平均值，对比自身历史估值（自 2020 年开始 PS 一直处于 5-6 倍）则上修可期。我们给予 AMD 24 年 8.5x PS，对应目标价 150 美元，首次覆盖给予买入评级。

AMD 以数据中心为茅，游戏和嵌入式为盾，客户端逐渐恢复，毛利率提升。AMD 四大业务：1）数据中心：AI 之风继续吹，CPU 制程仍领先英特尔，MI300 系列有力冲击英伟达，看好 CPU 和 GPU 均能提升份额，预计 23/24/25 年营收 76.0/104.1/120.9 亿美元，同比 26%/37%/16%；2）游戏：疫情间高基数开始消化、主机“半代升级”和高期待新游戏带动维持增长，预计 23/24/25 年营收 66.0/68.3/70.7 亿美元，同比-3%/3.5%/3.5%；3）客户端：随 PC 市场下滑收窄而回暖，预计 23/24/25 年营收 43.0/47.3/52.0 亿美元，同比-31%/10%/10%；4）嵌入式：赛灵思并表效应消退后进入平稳阶段，预计 23/24/25 年营收 56.9/65.4/75.3 亿美元，同比 25%/15%/15%。

图表168：2016 年至今 AMD 历史股价、Forward PE 和 Forward PS（数据截至 2023 年 9 月 20 日）



资料来源：Bloomberg、华泰研究

图表169：报告中提及上市公司一览（除英伟达和 AMD）

公司名称	股票代码	公司名称	股票代码	公司名称	股票代码
谷歌(Google)	GOOS US	奥特斯(AT&S)	ATS AV	3M	MMM US
亚马逊(Amazon)	AMZN US	南亚科技(Nanya)	2408 TT	JSR	4185 JT
微软(Microsoft)	MSFT US	景旺电子	603228 CH	Nissan	7201 JT
台积电(TSMC)	TSM US/2330 TT	沪电股份	002463 CH	Sumitomo Chemical	4005 JT
Tazmo	6266 JT	生益科技	600183 CH	东京应化工业(T.O.K)	8035 JT
Shibaura	6590 JT	SK 海力士(SK hynix)	000660 KS	SK Materials	036490 KS
博通(Broadcom)	AVGO US	美光(Micron Technology)	MU US	Air Products	APD US
Meta	META US			Central Glass	4044 JT
百度	BIDU US	AVC	3017 TT	Nippon Sanso Taiwan	4091 JT
苹果(Apple)	AAPL US	双鸿科技	3324 TT	Cabot Microelectronics	CBT US
阿里巴巴	BABA US	健策精密	3653 TT	AGC	5201 JT
寒武纪	688256 CH	胜宏科技	300476 CH	Fujibo	3104 JT
英特尔(Intel)	INTC US	奥士康	002913 CH	Fujimi	5384 JT
爱彼迎(Airbnb)	ABNB US	浪潮	600756 CH	英特格公司(Entegris)	ENTG US
Snap	SNAP US	联想	0992 HK	富士电子材料集团(Fujifilm)	4901 JT
Sprinklr	CXM US	IBM	IBM US	RASA 工业	3023 JT
Money Forward	3994 JT	广达	2382 TT	德山株式会社(Tokuyama)	4043 JT
特斯拉(Tesla)	TSLA US	纬创	3231 TT	胜一化工(Shiny)	1773 TT
金山云	KC US	英业达	2356 TT	信越半导体(Shin-Etsu Handotai)	4063 JT
腾讯	00700 HK	美超微(Super Micro Computer)	SMCI US	世创电子(Siltronic)	WAF GY
戴尔(DELL)	DELL US	惠与(HPE)	HPE US	胜高(SUMCO)	3436 JT
神州数码	000034 CH	思科	CSCO US	斐电(Ibiden)	4062 JT
大联大	3702 TT	高通(QUALCOMM)	QCOM US	景硕(Kinsus)	3189 TT
日月光	ASX US	Mobileye	MBLY US	环球晶圆(Global Wafers)	6488 TT
联华电子	UMC US	ASML	ASML US	杜邦公司(DuPont)	DD US
Amkor Technology	AMKR US	Oracle	ORCL US	台塑胜高(Formosa Sumco Technology Corporation)	3532 TT
通富微电	002156 CH	巴斯夫	BAS GY	液化空气集团(Air Liquide)	AI FP
三星电子(Samsung)	009150 KS	默克集团(Merck)	MRK US		
欣兴电子(Unimicron)	3037 TT	华立企业股份(Wah Lee)	3010 TT		

资料来源：Bloomberg、华泰研究

图表170：报告中提及未上市公司一览

公司名称		公司名称		公司名称	
Cerebras	未上市	BCG	未上市	亚泰半导体	未上市
Graphcore	未上市	Midjourney	未上市	矽得荣(SK Siltron)	未上市
字节跳动	未上市	InstaDeep	未上市	关东鑫林科技股份(Kanto)	未上市
OpenAI	未上市	Allen Institute for AI	未上市	广明实业	未上市
Habana	未上市	Finch Computing	未上市	普莱克斯（Praxair）	未上市
SambaNova	未上市	Qualtrics	未上市	联华林德(Linde LienHwa)	未上市
Mosaic ML	未上市	SeaMicro	未上市	茂泰利科技股份有限公司	未上市
Hugging Face	未上市	Cirrascale	未上市		

资料来源：Bloomberg、华泰研究

风险提示

AI 技术落地和推进不及预期。自 ChatGPT 落地应用并取得一定成功，各科技巨头均加快和加大力度布局 AIGC 领域，如 Meta 于 23 年年初建立 AIGC 团队、微软也在其 Azure、Bing 等多项自有业务进一步整合 AI 技术。由于人工智能属于高新技术，需投入较大前期研发成本和时间，后续 AI 技术落地可能会受企业投入、宏观经济、政策和舆论等多方面影响，致使研发进度不及预期。

行业竞争激烈。目前生成式 AI 技术仍处行业发展前期，文字、图片、视频等单一及多模态大模型不断推出，赋能聊天、搜索引擎、编辑代码等多类应用，行业暂未形成较为稳定的竞争格局，竞争激烈。若后续市场竞争进一步加剧，部分企业未能及时推出相关产品或技术研发不及预期，可能会受激烈竞争影响而导致市场出清。

中美竞争加剧。中美两国作为人工智能领域发展较为领先的两个国家，其本土多家企业均积极部署 AI 领域相关技术和产品，推动 AIGC、LLM 等尖端技术落地应用。若后续中美两国间竞争加剧，可能会阻碍 AI 产业相关应用的进一步推广。

文中提及未覆盖个股相关信息数据来自于公开渠道，不代表对相关公司的研究覆盖和推荐。

AI 芯片龙头双护城河，能否继续一枝独秀？

华泰研究 首次覆盖
2023 年 9 月 22 日 | 美国 半导体

投资评级（首评）：买入
目标价（美元）：650.00

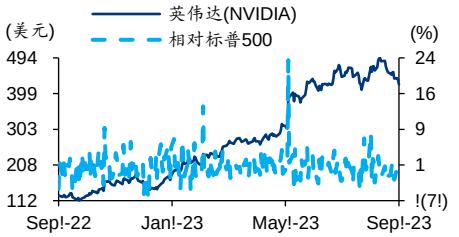
研究员 何翩翩
SAC No. S0570523020002 purdyho@htsc.com
SFC No. ASI353 +(852) 3658 6000

华泰证券研究所分析师名录

基本数据

目标价（美元）	650.00
收盘价（美元 截至 9 月 20 日）	422.39
市值（美元百万）	1,043,303
6 个月平均日成交额（美元百万）	18,947
52 周价格范围（美元）	108.06-502.62
BVPS（美元）	8.96

股价走势图



资料来源：S&P

AI 龙头软硬一体双护城河，首发给予买入评级，目标价 650 美元

英伟达作为全球数据中心 GPU 的龙头厂商，凭借高算力的硬件及高粘性的软件生态 CUDA，软硬一体平台布局充分受益于 AI 的喷发需求。我们认为，公司中短期内将主要由数据中心业务带动，长期成长性则取决于 AI 商业化应用落地及 AI 芯片竞争格局的发展。我们认为英伟达在 AI 芯片的龙头优势将持续，并带动毛利率和净利率上升，高估值得以支撑。我们预计 FY24/25/26 年营收为 517.3/828.2/1142.1 亿美元，同比 91.8%/60.1%/37.9%。我们给予公司 FY24-25 年动态营收 801 亿美元和 20 倍 PS，目标价 650 美元，首次覆盖给予买入评级。

英伟达主营业务已从游戏显卡转移到数据中心，带动毛利率和净利率上升

英伟达两大业务：1）数据中心营收占比已逾 75%，为公司主要营收和盈利贡献。台积电对未来五年 AI 服务器销售额 CAGR ~50%的预测，加上 2024 年先进封装 CoWoS 产能将迎来翻倍，以及美国加息步入尾声，均表明需求旺盛。我们预计 23/24/25 年营收 378/654/916 亿美元，同比 152%/73%/40%。2）游戏显卡逐渐从挖矿影响中恢复，叠加 PC 市场下滑收窄、疫情间高基数开始消化和高期待新游戏带动，我们预计 23/24/25 年营收 109/136/177 亿美元，同比 20%/25%/30%。公司 23-25 年毛利率和净利率受益于业务转移，将分别从 70%提升到 73%及 47%提升到 50%。

英伟达 AI 芯片供应端瓶颈或已消除，但需求端商业应用落地能否跟上？

台积电 CoWoS 先进封装产能瓶颈或已消除。据 Digitimes 在 7 月 14/21 日报道，台积电 23/24 年产能达 12/24 万片。英伟达或会将封装外溢到 Amkor 和联电。然而，AI 能否大规模商业落地，切实为企业降本增效，才是持续增长的关键。以英伟达三大模型为例，NeMo 赋能办公等生产力提升；BioNeMo 可促进医疗发展和医药研发；Picasso 则针对图像和传媒领域。目前在一些对精准度要求不高的行业，如电影，推进节奏明显较快，但对精准度要求高的行业，如金融，应用则较为谨慎。随着微软 Co-pilot 等商业应用软件的普及，我们看好在办公、教育、医疗和金融等领域的不断渗透。

多方入局 AI 芯片竞争已趋白热化，英伟达能否保持领先？

在 AI 训练端，英伟达的高算力 GPU 一直为首选。我们认为只有少数 AI 芯片能与其匹敌，如谷歌 TPU 和 AMD MI300。当算法开始稳定和成熟，ASIC 定制芯片凭着专用性和低能耗，也能承接部分算力。头部云及互联网厂商考虑到削减 TCO、提升研发可控性及集成自身生态圈等，均在推进自研芯片，我们认为或将成为英伟达最大的竞争对手。推理市场规模大，但对算力要求比训练较低，因此百花齐放，英伟达也希望能在当中复制其在训练端的成功。目前推理还是以 CPU 主导，但 GPU/FPGA/ASIC 等也能占到一席位。

风险提示：AI 落地缓慢、中美摩擦、芯片需求不及预期、宏观不确定性。

经营预测指标与估值

会计年度	2022	2023	2024E	2025E	2026E
营业收入（美元百万）	26,914	26,974	51,725	82,823	114,207
+/-%	61.00	0.22	91.76	60.12	37.89
归属母公司净利润（美元百万）	9,752	4,368	24,375	40,762	56,742
+/-%	125.12	(55.21)	458.04	67.23	39.20
EPS（美元，最新摊薄）	3.85	1.74	9.87	16.50	22.97
ROE (%)	44.83	17.93	71.09	60.97	49.08
PE（倍）	108.07	240.50	42.80	25.59	18.39
PB（倍）	39.60	47.53	22.45	11.96	7.25
EV EBITDA（倍）	76.24	99.94	39.35	23.77	16.36

资料来源：公司公告、华泰研究预测

核心投资逻辑

区别于市场的观点 1：市场普遍认为英伟达后续弹性的关键在于供应端的瓶颈，我们则认为需求端 AI 能否大规模商业落地，切实为企业降本增效，才是英伟达持续增长的关键。以英伟达三大模型为例，NeMo 赋能办公等生产力提升；BioNeMo 可促进医疗发展和医药研发；Picasso 则针对图像和传媒领域。目前在一些对精准度要求不高的行业，如电影，推进节奏明显较快，但对精准度要求高的行业，如金融，应用则较为谨慎。随着微软 Co-pilot 等商业应用软件的普及，我们看好在办公、教育、医疗和金融等领域的不断渗透。

区别于市场的观点 2：市场对 AI 芯片的竞争格局的讨论较少，我们认为多方入局下 AI 芯片竞争已趋白热化。在 AI 训练端，英伟达的高算力 GPU 一直为首选。我们认为只有少数 AI 芯片能与其匹敌，如谷歌 TPU 和 AMD MI300 系列。当算法开始稳定和成熟，ASIC 定制芯片凭着专用性和低能耗，也能承接部分算力。因此，头部云计算及互联网大厂出于削减 TCO、提升研发可控性及集成生态等考量，均陆续发力自研芯片，我们认为或将成为英伟达最大的竞争对手。另外，初创企业如 Cerebras、Graphcore 等，以及芯片行业以外的企业，包括特斯拉的 DOJO 等，正在异军突起。AI 推理市场规模大，但对算力要求比训练较低，因此百花齐放，英伟达也希望能在当中复制其在训练端的成功。在大模型和多模态趋势下 GPU 或能夺份额。但目前推理端还是以 CPU 主导，GPU/FPGA/ASIC 等也能占到一席位。

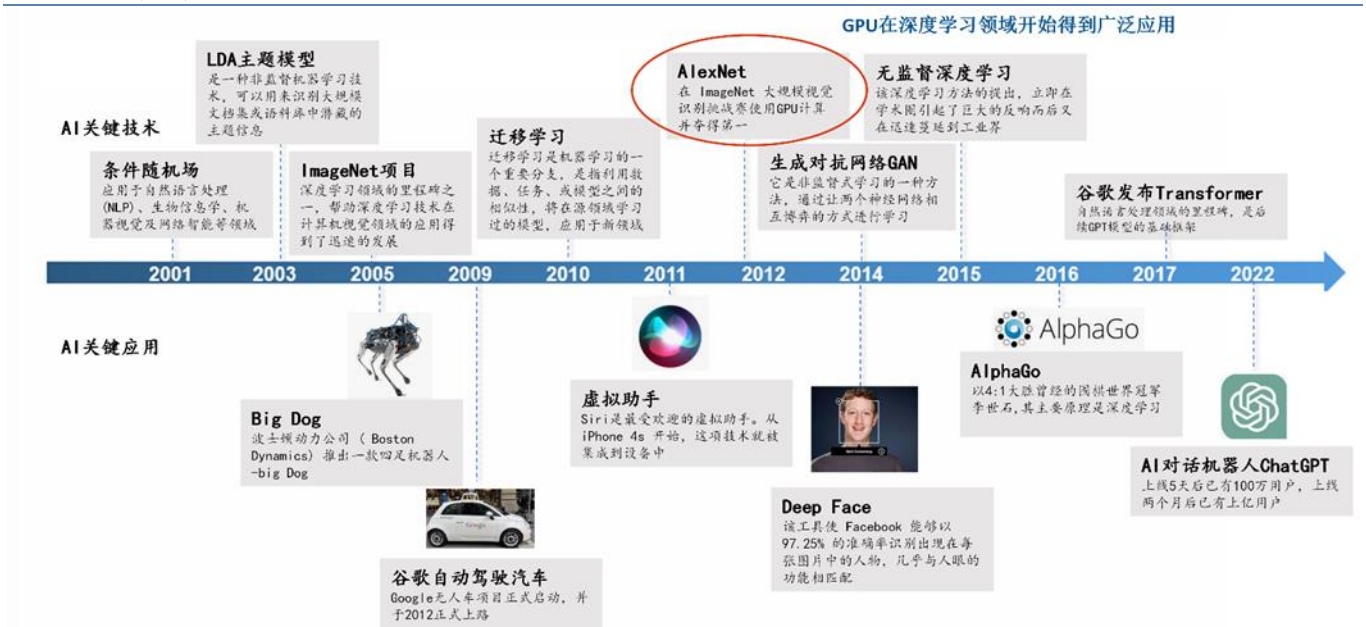
核心逻辑：我们认为英伟达作为全球数据中心 GPU 的龙头厂商，凭借高算力的硬件及高粘性的软件生态 CUDA，软硬一体平台布局充分受益于 AI 的旺盛需求。公司中短期内将主要由数据中心业务带动，长期成长性则取决于 AI 商业化应用落地及 AI 芯片竞争格局的发展。我们认为英伟达在 AI 芯片的龙头优势将持续，并带动毛利率和净利率上升，高估值得以支撑。英伟达主营业务已从游戏显卡转移到数据中心，带动毛利率和净利率上升。英伟达两大业务：1）数据中心营收占比已逾 75%，为公司主要营收和盈利贡献。台积电对未来五年 AI 服务器销售额 CAGR ~50%的预测，加上 2024 年先进封装 CoWoS 产能翻倍的扩产计划，以及美国加息步入尾声，均表明需求旺盛。2）游戏显卡逐渐从挖矿影响中恢复，叠加 PC 市场下滑收窄、疫情间高基数开始消化和高期待新游戏带动。

英伟达：从硬件到软件全面布局，引领 AI 辉煌时代的双护城河

GPU 是近代人工智能发展的重要推动力。在 AI 里深度学习的发展历程中，GPU (Graphics Processing Unit) 扮演了至关重要的角色。机器学习是 AI 的核心组成部分，而深度学习则是机器学习的重要分支。AI 始于上世纪五六十年代，但直到 21 世纪，从过去十年左右开始，才随着深度学习的应用走入寻常百姓家。人们开始在生活中享受到语音助手如 Siri 和 Alexa 等、人脸识别如手机开启和机场安检及 ChatGPT 等应用的便利。这一切科技发展都离不开英伟达的 GPU（硬件）和 CUDA（软件生态）技术的部署。GPU 的高算力以及并行计算能力，使得深度学习算法能够高效地处理海量数据，并提高了计算的效率，让深度学习得以获得快速发展。

英伟达 GPU，图像之花成长为 AI 计算的参天大树。英伟达成立于 1993 年，是图形处理器 GPU (Graphics Processing Unit) 的发明者。图形处理器最初的设计理念是用来进行图形渲染，其中涉及处理大量基于矩阵的多维度计算。而在深度学习，特别是在神经网络中的前向和后向传播，均涉及到大量矩阵运算。因此，拥有强大并行计算能力的 GPU 跟执行深度学习任务高度匹配。此外，公司还率先预见到 GPU 在人工智能市场的广泛应用前景，在 2006 年通过投入 CUDA 软件生态项目，大幅提升了 GPU 的可编程性，让 GPU 的通用计算时代得以展开。GPGPU (General-Purpose Graphics Processing Unit) 是专用于运算协作处理的通用图像处理器。跟传统的图形处理 GPU 相比，GPGPU 削减了图形显示的部分单元，并增加了专用向量、张量、矩阵运算指令等加速计算性能，以满足 AI 等加速计算的需求。英伟达自成立以来，不断进行技术创新和产品升级，不仅在游戏显卡领域取得了成功，还将 GPU 应用进一步拓展至汽车、数据中心以及专业视觉化领域，加速全球人工智能领域的研发与应用。英伟达也不断在加速软件和扩展计算，以三种芯片—GPU、CPU 和 DPU 为硬件基石，实现跨多 GPU 和多节点扩展，将平台从 PC 扩展到超级计算中心、数据中心、云和边缘环境，并通过芯片、软件、平台等全栈式布局，成为全球 AI 主导者。

图表171：从千禧年代开始的 AI 技术及应用梳理



资料来源：algotive.ai、华泰研究

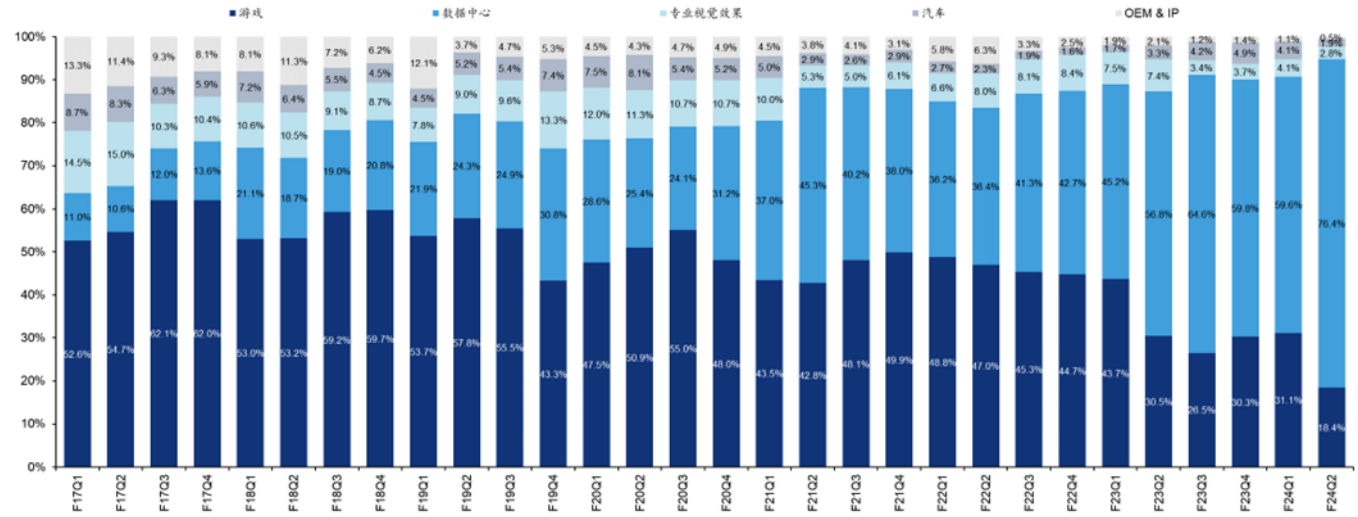
图表172：英伟达关键创新历程（数据取自 2016/3/11 至 2023/9/20）

资料来源：Wind、英伟达官网、华泰研究

FY24Q2 营收盈利大超市场预期，三季度指引再创新高。英伟达 FY24Q2 营收同比增长 101%至 135.1 亿美元，Non-GAAP EPS 同比增长 429%至 2.70 美元，远超彭博一致预期的 112.2 亿美元和 2.07 美元。公司发布下季度营收指引为 156.8-163.2 亿美元，大幅超过市场预期的 126.1 亿美元。数据中心营收同比高增 171%至 103.2 亿美元，环比增长 141%，高于彭博一致预期的 79.81 亿美元；游戏营收同比增长 22%至 24.9 亿美元，环比增长 11%，持续从挖矿影响中改善，高于彭博一致预期的 23.80 亿美元；汽车业务营收同比增长 15%至 2.53 亿美元，环比下降 15%，低于预期的 3.19 亿美元。展望 2024 年，微软、谷歌、亚马逊和 Meta 均在近一季度业绩会中提及到将加大 AI 方面的资本开支投入，其中 AI 基础设施将是重点投入的领域，故从宏观角度来看，若下半年美国加息步入尾声，各大云厂商的 Capex 也将迎来修复。

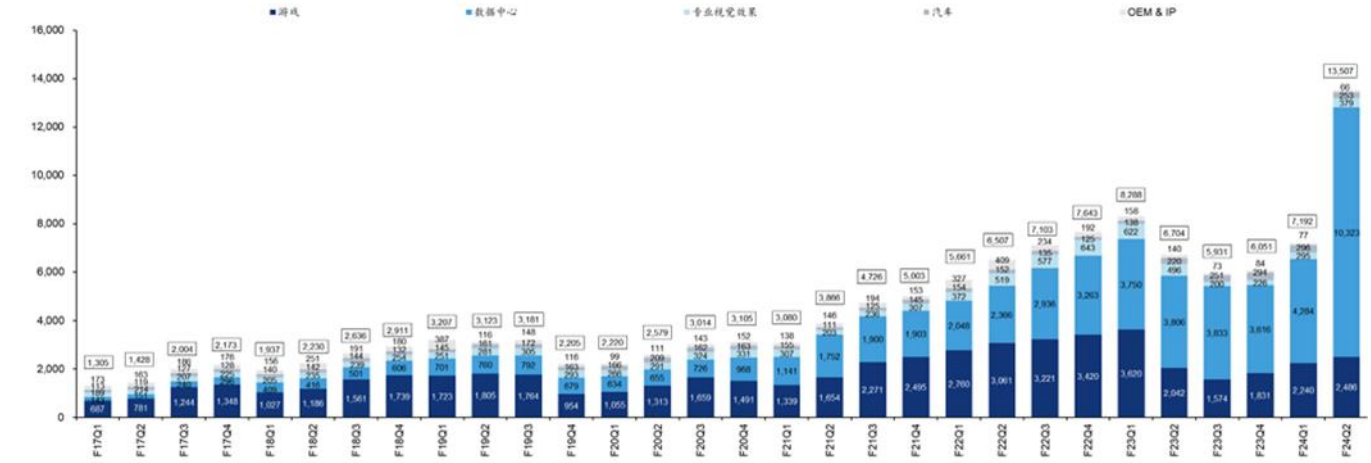
我们预计英伟达 FY24/25/26 年营收为 517.3/828.2/1142.1 亿美元，同比 91.8%/60.1%/37.9%。我们给予公司 FY24-25 年动态营收 801 亿美元和 20 倍 PS，目标价 650 美元，首次覆盖给予“买入”评级。我们预计数据中心业务 23/24/25 年营收为 378/654/916 亿美元，同比 152%/73%/40%。我们预计游戏显卡业务 23/24/25 年营收为 109/136/177 亿美元，同比 20%/25%/30%。我们也预计公司 23-25 年毛利率和净利率受益于业务转移，将分别从 70%提升到 73%及 47%提升到 50%。

图表173： 英伟达各业务收入占比



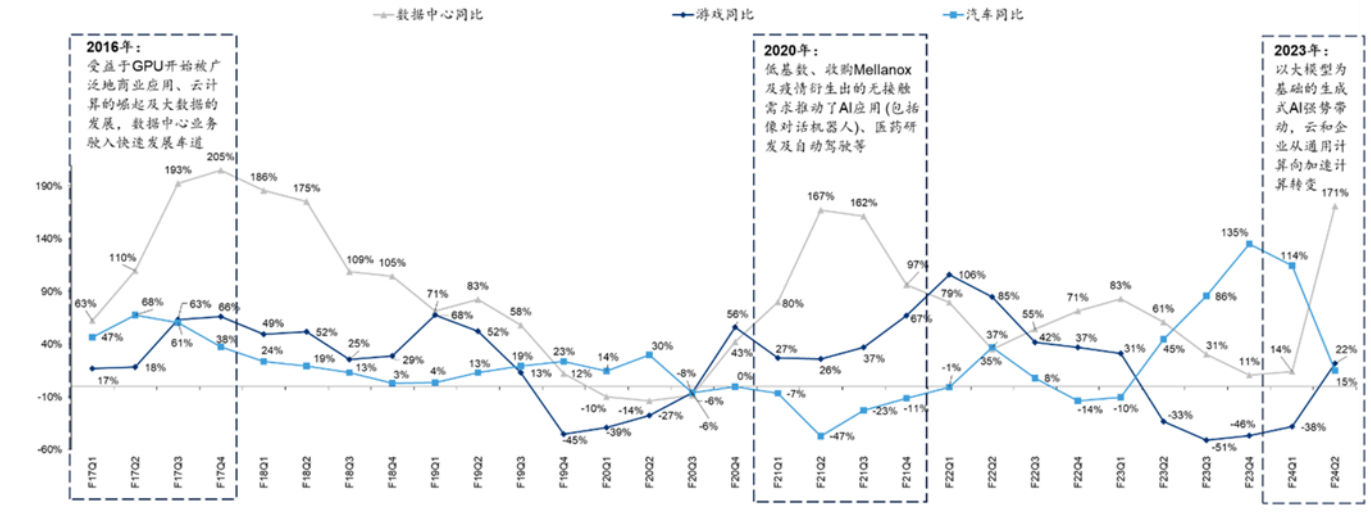
资料来源：Bloomberg、华泰研究

图表174： 英伟达分业务营收（单位：百万美元）



资料来源：Bloomberg、华泰研究

图表175： 英伟达游戏、数据中心、汽车业务同比增速

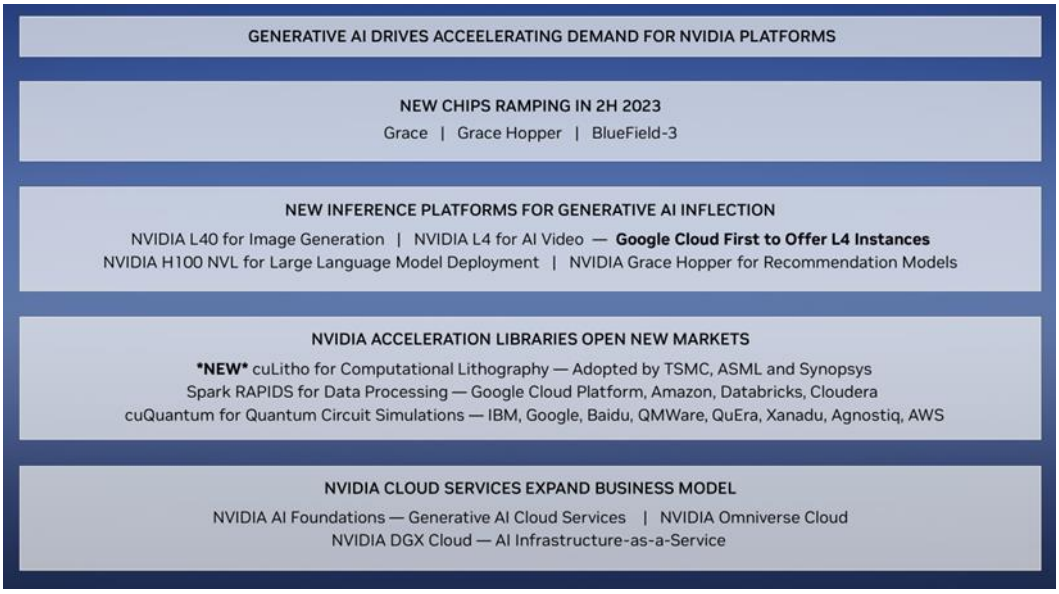


资料来源：Bloomberg、华泰研究

2023 以来推出多款重磅新品，引领加速计算性能突破

英伟达于 2023 GTC 里全面拥抱 AI 产业。2023 年 3 月 18 日的 GTC（GPU Technology Conference）被 NVIDIA 创始人兼首席执行官黄仁勋描述为“迄今为止最重要的一次 GTC”，大会的核心亮点可总结为 5 个方面：1）生成式 AI 将不断推动数据中心平台需求的增长；2）新芯片产品将在 2023 年下半年助推收入增长，在 2022 年 3 月推出的 Grace CPU 与用于推荐系统的 Grace Hopper 将在今年内正式发售，BlueField-3 芯片将在 2023 年全面投产；3）四大推理平台加速生成式 AI：于 2022 年 10 月发布的用于图像生成的 L40，两款推理端新产品（用于 AI 视频的 L4、用于大型语言模型部署的 H100 NVL），以及适用于推荐模型的 NVIDIA Grace Hopper；4）发布全新的加速库（如用于计算光刻的 cuLitho 等）；5）拓展了新的云服务商业模式，如 AI IaaS 的 DGX Cloud、针对生成式 AI 的 AI Foundations、以及 Omniverse Cloud。

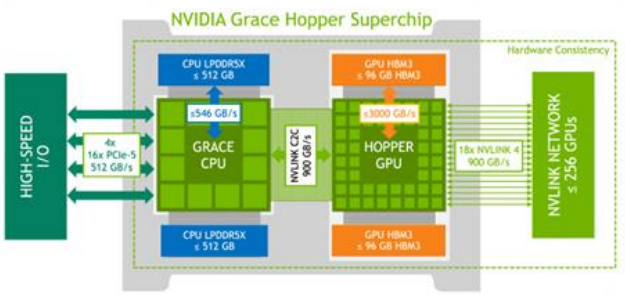
图表176： NVIDIA 2023 GTC 核心亮点



资料来源：英伟达官网，华泰研究

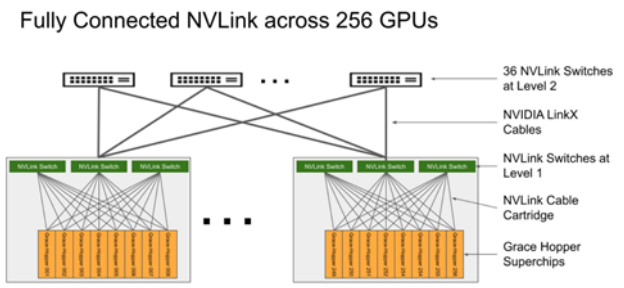
2023 COMPUTEX 大会继续发布 AI 新产品，先进的加速计算+网络技术，为吞吐量和可扩展性迎来新突破。英伟达在 2023 年 5 月 29 日的 COMPUTEX 大会上更新了多款 AI 算力产品，当中焦点落在 DGX GH200 超级计算系统上。该系统是通过 NVLink 互连技术及 NVLink Switch System，串联 32 台由 8 块 GH200 超级芯片（总计 256 块）合并而成的单一超级计算系统，存储器容量高达 144 TB，我们认为**大规模的共享内存能解决 AI 大模型训练的关键瓶颈，**将为生成式 AI 语言应用、推荐系统和数据分析工作负载的大模型增添动力。Google Cloud、Meta 与微软将是其首批用户。从具体参数上看 DGX GH200 性能优异，DGX GH200 可提供高达 1 exaFLOPS（1000 petaFLOPS）的算力。在 2023 年底，结合 Quantum-2 InfiniBand 技术与 4 台 DGX GH200 的 AI 超级计算机 NVIDIA Helios（含 1024 = 4*256 个 GH200 超级芯片）将会推出，我们认为或标志着英伟达在 AI 和数据分析工作负载加速计算的又一突破。

图表177： 英伟达 Grace Hopper 架构示意图



资料来源：英伟达官网、华泰研究

图表178： NVIDIA DGX GH200 整合了 256 个 GPU



资料来源：英伟达官网、华泰研究

图表179： 英伟达 DGX H100 VS DGX GH200

	DGX H100	DGX GH200
GPU and CPU	8x NVIDIA H100 Tensor Core GPUs + Dual Intel® Xeon® Platinum 8480C Processors	256x NVIDIA Grace Hopper Superchips (each Grace Hopper Superchip includes Grace Arm® CPU+ H100 Tensor Core GPU)
CPU Cores	112 Cores total, 2.00 GHz (Base), 3.80 GHz (Max Boost)	18,432 Arm® Neoverse V2 Cores with SVE2 4X 128b
GPU memory	640GB	144TB
Performance(FP8)	32 petaFLOPS	1 exaFLOPS
NVIDIA® NVSwitch	4x	96x L1 NVIDIA NVLink Switches 36x L2 NVIDIA NVLink Switches
Networking	4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet	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
Management network	10Gb/s onboard NIC with RJ45 100Gb/s Ethernet NIC Host baseboard management controller (BMC) with RJ45	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 3-year business-standard hardware and software support	

资料来源：英伟达官网、华泰研究

2023 年 8 月 SIGGRAPH 大会上，英伟达发布配备 HBM3e 内存的 GH200 Grace Hopper 超级芯片平台。该平台为全球首个采用 HBM3e 内存的超级芯片，由 72 核的 Grace CPU 和 Hopper 计算 GPU 组成。与 2023 年 5 月 Computex 上公开的配备 HBM3 内存的 GH200 相比，新款 GH200 采用双路配置，这使其内存容量和带宽分别是 GH200（HBM3 版）的 3.5 倍和 3 倍，其旨在满足包括大型语言模型、推荐系统和矢量数据库等在内的各种生成式人工智能工作负载要求，预计其将于 2024 年第二季度上市。

锚定下游应用企业厂商，投资加码助力生态渗透

2023 年来英伟达已投资十余家 AI 初创企业。伴随 2022 年末以 ChatGPT 为首的生成式 AI 走入大众视野，英伟达在 2023 年对各领域 AI 初创企业的投资持续推进。根据 Pitchbook 数据，自 2023 年起，英伟达至少已经投资 12 家初创企业，其中包括 InflectionAI、Cohere、Runway 等多家 AI 独角兽企业。我们认为此举或出于英伟达对拓宽业务渠道与巩固生态壁垒的需求。此外，主要锚准下游应用的投资布局也彰显英伟达对 AI 赛道下游应用企业成长价值的预期。

图表180： 英伟达 2023 年以来所投资的部分公司

行业	公司名称	交易时间	融资总金额（百万美元）	被投资公司主营业务	合作与投资方面其他情况
AI+生物	Recursion Pharmaceuticals	2023. 7	50	致力于使用 AI 模型识别和设计新疗法，模型客户为包括拜尔等在内的制药商	本次投资将用于加快生物及药物发现模型的开发
AI	Aleph Alpha	2023.7	109*	主要从事 AI 模型的研究、开发与运营，代表模型为 Luminous	英特尔也参与此轮投资
AI	Inflection AI	2023. 6	1300	致力于大型语言模型的研究、训练与开发，产品代表为 AI 聊天机器人 Pi	计划与英伟达、美国云计算公司 CoreWeave 一同合作开发 H100 集群
AI+多媒体	Runway	2023. 6	191	人工智能视频软件公司，致力于提供文生视频的 AI 服务	谷歌和 Salesforce 也参与此轮投资
AI+设计	NdotLight	2023. 6	7.7	提供基于网络的 3D 设计解决方案，代表解决方案为“Ndot CAD”	公司 3D 建模引擎技术在 CES 2023 上获得了软件和移动 app 类别创新奖
AI	Cohere	2023. 6	270	专注于 NLP 技术与语言 AI 的公司，帮助企业改进人机交互体验	Oracle 和 Salesforce 亦参与本轮投资
AI+多媒体	Synthesia	2023. 6	90	AI 视频创作平台，用户可借助其平台创建虚拟人物形象	资金将用于研发包括适配任务语言转换为用户需求的生成式 AI 功能
自动驾驶	Foretellix	2023. 5	43	公司主营业务为开发驾驶辅助和自动驾驶测试和验证平台的无人驾驶系统	曾在 2022 年宣布将英伟达模拟工具与 Foretellix 验证平台结合以创建解决方案的合作
AI+生物	CHARM Therapeutics	2023. 4	20	利用深度学习对 3D 分子结构型进行研究，开发针对过敏性蛋白质的分子治疗	由英伟达风险投资部门 Nventures 进行投资，将加速小分子疗法的开发
云计算	CoreWeave	2023.4	421	大规模 GPU 虚拟化云服务公司	CoreWeave 以大量英伟达 H100 芯片为抵押获得 23 亿美元债权融资
AI+软件	Adept AI Labs	2023. 2	350	主要负责开发 AI 软件以实现计算机任务自动化	微软亦参与投资，资金将用于开发可以根据人类提示实际执行命令的工具
光技术	Ayar Labs	2023.2	25	硅谷光子晶片新创企业，致力于研究取代电流处理的光技术	英伟达本轮投资为上轮追加，惠普和英特尔等也参与本轮投资，资金将用于加速其光学输入输出（I/O）解决方案的商业化

注：*号代表来自 Pitchbook 预测值
资料来源：Pitchbook 官网、CNBC 官网、华泰研究

与英伟达深度合作的 CoreWeave 为以太坊挖矿起家，现已转型 AI 算力专业云提供商。CoreWeave 是一家在 2018 年加密货币“矿难”之后，凭借其丰富的 GPU 储备从以太坊挖矿企业逐渐转型为云计算平台的公司。2022 年 9 月，CoreWeave 成为英伟达第一个精英级别云服务提供商，伴随 ChatGPT 为首的 AI 浪潮来临，公司于 2022 年底开始进行密集融资：在 2022 年 12 月、2023 年 4 月和 6 月分别完成 1 亿、2.21 亿（英伟达为主要跟投者之一）和 2 亿美元融资之后，8 月，CoreWeave 宣布完成以英伟达 H100 芯片为抵押的 23 亿美元债权融资，资金将用于支持 CoreWeave 在美国的数据中心拓展，且据 CoreWeave 官网信息，其机群中有超过 45,000 个可按需提供的高端 NVIDIA GPU。

深度绑定新创云服务商 CoreWeave，助力其数据中心建设。而在其丰富的英伟达芯片储备背后，离不开英伟达的支持。据 The Information 6 月报道，尽管英伟达 H100 十分紧缺，但在谷歌和亚马逊等云大厂的面前，英伟达将大量新卡分配给了 CoreWeave 和 Lambda Labs（据 The Information 7 月报道目前英伟达正在与其达成收购股权协议）两家初创云平台企业，且在 7 月公布的 MLPerf 基准评测数据中，正是 CoreWeave 和英伟达合作，使用基于 H100 的云服务实现只使用 11 分钟训练 GPT-3。我们认为，在谷歌、亚马逊和微软等云大厂如火如荼进行 AI 芯片自研的当下，英伟达对 CoreWeave 的深度绑定或出于希望与下游云平台厂商形成协同效应，积极应对来自云厂商自研芯片的挑战。

图表181： CoreWeave 部分融资轮次及情况

时间	融资轮次	融资总金额（美元）	估值金额（美元）	投资方
2023 年 6 月	B+轮	2.00 亿	超过 20 亿	Magnetar Capital
2023 年 4 月	B 轮	2.21 亿	20 亿	Magnetar Capital, Nvidia, DanielGross, NatFriedman
2022 年 12 月	战略融资	1.00 亿	-	Magnetar Capital
2021 年 11 月	战略融资	5000 万	-	Magnetar Capital

资料来源：Bloomberg、Business Wire 官网、华泰研究

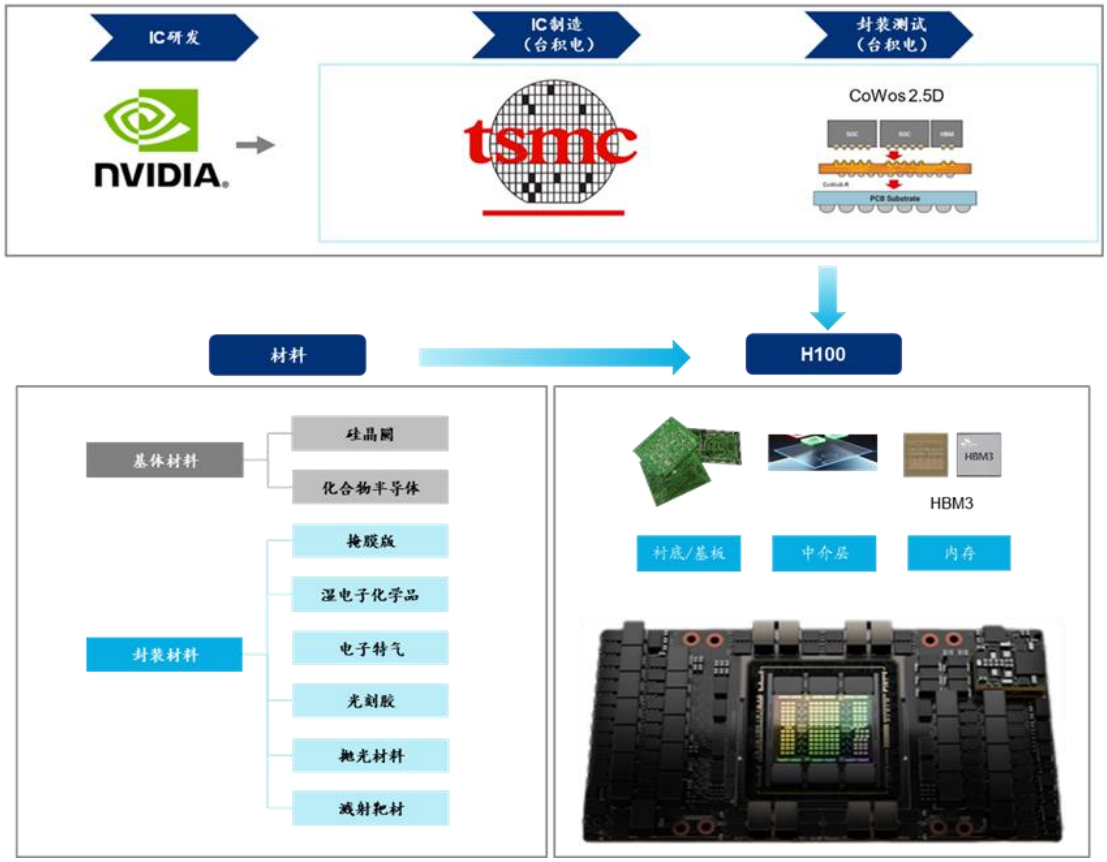
数据中心：AI 芯片风再起时，英伟达乘风而行

预计英伟达数据中心业务 FY2025 营业收入将达 654 亿美元

我们预计数据中心业务 FY2024/FY2025/FY2026 的营业收入同比增速 152%/73%/40%，对应营业收入为 378.13/654.16/915.82 亿美元。

我们对英伟达数据中心业务的营收做了三个情景假设：乐观、中性、保守。在中性情景下（也是我们财务模型和目标价格的依据）：按照 Digitimes 7 月 14 的报道，台积电 2023 年 CoWoS 产能为 12-13 万，英伟达在当中取得约 40%的产能。我们假设每片能切割出 25-30 颗来算，即 23 年出货量约 130 万颗，当中鉴于 H100 的性价比较 A100 高，我们预计 H100 占比也较高，因此 ASP 应更接近 H100 的单价，我们以 2.3 万美元 ASP 来算，加上其他产品，得到**英伟达 FY2024 年数据中心业务约 378 亿美元的营收**。而 FY2025 年我们预计英伟达将于台积电流片 7 万多片，加上少量外溢到 Amkor 等的流片，一共将接近 200 万颗。ASP 方面，我们认为虽然会受到 AMD MI300 等竞品放量的影响，但由于 H100 占比会更高，因此 ASP 将在 25000 美元左右，加上其他产品，得到**英伟达 FY2025 年数据中心业务约 654 亿美元的营收**。2025 年在 AI 的商业应用落地开始稳定，加上台积电和其他封装产能扩产渐趋稳定，我们认为同比增速将放缓至 40%，因此我们预计英伟达 FY2026 年数据中心营收将达 916 亿美元。

图表182：英伟达 H100 的制造流程和硬件结构示意图



资料来源：《半导体制造技术》（Quirk & Serda, 2015）、英伟达官网，台积电官网，华泰研究

乐观情境下，假设 2024 年台积电 CoWoS 产能将达每月 2 万片（即一年 24 万片）。按照英伟达在当中取得约 40%的产能，以及每片能切割出 25-30 颗来算，即 24 年英伟达在台积电的出货量约 240 万颗，加上外溢到 Amkor 等的流片，一共将接近 340 万颗。我们也预计当中 H100 占比会更高，因此 ASP 将在 25000 美元或以上，加上其他产品，得到**英伟达 FY2025 年数据中心业务约 1000 亿美元的营收，同比 208%**。FY2024 年预测不变，而 FY2026 年维持 40%的同比增长，因此我们预计英伟达 FY2026 年数据中心营收将达 1400 亿美元。

保守情景下，如果宏观经济的不确定性及 AI 芯片下游需求释放不及预期，我们预计英伟达 FY2025 年出货量约 160 万颗，以 ASP 25000 美元或以上来算，连同其他产品，FY2025 数据中心业务营收将达 454 亿美元，同比增速 40%。FY2024 年预测不变，而 FY2026 年同比增长放缓至 15%，因此我们预计英伟达 FY2026 年数据中心营收将达 522 亿美元。

图表183：英伟达数据中心业务盈利预测（单位：百万美元）

		FY2021A	FY2022A	FY2023A	FY2024E	FY2025E	FY2026E
乐观情景	数据中心营收	**$6,696**	**$10,613**	**$15,005**	**$37,813**	**$99,825**	**$139,755**
	YoY	*124.47%*	*58.50%*	*41.38%*	*152.00%*	*208.00%*	*40.00%*
中性情景	数据中心营收	**$6,696**	**$10,613**	**$15,005**	**$37,813**	**$65,416**	**$91,582**
	YoY	*124.47%*	*58.50%*	*41.38%*	*152.00%*	*73.00%*	*40.00%*
悲观情景	数据中心营收	**$6,696**	**$10,613**	**$15,005**	**$37,813**	**$45,375**	**$52,181**
	YoY	*124.47%*	*58.50%*	*41.38%*	*152.00%*	*40.00%*	*15.00%*

资料来源：Bloomberg、华泰研究

从短期（今年和明年）维度来看，我们认为英伟达将凭借其稳固的市场地位，在 AI 大模型的强劲需求下迎来高速增长。自 2022 年底 ChatGPT 发布以来，海内外云厂商和互联网企业如火如荼推进大模型业务开展。2023 年 8 月，据英国《金融时报》报道，百度、字节跳动、腾讯和阿里巴巴已下单 10 亿美元，采购约 10 万颗英伟达 A800 处理器，另外 40 亿美元订单将于 2024 年交付。我们认为，这或出于限制政策可能收紧而进行“囤货”。在综合考虑到英伟达对 FY2Q24 的乐观指引、台积电在 2Q23 电话会上对未来五年 AI 服务器销售额 CAGR~50%的预测及台积电 CoWoS 扩产的情况，我们认为未来 GPU 供给瓶颈将得到一定程度的缓解，GPU 的出货节奏将迎来加速。

图表184：2022 年底以来生成式 AI 部分大事件汇总

时间	事件	时间	事件
2022 年 11 月	OpenAI 推出 ChatGPT 模型	2023 年 4 月	毫末智行推出自动驾驶生成式大模型 DriveGPT
2022 年 12 月	DeepMind 公布 AI 剧本写作工具 Dramatron	2023 年 4 月	Databricks 推出了 120 亿参数的 Dolly-2 指令遵循模型
2023 年 2 月	OpenAI 推出 ChatGPT Plus 订阅计划	2023 年 4 月	腾讯云推出新一代 HCC 高性能计算集群
2023 年 2 月	谷歌推出 Bard 大模型	2023 年 4 月	字节旗下火山引擎推出大模型训练云平台
2023 年 2 月	微软正式推出新版 Bing 搜索引擎和 Edge 浏览器	2023 年 4 月	StabilityAI 推出并开源 StableLM
2023 年 2 月	复旦大学推出 MOSS	2023 年 4 月	英伟达推出 VideoLDM 论文
2023 年 3 月	OpenAI 推出 ChatGPT API	2023 年 4 月	复旦 MOSS 模型开源
2023 年 3 月	谷歌与柏林工业大学团队推出 PaLM-E	2023 年 5 月	科大讯飞推出“星火”认知大模型
2023 年 3 月	OpenAI 推出 GPT-4 模型	2023 年 5 月	谷歌发布 PaLM 2 人工智能语言模型
2023 年 3 月	百度正式推出文心一言大模型	2023 年 5 月	云从科技推出从容大模型
2023 年 3 月	中科院提出最新女娲大模型 NUWA-XL	2023 年 5 月	北京智源推出通用视觉 AI 模型 SegGPT
2023 年 3 月	OpenAI 推出 ChatGPT Plugins	2023 年 6 月	微软推出 LLaVA-Med AI 模型
2023 年 3 月	微软推出 SecurityCopilot	2023 年 6 月	普林斯顿大学推出 Infinigen AI 模型，
2023 年 3 月	360 推出 GPT 大模型产品矩阵 360 智脑	2023 年 6 月	Stability AI 推出文字生成图片模型 SDXL 0.9
2023 年 3 月	斯坦福开源机器人小羊驼 Vicuna	2023 年 6 月	Unity 推出面向开发者的 AI 软件市场 AI Hub
2023 年 4 月	Meta 推出了人工智能模型 SAM	2023 年 7 月	华为盘古大模型 3.0 正式推出
2023 年 4 月	阿里云大模型通义千问开始邀请测试	2023 年 7 月	上海交大推出开源“白玉兰科学大模型”
2023 年 4 月	上海人工智能实验室等推出气象预报大模型“风乌”	2023 年 7 月	Meta 和微软近日合作推出 Llama 2
2023 年 4 月	商汤科技宣布推出大模型体系“日日新”	2023 年 7 月	英特尔与埃森哲合作推出 34 个开源 AI 参考套件
2023 年 4 月	阿里在 2023 年云峰会上正式推出了通义千问大模型	2023 年 7 月	微软宣布新 Microsoft Store AI Hub 现已开始推出

资料来源：各公司官网、InfoQ 官网、雅虎官网、IT 之家、财联社、panewslab、华泰研究

中长期来看，AI 商业化落地是决定英伟达成长性的关键因素。我们认为英伟达推出的三个生成式 AI 模型，分别代表了三个最具前景的 AI 商业化领域：用于大型语言模型的 NeMo 赋能商业办公等生产力领域；用于生命科学的大型语言模型 BioNeMo 赋能生物医药领域；用于图片、视频和 3D 内容的 Picasso 赋能传媒领域。在推进节奏上，我们认为鉴于传媒领域对逻辑性与精准性要求不高，故为率先商业化落地的方向；反观，商业办公及医疗领域决策过程准确性要求较高，且往往伴随着法律责任风险，因此企业的应用发展应会较为谨慎。目前，AI 商业的渗透和融合已开始有进展，如 Runway（电影《瞬息全宇宙》幕后 AI 特效公司）、微软 365 Copilot（已进一步扩大适用范围，宣布定价模型）和 BloombergGPT 为等，但长期来看 AI+能否切实地为企业降本增效并得到大规模应用仍有待观察，这将是影响英伟达需求的重要指标。展望 2024 年，伴随美联储加息步入尾声， 微软和谷歌等云厂商巨头的 Capex 将迎来修复，AI 基础设施为重点领域。我们看好 AI 与办公、医疗和教育等下游应用领域的继续渗透与结合，且认为英伟达凭借其高算力的优势与高粘性的生态壁垒，其数据中心收入将受益于下游各项 AI 商业化应用的逐步落地与云计算厂商资本开支的持续投入，驶入持续快速发展轨道。

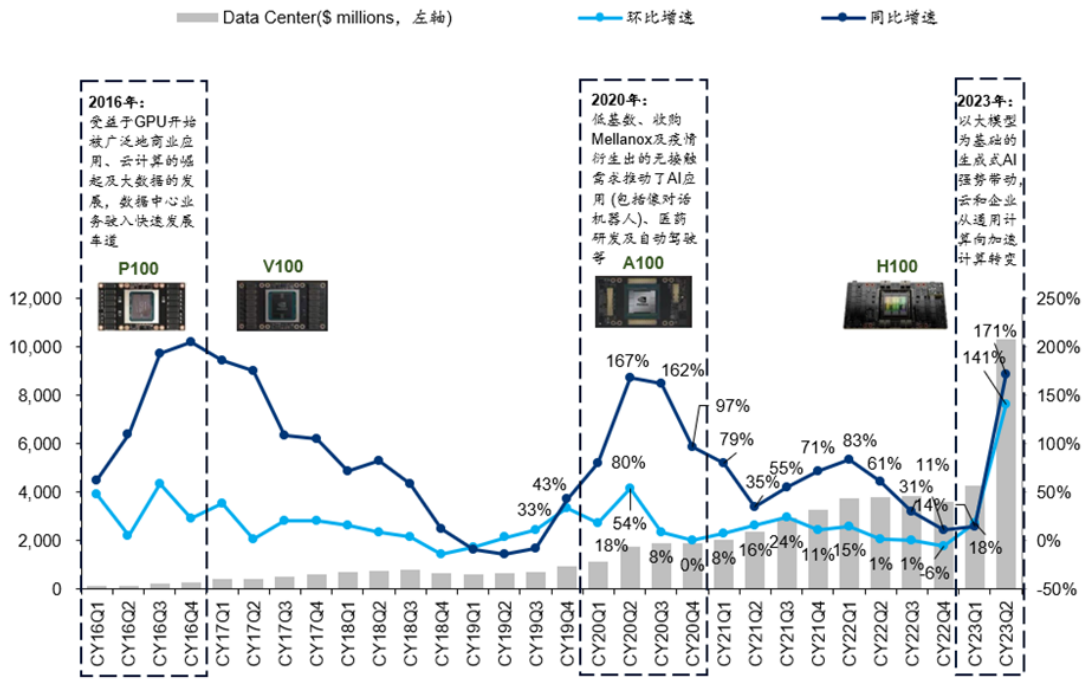
图表185： AI 应用相关落地方向及代表案例

应用领域	公司名称	大模型产品/所使用模型	产品功能
AI+传媒	Runway	Runway Gen-1/2	为视频内容创作者构建渲染、方案生成等生成式 AI 工具
	Adobe	Firefly	通过文字提示快速扩展图像、添加或删除对象
	BuzzFeed	ChatGPT	使用 ChatGPT 为网站生成问卷
AI+办公	Microsoft	365 Copilot（GPT-4）	嵌入 Office，提供文档汇总、PPT 编辑与美化等工作
	Google	Workspace	将生成式 AI 功能结合进入 Gmail 和 Google Docs 等办公组件
AI+教育	Duolingo	Duolingo Max（GPT-4）	以 GPT-4 为支撑，帮助完成语言学习中的多轮对话模拟和问题解析
	可汗学院	Khanmigo（GPT-4）	定制化学习方案、辅助教案编写和课后学科疑问解答
AI+医疗	Amgen	Nvidia BioNeMo	使用自身数据进行预训练模型微调，用于药物发现等领域
	谷歌	Med-PaLM2	第一个在美国医学执照考试(USMLE)式问题的 MedQA 数据集上以“专家”应试者水平表现的大语言模型
	Nuance	DAX Express（GPT4）	全自动临床文档应用程序，可在病人就诊后几秒内自动生成临床笔记
AI+金融	Bloomberg	BloombergGPT	金融知识问答、金融新闻及相关概念舆情分析等

资料来源：各公司官网、华泰研究

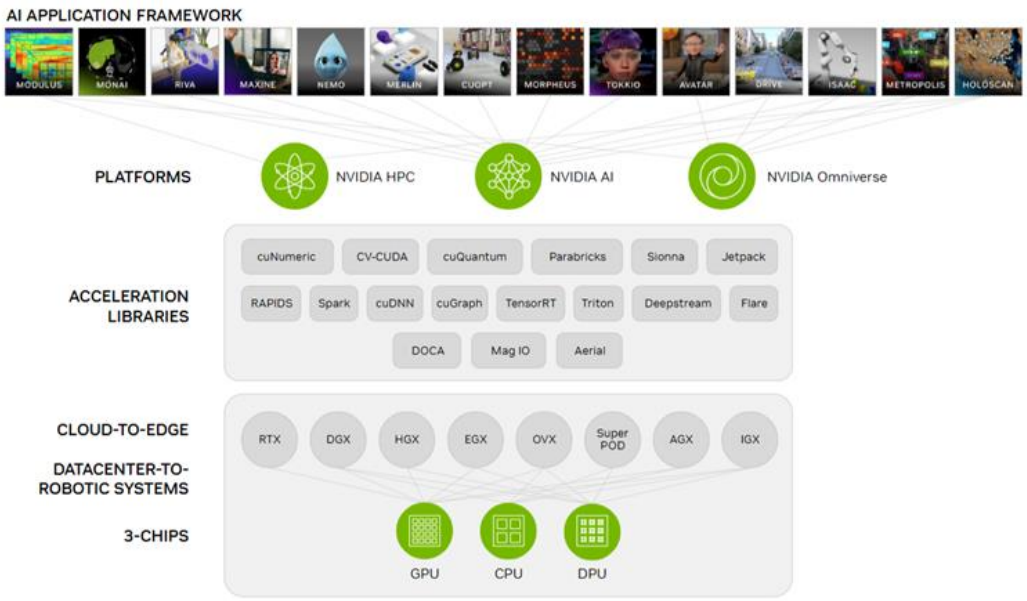
复盘近年来数据中心的业绩变化，我们认为英伟达每轮崛起均受益于不同的人工智能应用。2012 年英伟达 GPU 助攻深度学习模型 AlexNet，以低至 15.3%的错误率识别了上百万张谷歌 YouTube 里已标记图片，夺魁国际计算机视觉顶级赛事 ImageNet 竞赛，GPU 在人工智能的应用就此给谷歌发扬光大。到 2016 年，受益于 GPU 开始被广泛商业应用于语音识别及人脸识别等安防领域、云计算的崛起及大量铺设、以及海量数据的存在，让英伟达数据中心业务驶入快速发展车道。2018-2020 年，OpenAI 以每年一代的速度迭代将 GPT-1 迭代至 GPT-3；2020 年，英伟达数据中心业务受益于 Mellanox 并表、疫情衍生出的无接触 AI 需求（如基于 AI 模型的对话机器人）、医药研发、金融行业、客服及自动驾驶等，再加上前期数据中心营收的低基数效应，让英伟达数据中心再次步入了高速增长。

图表186： 英伟达数据中心业务营收情况（百万美元）



资料来源：公司公告、英伟达官网、华泰研究

图表187： 英伟达 AI 应用框架



资料来源：英伟达官网、华泰研究

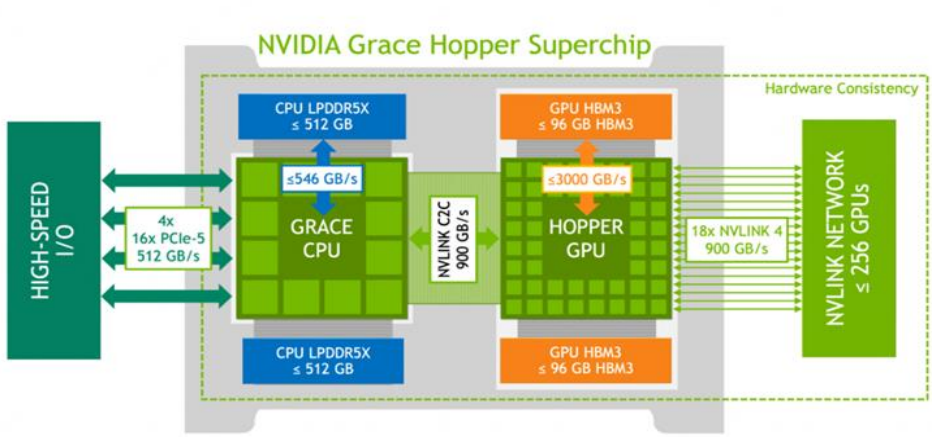
英伟达持续迭代高性能计算芯片，在产品工艺 、计算能力和存储带宽等不断创新。面向高性能计算和深度学习场景，英伟达打造了一系列支持提升张量核心和稀疏矩阵计算等能力的 GPU 产品。2022 年，英伟达已不仅满足于单 GPU 的更新换代，重磅推出结合 Grace CPU 与 Hopper GPU 的 GH200 Grace Hopper 超级芯片，实现了高达 900GB/s 的总带宽，加速大规模 AI 和 HPC 应用计算。在一年后的 SIGGRAPH 大会上，英伟达的 AI 芯片再迎升级，推出了全球首次采用 HBM3e 内存的 Grace Hopper 超级芯片。该芯片的带宽高达每秒 5TB 并能提供 141GB 的内存容量。我们认为其出色的吞吐量进一步证实了英伟达在 AI 芯片领域的霸主地位。

图表188： 英伟达主要 AI 芯片参数对比情况

	V100	A100	H100	GH200 (HBM3)	GH200 (HBM3e)
FP32 CUDA Cores	5120	6912	16896	-	-
Tensor Cores	640	432	528	528	528*
Boost Clock	1.53GHz	1.41GHz	~1.78GHz	-	-
Memory Clock	1.75Gbps HBM2	3.2Gbps HBM2e	4.8Gbps HBM3	HBM3	HBM3e
Memory Bus Width	4096-bit	5120-bit	5120-bit	-	-
Memory Bandwidth	900GB/sec	2TB/sec	3TB/sec	<=4TB/sec	5TB/sec
VRAM	16GB/32GB	80GB	80GB	96GB	141GB
FP32 Vector	15.7 TFLOPS	19.5 TFLOPS	60 TFLOPS	-	-
FP64 Vector	7.8TFLOPS	9.7TFLOPS	30TFLOPS	-	-
INT8 Tensor	N/A	624 TOPS	2000 TOPS	-	-
FP16 Tensor	125 TFLOPS	312TFLOPS	1000TFLOPS	-	-
TF32 Tensor	N/A	156 TFLOPS	500 TFLOPS	-	-
FP64 Tensor	NA	19.5TFLOPS	60 TFLOPS	-	-
Interconnect	NVLink 2	NVLink 3	NVLink 4	NVLink 4	NVLink 4
	6 Links (300GB/sec)	12 Links (600GB/sec)	18 Links (900GB/sec)	18 Links (900GB/sec)	18 Links (900GB/sec)
Transistor Count	21.1B	54.2B	80B	200B	-
TDP	300W/350W	400W	700W	450W - 1000W	450W - 1000W
Manufacturing Process	TSMC 12nm FFN	TSMC 7nm	TSMC 4nm	TSMC 4nm	TSMC 4nm
Interface	SXM2/SXM3	SXM4	SXM5	-	-
GPU Architecture	Volta	Ampere	Hopper	Hopper	Hopper

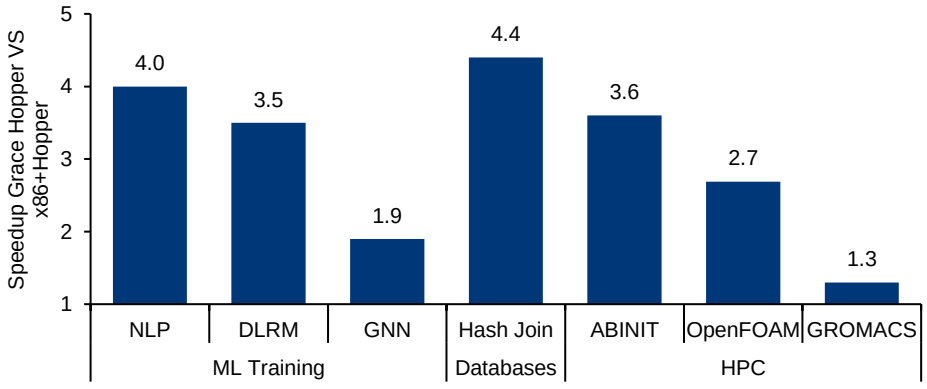
资料来源：英伟达官网、ANANDTECH、tom's Hardware、华泰研究

图表189： 英伟达 Grace Hopper 架构示意图



资料来源：英伟达官网、华泰研究

图表190： 英伟达 Grace Hopper 与 x86 + Hopper 的最终用户应用程序性能模拟（单位：倍数）



资料来源：英伟达官网、华泰研究

大模型的训练到底需要多少英伟达 A100?

我们根据英伟达 2021 年发表的论文《Efficient large-scale language model training on GPU clusters using megatron-LM》，计算得出在结合数据并行、流水线模型并行、张量模型并行及服务器通信优化等，多种加速方式下不同参数的单个大模型所需要的最少 GPU 个数：N=8TP/Xt（其中 T 为模型训练数据 Token 总数，X 是单 GPU 的有效算力，t 是训练时间，P 是模型参数量）。

图表191： 大模型训练所需 A100 测算

	参数数量(billion)	Training token(billion)	模型训练总计算量(FLOPs)	单 GPU 峰值算力 (TeraFLOP/s)	优化后算力利用率	有效算力 X (TeraFLOP/s)	GPU 数量(个)
GPT-3	175	300	4.1904E+23	312	44%	137	1180
GPT-4	1750	500	1.2E+24	312	44%	137	19672

注：GPT-4 的数据参考 semianalysis 估计值，此处 GPU 指代 A100

资料来源：OpenAI 官网、《Efficient large-scale language model training on GPU clusters using megatron-LM》（Deepak Narayanan et al.，2021）、NVIDIA 官网、semianalysis 官网、华泰研究

训练测算关键假设：

1) 芯片选择为业内计算 Transformer 类模型常用的英伟达 A100，精度为 FP16 TensorCore。
2) 按照以上英伟达论文里，在采取多种优化方式后，单 Token 的计算次数可大幅降低至 8 次，算力有效利用率可提升至 40%-50%。
3) 我们也发现，坊间对于所需 GPU 的数量众说纷纭。我们解释，使用更多 GPU 是完全可行：如 V100 的单卡 FP16 TensorCore 峰值算力较低，为 125 TFLOP/s，若将此带入我们的计算方法，则需要约为 3281 个 GPU；又或者 A100 若使用 FP32（更高精度），所需的 GPU 也会大量增加。值得一提的是，理论上精度越高，计算的准确度应该也越高，但在 AI 训练里，我们也需注意精度 vs 准确度的取舍 (precision vs accuracy tradeoff)，也就是说，如果使用了更高的精度，但准确度只有微量的增加，那就没必要提升精度而浪费算力和能源。此外，本测算基于 1 个月的计算时间，且使用了数据并行、流水线模型并行、张量模型并行和服务器通信优化等完全优化的方式去训练，若训练要求时间更短，且采取了较少的模型并行化的策略，需要的 GPU 数目也会更多。

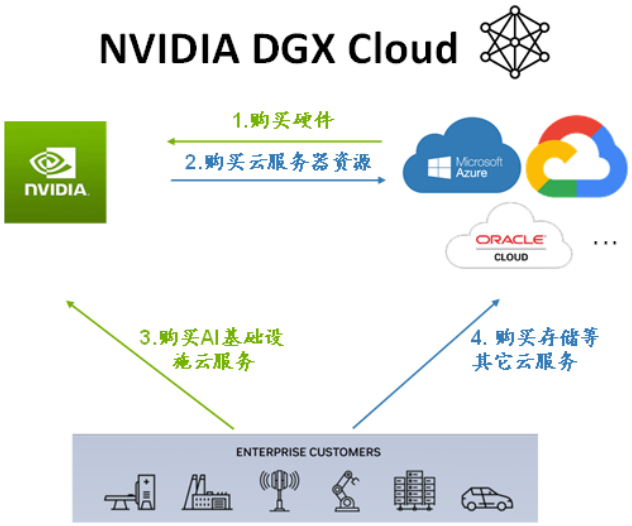
高歌入云拓展业务模式，算力租赁、模型部署到元宇宙应用三箭齐发

英伟达凭自身对 AI 技术的深刻理解将商业模式扩展到云端。我们认为这一举措将助力英伟达进一步提升在企业 AI 市场的渗透率。具体来看，英伟达在 GTC 2023 上共推出了三大云服务平台：

1) DGX Cloud：算力租赁服务

DGX Cloud 为企业提供训练生成式 AI 模型的全栈服务。目前，英伟达的 DGX 云服务已在一众云厂商上运行，如 Oracle 云、微软 Azure 和谷歌云。英伟达的云服务客户已涵盖 Amgen、Adobe、Getty Images、Deloitte 和 Morning Star 等各大知名企业。Oracle 是首家大规模提供 DGX Cloud 托管的云厂商，通过 Oracle 远程直接内存访问(RDMA)网络，企业用户可通过 OCI Supercluster 使用 NVIDIA NeMo 语言服务和 NVIDIA Picasso 图像、视频和 3D 服务来构建特定领域的生成式 AI 应用程序。在微软 Azure 云上，二者的合作计划已于 23Q2 展开，企业用户可按月租用 DGX AI 超级计算和其他应用软件的专用集群服务。

图表192： NVIDIA DGX Cloud 商业模型示意图



资料来源：NVIDIA 官网、GOOGLE 官网、Azure 官网、华泰研究

图表193： DGX Cloud 与 Azure ND96asr 参数对比

	DGX Cloud	Azure ND96asr
Software	>NVIDIA Base Command Platform >NVIDIA AI Enterprise software suite >Hybrid-cloud support	-
Hardware	> 8 NVIDIA A100 Tensor Core GPUs or 8 NVIDIA H100 Tensor Core GPUs per node (640GB total) > Access to the latest GPU technology > 10TB storage per instance > 10TB egress per month per tenant > The ability to scale up storage and egress bandwidth as needed	>8 NVIDIA A100 Tensor Core GPUs > 96 physical 2nd-generation AMD Epyc™ 7V12 (Rome) CPU cores > 6500GB Temp Storage >900GB RAM > The ability to scale up storage and egress bandwidth as needed
Services	> Access to NVIDIA AI experts > 24/7 business-critical support > Technical account manager > Customer service success manager > Single-point-of-contact support	>The ability to opt Professional Direct/ Standard/ Developer Support at an additional charge >Microsoft Customer Agreement
Prices	36999 USD+	19854 USD+

资料来源：英伟达官网、微软官网、华泰研究

从商业模式上来看，英伟达 DGX 云是将基础硬件设施出售给云厂商，再向他们购买云计算资源，最后把云服务出售给企业客户并自留全部收入。因此，英伟达可通过两种方式变现：售卖硬件和售卖云服务。而云厂商的收入则来自英伟达从他们购买云计算资源，同时英伟达也会为他们带来新客户，产生协同效应。DGX Cloud 实例的起步价为每个实例 36999 美元/月，其每个实例都具有 8 个 NVIDIA H100 或 A100 80GB Tensor Core GPU，每个节点共有 640GB 的内存。从价格上看，DGX Cloud 差不多是微软同档云服务产品 Azure ND96asr 的两倍。高定价或源于英伟达 AI 软件服务的加持，尤其是可提供 100 多个框架、预训练模型和开发工具 NVIDIA AI Enterprise。

2) NVIDIA AI Foundations：提供模型定制服务

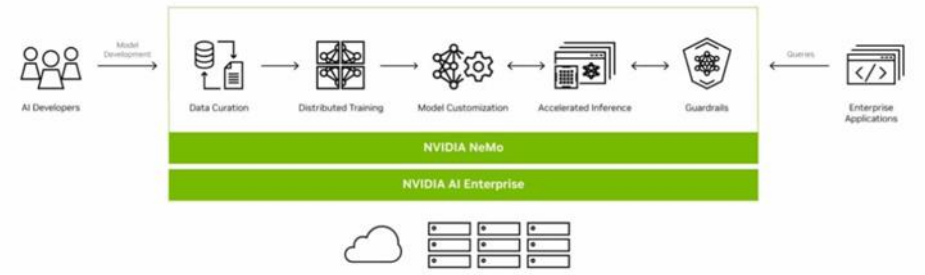
NVIDIA AI Foundations 主要面向具有构建、优化和运营定制大语言模型和生成式 AI 需求的客户，为其提供语言、图像和生物学方面的预训练模型与模型定制化工具。具体包括用于大型语言模型的 NeMo；用于生命科学的大型语言模型 BioNeMo；用于图片、视频和 3D 内容的 Picasso。

图表194： NVIDIA AI Foundations 为三个领域的生成式 AI 提供云服务

资料来源：英伟达官网、华泰研究

- **大语言模型 NeMo 适用于生产力的提升**，主要面向提供构建和部署文件分类、代码编写和内容优化模型的端对端工作流，客户可将 NeMo 框架中的指令学习、监督式微调或 RLHF (Reinforcement Learning Human Feedback)，结合知识库进行实时的决策辅助。举例来说，内容生成 AI 公司 Writer（客户包括德勤、欧莱雅和财捷等世界 500 强企业）使用 NeMo 将其大语言模型参数从 30 亿快速提高至 400 亿，并使用 NeMo 配套的高吞吐量 Triton 推理服务器支撑其模型每月大万亿次的 API 调用。

图表195： NVIDIA NeMo 支持从数据整理到推理的整个端到端大语言模型生成流程



资料来源：英伟达官网、华泰研究

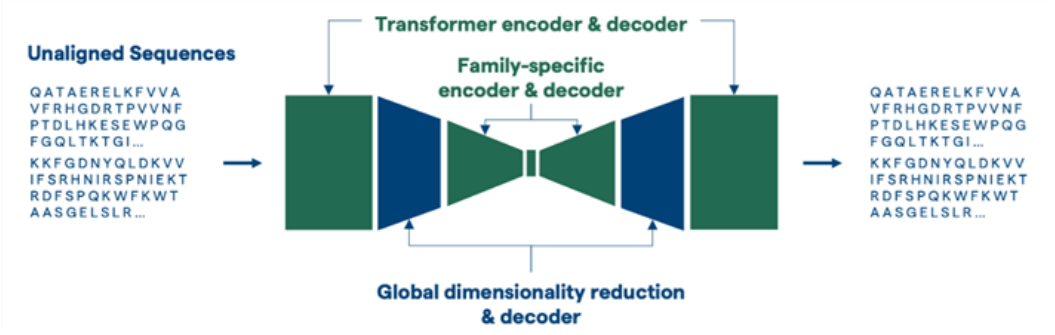
- **Picasso 适用于传媒领域，**可利用文本提示词对视觉设计生成式 AI 模型进行定制，从产品的风格、品牌和特定 IP 等方面对视觉内容创作进行高分辨率图片、时间一致性视频编辑、高逼真的 3D 网格模型和 360 度全景地图的编辑与生成。Picasso 目前已在 Shutterstock 中进行融合，帮助创作者使用文本提示词，快速创建与自定义其 3D 场景背景，生成自定义的 360 度、8K 分辨率、高动态范围成像(HDRi)的环境地图。

图表196：NVIDIA Picasso 帮助 Shutterstock 客户使用文本提示词快速创建与自定义其 3D 场景背景

资料来源：英伟达官网、华泰研究

- **BioNeMo 适用于生物医药领域**，向生物制药企业等客户提供从分子对接到蛋白质 3D 结构发现相关的 9 个预训练模型（截止 2023 年 8 月）以供其定制化其解决方案。Evozyne 基于其专有蛋白质数据与 BioNemo ProtT5 模型建立其蛋白质发现模型 ProT-VAE；Amgen 使用其专有抗体数据对 BioNeMo 的 ESM 模型架构与英伟达 DGX Cloud 进行模型预训练和微调，缩短其分子筛选模型的训练时间。

图表197：Evozyne 使用 BioNeMo ProtT5 模型 开发的 ProT-VAE 模型基本架构



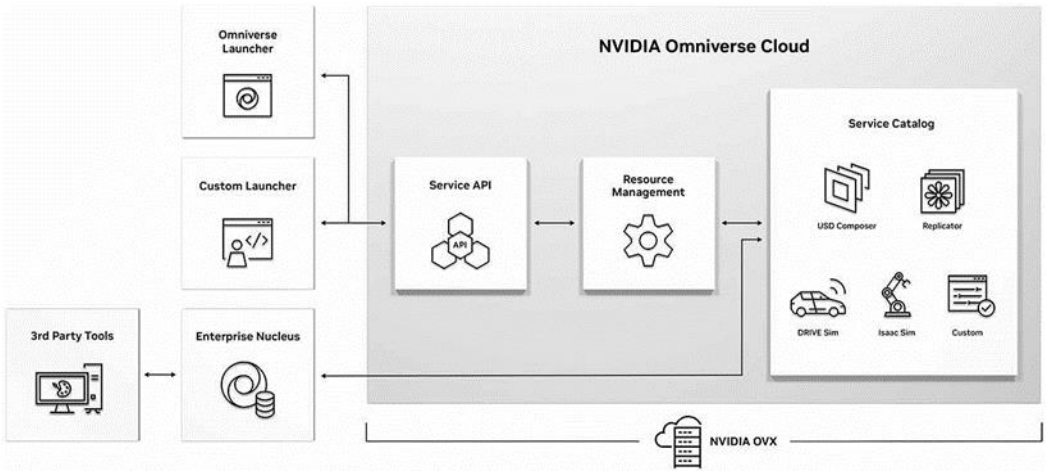
资料来源：《ProT-VAE: Protein Transformer Variational AutoEncoder for Functional Protein Design》(Emre Sevgen et al., 2023)、华泰研究

通过与 DGX Cloud 深度绑定，将其算力能力渗透至工具及平台层。英伟达 AI Foundation 服务均运行在可通过浏览器访问的 DGX Cloud 上，供开发人员通过简单的 API 进行使用。模型在部署完成之后，企业即可在英伟达数据中心运行大规模的推理工作负载，从而无需自行配备高性能硬件，通过订阅方式即可访问云端算力。

3) Omniverse Cloud 平台：构建和运营工业元宇宙

Omniverse Cloud 的用户可通过连接在云端、边缘设备或本地运行的 Omniverse 应用进行 3D 设计协作和模拟，打破物理与数字的壁垒，可应用于 5G、交通、物流、培训、自动驾驶、建筑、零售等各个领域。例如，宝马集团是第一家采用 Omniverse 打造全数字化智能工厂的汽车制造商，借助 Omniverse Cloud，他们可以通过虚拟方式对计划建造的现实工厂进行设计和优化，除 BMW 集团外，Omniverse Cloud 已和吉利路特斯（Geely Lotus）和捷豹路虎（Jaguar Land Rover）等建立了合作关系。

图表198： 英伟达 Omniverse Cloud



资料来源：英伟达 Blog、华泰研究

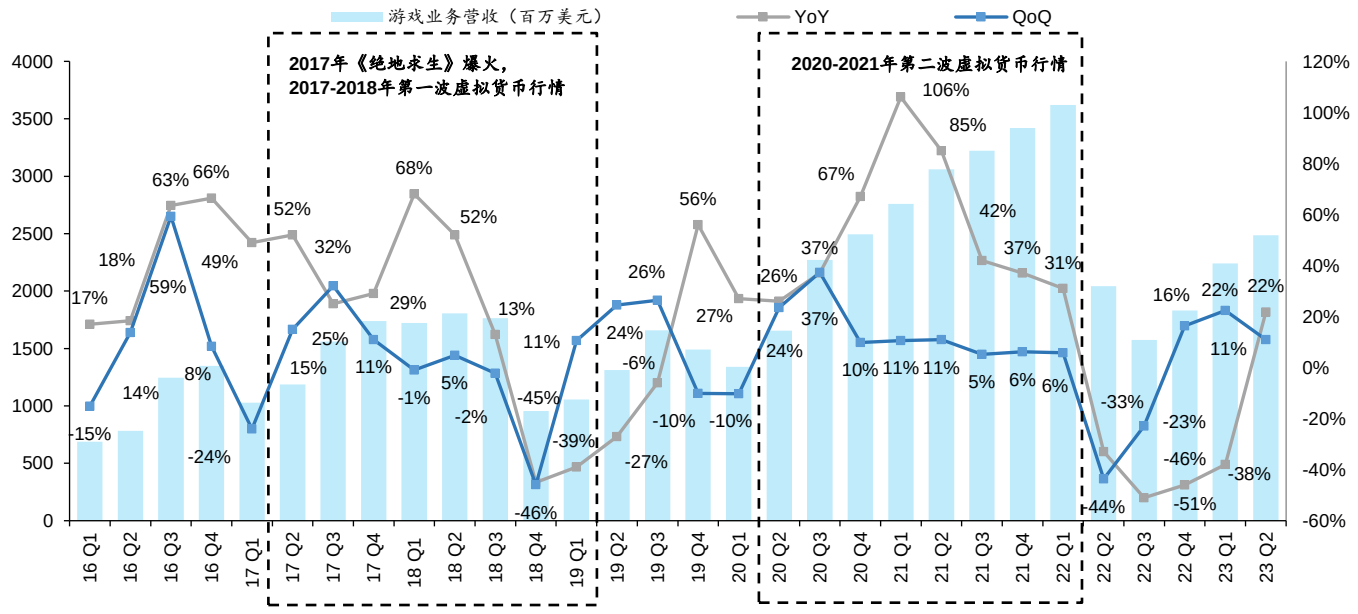
游戏显卡业务：虚拟货币扰动后重回正常发展轨道

游戏显卡业务：我们预计游戏显卡业务 FY2024/FY2025/FY2026 的营业收入同比增速 20%/25%/30%，对应营业收入 108.80/136.01/176.81 亿美元。FY2024 年方面，我们认为 2022 年以来以太币机制转换，以及“第二次矿难”对公司游戏显卡业务所造成的影响已逐渐消退。同时，我们也从 FY2023 年 4 季度开始看到拐点，认为游戏显卡业绩下滑趋势将逐渐收窄。故我们预计，FY2024 年英伟达游戏显卡业务同比增长为 20%，营收达 109 亿美元。**FY2025 年方面，**就 PC 市场来看，IDC 预计全球 PC 出货总量继今年 Q1 出现同比三成跌幅后，将在 Q2 和 Q3 将跌幅收敛至 17%和 7%，并在 Q4 重现上升趋势；在游戏方面，伴随 2024 年《黑神话：悟空》和《星球大战：亡命之徒》等 3A 游戏巨作预计全面上线带动游戏市场活力释放，结合明年 PC 销量将开始恢复的前瞻预期，我们认为，作为 Switch 游戏机芯片的独家供应商，任天堂 Switch 2 在 2024 年的发布将为英伟达游戏业绩重回健康轨道提供支撑。故我们预计英伟达 FY2025 年英伟达游戏显卡业务同比增长为 25%，营收达 136 亿美元。**FY2026 年方面，**鉴于英伟达 6 月简报路线图显示，Ada Lovelace Next 游戏架构相关芯片有望发售，我们认为新系列游戏显卡性能优势将为游戏业务新成长提供支撑，预计 FY2026 年英伟达游戏显卡业务同比增长为 30%，营收达 177 亿美元。

英伟达显卡成长路径复盘：PC 游戏与挖矿需求双主线

PC 独显市场中，英伟达龙头地位稳固。英伟达游戏业务核心产品为 PC 端独立显卡，主要面向个人电脑和游戏主机出售。该产品主要功能为通过图形处理功能为游戏玩家和创作者提供高游戏分辨率和低延迟体验。凭借性能、工艺与架构迭代，英伟达产品已在该领域具备较强的竞争优势和市场地位，且由于游戏显卡为英伟达的发家业务，其先发优势较强。

图表199： CY2016Q1-FY2024Q2 英伟达游戏显卡收入



资料来源：Bloomberg、华泰研究

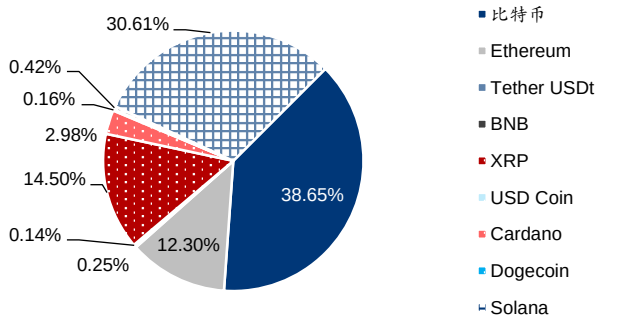
作为业内龙头，英伟达显卡业务受游戏及虚拟货币两大市场同频共振。在 PC 游戏市场和游戏图形质量要求的快速提升下，全球 PC 显卡销量在 2017 年前处于高速发展通道，此后到 2017 年虚拟货币迎来第一波爆发浪潮，大量 GPU 显卡被用于挖矿；2020-2021 年疫情期间游戏用户快速增长，叠加虚拟货币迎来第二波浪潮，导致挖矿需求旺盛，再次刺激显卡销量虚假繁荣。直到 2022 年起随着以太坊的挖掘算法出现结构性变化，导致显卡需求大幅减少。

阶段一（1999-2016）：伴随互联网普及和 3D 游戏兴起，PC 游戏及终端游戏设备用户持续增长，为英伟达的显卡带来发展契机。1999 年，英伟达推出 GeForce 256 显卡，作为第一款采用 T&L（Transform and Lighting）引擎的显卡，可分担 CPU 在 3D 游戏中的运算负载，让 CPU 进行更为高效的逻辑与数据运算，故 GeForce 256 也被英伟达称为世界上第一款 GPU 产品，直接为其在 PC 图形界及游戏市场上的成功奠定了基础。2001 年和 2005 年，公司分别宣布为微软首款 Xbox 和索尼 PlayStation 3 游戏机开发处理器，且成立了游戏事业部，专门负责开发和发布游戏相关技术和产品。在此之后，英伟达顺应互联网的普及度和 PC 游戏快速发展的市场趋势，不断提高显卡性能及质量的同时，致力于与商业伙伴开展合作，打造相关配套软硬件，引领游戏市场的发展潮流。2013 年，英伟达发布支持显卡驱动优化的 GeForce Experience 软件，推出游戏和娱乐便携设备 NVIDIA SHIELD。

阶段二（2017-2021）：2017 年，挖矿+“吃鸡”催化下显卡供不应求，英伟达显卡因此受益。2017 年，虚拟货币价格连创新高。根据 Coinmarketcap 数据，2017 年虚拟货币市场总值达 5724.8 亿美元，全年累计增长 3028%，比特币(BTC)价格上涨超过 13 倍，一度突破 20089 美元。以太坊（Ethereum）的数字代币以太币（Ether）在 2017 年涨幅达 111%，成为第三大市值的虚拟货币。比特币采用的 SHA-256 算法为计算密集型，这意味着其更多地依赖于处理器的计算能力，因此适配为特定的计算任务而设计的 ASIC 专用芯片（执行 SHA-256 哈希函数时比通用硬件，如 CPU 或 GPU，更加高效）。而以太坊采用名为 Ethash 的硬内存（memory-hard）算法，要求在挖矿过程中频繁地从一个大型数据集（称为 DAG）中读取数据，该操作主要受到内存带宽而不是计算能力的限制，因此降低了 ASIC 设备的优势，并使得拥有大量内存的 GPU 成为以太坊挖矿的首选硬件。即使存在针对以太坊的 ASIC 矿机，GPU 仍然是以太坊挖矿的主流选择。故在此时期中，以太币挖矿的热度提升直接带动 GPU 矿机需求高增。

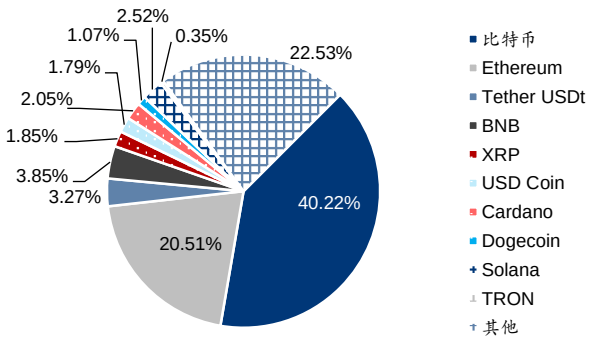
2017 年，《绝地求生》横空出世，催化显卡购置热潮。此游戏在不到一年时间便突破 300 万玩家同时在线，推动 Steam 最高在线人数从 1370 万增长到 1800 万。《绝地求生》官网显示，其为用户的建议推荐显卡为 NVIDIA GeForce GTX 1060 3GB 和 AMD Radeon RX 580 4GB；而最佳配置要求推荐为和 NVIDIA GeForce RTX 2060 Super 和 AMD Radeon RX 5700，以获得 144 fps 帧数的竞技体验。且在 2017 年最初发布时，由于其显卡优化不足，需要相对较高的硬件配置才能流畅运行，这直接导致许多玩家为获得高帧率和稳定流畅的画面显示所带来的竞技优势，选择升级或购买新的显卡，掀起了新一轮的购置显卡热潮。

图表200： 截至 2017 年底 ETH 在数字货币市场中的市值占比



资料来源：Coinmarketcap 官网、华泰研究

图表201： 截至 2021 年底 ETH 在数字货币市场中的市值占比

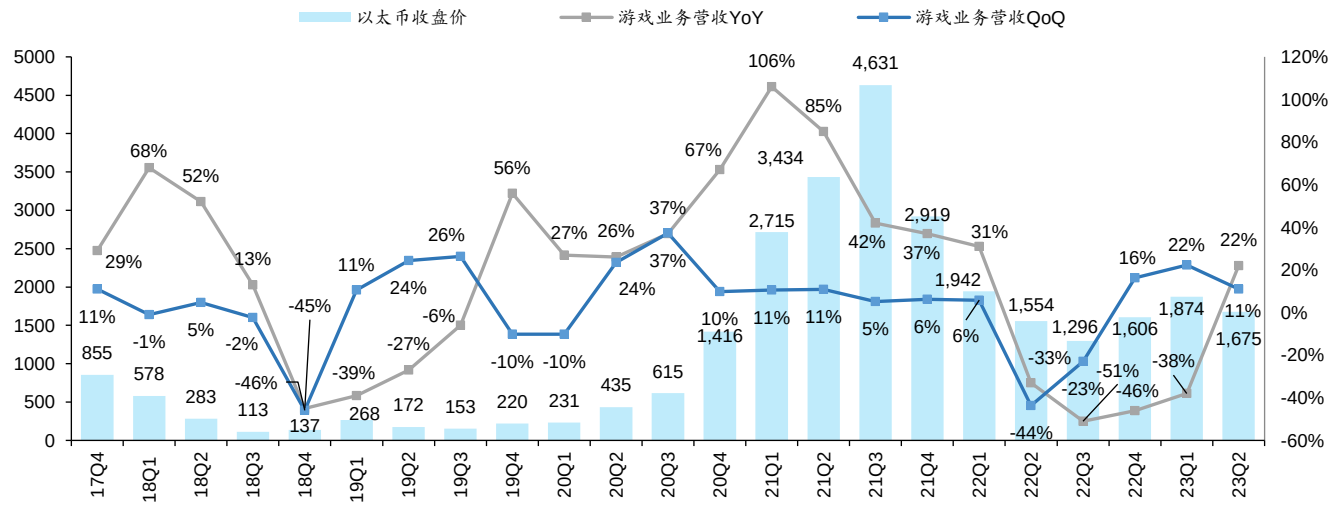


资料来源：Coinmarketcap 官网、华泰研究

2018 年矿难：泡沫破裂所带来的显卡出货承压。自 2017 年虚拟货币市场大牛市以来，2018 年初，比特币价格遭遇断崖式下跌，随之而来的是以太币在 1 月初突破 1400 美元单价大关后，受比特币影响持续下跌。2018 年 2 月 3 日凌晨，全球加密货币市值在 24 小时内缩水超过 1000 亿美元，且以太币在接下来的半年中一度跌至 85 美元水平。以太坊挖币收入作为矿工的主要收入之一，其盈利不及预期直接导致大量挖矿者对显卡需求大幅下降，二手高性能显卡（包括英伟达和 AMD）大量流入市场且价格持续受压，这与显卡厂商直接售卖产生冲突，也导致他们库存积压，英伟达后续的 4 个季度游戏显卡业务同比均为负数。

尽管公司未有详细披露游戏显卡业务中挖矿需求的占比，但回顾历史数据，我们发现该业务的环比增速和以太币价格呈现较强的相关性。18Q4，英伟达游戏显卡业务收入同比下降 45%，环比下降 46%，此时，虚拟货币热度泡沫所导致的市场显卡过剩，以及年初矿难所带来的显卡销量下降，逐步体现在 2018 全年的游戏显卡业务营收增速低迷中。结合公司存货水平来看，自 18Q2 开始，英伟达库存水平偏离正常范围，此后三个季度均保持高库存水平，而到了 19Q2 才开始伴随以太币价格上升而回落。

图表202： 英伟达游戏显卡业务营收情况及以太币价格季度走势 CY2017Q4-FY2024Q2（收盘价单位：美元）



资料来源：Yahoo Finance、公司公告、华泰研究

英伟达推出 CMP 系列专用显卡应对矿潮。故在“矿潮”期间，显卡制造商，包括英伟达和 AMD 的 GPU 销量大幅增长，出货价格长期呈上升趋势，这直接导致二手显卡也时有溢价。为缓解挖矿和游戏之间的需求冲突，避免二手矿卡对市场的扰动，英伟达于 21Q1 推出了 CMP（Cryptocurrency Mining Processor）系列挖矿专用显卡，其在阉割显示接口的基础上，考虑到挖矿任务拥有更低的核心峰值电压和频率的特定需求，在功耗、性能与散热设计等方面进行优化，并在当季度快速实现 CMP 营收 1.55 亿美元。另外，公司也尝试去杜绝游戏显卡用家把显卡用来挖掘，采取包括推出新驱动限制普通游戏显卡挖矿性能等方式。不过，话虽如此，一般游戏的重度玩家也依然可以破解。在 2018 年，英伟达曾经表示挖掘占公司营收不到 10%，然而，在 2022 年英伟达却被美国证监会（SEC）罚款 550 万美元，指控公司在当年对于挖矿带来的收入对投资者有误导，自此之后，公司在后续财务报告中甚少披露数字货币挖矿对公司财务的影响。

图表203：2021 年上半年推出的 4 款 CMP 系列显卡性能参数情况

	30HX	40HX	50HX	90HX
以太坊算力	26 MH/s	36 MH/s	45 MH/s	86 MH/s
额定功率	25 W	185 W	250 W	320 W
电源连接器	1x 8-pin	1x 8-pin	2x 8-pin	2x 8-pin
内存大小	6GB	8GB	10GB	10GB
发行时间	2021Q1	2021Q1	2021Q2	2021Q2

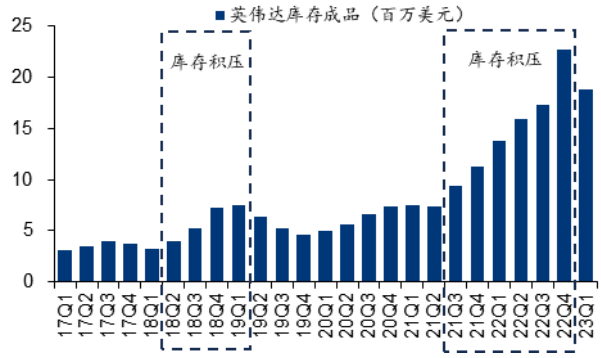
资料来源：英伟达官网、华泰研究

2021 年矿难+2022 年 ETH 挖掘机制结构性转变：大量算力流出带来显卡需求骤降。2021 年开始，伴随美国宏观利率上升，以及中国和美国等加大对虚拟货币的监管，包括马斯克也暂停特斯拉接受比特币支付，在一定程度上影响了市场对虚拟货币的信心，导致以太币在内的虚拟货币价格持续下跌。叠加 **2022 年 9 月 15 日，以太坊运行机制全面升级，从以太坊 1.0 的工作量证明机制（Proof of Work, PoW）转向以太坊 2.0 的权益证明机制（Proof of Stake, PoS）（即以太坊的全面合并）。**PoW 机制要求高性能矿机作为工作设备，鉴于算力决定了挖矿的收益，GPU 凭借其高算力成为以太币矿机首选芯片；而 PoS 机制中，只需通过质押虚拟货币获得收益，就是说，质押的虚拟货币数量越大获得记账收益的概率越高，因此对 GPU 的高算力要求大量减少。据 MESSARI 数据，在采用 GPU 挖矿的前 7 名虚拟货币中，以太坊 ETH 挖矿收入占矿工总收入的 97%，因此以太坊全面合并后不再需要购入大量显卡和投入高算力的计算资源用于挖矿，成为显卡挖矿市场的重要转折点。这是继 2018 年之后，英伟达的游戏显卡业务在 2022 年再次陷入低谷。

图表204：ETH 市场曾几度面临风波（数据截至 2023 年 9 月 20 日）

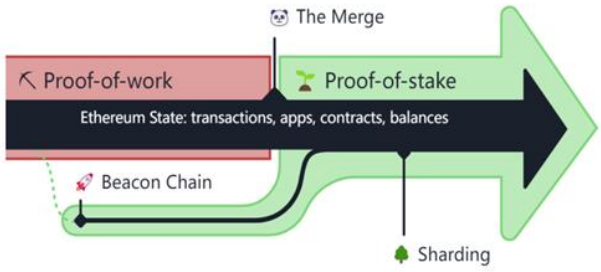
资料来源：Yahoo Finance、华泰研究

图表205：英伟达库存成品情况



资料来源：Bloomberg、华泰研究

图表206：以太坊由 PoW 转向 PoS



资料来源：CoinList、华泰研究

英伟达高端产品占据技术高地，持续保持独显市场领先地位

游戏显卡产品方面，英伟达产品目前主要包括面向 PC 和笔记本电脑的 GeForce RTX/GTX 系列显卡、面向移动端及线上游戏库开发的 GeForce NOW 云游戏平台和为 Switch 提供的 Tegra X1 处理器。英伟达不断深化其显卡的护城河，得益于产品技术的优势，暴雪和世嘉等游戏巨头纷纷为英伟达游戏显卡提供点对点支持，许多游戏制作商在渲染画面时也会针对英伟达显卡产品进行优化，实现发展的良性循环，因此新品系列价格一路走高。

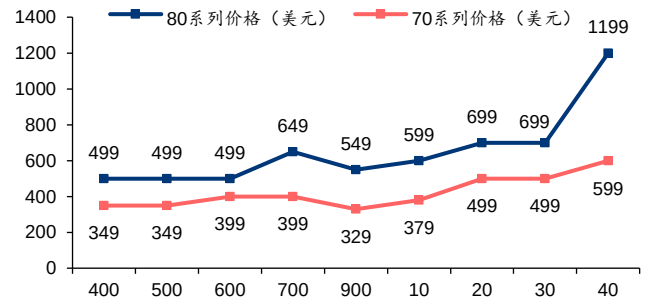
英伟达游戏显卡业务全面覆盖低中高端市场，RTX 50 系计划 2025 年到来。盘点英伟达显卡的技术迭代及市场表现，2016 年推出的 GTX 10 系列显卡相较于 9 系在性能上有显著提升，尽管因为架构和制程技术略显陈旧，但在玩家群体中依然保持较高使用率（比如说，截至 2023 年 6 月，GTX 1060 的 Steam 玩家使用占比长期位列 GPU 榜单 TOP3）。2018 年推出的 RTX 20 系列显卡引入了光线追踪（Ray Tracing）和深度学习超采样（Deep Learning Super Sampling）技术，带来更加真实流畅的游戏画面表现。在推出 20 系产品后，为填补产品线空缺并满足市场需求，英伟达于 2019 年又推出了基于 Turing 架构的 GTX 16 系列显卡，采用与 RTX 20 系列相同的 12nm 制程工艺，但不具备深度学习和硬件光线追踪功能，主打中低端路线。2020 年推出的 RTX 30 系列显卡则基于 Ampere 架构和更先进的 8nm 制程工艺，相较于 20 系性能有显著提升（如 RTX 3080 性能超越前代旗舰 RTX 2080 Ti），能满足高性能计算需求并应对 AMD RX 6000 系列的 PC 游戏显卡市场竞争。2022 年秋季的 GTC 大会，英伟达推出的最新一代 RTX 40 系列显卡性价比饱受争议，但不可否认的是，40 系显卡采用 Ada Lovelace 架构，在性能、功耗、静音和散热等方面均体现出较前代产品的显著提高，其中，RTX 4090 为市场上的顶配独显，产品性能相较上一代提升最高可达 4 倍。2023 年 6 月 28 日，英伟达在路线图中公布了下一代游戏架构“Ada Lovelace Next”于 2025 年登场的计划，该架构将被应用于 RTX 50 系列。

图表207： 英伟达在售热门显卡性能介绍

产品名称	发售年份	架构	流多处理器	RT 核心（光线追踪）	Tensor 核心（AI）	热门型号	CUDA 核心数量（个）	显存规格	发售价格（美元）
GeForce GTX 10 系列	2016-2017	Pascal	1×FP32	-	-	1050 Ti	768	4GB	$139
						1060	1280 / 1152	6GB/3GB	$249
						1070	2432 / 1920	8GB	$379
						1080	3584 / 2560	11GB/8GB	$549
GeForce RTX 20 系列	2018	Turing	1×FP32	第 1 代	第 1 代	2060	2176/1920	12GB/6GB	$349
						2070	2304	8GB	$499
						2080	2944	8GB	$699
GeForce GTX 16 系列	2019	Turing	1×FP32	-	-	1650	896	4GB	$149
						1660	1408	6GB	$219
						1660 Ti	1536	6GB	$279
GeForce RTX 30 系列	2020-2021	Ampere	2×FP32	第 2 代	第 3 代	3050	2560/2304	8GB	$249
						3060	3584	12GB/8GB	$329
						3060Ti	4864	8GB	$399
						3070	5888	8GB	$499
						3070Ti	6144	8GB	$699
						3080	8960/8704	12GB/10GB	$699
						3090	10496	24GB	$1,499
GeForce RTX 40 系列	2022	Ada Lovelace	2×FP32	第 3 代	第 4 代	4080	9728	16GB	$1,199
						4090	16384	24GB	$1,599

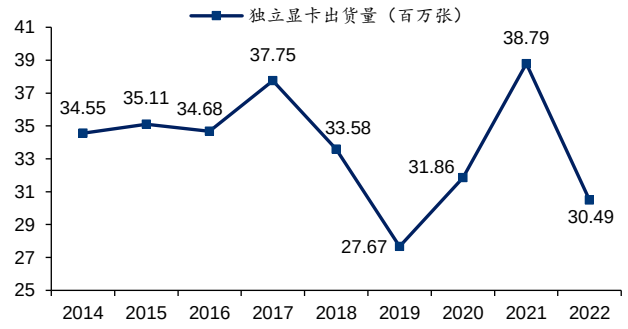
资料来源：Tom's Guide、TechPowerUp、The Verge、英伟达官网、华泰研究

图表208： 英伟达历代 GeForce 显卡价格走势



注：横轴为 GeForce 系列名，如 GeForce 400
资料来源：AROGED、华泰研究

图表209： 英伟达历年 GeForce 独立显卡出货量

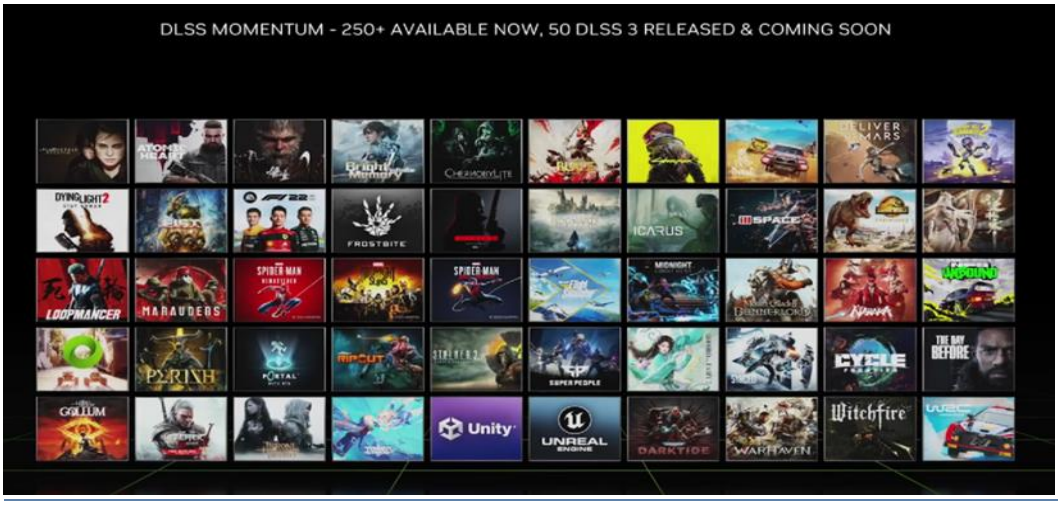


资料来源：Tom's Hardware、Jon Peddie Research、华泰研究

"光追"技术提高产品附加值。英伟达在 2018 年 RTX20 系上首发"光线追踪"技术，也是这种技术首次出现在民用级显卡上。得益于显卡性能大幅提升，光线追踪可以通过模拟光线的物理行为，实现更为逼真的阴影和反射效果，改善了半透明度和散射，将电影级质量的实时渲染带入游戏中，为玩家提供更贴近真实世界的体验。较多新游戏会在发布时特意标明"支持 NVIDIA RTX"吸引玩家。我们认为 PC 游戏玩家粘性高且付费意愿强，目前光线追踪技术还有继续提升空间，我们预计未来产品更新能够持续刺激玩家购买需求。

GeForce RTX 40 系列+新一代 DLSS 技术，引领游戏领域变革。DLSS 是英伟达推出的一项图像增强技术，它通过深度学习算法实现图像超采样，从而在游戏中提高分辨率并提高帧率，同时减少图像锯齿和噪点。DLSS 3 是由 2022 年推出的 GeForce RTX 40 系列显卡所搭载的全新第四代 Tensor Core 和光流加速器提供动力支持，可在不影响画质和响应速度的前提下，利用 AI 创造更多高质量帧。2023 年 8 月 22 日，英伟达宣布推出 DLSS 3.5，其采用光线重建技术。与 DLSS 3 仅限于 RTX 40 系列图形处理器的帧生成能力不同，DLSS 3.5 将增强追溯到 RTX 20 系列的所有 RTX 图形处理器的光线跟踪。新一代游戏的高质量图形效果、超高分辨率显示、大规模开放世界和"光追"技术得以普遍应用，但也显著增加了 GPU 的负载，降低了游戏帧率，而 DLSS 技术可提升游戏流畅度，改善游戏体验。目前已有超过 330 款游戏和应用支持 DLSS，包括《赛博朋克 2077》(Cyberpunk 2077)和《漫威蜘蛛侠：重制版》(Marvel's Spider-Man Remastered)等热门游戏，以及《深岩银河》(Deep Rock Galactic)等独立游戏大作，且《赛博朋克 2077》和《心灵杀手 2》也将在 9 月和 10 月配备 DLSS 3.5 的支持。我们认为，主流爆款的大制作游戏广泛适配 DLSS 技术，能通过提升游戏体验促使玩家购买新的游戏显卡。

图表210： DLSS 技术应用游戏一览（部分）

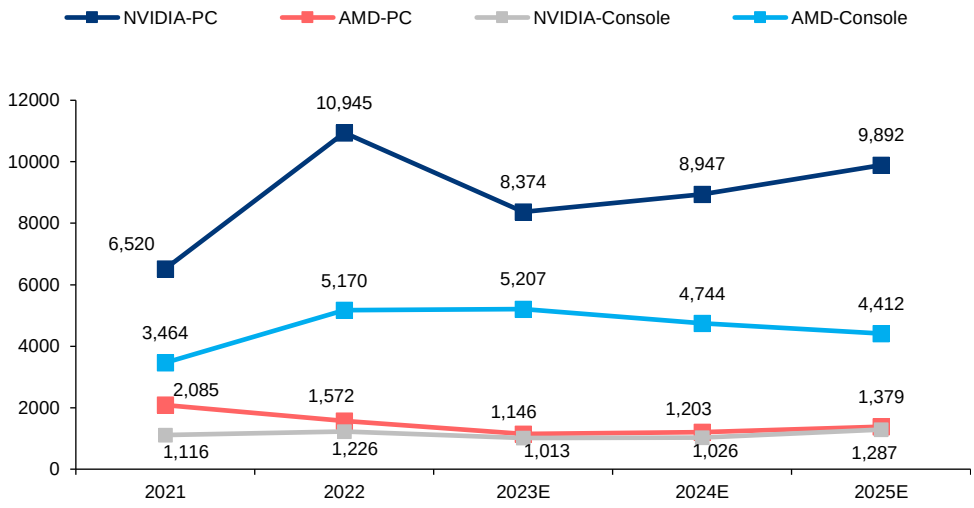


资料来源：英伟达官网、华泰研究

英伟达与 AMD 的较量：独显与主机市场各有优势。23Q1，英伟达在 PC 独显的市场份额达到 84%，相对于主要竞争对手 AMD 的 12%，以绝对优势领跑独显市场。AMD 游戏主机则为 Xbox 和 Play Station 单独提供 GPU 和 CPU，在主机市场领先英伟达。2022 年 AMD 在主机硬件领域的收入达英伟达的 422%。在主机市场，英伟达单独为 Switch 提供集成 CPU 和 GPU 的 Tegra X1 处理器，虽然推出时间较早（2015 年），但升级换代较少，且主要用于移动设备等轻量级游戏主机，贡献收入能力小于 AMD 主机硬件。

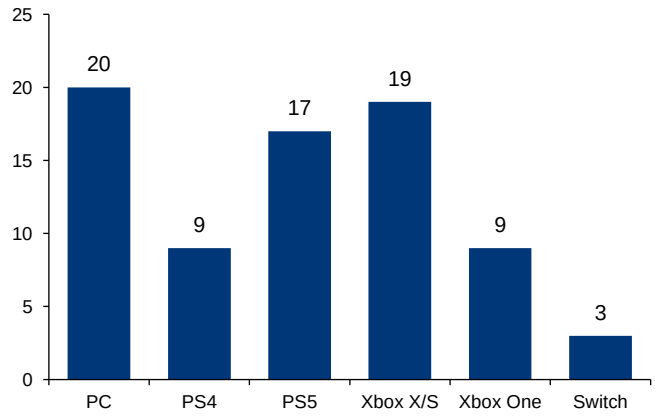
兼具高盈利性与高灵活性，PC 端独显市场长坡厚雪。对比英伟达的 PC 独显和 AMD 的 Xbox、Play Station 主机 CPU+GPU 的商业模式，由于主机制造商常常采用亏本销售的方式出售主机（在硬件的销售价格低于生产成本的情况下销售主机，并通过游戏软件销售、订阅服务等获得利润），因此尽管主机市场每年游戏收入大于 PC 游戏市场，英伟达 PC 独显收入仍远超 AMD 的主机游戏硬件收入。同时，对于 PC 显卡玩家来说，其升级灵活性更高，无需等待主机更新换代，其迭代更快且可针对不同类型游戏负载进行定制化产品推出。另外，我们对 IGN（权威游戏打分网站）评选出最受关注的 2023-25 年将发布的游戏所兼容的游戏平台进行了统计，可以看到，现今大部分游戏都兼容 PC 平台，包括 PC 独占游戏《博德之门 III》（被 IGN 评为满分作品，并获得了“编辑之选”称号），以及备受期待的《星空》、《刺客信条:幻景》、《黑神话:悟空》、《星球大战:亡命之徒》等游戏，或将进一步促进 PC 游戏市场发展，拉动英伟达游戏显卡销售。

图表211：英伟达、AMD 游戏硬件收入（单位：百万美元）



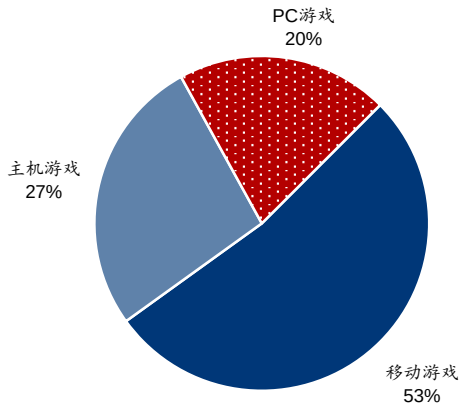
资料来源：Visible Alpha 官网、华泰研究

图表212：2023-25 年各平台高期待值游戏数量（单位：个数）



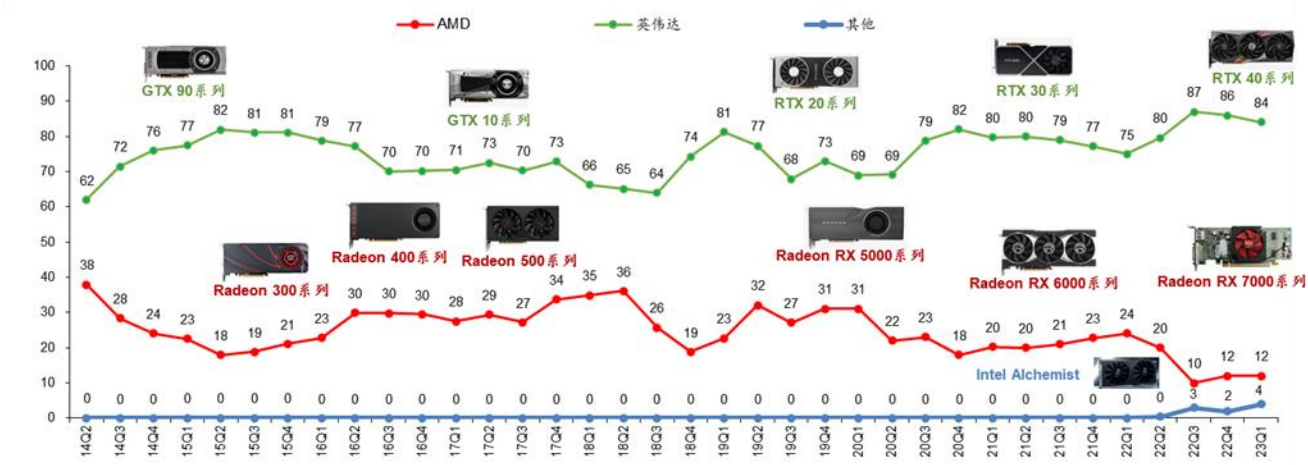
资料来源：IGN、各工作室官网、华泰研究

图表213：2023 年游戏细分市场收入占比



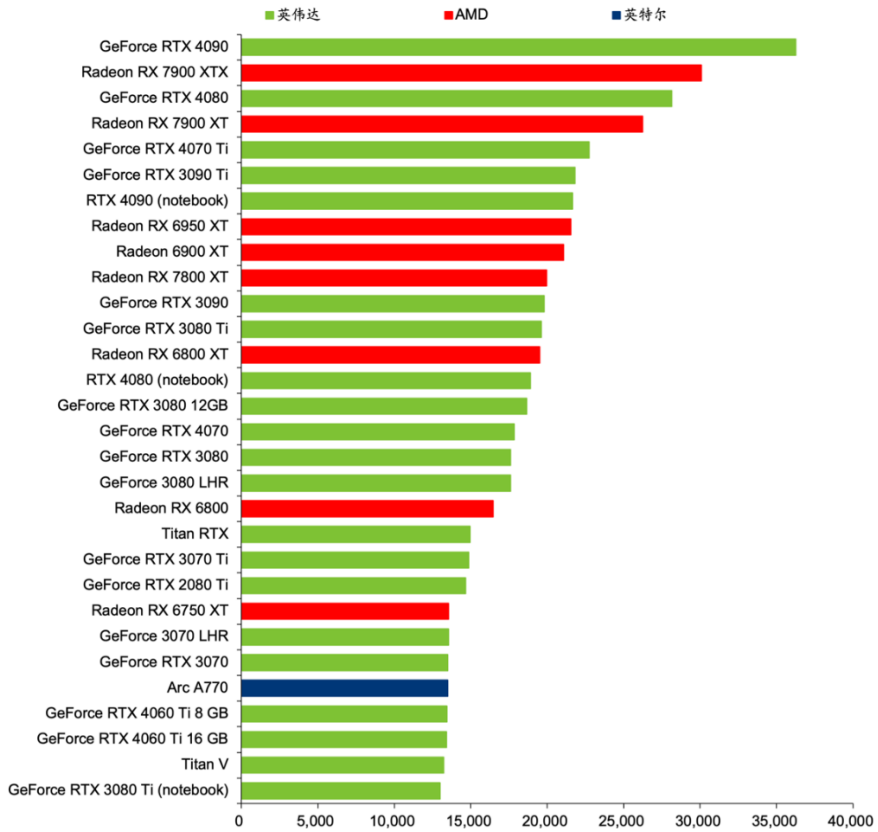
资料来源：Bankmycell、Statista、Sensor Tower、华泰研究

图表214： 全球独显 GPU 市场份额及英伟达/AMD/Intel 游戏业务主要产品系列（单位：%）



资料来源：Aroged、英伟达官网、华泰研究

图表215： 显卡 3D Mark Time Spy 分数天梯图（数据截止 2023 年 9 月 20 日）



注：Mark Time Spy 是一款由 Futuremark 开发的基于 DirectX 12 的 3D 图形性能测试工具，用于测试计算机的显卡和处理器性能，Time Spy 天梯图则是根据全球用户提交的测试成绩数据，按照排名顺序绘制的图表，展示了各款显卡在 Time Spy 测试中的相对性能表现

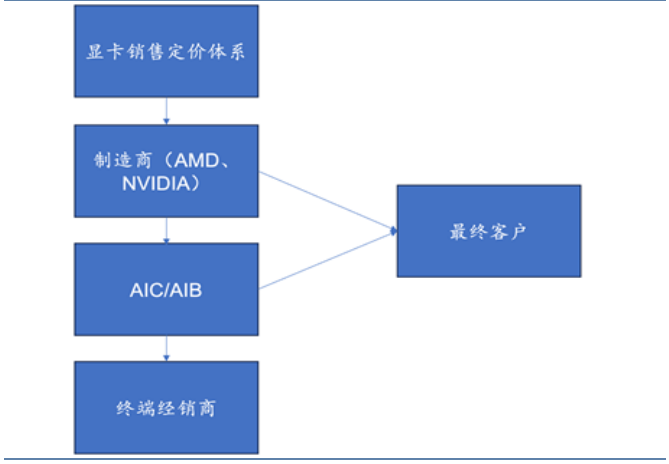
资料来源：UL Benchmarks、华泰研究

游戏显卡业务有望进入修复进程

我们认为，英伟达游戏显卡业务的焦点在于关注“矿潮”之后，公司显卡业务重回正轨，未来游戏的景气度，需要结合下游消费电子（游戏主机）和游戏增长预期、以及公司与游戏主机企业的战略合作，叠加云游戏平台的获客及盈利能力等方面。

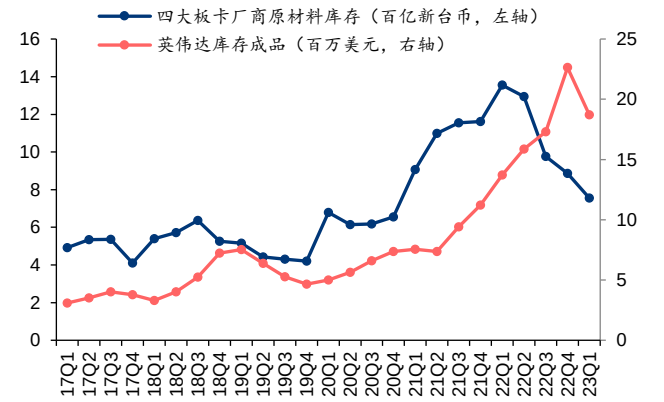
矿难带来的至暗时刻已过，虚拟货币影响渐消，沿显卡销售定价体系判断显卡恢复动力。显卡的定价与制造商、AIC (Add-in Card)/AIB (Add-in Board)（显卡授权生产厂商）和终端经销商三个参与者密不可分。其中，AIC/AIB 可以自主选择如何包装和设计风扇、显存等显卡其他核心部分，将不同显卡成品推向市场，并按照用料的高低配置进行档次划分，进而区分价格。华硕（ASUS）、技嘉（Gigabyte）、微星（MSI）和宏碁（Acer）等厂商均为英伟达主要旗舰家族。通过高频跟踪台湾 AIC 厂商的营收及库存情况，我们发现，2021 年 5 月开始，在经过一轮低价清仓后，板卡商库存金额自去年三季度起已见明显下降。且由于清库存带来的积极效应传导至上游需要一至两季度的时间，英伟达库存成品在 22Q4 已经出现下降的拐点。故我们认为显卡销售应能重归较健康轨道，英伟达 22Q4-23Q2 的财报也显示，游戏显卡业务环比正在恢复。

图表216： 显卡销售定价体系



资料来源：虎嗅智库、华泰研究

图表217： 英伟达 AIC 厂商营收及库存情况



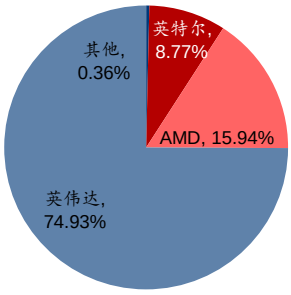
资料来源：Bloomberg、华泰研究

Steam 玩家日活数据或是为游戏显卡业务提供重要指引。当显卡主力购买者由矿工回归至游戏玩家，观察在线游戏平台 Steam 玩家活跃数量可判断玩家需求的变化，为显卡需求修复预期提供重要指引。如 20Q1 新冠疫情期间在线玩家数量激增 42.6%，全球在线日活玩家在 2020 年 4 月 2 日达到 817 万人次，传导至上游，公司 2021 年显卡收入也被带动实现显著上升。故我们认为，22Q4 到今年 Q2 的 Steam 月均游戏日活不改总体上升趋势，曾在 3 月一度超过 1000 万人次，这对英伟达下游需求的恢复是较为积极的信号。

图表218： STEAM 月均 In-game DAU 数量（单位：百万人次）

注：数据截止至 2023.9.20
资料来源：Steam、华泰研究

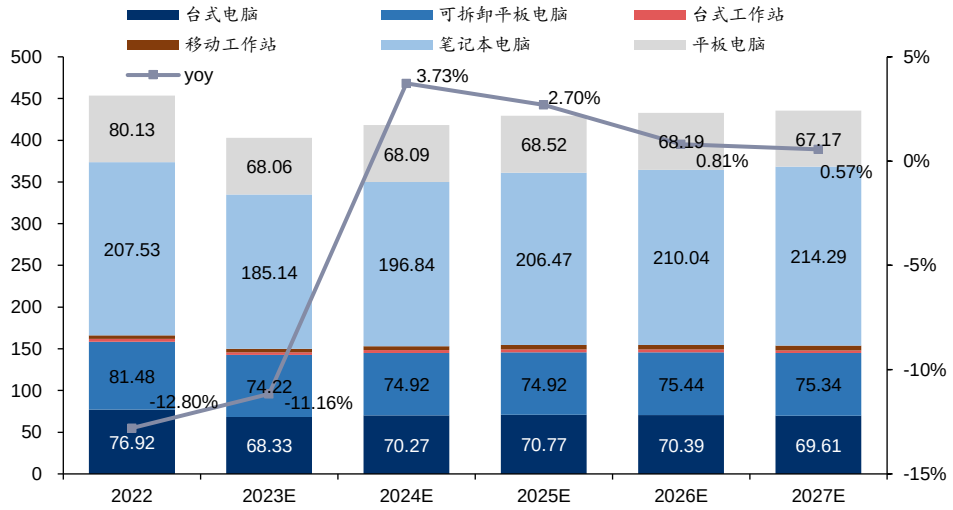
图表219： 英伟达在 Steam 硬件和软件调查中的占比（8 月数据）



资料来源：Steam、华泰研究

下行周期尾声将至，我们预计 PC 市场至 2024 年将缓慢恢复。英伟达回归消费级显卡市场，PC 出货量仍是预测其显卡业务长期走势的重要指标。2022 年，笔记本电脑和台式机出货量受疫情期间高基数影响，下降趋势尤为明显。英特尔和 AMD CEO 均预测，个人电脑将在 2023 年下半年开始复苏，且据 IDC 预测，台式电脑与笔记本市场均将在 2024 年恢复增长。故我们对 PC 渠道库存逐步消化呈谨慎乐观的观望态度，虽然短期内 PC 出货量或继续受压，但长期来看，伴随疫情复常以及挖矿影响消退，PC 市场在下半年有望迎来修复。

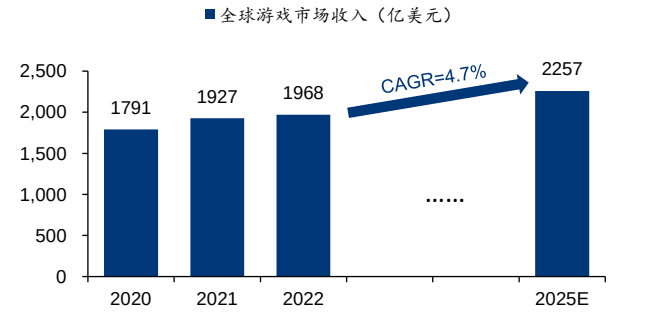
图表220：全球头部 PC 厂商出货量及预测（单位：百万台）



资料来源：IDC 官网、华泰研究

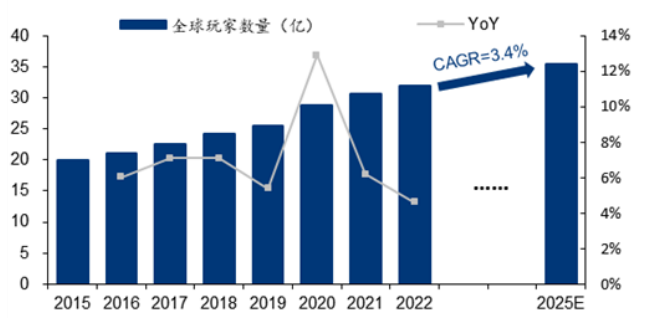
庞大玩家群体提供 PC 游戏市场稳健增长保障。根据 NewZoo 的统计，2022 年是游戏市场连续两年疫情经济增长后的修正年，全球游戏收入同比下降 4.3%。此前高速扩张的手游市场陷入疲态，作为手游的替代品，PC 游戏逆势上涨，2022 年产生约 404 亿美元的收入，同比增长 1.6%，占游戏市场总收入的 21%，而实体版及数字版 PC 游戏的下降态势对 PC 游戏市场的增长起到抵消作用。其中，Steam 作为目前全球最大的 PC 游戏分发商，为玩家提供单机和线上游戏，占据超半数的全球 PC 游戏下载量，2021-22 年 AMU 增幅达近 30%。我们认为，更大的玩家基数表示未来每年将产生可观的更新游戏 GPU 需求，为英伟达后续显卡销售带来稳定的保障。NewZoo 预计，未来几年游戏市场将会重回增长轨道，在 2025 年达到 2257 亿美元，五年间复合增长率 4.7%。

图表221：2020-2025 年全球游戏市场收入增速（单位：亿美元）



料来源：Newzoo 官网、华泰研究

图表222：2015-2025 年全球游戏玩家增速及预测（单位：亿）



资料来源：Newzoo 官网、华泰研究

主流游戏配置要求升级，带动玩家显卡更新。基于 Steam 在 2019 和 2022 年的畅销游戏推荐配置进行统计分析，排除了以游戏玩法而非画面表现为主的电竞游戏如 DOTA 2 和 CSGO 后发现，在不同年份的统计中，2019 年不需要独立显卡的畅销游戏（如支持 DirectX 10+ 的 Warframe）在 2022 年已经不再存在。同样，2019 年高居推荐配置顶部的 GTX 1060 在 2022 年则居于中枢位置，顶部位置则由 GTX 2060 取代。因此我们认为随着游戏行业的发展，游戏配置要求会逐步提升，从而激发存量游戏用户更新显卡的需求，而英伟达作为 PC 游戏显卡市场份额最大的制造商将因此受益，满足主流玩家对更高画质和流畅度的追求。

图表223：2019 和 2022 年 steam 畅销榜游戏推荐配置

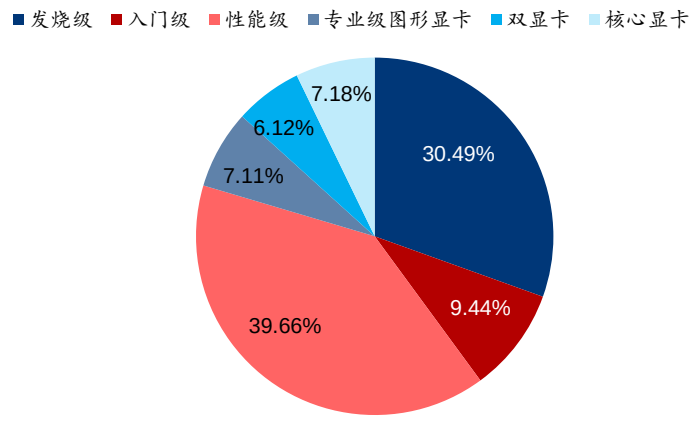
2019 畅销榜游戏	推荐英伟达显卡	2022 畅销榜游戏	推荐英伟达显卡
Monster Hunter: World	GTX 1060/GTX 1650	消逝的光芒 2	RTX 2060
PUBG: BATTLEGROUNDS	GTX 1060	艾尔登法环	GTX 1070
命运 2	GTX 970/GTX 1060	Yu-Gi-Oh! Master Duel	GTX 1650
Total War: THREE KINGDOMS	GTX 970	PUBG: BATTLEGROUNDS	GTX 1060
Sekiro: Shadows Die Twice	GTX 970	MONSTER HUNTER RISE	GTX 1060
Tom Clancy's Rainbow Six Siege	GTX 670/GTX 760/GTX 960	永劫无间	GTX 1060
Sid Meier's Civilization VI	GTX 770	使命召唤	GTX 1060
上古卷轴 OL	GTX 750	命运 2	GTX 970/GTX 1060
Grand Theft Auto V	GTX 660	Apex Legends	GTX 970
Warframe	支持 DirectX 10+		

资料来源：ZDC 互联网消费调研中心、华泰研究

我们认为，在此前虚拟货币爆火时，导致 PC 游戏玩家的硬件购置需求，长期受到高价显卡与芯片短缺等因素抑制，GTX 系列低价显卡的消费周期也给拉长。矿潮过后，PC 端游戏重掌独显主流市场，显卡紧张的供需关系渐渐恢复，此前被压制的玩家需求逐步释放，加速显卡去库存进度。根据 ZDC，用户在购买笔记本电脑时，具有发烧级和性能级的显卡是用户的主要关注点，因此考虑到 23 年下半年 PC 市场开始复苏，高端显卡销量有望增长。

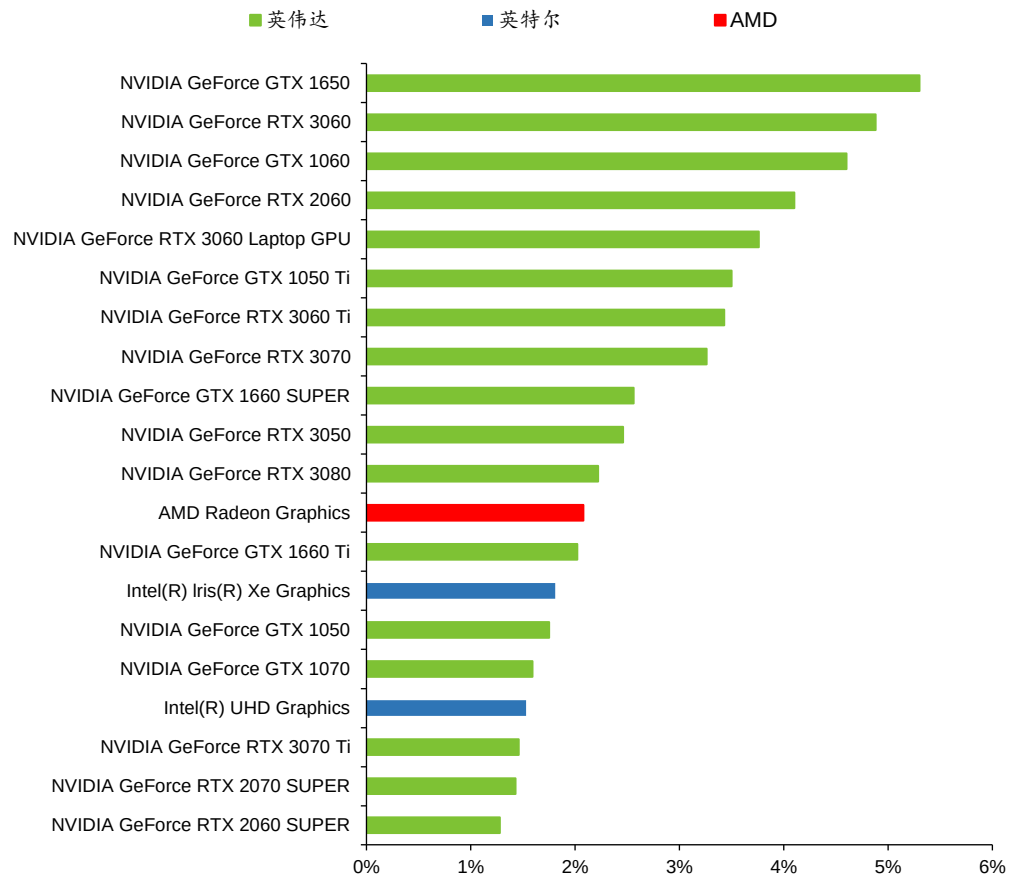
在 Steam 硬件数据统计中，2023 年 6 月共有约 40%玩家使用的显卡支持光线追踪功能（RTX 20 系列及之后），相对于去年年底不到 30%的比例而言，已见逐月提升趋势。而据英伟达在 23Q1 财报电话会上的表述，目前 RTX 装机量仅有不到 GeForce 游戏 GPU 1/3 的水平，考虑到目前主流显卡 RTX 3060 在支持光线追踪和 DLSS，性能超 GTX 1060 70%的前提下，价格仅比 GTX 1060 高 50 美元，因此玩家存在较大的替换动力。DFC intelligence 将 3 亿 PC 与主机玩家定义为“硬件驱动型消费者”，在去年《艾尔登法环》和《永劫无间》等爆款游戏的引领下，RTX 3060 在 Steam 的占有率从 2022 年 1 月的 3.72%增长到 2023 年 1 月的 8.14%，成为目前最受欢迎的 PC 显卡。随着今年更多新款 3A 大作推出，预计会有更多玩家从原来的低价显卡进一步向 40 系高性能贵价卡跨越。

图表224：2022 年笔记本显卡关注度



资料来源：ZDC 互联网消费调研中心、华泰研究

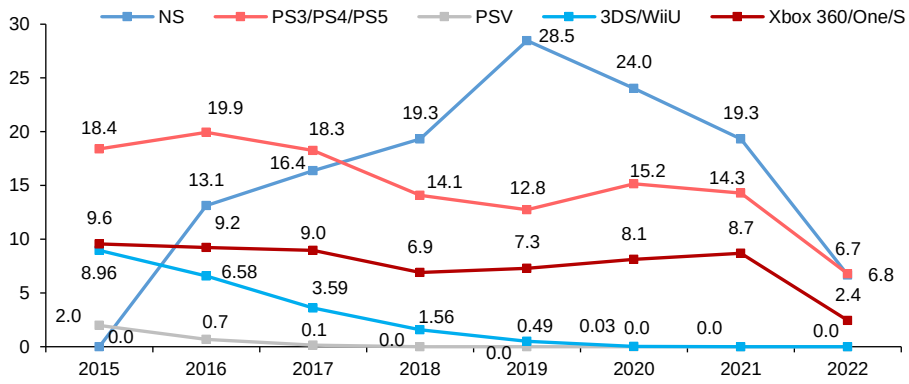
图表225：2023 年 8 月 Steam 游戏玩家显卡使用占比（TOP20）



注：数据截止至 2023.8.31
资料来源：Steam 官网、华泰研究

英伟达为 Switch 系列处理器的独家显卡供应商，或受益于任天堂 Switch 换代。目前，任天堂旗舰型主机产品 Switch 系列主要包括 Switch Original、Switch Lite 和 Switch OLED 三种类型，自推出起均选用英伟达 Tegra X1 系列芯片进行搭载。据 VGChartz 数据显示，Switch 自 2017 年发布以来，凭借其旗下马力欧、塞尔达传说和精灵宝可梦等忠实粉丝众多的游戏 IP 独家授权，在 2018-2021 年实现每年硬件产品销售量第一。2023 年 2 月，任天堂与微软签订十年协议，协议约定在合约期间，微软将 Xbox 旗下工作室及动视暴雪所开发的游戏带到任天堂的主机平台上，此合作或有利于任天堂将主机生态扩展至 Xbox 游戏。据 2023 年 7 月 31 日 VGC 新闻报告，**任天堂下一代 Switch 游戏机计划于明年推出，产品开发套件已交付给主要合作伙伴进行适配，第二代 Switch 游戏机会采用定制的英伟达 Tegra T239 处理器。**我们认为，考虑到 Switch 广大的用户基数、每年领先的销量及上一代推出时间距今已较久，新一代产品推出后将会显著带动英伟达主机游戏硬件收入。

图表226： 各游戏主机产品年销售量（单位：百万台）



资料来源：VGChartz 官网、华泰研究

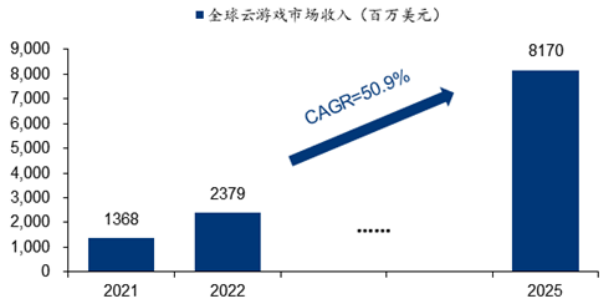
英伟达积极探索与合作伙伴布局，开辟多元化增长空间。在 CES 2023 展会上，英伟达宣布将与现代、比亚迪以及极星合作，在汽车系统上搭载 GeForce NOW，意在探索云游戏在车载终端的新流量入口。另外，公司于今年 2 月份宣布与微软建立为期 10 年的合作关系，将《我的世界》(Minecraft)、《光环》(Halo)和《微软模拟飞行》(Microsoft Flight Simulator)等 Xbox PC 游戏引入 GeForce NOW 云游戏服务。待微软完成收购动视之后，GeForce NOW 还将新增《使命召唤》(Call of Duty)和《守望先锋》(Overwatch)等游戏，为英伟达游戏云服务带来进一步成长契机。

GeForce Now 云服务商模式积极探索游戏新增长点。作为具备完善软硬件的技术型厂商，英伟达推出了云游戏平台 GeForce Now，在 2015 年开启内测，2020 年 4 月正式公测。GeForce Now 采用分级订阅制，通过为用户提供快速稳定的即时流式传播服务，帮助玩家在云端畅玩 Steam、EPIC、育碧、GOG 和 EA 等游戏库中多款大作与新游，4K 分辨率、超低延迟以及 5.1 和 7.1 环绕声支持等功能。公司在过去 3 年内升级多项游戏服务，目前订阅会员遍布全球，英伟达 2023 财年订阅会员人数已突破 2500 万。根据 Newzoo 数据，2022 年全球云游戏付费用户数量达 3170 万，云游戏的市场规模有望从 2022 年的 23.79 亿美元快速增长至 2025 年的 81.7 亿美元。在云游戏市场的蓬勃发展下，我们认为英伟达作为游戏显卡龙头，在硬件优势和游戏内容方面的积累丰厚，且多年来与游戏厂商合作广泛和深入，均为其积极探索云游戏服务提供了坚实的支撑。

图表227： GeForce Now 分级收费服务模式

料来源：英伟达官网、华泰研究

图表228： 云游戏市场规模增速（单位：百万美元）



资料来源：Newzoo 官网、华泰研究

图表229： GeForce Now 库支持的云游戏列表

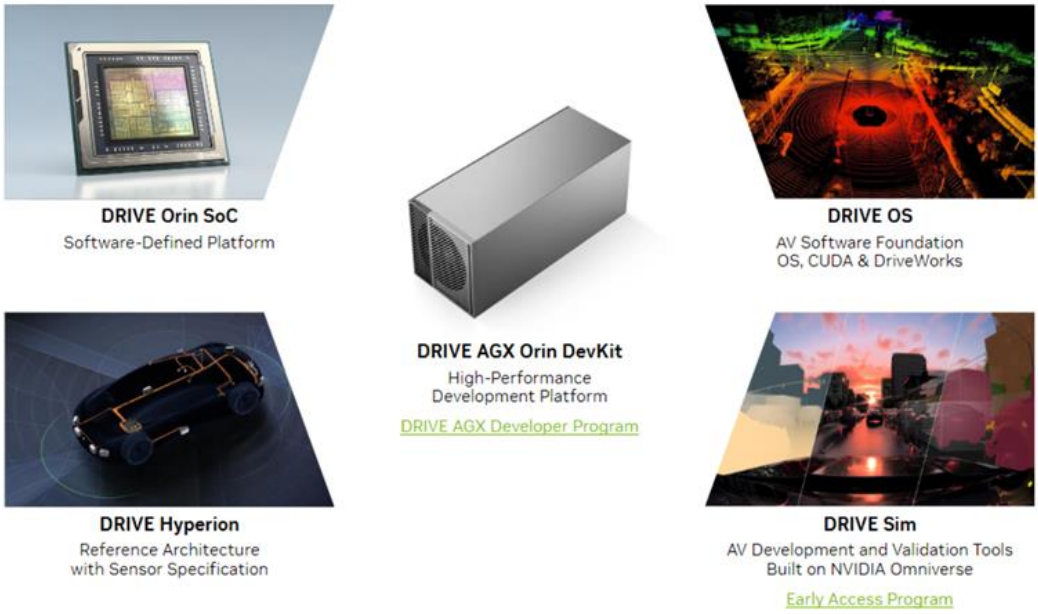
资料来源：英伟达官网、华泰研究

智能驾驶业务：从算力逐步延展至全栈智能驾驶工具链

我们预计英伟达智能驾驶业务 FY2024/ FY2025/ FY2026 的营收同比增速为 40%/30%/43%，对应营收 12.64/16.43/23.50 亿美元。英伟达自动驾驶芯片产品主要定位于 L3 及以上的高端市场。2023 年 6 月 21 日国新办吹风会提到，将启动智能网联汽车准入和上路通行试点，支持 L3+自动驾驶功能商业化。我们认为，伴随高等级自动驾驶逐步渗透，高阶自动驾驶的商业化落地将得到关键助力。根据麦肯锡在 2023 年 1 月 6 日发布报告中的预测数据，到 2030 年全球 L3 和 L4 自动驾驶市场将分别达到约 600 亿美元和 200 亿美元的规模。我们认为英伟达凭借高算力产品与生态壁垒，在 L3+市场的领先地位有望进一步扩展，预计英伟达智能驾驶业务增速将维持 30-40%左右。

从算力逐步横向延展至全栈智能驾驶工具链。英伟达自 2015 年进入自动驾驶领域，依托自身高算力 GPU 不断拓展智能驾驶业务。目前,英伟达智能驾驶业务已形成覆盖从芯片（Xavier/Orin/Atlan/Thor）、硬件计算平台（Driver AGX）、车辆操作系统平台（DRIVE OS）、中间件（DriveWorks）和应用软件（自动驾驶软件 DRIVE AV、驾驶员监控与可视化平台 DRIVE IX、以及多模态地图引擎 DRIVE Map）等的开放式平台（Drive SDK），到提供可扩展式算力支持的数据中心（DGX）和仿真平台（NVIDIA DRIVE Sim），实现全栈智能驾驶工具链。

图表230： NVIDIA DRIVE 自动驾驶汽车端到端解决方案



资料来源：英伟达官网、华泰研究

硬件层面，基于英伟达强算力基因，聚焦算力跃升。英伟达每次迭代芯片，算力均保持高增长。英伟达 2015 年所推出的第一代自动驾驶平台 DRIVE PX 所搭载的 Tegra X1 SoC 和其在 2018 年推出的第二代 SoC 芯片 Parker 算力仅约 1 TOPS，随后 30 TOPS 算力的 Xavier 芯片，到 254 TOPS 算力的 Orin 芯片陆续推出。对标业内可量产落地的车规级芯片产品，如 Mobileye Eye Q5、地平线 J5 和和芝麻 A1000 等，多数算力集中在 100 TOPS 以下。在 2021 年推出首次集成 Bluefield 数据处理单元且算力达到 1000 TOPS 的 Atlan 芯片后，公司于 2022 GTC 大会发布车载芯片 Thor，其算力达 2000TOPS。Thor 基于英伟达 Grace CPU、Hopper GPU 和 Ada Lovelace GPU 进行升级，内部拥有 770 亿个晶体管，FP8 格式下单颗算力达 2000 TFLOPs+2000 TOPS，是 Altan 的 2 倍，Orin 的 8 倍。

图表231：英伟达自动驾驶芯片

	发布时间	CPU	GPU	算力	功耗	制程	商用进度
Tegra X1	2015	4×ARM CortexA57	Maxwell GPU（256 个 CUDA 核）	1 TOPS	10W	20nm	未商用
Parker（Tegra X2）	2016	2×Denver 4×ARM CortexA57	Pascal GPU（256 个 CUDA 核）	1 TOPS	15W	16nm	2018 年量产，对应 L3 级别自动驾驶，合作特斯拉、沃尔沃等
Xavier	2017	8× NVIDIA custom Carmel ARM64	Volta GPU（512 个 CUDA 核）	30 TOPS	30W	12nm	2020 年量产，对应 L4/5 级别自动驾驶，合作小鹏、上汽等
Orin	2019	12×Arm Cortex-A78AE（Hercules）	Ampere GPU（2048 个 CUDA 核和 64 个 Sensor 核）	254 TOPS	60W	8nm	2022 年量产，对应 L3 级别自动驾驶合作梅赛德斯奔驰、丰田、小鹏、威马、蔚来、理想等
Atlan	2021	Grace-Next	Ampere-Next GPU（Lovelace）	1000 TOPS	-	-	自 Thor 发布后终止，未商用
Thor	2022	ARM Poseidon AE 核	Hopper GPU（Ada Lovelace）	2000 TOPS	-	-	计划于 2025 年投入生产,合作吉利旗下豪华电动汽车初创公司极氪

资料来源：英伟达官网、ANANDTECH、华泰研究

基于强 SoC 算力持续迭代，支撑定义软件的 AI 计算平台。英伟达于 2015 年首次推出 DRIVE 系列产品，以 Tegra X1 SoC 为自动驾驶芯片，其中 DRIVE CX 平台面向数字座舱，DRIVE PX 平台面向自动驾驶。DRIVE PX 平台构建开放和高效的研发生态，直接为汽车客户打造端到端的解决方案。此后八年时间，英伟达以平均 1~2 年更新一次的节奏迅速迭代出 DRIVE PX2、DRIVE PX Xavier、DRIVE PX Pegasus、DRIVE AGX Orin 等多个自动驾驶计算平台，在各代 SoC 基础上集成深度学习和计算机视觉加速器，充分利用 SoC 底层能力的同时，提供开发生产级自动驾驶汽车应用程序所需的硬件接口、软件和示例应用程序，以帮助车企开发大规模且复杂的软件产品。

最新芯片产品 Thor 作为行业首个中央计算单元，具有足够大的舱驾一体能力。随着汽车由传统燃油车时代的分布式架构转向电动车的域控架构，数量更少的 ECU 逐步取代几百上千规模的分布式 MCU，从而简化电子电气架构，实现中央集中计算。Thor 作为行业首个中央计算单元，标志着汽车领域由分布式 ECU 和 DCU (Domain Control Unit) 转向完全集中的功能融合型单芯片。同时，Thor 顺应“舱驾一体”的行业大趋势，具备多种配置模式，可以将其 2000 TOPS AI 算力和 2000 TFLOPs 浮点算力用于整合全车的智驾和座舱功能，能够在提供智能驾驶的同时，将一部分算力用于实现仪表盘、中控大屏等座舱计算功能，音乐、广播等信息娱乐功能，以及开空调、调节座位等舱内控制功能。车企可以用一个芯片实现未来高阶智能驾驶的硬件预埋，为未来的算法和应用发展预留充足空间。

图表232：英伟达自动驾驶计算平台

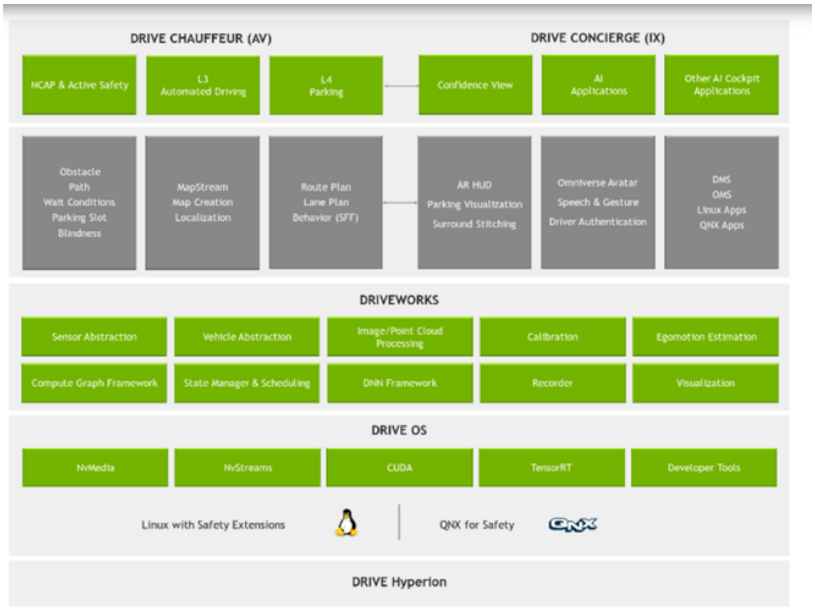
产品名称	发布时间	SoC 芯片	单浮点算力	深度学习算力（INT8）	功耗	自动驾驶级别
DRIVE PX	2015.1	Tegra X1	2.3 TFLOPS	-	-	L2
DRIVE PX2	2016.1	Tegra X2（Parker）	8 TFLOPS	24 TOPS	250W	L2/L3
DRIVE PX Xavier	2017.1	Tegra Xavier（Xavier）	-	30 TOPS	30W	L3/L4
DRIVE PX Pegasus	2017.10	Tegra Xavier	-	320 TOPS	500W	L4/L5
DRIVE AGX Pegasus	2019.1	Tegra Xavier	-	320 TOPS	-	L2/L3
DRIVE AGX Orin	2019.12	Orin	-	2000 TOPS	750W	L2/L3/L4/L5
Drive Thor	2022.9	Thor	2000 TFLOPS	2000 TOPS	-	L2/L3/L4/L5

资料来源：英伟达官网、ANANDTECH、AnyConnect、华泰研究

英伟达高度开放软硬件生态，DRIVE Hyperion 彰显进军全流程解决方案的野心。英伟达将 DRIVE Hyperion 定位为自动驾驶汽车参考框架平台，除英伟达自动驾驶 SoC AI 的计算平台外，还集成可扩展的软件堆栈 SDK 和传感器硬件设备组为一体，并在底层架构上为车企和 Tier 1 等厂商保留了较高的灵活度。

以 Hyperion 8.1 为例，除 DRIVE AGX Orin 计算平台开发者套件外，还包括：
1）由英伟达 Tier 1 及传感器产业合作伙伴提供的 12 个摄像头、9 个雷达、1 个激光雷达和 12 个超声波的传感器组；
2）帮助在车端 DRIVE OS 操作系统上构建应用软件的模块化堆栈 SDK，它包括基础 NVIDIA DRIVE® OS 和 DriveWorks SDK，以及高级应用，如高度自动化监督驾驶 DRIVE AV、AI 驾驶舱 DRIVE IX、AI 行车助手 DRIVE Concierge 和用于自动驾驶行泊一体的 DRIVE Chauffeur 等。

图表233：英伟达 DRIVE Hyperion 参考架构概览



资料来源：英伟达官网、华泰研究

高低阶自动驾驶市场竞争有别，英伟达能否实现双丰收？

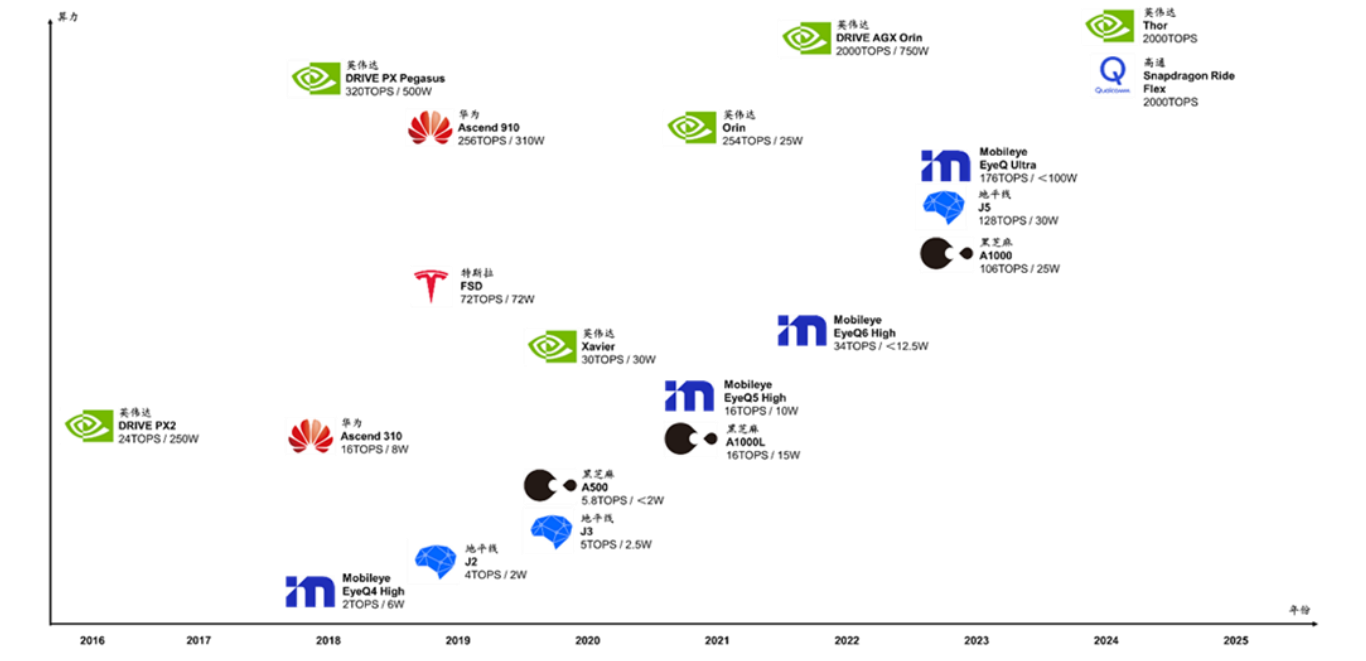
高低阶自动驾驶市场竞争有别。自动驾驶从 L1 到 L5，伴随感知、决策和执行三大任务的复杂程度逐步提高。据华为官网数据，一般来说，L2 自动驾驶等级所需的计算能力低于 10 TOPS，L3 等级需求在 30-60 TOPS，L4 等级需求超过 100 TOPS，而 L5 等级需求超过 1000 TOPS。根据这一计算能力需求，我们认为，以 L3 级自动驾驶为界，自动驾驶计算芯片市场可以根据芯片算力水平分为以下两部分：

- **ADAS 至 L2+ 级的低阶自动驾驶芯片市场：此市场主要集中于乘用车**，车辆系统被定义为驾驶员辅助系统，对芯片算力要求相对较低。目前，L2 以下市场渗透率逐步提升，需综合考虑算力、功耗、成本和量产等多重因素。在满足算力要求的范围下，低功耗、低成本和可提供成熟方案并快速装车量产的芯片，有望加速渗透自动驾驶功能相对初级的入门/基础款车型。
- **L3 及以上的高阶自动驾驶芯片市场：此市场目前更多是针对商用车**，车辆系统需具备在特定情况下完全接管驾驶任务的能力。在没有人类干预的情况下进行车辆自主驾驶，意味着复杂环境下更为精准的自主感知与决策能力和突发状态中更为智能的故障监测与处理要求。故对比低阶市场，算力的高低相较功耗、成本和量产等其他因素在芯片选取时更为重要。根据麦肯锡的预测数据，到 2030 年全球 L3 和 L4 自动驾驶市场将分别达到约 600 亿美元和 200 亿美元的规模。另外，目前部分 ADAS 到 L2+级市场的中高端车型已开始提早配置大算力芯片，瞄准 L3+自动驾驶，以“硬件预埋”方式提供灵活算力和可拓展性，支持后续高算力要求下“软件定义车辆”的开发。

英伟达瞄准高算力芯片，核心技术能力突出

从计算能力来看，英伟达自动驾驶产品大部分定位在 L3 及以上等级，并在 L4/L5 高端自动驾驶竞技场中占据领先地位。根据以上计算能力标准，英伟达的 Xavier 芯片能够满足 L3 自动驾驶等级的需求，而 Orin 和 Thor 平台则能够达到 L4/5 级别。我们梳理了英伟达芯片产品线相比国内外主流厂商，英伟达定位在高端自动驾驶市场（L3 及以上），相对较少涉及 L1-L2 领域，且具备 2000 TOPS 算力的旗舰产品 Thor 发布时间 2022 年也早于其他厂商，具备技术上的先发优势。根据黑芝麻数据显示，截至 2022 年，英伟达在高算力（50 TOPS 以上）SoC 方面市场占有率（按出货量）超过 80%，而仅次于其的地平线与黑芝麻在此市场有 5%左右市场份额。

图表234：英伟达及其他厂商的自动驾驶芯片及 DCU 方案算力对比



资料来源：各公司官网、华泰研究

得益于技术上的先发优势与成熟生态，英伟达自动驾驶合作伙伴遍布全球，涵盖国内外各大厂商，其自动驾驶合作项目可以从商用车和乘用车两方面进行划分。在商用车领域，英伟达的合作伙伴主要针对 L3 以上的自动驾驶，如 Robotruck 和 Robotaxi 相关公司，国内外知名企业包括滴滴、小马智行和 Aurora 等。乘用车方面，英伟达同样聚焦于较高级别的项目，合作伙伴包括中国新能源汽车制造商蔚来、小鹏和理想等，以及国际知名品牌奔驰、捷豹路虎等。

与本土 Tier 1 德赛西威深度合作对英伟达进军中国市场至关重要。自动驾驶芯片是自动驾驶域控制器的上游产品，由于 Tier 1 厂商在汽车系统的集成、合规和渠道推广等多方面经验丰富，自动驾驶芯片厂商往往在早期产品扩张时期会选择与当地 Tier 1 厂商合作从而打开市场并进行推广。目前，英伟达所合作的 Tier 1 供应厂商包括博世、采埃孚、海拉、奥托立夫、大陆、德赛西威等。对于中国市场，德赛西威是英伟达自动驾驶 SoC 最早期也是最重要的合作伙伴之一。德赛西威为英伟达在中国选择的第一家本土自动驾驶 Tier 1 厂商，其搭载英伟达 Xavier SoC 的域控制器 IPU03 于 2020 和 2021 年在小鹏 P7 和 P5 实现量产上车。2021 年，小鹏全年交付量达到 98155 辆，其中小鹏 P7 达到 60569 辆，为小鹏汽车全年交付量的 62%，全球销量榜排名第 19。小鹏强劲销量帮助英伟达打响了其芯片在中国量产上车的“第一枪”。目前，搭载英伟达 Orin SoC 的 IPU04 域控制器已在 2022 年量产的理想 L9 上线。

图表235：英伟达合作客户主要针对高端自动驾驶

定点类型		公司	服务目标等级	代表车型	平台算力	量产年份
乘用车		比亚迪 BYD	-	腾势 N7	NVIDIA DRIVE Orin	2023
		蔚来 NIO	L3-L4	ET5/ET7	NVIDIA DRIVE Orin（*4）	2022
		理想 Li	L3-L4	L9/X01	NVIDIA DRIVE Orin	2022/2023
		小鹏 XPeng	L3-L4	P5/P7/G9	NVIDIA DRIVE AGX Xavier/Orin	2021/2020
		上汽集团 SAIC	L3-L4	上汽智己	NVIDIA DRIVE Xavier/Orin	2022
		威马汽车 WM	-	M7	NVIDIA DRIVE Orin（*4）	2022
		集度 JIDU	L4	Robo-01	NVIDIA DRIVE Orin	2023
		极氪 ZEEKR	-	-	-	2025
		路西德汽车 Lucid	-	-	-	-
		捷豹路虎	-	-	NVIDIA DRIVE Orin	2025
		瑞马克 Rimac	L4	-	-	-
		法拉第 Faraday Future	-	FF91/FF81	NVIDIA DRIVE Orin	2023/2024
		越快 VinFast	-	-	NVIDIA DRIVE Orin	-
		沃尔沃 Volvo	-	XC90/EX90	NVIDIA DRIVE Orin	-
		梅赛德斯奔驰	L2-L4	-	-	2024
商用车	RoboTaxi	滴滴 DiDi	L4	-	-	-
		小马智行 Pony.AI	L4	-	NVIDIA DRIVE Orin	-
		文远知行 WeRide	L4	-	NVIDIA DRIVE AGX Pegasus/Hyperion（下一代）	-
		元戎启行 DeepRoute	L4	-	DRIVE Hyperion/Orin	-
		Oxbotica	L4	-	NVIDIA DRIVE Orin	-
		安途 AutoX	L4/L5	-	-	-
		Cruise	L4/L5	-	NVIDIA DRIVE GPU	-
		Zoox	L5	-	-	-
		奥兹莫比尔 Aurora	L4/L5	-	NVIDIA DRIVE Xavier	-
	RoboTruck	图森未来 TuSimple	L4	-	NVIDIA DRIVE Orin	-
		Einride	L4/L5	-	NVIDIA DRIVE Orin	-
		Plus	L4/L5	-	NVIDIA DRIVE Orin	-
		Locomation	L4/L5	-	NVIDIA DRIVE Orin	-
		纳威司达 Navistar	L4/L5	-	-	24 年开始生产
		Embark	L4/L5	-	-	-
		Nuport	L4/L5		NVIDIA DRIVE Orin	-
		Kodiak Robotics	L4	-	NVIDIA DRIVE Orin	-
		Outrider	L4/L5	-	-	

注：-表示暂未披露

资料来源：英伟达官网、各厂商官网、华泰研究

值得一提的是，鉴于较低端的乘用车自动驾驶市场规模大，英伟达也希望能于当中分一杯羹。但不可忽视的是，部分 ADAS 至 L3 级的自动驾驶和座舱芯片玩家已早在该市场较深度布局。我们认为，英伟达虽在高阶赛道具有先发优势，但低阶赛道对算力的要求不高，加上竞争对手众多，竞争格局类似 AI 推理端。因此我们认为，低阶赛道市场或会是百花齐放，而英伟达或也能占一席位，但不一定能复制其在高阶赛道的主导地位。

ADAS 至 L3 以下的芯片市场多方逐鹿，竞争趋于白热化。此市场主要参与方包括海外厂商，如 Mobileye、高通和特斯拉等，以及国内厂商，如地平线、黑芝麻和华为等。Mobileye 软硬件一体的完整解决方案可装车即用，帮助车企快速实现 ADAS 功能，但其软硬件耦合且算法相对封闭的“黑盒”模式较为局限。国内芯片厂商地平线和黑芝麻产品矩阵已逐步形成，并具备本土化服务能力，有助于响应国内车企功能快速迭代的需求。消费电子龙头高通也积极布局自动驾驶芯片，结合智能座舱域经验以提供功耗比和性能较高的芯片产品，定位中高端自动驾驶。

图表236：英伟达及其他厂商的自动驾驶芯片方案

芯片	英伟达			Mobileye							Tesla
	Xavier	Orin	Thor	EyeQ2	EyeQ3	EyeQ4	EyeQ5	EyeQ6 Light	EyeQ6 High	EyeQ Ultra	FSD
AI 算力（TOPS）	30	254	2000	0.026	0.0256	2	15	5	34	176	72
功耗（W）	30	60	-	2.5	2.5	3	10	3	-	<100W	72
量产时间	2020	2022	2025	2010	2014	2018	2021	2023	2024	2025	2019
适配场景	L2/L3	L2-L5	L2-L5	L1	L2	L2	L4	L1-L2	L4	L4	L3
业务模式			Tier 2							Tier 2	车企

芯片	高通	华为		地平线			黑芝麻			
	Snapdragon Ride Flex	310	Ascend 910	J2	J3	J5	A500	A1000L	A1000	A1000Pro
AI 算力（TOPS）	2000	16	512	>4	5	128	5.8	16	58	>106
功耗（W）	-	8	310	2	2.5	30	<2	<5	<8	25
量产时间	2024	2018	2019	2019	2020	2023	2020	2021	2021	2023
适配场景	L4/L5	L2-L4	L4	L1-L2	L1-L2	L3-L4	L0-L2	L0-L2	L2+/L3	L3-L4
业务模式	Tier 2		Tier 1			Tier 2				Tier 2

备注：-表示暂未披露
资料来源：各公司官网、华泰研究

图表237：英伟达及其他厂商的自动驾驶 DCU 方案

DCU	英伟达		特斯拉	高通	华为		地平线	
	DRIVE PX Pegasus	DRIVE AGX Orin	FSD HW3.0	Snapdragon Ride	MDC 610	MDC 810	Matrix 2	Matrix 5
算力(TOPS)	320	2000	144	>700	160	>400	16	512
功耗（W）	460	750	72	130	-	-	20	-
算力功耗比(TOPS/W)	0.7	2.67	1.44	5.4	-	-	0.8	-
适用等级	L2-L5	L5	L3	L2-L3	L3-L4	L4-L5	L2	L2-L4
传感器支持	13 摄像头 +5 毫米波雷达 +12 超声波雷达 +1 车内摄像头	12 外部摄像头 +9 毫米波雷达 +12 超声波雷达 +2 LiDAR +3 内部摄像头	8 摄像头 +1 毫米波雷达 +12 超声波雷达	8 摄像头 +6 雷达	16 摄像头 +6 毫米波雷达 +16 超声波雷达 +8 LiDAR	13 摄像头 +6 毫米波雷达 +12 超声波雷达 +3 LiDAR	单路 1080P@30fps 或 4 路 1080P@30fps +激光雷达	8 视频接入通路（最高 8MP）+多路毫米波雷达、4D 成像雷达，激光雷达、超声波雷达
计算处理器	2 Xavier + 2 GPU	2 Orin + 2 GPU	2 FSD + 1 GPU	SoC + ASIC	昇腾 310	-	基于 J2	基于 J5
量产时间	2020	2022	2019	2022	2019	2021	2020	2023

备注：-表示暂未披露
资料来源：各公司官网、华泰研究

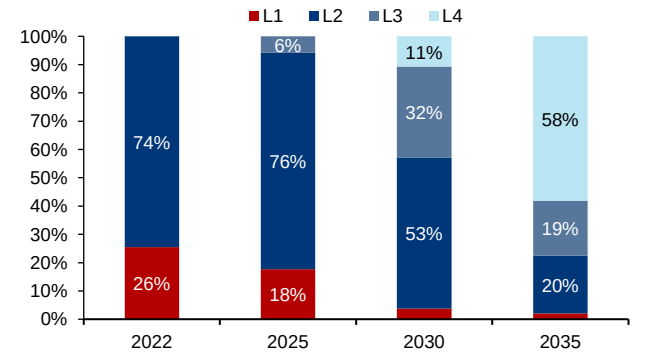
ADAS 市场成熟玩家众多，英伟达优势或不显著。目前，英伟达 Orin 系列产品已配置于蔚来 ET7、小鹏 P7、理想 L9 和沃尔沃 EX90 等高端车型，英伟达并于 GTC 2023 中表示装配于比亚迪下一代王朝和海洋系列的多款车型中。英伟达也与富士康达成战略合作，提供基于 Orin 芯片的 ECU 产品，进一步提升 Orin 芯片的用户规模，扩大生态圈。布局 L3 以下级别市场或可帮助英伟达培养用户习惯并建立合作基础，但我们认为目前低级别赛道的玩家较多，英伟达若布局这一赛道，需面临众多具备先发优势的成熟玩家，竞争激烈，另外，此市场对于计算能力的要求也较低，因此标榜高算力的英伟达芯片也较难突出优势。ADAS 至 L3 以下的市场中众多厂商已与车企构成一定的合作基础，如 Mobileye 自 2014 年开始，其 EyeQ2 CSoC（Vision System on a Chip）便已在沃尔沃、宝马和日产等车辆上进行配套，具备一定的先发优势。从前装上车情况来看，Mobileye 受益于其成熟的完整解决方案，截至 2023 年 6 月，其在全球 ADAS 市场渗透率高达 70%+，依旧是 ADAS 市场的领头羊。地平线深耕国内 ADAS 市场，进入上汽、长城和比亚迪等头部车企供应体系，在国内整车项目合作进度较为领先。

图表238： 英伟达及其他厂商的合作车企情况

资料来源：各公司官网、华泰研究

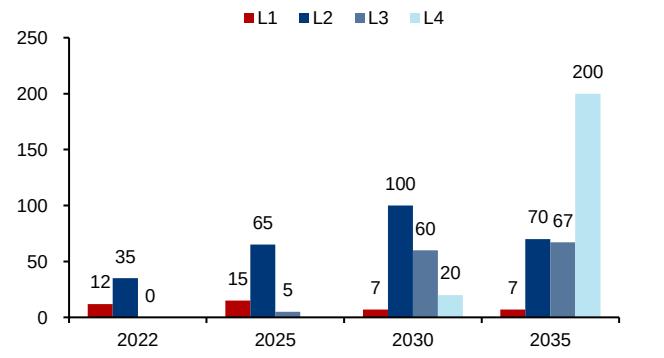
我们看好英伟达在 L4/L5 领域的绝对领先地位及其长期发展。英伟达自动驾驶产品主要定位于高端自动驾驶领域。因此对英伟达而言，L3 及以上市场的放量将是其需求端的主要增长动力之一。但目前全球市场中，L3 及以上在乘用车的渗透率仍然偏低，并以商用车为主。2023 年 6 月 21 日，国新办吹风会提到，将启动智能网联汽车准入和上路通行试点，支持 L3+自动驾驶功能商业化。政策的施行将为高阶自动驾驶的商业化落地提供关键助力。根据麦肯锡的预测数据，到 2030 年全球 L3 和 L4 自动驾驶市场将分别达到约 600 亿美元和 200 亿美元的规模，英伟达自动驾驶业务将凭借其在此市场中的领先地位迎来快速发展。

图表239： 全球自动驾驶市场各级别营收规模占比 (单位:十亿美元)



资料来源：高通官网、华泰研究

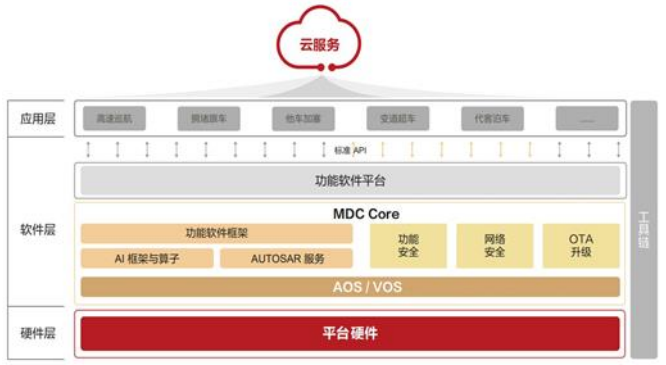
图表240： 全球自动驾驶市场营收规模 (单位:十亿美元)



资料来源：mckinsey 官网、华泰研究

华为和高通也希望在高端自动驾驶分一杯羹。目前，华为推出了计算平台 MDC 系列，并借力 5G 网络和 ICT 技术实现“车-路-云协同”，加快如露天矿和港口等场景的 L4 级商用落地。高通作为座舱芯片的头部玩家，于 2020 年推出 Snapdragon Drive 平台，并于 2023 年 1 月发布可同时支持数字座舱、ADAS 和 AD（Automated driving）功能的 Snapdragon Ride Flex SoC 芯片，最高算力达 2000 TOPS。

图表241：华为 MDC 整体架构

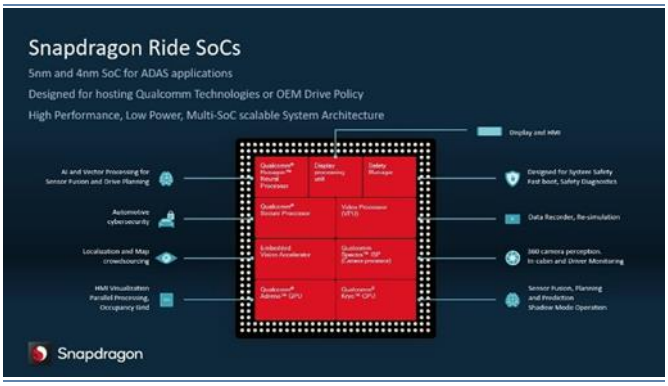


资料来源：华为官网、华泰研究

图表242：华为露天矿无人驾驶解决方案总体架构

资料来源：华为官网、华泰研究

图表243：高通 Snapdragon Ride 平台



资料来源：高通官网、华泰研究

图表244：高通 Snapdragon Ride Flex 芯片

资料来源：高通官网、华泰研究

综合对比下，我们认为英伟达在自动驾驶的核心竞争优势，主要包括以下四方面：

1）硬件层面长期积累优势明显。自动驾驶大算力芯片龙头，由针对 L3 以上车规级平台产品开始，并向 L3 以下的庞大市场布局；

2）软件层面生态完善，增强客户粘性。芯片算力是英伟达的第一护城河，而软件生态就是英伟达第二护城河。英伟达的全套自动驾驶软件工具链，包括功能软件层 DriveWorks、应用软件层 DRIVE AV 和 DRIVE IX，逐步培养用户的粘性，提高了英伟达产品的替代难度；

3）从软硬一体形成生态闭环，到软硬解耦实现各自升级，精准踩中不同客户需求。英伟达提供端到端的解决方案，以及车端到云端训练。从软件角度，作为整个生态的基石，深度学习算法基于 CUDA 和 TensorRT 工具进行，竞品目前较难绕过 CUDA 去兼容英伟达的生态；从硬件角度，CUDA 的核心设计都和公司 GPU 的硬件相对应，使公司具备提供软硬件全栈解决方案的能力。同时，公司“白盒子”方案也能够“分而治之”，硬件通用化的同时，软件支持下游二次开发的编程平台，可以在不同车型上灵活复用，适应车企快速迭代和优化成本的需求；

4）成熟的客户群和稳定的订单。一款自动驾驶芯片从交样到最终 SOP 量产交付，所需时间至少要 1~2 年（如 Xavier 2018 年交样，2020 年量产上车）。目前为自动驾驶领域抢跑期，竞争激烈，车企的供应商替换成本高。英伟达广泛对接造车新势力、传统车企与自动驾驶公司，自首个自动驾驶计算平台 DRIVE PX 发布起就与下游厂商建立了稳固的“车云一体”合作关系。

盈利预测与估值：首次覆盖给予“买入”评级，目标价 650 美元

数据中心业务：我们预计数据中心业务 FY2024/FY2025/FY2026 的营收同比增速为 152%/73%/40%，对应营收为 378.13/654.16/915.82 亿美元。我们认为英伟达在 FY2Q24 给的乐观指引，以及台积电在 FY2Q23 电话会上对未来五年 AI 服务器销售额 CAGR~50%的预测，均表明 AI 应用需求在未来数年将持续旺盛，英伟达数据中心业务有望驶入快速发展轨道。我们认为目前限制 GPU 出货量的主要瓶颈在于台积电的先进封装产能。台积电在 23Q2 法说会上预计 CoWoS 产能持续紧张，并将维持至 2024 年底。据 Digitimes 在 7 月 14 和 21 日的报道中提到，台积电正积极扩产，包括竹南、龙潭和台中三地；2023 年产能至少 12 万片，2024 年将达 24 万片，而英伟达将一共取得约 15 万片（1 片可切约 25 颗）。此外，报道也提到，英伟达也在考虑将封装需求外溢到 Amkor 和联电。因此，我们认为未来 GPU 供给瓶颈将得到一定程度的缓解，GPU 的出货节奏将迎来加速。

游戏显卡业：我们预计游戏显卡业务 FY2024/FY2025/FY2026 的营收同比增速为 20%/25%/30%，对应营收为 108.80/136.01/176.81 亿美元。我们认为 2022 年以来以太币机制转换，以及其他挖矿事宜对公司游戏显卡业务所造成的影响将逐渐消退。我们也从 FY23Q4 开始看到拐点，我们认为游戏显卡业绩下滑趋势将逐渐收窄。因此，我们预计 FY2024 年英伟达游戏显卡业务同比增长为 20%，营收达 109 亿美元。PC 方面，IDC 预计全球 PC 出货总量继今年第一季度出现同比三成跌幅后，将在第二和三季度跌幅收敛至 17%及 7%，在第四季度重现上升趋势。游戏方面，2024 年《黑神话：悟空》和《星球大战：亡命之徒》等 3A 游戏巨作预计全面上线。在明年 PC 销量将开始恢复，以及游戏市场活力释放的前景下，我们认为，作为 Switch 芯片供应商，FY2025 年任天堂 Switch 2 的发布将为英伟达游戏业绩恢复提供一定支撑，重回健康增长。因此，我们预计 FY2025 年英伟达游戏显卡业务同比增长为 25%，营收达 136 亿美元。FY2026 年，根据英伟达 6 月简报路线图，鉴于 Ada Lovelace Next 游戏架构相关芯片有望发售，新系列游戏显卡性能优势将为游戏业务新成长提供支撑，我们预计 FY2026 年英伟达游戏显卡业务同比增长为 30%，营收达 177 亿美元。

智能驾驶业务：我们预计智能驾驶业务 FY2024/FY2025/FY2026 的营收同比增速为 40%/30%/43%，对应营收为 12.64/16.43/23.50 亿美元。英伟达的自动驾驶芯片定位于高端市场，目前在 ADAS 到 L2 市场布局较少，故短期来看，其在该市场面临众多成熟玩家，因此或难以在此市场有所突破。但国新办吹风会提到，将启动智能网联汽车准入和上路通行试点，支持 L3+自动驾驶功能商业化。我们认为，政策的推行将为高阶自动驾驶的商业落地提供关键助力。英伟达凭借高算力硬件与软件生态壁垒，在 L3+市场的领先地位有望伴随高等级自动驾驶逐步渗透得到巩固，我们预计公司智能驾驶业务增速将维持在 30-40%左右。

专业可视化业务：我们预计专业可视化业务 FY2024/FY2025/FY2026 的营业收入同比增速-12%/20%/15%，对应营业收入 13.59/16.30/18.75 亿美元。2021 年，公司在推出 Omniverse 虚拟协作平台，专业可视化业务取得近一倍增长。2022 年伴随“矿难”来临，以及宏观经济不振所导致 B 端企业付费意愿萎靡，游戏业务与专业可视化业务双双受挫。但由于 2022 年上半年，Omniverse 已经在供应链数字化优化与设计等方面获得亚马逊和百事公司等稳定大标杆客户，我们认为，伴随经济形势伴随 Ominiverse 生态系统迅速扩大，将会在 FY2025 扭转同比下降的趋势，且在长期来看，Omniverse 立足英伟达底层图像加速计算能力，将受益于全真模拟数字化设计的市场需求扩大。

OEM 及其他业务：我们预计 OEM 及其他业务 FY2024/FY2025/FY2026 的营收同比增速为-10%/30%/35%，对应营收为 4.10/5.32/7.19 亿美元。受虚拟货币进入泡沫阶段导致 CMP 专用矿卡销量受挫影响，英伟达 OEM 及其他部分收入自 FY22Q3 起大幅下降。据前文分析，我们预计 2022 年矿难已出现拐点，影响将逐渐消退。且考虑到 PC 厂商库存积压情况将逐渐得到缓解，消费电子端 PC 出货量将实现稳定修复，我们认为公司 OEM 及其他业务将逐渐步入正常发展轨道。

费率及利润端：公司销售、行政及一般费用率基本保持稳健，或随产品上量继续保持在目前水平；公司新品迭代较快，我们预计研发费率后续将随着收入规模扩大而有所下降，但仍会维持在 16%的较高水平。随着高性能、高毛利及高单价的 H100 等数据中心产品占比提升，我们预计公司毛利率和净利率也会随之提高，FY24-26 年毛利率和净利率将分别从 70%提升到 73%及 47%提升到 50%。

图表245：英伟达盈利预测（单位：百万美元）

	FY2021A	FY2022A	FY2023A	FY2024E	FY2025E	FY2026E
数据中心	**$6,696**	**$10,613**	**$15,005**	**$37,813**	**$65,416**	**$91,582**
YoY	*124.47%*	*58.50%*	*41.38%*	*152.00%*	*73.00%*	*40.00%*
游戏业务	**$7,759**	**$12,462**	**$9,067**	**$10,880**	**$13,601**	**$17,681**
YoY	*40.61%*	*60.61%*	*-27.24%*	*20.00%*	*25.00%*	*30.00%*
专业可视化	**$1,053**	**$2,111**	**$1,544**	**$1,359**	**$1,630**	**$1,875**
YoY	*-13.12%*	*100.47%*	*-26.86%*	*-12.00%*	*20.00%*	*15.00%*
智能驾驶	**$536**	**$566**	**$903**	**$1,264**	**$16,43**	**$2,350**
YoY	*-23.43%*	*5.60%*	*59.54%*	*40.00%*	*30.00%*	*43.00%*
OEM 及其他	**$631**	**$1,162**	**$455**	**$410**	**$532**	**$719**
YoY	*24.95%*	*84.15%*	*-60.84%*	*-10.00%*	*30.00%*	*35.00%*
公司总营收	**$16,675**	**$26,914**	**$26,974**	**$51,725**	**$82,823**	**$114,207**
YoY	*52.73%*	*61.40%*	*0.22%*	*91.76%*	*60.12%*	*37.89%*
毛利率	**62.34%**	**64.93%**	**56.93%**	**70.00%**	**71.00%**	**73.00%**
销售、行政及一般费用率	**11.63%**	**8.05%**	**9.05%**	**6.00%**	**5.50%**	**7.00%**
研发费用率	**23.53%**	**19.57%**	**27.21%**	**16.00%**	**16.00%**	**16.00%**
净利率	**25.98%**	**36.23%**	**16.19%**	**47.12%**	**49.22%**	**49.68%**

资料来源：Visible Alpha 官网、华泰研究

和彭博一致预期相比，我们对公司 FY2024/FY2025/FY2026 的营业收入的预期差额为-25.13/+11.92/+132.07 亿美元。这一差别主要是由数据中心、智能驾驶、OEM 及其他三项业务导致：1）在 AI 应用的铺开节奏上，我们认为不是一阵高增而后大幅缩减，应是逐渐渗透的细水长流，因此我们预计在 FY2026 AI 的商业应用落地开始进入平稳发展阶段，加上台积电和其他封装产能扩产渐趋稳定，预计英伟达 FY2026 年数据中心营收将达 916 亿美元，同比增速达 40%；2）考虑到高阶自动驾驶政策利好下渗透率的逐渐提升，我们预测智能驾驶业务将持续高增；3）由于 2022 年矿难已出现拐点，且消费电子端 PC 出货量修复态势显著，我们认为公司 OEM 及其他业务将逐渐恢复正常，保持 30%以上的年增长率。费率及利润端，我们在研发费用率，销售、行政及一般费用率方面的预测与市场一致预期基本相符，但对毛利率和净利率的预测则高于市场，我们认为随着高性能、高毛利及高单价的 H100 等数据中心产品占比提升，公司的毛利率和净利率也将进一步提高。

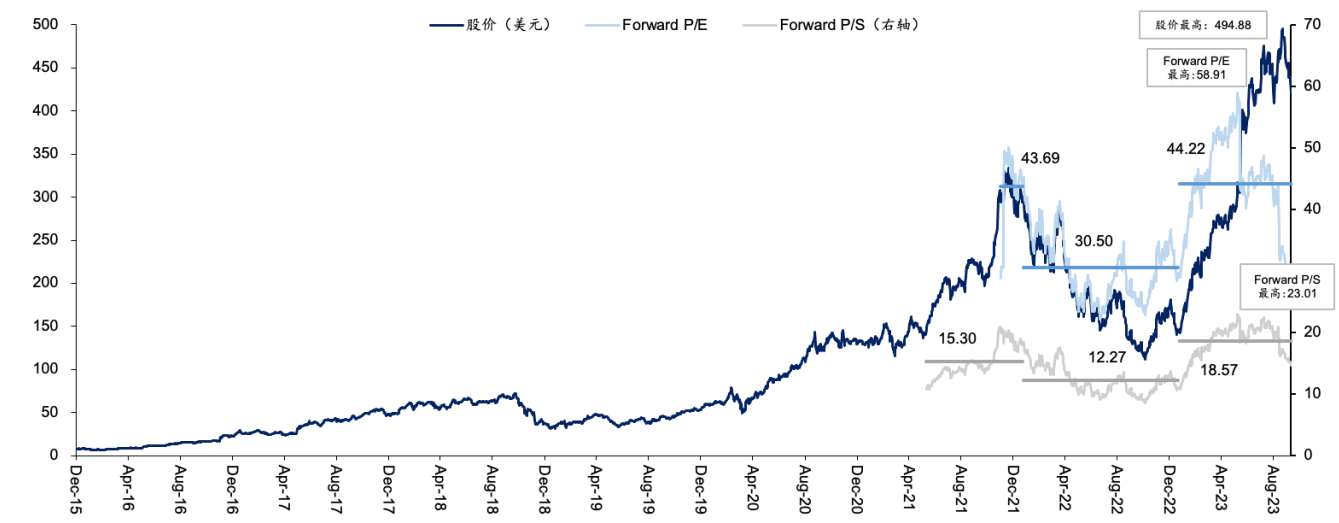
图表246：英伟达关键财务指标彭博一致预期 VS 华泰研究预测

	FY2024E			FY2025E			FY2026E		
	彭博一致预期	华泰预测	差额	彭博一致预期	华泰预测	差额	彭博一致预期	华泰预测	差额
公司总营收	$54,238	$51,725	-$2,513	$81,631	$82,823	$1,192	$101,000	$114,207	$13,207
YoY	101.07%	91.76%	-9.31%	48.91%	60.12%	11.21%	18.71%	37.89%	19.18%
数据中心	$41,223	$37,813	-$3,410	$64,054	$65,416	$1,362	$78,184	$91,582	$13,398
YoY	174.73%	152.00%	-22.73%	55.39%	73.00%	0.18	22.06%	40.00%	17.94%
游戏业务	$10,129	$10,880	$751	$14,065	$13,601	-$464	$18,419	$17,681	-$738
YoY	11.72%	20.00%	8.28%	38.85%	25.00%	-13.85%	30.96%	30.00%	-0.96%
专业可视化	$1,503	$1,359	-$144	$1,815	$1,630	-$185	$2,028	$1,875	-$153
YoY	-2.63%	-12.00%	-9.37%	20.75%	20.00%	-0.75%	11.71%	15.00%	3.29%
智能驾驶	$1,090	$1,264	$174	$1,386	$1,643	$257	$2,022	$2,350	$328
YoY	20.76%	40.00%	19.24%	27.13%	30.00%	2.87%	45.83%	43.00%	-2.83%
OEM及其他	$293	$410	$117	$310	$532	$222	$348	$719	$371
YoY	-35.51%	-10.00%	25.51%	5.75%	30.00%	24.25%	12.11%	35.00%	22.89%
毛利率	71.02%	70.00%	-1.02%	72.00%	71.00%	-1.00%	71.60%	73.00%	1.40%
销售、行政及一般费用率	5.24%	6.00%	0.76%	4.52%	5.50%	0.98%	4.53%	7.00%	2.47%
研发费用率	16.68%	16.00%	-0.68%	14.63%	16.00%	1.37%	15.63%	16.00%	0.37%
净利率	44.09%	47.12%	3.03%	46.62%	49.22%	2.60%	46.69%	49.68%	2.99%

资料来源：Bloomberg、华泰研究

综上讨论，我们预测英伟达 FY2024/FY2025/FY2026 的总营收同比增速为 92%/60%/38%，对应营收为 517/828/1142 亿美元，EPS 为 9.87/16.50/23.24 美元。伴随着 AI 需求喷发，英伟达作为全球数据中心 GPU 的龙头厂商，将凭借其高算力的芯片及高粘性的软硬一体生态布局将充分受益。英伟达 2021 年至今的历史 PS 均值为 14.98X，考虑到公司作为行业龙头具有显著的软硬壁垒和高成长性，以及利润率的数据中心业务占比不断上升，我们认为应享受估值溢价，我们预计 FY24-25 年动态营收为 801 亿美元，并给予 20 倍 PS，目标价 650 美元，首次覆盖给予“买入”评级。

图表247：2016 年至今英伟达历史股价、Forward PE 和 Forward PS（数据截至 2023 年 9 月 20 日）



资料来源：Bloomberg、华泰研究

风险提示

技术落地缓慢：公司的生产技术推进和产品落地可能达不到预期，或影响营收及利润。

中美政治摩擦：中国是半导体产业的重要市场之一，如果中美局势再次升级，将对宏观因素和板块产品销售产生影响。

芯片需求不及预期：市场的芯片需求规模可能不及预期，影响行业营收及利润。

宏观经济不确定性：宏观经济的下行压力和不确定性可能影响公司主营业务及合作进展。

盈利预测

利润表

会计年度 (美元百万)	2022	2023	2024E	2025E	2026E
营业收入	26,914	26,974	51,725	82,823	114,207
销售成本	9,439	11,618	15,518	24,019	30,836
毛利润	17,475	15,356	36,208	58,804	83,371
销售及分销成本	(2,166)	(2,440)	(3,104)	(4,555)	(7,994)
管理费用	(5,268)	(7,339)	(8,276)	(13,252)	(18,273)
其他收入/支出	0.00	(1,353)	0.00	0.00	0.00
财务成本净额	(100.00)	(43.00)	19.40	555.30	751.48
应占联营公司利润及亏损	0.00	0.00	0.00	0.00	0.00
税前利润	9,941	4,181	24,848	41,552	57,855
税费开支	(189.00)	187.00	(472.41)	(790.00)	(1,113)
少数股东损益	0.00	0.00	0.00	0.00	0.00
净利润	9,752	4,368	24,375	40,762	56,742
折旧和摊销	(1,174)	(1,544)	(1,377)	(1,630)	(2,206)
EBITDA	13,865	10,584	26,372	42,374	58,734
EPS (美元，基本)	3.91	1.76	9.87	16.50	22.97

资产负债表

会计年度 (美元百万)	2022	2023	2024E	2025E	2026E
存货	2,605	5,159	5,587	9,656	11,749
应收账款和票据	4,650	3,827	8,138	12,390	17,526
现金及现金等价物	1,990	3,389	22,098	51,567	97,942
其他流动资产	19,584	10,698	10,698	10,698	10,698
总流动资产	28,829	23,073	46,520	84,311	137,916
固定资产	2,778	3,807	5,073	7,874	12,438
无形资产	2,339	1,676	1,158	799.64	552.34
其他长期资产	10,241	12,626	12,626	12,626	12,626
总长期资产	15,358	18,109	18,857	21,299	25,616
总资产	44,187	41,182	65,377	105,611	163,532
应付账款	1,783	1,193	2,262	2,984	4,163
短期借款	0.00	1,250	1,250	0.00	0.00
其他负债	2,552	4,120	4,120	4,120	4,120
总流动负债	4,335	6,563	7,632	7,104	8,283
长期债务	11,687	10,605	9,355	9,355	9,355
其他长期债务	1,553	1,913	1,913	1,913	1,913
总长期负债	13,240	12,518	11,268	11,268	11,268
股本	3.00	2.00	2.00	2.00	2.00
储备/其他项目	26,609	22,099	46,474	87,237	143,979
股东权益	26,612	22,101	46,476	87,239	143,981
少数股东权益	0.00	0.00	0.00	0.00	0.00
总权益	26,612	22,101	46,476	87,239	143,981

估值指标

会计年度 (倍)	2022	2023	2024E	2025E	2026E
PE	108.07	240.50	42.80	25.59	18.39
PB	39.60	47.53	22.45	11.96	7.25
EV EBITDA	76.24	99.94	39.35	23.77	16.36
股息率 (%)	0.00	0.00	0.00	0.00	0.00
自由现金流收益率 (%)	0.64	0.14	1.96	2.97	4.48

现金流量表

会计年度 (美元百万)	2022	2023	2024E	2025E	2026E
EBITDA	13,865	10,584	26,372	42,374	58,734
融资成本	100.00	43.00	(19.40)	(555.30)	(751.48)
营运资本变动	(3,555)	(2,459)	(3,669)	(7,601)	(6,050)
税费	(189.00)	187.00	(472.41)	(790.00)	(1,113)
其他	(1,113)	(2,714)	(128.62)	1,364	2,078
经营活动现金流	9,108	5,641	22,083	34,792	52,898
CAPEX	(976.00)	(1,833)	(2,124)	(4,073)	(6,522)
其他投资活动	(8,854)	9,208	0.00	0.00	0.00
投资活动现金流	(9,830)	7,375	(2,124)	(4,073)	(6,522)
债务增加量	3,977	0.00	(1,250)	(1,250)	0.00
权益增加量	(1,706)	(11,217)	0.00	0.00	0.00
派发股息	(399.00)	(398.00)	0.00	0.00	0.00
其他融资活动现金流	0.00	0.00	0.00	0.00	0.00
融资活动现金流	1,865	(11,617)	(1,250)	(1,250)	0.00
现金变动	1,143	1,399	18,709	29,469	46,376
年初现金	847.00	1,990	3,389	22,098	51,567
汇率波动影响	0.00	0.00	0.00	0.00	0.00
年末现金	1,990	3,389	22,098	51,567	97,942

业绩指标

会计年度 (倍)	2022	2023	2024E	2025E	2026E
增长率 (%)					
营业收入	61.00	0.22	91.76	60.12	37.89
毛利润	68.00	(12.00)	135.79	62.41	41.78
营业利润	122.00	(58.00)	487.79	65.12	39.29
净利润	125.12	(55.21)	458.04	67.23	39.20
EPS	122.59	(55.07)	461.88	67.23	39.20
盈利能力比率 (%)					
毛利润率	64.93	56.93	70.00	71.00	73.00
EBITDA	51.52	39.24	50.98	51.16	51.43
净利润率	36.23	16.19	47.12	49.22	49.68
ROE	44.83	17.93	71.09	60.97	49.08
ROA	26.73	10.23	45.75	47.68	42.17
偿债能力 (倍)					
净负债比率 (%)	0.00	0.00	0.00	0.00	0.00
流动比率	6.65	3.52	6.10	11.87	16.65
速动比率	6.05	2.73	5.36	10.51	15.23
营运能力 (天)					
总资产周转率 (次)	0.74	0.63	0.97	0.97	0.85
应收账款周转天数	47.34	56.57	41.64	44.61	47.15
应付账款周转天数	(56.90)	(46.11)	(40.08)	(39.32)	(41.72)
存货周转天数	(84.50)	(120.29)	(124.65)	(114.23)	(124.95)
现金转换周期	19.75	(17.61)	(42.93)	(30.30)	(36.08)
每股指标 (美元)					
EPS	3.91	1.76	9.87	16.50	22.97
每股净资产	10.67	8.89	18.82	35.32	58.29

资料来源：公司公告、华泰研究预测

华泰证券 HUATAI SECURITIES

AI 带来重估之钥，AMD 能否分庭抗礼？

华泰研究 首次覆盖

2023 年 9 月 22 日 | 美国 电子

AI 新赛道为重估之钥，首发给予买入，目标价 150 美元，24 年 8.5x PS

2016 年开始 AMD 突围英特尔，在抢占份额中估值不断攀升，从 3x PS 到 21 年 5-6x。如今突围二战一触即发，面对 AI 新机遇 AMD 有望再次来到重估分水岭。我们预测 AMD 23/24/25 年营收为 241.9/285.1/318.9 亿美元，同比 2.5%/17.9%/11.8%。对比竞争对手英特尔和英伟达 2024 年的 PS 分别为 2.6 倍和 14.6 倍，AMD 在 CPU 制程上仍领先英特尔，但在 GPU 却奋力追赶英伟达，因此我们认为估值应在两家之间且略低于两家平均值，对比自身历史估值（自 2020 年开始 PS 一直处于 5-6 倍）则上修可期。我们给予 AMD 24 年 8.5x PS，对应目标价 150 美元，首次覆盖给予买入评级。

AMD 以数据中心为矛，游戏和嵌入式为盾，客户端逐渐恢复，毛利率提升

AMD 四大业务：1）数据中心：AI 之风继续吹，CPU 制程仍领先英特尔，MI300 系列有力冲击英伟达，看好 CPU 和 GPU 均能提升份额，预计 23/24/25 年营收 76.0/104.1/120.9 亿美元，同比 26%/37%/16%；2）游戏：疫情间高基数开始消化、主机“半代升级”和高期待新游戏带动维持增长，预计 23/24/25 年营收 66.0/68.3/70.7 亿美元，同比-3%/3.5%/3.5%；3）客户端：随 PC 市场下滑收窄而回暖，预计 23/24/25 年营收 43.0/47.3/52.0 亿美元，同比-31%/10%/10%；4）嵌入式：赛灵思并表效应消退后进入平稳阶段，预计 23/24/25 年营收 56.9/65.4/75.3 亿美元，同比 25%/15%/15%。

突围二战能否凯旋？且看软件生态会否成为 AMD 的阿克琉斯之踵

AMD 曾凭台积电的领先制程颠覆了一家独大的英特尔，至 22 年 2 月 AMD 市值更超越英特尔，如今在 AI 领域面对英伟达的突围战似曾相识。AI 已成为 AMD 的战略首位，MI300 系列蓄势待发切入 AI 赛道。我们认为此产品应是 TPU 外最能与英伟达匹敌的 AI 芯片。不过，对比 CUDA，ROCm 生态起步较晚，虽能完全兼容 CUDA，但亦只属权宜之计，开发者数量相差也较大。为破解此困境，AMD 除持续兼容 CUDA 外，也积极完善生态圈，并与云厂商直接合作，分而治之与 CUDA 脱钩。我们也认为云厂商热切期待二供的出现，MI300 恰逢其时为市场提供了英伟达以外的选择。

CPU 依然是 AI 推理主力，面对众多低功耗产品冒起，AMD 能否站稳阵脚？

AI 推理端市场规模大，对算力要求比训练端较低，各类芯片，包括 GPU、CPU、ASIC 等均能占一席位，主要取决于 AI 负载运行的性价比。AMD 的秘密武器是“每瓦性能”，而英特尔当年份额被撬动的主因是其更重视“每核性能”，如今痛定思痛，准备于明年推出 Intel 3 制程的低功耗服务器 CPU Sierra Forest，试图扭转局势。但不能忽视的是，一直深耕移动端、标榜低功耗的 ARM 对于服务器市场也来势汹汹。另外，英特尔的“四年五节点”狂奔会否反超制程的领导地位，我们认为将是 AMD 的一大挑战。

风险提示：新产品落地进度推迟、PC 恢复和 AI 技术落地不及预期等。

经营预测指标与估值

会计年度	2021	2022	2023E	2024E	2025E
营业收入 (美元百万)	16,434	23,601	24,188	28,510	31,887
+/-%	68.33	43.61	2.49	17.87	11.84
归属母公司净利润 (美元百万)	3,162	1,320	1,483	3,579	5,485
+/-%	26.99	(58.25)	12.39	141.23	53.27
EPS (美元，最新摊薄)	2.57	0.82	0.92	2.21	3.39
ROE (%)	47.43	4.24	2.67	6.17	8.77
PE (倍)	41.98	133.97	119.20	49.41	32.24
PB (倍)	17.71	3.23	3.14	2.96	2.71
EV EBITDA (倍)	43.72	33.31	39.36	23.82	16.74

资料来源：公司公告、华泰研究预测

投资评级(首评)：	买入
目标价(美元)：	150.00

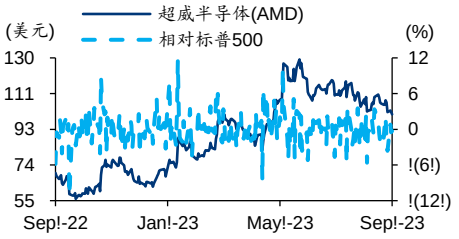
研究员 何翩翩
SAC No. S0570523020002 purdyho@htsc.com
SFC No. ASI353 +(852) 3658 6000

华泰证券研究所分析师名录

基本数据

目标价 (美元)	150.00
收盘价 (美元 截至 9 月 20 日)	100.34
市值 (美元百万)	162,116
6 个月平均日成交额 (美元百万)	6,870
52 周价格范围 (美元)	54.57-132.83
BVPS (美元)	33.96

股价走势图



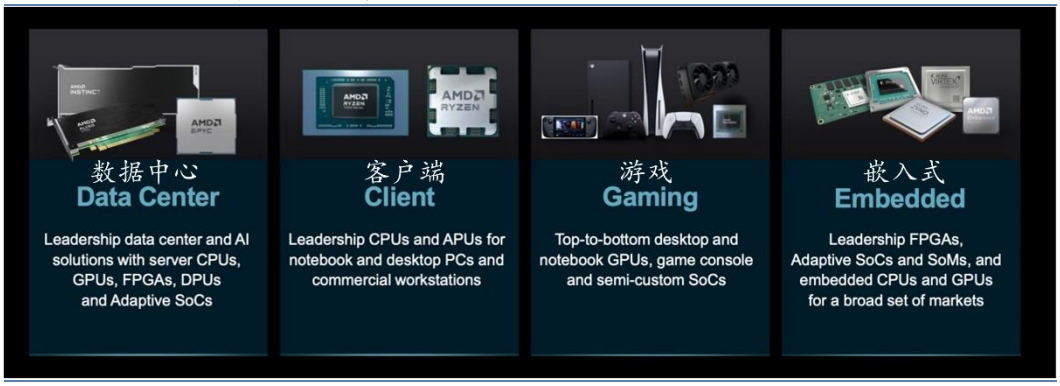
资料来源：S&P

AMD 四大业务板块全景图与增长动力来源

区别于市场的观点：市场对公司在 AI 领域的突围潜力存在认知差。AMD 曾凭台积电的领先制程颠覆了一家独大的英特尔，至 22 年 2 月 AMD 市值更超越英特尔，如今在 AI 领域面对英伟达的突围战似曾相识，面对 AI 新机遇 AMD 有望再次来到重估分水岭。AI 已成为 AMD 的战略首位，MI300 系列蓄势待发切入 AI 赛道。我们认为此产品应是 TPU 外最能与英伟达匹敌的 AI 芯片。对比 CUDA，ROCm 生态起步较晚，虽能完全兼容 CUDA，但亦只属权宜之计，开发者数量相差也较大。为破解此困境，AMD 除持续兼容 CUDA 外，也积极完善生态圈，并与云厂商直接合作，分而治之与 CUDA 脱钩。我们也认为云厂商热切期待二供的出现，MI300 恰逢其时为市场提供了英伟达以外的选择。

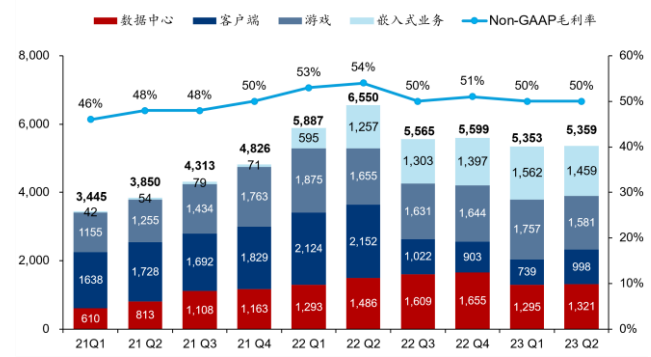
AMD 的四大业务板块清晰地展现了公司全景，我们将分别对数据中心、客户端、游戏和嵌入式业务的表现和未来潜力进行分析，以此作为对 AMD 进行合理估值的基础。AMD 从 22Q2 开始在财报中采用了新的业务划分。在此之前，AMD 的业务划分包括两部分：1）计算与图形（Computing and Graphics）；2）企业、嵌入和半定制（Enterprise, Embedded and Semi-Custom）。而 22Q2 后变为四部分：1）数据中心（Data Center）；2）客户端（Client）；3）游戏（Gaming）；4）嵌入式业务（Embedded）。**我们认为：**1）**数据中心业务**以制程领先和较低能耗的 CPU，以及新产品 MI300 系列的 GPU 携 AI 浪潮前行，或为未来三年核心增长动力源泉；2）**嵌入式业务**在赛灵思并表后进入平稳贡献营收阶段；3）**游戏业务**将受益主机“半代升级”和高期待值游戏维持增长；4）**客户端业务**虽面临疫情后 PC 市场的高基数，但由于最大竞争对手英特尔仍在追赶制程，因此我们认为 AMD 大概率继续占优。总结来说，AMD 以数据中心业务为矛，以游戏业务和嵌入式业务为盾，而客户端业务则开始恢复。

图表248： AMD 四大业务板块：数据中心、客户端、游戏和嵌入式



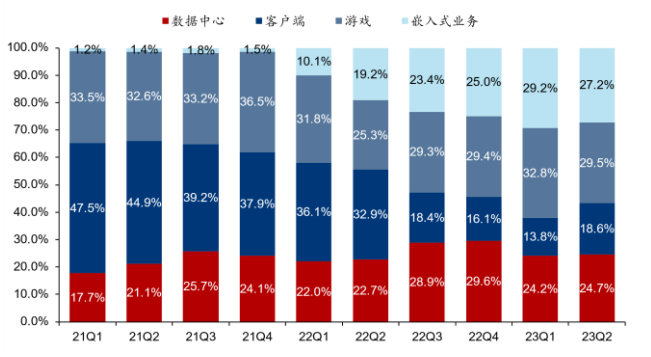
资料来源：AMD 财报、华泰研究

图表249： AMD 2021Q1-2023Q2 分业务同比增速（单位：百万美元）



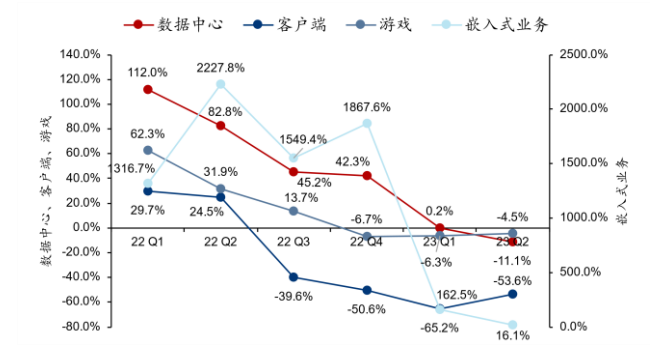
资料来源：AMD 财务公告，华泰研究

图表250： AMD 2020Q1-2023Q2 分业务营收占比



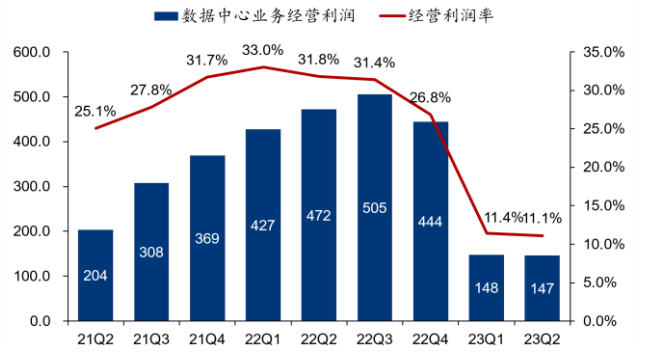
资料来源：AMD 财务公告，华泰研究

图表251： AMD 2021Q1-2023Q2 分业务同比增速



注：嵌入式变化较大，右轴；21Q1/Q3 同比营收数据为 0，增速暂以 100%标记
资料来源：Visible Alpha 官网，AMD 财务公告，华泰研究

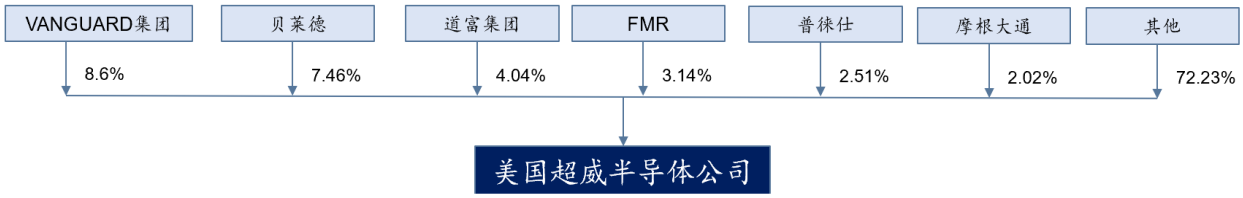
图表252： AMD 2020Q1-2023Q2 数据中心业务营业利润（单位：百万美元）



资料来源：AMD 财务公告，华泰研究

公司股权结构较为分散，机构投资者占比较高。截止 2023 年 9 月，机构投资者持有流通股占比 69.86%，其中 VANGUARD 集团持股 8.6%，为公司第一大股东；贝莱德持股 7.46%，为第二大股东。

图表253： AMD 股权结构情况（根据 2023 中报）



资料来源：Nasdaq，Bloomberg，华泰研究

数据中心业务：EPYC 有望实现长期回报，MI300 携 AI 浪潮前行

AMD 数据中心业务的核心客户包括：1）云服务厂商，如 AWS、谷歌云、百度云、微软 Azure 和甲骨文；2）服务器厂商，如戴尔、HPE（惠与）、联想、美超微、思科等，这些服务器公司将搭载 AMD EPYC CPU 的服务器提供给终端企业。例如 AMD 在 23Q1 财报电话会上提到包括汽车、科技和金融等行业的终端公司，这些公司由于行业特点或较大的规模，需要更高的安全性或更强的可扩展性，因此或会选择使用自有数据中心，而非全部使用云服务；3）有 HPC（High Performance Computing）需求的科研机构，如橡树岭国家实验室（Oak Ridge National Laboratory，MI250X）、劳伦斯利弗莫尔实验室（Lawrence Livermore National Laboratory，MI300）搭载 AMD MI 系列 GPU 搭建的超级计算机等。数据中心业务来自云厂商和终端企业资本支出预算，受经济景气度影响较大，且随终端客户的采购支出节奏而具有一定季节性。

图表254： AMD 数据中心业务营收历史数据与预测值

百万美元（Million USD）	2020（A）	2021（A）	2022（A）	2023（E）	2024（E）	2025（E）
数据中心业务营收	1685	3694	6043	7601	10406	12089
YoY	-	119.2%	63.6%	25.8%	36.9%	16.2%
数据中心业务占总营收比重	17.3%	22.5%	25.6%	31.4%	36.5%	37.9%

注：AMD 从 2022 年 Q2 开始采用新的业务划分，因此历史数据从 2020 年度数据开始
资料来源：AMD 2020-2022 年报，华泰研究预测

AMD 数据中心业务的主要产品包括：1）EPYC（霄龙）CPU；2）Instinct GPU 和异构 AI 芯片如 MI300A（CPU+GPU）；3）FPGA（主要来自 2022 年完成收购的赛灵思）；5）DPU（主要来自 2022 年收购的 Pensando）。我们认为 EPYC CPU 能为 AMD 的数据中心业务带来长期回报，与 MI300 系列产品一同在 AI 浪潮中继续受益。**本小节中我们将首先对 EPYC CPU 业务进行分析，并在本报告的第三章节对 AMD 的 AI 芯片能力进行详细解读。**

AMD EPYC 是 x86 架构的服务器端 CPU。每一代产品代号都为意大利城市名，并且从南往北，如 2017 年第一代 EPYC 7001 代号“Naples（那不勒斯）”，2019 年 EPYC 7002 代号为“Rome（罗马）”，2021 年的 7003 为“Milan（米兰）”。2022 年推出了“Genoa（热那亚）”，到 2023 年 6 月 13 日，AMD 正式推出代号为“Bergamo（贝加莫）”的第四代 EPYC，且已经向其云客户 Meta 等批量出货。

AMD 第四代 EPYC CPU 均采用台积电 5nm 制程，目前在制程上仍领先于英特尔。Bergamo 和 Genoa-X 作为最新的 EPYC 均采用台积电 5nm 制程，对比 Intel 最新的服务器端 CPU Sapphire Rapids 采用的 Intel 7 节点（10nm，相当于台积电 7nm）尚处领先。但根据 Intel 技术路线图，如果英特尔按照“四年五节点”计划顺利推进，AMD 依赖台积电所获得的制程优势或将缩小，在服务器端可能会出现英特尔通过预计在 2024 年下半年就绪的 18A（1.8nm）超过台积电 2025 年的 N2（2nm）的情况。但英特尔究竟能否在 2025 年或以前顺利推进制程计划还需进一步观察，因此 2024 年 Intel 3 的落地情况将是关键一步。

AMD 2023 年 6 月发布的 Bergamo 和 Genoa-X（Genoa 的后代产品，使用 3D V-Cache）分别具备以下优势：1）Bergamo 基于 Zen 4c 架构，内核+L2 区域合计 2.48 平方毫米，比 Zen 4 的 3.84 平方毫米减少了 35%，采用了 8 个 16 核的 CCD，内核数最高可至 128 个，可提高多线程能力，对标的是其他多核数的服务器产品如英特尔下一代 144 核的 Sierra Forest；2）Genoa-X 内核数最高 96 个，采用 1.3GB 的 L3 缓存，并集成了 3D V-Cache 技术，适用于缓存需求较高的技术计算。

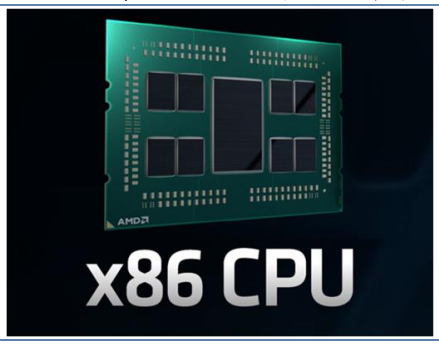
AMD 在数据中心 CPU 的竞争环境渐趋激烈，除英特尔的 x86 架构 CPU 外，ARM 架构的低能耗优势从移动端逐渐向服务器端蔓延。ARM（安谋）已于美东时间 9 月 14 日在纳斯达克上市，苹果、谷歌、英伟达、AMD、英特尔和台积电等科技巨头均是此次发行的基石投资者。亚马逊在 AWS 中大量运用到 ARM 架构的 Graviton CPU（目前最新版本为 Graviton3）；英伟达 Grace Hopper 中的 Grace CPU 也采用了 ARM 架构。**我们认为 ARM 和 x86 CPU 架构在 AI 应用里各有优势。由于 ARM 架构从性能层面上较难与 x86 相比，我们不认为其能有朝一日完全取代 x86 架构。但我们也强调，在云计算任务类型和数据模态多元化的趋势下，低能耗的 ARM 架构和高性能的 x86 架构可以分别负责较轻和较重的工作负载。而随着 AI 和云计算领域对节能的要求提高，ARM 架构本身的能耗优势凸显。因此，我们认为 ARM 架构的 CPU 将逐渐在 AI 推理市场占一席之地。**AMD 在 2023 年 6 月发布 Bergamo 时也称，目前绝大多数 AI 工作负载仍以 CPU 运行。我们对此观点部分认同：由于在面对不同模态数据的推理时，CPU 与 GPU 的分工各有不同，例如在处理语音、语言和文本数据的推理时，或更适合使用擅长串行运算的 CPU 进行，故 CPU 在 AI 推理领域仍扮演重要角色。

x86 和 ARM 架构的最根本区别在于，前者的设计模式为 CISC（复杂指令集，Complex Instruction Set Computer），而后者是 RISC（精简指令集，Reduced Instruction Set Computer）。x86 架构在早期设计时主要针对 PC 端追求高性能，希望通过最少的指令来完成计算任务，但芯片面积较大、晶体管数量较多，因此能耗较高，不过鉴于 PC 的体积较大，可安装散热装置；而 ARM 架构则针对体积较小的移动端设备，能耗较低。ARM 架构相比 x86 能耗较低，在较重视低能耗要求的云里以及能耗不断提升的 AI 应用中或较合适，且在 CPU 和 GPU 搭配使用的架构里，CPU 或仅需发挥其部分功能，如向 GPU 发出指令等，因此 ARM 架构或已足够。而 x86 架构追求高性能，拥有丰富的指令集，在 AI 推理里可跟 GPU 在功能上有所互补。能耗已经成为芯片设计的关注重点，英特尔将于 2024 年上半年量产交货的第一款数据中心 E-Core（Efficiency Core，高效能核）CPU Sierra Forest 同样是为向客户提供低能耗选择设计，目的是与 AMD 产品的低能耗特点相比，以抢回失去的 CPU 市场份额。

图表255： 英伟达 Grace Hopper 中的 Grace CPU 采用 ARM 架构

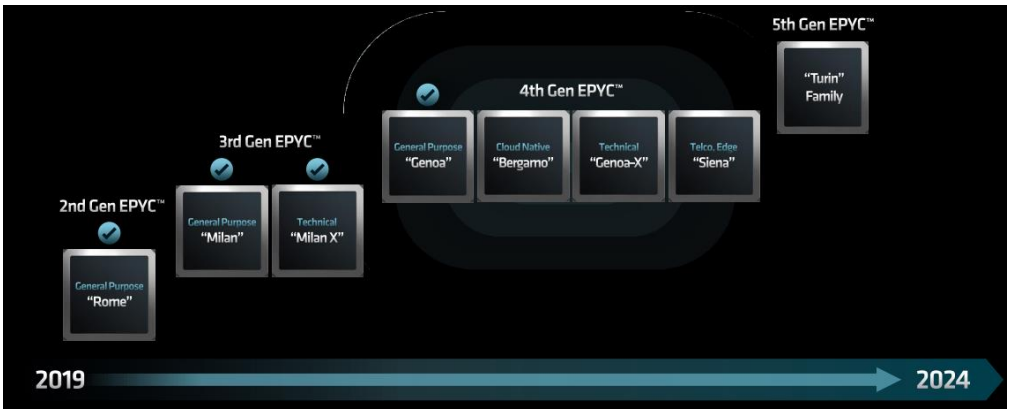
资料来源：AnandTech 官网、华泰研究

图表256： AMD MI300 中的 Zen 4 CPU 采用 x86 架构



资料来源：Wccftech 官网、华泰研究

图表257： AMD EPYC 系列 CPU 产品路线图



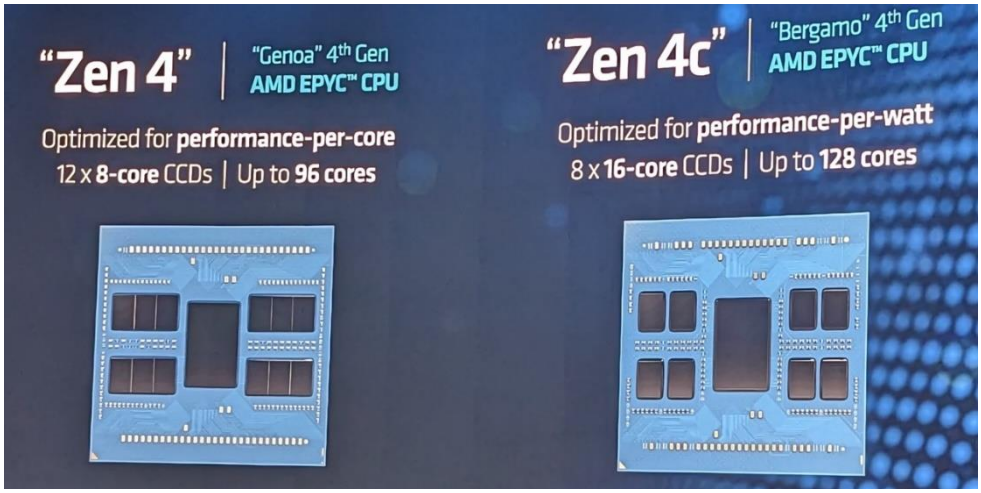
资料来源：CES 2023、华泰研究

图表258： AMD EPYC CPU 产品矩阵

系列名称	Venice	Turin	Siena	Bergamo	Genoa-X	Genoa	Milan	Rome	Naples
系列编号	EPYC 11K*	EPYC 10K*	EPYC 8004（Sienna）	EPYC 9004（Bergamo）	EPYC 9004（Genoa-X）	EPYC 9004（Genoa）	EPYC 7003（Milan）	EPYC 7002（Rome）	EPYC 7001（Naples）
发布年份	2025+	2024	2023	2023	2023	2022	2021	2019	2017
CPU 架构	Zen 6*	Zen 5	Zen 4	Zen 4C	Zen 4 V-Cache	Zen 4	Zen 3	Zen 2	Zen 1
制程	NA	3/4nm TSMC	5nm TSMC	4nm TSMC	5nm TSMC	5nm TSMC	7nm TSMC	7nm TSMC	14nm GloFo
插座	NA	SP5 / SP6	SP6	SP5	SP5	SP5	SP3	SP3	SP3
插槽	NA	LGA 6096 (SP5)	LGA 4844	LGA 6096	LGA 6096	LGA 6096	LGA 4094	LGA 4094	LGA 4094
最大核心数	384*	128*	64	128	96	96	64	64	32
最大线程数	768*	256*	128	256	192	192	128	128	64
最大三级缓存	NA	NA	NA	NA	1152 MB	384 MB	256 MB	256 MB	64 MB
Chiplet 设计	NA	NA	8 CCD's (1CCX/CCD) + 1 IOD	12 CCD's (1 CCX/CCD) + 1 IOD	3D V-Cache 12 CCD's (1 CCX/CCD) + 1 IOD	12 CCD's (1 CCX/CCD) + 1 IOD	8 CCD's (1 CCX/CCD) + 1 IOD	8 CCD's (2 CCX's/CCD) + 1 IOD	4 CCD's (2 CCX's/ CCD)
内存支持	NA	DDR5-6000*	DDR5-5200	DDR5-5600*	DDR5-4800	DDR5-4800	DDR4-3200	DDR4-3200	DDR4-2666
存储器通道	TBD	12 Channels (SP5) 6-Channels (SP6)	6-Channels	12 Channels	12 Channels	12 Channels	8 Channels	8 Channels	8 Channels
PCIe Gen 支持	TBD	TBD	96 Gen 5	160 Gen 5	128 Gen 5	128 Gen 5	128 Gen 4	128 Gen 4	64 Gen 3
最大热设计功耗	TBD	480W (cTDP 600W)	70-225W	320W (cTDP 400W)	400W	400W	280W	280W	200W

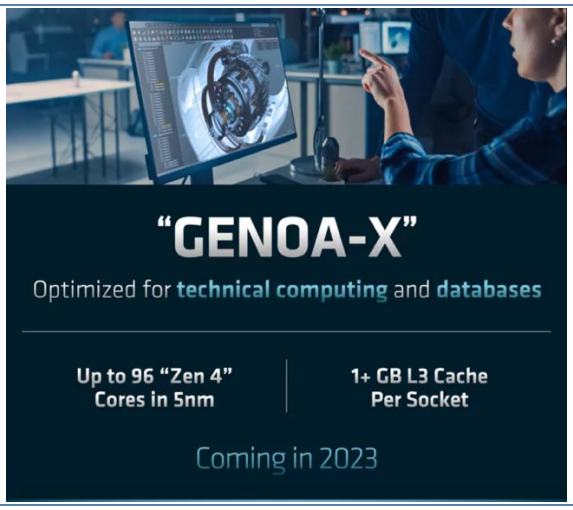
*注：具体参数官方尚未确认
资料来源：Wccftech、华泰研究

图表259： AMD 6 月 13 日发布 EPYC 处理器系列新产品 Bergamo 属 Zen 4c 代际（图右）



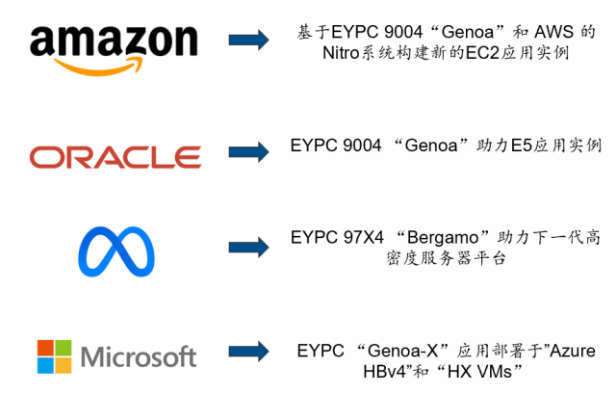
资料来源：AMD 2023 年 6 月 AI 与数据中心发布会、华泰研究

图表260：EPYC 处理器系列新产品 Genoa-X 预计在 2023 年内上市



资料来源：AMD 官网、华泰研究

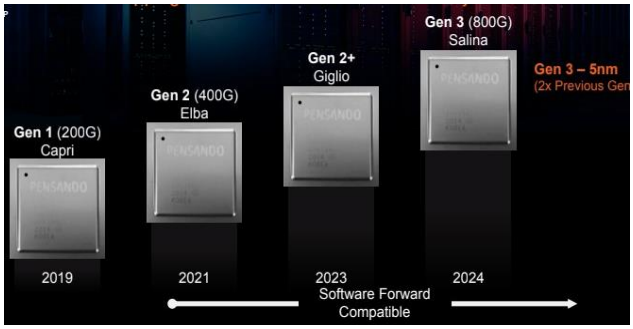
图表261：AMD 在 CES2023 披露四家云厂商为第四代 EPYC 客户



资料来源：CES 2023、华泰研究

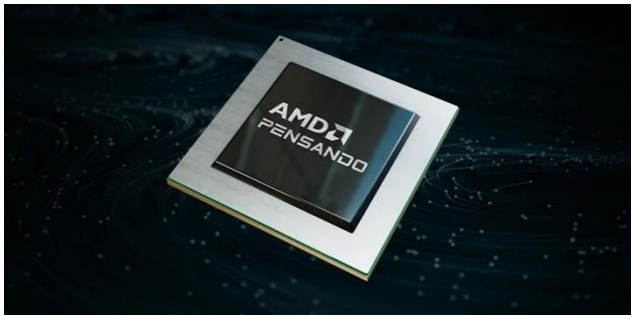
除 EPYC CPU 和 MI 系列之外，在 AMD 的数据中心业务中还包括来自 AMD 收购的 Pensando 的 DPU 产品，与其他数据产品产生交叉销售效应（cross-selling）。AMD 于 2022 年 4 月宣布了对 Pensando（网络芯片公司，生产 DPU，Data Processing Unit 数据处理单元）的收购，并于 2022 年 5 月 16 日完成了收购。完成收购当日，AMD 称通过收购 Pensando，有效拓展了 AMD 的数据中心解决方案：鉴于 DPU 的主要任务是为数据中心部署的 CPU 减轻负载，承担包括数据下载、加密和解密等，会占用 CPU 处理能力的其他工作。因此，DPU 与 AMD 的其他数据中心产品就可以成为互补，也实现交叉销售。据 Allies Market Research，DPU 市场规模将在 2031 年达 55 亿美元，从 2022 年到 2031 年，CAGR 达 26.9%。

图表262：2022 年 8 月 AMD Pensando DPU 产品路线图



资料来源：AMD 官网、华泰研究

图表263：AMD Pensando 产品示意图



资料来源：AMD 官网、华泰研究

游戏业务：PS5、Xbox 游戏主机"半代升级"已至，显卡落后英伟达

AMD 的游戏业务可以分为两部分：游戏显卡和游戏主机。**AMD 的游戏主机半定制芯片业务客户包括：Play Station（索尼）、Xbox（微软），以及最大游戏平台 Steam 的主机 Steam Deck，在目前三家主流游戏主机中（索尼 PS、任天堂 Switch 和微软 Xbox），AMD 三中占二（任天堂 Switch 使用英伟达）。**其中，据 Ampere Analysis 测算，2022 年索尼 PS 在全球游戏主机硬件及软件销售占比为 45%，任天堂 Switch 为 27.7%，微软 Xbox 为 27.3%。23Q2，AMD 游戏业务（游戏主机半定制芯片和 PC 端 GPU 显卡）营收 15.8 亿美元，同比下降 4%，环比下降 10%。在四大业务板块中，游戏业务的营收占比 2020/2021/2022 年分别为 28.1%/34.1%/28.8%，稳定在 30%左右。

图表264： AMD 游戏业务营收历史数据与预测值

百万美元 (Million USD)	2020（A）	2021（A）	2022（A）	2023（E）	2024（E）	2025（E）
游戏业务营收	2746	5607	6805	6601	6835	7074
YoY	-	104.2%	21.4%	-3.0%	3.5%	3.5%
游戏业务占总营收比重	28.1%	34.1%	28.8%	27.3%	24.0%	22.2%

注：AMD 从 2022 年 Q2 开始采用新的业务划分，因此历史数据从 2020 年度数据开始
资料来源：AMD 2020-2022 年报，华泰研究预测

我们认为，由于两款主要游戏主机“半代升级”时点已至，AMD 游戏主机业务动力不缺。AMD 游戏主机业务半定制芯片的销量支撑之一是主机版本的迭代，而主机的迭代升级又可以分为“整代升级”和“半代升级”两种，其中“整代升级”如从 PS4 到 PS5、从 PS5 到 PS6，或从 Xbox One 到 Xbox Series S/X，而“半代升级”则是指推出 PS5 Pro 版，或推出新版本的 Xbox Series S/X。梳理 AMD 和 Xbox 的结缘历史，Xbox 使用 AMD 的半定制芯片是从 2013 年首次上市的第三代 Xbox 开始的，在此之前，2001 年的第一代 Xbox 搭载的是英特尔的芯片，2005 年的第二代 Xbox 则使用的是 IBM 的芯片。第三代 Xbox，同时也是最新系列，为 Xbox Series S 和 Xbox Series X，上市时间为 2020 年 11 月 10 日，与 PS5（2020 年 11 月 12 日）几乎相同。如今 2023 年下半年，距离上一次主机版本更新已经接近三年。

虽然整代升级尚且遥远，但半代升级近在眼前。据索尼于 2023 年 3 月披露的文件，PS6 的推出时间很可能在 2026-27 年，但据 Insider Gaming 在 2023 年 5 月的预测，PS5 的半代升级版本 PS5 Pro 级将于 2024 年底上市。Xbox 的整代升级按照上一次从 2013 年的第三代 Xbox One 到 2020 年的第四代 Xbox Series S/X，也需 7 年左右，但是据微软 6 月 12 日发布的最新预告，新版 Xbox Series S Carbon Black 将在今年 9 月 1 日推出。这两款半代升级的主机将是 AMD 游戏业务的接下来的增长动力之一。同样，英伟达的主机客户任天堂或将于 2024 年推出新款 Switch，届时三款主机又将正面相遇。

图表265： Play Station 5 游戏主机

资料来源：索尼官网、华泰研究

图表266： Xbox Series S 和 Xbox Series X

资料来源：Xbox 官网、华泰研究

图表267： 半代升级的 Play Station 5 Pro

资料来源：索尼官网、华泰研究

图表268： Xbox Series S Carbon Black

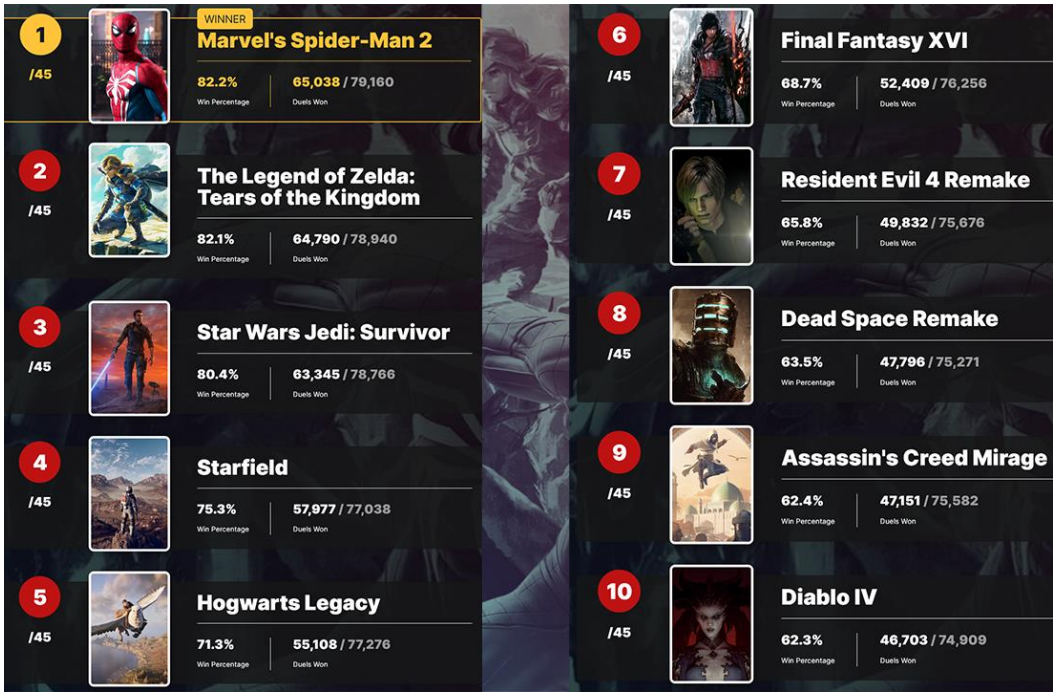
资料来源：Xbox 官网、华泰研究

Play Station 5是索尼在2020年11月12日上市的最新一代PS,截至目前已上市接近三年。**在PS6于2026-27年到来前，PS5和PS5 Pro将在一段时间内具备较强的购买吸引力的原因有二：1）游戏主机迭代周期较长：**PS4上市年份为2013年，到2020年PS5上市有七年之久，因此玩家无需担心PS像其他消费电子一样迭代飞快，让购买者措手不及；2）**PS5购买难度较高，玩家刚刚入手：**PS5存货从上市开始一年多内都非常有限，难以购得，直到2022年末PS5的购买难度才有所下降,玩家直到2023年才可以随时自由购买PS5,因此即便PS5上市已有三年，部分玩家拿到PS5的时间却并不长。并且，半代升级的版本也将为还未购买PS5的玩家提供新的购买动力。

Xbox的半代升级虽提升不多，但踩点配合热门游戏《Starfield》发售时间，或能借力。半代升级的Xbox Series S Carbon Black的特点是1TB的SSD内存，原版的Xbox Series S只有512GB，其他参数维持不变，但是这款半代升级主机的发售优势在于其上市时间为9月1日,与高期待值的游戏Starfield(IGN评分网站排名第四,暂定2023年9月6日发布)发布时间重合。

高期待值且兼容平台有限的游戏作品或能有效拉动游戏主机销量。Xbox Series S Carbon Black内存更新本身或不能为玩家提供足够的购买理由,但9月的发布时间恰好踩点高期待值游戏Starfield，两相结合，或被期待带来新销量。为此，Xbox已做好准备，营销和产品配套方面，还同步推出了配合Starfield的摇杆和耳机。此外，我们对IGN（权威游戏打分网站）评选出最受关注的2023到2025年将发布的游戏所兼容的游戏平台进行了统计，可以看到，**现今大部分游戏都兼容PC平台，但是期待值最高的《蜘蛛侠2》仅能兼容PS5一款游戏主机，同样仅上PS5的还有排名第6的《最终幻想7》。**

图表269：IGN游戏期待值排名：第一名为《蜘蛛侠2》，第六名为《最终幻想7》



资料来源：IGN、华泰研究

图表270： 2023-2025 年高期待值游戏：重点关注在 PS5 上线而不兼容 PC 的《蜘蛛侠 2》和《最终幻想 7》

游戏	中文名	出品公司	PC	PS4	PS5	Xbox X/S	Xbox One	Switch	预计发售时间
Crimson Desert	红色沙漠	珍艾碧丝	✓	✓	✓	✓	✓		2023-2025
Instinction	本能	Hashbane	✓	✓	✓	✓	✓	✓	2025
Judas	犹大	鬼故事游戏	✓		✓	✓			2025
Avatar: Frontiers of Pandora	阿凡达：潘多拉边境	育碧	✓		✓	✓			2023-2025
Marvel's Spider-Man 2	**蜘蛛侠 2**	**索尼**			**✓**				**2023Q4**
Starfield	**星空**	**贝塞斯达**	**✓**			**✓**			**2023（9 月）**
Assassin's creed: Mirage	刺客信条：幻景	育碧	✓	✓	✓	✓	✓		2023
Black Myth Wukong	黑神话：悟空	Game Science	✓		✓	✓			2024Q3
The Lost Wild	迷失荒野	Great Ape Games	✓		✓	✓			2024
Greedfall2	贪婪之秋 2	Nacon	✓	✓	✓	✓	✓		2024
Baldur's Gate 3	博德之门 III	拉瑞安工作室	✓		✓				2024（8 月）
Hollow Knight: Silksong	空洞骑士：丝之歌	Team Cherry	✓	✓	✓	✓	✓	✓	2024
Persona 3 Reload	女神异闻录 3 重制版	Atlus/ SEGA	✓	✓	✓	✓	✓		2024
Like A Dragon: Infinite Wealth	如龙 8	SEGA	✓	✓	✓	✓	✓		2024
Prince Of Persia: The Lost Crown	波斯王子：失落王冠	育碧	✓	✓	✓	✓	✓	✓	2024
Palworld	幻兽帕鲁	Pocket pair	✓			✓			2023
The Wolf Among Us 2	人中之狼 2	诉说游戏	✓	✓	✓	✓	✓		2024
Microsoft Flight Simulator 2024	微软模拟飞行	Xbox/Asobo	✓			✓			2024
Avowed	宣誓	Xbox	✓			✓			2024
Senua'S Saga: Hellblade Ii	地狱之刃 2：赛怒雅传奇	Xbox	✓			✓			2024
Final Fantasy 7 Rebirth	**最终幻想 7：重生**	**史克威尔艾尼克斯**			**✓**				**2024**
Starwars Outlaws	星球大战：亡命之徒	育碧	✓		✓	✓			2024

资料来源：IGN、各工作室官网、华泰研究

图表271： 配合游戏 Starfield 的 Xbox 定制摇杆

资料来源：Xbox 官网、华泰研究

图表272： 配合游戏 Starfield 的 Xbox 定制耳机

资料来源：Xbox 官网、华泰研究

正如 AMD 在 2023 年 Q1 财报电话会上指出，游戏主机业务的“较高收入被较低的显卡业务抵消（offset by lower gaming graphics sales）”。显卡业务难有突破的原因有二：1）PC 市场仍然萎缩，据 Gartner，2023Q2 全球 PC 出货量总计 5970 万台，同比下降 16.6%；2）市占率难有突破，独立显卡英伟达一骑绝尘。

AMD 目前最新代际的显卡为 Radeon 7000 系列，与此同时 Radeon 6000 系列也依然在售。目前，Radeon 7900 XTX 为 AMD 的旗舰款高端显卡，与英伟达 GeForce 4090 对标。**Radeon 7000 系列采用了 RDNA 3 架构，内置 AI 加速单元和光线追踪加速单元，并使用 Chiplets 技术，是全球第一款使用 Chiplet 的游戏显卡。**AMD Radeon 显卡的发售价格显著低于英伟达 GeForce，尤其是旗舰款 7900 XTX 和 4090 对比，价格相差 600 美元，即便在如此巨大的价格差别之下，独立显卡市场依然是英伟达领军。

图表273： AMD Radeon 7900 XTX

资料来源：AMD 官网、华泰研究

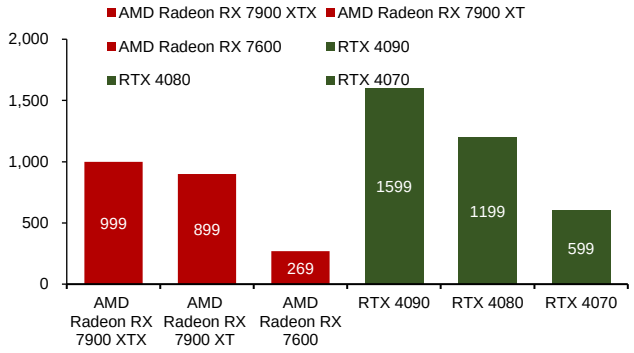
图表274： 英伟达 GeForce 4090

资料来源：英伟达官网、华泰研究

图表275： AMD Radeon 7000 系列使用 Chiplet 技术

资料来源：Xbox 官网、华泰研究

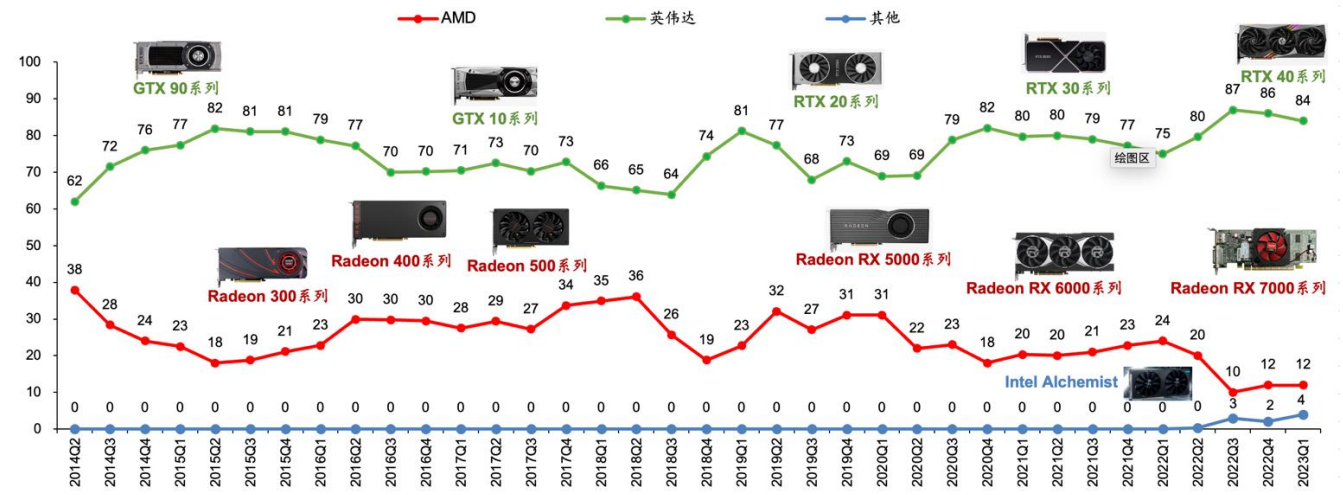
图表276： AMD Radeon 与英伟达 GeForce 发布价格比较（美元）



资料来源：AMD 官网、英伟达官网、华泰研究

AMD 在独立显卡（dGPU，discrete GPU）业务的市占率与英伟达难以相比，且英特尔也渐能分一杯羹，蚕食了 AMD 本就不多的市场份额：据 Jon Peddie Research 统计，2023 年 Q1 英伟达的桌面端独立显卡全球市占率为 84%，AMD 仅为 12%，另外 4%推测为英特尔通过其 2022 年 Q1 开始发售至今的独立显卡产品 Arc Alchemist GPU 所获得的市场份额。

图表277： 全球独显 GPU 市场份额 2014Q2-2023Q1（单位：%）



资料来源：Jon Peddie Research、英伟达官网、华泰研究

嵌入式业务：赛灵思并表一年有余如虎添翼，营收贡献进入稳定阶段

AMD 的嵌入式业务营收从 2022 年 Q1 开始陡增，占 AMD 总营收之比从不到 2%一举提高到超过 10%，2022 年较 2021 年增长 1750%，这是因为 2022 年 Q1 AMD 与赛灵思开始并表。从 2022 年 Q1 到 Q2 的绝对值高增长可以看出，Q1 时并表并没有完成，到 Q2 开始大致完成，此后一直到 2023 年 Q1，嵌入式业务的绝对值变化都不大。2023 年 Q1 是 AMD 与赛灵思并表后的第一个可比季度，这也是嵌入式业务 23Q1 的营收从去年的 5.95 亿美元激增至 15.6 亿美元且同比大幅增长 163%的原因，由此我们也认为 2023 年下半年嵌入式业务将进入同比稳定阶段。

此外，观察被 AMD 收购前赛灵思的营收情况，其 FY2020/FY2021 的年同比变化都很小，分别为 3.39%和-0.48%，而再往前 FY2019（日历时间为 2018 年 4 月到 2019 年 3 月）的 24%增幅则是因为华为在 2018 年从赛灵思采购 FPGA 网络芯片用于 5G 建设，但 2018 年末开始华为给赛灵思带来的增长动力消退。另外，我们也在此指出，全球 FPGA 市场竞争格局清晰且稳定，赛灵思、Altera（英特尔于 2015 年收购）分别为龙头和第二（据 Frost&Sullivan 2019 年数据赛灵思当时市占率超过 50%，Altera 则接近 40%），另外有 Lattice 和 Microchip。

图表278： AMD 嵌入式业务营收历史数据与预测值

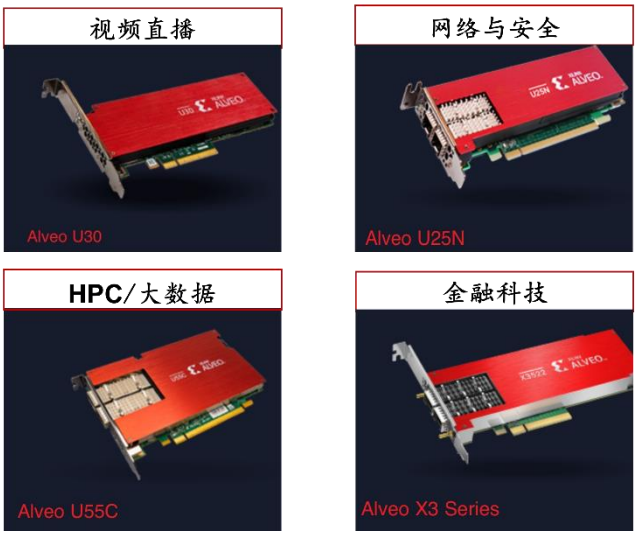
百万美元（Million USD）	2020（A）	2021（A）	2022（A）	2023（E）	2024（E）	2025（E）
嵌入式业务营收	143	246	4552	5690	6544	7525
YoY	-	72.0%	1750.4%	25.0%	15.0%	15.0%
嵌入式业务占总营收比重	1.5%	1.5%	19.3%	23.5%	23.0%	23.6%

注：AMD 从 2022 年 Q2 开始采用新的业务划分，因此历史数据从 2020 年度数据开始
资料来源：AMD 2020-2022 年报，华泰研究预测

2020 年 10 月 27 日，AMD 宣布以 350 亿美元收购 FPGA 龙头赛灵思，目的是通过赛灵思的 FPGAs、Zync RF SoC（通信）、Kria（可用于自动驾驶中的智能摄像头，搭载车规级 Zynq SoC）、Versal 自适应 SoC（可用于 AI）、以及 Alveo 加速卡（用于数据中心，搭载 FPGA）等产品扩充产品线，这样一来，AMD 从产品线上与早在 2015 年就收购了 Altera 的英特尔看齐。赛灵思产品与 AMD 产生协同效应，在赛灵思并表带来的无机营收增长之外，丰富产品组合，实现有机增长，如虎添翼。这与上文所述 Pensando DPU 情况有相似之处，AMD 与赛灵思的产品同样可以产生交叉销售。

嵌入式业务的客户包括工业、视觉、医疗保健、通信、航空航天、国防、测试和仿真以及汽车行业内客户，例如数据中心 FPGA Alveo 为不同使用场景打造不同产品，目前有约 10 种以上型号，为不同的使用场景设计，例如 Alveo U30 的使用场景为视频直播，另外有高算力的型号 U55C 为 HPC 和大数据场景设计，X3 则主要辅助金融交易和风险管理，可以有效丰富 AMD 为不同行业客户提供的方案，产生交叉销售。

图表279： 赛灵思数据中心产品 Alveo 部分型号与使用场景

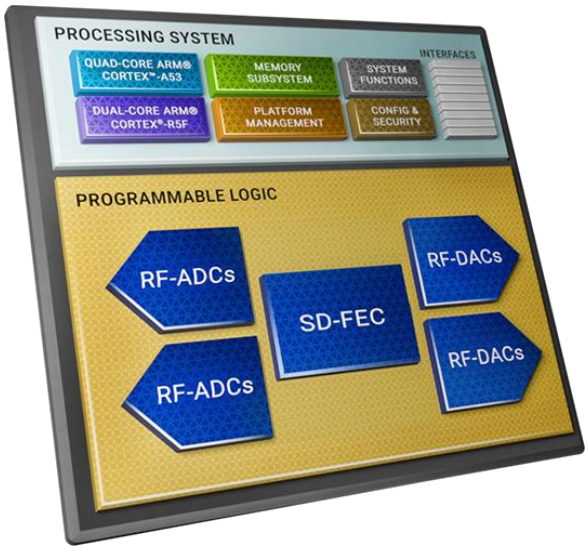


资料来源：AMD 官网、华泰研究

图表280： 赛灵思 Versal Adaptive SoC

资料来源：AMD 官网、华泰研究

图表281： 赛灵思 Zynq RF SoC



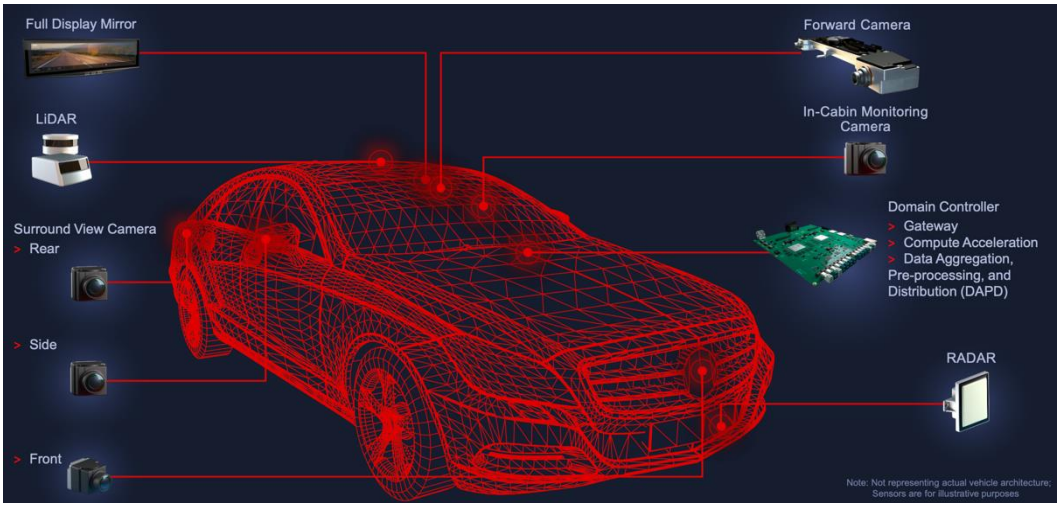
资料来源：AMD 官网、华泰研究

图表282： 赛灵思车规级 Zynq 产品示意图

资料来源：AMD 官网、华泰研究

赛灵思车规级 FPGA 专攻感知环节如雷达、摄像头等，客户覆盖 Tier1 供应商、OEMs 整车厂和自动驾驶新兴企业。赛灵思的汽车产业链上的客户包括：1）Tier1 供应商如 Continental、Magna、Aisin、HITACHI、Veoneer；2）OEMs 如比亚迪、斯巴鲁（EyeSight 系列，最近在 2023 年 4 月推出的 EyeSight 4 平台搭载了赛灵思的车规级 Zynq）、戴姆勒等；3）其他自动驾驶企业，如小马智行、速腾聚创、Minieye、Ouster 等。我们重申自动驾驶是 AI 的多种应用中更全面、发展更快速、落地商业化能力更强的应用场景，自谷歌 Waymo 将 AIGC 应用于自动驾驶领域开始，Transformer 应用愈发广泛，在其感知+决策环节中，感知的信息来自激光雷达、雷达和摄像头等，这正是赛灵思车规级产品的切入点，例如 Zynq 对图形和视频处理等流程有专用处理模块，可以实现数据处理的更低延迟，用于自动驾驶必备的前置摄像头、激光雷达、4D 成像雷达等各类传感器。

图表283：赛灵思 Zynq 自动驾驶产品应用场景包括雷达、激光雷达、摄像头等



资料来源：AMD 官网、华泰研究

客户端业务：清库存或还需 1-2 季度，AMD 市占率持续爬升

AMD 的客户端业务产品结构包括 PC 端 CPU 和集成了 CPU 和显卡的 APU（消费级 CPU+消费级 GPU）。我们认为，由于 PC 市场在疫情间录得较大增速，因此基数较高，加上随着疫情后用户的消费电子使用习惯改变（使用个人电脑的场景减少）而不断萎缩，对 AMD 的客户端业务来说，在四业务中的占比缩小不可避免，往后需期待疫情后的需求复苏，以及要看能否维持相对英特尔的市占率，并继续爬升。

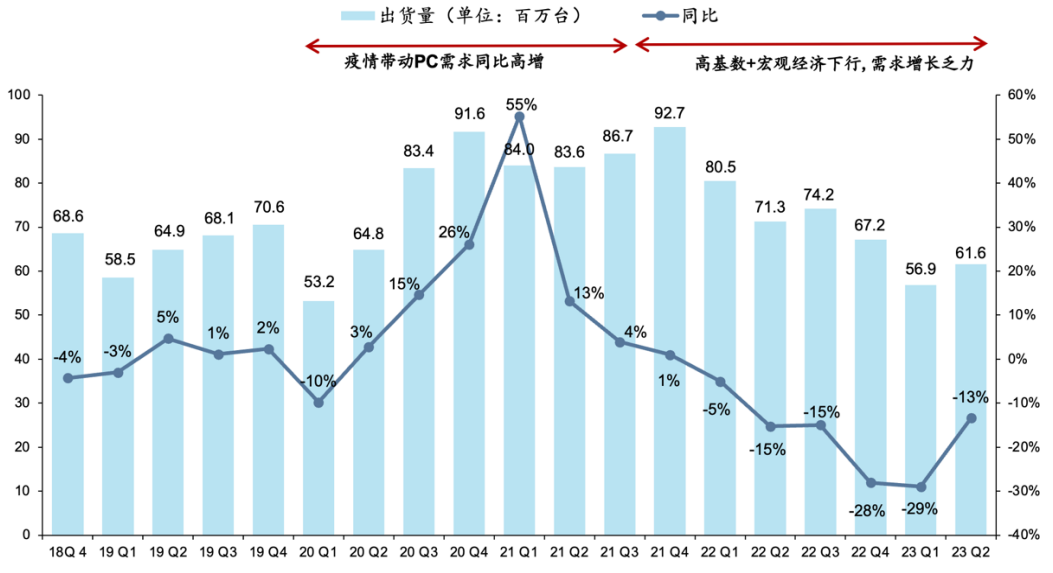
我们认为 PC 厂商的库存或还需要消化 1-2 个季度，或将在下半年缓慢回暖，与 AMD 和英特尔在 2023 年中报会议上预测 PC 业务将于 2023 年下半年开始回弹相符；另外，AI 软件应用广泛也将带动 PC 端芯片需求。AMD 客户端业务的下游客户 PC 厂商先前积累了较多库存，因此在需求走弱后先消化额外库存。23Q2 AMD 客户端业务营收 9.98 亿美元，超过彭博一致预期的 8.4 亿美元，同比下降 54%，行业疲软，但环比大幅上升 35%。公司预期在下半年传统旺季里 PC 业务将凭着 Ryzen 7000 系列 CPU 的增量以及 PC 厂商库存消化见底而回暖。此外，我们也认为，随着 PC 端 AI 软件应用（如微软 Copilot 等）更加广泛，或将带来 PC 端芯片要求上升，AMD 大概率将受惠。

图表284：AMD 客户端业务营收历史数据与预测值

百万美元（Million USD）	2020（A）	2021（A）	2022（A）	2023（E）	2024（E）	2025（E）
客户端业务营收	5189	6887	6201	4296	4726	5199
YoY	-	32.7%	-10.0%	-30.7%	10.0%	10.0%
客户端业务占总营收比重	53.1%	41.9%	26.3%	17.8%	16.6%	16.3%

注：AMD 从 2022 年 Q2 开始采用新的业务划分，因此历史数据从 2020 年度数据开始
资料来源：AMD 2020-2022 年报，华泰研究预测

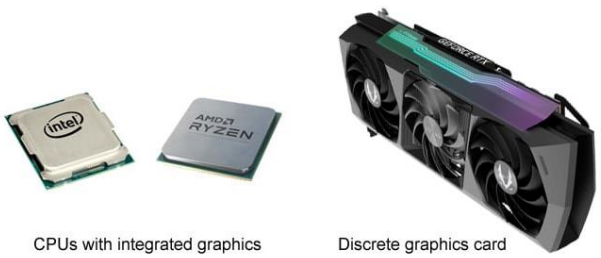
图表285：全球 PC 出货量及同比和环比变化（单位：百万台）



资料来源：IDC、华泰研究

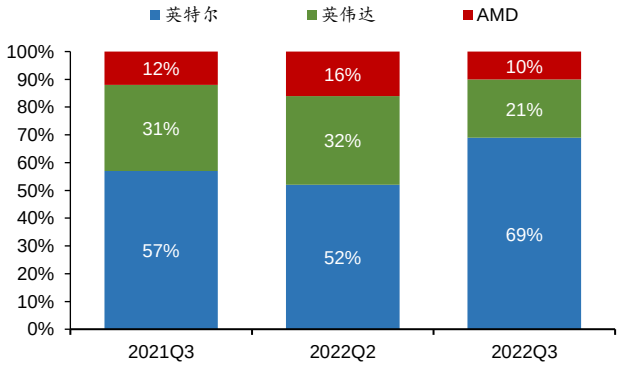
PC 端集成显卡方面，英特尔依然领先：据 Jon Peddie Research2022 年末公布的数据（2022Q4/2023Q1 数据暂未公布），包含集成显卡（iGPU，integrated GPU）的统计口径下，英特尔的市占率为三家巨头中的第一：最近有数据的季度 2022 年 Q3，桌面端 GPU AMD 市占率 10%，相比 2021 年同期的 13%下降了 3%；笔记本端 GPU AMD 市占率为 13%，相比 2021 年同期的 20%下降了 7%，英特尔占比则从 65%提高到了 74%，英伟达略有下跌，从 15%下降到 13%；全部 PC GPU（笔记本+桌面）的情况类似，22Q3 AMD 在全部 PC GPU 占比从 2021 年同期的 18%降低到 12%，而英特尔则从 62%提升到了 72%。目前来看，虽然不如独立显卡备受瞩目，但 PC 端集成显卡依然是英特尔掌控的领域。

图表286：iGPU（集成显卡）和 dGPU（独立显卡）



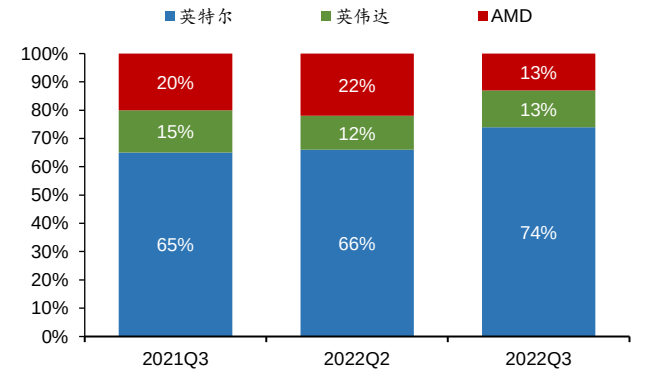
资料来源：英特尔官网、AMD 官网、华泰研究

图表287：22Q3 桌面 PC GPU：AMD 占比 10%，英特尔占比 69%



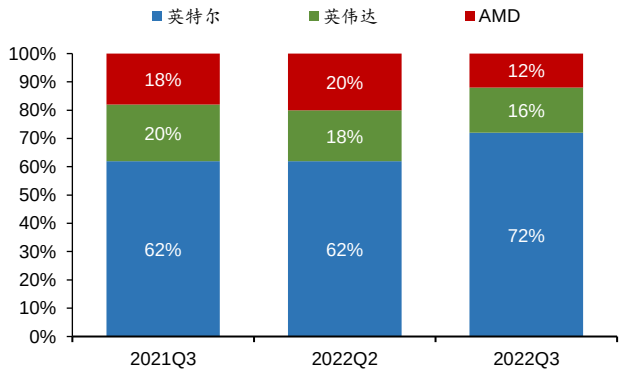
资料来源：Jon Peddie Research、华泰研究

图表288： 22Q3 笔记本端 GPU：AMD 占比 13%，英特尔占比 74%



资料来源：Jon Peddie Research、华泰研究

图表289： 22Q3 全部 PC GPU：AMD 占比 12%，英特尔占比 72%



资料来源：Jon Peddie Research、华泰研究

AMD 在 CPU 市场曾突围英特尔，制程优势能否保持？

回顾 CPU 领域中 AMD 与英特尔自 2016 年开始的争锋历程，AMD 曾凭借台积电的领先制程，一举颠覆在制程上屡遇障碍的英特尔一家独大的局面，成为其 CPU 业务及股价的拐点。在 2020 年 7 月，AMD 和英特尔的股价首次出现倒挂，2022 年 2 月 15 日，AMD 的市值达 1977.5 亿美元，首度超越英特尔的市值。**我们认为，AMD 在 2016 年前在 CPU 领域与如今在 GPU 领域面临的局面有相似也有不同，以下我们对 AMD 在 CPU 领域突围英特尔的过程进行复盘，以此为评估 AMD 的破局能力带来启示。**

我们也要指出，虽然 AMD 在过去的 7 年间分别在 PC 端 CPU 和服务器端 CPU 市场取得了突破性的进展，但是截至 2023 年第一季度，英特尔依然占据着 CPU 市场约 80%以上的市场份额。随着英特尔 IDM 2.0 战略的推进，加上其“四年五节点”的步步落实，或将为其带来赶超台积电先进制程的可能。因此，我们认为，英特尔将持续作为 AMD 最大的竞争对手之一。AMD 目前抢夺到的市占率能否稳固？制程优势又能否维持？

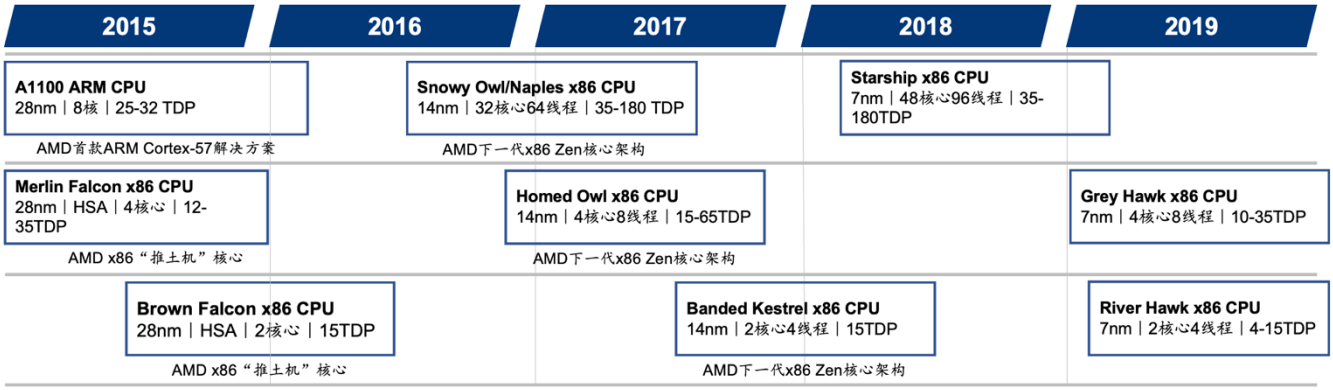
图表290： 2016-2023AMD 及英特尔股价对比图及主要事件（美元）

注：截至 2023 年 7 月数据
资料来源：彭博、AMD、英特尔、华泰研究

AMD 在 CPU 市场突围英特尔全过程复盘

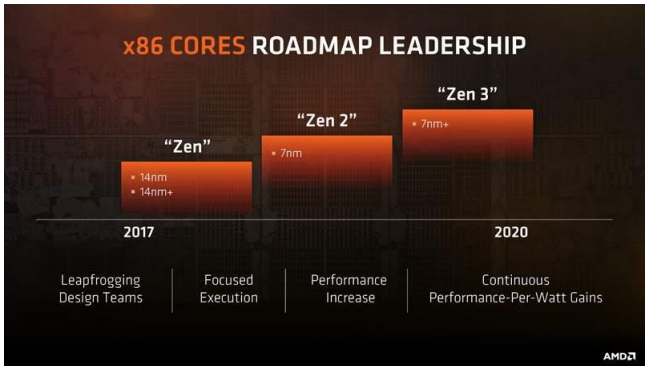
在英特尔正深陷良率等问题不断推迟 10nm 量产的同时，AMD 联手台积电，在制程上不断取得突破，在 PC 及服务器端的制程上纷纷弯道超车英特尔，为市场份额的提升开了绿灯。2016 年上半年，AMD 发布了企业端 CPU 技术路线图，其中明确表示了制程上的突破，基于台积电 7nm 的 CPU 将于 2018/19 年推出。随后 6 月，AMD 发表了 Zen 架构，涵盖 PC 端及服务器端 CPU 产品，并在 2017 年宣布以 Zen 架构重新整合其 PC 及服务器产品。在该 Zen 架构技术路线图中，AMD 进一步明确了 2018/19 年将有 7nm 产品推出，2020 年将向更先进制程迈进。反观彼时的英特尔，由于在制造更先进制程芯片的过程中遭遇技术困难，10nm 芯片良率不佳，导致原定于 2016 年下半年的 10nm（相当于台积电 7nm）量产多翻推迟至 19 年下半年。目前，AMD 的 Zen 架构已发展到台积电 5nm 制程，而 2022 年新版技术路线图进一步更新了其进入台积电 3/4nm 制程的计划。

图表291：AMD 在 2016 年发表的企业端 CPU 技术路线，明确指出台积电 7nm 制程 CPU 将于 18/19 年推出



资料来源：Videocardz、华泰研究

图表292：2017 年 AMD 表示 Zen 架构 CPU 将在 2020 年前超越 7nm

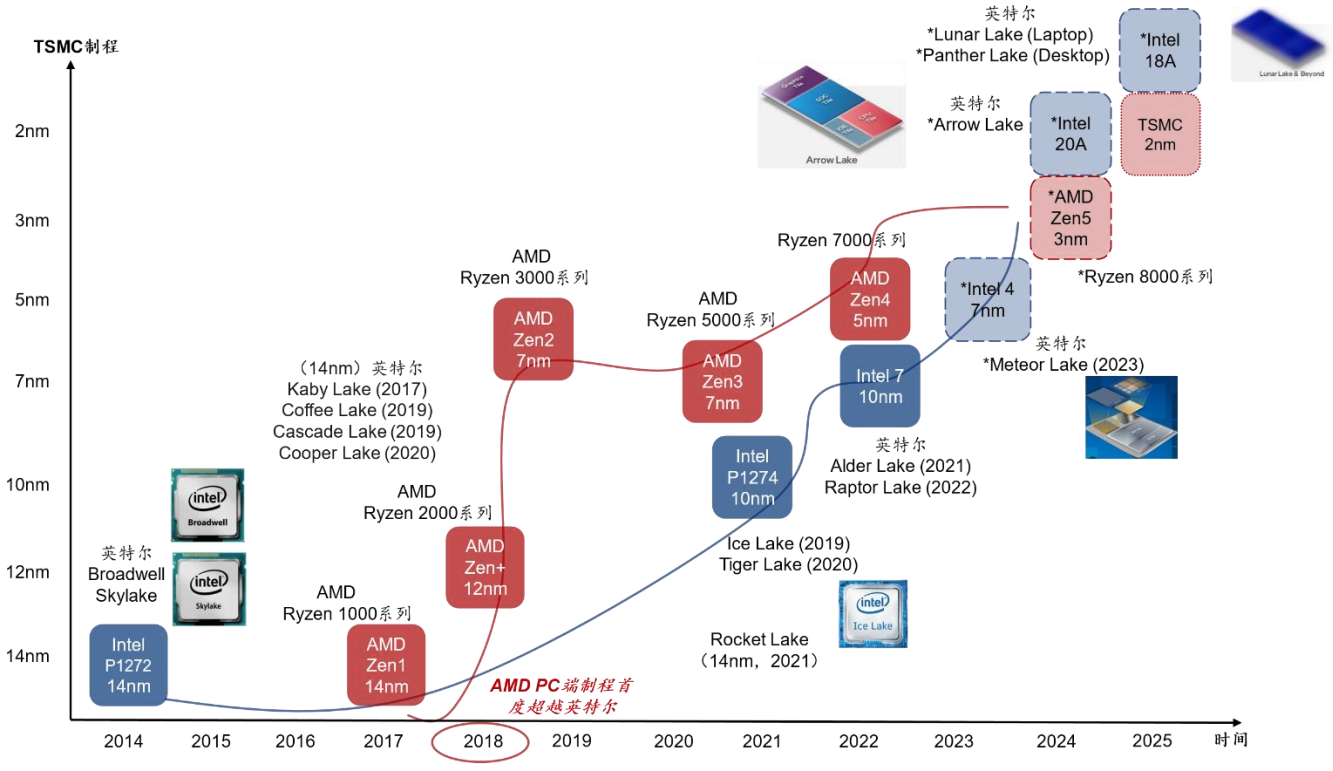


资料来源：AMD 官网、华泰研究

图表293：2022 年 AMD 表示 Zen 架构 CPU 将在 2024 年进入 3nm

资料来源：AMD 官网、华泰研究

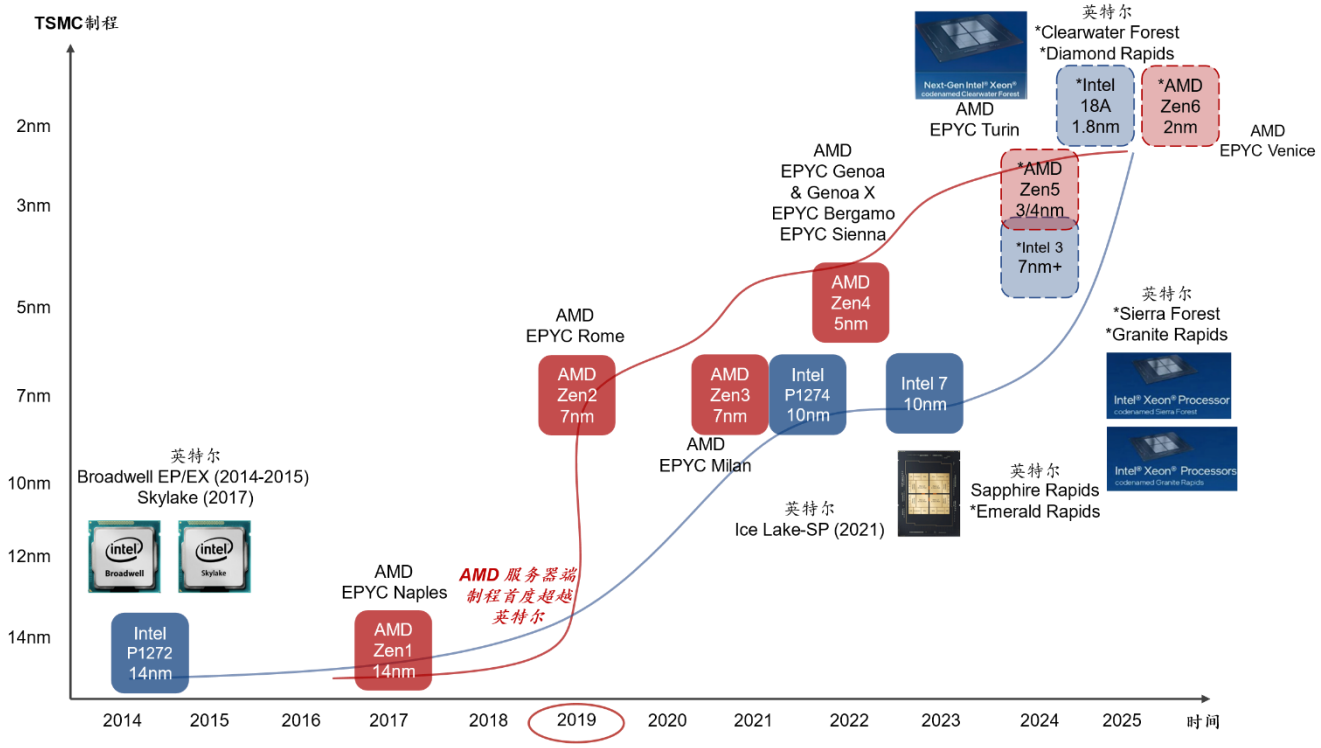
图表294：AMD 与英特尔 PC 端制程发展进度对比及相关产品举例，2018 年 AMD PC 端制程首度超越英特尔



注：*为还未正式推出产品的制程

资料来源：AMD 官网、英特尔官网、华泰研究

图表295： AMD 与英特尔服务器制程发展进度对比及相关产品举例，2019 年 AMD 服务器制程首度超越英特尔



注：*为还未正式推出产品的制程

资料来源：AMD 官网、英特尔官网、华泰研究

2016 年 6 月，AMD 宣布推出 Zen x86-64 微架构。对比彼时英特尔的 Skylake 架构，Zen 的 CPU 部分面积较小，缓存空间有所提升，且散热片间距加宽，默认频率更高，功耗更低，价格也较低。随后在 17 年初上市，同样基于 Zen 架构的 Ryzen 7 系列 PC 端 CPU，采用了 14nm 制程及 8 核 16 线程工艺，对标同为 14nm 制程的英特尔酷睿 i7-6900K（基于 16 年 6 月推出的 Broadwell-E 升级架构）。从此时开始，AMD 的 PC 端产品已逐渐逼近英特尔，而**蚕食份额的趋势也初见苗头。**服务器端方面，AMD 于同年也推出了同样基于 Zen 架构的 EPYC CPU 产品，对标英特尔 Xeon CPU（同为 14nm 制程）。EPYC 凭借高性能表现及高能耗效率，也**开始在数据中心的市场份额上攻城略地。**

2018 年，AMD 推出了采用当时 Global Foundry 12nm 半节点制程的 Zen+架构，并基于此推出 PC 端 Ryzen 5 2600 及 Ryzen 7 2700 CPU，**首度在制程上超越当时还是 14 nm 的英特尔**，加速抢占英特尔的 CPU 市场份额。随后在 2019 年，AMD 在 PC 和服务器端的制程均更上一层楼。AMD 推出了基于台积电 7nm 制程（相当于 Intel 10nm 制程）的 Zen 2 架构，当中包括 Ryzen 9 系列的 PC 端 CPU，以及 EPYC Rome 服务器端 CPU。Zen 2 架构采用了 Chiplet 设计，通过 CPU 和 IO 核心分离，解决超多核心并行问题的同时也降低了生产成本，并达到降低延迟的效果。而在差不多时间里，英特尔的 Ice Lake PC 产品也终于上市量产，该产品采用了英特尔早在 2015 年发表的 Intel 10nm 制程。英特尔虽勉强追赶上 AMD 在 PC 端的制程进度，但在制程上的落后也并没有改善。

图表296： Zen 架构与英特尔 Skylake 架构参数对比

	Zen 架构	Skylake
制程	14nm	14nm
核心/线程	4/8 核心, 8/16 线程	4/8 核心, 8/16 线程
面积	44mm²	49mm²
二级缓存（每核心）	512KB, 1.5mm²	256 KB, 0.9mm²
三级缓存（每核心）	8MB, 16mm²	8MB, 19.1mm²
CPP（nm）	78	70
散热片间距（nm）	48	42
1x 金属片间距（nm）	64	52
标准 6t SRAM（mm²）	0.0806	0.0588
金属片层数	12 w/ MiM	13 w/ MiM

资料来源：AMD 官网、英特尔官网、华泰研究

图表297： 2017 年 AMD 与英特尔数据中心 CPU 产品参数对比

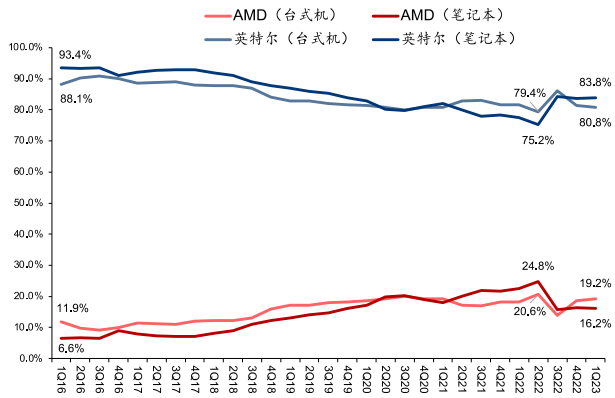
	Intel Xeon E5 Bronze/Silver	Intel Xeon E7 Gold/Platinum	AMD Naples Platform
系列代号	Skylake-SP	Skylake-SP	AMD EPYC
工艺节点	14nm	14nm	14nm
PCH	Lewisburg PCH	Lewisburg PCH	SOC
插槽	Socket P LGA 3647	Socket P LGA 3647	SP3 LGA socket
最大核心数	26	28	32
最大线程数	52	56	64
最大三级缓存	16.5 MB L3	38.5 MB L3	64 MB L3
DDR4 内存支持	6 通道 DDR4	6 通道 DDR4	8 通道 DDR4
热设计功耗	70-85W	85-205W	120-180W

资料来源：AMD 官网、英特尔官网、华泰研究

2020 年 7 月底，英特尔宣布将推迟 Intel 7nm（对标台积电 5nm）制程至 2022 年以后。反观 AMD 在 20Q2 财报中 PC 端业务营收大涨 45%，并进一步上调了全年营收预期。当月 AMD 股价大涨 47%并首度超越英特尔的股价。同年 10 月，AMD 宣布收购头部可编程逻辑器件（FPGA）生产商赛灵思（Xilinx），并于 22Q1 完成并表。对比英特尔在 2015 年收购了 FPGA 生产商 Altera，收购赛灵思能为 AMD 带来 FGPA、可编程 SoC 及自适应计算加速平台产品，并将 AMD 的产品矩阵扩充至与英特尔看齐，为 AMD 数据中心及嵌入式业务如虎添翼。

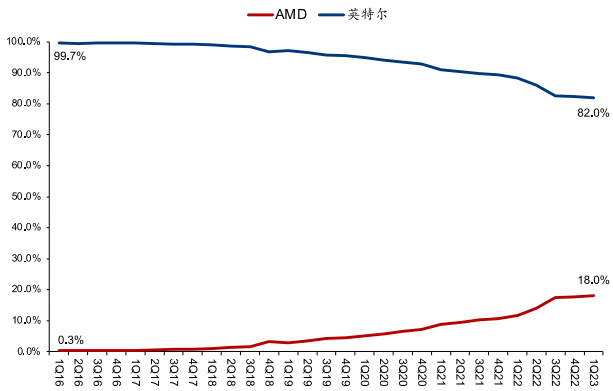
2021 年，AMD 推出了基于台积电 7nm 制程的 Zen 3 架构，并推出了 EPYC Milan 服务器 CPU，对比姗姗来迟的英特尔，此时才推出采用 Intel 10nm 制程的第三代 Xeon 可扩展 CPU。2022 年，AMD 基于台积电 5nm 制程的 Ryzen 7000 系列产品顺利量产，再次拉开与英特尔 PC 端制程的距离（当时英特尔仍处于 Intel 7 阶段，对标台积电 7nm）。

图表298： 16Q1-23Q1 AMD 与英特尔在 PC 端 CPU 市场份额



资料来源：Mercury Research、华泰研究

图表299： 16Q1-23Q1 AMD 与英特尔在服务器端 CPU 市场份额



资料来源：Mercury Research、华泰研究

在 16Q1，AMD 在台式机和笔记本的 CPU 份额仅为 11.9%和 6.6%。而 Zen 架构及相关产品推出后，AMD 开始逐步蚕食英特尔的 PC 份额，并在 18 年制程反超英特尔后，份额提升加速。AMD 的服务器 CPU 在 16Q1 份额仅为 0.3%，市场基本被英特尔所垄断，但 2017 年 EPYC 推出后，服务器 CPU 的份额也开始一路上涨。至 22Q2，AMD 在 PC 端（台式机/笔记本）CPU 份额分别攀升到 20.6%和 24.8%，为史上最高，但随后 Q3 出现回落。截止 23Q1，AMD 台式机和笔记本 CPU 份额为 19.2%和 16.2%，服务器端 CPU 市场份额为 18.0%；英特尔台式机和笔记本 CPU 份额为 80.8%和 83.8%；服务器端 CPU 市场份额为 82.0%。

英特尔在服务器端 CPU 的反击：Xeon 在 AI 推理端的应用

我们观察到，2023 年第一季度英特尔和 AMD 在服务器 CPU 市场的市占率维持了稳定。英特尔在 23Q1 业绩电话会上称 2023Q1 服务器市场份额维持稳定的原因之一是“Sapphire Rapids 带来的贡献”，这也是 CEO Pet Gelsinger 掌舵英特尔后第一次维持住服务器市场占有率大致平稳不降。

Intel 在数据中心业务的布局广泛，包括未来几代 P-Core 及 E-Core Xeon CPU、ASIC 产品 Habana Gaudi 2&3、Falcon Shores GPU 及多款 FPGA。**其中，我们认为 Xeon CPU 产品在推理端具备较强竞争力**，这也是本节的讨论重点。另外，Habana Gaudi 系列或能与 GPU 在训练端有一战之力，而 FPGA 则囿于功耗表现和经济效益不佳，市场空间较为有限，Falcon Shores 曾经规划为 CPU+GPU 架构，本有可能与 AMD 的 MI300A 及英伟达 GH200 在 AI 端展开较量，但现在已正式改为纯 GPU 产品。

英特尔认为，AI 推理端的 TAM（Total Available Market，潜在市场规模）应该比训练端的 TAM 要大，因此公司的 CPU 将受惠于未来 AI 推理的可触及市场规模，同时认为 CPU 在推理端的位置不会轻易被 GPU 取代。推理端对于算力的要求相对训练端较低，目前各类芯片，包括 GPU、CPU、ASIC 等，都在此领域获得一席之地，因此 AI 推理方面英特尔存在胜出机会。从发展历程来看，传统推理端主要依赖 CPU 处理大多数相对简单的推理任务。然而目前 AI 模型的规模、模态和复杂度跟过去相比提升，随着更多工作负载将逐渐纳入到推理领域，对于算力的要求也会提高。但我们需强调，**推理所需要的算力本身比训练所需要的算力要低**，与此同时，考虑到采用英伟达最高性能的 GPU 来进行推理工作成本或过高，因此 GPU、CPU、以及 ASIC 等依然在相互竞争。

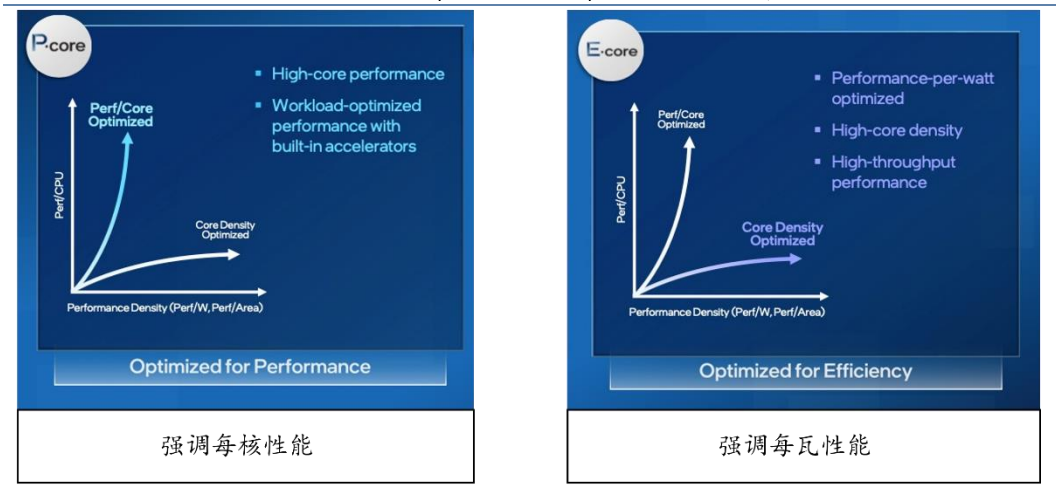
“从云端到边缘”是英特尔对 AI 市场前景的又一预测。英特尔认为，鉴于延迟问题和成本问题，AI 将向边缘转移。据 Gartner 预测，2025 年，将有多达 75%的企业数据会在传统数据中心以外生成。**对于边缘 AI 运算来说，面对的应用场景更为丰富，垂直行业会产生不同类型的数据，因此同样适用各种类的芯片。**例如，智慧城市需要大量处理视频（监控、交通录像）、零售业需要处理语言和文本数据（与客户的对话、广告与营销）、医疗行业则需要大量处理图像数据（造影诊断等）。

图表300：英特尔希望切入的边缘计算目标应用场景

资料来源：英特尔官网，华泰研究

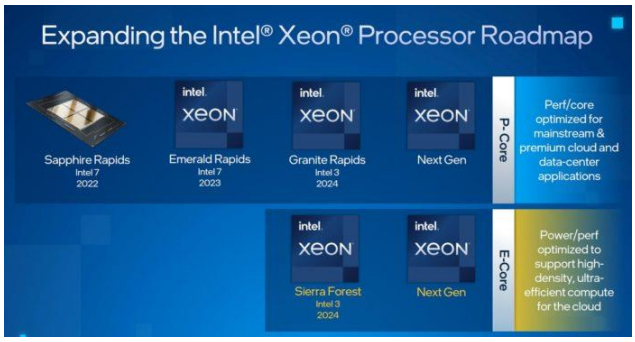
性能与功耗双管齐下，Xeon Scalable“大核”（P-Core）与“小核”（E-Core）策略能否反击 AMD？P-Core（Performance Core）强调高性能；E-Core（Efficiency Core）着眼低能耗，我们认为，英特尔对服务器端产品的战略规划思路或来自其遭 AMD 撬动份额的反思。英特尔在服务器方面落后 AMD 的重要一环或许正在于英特尔更关注“每核性能（performance per core）”，而 AMD 则将重点放在“每瓦性能（performance per watt）”，前者指的是提高 CPU 单核的性能，而后者则考虑的是产品的能耗效率。面对数据中心降本增效及解决高能耗问题的发展趋势，单位能源带来的经济效益越来越重要，在这个过程中，英特尔认为 AMD 抓住了市场对于降低能耗的需求。根据 AMD 在 2022 年 11 月发布第四代 EPYC 时的测算，其每瓦性能比第三代 Xeon 高 1.7 倍。

图表301： P-Core 与 E-Core 的关注重点不同，前者关注性能，后者关注能耗效率



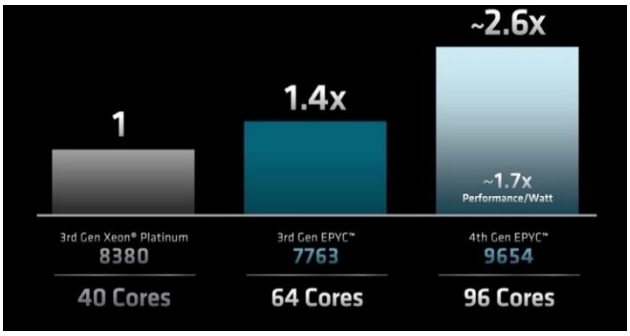
资料来源：英特尔官网、华泰研究

图表302： P-Core 和 E-Core 产品布局



资料来源：英特尔官网、华泰研究

图表303： AMD 第四代 EPYC 比英特尔第三代 Xeon Performance/Watt 高 1.7 倍



资料来源：AMD 官网、华泰研究

图表304： 英特尔在 ISC 2023（2023 年 5 月 21 日至 24 日）公布的产品路线图

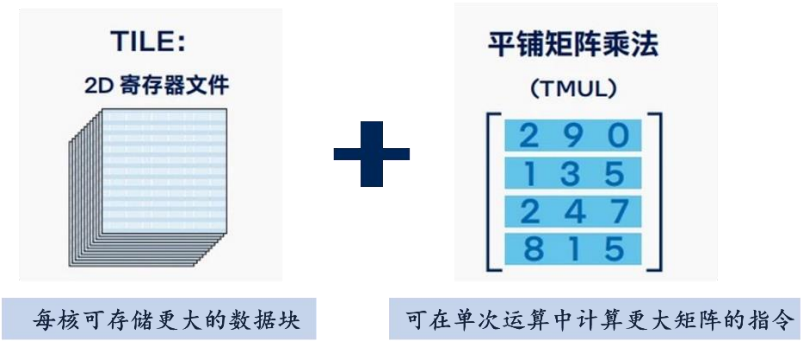
资料来源：英特尔，华泰研究

对此，英特尔针对高性能和低能耗两种需求提供不同的产品，既要覆盖更注重功率能耗的应用场景，又不放弃高性能需求。根据 Xeon Scalable 现有的推出计划，英特尔将在 2023 年 Q4 推出 Emerald Rapids（Intel 7，P-Core），2024 上半年将推出 Sierra Forest（Intel 3，E-Core），随后同年推出 Granite Rapids（Intel 3，P-Core），2025 年推出 Clearwater Forest（18A，E-Core）。**这种全面的产品策略符合此前英特尔对其产品战略的总结：提供足够多的选择，满足更广泛的用户光谱。**

2023 年 1 月推出的 Sapphire Rapids 引入的 AMX（Advanced Matrix Extensions，矩阵运算扩展）提高了 Xeon 的深度学习能力，为英特尔的 CPU 产品在 AI 推理端的进一步应用带来了更多可能。Xeon 是在 AI 推理端广泛应用的头把交椅，早在 2021 年 12 月，英特尔曾对全球运行 AI 推理工作负载的数据中心服务器装机情况进行市场建模，英特尔当时称，推理所用的已装机 CPU 中高达 70%是英特尔 Xeon 产品。尽管如此，面对瞬息万变的 AI 市场，面临 GPU、ASIC 等产品的竞争，Xeon 也在提升其 AI 能力：AMX 在 2020 年就已经发布，2021 年就已确定会被引入 Sapphire Rapids，但由于 Sapphire Rapids 的延后，2023 年才与其一同推出。AMX 包含两个组件，分别是：1）TILE，由 8 个 1KB 大小的 2D 寄存器组成，可存储大数据块，2）平铺矩阵乘法（TMUL），它是与 TILE 连接的加速引擎，可执行用于 AI 的矩阵乘法计算，其中 TMUL 是 AMX 的核心，可以使 Xeon 具备更强的 AI 学习与推理所需要的矩阵运算能力。

根据英特尔以第三代 Xeon 为基准的测试，第四代 Xeon（Sapphire Rapids）可以在多个大模型里实现训练端 3.5-10 倍的性能提升，推理端可以实现 5.7-10 倍的性能提升。**根据基准测试，AMX 带来了 Xeon 代际间的 AI 能力提升。但是需要注意的是，英特尔没有以其他厂商的产品为基准进行对比测试，因此对于 AMX 能为 Xeon 带来何种的横向对标暂未可知。**

图表305：AMX 架构由 TILE 和 TMUL（平铺矩阵乘法）组成



资料来源：英特尔，华泰研究

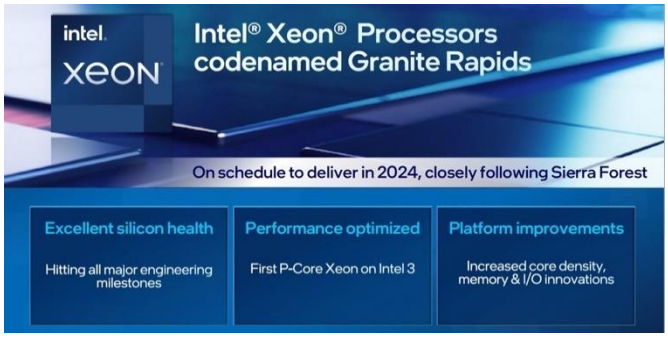
图表306：2023 年英特尔 DCAI 会议中公布的 Xeon 路线图

资料来源：英特尔，华泰研究

图表307： Sierra Forest 将于 2024 年一季度开始出货

资料来源：英特尔，华泰研究

图表308： Granite Rapids 将于 2024 年在 Sirerra Forest 后推出



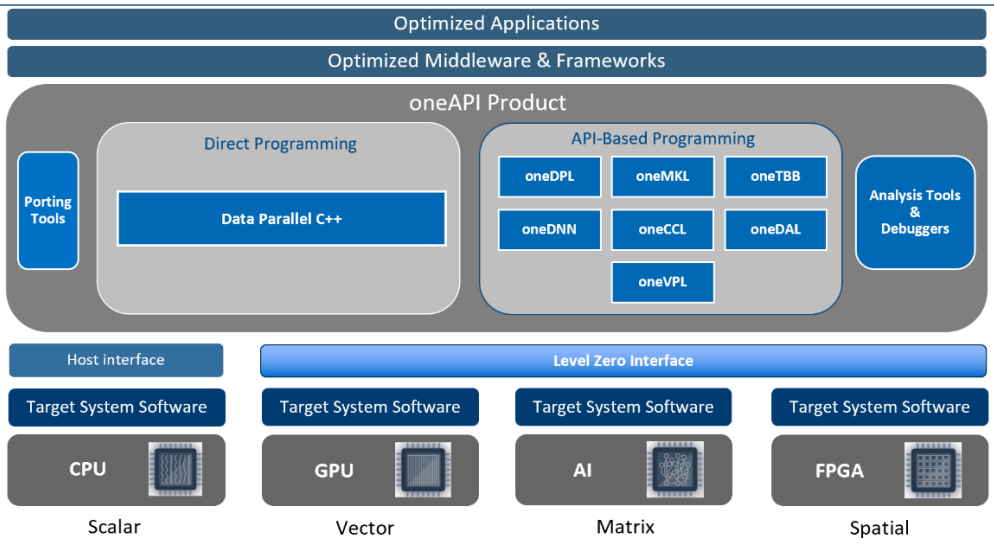
资料来源：英特尔，华泰研究

与 AMD 一样，英特尔推进 AI 产品布局成功的另一大障碍或许是软件生态。英特尔的软件框架是 oneAPI，于 2019 年底开始测试，2020 年 9 月推出了 1.0 正式版，比 CUDA 晚了 13 年。对此，英特尔也承认 oneAPI 推出较晚，但同时也认为目前只是 AI 的起步点。对比英伟达 CUDA（Compute Unified Device Architecture）于 2007 年发布，通过先发优势和长期耕耘，生态圈已较为成熟，为英伟达 GPU 开发、优化和部署多种行业应用提供了独特的护城河。全球 CUDA 开发者 2020 年达 200 万，2023 年已达 400 万。

不过，oneAPI 不直接与 GPU 通用运算生态圈的领军 CUDA 竞争，而是横跨 CPU、GPU、FPGA、NPU 等多种硬件，以及 CUDA、ROCm 等不同软件平台，试图建立统一的生态圈，但这种兼容所有软件和硬件的思路，落地效果如何，能否突出 CUDA 重围，目前看还需进一步判断。

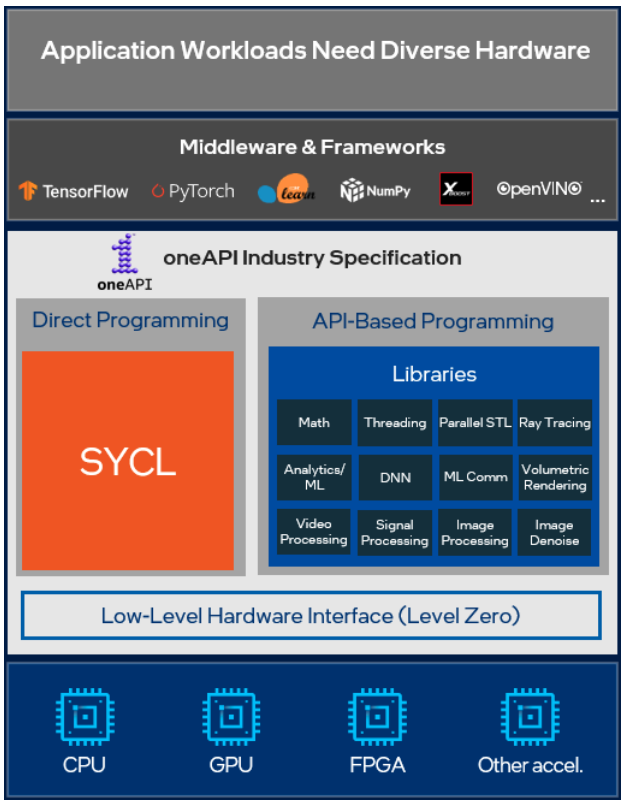
在英特尔最初发布 oneAPI 时，口号是“No transistor left behind（所有晶体管都不能落下）”，这是指英特尔希望 oneAPI 能提供覆盖多种硬件的异构编程模型，并且成为行业标准。oneAPI 面对“统一”的目标方向进行了诸多努力：1）软件方面，容纳其他软件生态圈的代码，例如 oneAPI 提供的 DPCT（Data Parallel C++ Compatibility Tool，即 DPC 兼容工具，其中 Data Parallel C++是 OneAPI 的核心编程语言）工具，可移植 95%以上的 CUDA 程序；2）硬件方面，oneAPI 本就基于科纳斯组织（Khronos Group）的 SYCL 规范（基于 C++的异构平行编程框架，发音为 sickle，是一个名称而并非缩写），支持异构硬件编程，而为了更有效经营 oneAPI 的硬件统一生态，2022 年 6 月英特尔宣布收购 Codeplay（一家 SYCL 编程机构），这家公司编写的库在英伟达、AMD 和英特尔的硬件上都可以运行，利用了 SYCL 的可移植性，这被称为“一次写入，多次调整”，类似于英特尔提到的“write the code once and it works everywhere”。然而，移植代码本身，与移植后的代码在其他硬件和软件平台上运行的效率如何是两个不同的命题，目前 oneAPI 发展时间不过三年，能否突出 CUDA 重围，我们认为还需要更多基准测试才能判断。

图表309： oneAPI 架构示意图



资料来源：英特尔，华泰研究

图表310： 基于 SYCL 的 oneAPI 支持多种硬件和多种框架

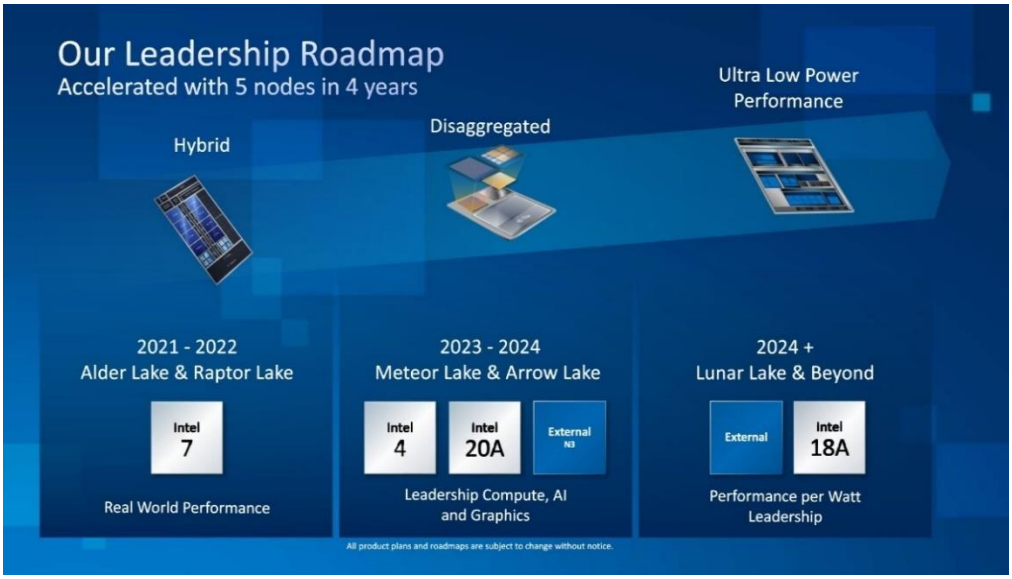


资料来源：英特尔，华泰研究

赛点时刻到来：英特尔的“四年五节点”先进制程赶超计划

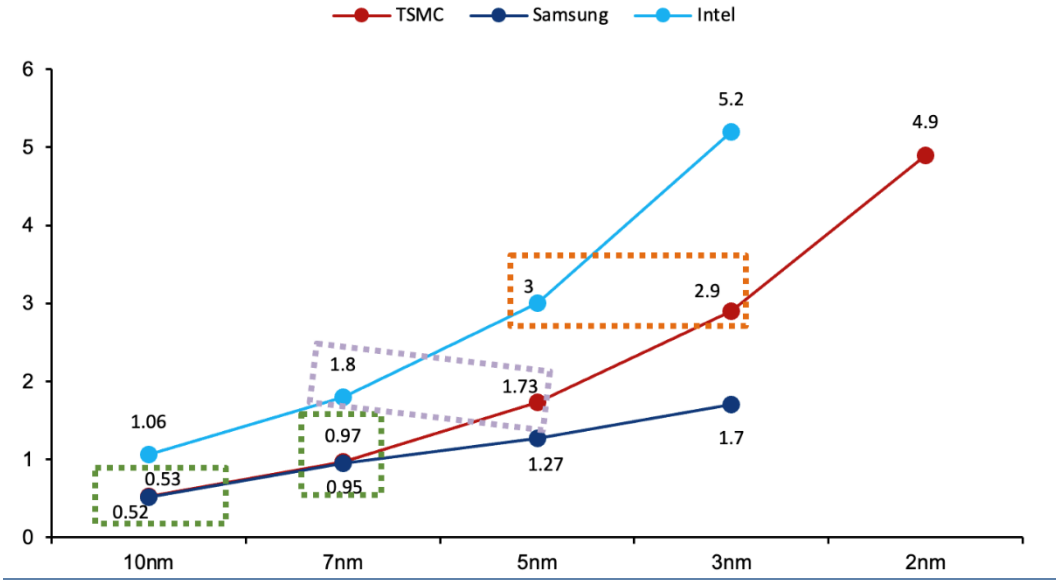
英特尔的产品路线图能否实施成功，有赖于公司较为进取的先进制程路线图能否顺利落地。早在 14nm 工艺时，英特尔占据着领先优势，却在 10nm 的关口停滞不前。2019 年，AMD 在服务器端采用的工艺制程首次超越英特尔：采用 10nm 工艺的服务器版 Ice Lake 于 2021 年 4 月姗姗来迟，此时距离 AMD 在 2019 年推出 7nm 工艺的 EPYC Rome 已有两年（AMD 采用的台积电工艺 7nm 大致相当于英特尔 10nm）。2018 年，AMD 在 PC 端采用的制程开始超越英特尔，而英特尔直到 2019 年 Ice Lake 推出才进入了 10nm 阶段，2021 年还发布了 14nm 的 Rocket Lake。

图表311：四年五节点规划图：2024 年进入 Intel 18A 节点



注：2023 年 6 月发布
资料来源：英特尔，华泰研究

图表312：晶体管密度对比，台积电工艺 7nm 相当于英特尔 10nm，台积电 5nm 为英特尔 7nm（Intel 3），台积电 N2 相当于英特尔 20A（单位：100 million 晶体管/mm²）



注：虚线框中为大致相同的晶体管密度
资料来源：Digitimes，华泰研究

此后，英特尔一直在艰难追赶。英特尔于 2021 年 7 月公布了工艺制程的赶超战略“四年五节点”，指的是公司希望能在 2025 年以前实现 Intel 7（10nm）、Intel 4（7nm）、Intel 3（7nm+）、Intel 20A （2nm）及 Intel 18A（1.8nm）5 代工艺节点。在这五代节点中，前三个节点的目的是追赶上台积电的进度，后两个则是英特尔在 2025 年超越台积电重返制程领先地位的关键。2021 年计划发布时，英特尔预计 Intel 18A 将于 2025 年推出；2022 年，英特尔将 18A 时间点提前到了 2024 年下半年。目前看到的服务器端进度显示，基于 Intel 7 的 Sapphire Rapids 今年继续批量出货，同样基于 Intel 7 的 Emerald Rapids 有望在今年推出，基于 Intel 3 的 Sierra Forest 和 Granite Rapids 在 2024 年相继推出，18A 方面已开展内部测试以及与潜在代工客户的产品测试。

CEO 基辛格极力显示了公司实现“四年五节点”先进制程赶超计划的决心，英特尔意识到，能否在先进制程上赶超台积电，是决定英特尔未来最重要的一环。在 2023 年 3 月的 Morgan Stanley Technology, Media and Telecom 会议上，基辛格称英特尔将开源节流，但唯一不会裁减开支的领域就是先进制程计划，“四年五节点必须实现（five nodes in four years has to happen）”，而英特尔也会为了这个目标，进行所有必须的资本投资。

如果英特尔能够按照计划顺利推进，则台积电在制程上的领先优势或将大幅缩小，甚至在 2025 年会出现英特尔超过台积电的情况，但其能否在 2025 年及以前顺利落地“四年五节点”赶超台积电还需要进一步观察。根据台积电的规划，N2 制程（2nm）将在 2025 年开始量产，而英特尔目前的计划是在 2024 年开始量产 20A（2nm）和 18A（1.8nm），如果二者分别都实现自己的计划，那么 2025 年英特尔在先进制程方面应该超过台积电。

从现在的进展看，五节点中的前两个已经基本完成：Intel 7 的 Sapphire Rapids 批量出货，Intel 4 的 Meteor Lake（PC 端）计划下半年推出。这样一来，关键就在于英特尔对接下来三个节点进度的掌控能力，而根据公司一季度报告，Intel 3 的 Sierra Forest 将于 2024 年上半年开始出货，Granite Rapids 紧跟其后。暂到目前为止，计划的确如公司所说在按部就班进行（“squarely on track”）。但是，我们也要将公司以往在 10nm 进程上的一再延后纳入考量，不过，现在公司的先进制程推进能力相比当时却有了多大程度的提高，但还需要结合 Intel 3 在 2024 年的落地情况进行观察，能否实现 2024/2025 赶超台积电的战略规划，尚需拭目以待。

向外寻求代工，与台积电的合作进展或能缓解英特尔的先进工艺及产能压力。IDM2.0 包括三个主题，分别是：1）一部分产品继续在内部工厂网络生产；2）一部分产品寻求外部代工产能，以优化成本、推动进度和提高供货能力；3）为客户提供代工服务（IFS 业务）。其中第一条，内部工厂网络的生产能力和技术水平，有赖于英特尔先进工艺的进展，上文已有分析，而第二条的核心则是与台积电的合作。

据《联合报》，今年 5 月下旬 Pet Gelsinger 再次到访台积电，其访问的目的是寻求台积电 3nm 制程（N3，大致相当于 Intel 4）的代工产能去生产将在 2023 年底推出的 PC 端 CPU Meteor Lake 中的一个 chiplet（tGPU，tiled GPU，英特尔称其为 Next Gen Graphics Engine，是 Meteor Lake 中的 GPU chiplet）。Meteor Lake 中的 SoC 和 IOE 单元则采用了台积电 6nm。另外，据 Wccftech 2022 年 10 月 21 日报道称英特尔将于 2024 年推出的 Arrow Lake 中或也将使用台积电 3nm，或也将是在其中的某个 chiplet 上。此前，基辛格曾在 2022 年 4 月访台积电参加闭门会议，据 TechWeb 同时间报道称讨论了 7nm 及以下制程产能。从这两次访台积电的讨论内容及新品消息来看，目前与台积电的合作似主要在 PC 业务，未来如在 DCAI 业务的产品中也如此般在部分 chiplet 中或全部采用台积电工艺，或能进一步缓解制程进步和产能的压力。

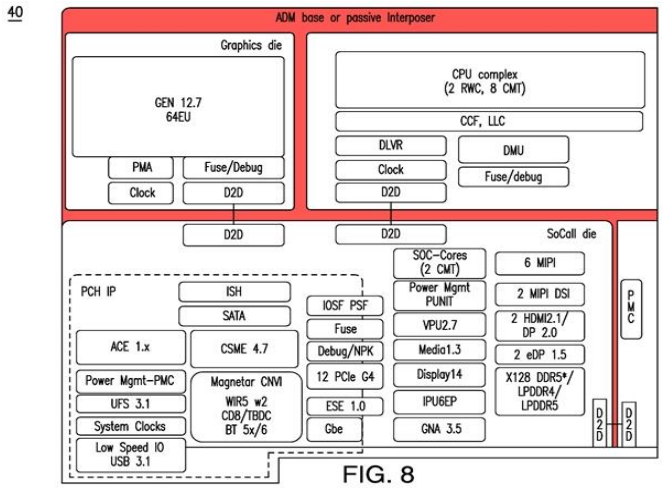
Meteor Lake 将于 2023 年下半年推出，tGPU 或将打开 Intel 笔记本端 CPU 新阶段。5 月初，英特尔全球传播总监 Bernard Fernandes 称公司目前已处在客户端产品拐点，Meteor Lake 即将推出（消息称仅有 laptop 版，desktop 版或不会发布），而在 2023 年一季度财报会议上，公司宣布基于 Intel 4 工艺节点的 Meteor Lake 正在加速量产，将于 2023 二季度推出。1）Meteor Lake 采用 Tile 架构，让不同的单元可以采用不同的制程，让英特尔借助台积电成熟工艺和低成本成为可能；2）Meteor Lake 中 tGPU 单元首次采用了台积电 5nm 制程，SoC 和 IOE 采用了台积电 6nm 制程，而 CPU 则采用英特尔自己的 Intel 4 制程，封装技术采用归属 Intel 16 的 3D Foveros。

我们认为英特尔与台积电在 Meteor Lake 的合作（及未来的更多合作）是重要看点，符合英特尔 IDM2.0 扩大第三方代工产能的战略，或将能使公司更有效抗衡早早借助台积电东风的 AMD。除此之外，第三方 mooreslawisdead 官网认为 Meteor Lake 中的 tGPU 单元（核显）使用的 Iris Xe 或将足以替代笔记本显卡，将为笔记本集成显卡市场带来威胁。Meteor Lake 的另一个特点在于其或将支持 L4 四级缓存（Adamantine，架构图中简称为 ADM），可以提供比 L3 缓存更快的访问速度，缩短加载时间。Meteor Lake 的具体信息（如功耗降低等其它性能上的提升，是否会有桌面版等）还需要继续关注英特尔公布的有关产品内容，仅从目前已有的消息，市场及评测机构对 Meteor Lake 已经拉高期待，产品或为英特尔 PC 端业务带来新增长。

图表313： Meteor Lake 单元结构

资料来源：英特尔，华泰研究

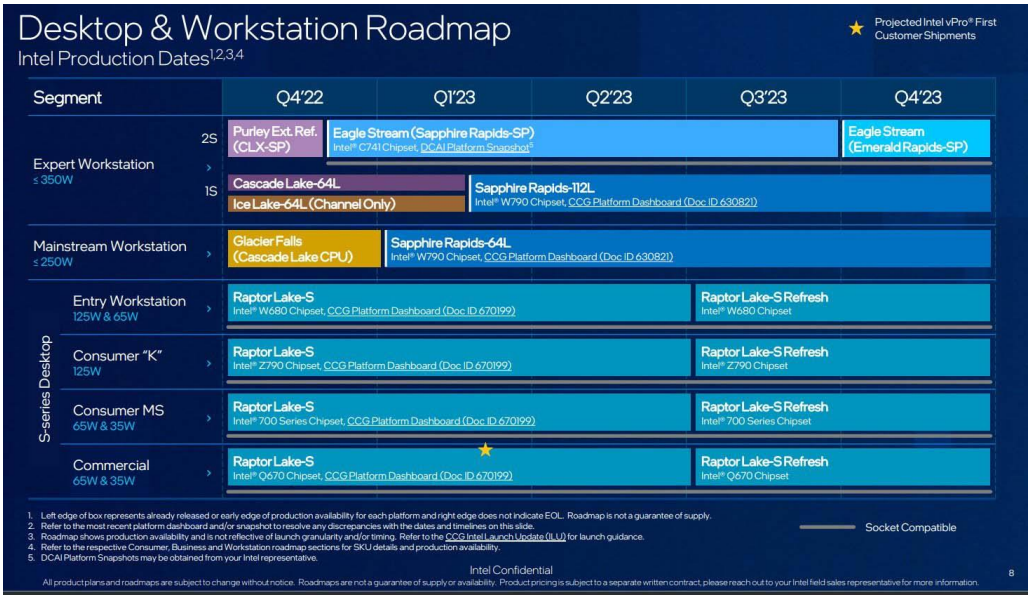
图表314： Meteor Lake 内部结构图（ADM 即 Adamantine，四级缓存）



资料来源：英特尔专利文件，华泰研究

在 desktop 桌面端 CPU 方面，面对二月上市的 AMD Ryzen 7950X3D，英特尔将于 2023 年第三季度推出 Raptor Lake-S Refresh，能否对打 AMD 还需进一步判断。Raptor Lake-S Refresh 官方加速频率达 6.2+GHz，市场期待其可以应对 AMD 在一季度推出的 5.7GHz 加速频率的 Ryzen 7000X3D 系列处理器。据 2023 年 8 月 21 日 Hardware Times 报道，Raptor Lake-S Refresh 应将于 2023 年底推出。Raptor Lake-S Refresh 依然属于 13 代酷睿，在此之前，AMD Ryzen 7000X3D 已将 Intel Raptor Lake CPUs 在游戏用高性能 CPU 领域击败：根据 Notebook Check 测评，Ryzen 7950X3D 较 Core i9-13900K 在游戏性能表现上高 3.8%，同时能耗效率高两倍。在 Raptor Lake-S Refresh 消息确定之前，第三方 mooreslawisdead 官网认为桌面 CPU 领域英特尔无法匹敌 AMD，除非尽快推出 Arrow Lake，然而消息确定后 MLID 认为也许 13 代酷睿的新亮相可以与 Ryzen 7000X3D 系列一战。

图表315： 英特尔桌面（及工作站）产品路线图（至 2023Q4）



资料来源：英特尔，华泰研究

突围战 II：逐鹿 AI 芯片，AMD 几成胜出机会？

上一节中，我们对 AMD 在 CPU 市场成功突围英特尔进行了复盘。我们认为未来三年内 AMD 的重估机会之钥应在 AI 芯片，并希望能通过对第一次突围的理解，判断 AMD 在 AI 浪潮中能否再次突围。表面上看，目前 AMD 面临的局面与 2016 年在 CPU 市场类似，但事实上，如今局势较此前更为复杂：1）2016 年时，AMD 在 CPU 市场的唯一对手是英特尔，但如今在 GPU 市场，虽表面上看 AMD 前方只有英伟达一家独大，但实际还需面对来自云厂商自研、初创 AI 芯片企业新技术探索等多方压力；2）英特尔被 AMD 攻破的最大原因是其在制程上的一再落后，但如今 AMD 和英伟达都采用台积电代工，在制程上不会出现很大差别。对于 AMD 箭在弦上的第二次突围，我们认为，虽然其市占率的爬升过程或将类似其在 CPU 市场的蚕食方式，但是由于 AI 浪潮本身的高速，如能成功突围，提升市占率到 5%以上所需时间应较短，我们认为或能抓住 2 年之机（在 CPU 市场从 0.3%到 20%则用了 4 年）。

AMD 的 AI 战略主要包括三个方向：1）广泛的 CPU 和 GPU 产品组合；2）开放稳定且已证明（open, steady and proven）的软件平台；3）深入完善的生态系统。AMD CEO Lisa Su 于 5 月 31 日的《福布斯》采访时强调“放眼 5 年，将在 AMD 每一个产品中看到 AI”，AI 是公司的战略首位。目前英伟达领军 AI 训练端，但随着 AMD 在 AI 芯片上逐步发力，我们认为或能开始撼动英伟达在行业里独占鳌头的地位。我们也认为云厂商应不希望 AI 芯片呈现一家独大的局面，MI300 恰逢其时地出现，为市场提供了英伟达以外的选择。**然而，MI300 虽备受瞩目，但目前在客户方面几乎未有正式披露，这意味着一旦正式公布客户，将有效提振市场信心。目前，云厂商应还在对 MI300 进行测试和下单阶段，我们将对后续公布的订单情况保持关注。**

MI300 系列能否“虎口夺食”英伟达

AMD 在今年 1 月的 CES 2023 大会里介绍了 MI300（也就是现今的 MI300A），它是 CPU 与 GPU 结合架构，聚焦 AI 和 HPC（high performance computing），对标英伟达 Grace Hopper（Grace CPU + Hopper H100 GPU）。通过 MI300，AMD 也一改过去 GPU 产品主要应用在图像处理及较简单的 AI 推理领域的局限，**我们认为，MI300A 应是除了谷歌的 TPU 之外，能与英伟达在 AI 训练端上匹敌的产品。**

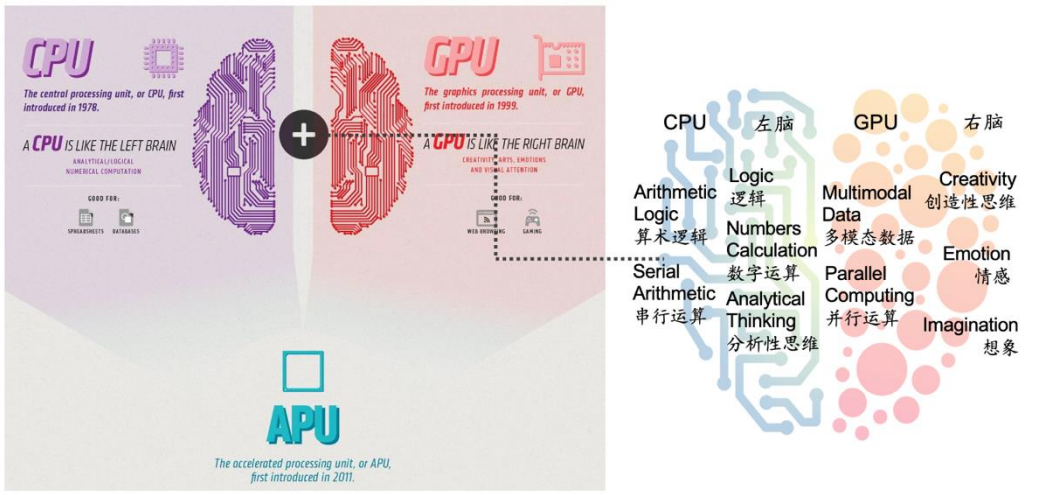
MI300 虽是 AMD 的首个 CPU+GPU 架构的数据中心产品，但其实 AMD 早在 2011 年发布的 APU（Accelerated Processing Unit）就已经是 CPU+GPU 架构，但当时只用于 PC 端。因此，AMD 在 CPU+GPU 架构具备深厚的 know-how。AMD 的 CPU+GPU 架构理念可以追溯到 2006 年。当时 AMD 通过收购 ATI 获得 GPU 以及芯片组技术，并在同年开展 AMD Fusion 项目（即后来的 APU），提出将 CPU 和 GPU 集成到一颗芯片上的理念，但当时 AMD 的 CPU 和 GPU 采用 45nm 制程，将两者放在同一芯片上的难度较大。直到 2011 年，AMD 发布首款 APU 产品 Llano。目前，MI 系列在 MI300 之前的产品都是纯 GPU，比如说 MI200 家族（包括 MI250 及 MI250X）均是纯 GPU。2017 年，AMD 发布的论文《Design and Analysis of an APU for Exascale Computing》中讨论了包含 CPU、GPU 和 HBM 内存堆栈的 APU 芯片设计。

AMD 对芯片与人类大脑的操作理解较为超前。早在 2011 年，AMD 产品构想中就以 CPU 和 GPU 分别类比人类左右脑，并基于此提出了 CPU+GPU 的异构产品策略。类比人脑，AMD 认为左脑更像 CPU，负责对信息的逻辑处理，如串行运算、数字和算术、分析思维、理解、分类、整理等，而右脑更像 GPU，负责并行计算、多模态、创造性思维和想象等。

人类大脑神经网络的运作模式，始终是人工智能追求的终极形态，因此，我们认为 CPU+GPU 的异构集成，对比人类可实现左右脑协同工作，整体调动神经网络，或将成为 AI 芯片的主流技术方向。目前 AMD 的 MI300 和英伟达的 Grace Hopper 在此均有布局。GPU 的算力高，并针对并行计算，但须由 CPU 进行控制调用，发布指令。在 AI 训练端，CPU 可负责控制及发出指令，指示 GPU 处理数据和完成复杂的浮点运算（如矩阵运算）。

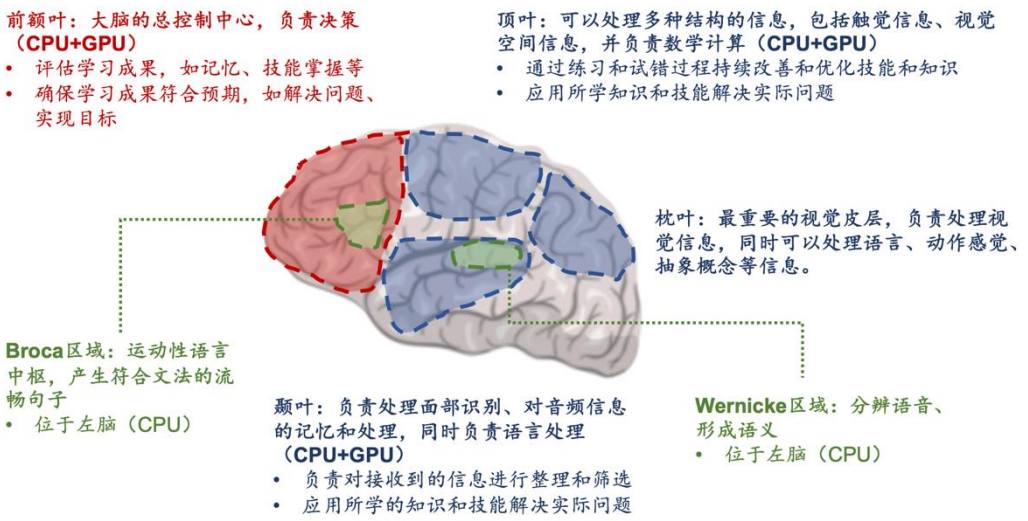
在面对不同模态数据的推理时，我们认为，CPU 与 GPU 的分工也各有不同，因此，同时部署 CPU 和 GPU 能提供最大的运算支撑。例如，在处理语音、语言和文本数据的推理时，AI 模型需逐个识别目标文字，计算有序，因此或更适合使用擅长串行运算的 CPU 进行运算支持；但在处理图像、视频等数据的推理时（对比人类的操作，每一个像素是同时进入眼睛），需要大规模并行运算，或更适宜由 GPU 负责，例如英伟达 L4 GPU 可将 AI 视频性能提高 120 倍，据英伟达测试，L4 与基于 CPU 的传统基础设施相比能源效率提高 99%。

图表316：2011 年 AMD 已提出 APU 概念，将结合 CPU 与 GPU 在左右脑层面的分工区别和组合构想



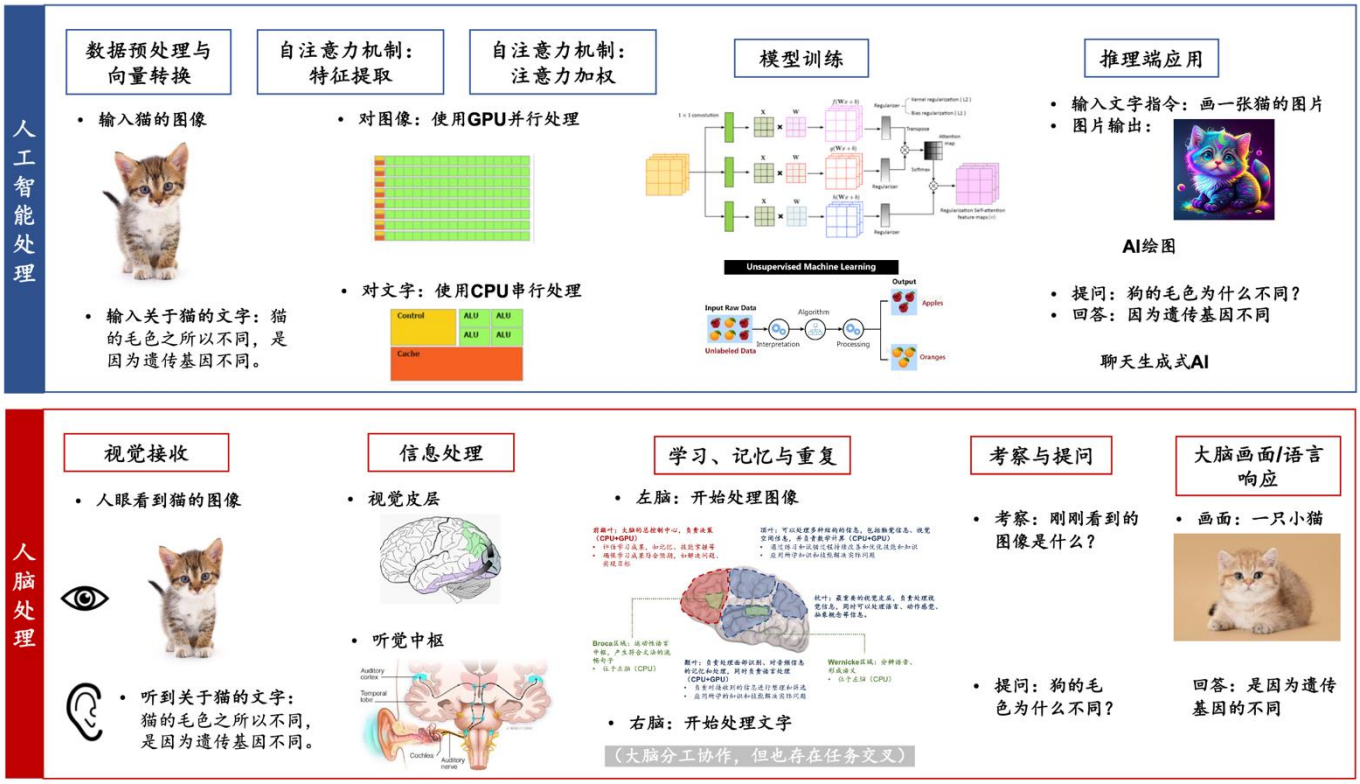
资料来源：AMD 推特官方、华泰研究

图表317：人类大脑不同部分功能及对应芯片种类



资料来源：Hari R. From brain–environment connections to temporal dynamics and social interaction: principles of human brain function[J]. Neuron, 2017, 94(5): 1033-1039.、BrainFacts、华泰研究

图表318： 人脑处理信息与人工智能训练和推理的流程对比



资料来源：CSDN、谷歌官网、Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.、华泰研究

图表319： 2017 年 AMD 发布的论文中讨论了包含 CPU、GPU 和 HBM 内存堆栈的 APU 芯片设计

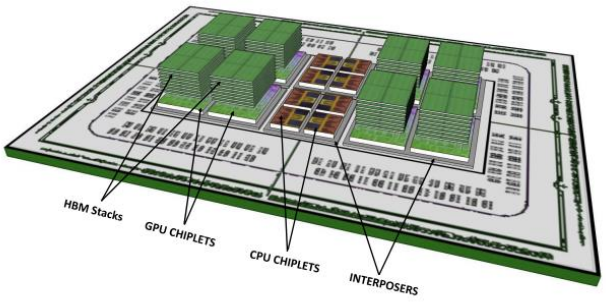
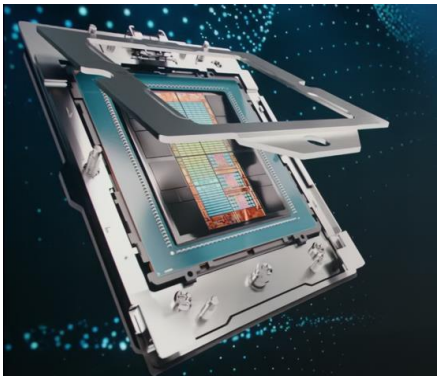


Figure 2. Exascale Heterogeneous Processor (EHP)

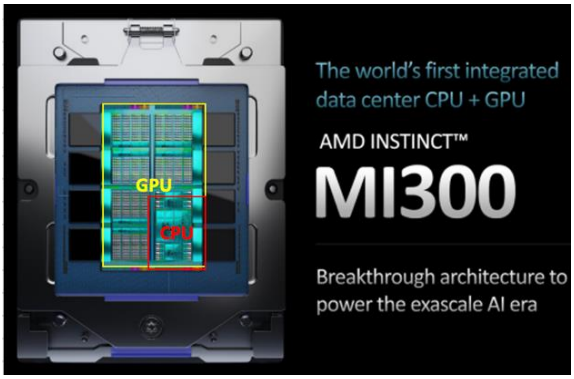
资料来源：T. Vijayaraghavan et al., "Design and Analysis of an APU for Exascale Computing," 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA), Austin, TX, USA, 2017, pp. 85-96, doi: 10.1109/HPCA.2017.42.、华泰研究

图表320： AMD MI300A 产品实物图



资料来源：AMD 数据中心&AI 首映式、华泰研究

图表321： AMD MI300A 产品示意图

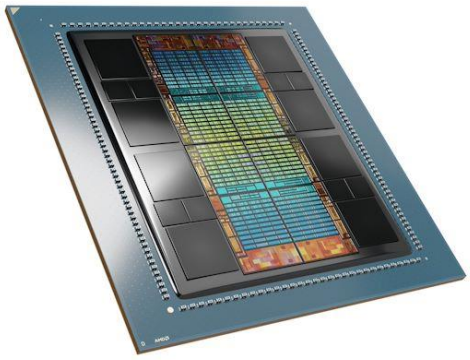


资料来源：CES 2023、华泰研究

MI300 系列目前包括两种产品：MI300X 是纯 GPU 产品，由 12 个 chiplets（8 个 GPU+4 个 IO+Cache）构成，专为生成式 AI 和大型语言模型（LLM）而设；对比 MI300A 由 13 个 chiplets（6 个 GPU+3 个 CPU+4 个 IO+Cache）构成。MI300X 作为纯 GPU 产品对标 H100，而 MI300A 为 APU 架构（Zen 4 CPU + CNDA 3 GPU），与英伟达的 GH200 正面交锋。

MI300 系列在参数上值得关注的亮点包括： 1）MI300X 的 192GB HBM3 内存，领先英伟达 H100 双卡 NVL 的 188GB HBM3，更远超 H100 PCIe 和 SMX 的 80GB HBM3，MI300A 的 128GB HBM3 内存也不遑多让；2）MI300X 晶体管数量 1530 亿，MI300A 晶体管数量 1460 亿，远超 H100 的 800 亿；3）内存带宽 5.2TB/s 与英伟达 H100 的 2-7.2TB/s 相近；4）Infinity Fabric 互联带宽的 896GB/s 与 NVLink 的 900GB/s 也相差无几；5）比 H100 高 2.4X 的 HBM 密度，以及 1.6X HBM 带宽。我们认为，MI300A 和 X 是客户在英伟达 GPU 之外的有力选择，或也可对 AI 芯片定价造成一定影响。

图表322： AMD MI300X 产品实物图（图中可见共 8 个 GPU chiplets）



资料来源：AMD 数据中心&AI 首映式、华泰研究

图表323： 搭载 8 个 MI300X 的 Instinct Platform

资料来源：AMD 数据中心&AI 首映式、华泰研究

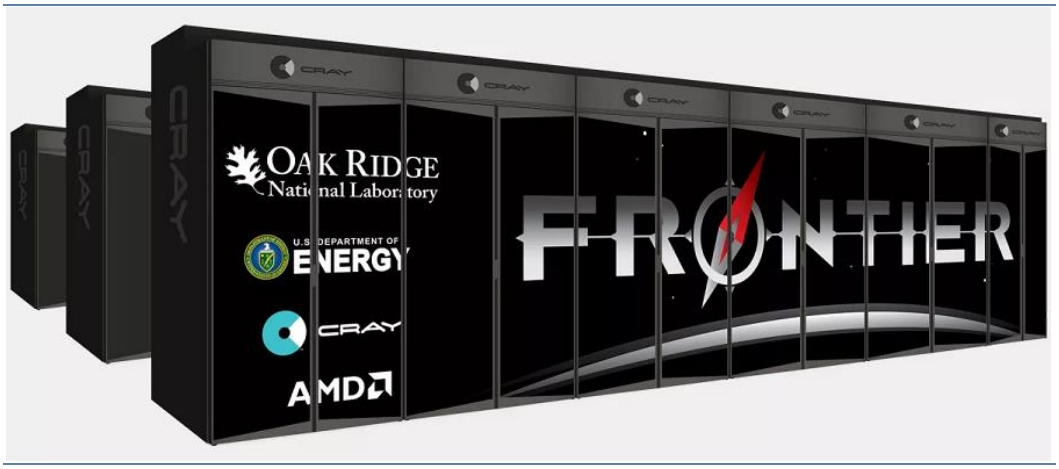
AMD MI300X 在 400 亿参数的 Falcon 模型上进行实时推理的能力，公司称这是此等参数规模的大语言模型第一次在单 GPU 上运行。公司表示 MI300X 还可运行规模更大的模型，比如 Meta 的 OPT 模型（660 亿参数版本）和 LLaMA（650 亿参数），如果使用 FP16 精度在单一 MI300X 上推理，最高可以运行 800 亿参数的模型。

图表324： 单片 MI300X 运行 400 亿参数的 Falcon 模型

资料来源：AMD 官网、华泰研究

2022 年惠与制造了搭载 AMD MI250X 的超级计算机 Frontier，这是世界上第一台 exascale 百亿亿次级的超级计算机；2023 年惠与制造了搭载 AMD MI300 的超级计算机 EL Capitan。

图表325：橡树岭国家实验室 Frontier 超级计算机搭载 AMD MI250X



注：图中 Cray 为 HPE 惠与公司 2019 年收购的超级计算机制造商
资料来源：HPE 惠与公司官网，华泰研究

图表326：劳伦斯利弗莫尔实验室 EL Capitan 超级计算机搭载 AMD MI300

资料来源：HPE 惠与公司官网、华泰研究

AMD 和英伟达在异构 AI 芯片中所用 CPU 架构各有千秋

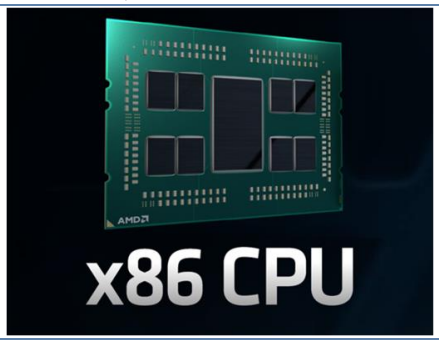
MI300 中的 Zen CPU 为 x86 架构，英伟达 GH200 中 Grace CPU 则采用 ARM 架构。如前文所述，我们认为 ARM 和 x86 CPU 架构在 AI 应用中各有优势。ARM 架构相比 x86 能耗较低，在能耗不断提升的 AI 应用中或较合适，且在这类 CPU+GPU 架构里的 CPU 或仅需发挥其部分功能，如向 GPU 发出指令等，因此 ARM 架构或已足够。而 x86 架构追求高性能和拥有丰富的指令集，在 AI 推理里也可跟 GPU 在功能上作互补。

AMD 虽同时拥有 ARM 和 x86 架构，但我们认为其在 MI300 里采用 x86 架构的 CPU，主要鉴于以下 3 点：1）AMD 在基于 x86 架构的 APU 设计已有深厚的积累；2）x86 性能高，指令集丰富，可跟 GPU 在 AI 推理时在功能上有效互补；3）AMD 采用的 3D Chiplet 封装，可通过更紧密的元器件间距，缩短传输距离，减少能耗，能一定程度弥补 x86 架构相较于 ARM 在能耗方面的不足。反观，英伟达在 GH200 里使用 ARM 架构的 CPU，我们认为主要鉴于 AI 应用和数据中心均追求低能耗。另外，在这类架构中，CPU 更多用于发布指令、协调控制等，因此 ARM 架构已经足够；4）AMD 的 CPU 从设计层面本来就更注重考虑能耗，这也是其能抢夺英特尔市占率的基础，ARM 架构能带来的能耗优势或并不足以让 AMD 放弃积累丰厚的 x86。

图表327： 英伟达 Grace Hopper 中的 Grace CPU 采用 ARM 架构

资料来源：AnandTech 官网、华泰研究

图表328： AMD MI300 中的 Zen 4 CPU 采用 x86 架构



资料来源：Wccftech 官网、华泰研究

AMD 和英伟达封装方案对比：AMD 发明 Chiplet 技术，原生优势凸显

从封装方案看，MI300 系列使用了台积电 SoIC（3D）和 CoWoS（2.5D）两种封装技术，相比英伟达的 H100 和 GH200 暂只采用 CoWoS 封装技术。我们认为，AMD 发明 chiplet 技术，又与台积电合作开发 3D Chiplet，在先进封装方面有丰富的原生优势。

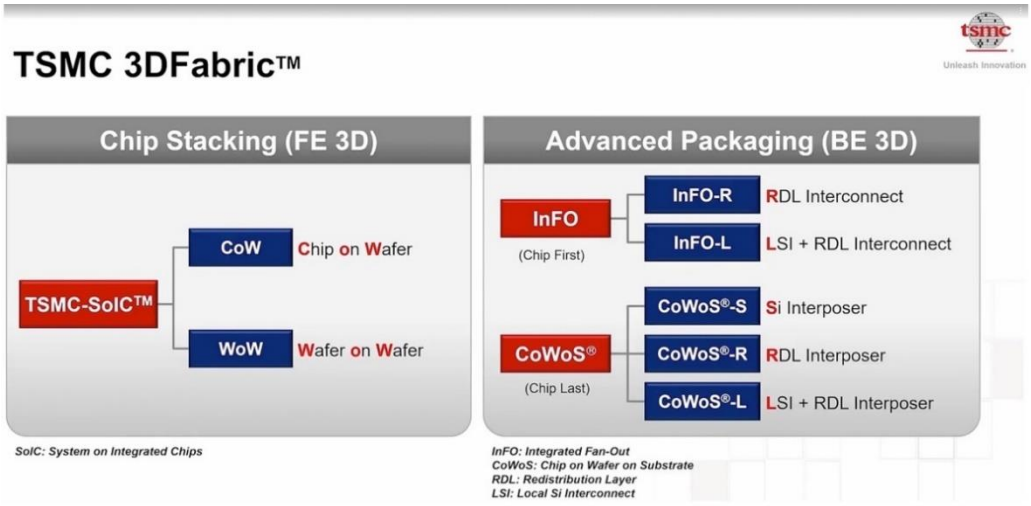
Chiplet 技术本身由 AMD 于 2015 年发明。2021 年，AMD 宣布与台积电合作开发 3D Chiplet，因此，相对于英伟达，AMD 在 Chiplet 方面应具备更丰富的 Know-how 经验。3D Chiplet 在提高性能及降低能耗方面有一定优势，另外相比 H100 和 GH200 采用 CoWoS 2.5D 的工艺，AMD 3D Chiplet 封装也进一步缩小了单元之间的距离。根据 DigiTimes Asia 在 2023 年 4 月 7 日的报道，英伟达预计将在 2024-25 年左右开始采用台积电 3D SoIC，因此，英伟达在积累一定的 Know-how 后，未来或也会采用 3D 封装技术。

MI300 是 MI 系列中第一个采用 3D V-Cache 技术的产品。AMD 在 2021 年推出的 3D V-Cache 封装技术是将缓存层堆叠到 CPU 上，也是一种 3D 封装技术。PC 方面，2022 年推出的 Ryzen 7 5800X3D 是首个使用 3D V-Cache 的产品，2023 年 Q1 推出的 Ryzen 7000X3D 系列也采用该技术。服务器方面，从第三代 EPYC CPU Milan-X 开始采用 3D V-Cache 技术，而即将推出的第四代 EPYC CPU Genoa-X 也将采用该技术。

Chiplet 技术在 IC 设计阶段将大型芯片按照不同功能模块的独立芯粒分开进行制造，可有效减小芯片的面积，提高良率和降低生产成本。在 Chiplet 工艺的基础上进行 3D 封装，一方面可进一步提升封装密度（在同样的面积上铺设更多数量的元器件），减少芯片整体面积；另一方面，3D 封装通过采用更细小、更密集的电路，缩短线路之间的传输距离，能提升指令的响应速度和降低功耗。然而，虽然 3D 封装通过缩短传输距离降低能耗，但由于该技术需在较小的封装体内堆叠多个芯粒，或会导致其存在散热问题，这也是 3D 封装技术领域正致力于攻克的主要问题。对此，AMD 通过对架构和硅平面的优化来缓解散热问题，例如 AMD 在 SRAM 单元中设计了 3D 64MB SRAM（Static Random Access Memory），并避免其堆叠以保持低热密度，也采用了结构硅从高密度核心中逸出热量。

3D Chiplet 封装技术具有提高性能、降低延迟和功耗的优点，也可弥补 x86 架构相较 ARM 架构在 AI 应用中的能耗问题。在 MI300A 中，有 6 个 GPU、3 个 CPU 和 4 个 I/O+cache 共 13 个 chiplets；而在 MI300X 中，有 8 个 GPU 和 4 个 I/O+cache 共 12 个 chiplets。MI300A 将 13 个 chiplets 分为两层：上层为 9 个基于 5nm 制程的 chiplets （据 PCgamers 推测，包括 3 个 CPU 和 6 个 GPU），而下层为 4 个基于 6nm 制程的 I/O+cache chiplets，芯片两侧围绕 8 个总计 128GB 容量的 HBM3 高带宽内存。MI300X 与 MI300A 相比，去掉了 3 个 CPU，换为 2 个 GPU，并将 HBM3 内存从 128GB 提高到了 192GB。

图表329： 台积电先进封装 3DFabric



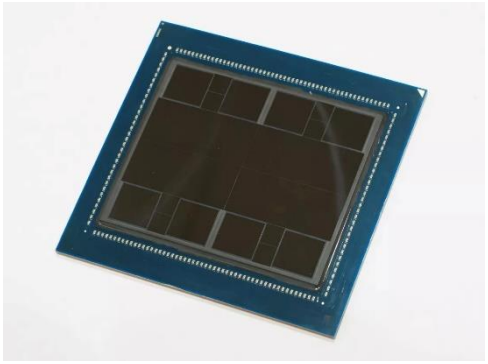
资料来源：台积电官网、华泰研究

图表330： 英伟达及 AMD 主要 GPU 相关产品参数对比

	英伟达		AMD		
产品名称	A100 PCIe丨SXM	H100 PCIe丨SXM丨NVL	MI250X	MI300A	MI300X
发布时间	2020.6	2022.3	2021.11	2023.1	2023.6
峰值算力（TFLOPS）	FP16: 312丨624 FP32: 19.5 FP64: 19.5	FP8: 3,026丨3,958丨7,916 FP16: 1,513丨1,979丨3,958 FP32: 51丨67丨134 FP64: 51丨64丨134	FP16: 383 FP32/64: 47.9 FP32/64 Matrix: 95.7	暂无	暂无
工艺制程及芯片面积	7nm, 826mm^2	4nm, 814 mm^2	6nm, 724 mm^2	5nm, 1017 mm^2	5nm, 1017 mm^2
晶体管数量（亿）	540	800	582	1460	1530
内存容量	80 GB HBM2e	80丨80丨188 GB HBM3	128 GB HBM2e	128 GB HBM3	192 GB HBM3
内存带宽	1,935丨2,039 GB/s	2丨3.35 TB/s丨7.8TB/s	3.2 TB/s	暂无	5.2TB/s
Interconnect	600 GB/s NVLink for 2 GPUs 64 GB/s PCIe Gen4	600丨900丨600 GB/s NVLink 125 GB/s PCIe Gen5	100GB/s	约 800GB/s	896GB/s
热设计功耗 TDP（W）	300丨400	300-350丨700丨2x350-400	500	600	暂无

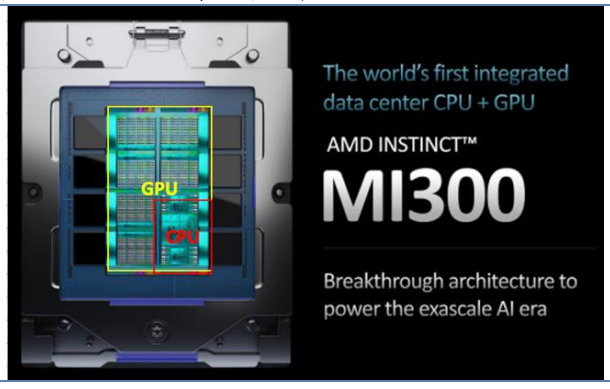
资料来源：AMD 官网、英伟达官网、华泰研究

图表331： AMD MI300 产品实物图



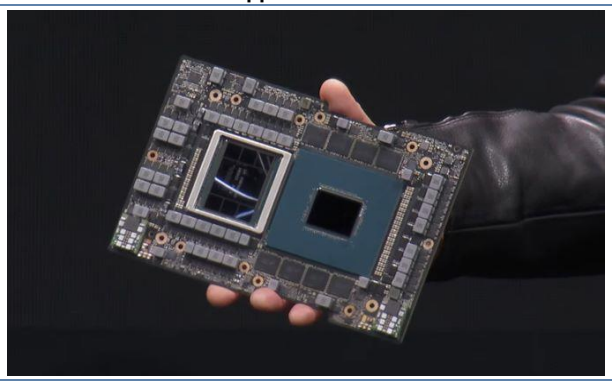
资料来源：Tom's Hardware 官网、华泰研究

图表332： AMD MI300 产品示意图



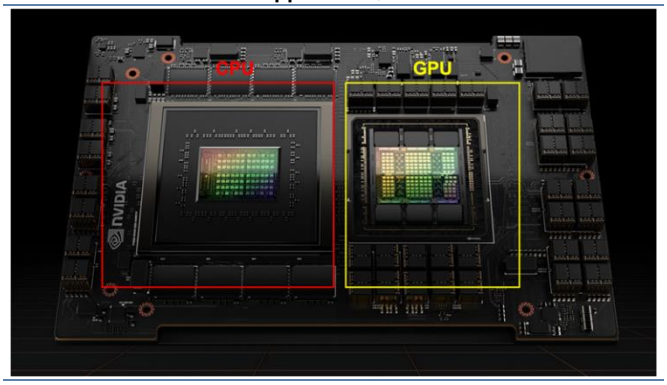
资料来源：CES 2023、华泰研究

图表333： 英伟达 Grace Hopper 芯片实物图



资料来源：AnandTech 官网、华泰研究

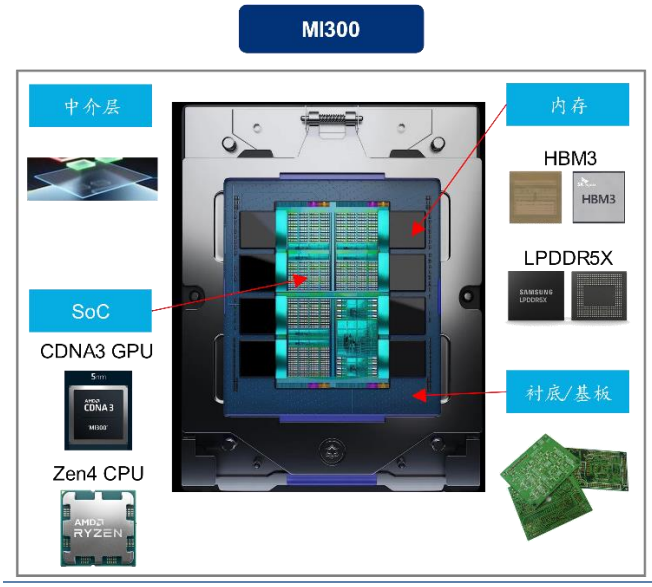
图表334： 英伟达 Grace Hopper 芯片中同时包含 CPU 和 GPU



注：对照英伟达官网 Grace CPU 和 Hopper GPU 照片，我们推测左侧为 CPU，右侧为 GPU

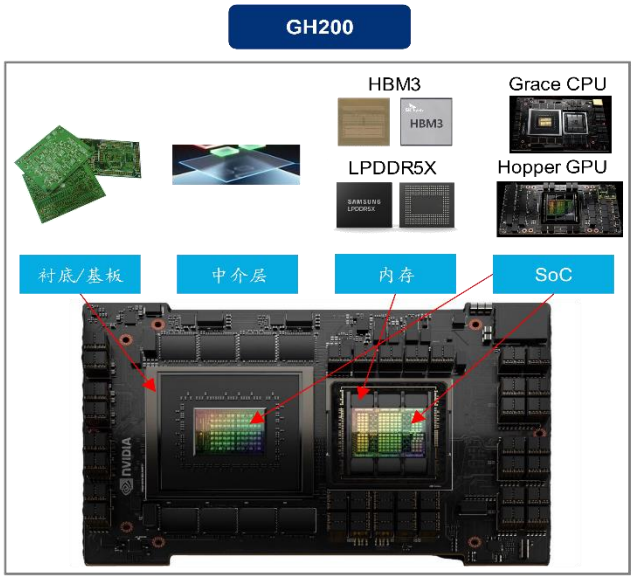
资料来源：英伟达官网、华泰研究

图表335： MI300 芯片结构示意图



资料来源：AMD 官网、华泰研究

图表336： GH200 芯片结构示意图

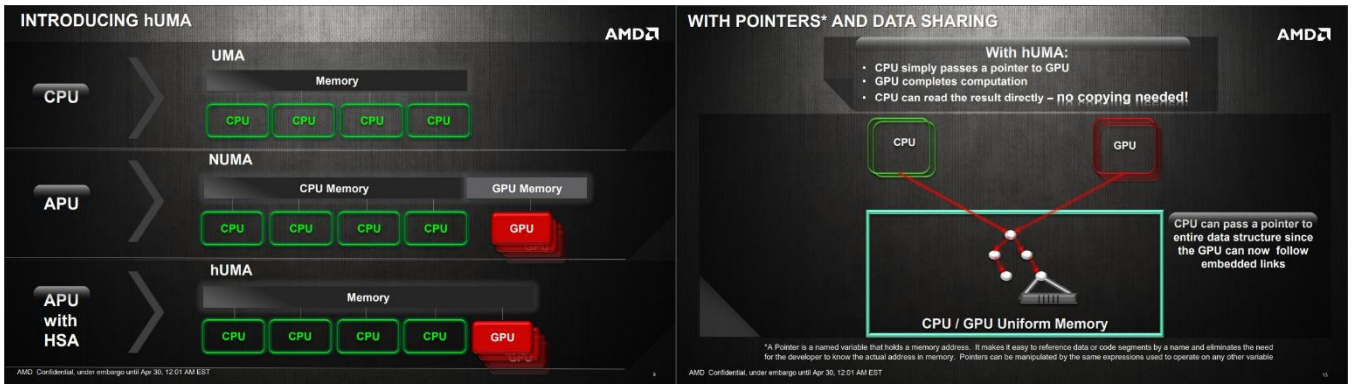


资料来源：英伟达官网、华泰研究

AMD 统一内存架构与英伟达 NVLink

我们认为，AMD MI300 的统一内存架构优点在于芯片内的数据传输性能较高效，可绕过一般传输协议速率限制，加快 CPU-GPU 之间数据交流，也可降低不同内存间来回复制、同步和转换数据所导致的编程复杂性，实现单芯片上更高性能。而英伟达 GH200 里的 NVLink 本质上仍是通过通信协议/链路实现 CPU 和 GPU 之间的数据传输，受限于协议速率，不过可实现 PCIe 5.0 的 7 倍带宽。在芯片以外，NVLink 支持跨节点多 GPU 间的数据访问与高速传输，达到较大的内存容量和较强的扩展性；而 MI300 则将采用第四代 Infinity Fabric 架构，具体情况在下文解释。

图表337： AMD 早在 2013 年便提出 hUMA（异构统一内存访问）架构，希望实现 CPU 与 GPU 共用同一内存

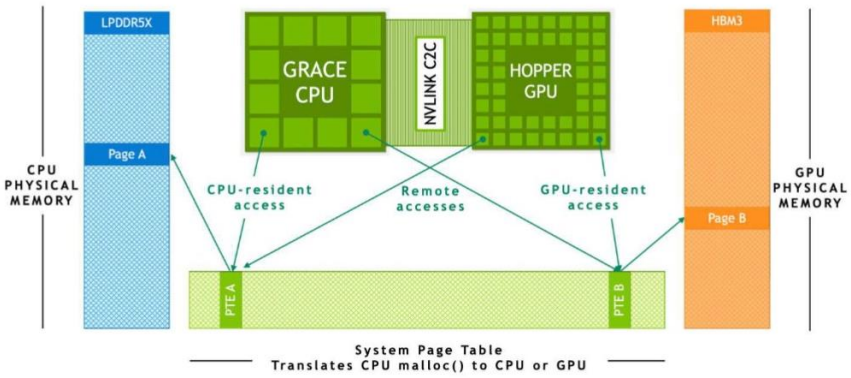


资料来源：AMD 官网、华泰研究

MI300 通过统一内存架构（CPU 与 GPU 使用同一个内存空间）实现两者在物理上的内存共享。此概念最早可追溯到 AMD 在 2013 年提出的 hUMA（Heterogeneous Uniform Memory Access，异构统一内存访问）。CPU 和 GPU 可针对相同数据同时展开计算，实现“zero-copy”（即执行计算时 CPU 无需先将数据从某处内存复制到另一个特定内存区域），越过一般传输协议的速率限制，也可降低 CPU 和 GPU 共享数据中必须在两个内存分配、来回复制、同步和转换所导致的编程复杂性。一般来说，GPU 在进行数据处理时，需要将数据从 CPU 内存复制到 GPU 内存，并在处理完成后再次返回。此举通常由 CPU 执行，因此既占用 CPU 性能，又受限于中间传输协议的带宽。据 MI300 客户劳伦斯利弗莫尔国家实验室在 2023 年 5 月 22 日的 ISC 2023 大会上表示，MI300 统一的内存层可简化编程，降低了不同类型计算和不同内存池之间数据移动的复杂性，从而减少延迟并提高性能和功效。

NVLink 通过地址转换服务（ATS, address translation service）使 CPU 和 GPU 可访问对方的内存，即达成虚拟意义上的内存统一。同时，为解决软件编程难度，英伟达在 CUDA 6.0 引入了统一内存模型，系统会自动在 GPU 和 CPU 之间迁移在虚拟统一内存中分配的数据，简化数据分配操作。但整体上看，尽管相比 PCIe 而言，NVLink 具备更高带宽（实现了 PCIe 5.0 总线的 7 倍带宽），但 NVLink 本质仍是通过协议传输，数据仍需在两者内存间分配和复制，与 AMD 的统一内存架构物理上内存统一、无需协议传输和数据复制同步的方法仍有所不同。

图表338： 英伟达 Grace Hopper 芯片通过地址转换服务（ATS）实现统一的虚拟内存

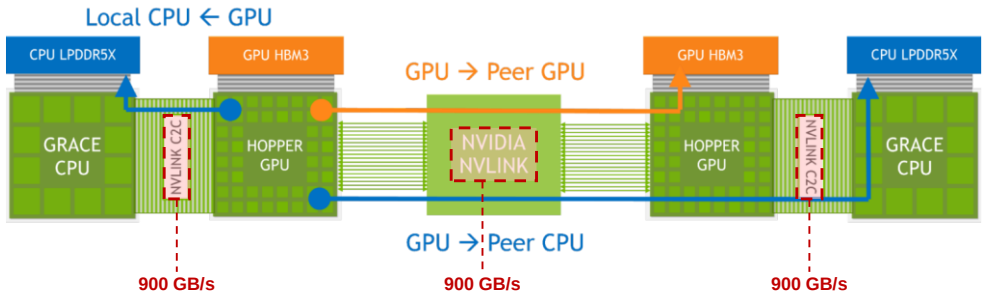


资料来源：英伟达官网、华泰研究

芯片以外，NVLink 可支持跨节点多 GPU 间的数据访问与高速传输，具备较高的内存容量和较强的扩展性。而 MI300 所用的第四代 Infinity Fabric 方案尚未发布。据 AMD 于 2023 年 5 月 13 日在 ITF World 2023 上公布，MI300 将使用第四代 Infinity Fabric 总线架构进行跨节点传输，目前虽暂无第四代 Infinity Fabric 的更多信息，但在 MI250X 上使用的第三代 Infinity Fabric 已支持 800GB/s 的 GPU 互连，第四代或能实现更高速率。从扩展性来看，第三代架构最多可实现 8 个 GPU 互联，第四代架构或为对比 AMD 和英伟达两者内存和扩展性的关键。

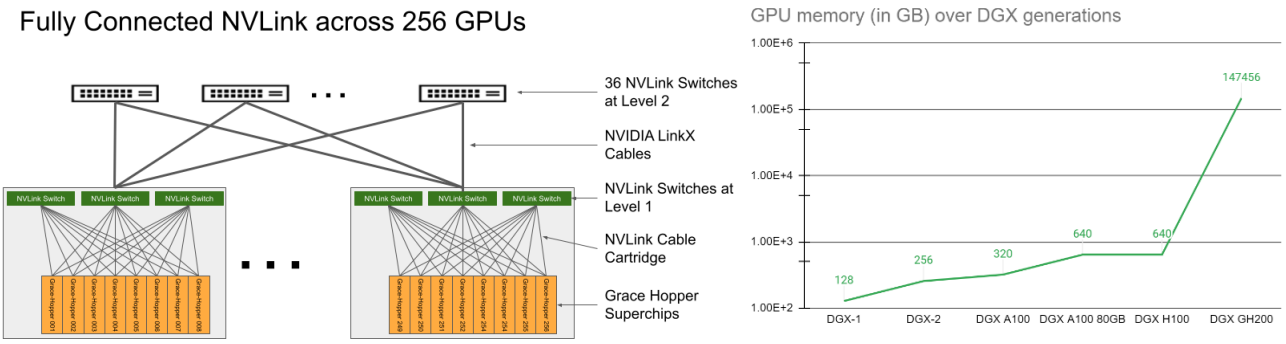
而英伟达通过其独家的 NVLink 连接实现较强扩展性，允许多芯片高速互联，构成高带宽的超级计算系统，解决了因数据处理量增大而 GPU 本地内存容量无法满足存储的痛点。比起传统的替代方案（即通过 PCIe 总线从 CPU 系统内存中提取数据），NVLink 可提供高带宽和低延迟的数据传输，保持 GPU 以最高性能运行。据英伟达在 2023 年 5 月 29 日上的 Computex 2023 大会公布，DGX GH200 超算通过 NVLink 互连技术及 NVLink Switch 系统串联由 32 台 8 块 GH200 超级芯片组成的系统（pod），将总计 256 块 GH200 芯片合并成单一超级计算机。故 NVLink 使 DGX GH200 中的每个 GPU 都能以最高 900 GB/s 的速度任意访问其他 GPU 和 CPU 的内存，实现总计高达 144 TB 内存。单片 MI300X 配置的显存为 182GB HBM3，MI300A 为 128HBM3，而 Grace Hopper 芯片中的 96GB HBM。

图表339： 英伟达 Grace Hopper 通过 NVLink-C2C 和 NVLink Switch 实现 CPU-GPU 和 GPU-GPU 互连



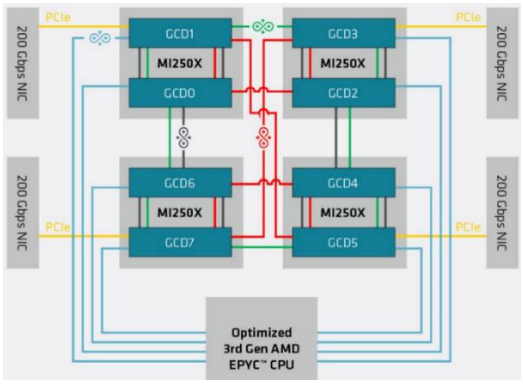
资料来源：英伟达官网、华泰研究

图表340： NVLink Switch 系统通过两级 NVLink 结构，使得单个 DGX GH200 系统可连接 256 个 GH200 芯片，实现高达 144 TB 内存的容量



资料来源：英伟达官网、华泰研究

图表341： MI250X 通过第三代 Infinity Fabric 实现 800GB/s 的 GPU 互联（红线、绿线、灰线和蓝线为 Infinity Fabric）

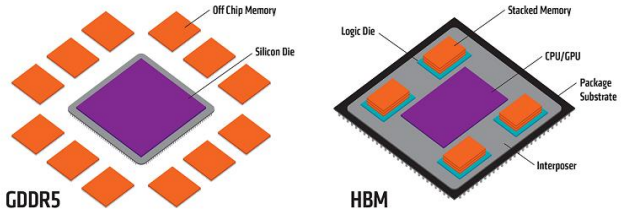


资料来源：AMD 官网、华泰研究

MI300A 与 GH200 中使用内存：HBM3e、HBM3，LPDDR5X、DDR5

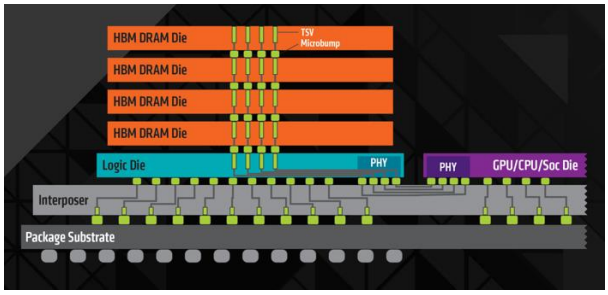
HBM 是 RAM（Random Access Memory，随机存取存储器）的一种，是海力士和 AMD 合作开发的 3D 堆叠工艺的高带宽存储器，因为其更高的带宽，适合应用于 GPU。值得一提的是，第一款使用 HBM 的产品就是 AMD 的 Radeon Fury 系列 GPU。

图表342： HBM 与 DDR 的区别



资料来源：SK 海力士官网、华泰研究

图表343： AMD 和 SK 海力士合作设计了



资料来源：华泰研究

GH200 和 MI300 都使用 HBM3 系列内存。回顾在 2023 年 6 月时，HBM3 的唯一生产商是 SK 海力士，但在 2023 年 8 月 23 日，据 Digitimes 消息三星也将开始为 AMD 提供 HBM3，并已通过英伟达测评。这符合我们之前的预测，即为了满足英伟达和 AMD 拓展二供、三供的需求，三星将提速进入 HBM3 的供应商列表。此外，美光于 2023 年 7 月也发布了新的 HBM3 产品并开始同步送样。

2023 年 4 月 20 日，SK 海力士发布了业界目前最大内存（24GB）的 HBM3 产品，同时宣布已给客户送样；2023 年 6 月 1 日，海力士发布 HBM3e，其相比 HBM3 传输速率提高 25%（6.4GT/s 到 8.0GT/s），内存带宽从 819GB/s 提升到 1TB/s。而英伟达在 2023 年 8 月 8 日发布了使用 HBM3e 版本的 GH200。

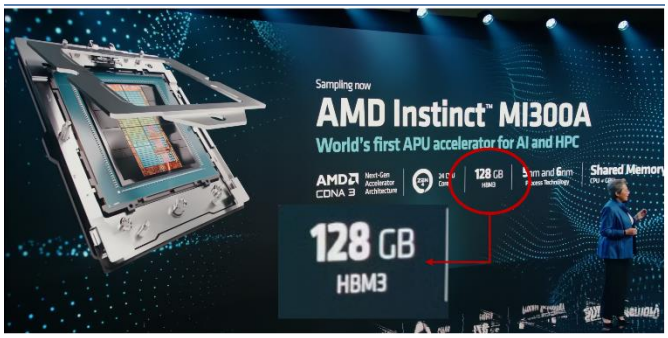
我们通过计算知道，AMD MI300X 的 192GB HBM3（24*8=192），MI300X 搭载的 HBM3 或即是 24GB 的海力士 HBM3 产品；而 MI300A 的 128GB HBM3（16*8=128）则或对应的是海力士于 2022 年 6 月发布的 16GB 的 HBM3 产品。Hopper GPU 的 96GB HBM3 也同样对应 16GB 的 HBM3（16*6=96）。据中国台湾《科技新报》2023 年 4 月 18 日报道，SK 海力士在 2022 年 6 月发布 16GB HBM3 的同时就立即供给英伟达。

除三家内存巨头外，Rambus Inc 作为老牌内存厂商，早在 2021 年就发布了 HBM3 产品，并同时称将在 2022 年末到 2023 年初流片量产，但其从发布后直到目前都没有关于 HBM3 产品的新消息流出，截至目前或已失去参与 HBM3 产品的意图。

图表344： MI300X 使用 HBM3 内存

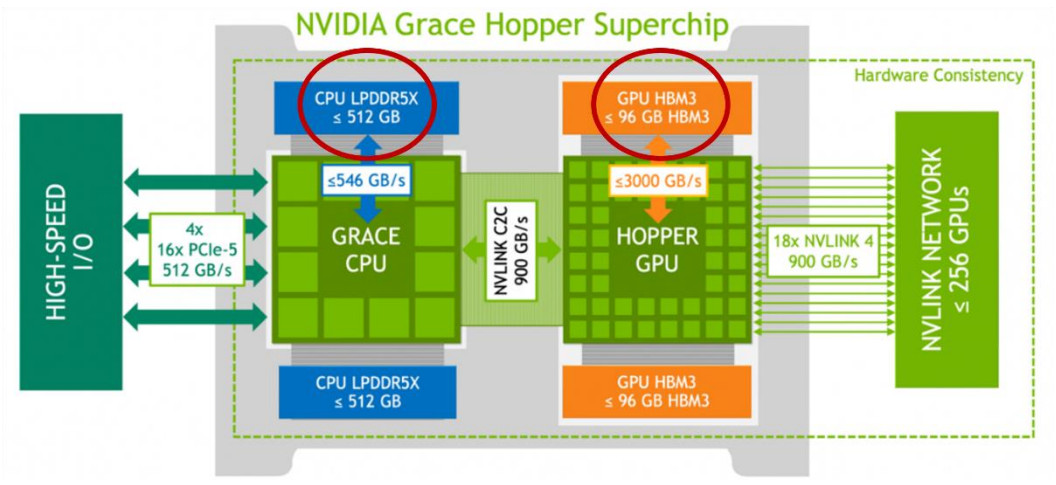
资料来源：AMD 官网、华泰研究

图表345： MI300A 使用 HBM3 内存



资料来源：ADM 官网、华泰研究

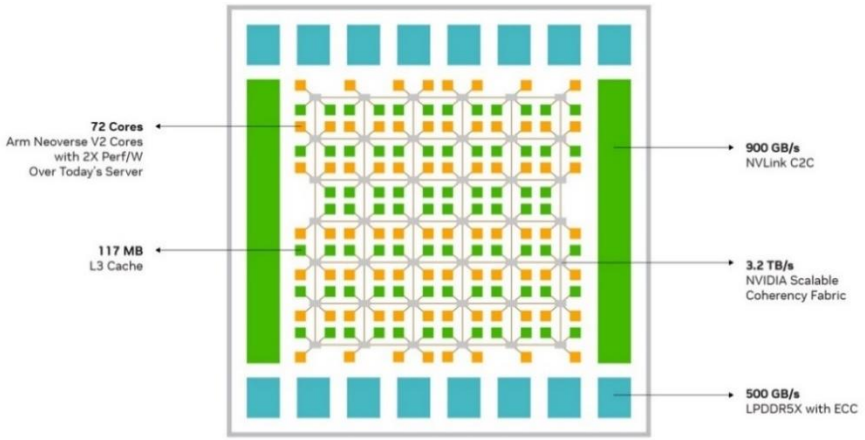
图表346： 英伟达 Grace Hopper 中 CPU 对应内存 LPDDR5X（512GB），GPU 对应内存为 HBM3（96GB）



资料来源：英伟达官网、华泰研究

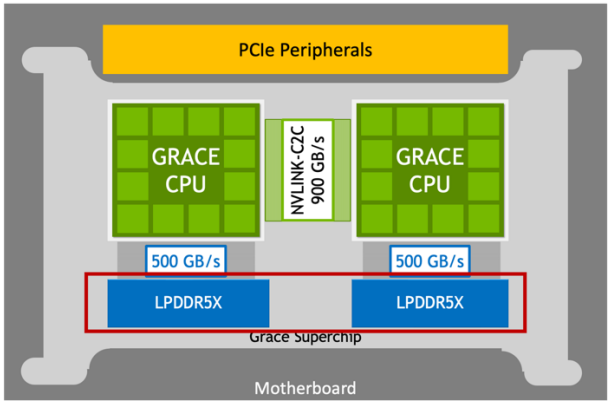
AMD MI300A 中使用的 Zen 4 CPU 的内存是 DDR5；而据英伟达 Grace CPU 白皮书，GH200 中的 Grace CPU 则选用了 LPDDR5X。DDR 内存即双倍数据率（Double Data Rate）内存，是 DRAM（Dynamic Random Access Memory）中占比最大的种类，负责暂存 CPU 的运算数据，并与外部进行数据交换，DDR5 是 DDR 最新的第五代。LPDDR5X（Low Power DDR5X，低功耗 DDR5X，其中 X 指第二代的 LPDDR5）。正如其名，能效是 LPDDR5X 的主要优势，而根据 Grace CPU 白皮书，能效也是 Grace 选用 LPDDR5X 内存的主要原因，通过使用 LPDDR5X，可在低功耗的前提下保证 Grace CPU 的容量和带宽。DDR5 和 LPDDR5X 供应商也是 SK 海力士、三星和美光三家内存巨头。

图表347： 单颗英伟达 Grace CPU 的基本构成



资料来源：英伟达官网、华泰研究

图表348： Grace CPU 使用的 LPDDR5X 内存



资料来源：英伟达官网、华泰研究

图表349： Zen 4 CPU 使用的 DDR5 内存

AMD Ryzen 7000 Zen 4 Specifications

	Cache (L2+L3)	TDP / Max	Memory
Ryzen 9 7950X	80MB	170W / 230W	DDR5-5200
Ryzen 9 7900X	76MB	170W / 230W	DDR5-5200
Ryzen 7 7700X	40MB	105W / 142W	DDR5-5200
Ryzen 5 7600X	38MB	105W / 142W	DDR5-5200

资料来源：AMD 官网、华泰研究

图表350： HBM2e VS. DDR5 VS. LPDDR5X

	HBM2e (4-sites)	DDR5 (8-channel)	LPDDR5x (32-channel)
Capacity	64GB	Up to 4TB	Up to 512GB
BW	Up to 1.8TB/s	Up to 358GB/s	Up to 546GB/s
Power/GB	1x	8x	1x
Cost/GB	>3x	1x	1x

资料来源：英伟达官网、华泰研究

AMD 与英伟达面临的共同境况之一：芯片出口受限

在中国市场方面，AMD 和英伟达的高算力 AI 芯片的出口均受限，MI300 和英伟达 A100、H100 等面临的境况相似：根据美国商务部 2022 年 10 月宣布的出口管制更新（107 新规），目前美国对芯片出口的限制条件是：I/O 传输接口传输速率大于或等于 600GB/s，且总算力超过 4800 TOPS（INT8 超过 600TOPS 或 FP16 超过 300TFLOPS）；当中以传输速率为前提，如传输速度低于 600GB/s，就算总算力超过 4800TOPS 也不受限。因此，英伟达和 AMD 或需降低产品的传输速率才能符合供应中国市场的条件。反过来，若传输速率上不满足条件，即使降低算力也依然属于受限范畴。

107 新规从 2022 年 10 月开始生效，但英伟达在 2022 年 9 月已对此表示，其获得美国政府批准，可在 2023 年 3 月前为美国客户提供所需的 A100 出口，9 月前可以继续通过其香港办事处履行 A100 和 H100 的订单和物流，也就是说，A100 和 H100 可向中国运输至 2023 年 9 月，可供货时间覆盖了 2023 年的 3/4。

以此前英伟达和 AMD 受限的产品举例来说，英伟达 A100 传输速率为 600 GB/s，传输速率超过限制。MI250 传输速率为最高 800GB/s，也超过 600GB/s。H100 和 MI250X 在算力和传输速率上大于或等于前序产品，因此也受限。英伟达为 A100 和 H100 分别提供可出口的 A800 和 H800。A800 将 NVLink 传输速率从 600GB/s 降低到 400GB/s；而据路透社 2023 年 5 月 22 日消息，H800 则将传输速率降至约 H100 的一半左右。

我们推断，英伟达和 AMD 的两款 AI 异构芯片也应落入管制范畴：从总算力限制角度看，GH200 和 MI300 均超过限制：1）GH200 内使用了 H100，以 H100 Tensor Core 在 FP16 精度上的算力为例，非稀疏性算力是 990 TFLOPS，超过 300 TFLOPS。2）MI300 的具体参数尚未公布，但它的前序产品 MI250X 的 FP16 算力为 383 TFLOPS，超过 300 TFLOPS，因此 MI300 或也受限。另外，GH200 和 MI300 的传输速率均落入限制范围：1）GH200 的传输速率最高可达 900GB/s，远高于 600GB/s；2）MI300 的具体参数尚未公布，但它的前序产品 MI250 和 MI250X 的最高传输速率均达 800GB/s，也超过了 600GB/s。

2023 年 8 月 1 日，AMD 正式宣布将考虑推出中国特供版芯片。Lisa Su 在 2023 年二季度财报电话会上表示，此前英伟达推出 H800 和 A800 以避开 107 新规，AMD 也在认真思考 MI300 和 MI250 的类似策略。

图表351：2022 年 10 月美国商务部更新的限制条件（107 新规）

List of Items Controlled

Related Controls: See ECCNs 3D001 and 3E001 for associated technology and software controls.

Related Definitions: N/A

Items:

a. Integrated circuits that have or are programmable to have an aggregate bidirectional transfer rate over all inputs and outputs of 600 Gbyte/s or more to or from integrated circuits other than volatile memories, and any of the following:

a.1. One or more digital processor units executing machine instructions having a bit length per operation multiplied by processing performance measured in TOPS, aggregated over all processor units, of 4800 or more;

a.2. One or more digital `primitive computational units,' excluding those units contributing to the execution of machine instructions relevant to the calculation of TOPS for 3A090.a.1, having a bit length per operation multiplied by processing performance measured in TOPS, aggregated over all computational units, of 4800 or more;

a.3. One or more analog, multi-value, or multi-level `primitive computational units' having a processing performance measured in TOPS multiplied by 8, aggregated over all computational units, of 4800 or more; or

a.4. Any combination of digital processor units and `primitive computational units' whose calculations according to 3A090.a.1, 3A090.a.2, and 3A090.a.3 sum to 4800 or more.

资料来源：美国商务部、华泰研究

图表352：英伟达及 AMD 过往部分受影响产品

产品名称	A100 SXM	H100 SXM	MI250	MI250X
峰值算力（TFLOPS）	FP16: 312 | 624* FP32: 19.5 FP64: 9.7 | 19.5*	FP8: 1979 | 3,958* FP16: 990 | 1,979* FP32: 67 FP64: 34 | 67*	FP16: 362.1 FP32/64 Matrix: 90.5 FP32/64 Vector: 45.3	FP16: 383 FP32/64 Matrix: 95.7 FP32/64 Vector: 47.9
工艺制程及芯片面积	7nm, 826mm^2	4nm, 814 mm^2	6nm, 724 mm^2	6nm, 724 mm^2
晶体管数量（亿）	540	800	582	582
内存容量	80 GB HBM2e	80 GB HBM3	128 GB HBM2e	128 GB HBM2e
内存带宽	2,039 GB/s	3.35 TB/s	3.2 TB/s^2	3.2 TB/s^2
互联	600 GB/s NVLink for 2 GPUs 64 GB/s PCIe Gen4	900 GB/s NVLink 125 GB/s PCIe Gen5	Infinity Fabric Links 最高 800 GB/s	Infinity Fabric Links 最高 800 GB/s
热设计功耗 TDP（W）	400	700	500	500

注：*代表稀疏性计算情况下

资料来源：AMD 官网、英伟达官网、华泰研究

图表353：限制文件中关于算力转换的参考：条件为 INT8 超过 600TOPS 或 FP16 超过 300TFLOPS

4. The rate of TOPS is to be calculated at its maximum value theoretically possible when all processing elements are operating simultaneously. The rate of TOPS and aggregate bidirectional transfer rate is assumed to be the highest value the manufacturer claims in a manual or brochure for the integrated circuit. For example, the threshold of 4800 bits x TOPS can be met with 600 tera integer operations at 8 bits or 300 tera FLOPS at 16 bits. The bit length of an operation is equal to the highest bit length of any input or output of that operation. Additionally, if an item specified by this entry is designed for operations that achieve different bits x TOPS value, the highest bits x TOPS value should be used for the purposes of 3A090.a.

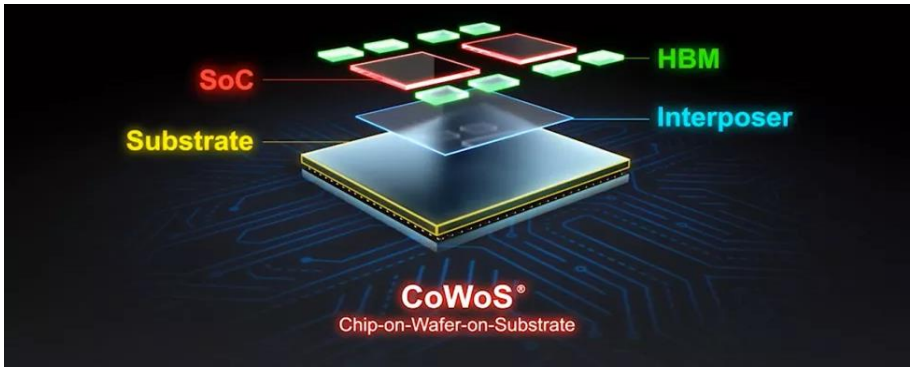
资料来源：美国商务部、华泰研究

AMD 与英伟达面临的共同境况之二：台积电 CoWoS 产能瓶颈

在市场对 AI 芯片的旺盛需求下，先进封装产能或将供不应求。我们认为，CoWoS 是限制 AI 芯片出货量的主要瓶颈之一。英伟达 H100 采用台积电 CoWoS（2.5D）封装技术，AMD MI300 采用台积电 CoWoS（2.5D）和 SoIC（3D）技术，二者都依赖台积电先进封装产能。由于当下 AI 芯片需求喷发，台积电的 CoWoS 产能已现不足。据 DigiTimes 在 7 月 14 日报道，台积电正扩增包括竹南、龙潭和台中三地在内的 CoWoS 产能，因此，2023 年台积电 CoWoS 产能至少 12 万片，2024 年可扩至 17.5 万片。之后在 7 月 25 日 DigiTimes 又称台积电扩产后全年 CoWoS 产能加总或将超 24 万片，当中英伟达能拿到约 15 万片。为了实现扩产，台积电或将把部分 oS（on Substrate）释放给其他封装厂商，订单或会外溢到包括中国台湾封测龙头日月光、矽品精密（2018 年被日月光控投收购）、台湾的联华电子、美国的 Amkor Technology、中国大陆的通富微电（6 月 27 日披露）等。目前，台积电 CoWoS 的三大客户包括：英伟达、博通和赛灵思；据上文报道，由于 MI300，AMD 在 4 季度后或将跻身前五大客户；而亚马逊在 2024 年则或将成为第三大客户。我们认为，对 AMD 来说，在英伟达的大量订单之下，获得台积电 CoWoS 产能排期至关重要。

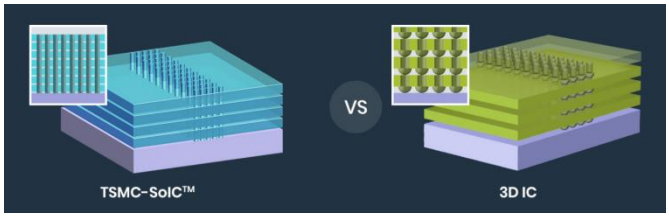
按照上文，台积电已在改装一些厂房来应对供不应求的 CoWoS 产能，我们认为，改装厂房其实还相对简单，CoWoS 的真正瓶颈之一，也许是封装用的机械设备所需的交货周期较长。据 DigiTimes 在 7 月 25 日报道，CoWoS 产能扩充缓慢的原因在于其使用的设备交货延迟，关键设备如研磨液供给设备（用于在晶圆的研磨过程中所需用于化学腐蚀液体的供给）、半导体清洁装置（同样为晶圆的湿制程设备）等，主要供应厂商包括日本的 Tazmo 和 Shibaura，设备的完整交货周期往往在 6-8 个月（6 个月交货+2 个月调试参数）。同样可以提供这类设备的还有如台湾本地供应商亚泰半导体，因此台积电或能通过其他供应商缓解一定压力。

图表354：台积电 CoWoS 封装示意图



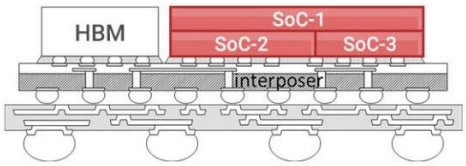
资料来源：台积电官网、华泰研究

图表355： TSMC-SoIC 与传统 3D IC 封装对比示意图



资料来源：TSMC、华泰研究

图表356： SoIC 封装的芯片可利用 CoWoS 封装整体集成到其他芯片（中介层下方为封装衬底）



资料来源：TSMC、华泰研究

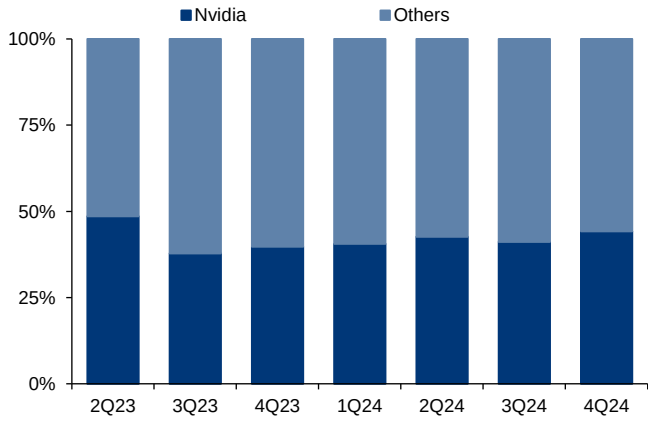
图表357： Tazmo 半导体制造设备产品：CoWoS 瓶颈或为研磨液供给系统设备

资料来源：Tazmo 官网、华泰研究

图表358： Tazmo 研磨液供给系统设备

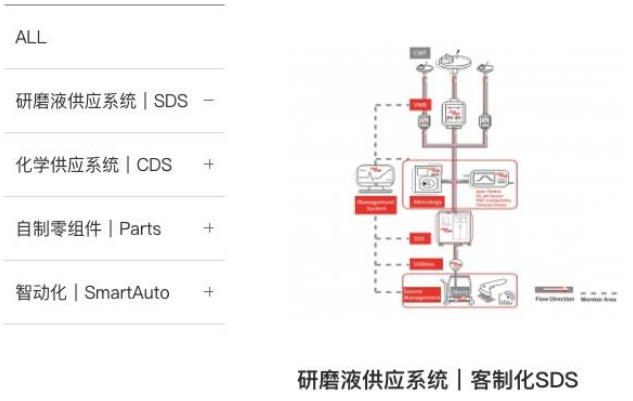
资料来源：Tazmo 官网、华泰研究

图表359： 英伟达在 CoWoS 季度 Output 的占比



资料来源：semianalysis 官网、华泰研究

图表360： 亚泰半导体的研磨液供应系统



资料来源：Tazmo 官网、华泰研究

ROCm 生态圈：AMD 的“阿克琉斯之踵”，分而治之或可解困

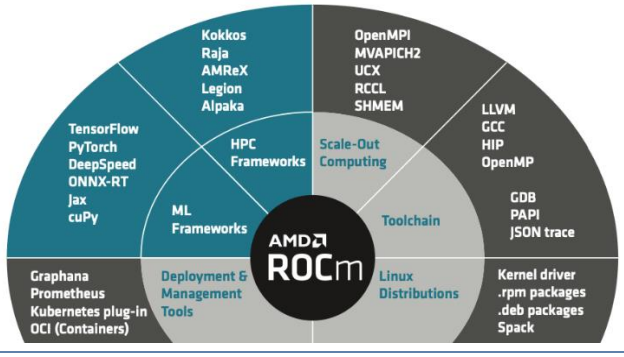
AMD 的软件生态圈 ROCm（Radeon Open Compute Ecosystem）于 2016 年 4 月发布，相比 2007 年发布的英伟达 CUDA 起步较晚。目前，ROCm 具备完全兼容 CUDA 的能力，为 AMD 提供了说服客户迁移的条件和理由，然而一味兼容只会导致 ROCm 受 CUDA 的掣肘，加上需应对 CUDA 的每一次更新迭代，或会导致 ROCm 陷入长期被动的局面，这使得生态圈已经成为 AMD 的“阿克琉斯之踵”。为了更有效地破解此困境，AMD 进行了三类努力：1）如上所述持续兼容 CUDA；2）继续完善 ROCm 生态圈；3）与云厂商分别进行直接合作，分而治之（divide and conquer）与 CUDA 脱钩。

目前，ROCm 有以下三点需完善：1）操作系统：长期只支持 Linux，在 2023 年 4 月才宣布登录 Windows；反观，CUDA 从 1.0 版就同时支持 Linux 和 Windows；**2）产品支持：**ROCm 长期只支持 AMD 的高端 GPU，例如 Radeon Pro 系列等，直到 2023 年 4 月才宣布扩展至一些消费级 GPU 如 Radeon RX 6900 XT、Radeon RX 6600、以及 Radeon R9 Fury；反观 CUDA，2006 年发布的 G80 系列及以后所有的英伟达 GPU 都能支持；**3）开发者数量和生态扩展：**CUDA 在 2023 年达 400 万以上的开发者，这是 ROCm 目前无法相比。我们发现，截至 2023 年 8 月 9 日，全球最大的开发者社区之一 StackOverflow 中“CUDA”的标签下已有超过 14000（14259）个问题讨论，而 ROCm 在“amd-rocm”的标签下只有 120 个问题讨论；在 Github 上，CUDA 已有超过 33400 个开发者贡献的软件包仓库，而 ROCm 只有不到 600（559）个。更多的开发者意味着不断迭代的工具和更广泛的多行业应用，因此 ROCm 需要更多的开发者以形成生态的良性循环。

为了破解“阿克琉斯之踵”，AMD 进行了三类努力：

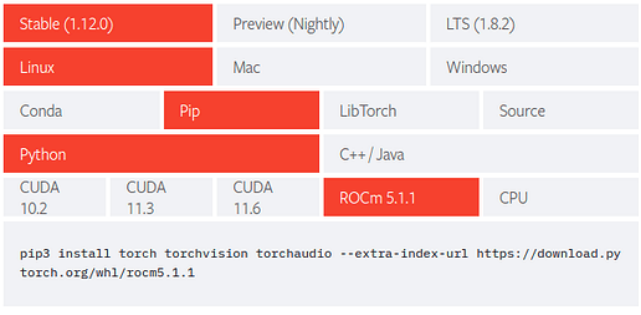
1）AMD 正积极拓展 ROCm 的生态圈：首先，ROCm 从 2023 年 4 月 14 日开始支持 Windows 操作系统，终于迈出了来迟的一步；其次，ROCm 在 AI 领域进行了更广泛的框架和软件兼容，已支持 TensorFlow 和 PyTorch 等主流机器学习框架，而且与英伟达和英特尔同属 PyTorch 的 Founding Member（PyTorch 在 2022 年 12 月成立的联盟，包括四大云厂商、三大芯片巨头等）一员；软件库方面，优化深度学习库 MIOpen 和计算机视觉和机器智能库 MIVisionX，PyTorch 2.0 也在 2023 年一季度开始提供对 ROCm 的支持。据 TechGoing 官网在 2023 年 9 月 3 日的报道，OpenAI 的 Triton（一种类似于 Python 的开源编程语言，用于编写 GPU 代码）已经开始合并 AMD ROCm 代码，结束 Triton 对英伟达 CUDA 的单一支持，未来或会有更多平台选择适配 AMD 硬件。

图表361：ROCm 支持主流机器学习框架



资料来源：AMD ROCm 手册 2022 版、华泰研究

图表362：PyTorch 中可以选择 ROCm



资料来源：AMD ROCm 手册 2022 版、华泰研究

丰富 ROCm 软件栈并非 AMD 的一厢情愿，AI 初创企业和 AI 开发者社区都愿为了更低的算力成本和更多的芯片选择而助 AMD 一臂之力。我们认为，AI 初创企业对于获取英伟达之外的其他可选算力意愿较高。2023 年 6 月 30 日 **Mosaic ML**（初创生成式 AI 公司，MIT 背景）发布了基于 ROCm 使用 AMD 的 MI250 GPU 进行大语言模型的训练日志，称其希望在“这个全由英伟达供应的世界里”提高选择性。日志中，MosaicML“无需转码（no code changes were needed）”基于 AMD 的 MI250 和 ROCm 实现了模型训练。更多类似的尝试将推动 ROCm 的边界向外拓展。2023 年 6 月 14 日 AMD 数据中心与人工智能发布会上，**HuggingFace**（人工智能开发者社区，开源共享模型和数据集，可认为是 AI 的 Github）宣布与 AMD 建立合作，这项合作的重点正是把 Hugging Face 的 Transformer 库集成进 ROCm 中，目的是让用户在 AMD 的芯片上训练和推理在库中的模型时无需其他操作，正如 Hugging Face CEO Clement Delangue 在会上直言“我们希望所有人都能在 AMD 的芯片上运行模型（we want everyone to be able to run their models on AMD hardware）”。

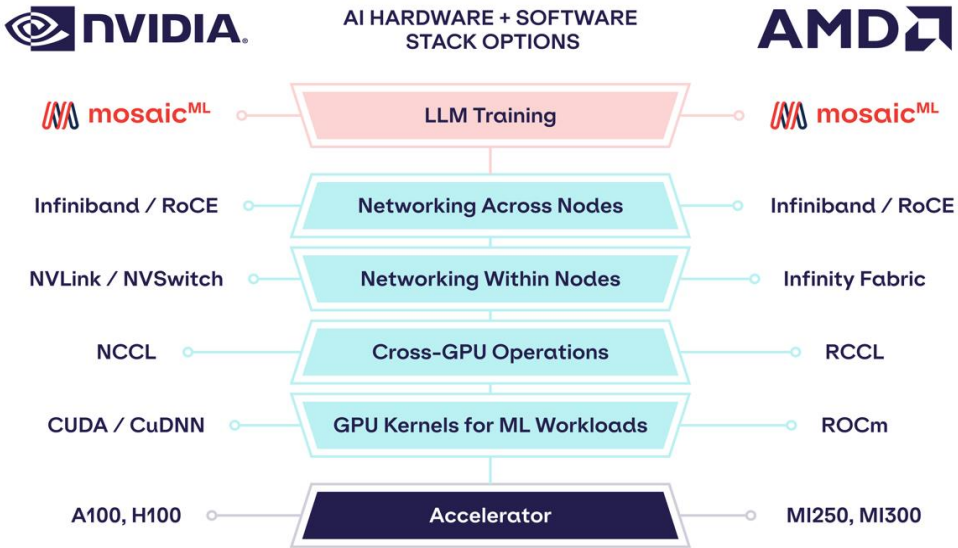
图表363： AMD 与 Hugging Face 的合作伙伴关系

资料来源：AMD 数据中心与人工智能首映式、华泰研究

图表364： AMD 与 Mosaic ML 的合作伙伴关系

资料来源：MosaicML 官网、华泰研究

图表365： MosaicML 希望能同时使用英伟达和 AMD 两套硬件+软件

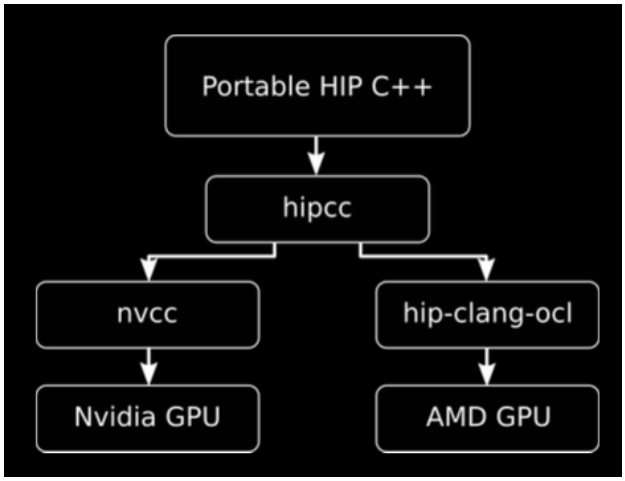


资料来源：MosaicML 官网、华泰研究

2）进一步兼容 CUDA：ROCm 可以通过 HIP（Heterogeneous-Computing Interface for Portability）完全兼容 CUDA。HIP 是 AMD 的 GPU 软件开发框架，它提供的 HIPify-perl 和 HIPify-clang 工具，可用于 CUDA 到 HIP 的代码转换，转码后可在 AMD GPU 上编译运行，而基于 HIP 开发的应用也可同时用于 AMD 和英伟达的 GPU 上。虽然这种兼容方式依然需要开发者进行一定的转码工作，不过已可大量节省开发者针对 AMD 产品进行代码重新开发的时间。**这为 AMD 提供了说服客户进行迁移的条件和理由。兼容 CUDA 属权宜之计，能让 AMD 在短期内争取客户和抢占市场。**

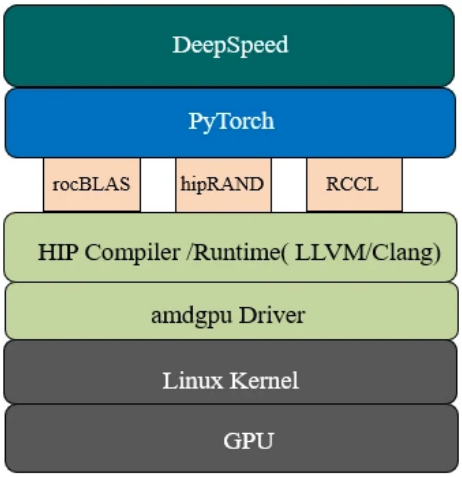
3）与云厂商等重要客户分别合作，重构自己的库，分而治之，以此与 CUDA 脱钩：长期一味兼容 CUDA 只会导致 ROCm 受 CUDA 的掣肘，加上需应对 CUDA 的每一次更新迭代，会导致 ROCm 陷入长期被动的局面。对云厂商来说，培育 AMD 与英伟达分庭抗礼，能有效影响芯片的定价权力，对重成本的云厂商来说也符合利益。因此，AMD 和云厂商有一拍即合的基础，可通过分别与云厂商客户合作构建兼容度更高的生态，分而治之与 CUDA 脱钩。例如 2022 年 3 月，微软在其开发的深度学习最优化函数库 DeepSpeed 中支持 ROCm，使开发者无需修改代码，就可以直接在 AMD 的 GPU 上运行。

图表366： AMD HIP 使 ROCm 可以部署英伟达和 AMD 的 GPU



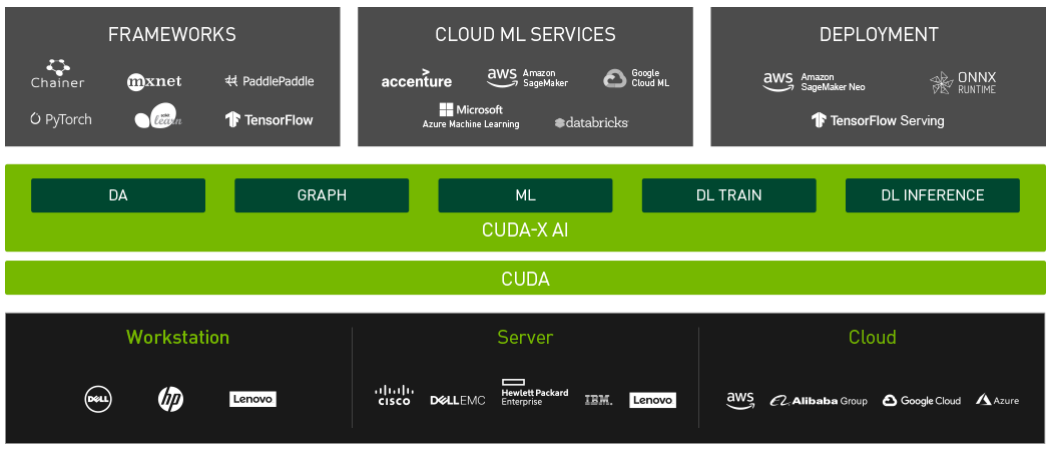
资料来源：AMD 官网、华泰研究

图表367： 微软 DeepSpeed 函数库支持 ROCm 和 AMD GPU



资料来源：微软官网、华泰研究

图表368： 英伟达 CUDA-X AI 生态圈及相关客户矩阵



资料来源：英伟达官网、华泰研究

图表369： 英伟达 CUDA 软件库及应用场景

种类	软件包组成	功能/应用
数学库	cuBLAS 、 cuFFT 、 cuRAND 、 cuSOLVER 、 cuSPARSE、cuTENSOR、AmgX	为分子动力学、计算流体学、计算化学、医学成像和地震勘探等领域的计算密集型应用奠定基础
并行算法库	Thrust	用于 C++中的多项运算，并在研究自然科学、物流、旅行规划等领域的关系时与图形一起使用
图像和视频库	nvJPEG、NVIDIA 性能基元、NVIDIA 视频编解码器 SDK、NVIDIA 光流 SDK	用于通过 CUDA 和 GPU 的专用硬件组件来进行图像和视频的解码、编码和处理
通信库	NVSHMEM、NCCL	性能经过优化的多 GPU 和多节点通信基元
深度学习库	NVIDIA cuDNN、NVIDIA TensorRT、NVIDIA Riva、NVIDIA DeepStream SDK、NVIDIA DALI	用于利用 CUDA 和专用硬件组件的深度学习应用
合作伙伴库	OpenCV、FFmpeg、ArrayFire、MAGMA、IMSL Fortran 数值库、Gunrock、CHOLMOD、Triton Ocean SDK、CUVIlib	包含 GPU 加速的开放源代码库等，覆盖矩阵、信号、图像、音频、视频等多种数据类型处理

资料来源：英伟达官网、华泰研究

图表370： ROCm 系统对应英伟达 CUDA 部分名称

CUDA	ROCm
CUDA API	HIP
NVCC	HCC
CUDA 函数库	ROC 库、HC 库
Thrust	Parallel STL
Profiler	ROCm Profiler
CUDA-GDB	ROCm-GDB
NVIDIA-smi	ROCm-SMI
Direct GPU RDMA	ROCm RDMA
TensorRT	Tensile
CUDA-Docker	ROCm-Docker
cuDNN	MIOpen

资料来源：CSDN、华泰研究

盈利预测与估值：AI 带来重估机会，24 年 PS 8.5x，TP 150 美元

2016 年 AMD 突围英特尔为其带来了估值的攀升。从 2016 年开始到 2021 年，AMD 的 Forward PS 从 3 倍左右增长到了 5-6 倍左右，估值提升的驱动力是 AMD 的 CPU 业务在不断抢夺英特尔的市场份额。我们认为，如今 AMD 面对 AI 浪潮带来的新增长机会，有望再次来到重估的转折点。

我们预测 AMD 2023/2024/2025 年收入为 241.88/285.10/318.87 亿美元。对比竞争对手英特尔和英伟达，2024 年的 PS 分别为 2.6 倍和 14.6 倍，我们认为 AMD 的业务跟两家的重叠度较大，在 CPU 业务方面制程一直领先英特尔，因此值得给予较英特尔高的估值；但在 GPU 业务方面，AMD 却处于奋力追赶追英伟达的状态，因此给予较英伟达低的估值。我们认为最终 AMD 的估值在两家之间较为合理，自 2020 年开始 AMD 的 PS 一直处于 5-6 倍之间，我们看好 AMD 在 AI 芯片的奋力一搏，估值上修可期，因此给予 2024 年 8.5 倍的 PS 估值，对应目标价为 150 美元，对应市值为 2423 亿美元。本节中，我们还将给出对 AMD 的盈利预测逻辑。

图表371： AMD 分板块盈利预测

百万美元（Million USD）	2020（A）	2021（A）	2022（A）	2023（E）	2024（E）	2025（E）
主营业务收入	9763.00	16434.00	23601.00	24188.49	28510.28	31886.66
YoY	*0.0%*	*68.3%*	*43.6%*	*2.5%*	*17.9%*	*11.8%*
数据中心	*1685.00*	*3694.00*	*6043.00*	*7601.24*	*10406.09*	*12089.13*
YoY	*0.0%*	*119.2%*	*63.6%*	*25.8%*	*36.9%*	*16.2%*
客户端	*5189.00*	*6887.00*	*6201.00*	*4296.41*	*4726.05*	*5198.65*
YoY	*0.0%*	*32.7%*	*-10.0%*	*-30.7%*	*10.0%*	*10.0%*
游戏	*2746.00*	*5607.00*	*6805.00*	*6600.85*	*6834.65*	*7073.86*
YoY	*0.0%*	*104.2%*	*21.4%*	*-3.0%*	*3.5%*	*3.5%*
嵌入式业务	*143.00*	*246.00*	*4552.00*	*5690.00*	*6543.50*	*7525.03*
YoY	*0.0%*	*72.0%*	*1750.4%*	*25.0%*	*15.0%*	*15.0%*
毛利润	**4347**	**7929**	**10603**	**11,369**	**14,540**	**17,219**
毛利率	*44.5%*	*48.2%*	*44.9%*	*47.0%*	*51.0%*	*54.0%*
销售及管理费用率	*10.2%*	*8.8%*	*9.9%*	*10.0%*	*9.0%*	*8.0%*
研发费用率	*20.3%*	*17.3%*	*21.2%*	*24.0%*	*23.0%*	*22.0%*
其他费用率	*0.0%*	*-0.1%*	*8.5%*	*6.0%*	*5.0%*	*5.0%*
净利率	*25.5%*	*19.2%*	*5.6%*	*6.1%*	*12.6%*	*17.2%*

资料来源：华泰研究预测

图表372：AI 芯片行业产业链的相关公司估值（数据截至 2023 年 9 月 20 日）

		PS			PE		
		2023E	2024E	2025E	2023E	2024E	2025E
芯片设计/晶圆代工/半导体设备							
NVIDIA CORPORATION	NVDA US	19.9	13.1	10.8	39.9	25.1	21.0
ADVANCED MICRO DEVICES	AMD	7.2	6.1	5.4	37.1	24.8	19.7
INTEL CORPORATION	INTC	2.9	2.6	2.4	56.7	20.6	15.6
BROADCOM	AVGO	10.1	9.4	8.8	25.3	22.5	19.4
QUALCOMM	QCOM	3.5	3.3	3.0	13.0	11.8	10.6
MOBILEYE	MBLY	14.8	11.6	8.4	54.4	44.5	29.8
TSMC	2330 TT	6.6	5.5	4.6	18.5	14.6	12.1
ASML Holding N.V.	ASML	8.0	7.4	6.2	28.8	25.5	19.7
均值		9.1	7.4	6.2	34.2	23.7	18.5
大型云厂商							
MICROSOFT	MSFT	11.0	9.7	8.5	31.7	28.0	24.1
ALPHABET	GOOGL	5.8	5.1	4.5	24.3	20.5	17.7
AMAZON	AMZN	2.5	2.3	2.0	61.6	41.8	29.8
ORACLE	ORCL	6.1	5.7	5.3	21.2	19.2	17.0
均值		6.3	5.7	5.1	34.7	27.4	22.2

注：预测值参考 Visible Alpha 一致预期
资料来源：Visible Alpha 官网、华泰研究

基于报告前序章节对 AMD 四大板块的业务历史、现状和前景的分析，以及对四业务终端市场态势的判断，我们分别对不同业务板块进行了评估，并进行盈利预测。在此我们首先解释对已经过去大半的 2023 年给予的较低的总营收增速 2.5%：1）数据中心营收方面，已发布的 2023Q1 和 Q2 相比 2022 年同比分别为 0.15%和-11.1%，而 Q3 根据公司 Q2 给出的展望将基本持平，但 Q4 因 MI300 系列或将于四季度量产，或能弥补前三季度的乏力。另外，据 Digitimes 2023 年 8 月 7 日最新预测，2023 年 Q2 全球服务器出货量环比下滑 5.7%，但 Q3 将迎来一定回弹（环比不超过 10%），综合以上，我们对 23 年数据中心营收给出预测增速为 25.8%；**2）客户端营收方面**，2023Q1 和 Q2 相比 2022 年同比大幅下降 65.2%和 53.6%，但 Q3 公司展望为同比持平，而 Q4 结合 PC 市场去库存或延续到四季度，我们最终给出 2023 客户端收入下降 30.7%的预测；**3）游戏业务方面**，虽然游戏主机 PS5 和 X-box S/X 的购买难度终于降低且存在游戏大作带动作用，但距离二者 2020 年上市已有三年，又有 AMD 较为弱势的显卡业务或与游戏主机的贡献相抵，我们给出 2023 收入端下降 3%的判断；**4）嵌入式业务方面**，考虑到全球 5G 布建放缓，大运营商减少基础设施投入，AMD 嵌入式业务从通信客户处获得的营收减少，工业、医疗、汽车等终端客户的贡献被通信客户的疲软抵消了，这种疲软或仍将持续，且赛灵思并表已进入平稳期，2022 年此板块因并表带来的大幅增长 1750%自然也不会再出现，我们对 2023 年嵌入式业务收入给出 25%的涨幅预测。

数据中心业务：我们预计 AMD 数据中心业务 2023/2024/2025 的营业收入同比增速为 25.8%/36.9%/16.2%，对应营业收入为 76.01/104.06/120.89 亿美元。**1）公司对 2023 年接下来的数据中心业务指引乐观：**2023Q2 财报会议上，公司称数据中心 Q3 业绩将同比持平，2023Q1 财报会议上，公司则称下半年数据中心业务较上半年增长 50%；**2）AI 风继续吹，AMD 产品对两大竞争对手都有可看战力：**大模型和生成式 AI 对算力需求的带动依然蓬勃，且考虑到 AMD 的数据中心 CPU 产品制程依然领先英特尔，且 MI300 系列从性能上又有力冲击英伟达，从竞争层面我们也认为 AMD 在这场 AI 争锋战中在数据中心 CPU 和 GPU 方面都能夺得份额；**3）但是，我们也考虑到特定不明朗因素：**宏观经济不确定性或导致市场需求信号不清，且地缘政治产生的出口限制或将对中国地区大型云客户及企业客户大算力 AI 芯片的需求产生一定影响，同时也考虑到了 ARM 架构的甚嚣尘上，以及英特尔的“五年四节点”奋起直追。另外，对于数据中心业务，我们使用的驱动因子包括：1）大语言模型和生成式 AI 的发展速度（我们认为 2 年内都将带来增长动力）；2）主要云厂商的资本支出增速（或将有所放缓）；3）AMD 在 2023 年二季度业绩会上预测称到 2027 年，人工智能加速器（AI 芯片）市场将达到 1500 亿美元以上。

客户端业务：我们预计 AMD 客户端业务 2023/2024/2025 的营业收入同比为-30.7%/10%/10%，对应营业收入为 42.96/47.26/51.99 亿美元。**1）公司对 2023 年接下来的客户端业务指引较为乐观：**2023Q2 财报会议上，公司称 Q3 客户端同比持平，接下来还将现两位数增长，公司预期在下半年传统旺季里 PC 业务将凭着 Ryzen 7000 系列 CPU 的增量以及 PC 厂商库存消化见底而回暖；**2）PC 端 AI 应用扬帆：**我们认为，随着 PC 端 AI 软件应用（如微软 Copilot 等）更加广泛，将带来 PC 端芯片要求上升，AMD 有望受惠；**3）2023 年 PC 厂商继续去库存：**2023 年 2/3 季度需求持续疲软，但或于年末后有所恢复，并或将于 2024 年开始回暖，逐步恢复至疫情前的水平。

游戏业务：我们预计 AMD 游戏业务 2023/2024/2025 的营业收入同比变化为-3.0%/3.5%/3.5%，对应营业收入为 66.0/68.35/70.74 亿美元。**1）疫情居家配合新主机的周期业已结束：**观察游戏业务板块历史增幅，2021 年增长 104.2%，2022 年增长 21.37%，但当时正值疫情期间，居家活动遇到最新款主机，如今疫情结束又处于主机代际之间，较难达到类似水平的高涨幅；**2）半代升级版本主机即将在 2024/2025 到来，或能略微拉动游戏业务增长：**游戏主机即将在 2024-2025 开启半代升级，如 PS5 的半代升级版本 PS5 Pro，和微软的新版 Xbox Series S Carbon Black 将在今年 9 月 1 日推出；**3）但是考虑到 2026-2027 年整代升级 PS6 和新版 Xbox 或将至**，我们认为 2025 年的游戏主机销量和 AMD 的游戏业务收入不会出现明显提升。

嵌入式业务：我们预计 AMD 嵌入式业务 2023/2024/2025 的营业收入同比变化为 25%/15%/15%，对应营业收入为 56.9/65.4/75.25 亿美元。**1）赛灵思并表效果已经止歇：**2022Q1 AMD 与赛灵思开始并表，从 2022 年 Q1 到 Q2 的绝对值高增长可以看出，Q1 时并表并没有完成，到 Q2 开始大致完成，此后一直到 2023Q1，嵌入式业务的绝对值变化都不大。2023Q1 是 AMD 与赛灵思并表后的第一个可比季度，这也是嵌入式业务 23Q1 的营收从去年的 5.95 亿美元激增至 15.6 亿美元且同比大幅增长 163%的原因，由此我们也认为 2023 年下半年嵌入式业务将进入同比稳定阶段；**2）FPGA 市场格局和赛灵思收入稳定可持续：**全球 FPGA 市场竞争格局及赛灵思市占率清晰稳定，且观察赛灵思与 AMD 并表前的营收，华为采购结束后，FY2020/FY2021 的年同比变化很小，分别为 3.39%和-0.48%，客户和业务变化也较少；**3）赛灵思与 AMD 能产生协同效应和交叉销售**，在赛灵思并表带来的无机营收增长之外，丰富产品组合，实现有机增长，如虎添翼。

费率及利润端：毛利率方面，我们参考了公司在 23Q2 业绩电话会上对下一季度毛利率 51%的指引，并且我们认为后续公司的毛利率或随 MI300 等新品规模释放及业务结构从 C 端到 B 端转移而持续改善，预计 23/24/25 年毛利率分别为 47.0%/51.0%/54.0%；费用率方面：20-22 年公司期间费率为 30.5%/26.0%/39.6%，22 年其他费用率大幅度上升主要系并购赛灵思后形成的无形资产摊销。往后我们认为公司将保持稳定的销售及管理开支水平，预测 23-25 年销售及管理费用率 10%/9%/8%；公司将继续加大对下一代 CPU、GPU 和 FPGAs 等的研发投入，预测研发费用率 24%/23%/22%；伴随着收入的扩张，预测其他费用率下降至 6%/5%5%；净利率方面，我们后续收入的增长、库存水平正常化和数据中心等高利润业务占比提升将带动公司净利率的回升，预计 23/24/25 年净利率分别为 6.1%/12.6%/17.2%。

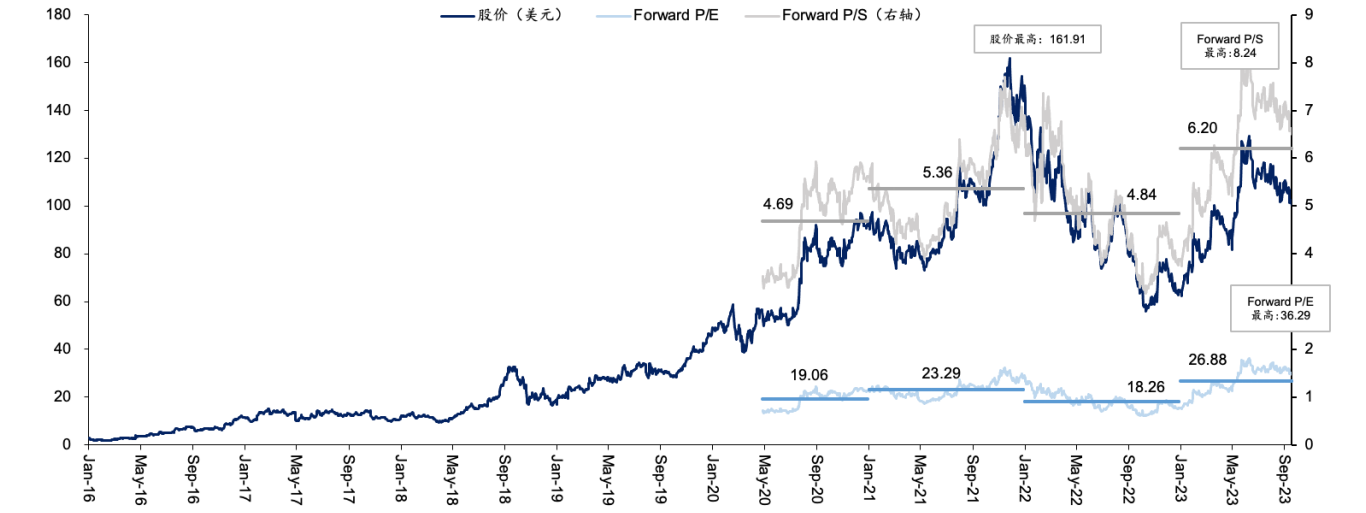
我们对公司 FY2023/FY2024/FY2025 的营业收入预测相对市场一致预期差额为+13.56/+10.03/+8.35 亿美元，略高于彭博一致预期，主要由于我们对数据中心业较为乐观。AMD CPU 制程仍领先英特尔，MI300 系列有力冲击英伟达，我们看好其在 CPU 和 GPU 市场份额的提升。

图表373： AMD 关键财务指标彭博一致预期 VS 华泰研究预测（单位：百万美元）

	FY2023E			FY2024E			FY2025E		
	彭博一致预期	华泰预测	差额	彭博一致预期	华泰预测	差额	彭博一致预期	华泰预测	差额
公司总营收	$22,833	$24,188	$1,356	$27,507	$28,510	$1,003	$31,052	$31,887	$835
YoY	-3.26%	2.50%	5.76%	20.47%	17.90%	-2.57%	12.89%	11.80%	-1.09%
数据中心	$6,514	$7,601	$1,087	$9,900	$10,406	$506	$12,654	$12,089	-$565
YoY	7.79%	25.80%	18.01%	51.98%	36.90%	-15.08%	27.82%	16.20%	-11.62%
客户端	$4,340	$4,296	-$44	$5,629	$4,726	-$903	$5,786	$5,199	-$587
YoY	-30.01%	-30.70%	-0.69%	29.71%	10.00%	-19.71%	2.78%	10.00%	7.22%
游戏	$6,317	$6,601	$284	$5,908	$6,835	$926	$6,029	$7,074	$1,045
YoY	-7.18%	-3.00%	4.18%	-6.46%	3.50%	9.96%	2.04%	3.50%	1.46%
嵌入式业务	$5,496	$5,690	$194	$5,390	$6,544	$1,154	$6,228	$7,525	$1,297
YoY	20.74%	25.00%	4.27%	1.93%	15.00%	13.07%	15.56%	15.00%	-0.56%
毛利率	48.65%	47.00%	-1.65%	51.89%	51.00%	-0.89%	53.44%	54.00%	0.56%
销售及管理费用率	9.61%	10.00%	0.39%	8.44%	9.00%	0.56%	8.30%	8.00%	-0.30%
研发费用率	24.84%	24.00%	-0.84%	22.14%	23.00%	0.86%	20.23%	22.00%	1.77%
净利率	6.50%	6.10%	-0.40%	15.38%	12.60%	-2.78%	19.14%	17.20%	-1.94%

资料来源：Bloomberg、华泰研究

图表374： 2016 年至今 AMD 历史股价、Forward PE 和 Forward PS（数据截至 2023 年 9 月 20 日）



资料来源：Bloomberg、华泰研究

风险提示

新产品落地进度推迟：MI300、Bergamo 等新产品的发售或受到市场需求波动、供应链扰动、技术挑战无法及时攻克等因素的影响，无法按照预期进度落地放量，使得营收提升不及预期。

PC 出货量恢复不及预期：若全球 PC 需求及出货量持续下行，恢复不及预期，客户端及游戏业务营收跌幅可能会继续扩张。

AI 技术落地和推进不及预期：AMD 重点发力 AI 领域，发布 MI300 进军 AI 训练端，而若 AI 市场技术落地和推进受阻，AMD AI 产品的需求及营收可能受到影响。

盈利预测

利润表

会计年度 (美元百万)	2021	2022	2023E	2024E	2025E
营业收入	16,434	23,601	24,188	28,510	31,887
销售成本	8,505	12,998	12,820	13,970	14,668
毛利润	7,929	10,603	11,369	14,540	17,219
销售及分销成本	1,448	2,336	2,419	2,566	2,551
管理费用	2,845	5,005	5,805	6,557	7,015
其他收入/支出	0.00	1,998	1,451	1,426	1,594
财务成本净额	21.00	(80.00)	(56.89)	(23.92)	29.94
应占联营公司利润及亏损	6.00	14.00	15.40	16.94	18.63
税前利润	3,675	1,198	1,652	3,984	6,107
税费开支	513.00	(122.00)	(168.20)	(405.76)	(621.92)
少数股东损益	0.00	0.00	0.00	0.00	0.00
净利润	3,162	1,320	1,483	3,579	5,485
折旧和摊销	(407.00)	(4,174)	(2,857)	(3,433)	(4,258)
EBITDA	4,061	5,452	4,565	7,441	10,335
EPS (美元，基本)	2.61	0.82	0.92	2.21	3.39

资产负债表

会计年度 (美元百万)	2021	2022	2023E	2024E	2025E
存货	1,955	3,771	3,809	4,456	4,991
应收账款和票据	2,706	4,126	4,250	4,972	5,569
现金及现金等价物	2,535	4,835	6,423	8,573	12,530
其他流动资产	1,387	2,287	2,744	3,293	3,952
总流动资产	8,583	15,019	17,226	21,295	27,042
固定资产	702.00	1,513	2,313	3,382	4,713
无形资产	289.00	48,295	46,451	44,606	42,762
其他长期资产	2,845	2,753	3,028	3,331	3,664
总长期资产	3,836	52,561	51,792	51,319	51,139
总资产	12,419	67,580	69,018	72,614	78,181
应付账款	1,406	2,956	3,252	3,577	3,934
短期借款	312.00	0.00	0.00	0.00	0.00
其他负债	2,522	3,413	3,072	2,765	2,488
总流动负债	4,240	6,369	6,323	6,341	6,423
长期债务	1.00	2,467	2,467	2,467	2,467
其他长期债务	681.00	3,994	3,994	3,994	3,994
总长期负债	682.00	6,461	6,461	6,461	6,461
股本	12.00	16.00	16.00	16.00	16.00
储备/其他项目	7,485	54,734	56,217	59,796	65,281
股东权益	7,497	54,750	56,233	59,812	65,297
少数股东权益	0.00	0.00	0.00	0.00	0.00
总权益	7,497	54,750	56,233	59,812	65,297

估值指标

会计年度 (倍)	2021	2022	2023E	2024E	2025E
PE	41.98	133.97	119.20	49.41	32.24
PB	17.71	3.23	3.14	2.96	2.71
EV EBITDA	43.72	33.31	39.36	23.82	16.74
股息率 (%)	0.00	0.00	0.00	0.00	0.00
自由现金流收益率 (%)	2.60	3.68	1.08	1.40	2.41

现金流量表

会计年度 (美元百万)	2021	2022	2023E	2024E	2025E
EBITDA	4,061	5,452	4,565	7,441	10,335
融资成本	(21.00)	80.00	56.89	23.92	(29.94)
营运资本变动	314.00	1,618	(664.59)	(1,901)	(1,709)
税费	513.00	(122.00)	(168.20)	(405.76)	(621.92)
其他	(1,346)	(3,463)	(72.29)	(40.86)	11.30
经营活动现金流	3,521	3,565	3,717	5,118	7,986
CAPEX	(301.00)	(450.00)	(1,812)	(2,658)	(3,745)
其他投资活动	(379.00)	2,463	(259.90)	(285.89)	(314.48)
投资活动现金流	(686.00)	1,999	(2,072)	(2,943)	(4,059)
债务增加量	0.00	679.00	0.00	0.00	0.00
权益增加量	(1,895)	(3,941)	0.00	0.00	0.00
派发股息	0.00	0.00	0.00	0.00	0.00
其他融资活动现金流	0.00	(2.00)	(56.89)	(23.92)	29.94
融资活动现金流	(1,895)	(3,264)	(56.89)	(23.92)	29.94
现金变动	940.00	2,300	1,588	2,150	3,957
年初现金	1,595	2,535	4,835	6,423	8,573
汇率波动影响	0.00	0.00	0.00	0.00	0.00
年末现金	2,535	4,835	6,423	8,573	12,530

业绩指标

会计年度 (倍)	2021	2022	2023E	2024E	2025E
增长率 (%)					
营业收入	68.33	43.61	2.49	17.87	11.84
毛利润	82.40	33.72	7.22	27.90	18.42
营业利润	165.60	(65.24)	33.95	135.73	51.79
净利润	26.99	(58.25)	12.39	141.23	53.27
EPS	23.90	(68.66)	12.39	141.23	53.27
盈利能力比率 (%)					
毛利润率	48.25	44.93	47.00	51.00	54.00
EBITDA	24.71	23.10	18.87	26.10	32.41
净利润率	19.24	5.59	6.13	12.55	17.20
ROE	47.43	4.24	2.67	6.17	8.77
ROA	29.58	3.30	2.17	5.05	7.27
偿债能力 (倍)					
净负债比率 (%)	(29.64)	(4.33)	(7.04)	(10.21)	(15.41)
流动比率	2.02	2.36	2.72	3.36	4.21
速动比率	1.56	1.77	2.12	2.66	3.43
营运能力 (天)					
总资产周转率 (次)	1.54	0.59	0.35	0.40	0.42
应收账款周转天数	52.27	52.11	62.33	58.22	59.50
应付账款周转天数	(41.31)	(60.41)	(87.16)	(87.98)	(92.17)
存货周转天数	(70.98)	(79.30)	(106.42)	(106.49)	(115.93)
现金转换周期	22.60	33.22	43.06	39.71	35.75
每股指标 (美元)					
EPS	2.61	0.82	0.92	2.21	3.39
每股净资产	6.18	33.89	34.81	37.02	40.42

资料来源：公司公告、华泰研究预测

免责声明

分析师声明

本人，何翩翩，兹证明本报告所表达的观点准确地反映了分析师对标的证券或发行人的个人意见；彼以往、现在或未来并无就其研究报告所提供的具体建议或所表达的意见直接或间接收取任何报酬。

一般声明及披露

本报告由华泰证券股份有限公司（已具备中国证监会批准的证券投资咨询业务资格，以下简称“本公司”）制作。本报告所载资料是仅供接收人的严格保密资料。本报告仅供本公司及其客户和其关联机构使用。本公司不因接收人收到本报告而视其为客户。

本报告基于本公司认为可靠的、已公开的信息编制，但本公司及其关联机构(以下统称为“华泰”)对该等信息的准确性及完整性不作任何保证。

本报告所载的意见、评估及预测仅反映报告发布当日的观点和判断。在不同时期，华泰可能会发出与本报告所载意见、评估及预测不一致的研究报告。同时，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。以往表现并不能指引未来，未来回报并不能得到保证，并存在损失本金的可能。华泰不保证本报告所含信息保持在最新状态。华泰对本报告所含信息可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。

本公司不是 FINRA 的注册会员，其研究分析师亦没有注册为 FINRA 的研究分析师/不具有 FINRA 分析师的注册资格。

华泰力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，不构成购买或出售所述证券的要约或招揽。该等观点、建议并未考虑到个别投资者的具体投资目的、财务状况以及特定需求，在任何时候均不构成对客户私人投资建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。对依据或者使用本报告所造成的一切后果，华泰及作者均不承担任何法律责任。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。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现，过往的业绩表现不应作为日后回报的预示。华泰不承诺也不保证任何预示的回报会得以实现，分析中所做的预测可能是基于相应的假设，任何假设的变化可能会显著影响所预测的回报。

华泰及作者在自身所知情的范围内，与本报告所指的证券或投资标的不存在法律禁止的利害关系。在法律许可的情况下，华泰可能会持有报告中提到的公司所发行的证券头寸并进行交易，为该公司提供投资银行、财务顾问或者金融产品等相关服务或向该公司招揽业务。

华泰的销售人员、交易人员或其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。华泰没有将此意见及建议向报告所有接收者进行更新的义务。华泰的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。投资者应当考虑到华泰及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。有关该方面的具体披露请参照本报告尾部。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布的机构或人员，也并非意图发送、发布给因可得到、使用本报告的行为而使华泰违反或受制于当地法律或监管规则的机构或人员。

中国香港

本报告由华泰证券股份有限公司制作,在香港由华泰金融控股（香港）有限公司向符合《证券及期货条例》及其附属法律规定的机构投资者和专业投资者的客户进行分发。华泰金融控股（香港）有限公司受香港证券及期货事务监察委员会监管，是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。在香港获得本报告的人员若有任何有关本报告的问题,请与华泰金融控股（香港）有限公司联系。

香港-重要监管披露

- 华泰金融控股（香港）有限公司的雇员或其关联人士没有担任本报告中提及的公司或发行人的高级人员。
- 有关重要的披露信息，请参华泰金融控股（香港）有限公司的网页 https://www.htsc.com.hk/stock_disclosure

其他信息请参见下方 **“美国-重要监管披露”**。

美国

在美国本报告由华泰证券（美国）有限公司向符合美国监管规定的机构投资者进行发表与分发。华泰证券（美国）有限公司是美国注册经纪商和美国金融业监管局（FINRA）的注册会员。对于其在美国分发的研究报告，华泰证券（美国）有限公司根据《1934 年证券交易法》（修订版）第 15a-6 条规定以及美国证券交易委员会人员解释，对本研究报告内容负责。华泰证券（美国）有限公司联营公司的分析师不具有美国金融监管（FINRA）分析师的注册资格，可能不属于华泰证券（美国）有限公司的关联人员，因此可能不受 FINRA 关于分析师与标的公司沟通、公开露面和所持交易证券的限制。华泰证券（美国）有限公司是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。任何直接从华泰证券（美国）有限公司收到此报告并希望就本报告所述任何证券进行交易的人士，应通过华泰证券（美国）有限公司进行交易。

美国-重要监管披露

- 分析师何翩翩本人及相关人士并不担任本报告所提及的标的证券或发行人的高级人员、董事或顾问。分析师及相关人士与本报告所提及的标的证券或发行人并无任何相关财务利益。本披露中所提及的“相关人士”包括 FINRA 定义下分析师的家庭成员。分析师根据华泰证券的整体收入和盈利能力获得薪酬，包括源自公司投资银行业务的收入。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或不时会以自身或代理形式向客户出售及购买华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或其高级管理层、董事和雇员可能会持有本报告中所提到的任何证券（或任何相关投资）头寸，并可能不时进行增持或减持该证券（或投资）。因此，投资者应该意识到可能存在利益冲突。

评级说明

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力（含此期间的股息回报）相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数），具体如下：

行业评级

增持：预计行业股票指数超越基准
中性：预计行业股票指数基本与基准持平
减持：预计行业股票指数明显弱于基准

公司评级

买入：预计股价超越基准 15%以上
增持：预计股价超越基准 5%~15%
持有：预计股价相对基准波动在-15%~5%之间
卖出：预计股价弱于基准 15%以上
暂停评级：已暂停评级、目标价及预测，以遵守适用法规及/或公司政策
无评级：股票不在常规研究覆盖范围内。投资者不应期待华泰提供该等证券及/或公司相关的持续或补充信息

法律实体披露

中国：华泰证券股份有限公司具有中国证监会核准的“证券投资咨询”业务资格，经营许可证编号为：91320000704041011J

香港：华泰金融控股（香港）有限公司具有香港证监会核准的“就证券提供意见”业务资格，经营许可证编号为：AOK809

美国：华泰证券（美国）有限公司为美国金融业监管局（FINRA）成员，具有在美国开展经纪交易商业务的资格，经营业务许可编号为：CRD#:298809/SEC#:8-70231

华泰证券股份有限公司

南京
南京市建邺区江东中路 228 号华泰证券广场 1 号楼/邮政编码：210019

电话：86 25 83389999/传真：86 25 83387521
电子邮件：ht-rd@htsc.com

北京
北京市西城区太平桥大街丰盛胡同 28 号太平洋保险大厦 A 座 18 层/邮政编码：100032
电话：86 10 63211166/传真：86 10 63211275
电子邮件：ht-rd@htsc.com

深圳
深圳市福田区益田路 5999 号基金大厦 10 楼/邮政编码：518017
电话：86 755 82493932/传真：86 755 82492062
电子邮件：ht-rd@htsc.com

上海
上海市浦东新区东方路 18 号保利广场 E 栋 23 楼/邮政编码：200120
电话：86 21 28972098/传真：86 21 28972068
电子邮件：ht-rd@htsc.com

华泰金融控股（香港）有限公司
香港中环皇后大道中 99 号中环中心 58 楼 5808-12 室
电话：+852-3658-6000/传真：+852-2169-0770
电子邮件：research@htsc.com
http://www.htsc.com.hk

华泰证券（美国）有限公司
美国纽约公园大道 280 号 21 楼东（纽约 10017）
电话：+212-763-8160/传真：+917-725-9702
电子邮件: Huatai@htsc-us.com
http://www.htsc-us.com