

电子

专业云厂商扩张趋势逐渐明确

英伟达、微软投资 CoreWeave，云服务助力算力布局。CoreWeave 是一家专为大规模 GPU 加速工作负载而架构的专业云提供商。1) 产品：CoreWeave 拥有七大产品模块，通过并购 Conductor Technologies 赋能产品及扩张。2) 核心优势：与传统云提供商相比，采用 Kubernetes-native 云计算平台基础机构，通过 Kubernetes 的 API 和控制面板来管理和调度容器和 GPU 资源，计算成本比其他竞争对手产品便宜 80%（无需基础架构开销）（英伟达 HGX H100 GPU 组件成本 4.76 美金每小时），速度提升了 35 倍。其重点专利包括 GPU 云计算平台、GPU 资源管理、容器化 GPU 加速等。3) 合作：英伟达（参与公司 B 轮融资）、微软（投入十亿美元用于云计算基础设施）与 CoreWeave 合作以应对日益增长的算力需求和高昂的算力成本。此外，公司与 Tarteel AI、Anlatan（NovelAI 创建者）、Stable diffusion、EleutherAI（机器学习项目和开源人工智能）、Spire Animation 以及 PureWeb(视觉效果和动画公司)建立了合作关系。

AIGC 显著提速，算力及云服务需求呈向上弹性。AIGC 大模型、多模态、商业化发展推动算力需求持续扩大。AIGC 大模型的数据规模和算法模型的双层叠加下，算力需求将会越来越大。国际科技巨头纷纷推动 AI 模型商业化，进一步刺激算力需求。据中国信息通信研究院报告，预计 2030 年全球算力规模将达到 56 ZFLOPS。AIGC 大模型时代的到来使得智能算力成为普遍需求，影响云计算服务的模式和格局。根据中商产业研究所数据，2022 年全球云计算规模达到 3566 亿美元，预计 2023 年将突破 4000 亿美元。

AIGC 基础设施层发展成熟，云服务与芯片为核心资源。在所有层级中，基础设施层通常被认为是最成熟、稳定和商业化的。该层级中重点关注云服务与芯片两个核心资源：1) 云服务提供商：云服务提供商通过提供超大规模和特定目的的计算、储存和网络技术在基础设施层占据了市场的主导地位。AI 需要机器庞大的计算能力，很多公司转向云服务，通过云服务基础设施来解决算力问题。2) 芯片：AI 算力芯片是类 ChatGPT 模型的基石。我们认为，短期内具有大算力、通用性的 GPU 芯片或成为大算力应用首选，未来 GPU 与 ASIC 的界限可能会在较大程度上模糊。

投资建议：建议关注云端 AI 相关企业：寒武纪、海光信息（天风计算机团队覆盖）、龙芯中科、紫光国微、复旦微电、安路科技、亚马逊、微软公司（天风海外组覆盖）、谷歌、甲骨文等；建议关注算力芯片相关企业：英伟达（天风海外组覆盖）、AMD、Intel、景嘉微（天风计算机团队联合覆盖）等。

风险提示：AI 发展及商业化不及预期、AI 行业竞争加剧、政策不确定性。

证券研究报告

2023 年 06 月 20 日

投资评级

行业评级

强于大市(维持评级)

上次评级

强于大市

作者

潘暕

分析师

SAC 执业证书编号：S1110517070005
panjian@tfzq.com

俞文静

分析师

SAC 执业证书编号：S1110521070003
yuwenjing@tfzq.com

行业走势图



资料来源：聚源数据

相关报告

- 1 《电子行业深度研究：AI 算力需求持续释放，重点看好 AI 服务器产业链》 2023-06-14
- 2 《电子行业点评：苹果 MR 发布在即，重点推荐相关产业链》 2023-05-31
- 3 《电子行业深度研究：电子行业 23Q1 总结：有望进入复苏周期》 2023-05-21

内容目录

1. 英伟达、微软投资 CoreWeave，云服务助力算力布局	3
1.1. 拥有七大产品模块，并购赋能产品及扩张	3
1.2. 三大优势助力，生成式 AI 市场中脱颖而出	4
1.3. 英伟达、微软 AIGC 亮眼，与 CoreWeave 进行战略合作	5
2. AIGC 显著提速，算力及云服务需求呈向上弹性	6
2.1. AIGC 大模型、多模态、商业化发展，算力需求持续扩大	6
2.2. AIGC 模型及算力需求提升，云计算服务格局及规模有望改善	7
3. AIGC 基础设施层发展成熟，云服务与芯片为核心资源	7
4. 投资建议	8
5. 风险提示	8

图表目录

图 1：CoreWeave 云架构	3
图 2：Conductor Technologies 核心优势	4
图 3：AIGC 多模态大模型生成结果图	6
图 4：全球云计算市场规模	7
图 5：AIGC 技术栈	7
表 1：CoreWeave GPU 云定价	5
表 2：CoreWeave 融资情况（截至 2023 年 5 月 31 日）	6
表 3：国际科技巨头 AI 模型商业化	6
表 4：基础设施层提供商	8

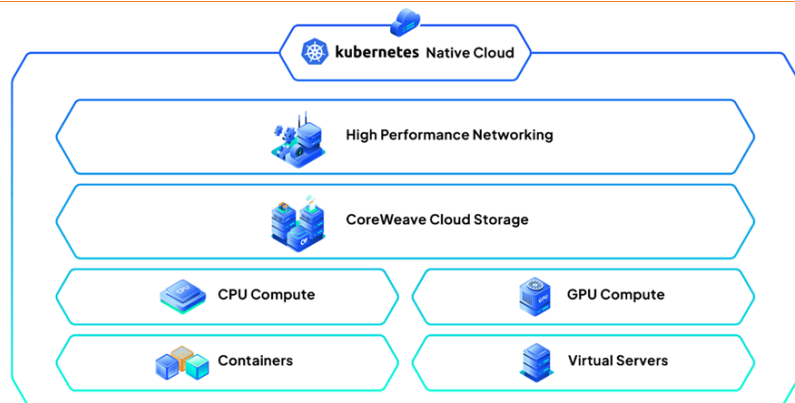
1. 英伟达、微软投资 CoreWeave，云服务助力算力布局

CoreWeave 是一家专为大规模 GPU 加速工作负载而架构的专业云提供商。CoreWeave 由 Michael Intrator、Brian Ventura 和 Brannin McBee 于 2017 年成立，曾是一家以太坊挖矿企业，后转型为云计算平台公司。CTO Brian Ventura 是一名以太坊挖矿爱好者，曾选择英伟达硬件来增加内存（英伟达后成为 CoreWeave 的投资方）。CoreWeave 的快速发展也受到融资团队的大力支持，融资团队包括 Magnetar Capital、Nvidia、前 GitHub 执行官 Nat Friedman 和前苹果高管 Daniel Gross。

1.1. 拥有七大产品模块，并购赋能产品及扩张

CoreWeave 拥有七大板块产品。主要产品包括 NVIDIA HGX H100、GPU Compute、CPU Compute、Kubernetes、Virtual Servers、Storage 和 Networking：

图 1：CoreWeave 云架构



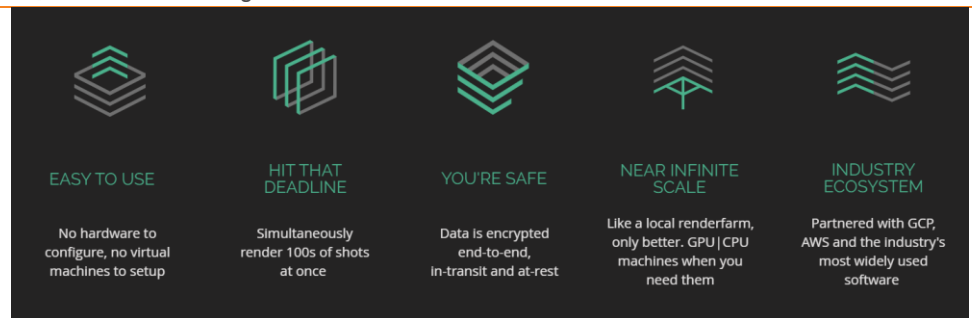
资料来源：CoreWeave 官网，天风证券研究所

- 1) **NVIDIA HGX H100：**为大规模 HPC 和 AI 工作负载而设计，与 NVIDIA HGX A100 相比，高性能计算 (HPC) 应用程序的效率提高了 7 倍，最大模型的 AI 训练速度提高了 9 倍，AI 推理速度提高了 30 倍。CoreWeave HGX H100 分布式训练集群采用轨道优化设计，使用 NVIDIA Quantum-2 InfiniBand 网络支持 NVIDIA SHARP 网络内收集，每个节点提供 3.2Tbps 的 GPU Direct 带宽。CoreWeave 针对 NVIDIA GPU 加速工作负载进行了优化，能够轻松地运行现有工作负载，而只需进行极少更改或无需更改。CoreWeave 的快速、灵活的基础架构也可以帮助实现最佳性能。
- 2) **GPU Compute：**CoreWeave 是 GPU 加速工作负载的主要云提供商，核心产品是各类别的 NVIDIA GPU。作为一家专业的云提供商，CoreWeave 在使用户能够扩展的基础架构上提供与用户的工作负载复杂性相匹配的计算服务。CoreWeave 拥有 10 多个专为计算密集型用例设计的 NVIDIA GPUs 的使用案例，使客户能够适当调整工作负载的性能和成本。CoreWeave 是为大规模弹性、实时消费而构建的，可以在客户需要时提供所需的 GPU 并且拥有可配置实例、透明定价和直观计费。
- 3) **CPU Compute：**CoreWeave 的 CPU 服务器群独立存在。凭借针对最终帧渲染、数据分析或视频转码显著的扩展能力，CoreWeave 的纯 CPU 实例提供了通用计算所需的规模、范围和灵活性。CoreWeave CPU 可以轻松扩展用户的应用程序，从同一控制平面安排和管理客户的 CPU 工作负载。CoreWeave 的 CPU 计算产品组合为任何用例提供了经过成本调整的出色性能选项，按需启动数以万计的 CPU 内核以满足紧迫的渲染期限，或以惊人的规模进行数据分析。
- 4) **Kubernetes：**与传统的基于 VM 的部署相比，通过单个编排层管理所有资源，CoreWeave 的客户可以充分利用更高的可移植性、更少的开销和更低的管理复杂性。得益于容器图像缓存和专门的调度程序，CoreWeave 的工作负载可以在短短 5 秒内启动并运行。CoreWeave 可以即时访问同一集群中的大量资源，只需用户请求所需要的 CPU 内核和 RAM，以及可选数量的 GPU。CoreWeave 处理所有控制平面基础设施、集群操作和平台集成，因此客户可以将更多时间用于构建产品。

- 5) **Virtual Servers:** CoreWeave 的虚拟服务器建立在 Kubernetes 之上, 使用开源项目 KubeVirt 来处理不易容器化的工作负载。从 UI 或通过 CoreWeave Kubernetes API 在几秒钟内启动虚拟服务器。CoreWeave 通过 PCI 直通专用的 GPU 的裸机性能, 没有 GPU 虚拟化或共享资源。与 CoreWeave 中的所有内容一样, 虚拟服务器是可定制的, 可以将客户的工作负载与 NVIDIA GPU 相匹配, 在几秒钟内实现类型切换, 并且完全支持 Linux 和 Windows 虚拟服务器。
- 6) **Storage:** 根据用户的工作负载, CoreWeave 提供一系列存储选项。CoreWeave Cloud Storage Volumes 建立在 Ceph 之上, Ceph 是一个软件定义的横向扩展企业级存储平台, 旨在为客户云原生工作负载提供高可用性、高性能存储。CoreWeave 使用三重复制, 分布在多个服务器和数据中心机架上, 专为高可用性而构建。所有存储卷都可以由容器化工作负载和虚拟服务器安装, 从而可以灵活地更改底层计算资源或部署方法, 并且可以将容量从 1GB 快速扩展到 PB (1000 TB) 规模。
- 7) **Networking:** CoreWeave 的 Kubernetes 原生网络设计将功能转移到网络结构中, 因此客户可以花更少的时间管理 IP 和 VLAN 来获得所需的性能和安全性。通过区域优化的交通提供商, CoreWeave 的公共连接为美国超过 5100 万人提供低延迟访问。使用 Kubernetes 网络策略管理防火墙, 或为第 2 层本机环境部署 VPC 网络。CoreWeave 可以为客户的应用程序部署负载均衡器服务, 以免提供高度可用、可扩展的基础架构。当需要管理第 2 层环境时, CoreWeave 虚拟私有云 (VPC) 将网络控制权交还给用户。

CoreWeave 积极通过并购来增强自己的产品并实现扩张。2023 年 1 月, CoreWeave 宣布对于 Conductor Technologies 的收购。Conductor 是基于云的任务管理服务开发商。CoreWeave 对 Conductor 的收购将增强产品在媒体和娱乐行业的应用, 将帮助 CoreWeave 扩展 VFX 和动画工作室的功能, 轻松地将工作负载到云端。同时, 这一收购也使得 Core Weave 的员工人数快速增加, 截至 23 年 1 月 25 日, 员工数量已超过 90 人, Conductor 的 CEO Mac Moore 现在在 CoreWeave 管理媒体和娱乐部门。

图 2: Conductor Technologies 核心优势



资料来源: Conductor Technologies 官网, 天风证券研究所

1.2. 三大优势助力, 生成式 AI 市场中脱颖而出

CoreWeave 与传统云服务商相比聚焦于生成式 AI、深耕 GPU 加速技术并且具有价格优势:

1) **聚焦于生成式 AI:** AWS、微软和谷歌云等传统超大规模云计算服务商, 形成了一系列大规模的云计算服务并建立庞大的数据中心, 目的在于针对几乎所有的潜在客户需求。而 CoreWeave 则采用完全相反的方法, 聚焦于以极具竞争力的价格为生成式 AI 提供平台。CoreWeave 在生成式 AI 领域的合作表现亮眼。CoreWeave 与知名生成人工智能公司 Tarteel AI、Anlatan (NovelAI 的创建者), 机器学习及开源人工智能公司 Stability AI 的 Stable Diffusion 和 EleutherAI 进行合作。同时, Spire Animation 和 PureWeb 等视觉效果 (VFX) 和动画公司已与 CoreWeave 建立合作关系。

2) **深耕 GPU 加速技术:** CoreWeave 云架构是专为大规模 GPU 加速工作负载构建的 Kubernetes 原生云。Kubernetes 是一种容器编排引擎, 支持容器自动化部署、大规模弹性扩展及容器化应用的统一管理。在 Kubernetes 统一管理和使用 GPU 资源可以提高部署

效率、实现租户隔离和进行统一资源调度和管理。现在对 GPU 加速技术的关注使 CoreWeave 在涉及更专业的用例，尤其是 AI 特定需求时，能够超越其他云提供商。生成式 AI 技术，例如 ChatGPT 聊天机器人和 Stable Diffusion 的艺术生成 AI，需要大规模运行大量几乎相同的任务。由于 GPU 擅长执行此操作，从而大大提高了速度和功率。

3) 价格优势：云基础设施可用于众多用例，包括视觉效果渲染、机器学习和人工智能、大规模批处理和像素流，根据公司官网数据，与通用技术相比，处理速度最高可提高 35 倍，成本降低 80%。一方面，CoreWeave 采用的是 Kubernetes 原生云可以实现可移植性，即可以充分利用混合云并部署到任何云提供商，可以帮助客户降低基础设施的构建成本。另一方面，CoreWeave 使用基于资源的定价，客户只需在使用资源时为使用的资源付费。除此之外，CoreWeave 提供所有大型云提供商中最低的按需价格和业界最广泛的 NVIDIA GPU。以 CoreWeave 的 GPU 云定价为例，定价为单点定价，其中总实例成本是 GPU 组件、vCPU 数量和分配的 RAM 量的组合。为简单起见，每个基本单元的 CPU 和 RAM 成本相同，唯一的变量是为客户的工作负载或虚拟服务器选择的 GPU。

表 1: CoreWeave GPU 云定价

GPU 型号	VRAM(GB)	每个 GPU 的最大 vCPU (\$0.01/每小时)	每个 GPU 的最大 RAM (GB) (\$0.005/每小时)	每小时 GPU 组件成本
NVIDIA HGX H100	80	48	256	\$4.76
NVIDIA H100 PCIe	80	48	256	\$4.25
A100 80GB NVLINK	80	48	256	\$2.21
A100 80GB PCIe	80	48	256	\$2.21
A100 40GB NVLINK	40	48	256	\$2.06
A100 40GB PCIe	40	48	256	\$2.06
A40	48	48	256	\$1.28
RTX A6000	48	48	256	\$1.28
RTX A5000	24	36	128	\$0.77
RTX A4000	16	36	128	\$0.61
Quadro RTX 5000	16	36	128	\$0.57
Quadro RTX 4000	8	36	128	\$0.24
Tesla V100 NVLINK	16	36	128	\$0.80

资料来源：CoreWeave 官网，天风证券研究所

1.3. 英伟达、微软 AIGC 亮眼，与 CoreWeave 进行战略合作

英伟达、微软等巨头 AIGC 表现靓丽。英伟达掌握 AI 算力命脉。NVIDIA H100 被黄仁勋称为“全球首个为 AIGC 设计的计算机芯片”，产品可以帮助 AI 系统更快输出顺畅自然的文本、图像和内容。当前 AI 需求高涨，AI 算力芯片赛道竞争激烈，但英伟达凭借通用性和易用性具有稳定优势。2023 年第一季度英伟达总收入达到 71.9 亿美元。美东时间 5 月 30 日，英伟达成为全球首家市值超过 1 万亿美元的芯片公司。微软本季度发布一系列“AI 全家桶”，AI 算力需求吸引部分新客户，Bing 搜索引擎崛起有望挤压谷歌部分市场份额和营收。微软第一季度营收为 528.6 亿美元，收入主要来源于云服务，其中 Azure 和其他云服务收入增长 27%。

英伟达、微软与 CoreWeave 合作提升算力储备。AIGC 受到投资者追捧。根据 PitchBook 数据，2023 年第一季度，AIGC 初创企业进行的 46 项交易总价值约 17 亿美元，另外还有 106.8 亿美元交易在该季度宣布。截至 2023 年 5 月 31 日，CoreWeave 总融资达到 5.765 亿美元，英伟达也参与了 B 轮融资。根据 cbinsights 网站数据，CoreWeave 的估值在 2023 年 4 月就已达 20-22.21 亿美元。以 CoreWeave 为代表的新一代云服务提供商通过定制硬件和更低价格来针对可互换的 AI 工作负载，可以与传统的超大规模云服务提供商形成一定的竞争。根据 CNBC 消息，微软将在未来数年内向 CoreWeave 投资数十亿美元，用于 GPUs 驱动的云计算基础设施，以确保 OpenAI 有足够的算力运营，体现了科技巨头通过将基础设施、模型和应用程序结合在一起的解决方案应对日益增长的算力需求和高昂的算力成本。

表 2: CoreWeave 融资情况 (截至 2023 年 5 月 31 日)

轮次	投资方	投资金额
种子轮	-	3 百万美元
A 轮	-	2.5 百万美元
私募股权融资	Magnetar Capital	5 千万美元
债务融资	Magnetar Capital	1 亿美元
B 轮	Magnetar Capital	1.11 亿美元 (Series B) +2 亿美元 (Series B extension)
	Nvidia 、 Nat Friedman 、 Daniel Gross	1.1 亿美元
总计	-	5.765 亿美元

资料来源: Businesswire, Tech 商业公众号, Crunchbase, 天风证券研究所

2. AIGC 显著提速，算力及云服务需求呈向上弹性

2.1. AIGC 大模型、多模态、商业化发展，算力需求持续扩大

AIGC 大模型推动算力需求增长。算力是数字经济时代的核心生产力，也是人工智能技术发展的重要支撑和驱动力之一。以 AIGC 大模型 ChatGPT 为例，算力需求场景可以分为训练和推理两大类，根据实际应用可以进一步拆分为预训练、Finetune 和日常运营三个阶段。根据 OpenAI 论文，GPT-3 模型参数约 1750 亿个，预训练数据量为 45TB，折合为训练集约为 3000 亿 tokens，训练阶段算力需求约为 3.15×10^8 PFLOPS。除了训练，在推理方面也需要强大的算力支撑。在数据规模和算法模型的双层叠加下，算力需求将会越来越大。

多模态 AIGC 或成算力需求新驱动力。2021 年后，人工智能逐渐从单模态 AI 转向了多模态 AI。作为人工智能最受瞩目的发展方向之一，AIGC 是以人工智能为核心，多模态交互技术等技术共同整合而成的。随着算法的不断迭代，AIGC 可生成的内容形式已囊括文本、图像、音频和视频。今年 3 月 OpenAI 发布 GPT-4 模型，接受文本和图像输入信息。以谷歌发布的 PaLM-E 多模态模型为例，参数量最高可以达到 5620 亿个，需要集成各类模型对信息流的嵌入处理，使得模型整体更为庞大，对算力资源的需求进一步提升。

图 3: AIGC 多模态大模型生成结果图



资料来源: 中国信息通信研究院和京东探索研究院, 天风证券研究所

算力赋能 AIGC 商业化。国际科技巨头纷纷推动 AI 模型商业化。微软发布 Microsoft Copilot，将包括 GPT-4 在内的 LLM 与 Microsoft 365 应用程序和 Microsoft Graph 中的业务数据相结合，将带给用户一种全新的工作方式。同时，GPT-4 通过开放 API 接口，尝试接入更多的商业合作伙伴，以创造出更多的商业化的应用。Google 推出的 PaLM 2，Meta 的 LLaMA 以及 Amazon 的 Bedrock 均体现 AIGC 的商业化。随着 AIGC 商业化发展，强算力资源的企业将拥有更多的商业可能，我们认为这将进一步刺激算力需求。根据中国信息通信研究院报告，预计 2030 年全球算力规模将达到 56 ZFLOPS。

表 3: 国际科技巨头 AI 模型商业化

公司	信息
Microsoft	推出 OpenAI 企业级服务以及面向 Microsoft 365、Dynamics365、PowerPlatform 等一系列产品的“Copilot”服务，通过大语言模型（LLM）实现基于自然语言理解和生成的人

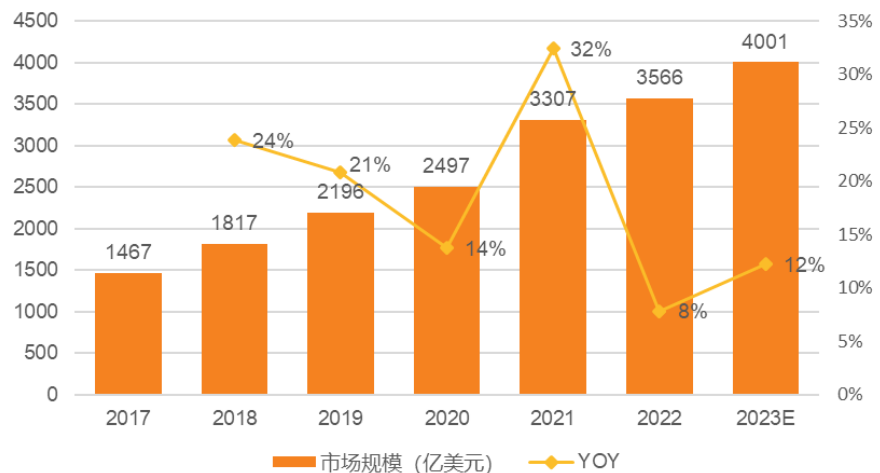
Google	机交互，并帮助用户完成各种复杂任务。 发布 PaLM 2，一种新的语言模型，具有改进的多语言、推理和编码功能。PaLM 2 支持超过 25 种 Google 产品和功能，被直接引入 Google 的产品和用户（包括消费者、开发人员和各种规模的企业），广泛应用于办公（Gmail、Google Docs）、医学（Med-PaLM 2）、网络安全（Sec-PaLM）、开发人员交流（Vertex AI、Duet AI）。
Meta	发布 LLaMA，是一个包含 650 亿参数的基础大型语言模型，作为封闭版本共享给研究社区。目标是供一个更小、性能更高的模型，研究人员可以在不需要大量硬件的情况下研究和微调特定任务。
Amazon	发布新服务 Bedrock，是客户使用 FM 构建和扩展基于 AI 的生成式应用程序，让所有构建者都能获得大众化访问权限的最简单方法。Bedrock 提供了通过可扩展、可靠和安全的 AWS 托管服务访问一系列功能强大的文本和图像 FM（包括 Amazon Titan FM）的功能。

资料来源：各公司官网，天风证券研究所

2.2. AIGC 模型及算力需求提升，云计算服务格局及规模有望改善

AIGC 发展刺激云计算服务需求。随着 AIGC 大模型时代的到来，智能算力成为普遍需求，进一步影响云计算服务的模式和格局。AI 云服务为 AIGC 开发提供了平台支撑。具体来看，人工智能预训练模型开发对于云服务有较大需求，AI 云服务可以提供人工智能开发模块，通过多元化的服务模式，降低开发者的开发成本和产品开发周期，为模型开发提供 AI 赋能。AIGC 大模型的逐渐成熟将推动云计算格局逐步从算力为基础的平台 IaaS，走向以模型能力为主的平台 MaaS。云计算 AI 能力的逐步放大也将刺激云计算服务需求。根据中商产业研究所数据，2022 年全球云计算规模达到 3566 亿美元，预计 2023 年将突破 4000 亿美元。

图 4：全球云计算市场规模



资料来源：中商产业研究院，中国信息通信研究院，深圳市电子商会，天风证券研究所

3. AIGC 基础设施层发展成熟，云服务与芯片为核心资源

基础设施层领 AIGC 技术栈成熟发展。生成式人工智能技术栈由三层组成，包括基础设施层、模型层和应用层。基础设施层包括超大规模计算及芯片两大部分，分别作为 AIGC 的基础设施和硬件基础。基础设施层现有龙头企业主要提供算力、网络、储存和中间件基础设施。芯片方面，厂商提供专门为人工智能工作负载优化的芯片。模型层到应用层的实现主要为两种方式，包括垂直整合基础模型以及在基础模型和微调模型基础上进行应用程序开发两种方式，相当于 AIGC 的平台。在所有层级中，基础设施层通常被认为是最成熟、稳定和商业化的。

图 5：AIGC 技术栈



资料来源：德勤人工智能研究院《人工智能的新篇章：生成式人工智能对企业的影响和意义》，天风证券研究所

表 4：基础设施层提供商

供应商	描述	案例
云服务提供商	超大规模的和特定目的的计算、储存和网络技术	Amazon 、 Baidu 、 Google 、 Microsoft
生成式人工智能服务提供商	提供专业化服务，加速部署（例如安全、监控、测试、模型隔离）	Amazon、Co:here、Google
芯片供应商	专用的芯片半导体，包括 GPU 和 CPU	AMD、Nvidia

资料来源：德勤人工智能研究院《人工智能的新篇章：生成式人工智能对企业的影响和意义》，天风证券研究所

该层级中重点关注云服务与芯片两个核心资源：

- 1) 云服务提供商**：云服务提供商通过提供超大规模和特定目的的计算、储存和网络技术在基础设施层占据了市场的主导地位。云服务提供商的商业模式通过提供可扩展的计算资源，并采用按消费计价的定价策略被证明是有效的。为了使得 AIGC 的工作负荷更加稳定，云服务提供商已经与模型提供商签署相关承诺，以保证未来的工作。AI 需要机器庞大的计算能力，很多公司转向云服务，通过云服务基础设施来解决算力问题。从市占率来看，目前亚马逊是云服务市场的领头羊，微软、IBM、谷歌和阿里云也具有较高的市场份额。具体来看，Azure 与 OpenAI、Google 与 Anthropic 以及 AWS 与 Stability.ai 已形成重要合作。
- 2) 芯片**：基础设施中另一个快速发展的关键层次是芯片。AI 算力芯片是类 ChatGPT 模型的基石，支撑类 ChatGPT 模型需要大量的算力芯片，其中对 GPU、FPGA、ASIC 需求较大。在这方面，英伟达和 AMD 是行业的领导者。英伟达的 Ampere 和 Hopper 系列 GPU、分别为训练和推理工作负载专门设计，加上英伟达的 Selene 超级电脑计算集群，可以加速训练时间。同时，AMD 的 CDNA2 架构同样也是专门为机器学习应用的超级计算而设计，推动了高性能计算市场的竞争。我们认为，短期内具有大算力、通用性的 GPU 芯片或成为大算力应用首选，未来 GPU 与 ASIC 的界限可能会在较大程度上模糊，形成替代竞争。

4. 投资建议

建议关注云端 AI 相关企业：寒武纪、海光信息（天风计算机团队覆盖）、龙芯中科、紫光国微、复旦微电、安路科技、亚马逊、微软公司（天风海外组覆盖）、谷歌、甲骨文等；

建议关注算力芯片相关企业：英伟达（天风海外组覆盖）、AMD、Intel、景嘉微（天风计算机团队联合覆盖）等。

5. 风险提示

AI 发展及商业化不及预期：AIGC 预期技术迭代及商业化进程可能受到软硬件研发和市场

反馈的影响不达预期。

AI 行业竞争加剧：国内外科技企业布局 AIGC 产业链，可能导致 AI 行业供给快速增加，导致行业竞争超出预期。

政策不确定性：AIGC 行业未来可能受到监管对于数据安全、版权等方面的限制。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

除非另有规定，本报告中的所有材料版权均属天风证券股份有限公司（已获中国证监会许可的证券投资咨询业务资格）及其附属机构（以下统称“天风证券”）。未经天风证券事先书面授权，不得以任何方式修改、发送或者复制本报告及其所包含的材料、内容。所有本报告中使用的商标、服务标识及标记均为天风证券的商标、服务标识及标记。

本报告是机密的，仅供我们的客户使用，天风证券不因收件人收到本报告而视其为天风证券的客户。本报告中的信息均来源于我们认为可靠的已公开资料，但天风证券对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，天风证券及/或其关联人员均不承担任何法律责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，天风证券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。天风证券的销售人员、交易人员以及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。天风证券没有将此意见及建议向报告所有接收者进行更新的义务。天风证券的资产管理部门、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。

特别声明

在法律许可的情况下，天风证券可能会持有本报告中提及公司所发行的证券并进行交易，也可能为这些公司提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。因此，投资者应当考虑到天风证券及/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一参考依据。

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	买入	预期股价相对收益 20%以上
		增持	预期股价相对收益 10%-20%
		持有	预期股价相对收益 -10%-10%
		卖出	预期股价相对收益 -10%以下
行业投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	强于大市	预期行业指数涨幅 5%以上
		中性	预期行业指数涨幅 -5%-5%
		弱于大市	预期行业指数涨幅 -5%以下

天风证券研究

北京	海口	上海	深圳
北京市西城区佟麟阁路 36 号	海南省海口市美兰区国兴大道 3 号互联网金融大厦	上海市虹口区北外滩国际客运中心 6 号楼 4 层	深圳市福田区益田路 5033 号平安金融中心 71 楼
邮编：100031	A 栋 23 层 2301 房	邮编：200086	邮编：518000
邮箱：research@tfzq.com	邮编：570102	电话：(8621)-65055515	电话：(86755)-23915663
	电话：(0898)-65365390	传真：(8621)-61069806	传真：(86755)-82571995
	邮箱：research@tfzq.com	邮箱：research@tfzq.com	邮箱：research@tfzq.com