

电子行业深度报告

AI 系列深度: ChatGPT 掀起算力潮, 芯片及服务器上游投资价值凸显

增持 (维持)

2023 年 04 月 03 日

证券分析师 张良卫

执业证书: S0600516070001

021-60199793

zhanglw@dwzq.com.cn

证券分析师 马天翼

执业证书: S0600522090001

maty@dwzq.com.cn

证券分析师 唐权喜

执业证书: S0600522070005

tangqx@dwzq.com.cn

证券分析师 周高鼎

执业证书: S0600523030003

zhougd@dwzq.com.cn

投资要点

■ **ChatGPT 爆火, 其能力已接近人类水平。**GPT 升级至四代, 模型能力高速提升。2023 年 3 月 15 日, OpenAI 正式官宣了多模态大模型 GPT-4, ChatGPT4 将输入内容扩展到 2.5 万字内的文字和图像, 较 ChatGPT 能够处理更复杂、更细微的问题。最新版的 GPT-4 在 ChatGPT 的 GPT-3.5 基础上主要提升了语言模型方面的能力, 并添加了多模态功能, 在不同语言情景和内部对抗性真实性评估的表现都显著优于 GPT-3.5, 在各种专业和学术基准上已经表现出人类水平。

■ **以搜索引擎和办公软件为支点, 下游应用多点开花。**OpenAI+传统搜索引擎=AI 智能搜索引擎。ChatGPT 能生成和理解类人类思维方式与上下文联系的结果, ChatGPT 都对传统搜索引擎是一种颠覆式进化。Office 方面: OpenAI+办公软件, 以 Office 全家桶与微软 Teams premium ChatGPT 为例。ChatGPT+Office 三大件 Word、Excel、PowerPoint 可完成从文字创作到可视化全程服务。ChatGPT 可以根据用户的描述需求自动撰写文档内容, 缩短文档的创作时间; 根据已形成文档内容智能推荐文档格式、排版等; 并根据用户描述制作个性化 PPT, 同时智能生成朗读语言助力 PPT 演示。

■ **算力需求带动 AI 服务器需求量上涨。**GPT4 作为大型多模态模型支持图像和文本的输入, 背后是大语言模型的演进与参数量、训练数据量的指数级增长。GPT4 频频下调提问限制次数传达算力不足信息, 大语言模型的快速迭代催生大量算力需求。满足算力需求依托叠加 AI 服务器数量与提升服务器算力性能两种途径, 从而支撑大语言模型发挥预期功效。我们率先提出建立推理/训练服务器需求测算模型: 1) 考虑大语言模型迭代呈现参数量指数级增长趋势, GPT-4 参数量为万亿, 此基础上我们测算得出 **GPT-4 推理发挥预期功效所需服务器数量为 6652 台**, 百万亿参数基础上, 所需服务器数量将上升至 **66 万台**; 2) 根据硅谷披露 GPT-4 训练 Token 数 9 万亿, 我们测算得出 **GPT-4 训练所需服务器数量为 1391 台**。未来随着大语言模型不断迭代、GPT 等语言模型渗透率不断提高, 算力时代下服务器呈现基建化趋势。

■ **AI 服务器核心依托高性能芯片, 国产芯片自主发展潜力大。**AI 芯片中 GPU 凭借并行计算具备强大算力, 占据 AI 芯片大部分市场; 美企英伟达独占鳌头, 占据 80%GPU 市场。相比英伟达产品, 国产芯片性能存在较大差距, 仍处于追赶阶。受美实体清单影响, 高端芯片采购受阻, 倒逼国内服务器厂商大力推进国产替代, 利好国产芯片自主发展。

■ **投资建议:** 我们看好算力芯片及服务器需求剧增下相关零部件的成长空间。**CPU\GPU 等:** 寒武纪、景嘉微、海光信息; **存储:** 江波龙、聚辰股份; **电源芯片:** 晶丰明源、杰华特; **封测:** 长电科技、通富微电; **PCB 及其他:** 深南电路、沪电股份、鼎通科技、裕太微、澜起科技、工业富联

■ **风险提示:** LLM 模型发展不及预期、行业空间测算不及预期、AI 芯片发展不及预期。

行业走势



相关研究

《技术创新系列深度-苹果 MR 搭载前沿技术, 引领产业风向》

2023-03-16

《充电桩行业高速增长, 上游元器件需求弹性提升》

2023-03-10

内容目录

1. ChatGPT 快速迭代，OpenAI 模型演进，下游应用百花齐放.....	5
1.1. GPT 迭代更新，人工智能掀起科技潮。.....	5
1.2. OpenAI 模型演进，人工智能向应用拓展.....	6
1.3. Open AI 赋能下游应用，AI+降本增效	7
2. 云服务器优化配置算力资源，是主要算力供给方式	9
3. ChatGPT 引爆算力需求，拉动高端芯片市场扩张.....	11
3.1. 人工智能架构以 AI 芯片为基	11
3.2. GPT 拉动算力提升，带动高算力芯片市场量价齐升	13
3.3. 构建推理/训练服务器需求计算模型，测算未来空间广阔	15
4. 算力时代凸显芯片价值，国产芯片仍处追赶阶段	19
4.1. AI 服务器依赖高性能芯片供给.....	19
4.2. 主流芯片限制进口，利好国产化芯片自主发展.....	22
5. 标的梳理及投资建议	23
6. 风险提示	26

图表目录

图 1: GPT-4 和 GPT-3.5 的考试表现	5
图 2: OpenAI 网站 12-2 月访问量	5
图 3: GPT-4 理解图片含义	6
图 4: GPT-4 理解图表中数据的含义并做进一步计算	6
图 5: AIGC 相关技术场景及成熟度分类	7
图 6: Meta AI 的 Make-A-Video 根据文字生成视频	7
图 7: Satya 宣布 ChatGPT 加入 Azure OpenAI	8
图 8: Azure OpenAI 全家桶	8
图 9: ChatGPT+Bing	8
图 10: Teams Premium ChatGPT 智能回顾	9
图 11: Teams Premium 生成任务	9
图 12: 云计算系统	9
图 13: 云计算服务模式	9
图 14: 云计算的 IAAS 与传统服务器资源配置对比	10
图 15: 全球云计算细分市场规模及增速 (单位: 亿美元)	11
图 16: 云计算下游应用 (单位: 亿元)	11
图 17: 人工智能 IT 技术栈	11
图 18: CPU 架构及优缺点	12
图 19: GPU 架构及优缺点	12
图 20: GPU、FPGA 和 ASIC 芯片对比	13
图 21: 训练 Transformers 的算力要求	14
图 22: 模型算力预测	14
图 23: 2020-2027 全球 GPU 市场规模 (单位: 亿美元)	14
图 24: 2019-2025 中国 AI 芯片市场规模 (单位: 亿元)	14
图 25: 推理及训练所需服务器计算框架	15
图 26: the decoder 报告 GPT-4 模型参数数据	16
图 27: 硅谷猜测 GPT-4 训练数据 Token 量	16
图 28: 推理所需服务器测算	17
图 29: 推理每秒处理 Token 测算	17
图 30: 训练所需服务器测算	18
图 31: 每秒训练 Token 测算	18
图 32: 2021 年中国 AI 服务器市场份额	19
图 33: 2021 年全球 AI 服务器市场份额	19
图 34: GPU 并行计算架构适配 AI 模型建构	20
图 35: 2021 年 H1 中国 AI 芯片市场占比	20
图 36: 2021-2022Q2 全球独显 GPU 市场各产商份额占比	21
表 1: 历代 GPT 表现情况	5
表 2: 英伟达主要 GPU 芯片价格	15
表 3: ChatGPT 所需推理服务器数量计算	16

表 4: 每秒处理 token 测算过程.....	17
表 5: GPT-4 训练所需服务器数量计算.....	17
表 6: 每秒训练 token 测算过程.....	18
表 7: 主要 LLM 基本信息表.....	18
表 8: 英伟达个芯片性能指标均为世界领先.....	21
表 9: 海内外 GPU 产品性能对比.....	22
表 10: 美国 BIS《对向中国出口的先进计算和半导体制造物项实施新的出口管制》.....	22
表 11: 国内服务器 CPU、GPU 芯片厂商业务情况.....	23
表 12: 国内服务器内存和硬盘相关厂商业务情况.....	24
表 13: 国内服务器电源芯片相关厂商业务情况.....	25
表 14: 国内封测相关厂商业务情况.....	25
表 15: 服务器其他相关零部件厂商业务情况.....	26

1. ChatGPT 快速迭代，OpenAI 模型演进，下游应用百花齐放

GPT-4 是一个大型的多模态模型，相比上一代 ChatGPT 新增了图像功能，同时具备更精准的语言理解能力。GPT 的升级背后是 OpenAI 的大语言模型的进一步演进，同时带动下游应用的拓展，涌现出新一批应用场景。

1.1. GPT 迭代更新，人工智能掀起科技潮。

GPT 升级至四代，模型能力高速提升。ChatGPT 是由 OpenAI 开发的自然语言生成模型，采用 Transformer 神经网络架构（又称 GPT-3.5 架构），基于大量的语料库使用指示学习和人工反馈的强化学习（RLHF）来指导模型训练。模型可理解并生成对各种主题的类似人类的响应，是 AIGC 技术进展的成果。2023 年 3 月 15 日，OpenAI 正式官宣了多模态大模型 GPT-4，ChatGPT4 将输入内容扩展到 2.5 万字内的文字和图像，较 ChatGPT 能够处理更复杂、更细微的问题。

表1: 历代 GPT 表现情况

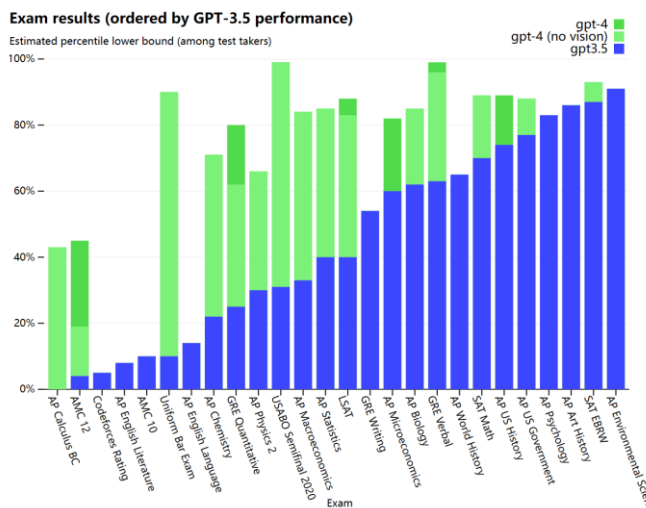
模型	发布时间	参数量	预训练数据量	Tokens	学习目标
GPT-1	2018 年 6 月	1.17 亿	约 5GB	1.3B	结合无监督学习及有监督的微调
GPT-2	2019 年 2 月	15 亿	40GB	15B	学习在无明确监督情况下执行多种任务
GPT-3	2020 年 5 月	1,750 亿	45TB	499B	结合少样本学习和无监督学习
GPT-4	2023 年 3 月		待公布		基于规则的奖励模型

数据来源：OpenAI，Medium,东吴证券研究所

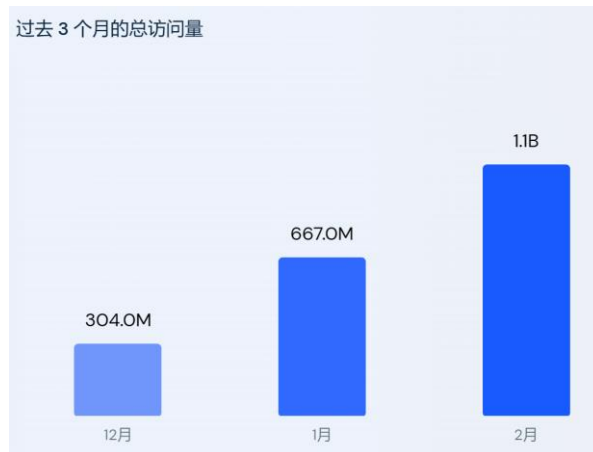
ChatGPT 提供变革性的用户体验，用户数量飙升。ChatGPT 发布后爆火，仅用 5 天时间用户量便破百万，推出 2 个月后用户量破亿，访问量从 1830 万增长到 6.72 亿，成为史上用户增长速度最快的消费级应用程序。最新版的 GPT-4 在 ChatGPT 的 GPT-3.5 基础上主要提升了语言模型方面的能力，并添加了多模态功能，可以接受图像输入并理解图像内容，可接受的文字输入长度也增加到 3.2 万 token，在不同语言情景和内部对抗性真实性评估的表现都显著优于 GPT-3.5，在各种专业和学术基准上已经表现出人类水平，为用户提供变革性的使用体验。

图1: GPT-4 和 GPT-3.5 的考试表现

图2: OpenAI 网站 12-2 月访问量



数据来源：OpenAI，东吴证券研究所



数据来源：Similarweb、东吴证券研究所

ChatGPT 带动 AI 潮流，多种相关产品推出。办公领域，微软将 GPT-4 整合到 Office 应用程序，Teams Premium 中接入 Chat GPT 提供人工智能生成章节和字幕实时翻译等功能；编程领域中，Viva Sales 将利用 ChatGPT 为电子邮件中客户问题生成回复建议，Stripe 使用 GPT-4 扫描商业网站并向客户支持人员提供摘要；软件领域，Duolingo 将 GPT-4 构建到新的语言学习订阅层中，国内百度“文心一言”也正式推出，AI 的潮流开始遍布国内外多行业。

1.2. OpenAI 模型演进，人工智能向应用拓展

OpenAI 模型可以完成绝大多数文本和图像任务，具有成熟的生产力。根据 OpenAI 的官方介绍，GPT-4 可以接受文本和图像的信息，并允许用户指定任何图像或语言任务，处理超过 25000 个单词的文本。除了在各种标准考试和不同语言情景下都有突出的表现外，在图像的处理分析上，GPT-4 能够直接阅读并分析带有图片的论文，承担文本、音频、图像的生成和编辑任务，并能与用户一起迭代创意和技术写作任务，例如创作歌曲、编写剧本或学习用户的写作风格等。

图3：GPT-4 理解图片含义

图4：GPT-4 理解图表中数据的含义并做进一步计算

GPT-4 visual input example, Chicken Nugget Map:

Can you explain this meme?

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



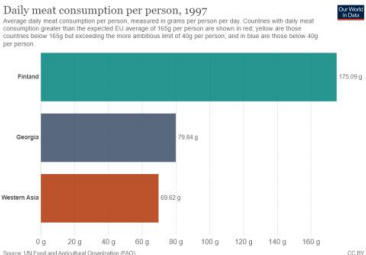
This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the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

Table 18: Example prompt demonstrating GPT-4's visual input capability.

GPT-4 visual input example, Chart Reasoning:

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

Daily meat consumption per person, 1997



Region	Daily meat consumption per person (g)
Finland	179.09
Georgia	79.84
Western Asia	69.62

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

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

Step 1: From the provided chart, we see that the average daily meat consumption for Georgia is 79.84 grams per person per day.
Step 2: From the chart, we see that the average daily meat consumption for Western Asia is 69.62 grams per person per day.
Step 3: Add the values from steps 1 and 2.
 $79.84\text{ g (Georgia)} + 69.62\text{ g (Western Asia)} = 149.46\text{ g}$

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

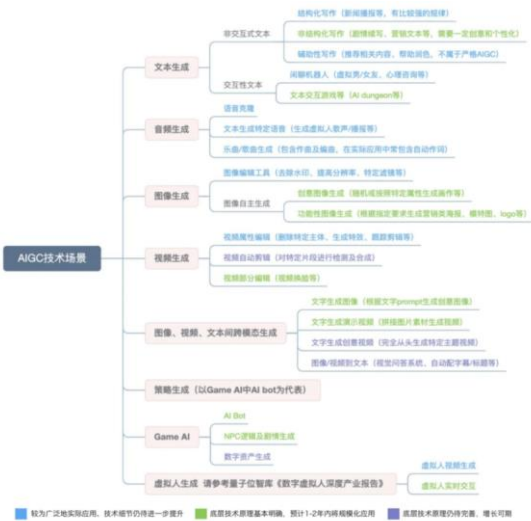
Table 14: Example prompt demonstrating GPT-4's visual input capability. The prompt consists of a question which requires reading information from a chart and performing computation.

数据来源：OpenAI，东吴证券研究所

数据来源：OpenAI，东吴证券研究所

AIGC 发展改革生产力，已具备多领域应用能力。AIGC 可以利用大量无标注数据进行自监督学习再利用少量的标注数据进行迁移学习，能够持续生成规模大、质量高、单位成本低的内容，在生产力上具有革命性的增长。在应用方面，按场景分类 AIGC 已经较为成熟地应用于文本和代码撰写、图像识别和生成，在影视、传媒、电商、C 端娱乐领域规模应用，未来将逐步拓展到视频和游戏等其他领域，为各个行业和领域的创新和发展提供更多可能性。

图5：AIGC 相关技术场景及成熟度分类



数据来源：量子位，东吴证券研究所

图6：Meta AI 的 Make-A-Video 根据文字生成视频

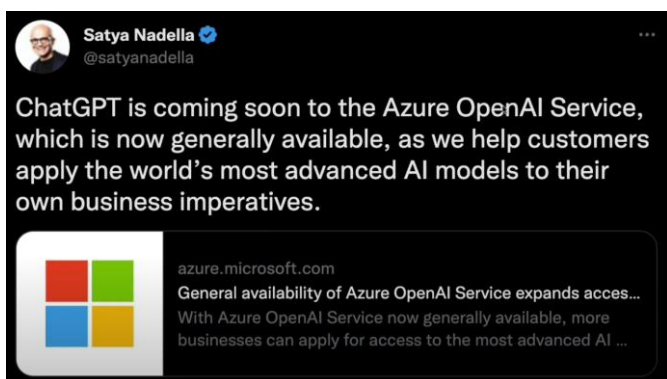


数据来源：OpenAI，东吴证券研究所

1.3. Open AI 赋能下游应用，AI+降本增效

以 Azure OpenAI 为例：AI 技术模型全覆盖。AI 技术模型全家桶，包含 GPT4.0 以及后续版本、DALL-E2、Codex、Embeddings 神经网络、VALL-E、认知学习成长、机器学习等，可实现 4 种用途：1) 企业及客户可利用 Azure OpenAI 全家桶构建自己的大型预训练 AI 模型；2) 利用全家桶对 AI 模型进行数据和参数的微调与提升；3) 通过内置全家桶检测错误与潜在风险；4) 利用全家桶实现角色访问控制和企业级云安全合规。

图7: Satya 宣布 ChatGPT 加入 Azure OpenAI



数据来源: Twitter, 东吴证券研究所

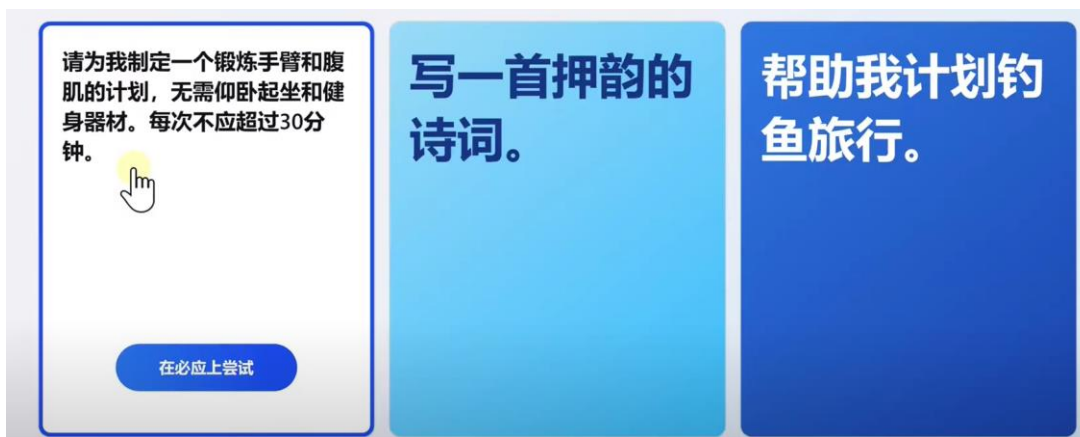
图8: Azure OpenAI 全家桶



数据来源: OpenAI, 东吴证券研究所

OpenAI+传统搜索引擎=AI 智能搜索引擎。ChatGPT 能生成和理解类人类思维方式与上下文联系的结果，而传统的搜索引擎只是给出网页的链接，需要用户自己阅读、点击和判定。无论是知识的获取效率、回答问题的深度，还是交互的延展性，ChatGPT 都对传统搜索引擎是一种颠覆式进化。

图9: ChatGPT+Bing

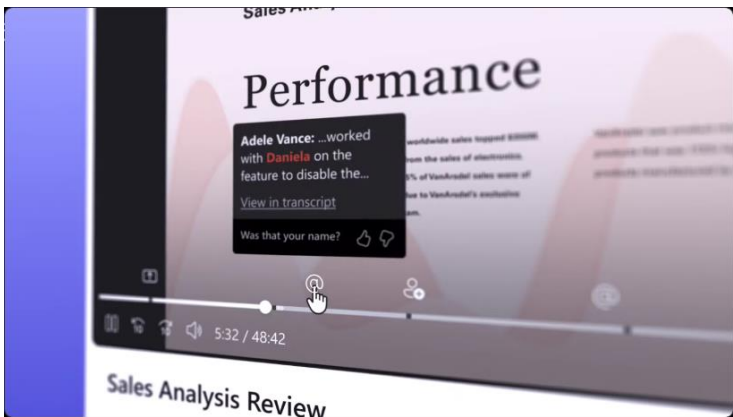


数据来源: Bing, 东吴证券研究所

OpenAI+办公软件，以 Office 全家桶与微软 Teams premium ChatGPT 为例。ChatGPT+Office 三大件 Word、Excel、PowerPoint 可完成从文字创作到可视化全程服务。ChatGPT 可以根据用户的描述需求自动撰写文档内容，缩短文档的创作时间；根据已形成文档内容智能推荐文档格式、排版等；并根据用户描述制作个性化 PPT，同时智能生成朗读语言助力 PPT 演示。Teams Premium ChatGPT 同样具有多样智能服务：1) 自动

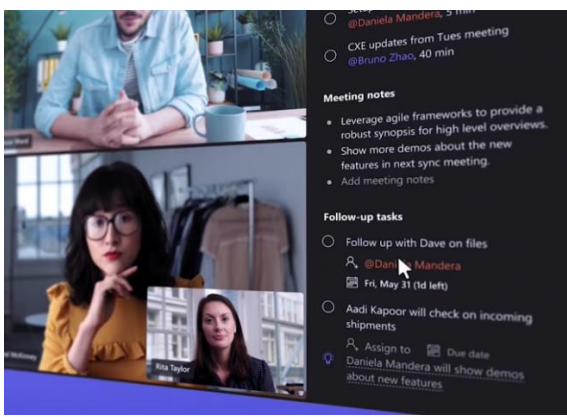
提供实时翻译；2）自动生成文本型会议纪要；3）自动记录发言人员对话和内容；4）根据讨论上下文自动生成任务等。

图10: Teams Premium ChatGPT 智能回顾



数据来源：O365，东吴证券研究所

图11: Teams Premium 生成任务



数据来源：O365，东吴证券研究所

2. 云服务器优化配置算力资源，是主要算力供给方式

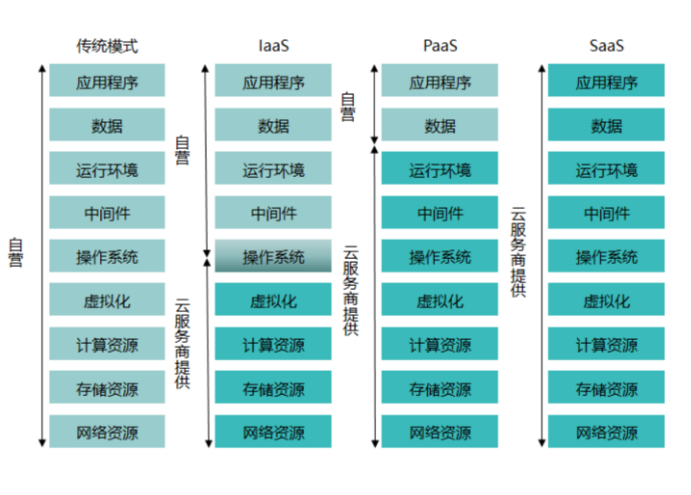
云计算属于分布式计算，满足各类算力需求。云计算可以通过网络“云”将所运行的巨大的数据计算处理程序分解成无数小程序，再交由计算资源共享池进行搜寻、计算及分析后，将处理结果回传给用户，通过这种方式可以实现随时随地、便捷地、按需应变地从可配置计算资源共享池中获取所需的资源与架构，包括存储、数据库、服务器、应用软件及网络等，灵活调配计算机相关算力存储资源。

图12: 云计算系统



数据来源：Wiki，东吴证券研究所

图13: 云计算服务模式



数据来源：CSDN，东吴证券研究所

AI 语言模型扩大算力需求，云计算进一步发挥资源优化配置功效。面对 AI 的需求提升，云计算可带来 4 个方面的提升：

1.资源共享和优化：云计算允许多个用户共享同一台服务器或一组服务器的计算资源。这意味着，与每个用户都拥有自己的独立服务器相比，云计算可以通过提高资源利

用率来减少服务器的总数量。

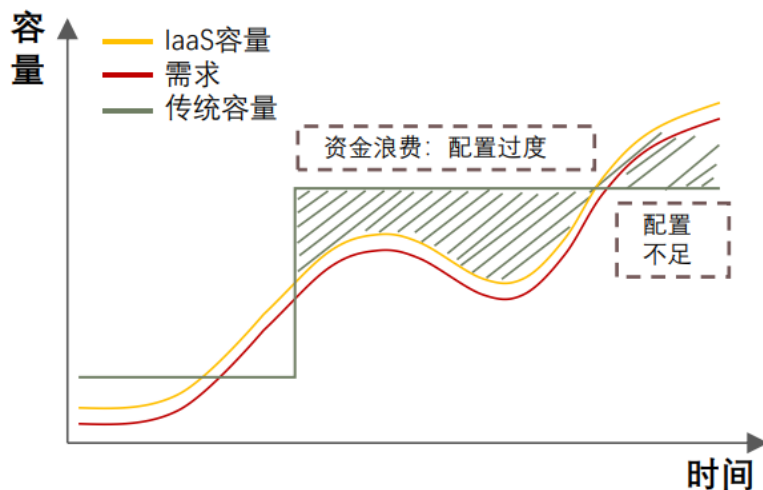
2.弹性可伸缩：云计算服务提供了高度的弹性和可伸缩性，使得用户可以根据实际需求调整所需的计算资源。在计算需求降低时，用户可以减少资源使用，从而减少服务器的数量。而在计算需求增加时，用户可以快速扩展资源，而无需购买新的服务器。

3.成本效益：云计算通常采用按需付费的模式，用户只需支付实际使用的计算资源。这种支付模式降低了用户购买和维护服务器的成本，使得更多企业倾向于使用云计算服务而不是购买自己的服务器。

4.更快的创新和部署：云计算使得企业能够更快地开发和部署新的应用程序和服务。这减少了对于购买、配置和维护自有服务器的需求，从而抑制了服务器数量的增长

云计算为 AI 提供算力支持，实现降本增效。由于 ChatGPT 等一系列大规模模型对算力要求较高，厂商自建足够算力的数据中心需要耗费巨大的成本以及昂贵的后续运维成本，而云计算服务可以有效节省早期支出，并且大规模的算力集群还具有集群优势以提升计算性能。OpenAI 的 ChatGPT 和其他关键 AI 产品依赖于微软 Azure 云计算服务，通过 Azure 的 HPC、弹性、数据存储和计算服务来完成模型计算和算法调试，没有云计算，ChatGPT 的训练时间将被一再拉长，运用云计算后 GPT-3 等模型的训练成本下降了 80% 以上，从首次训练开销 1200 万美元降至 140 万美元。

图14：云计算的 IAAS 与传统服务器资源配置对比

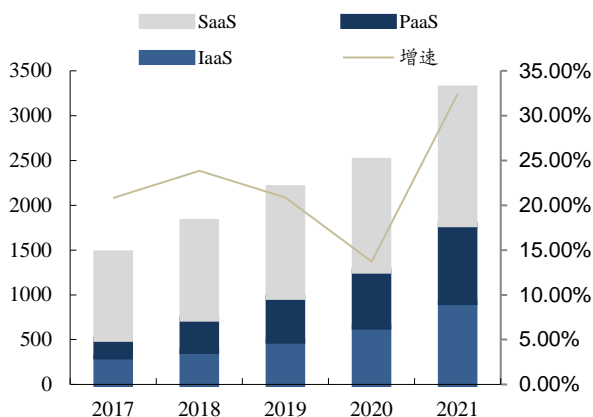


数据来源：头豹研究院，东吴证券研究所

云算力市场规模持续扩大，下游应用渗透提升。云算力市场根据 Gartner 统计，2021 年以 IaaS、PaaS、SaaS 为代表的全球公有云市场规模达到 3307 亿美元，增速达 32.5%。其中，IaaS、PaaS、SaaS 市场规模分别同比增长 42.7%、39.3%、23.6%至 916 亿美元、869 亿美元、1522 亿美元。2021 年亚马逊、微软、阿里云为全球 IaaS 前三厂商，占据 69.54% 市场份额，国内厂商阿里云、华为云、腾讯三家合计占全球 17% 的市场份额。目前国内游戏、视频、电商和金融是主要应用领域，医疗、传统工业和政务领域渗透率逐

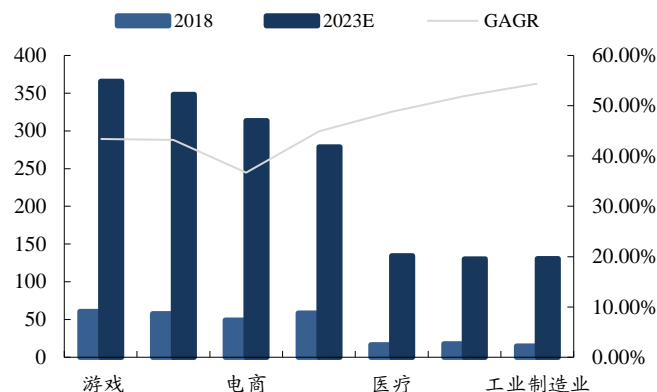
步增加。

图15: 全球云计算细分市场规模及增速(单位: 亿美元)



数据来源: 信通院, 东吴证券研究所

图16: 云计算下游应用(单位: 亿元)



数据来源: 灼识咨询, 东吴证券研究所

AI 与云计算结合, 提升大数据分析计算能力。AI 与云计算结合有助于通过自动化冗余活动、识别、排序和索引各种类型的数据、管理云中的数据事务、识别整个云存储基础设施中的任何故障, 帮助改善数据管理, 优化管理流程。例如阿里云推出的飞天智算平台提供公共云和专有云两种模式, 为各类科研和智能企业机构提供强大的智能计算服务, 可将计算资源利用率提高 3 倍以上, AI 训练效率提升 11 倍, 推理效率提升 6 倍。

云计算的本质是将计算资源进行集中管理和调度, 从而满足不同用户的需求, 而这些计算资源的核心就是**服务器**。因此, 随着云计算需求的提升, 服务器数量的提升也是必然的趋势。AI 为云计算所带来的巨大计算量, 将直接带动服务器数量的提升。

3. ChatGPT 引爆算力需求, 拉动高端芯片市场扩张

3.1. 人工智能架构以 AI 芯片为基

人工智能 IT 技术栈分为四层架构, 算力是基础。人工智能架构由芯片层、框架层、模型层和应用层四层架构组成。其中, 芯片层主要是指人工智能芯片, 为整个架构提供算力基础支撑; 框架层主要包括深度学习训练和推理框架平台和数据平台等; 模型层主要是自然语言处理 (NLP)、计算机视觉 (CV)、音视频、多模态等各种大模型; 应用层则是面向各种应用场景的人工智能专业服务。下游应用体验的升级需要大模型不断调优迭代, 在深度学习框架内进行大规模模型的训练和推理, 每一次都对芯片提供的算力基础提出要求, AI 芯片决定了平台的基础架构和发展生态。

图17: 人工智能 IT 技术栈



数据来源：甲子光年智库，东吴证券研究所

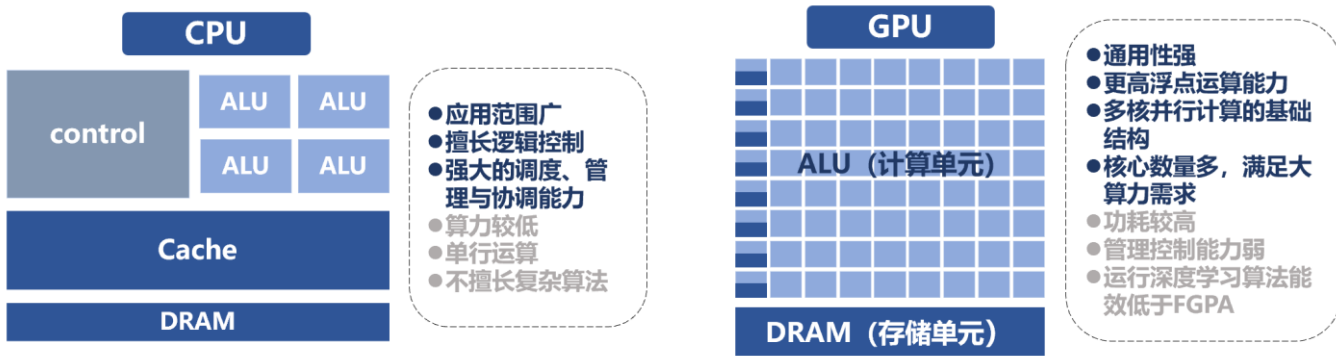
AI 芯片主要有 CPU、GPU、FPGA 及 ASIC，以 GPU 为主的并行计算为 AI 提供了主要生产力

1) CPU (中央处理器) 擅长逻辑控制，用于推理及预测。CPU 主要由 ALU(算术逻辑单元)、CU (控制单元) 与 Cache (临时指令存储器) 构成。其中，ALU 由"And Gate" (与门) 和"Or Gate" (或门) 构成的算术逻辑单元，主要功能是进行二位元的算术运算，约占 CPU 空间 20%；CU 则负责程序的流程管理。CPU 的执行周期是从内存中提取第一条指令、解码并决定其类型和操作数，执行，然后再提取、解码执行后续的指令，重复循环直到程序运行完毕。CPU 具备强大的调度、管理与协调能力，但受限于单行运算导致算力较低。

2) GPU 图形处理器，并行计算满足强大算力需求。在结构方面，与 CPU 相比，GPU 中同样具备 ALU 与 CU 等架构，但 ALU 数量与体积占比更多，且采用数量众多的计算单元和超长流水线，具备高并行结构，通过多核并行计算支撑大算力需求，且拥有更高浮点运算能力，在处理图形数据和复杂算法方面拥有比 CPU 更高的效率，满足深度学习领域海量数据运算的需求。但 GPU 管理控制能力弱，无法单独工作，需由 CPU 进行控制调用。

图18: CPU 架构及优缺点

图19: GPU 架构及优缺点



数据来源：东吴证券研究所绘制

数据来源：东吴证券研究所绘制

3) **FPGA 半定制化芯片，无限次编程支持高灵活性。**FPGA 称为现场可编程门阵列，基本结构包括可编程输入输出单元和可配置逻辑块等。基于其结构，用户可根据自身的需求进行重复编程重组电路，**具有高度实时性和灵活性，可以同时进行数据并行和任务并行计算**，在处理特定应用时有更加明显的效率提升，可以实现底层硬件控制操作技术，为算法的功能实现和优化留出了更大空间。

4) **ASIC 全定制化芯片，高速处理速度叠加低功耗表现。**ASIC 是为专门目的为设计的集成电路，为实现特定要求而定制的专用 AI 芯片，能够在**特定功能上进行强化，具有更高处理速度与更低能耗**。缺点是研发成本高、前期研发投入周期长，且由于定制化属性，可复制性一般，缺乏一定灵活性。深度学习算法稳定后，AI 芯片可采用 ASIC 设计方法进行全定制，使性能、功耗和面积等指标面向深度学习算法做到最优。

图20: GPU、FPGA 和 ASIC 芯片对比

芯片	定制化程度	功耗	优点	缺点	主要厂商
GPU	通用性	大	峰值计算能力强 产品开发生态成熟	单价高、功耗高	英伟达、AMD
FGPA	半定制化	较大	可编程灵活度高、落地周期短迭代快	单价较高、峰值计算能力较低、编程语言难度较大	Xilinx、英特尔
ASIC	全定制化	小	平均性能强、功耗低	前期投入大、不可编程、灵活度低、研发周期较长	谷歌、寒武纪、地平线

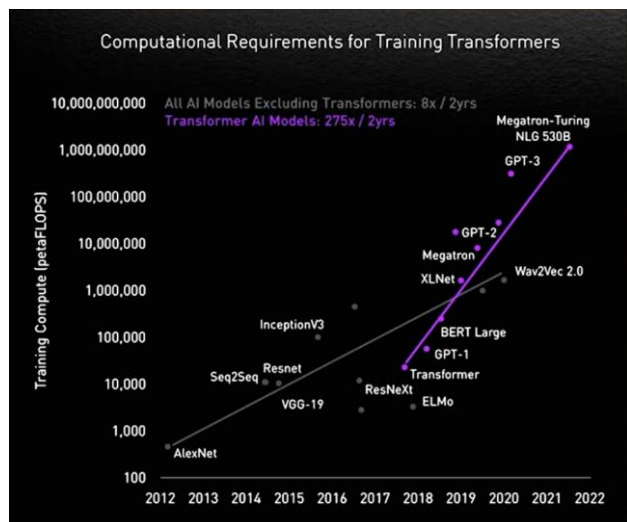
数据来源：艾瑞咨询，东吴证券研究所整理

3.2. GPT 拉动算力提升，带动高算力芯片市场量价齐升

ChatGPT 数据量扩张，算力需求持续增加。历代 GPT 的参数量呈现指数级增长，ChatGPT 的总算力消耗约为 3640PF-days(每秒计算一千万亿次，需要计算 3640 天);GPT-4 在 ChatGPT 的基础上增加了图像、视频等交互信息类型，内容容量扩大到 2.5 万字，

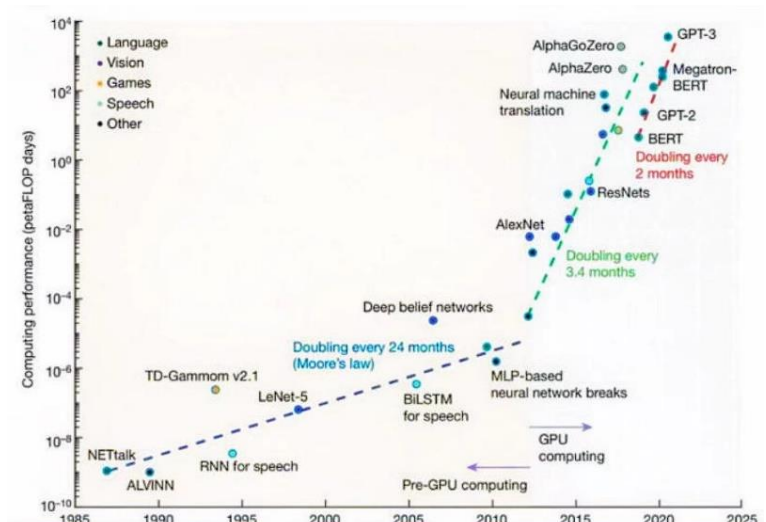
所需的算力规模远大于单纯的文字交互。OpenAI 首席执行官 Sam Altman 接受公开采访时表示，GTP-5 在 2024 年底至 2025 年发布，它的参数数量为 GTP-3 的 100 倍，需要的计算量为 GTP-3 的 200-400 倍。随着 ChatGPT 的用户和应用范围的持续扩大，数据处理的压力增大，将提出庞大的算力需求。

图21: 训练 Transformers 的算力要求



数据来源：英伟达官网，东吴证券研究所

图22: 模型算力预测



数据来源：Nature，东吴证券研究所

算力需求带动高算力芯片市场，AI 芯片市场规模持续扩张。机器所产生的数据量在 2018 年首次超越人类所创造的数据量，从 2019 年，每年几乎以倍数的幅度来增加，从 2020 年到 2025 年，全球数据增量将达到 157Zetabytes，5 年 GAGR 高达 89%。庞大的数据增量，必须运用各种具备高速运算的人工智能芯片来过滤、处理分析、训练及推理，这将持续带动 AI 芯片，尤其是高算力芯片需求。根据华经产业研究院发布的《2021-2026 年中国 GPU 行业发展监测及投资战略规划研究报告》，2020 年全球 GPU 行业市场规模达 254.1 亿美元，预计 2027 年将达到 1853.1 亿美元，复合年均增长率 32.82%，GPU 市场保持高速增长态势。同时根据亿欧智库预测，2025 年我国 AI 芯片市场规模将达到 1780 亿元，2019-2025GAGR 可达 42.9%。

图23: 2020-2027 全球 GPU 市场规模 (单位: 亿美元)

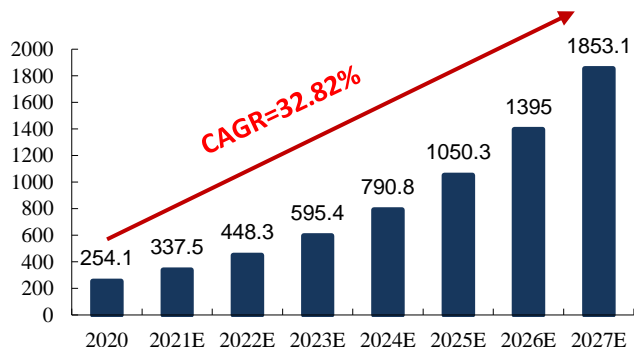
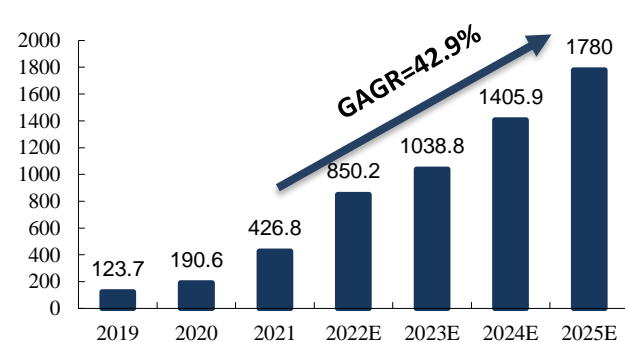


图24: 2019-2025 中国 AI 芯片市场规模 (单位: 亿元)



数据来源：Verified Market Research，东吴证券研究所

数据来源：亿欧智库，东吴证券研究所

AI 风潮席卷，芯片价值量有望提升。AI 技术的蓬勃发展和广泛应用，导致对高性能计算能力的需求空前旺盛。AI 芯片作为行业的核心组件，其价格也随之攀升，成为科技产业的新增长点。随着 AI 模型复杂度的增加和参数量的指数级扩张，对计算能力的要求不断提高。例如预计在 2024 年底至 2025 年发布的 GPT-5，其参数量将是 GPT-3 的 100 倍，所需算力为 GPT-3 的 200-400 倍。由于高性能 AI 芯片在满足这种日益增长的算力需求方面具有不可替代的作用，算力快速增长需求下，芯片供给的不确定性有望提升芯片价格提升。

表2：英伟达主要 GPU 芯片价格

产品	单价（万元）	产品	单价（万元）
英伟达 V100 16G	3.9	英伟达 A100 40G	6.4
英伟达 V100 32G	6.2	英伟达 A100 80G	10.4
英伟达 A800 80G	8.3	英伟达 H100 PCIe	25.2

数据来源：Thinkmate，东吴证券研究所

3.3. 构建推理/训练服务器需求计算模型，测算未来空间广阔

算力紧缺，ChatGPT 频繁下调提问限制次数。鉴于使用人数过多，GPT-4 提问上限次数先后从每 4 小时的使用上限从 100 次下调到 50 次，近日 ChatGPT Plus 又一次将 GPT-4 提问次数下调至每 3 小时 25 次，同时预告再会继续下调提问次数。

GPT-4 发挥预期效果需服务器数量与芯片所提供算力支撑。GPT4 推理与运算发挥预期能力所需服务器数量主要依赖于 2 个核心变量，模型参数与训练/推理 Token 量，不考虑 GPT4 当前回答次数限制与图像处理功能，在峰值容纳一切原则下，计算框架大致如下：

推理所需服务器=每秒处理峰值 Token 所需 FLOPS/（一台服务器提供的 FLOPS*推理下 FLOPS 利用率）；

训练所需服务器=规定时长内训练完所有 Token 所需 FLOPS/（一台服务器提供的 FLOPS*训练过程 FLOPS 利用率）

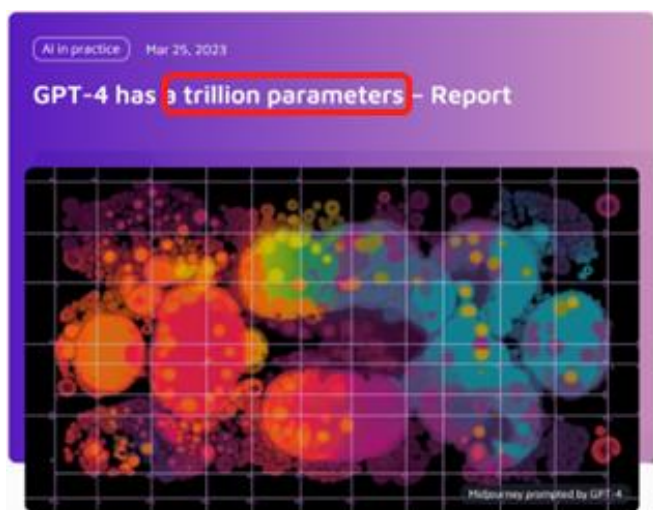
图25：推理及训练所需服务器计算框架



数据来源：东吴证券研究所整理

OpenAI 官方尚未完全公开 GPT4 数据，根据 the Decoder 报告，此次 GPT-4 是模型参数数量为 1 万亿（GPT3 参数为 1750 亿）；据硅谷猜测，GPT-4 用于训练数据量为 9 万亿 Token。

图26: the decoder 报告 GPT-4 模型参数数据



数据来源：the Decoder，东吴证券研究所

图27: 硅谷猜测 GPT-4 训练数据 Token 量

硅谷王川 Chuan @Svwang1
GPT-4 用于训练的数据量据说在 5 到 9 万亿 token 之间。假设是 9 万亿，这就等价于 45 terabyte. 普通一本英文书的数据量大约在 1 MB (一个英文字母相当于一个 byte), 所以 45 TB 相当于四千五百万本书。而美国国会图书馆收藏的书籍到2021年底不到四千万。另外一个参考数据，全世界每年出版的新书，大约两百万左右。

数据来源：Twitter，东吴证券研究所

基于英伟达 A100 进行测算，GPT4 发挥预期能力下推理过程所需服务器数量约为 6652 台。我们通过单个模型参数单个 token 所需要的 FLOP 次数、模型参数、每秒处理 token 数量、服务器所需 GPU 数量（按照 DGX A100 进行测算）、单个 GPU 峰值 FLOP 次数以及利用率为 21.6%进行测算，得到推理所需服务器结果。

表3: ChatGPT 所需推理服务器数量计算

单个模型参数、 单个 Token 需要 的 FLOPs*	模型参数*	每秒处理 token*	单个服务器 GPU 数量/	单个 GPU FLOPs/	FLOPs 利用率/	推理所需服 务器数量
2	1E+12	1,768,261	8	3.12E+14	21.3%	6652 台

*OpenAI 论文给出：Transformer 模型推理过程中每 token 计算所需 FLOPs 为 2N，N 为参数数量；

*据 the Decoder 报告，GPT-4 模型参数数量为 1 万亿；

*一台英伟达 DGX A100 搭载 8 片 GPU；

*根据英伟达官网信息：每颗 A100GPU 提供峰值 312 TFLOPS FP16/FP32 混合精度吞吐量。由于 FP16 的开销较低，混合精度既支持 FLOPS 吞吐量较高，而维持精确结果要求的数值稳定性将是恒定的；

*假定模型的训练 FLOPS 利用率为 21.3%，与训练期间的 GPT-3 保持一致

数据来源：OpenAI、英伟达、东吴证券研究所测算

其中关键假设为每秒处理 token 数量，其测算过程主要为每日提问人数*每人提问数量*每个问题回答字数，同时按照中英文翻译比重 1.6: 1 为基准，在每个单词 token 为 1.3 前提下计算，得到每秒处理 177 万 token。且当前数据建立在 GPT4 提问次数受限的现实条件与 OpenAI 网站 2 月份日活数据下，等待 GPT4 发挥预期功能，开放图像处理能力，日活、提问数量、回答字数等假设量将迅速增长，所需服务器仍存在增长空间。

表4：每秒处理 token 测算过程

每日提问人数*	每人提问数量*	每个问题+ 回答中文 字数*	中翻英比/ 每单词 token*	每日秒数/	每秒处理 token
36,666,667	10	500	1.6	86400	1,768,261

*根据 Similarweb 统计，OpenAI 网站 2 月总访问量达 1.1B，平均每天访问量 36,666,667；

*假设每人单日提问为 10 次；

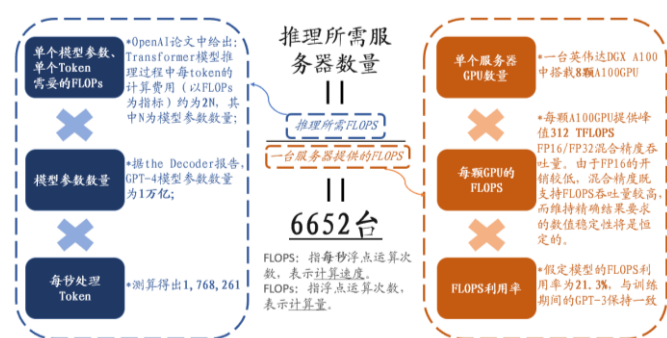
*假设每轮问答共生成 500 字；

*在一般的中英翻译文档中，中文文档的字数与英文文档单词数比例在 1.6: 1 左右；

* ChatGPT 的 GPT3.5 模型推理的价格约为 0.02 美元/750 个单词，其中每 1000token 约等于 750 个单词，即每单词对应 1.33Token

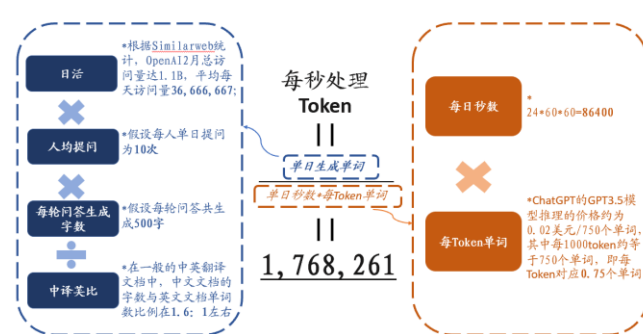
数据来源：Similarweb，东吴证券研究所测算

图28：推理所需服务器测算



数据来源：东吴证券研究所整理

图29：推理每秒处理 Token 测算



数据来源：东吴证券研究所整理

在 6 个月训练时间假设下 GPT4 训练需要服务器数量为 1391 台。同样根据通过单个模型单 token 所需要的 FLOP 次数、模型参数、每秒处理 token 数量、服务器所需 GPU 数量（按照 DGXA100 进行测算）、单个 GPU 峰值 FLOP 次数以及利用率为 46.2%进行测算，得到推理结果。

表5：GPT-4 训练所需服务器数量计算

单个模型参数、 单个 Token 需要 的 FLOPs*	模型参数*	每秒训练 token*	单个服务器 GPU 数量/	单个 GPU FLOPs/	FLOPS 利用率/	训练所需服务 器数量
6	1E+12	578704	8	3.12E+14	46.2%	1391 台
* OpenAI 论文中给出：Transformer 模型推理过程中每 token 的计算 FLOPs 约为 6N，其中 N 为参数数量； *假定模型的训练 FLOPS 利用率为 46.2%，与在 TPUv4 芯片上进行训练的 PaLM 模型一致						

数据来源：OpenAI、东吴证券研究所测算

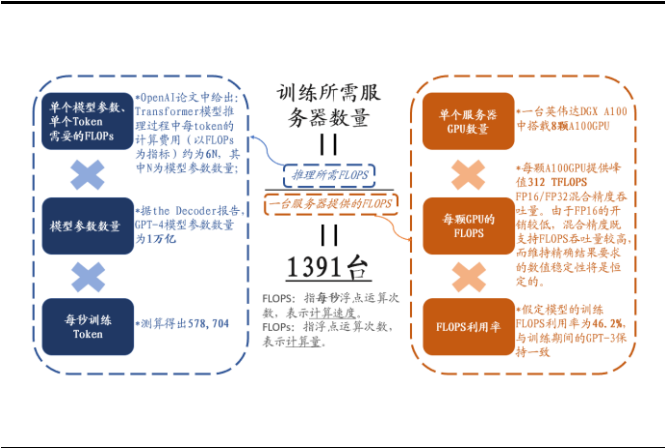
测算过程关键量仍在每秒训练 Token 量，这取决于训练时间。OpenAI 在 6 至 8 个月之前就开始训练 GPT-4，在总数 9 万亿 Token 量+6 个月训练时间假设下，平均每秒训练量为 578,704Token。根据一台服务器每秒提供的 FLOPS 得出训练所需服务器数量。

表6：每秒训练 token 测算过程

总训练 Token 量*	训练天数/	每日秒数/	每单词 token*	每秒处理 Token
9E+12	180	86400	1.33	578,704
* 硅谷称 GPT4 训练 Token 量在 5-9 万亿间，假设为 9 万亿； *OpenAI 公司自 6-8 月前已开始训练 GPT4，即训练天数 30*6=180 天				

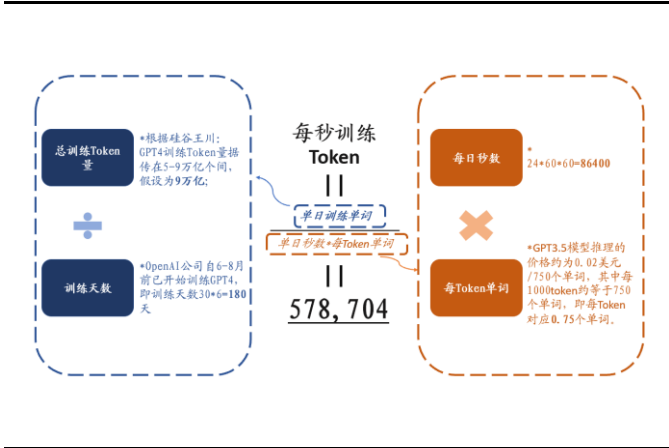
数据来源：OpenAI、东吴证券研究所测算

图30：训练所需服务器测算



数据来源：东吴证券研究所整理

图31：每秒训练 Token 测算



数据来源：东吴证券研究所整理

未来大规模参数模型持续叠加算力需求，算力供需缺口亟待高性能芯片补充。考虑过去 5 年中模型参数呈指数增长，LLM（大语言模型）每次迭代都意味着模型训练量与推理量的指数级增长。随着算力时代到来，未来对服务器数量与芯片算力均存在巨大需求，市场空间难以探底。

表7：主要 LLM 基本信息表

模型名称	发布时间	发布机构	语言	参数规模	Tokens 规模	模型结构	是否开源
T5	2019.10	Google	英	13B		T5-style	✓
GPT-3	2020.05	OpenAI	英	175B	300B	GPT-style	×

LaMDA	2021.05	Google	英	137B	2.8T	GPT-style	×
Jurassic	2021.08	AI21	英	178B	300B	GPT-style	×
MT-NLG	2021.10	Microsoft-NVIDIA	中	530B	270B	GPT-style	×
ERNIE 3.0 Titan	2021.12	Baidu	中	260B	300B	Multi-task	×
Gopher	2021.12	DeepMind	英	280B	300B	GPT-style	×
Chinchilla	2022.04	DeepMind	英	70B	1.4T	GPT-style	×
PaLM	2022.04	Google	多语言	540B	780B	GPT-style	×
OPT	2022.05	Meta	英	125M-175B	180B	GPT-style	✓
BLOOM	2022.07	BigScience	多语言	176B	366B	GPT-style	✓
GLM-130B	2022.08	Tsinghua	中、英	130B	400B	GPT-style	✓
LLaMA	2023.02	Meta	多语言	7B-65B	1.4T	GLM-style	✓
GPT-4	2023.03	OpenAI	多语言	1-100T	5-9T	GPT-style	×

数据来源：arXiv, OpenAI, 东吴证券研究所

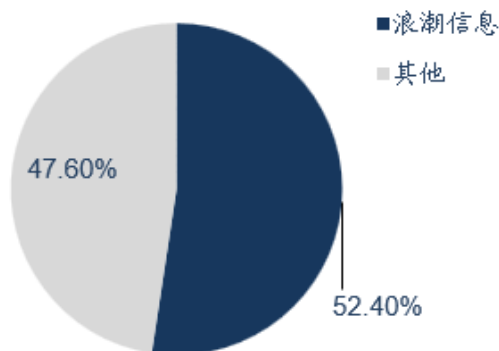
4. 算力时代凸显芯片价值，国产芯片仍处追赶阶段

4.1. AI 服务器依赖高性能芯片供给

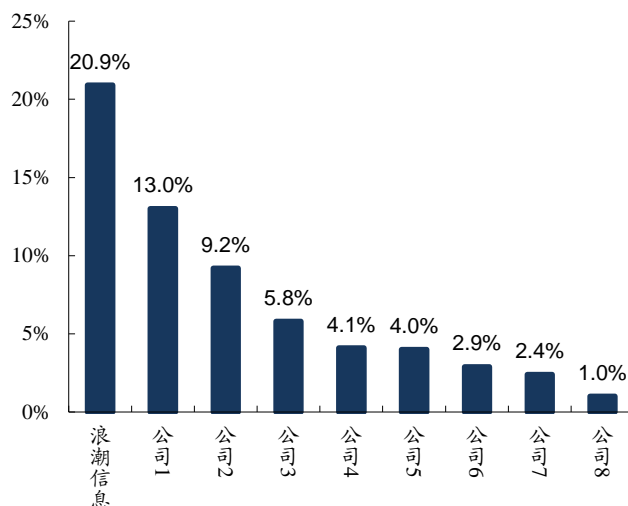
算力时代依托 AI 服务器，浪潮信息服务器蝉联国内第一。根据 IDC 最新发布的《中国加速计算市场（2021 年下半年）跟踪报告》，2021 全年浪潮信息 AI 服务器中国市场占有率达 52.4%。自 2017 年以来，浪潮信息已连续 5 年保持中国 AI 服务器市场份额超过 50%，全球范围内，人工智能服务器市场规模达 156 亿美元，约合人民币 1006 亿元，而浪潮信息市场占有率达 20.9%，保持全球市场第一。

图32: 2021 年中国 AI 服务器市场份额

图33: 2021 年全球 AI 服务器市场份额



数据来源：IDC，东吴证券研究所

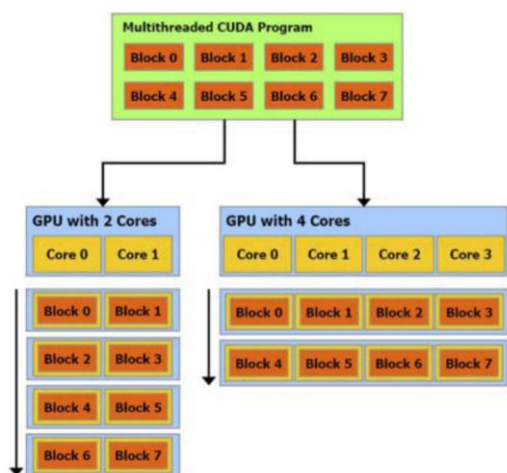


数据来源：IDC，东吴证券研究所

AI 服务器核心在于高性能芯片。参数规模、训练数据随着模型多模态发展呈现量级增长，而消化、处理数据的能力，即算力，才是制约模型发挥与其能力的关键因素。在满足算力的最终诉求下，相较于大量叠加服务器台数，提升算力芯片性能同样不可或缺。以英伟达 GPU 芯片 H100SXM 为例，FP16 Tensor Core 算力为 1979TFLOPS，一台搭载 8 片 H100SXM 的服务器算力是之前所假设的 GPU A100 服务器 **50 倍**。

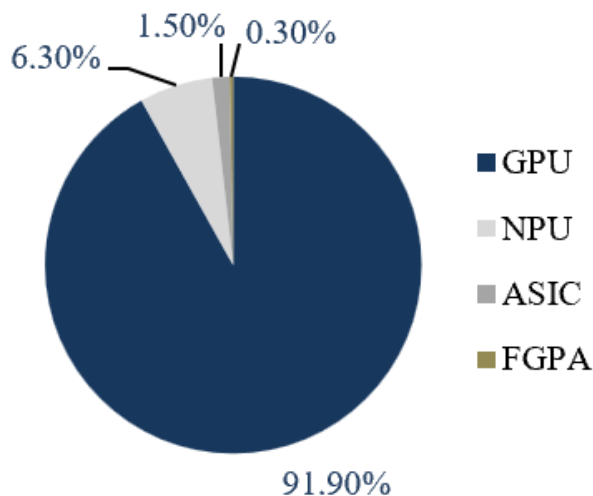
GPU 占据 AI 芯片大类市场，全球应用最为广泛。AI 芯片在不同领域对于指标的需求不同。训练过程中，GPU 由于具备并行计算高度适配神经网络，支持高速解决巨额工作量；推理过程中仅稀疏结构支持一项便能带来 2 倍性能提升。由于 GPU 可兼容训练和推理，高度适配 AI 模型构建，在全球 AI 芯片中应用最为广泛。2021 年中国 AI 芯片市场中，GPU 占比达到 91.9%。据 IDC 数据，预计到 2025 年 GPU 仍将占据 AI 芯片 8 成市场份额。

图34：GPU 并行计算架构适配 AI 模型建构



数据来源：英伟达官网，东吴证券研究所

图35：2021 年 H1 中国 AI 芯片市场占比



数据来源：中商情报局，东吴证券研究所

英伟达主要 GPU 产品技术指标表现领先，是全球 GPU 算力芯片龙头。英伟达的 GPGPU 目前在全球应用最为广泛，其 GPU 产品的典型代表 V100、A100、H100GPU 分别采用 Volta、Ampere、Hopper 架构。A100GPU，INT8 算力达 624TOPS；H100GPU，INT8 算力达到 1513TOPS。此外，2006 年英伟达发布的 CUDA 平台是现今全球应用最为广泛的 AI 开发生态系统。通用 GPU 与 CUDA 生态系统奠定英伟达引领 AI 芯片的根基。当前全球主流深度学习框架均使用 CUDA 平台，2021-2022 年全球独立 GPU 市场中英伟达份额基本维持在 80% 左右。

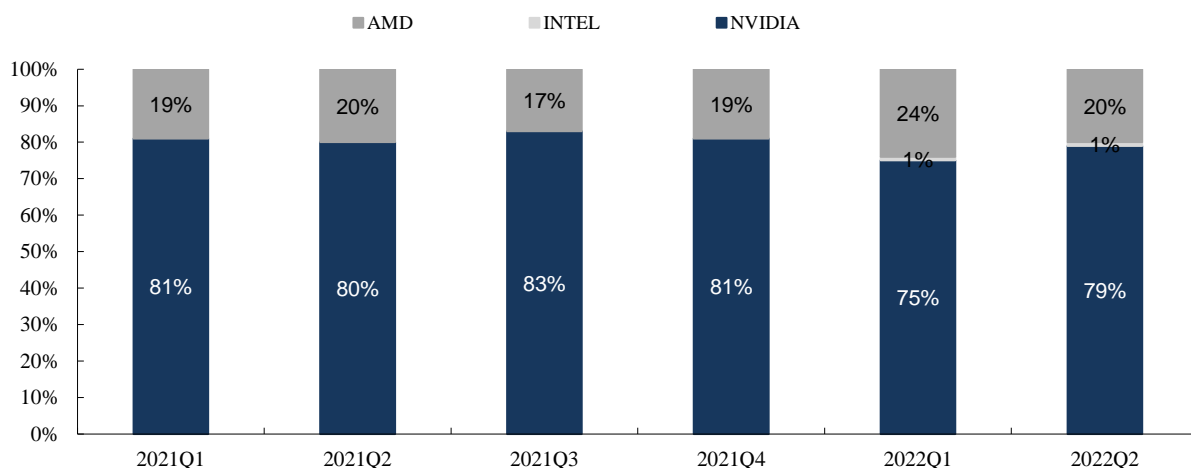
表8：英伟达个芯片性能指标均为世界领先

性能参数	英伟达 H100SXM	英伟达 H100PCle	英伟达 A10080GBPCle	英伟达 V100PCle	AMD InstinctMI250X
FP64	34TFLOPS	26TFLOPS	9.7TFLOPS	7TFLOPS	47.9TFLOPS
FP32	67TFLOPS	51TFLOPS	19.5TFLOPS	14TFLOPS	47.9TFLOPS
FP16 Tensor Core	1979TFLOPS	1513TFLOPS	624 TFLOPS	/	/
INT8 Tensor Core	3958TOPS	3026TOPS	1248TOPS	62TOPS	383TOPS
GPU 显存	80GB	80GB	80GBHBM2	16GBHBM2	128GB
GPU 显存带宽	3.35TB/s	2TB/s	1935GB/s	900GB/s	3276.8GB/s
最大热设计功率	高达 700w	300-350w	300W	300W	500W -560W
制程工艺	TSMC 4nm FinFET	TSMC 4nm FinFET	TSMC 7nm FinFET	TSMC 12nm FinFET	TSMC 6nm FinFET

数据来源：英伟达官网，AMD 官网，东吴证券研究所

注：TOPS（tera Operations Per Second）、TFLOPS（Floating-point Operations Per Second）分别代表芯片每秒进行多少万亿次定点运算和浮点运算，算力越高，运算速度越快，性能越强。

图36：2021-2022Q2 全球独显 GPU 市场各产商份额占比



数据来源：Verified Market Research，东吴证券研究所

服务器同样依赖芯片供给，把握芯片才能把握算力时代。Counterpoint 发布的报告显示，2022 年服务器 CPU 的市场份额中，英特尔占超七成，位列第一，AMD 拿下约两成，排名第二，两家美企联手占据 9 成市场，而浪潮信息需要的 AI 服务器芯片也来自美企英伟达。尽管自 2019 年起，浪潮信息前五大供应商的名字均隐去，但历史数据显示，英特尔仍稳坐最大供应商之席。上游芯片高度依赖外部企业成为浪潮信息发展的关键制约因素，尤其在美国商务部 3 月份将浪潮列入实体清单后，上游芯片进行国产替代显得更为紧迫。算力时代背景下，自主掌握高端芯片技术成为兵家必争之地。

4.2. 主流芯片限制进口，利好国产化芯片自主发展

海外龙头垄断高端芯片技术，国产芯片仍处追赶阶段。以 GPU 细分赛道来看，目前国内自研 GPU 的领军企业主要是寒武纪、景嘉微、华为昇腾等，其中成立最早的是景嘉微，主打产品有 JH920 独立显卡。行业内专家称，从产品参数来看，景嘉微的 JH920 的性能与英伟达在 2016 年发布的 GTX 1050 相仿，虽然仅相差 6 年，但考虑到模型与算力发展之迅速，整体而言国产 GPU 的现状并不算乐观，虽然在特殊领域可满足部分的需求，但是在中高端领域及硬核的算力需求仍存着较长的追赶道路。

表9：海内外 GPU 产品性能对比

公司	英伟达	景嘉微	寒武纪	沐曦半导体	华为昇腾	昆仑芯
产品	H100SXM	JM92 系列	思源 370	曦思®N100	昇腾 910	昆仑 2 代 AI 芯片
单精度性能 FP32 (TFLOPS)	667	1.5	24		640	
半精度性能 FP16 (TFLOPS)	1979		92	80	320	128
整数运算能力 INT8 (TOPS)	3958		256	160	640	256
生产工艺 (nm)	7	14	7		7	7

数据来源：各公司官网产品图，东吴证券研究所

主流的 AI 芯片限制进口,国内 AI 产业发展面临阻碍。2022 年 10 月美国对中国半导体进行三方面限制，在 AI 领域限制中国获取等效 8 Int 600 TOPS 算力的芯片。英伟达针对中国市场推出了符合新规的 A800 芯片，相比 A100 芯片，A800 在搭载 2 个 GPU 的 NVIDIA “NVLink” 桥接器连接下，互联标准由 600GB/s 降为 400GB/s。面临 V100/A100/H100 芯片都进口受限，互联标准下降，英伟达等头部企业维持垄断地位，国产大模型算力需求和国内 AI 产业发展面临阻碍。

表10：美国 BIS《对向中国出口的先进计算和半导体制造物项实施新的出口管制》

涉及方面	主要限制	限制产品
先进计算芯片	对条件内的先进计算半导体芯片、超级计算机最终用途交易以及涉及	3A090 先进计算芯片产品：基础计算单元算力之和超过 4800TOPS 并且 I/O 传输接口的传输速率大于 600GB/s

实体清单上某些实体的交易实施了 限制性出口管制	4A090 超级计算机产品：具有 100 或以上双精度（64 位）每秒千万亿次浮点运算，或 200 或以上单精度（32 位）每秒千万亿次浮点运算的集体最大理论计算能力
先进制程设备	晶圆加工企业为中国生产加工满足 3A090、4A090 的芯片和计算机产品，
	新增 ECCN 3B090 编码，编码涵盖 14nm 先进制程下晶圆抛光、光刻、化学刻蚀、薄膜沉积等全流程的设备
对条件内的半导体制造物项以及某些集成电路 (IC) 最终用途的交易 实施新的管制	新增“最终用途和最终用户规则”，禁止了满足以下条件先进制程设备、零部件、元器件、软件技术的出口： <ol style="list-style-type: none"> 1) 用于生产 16/14nm 以下制程的非平面晶体管结构（即 FinFET 和 GAAFeT）逻辑芯片； 2) 用于生产 128 层或以上 NAND； 3) 用于生产 18nm 以下制程的 DRAM。
从事先进芯片和先进制程的美国实体	美国籍、美国绿卡、美国法律下的法人甚至身处美国的个人/公司都被禁止从事中国境内的先进芯片相关工作。

数据来源：BIS，东吴证券研究所

自主可控趋势下，利好国内服务器供应厂商。近年来中美之间的贸易摩擦不断升级，导致双方在技术领域的竞争愈发激烈。例如美国政府对华为等中国科技企业实施了一系列制裁措施，试图限制这些企业获取关键技术和先进芯片。这些制裁措施无疑给中国高端芯片产业的发展带来了巨大压力。目前自主可控进程中，国内企业在设计、制程、封装和测试等各个环节取得了显著进步，部分产品性能已经接近国际先进水平。另一方面，国家和地方政府也在积极出台各种政策措施，支持芯片产业的创新和发展，培育具有国际竞争力的芯片企业。目前 ChatGPT 引爆的 AI 潮将拉动云服务器数量的增长，其国内相关供应厂商将有望受益于下游 AI+ 级应用所带来的算力需求增长，国内云服务器与算力相关厂商有望直接受益。

5. 标的梳理及投资建议

我们看好算力芯片及服务器需求剧增下相关零部件的成长空间。

CPU、GPU 是服务器最主要的部件，是衡量服务器性能的首要指标，需具有大数据量的快速吞吐、超强的稳定性、长时间运行等能力，由算力芯片提供算力支持。根据 IC Insights 统计，2026 年预计全球 CPU 出货量达到 29 亿颗，市场规模达到 1336 亿美元。行业龙头集中效应显著，Intel 占据服务器 CPU 市场 80% 以上的份额，国内厂商与其技术差距较大，关注重点公司的技术突破进展。

表11：国内服务器 CPU、GPU 芯片厂商业务情况

厂商	主营业务	服务器相关
寒武纪	主要提供各类云服务器/边缘计算/终端设备中人工智能核心芯片的研发、设计和销售，主要产品包括云端智能芯片及加速卡、训练整机、边缘智能芯片及加速卡、终端智能处理器 IP 及相应的配套软件开发平台。	云端智能芯片及加速卡产品应用于服务器的 CPU、GPU 中，提供算力支持。
景嘉微	业务聚焦于图形显控，小型专用化雷达领域的核心模块及系统级产品，主要产品有初代低功耗 GPU 芯片 JM5400、第二代芯片 JM7200 以及新研发的 JM9 系列图形处理芯片等。	GPU 芯片产品已完成与国内主要的 CPU 和操作系统厂商的适配，JM9 系列可以应用于人工智能、云计算等高端应用领域，提供算力支持。
海光信息	公司主营业务为研发、设计和销售应用于服务器等计算、存储设备中的高端处理器，目前拥有海光通用处理器（CPU）和海光协处理器（DCU）两条产品线。主要产品有 CPU 产品海光系列以及 DCU 产品深海系列。	海光 CPU 产品可以应用于高端服务器、数据中心、中低端服务器以及边缘计算服务器等计算领域中，DCU 产品专门用于为人工智能和深度学习提供算力支持。

数据来源：公司官网，东吴证券研究所

内存、硬盘是服务器中其他的重要部件，内存是与 CPU 进行通信的桥梁，硬盘是服务器数据的仓库。计算机中的所有程序都在内存中执行，服务器的硬盘需要满足速度快和高可靠性，内存和硬盘都对服务器的性能有较大的影响。据 Yole 统计，存储器总体市场空间将从 2021 年的 1670 亿美元增长至 2027 年的 2630 亿美元，年复合增长率为 8%。市场呈现呈垄断竞争格局，DRAM 市场 CR3 超 90%，目前国内厂商与国外的技术，规模等差距较大，自主产品亟待突破。

表12：国内服务器内存和硬盘相关厂商业务情况

厂商	主营业务	服务器相关
江波龙	公司主要从事 Flash 及 DRAM 存储器研发、设计和销售，提供消费级、工规级、车规级存储器以及行业存储软硬件应用解决方案，目前有嵌入式存储、固态硬盘、移动存储及内存条四大产品线。	公司的 DDR4RDIMM 内存条、SSD 系列产品可用于服务器中的内存模块，与服务器中 CPU 直接交换数据。
聚辰股份	致力于为客户提供存储、模拟和混合信号集成电路产品并提供应用解决方案和技术支持服务，主要有非易失性存储业务包括 EEPROM 以及 NorFlash 两条产品线。	公司 EEPROM 产品线中标准接口的产品用于服务器的内存模块，专有的 SPDEEPROM 产品应用于服务器内存条的温度传感器部分。

数据来源：公司官网，东吴证券研究所

电源负责各个模块的供电和电路控制，由电源管理芯片实现。模拟芯片市场呈现国外企业主导的竞争格局，据 IC Insights 数据，21 年中国模拟 IC 市场需求占据全球规模的 43%，为最大需求市场。海外占据主导地位，根据 MPS 预测，MPS 在服务器主板 CPU/GPU 电源管理芯片领域可触及的市场规模达 20 亿美元，但国内厂商的自给率极低，成长潜力可期。

表13: 国内服务器电源芯片相关厂商业务情况

厂商	主营业务	服务器相关
晶丰明源	公司专注于电源管理和电机控制芯片的研发和销售，产品覆盖 LED 照明驱动芯片、AC/DC 电源管理芯片、DC/DC 电源管理芯片、电机控制驱动芯片等。	推出的多相数字控制电源管理芯片可用于为服务器中 CPU、GPU、内存和芯片组等供电。
杰华特	以虚拟 IDM 为主要经营模式的模拟集成电路设计企业，公司产品分为电源管理芯片和信号链芯片两大类，电源管理芯片产品包括 AC-DC 芯片、DC-DC 芯片、线性电源产品等子类别。	公司 DC-DC 芯片产品是服务器中的 CPU/GPU/DRAM 等供电和电路控制的核心模块，公司目前与 Intel 深度合作。

数据来源：公司官网，东吴证券研究所

服务器芯片经由封装形成模块，进入整机组装环节。根据 Yole 数据预计，2026 年先进封装全球市场规模 475 亿美元，2020-2026ECAGR 约为 7.7%。目前，全球封测产业正逐步向中国大陆转移，内资企业与外资厂商技术差距持续缩小，中国台湾、中国大陆和美国占据主要市场份额，同时积极布局 Chiplet 等先进技术。

表14: 国内封测相关厂商业务情况

厂商	主营业务	服务器相关
长电科技	公司在 HPC、存储类、5G 通信类、消费类、汽车和工业等重要领域拥有行业领先的先进封装技术（如 SiP、WL-CSP、FC、eWLB 系列等），业务覆盖高、中、低端半导体封测类型。	公司的 2.5D/3D 集成技术适用于对集成度和算力有较高要求的 FPGA、CPU、GPU 等芯片的集成封装，银线引线类封装产品可应用于存储芯片的封装，是芯片和服务器成品间的中间环节。
通富微电	是集成电路封装测试服务提供商，目前已掌握一系列高端集成电路封装测试技术，WLCSP、FC、SiP、BGA 基板设计及封装技术及功率器件等产品已全部实现产业化，具有 FCBGA、FCPGA 等高端封装技术和大规模量产平台。	公司的 2.5D/3D 封装平台和超大尺寸 FCBGA 研发平台可为客户提供晶圆级和基板级 Chiplet 封测解决方案，适用于服务器高性能计算领域芯片的封测。

数据来源：公司官网，东吴证券研究所

服务器其他零部件包括 PCB、连接器和接口等，成本占比低于 20%。

PCB 在高端服务器中的应用主要包括背板、高层数线卡、HDI 卡、GF 卡等，其特点主要体现在高层数、高纵横比、高密度及高传输速率，2016 年以来，中国大陆 PCB 产值规模在全球占比均超 50%。

服务器内部通过连接器实现印刷电路板和电子元件之间的连接，根据 Bishop& Associates 数据，2019 年全球连接器公司以欧美和日本企业主导，全球前十连接器厂商市场份额合计为 60.8%，市场集中度较高。

服务器通过接口芯片以实现内外数据交换。目前全球接口芯片龙头厂商依旧为 TI、美信及凌力尔特等，从 DDR4 世代开始，全球内存接口芯片厂商仅剩澜起科技、瑞萨（原 IDT）和 Rambus 三家厂商。

表15：服务器其他相关零部件厂商业务情况

厂商	主营业务	服务器相关
工业富联	公司是全球领先的高端智能制造及工业互联网解决方案服务商，主要业务包含云计算、通信及移动网络设备、工业互联。产品实现了对云及边缘计算、工业互联网、智能家居、5G 及网络通讯设备、智能手机及智能穿戴设备等数字经济产业五大类范围的全覆盖。	公司在云计算服务器出货量持续全球第一，与全球主要服务器品牌商、国内外 CSP 客户深化合作，推出新一代云计算基础设施解决方案，包括模块化服务器、高效运算（HPC）等，重点解决因 ChatGPT 持续升温而引发 AIGC 算力井喷需求。
深南电路	内资 PCB 领先企业，打造“3-In-One”业务布局，拥有印制电路板、封装基板、电子装联三项业务，在背板、高速多层板、多功能金属板等高中端 PCB 的加工工艺方面拥有领先技术实力。	公司的 PCB 应用于服务器主板、硬盘等部件中，封装基板覆盖模组类封装基板、存储类封装基板、应用处理器芯片封装基板等，应用于服务器中各模块封装。
沪电股份	公司立足于印制电路板的研发设计和生产制造，现已发展成为印制电路板行业内的重要品牌之一。公司主导产品广泛应用于通讯设备、汽车、工业设备、数据中心、网通、微波射频、半导体芯片测试等多个领域。	公司的背板、HDI、通孔板、背钻等产品应用于服务器中的背板等，应用于 EGS 级服务器领域产品已实现规模化量产，应用于 GPU 等加速模块类的产品已批量出货
鼎通科技	公司专注于研发、生产、销售通讯连接器精密组件和汽车连接器精密组件，具备精密模具设计开发、产品制造一体化能力，直接客户为安费诺、莫仕和中航光电等连接器模组制造商。	公司高速背板连接器用于服务器中单板和背板的连接和信号传输，I/O 连接器用于服务器的内外信号传输。
裕太微	公司专注于高速有线通信芯片的研发、设计和销售，是国内以太网芯片龙头，公司以以太网物理层芯片作为市场切入点，不断推出系列芯片产品，目前已推出单/多端口百兆、千兆、2.5G 物理层芯片。	以太网物理层芯片用于实现以太网协议中的物理层功能，应用于大数据中心和云计算数据中心的局域网中以保证网络链路中数据传输的稳定性和快速性。
澜起科技	目前公司的主营业务是为云计算和人工智能领域提供以芯片为基础的解决方案，目前主要产品包括内存接口芯片、津逮服务器 CPU 以及混合安全内存模组，可提供从 DDR2 到 DDR4 内存全缓冲/半缓冲完整解决方案。	公司内存接口芯片是内存模组的核心器件。内存模组配套芯片产品用于存储内存模组的相关信息和实现内存模组的电源及温度管理。

数据来源：公司官网，东吴证券研究所

6. 风险提示

1. 本文对服务器、芯片需求量等测算是基于自身对于 ChatGPT 使用经验的前提假设，

存在假设条件不成立或假设出现偏差导致计算结果出现出入。

2.AI 服务器渗透率提升低于预期。AI 服务器渗透率提升的速度存在低于预期的风险，主要与企业投资金额，产品产能扩张，以及供给格局有关，以上因素均会导致 AI 服务器发展不及预期。

3.AI 应用推广不及预期。AI 技术在应用推广的过程可能面临数据质量、资源限制和技术能力等因素的制约，导致相关企业的 AI 应用存在推广进度不及预期的风险。

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司不对任何人因使用本报告中的内容所导致的损失负任何责任。在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。如引用、刊发、转载，需征得东吴证券研究所同意，并注明出处为东吴证券研究所，且不得对本报告进行有悖原意的引用、删节和修改。

东吴证券投资评级标准：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对大盘在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对大盘介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对大盘介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对大盘介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对大盘在-15%以下。

行业投资评级：

增持：预期未来 6 个月内，行业指数相对强于大盘 5%以上；

中性：预期未来 6 个月内，行业指数相对大盘-5%与 5%；

减持：预期未来 6 个月内，行业指数相对弱于大盘 5%以上。

东吴证券研究所

苏州工业园区星阳街 5 号

邮政编码：215021

传真：（0512）62938527

公司网址：<http://www.dwzq.com.cn>