

# 从美股科技看海外AI进展

——海外科技24年春季投资策略暨GenAI系列33

证券分析师：

林起贤 A0230519060002 闫海 A0230519010004

李国盛 A0230521080003 杨海晏 A0230518070003

洪依真 A0230519060003 刘洋 A0230513050006

研究支持：黄俊儒 A0230123070011

2024.03.29



# 主要内容

---

1. AI算力：业绩正兑现，空间看应用爆发
2. AI+云计算：MaaS切入，增长提速
3. AI应用：仍在萌芽，收入初步体现

# 占比提升接续训练需求

- 基于Transformer架构的大模型军备竞赛仍在持续, 算力需求仍强劲: Scaling Laws+多模态, 仍是迭代路径。

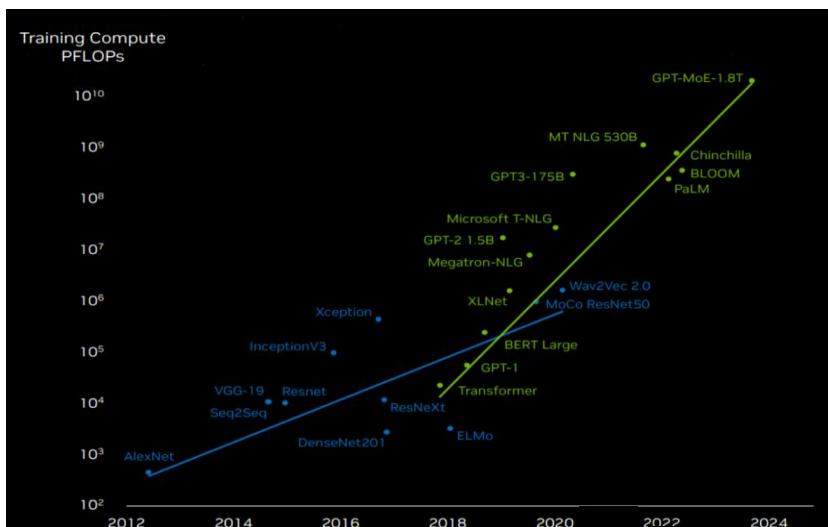
- 英伟达财报验证推理需求接续训练算力需求的增长, 未来市场空间取决于AI应用爆发。

英伟达Data Center业务FY24Q4营收184美元, 环比+27%, 同比+409%, 本季来自大型云厂商营收占比超50%, 公司指引FY25Q1收入240亿美元(上下浮动 2%), 英伟达在FY24Q4财报交流中指出FY24 AI推理需求占收入的比例已达到40%, 高于此前市场预期的判断。目前, 英伟达数据中心总需求的大约70%来自推理, 而AMD则指出到2027年, 其4000亿美元的AI总潜在市场中大约65%至70%来自推理。

- 大模型推理仍然需要高性能功能算力网络, 对AI芯片的性能需求持续提升。

1) 预训练 – 微调 – 推理密不可分; 2) 单位性能的成本差异巨大; 3) 容纳千亿以上参数。

图：Transformer后Scaling Laws仍在持续验证



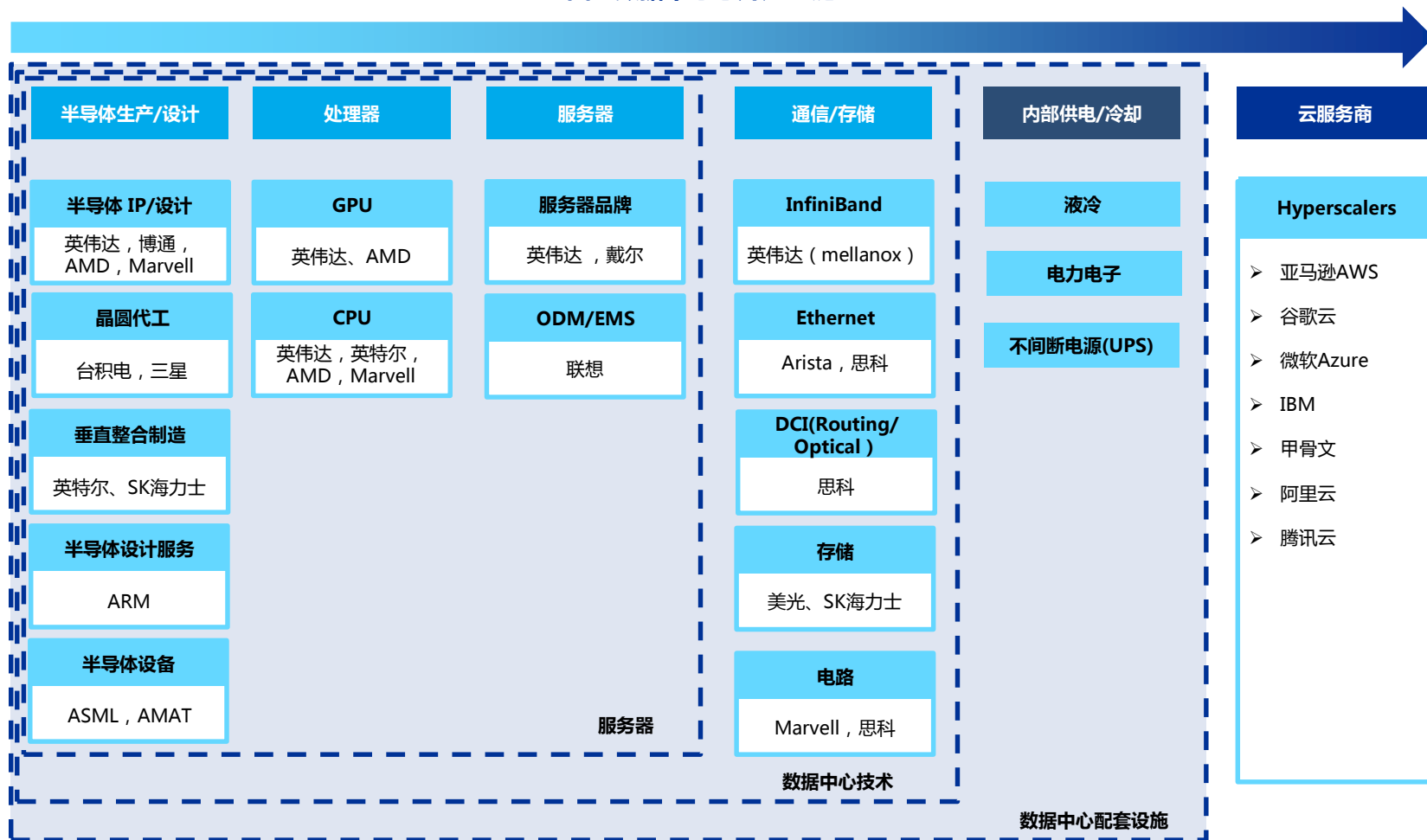
表：2024 LLM进展重要时间点

| 时间 |    | 事件                        | 备注 |
|----|----|---------------------------|----|
| Q1 | 3月 | Anthropic发布Claude 3       |    |
|    |    | xAI发布Grok1.5              |    |
|    |    | 月之暗面新产品发布                 |    |
|    | 6月 | OpenAI发布GPT-4.5           | 预测 |
| Q2 | 6月 | 科大讯飞发布达到GPT-4 Turbo的下一代模型 |    |
|    | H1 | Mistral开源GPT-4水平的MoE模型    | 预测 |
| Q3 | 7月 | Meta开源LLaMa3              |    |
| Q4 | H2 | OpenAI发布GPT-5             | 预测 |

## 1.2 产业链：24年GPU供给仍偏紧，GPU厂商欲拓展价值链

- GPU为数据中心核心，英伟达、AMD正依托GPU向CPU+网络+存储系统级解决方案延伸，旨在拓展价值链及提升竞争力。
- 24年供给仍偏紧，台积电COWOS产能、HBM3产能为当前GPU供给端的约束。

图：数据中心芯片产业链



## 1.2 产业链-AI算力网络:英伟达携非以太网 VS 超以太网阵营

- 当前海外AI算力网络市场，已呈现“二元格局”：
- **EtherNOT（非以太网）**：Nvidia（Mellanox）代表的InfiniBand网络阵营，是AI网络架构的新力量，优势是性能与生态；
- **EtherNET（以太网）**：博通为代表的“超以太网联盟UEC”（另以传统硬件与科技大厂为代表），则形成基于以太网底座的高性能网络阵营，优势是通用性与成本。从边际变化以及Nvidia生态角度出发，2023年产业与资本市场近期较多关注InfiniBand等阵营的机会；而实际上自2024年起，以太网阵营亦代表了行业放量的重要方向。

图：博通、AMD是超以太网联盟的代表厂商



图：海外AI算力网络架构呈现“二元格局”



# 1.3 英伟达：硬件加速迭代，软件日渐丰富向商业化迈进

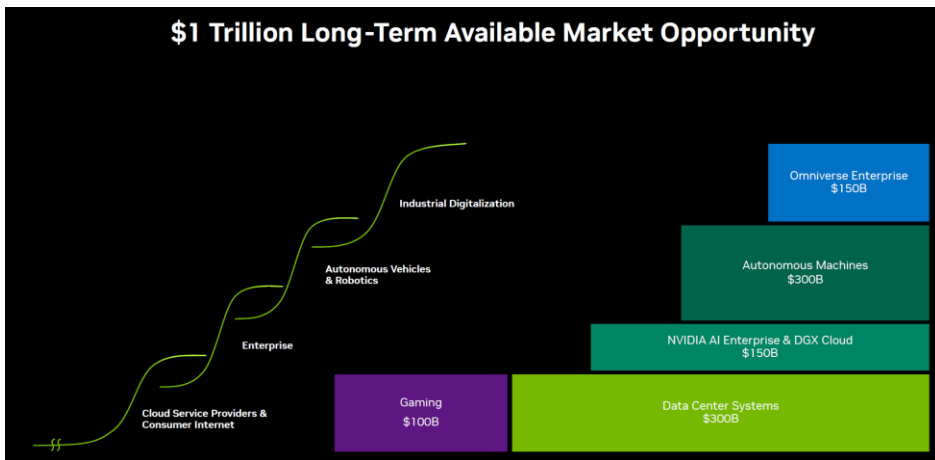
## ■ 竞争压力下硬件加速迭代提升产品性价比，从单卡性能竞争升维至计算集群能力竞争。

- 面对AMD及未来云厂商的AI自研芯片竞争，英伟达明显加快了产品迭代速度并提升性价比，24年3月推出GB200，推理/训练性能分别为H100的30x/4x，而价格预计仅小幅提升，新品强调性价比路线也引发竞争影响高毛利产品份额的讨论。
- 此外，GB200推出NVL72算力集群，依托NVLink等组件从组网层面进行系统性优化，从集群层面而非单卡层面提升竞争力。其中网络产品FY24Q4收入YoY+217%，全年收入YoY+133%，年化收入超130亿美元；其中Quantum InfiniBand 产品营收同比增长超5倍。

## ■ 软件业务不仅是护城河，更有望提供商业化收入。

- 公司FY24Q4软件业务ARR达到10亿美元，将进一步丰富软件生态，公司预计未来DGX Cloud和AI Enterprise合计TAM将达到1,500亿美元。
- 英伟达在GTC2024更新AI Enterprise功能，并发布NIM ( Nvidia Inference Microservice ) 服务，每个GPU 费用4500美元/年。

图：英伟达正积极拓展硬件外业务



图：竞争压力下，英伟达产品迭代周期从两年缩短至一年



## 1.4 AMD：最接近英伟达的追赶者

- **布局向英伟达看齐**：不仅是GPU，AMD还在通信（Infinity fabric）、开发生态（ROCm）上全面对标英伟达。
- **牵手众多合作伙伴，MI300系列供货量正快速爬坡**：数据中心 23Q4 营收为 22.82 亿美元，同比+38%，公司指引2024年数据中心GPU营收将达35亿美元，AI巨头出于降低依赖英伟达芯片的需求扶持AMD为第二供应商。
- **ROCm开发生态正走向丰富**：1）加强对CUDA生态的兼容，培养开发者群体；2）争取合作伙伴，加速构建自主开发生态。

表：AMD与英伟达GPU产品对比

|                 | AMD MI300X              | 英伟达 H100 SXM           | 英伟达 H200 SXM           |
|-----------------|-------------------------|------------------------|------------------------|
| 制程              | TSMC 5nm   6nm FinFET   | TSMC 4nm               | TSMC 4nm               |
| FP64 ( TFLOPS ) | 81.7                    | 67                     | 67                     |
| BF16 ( TFLOPS ) | 1300                    | 1979                   | 1979                   |
| INT8 ( TOPS )   | 2600                    | 3958                   | 3958                   |
| 内存容量及类型         | 192 GB HBM3             | 80GB HBM3              | 141GB HBM3e            |
| 内存带宽 ( TB/s )   | 5.3                     | 3.35                   | 4.8                    |
| Interconnect    | PCIe Gen5.0 : 128 GB/s  | PCIe Gen5.0 : 128 GB/s | PCIe Gen5.0 : 128 GB/s |
|                 | Fabric Links : 896 GB/s | NVLink : 900 GB/s      | NVLink : 900 GB/s      |
| 生态              | ROCm                    | CUDA                   | CUDA                   |

表：AMD MI300合作伙伴进展

| 合作伙伴 | 产品进展                                                                         |
|------|------------------------------------------------------------------------------|
| Meta | 新增AMD Instinct MI300X加速器至其数据中心                                               |
| 微软   | 发布Azure ND MI300x v5虚拟机（VM）系列产品                                              |
| 甲骨文  | 计划引入搭载AMD MI300X的裸机实例用于其高性能计算，且该基于MI300X的实例预计将支援配备高速RDMA网络的OCI Supercluster。 |
| 戴尔   | 展示了Dell PowerEdge XE9680服务器，其配备8个MI300系列加速器。                                 |
| 惠普   | 发布首款搭载AMD MI300A APU的超算加速器HPE Cray Supercomputing EX255a，预计于2024年上半年上市。      |
| 联想   | 宣布计划将在2024年上半年推出基于AMD MI300系列加速器的创新设计。                                       |
| 超微电脑 | 推出基于第4代AMD EPYC处理器和AMD Instinct MI300系列加速器的H13系列新产品                          |

表：AMD近期重要拓展收购，与英伟达较为类似

| 时间       | 价格        | 标的               | 方向        | 明细        | 对标              |
|----------|-----------|------------------|-----------|-----------|-----------------|
| 2022年2月  | 490亿美元全股票 | Xilinx           | 数据中心，异构计划 | 拓展异构计算。   | 英伟达尝试收购ARM（但受挫） |
| 2022年5月  | 19亿美元     | Pensando Systems | 分布式计算     | 补IB和高速背板。 | 英伟达收购mellanox   |
| 2022年10月 | --        | Nod.AI           | 推理框架      | 拓展推理软件框架， | 英伟达TensorRT-LLM |

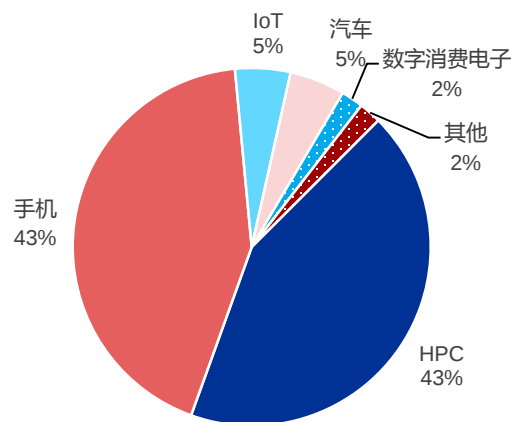
# 先进封装+代工-台积电

- **博通：巨头自研ASIC合作伙伴+AI算力网络需求强劲。** FY24Q1法说会公司调整网络收入全年增速提升5pct至35%，并指引AI业务营收为100亿美元，根据外媒IBD，24年3月博通获得第三个ASIC业务客户，此前已有包括谷歌（TPU）在内的两个大规模客户。
- **HBM（美光/SK海力士）：HBM3E量产后产能紧俏，传统DRAM处于周期回暖阶段。** 其中美光HBM3E已量产，且较竞品功耗低30%、将导入H200，24年产能、25年大部分产能已分配完毕；海力士指引2024年TSV产能同比翻倍，HBM长期需求增长率超60%。21H2存储进入下行周期，当前处于底部回暖，FY2Q24美光DRAM和NAND出货量环比提升，且ASP均有两位数环比涨幅。
- **台积电：AI需求旺盛带动业绩回暖。公司预计全年手机恢复增长，尤其HPC增长显著。** 根据公司指引，预计AI相关营收可逐年复合增长50%，5年后营收占比将达17%-19%（200-270亿美元）；未来营收增长目标CAGR15%-20%，毛利率53%+，保持不变。

表：24年博通AI业务营收将超越此前单季度总营收水平

| 单位：百万美元   | FY2Q22A | FY3Q22A | FY4Q22A | FY1Q23A | FY2Q23A | FY3Q23A | FY4Q23A | FY1Q24A |
|-----------|---------|---------|---------|---------|---------|---------|---------|---------|
| 营业收入      | 8103    | 8464    | 8930    | 8915    | 8733    | 8876    | 9295    | 11961   |
| YoY       | 22.60%  | 24.90%  | 20.60%  | 15.70%  | 7.80%   | 4.90%   | 4.10%   | 34.20%  |
| 毛利润       | 6186    | 6427    | 6673    | 6578    | 6606    | 6665    | 6908    | 9014    |
| 毛利率       | 76.30%  | 75.90%  | 74.70%  | 73.80%  | 75.60%  | 75.10%  | 74.30%  | 75.40%  |
| 净利润       | 4000    | 4241    | 4544    | 4483    | 4489    | 4596    | 4810    | 5254    |
| 同比增速      | 34.20%  | 35.80%  | 29.80%  | 19.80%  | 12.20%  | 8.40%   | 5.90%   | 17.20%  |
| 净利润率      | 49.40%  | 50.10%  | 50.90%  | 50.30%  | 51.40%  | 51.80%  | 51.70%  | 43.90%  |
| 分部门业绩     |         |         |         |         |         |         |         |         |
| 半导体解决方案业务 | 6229    | 6624    | 7092    | 7107    | 6808    | 6941    | 7326    | 7390    |
| YOY       | 29.20%  | 31.90%  | 25.90%  | 21.00%  | 9.30%   | 4.80%   | 3.30%   | 4.00%   |
| 基础设施软件业务  | 1874    | 1840    | 1838    | 1808    | 1925    | 1935    | 1969    | 4571    |
| YOY       | 4.70%   | 4.70%   | 3.70%   | -1.40%  | 2.70%   | 5.20%   | 7.10%   | 152.80% |

图：台积电FY23Q4分应用收入



# 主要内容

---

1. AI算力：业绩正兑现，空间看应用爆发
2. AI+云计算：MaaS切入，增长提速
3. AI应用：仍在萌芽，收入初步体现

## 2.1 云厂商+META：全面拥抱AI，各具禀赋优势

1) 大模型：OpenAI、谷歌、Anthropic为第一梯队。2) 算力布局：自研芯片进度，谷歌>亚马逊>微软、Meta。3) 现有业务生态协同、AI应用落地及商业化进度：微软>谷歌>Meta、亚马逊。

表：主要科技公司AI布局

|      | AI芯片布局                                                                                                                                                            | 大模型及开发框架                                                                                                                                                                     | AI研发布局模式                                                                                                                                           | 已推出的核心AI应用                                                                                                                                                             | 云服务的主要AI客户                                                                                                       |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|
| 微软   | <p><b>外购：</b>根据Omdia，23年购买约15万张H100芯片</p> <p><b>自研：</b>2023年11月发布Maia 100芯片，预计24年投产，用于训练和推理</p>                                                                   | <p><b>大模型：</b>OpenAI推出GPT系列模型，2023年3月推出GPT-4，后续推出升级的GPT4 Turbo、GPT4v版本；下一代模型已在研发，将于2024年推出</p> <p><b>开发平台：</b>Azure AI Studio，包括GPT4-Turbo等</p>                              | <p><b>大比例持股体外公司+深度合作。</b>在2019年、2023年向OpenAI分别投资10亿、100亿美元，持有OpenAI49%的股权，OpenAI为独立公司运营模式，与微软合作关系紧密；</p> <p>公司内部设立微软研究院，探索AI领域应用落地</p>             | <p><b>办公：</b>推出Microsoft 365 Copilot</p> <p><b>CRM/ERP：</b>推出Dynamic 365 copilot</p> <p><b>编程工具：</b>Github Copilot</p> <p><b>搜索引擎：</b>必应集成ChatGPT，提供Copilot功能</p>      | <p><b>OpenAI、Mistral、Hugging Face、Perplexity、Databricks等</b></p>                                                 |
| 谷歌   | <p><b>外购：</b>根据Omdia，23年购买约5万张H100芯片；</p> <p><b>自研：</b>2016年推出第一代TPU芯片，2023年推出TPUv5系列，性能出色，可基本支撑自研大模型的训练和推理</p> <p><b>通信：</b>自研OCS通信系统，通信性能出色</p>                 | <p><b>大模型：</b>推出LaMDA、PaLM、PaLM2；2023年12月推出首个多模态大模型Gemini1.0，24年2月推出Gemini1.5Pro</p> <p><b>深度学习框架：</b>TensorFlow（两大主流框架之一）、JAX</p> <p><b>开发平台：</b>Vertex AI，已搭载超过130个大模型</p> | <p><b>旗下部门自研：</b>此前有Google Brain、Deepmind等多个AI研发部门/全资子公司，分立运营；2023年4月起整合为单一AI研发部门Google Deepmind</p>                                               | <p><b>办公：</b>推出Duet AI，定价30美元/月</p> <p><b>搜索：</b>AI搜索功能SGE已在内测，未来将推出</p> <p><b>医疗：</b>依托医疗行业垂直大模型，辅助医药研发和诊断</p> <p><b>其他：</b>AI编程工具；自动驾驶领域的Waymo；Youtube、谷歌地图的AI工具</p> | <p>Anthropic、Mistral、Character.AI、Cohere、Jasper、MosaicML、Replit、Runway等。包括超过50%的AI初创公司、超过70%的生成式AI独角兽为谷歌云客户。</p> |
| Meta | <p><b>外购：</b>根据Omdia，2023年购买约15万张H100芯片。公司计划2024年底总共拥有大约35万颗H100 GPU，总共将近60万颗AI芯片；</p> <p><b>自研：</b>公司预计于2024年推出Artemis AI芯片，将用于推理侧；预计于2025年Mtia将投入使用，用于训练和推理</p> | <p><b>大模型（开源）：</b>2023年7月开源Llama2，公司计划于2024年推出Llama3，目前正在训练</p> <p><b>深度学习框架：</b>Pytorch（两大主流框架之一）</p>                                                                       | <p><b>旗下部门自研：</b>AI业务均由旗下AI部门进行研发，为直属部门模式</p>                                                                                                      | <p><b>社交软件：</b>提供AI功能增加用户体验</p> <p><b>广告：</b>辅助广告内容生成工具；通过AI实现广告智能投放提升效率</p>                                                                                           | -                                                                                                                |
| 亚马逊  | <p><b>外购：</b>根据Omdia，2023年购买约5万张H100芯片</p> <p><b>自研：</b>2020年推出Trainium，23年11月推出Trainium2，用于训练侧；2023年4月推出Inferentia2，用于推理侧</p>                                    | <p><b>自研大模型：</b>2023年12月推出Titan系列AI模型；根据路透社，下一代大模型Olympus已在研发，参数量或达到2万亿</p> <p><b>大模型（Anthropic）：</b>推出Claude系列模型，24年3月推出Claude3</p> <p><b>开发平台：</b>Bedrock AI搭载自研及第三方模型</p> | <p><b>旗下部门自研+持股重点公司：</b>旗下AI部门完成自研大模型研发；</p> <p>亚马逊重点投资Anthropic，2023年9月投资40亿美元，但与亚马逊的绑定关系不够深；谷歌也参与Anthropic多轮投资，23年10月谷歌宣布向Anthropic投资最高20亿美元</p> | <p><b>电商：</b>为电商运营提供一系列AI功能支持，以及导购助手Rufus；</p> <p><b>生成式助手：</b>面向企业端的Amazon Q；</p> <p><b>广告：</b>辅助广告内容生成工具；通过AI实现广告智能投放提升效率</p>                                        | <p>Anthropic、Mistral、Perplexity AI、Nomic AI、Stability AI、Grammarly、Hugging Face等</p>                             |

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/636045214000010112>