国金证券 SINOLINK SECURITIES

2024 年 12 月 09 日

AI 周观察

行业深度研究(深度)

证券研究报告

分析师：刘道明（执业 S1130520020004） liudaoming@gjzq.com.cn
联系人：黄晓军（执业 S1130122050092） huangxiaojun@gjzq.com.cn
联系人：麦世学（执业 S1130123100111） maishixue@gjzq.com.cn

AI 眼镜销量超预期，自研芯片加速部署，市场静待 Sora 发布

报告摘要

- AI 应用热度仍在上升，聊天助手类应用活跃度显著提升，亚马逊推出 Amazon Nova 多模态模型系列，支持文本、图像与视频处理，Meta 发布高效模型 LLaMA 3.3 70B。视频生成领域迎来腾讯开源混元模型、谷歌 Genie 2 世界模型和 World Lab 的世界模型，强化 3D 场景与动态交互能力。OpenAI 启动“12 天发布计划”，发布 o1 正式版、Pro 版订阅与强化微调（RFT）技术，提升模型推理与应用潜力，市场静候 OpenAI 后续发布 Sora。
- 博通将在 12 月 12 日发布 FY24Q4 财报。回顾 FY24Q3，公司营收强劲增长，主要受 AI 半导体业务和 VMware 整合推动。基础设施软件和半导体解决方案部门表现突出，AI 产品销售尤为亮眼。展望 Q4，公司预计营收将继续增长，全年 AI 收入高于此前预期。建议关注公司 FY24Q4 自研芯片及数据中心网络连接产品的发展情况。
- 亚马逊推出的新一代自研 AI 加速器 Trainium2，并在大规模部署中得到了实际应用。这款芯片针对推理和特定 AI 工作负载进行了优化，体现了云厂商在加速器采购决策中对总拥有成本（TCO）和工作负载特性匹配的高度重视。我们认为，随着 AI 训练和推理需求的持续增长，自研芯片的应用将加速推进，为自研芯片供应商带来长期的业绩增长机遇。
- 2024 年三季度，Meta Rayban AI 眼镜全球出货量达到约 48 万部，环比增长 83%，2023 年四季度发布至今累计销量达到约 122 万部。一季度 Meta AI 功能上线后对销量提振效果显著，24 年 Q4 双节大促对销量会有更强的提振效应，AI 眼镜的接受度得到提升。我们认为，在 2025 年具备音频和摄像头的 AI 眼镜是当下 AI 模型应用落地的最佳可穿戴设备，随着多模态模型能力的提升和 AI Agent 的成熟，产品功能性和应用场景将获得极大提升，持续看好 2025 年 AI 眼镜大规模放量。
- 短期内 VR/MR 设备受制于长期佩戴体验不佳&部分性能优异产品价格过高，难以在数量上爆发。但长期来看，我们看好 VR/MR 设备成为元宇宙的入口。在 VR/MR 产品技术更为成熟后，我们认为市场空间将被进一步打开。

风险提示

芯片制程发展与良率不及预期

中美科技领域政策恶化

AI 硬件销量不及预期

内容目录

财报前瞻

关注博通自研芯片业务

博通将于 12 月 12 日周四盘后发布其 FY24Q4 财报，回顾上一财季，博通在 FY24Q3 实现营收 131 亿美元，同比增长 47%，运营利润 79 亿美元，同比增长 44%。业绩主要受 AI 半导体业务的强劲需求和 VMware 整合带来的增长推动。

基础设施软件部门营收 58 亿美元，同比增长 200%，得益于 VMware 的贡献和成本控制，季度运营支出从 Q2 的 16 亿美元降至 13 亿美元。VMware 的年化预订价值(ABV) 增长 32%至 25 亿美元，反映了 VMware Cloud Foundation 的持续需求。

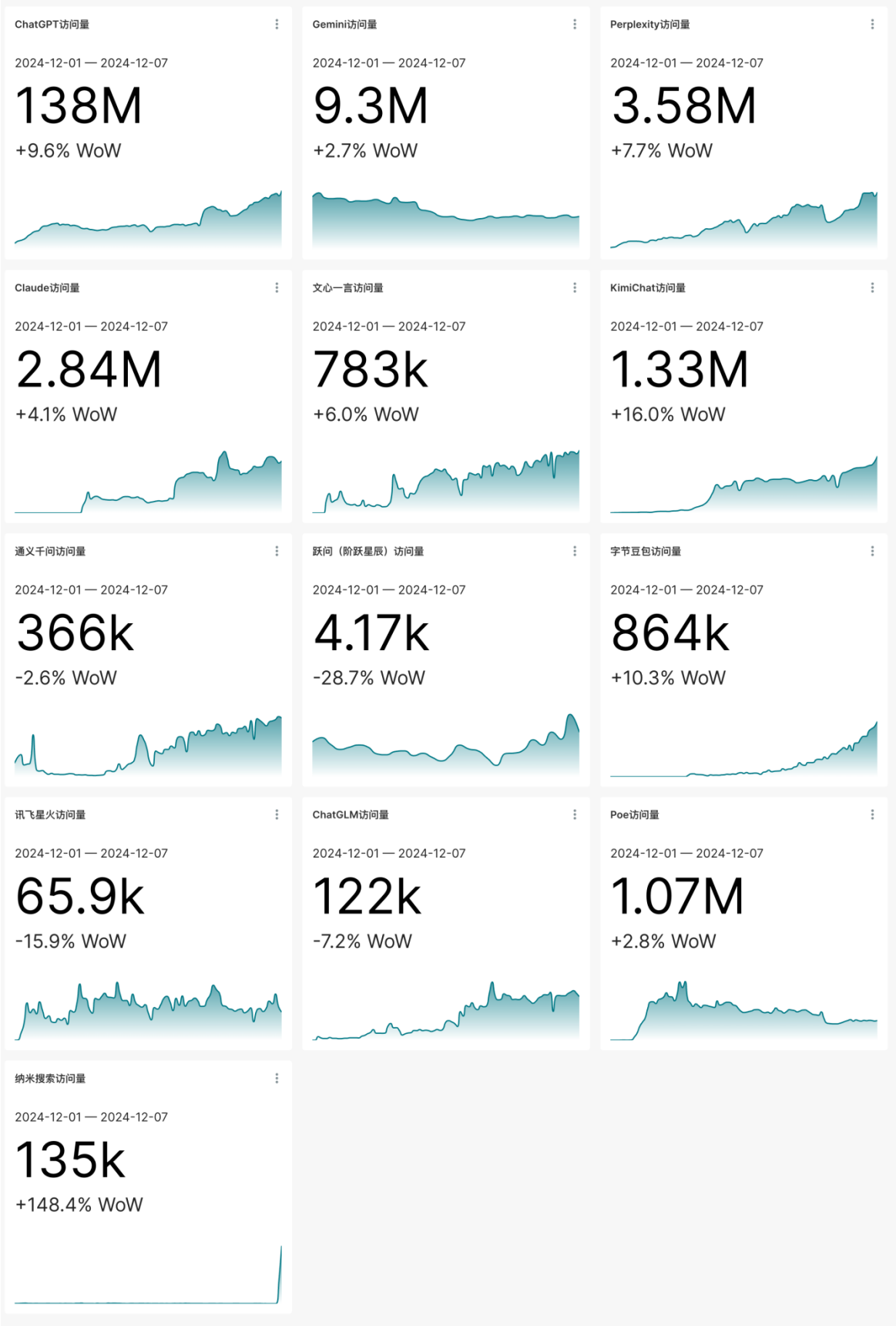
半导体解决方案部门营收 73 亿美元，AI 产品销售表现亮眼，自研芯片营业额同比增长 3.5 倍，以太网交换解决方案营业额增长超 4 倍。非 AI 网络业务环比增长 17%，但同比下降 41%，预计将在 Q4 稳定。

展望 Q4，博通指引营收 140 亿美元，同比增长 51%，其中半导体营收 80 亿美元，软件营收 60 亿美元。全年 AI 收入预计 120 亿美元，高于此前指引。公司对 FY2025 保持乐观，预期 AI 和 VMware 将持续增长，非 AI 半导体业务将逐步复苏。

当前市场预期 FY24Q4 公司 AI 相关收入为 34.6 亿美元，低于先前公司全年展望中隐含的四季度 35 亿美金的 AI 相关收入指引。我们认为预期下调为公司股价提供了更多安全边界，美满电子三季度自研芯片业务板块的强劲体现出该细分赛道的高景气度，博通作为该细分赛道龙头，将持续受益。

AI 模型与应用动态

图表1：聊天助手类 AI 应用日活跃度



来源：SimilarWeb、国金证券研究所

AI 聊天助手应用活跃度持续上升，模型竞争加剧

从 AI 应用活跃度看，上周聊天助手类应用热度仍在持续上升，ChatGPT 周均日访问量环比上升接近 10%，达到 1.38 亿。国内 AI 应用如豆包和 KimiChat 有超过 10%的环比增速，纳米搜索在发布两周后仍有约 150%的环比增速。

在基础模型领域，亚马逊推出一系列模型，Amazon Nova 系列基础模型包括 Micro、Lite、Pro 和 Premier 四款。Micro 是纯文本模型，拥有 128k 的上下文窗口，响应速度快、性价比高。Lite 和 Pro 是多模态模型，支持高达 300k 的上下文窗口，能处理文本、图像和视频。Pro 模型在 20 个基准测试中有 17 个表现相当或优于市场上领先的 GPT-4o。Premier 模型设计用于更复杂的推理任务，将于 2025 年第一季度推出。此外，亚马逊还推出了 Amazon Nova Canvas 和 Amazon Nova Real 两个新模型，分别用于图像和视频的生成。Nova Real 展示了制作 6 秒短视频的能力，未来几个月将扩展至支持制作最长 2 分钟的视频。Meta 发布了 LLaMA 3.3 70B 作为 70B 参数规模的模型，在性能上可与更大规模的 LLaMA 3.1 405B 相媲美。大幅降低了推理成本和使用硬件门槛，使更多的个人开发者和小型团队能够使用。

OpenAI 2/12 Days 发布，o1 正式版、Pro 版订阅和强化微调发布，静候 Sora 发布

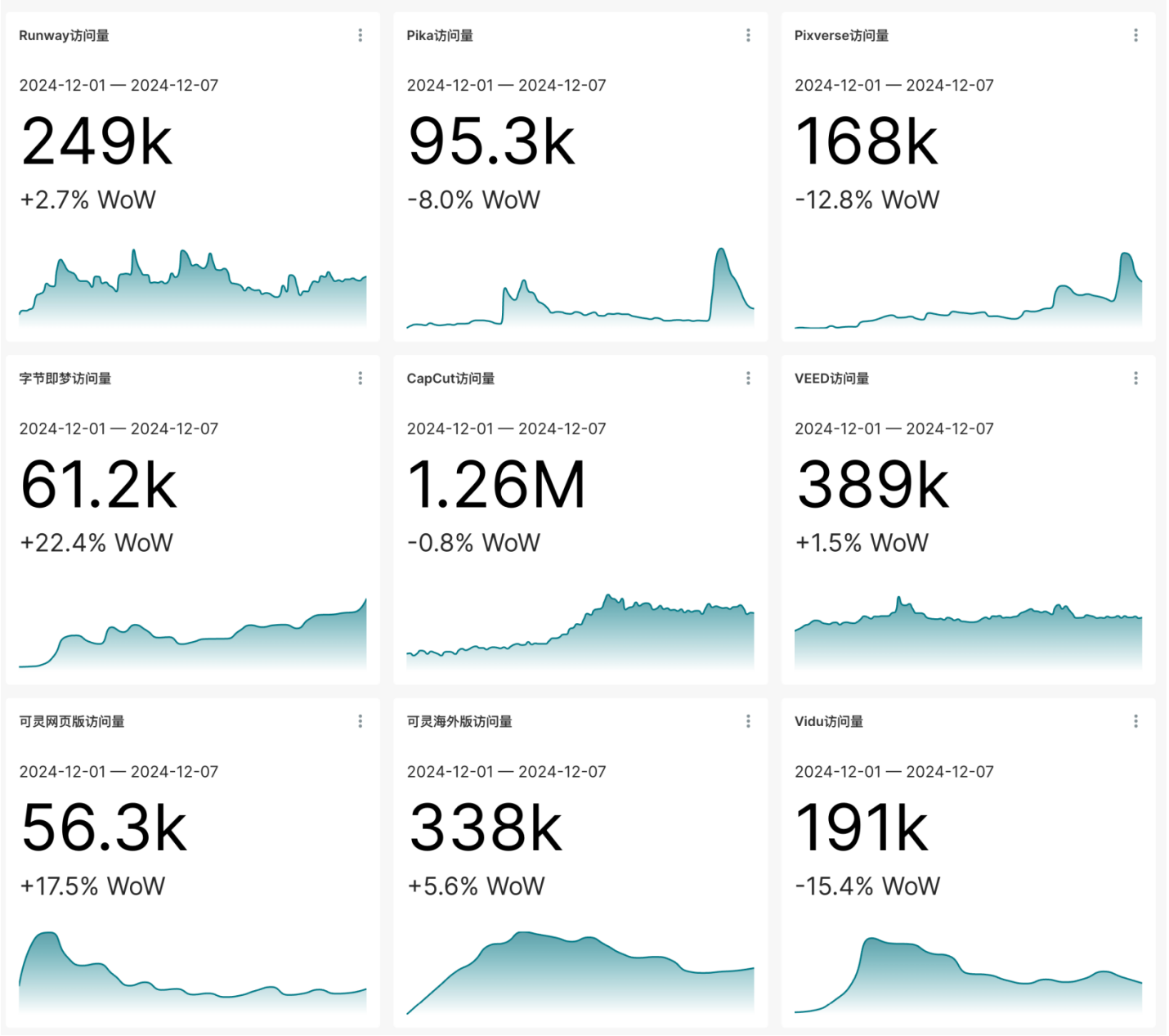
OpenAI 的 12 天发布计划已经进行了两天。第一天，OpenAI 发布了正式版 o1 。该版本新增了读取图片和文件的能力，为用户提供了更加丰富的交互方式。此外，OpenAI 还推出了每月 200 美元的 Pro 会员服务。Pro 会员权益包括能够使用思考深度更深的 o1 pro mode，并且不限量使用，能更好的满足对人工智能有更高需求的专业用户。

第二天，OpenAI 发布了强化微调（RFT），强化微调与传统微调不同，它并非仅仅依赖于微调数据，而是从微调数据中学习如何推理。这一创新技术采用两个不同的数据集，一个是微调数据集，另一个是测试数据集。模型首先基于微调数据集进行训练，随后用测试数据集进行验证，通过反复的自我推理训练验证过程，不断提升自身能力。强化微调功能的推出，为人工智能在各个领域的应用提供了新的可能性。例如在医疗领域，通过强化微调可以让模型更好地根据病例症状进行推理，找出相关病因。在法律领域，也能帮助模型更准确地分析法律文本，提供专业的法律建议。而 o1 正式版的新功能和 Pro 会员服务，则进一步提升了用户体验和使用价值。OpenAI 的这些创新举措，无疑将推动人工智能技术的发展，为各个行业带来更多的机遇和挑战。

OpenAI 在伦敦 C21Media 大会上透露，公司的 Sora 视频生成器即将推出更新版本。OpenAI 的代表查德·尼尔森在会上表示，这款新版本将会支持三种视频生成方式，具体包括：文字生成视频、文字和图像生成视频，以及文字和视频生成视频，每个视频的时长可达一分钟。Sora V2 版本的发布预示着开放给高级用户的 Sora 很可能在 OpenAI12 天发布会接下来的几天公开。

视频生成模型持续开源，世界模型热度再次上升

图表2：视频类 AI 应用日活跃度



来源：SimilarWeb、国金证券研究所

从视频生成应用活跃度看，目前一梯队的视频生成应用如 Runway、可灵，活跃度增速下降，活跃度略高于上周，而二梯队借助于新模型和新功能发布的应用如 Pika、Pixverse，活跃度开始回落，环比下降约 10%。市场仍在等待 Sora 的发布，对视频生成应用的整体热度会有促进作用。

腾讯于 2024 年 12 月 3 日正式推出开源的混元视频生成模型，参数量 130 亿，是当前最大的视频开源模型之一。该模型基于 DiT 架构，采用了新一代文本编码器、统一的全注意力机制、3D 形状变分自编码器等技术，具备智能场景理解、真实动作捕捉等能力，能够实现超写实画质，生成的视频画面流畅、不易变形，且光影反射基本符合物理规律，尤其在人物、人造场所等场景下表现出色。混元视频生成模型已在 HuggingFace 平台及 GitHub 上发布，包含模型权重、推理代码、模型算法等完整内容，可供企业与个人开发者免费使用和开发生态插件，有助于加速行业创新步伐.

谷歌 DeepMind 发布了大型基础世界模型 Genie 2，能够根据单张图片和文字描述生成具有交互功能的 3D 世界。该模型生成的虚拟世界具有丰富的动态效果和多样化环境，能够模拟对象交互、动画、照明、物理反射和 NPC 行为等，具备空间记忆与反设事实能力，能够在长达一分钟内维持世界的连贯性与一致性，广泛应用于游戏制作与 AI 智能体训练等领域。与此同时，World Lab 推出了基于单张图片生成交互性 3D 虚拟世界的世界模型。该模型遵循 3D 几何与物理基本规则，展现出逼真的深度与空间感，可作为专业创作工具，助力 VR 数字空间的内容填充，在 3D 场景重建与视觉效果生成等领域展现出重要应用价值。

图表3：字节即梦 2.1 生成的包含中文的图片

来源：沃垠 AI、国金证券研究所

字节即梦 2.1 开始灰度测试，解决了 AI 图像生成中文字体的问题，支持直接在图片上画出中英文字体。目前，它虽在灰度测试阶段，但已经展现出了巨大的潜力。用户可以用它制作各种类型的海报，如双 12 购物节海报、元旦祝福海报、影视海报等。虽然在生成过程中还存在一些瑕疵，如字体出现锯齿、写法问题，以及会重复或多出一些语句等，但对于制作一些文案不太复杂的海报效果较为稳定。

亚马逊发布 Trainium2，看好 2025 年自研芯片大规模部署

Amazon Web Services 本周于 re:Invent 大会上推出了其新一代人工智能加速器 Trainium2，与其前代产品相比，这款加速器的性能显著提升，使 AWS 能够对具有数万亿参数的基础模型和大型语言模型进行训练。此外，AWS 设定了雄心勃勃的目标，为其客户提供 65 ExaFLOPS 的强大性能来支持 AI 工作负载。

AWS 的 Trainium2 芯片采用先进的多芯片集成(System-in-Package,SiP)设计，由两个计算核心和四组 HBM3e 高带宽内存堆叠组成。每个计算核心通过 CoWoS-S/R 封装技术与其相邻的两组 HBM3e 内存进行通信。此外，这两个计算核心之间通过 ABF 基板相互连接，实现高效数据传输。

图表4：Trainium2 结构示意图

来源：Anandtech、Semianalysis、国金证券研究所

AWS 的 Trainium2 加速器在网络拓扑和算术强度方面与其他主流 AI 加速器如 Google TPU 和 NVIDIA H100 存在显著差异。Trainium2 采用 2D/3D 环形网络拓扑结构，16 芯片型号使用 2D 环形网络，而 64 芯片型号则采用 3D 环形网络。这种点对点连接方式与 Google TPU 相似，而 NVIDIA 的 NVLink 拓扑则通过交换结构实现全互联。

在算术强度方面，Trainium2 的算术强度为 203 BF16 FLOP/字节，低于 TPU 和 H100 的 300 到 560 BF16 FLOP/字节。算术强度反映计算吞吐量与内存带宽的比率，适用于内存带宽受限的任务。Trainium2 选择较低的算术强度是为了适应机器学习模型的发展趋势，例如专家混合模型在加载权重时对内存的需求较大。

此外，Trainium2-Ultra 的最大扩展规模为 64 颗芯片，而 TPU 的最大规模为 256 颗芯片。因此，Trainium2 的整体峰值计算性能较低，但其在内存带宽受限的任务中表现更优，适合推理和专家混合模型等应用场景。

Trainium2 已经获得了大规模应用，AWS 正在印第安纳州部署一个包含 40 万颗 Trainium2 芯片的集群，为 Anthropic 的 AI 训练提供算力支持。该数据中心园区规划总 IT 功率达到 1,040MW，展现了 AWS 对 Trainium2 的高需求和大规模算力投入。尽管面临同步计算挑战，Anthropic 通过异步训练等技术创新来充分利用这一大规模集群，显示出 Trainium2 在实际 AI 任务中的应用潜力和市场认可。

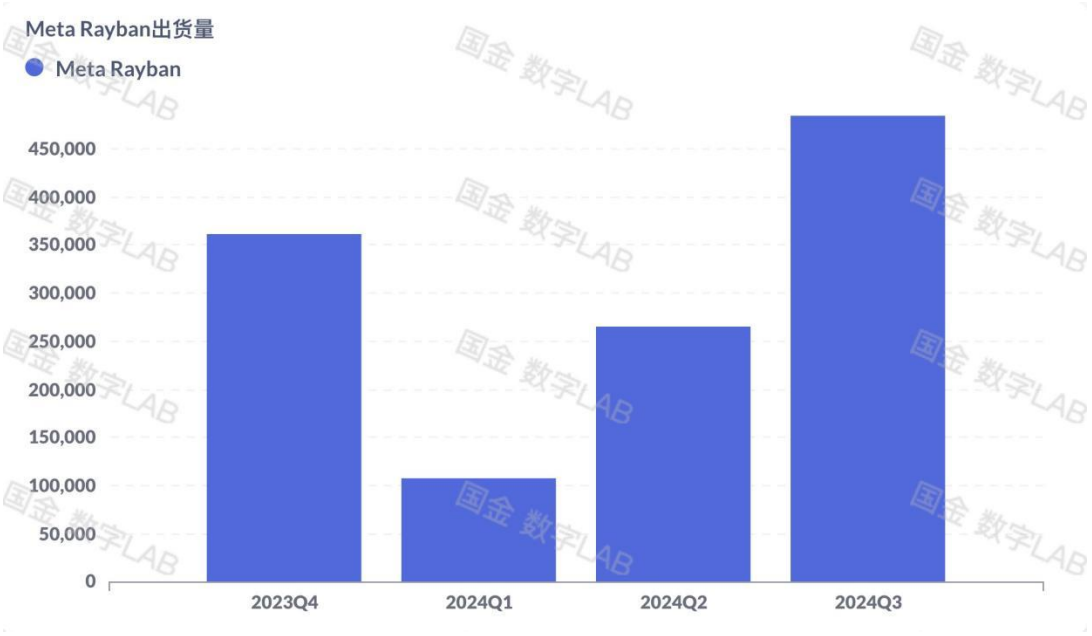
随着加速器部署规模的扩大和推理需求的增长，大型云厂商在加速器采购决策中将愈加注重总拥有成本（TCO）和工作负载特性的匹配。Trainium2 所具备的低算术密度正是 AWS 针对推理负载进行精细化优化的体现。在这一趋势下，自研芯片的部署有望加速推进，而自研芯片供应商也将迎来持续的业绩增长。

AI 硬件

Meta Rayban 三季度销量环比大涨，看好 AI 眼镜未来发展

根据 IDC 数据，2024 年三季度，Meta Rayban AI 眼镜全球出货量达到约 48 万部，环比增长 83%，2023 年四季度发布至今累计销量达到约 122 万部。一季度 Meta AI 功能上线后对销量提振效果显著，24 年 Q4 双节大促对销量会有更强的提振效应，AI 眼镜的接受度得到提升。我们认为，在 2025 年具备音频和摄像头的 AI 眼镜是当下 AI 模型应用落地的最佳可穿戴设备，随着多模态模型能力的提升和 AI Agent 的成熟，产品功能性和应用场景将获得极大提升，持续看好 2025 年 AI 眼镜大规模放量。

图表5：分季度 Meta Rayban 全球出货量（台）



来源：IDC、国金数字未来实验室、国金证券研究所

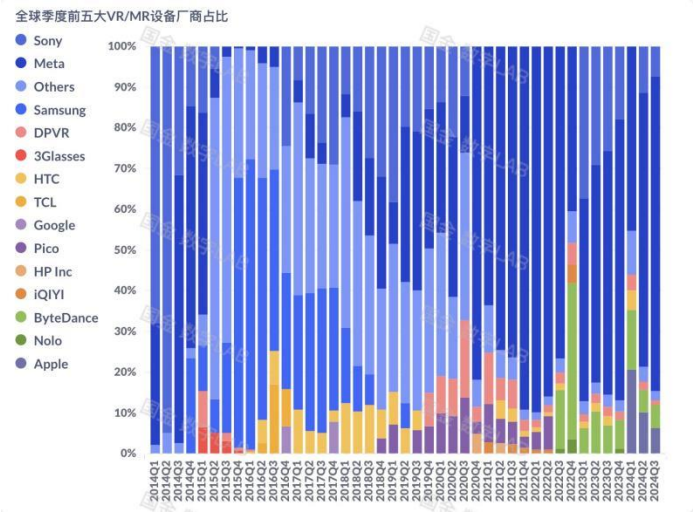
Meta Quest 3 占据 AR/VR 主要市场

根据 IDC 数据，2024 年三季度全球 VR/MR 设备销量约为 152 万台，同比增长 13%，其中 Meta Quest 3 销量约为 118 万台，占比约为 77%；索尼 PlayStation VR2 销量约为 11 万台，占比约为 7%；苹果 Vision Pro 销量约为 9.6 万台，占比约为 6%；Pico 4 系列销量总计约为 7.5 万台，占比约为 5%。

图表6：全球 VR/MR 设备销量（台）及增速

来源：IDC、国金数字未来实验室、国金证券研究所

图表7：全球季度前五大 VR/MR 设备厂商占比



来源：IDC、国金数字未来实验室、国金证券研究所

由于长期佩戴体验不佳&部分性能优异产品价格过高，短期内 VR/MR 产品难以在数量上爆发。但长期来看，得益于身临其境般的视觉体验，VR/MR 设备非常适合作为元宇宙的入口打开新的空间维度。目前 Meta 凭借其 Reality Lab 在 VR 上多年的技术积累与产品低售价的定位成功抢夺大部分 VR/MR 市场，而高端市场则被苹果 Vision Pro 系列掌控。我们认为，在代表高性能的苹果 MR 产品价格逐渐下沉后并且 Meta MR 设备性能逐渐上升后，VR/MR 设备的市场将会进一步打开，最终进入传统消费电子的逻辑：性能&性价比。

风险提示

1. 芯片制程发展与良率不及预期：半导体工艺的发展面临诸多挑战，主要包括技术瓶颈、良率提升难度、研发成本高企以及供应链不确定性等问题。随着工艺节点微缩变得愈发复杂，先进制程的实现难度和成本不断攀升，可能导致量产延迟，甚至影响产品性能和成本控制。此外，地缘政治风险和出口管制可能扰乱供应链，进一步拖累产能扩张。

2. 中美科技领域政策恶化：中美在 AI 领域竞争激烈，美国限制先进芯片和半导体对中国的出口，随着竞争的加剧，未来可能会推出更严格的限制政策，限制国内 AI 模型的发展。

3. AI 硬件销量不及预期：AI 硬件销量与产品本身质量关系紧密，若产品本身有缺陷则 AI 硬件销量可能收到影响。同时宏观经济变化也有可能导致消费者消费意愿发生变化从而影响 AI 硬件销量。

特别声明：

国金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。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

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、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

上海
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究