国金证券 SINOLINK SECURITIES

2024 年 10 月 28 日

AI 周观察

数据专题研究(动态)

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 应用与端侧设备热度持续，英伟达新品延迟问题已解决

报告摘要：

- 本周发布季报的海外 AI 相关的重点公司有：芯片相关的 AMD 和英特尔，重点关注 AI 芯片的营收贡献和指引；云服务和 AI 应用相关的谷歌、微软、亚马逊，重点关注 CapEx 的走向与 AI 对收入的贡献情况；AI 硬件相关的苹果和三星电子，重点关注端侧 AI 对手机等消费电子销售的带动作用。
- 从 AI 应用的日活跃度数据看，AI 应用的热度仍然在上升，大语言模型的聊天应用如 ChatGPT 日活跃度达到了新高，国内的 AI 聊天应用的活跃度也维持在高位。新的模型与功能的推出速度也在加快，Anthropic 发布了新版 Claude，支持 Computer Use 控制电脑，The Verge 透露 OpenAI 年底之前下一代模型将会出现。视频模型在快速发展阶段，闭源模型如 Runway 和可灵的活跃度较为稳定，新发模型对应用活跃度仍然有较大的提升，可用的开源视频生成模型也开始出现，对算力提出更高的要求。
- 英伟达 Blackwell 的延迟问题已解决，且台积电的 CoWoS 产能持续扩张，为 2025 年市场需求的稳步增长提供了保障，Blackwell 供应稳定。短期来看，市场供需关系较为健康，但我们认为需关注 Rubin 系列未来迭代速度可能放缓的风险。
- 科技巨头出于成本控制及多元化供应链的需求，正在积极尝试使用 AMD 集群。然而，受限于互连能力和较为薄弱的软件生态，短期内 AMD 在大规模集群市场的拓展空间有限，其主要机会仍集中在小规模模型训练和推理负载领域。鉴于 MI 系列 GPU 的营收基数较低以及科技公司业务规模的带动，AMD 的数据中心业务有望在未来数个季度实现显著的营收和利润增长。建议关注即将发布的三季报中 MI 系列 GPU 部署情况及营收增长趋势。
- 亚马逊 Trainium2 实现了大规模应用，标志着定制芯片发展迎来新里程碑。Databricks 与 AWS 达成战略合作，将使用 Trainium2 芯片进行 Mosaic AI 训练，我们认为这为定制芯片的放量奠定基础。
- 传统存储市场需求依旧疲软，DRAM 和 Flash Wafer 价格持续下滑，存储市场整体处于下行趋势中。AI 应用带来结构性增长机会，海力士第三季度表现突出，其 DRAM 和 NAND 部门受 AI 相关的 HBM 和 eSSD 需求推动，保持了高速增长。我们认为 HBM 需求将继续强劲，持续带来存储结构性成长机会。
- 手机厂商旗舰发布季，竞争愈发激烈&消费者买单。10 月，vivo、OPPO、荣耀、小米等先后发布旗舰机。虽然由于手机 SoC 及部件价格增长导致手机价格有所增长，但消费者对高端手机热情依旧。根据 vivo 公告，X200 系列手机全渠道销售金额已经突破了 20 亿元，这一数据打破了 vivo 历史上所有新机销售记录。随着今年联发科天玑 9400 性能的再一次大提升，安卓系手机厂商旗舰机 SoC 再不是高通一家独大。我们认为联发科的加入将会降低手机厂商对于高通的依赖，刺激手机供应链有利竞争，有助于更加自由的研发满足消费者需求的产品，促进高端机型销量的提升。同时今年各品牌手机 SoC 性能整体提升幅度较大叠加配件提升与 AI 功能的加入，我们认为消费者对于换机的意愿有所增强。

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

智能手机销量不及预期

内容目录

图表目录

财报日历与前瞻

图表1：美股 AI 相关季报日历与一致预期

来源：Reuters、国金证券研究所

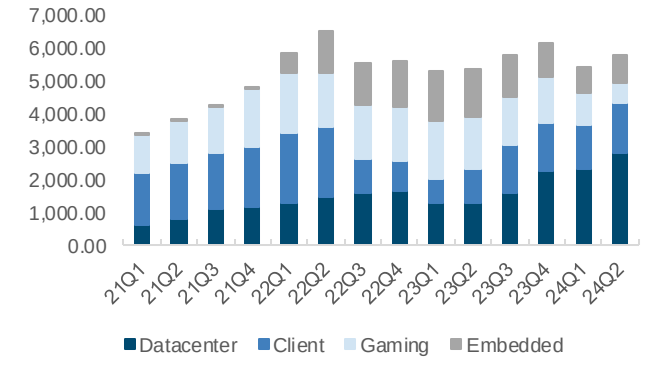
市场预期高速盈利增长，关注 AMD MI 系列 GPU 营收贡献表现

2024 年第二季度，AMD 实现了 9%的同比收入增长和 18%的非 GAAP 每股收益（EPS）增长，表现超出市场预期。尽管游戏和嵌入式业务仍面临压力，AMD 在数据中心和 AI 市场的拓展取得显著进展，尤其是 Instinct、Ryzen 和 EPYC 处理器的强劲表现使其获得了主要云计算和企业客户的青睐。展望 2024 年第三季度，预计 AMD 的增长动力将主要来源于数据中心和 AI 产品线的持续扩展。

根据公司上季度财报电话会议，管理层对未来增长前景保持乐观，尤其是在数据中心和客户端市场。预计第三季度收入将达到 67 亿美元，并且数据中心业务的高毛利 AI 产品的推出可能将毛利率提升至 50%以上。公司已于近日推出 MI325X 加速器，并在 2025 年发布 MI350 系列 AI 加速器，这将有助于其在未来 1-2 年内更具竞争力。此外，ZT System 的收购为 AMD 将加强其在软件方面的开发和支持能力，有望加速其在数据中心市场的渗透。

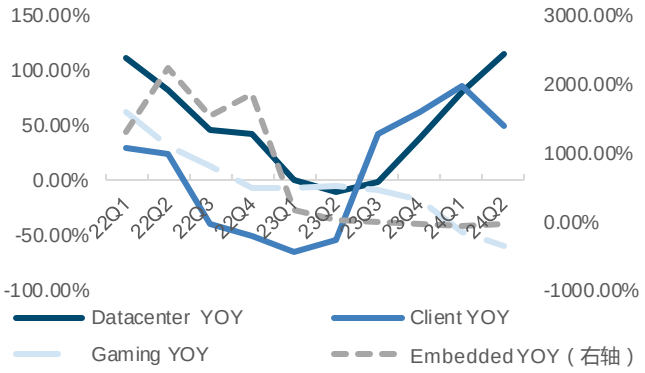
我们认为当前 AI 是公司发展的核心叙事，建议重点关注本季度公司 MI 系列 GPU 在下游厂商中的部署进展及其带来的营收增长趋势。

图表2：AMD 各部门营业收入



来源：Bloomberg、国金证券研究所

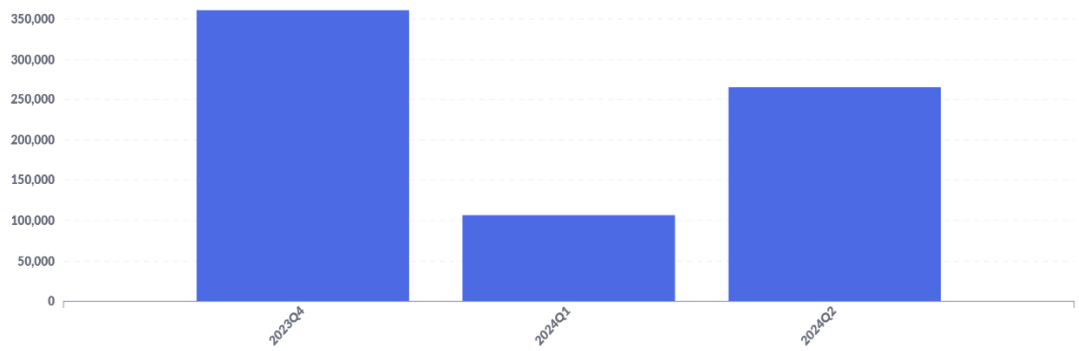
图表3：AMD 各部门营业收入同比增速



来源：Bloomberg、国金证券研究所

Meta 重点关注 AI 驱动的推荐与广告效果以及 Meta Rayban 的销售情况

图表4：Meta Rayban AI 眼镜季度销量



来源：IDC、数字未来实验室、国金证券研究所

Meta AI 随着 Llama3.1 的发布，模型能力得到了较大的提升，需关注其用户增长和在 Meta 的 App 全家桶中的使用情况，包括用户日活月活和 Meta AI 使用次数；广告业务作为 Meta 的营收主要增长来源，需关注由 AI 驱动的广告工具 Advantage+的采用率和对广告效果的影响；Reality Lab 业务中，Meta Rayban 是短期内唯一有盈利希望的产品，需关注其 Q3 的出货量与 Q4“黑五”大促的销售指引。

云厂商重点关注 CapEx 的走向与 AI 对收入的贡献情况，Google 需关注 GCP 的增速能否持续，以及 AI 相关工作负载的需求变化；Amazon 重点关注 Bedrock AI 模型 API 业务的发展情况以及为 Anthropic 提供的算力需求的变化。

苹果 Apple Intelligence 进展略缓，M4 芯片 Mac 将加速 AI PC 渗透

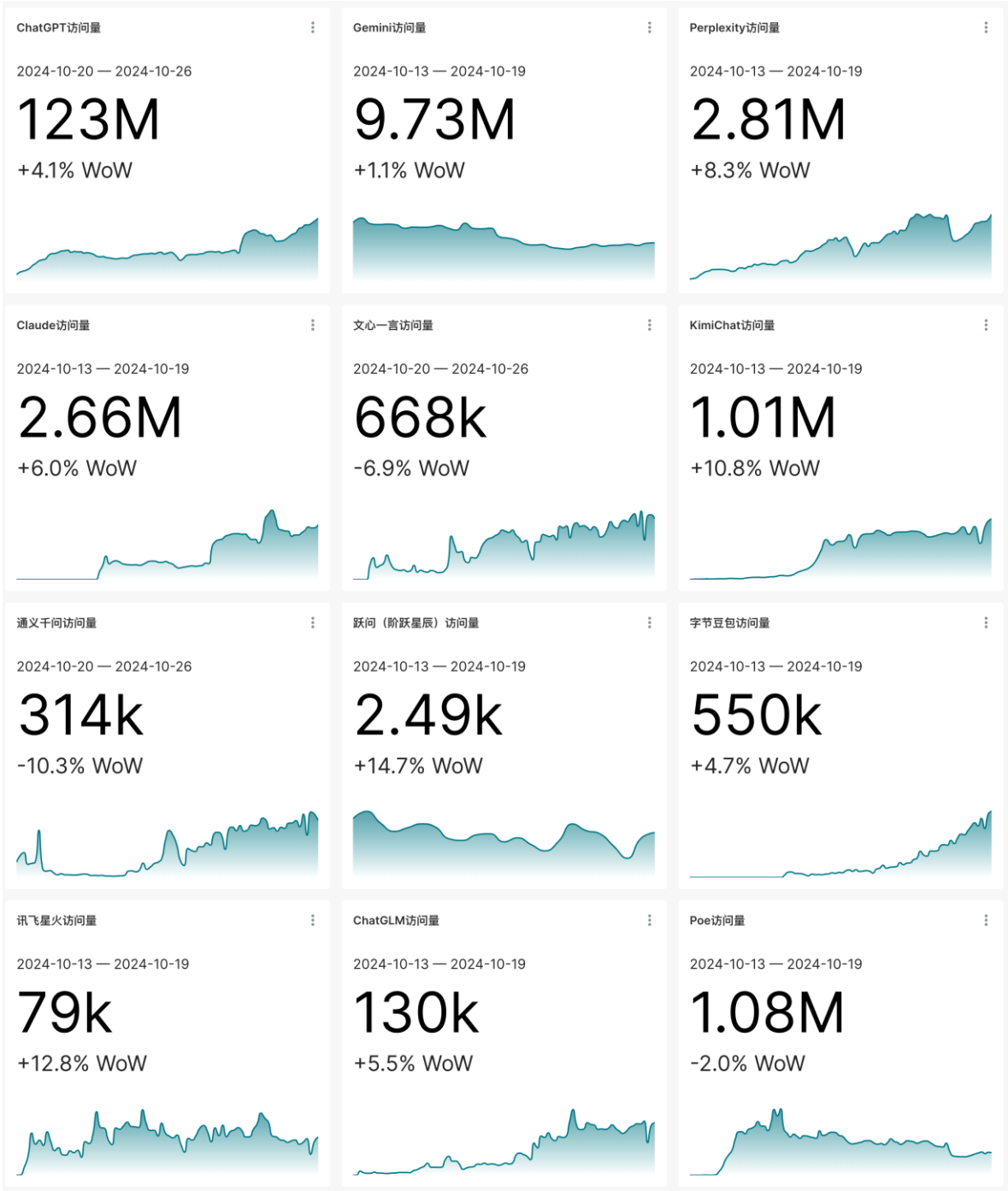
10 月 23 日，苹果宣布 Apple Intelligence 的正式公开版本将于下周在 iOS 18.1 中上线。iOS 18.1 中引入 Apple Intelligence 的部分功能，包括支持文本校对与改写的写作工具套件、照片清理、通知摘要，以及 Siri 增强功能。同一天，苹果还面向开发者推送了 iOS 18.2 版本的预览版，AI 除了有文本改写工具、智能表情包 Genmoji、AI 图像生成器 Image Playground，和图像处理工具 Image Wand，还集成了 OpenAI 的 ChatGPT 功能。我们认为目前 iOS18.1 中许多 AI 工具尚不成熟，没有让人眼前一亮的 APP 在此次更新中展现。但随着 iOS18.2 的推出，苹果直接将 GPT 嵌入到操作系统中后，我们认为“Apple Intelligence”将会变得完善。由于“Apple Intelligence”进展不及预期，iPhone 16 系列销量受到影响。随着苹果 AI 逐步推进，新一代 iPhone 销量有望回暖。

除此之外，苹果在 X 发布 Mac 预告，预计下周 Mac 相关资讯将会公布。预计下周发布的新产品主要有：M4 芯片 14 英寸 MacBook Pro、M4 Pro&M4 Max 芯片 14/16 英寸 MacBook Pro、、M4、M4 Pro 芯片的 Mac Mini、搭载 M4 芯片的 iMac 等。其中部分产品有望在 11 月 1 日正式上市发售。我们认为苹果新一代 PC 的发布有助于进一步加速 AI PC 的渗透。在高通 X

Elite 芯片遭遇瓶颈的情况下，ARM 架构 AI PC 苹果一家独大，将占据整个 AI PC 市场的一半。其余的市场将会由 AMD、英特尔等联合传统的 X86 PC 厂商抢占。

AI 模型与应用动态

图表5：聊天助手类 AI 应用日活跃度



来源：SimilarWeb、数字未来实验室、国金证券研究所

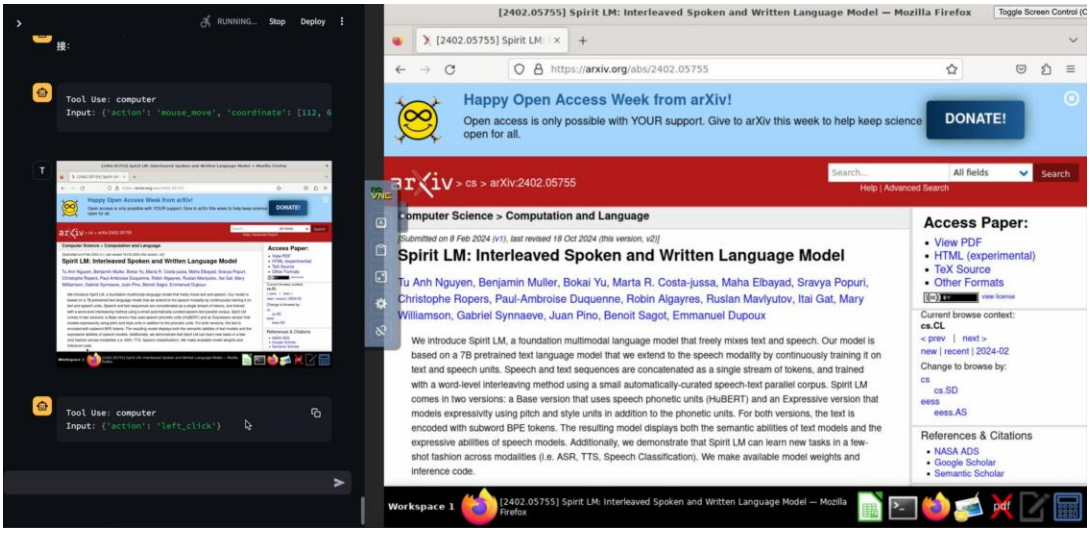
AI 聊天助手应用热度持续增长，新功能不断推出

从 AI 应用的日活跃度数据看，AI 应用的热度仍然在上升，大语言模型的聊天应用如 ChatGPT 日活跃度达到了新高，国内的 AI 聊天应用的活跃度也维持在高位。

Anthropic 发布一系列更新 1）新版 Claude Sonnet 3.5 推理和代码能力进一步提升，超

过 OpenAI o1。2）新模型 Haiku 3.5，编程能力超过原版 Sonnet 3.5，而且规模更小速度快得多。3）Computer Use：通过调用 Claude API 来自动化操作电脑，完成指定任务。对 Computer Use 实测后，我们发现，该功能对错误操作的反思能力超过预期，对于大多数工作的完成度较高，但是仍然存在反应慢，对网页的解析不够精准等问题。AI 系统级别的助手开始成为现实，后续增加端到端训练操作系统的专用模型会改进目前的缺点。

图表6：Claude Computer Use 功能实测

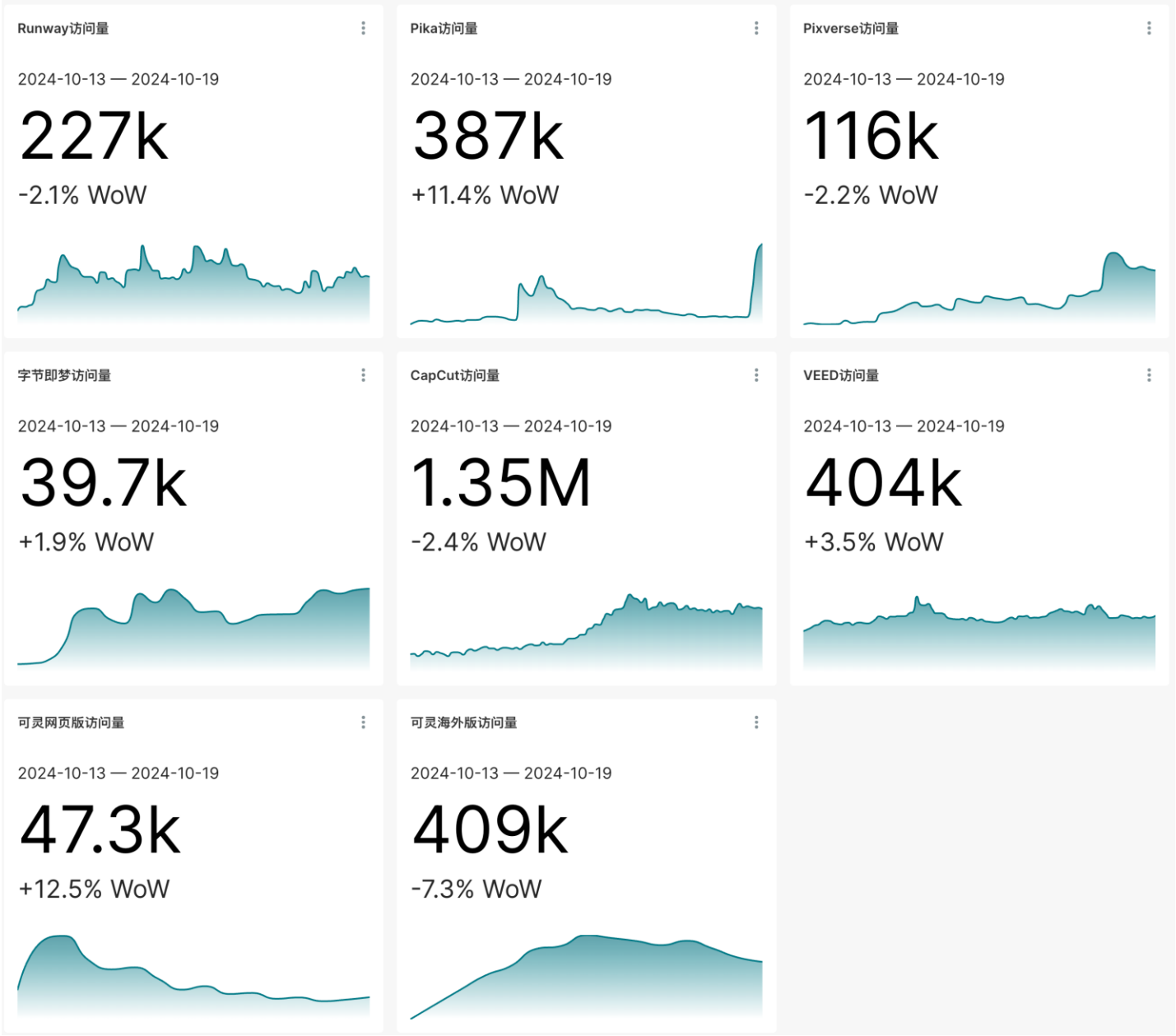


来源：数字未来实验室、国金证券研究所

据报道，OpenAI 计划在 12 月前推出其下一个前沿模型 Orion。与 OpenAI 上两个模型 GPT-4o 和 o1 的发布不同，Orion 最初不会通过 ChatGPT 广泛发布。

Meta 发布了 Spirit LM，一个语音多模态模型，支持输入输出文本和语音，分 Base 和 Expressive 两个版本。对语音进行续写的同时也能学习到输入语音的语气语调和音色，有潜力成为文本/音频模型中的基底模型，经过微调可以实现语音识别或者语音生成等任务。

图表7：视频类 AI 应用日活跃度



来源：SimilarWeb、数字未来实验室、国金证券研究所

视频生成模型快速发展，开源高质量模型开始出现

视频模型在快速发展阶段，闭源模型如 Runway 和可灵的活跃度较为稳定，新发模型对应用活跃度仍然有较大的提升。快手的可灵国际版实现了 AI 模型出海，属于现在可用模型中在海外的评价较高的视频生成模型。开源的视频模型也在出现，包括 Meta 的 Movie Gen 和 Mochi 1。视频模型对算力需求的提升符合我们的预期，比如未量化版本的 Mochi 需要 4 个 H100 才能进行推理。Runway 推出 Act-One 功能，可以同步表情和嘴型，自由更换角色、画风和背景，输入视频就可以拍动画片或者特效电影。

智谱推出了 Emu3，与目前使用扩散模型或者 DiT 的图像/视频生成模型不同，Emu 使用单一 Transformer 进行下一 token 的预测来生成图像/视频，为图像/视频生成探索出了新的路线，效果超出预期但是生成时间很长，使用 L40S 显卡推理生成 1 张图需要十几分钟；Stable Diffusion 3.5 Large 发布 8B 参数，支持图里写英文，年收入小于 100 万的情形可以商用，是目前开源的领先基底模型，为后续的微调提供了更好的基础；Comfy Org 发布了 ComfyUI 官方客户端，跨平台、自动更新，致力于解决自定义节点的适配问题。

GPGPU 市场动态

Blackwell 设计缺陷情况及料号更新

据悉，Blackwell 系列的设计缺陷主要发生在连接两个 die 的环节。根据 Semianalysis 的报告，这一问题主要是由于 Blackwell 的热膨胀系数与封装材料不匹配，导致 CoWoS-L 的良率不足。不过，目前这一问题已通过重制掩膜完全解决。英伟达预计将在 2024 年第四季度开始小批量生产 GB200。

料号方面，近日英伟达对其 Blackwell 系列产品名称进行了更改，取消了 B200A，新增了 B300，B300A，GB300 和 GB300A，在此做个梳理：

1）取消 B200A，保留 B200：在这一代产品中，尾缀 “A” 代表仅配备一个 GPU die，相应的 HBM die 数量减半。B200A 实际上是配置减半的 B200 产品，采用 CoWoS-S 封装形式。但目前这一料号已被取消。

2）B300A（原 B200A Ultra）：B300A 原先命名为 B200A Ultra，该版本相较于普通料号使用了 12 层堆叠的 HBM3E，而非 8 层堆叠，使单个堆叠的 HBM 容量从 24GB 提升至 36GB。更名后的 B300A 配备了一个 TSMC N4P Blackwell 架构计算 die，加上 4 个 12 层堆叠的 36GB HBM3E die。

3）B300（原 B200 Ultra）：更名前为 B200 Ultra，该版本配备两个 Blackwell 架构计算 die，并搭配 8 个 12 层堆叠的 36GB HBM3E die，可以视为 B200 的 HBM 增配版本。

4）GB300 和 GB300A（原包含 Grace CPU 的 B200 Ultra 和 B200A Ultra）：这两款产品包含 Grace CPU，其中 GB300A 对应于 B200A Ultra，而 GB300 对应于 B200 Ultra。

英伟达此次更名使产品识别更加简洁明了，数字越大代表更高的 HBM 配置，尾缀 “A” 则表示单 die 版本。由于单 die 版本的面积较小，且无需连接两个 die，因此这些版本将全部采用 CoWoS-S 封装形式。此外，在包含 Grace CPU 的版本中，CPU/GPU 的比例将从 1/2 降至 1/4。预计从 2025 年起，单 die 版本将逐步取代 CoWoS-S 产线上原有的 Hopper 产能。

图表8：Blackwell 系列料号整理

料号名称	计算托盘配置	GPGPU 计算/存储配置	封装形式
B200	8 x Blackwell	Dual-die + 8 x 8-Hi HBM3E	CoWoS-L
GB200	2 x Grace + 4 x Blackwell	Dual-die + 8 x 8-Hi HBM3E	CoWoS-L
B300	8 x Blackwell	Dual-die + 8 x 12-Hi HBM3E	CoWoS-L
GB300A	1 x Grace + 4 x Blackwell	Single-die + 4 x 4-Hi HBM3E	CoWoS-S
GB300	2 x Grace + 4 x Blackwell	Dual-die + 8 x 12-Hi HBM3E	CoWoS-L

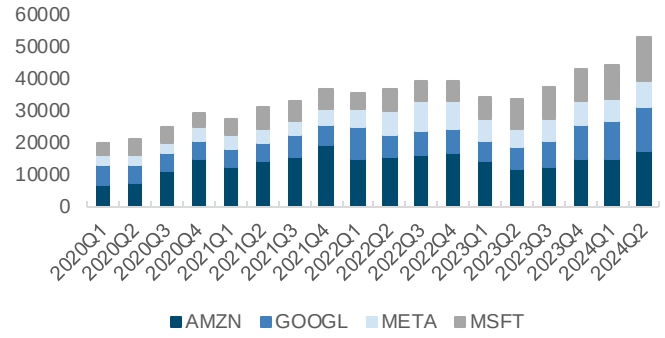
来源：TrendForce、国金证券研究所

Blackwell 需求旺盛，CoWoS 产能扩张进展顺利

英伟达的传统客户正大力推动对 Blackwell GPU 的需求，包括 AWS、谷歌、Meta、微软和 CoreWeave 等科技巨头。据悉，2025 全年英伟达的所有 Blackwell 芯片已被下游厂商预订，海外科技巨头在 AI 领域的资本支出依然保持高强度。根据趋势，2024 年第二季度，亚马逊、谷歌、Meta 和微软的合计资本支出同比增长率已提升至 58%，我们预计这一增长趋势将持续。

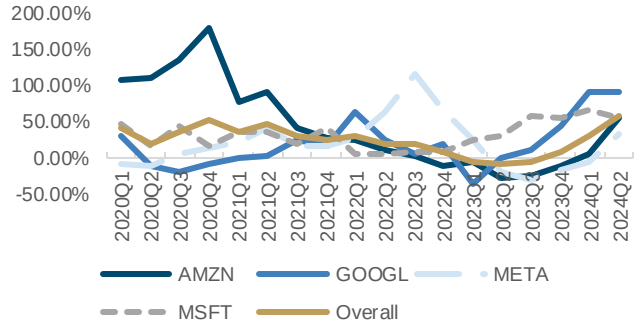
在科技巨头持续增长的资本支出背景下，Blackwell 系列的供应紧张情况将日益加剧，英伟达未来的营收将高度依赖 CoWoS 产能。台积电正在加速扩展其 CoWoS 产能，月产能预计在 2024 年将达到 35,000 至 40,000 片晶圆，并将在 2025 年大幅提升至 80,000 片/月，为英伟达后续 GPU 的出货提供一定的产能保障。

图表9：海外科技大厂资本开支持续增长



来源：Reuters、国金证券研究所

图表10：海外科技大厂资本开支保持高增速

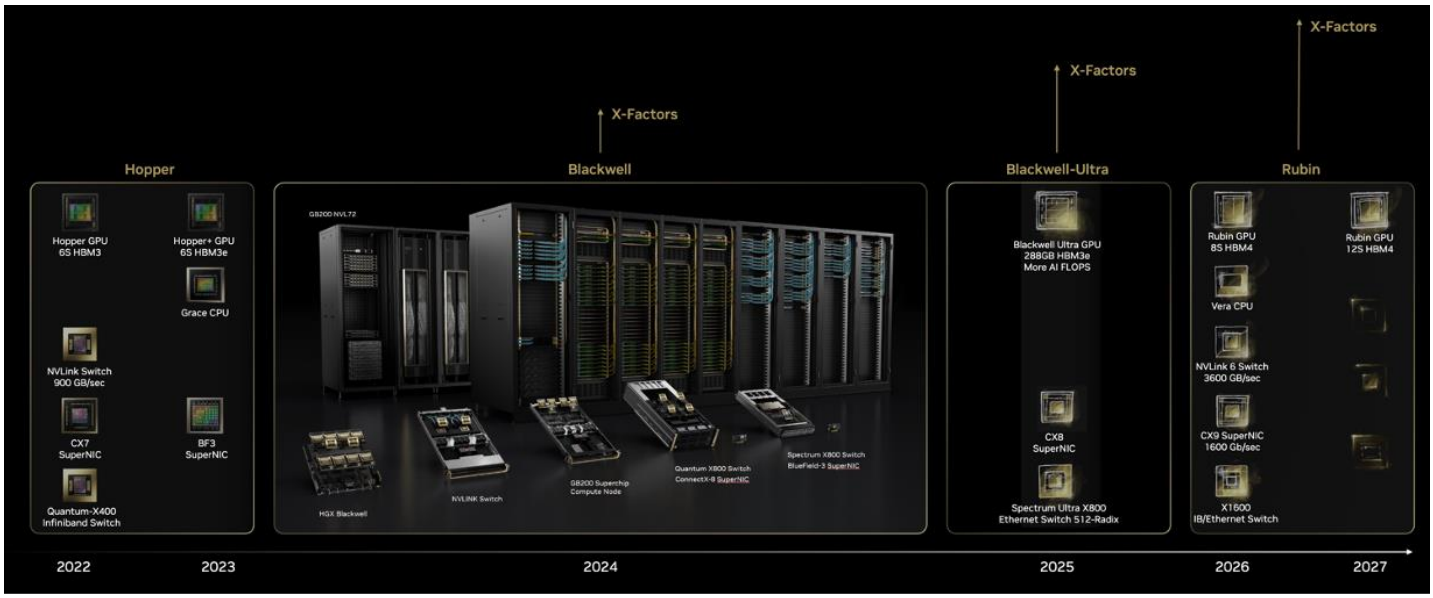


来源：Reuters、国金证券研究所

关键风险点：后续英伟达 GPGPU 产品更新节奏放缓

Blackwell 的设计延迟在一定程度上反映出当前半导体在先进制程迭代上的挑战。本世代采用双计算芯片（die）集成方案，侧面体现了晶体管缩放的难度日益加剧。未来的持续迭代将依赖于更大面积的硅中介层和更精密的晶体管技术。英伟达自 Blackwell 发布起，将产品周期从两年缩短至一年，提升了市场预期。我们认为应关注 Rubin 产品可能无法如期推出或性能提升不及预期的风险。

图表11：英伟达 GPGPU 迭代节奏提速



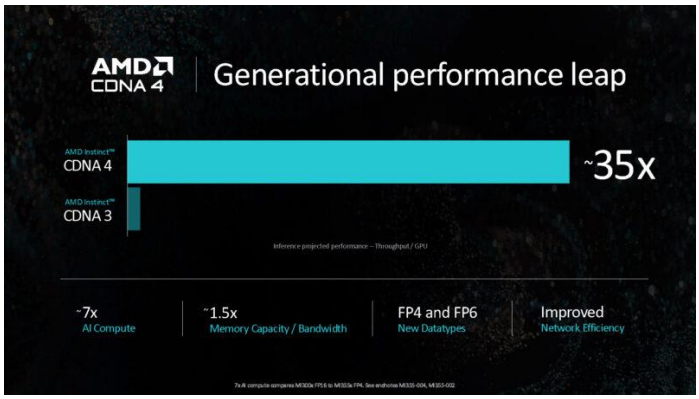
来源：英伟达、国金证券研究所

AMD 生态尚不支持其打入大集群市场，英特尔 Gaudi3 尚未现大规模应用

近日，AMD 推出了最新的 GPGPU 产品 MI325X，在 MI300X 基础上，进一步提升了性能。作为后起之秀，MI 系列在单卡算力方面已经超越了英伟达的 H200 SXM。MI325X 搭载了 256GB 的 HBM3E 内存，带宽高达 6TB/s，计算性能方面在 FP16 数据格式下达到 1.3 PetaFlops，而在 FP8 数据格式下更是达到 2.6 PetaFlops。

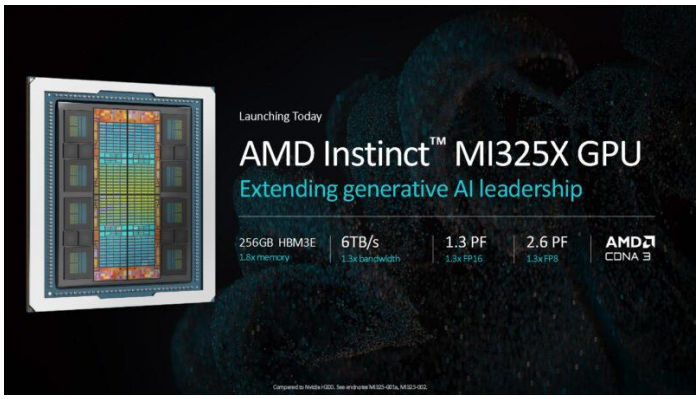
AMD 计划于 2025 年中期发布下一代 MI350 系列 GPU，采用全新 CDNA4 架构，并将使用台积电的 N3 工艺。此外，MI350 将开始支持 FP4/FP6 数据格式，以进一步提升其在高性能计算和 AI 领域的竞争力。

图表12：AMD 预计将于 CY25 年中发布 CDNA4 架构 GPU



来源：AMD、国金证券研究所

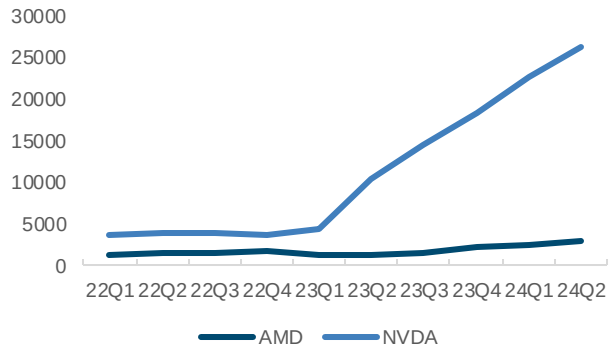
图表13：AMD 发布全新 MI325X GPU



来源：AMD、国金证券研究所

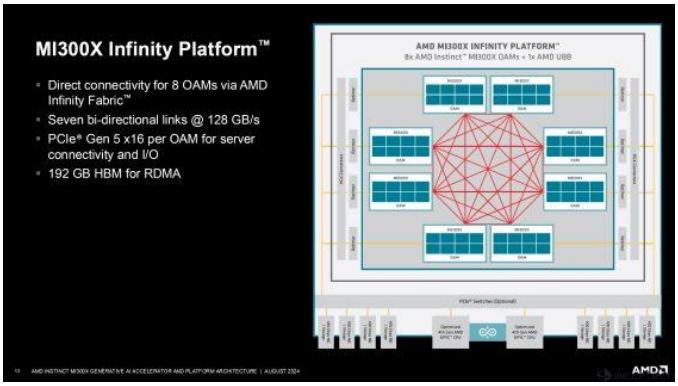
回顾 AMD 过去几代 MI 系列 GPU，其单卡性能的提升速度通常快于英伟达，尤其在存储方面表现突出。然而，市场反馈却不如预期理想。AMD 于 2020 年底发布了首款 MI 系列 GPU MI100，但其后续产品的市场表现有限。尽管旗舰级 MI300X 在 2024 年第二季度仅实现了 10 亿美元的营收。

图表14：英伟达和 AMD 数据中心营业收入差距持续扩大



来源：Reuters、国金证券研究所

图表15：AMD Infinity 平台



来源：AMD、国金证券研究所

AMD 的 MI 系列 GPU 在市场竞争中面临的主要制约因素，除了其软件生态上的 ROCm 显著落后于英伟达的 CUDA 外，集群互联能力和系统稳定性也不及英伟达。根据我们的调研，目前 AMD 的 Infinity Fabric 已迭代至第四代，双向带宽约为 900GB/s，仅为英伟达 NVLink 5.0 带宽的一半。受限于带宽不足及互联性能，MI GPU 集群的规模受到一定限制，在同等规模集群下，MI GPU 集群的稳定性和可维护性也不如英伟达的集群系统。

这种集群规模的限制直接影响了其承载负载的能力。目前，MI GPU 集群在大规模模型训练中的表现尚不理想，主要用于推理任务和小规模模型训练。然而，出于降低成本和寻找第二供应商的考虑，部分大型科技公司正在积极尝试使用 MI GPU 集群，但其在超大规模模型训练中的应用仍受到较大限制。

结合 AMD 过去在 ROCm 和 Infinity Fabric 上的表现，我们认为在未来两个架构更新周期（CDNA4 和潜在的 CDNA5）内，AMD 仍难以赶超甚至接近英伟达在这些技术上的领先地位。然而，鉴于当前科技巨头对低成本解决方案的强烈需求，我们预计 AMD 将从 2024 年第二季度起逐步在小规模集群市场中扩大份额，尤其是在以科技公司为主要客户的终端市场中取得进展。需要强调的是，鉴于这些科技公司本身的业务体量及其对第二供应商的高度需求，加之 AMD 在 GPGPU 市场的收入基数较低，即便 AMD 当前仅能突破小集群市场，其带来的收入增量依然有望相当可观。

对于企业端客户，由于其更为注重系统的稳定性和易用性，且通常不具备科技公司那样的技术研发和维护能力，倾向于选择完整的交钥匙（turn-key）解决方案。因此，我们认为这一部分市场的绝大多数份额仍将由英伟达占据。

图表16：Infinity Fabric 目前双向带宽仅为 900GB/s，仅为 NVLink5.0 的一半

AMD Instinct GPUs	MI25	MI50	MI60	MI100	MI210	MI250	MI250X	MI300X	MI300A	"Antares-A"
Chip	"Vega 10"	"Vega 20"	"Vega 20"	"Arcturus"	"Aldebaran"	"Aldebaran"	"Aldebaran"	"Antares"	GPU	CPU
Architecture	GCN	GCN	GCN	CDNA	CDNA-2	CDNA-2	CDNA-2	CDNA-3	CDNA-3	Zen 4
Process	14 nm GF	7 nm TSMC	7 nm TSMC	7 nm TSMC	6 nm TSMC	6 nm TSMC	6 nm TSMC	5 nm TSMC	5 nm/6 nm TSMC	5 nm TSMC
Transistors (Billion)	12.5	13.23	13.23	25.6	29.1	2 x 29.1	2 x 29.1	153, 12 chiplets	146, 13 chiplets	
Compute Units (Active)	64	60	64	120	104	208	220	304	228	24
Streaming Processors (Active)	4,096	3,840	4,096	7,680	6,656	13,312	14,080	19,456	14,592	–
Matrix Cores	–	–	–	–	–	–	–	1,216	912	–
Engine Clock Peak	1,500 MHz	1,746 MHz	1,800 MHz	1,520 MHz	1,700 MHz	1,700 MHz	1,700 MHz	2,350 MHz	2,350 MHz	3,700 MHz
FP64 Peak (Teraflops)	0.77	6.71	7.37	11.54	22.6	45.3	47.9	81.7	61.3	0.4
FP32 Peak (Teraflops)	12.29	13.41	14.75	23.07	22.6	45.3	47.9	163.4	122.6	0.8
FP16 Peak (Teraflops)	24.58	26.82	29.49	–	–	–	–	–	–	–
FP64 Matrix Peak (Teraflops)	–	–	–	–	45.3	90.5	95.7	163.4	122.6	–
FP32 Matrix Peak (Teraflops)	–	–	–	46.1	45.3	90.5	95.7	163.4	122.6	–
TF32 Matrix Peak (Teraflops)								653.7 / 1,307.4	490.3 / 980.6	
FP16 Matrix Peak (Teraflops)	–	–	–	184.6	181	362.1	383	1,307.4 / 2,614.9	980.6 / 1,961.2	–
Bfloat16 Peak (Teraflops)	–	–	–	92.3	181	362.1	383	1,307.4 / 2,614.9	980.6 / 1,961.2	–
FP8 Matrix Peak (Teraflops)	–	–	–	–	–	–	–	2,614.9 / 5,229.8	1,961.2 / 3,922.3	–
INT8 (Teraops)	–	–	–	–	181	362.1	383	2,614.9 / 5,229.8	1,961.2 / 3,922.3	–
Infinity Cache	–	–	–	–	–	–	–	256 MB	256 MB	
Memory Capacity	16 GB HBM2	16 GB HBM2	32 GB HBM2	32 GB HBM2	64 GB HBM2e	2 x 64 GB HBM2e	2 x 64 GB HBM2e	192 GB HBM3	128 GB HBM3	
Memory Clock	852 MHz	1.0 GHz	1.0 GHz	1.2 GHz	1.6 GHz	1.6 GHz	1.6 GHz	2.0 GHz	2.0 GHz	
Memory Bandwidth	436.2 GB/sec	1 TB/sec	1 TB/sec	1.23 TB/sec	1.6 TB/sec	2 x 1.6 TB/sec	2 x 1.6 TB/sec	5.3 TB/sec	5.3 TB/sec	
Memory Coherency	No	No	No	No	No	No	Yes	Yes	Yes	
Interface	PCI-E 3.0	PCI-E 4.0	PCI-E 4.0	PCI-E 4.0	PCI-E 4.0	OAM	OAM	OAM	Shared OAM	
Infinity Fabric Links (x16)	–	2	2	3	4	6	8	7 * IF + 1 * PCI-E	4 * IF, 4 * PCI-E / IF	
Infinity Fabric Bandwidth	–	184 GB/sec	184 GB/sec	276 GB/sec	400 GB/sec	600 GB/sec	800 GB/sec	896 GB/sec	512 / 1,024 GB/sec	
Max Power	300 W	300 W	300 W	300 W	300 W	500W/560 W	500W/560 W	750 W	550 W / 760 W Board Level	

来源：nextplatform、国金证券研究所

Databricks 宣布使用亚马逊 Trainium 运行其 Mosaic AI 模型，定制芯片放量在即

数据与 AI 公司 Databricks 近日宣布与亚马逊云服务（AWS）达成战略合作协议（SCA），以加速在 AWS 上构建 Databricks Mosaic AI 定制模型的发展。Databricks 将采用 AWS 的 Trainium 芯片作为首选的 AI 芯片，用于在 AWS 上为 Mosaic AI 模型提供训练和服务功能。双方的共同客户可以利用 Mosaic AI 在其私有数据上预训练、微调、增强和部署大语言模型（LLMs），并享受 AWS 的规模、性能和安全保障。此次扩展的合作伙伴关系还将包括 Databricks 在 AWS Marketplace 中的新集成。我们认为这标志着亚马逊定制芯片正走向正轨，未来有望迎来显著放量。

存储市场动态

传统存储价格持续下行，需求疲软难止跌势

随着上游部分产能的持续扩张，部分资源供应出现过剩，渠道现货资源价格已连续一段时间缓慢下滑，而需求端整体依然疲弱。近期，存储现货市场多数产品价格仍处于下行通道，十月 DRAM 晶圆价格继续环比下降，市场情绪短期内偏向悲观。华邦电子台中厂将从四季度起开始减产，稼动率下调至八成。

图表17：DRAM Wafer 月度涨跌幅

料号	Mar-24	Apr-24	May-24	Jun-24	Jul-24	Aug-24	Sep-24	Oct-24
DRAM:DDR3 2Gb 128Mx16 1600/1866	-3.03%	-1.04%	-2.11%	-3.23%	-4.11%	-4.49%	-4.71%	-2.47%
DRAM:DDR3 2Gb 256Mx8 1600/1866	-2.65%	0.91%	-2.70%	-0.93%	-2.80%	0.00%	-4.81%	-4.04%
DRAM:DDR3 4Gb 256Mx16 1600/1866	-1.90%	0.00%	-1.94%	-3.96%	-2.06%	-5.26%	-4.44%	-4.65%
DRAM:DDR3 4Gb 512Mx8 eTT	0.00%	3.23%	0.00%	-6.25%	-13.33%	-15.38%	-13.64%	0.00%
DRAM:DDR4 16Gb(1Gx16)3200	-4.77%	2.23%	-0.54%	-4.11%	1.14%	-4.24%	-4.13%	-1.23%
DRAM:DDR4 16Gb(2Gx8)2666 Mbps	-0.55%	0.82%	0.82%	-0.27%	4.88%	-1.55%	-0.52%	0.00%
DRAM:DDR4 16Gb(2Gx8)3200	-1.68%	1.42%	-2.53%	-3.17%	3.27%	-3.46%	-2.99%	-0.62%
DRAM:DDR4 16Gb(2Gx8)eTT Mbps	0.00%	0.72%	-0.36%	-1.08%	0.73%	-1.44%	-6.96%	-9.45%
DRAM:DDR4 4Gb 512Mx8 eTT	1.92%	1.89%	1.85%	0.00%	9.09%	1.67%	-3.28%	-5.08%
DRAM:DDR4 4Gb(256Mx16)2400/2666	-2.38%	0.81%	-1.61%	2.46%	3.20%	-2.33%	-4.76%	-2.50%
DRAM:DDR4 4Gb(512Mx8)2400/2666	-0.82%	0.83%	-1.64%	1.67%	4.10%	0.00%	-3.15%	0.00%
DRAM:DDR4 8Gb(1Gx8)2666 Mbps	-1.54%	1.56%	-2.05%	-1.05%	5.29%	-1.01%	-2.03%	-1.55%
DRAM:DDR4 8Gb(1Gx8)3200	-3.28%	0.56%	-3.37%	1.74%	4.57%	-3.28%	-3.95%	-4.71%
DRAM:DDR4 8Gb(1Gx8)eTT	-5.04%	-2.27%	-4.65%	-9.76%	0.00%	-5.41%	-8.57%	-7.29%
DRAM:DDR4 8Gb(512Mx16)2666 Mbps	-0.53%	1.06%	-3.14%	0.54%	3.23%	-6.25%	-5.56%	-3.53%
DRAM:DDR4 8Gb(512Mx16)3200	-4.37%	0.57%	-1.70%	-2.31%	3.55%	-1.71%	-2.91%	-1.20%
DRAM:DDR5 16G(2Gx8)4800/5600	-0.42%	4.48%	-1.22%	0.83%	5.33%	-0.58%	-2.74%	-3.42%

来源：DramExchange、数字未来实验室、中国闪存市场、国金证券研究所

Flash Wafer 市场同样疲软，主要原因在于 PC 市场需求低迷。部分厂商在低需求的拖累下，库存水平依然偏高且库存成本较高。为进一步加速库存流通，厂商可能通过降价促销的方式来消化库存，尽管这可能带来盈利压力。尽管上半年多数存储厂商收益表现良好，在一定程度上缓解了下半年业绩压力。然而，尽管现货资源价格持续下跌，整体市场成交动力依然不足，备货情绪普遍较为观望。

图表18：Flash Wafer 月度涨跌幅

料号	May-24	Jun-24	Jul-24	Aug-24	Sep-24	Oct-24
MLC 128Gb	-0.16%	-0.47%	-0.63%	2.22%	5.12%	
MLC 256Gb	-0.61%	-0.26%	-0.44%	0.00%	-1.50%	
MLC 32Gb	0.00%	0.48%	0.00%	-1.92%	5.88%	
MLC 64Gb	0.00%	0.00%	0.00%	0.72%	-0.96%	
QLC 1Tb	-1.39%	-1.41%	-2.86%	-8.82%	-6.45%	-10.34%
SLC 16Gb	0.40%	0.00%	0.00%	-0.94%	6.37%	
SLC 1Gb	1.27%	-1.88%	0.00%	0.00%	1.91%	
SLC 2Gb	0.83%	-0.82%	-0.83%	-0.83%	0.84%	
SLC 4Gb	1.08%	1.60%	-1.05%	-1.06%	3.21%	
SLC 8Gb	0.00%	-0.57%	0.00%	-0.57%	3.45%	
TLC 128Gb					12.26%	
TLC 1Tb	-1.94%	-1.32%	-1.33%	-4.05%	-5.63%	-7.46%
TLC 256Gb	-7.14%	-2.56%	0.00%	-15.79%	-9.06%	-3.78%
TLC 512Gb	-2.38%	-2.44%	0.00%	-5.00%	-16.97%	4.60%

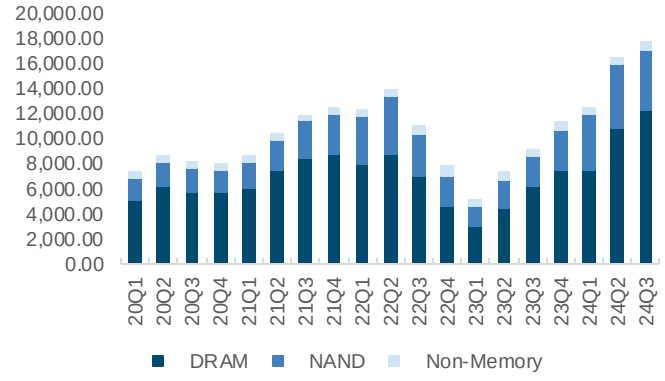
来源：DramExchange、数字未来实验室、中国闪存市场、国金证券研究所

海力士季报三季报亮眼，HBM 强劲需求带来结构性成长机遇

根据 SK 海力士最新发布的 2024 年第三季度财报（截至 9 月 30 日），公司实现营收 17.57 万亿韩元（约合 127.2 亿美元），环比增长 7%，同比大幅增长 94%；营业利润达 7.03 万亿韩元（约合 50.9 亿美元），环比增长 29%，营业利润率达到 40%，较上季度提高 7 个百分点；净利润为 5.75 万亿韩元（约合 41.6 亿美元），环比增长 40%，净利润率为 33%，环比提升 8 个百分点。公司营业利润与净利润大幅超越了半导体超级繁荣期的 2018 年第三季度水平（营业利润为 6.4724 万亿韩元，净利润为 4.6922 万亿韩元）。

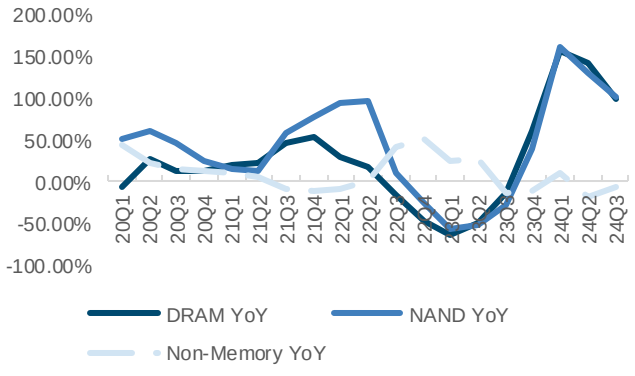
本季度业绩显著增长主要得益于 DRAM 和 NAND 平均销售价格（ASP）的环比上升，以及高盈利、高附加值产品的销售增长。SK 海力士表示，HBM（高带宽存储）和 eSSD 等应用于 AI 的存储产品需求表现强劲。其中，HBM 销售额环比增幅超过 70%，占 DRAM 总销售额的 30%，公司预计第四季度这一比重将进一步提高至 40%。eSSD 销售额则环比增长约 20%。同时，DRAM 和 NAND 产品的整体盈利能力增强，带动收入环比增长 7%。高端产品销售的扩大推动了第三季度经营利润率提升至 40%。

图表19：SK Hynix 季度部门营收（十亿韩元）



来源：Bloomberg、国金证券研究所

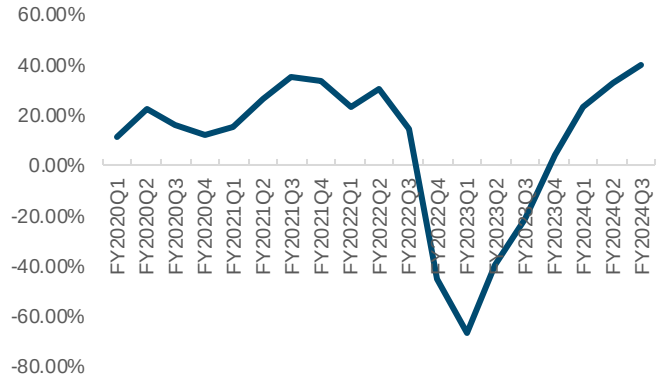
图表20：SK 海力士 DRAM 和 NAND 保持高同比增速



来源：Bloomberg、国金证券研究所

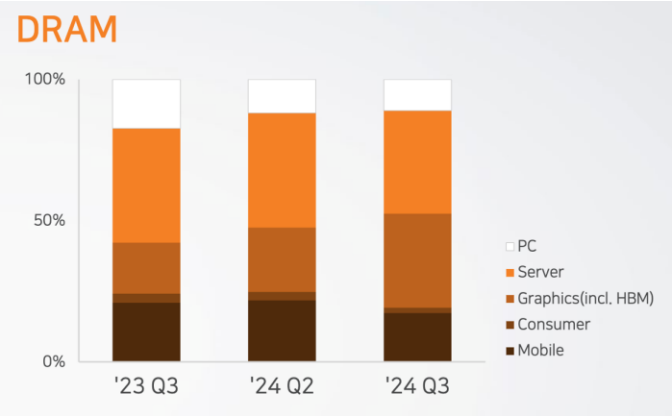
展望第四季度，SK 海力士计划通过增加 HBM 和服务器 DRAM 的销量，推动 DRAM 位出货量环比提升约 5%。在 NAND 业务方面，公司计划加大 eSSD 产品的销售，预计四季度 NAND 位出货量环比将增长 10-13%（包含 Solidigm 业务）。此外，在财报会议上，海力士指出，HBM 市场的供给将继续大于需求。公司判断，2025 年 GPGPU 市场需求仍将保持强劲，这将进一步推动 HBM 技术发展，优化 HBM 市场竞争格局，有望为当前核心的 HBM3E 供应商 SK 海力士和美光科技带来利好。

图表21：SK 海力士运营利润率提升至 40%



来源：Bloomberg、国金证券研究所

图表22：SK 海力士 DRAM 营收中 HBM 占比持续提升



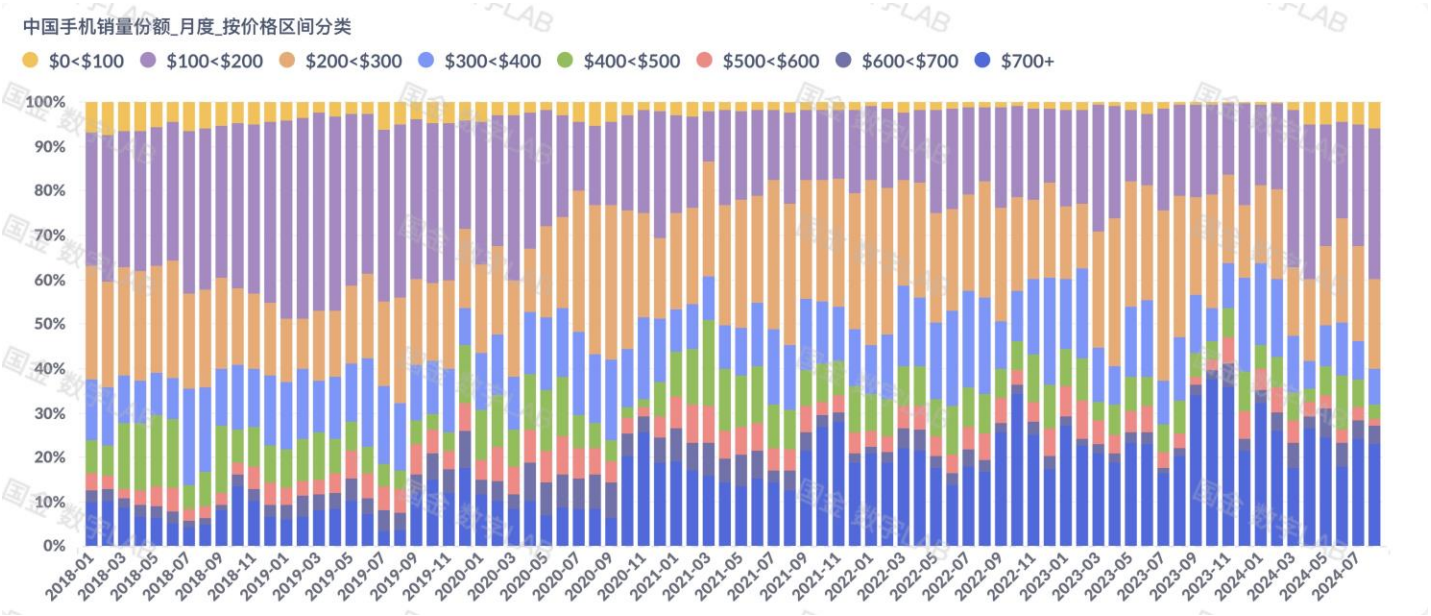
来源：SK 海力士、国金证券研究所

智能手机市场动态

各厂商手机发布季，性能&价格的提升并没有减少消费者的热情

10 月，主要安卓系厂商已经或者即将发布自己的旗舰机型。目前来看，在手机价格上涨的情况下，消费者对于高端智能手机的热情没有降低。根据 vivo 官方公告，X200 系列手机全渠道销售金额已经突破了 20 亿元，这一数据打破了 vivo 历史上所有新机销售记录，显示了消费者对这一系列新品的热烈欢迎。截至 2024 年 10 月 19 日，X200 系列的销量估计在 29.4 万到 46.5 万台之间（按最低价与最高价估计）。我们认为，消费者对于高端手机性能的追求以及其长换机周期的预期推动高端智能手机市场的发展。在苹果、高通、联发科今年旗舰芯片均有大提升的情况下，高端手机市场虽然竞争将会更加激烈，但整体销量将会进一步提升。

图表23：中国月度手机销量份额



来源：IDC，国金证券研究所

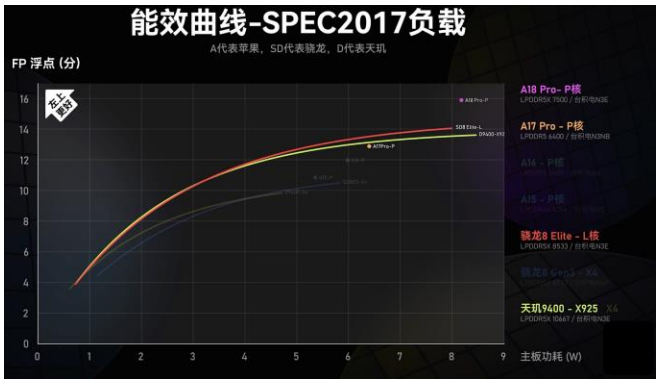
目前 vivo、荣耀、OPPO、小米、荣耀等厂商均选用高通或联发科的旗舰芯片作为首选。而今年芯片厂商的角力也将会影响未来智能手机市场的格局。以往，高端手机除了苹果外，高通旗舰芯片几乎是唯一选择。但在 2023 年联发科使用天玑 9300 的 ARM 公版 4 超大核+4 大核的“加量策略”以来，凭借优异的性能，联发科杀入高端智能手 SoC 市场成为有力的竞争者。我们认为，安卓系手机厂商在有了更多的旗舰 SoC 选择后对于高通的依赖性将

会有所降低，更有效的策略会促使厂商研发更能满足消费者的产品，使得高端手机市场保持稳定增长。

图表24：手机旗舰 SoC CPU 能耗曲线（整数）

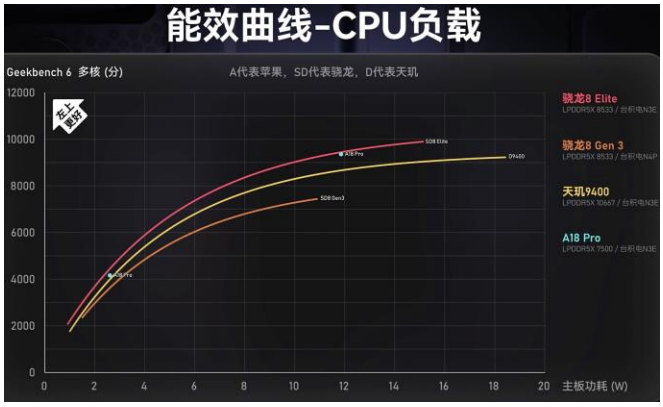
来源：极客湾，国金证券研究所

图表25：手机旗舰 SoC CPU 能耗曲线（浮点）



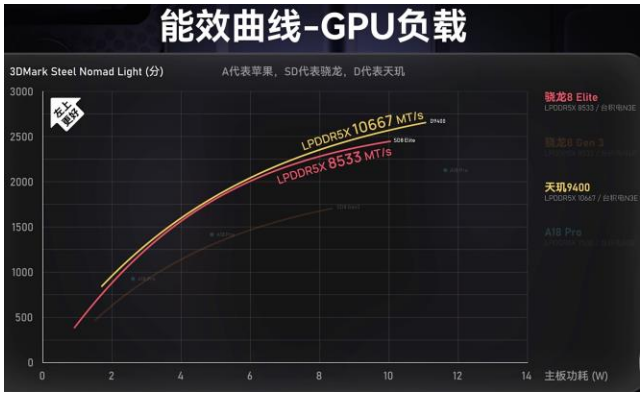
来源：极客湾，国金证券研究所

图表26：手机旗舰 SoC CPU 负载能耗曲线



来源：极客湾，国金证券研究所

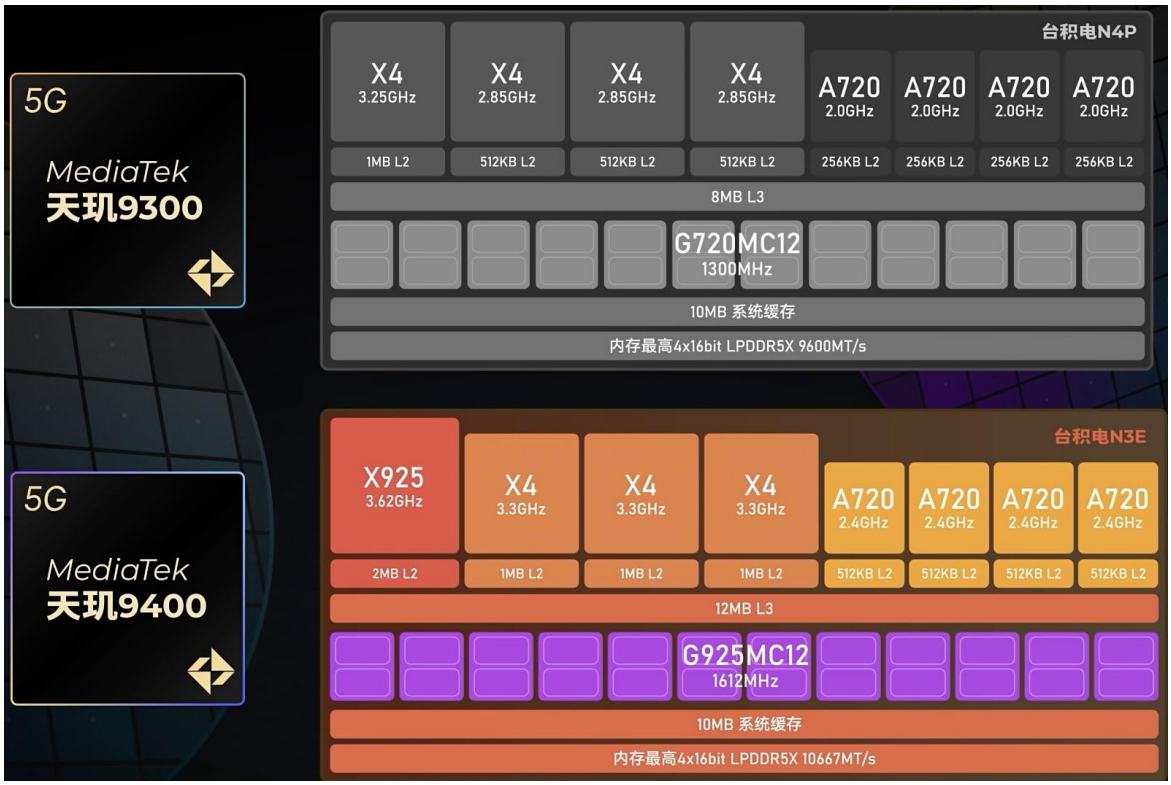
图表27：手机旗舰 SoC GPU 负载能效曲线



来源：极客湾，国金证券研究所

天玑 9400 采用台积电 N3E 工艺，这也是安卓平台首颗 3nm 芯片，晶体管规模来到了 291 亿。相比于天玑 9300 4 颗 ARM Cortex-X4 超大核，天玑 9400 超大核升级到 ARM 最新的 Cortex-X925 + 3 颗 Cortex-X4。天玑 9400 并没有一味提高主频，而是大幅增加关键的核心缓存容量。L2 缓存直接增加 100%，超大核 L2 堆到了 2MB。而 L3 缓存也提升 50%，达到了 12MB。另外 SVE2 指令集的支持，也是天玑 9400 提升性能同时降低功耗的关键。它增加了多条针对计算机视觉、5G、多媒体加速的指令，这使得天玑 9400 在更多运算中得以提高运行效率。GPU 领域，天玑 9400 升级到 12 核的 Immortalis-G925 GPU，峰值性能相较上一代提升 41%，功耗节省 44%。

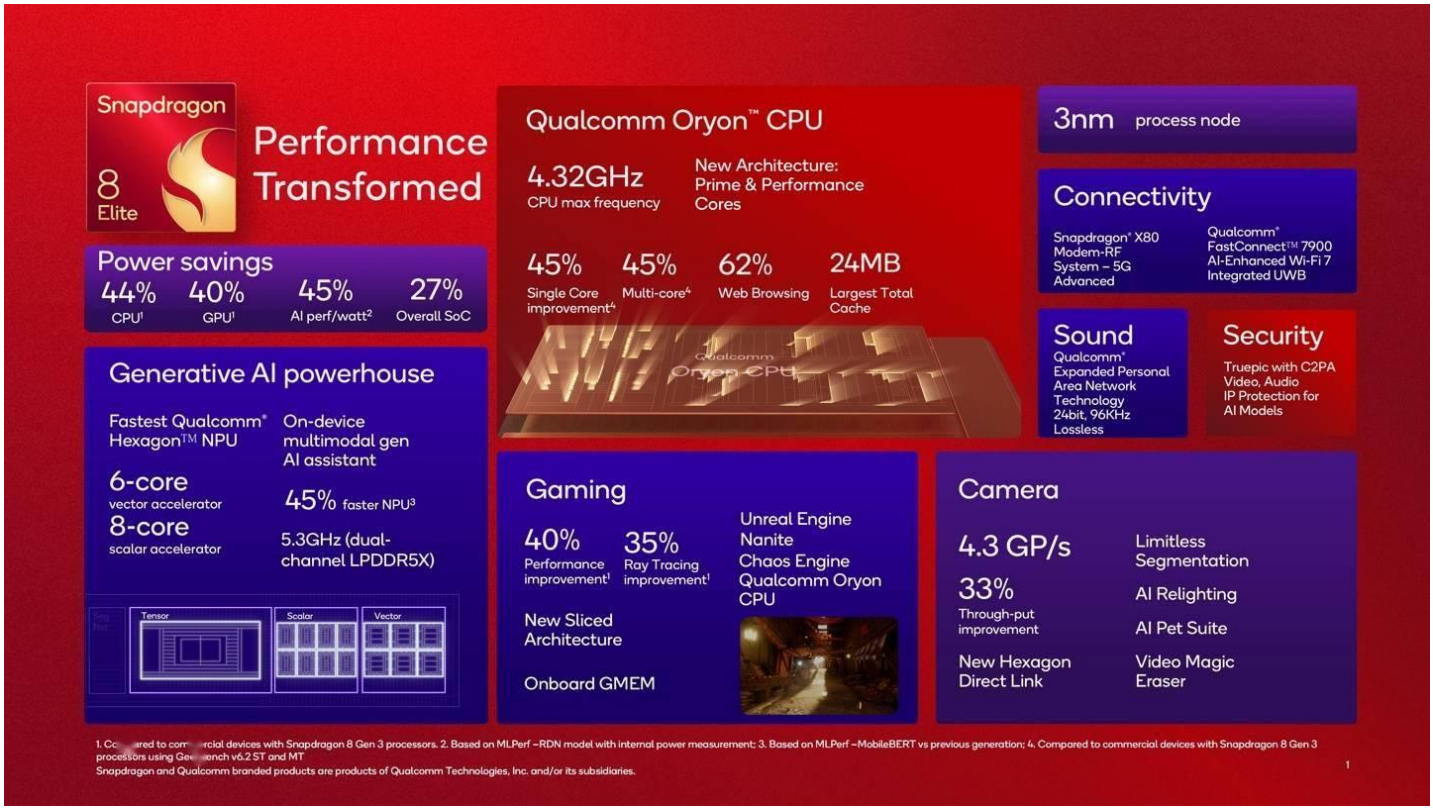
图表28：联发科天玑 9400 架构



来源：联发科官网，极客湾，国金证券研究所

高通 8 Elite 基于台积电 N3E 工艺，采用“2+6”设计，拥有 2 颗 4.32GHz Prime 超级内核和 6 颗 3.53GHz Performance 性能内核，同时配备 12MB 的 L2 缓存，并取消能效内核。此款芯片提供高达 24MB CPU 缓存，并特别为 GPU 预留 12MB 内存，以减少数据传输时对系统内存的依赖，进而降低功耗和延时。在 GPU 方面，骁龙 8 Elite Adreno GPU 采用全新的切片架构设计，频率为 1100MHz，具备 12MB 独立图形缓存。

图表29：高通 8 Elite 表现



来源：高通，sohu，国金证券研究所

在智能手机 AI 功能的开发上，芯片厂商已经做好了准备。苹果 A18 Pro 搭载了和 A17 Bionic 相同的 16 核 NPU，支持每秒高达 35TOPS 的计算能力。高通骁龙 8 Elite 采用增强的 Hexagon NPU 技术，具备 80TOPS 算力，性能提升了 45%，能效提升了 45%，支持更长的 token 输入、多模态 AI 助手的本地部署，综合 AI 性能增强达到 45%。天玑 9400 凭借全新第八代 NPU 890，不仅 AI 跑分再夺得苏黎世理工学院的 AI Benchmark 测试第一，同时还首发带来了天玑 AI 智能体化引擎，端侧视频生成及端侧 LoRA 训练，全面提升了端侧 AI 的体验。我以目前手机端侧算力需求来说，目前的旗舰 SoC 都可满足。AI 在手机端的爆发更多的在软件端。硬件、软件、系统、生态等良好适配在一起的环境更容易孕育出 AI 爆款。我们认为苹果凭借最好的软硬件结合将更有机会做到。

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. 智能手机销量不及预期：智能手机销量与产品本身质量关系紧密，若产品本身有缺陷则智能手机销量可能收到影响。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智能手机销量。

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。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。

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