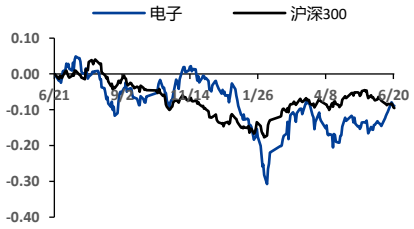
电子

端侧 AI 风起扬帆，软硬生态革新交相辉映

强于大市（维持评级）

一年内行业相对大盘走势



团队成员

分析师： **陈海进(S0210524060003)**
chj30590@hfzq.com.cn
联系人： **陈妙杨(S0210124040080)**
cmy30509@hfzq.com.cn

相关报告

1、鸿海获得英伟达 GB200 NVLink Switch 独家大单，800G 光模块需求再超预期-算力系列跟——2024.06.20
2、上游晶圆厂报价看涨再传利好，半导体超跌板块走出反弹行情-半导体系列跟踪——2024.06.16

投资要点：

➢ **端侧是 AI 场景落地关键，混合 AI 架构是未来，手机/PC 有望率先落地，后续将延伸到 XR/耳机等更多硬件。**

ChatGPT 带来的生成式 AI 热潮席卷全球，如何结合硬件实现应用领域拓展成为下一个课题，其又分为端侧场景云端计算以及本地运行两种路线。我们认为端侧运行 AI 模型，具有成本、可用性、隐私安全等优势，云端可弥补端侧设备算力的不足。从历史看，此前 AI 任务在端侧运算能力提升后，也经历了任务重心向终端转移的过程。由于高价值用户场景较易探索，手机和 PC 有望成为端侧 AI 率先落地的载体，后续 AI 在 XR/耳机/智能汽车/机器人等领域也有很大的应用潜力。

➢ **行业欣欣向荣，巨头入局引领生态发展**

品牌端：大力推进端侧 AI，单个设备通过 MoE 架构搭载大量模型以完成复杂任务，如微软的 Surface Pro 就深度整合了 40 多个本地大模型和 GPT-4o 等云端大模型，共同构造系统级 AI 能力。当前手机主流参数量在 7B，未来将持续增长并不断拓展可处理任务类型。**系统厂商：**微软/谷歌/苹果行业号召力巨大，是 AI 生态的关键玩家，其在各自系统中更新了一系列 AI 功能供开发者使用。**芯片厂商：**各厂商新发布的产品均在 AI 性能上大幅提升，布局激进，如苹果罕见的在 M3 推出半年后就推出了 M4 芯片，NPU 算力从 18 TOPS 飙升到 38 TOPS。英特尔下一代芯片 Lunar Lake 总算力将超过 100 TOPS，比 Meteor Lake 高出 3 倍，其中 NPU 算力将超过 40 TOPS，GPU 具备超过 60 TOPS 的算力。

➢ **AI 处理能力成新赛道，软硬件革新进化**

1、独立 NPU 成大势所趋：XPU 协作处理不同类型 AI 任务，其中 NPU 是低功耗长续航的高效能算力底座，是执行泛在型 AI 负载的最佳处理器，同时 NPU 将在更多终端消费电子上得到应用，1TOPS 或成是否增加 NPU 单元的分水岭；**2、ARM 搅动 PC 处理器市场：**ARM 架构功耗优势明显，尤其适用于 AI 场景。苹果推出 M 系列芯片后，苹果 PC 市占率从 2018/19 年的约 7.1%快速提升至 22 年的 9.5%。ARM 架构在取得制程优势后，依靠转译的生态过渡方式被苹果走通，微软快速跟进。微软力挺高通进军 PC 处理器市场，不仅硬件上 Copilot+PC 首发搭载高通 X 系列芯片，还推出全新转译工具，转译后在相同的 ARM 硬件上应用运行速度将提高 10%-20%；**3、整机设计或变更：**对内存/带宽的需求增长呼唤近存计算，通过更高程度的集成以减少功耗和延迟，或进一步加速 ARM PC 渗透和散热需求提升，PC 主板设计预计迎来大规模革新；**4、端侧 AI 带领存储需求快速增加：**AI 对单个设备存储消耗巨大。7B 参数模型需要占用 4GB 内存还对内存带宽提出了更高的要求，模型搭载和运行都将需要大量硬盘空间。DRAM 和 NAND Flash 下游中手机+PC 占比均超过 50%，端侧 AI 将深刻改变存储市场的中长期需求增速和供需格局。

➢ **投资建议**

1、硬件 AI 化有望带来换机潮，有望率先在手机/PC 落地，建议关注：手机产业链——立讯精密、东山精密、鹏鼎控股、长盈精密、领益智造、蓝思科技等；AI PC 产业链——华勤技术、光大同创、春秋电子、隆扬电子、中石科技、思泉新材等。

2、XR/耳机/音响等其它硬件载体：AI 带来人机交互的变化在新硬件形式上潜力巨大。

建议关注：智能音箱——国光电器，漫步者，恒玄科技，炬芯科技等；XR——歌尔股份、水晶光电、兆威机电、天键股份等。

3、存储：AI 大模型搭载和运行都将占用大量存储，同时存储的封装形式有望迎来变革。建议关注：澜起科技、聚辰股份、兆易创新、江波龙等。

4、散热：AI 负载下功率增加、近存计算增加散热挑战，单价提升叠加换机周期。建议关注：中石科技、思泉新材、安洁科技、飞荣达等。

- **风险提示**

技术发展不及预期、场景落地不及预期、市场竞争

正文目录

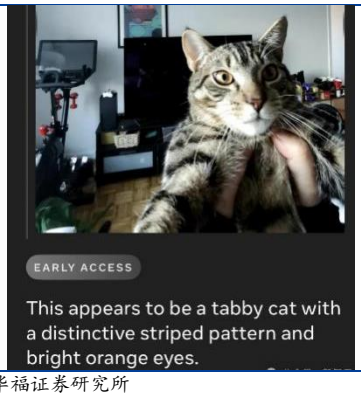
图表目录

1、 端侧 AI 是混合 AI 生态成熟的关键

1.1 不止于云端，端侧成 AI 场景落地关键

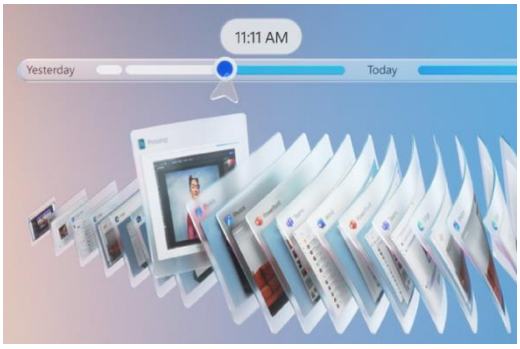
端侧场景是 AI 应用拓展领域的关键，模型根据算力部署可分为两类落地路线。ChatGPT 带来的生成式 AI 热潮席卷全球，但是当前更多还是对话框形式的应用类交互为主。如何结合终端硬件实现应用领域的拓展成为 AI 发展的下一个课题。我们认为，根据设备端仅提供感知和输入（算力全部署在云端）与设备端提供 AI 算力（模型运行在端侧）可分为两类路线。第一条路线，如 Meta 的 AR 眼镜支持多模态 Llama 实现实时互动，微软的 Copilot 也搭载 OpenAI 的 GPT-4o 实现读懂屏幕等功能。第二条路线，如微软的 Recall 功能就全部在端侧处理，不上云。

图表1：Meta 的雷朋眼镜通过 Llama 生成物体描述



来源：36 氪，华福证券研究所

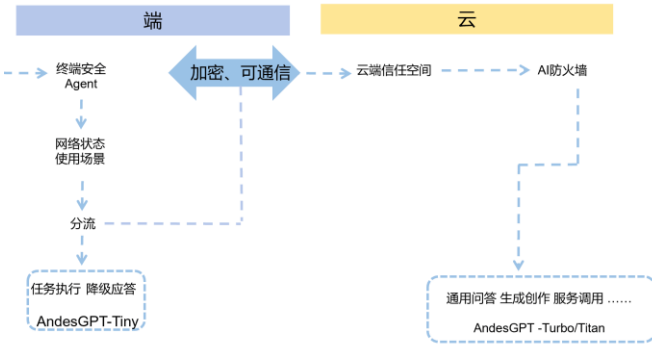
图表2：微软 Recall 功能完全运行在本地



来源：微软，华福证券研究所

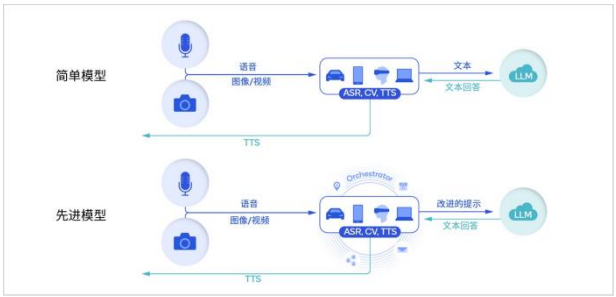
端侧场景下，混合 AI 是 AI 的未来，云端用于弥补端侧设备处理能力的不足。在云端和边缘终端之间分配并协调 AI 工作负载形成混合 AI，支持大模型的本地部署，或者通过端云协同的方式执行复杂的生成式 AI 任务。例如，如果模型大小、提示和生成长度小于某个限定值，推理即可完全在终端侧进行。如果是更复杂的任务，模型则可以跨云端和终端运行；混合方式也分为多种，在以终端为中心的混合 AI 架构中，云端仅用于分流处理终端无法充分执行的任务。在基于终端感知的混合 AI 场景中，在边缘侧运行的模型还将充当云端大语言模型（类似大脑）的传感器输入端（类似眼睛和耳朵）。

图表3：端侧和云端 AI 处理的分工示意



来源：Oppo，华福证券研究所绘制

图表4：基于终端感知的混合 AI 架构



来源：高通，华福证券研究所

1.2 端侧运行优势明显，AI 负载正在向端侧转移

端侧运行 AI 模型，具有成本、可用性、隐私安全、性能和个性化等优势。

1、成本：推理规模远高于训练，在云端进行推理的成本极高。随着用户数量的增加，硬件、场地、能耗、运营、额外带宽和网络传输成本将持续增加。比如生成式 AI 搜索查询（query）的成本是传统搜索方法的十倍。将一些处理从云端转移到端侧，充分利用 10 亿级的终端碎片化算力，可以支持开发者基于终端算力开发应用程序。仅仅手机端，Counterpoint 认为 **2027 年生成式 AI 手机计算资源将超过 50000EOPS，对应算力超过 1200 万张 H100**，经济效益巨大。

2、低延时、减少对网络的依赖：云端运行对网络要求较高，端侧运行甚至在无连接的情况下，依然可以运行 AI 应用，将时延控制在秒级甚至毫秒级别。

3、隐私和安全：因为查询信息完全保留在终端上，终端侧 AI 有助于保护用户隐私。同时还可以结合芯片公司提供的基于底层硬件的防护机制，最大程度保护用户数据和隐私安全。

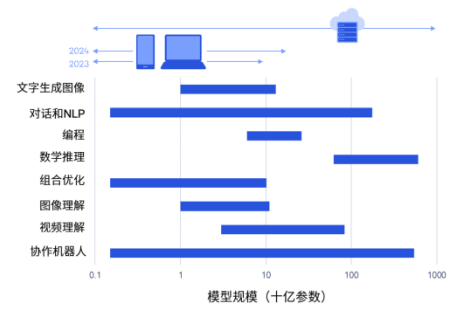
4、性能：当 AI 需求达到高峰期的时候，会产生大量排队和高时延，甚至出现拒绝服务的情况。算力达到要求的情况下，端侧运行模型时无需排队时延较低。

5、个性化：具有自学习能力的本地大模型可以成长为每个用户专属的智能体，从而有能力为用户提供个性化的服务。

6、跨应用功能：App/网页形式的云端 AI 入口，无法调用自身以外的其他应用功能，而端侧部署的模型可以支持终端相关功能，涵盖系统应用等全部场景。

基于端侧运行的优势，随着终端 AI 算力的提升，能运行的模型更加多样，生成式 AI 负载将逐渐向终端转移，此前其它 AI 任务也经历了向终端转移的过程。

图表5：端侧能运行的生成式 AI 模型将持续增加



来源：高通，华福证券研究所

图表6：AI 处理任务重心向终端侧转移

来源：高通，华福证券研究所

1.3 手机/PC 有望率先落地，后续将延伸到 XR/耳机等更多硬件

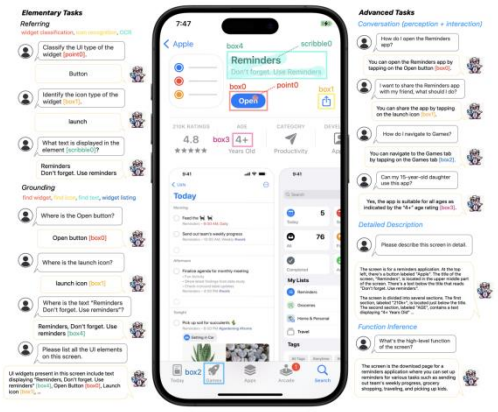
高价值用户场景较易探索，方向明确，手机和 PC 有望成为端侧 AI 率先落地的载体。手机和 PC 规模巨大，软硬件生态成熟，拥有丰富的使用场景，容易找到对用户有价值的 AI 应用，微软、高通、谷歌、各家品牌厂商都在拥抱生成式 AI 趋势。如

AI 个人助理理解复杂需求，提供更聪明、个性化的服务；在系统应用中解决用户高频高感知痛点，向生态级 AI 体验演进等方向上产业一致性较高，各方正形成合力。**因此我们认为手机和 PC 的端侧 AI 创新有望率先成熟。**

图表7：AI 手机的用户价值

来源：Oppo，华福证券研究所

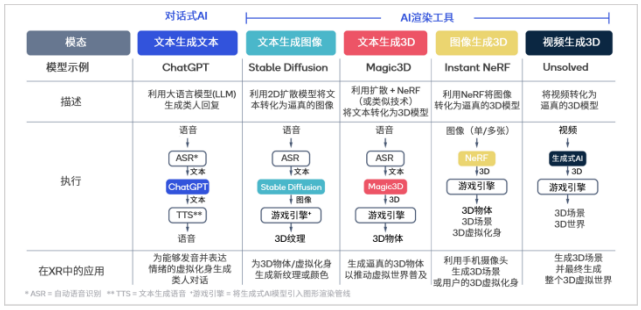
图表8：通过 AI 加强对 UI 屏幕的理解



来源：Apple，华福证券研究所

AI 在 XR/耳机/智能汽车/机器人等领域应用潜力更大，AI 后续将延伸到更多终端。XR、智能汽车等软硬件生态成熟度相对较低，根据生成式 AI 进行改进的空间也更大。如 XR 设备痛点的用户操控和输入有望被 AI 自然语言交互大幅改善，基于文本或图像生成 3D 物体点云等；多模态 AI 有望基于强大的视觉理解能力在自动驾驶中发挥重要作用，而自动驾驶对实时处理的要求正是端侧 AI 部署的强项。我们认为 AI 在各类终端落地时间有先后，但是端侧搭载是大势所趋。

图表9：AI 全面赋能 XR 体验和内容创造



来源：高通，华福证券研究所

图表10：利用 AI 理解路况并预测主体行为轨迹

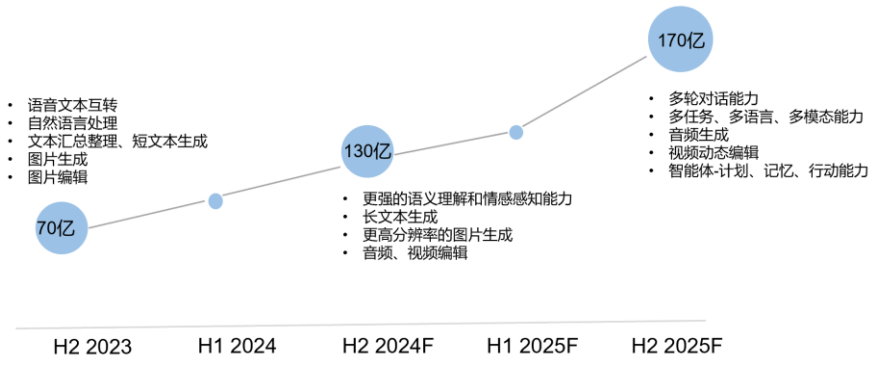
来源：高通，华福证券研究所

2、 端侧 AI 欣欣向荣，巨头入局引领生态发展

2.1 品牌大力推进端侧 AI，单个设备搭载大量模型以完成复杂任务

品牌厂商纷纷推出端侧运行的 AI 模型搭载至旗舰机型，当前参数量以 7B 为主流。整机厂商是端侧 AI 的重要推动者，一方面品牌厂商希望通过 AI 推动功能革新提升用户体验，一方面也希望提升品牌科技内涵和用户粘性。截至目前，多数头部品牌厂商已经发布 AI 大模型，如三星的 Galaxy AI、Vivo 的蓝心大模型等。目前端侧落地的参数集中在 7B 级别。随着每年下半年手机核心处理器的更新，预计本地算力能支持的模型参数将逐渐增加，本地能够处理的 AI 任务也更加多样和复杂。

图表11：本地模型参数预计持续增长并不断拓展可处理任务类型



来源：Counterpoint，华福证券研究所

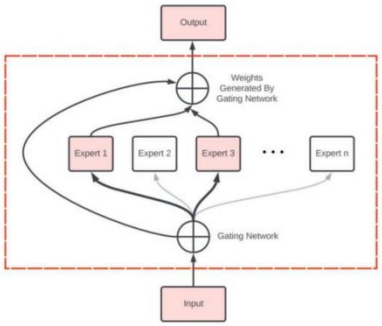
MoE 大模型架构能提升端侧 AI 响应速度，单终端有望集成数十个模型。MoE 是一种神经网络架构，可以集成进 Transformer 的结构中。MoE 增加了专家网络层，当数据流经 MoE 层时，每个输入（token）都会通过门控网络动态分配给某个专家模型进行计算，最后进行加权融合得到最终输出，更加高效和专业，处理复杂任务时性能显著变好。MoE 架构让模型拥有**更高的推理计算效率，从而让用户获得更快的 AI 响应速度**。同时模型构建更为灵活、多样、可扩展。**微软的 Surface Pro 就深度整合了 40 多个本地大模型和 GPT-4o 等云端大模型，共同构造系统级 AI 能力。**

图表12：整机环节的生成式 AI 进展

厂商	事件
Vivo	2023年11月，推出蓝心大模型（BlueLM），1B和7B支持高通和联发科双平台，面向端侧；70B、130B和175B面向云端和复杂逻辑推理场景；7B版本2.6T语料、支持32K上下文
Oppo	2023年11月，推出安第斯大模型（AndesGPT），分为Tiny、Turbo和Titan版本。其中Tiny版本为面向端侧的7B参数模型，1024tokens输入首字响应2.1s，文生图单张图片耗时5.5s，图生图6.6s
小米	小米大模型MiLM 23年推出，24年5月通过备案，有1.3B和6.4B版本，后续将逐步应用于汽车、手机、智能家居产品，轻量化和本地部署，通过端云结合实现设备和场景之间的互联
荣耀	2024年1月推出"魔法大模型"，7B参数，端侧运行，云端与百度千帆协作，24Q2与高德地图和航旅纵横等合作，发力对接垂类大模型，实现个人化服务
三星	2024年1月推出Galaxy AI大模型首发搭载于S24系列手机，通话实时翻译、即圈即搜、写作助手、笔记助手等诸多功能致力于融入用户生活
联想	2024年4月发布个人智能体"联想小天"，具备个人知识库，能在与用户的交互中不断总结经验并自我完善，熟悉用户的习惯和偏好。核心应用包括包括 AI 画师、AI PPT、文档总结、知识问答、AI 识图、会议纪要等，且还在持续增加
微软	2024年5月，推出第一款Copilot+PC——Surface pro，搭载高通ARM架构处理器，45TOPS NPU算力，搭载GPT-4o，一个PC用40多个本地大模型

来源：Vivo、三星、Oppo 等，华福证券研究所

图表13：MoE 架构示意图



来源：36 氪，华福证券研究所

2.2 系统厂商是生态核心推动者，微软/谷歌/苹果轮番上场

系统环节是 AI 生态繁荣的关键。系统厂商生态号召力巨大，在整体的生态中不

仅需要在系统层面深度整合 AI 功能，在第一方工具和应用中搭载 AI 功能，还需要为开发者降低开发 AI 应用的成本，提供 AI 系统服务管理模型运行和安全问题等。

除了云端 AI，系统厂商在端侧 AI 进展不断。微软在 Windows 操作系统中大力推广 Copilot，并将其定位从个人助手不断拓展至团队成员等。Copilot+PC 的重要组成部分就是众多端侧运行的模型，Recall 等 AI 功能在端侧运行将会搭载至 Windows 阵营所有 AI PC 中，ARM to x86 转译也会大幅提高 ARM 架构的优势。手机端，谷歌和苹果在 AI 技术布局、系统整合上都持续加码，Android 15 和 iOS 18 等系统将更新一系列生成式 AI 功能。

图表14：系统厂商在端侧 AI 的进展

厂商	事件	事件概述
微软	2023年9月	Windows 11引入AI助手Copilot，定位日常人工智能伴侣，引入到常用的微软产品，包括GitHub编程工具，Microsoft 365工具箱、Bing搜索引擎、Edge浏览器和Windows操作系统，并将个人隐私和数据安全放在首位
	2024年1月	推出实体Copilot键，30年来首次改变键盘布局，更加轻松使用AI
	2024年1月	将Copilot的全部功能开放给更多个人和企业
	2024年5月	1、Copilot+PC新增Recall功能，全部端运行；2、定义NPU 40TOPS的算力是AIPC算力及格线；3、推出全新的ARM to x86兼容层Prism，转译效率提高10-20%；4、推出Team Copilot，功能从个人助理延伸至团队成员
谷歌	2023年10月	安卓14系统正式发布，新增一项新系统服务Android AICore供开发者使用，可处理模型管理、运行、安全功能等，简化用户将AI融入应用程序的工作
	2023年12月	推出Gemini 1.0模型，融入谷歌更多产品和服务中，分为Ultra、Pro和Nano版本。其中Nano版本为用于设备端的高效模型
	2024年5月	I/O大会宣布，Android 15将深度整合其Gemini大模型，为用户带来一系列AI功能。例如即圈即搜增加截图与题目解答的智能化处理，自动总结文件和视频内容。
苹果	2023年12月	1、向开发者推出MLX机器学习框架，为M系列芯片优化AI处理能力；2、发布LLM in a Flash技术，利用闪存解决内存不足的问题。
	2023年3月&4月	1、3月和4月分别发布ReLAM和Ferret-UI模型，共同提出全新的UI界面的AI人机交互方案；2、发布OpenELM系列小模型，提出分层缩放策略提高准确率，并降低50%预训练Token数，大幅降低开发者开发AI应用成本亿构建更繁荣的AI生态。
	2024年6月	6月WWDC大会，iOS、iPadOS、macOS等系统将更新，据彭博社预计，苹果各类系统将更新一系列生成式AI应用

来源：IT之家，谷歌、微软、苹果等，华福证券研究所

2.3 芯片环节：端侧 AI 的算力底座，发力总体 AI 算力提升

芯片厂商大力发展 AI 算力，综合 AI 算力提升明显。无论是手机还是 PC 处理器，各厂商新发布的产品均在 AI 性能上大幅提升。

1、从架构上，英特尔首次引入 NPU，高通将过去的 Hexagon 张量加速器升级为 Hexagon NPU，X Elite 上首次推出自研 Oryon 架构的 CPU，苹果也是罕见的在 M3 推出半年后就推出了 M4 芯片，NPU 算力从 18 TOPS 飙升到 38 TOPS。

2、从最终的产品算力看：手机 NPU 算力集中在 30+ TOPS，高通整体 AI 算力超过 70 TOPS；PC 端骁龙 X Elite 优势明显，NPU 算力 45 TOPS，总算力超过 75 TOPS。

3、从后续产品布局看：英特尔 2024 年 H2 上市的 Lunar Lake 芯片总算力将达到

120 TOPS，比 2023 年发布 Meteor Lake 高出 3 倍，其中 NPU 算力达到 48 TOPS，GPU 具备超过 60 TOPS 的算力。AMD 的 AI 300 处理器 NPU 算力也是 2023 年锐龙 8040 的 3 倍多。当前苹果 M4 Pro 和 Max 款尚未发布，NPU 提升尚不明确，但从 M3 系列看，Pro 和 Max 的 GPU 将会带来总算力的大幅增加。

图表15：芯片厂商在端侧 AI 的进展

芯片类型	厂商	芯片名称	发布时间	算力参数	其它
手机芯片	高通	骁龙8Gen3	2023年10月	AI算力超过73TOPS，其中NPU算力34TOPS	首款为生成式AI设计的芯片，支持100亿参数模型，支持端侧多模态，Llama2-7B每秒20tokens
	联发科	天玑 9300	2023年11月	NPU算力33TOPS	支持130亿参数模型，70亿参数模型20 tokens/秒
	苹果	A17 Pro	2023年9月	NPU算力35TOPS	-
PC级芯片	高通	骁龙X ELite	2023年10月	AI算力75TOPS，其中NPU 45TOPS	12核OryonCPU，支持130亿参数模型，Llama2-7B 30tokens/秒
	英特尔	Meteor Lake	2023年12月	AI总算力34TOPS，其中NPU 11TOPS	引入NPU单元
	英特尔	Lunar Lake	2024年H2	AI总算力120TOPS，其中NPU 48 TOPS	首次使用封装级内存，采用LPDDR5X，节省40%功耗和250mm^{2}主板空间
	AMD	锐龙8040	2023年12月	AI总算力39TOPS，其中NPU 16TOPS	-
	AMD	AI 300	2024年7月	NPU 50 TOPS	-
	苹果	M3系列芯片	2023年10月	NPU算力18TOPS	M3/M3 Pro/M3 Max分别拥有10/18/40核GPU
	苹果	M4	2024年5月	NPU算力38TOPS	升级节奏超常规。10核GPU，相比M2，CPU性能提升最高1.5倍，GPU渲染性能提升最高4倍

来源：苹果、英特尔、高通、联发科、AMD 等，华福证券研究所

3、 AI 处理能力成新赛道，软硬件革新进化

3.1 独立 NPU 增强整体算力，适配泛在 AI 负载

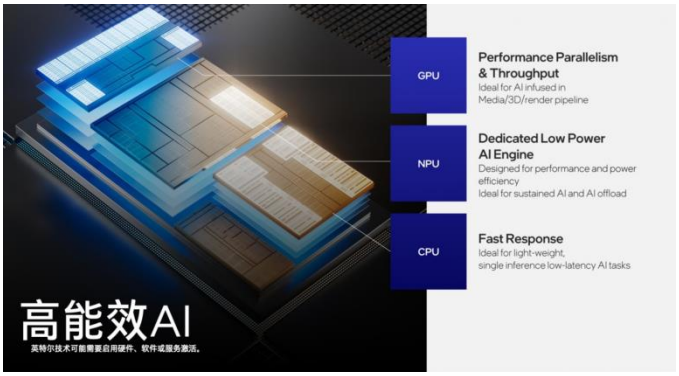
英特尔、AMD、高通纷纷推出集成 AI 加速引擎的处理器，内置 NPU 为 PC 端 SoC 架构 40 年来最大变化。端侧大模型对设备载体的算力要求较高，核心处理器的 AI 任务处理能力至关重要。2023 年 9 月的创新大会上，英特尔推出 PC 处理器 Meteor Lake，首次内置 NPU 以增强 AI 任务处理能力，英特尔称其为客户端 SoC 架构 40 年来最大的改变。AMD 在部分处理器加入了 Ryzen AI 技术，集成 AI 加速引擎。高通也在发布的 PC 和手机处理器中加大了 AI 能力的分配。

图表16：英特尔首款内置 NPU 的消费级 CPU

来源：英特尔，华福证券研究所

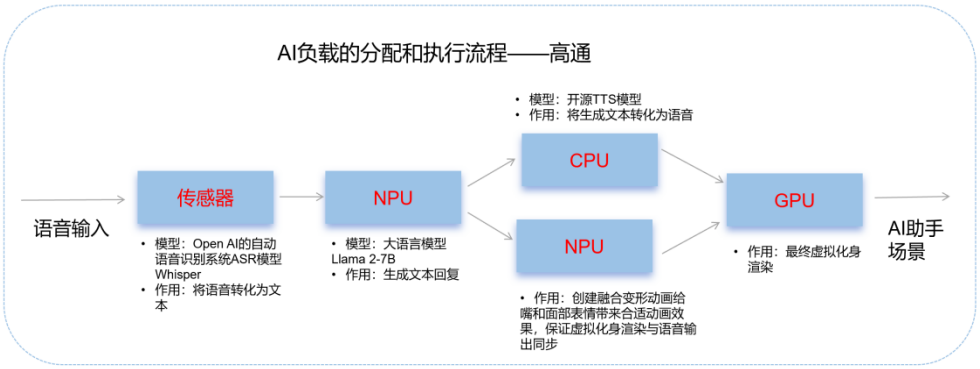
CPU/GPU/NPU 各有所长，共同处理不同类型 AI 任务和环节，综合 AI 处理能力的提升需要协同升级。CPU、GPU、NPU 都可以提供 AI 算力，但是针对不同使用场景匹配程度并不相同。每个处理器擅长不同的任务：CPU 擅长顺序控制和即时性，GPU 适合并行数据流处理，NPU 擅长标量、向量和张量数学运算。对应到 AI 用例，CPU 适合处理轻量级 AI 并实现快速响应，GPU 适合需要高吞吐量的 AI 应用。**NPU 专为 AI 而设计，是执行神经网络算法时性能、功耗和面积效率最高的处理器，尤其适用于持续性的 AI 负载，如视频会议、屏幕理解等涉及长时间处理的任务。**NPU 加入后可以降低 AI 负载对 CPU、GPU 的调用，从而让轻薄本拥有更持久的续航。在同一个任务中，不同的处理核心也将分工以更高效完成不同的环节。

图表17：英特尔对 CPU、GPU 和 NPU 在 AI 负载下的定位



来源：英特尔，华福证券研究所

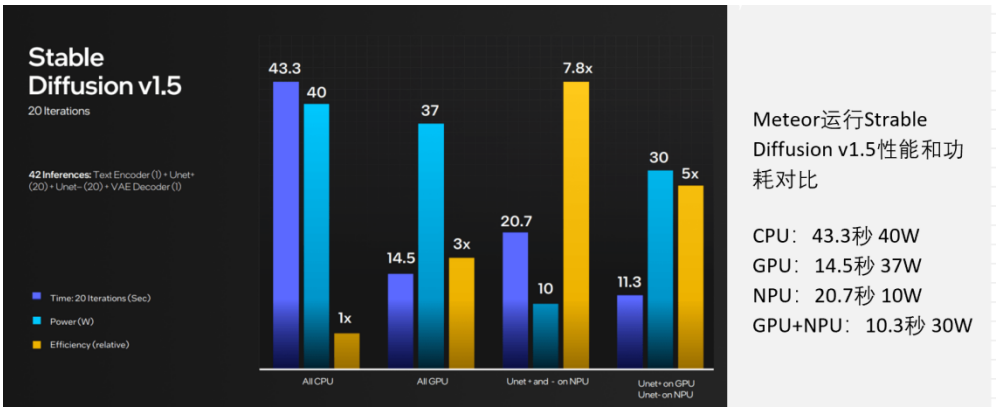
图表18：一个 AI 助手执行流程需要各类处理器分配 AI 负载



来源：高通，华福证券研究所绘制

NPU 大幅提升综合 AI 算力表现，是低功耗长续航的高效能算力底座。在 2024 年 5 月，微软认为内置 NPU 是 AIPC 的两大特性之一，同时划定 NPU 算力 40 TOPS 是“及格线”。英特尔在 23 年技术创新大会上分享了使用 Meteor Lake 运行文生图模型 Stable Diffusion 的内部测试数据。负载全部跑在 CPU 上用时为 43.3 秒，功耗 40W；全部跑在 GPU 上用时为 14.5 秒，功耗 37W；将部分负载（Unet+与 Unet-）交由 NPU 执行，其余交由 CPU 执行，用时为 20.7 秒，功耗 10W；Unet+由 GPU 执行，Unet-由 NPU 执行，用时为 11.3 秒，功耗为 30W。

图表19：灵活调配 XPU 兼顾功耗和效率



来源：英特尔，华福证券研究所

不止于 PC、手机，NPU 将在更多终端消费电子上得到应用，1TOPS 或成是否增加 NPU 单元的分水岭，产业界快速跟进。在手机和 PC 以外，耳机、音箱、XR、智能驾驶等也将提升对 AI 算力的需求，算力需求从 GOPS 到上千 TOPS。据新思，若 AI 性能需求小于 1 TOPS，CPU+DSP 可满足需求。当运算能力要求高于 1TOPS 时，NPU 的 AI 性能效率、功耗效率和面积效率（影响制造成本）优势明显。如炬芯科技在 2023 年报披露，其正在推进产品架构升级为 CPU+DSP+NPU 三核异构的 AI SoC 架构，为端侧智能音频、智能穿戴产品在低功耗前提下提供更大 AI 算力，探索 AI 驱动下的音频芯片创新。

图表20：不同应用对算力的需求

图表21：各类处理器大致性能范围

来源：新思官网，华福证券研究所

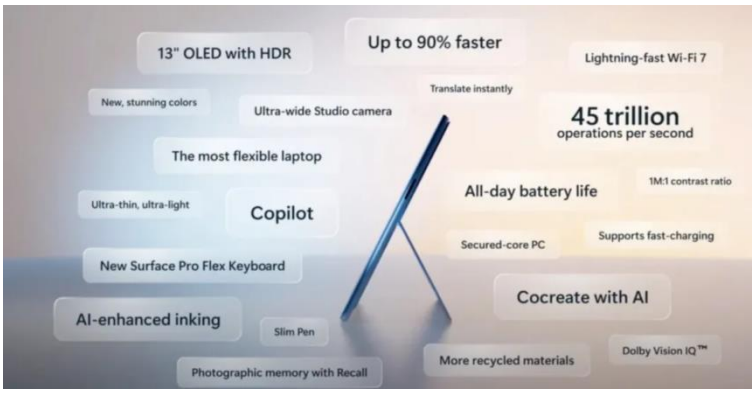
处理器类型	MAC次数/周期	最大频率	理想的TOPS
具有DSP拓展的CPU	1	2GHz	2 GOPS
适量DSP	512	1.2GHz	1.2TOPS
NPU（低端）	4019	1.3GHz	10.6 TOPS
NPU（高端）	96304	1.3GHz	255.6 TOPS

来源：新思官网，华福证券研究所
注：1TOPS 约等于 1000GOPS

3.2 ARM 架构低功耗优势明显，AI 时代优势进一步放大

微软力挺高通入局 PC 处理器市场，ARM 架构处理器有望进军 Windows 笔记本市场。微软 2024 年 5 月发布的新款 Surface Pro 用高通骁龙 X Elite 和骁龙 X Plus 替换了英特尔的酷睿 Ultra 处理器，相比上代快了 90%。而采用英特尔和 AMD 芯片的版本将会在更晚的时候发售。同时微软在软件方面也再次发力，推出全新的 ARM to x86 兼容层 Prism，转译后的应用在相同的 ARM 硬件上运行速度将提高 10%-20%。

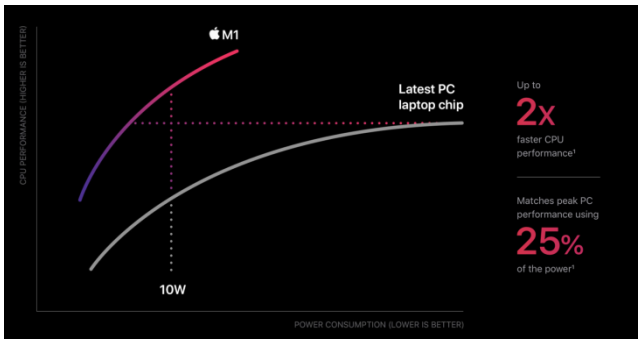
图表22：第一款 Colipot+PC——微软新款 Surface Pro 主要参数



来源：微软 Build 2024 开发者大会，华福证券研究所

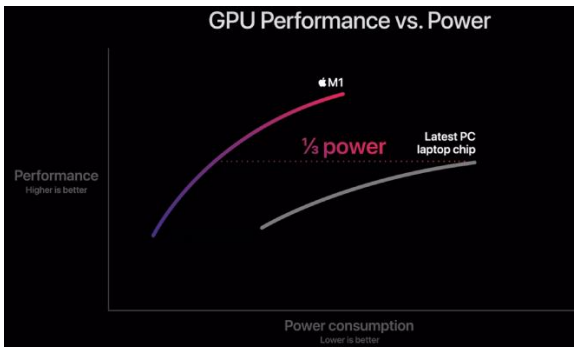
2020 年苹果在 Macbook 上推出自研 M 系列 SoC 芯片，综合性能大幅提升，尤其是低功耗优势明显。M1 芯片首次采用 5nm 制程封装了 160 亿晶体管，在处理器和内存架构上都大幅革新。M1 芯片将中央处理器速度提升至最高 3.5 倍，图形处理器速度提升至最高 6 倍，机器学习速度提升至最高 15 倍。相比性能提升，功耗降低更为惊人。对比当时市面最新的 PC 处理器，CPU 在同样性能表现下功耗仅为 25%，GPU 在同样性能表现下功耗仅为 1/3。反映到整机设备上，搭载 M1 的 Macbook Air 续航最长达 18 小时，比之前多出 6 小时。

图表23：M1 芯片 CPU 能效曲线



来源：Apple，华福证券研究所

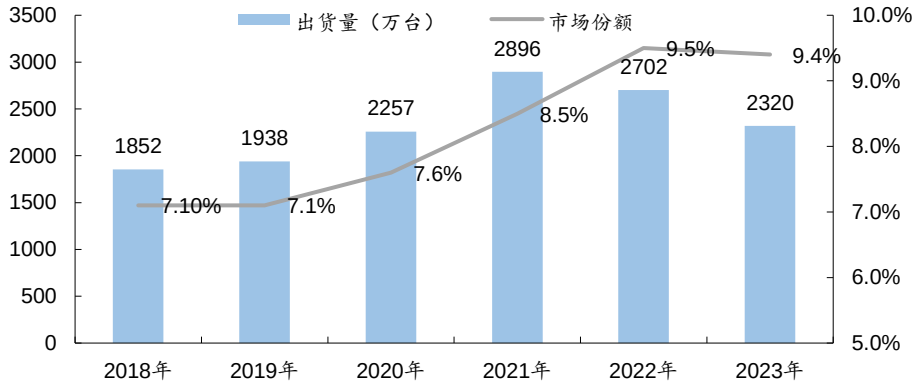
图表24：M1 芯片 GPU 能效曲线



来源：Apple，华福证券研究所

推出 M 系列芯片后，苹果 PC 市占率大幅提升，AI PC 或成为 ARM PC 增长关键节点。尽管前期 ARM 生态不如 x86 成熟，但是凭借功耗和性能优势，搭载 M 系列芯片的 Macbook 大受欢迎。苹果 PC 市占率从 2018/19 年的约 7.1%快速提升，M1 推出后仅 2 年时间就达到 9.5%。随着系统级 AI 在端侧运行带来的 AI 负载增加，处理器低功耗要求增加后，ARM 架构必要性进一步提升，可能成为 ARM PC 快速增长的一个关键节点。

图表25：推出 M 系列芯片后苹果 PC 市占率快速提升



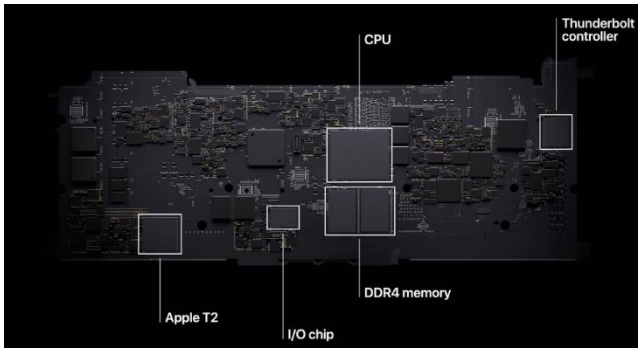
来源：Canalys，华福证券研究所

3.3 内存/带宽需求增加或带来整机设计变更，散热需求增加

近存计算将处理和存储单元以更高集成度集成，减少功耗和延迟。在冯·诺依曼架构中，数据存储和计算分离，存储器和处理器之间通过总线进行数据传输，频繁的数据传输导致大量的时间与功耗开销，称为“存储墙”和“功耗墙”。存内计算通过消除“存”与“算”的界限减少了数据搬运开销，在提升性能降低功耗上优势明显。但是受限于关键技术尚未成熟，通过先进封装达到 HBM 等的近存计算效果的方式有望更快落地。尽管电路结构上仍然是独立的实体，但是通过将数据尽可

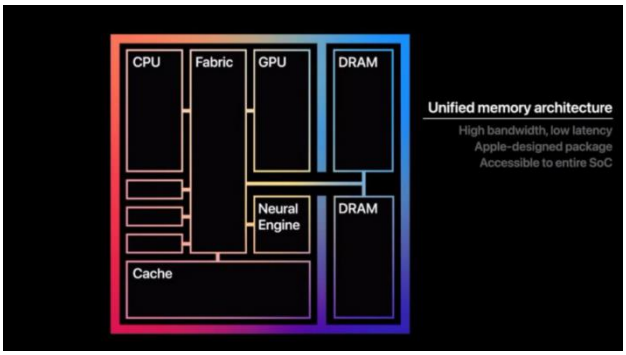
能的靠近计算单元，也可大幅减少数据移动的延迟和功耗。**如苹果 M 系列芯片就已推动存储和处理单元从板卡组装级别的集成程度向芯片封装水平提升。**

图表26：原有 Mac 上处理单元和存储在主板级集成



来源：Apple，华福证券研究所

图表27：处理和存储单元共同封装成 M1 芯片



来源：Apple，华福证券研究所

苹果 M 系列芯片大内存、高带宽优势明显。M3/M3 Pro/M3 Max 的内存分别为 24/36/128GB，作为对比，目前英伟达最高端的笔记本 GPU 产品 RTX 4090 的显存容量为 16GB。M3/M3 Pro/M3 max 的带宽分别为 100/200/400GB/s，我们选取英特尔 2023 年发布的 Meteor Lake 芯片作为对比，其最高带宽为 120GB/s。因而 M 系列芯片在传输带宽、降低延时、降低整体功耗上表现优异。

图表28：M3 系列芯片信息

	M3	M3 pro	M3 max
制程，晶体管数量	3nm，250亿	3nm，370亿	3nm，920亿
CPU	8核，4*高性能核心+4*高效能核心	12核，6*高性能核心+6*高效能核心	16核，12*高性能核心+4*高效能核心
GPU	10核，硬件加速光线追踪	18核，硬件加速光线追踪	40核，硬件加速光线追踪
NPU	16核神经网络引擎，18TOPS	16核神经网络引擎，18TOPS	16核神经网络引擎，18TOPS
支持内存	24GB	36GB	128GB
内存带宽	100GB/s	200GB/s	400GB/s

来源：Apple 官网，cpu-monkey，华福证券研究所
注：Pro 和 Max 选取可选最高规格

对内存/带宽的需求增长呼唤近存计算，或进一步加速 ARM PC 渗透和散热需求提升，AI PC 时代主板设计预计迎来大规模革新。AI 负载需要存储和处理单元大量的数据传输，同时异构计算架构需要在不同处理单元间通信。高内存和高带宽需求大幅提升，近存计算的必要性提升。

将存储和计算靠近后，整机散热区域更加集中，对散热挑战性较大。因此，我们认为除了进一步提升制程外，ARM 的低功耗为治本之策，x86 架构下对散热的需求将会大幅提升。

3.4 端侧 AI 带领存储需求快速增加

3.4.1 DRAM：容量和带宽加速提高以免成为 AI 模型性能限制短板

高规格内存对生成式 AI 硬件不可或缺，促进 DRAM 加速提高容量和带宽，7B 参数模型占用约 4GB 内存，推荐 60GB/s 以上带宽。大模型在运行时，需要驻留在内存中，每次处理生成式 AI 任务，都可能涉及到海量的数据搬运。限制 AI 模型的最短板可能是内存限制（即性能表现受限于内存带宽）或者计算限制（即性能表现受限于处理器性能）。当前的大语言模型在生成文本时受内存限制，因此需要关注 CPU GPU 或 NPU 的内存效率。以目前主流的 7B 参数模型为例，模型运行需要占用约 4GB 的内存空间，建议采用至少 8GB 的 LPDDR5x（推荐 60GB/s 以上的 I/O 带宽）。

3.4.2 NAND Flash：搭载和运行都将需要大量硬盘空间

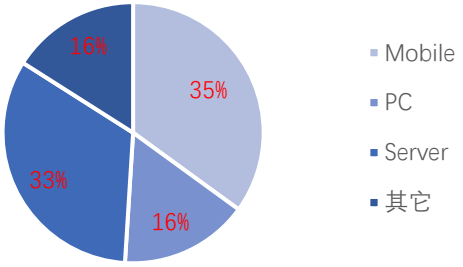
设备需要搭载多达数十个 AI 模型，将会占用大量硬盘空间。AI 大模型的能力和性能，很大程度上是由模型包含的参数数量和质量决定的。AI 手机、AI PC 等需要同时运行多个 AI 模型，包括参数量大的多模态模型，也有体量较小的影像计算和图像生成模型。如微软 Copilot+PC 就搭载了 40 多个本地大模型，将大幅占用硬盘空间。24 年初英伟达发布了 Chat with RTX，里面有两个模型，分别是 130 亿参数的 LLama2 和另一个 70 亿参数的模型。Chat with RTX 仅安装包大小就有 35G，实际安装使用后会占用更多硬盘空间。再加上一些完成细分功能的模型，仅仅搭载本地模型就会占用大量的硬盘空间。

模型运行过程中将产生大量数据，进一步占用硬盘空间。微软推出的 Recall 在屏幕内容发生变化时，5 秒截图一次存储在时间线中。而要启用 Recall，至少需要 50GB 的可用空间，一旦设备的存储空间小于 25GB，快照捕获就会自动暂停。我们认为随着 AI 功能的增加，运行过程中产生的硬盘占用也会大幅增加。

3.4.3 端侧 AI 存储需求对整体存储市场影响巨大

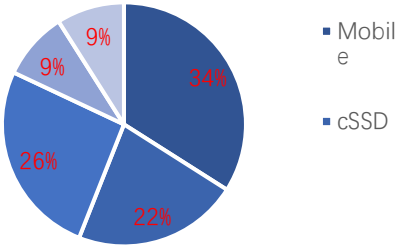
手机+PC 存储需求占比大，单机容量对存储市场有巨大拉动。DRAM 市场中，手机+PC 占比超过 50%；NAND Flash 中，手机+PC 占比约为 56%。手机和电脑上对 DRAM 和 NAND Flash 需求的提升将深刻改变存储市场的中长期需求增速和供需格局。同时服务器市场由于 AI 训练和云端推理的需求，存储需求也将快速增加。

图表29：DRAM 市场应用



来源：CFM 闪存市场，华福证券研究所

图表30：NAND Flash 市场应用



来源：CFM 闪存市场，华福证券研究所
注 1：cSSD 主要应用于 PC 市场
注 2：eSSD 主要应用于服务器市场

4、 投资建议

端侧 AI 硬件带来消费电子行业革新，当前巨头入场引领行业发展，AI 对消费电子带来了从指令集架构到芯片架构、整机功能和结构设计的全方位快速迭代，带来各个环节的投资机遇。

1、**手机/PC 整机换机：硬件 AI 化有望带来换机潮，有望率先在手机/PC 落地。**建议关注：手机产业链——立讯精密、东山精密、鹏鼎控股、长盈精密、领益智造、蓝思科技等；AI PC 产业链——华勤技术、光大同创、春秋电子、隆扬电子、中石科技、思泉新材等。

2、**XR/耳机/音响等其它硬件载体：**AI 带来人机交互的变化在新硬件形式上潜力巨大。建议关注：智能音箱——国光电器、漫步者、佳禾智能、恒玄科技、炬芯科技等；XR——歌尔股份、水晶光电、兆威机电、天键股份等。

3、存储：AI 大模型搭载和运行都将占用大量存储，同时存储的封装形式有望迎来变革。建议关注：澜起科技、聚辰股份、兆易创新、江波龙等。

4、散热：AI 负载下功率增加、近存计算增加散热挑战，单价提升叠加换机周期。建议关注：中石科技、思泉新材、安洁科技、飞荣达等。

图表31：行业重点公司（2024/6/21）

公司代码	公司名称	当前市值（亿元）	EPS(摊薄)				PE			
			2023A	2024E	2025E	2026E	2023A	2024E	2025E	2026E
002475.SZ	立讯精密	2,712	1.53	1.92	2.39	2.84	26.9	17.3	13.9	11.7
002384.SZ	东山精密	324	1.15	1.36	1.70	2.02	13.1	11.1	8.9	7.7
002938.SZ	鹏鼎控股	879	1.42	1.67	1.90	2.09	10.3	19.1	16.8	15.2
300115.SZ	长盈精密	140	0.07	0.59	0.70	0.87	374.8	18.8	15.4	12.4

002600.SZ	领益智造	399	0.29	0.34	0.45	0.52	29.7	13.9	10.4	9.0
300433.SZ	蓝思科技	878	0.61	0.81	1.02	1.24	26.9	19.6	15.4	12.4
603296.SH	华勤技术	580	3.74	4.08	4.71	5.60	22.6	20.2	17.4	15.0
301387.SZ	光大同创	36	1.51	2.57	3.33	2.69	45.2	17.5	12.7	13.9
603890.SH	春秋电子	39	0.06	0.39	0.65	0.85	32.0	23.6	14.8	11.1
301389.SZ	隆扬电子	45	0.34	0.71	0.97	1.07	32.0	22.4	18.5	14.2
002045.SZ	国光电器	72	0.63	0.60	0.74	0.88	43.7	19.3	15.5	13.1
002351.SZ	漫步者	113	0.47	0.62	0.76	0.86	63.7	20.0	16.5	14.5
300793.SZ	佳禾智能	48	0.39	0.52	0.61	0.72	42.3	24.4	20.7	17.5
688608.SH	恒玄科技	171	1.03	2.35	3.58	4.78	151.3	55.6	36.5	27.4
688049.SH	炬芯科技	34	0.53	0.73	0.97	1.25	86.4	33.8	25.4	19.8
002241.SZ	歌尔股份	633	0.32	0.72	0.84	1.05	41.1	24.1	20.7	16.6
002273.SZ	水晶光电	231	0.43	0.58	0.71	0.85	32.7	27.6	22.4	18.9
003021.SZ	兆威机电	118	1.05	1.39	1.72	2.14	106.8	49.8	38.8	31.5
301383.SZ	天键股份	43	1.17	1.27	1.59	2.09	113.0	22.1	16.7	12.0
300684.SZ	中石科技	52	0.25	0.43	0.68	0.96	33.3	38.8	24.4	17.2
301489.SZ	思泉新材	43	0.95	1.72	2.25	-	74.7	41.6	31.8	-
002635.SZ	安洁科技	101	0.46	0.57	0.78	-	46.7	25.2	18.3	-
300602.SZ	飞荣达	86	0.18	0.67	0.84	0.90	108.6	21.7	17.2	16.1

数据来源：iFind，华福证券研究所
注：EPS、PE 来自 iFind 一致预期

5 风险提示

技术发展不及预期：模型在设备端搭载需要软硬件技术的协同发展，若技术发展不及预期，相关产品推出时间也将延后。

场景落地不及预期：相比云端大模型，端侧更加重视和用户及场景结合，端侧模型在具体场景中的表现将影响整体产品力。

市场竞争加剧：智能终端市场若竞争加剧，相关供应链公司盈利水平或将承压。

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

一般声明

华福证券有限责任公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。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，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，该等公开资料的准确性及完整性由其发布者负责，本公司及其研究人员对该等信息不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，之后可能会随情况的变化而调整。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。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

在任何情况下，本报告所载的信息或所做出的任何建议、意见及推测并不构成所述证券买卖的出价或询价，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。投资者应自行决策，自担投资风险。

特别声明

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在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。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	评级	评级说明
公司评级	买入	未来 6 个月内，个股相对市场基准指数指数涨幅在 20%以上
	持有	未来 6 个月内，个股相对市场基准指数指数涨幅介于 10%与 20%之间
	中性	未来 6 个月内，个股相对市场基准指数涨幅介于-10%与 10%之间
	回避	未来 6 个月内，个股相对市场基准指数涨幅介于-20%与-10%之间
	卖出	未来 6 个月内，个股相对市场基准指数涨幅在-20%以下
行业评级	强于大市	未来 6 个月内，行业整体回报高于市场基准指数 5%以上
	跟随大市	未来 6 个月内，行业整体回报介于市场基准指数-5%与 5%之间
	弱于大市	未来 6 个月内，行业整体回报低于市场基准指数-5%以下

备注：评级标准为报告发布日后的 6~12 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准（另有说明的除外）

联系方式

华福证券研究所　上海
公司地址：上海市浦东新区浦明路 1436 号陆家嘴滨江中心 MT 座 20 层
邮编：200120
邮箱：hfyjs@hfzq.com.cn