# 电子行业深度报告

# 2026 年端侧 AI 产业深度：应用迭代驱动终端重构，见证端侧 SoC 芯片的价值重估与位阶提升

增持（维持）

2026 年 02 月 23 日

证券分析师 陈海进
执业证书：S0600525020001
chenhj@dwzq.com.cn
证券分析师 李雅文
执业证书：S0600526010002
liyw@dwzq.com.cn

## 投资要点

- **AI 应用迭代驱动端侧硬件需求持续攀升，端侧高算力升级推动传统手机和 PC 端侧存量市场格局重塑，行业巨头需依托 AI 软件需求驱动硬件创新以巩固地位。**AI 应用的落地离不开端侧硬件支撑，其快速发展也持续抬升端侧硬件需求，豆包手机形态、Openclaw 带火的 Mac Mini 均是端侧 AI 终端落地的标志性案例。AI 应用对端侧硬件算力与效率提出明确要求，推动传统手机、PC 芯片加速向高端化升级，也推动相关芯片在制程工艺与架构设计上持续革新。PC 和手机作为核心用户入口，是大模型从算力中心走向物理世界、触达 C 端与 B 端用户的第一入口，也是端侧 AI 最大的落地载体；该赛道亦吸引各大云厂商跨界布局，新兴力量的突围进一步重塑市场竞争格局。抓住端侧入口的大厂，以及积极适配新型 AI 应用、重新定义 PC 和手机芯片产品的公司将在 AI 竞争中占据优势。行业巨头虽坐拥深厚的端侧芯片技术壁垒，可满足低功耗与高端算力的核心要求，但仍需与时俱进，以软件模型驱动硬件产品创新，方能持续稳固行业领先地位。

- **车载场景是端侧 AI 落地的最佳实践场景，车载芯片的迭代升级与国产生态构建将迎来重要发展机遇。**汽车产品形态天然搭载智驾所需的超高算力芯片、人机交互界面以及物联互联控制所需的车载芯片，同时车载大电瓶供电可在一定程度上弥补端侧功耗瓶颈，是最适合端侧 AI 硬件应用实践的理想场景，英伟达提出的智驾世界模型也进一步强化了对端侧算力的需求。车载芯片主要分为座舱芯片与智驾芯片两大品类，座舱芯片一方面迎来国产芯片的强势追赶与替代，IoT 芯片向上迭代，智驾、手机芯片也在尝试降维打击，另一方面在技术上持续向智能化方向演进，硬件端支撑手机与车端互联的软件生态发展。智驾芯片则在技术上正经历从 L2 到 L4 的持续算力跃迁，由感知智能向世界模型跨越，同时国产芯片凭借智驾平权与多价位全布局实现全方位突围，并通过与终端车企紧密合作、适配软件生态重新定义汽车智驾的产业模式，与此同时座舱芯片与智驾芯片正逐步向单芯片的终极形态融合演进，这一过程仍需一定时间，而国产芯片通过与新能源车企深度合作、伴随国产新能源车型出海，叠加依托软件能力构筑自身生态壁垒，将迎来核心发展机会。

- **IoT 市场是当前规模最大的蓝海市场，也是国产替代的核心机遇所在。**IoT 覆盖穿戴、家居、工业等多元场景，不仅对硬件技术能力提出要求，更需要适配具体场景与终端的定制化解决方案和软件生态。国产芯片依托国内丰富的终端消费电子产业基础，拥有广阔的合作开发空间。其中，AI 眼镜仍是当下尚未被证伪的优质端侧场景，无论是作为手机的衍生产品，还是探索替代手机的产业方向，均在持续寻求最优解决方案。具身智能有望实现与 IoT、智能驾驶领域的技术能力平滑迁移，其适配场景的需求逐步清晰，这些尚未完全定型的终端 AI 新场景，均为国产 IoT 芯片带来重要发展机遇。

- **持续关注大厂硬件建设的核心供应链企业。**互联网与云算力大厂加速布局端云协同硬件生态、筑牢 AI 转型硬件底座，云算力企业与互联网大厂均在持续加大端侧布局力度，通过搭建端云协同的闭环硬件生态体系，夯实自身向 AI 全面转型的底层硬件支撑。从投资视角出发，紧密跟踪国内外科技大厂的端侧战略布局动向，深度绑定大厂产业链、成功跻身其硬件核心供应链的相关企业，将充分受益于行业发展红利，迎来

## 行业走势


## 相关研究

《格局落定，价值归真：从周期波动走向技术溢价》

2026-02-06

《Moltbot 重构个人 AI 助理：边缘算力硬件新赛道》

2026-02-01

清晰可观的投资机会。

- **端侧 AI 的产业发展趋势明确且确定，存储涨价压制只是短期扰动。**AI 云算力与 AI 应用向物理世界延伸，是技术与产业演进的必然阶段；端侧需具备 AI 和互联能力，方能有效承接云端算力与应用的落地需求。端侧 AI 存在的核心合理性，包括隐私保护、安全、低延迟，以及依托端侧算力实现的多模态初步处理能力等。存储价格自 2025 年二季度启动上涨，下半年进入全面加速涨价阶段，阶段性掩盖了端侧产业基本面向上的趋势。展望 2026 年，若存储市场涨价压力逐步缓解，或相关企业通过自身举措有效对冲成本端压力，AI 赋能端侧硬件所带来的基本面改善有望充分显现。短期的周期性扰动，无法掩盖端侧 AI 长期的技术发展大趋势。

- **产业链相关公司：**
  - 端侧 AI SoC 芯片：晶晨股份、瑞芯微、星宸科技、恒玄科技、乐鑫科技、地平线机器人（汽车&传媒互联网海外组覆盖）、黑芝麻智能（汽车&传媒互联网海外组覆盖）等
  - 端侧存储芯片：兆易创新等
  - 消费电子终端与供应链：立讯精密、歌尔股份、领益智造、东山精密、绿联科技（商社组覆盖）等
  - 端云生态：阿里巴巴（传媒互联网海外组覆盖）等

- **风险提示：**国内端侧算力互联及端云协同建设需求不及预期；手机、PC 等消费终端换机周期复苏不及预期；端侧 AI 芯片国产替代进度及产品量产落地不及预期；行业竞争加剧；国内数据安全与合规相关政策调整带来经营不确定性风险。

表 1：重点公司估值（数据截止于收盘时间 2 月 13 日）

| 代码 | 公司 | 总市值（亿元） | 收盘价（元） | EPS（元/股） | | | PE | | | 投资评级 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 2024A | 2025E | 2026E | 2024A | 2025E | 2026E | |
| 688099 | 晶晨股份 | 397.07 | 94.28 | 1.95 | 2.07 | 3.67 | 48.31 | 45.55 | 25.69 | 买入 |
| 603893 | 瑞芯微 | 776.04 | 184.35 | 1.41 | 2.52 | 3.47 | 130.46 | 73.15 | 53.13 | 买入 |
| 301536 | 星宸科技 | 299.25 | 70.96 | 0.61 | 0.63 | 0.96 | 116.78 | 112.63 | 73.92 | 买入 |
| 688608 | 恒玄科技 | 352.59 | 209.01 | 2.73 | 5.13 | 7.40 | 76.57 | 40.74 | 28.24 | 买入 |
| 688018 | 乐鑫科技 | 279.63 | 167.30 | 2.03 | 3.24 | 4.84 | 82.41 | 51.64 | 34.57 | 买入 |
| 603986 | 兆易创新 | 2,164.30 | 308.70 | 1.57 | 2.48 | 3.36 | 196.30 | 124.48 | 91.88 | 买入 |
| 002241 | 歌尔股份 | 927.54 | 26.16 | 0.75 | 0.99 | 1.26 | 34.80 | 26.42 | 20.76 | 买入 |

数据来源：Wind，东吴证券研究所

# 内容目录

# 图表目录

# 1. 端侧 AI 开启物理世界入口，国产供应链迎来跨越性机遇

2026 年，端侧人工智能已从技术构想转变为高度确定的产业演进路径，标志着大模型正由云端算力中心向物理世界入口实现战略级重心转移。这种演进趋势植根于隐私安全保障、毫秒级极致响应时延以及带宽流量成本等物理规律对云端算力的刚性约束，使得终端设备不再仅是云端的延伸，而是逐步演变为承载产业价值落地，并实现感知、记忆与执行全链路实时闭环的物理衍生形态。随着端云协同架构成为主流，人工智能在手机、个人电脑、座舱及机器人等终端的本地化推理，已成为驱动全球电子产业由周期性波动转向技术溢价，并重塑存量市场格局的核心确定性支柱。

在此技术趋势下，端侧硬件的深度重构为国产供应链提供了系统级的位阶提升机遇。依托异构计算架构革新，特别是神经网络处理单元算力的显著突破，配合模型轻量化技术的深度协同，产业竞争范式正由单纯的硬件参数博弈转向由软件定义硬件主导的系统级生态博弈。凭借本土供应链的成熟度与敏捷的生态响应能力，国产厂商在人工智能眼镜、具身智能机器人等新型终端赛道已主导部分芯片创新与量产方案，其硬件适配生态正展现出抢占蓝海市场高份额的确定性潜力。这种全链路实时交互能力不仅推动了国产供应链的价值跃迁，更为全球端侧人工智能赛道确立了以中国市场为核心的增长逻辑。

# 2. AI 赋能核心基本盘手机与 PC 市场的存量革新

就 OpenClaw 和豆包手机的不俗表现来看，PC 与手机依然是大模型实现端侧全链条执行的主要物理载体。我们观察到，用户为了获取自动化带来的效率提升，对让渡底层硬件控制权限的接受度较高，这为端侧智能体的落地提供了需求基础。

然而如果考虑商业化，当前无论是 GUI 模拟还是 API 协同技术路线，都面临数据安全、隐私合规及生态兼容的现实挑战。这种从 Demo 验证向规模化商用演进的矛盾，正驱动终端产业链在底层架构上进行创新，比如引入硬件级隔离与分级权限管控。“端云协同”的混合计算架构，仍是目前平衡执行效率与隐私安全的有效方案。

## 2.1 手机芯片高端化与 AI 驱动：性能体验升级及市场格局重塑

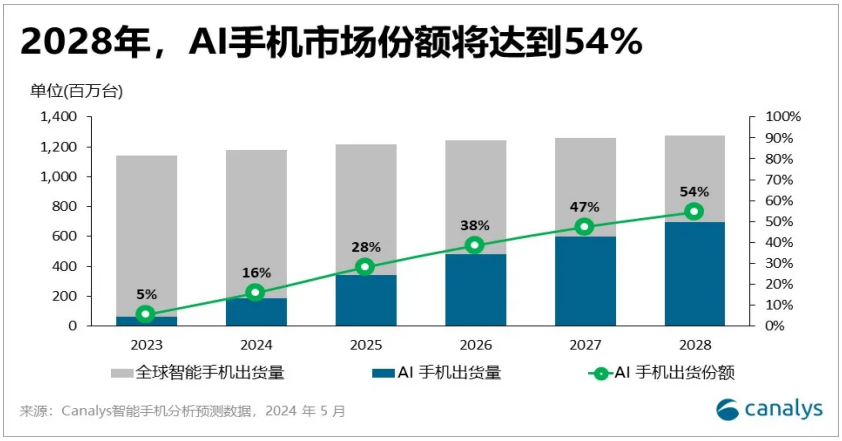
### 2.1.1 手机芯片产品趋于高端化，市场增长锚定 ASP 提升

**人工智能技术的快速普及正在刺激高端智能手机需求，端侧 AI 加速渗透手机市场。**未来几年，AI 手机的出货量和渗透率将呈现持续增长态势。据 Canalys 数据，渗透率预计将于 2028 年达到 54%，实现市场中超过一半的智能手机的端侧 AI 部署。据 Counterpoint 预测，2026 年 90%的高端智能手机将支持端侧 AI 功能。相比之下，售价在 100 至 500 美元间的中端智能手机，在内存价格持续上涨的压力下，或更多依赖云端 AI 处理以控制成本。端侧 AI 手机渗透率的上升，也表明智能手机市场需求和结构向高端化方向发展的趋势。

**全球智能手机市场扩张逻辑聚焦于“卖得更贵”，而非“卖得更多”。**手机高端化趋势加速，存储等物料成本的上涨进一步抬升智能手机的整机成本，推动智能手机 25 年四

季度平均售价同比上涨 8%，当季度售价首次突破 400 美元。

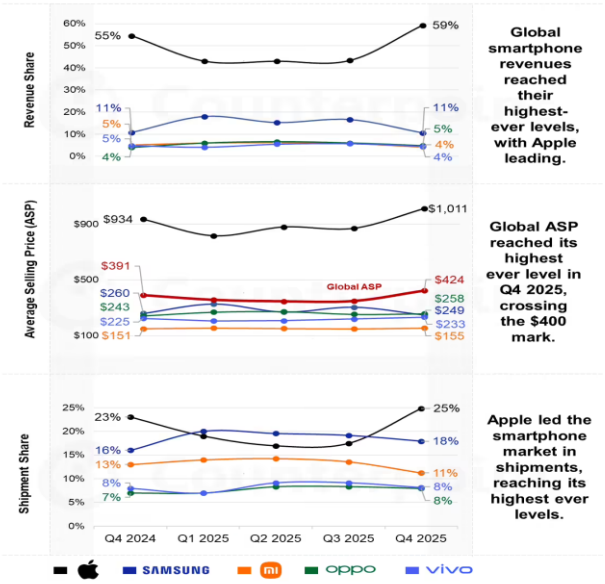
**图1：全球 AI 手机出货及渗透率**




数据来源：canalys，东吴证券研究所

**在端侧 AI 性能日益增长的需求下，市场增长锚定平均售价（ASP）的提升，此趋势有望带动厂商销售额实现逆势增长。随着** DRAM、NAND 及其他半导体供应端压力不断升级，将持续压缩 2026 年的出货量增长空间和厂商利润空间，导致 26 年智能手机出货量的下降趋势。同时，市场结构分化趋势将加快，市场将持续向高端化方向发展。受高端化趋势和 AI 功能需求的进一步提升，设备均价将进一步提升。据 Counterpoint，26 年近三分之一的手机售价预计将超过 500 美元。厂商将更加重视价值增长和产品的结构调整，高端机型和中低端机型的分化将进一步提升。

**图2：全球智能手机前五大厂商市场收入与平均售价**

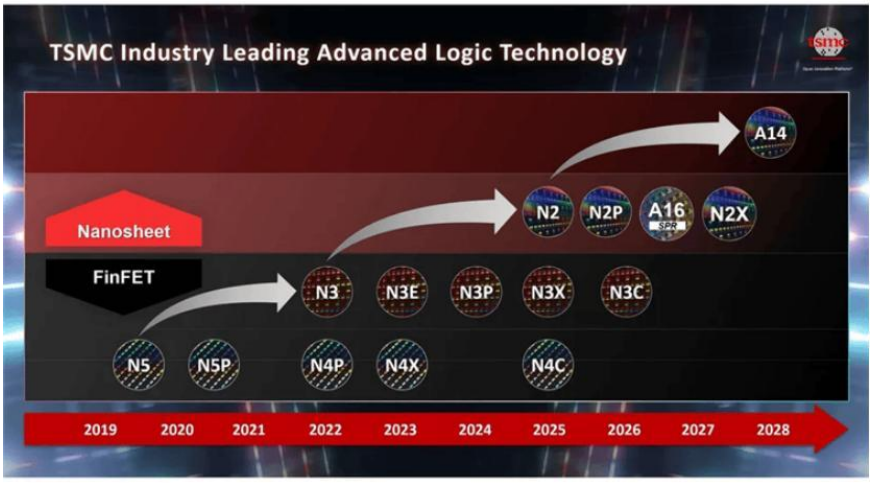



数据来源：Counterpoint Research，东吴证券研究所

### 2.1.2 驱动制程迭代与架构革新，双轮赋能算力升级

**芯片制程工艺由 3nm 向 2nm 迭代，加速片内架构算力升级。**25 年第四季度，台积电 2nm 制程（N2）在新竹和高雄同步量产，良率突破 80%，远超行业预期，2nm 工艺制程趋于成熟。台积电也同步发展低阻值重置导线层与超高效能金属层间电容，以持续进行 2nm 制程技术效能提升。据半导体产业纵横数据，26 年下半年将进行 N2P 制程技术的量产。相比 3nm 制程，2nm 在相同功耗下性能提升 10%-15%，或在相同性能下降低功耗 25%-30%，晶体管密度提升 15%（混合设计）或 20%（纯逻辑电路），具有能效更低和 AI 处理能力更好的优势。

**图3：台积电制程工艺路线图**




数据来源：半导体产业纵横，东吴证券研究所

手机 SoC 市场中，台积电为苹果、高通、联发科等头部厂商的核心流片服务商。预期苹果的 iphone18 系列高端机型搭载的 A20 处理器、以及预期 26 年 H2 发布的高通旗舰手机芯片骁龙 8 Elite Gen 6 系列与联发科旗舰芯片天玑 9600 都将采用台积电 2nm 制程生产，市场需求旺盛。为应对巨大的市场需求，台积电计划在中国台湾和美国共同建设 10 座 2nm 制程工厂。据半导体产业纵横数据，2026 年底产能预计将达到 8 万至 10 万片晶圆。

**而架构革新筑牢底层支撑，软硬件协同推动端侧 AI 能力持续升级。**架构革新是端侧 AI 能力提升与落地的底层核心支撑，为模型架构优化、硬件算力升级提供技术赋能，是推动端侧 AI 能力进阶的关键抓手。为适配端侧设备部署需求，端侧 AI 模型正向稀疏化、轻量化迭代，架构层面通过 MOE、GQA 等技术精简模型规模、降低内存占用，弱化设备性能要求；模型层面借助量化、知识蒸馏提升知识密度，实现大模型“瘦身”适配。硬件算力突破是端侧 AI 落地的前置条件，其中 NPU 算力与内存能力升级尤为关键。当前主流旗舰手机 SoC NPU 算力已普遍突破 50TOPS，可满足 7-13B 量级端

侧模型的落地算力标准，这一硬件升级的实现，正是依托架构革新的技术支撑。架构层面的核心跃迁，是从冯·诺依曼架构到异构计算架构的升级，这一革新打破了传统架构中“存储墙”与“功耗墙”的双重瓶颈。传统冯氏架构因计算与存储单元物理分离导致效率低下，制约端侧 AI 算力提升，而存算一体技术通过深度融合计算与存储功能，直接在存储阵列中完成核心计算任务，实现超高算力与能效比，为端侧 AI 能力升级扫清架构障碍。目前，异构计算架构已成为高端智能手机 SoC 的主流选择，其可根据不同计算任务特性精准调配专用核心，适配 AI 计算的多元化需求；叠加云端大模型端侧下沉带来的算力需求提升，以及大型语言模型、多模态模型运算复杂度的升级，NPU 作为核心 AI 算力单元，规模持续扩展、可编程特性不断丰富，其算力革新成为厂商提升端侧 AI 能力的核心发力点，且全程依托异构计算架构的革新赋能。

**端侧 AI 将以架构革新为底层支撑、硬件升级为主线、模型优化为辅助实现能力升级，长期达成复杂推理本地运行并与云端协同赋能复杂场景。**当前端侧 AI 能力虽已显著提升，但与云端算力仍存较大差距，依据 OpenAI《Scaling Laws for Neural Language Models》论文中的算力公式（算力需求 =2× 参数量 ×token 数），相同 1000tokens 文本推理任务中，端侧 7B 小模型算力需求仅约 14TOPS，远低于云端 GPT4 大模型的 560TOPS，因此复杂场景下端侧仍需外接云端算力，端云协同成为行业过渡期核心解决方案。展望未来，端侧 AI 能力的持续升级需架构、硬件、模型三者协同发力，架构革新将持续为硬件升级、模型优化提供核心技术支撑，而 AI 端侧需求的强劲增长对本地运算能力提出更高要求，手机 SoC 的硬件算力升级也将成为行业长期发展主线；同时针对主流 AI 功能开展模型定向优化，可进一步减轻硬件算力负载，助力端侧 AI 能力高效提升，持续推动端侧 AI 从“可用”向“好用”迭代升级，最终依托多维度升级稳步实现复杂推理的本地自主运行。

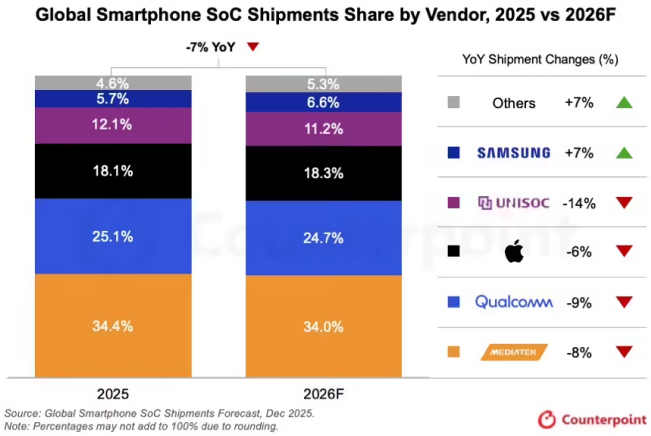
### 2.1.3 增量重塑，打破手机存量僵局

**联发科在智能手机 SoC 出货量方面以领先优势占据榜首，凸显规模优势。**数据显示，在 2025 年的全球智能手机 SoC 市场，联发科以 34.4%的出货量份额位居第一，高通（25.1%）、苹果（18.1%）、紫光展锐（12.1%）和三星（5.7%）紧随其后。但存储价格的暴涨或对各手机 SoC 厂商，尤其是手机芯片营收占比大的联发科带来较为严重的影响。据 Counterpoint 数据，2026 年联发科的手机 SoC 出货量预计同比下滑 8%，市占率同比下滑 0.4 个百分点至 34%，但联发科仍旧将保持较为稳固的出货量领先优势。

截止于 2025 年，联发科的优势来自于精准的市场定位，以性价比的差异化优势开拓中高端手机 SoC 市场。联发科的崛起在于精准踩中安卓阵营需求痛点，高通骁龙芯片溢价过高，联发科主打“性价比”战略，将中国厂商作为主要客户，对国内手机厂商几乎有求必应。不仅开放芯片底层调度权限，还能根据厂商需求定制专属版本，这与高通的“强势管控”形成鲜明对比。联发科的天玑系列填补厂商对于性价比突出、性能够用的中高端芯片的市场需求，实现中高端市场的稳健布局。依托中国台湾成熟的半导体供

应链，联发科在芯片设计环节严控成本，同款性能芯片售价比高通低 15%-20%，从而在中高端领域通过“低价高配”吸引小米、realme、一加等品牌，2026 年还发布 6nm 制程的天玑 7100 芯片，稳固扎根中端市场。

**图4：全球手机 SoC 厂商出货量情况**




数据来源：Counterpoint，东吴证券研究所

**而聚焦 2026 年，联发科则在积极突围高端市场。**其旗舰芯片天玑 9500 转变竞争思路，从参数竞赛走向体验深耕。天玑 9500 芯片基于台积电第三代 3nm 制程工艺，并采用创新的“1+3+4”全大核 CPU 架构，包含一颗主频高达 4.21GHz 的 C1-Ultra 超大核、三颗 3.5GHz C1-Premium 大核及四颗 2.7GHz C1-Pro 大核。同时搭载第二代天玑调度引擎，具备场景感知与动态资源分配能力。其单核性能较天玑 9400 提升 32%，多核性能提升 17%。相比于纸面参数的提升，天玑 9500 更重视真实体验的优化。搭载天玑 9500 芯片的旗舰手机连续运行多款高负载游戏后，机身温度依然维持清凉水平。AI 体验方面，采用全新超性能 NPU990+超能效 NPU 双架构设计，可支持百亿参数大语言模型的本地化运行，“AI 修图”、“实时转写”、“离线摘要”等功能相应迅速，保护用户隐私安全。

**图5：高通与联发科旗舰芯片架构对比**

| 芯片 | CPU | GPU | NPU |
|---|---|---|---|
| 联发科天玑9500 | 1）采用台积电N3P制程工艺<br>2）采用“1+3+4”CPU架构，超大核主频高为4.21GHz、三颗3.5GHz C1-Premium大核及四颗2.7GHz C1-Pro大核，单核性能较天玑9400提升32%，多核性能提升17%<br>3）采用PC级架构ARMV9.3，集成最新矩阵运算指令集SME2 | 1）采用基于Dynamic Cache架构的GPU G1-Ultra<br>2）峰值性能较上一代提升33%，功耗较上一代下降42%，光追性能较上一代提升119% | 1）采用全新超性能NPU990+超能效NPU双架构设计<br>2）超性能NPU990峰值性能较上一代提升111% |
| 高通骁龙8 Elite Gen5 | 1）采用台积电N3P制程工艺<br>2）采用“2+6”的CPU架构，超大核主频为4.6GHz,性能核为3.62GHz<br>3）单核性能提升20%，多核性能提升17%，能效提升35% | 1）采用频率为1.2GHz的GPU Adreno 840<br>2）首次配备18MB独立高速显存<br>3）图形性能提升23%，光追性能提升25% | 1）Hexagon NPU AI性能提升约37%，每瓦性能提升16%<br>2）支持端侧大语言模型（LLMs)和个性化AI助手 |

数据来源：联发科，高通，东吴证券研究所

**高通一直以来主营高端化市场，其高质量芯片深受各大手机厂商信赖。**高通在 2025

年的智能手机 SoC 的出货量份额占比为 25.1%，低于联发科居市场第二位。但高通手机芯片业务的毛利率高达 52%，远高于同期联发科的 35%。其主要原因在于高通不仅靠芯片赚钱，还通过其通信技术层面的专利，向各大手机厂商和运营商收取授权费，其专利授权收入占比超 30%，利润率高达 80%以上，构建“芯片 + 专利 + 生态” 护城河。其高端芯片的稳定性以及与厂商长期合作关系，在高端 SoC 市场依旧是小米、荣耀、oppo、vivo 等厂商旗舰手机芯片的第一选择。

**图6：安卓旗舰手机性能测试排行**


数据来源：安兔兔，东吴证券研究所

高通骁龙 8 Elite Gen5 树立全新性能标杆，高通在高端市场延续产品领先态势。骁龙 8 Elite Gen5 采用台积电 N3P 制程工艺和最新一代 Oryon 架构，与上一代相比 CPU 单核性能提升约 20%，能效提升 35%。GPU 部分首次引入 Adreno 高性能显存（HPM）架构，芯片配备 18MB 独立高速显存，实现 23%的 GPU 性能提升和约 20%的功耗下降。在 AI 算力方面，NPU 整体性能提升约 37%，每瓦性能提升约 16%，终端侧 AI 算力高达 220TOPS。Geekbench 6 测试显示，骁龙 8 Elite Gen 5 的每瓦性能远高于联发科同期对标产品天玑 9500。最新公布的安卓旗舰手机性能榜单显示，前十名有 8 款机型搭载骁龙 8 Elite Gen5 芯片，高通在旗舰级芯片性能端形成压倒性优势。

**聚焦华为，麒麟芯片的稳定迭代助力华为实现国产突围，并在高端市场竞争中强势回归。**2025 年华为重返中国智能手机市场出货榜首，其旗舰芯片麒麟 9030Pro 精简了传统制程束缚，NPU 采用达芬奇架构，AI 算力达到 40TOPS 。华为也通过实现从芯片设计到操作系统的垂直整合，利用 AI 算力的提升优化了性能功耗表现，成功重掌高端市场话语权。而麒麟芯片的这一成功，源于华为对卡脖子困境的艰难突破。2020 年美国的极端封锁使华为被迫停止与 ARM 的合作，同时切断了麒麟芯片的台积电代工渠道，在这种困境下，华为麒麟坚持全栈自研路线，实现从芯片设计到操作系统（鸿蒙）的垂直整合。2023 年搭载 7nm 工艺的麒麟 9000S 在 Mate 60 系列中强势回归，更以核心零部件 100% 国产化率进一步稳固高端市场布局，2025 年年底推出的麒麟 9030Pro 更在缺少 EUV 的情况下，利用落后两代的 DUV 设备和自对准四重曝光（SAQP）工艺，在工艺密度上追平三星 5nm 旗舰芯片，性能推进到骁龙 8Gen2 水平，打开了高端化市场的国产突围

新格局。

**图7：中国市场主要厂商智能手机出货量**

| 厂商 | 2025年出货量（单位：百万台） | 2025年市场份额 | 2024年出货量（单位：百万台） | 2024年市场份额 | 同比增幅 |
|---|---|---|---|---|---|
| 华为 | 46.7 | 16.40% | 47.6 | 16.60% | -1.90% |
| 苹果 | 46.2 | 16.20% | 44.4 | 15.50% | 4.00% |
| vivo | 46.1 | 16.20% | 49.3 | 17.20% | -6.60% |
| 小米 | 43.8 | 15.40% | 42 | 14.70% | 4.30% |
| OPPO | 43.4 | 15.30% | 42.5 | 14.80% | 2.10% |
| 其他 | 58.2 | 20.50% | 60.4 | 21.10% | -3.50% |
| 合计 | 284.4 | 100% | 286.2 | 100% | -0.60% |

数据来源：IDC，东吴证券研究所

华为麒麟 9030Pro 制程差距客观存在，但与行业顶尖性能水平的差距缩小。麒麟 9030Pro 采用 N+3 工艺，晶体管密度这一硬指标已追平三星 5nm 芯片，但由于是依靠 SAQP 的复杂工艺技术实现，因此在性能功耗表现上更接近台积电 N7P 水平，同时步骤的繁琐也让芯片良率和产能成为客观存在的隐忧。此外该芯片还采用 1+4+4 的核心架构，包含 1 个 2.75GHz 的超大核、4 个 2.27GHz 的大核和 4 个 1.72GHz 的小核，GPU 为 Maleoon 935，计算单元从上代的 5 个提升为 6 个，NPU 的 AI 算力则提升至 40TOPS，根据 Geekbench 的跑分结果，其单核成绩接近 2000 分，多核成绩接近 6000 分，在性能端成功追平骁龙 8 Gen 2，将华为芯片与行业顶尖水平的性能差距从遥不可及拉回至三年内的水平。

**图8：麒麟芯片回归后的产品演进**

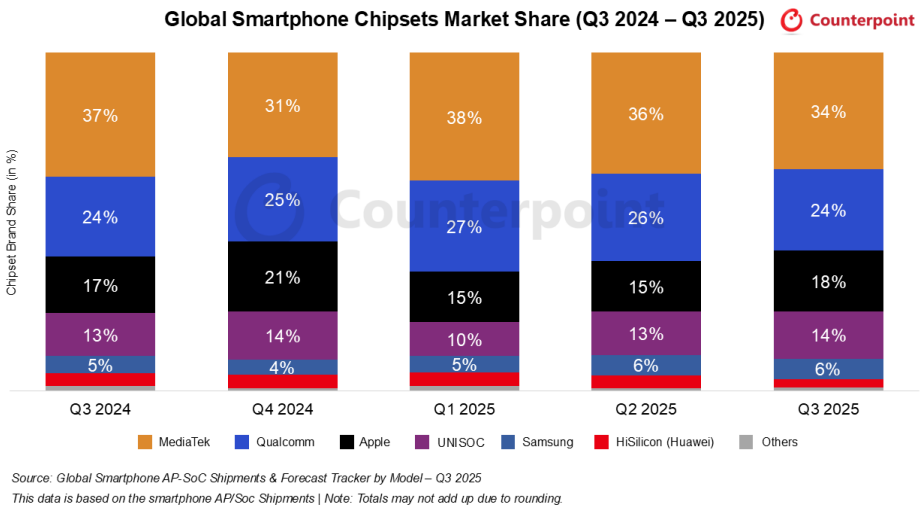
| 芯片型号 | 首发时间 | 首发机型 | 制程工艺 | 性能对标友商处理器 | 历史节点 |
|---|---|---|---|---|---|
| 麒麟9000S | 2023年8月 | 华为Mate60Pro | N+2 | 高通骁龙888 | 制裁后华为首款自主制程芯片，标志芯片自主化回归 |
| 麒麟9010 | 2024年4月 | 华为Pulra70 Ultra | N+2 | 高通骁龙8 Gen 1 | 架构优化后性能反超麒麟9000S |
| 麒麟9020 | 2024年11月 | 华为Mate70Pro | N+2 | 高通骁龙8 Gen 2 | 首款支持3GPP R18的5G-A Soc，摆脱部分Arm公版IP |
| 麒麟9030 | 2025年11月 | 华为Mate80Pro | N+3 | 高通骁龙8 Gen 3 | 搭载鸿蒙OS 6,整机性能较上代提升明显 |

数据来源：华为，东吴证券研究所

**紫光展锐深耕 4G 中低端市场，以性价比优势维持中国大陆第一手机芯片厂商定位。**公司利用成熟制程的性价比优势，着眼拉美、非洲、东南亚等新兴市场，在中低端市场具有较强竞争力。25 年紫光展锐手机 SoC 出货量稳步提升，25 年 Q3 市场份额达到 14%，位列全球市场第四位。紫光展锐的产品已覆盖全球 140 多个国家和地区，在传音、中兴

等低端机型中受到广泛应用。产品矩阵低端化倾向同样衍生出收入份额与出货量背离的问题，公司芯片产品的平均售价仅为行业竞争对手的 1/3，在 400 美元以上的高端手机市场其 24 年营收占比不足 1%。

**图9：25Q3 全球智能手机 SoC 市场份额**




数据来源：Counterpoint，东吴证券研究所

依托通信技术优势，紫光展锐加速 5G 产品布局，实现从入门到中高端的全系列布局。紫光展锐作为全球公开市场三家 5G 手机芯片企业之一，长期深耕通信半导体产业，全面掌握 2G-5G、蓝牙、卫星通信、Wi-Fi 等全场景通信技术。紫光展锐在 5G 领域的强劲技术实力和公司产品的性价比成为公司快速实现多元化布局的关键。

**图10：紫光展锐 T9300 的 5G 通信服务**


数据来源：紫光展锐，东吴证券研究所

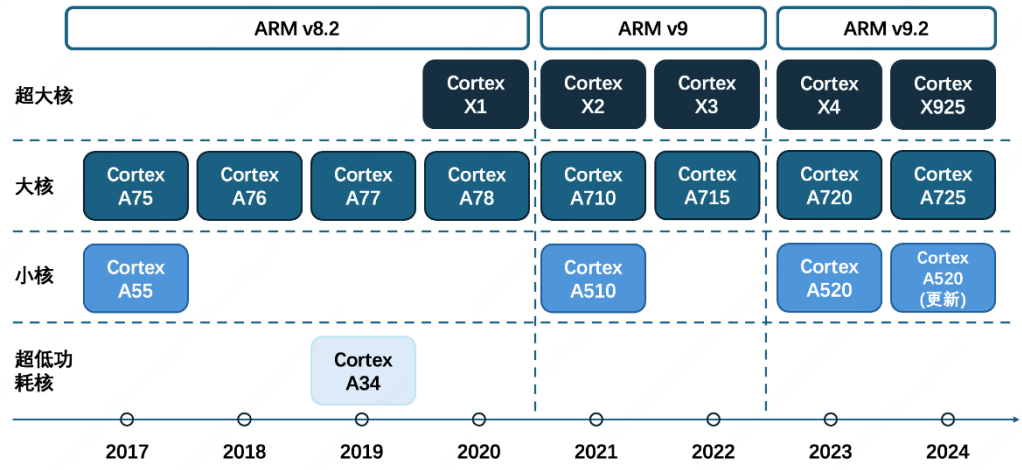
目前，紫光展锐的 5G 芯片已在全球范围内规模商用，搭载展锐芯片的 5G 终端产品已在全球 88 个国家和地区实现规模出货，并在 122 个国家和地区完成网络适配和场测，为全球用户带来便捷的 5G 智能体验。

紫光展锐最新 5G SoCT9300 上市，5G 体验再升级，展锐 5G 产品矩阵愈加丰富。紫光展锐 T9300 采用 6nm 制程与八核架构，由 2 个 2.4GHz 主频的 A78 和 6 个 2.2GHz 的 A55 组成，整体能效相比上一代产品提升 38%。同时搭载展锐第 7 代 Vivimagic 影像引擎，升级 2 亿像素摄像头，影像效果全面提升。紫光展锐 T8300 全面支持最新 3GPP R17 标准，融合 5G NR NTN 卫星通信及 5G MBS 广播功能，为用户带来更多元化、更便捷的 5G 使用体验。但其性能表现依然无法和华为的麒麟芯片相比，在市场定位方面依旧瞄准 5G 入门级市场，致力于为用户提供轻旗舰级性能体验。

**端侧 AI 对先进制程依赖引发了上游供应链的激烈博弈，2nm 产能已成为联发科、高通等厂商旗舰竞争的制约因素。**苹果、高通、联发科的下一代 AI 旗舰芯片均计划采用台积电 2nm 工艺，导致产能遭受疯抢并可能供不应求。台积电正通过扩大资本开支和新建工厂来应对激增的 AI 芯片需求，产能争夺战有望成为决定未来市场竞争格局的关键力量。

**值得一提的是，ARM 架构的灵活授权策略对厂商在 AI 性能与成本之间的权衡产生了深远影响。**公版 IP 内核授权方案助力联发科、紫光展锐等厂商降低设计成本、缩短开发周期，快速铺开 AI 产品线；而私版指令集架构授权则赋予高通、苹果更高的自主性，使其能针对端侧 AI 算法进行深度内核定制。架构选择带来的不同特点，很大程度决定了厂商在端侧 AI 生态中的差异化竞争优势。

**图11：ARM 的指令集与内核架构**




数据来源：Arm 官网，东吴证券研究所

## 2.2 AI PC SoC：架构之争与算力重构

### 2.2.1 模型端侧部署与硬件进化共同引爆 AI PC 换机潮

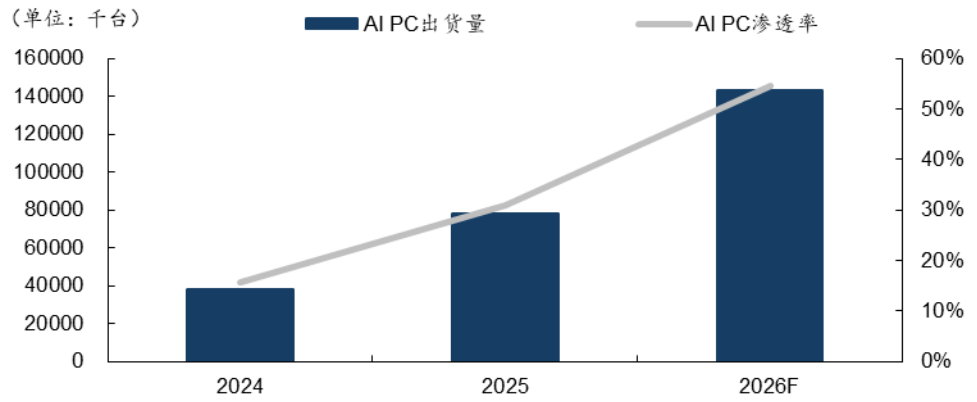
**AI 技术发展加速，发展重心由云端向端侧转移。**一方面，端侧 AI 模型迭代为端侧部署奠定基础。DeepSeek 模型完成从 DeepSeek-V3 到 DeepSeek-R1 的演变过程，通过架构的优化与创新大幅降低训练成本，从而降低行业门槛，推动端侧硬件智能化的普及。

OpenAI 的 GPT 系列模型和谷歌的 Gemini 大模型不断升级模型压缩技术，实现 AI 大模型的端侧部署。另一方面，硬件设备的协同进化成为端侧 AI 落地关键。SoC 芯片向异构多核架构转变，采用“通用核+专用核”模式提升性能，利用异构架构动态分配负载，显著提高能效比。DRAM 在微观结构和封装形式进行内存技术的双重变革，从而提升频率、带宽和能效比以满足 AI 对数据吞吐量的指数级需求。电池电芯数量增加、能量密度提升，通过续航和补能体验的优化满足高强度 AI 应用下的续航体验。

**相比云端，PC 在端侧部署 AI 大模型具备多方面优势。**时延方面，AI PC 依靠本地推理实现即时响应。端侧 AI 将 AI 算法和模型直接部署在 PC 上，能够在本地实时响应用户需求，且其集成化的特点能减少因数据传输而引起的延迟，提供更为流畅和即时的用户体验。本地化大模型部署不再受制于网络条件约束，减少因网络和信号质量对于系统稳定性的影响。安全方面，AI PC 能帮助解决数据安全问题，保护客户隐私。AI PC 通过本地化处理数据，增强终端设备自身独立性。由于不需将数据上传至云端，能有效减少数据的暴露和篡改风险。个性化方面，本地大模型通过长期学习提供个性化服务。在利用本地存储信息和自身长期学习的基础上，可以给出更贴合用户生活习惯和需要的个性化服务和建议。

**Windows 10 停服换机潮或促使 AI PC 于 2026 年大规模普及。**本地 AI 计算设计的新一代处理器的商用化，推动全球 PC 市场逐步走向 AI 赋能设备时代。据 Gartner 预测，截止 2025 年底，全球 AI PC 的出货量将达到 7780 万台。全球市场渗透率从 2024 年的 15.6%快速上升至 2025 年的 31%，并于 2026 年进一步加速渗透，据 Gartner 分析师预测，2026 年 AI PC 出货量将达到 1.43 亿台，到 2029 年，AI PC 将成为常态。

**图12：全球 AI PC 出货量和渗透率情况**




数据来源：Gartner，东吴证券研究所

**国内 AI PC 市场呈现繁荣发展态势，AI PC 渗透率逐年攀升。**中国的 AI PC 市场呈现稳步增长态势，2025 年第二季度中国大陆 AI PC 出货量占中国大陆个人电脑总出货量的 28%，消费者和企业对更高硬件的需求不断加强，中国本土 AI 系统加速发展。据

Canalys 数据预测，到 2027 年，支持人工智能的个人电脑的渗透率将增长至 60%，2029 年底大中华区累计出货预计约 1.07 亿台 AI PC，为 AI 广泛使用奠定装机基础。

**图13：中国 AI PC 出货量和渗透率情况**


数据来源：Canalys，东吴证券研究所

### 2.2.2 架构变革：高算力诉求引发 ARM 高能效路线与 x86 传统生态的深度博弈

**高算力 AI PC 凸显低能耗诉求，与 x86 架构相比，ARM 架构进一步发挥高能效优势。**根据 IDC 新数据，高阶 AI PC 需至少具备 60TOPS 以上算力。处理器的高算力特征大幅提升续航和散热能力要求，低能效、高性能路线将在 AI PC 竞争中成为重要衡量因素。基于 RISC 的 ARM 架构能有效降低指令复杂性，相比 x86，ARM 架构在相同性能下功耗比 x86 低 40%，其低功耗、高能效优势在未来将被进一步放大。

**立足能耗优势，ARM 架构 CPU 快速迭代，性能追赶 x86。**ARM 指令集和微结构的持续优化，从根本上提升 CPU 计算能力。于 2025 年新公布的 ARMv9.3 指令集引入对 SME2 的支持，允许在矩阵和矢量运算中复用架构状态，提升矢量处理能力。主流芯片先进制程工艺向 3nm 工艺迭代，实现晶体管密度的提升，增加逻辑密度，降低功耗，进一步提升 CPU 的运算能力和能效比。

**ARM 架构在边缘计算和低功耗场景有所突破，但 x86 在高性能计算和企业级应用中仍不可替代。**生态系统与兼容性方面，Windows PC 和大部分 Linux PC 主要基于 x86 架构，大部分的桌面应用程序原生支持 x86，具有成熟且兼容性强的软件生态系统，使得 x86 架构在企业级应用和专业软件领域具有不可替代的地位。性能上，x86 架构处理器基于 CISC 指令集，能够为复杂操作系统和应用程序提供全面功能支持，在高性能计算方面表现出色。

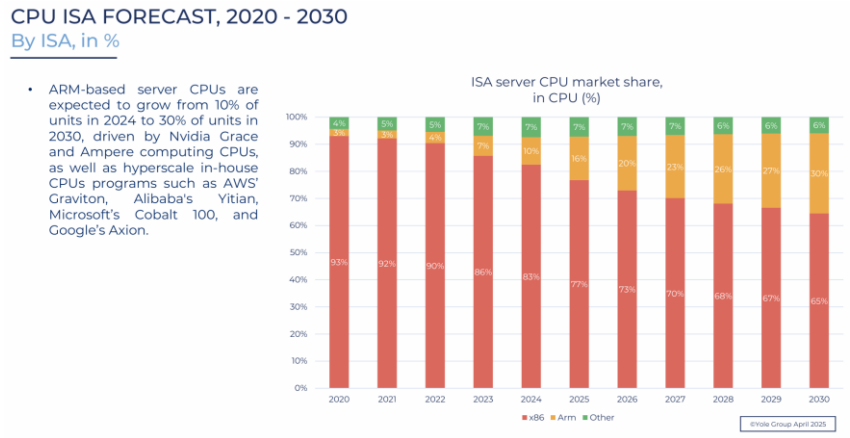
**图14：不同架构的特点对比**

| 种类 | 主要架构 | 架构特征 | 架构优势 | 应用场景 |
|---|---|---|---|---|
| CISC | x86 | •指令系统庞大，功能复杂，寻址方式多，且长度可变，有多种格式<br>•各种指令均可访问内存数据<br>•一部分指令需多个机器周期完成<br>•复杂指令采用微程序实现<br>•系统兼容能力较强 | x86架构兼容性强，配套软件及开发工具相对成熟，且x86架构功能强大，高效使用主存储器，因此在处理复杂指令和商业计算的运用方面有较大优势 | 服务器、工作站和个人计算机等 |
| RISC | ARM | •指令长度固定，易于编码执行<br>•大部分指令可无条件式地执行，降低在分支时产生的开销，弥补分支预测器的不足<br>•算数指令只会在要求时更改条件编码 | ARM架构具有低功耗、小体积的特点，聚焦移动端市场，在消费电子产品中具有优势 | 智能手机、平板电脑、工业控制、网络应用、消费类电子产品等 |
| | MIPS | •采用32位寄存器<br>•大多数指令在一个周期内执行<br>•所有指令都是32位，且采用定长编码的指令集和流水线模式执行指令<br>•具有高性能高速缓存能力，且内存管理方案相对灵活 | MIPS结构设计简单、功耗较低，在嵌入式应用场景具有优势 | 桌面终端、工业、汽车、消费电子系统和无线电通信等专用设备等 |
| | Alpha | •采用32位定长指令集，使用低字节寄存器占用低内存地址线<br>•分支指令无延迟槽，使用无条件分支码寄存器 | Alpha结构简单，易于实现超标量和高主频计算 | 嵌入式设备、服务器等 |

数据来源：海光招股说明书，东吴证券研究所

**X86 架构积极布局 AI PC 领域，持续提升性能和功耗。**在性能端，英特尔通过混合架构设计（大小核）提升能效，AMD 则通过 Zen 架构的优化持续改进 IPC 性能。能效优化上，英特尔的 SpeedStep 技术和 AMD 的 PowerNow!™技术，通过动态调整 CPU 电压和内核频率达到降低功耗的作用。

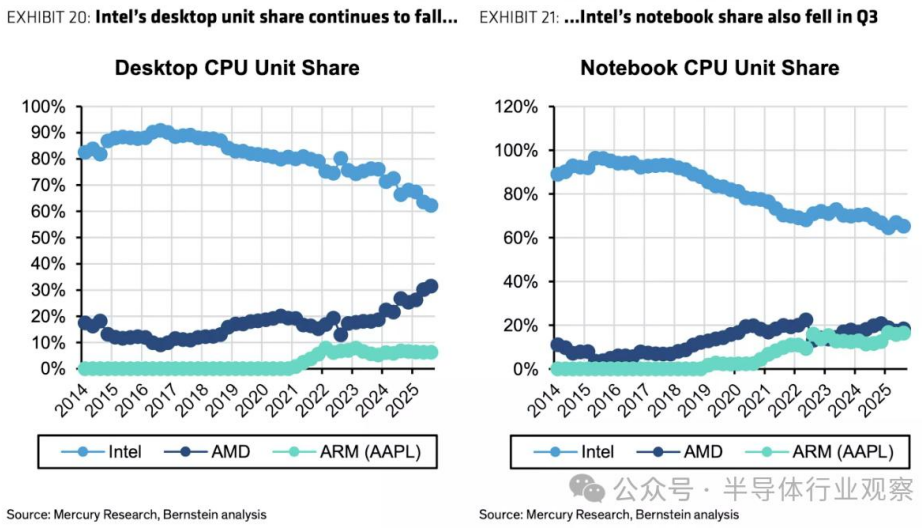
**图15：全球服务器 CPU 市场份额占比**




数据来源：Yole Group，东吴证券研究所

**X86 架构依然在 PC 和服务器市场上占据主导地位。**根据 2025 年数据，x86 架构在桌面 CPU 占据 90%以上的市场份额，在笔记本市场占据 80%的市场份额，市场地位依旧稳固。在服务器 CPU 领域，2025 年 x86 架构市场份额高达 77%。但由于 ARM 架构的强势入局和快速发展，据 Yole Group 数据预测，x86 架构未来几年市场份额或有下降趋势。

**图16：全球桌面和笔记本 CPU 市场份额情况**



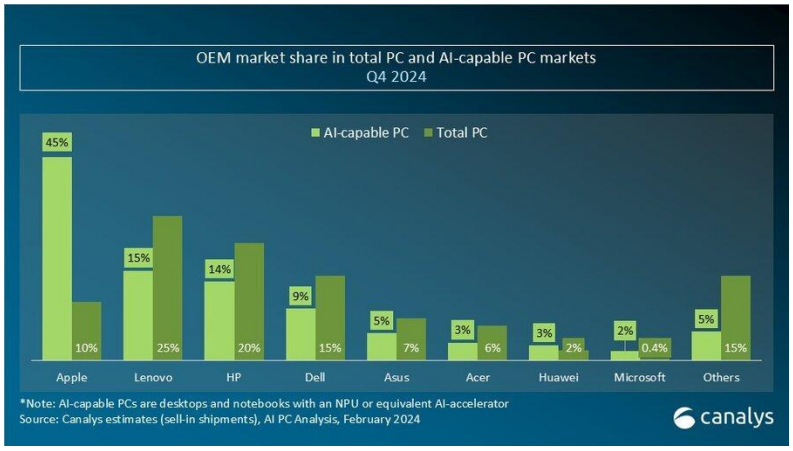

数据来源：半导体行业观察，Mercury Research，东吴证券研究所

**市场格局方面，英特尔市场份额持续下滑，长期霸主地位有所动摇。**近年来，AMD 锐龙系列凭借 Zen 架构（Zen3、Zen4、Zen5）的性能提升和性价比优势，持续侵蚀英特尔市场份额，在桌面 CPU 市场表现强劲，市场份额稳步提升。根据 Mercury Research 数据，2025 年，在桌面 CPU 市场，AMD 的市场份额增长至 30%左右。而英特尔由于制造良率和第 13 代和第 14 代芯片稳定性不足问题，品牌信誉受损，市场份额不断下滑，AMD 和 ARM 架构处理器持续施压，行业竞争生态开始重塑。截止 2025 年，英特尔在桌面和笔记本 CPU 市场份额已跌至 60%，处于近五年最低值。

### 2.2.3 跨界厂商生态突围与传统巨头先进制程反击重塑市场格局

**依靠 M 芯片和软硬件一体化生态优势，苹果成为 AI PC 市场最大赢家。**根据 Canalys 2024 年 Q4 数据，苹果在 AI-capable PC 市场份额达 45%，全年份额为 54%，显著领先于联想、惠普、戴尔等厂商。展望 2025 年，随着 AI PC 渗透率提升，苹果凭借端侧 AI 算力与系统深度优化构筑的壁垒，有望继续领跑市场。

**图17：24Q4 各厂商 PC 和 AI PC 的市场份额**




数据来源：canalys，东吴证券研究所

**M 系列芯片性能强劲，AI 算力再升级。**苹果于 2025 年 10 月推出 M5 芯片，全方位提升芯片表现，实现 AI 性能的又一次跃升。M5 芯片采用台积电 3nm 制程工艺打造，搭载最多达 10 核的中央处理器，包括 6 颗能效核心和最多 4 颗性能核心，多线程性能提速相比 M4 芯片最高可达 15%。M5 芯片采用新一代图形处理器 10 核架构，每颗核心内皆配备专用神经网络加速器，GPU 峰值计算性能较 M4 芯片提升 4 倍以上，基于 AI 任务的 GPU 峰值性能较 M1 芯片提升 6 倍以上，依托 M5 芯片，新款 MacBook Pro 处理 AI 工作流速度显著提升。内存方面，M5 芯片提供 153GB/s 的统一内存带宽，较 M4 芯片提升近 30%。M5 芯片更快的 16 核神经网络引擎能效卓越，搭配 CPU 和 GPU 的神经网络加速器，能实现强大 AI 性能。

**图18：Apple 发布 M5 芯片**

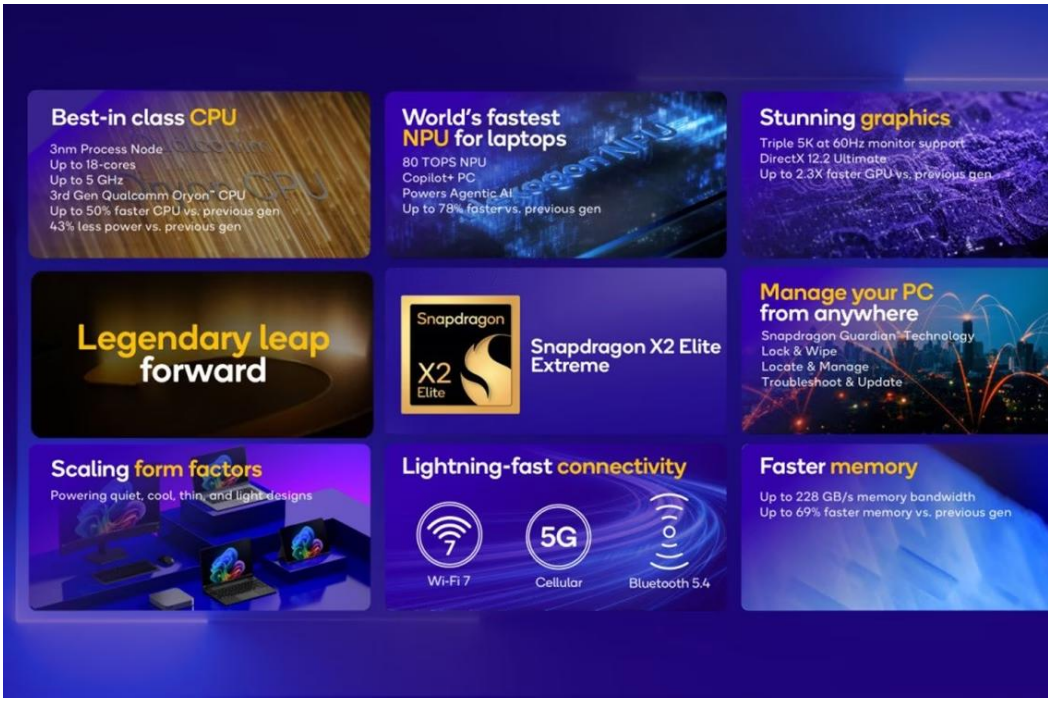

数据来源：苹果，东吴证券研究所

**而对于高通，骁龙 X Elite 系列符合微软算力标准，高通建立初期竞争优势。**微软于 2024 年 5 月公布 AI PC 硬件标准，要满足 AI PC 定义需至少有 16GB 内存和 256GB 硬盘，NPU 算力需达到 40TOPS 以上，在当时现有的 AI PC SoC 中，仅高通的骁龙 X Elite 系列符合微软算力要求。高通从移动端顺利进军 AI PC 市场，并与微软、联想、惠普、戴尔等 PC 厂商建立积极合作关系，推动微软 WOA 生态拓展战略的实施，实现早期 AI PC 市场的精准卡位。

**骁龙 X2 Elite 系列专为 WOA 生态打造，实现性能和能效的大幅提升。**骁龙 X2 Elite 系列采用台积电 3nm 工艺打造，搭载高通第三代自研 Oryon™核心，CPU 的核心数量最高可达 12 核，包括 6 个性能核心和 6 个能效核心，性能核心最高频率可达 3.4GHz。与上一代相比，新款处理器在功耗端降低 43%，相同功耗下 CPU 性能提升 31%。骁龙 X2 Elite 系列集成升级版 Adreno GPU，每瓦性能和能效均较上一代提升 2.3 倍。AI 性能上搭配 Hexagon NPU，NPU 算力高达 80TOPS。产品定位上，标准款骁龙 X2 Elite 面向资源密集

型工作负载的高端 PC，骁龙 X2 Elite Extreme 为专家级工作负载的超高端 Windows 11 PC 专门设计，市场定位明确。

**图19：骁龙 X2 Elite 性能**




数据来源：高通，东吴证券研究所

**AI PC SoC 的新一轮竞争格局下，英特尔的产品力与市场话语权有所弱化，PC 领域的既有格局正迎来结构性变化。**作为 AI PC 概念的提出者，英特尔于 2023 年 H2 推出 Meteor Lake 处理器，占据 AI PC SoC 市场先发优势。但由于芯片制造工业落后与产品延迟等问题，英特尔在市场上缺乏持续竞争力。一方面，WOA 生态强势扩张，高通在算力和能耗上实现对英特尔的超越，与微软进行深度合作。另一方面，AMD 基于 Zen 架构和率先集成的 NPU 迅速崛起，在 x86 阵营内部给英特尔带来强大竞争压力。

**英特尔加速产品迭代，持续提高处理器性能。**为应对市场竞争，英特尔加速推进 AI PC SoC 产品的改良升级。在 Meteor Lake 处理后，于 2024 年推出酷睿 Ultra 处理器（系列二），并且针对不同的市场需求进行产品设计的分化。Lunar Lake 面向高端轻薄本市场，以低功耗作为主要卖点；Arrow Lake 应用于台式机及常规与高性能笔记本市场，面向更为广阔的市场。新产品对于 CPU 架构和 NPU 算力进行持续优化，精进制程工艺，实现性能和能效比的双重提升。

**图20：英特尔 AI PC SoC 的产品迭代**

| 处理器 | 芯片 | CPU | GPU | NPU | 其他 |
|---|---|---|---|---|---|
| 酷睿Ultra处理器（系列一） | Meteor Lake | 1）采用Intel4（7nm)制程工艺；<br>2）具有快速响应能力，适合需要快速决策和低延迟的轻量级AI任务 | 1）采用台积电N5工艺构建的Xe-LPG架构GPU；<br>2）具有性能并行性和高吞吐量；<br>3）擅长处理与媒体、3D应用程序和图形渲染有关的任务，能够同时运行处理大量AI任务； | 1）Intel首款集成式NPU；<br>2）节能、可持续运行和处理AI任务； | 1）采用Foveros封装技术，在芯片内实现极低功耗和高密度的晶片连接 |
| 酷睿Ultra处理器（系列二） | Lunar Lake | 1) 采用台积电N3B制程工艺；<br>2)CPU内核架构更新，搭载4个Lion Cove性能核心和4个Skymont能效核心，P核相比上一代IPC提升14%，E核单核性能是上一代E核的两倍；<br>3）取消超线程设计，采用8线程，提升能效比； | 1）采用基于 Xe2 架构的GPU；<br>2）图形性能相较上代提升1.5倍，AI性能是前代的3.5倍 | 1）采用第四代NPU，NPU算力提高3倍，到达48TOPS<br>2）极低功耗加速AI计算，带来出色用户体验 | 1）延续芯粒（chiplet）设计，采用Foveros先进封装技术，将两颗LPDDR5封装到CPU的PCB上，提升内存频率，降低功耗；<br>2）每瓦性能角度，相较上代提升超过2倍，比骁龙X Elite高出约20% |
| | Arrow Lake | 1) 采用台积电N3B制程工艺；<br>2）CPU性能与能效核心与Lunar Lake相同，但核心规模更大，最高可搭载8颗性能核心和16颗能效核心<br>3）取消超线程设计，追求低功耗 | 1）采用基于Xe-LPG架构的GPU; | 1）NPU核心与Meteor Lake相同，采用第三代NPU，NPU算力为13TOPS | 1）封装设计上类似Meteor Lake的分离式模块化设计，采用先进Foveros封装技术<br>2）36TOPS的台式机AI PC性能，带来卓越计算吞吐性能 |

数据来源：英特尔，东吴证券研究所

**近期基于 Intel 18A 制程工艺的 Panther Lake 上市，英特尔有望重构市场竞争格局。**Panther Lake 是首款基于 Intel 18A（2nm）制程工艺打造的客户端系统级芯片，与 3nm 制程工艺相比，其每瓦性能提升高达 15%，芯片密度提升约 30%。CPU 配备 16 个核心，包括 4 个性能核心和 12 个能效核心，相较上代性能提升约 50%。GPU 上采用全新 Xe3 架构，核显规模达到 12 个 Xe 核心，凭借 XMX 技术可提供 120TOPS 算力，显卡性能相较上一代大幅提升 77%。NPU 采用全新的第五代 NPU，NPU 算力达到 50TOPS，整体可实现 180TOPS 的超高算力提升，大语言模型推理性能较上一代提升 2 倍。

**图21：Panther Lake 的性能提升**

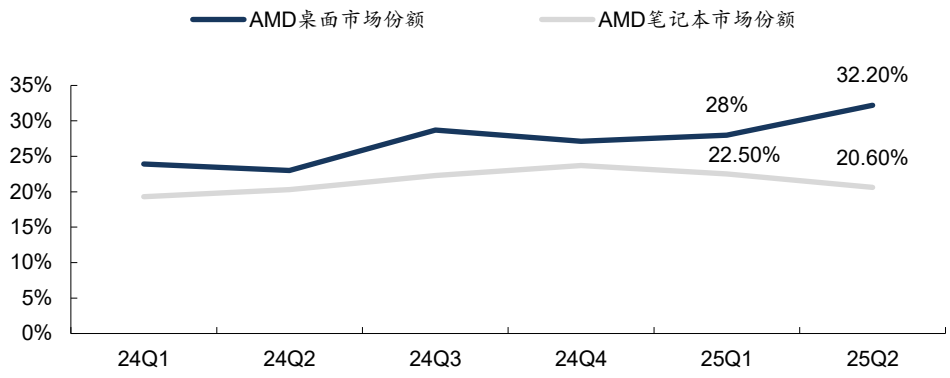

数据来源：英特尔，东吴证券研究所

在 AI 引擎的对比测试中，Panther Lake 都展现出了一定优势，英特尔在产品性能

端有望重返市场前列。

**AMD 在台式机 PC 市场份额创下新高。**2025 年二季度里，在消费类 PC 处理器市场，AMD 的营收占比增长至 26.5%，同比增长 9.8 个百分点。尤其是在台式机市场，相比 24 年同期增长 9.2 个百分点，市场份额占比达到 32.2%，位居历史最高点。与英特尔的销售比由过去的 9：1 提升至 2：1 左右，销售额增长明显。

**图22：AMD 桌面与笔记本 CPU 市场份额**




数据来源：EXPreview，东吴证券研究所

**AMD 在 AI PC 领域的竞争优势，源于其 Zen 架构带来的显著性能提升。**Zen 架构带来的显著 IPC 提升，使 AMD 扭转过去完全无法在高性能处理市场与英特尔竞争的市场格局。作为一种可持续改进的可扩展结构，从“Zen”到“Zen5”架构，AMD 不断进行制程工艺的改进和架构设计的优化，不断提升单线程性能。借助“Zen”架构，AMD 可为消费者和商业客户不断打造具有非凡性能、可扩展性和高能效的 AI PC 处理器。

**AMD 发布全新锐龙 AI 400 系列处理器，全面扩展其在客户端、图形和软件领域的 AI 领先地位。**锐龙 AI 400 系列基于先进的“Zen 5”架构打造，并搭载采用第二代 AMD XDNA 2 架构的 NPU，NPU AI 算力至高可达 60 TOPS，全面超越 Windows 11 AI + PC 的 TOPS 性能要求，实现无缝流畅的本地 AI 计算体验。凭借最高 12 个高性能 CPU 核心，集成 AMD Radeon 800M 系列显卡，并支持更高的内存频率，处理器可在多种系统与设备形态中释放领先性能、实现超长电池续航，并提供更智能的计算性能。

## 2.3 AI NAS 方案：存算一体破解核心痛点，全场景渗透驱动 AI NAS 迈入规模化落地的关键窗口期

**AI NAS 凭借“本地智能+高效存储”的存算一体架构，精准解决 AI 基础设施中存储缺位与数据利用效率低的核心痛点，同时以隐私安全、成本可控、场景适配多重优势，实现从消费级到企业级的全面渗透，成为 AI 时代连接数据与生产力的关键枢纽。**AI NAS 的核心变革在于彻底打破传统 NAS“数据仓库”的单一属性，通过“存储模块+AI 算力

模块+智能调度模块”的架构升级，构建起“存算一体”的闭环系统，其性能指标已从单纯的读写速度与容量扩展至算力水平、模型支持能力等核心维度，从根源上重塑了存储设备的价值逻辑。在数据处理全流程中，AI NAS 在数据采集环节支持多终端智能同步与自动分类，解决了传统 NAS 手动操作的低效问题；在数据检索环节突破文件名限制，实现文本、语音、图片的多模态语义级检索，大幅提升数据查找效率；在数据处理环节内置 AI 工具，达成“存储即处理”的即时性价值输出；在数据共享环节则将原始文件转化为可直接调用的智能服务，远超传统 NAS 的文件共享功能。此外，AI NAS 通过“内置模型+开源兼容+SDK 支持”的灵活模式实现模型自由，既适配英特尔 AI SDK 等基础模型开发的场景化应用，又支持 Ollama、llama.cpp 等开源框架，允许用户自行部署 Llama 3、Mistral 等大模型，满足不同层级的智能需求。

**AI NAS 已形成个人家庭、中小企业、垂直行业三大核心应用场景，需求从消费级向企业级持续扩张，市场进入规模化落地的关键窗口期。**个人家庭场景中，AI NAS 凭借易用性突破与个人媒体数据加速增长的需求，成为家庭智能计算中心，飞牛 OS AI 相册凭借本地人脸识别数据保护积累 80 万高频用户，绿联 AI NAS 在日本市场 50 岁以上用户占比达 32%，远超传统 NAS 的 15%。中小企业场景下，AI NAS 精准解决算力成本高、IT 人才缺的困境，通过构建统一知识库、支撑 AI Agent 任务执行、作为 AI PC 算力后端等功能，成为企业 IT 架构的新组成部分，铁威马 T 系列 AI NAS 2024 年销量同比增长 180%，核心客户覆盖电商、制造、律所等多个领域。垂直行业中，医疗、制造、安防等领域已实现从定制化方案到标准化产品的跨越，某县级医院采用 AI NAS 存储百万级 CT 影像，本地肺结节检测模型将阅片时间从 30 分钟缩短至 5 分钟，灵敏度达 95%，有效应对了全球医疗影像数据量的加速渗透。

**绿联科技凭借消费级 AI NAS 市场的领先地位、AI NAS 旗舰产品的软硬件抢先配置以及场景生态的持续拓展，稳固确立了其高端 AI NAS 赛道的领军地位。**绿联科技作为“NAS 第一股”，已在消费级 NAS 市场建立稳固的领先地位。2025 年 618 大促期间拿下京东平台超 50%市占率，斩获天猫、京东、抖音三平台销量与销售额双冠王，并获弗若斯特沙利文“消费级 NAS 产品全国销量第一”认证，为 AI NAS 的落地奠定了坚实的市场基础。在 AI NAS 赛道，公司以 iDX6011 系列为核心旗舰，首次亮相后的 14 天内在 Kickstarter 平台筹集超 660 万美元，创下该平台 NAS 产品众筹金额纪录，印证了市场对其“AI+NAS”形态的高度认可。硬件配置上，iDX6011 Pro 搭载英特尔酷睿 Ultra 7 255H 处理器，AI 总算力达 96 TOPS，内置 NPU 神经处理单元，可高效执行本地 AI 推理任务；最高支持 64GB LPDDR5x 内存与 196TB 存储容量，满足大模型运行与海量数据存储需求；配备 OCuLink 接口支持外接独立显卡，双 10GbE 网口与双雷电 4 接口保障数据传输无瓶颈，构建了基础算力强劲、扩展潜力充足的硬件架构。软件层面，Uliya AI 助手深度融入系统底层，提供全语义搜索、离线对话、智能相册、语音备忘录转录、自动文件整理等多元化功能，实现了从文件存储到智能交互的全流程覆盖，同时支持接入 DeepSeek、Qwen 等第三方开源模型，形成开放演进的软件生态。场景拓展方面，公司与

腾讯游戏 MTGPA 团队达成合作，实现《王者荣耀》等游戏的局域网高速下载，推动“NAS+游戏”生态构建，未来还将向家庭健康数据分析等场景延伸，持续丰富 AI NAS 的应用边界。

# 3. 汽车电子的“算力军备竞赛”是端侧 AI 的第二增长极

国产车载算力赛道正迎来由“局部替代”向“全局引领”跨越的战略质变期。智驾芯片领域的竞争已从单纯的硬件参数博弈演进为系统级生态较量；座舱芯片则依托先进制程与集成化架构，加速推动端侧大模型向主流价位段普及。伴随“舱驾一体”与单芯片中央计算趋势的加速落地，国产厂商正凭借“智驾平权”的规模红利与软硬协同优势，深度重构全球智能汽车的底层算力架构，并有望在这一进程中实现向全球主流供应链体系的实质性突破。

## 3.1 智能座舱 SoC 面临一芯多屏与消费电子的降维打击

### 3.1.1 座舱 SoC 芯片正在经历从“功能机”到“智能机”的存量替代

**智能汽车座舱 SoC 芯片进入产品换代周期，面向 AI 的座舱 SoC 将成为未来 2-3 年主流。高性能 SoC 取代传统 MCU，是支撑座舱向“第三空间”演进的核心硬件基础。**随着汽车 E/E 架构向集中化演进，座舱功能从基础信息显示向多屏交互、智能感知与沉浸式体验深化，传统以低算力 MCU 为核心的座舱控制器在应对高并发数据处理、复杂图形渲染及实时 AI 计算时面临显著算力瓶颈。智能座舱 SoC 需高算力支撑，其核心指标已从单一主频向异构计算架构延伸，主要包括 CPU 算力、GPU 渲染能力以及 NPU AI 算力。

**图23：15-25 万标配智能座舱自主品牌新能源车交付量及其 SoC 芯片**

| 可比特征 | MCU | SoC |
|---|---|---|
| 主频范围 | 16MHz-300MHz | 1GHz-3GHz+ |
| 典型功耗 | 微瓦级（电池供电数月） | 瓦级（需主动散热） |
| 内存容量 | <10MB | >1GB |
| 典型操作系统 | FreeRTOS | Linux、Android、Windows |
| 处理能力 | 专一化、适合简单控制 | 多核 CPU、GPU 和专用加速器，支持复杂计算 |
| 安全性与可靠性 | 设计用于在严苛的环境中长期稳定运行 | 由于集成度高，设计和验证更为复杂，需要额外的安全机制来确保系统的可靠运行 |
| 成本 | 便宜 | 较高 |
| 典型应用场景 | 需求明确且固定；成本敏感项目；超低功耗要求等 | 多任务并发处理；高性能计算需求；复杂协议支持等 |

数据来源：智研咨询，东吴证券研究所

**座舱 SoC 芯片作为“座舱大脑”，其技术迭代能力直接决定车型智能体验的上限。**主流芯片制程正从 7nm 向 4nm 及以下节点快速演进。截至 2024 年底，7nm 及以下制程芯片市场占比已达到 36%，并在 2025 年通过座舱域控制器的大规模前装实现了渗透率的进一步跃升，据高工智能汽车研究院预计到 2030 年该占比将突破 65%。下一代产品将持续向 4nm、3nm 工艺升级，相较于当前主流的 7nm 和 5nm 芯片，4nm 制程在晶体管密度、计

算性能和功耗控制等方面实现显著提升，能够更高效地支撑 AI 座舱在多样化场景下的高吞吐量、持续并发的 AI 计算需求。

**GPU 性能是实现沉浸式视觉体验的关键。**随着车载屏幕数量、分辨率及 3D 渲染需求提升，GPU 算力亦呈指数级增长。以高通为例，从 SA6155P 到 SA8295P，GPU 算力由 430GFLOPS 跃升至 3000GFLOPS，增幅近 6 倍。

**图24：座舱芯片的性能分级**

| | 型号 | 制程工艺 | CPU 算力（kDMIPS） | GPU 算力（GFLOPS） |
|---|---|---|---|---|
| 入门 | 高通 6155(SA6155P) | 11nm | 40 | 430 |
| | 瑞芯微 RK3588 | 8nm | 93 | 450 |
| 主流 | 高通 8155(SA8155P) | 7nm | 105 | 1142 |
| 旗舰 | 高通 8295(SA8295P) | 5nm | 220 | 3000 |
| | 芯驰科技 X10 | 4nm | 200 | 1800 |
| | 三星 Exynos Auto V920 | 5nm | 250 | 1420 |
| | 联发科 MT8676 | 4nm | 250-300 | 1050 |
| | AMD Ryzen V2000 | 7nm | 394 | 1433 |

数据来源：高通，AMD，瑞芯微，半导体行业观察，佐思汽研，facetop 智能汽车，AutoLab，电车通，东吴证券研究所

**能座舱市场近年来呈现出快速增长态势，市场规模和增速均表现突出。**全球范围内，智能座舱市场规模从 2021 年的 331.6 亿美元提升至 2024 年的 706.3 亿美元，期间年复合增长率达 28.66%。据预测，2025 年全球市场规模将增至 797.7 亿美元，并有望在 2030 年达到 1484.1 亿美元，显示出持续强劲的增长潜力。

**中国市场在智能座舱领域的发展尤为显著。**市场规模由 2021 年的 76.3 亿美元增长至 2024 年的 173.8 亿美元，年复合增长率达 31.58%，领先于全球市场整体增速（28.66%）。展望未来，中国市场规模预计将在 2030 年进一步提升至 548.1 亿美元，期间年复合增长率预计保持在 21.14%的较高水平。

**图25：全球智能座舱市场规模与预测**

| 年份 | 市场规模（亿美元） | 增长率（%） |
|---|---|---|
| 2021 | 331.6 | - |
| 2022 | 426.5 | 28.6% |
| 2023 | 548.9 | 28.7% |
| 2024 | 706.3 | 28.7% |
| 2025E | 797.7 | 12.9% |
| 2030E | 1484.1 | 13.2% |

数据来源：汽车半导体情报局，东吴证券研究所

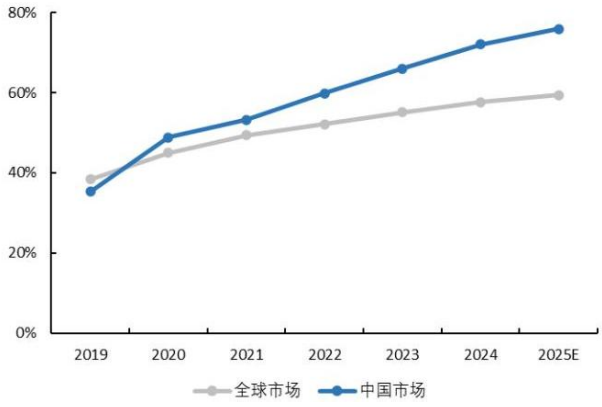
**图26：中国智能座舱市场规模与预测**

| 年份 | 市场规模（亿美元） | 增长率（%） | 中国占比 |
|---|---|---|---|
| 2021 | 76.3 | - | 23.0% |
| 2022 | 99.8 | 30.8% | 23.4% |
| 2023 | 128.7 | 29.0% | 23.5% |
| 2024 | 173.8 | 35.0% | 24.6% |
| 2025E | 210.1 | 20.9% | 26.3% |
| 2030E | 548.1 | 21.1% | 36.9% |

数据来源：汽车半导体情报局，东吴证券研究所

**在消费升级趋势深化、用户乘车体验需求提升以及消费电子产品应用场景持续向车载领域延伸的背景下，智能座舱技术加速市场渗透。**据普华有策咨询统计，自 2020 年起，中国新车市场的智能座舱渗透率已领先于全球平均水平。2025 年，中国市场渗透率达 53%；至 2026 年，这一比例将预计突破 80%，显著高于同期全球市场约 59%的渗透水平。基于当前发展态势，未来我国智能座舱市场增速有望持续超越全球均值，为本土汽车智能座舱领域的企业创造更广阔的发展机遇与市场空间。

**图27：国内及全球乘用车智能座舱渗透率**




数据来源：HIS Markit，博泰车联公司公告，东吴证券研究所

**现如今，智能座舱不再是高端车的专属，而是迅速向 10 万级平民车型普及，呈现出“配置多样、全面普及”的加速渗透态势。**2025 年 1-9 月的行业数据显示，智能座舱的核心部件正加速成为新车“标配”：中控屏渗透率接近 95%，几乎每辆新车都配备了智能交互入口；语音交互与车联网渗透率高达 85%，改变了“开车靠手调、联网靠手机”的状况。同时，数字钥匙、座舱域控、HUD（抬头显示）、DMS（驾驶员监测系统）等进阶配置的渗透率也在不断攀升，提升了智能座舱的“含金量”。

**图28：我国乘用车智能座舱核心产品/功能渗透率**


数据来源：盖世汽车研究院智能座舱配置数据库，盖世汽车研究院分析，东吴证券研究所

**智能座舱 SoC 芯片向下沉车市场渗透。**2024 年中国乘用车前装座舱域控制器搭载量达 673.19 万辆，搭载率从 2023 年的 17.56%提升至 29.37%。尽管 25-30 万元和 50 万元以上价位车型仍是标配主力（搭载率超 70%），但 10-25 万元价格区间呈现显著增长，域控搭载率从 2022 年的 9.01%跃升至 28.42%，增长达 2.58 倍。该价位车型占整体市场约 58%，而当前域控渗透率仍仅 28.42%，在 AI 技术持续赋能和智能座舱进一步下沉的推动下，座舱域控制器及 SoC 芯片市场预计将迎来持续扩张。

**图29：2025 年 15-25 万标配智能座舱自主品牌新能源车交付量及其 SoC 芯片**

| 车型 | 座舱域控芯片 | 交付量（辆） |
|---|---|---|
| 理想 L6 | 高通 8295 | 121,892 |
| 深蓝 S07 | 高通 8155 | 75,289 |
| 银河 L7 | 高通 8155 | 73,923 |
| 银河 E5 | 芯擎(龍鷹一号) | 73,522 |
| 小米 SU7 | 高通 8295 | 69,422 |
| 领克 08 | 芯擎(龍鷹一号) | 67,394 |
| 零跑 C10 | 高通 8295/8155 | 66,989 |
| 哈弗猛龙 | 高通 8155 | 63,037 |
| 零跑 C11 | 高通 8295/8155 | 58,333 |
| 银河 L6 | 高通 8155 | 56,562 |

数据来源：观研天下，东吴证券研究所

**国内燃油车智能座舱渗透率已进入稳步提升阶段。**随着产品竞争由动力性能向体验维度延伸，智能座舱正成为传统燃油车提升用户感知度与增强产品竞争力的核心抓手。

**图30：国内乘用车市场燃油车智能座舱搭载量及渗透率**


数据来源：盖世汽车研究院智能座舱配置数据库，盖世汽车研究院分析，东吴证券研究所

**随着汽车智能化进程加速，智能座舱芯片正经历一场价值跃迁。**其价值中枢正从传统汽车中单价仅几十元人民币的微控制器（MCU），快速跃升至高达数百甚至上千美元的高性能系统级芯片（SoC）。价格带全面上移，国产芯片具备性价比。主流智能车型的 SoC 采购价已普遍进入 100-500 美元区间，相较于传统 MCU 实现了数十倍的价值跃升。来自中国本土的芯擎科技、芯驰科技等厂商的座舱芯片解决方案日益成熟，在激烈的市场竞争下，厂商开始推动芯片向更低价格区间的车型渗透。

**图31：芯片价格**

| 芯片类型 | 价格（美元） | 市场 |
|---|---|---|
| 高通骁龙 8295 | 150 | 30 万元以上，下探至 15-20 万元市场 |
| 高通骁龙 8397 | 240-270（预计） | 中高端车型 |
| 英伟达 Orin | 300-500 | 高端车型 |
| 芯驰 X10 | 100-150 | 中高端车型 |
| 黑芝麻智能武当 C1296 | 80-120 | 15-30 万元价格区间车型 |

数据来源：facetop 智能汽车，东吴证券研究所

### 3.1.2 从手机芯片“魔改”到算力芯片下场“降维打击”再到国产芯片强势崛起

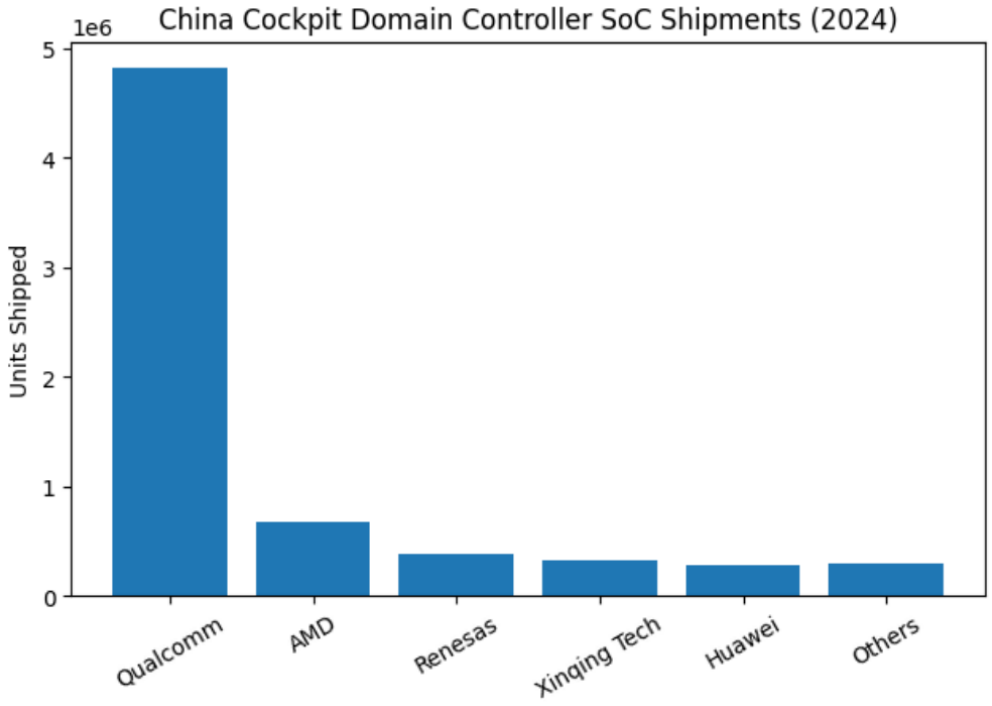
**从市场份额来看，海外巨头仍占据主导地位，其中高通凭借骁龙数字底盘生态以绝对优势领跑。**2025 年 1-11 月，国内乘用车座舱域控 SoC TOP10 的总安装量超过 1000 万颗，其中高通骁龙 8155 以超 370 万颗的安装量位居榜首，占比达 31.6%，依旧是市场最受欢迎的座舱芯片之一。高通旗舰芯片骁龙 8295 表现同样亮眼，安装量为 198.0 万颗，同比增长 112.8%，主要受奔驰、理想、小米、零跑等品牌车型拉动。

图32：2025 年 1-11 月国内座舱域控 SoC 供应商排名（按标配安装量）

| 排名 | 供应商 | 安装量（万颗） | 2025.1-11 市占率 | 2024.1-11 市占率 |
|---|---|---|---|---|
| 1 | 高通 | 871.7 | 74.4% | 73.2% |
| 2 | 海思 | 67.2 | 5.7% | 6.0% |
| 3 | AMD | 54.0 | 4.6% | 7.6% |
| 4 | 芯擎科技 | 51.8 | 4.4% | 3.8% |
| 5 | 联发科 | 36.6 | 3.1% | 1.1% |
| 6 | 瑞芯微 | 30.7 | 2.6% | 0.6% |
| 7 | 亿咖通 | 11.0 | 0.9% | 1.0% |
| 8 | 瑞萨 | 9.4 | 0.8% | 1.6% |
| 9 | 三星 | 7.7 | 0.7% | 2.1% |
| 10 | 芯驰 | 7.5 | 0.6% | 0.6% |

数据来源：观研天下，东吴证券研究所

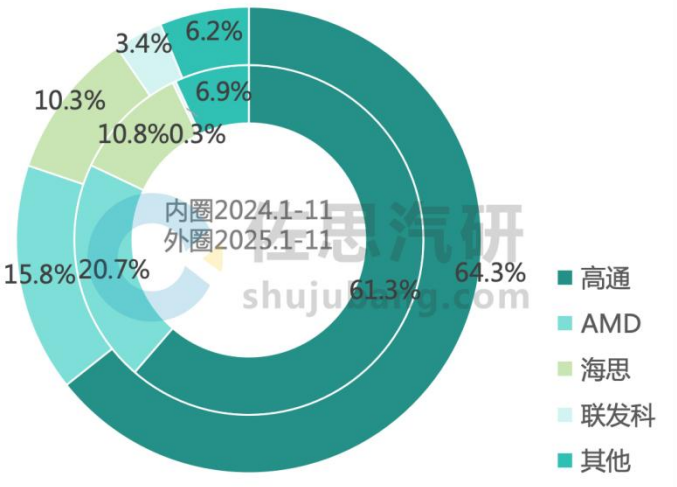
图33：座舱 SoC 在中国市场安装量（百万颗）




数据来源：facetop 智能汽车，东吴证券研究所

**在 20-35 万元的中高端市场区间，高通、联发科稳步增长。**2025 年 1-11 月，高通凭借 8155/8295 芯片，在该区间赢得 64.3%的份额。AMD 以 15.8%的市占率位居第二，但份额较上年同期有所下滑；海思以 10.3%的市占率排名第三，份额较上半年同期微幅下降；联发科则以 3.4% 的市占率排名第四，较上年同期增加 3.1 个百分点。

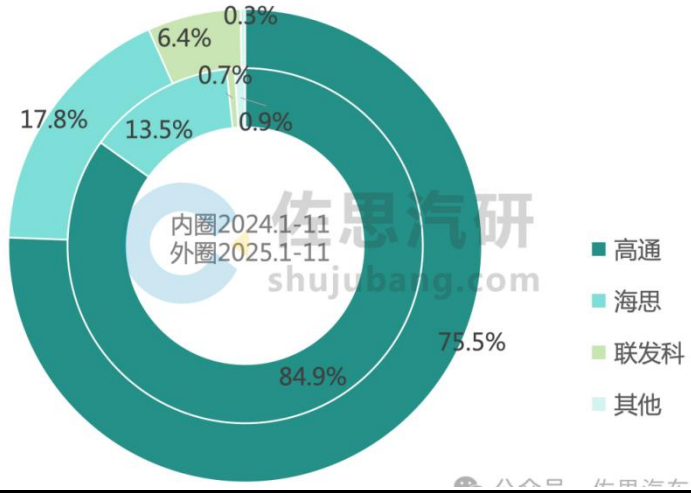
图34：2025 年 1-11 月中国 20-35 万元区间智能汽车座舱域控 SoC 市场供应商格局




数据来源：佐思汽研，东吴证券研究所

**在 35 万元以上的高端市场，高通市占率下降。**相反，搭载海思芯片的鸿蒙座舱系列车型销量高涨，2025 年 1-11 月其市占率为 17.8%，较上年同期增加 4.3 个百分点。

图35：2025 年 1-11 月中国 35 万元以上智能汽车座舱域控 SoC 市场供应商格局




数据来源：佐思汽研，东吴证券研究所

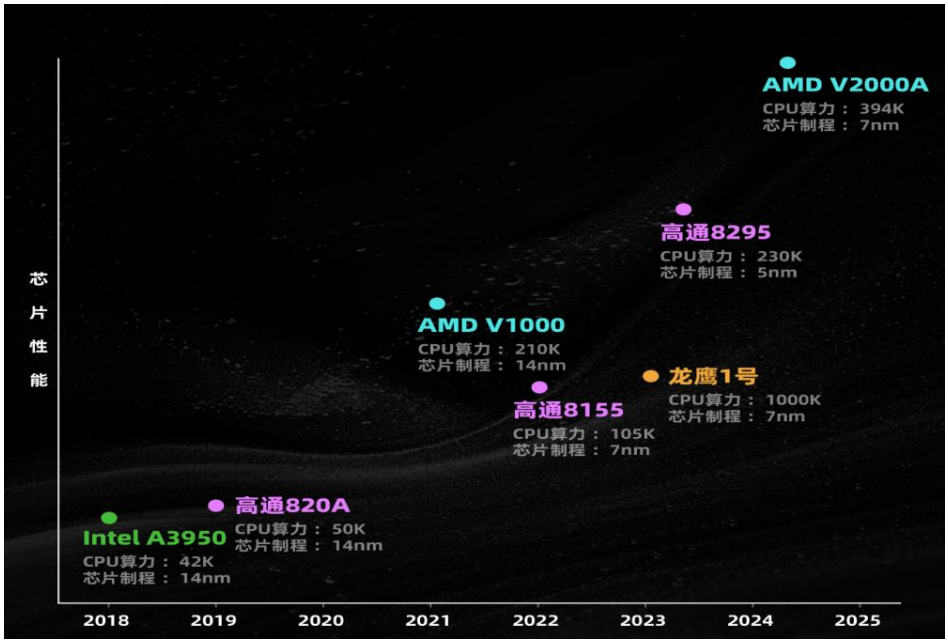
**高通在座舱领域的迅速崛起，很大程度得益于在智能手机领域积累的底层技术。**在高通进入市场前，汽车座舱电子生态处于“前智能时代”，其市场主要由恩智浦、瑞萨、德州仪器等传统汽车芯片供应商主导。这些厂商的产品多采用 28nm 及以上成熟制程，CPU、GPU 算力低，难以支撑复杂的智能交互与多任务处理需求。这套体系在功能车时代游刃有余，但无法支撑起智能车所要求的复杂多任务处理、高清多屏显示与流畅人机交互。

**在高端智能座舱 SoC 市场中，已形成以高通为引领、多家芯片厂商紧随其后的“一强多极”竞争格局。**随着智能座舱市场迈向千亿规模，在车载 AI 大模型部署、高品质

3A 游戏上车等高算力、高性能需求的驱动下，包括英伟达、英特尔、联发科及 AMD 在内的各家芯片供应商正加速技术布局与产品迭代，力图在快速扩张的市场中提升自身份额，推动行业竞争向更深层次演进。

**通过特斯拉进入座舱市场的 PC 芯片巨头 AMD 也在持续发力。**在游戏性能等高负载场景下，当前基于 AMD 架构的芯片展现出显著优于行业主流方案的算力表现。值得关注的是，当前多数智能座舱应用基于安卓系统开发并主要适配 ARM 架构芯片，而 AMD 推出的 V1000 与 V2000A 系列芯片均采用 X86 架构。2024 年初，AMD 正式发布车规级嵌入式处理器 V2000A 系列。该芯片采用 7nm 制程，搭载“Zen 2”CPU 核心和 AMD Radeon Vega 7 显卡，支持虚拟机管理程序以增强功能安全和车载软件兼容性，可同时运行汽车级 Linux 和 Android Automative 系统。V2000A 系列主打高性能图形与游戏处理能力，契合智能座舱日益提升的沉浸式娱乐需求，具备较高的性能与成本优势。性能方面，V2000A 系列 CPU 算力约 360-370kDMIPS，较上一代 V1000 系列提升约 88%，优于当前主流的高通的 SA8295P 芯片。产品落地方面，特斯拉已在部分车型中率先导入 AMD 锐龙 V1000 系列消费级芯片；而领克 Z10 车型实现了 AMD 锐龙 V2000A 系列车规级芯片的全球首次量产搭载，标志着 X86 架构在智能座舱领域向车规级、高性能方向迈出关键一步。

**图36：智能座舱芯片 SoC 性能**




数据来源：AutoLab，东吴证券研究所

**尽管 AMD 在性能维度表现亮眼，但其市场份额与高通仍存在一定差距。**2025 年 1-10 月，高通智能座舱域控 SoC 安装量 871.7 万颗，市占率 74.4%，较上年同期市占率 73.2%有所上升；AMD 智能座舱域控 SoC 安装量 54.0 万颗，市占率 4.6%，较上年同期市

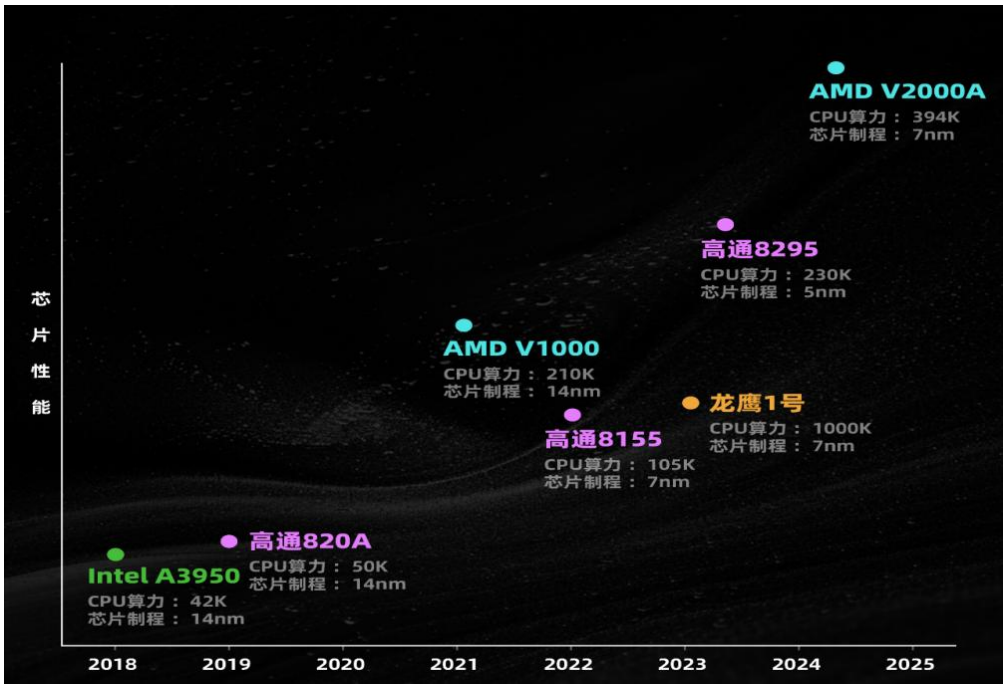
占率 7.6%有所下降。高通自带开发生态优势：基于 ARM 架构打造的芯片，接过了高通在手机行业对安卓系统长期适配的经验，相较于其他汽车半导体，其开发生态更加庞大，开发友好型更高。而 AMD 采用 X86 架构，与主流车机安卓生态整合需额外工作，在软件生态方面需要花更多精力进行开发和适配。

**图37：AMD V2000A 和高通 8155/8295 对比表**

| 对比维度 | AMD V2000 | 高通 8155/8295 |
|---|---|---|
| 性能 | CPU 算力：394kDMIPS<br>GPU 算力：1433GFLOPS | CPU 算力：105/230kDMIPS<br>GPU 算力：1142/3000GFLOPS<br>NPU 算力：8/30TOPS |
| 架构 | X86，在软件生态方面需要花更多精力进行开发和适配 | ARM，功耗低 |
| 发布时间 | 2024 年 1 月 4 日 | 2019 年/2021 年 1 月 |

数据来源：Autolab，facetop 智能汽车，东吴证券研究所

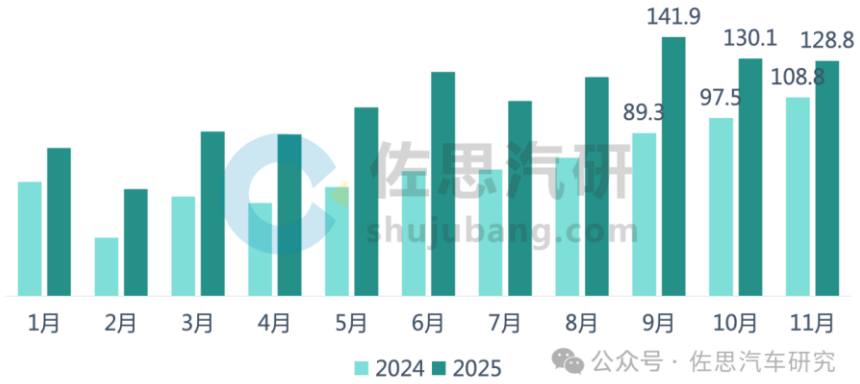
**图38：智能座舱芯片 SoC 性能**




数据来源：AutoLab，东吴证券研究所

**国产芯片 SoC 强势崛起，座舱芯片平台加速迭代升级。**中国智能汽车座舱 SoC 市场中，虽然高通、瑞萨、AMD 等厂商仍然占据主导地位，但同时国产化率也正在快速提升。佐思汽研数据显示，2025 年 1-11 月国内乘用车新车座舱域控 SoC 标配安装量达到 1171.9 万颗，同比增长 52.7%，其中 9-10 月单月均超过 125 万颗；全年总安装量将预计超 1300 万颗。

**图39：2025 年 1-11 月中国座舱域控 SoC 安装量（万颗）**




数据来源：佐思汽研，东吴证券研究所

**国产芯片加速渗透，多款芯片跻身榜单前列。**2025 年 1-11 月国内 TOP10 国产座舱域控 SoC（按标配安装量排名），芯擎龙鹰一号闯入 TOP6，安装量逼近 50 万颗，同比增长 65.8%；瑞芯微 RK3588M 跃居第八，安装量突破 30 万颗，同比增幅高达 533.5%；比亚迪 D9000 入围 TOP9，安装量较上年同期增长 2638.6%。

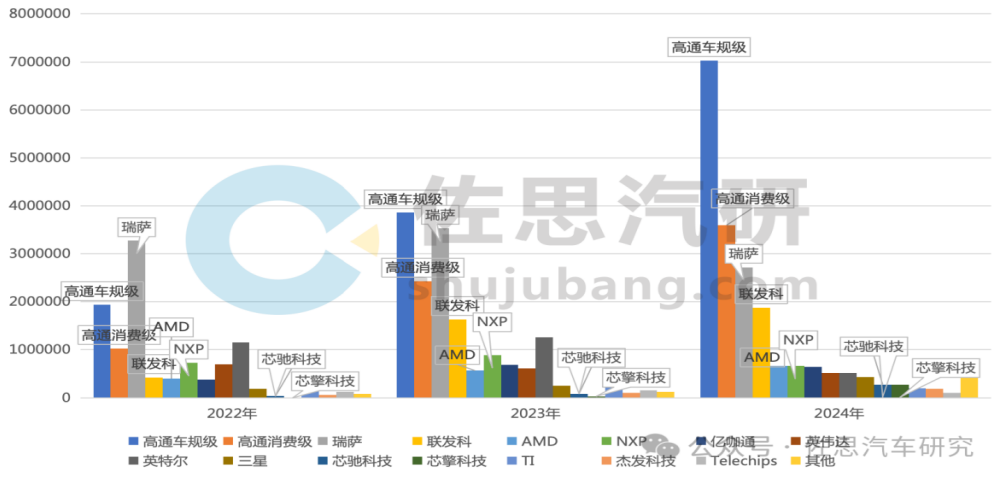
**图40：2025 年 1-11 月国内 TOP10 国产座舱域控 SoC（按标配安装量）**

| 排名 | 芯片 | 安装量（万颗） | 同比增长 | 市占率 |
|---|---|---|---|---|
| 6 | 芯擎龙鹰一号 | 48.1 | 65.80% | 4.10% |
| 7 | 海思麒麟 990A | 46.9 | 92.80% | 4.00% |
| 8 | 瑞芯微 RK3588M | 30.7 | 533.50% | 2.60% |
| 9 | 比亚迪 D9000 | 26.2 | 2638.60% | 2.20% |

数据来源：佐思汽车研究，东吴证券研究所

国产化率提速明显。根据佐思汽研统计，2025 年智能座舱 SoC 国产化率达到 18%，较上年同期相比，提升 4.9 个百分点。以芯驰科技、华为海思、芯擎科技等为代表的国产厂商正快速崛起。

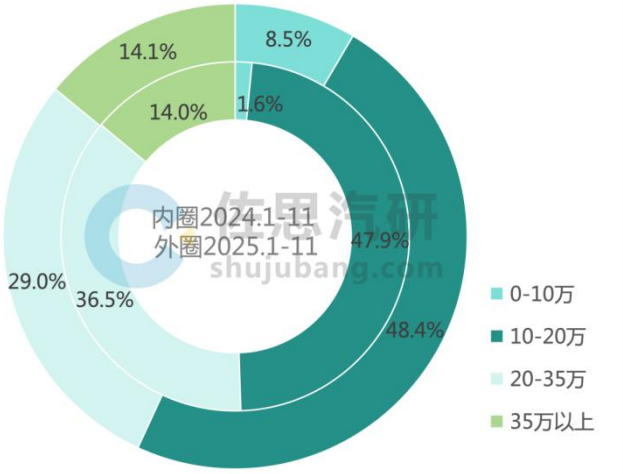
**图41：2022-2024 年中国智能汽车座舱 SoC 出货量趋势**




数据来源：佐思汽研，东吴证券研究所

**国产芯片加速渗透中低端市场。**从价格区间分布来看，2025 年 1-11 月国内座舱域控 SoC 集中安装在 10-20 万元价格区间乘用车上，占比为 48.4%；其次是 20-35 万区间，占比为 29.0%；35 万以上和 10 万以下区间占比相对偏低，分别为 14.1%和 8.5%。

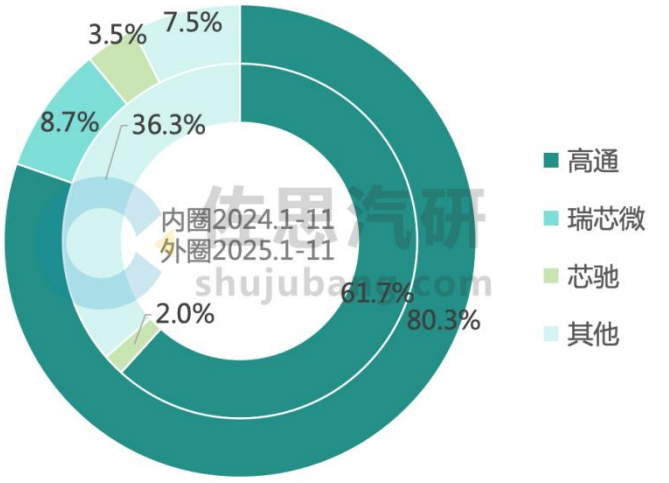
**图42：2025 年 1-11 月中国智能汽车座舱域控 SoC 安装量（分价格区间）**




数据来源：佐思汽研，东吴证券研究所

**瑞芯微深耕消费电子与汽车电子芯片领域，凭借高性价比方案在中端座舱芯片市场占据重要份额，产品以多屏适配能力强、兼容性高为核心优势。**在 10 万元以下的入门级市场，本土供应商积极突围，瑞芯微表现最为突出。其通过推动 RK3588M 芯片规模化量产落地，成功拿下 8.7% 的市场份额，成为该区间国产芯片的核心代表。

**图43：2025 年 1-11 月中国 10 万元以下智能汽车座舱域控 SoC 市场供应商格局**




数据来源：佐思汽研，东吴证券研究所

**瑞芯微以高性能车载计算平台为核心，展出一系列基于车规级智能座舱方案。**囊括了 RK3588M、RK3576M、RK3568M、RK3358M 的多场景乘/商用车应用，涵盖 3D 液晶仪表、智能中控、座舱域控、副驾娱乐、吸顶屏、移动扶手 PAD、头枕、AVM、VR 等。对于高性能系列，RK3588M 芯片作为当前主力量产产品，采用成熟的 8nm 制程工艺，集成 6TOPS NPU 算力，已成功应用于数十款车型，成为行业“舱泊一体”高性价比解决方案的核心驱动力。其搭载双 RK3588M 的座舱娱乐域控备受关注，提供强大的性能支撑和扩展能力，最高支持 9 个屏，除中控仪表外，还支持头枕屏、吸顶屏、扶手屏、AR 眼镜等应用。此外，头枕接入了 800w 像素摄像头并集成 HDR 和多帧降噪拍照算法，支持多品类游戏机 DP in 并享受低至 50ms 的游戏延迟，PCIe3.0 芯片级联更能拓展更多应用场景。

**图44：瑞芯微座舱 SoC 芯片**

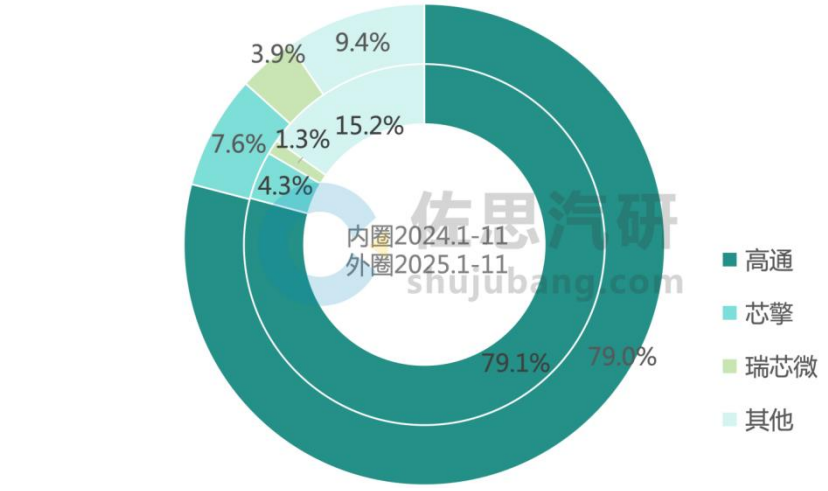
| 型号 | | 制程工艺 | 主要特性 |
|---|---|---|---|
| 高性能 | RK3588M | 8nm | 6TOPs NPU，赋能各类 AI 场景；内置多种显示接口，支持多屏异显；超强影像处理能力， 48MP ISP，支持多摄像头输入；支持 Android 和 Linux OS。 |
| | RK3576M | 12nm | 丰富的显示接口，高效率 GPU 处理器，支持 3 个显示屏异显（每屏内容不同）；强大的影像感知、视频编解码及音频处理能力，整合视觉和语音识别交互；标准 Android、Linux SDK 支持，适配各类国产 OS。 |
| 全国产化 | RK3568M | 22nm | NPU 支持 1T 算力；支持多屏异显。 |
| 入门级 | RK3358M | 12nm | 支持 1.2 TOPS NPU；应用于车载信息娱乐系统：中控屏、后排娱乐终端。 |

数据来源：瑞芯微电子官网，facetop 智能汽车，万联芯城，中积芯，东吴证券研究所

而规划中的 RK3688M 芯片则将制程提升至 4nm，NPU 算力大幅跃升至 32TOPS，旨在支持 Transformer 大模型的端侧运行。该系列芯片适配中控、仪表、娱乐屏、HUD、DLP、视觉等应用，IP 和软件的良好继承性可以进一步降低客户的开发投入，支持端侧大模型、轻量级 L2 和 APA，并可提供 QNX、RAITE、硬隔离、开源、自研 Type-I 等多系统方案。全国产化 SoC RK3568 系列：应用于带环视的中控或小域控。凭借成熟的 8nm 工艺和 6TOPS NPU 算力，以高性价比推动“舱泊一体”方案的规模化上车，目前已应用于数十款车型，并支持端侧大模型实现多模态融合交互。入门级座舱芯片 RK3358M 系列：应用于中控、扶手屏、头枕等独立组件。

**在 10-20 万元市场区间，芯擎、瑞芯微、联发科等本土企业持续发力，市场份额稳步提升。**其中芯擎同比提升 3.3 个百分点，瑞芯微同比提升 2.6 个百分点，国产化替代趋势进一步凸显。

图45：2025 年 1-11 月 10-20 万元区间智能汽车座舱域控 SoC 市场供应商格局




数据来源：佐思汽研，东吴证券研究所

**芯擎科技主要以“龙鹰一号”撬动市场。**该公司专注于高端车规级芯片研发，填补了国产高端车规处理器的空白。其核心产品龙鹰一号是国内首款量产上市的 7nm 车规智能座舱 SoC，于 2021 年底发布，2023 年实现量产。量产后该芯片迅速放量，2024 年出货量突破百万量级。截至 2025 年 8 月，其累计出货量已突破 150 万片，搭载于包括领克、吉利银河、一汽红旗、沃尔沃、长安启源等品牌的数十款量产车型中，2024 年稳居国产座舱芯片市场占有率第一。此外，该芯片于 2024 年底成为首个获得德国大众海外车型定点的国产座舱芯片。芯擎的成功源于精准的技术卡位与差异化策略。“龙鹰一号”芯片采用多核异构架构，内置 8 核 CPU、14 核 GPU 及 8 TOPS 算力的独立 NPU，支持 7 屏高清输出和 12 路视频接入，能满足多模态交互、3D 游戏等复杂场景需求，例如一汽红旗天工 05 就采用双“龙鹰一号”解决方案打造高端座舱。更关键的是，芯擎创新性推出“舱行泊一体”集成方案。通过双 NPU 设计，在单芯片上融合智能座舱、行车辅助与泊车控制功能，能够为车企节省 20%-30%的成本，较好地契合当前车企对高性能与成本控制的双重需求。市场反馈方面，搭载龙鹰一号“舱泊一体”解决方案的银河 E5 上市不到 1 个月，销量即突破 1 万辆。

**作为本土车规芯片的领军企业，芯驰科技已经在座舱芯片激烈的市场竞争中脱颖而出。**盖世汽车研究院最新数据显示，2025 年 1-11 月，芯驰科技装机量 178357 颗，市场份额 2.1%，排名第六。芯驰科技成立于 2018 年，彼时的国内汽车市场上，电动化、智能化的行业大势已经初显端倪，对车规级芯片的需求量，迅速井喷，市场潜力巨大。只不过，当时的车规级芯片市场，被国际芯片巨头强势垄断，他们凭借先发优势和已经形成的技术壁垒，控制着全球车规芯片市场绝大部分的份额，筑起了一道看似难以逾越的高墙。2019 年芯驰科技完成了中国第一颗 16nm 车规 SoC 芯片流片。此后几年时间里，随着中国汽车智能化进程的不断提速，本土芯片企业也实现了快速崛起，以芯驰科技为代表的本土企业通过技术创新和精准的定位，逐步打破了外资企业的市场垄断。芯驰科

技的一系列产品迅速量产落地，并大规模上车装载，成为近年来，国产车规主控芯片落地与量产速度最快的公司之一，尤其是在座舱领域，芯驰科技更是建立起了全面领跑的优势。截至 4 月 23 日上海车展，芯驰科技宣布，目前全系列产品累计出货量已突破 800 万片，搭载超 100 款主流量产车型，持续领跑智能座舱与智能车控量产领域。

**图46：2025 年 1-11 月座舱域控芯片供应商装机量排行**

| 排名 | 供应商 | 装机量 | 市场份额 |
|---|---|---|---|
| 1 | 高通 | 6369359 | 73.3% |
| 2 | 华为技术 | 587865 | 6.8% |
| 3 | 超威半导体 | 541602 | 6.2% |
| 4 | 芯擎科技 | 483304 | 5.6% |
| 5 | 瑞萨电子 | 195367 | 2.2% |
| 6 | 芯驰科技 | 178357 | 2.1% |
| 7 | 三星半导体 | 103827 | 1.2% |
| 8 | 联发科 | 96722 | 1.1% |
| 9 | 德州仪器 | 74352 | 0.9% |
| 10 | 英特尔 | 17181 | 0.2% |

数据来源：盖世汽车社区，东吴证券研究所

在 X9 系列智能座舱产品数百万片量产交付的基础上，芯驰以 X10 卓越的性能、创新的架构以及丰富的 AI 生态，率先引领座舱处理器的 AI 变革，打造出全民 AI 时代座舱处理器新标杆。2025 年 4 月 23 日，芯驰科技在上海国际车展期间重磅发布最新一代 AI 座舱芯片 X10。芯驰 X10 系列产品采用专为 AI 计算优化的 ARMv9.2 CPU 架构，CPU 性能高达 200K DMIPS；同时，X10 还集成 1800 GFLOPS GPU 和 40 TOPS NPU，并配置了高达 128-bit 的 LPDDR5X 内存接口，速度达到 9600 MT/s，为整个系统提供 154 GB/s 的超大带宽。制程工艺方面，X10 产品采用 4nm 先进制程，相较于当前主流高端车规芯片常用的 7nm/5nm 制程，4nm 在晶体管密度、性能、功耗控制上都有明显的提升，可更好地支持 AI 座舱在不同应用场景下的高吞吐量、持续运行的 AI 计算任务，确保产品在整个生命周期中保持领先性。开放多元的 AI 生态，加速落地应用。芯驰科技正围绕 X10 构建开放、多元的 AI 生态系统。X10 不仅可以支持 Deepseek、Qwen、Llama 等开源大模型，也将持续与斑马智行、面壁智能等生态合作伙伴完成车载 AI 大模型的深度适配。同时，针对车厂自研大模型，芯驰也可提供软硬件协同优化支持。

### 3.1.3 手车互联与操作系统壁垒成为智能座舱系统新趋势

**手机已经深度融入车内，打破设备间的壁垒。**随着汽车智能化的发展，手车互联的内置率在不断地提升，手车互联已经成为了大多数前装车型的标配，2027 年的内置率预计将超过七成，即年新增搭载量将会超过 2000 万台。在手车互联的基础上，车企逐步拓展与智能家居、无人机、平板等多终端产品的互联，构建更为完整的智能生态体系。这样的拓展过程不仅为用户提供了更加丰富的功能和服务，还加强了主机厂在智能生态领域的竞争力。

**图47：手车互联主要方式**

| 技术类型 | 手机投屏 | App 互联 | 远程控制 |
| --- | --- | --- | --- |
| 涉及技术 | 通过将手机屏幕内容投射到汽车的中控显示屏上实现手机与车机的互动。常见的投屏技术包括 Carplay、Carlife、HiCar、Android Auto、ICCOA Carlink 等 | 主要是针对导航、音乐等类型的第三方 APP，通过开发专门的车机版本或者适配接口，实现手机与车机之间的互联 | 主要依托车企开发的手机 APP，通过网络与车辆的车载通信模块（如 T-BOX）进行连接，实现对车辆的远程控制。 |
| 技术实现方案 | 由手机厂商主导开发和推进 | 由第三方 APP 开发者主导，进行应用的适配和功能开发 | 由车企自主研发和部署 |
| 功能特点 | 侧重于将手机内容无缝迁移到车机，提供一致的用户体验。允许用户在车机上操作手机应用，如导航、音乐播放、接打电话等。提供与手机一致的用户体验，减少用户的学习成本。依赖手机性能和网络连接。 | 通过第三方应用扩展了车机的功能。提供多样化的应用选择以满足用户的不同需求。可以通过车机或手机进行操作，灵活性高。需要车机和手机都安装相应的应用程序。 | 通过车企提供的 APP，实现对车辆的远程管理和控制。提供车辆远程控制功能提升用户便利性。可以实时监控车辆状态如电量、位置等。增强车辆安全性，如远程锁车、防盗报警等。 |

数据来源：佐思汽车研究，东吴证券研究所

**小米座舱系统有着强大的应用生态。**基于小米集团强大的生态能力，小米智能座舱带来了更强的生态体验。车机不仅深度适配行业主流车载应用，小米平板应用生态也可无缝上车，5000+款小米平板应用将逐步适配车机，体验达标才会上车，轻松移植。人车家全生态互联能力进阶。小米 Pad 本身和车机大屏一样，具备查看同账号手机镜像画面的能力。为了让后排乘客可以享受智能的后排屏幕控制同时又能保护车主个人隐私，小米为 Pad 带来了新的车上隐私保护功能。同样的镜像升级也应用在了手车互联中，手机画面镜像到车上后，当车主离车，车内其他乘客便无法通过车内的屏幕查看到车主的手机画面。基于人车家全生态打造的小米澎湃 OS，有着最先进的手车互联体验。连接丝滑，手机进入座舱，手机 dock 栏就会自动浮现“手机图标”，点击即可一键呼出妙享桌面，将手机镜像到车机，实现屏幕共享。此外，手机应用可通过“Pin 到车上”功能与车载系统无缝衔接，实现手机与车机生态的深度融合。该技术将移动端应用一键投射至车辆中控大屏，使其作为原生车机应用运行。例如，视频类应用可在大屏呈现全屏播放界面，且手机与车机操作相互独立——用户可在车机端持续观看视频的同时，于手机端进行其他操作，互不干扰，真正实现手机车机生态融合共享。

**多种地址流转方式，全方位满足高频用户需求。**导航作为车主高频刚需功能，小米

深入挖掘其在不同场景下的需求，打造了五种地址流转方案，无论是小米手机还是苹果手机用户，都能轻松将地址信息发送至车机系统，实现无缝导航体验。针对小米手机用户，在小红书、大众点评、微博等含有地址信息的页面，只需长按电源键唤醒小爱同学，说出“把这个地址发到车上”，即可将地址发送至车机。对于苹果手机用户，当用户在小红书等平台上搜索网红打卡地时，若笔记内容无法直接复制，只需将页面截图发送到小米汽车 APP，系统即可识别截图中的地点信息并发送到车机，方便快捷。

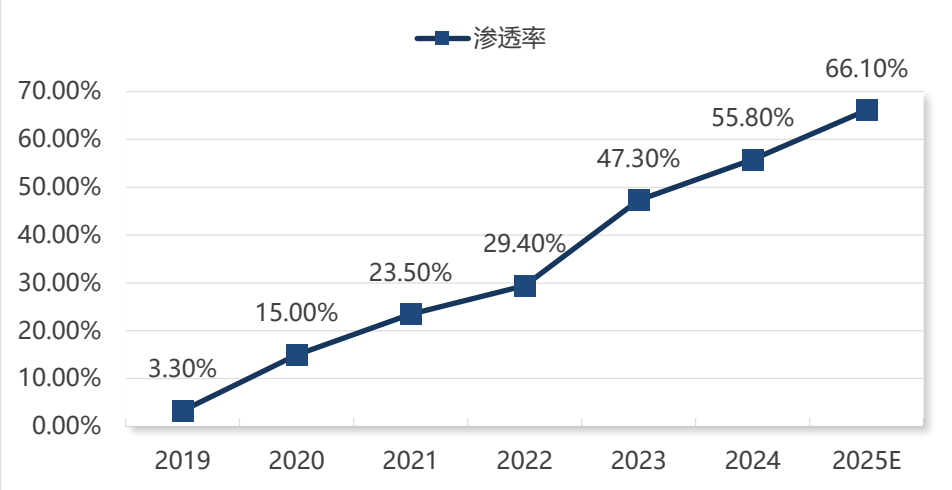
**华为终端为手车互联产业的发展壮大贡献了一份力量。**借助 HUAWEI HiCar 分布式技术能力，手机和车机通过分布式软总线互联后将会形成一个“超级终端”，让手机与车机可以互相调用彼此的硬件能力，从而实现无缝连接，实现用户体验的统一。截至 2026 年 1 月，华为 HiCar 已经搭载超过 500 款车，不仅国内车企，奥迪、大众、丰田、本田等一众合资车企也纷纷搭载，而且还有更多的车企陆续接入中。

## 3.2 自动驾驶 SoC 正在经历从 L2 到 L4 的算力跃迁

### 3.2.1 从“感知智能”向“世界模型”跨越

**L2+（行泊一体）产业规模快速提升。**智能网联汽车是汽车与人工智能、信息通信、云计算、大数据等技术融合发展的关键载体，已成为全球汽车转型升级的重要战略方向。工业和信息化部高度重视智能网联汽车产业发展，着力推进技术攻关、标准研制、应用推广，推动产业发展取得积极成效。2026 年 1 月 14 日，中国汽车工业协会汽车行业信息发布会上正式发布《2025 城市 NOA 汽车辅助驾驶研究报告》。数据显示，2025 年前三季度具备组合驾驶辅助功能（L2）的乘用车新车销量同比增加 21.2%，渗透率达 64%，预计 2025 年年底升至 66.1%，三分之二左右的新车型将搭载 L2 级辅助驾驶功能。

**图48：历年中国乘用车 L2 及以上渗透率**

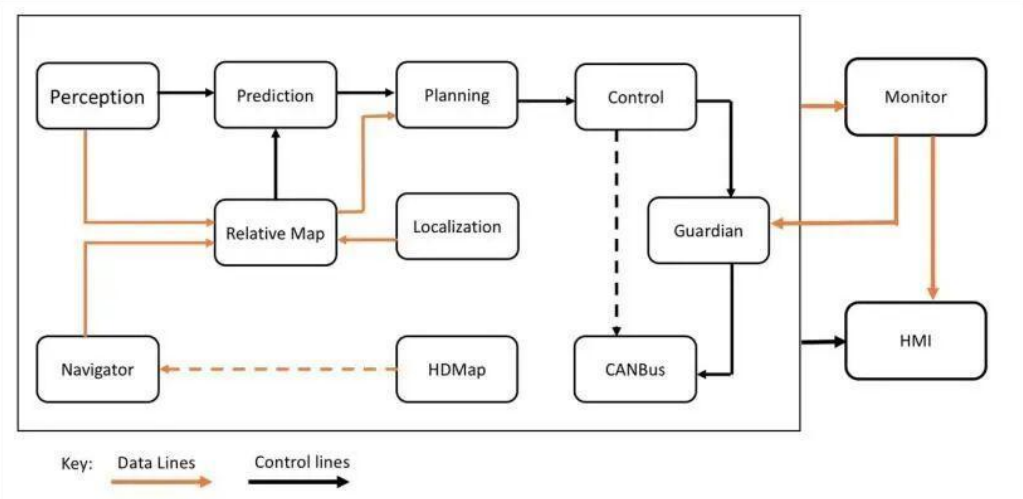



数据来源：光锥智能，东吴证券研究所

**L3/L4（城市 NOA）与世界模型在加速落地。**过去十年间，汽车行业正经历一场由技术驱动的深刻变革——从早期基于规则的辅助驾驶系统，逐步演进至当前以大模型为核

心驱动力的智能驾驶阶段。这一转型不仅表现为软件算法能力的指数级提升，更显著拉动了对车载芯片在算力规模、架构设计及能效水平等方面的系统性需求升级。规则驱动范式衍生于传统的机器人架构，由定位、感知、预测、规划、控制等子模块组成。该架构具备模块化程度高、可解释性强、问题易于追踪等优势，但由于其依赖人工规则制定，在系统拓展性与长尾问题覆盖方面存在明显瓶颈。面对高速、城区、泊车等多场景及跨区域部署需求时，往往需投入大量人力进行适配调整，难以实现高效规模化。

**图49：规则驱动范式**

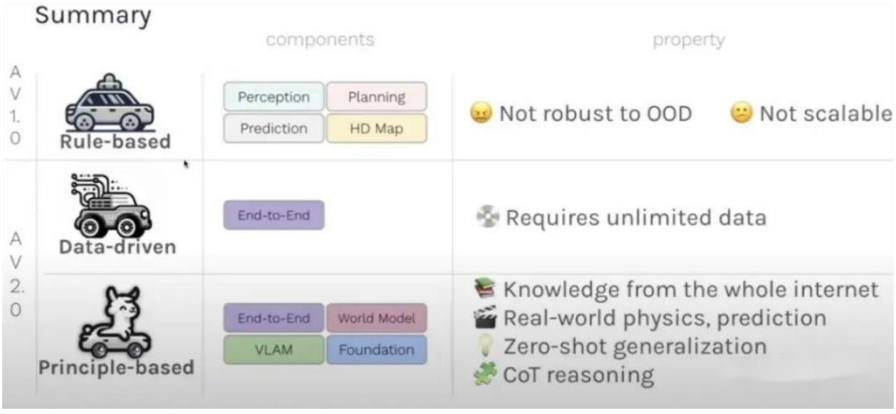



数据来源：深蓝 AI，东吴证券研究所

**端到端大模型已成为 NOA 辅助驾驶技术迭代的核心引擎。**通过构建统一的大模型，直接处理来自摄像头、激光雷达等多源传感器的原始数据，并输出车辆控制指令（如方向盘转角、油门及刹车信号）。该架构摒弃了传统模块化技术路径中感知、预测、规划等独立环节，理论上有利于减少信息传递过程中的误差累积与系统延迟。然而，该方案仍面临显著挑战，主要表现为模型决策过程可解释性不足，且在训练数据未覆盖的长尾场景中，系统行为存在不确定性风险。以华为乾崑智驾采用的全新一代架构 WEWA（World Engine+World Action Model）为例，这是典型的生成式端到端，通过“云端世界引擎+车端世界行为模型”的协同模式，可高效输出车控轨迹并有效提升长尾场景的适配能力。

**世界模型技术路径的核心在于“先推演后决策”的仿真机制**，通过构建具备环境动态预测能力的内部模型，系统能够学习并推断未来数秒内交通参与者（如车辆、行人）的行为轨迹及传感器观测状态变化。基于此，车辆可在虚拟仿真空间中对多种可能场景进行推演，并从中选择最优驾驶策略。该方案在复杂动态交互场景中具有突出优势，且支持在仿真环境中进行高效、低风险的迭代训练与验证。然而，该范式对模型预测精度与实时计算能力均有极高要求，若预测出现偏差，将直接影响决策的安全性与可靠性。

**图50：技术范式革命**



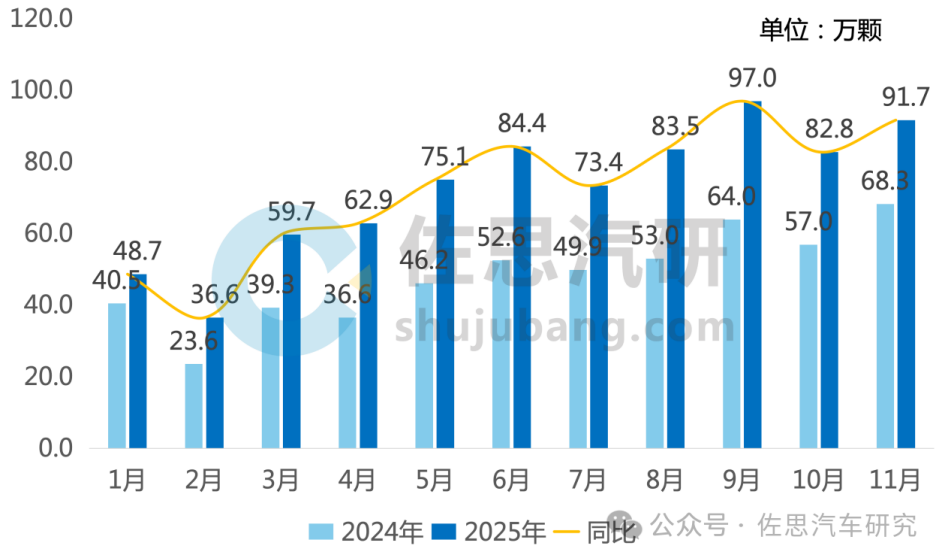

数据来源：深蓝 AI，东吴证券研究所

**近两年，NOA 技术市场规模快速扩容，高速 NOA 技术已进入成熟期，行业差异化竞争与战略布局的重点已转向城市 NOA。**2025 年作为中国城市 NOA 商业化落地的关键一年，市场规模快速增长。2025 年 1—11 月，我国搭载城市 NOA 功能的乘用车累计销量达 312.9 万辆，渗透率占乘用车上险量的 15.1%，较 2024 年全年提升 5.6 个百分点。起售价 30 万元以下的主流乘用车搭载城市 NOA 车型的销量占比超 68.9%，标志着高阶智驾已经脱离“奢侈品”身份。

### 3.2.2 智驾域控 SoC 市场规模高速扩张，本土“芯”势力加速多价位渗透

**智驾域控 SoC 安装量呈上涨趋势。**2025 年 1-11 月，中国乘用车新车整体智驾域控 SoC 安装量达 796.0 万颗，同比增长 49.9%。其中，自主品牌智驾域控 SoC 安装量达 601.1 万颗，同比增长 83.1%，主要受比亚迪、小鹏、小米汽车等拉动。合资品牌智驾域控 SoC 安装量达 194.9 万颗，同比下滑 3.8%，主要搭载品牌包括大众、丰田、日产等。

**图51：2025 年 1-11 月中国乘用车新车智驾域控 SoC 安装量**




数据来源：佐思汽研，东吴证券研究所

**智驾芯片市占率稳步提升。**2025 年 1-11 月，从智驾域控 SoC 供应商安装量排名情况看，英伟达和特斯拉占主要市场份额，合计达 63.5%。国内智驾域控 SoC 厂商看，其整体市占率达 20.7%，较上年同期增长 4.6 个百分点，主要受海思、地平线、黑芝麻智能等拉动。

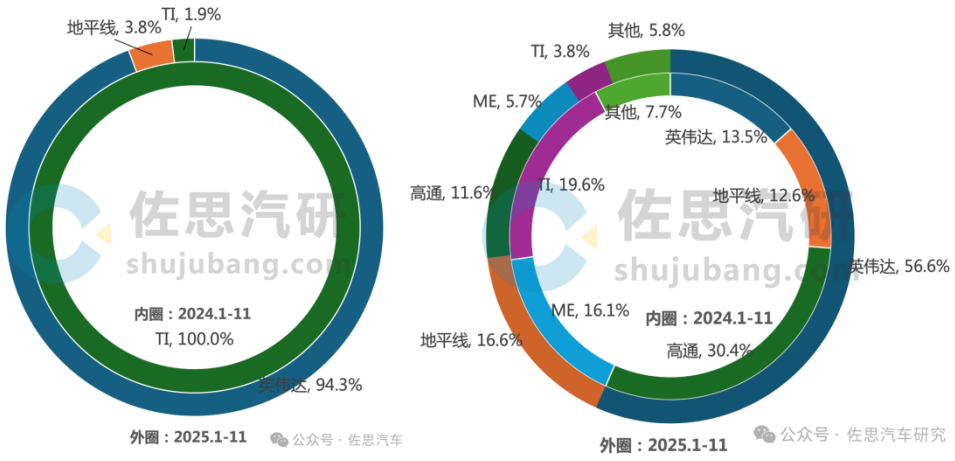
**图52：2025 年 1-11 月中国乘用车新车智驾域控 SoC 供应商 TOP10 竞争格局（按安装量）**

| 排名 | 供应商 | 装配量（万颗） | 市占率 |
|---|---|---|---|
| 1 | 英伟达 | 399.0 | 50.1% |
| 2 | 特斯拉 | 106.4 | 13.4% |
| 3 | 海思 | 76.5 | 9.6% |
| 4 | 地平线 | 64.5 | 8.1% |
| 5 | Mobileye | 46.8 | 5.9% |
| 6 | 高通 | 31.5 | 4.0% |
| 7 | 德州仪器 | 31.1 | 3.9% |
| 8 | altera | 16.1 | 2.0% |
| 9 | 黑芝麻智能 | 10.9 | 1.4% |
| 10 | 蔚来 | 8.5 | 1.1% |

数据来源：佐思汽车研究，东吴证券研究所

**0-10 万元市场英伟达绝对领跑，10-20 万元市场地平线逐渐上升。**0-10 万元区间，英伟达占主要份额，市占率为 94.3%。10-20 万元区间，地平线市占由上年同期 12.6%增至 16.6%。2026 年，地平线基于单征程 6M 的城区辅助驾驶方案将量产上车，该方案首批量产合作伙伴包括博世、卓驭、轻舟智航，及电装、酷睿程、智驾大陆 neueHCT 等。此外，地平线最新征程 7 系列芯片也将于 2026 年量产，该系列将搭载地平线第四代 BPU 架构—“黎曼”。黎曼可支持多芯片级联，未来将为 L4 自动驾驶及工业机器人提供千 TOPS 级算力支持。

**图53：2025 年 1-11 月 0-10、10-20 万元中国乘用车新车智驾域控 SoC 市场供应商格局**

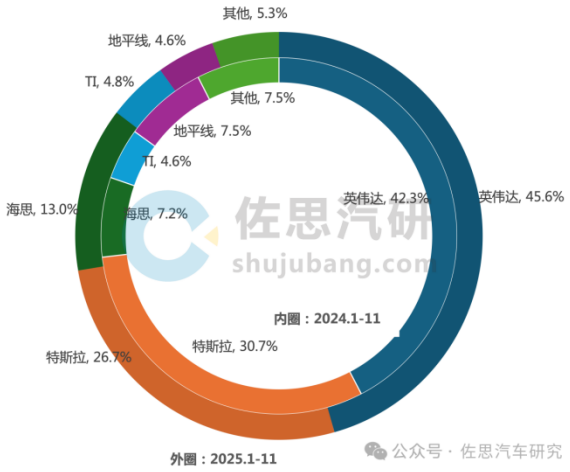



数据来源：佐思汽研，东吴证券研究所

**2025 年 1-11 月，在 20-40 万元价格区间的市场中，英伟达、海思及德州仪器（TI）**

**位列前五大供应商之列，且三者的市场份额均呈现上升态势。**海思市占 13.0%，较上年同期增长 5.8 个百分点。TI 市占 4.8%，较上年同期增长 0.1 个百分点。CES2026 上，TI 推出了采用 5nm 制程的 TDA5 SoC 系列。该系列产品可提供 10-1200 TOPS 的可配置边缘 AI 算力，能效比超 24 TOPS/W，引入 Chiplet（芯粒）设计，支持标准 UCIe 接口，覆盖从 L2 + 辅助驾驶到 L3 级自动驾驶需求。

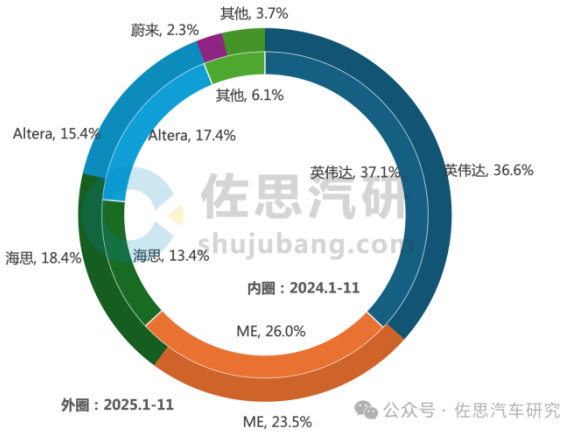
**图54：2025 年 1-11 月 20-40 万元中国乘用车新车智驾域控 SoC 市场供应商格局**




数据来源：佐思汽研，东吴证券研究所

**2025 年 1-11 月，40 万元以上高端市场，英伟达、Mobileye、海思市占分居前三。**其中 Mobileye 市占 23.5%，较上年同期下滑 2.5 个百分点。CES2026 上，Mobileye 展示了 EyeQ7（2027 年量产）与 Nvidia Orin-X 的实测对比，EyeQ7 的卷积神经网络 ResNet 50 和 ViT（900 万参数）推理延迟均低至 0.5ms，效率达 Orin-X 的 3 倍。

**图55：2025 年 1-11 月 20-40 万元中国乘用车新车智驾域控 SoC 市场供应商格局**




数据来源：佐思汽研，东吴证券研究所

### 3.2.3 智驾芯片格局重塑，英伟达筑起技术高墙，本土力量凭“智驾平权”与生态解构加速突围

**英伟达凭借 CUDA 生态和对 Transformer 的先发支持，垄断了高端市场 NOA 市场。**在 2025 年国际消费电子展 CES 2025 上，英伟达重磅发布了其最新一代车规级自动驾驶

芯片“Thor”，并同步展示了在智能汽车领域的技术进展与合作布局 DRIVE Thor 提供 1000 TFLOPS 的加速计算性能，以更高精度加速推理任务，例如模式识别、适应恶劣天气等。

DRIVE Thor 平台集成了先进的多域计算架构，能够将自动驾驶、智能座舱、车载信息娱乐等多个核心功能域进行硬件隔离与任务分区，实现并发运行且互不干扰。其多计算域隔离机制确保了对时间敏感的关键进程（如实时自动驾驶控制）在独立、确定性的环境中持续执行，同时支持非实时功能（如娱乐系统）的灵活运行。基于这一架构，该平台可实现在同一套硬件上同时部署并高效运行包括 Linux、QNX 和 Android 在内的多种操作系统，满足不同功能域在实时性、安全性与生态兼容性方面的差异化需求。在 CES 上，英伟达、Aurora 和大陆集团宣布建立长期战略合作关系，共同大规模部署由下一代 NVIDIA DRIVE Thor 系统级芯片驱动的自动驾驶卡车。NVIDIA 的 DRIVE Thor 和 DriveOS 将被集成到大陆集团 SAE L4 级自动驾驶系统——Aurora Driver 中，计划于 2027 年大规模量产。

**2025 年 1-11 月，Momenta 城市 NOA 搭载率高达 41.44 万辆，占第三方供应商比例约 61.06%，领跑行业。**在技术层面，算法是驱动 Momenta 持续发展的核心动能，其技术架构具备显著优势。一是端到端的深度学习框架，实现了感知与规划模块在统一模型中的深度融合。该框架基于数据驱动的闭环体系，能够高效完成模型迭代，并具备低成本、高迭代速度及规模化数据回流能力，为算法优化提供持续的数据支撑。二是依托闭环自动化工具链，构建了以数据流为驱动的算法自动演进机制。该工具链能够自动筛选、标注并回流“极端场景”等高价值数据，形成“数据-算法-性能”的增强循环，实现系统迭代效率的持续提升。三是其独特的“长短时记忆”学习结构，进一步强化了算法演进效能。短期记忆模块专注于实时响应与动态适应，长期记忆模块则系统沉淀最优决策策略，转化为系统的“本能反应”，从而显著提升算法学习效率与稳定性。

**图56：Momenta 辅助驾驶软件产品**


数据来源：Momenta 官网，东吴证券研究所

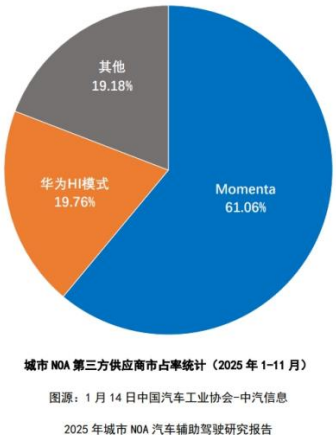
**各车企与供应商纷纷加大端到端技术布局，形成多元化竞争格局。**Momenta 以“强化学习+端到端架构”为核心技术路线，通过量产辅助驾驶(Mass Production)与自动驾驶(Scalable Robo)双轨并进的产品战略，依托海量数据闭环实现算法持续迭代。目前，Momenta 已与上汽、广汽、比亚迪、奔驰、宝马、奥迪、丰田等众多主流车企达成合作，

Momenta R6 强化学习大模型在行业中率先实现量产落地，已于 2025 年第三季度正式量产上车。

**高阶智驾成本逐渐降低。**2022 年，一套典型城市 NOA 级智驾的硬件成本约为 2 万元。据 Momenta 全球解决方案首席架构师饶庆预测，到 2026 年之后，高性价比的城市 NOA 方案成本有望进一步降至 4000-5000 元水平，从而推动智能辅助驾驶技术向大众市场加速普及。

**第三方城市 NOA 供应商的“双强”市场格局已经显现。**数据显示，聚焦第三方城市 NOA 供应商领域，Momenta 和华为 Hi 模式两者合计占第三方供应商比例约八成，并显著领先其他竞争者的市场份额。中国车企在高阶智驾供应商的选择上，达成了初步共识。

**图57：城市 NOA 第三方供应商市占率统计（2025 年 1-11 月）**




数据来源：光锥智能，东吴证券研究所

尽管 Momenta 已经把智驾方案拉低至了单 Orin-X 芯片加激光雷达的方案，但最便宜的城区 NOA 车型也依然要 15 万元以上。在中国市场，50%的销量来自于价格低于 13 万元的车型，还有 20%的销量价格带为 13-15 万元区间。换言之，只有占据市场约三分之一的高价位车型，才能享受到高算力芯片带来的城区 NOA。

**而在 2025 年地平线技术生态大会的主旨发言中，地平线创始人兼 CEO 余凯博士指出，地平线 HSD 在深蓝 L06、星途 ET5 上量产之后、接下来会把智驾带到 10 万元市场。**仅仅两周时间，这两款车型的智驾辅助激活量已达到 12000 辆，它标志着地平线从 ADAS 走向 L2+高速 NOA，现在终于入局城区辅助驾驶、城区 NOA。

征程 6B 是入门级 ADAS 主动安全的性价比芯片。拥有 10+TOPS 算力，配备车规级 CPU，达 20K + DMIPS。支持最新 NCAP 法规标准；高集成度，实现更小系统尺寸与更低系统功耗；还可支持功能扩展升级，涵盖 DMS、DVR 及数据回传等功能。而征程 6E/M 是针对中阶辅助驾驶市场对性能体验和系统成本差异化需求推出的芯片。同时，地平线进一步开放生态合作，协同生态伙伴打造面向中阶智驾量产的多元化解决方案，加速高速 NOA 及城区记忆行车等高频智驾应用普惠。

**图58：智驾芯片双雄对决表（高端 vs 性价比）**

| 对比对象 | | 单芯片算力(TOPS) | 主要特性 |
|---|---|---|---|
| NVIDIA Orin-X | | 254 | smart 精灵 #5<br>蔚来 ET7<br>长城汽车旗下豪华高端SUV——魏牌蓝山智驾版 |
| 地平线征程 5（J5） | | 128 | 理想 L6-9 Pro<br>比亚迪汉 EV 荣耀版 |
| 地平线征程 6（J6） | | | |
| | 征程 6P | 560 | 星途 ET5<br>iCAR V27<br>风云 T9L |
| | 征程 6E | 80 | 埃安霸王龙<br>名爵 MG4<br>荣威 M7 DMH |
| | 征程 6M | 128 | 深蓝 L06 |
| | 征程 6B | 18 | - |
| Momenta | | 254（估计） | 智己 LS6<br>别克至境 E7<br>风云 T11<br>奔驰纯电 CLA |
| 黑芝麻 | 华山 A1000 | 58 | 领克 08EM-P<br>合创 V09<br>东方奕派 007 |

数据来源：英伟达官网，地平线官网，黑芝麻智能官方公众号，东吴证券研究所

征程 6E 则聚焦高速 NOA 驾驶，其支持域控被动散热，推动高速 NOA 实现成本平价化和体验极致化（75℃开放无风环境 ）。芯片采用 BPU Nash，具备 80TOPS 算力，搭配车规级 CPU，达 100K DMIPS。Transformer 计算效率较上代产品提升 10 倍，拥有高能效比 FPS/Watt，且支持被动散热 。相较征程 6E，6M 定位普惠级城区场景，具备 128TOPS 算力，对比征程 6E 提升 60% ；车规级 CPU 达 137K DMIPS，对比征程 6E 提升 37% 。能支持轻量级城区 NOA 及记忆行车，还支持激光雷达接入。征程 6P 是针对全场景辅助驾驶应用推出的性能旗舰版方案。凭借强大的多核异构计算资源，征程 6P 能够全面发挥片上系统的计算性能，是支持当下先进端到端智能辅助驾驶路线的优选。该芯片采用 4 核 BPU®Nash ，具备 560TOPS 算力，搭配 18 核 ARM Cortex A78AE，达 410K DMIPS。拥有 200G FLOPS 算力，支持 3D 图像渲染，集成式全功能 MCU，具备 10K DMIPS 及 ASIL-D 算力。

MDC620 是华为承接旗舰技术，面向更广泛中高端市场的性能计算平台。它旨在以一个更具成本效益的方案，支持包括城市 NOA 在内的高阶智能驾驶功能，从而帮助合作车

企在 20-30 万元级别的主流市场建立智驾领域的竞争优势。MDC810 是华为面向 L4-L5 级别自动驾驶的旗舰计算平台。该平台不仅是华为前沿自动驾驶技术的展示，更是其全栈智能驾驶解决方案商业化落地的重要载体。在性能方面，MDC810 可提供高达 400 TOPS（INT8）的算力，能够支持多路高分辨率传感器（包括激光雷达）的实时数据处理，并可高效运行基于 BEV（鸟瞰图）的感知模型和 Transformer 的复杂算法。该平台主要面向高阶自动驾驶功能场景，如城市通勤 NOA、跨区域高速 NOA、AVP（代客泊车）等复杂应用。

**图59：华为 MDC**


| | MDC 300F | MDC 210 | MDC 610 | MDC 810 |
|---|---|---|---|---|
| AI 算力 | 64 Tops | 48 Tops | 160 Tops ~ 200 Tops | 400T ops |
| 接口 | Camera：12<br>车载以太：4<br>CAN/CAN FD:10 | Camera：8<br>车载以太：4<br>CAN/CAN FD:10 | Camera：16<br>车载以太：8<br>CAN/CAN FD:12 | Camera：16<br>车载以太：8<br>CAN/CAN FD:12 |
| 防尘防水 | IP52 | IP67 | IP67 | IP67 |
| 制冷 | 风冷 | 液冷/风冷 | 液冷/风冷 | 液冷/风冷 |

数据来源：facetop 智能汽车，汽车 ECU 开发公众号，东吴证券研究所

### 3.2.4 产业链协同：软件生态辅助重新定义汽车的“朋友圈”逻辑

**生态链条重构已成为智能驾驶 SoC 发展关键。**一方面，“芯片+软件栈+开发平台”一体化模式加速落地，头部企业通过整合工具链、算法适配能力，缩短车企开发周期，推动 “舱驾一体” 等技术方案普及；另一方面，国产化进程提速，国内厂商在智能驾驶 SoC 市场份额快速提升。同时，智驾成本下降（传感器、芯片迭代），车企自研成趋势，助力 10 万元级车型搭载高阶智驾，倒逼技术迭代与生态重构。

**自主车企“智驾平权”驱动智驾域控快速上量，合资车企跟进功能普及，进一步释放 Tier 1 方案增量空间。**2025 年以来，自主车企积极推进“智驾平权”，进一步下探高速 NOA 功能搭载车型的价格带，带动智驾域控搭载量快速成长，1-10 月渗透率已快速提升至 27.6%。为了兼顾系统性能、部署成本和量产落地速度，大部分自主车企将选择保留高效率的供应商方案，这为智驾 Tier 1 的短期增长提供了充足空间。同时，合资车企积极跟进智驾普及，同步推进“油电同智”，选择与实力强、量产经验丰富的方案商合作，释放出了具有潜力的增量市场。

**传统链条（芯片原厂->Tier 1->主机厂）瓦解，转向芯片原厂+算法 Tier 1.5+主机厂的三角合作。**博世凭借与星途星纪元 ES 的合作成为首个在中国市场落地城区 NOA

的国际 Tier-1，但其算力平台仍落后于华为 ADS 3.0 等本土方案；本土头部企业里，Momenta 以 60.1%的市场份额领跑，服务 130 款量产车型，轻舟智航则占据 NOA 市场 50.84%的份额；技术特色厂商方面，元戎启行专注 VLA 端到端模型，卓驭科技则以“成行平台”量产 17 款车型的业绩证明性价比路线的可行性。

**地平线客户粘性高，依赖大客户。**地平线还通过引入车企（上汽集团）作为资方，以及与车企成立合资公司（与大众集团成立酷睿程、与大陆集团成立大陆芯）等方式，进入对方的供应链中。不过，这种与车企的强绑定关系在给地平线带来大量订单的同时，也使得地平线更加依赖这些大客户。在 2022 年及 2023 年，上汽集团分别占地平线当期收入的 11.2%及 6.9%，位列当期的第二大客户及第四大客户。而在 2023 年和 2024 年，地平线与大众汽车的合资公司酷睿程已跃升为公司第一大客户。

**图60：地平线前五大客户产生收入（亿元）**

| 年份 | 前五大客户收入总额 | 占总收入比重 | 最大客户收入总额 | 占总收入比重 |
|---|---|---|---|---|
| 2021 | 2.83 | 60.7% | 1.15 | 24.7% |
| 2022 | 4.82 | 53.2% | 1.45 | 16.0% |
| 2023 | 10.67 | 68.8% | 6.27 | 40.4% |
| 2024 | 17.12 | 71.8% | 7.51 | 31.5% |

数据来源：新皮层 NewNewThing，东吴证券研究所

**图61：厂商 OEM 客户**

| 厂商 | 客户 |
|---|---|
| 地平线 Horizon | 广汽埃安、江淮汽车、广汽传祺、岚图汽车、理想汽车、哈弗、奇瑞汽车、上汽荣威、吉利汽车、领克、哪吒汽车、星途汽车、上汽大通、五菱汽车、北京汽车、启辰汽车、捷途汽车、iCAR、比亚迪、红旗、蔚来、深蓝汽车、均联智行 |
| 华为 MDC | 一汽奥迪、享界、阿维塔、尚界、问界、广汽传祺、深蓝汽车、猛士、上汽奥迪、岚图汽车、尊界、智界、方程豹 |

数据来源：地平线官网，华为官网，东吴证券研究所

## 3.3 智驾架构融合演进从“双脑”到“单脑”

**在传统汽车电子架构阶段，整车普遍采用高度分布式的电控单元（ECU）布局。**以高端车型为例，其电子系统通常由数十至上百个独立 ECU 模块组合而成，每个模块均搭载专用微控制器（MCU），并对应于一项明确且功能边界清晰的车辆控制任务——如发动机管理、车身控制、制动辅助或转向调节等，各系统间分工明确、互不干涉。在这一架构下，传统燃油车型平均需搭载约 70 颗 MCU 芯片，而新能源汽车因三电控制、能量管理及更多辅助功能的引入，其 MCU 需求进一步上升至 100-200 颗。每个 ECU 均基于独立

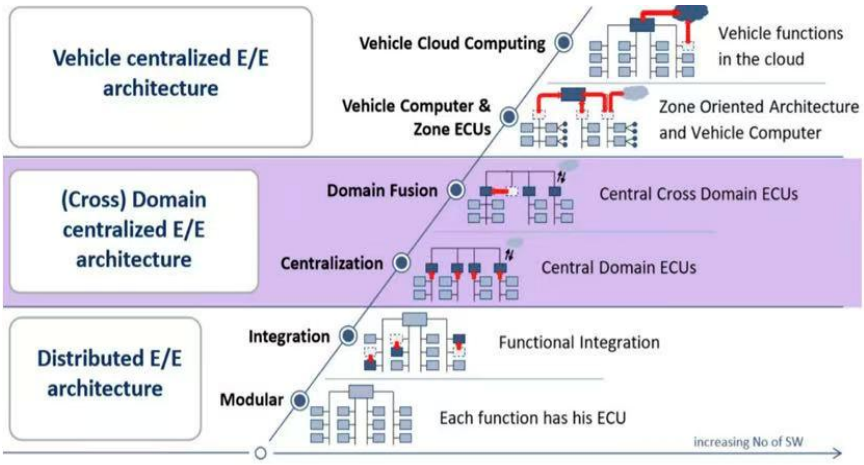
的微控制器运行，形成功能隔离、可靠稳定的嵌入式控制网络。

从市场格局看，汽车 MCU 长期呈现集中、稳定的产业生态。2020 年全球汽车 MCU 市场规模已达 60 亿美元，约占整体 MCU 市场份额的 40%。该领域主要由在汽车电子领域积淀深厚的传统芯片供应商主导，呈现出较高的行业壁垒与客户粘性。

随着智能驾驶、车联网和软件定义汽车（SDV）的加速渗透，传统分布式架构的算力冗余、通信瓶颈和开发效率问题愈发突出。架构的升级不仅是硬件拓扑的重新排列，更是整车设计逻辑的重构。由分布式架构演进成功能域架构。功能域集中式的集中化阶段引入车载以太网高速通讯，各域控制器之间采用以太网进行信息共享，打破了带宽的局限性。按照功能划分出不同的域，如底盘域、车身域、座舱域、辅助驾驶域等。每个域由一个域控制器（DCU）进行集中控制，负责处理本域内所有附属 ECUs 相关功能和数据运算。

**最新的架构引入了区域控制器和中央计算单元（ZCU+CC）。**区域控制器（ZCU）负责特定物理空间区域内的所有任务，同时运用中央计算单元（CCU）进行跨区域的集中运算和决策。不同车企划分车辆前、后、左、右等不同区域，可分别由相应 ZCU 管理，通过以太网实现中央与区域之间的可靠连接。

**图62：E/E 架构演进**




数据来源：facetop 智能汽车，东吴证券研究所

随着“中央+zonal”E/E 架构逐渐成为主流，计算架构进一步向中央计算机（CCU）发展演进，主要形式包括：

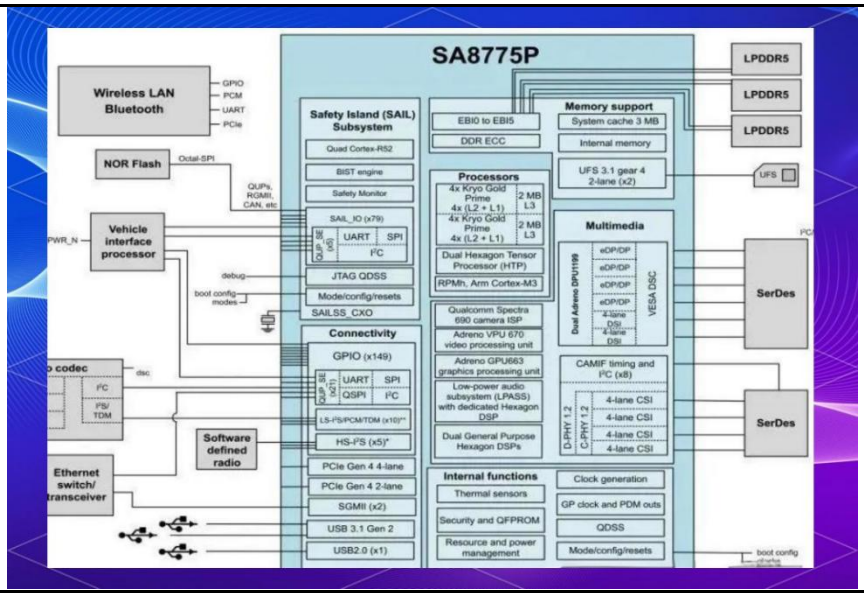
图63：多种“中央+zona1”计算架构形式

| 时间 | 适配车型 | 应用趋势 |
| --- | --- | --- |
| **形式一：高性价比、高灵活性** | ✓ 10-20万价格区间的主机厂及车型常见架构，如零跑、吉利、小鹏mona等高性价比电动车型 | ✓ 高中低配置阶梯灵活适配<br>✓ 舱+驾（低阶）Onechip，高阶智驾(选配）可以灵活选型<br>✓ 域间100M/1G以太网连接<br>✓ 2~3个区域控制器 |
| **形式二：高性能，多计算平台** | ✓ 20万+中高端车型广泛采用 | ✓ 使用座舱和智驾两个计算单元，多采用高通+英伟达/国产智驾芯片；<br>✓ 相对独立的软件平台，继承性高<br>✓ 域间100M/1G以太网连接<br>✓ 3~4个区域控制器 |
| **形式三：高性能，舱驾一体计算平台** | ✓ 20万+中高端车型正逐渐向舱驾一体计算平台发展 | ✓ AI计算中心（ThorX/U、ORIN-X、地平线J6、高通SA8397P等）部署智驾E2E模型和端侧大模型（多模态交互/LLM）<br>✓ AI交互中心（高通SA8295P芯片、高通骁龙Gen3、AMD Ryzen2等）作为智慧智能体，实现认知中枢和记忆中枢能力，部署大模型Proxy，意图识别模型，大模型推理架构等<br>✓ 通过PCIe片间通信，智驾和座舱之间跨域数据带宽从千兆大幅提升到10+Gbps，实现10倍以上的传输速率提升<br>✓ 这一阶段重点发展融合芯片（Fusion Chips），包括主机厂如小鹏、蔚来自研芯片，以及高通SA8775/SA8795，黑芝麻智能武当C1200系列等 |
| **形式四：高性能，舱驾控一体化中央计算平台，未来进一步引入Chiplet** | ✓ 30万+高性能旗舰车型 | ✓ 采用双Thor-U、双高通SA8795等高性能中央计算芯片实现舱驾控一体化<br>✓ 下一步，芯粒Chiplet技术将在汽车行业应用，瑞萨在业内率先推出采用车规3nm制程的多域融合SoC——R-Car X5系列，计划于2027年量产，提供通过Chiplet（小芯片封装）技术扩展人工智能（AI）和图形处理性能的选项。 |

数据来源：佐思汽研，东吴证券研究所

**而舱驾一体是汽车 E/E 架构从分布式向集中式演进过程中的产物。**使用一颗超高算力的 SoC 同时承载智能座舱和智能辅助驾驶，能够最大化地实现降本增效、优化算力利用率和提升整车性能。

图64：SA8775P




数据来源：facetop 智能汽车，东吴证券研究所

**SA8775P 是高通主推的一款面向主流市场的高性能、高性价比的“舱驾一体”融合芯片。**其核心战略是通过单芯片解决方案，整合智能座舱与智能驾驶功能，以替代传统的“座舱芯片+智驾芯片”分体式方案。该芯片的 CPU 采用两簇八核心设计（Kryo 680，基于 ARM Cortex-X1 架构），主频最高可达 2.35GHz，总算力达 230k DMIPS，并配备 4MB 三级缓存与 512KB 二级缓存。GPU 搭载 Adreno 663，图形算力为 1.1~1.3 TFLOPS，支持高质量图像渲染与视频处理，可满足复杂 3D 图形、AR/VR 及高分辨率车载显示的运行需求。

**Drive Thor 是英伟达面向下一代智能汽车推出的旗舰 SoC。**它整合了 L2+级到全自动驾驶、自动泊车、驾驶员与乘客监测、数字仪表盘、车载信息娱乐系统以及后排娱乐等多项智能功能，实现真正意义上的舱驾一体架构。这种集中式的设计不仅提升了整车系统的运行效率，也有效降低了整体硬件和软件成本。作为 DRIVE Orin 的继任者，Thor 不再是简单的性能迭代，而是在架构层面实现跨域融合，推动车载计算从多芯片协同走向单芯片整合。它可以将全部高达 1000 TOPS 用于 L2+级到全自动驾驶流程，也可以部分分配给信息娱乐和车内 AI，从而实现座舱与自动驾驶的算力平衡。

**图65：超级芯片参数**

| 芯片类型 | AI 算力(TOPS) | 替代关系 | 量产时间 | 预估售价(美元) |
|---|---|---|---|---|
| NVIDIA Thor | 2000 | 1 颗 Thor-U=2 颗 Orin-X | 2025 | 1000+ |
| Qualcomm Flex | 2000 | 1 颗 SA8775P=1 智驾+1 座舱 | 2025 | 8650+ |

数据来源：英伟达官网，高通官网，facetop 智能汽车，芯流汽车，东吴证券研究所

**作为中国本土智能辅助驾驶芯片的领军企业，地平线采取了“软硬协同，开放生态”的差异化竞争策略。**地平线计划在 2026 年内发布舱驾一体芯片。

黑芝麻武当 C1296 是行业首颗支持多域融合的芯片平台。武当 C1200 家族基于 7nm 车规级制程工艺打造，集成 CPU、GPU、NPU、DSP、ISP、MCU 及网关模块于一体，真正实现座舱、辅助驾驶、车身、网关四域融合。所有计算单元通过高速片上总线互联，算力由统一软件平台进行动态调配：高速巡航时，NPU 优先服务辅助驾驶感知；停车娱乐时，GPU 与 NPU 可以协同工作，提供极致的座舱游戏与 AI 交互体验；DMS 检测到驾驶员疲劳时，可联动空调、音响、导航自动调节环境，甚至规划至休息区。武当 C1236：国内首颗单芯片实现 NOA 行泊一体方案，集成高性能 ISP（每秒处理 2.4G 像素），支持 12 路高清摄像头接入，满足高速 NOA 与入门级城区 NOA 的感知需求；以高性价比助力 L2+辅助驾驶普及。武当 C1296：面向跨域计算架构，支持 DSI、eDP、LVDS 等多种显示接口，最多可驱动 5 个显示屏；支持 4K@60fps 视频编解码(H264/H265/VP8/VP9/JPEG)；满足多

屏联动与沉浸式娱乐需求；音频系统集成 HIFI5 DSP，支持 7.1 声道输出，可实现座舱语音交互与环绕声体验；内置万兆级网关交换模块，支持 2x 10GbE+2x 2.5GbE 以太网接口，数据转发容量达 40Gbps，作为智能辅助驾驶、座舱、MCU 多域之间的数据高速交换通道，全面支持舱驾一体、CMS 电子后视镜、整车数据交换等复杂应用场景。武当 C1200 家族采用差异化产品策略，形成覆盖主流市场的完整矩阵，且每款产品的参数配置都精准匹配目标场景。该策略支持车企基于不同车型定位灵活选型，实现从 10 万元级主流车型到 30 万元以上旗舰车型的平台化部署，显著降低研发与制造成本。

**图66：黑芝麻武当 C1200 家族**


数据来源：黑芝麻智能官方公众号，东吴证券研究所

## 4. AIoT 与具身智能是端侧 AI 的市场增量长尾与未来

具身智能产业化进程持续加快，整机量产环节呈现高度集中于国内的产业特征，机器人与智能硬件市场的快速渗透，带动端侧主控与边缘计算芯片方案进入规模化落地阶段。随着终端对本地 AI 算力需求持续提升，行业内针对算力扩展的硬件方案逐步成熟，为产业链带来明确的升级红利与业绩兑现空间。此外，在全球科技产业分工下，部分具备长期生态合作基础的边缘计算平台厂商，有望承接大厂释放的相关供应链份额。

端侧 AI 的产业景气度正从核心硬件向泛终端领域持续蔓延。消费端，在多模态大模型的赋能下，新型穿戴式算力终端逐步成为重要的入口形态；汽车端，本土计算芯片凭借性价比优势持续推进智能座舱领域的替代进程，并向舱驾融合等高价值环节延伸，不断打开行业成长空间。从产业演进趋势看，终端产业链正从单一产品放量，转向架构升级、生态协同与场景渗透的综合竞争。

### 4.1 消费级 AIoT（XR 与穿戴）：下一代计算平台的黎明

消费级 AIoT 涵盖智能眼镜、智能手表、智能音箱、TWS 耳机等多种终端形态。其中，AI 眼镜作为最具代表性的创新品类，正成为端侧 AI 落地的核心载体，未来或将引领整个穿戴设备市场的增量增长。

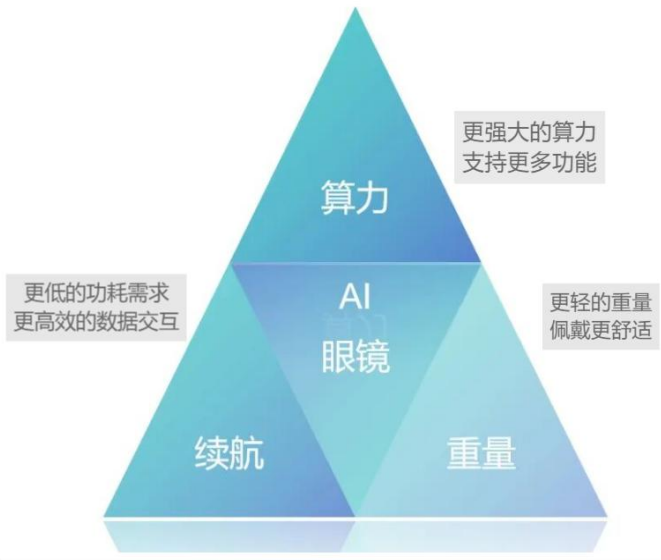
### 4.1.1 现状与痛点：多元终端并进，AR 眼镜引领突破

**消费级 AIoT 市场正呈现“多点开花”的格局。**智能手表、TWS 耳机、智能音箱等品类持续渗透，形成稳定基本盘，但以 AR 眼镜为代表的 XR 设备正凭借其作为“下一代计算平台”的潜力，成为驱动市场空间（TAM）扩张的核心变量。根据 IDC 数据，2024 年全球可穿戴设备出货量达 5.4 亿台，市场进入平稳增长期，市场增长引擎正从量的普及转向质的革新，即通过嵌入更复杂的本地 AI 能力来提升产品价值与用户体验。其中，智能手表和耳机占据主导，功能迭代集中于健康监测与音频体验。而 AR 眼镜因其能深度融合物理与数字世界，提供情境化、无感化的信息交互，被视为继智能手机后最具革命性的终端形态。然而，这一进化路径面临一个共通的底层约束：在消费电子严格的体积、续航与成本限制下，如何实现高性能 AI 计算。

**具体而言，不同品类应对其特有的“性能三角”困境。**智能手表手环的核心矛盾在于高精度健康监测（如连续血氧、心电图 ECG）、独立通信（eSIM）与超长续航（>7 天）之间的平衡。增加传感器精度与功能会显著增加功耗。TWS 耳机的痛点集中在**顶级主动降噪（ANC）/空间音频算力、高保真无损音频传输与单次续航（>6 小时）的兼顾，实现计算音频所需的高性能 DSP 与低延迟无线连接均是耗电大户。智能音箱的挑战来自远场语音识别的准确率、多设备协同的响应速度与始终在线（Always-On）的待机功耗。其需在离线环境下处理复杂的语音唤醒和指令理解，对端侧 SoC 的能效比要求苛刻。

**在 AIoT 终端中，AR 眼镜是其中一个关键品类，也面临“算力、续航、重量”的不可能三角挑战，技术突破成为规模化普及的关键前提。**该挑战即在普通眼镜的重量(<50g)与形态限制下，实现空间计算所需的高算力与全天续航，构成了“重量、算力、续航”的“不可能三角”。

**图67：性能、重量、续航构成 AI 眼镜“不可能三角”**




数据来源：OFweek 网，东吴证券研究所

**这一矛盾构成核心发展瓶颈：提升续航需增大电池容量，但电池增重会恶化佩戴体**

**验；提高算力需提升功耗，但主动散热方案（风扇/散热片）在<50g 重量约束下无实施空间。**在重量维度，全天候佩戴要求重量需接近普通眼镜（<50g，理想状态<30g），否则会导致耳廓压痛、鼻托疲劳，这是消费电子向医疗级佩戴体验进化的硬约束。现有产品典型续航仅 2-4 小时（Meta Ray-Ban 拍摄状态仅 30-45 分钟），而被动散热条件下，设备面临温度过高的问题，而耳部接触面温度需<43° C，否则有烫伤风险。从 2025 年产品表现来看，主流 AI 眼镜的续航时间仍集中在 3-4 小时，当前主流消费级 AR 眼镜产品，如 Ray-Ban Meta（约 50g，续航 4 小时）和 Rokid Max（约 75g，续航约 3 小时），仍在这一三角约束中艰难平衡，算力多依赖运行轻量模型。

### 4.1.2 技术破局：分体式计算与高速互联成为主流路径

**当前 AI 眼镜行业正在努力攻克这一不可能三角的极限，分体式（Split Processing）架构成为主流过渡方案。**分体式架构通过计算卸载重构了端侧 SoC 的设计逻辑：眼镜只负责感知和显示，重负载计算（AI/渲染）交给手机或主机，改变了市场对 SoC 的连接能力（Wi-Fi 7/UWB）需求。

**图68：AI 眼镜芯片的分布式架构**


数据来源：高通，青亭网，东吴证券研究所

**分体式架构将传统一体机式的计算任务拆解为“端-边-云”三级算力协同。**分体式方案将重负载的 AI 推理与渲染任务卸载至手机、主机或云端，从而重塑了端侧 SoC 的设计重点。该架构的本质是根据任务对延迟、带宽和算力的敏感性进行分层处理：眼镜端 SoC（如高通 AR1）专注于低功耗传感器融合（IMU，摄像头）、基础视觉 AI（如二维码识别）和显示驱动；配套设备（如手机、专用计算单元）则承担复杂的多模态大模型推理、操作系统运行等高算力任务。这一模式将端侧 SoC 的核心能力要求，从纯粹的峰值算力（TOPS）转向了极致的能效比（Perf/Watt）和强大的高速连接能力。眼镜端（1-10 TOPS）负责处理传感器采集、实时 SLAM 定位、初阶 AI 筛选（如人脸检测、二维码识别），要求延迟<20ms；外接设备端（5-10 TOPS）如手机/主机承担复杂 AI 推断（VLM 视觉语言模型）、渲染预处理，延迟<50ms；云端（50-100 TOPS）支持大模型推理、多模态生成，延迟<100ms。这种分布式架构既缓解了眼镜本身的功耗与散热限制，也扩展了其实际应用边界。

**随着架构将计算任务从眼镜端剥离后，前端与后端之间的高速数据通道成为整个系统的生命线。**高传输带宽方面，需实时传输未经压缩的高分辨率画面（单眼 1080p+）、

传感器数据流，Wi-Fi 6 的 9.6Gbps 理论峰值已难以满足无损传输需求；需要低传输延迟，运动到光子延迟（MTP）必须控制在 20 毫秒以内，最好低于 10 毫秒，否则会导致眩晕；以及需要高连接可靠性，防止数据丢包或波动而导致画面卡顿、定位漂移。

**图69：不同端侧算力需求**

| | 眼镜端 | 外接设备端 | 云端 |
|---|---|---|---|
| 主要任务 | 实时传感器处理（摄像头、IMU）、基础环境理解、低延迟显示 | 运行轻量化大模型、处理复杂 AI 任务（实时翻译、视觉搜索） | 复杂大模型训练、超大规模数据分析、为终端提供模型推理支持 |
| 核心需求 | 能效比(TOPS/W)远高于峰值算力。必须在 1-2W 的严格功耗限制下完成工作 | 在设备散热和续航容许范围内，提供尽可能强的本地推理能力，以减少对云端的依赖 | 算力以集群化形式提供，如 GPU/TPU 服务器集群，通过高速互联构成巨大算力池 |
| 算力范围 | 0.2-2 TOPS | 5-10TOPS | 50-100TOPS |

数据来源：ARAI 眼镜社区，东吴证券研究所

**因此，这直接催生了对新一代支持低延迟、高带宽的无线连接技术的刚性需求，Wi-Fi 7 和 UWB（超宽带），已成为下一代 XR SoC 的必备特性。**Wi-Fi 7 成为满足分体式 AR 超高带宽和低延迟需求的核心技术。其支持 320MHz 信道带宽和 4K QAM 调制，理论峰值速率可达 5.8Gbps 以上，多链路操作（MLO）允许设备同时通过 2.4GHz、5GHz 和 6GHz 多个频段传输数据，保障高清视频流与传感器数据的实时同步传输，大幅提升吞吐量和抗干扰能力。UWB（超宽带）则承担空间锚定功能，实现厘米级定位（精度 10cm 级）。UWB 的核心优势在于能够实现厘米级的高精度定位和测距，同时具备低功耗、高安全性和强抗干扰性，因此可用于眼镜与手机/手柄的精确相对定位，以及低功耗唤醒检测，构建无缝的跨设备交互体验。

**图70：Wi-Fi 6、Wi-Fi 7、Wi-Fi 8 区别**

| | Wi-Fi6 | Wi-Fi7 | Wi-Fi8 |
|---|---|---|---|
| IEEE 标准 | 802.11ax | 802.11be | 802.11bn |
| 最大传输速率 | 9.6Gbps | 23Gbps | 23Gbps |
| 安全协议 | WPA3 | WPA3 | WPA3 |
| 信道宽带 | 20/40/80/160/80+80 MHz | 最大可达到 320MHz | 320MHz |
| 调制方式 | 1024-QAM OFDMA | 4096-QAM OFDMA | 4096-QAM OFDMA |
| 频段 | 2.4GHz、5GHz、6GHz（仅 Wi-Fi6E） | 2.4GHz、5GHz、6GHz | 2.4GHz、5GHz、6GHz |

数据来源：华为官网，IT 之家，东吴证券研究所

**高通于 2025Q3 宣布 IEEE802.11bn 标准预计将于 2028 年完成，并将成为 Wi-Fi 8 的基础。**Wi-Fi 8 的目标是在复杂的现实环境中优先保障可靠的性能表现，即使在网络

拥塞、易受干扰且移动性强的场景中，也能提供出色的连接。Wi-Fi 8 引入超高可靠性（UHR）框架，在峰值速率与 Wi-Fi 7 持平的基础上，通过协调空间重用（Co-SR）、协调波束成形（Co-BF）、动态子信道操作（DSO）等技术，将边缘吞吐量提升 25%、P95 延迟降低 25%、漫游丢包减少 25%，实现复杂环境下的"有线级"无线体验。

**高通骁龙 AR2 等专用芯片的出现，配合端云协同计算模式，正逐步满足实时图像识别、自然语言处理等 AI 任务的算力需求。**高通在 2022 骁龙峰会推出第一代骁龙 AR2 平台，这是市场上首款专为 AR 设备打造的处理器。AR2 采用先进的 4nm 工艺制程；同时，为打造超轻薄、高性能 AR 眼镜，高通采用多芯片分布式处理架构并结合定制化 IP 模块。AR2 可满足骁龙分离式渲染方案，其处理器能够动态地将时延敏感性感知数据处理分配给眼镜终端，把更复杂的数据处理需求分流（如渲染）到智能手机、PC 或其他兼容的终端上。

**图71：高通骁龙 AR2 芯片分布式架构**


数据来源：高通官网，东吴证券研究所

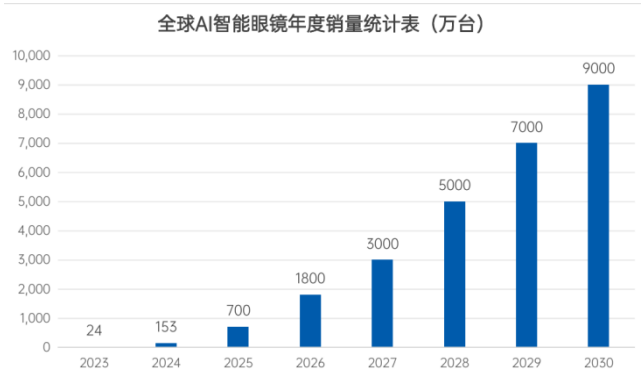
**骁龙 AR2 平台为多芯片组合模式分布式架构（此前 XR1、XR2 均为单芯片集成方案）。**包括 AR2 主处理器（负责感知、显示）、AR2 协处理器（负责摄像头聚合、AI 和计算机视觉）、 连接模组（低时延、低功耗 Wi-Fi 7）三个部分。分布式架构可以更好发挥各芯片的优势。高通骁龙 AR2 三个小芯片分别位于左镜腿、鼻托上方、右镜腿上。AR 处理器为实现动作到显示（M2P）的低时延而优化，同时支持多达九路并行摄像头进行用户和环境理解，其增强的感知能力包括能够改善用户运动追踪和定位的专用硬件加速引擎，用于降低手势追踪或六自由度（6DoF）等高精度输入交互时延的 AI 加速器，以及支持更流畅视觉体验的重投影引擎；AR 协处理器可聚合摄像头和传感器数据，支持面向视觉聚焦渲染的眼球追踪和虹膜认证，从而仅对用户注视的内容进行工作负载优化，以帮助降低功耗；通讯模块利用高通 Fast Connect 7800 连接系统，开启极速商用 Wi-Fi7 连接，使 AR 眼镜和智 能手机或主机终端之间的时延低于 2 毫秒。该芯片集成了对于 Fast

Connect XR 软件套件 2.0 的 支持，能够更好地控制 AR 眼镜的数据，以改善时延、减少抖动并避免不必要的干扰。

### 4.1.3 市场空间扩张，从“配件”到“必需品”的质变

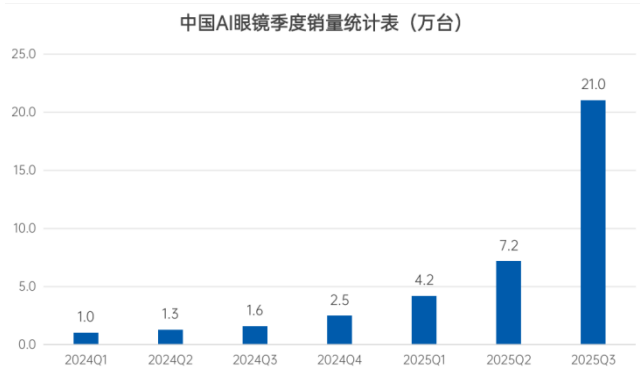
**智能眼镜正从“配件”向“必需品”质变，市场天花板有望向智能手机规模看齐。**未来智能眼镜将部分替代手机在导航、通信、信息获取、简单的 AI 交互等高频场景的功能。 当前市场仍处于早期起量阶段，其中消费级产品占比快速提升。行业预测显示，随着苹果 Vision Pro 平价版、Meta Ray-Ban 系列迭代以及更多手机厂商入局，市场将在 2026 年迎来加速拐点。长期来看，若产品形态和体验成熟，AR 眼镜的年出货量天花板有望向亿级乃至十亿级迈进，这为上游端侧 SoC 带来了确定性的增量空间。根据 wellsennXR 数据及预测，2035 年全球眼镜销量将达 20 亿副左右，市场规模将接近 2000 亿美元。据 IDC 统计，2025 年第三季度全球智能眼镜市场出货量 429.6 万台，同比增长 74.1%。其中全球音频和视频拍摄眼镜市场出货量 299.4 万台，同比增长 287.5%。IDC 预测，2026 年中国市场 AI 眼镜出货量将达 275 万台（同比增长 107%），全球市场规模于 2030 年突破 1170 万台。Wellsenn XR 预测 2035 年全球 AI 眼镜销量有望突破 14 亿台，年出货量将看齐当今智能手机的规模。

**图72：全球 AI 智能眼镜年度销量统计（万台）**




数据来源：WellsennXR，东吴证券研究所

**图73：中国 AI 智能眼镜销量统计（万台）**




数据来源：WellsennXR，东吴证券研究所

**2025-2027 年将成为智能眼镜产业从技术验证期向规模化普及期跃迁的历史拐点，Apple Vision Air、Meta Orion 消费版、三星 Project Moohan 三款标杆产品的集中上市，将触发产业链的快速起量。**苹果计划于 2027 年下半年量产 Vision Air，通过减重 40%并降价 50%的组合拳，将 XR 设备从极客奢侈品推向消费电子。市场普遍预期，Vision Air 的出货量有望远超 Vision Pro 上市初期的表现。Meta 凭借“双线并行”策略，同时巩固当下市场与定义未来形态。未来，Meta 在通过 Ray-Ban Meta 智能眼镜培育用户 AI 交互习惯的同时，其内部称为 “Project Nazare” 的 Orion 项目正致力于攻克技术难关。业内预计，该量产版 AR 眼镜有望在 2027-2030 年间问世，目标是实现真正的全天候佩戴，对光学、微显示和低功耗芯片产业链提出极高要求并产生巨大拉动。三星的入局为 Android 生态树立了高端标杆，加速了全局竞争与供应链迭代。它的出现与苹果

Vision 产品形成了直接竞争，迫使双方在显示、交互、芯片性能上持续加码，从而推动高端 XR 供应链的技术升级与成本优化。

**图74：Meta Orion 眼镜**

数据来源：Meta 官网，东吴证券研究所

**图75：Project Moohan 头显**

数据来源：三星官网，东吴证券研究所

**三大巨头的行动共同指向一个方向：产品形态从头显向轻型眼镜演进，价格向高端数码产品靠拢，共同推动市场向规模化普及迈进。**这个过程将推动价格下探与生态建立，苹果 Vision Air、三星 Moohan 等将高端 MR 头显价格从 3500 美元拉至 1500-2000 美元区间。同时，Meta Quest、苹果 Vision Pro/ Air、三星 Moohan 将分别围绕各自的 OS（Quest OS、Vision OS、Android XR）建立应用生态，吸引早期大众用户和开发者。随着苹果、Meta 的全天候 AI/AR 眼镜逐步成熟并上市。产品形态接近普通眼镜，主打无感交互和情境化 AI，旨在成为下一代个人计算中心。若能解决核心应用场景，市场天花板将被打开，进入更广阔的普及阶段。

**图76：主流消费级智能眼镜型号对比**

| 型号 | 苹果 Vision Air | Meta Orion 消费版 | 三星 Project Moohan |
|---|---|---|---|
| 发布时间 | 2027 Q3 | 2027（最早） | 2025.10 |
| 定价 | $1500-$2000 | $1500-$2000 | $2000 |
| 出货量级 | 100 万台（2027） | 待定 | 10 万台（2025） |
| 技术路线 | 分体式 | 一体式/分体式混合 | 分体式（外置电池） |

数据来源：苹果官网，Meta 官网，三星官网，东吴证券研究所

**除 AI 眼镜外，多终端协同拉动 SoC 需求升级。**智能手表领域，三星 Galaxy Watch6 采用 Exynos W950 芯片，支持 ECG +血糖联合监测；TWS 耳机方面，AirPods Pro 3 搭载 H17 芯片，实现实时翻译与噪音分类；智能音箱市场，小爱音箱 Pro 2026 款采用全志 A733 芯片，支持离线唤醒与多轮对话。这些终端的 AI 化升级推动 SoC 平均售价（ASP）从 2023 年的 18 美元升至 2026 年的 27 美元。

### 4.1.4 竞争格局：高通垄断生态，手机厂商开启破局尝试

高通在 XR/AI 眼镜芯片市场占据绝对主导地位，但三星、Google、小米等手机巨头正通过"自研+深度定制+生态绑定"的三线策略打破这一格局，存在产业变局。

**在 XR 专用 SoC 领域，高通凭借其 Snapdragon XR 系列（XR2 Gen 2/AR1/AR2）建立了近乎垄断的生态优势。**截至目前，除苹果采用自研 M 系列芯片外，Meta、Pico、HTC、DPVR 等主流品牌占据全球市场出货量的 90%，全部采用高通骁龙 XR 平台。市面上主流的独立 VR 头显和 AI 智能眼镜均采用高通骁龙 XR2 Gen 2 或 AR1 Gen 1 平台。Meta Quest 全系 Quest 3（骁龙 XR2 Gen 2）、Quest Pro（XR2+ Gen 1）及 Quest Pro 2（XR2+ Gen 3）均采用高通方案；Pico 与国产阵营中字节跳动 Pico 4、Pico Ultra，以及 Vivo Vision Discovery Edition 等中国主流头显清一色搭载 XR2+ Gen 2；AI 眼镜中 Ray-Ban Meta（AR1 Gen 1）、雷鸟 X2（XR2）、Rokid Glasses（AR1）、小米 AI 眼镜（AR1）等轻量级设备亦被高通覆盖。根据 Counterpoint Research 数据，截至 2024 年，其全球 XR 芯片 SoC 市场占有率高达 90%以上。2022 年，高通与 Meta 签署多年期战略协议，联合研发基于骁龙 XR 平台的定制化芯片组，形成"芯片-算法-内容"的闭环优化。通过 Snapdragon Spaces XR 开发者平台，高通向 OEM 厂商提供从光学模组选型、散热设计到 Tracker 布局的完整参考方案，使 Pico、DPVR 等厂商能在 6-9 个月内完成产品量产，反过来又强化高通标准的行业地位。

**图77：主流厂商芯片使用情况**

| 厂商 | 芯片方案 | 代表产品 |
|---|---|---|
| 苹果 | 自研 | 基于 Apple Watch 芯片定制开发，Vision Pro 使用 M 系列+R1 芯片 |
| Meta | 高通 | 核心产品全线采用，Quest 系列（XR2 Gen 2 等）、Ray-Ban Meta（AR1 Gen 1） |
| Pico（字节跳动） | 高通/自研 | Pico 4 等现售产品采用高通 XR2+ Gen 2。已公布将推出搭载全链路自研专用芯片的新一代头显 |
| 三星 | 高通 | Android XR 头显采用高通骁龙 XR2+ Gen 2 平台 |
| 小米 | 高通＋协处理器 | AI 眼镜采用高通 AR1 作为主控，并搭载恒玄科技的蓝牙音频处理器等协处理器芯片 |

数据来源：苹果官网，Meta 官网，三星官网，小米官网，东吴证券研究所

**高通的垄断并非单纯依靠先发优势，而是通过"硬件-算法-开发者"的三重绑定构建起生态壁垒：**

**芯片领先优势明显。**XR2+ Gen 2 支持单眼 4.3K@90fps，AI 性能较前代提升 8 倍，实现 12ms 全彩视频透视延迟；XR2+ Gen 3 首发 Oryon CPU 内核，支持 16GB LPDDR5X 内存，直接对标苹果 Vision Pro 的 M2 芯片；2025 年 6 月发布的 AR1+ Gen 1 进一步支持运行参数规模达 10 亿的 Meta Llama-3.2-1B 等端侧小型语言模型。高通专为眼镜设计的 AR1 平台集成了强大的 ISP 和 NPU（如第三代 Hexagon NPU），能本地运行数十亿参数的 AI 模型。

**战略客户深度绑定，将智能手机时代的“统一技术路线图”成功延伸至空间计算领域。**通过与 Meta 等头部厂商的深度绑定和持续迭代，高通构建了从高性能头显到轻量化眼镜的完整产品矩阵，主导了行业技术标准。2025 年 Meta 的 Ray-Ban Meta 累计出货

已突破 300 万副，由于 Ray-Ban Meta 等产品超预期销售，Meta 向高通大幅追加了超过 1200 万颗 AR1 系列芯片的订单。这形成了“爆款产品-芯片放量-生态强化”的闭环，主流品牌如 Meta、Pico、HTC 等在开发独立头显时，普遍将高通骁龙 XR 平台作为首选。

**图78：高通 AR1 与 W5 等对比**

| 型号 | Wear 4100+ | W5/W5+ Gen1 | AR1 Gen 1 |
|---|---|---|---|
| 工艺制程 | 12nm | 4nm | 4nm |
| CPU | 4×A53，最高 2.0GHz | 4×A53，1.7GHz | 4×A55，1.9+GHz |
| 内存 | LPDDR3，750MHz | 1×16 LPDDR4，2133MHz | 1×16 LPDDR4X，2.1GHz，LLC |
| GPU | Adreno 504@320MHz | Adreno 702@1GHz | Adreno 621，支持 OpenGL ES3.2 和 Vulkan 1.1 |
| 显示 | 1080p 30fps | 1080p 60fps | 1280×1280 60fps (Dual) |
| ISP | Dual ISP 16MP+16MP | Dual ISP 16MP+16MP，EIS 3.0，FNR，Pseudo ZSL | 2×12MP，IFE 和 IFE-lite，HW JPEG Enc，Pseudo ZSL |
| DSP | Hexagon QDSP6 v56 | Dual Hexagon QDSP V66K HiFi 5 DSP | Hexagon DSP，1.2GHz，2MB LPI，eNPU，Sensor Hub，Voice UI |
| 接口 | USB2.0 | USB2.0 | SPI-NOR，12 SE，1×USB3.1 Gen1，2×PCIe Gen3 1-lane |
| 蓝牙 | 蓝牙 5.0 | 蓝牙 5.3 | 蓝牙 5.3 |

数据来源：高通官网，三易，东吴证券研究所

**面对高通的生态壁垒，手机厂商以及国产芯片供应商等新进入者并未选择正面强攻通用主控 SoC，而是从外围和细分市场切入，聚焦生态延伸与方案整合。**手机厂商入局 XR，将其在移动生态、人机交互和供应链整合上的优势自然延伸，多采用“高通主 SoC + 专用协处理器”的整合方案，以快速推出有竞争力的产品。小米在其 AI 眼镜产品中，采用了高通 AR1 + 恒玄科技 BES2700 的双芯片架构。高通芯片处理复杂 AI 和视觉任务，恒玄芯片专攻高清音频和低功耗无线连接，通过明确分工优化整体能效。创维的 A6 系列 AI 眼镜也采用了类似的高通+恒玄（BES2800）方案。而理想汽车推出的 Livis AI 眼镜，则更聚焦车载联动场景，采用了恒玄 BES2800 主控 + 独立 ISP（图像信号处理器）的极简组合，在保证核心功能的前提下追求更低功耗和成本。

**国产芯片供应商聚焦差异化技术路径崛起，以恒玄科技、瑞芯微为代表的国产芯片厂商，正通过提供专用协处理器、主控平台或完整解决方案，而一些新兴专精厂商聚焦于填补高通等巨头未充分覆盖的细分需求。**恒玄科技已成为 AI 眼镜音频与低功耗连接方案的核心供应商。其旗舰芯片 BES2800 基于 6nm 工艺，单芯片集成多核 CPU、NPU、低功耗 Wi-Fi 与蓝牙。它不仅独家供应 Meta Ray-Ban 系列，更被小米、创维、理想等众

多品牌采用，全球市占率超 30%。其下一代 BES3000 系列据称功耗降低 40%，算力提升至 12TOPS，旨在适配更独立的智能眼镜终端。此外，万有引力发布专为 AI 眼镜打造的极眸 G-VX100 ISP 芯片，专注于超低功耗下的高清视频拍摄、空间视频及眼动追踪功能。瑞芯微凭借在多媒体处理领域的积累，其 RV 系列芯片已应用于多个 AI 眼镜项目，下一代芯片正重点向 AI 眼镜方向演进。

**图79：主流 AI/XR 设备 SoC 芯片方案对比**

| 型号 | | 制程工艺 | 定位与技术特点 | 典型应用场景/客户 |
|---|---|---|---|---|
| 高通 | 骁龙 XR2 Gen 2 | 4nm | CPU/GPU 性能提升，支持高级 VST， AI 算力约 20TOPS | Meta Quest 3 等 VR/MR 头显 |
| | 骁龙 AR1 Gen 1 | 4nm | 首代智能眼镜专用处理器。优化低功耗拍摄、分享与基础 AI | Ray-Ban Meta（第一代） |
| | 骁龙 AR1+ Gen 1 | 4nm | 高端独立智能眼镜主控 SoC。集成第三代 Hexagon NPU，支持本地运行 10 亿参数 SLM；尺寸更小功耗更低 | 2025 年后高端智能眼镜 |
| 恒玄科技 | BES2800 | 6nm | 低功耗无线音频与计算 SoC。集成多核 CPU、NPU、Wi-Fi/蓝牙，专注音频与连接 | Meta Ray-Ban（音频主控）、小米/创维 AI 眼镜（协处理器）、理想 Livis 眼镜 |
| 瑞芯微 | RK3588M | 8nm | 6TOPS NPU，支持多屏异显 | AI 眼镜 |
| 手机厂商整合方案 | 高通 SoC + 协处理器 | 4nm/3nm | 利用高通 SoC 处理复杂 AI 与视觉，搭配恒玄等协处理器专攻音频/连接，平衡性能与功耗 | 小米、创维等品牌 AI 眼镜 |
| 苹果 M2/M3 | Apple Vision Pro | 5nm/3nm | 超强统一内存架构，顶级 CPU/GPU 性能， 专用媒体处理引擎 | Apple Vision Pro |
| 垂直场景方案（君正） | 专用 ISP/MCU | 8nm | 如万有引力 G-VX100 ISP 芯片，专注低功耗高清视觉处理；或采用 MCU+独立 ISP 组合，追求极致能效与成本 | - |

数据来源：高通，瑞芯微，半导体行业观察，东吴证券研究所

## 4.2 工业与行业物联网（泛 IoT）：国产替代的“蚂蚁雄兵”

### 4.2.1 市场特征：“多品类、小批量”的极致碎片化格局

**工业与行业物联网（泛 IoT）市场呈现极其碎片化特征，单一品类出货规模有限但细分品类覆盖范围极广，形成典型的“分散化”市场结构。**碎片化体现在两个维度：一是品类分散，2024 年中国智能家居市场规模 7850 亿元，其中扫地机器人、智能音箱、智能门锁分别占比 18%、15%、12%；二是需求差异化，商显芯片需侧重高清显示驱动，扫地机芯片需强化路径规划与传感器融合，智能音箱芯片需优化语音识别算法，通用型 SoC 难以满足全场景需求。2024 年中国工业物联网解决方案市场规模已达 1.4 万亿元，2025 年该规模预计增至 1.5 万亿元，其中泛 IoT 细分领域贡献极大的市场规模，但该规模分散于超 20 个细分品类，包括商显设备、智能收银机、扫地机器人、智能音箱、安防 IPC、工业传感器、智能电表、二维码扫描设备等。

**图80：IoT 芯片下游应用**

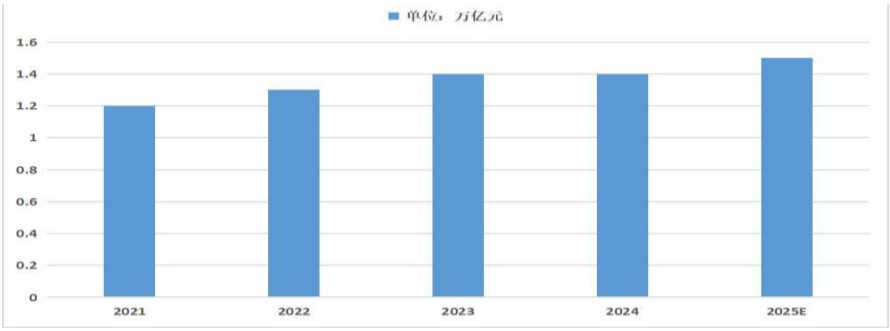

数据来源：与非网，东吴证券研究所

**中国智能家居市场规模和渗透率快速增长。**2024年中国智能家居市场规模达到7850亿元，同比增长 18.2%，中国一线城市智能家居渗透率已达 42%，一二线城市渗透率也超过 25%。但分散于扫地机、智能音箱、智能门锁、智能空调等多个小品类中，进一步凸显“品类多、单量小”的碎片化特点。同时，不同细分品类的技术需求差异显著：商显芯片需侧重高清显示驱动与多设备协同能力，扫地机芯片需强化传感器数据处理、路径规划与低功耗控制，智能音箱则需优化语音识别算法适配，这种场景化差异加剧了市场碎片化，也使得通用型 SoC 难以覆盖所有需求，为垂直品类厂商提供了生存空间。

### 4.2.2 核心驱动：国产替代的“低成本+成熟制程”双优势

**国产替代是推动泛 IoT 芯片市场发展的核心驱动因素，该领域凭借对制程要求低、成本敏感度高的特点，成为 SoC 国产替代进程中国产化率最高的板块之一。**从制程门槛来看，泛 IoT 芯片无需依赖 3nm/2nm 等先进制程，22nm/12nm 成熟制程已能满足多数场景的性能需求，很多智能家居芯片需支持基础数据运算与外设连接，22nm 制程即可实现。

**图81：中国工业物联网解决方案市场规模情况**




数据来源：中国报告大厅，《2025-2030 年中国工业互联网行业发展趋势及竞争策略研究报告》，东吴证券研究所

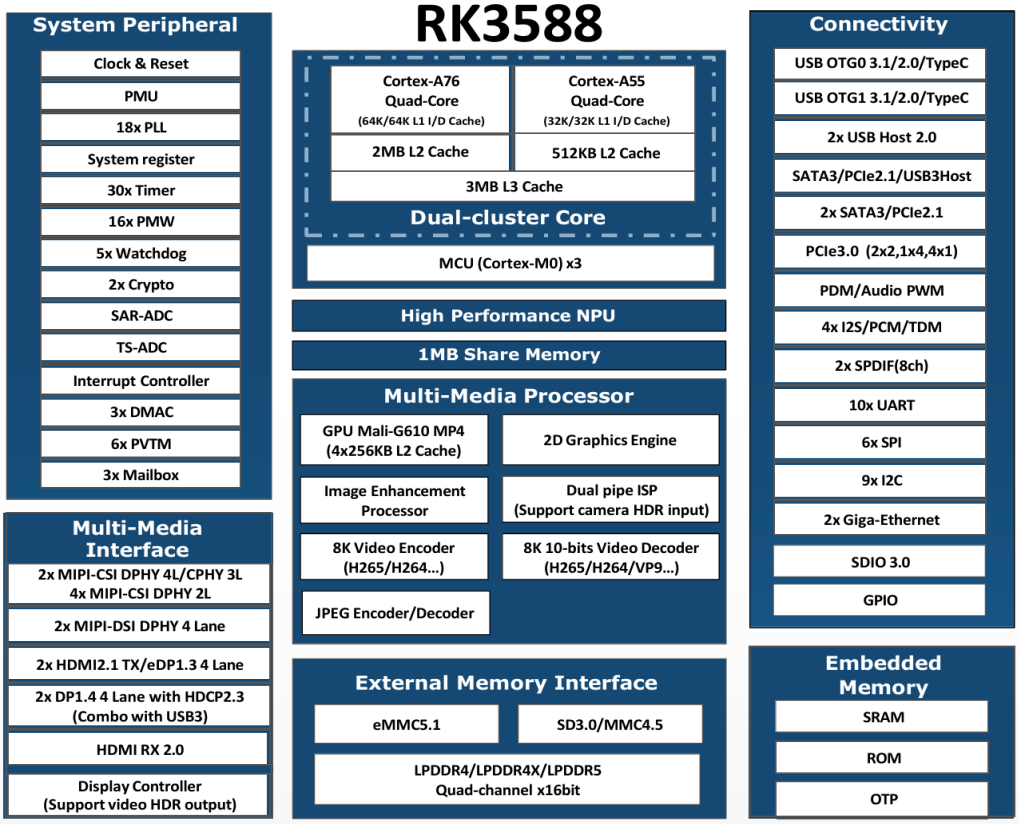
成本方面，国产芯片价格较海外品牌具有显著优势，瑞芯微 RK3568 系列采用 22nm 制程，市场单价约 20-30 美元，成为商显、智能家居等价格敏感型市场的首选。以 NB-IoT 模组为例，2025 年国产化率达 68%，较 2023 年提升 14 个百分点，显示国产替代在特定细分领域的快速突破。在智能座舱等复杂应用场景，国产芯片仍处于早期阶段，2024 年国产化率约 10%，但在入门级车型和特定应用场景中正快速渗透，本土厂商主导中低端市场的格局正在形成。

### 4.2.3 竞争格局：细分领域的“隐形冠军”

**瑞芯微（Rockchip）是高端商显与边缘计算芯片领域“高通平替”的核心国产厂商，凭借成熟制程与场景化性能实现对海外芯片的规模化替代。**从高端商显场景来看，其 8nm 制程的 RK3588 系列芯片的持续放量成为核心驱动力，其带动各 AIoT 算力平台在汽车电子、机器视觉、工业领域及各类机器人市场深度渗透，同时拉动产品结构升级，推动综合毛利率从 2023 年的 34.25%稳步提升至 2025 年前三季度的 41.77%。

**瑞芯微把机器人作为 AIoT 重要产品线。**当前公司 SoC 产品已经应用在多种形态机器人，拥有较高市占率，主要承担机器人“小脑”功能。依托技术积累和产品布局优势，公司以端侧算力协处理器布局机器人“大脑”，同时在机器视觉、音频等领域都有成熟方案，后续会快速在这些领域与客户展开更广泛合作。

图82：瑞芯微 RK3588 系列芯片架构图




数据来源：瑞芯微官方数据手册，东吴证券研究所

**晶晨股份（Amlogic）在机顶盒与智能电视 SoC 领域具备全球霸主地位，核心产品的市占率、技术壁垒与客户覆盖度均居全球前列。**在机顶盒 SoC 领域，晶晨股份大部分营收则主要于海外市场。招股书显示，截至 2025 年上半年，公司中国内地以外的市场收入占比高达 88.9%，而 24 年全年，这一比例高达 92.0%。以 2024 年相关收入计，晶晨股份在全球所有智能机顶盒 SoC 厂商中排名第一，市场占有率为 31.5%，全球每 3 台智能机顶盒即搭载 1 颗晶晨股份芯片，2025 年第二季度其系统级 SoC 芯片单季出货量近 4400 万颗，机顶盒芯片全球市占率排名第一，业务覆盖全球 250 余家主流运营商及 Netflix 等流媒体巨头。在智能电视 SoC 领域，在全球所有智能电视 SoC 厂商中排名第二，市场占有率为 16.8%。全球每 5 台智能电视中即有 1 台搭载其芯片，已与 TCL、创维、海信、小米等全球前 20 大电视品牌中的 14 家建立合作，且推出全球首颗 8K 超高清机顶盒 SoC 芯片 S928X（集成自研神经网络处理器），推动终端画质升级。同时，公司以高研发投入筑牢壁垒，2022-2024 年每年研发费用超 10 亿元，研发投入占营收比例均超 20%，在科创板同类企业中处于领先水平。

图83：晶晨股份产品矩阵


数据来源：公司官网，公司公告，scensmart，东吴证券研究所

**在稳固上述基本盘的同时，晶晨股份正凭借其深厚的多媒体处理与系统集成技术，加速向更广阔的 AIoT 市场进行平台化扩张，已成功在多个高增长细分领域建立起领先优势。**其智能影像解决方案面向安防监控、智能分析等场景，其芯片（如 A311D、C308X）支持高清视频处理与端侧智能分析（如人脸识别、对象追踪），实现数据本地处理，大幅减轻服务器负载。智能家居解决方案以智能音箱等产品为入口，其系列芯片（如 A113D、A113L、A113X）提供高性能、低功耗的远场语音交互能力，已广泛应用于小米小爱音箱、JBL Link Bar、Google Home Max、百度 Raven-H、Rokid Pebble 等全球主流品牌产品中。其智能商显解决方案针对教育、广告、金融等行业的交互显示需求，其高性能芯片（如 S905D3、T962X3、T972）结合了强大的 CPU/GPU/NPU 与丰富接口，为客户提供一站式解决方案。晶晨已成功演进为横跨家庭娱乐、智能视觉、语音交互、商业显示等多场景的平台型 AIoT 芯片设计公司，其多元化的业务布局为其打开了全新的增长空间。

**星宸科技（SigmaStar Technology）是在视觉感知这一更基础、更普适层级的定义者与领军者。**公司以全球视频安防芯片市场超过 40%的市占率稳居第一，其 AI SoC 芯片累计出货量已突破 5.5 亿颗，这标志着它已是端侧智能视觉领域无可争议的隐形冠军。与追逐通用算力竞赛的策略不同，星宸科技的成功在于其“深耕亿万小场景”的精准战略，即不盲目追求尖端制程的极限算力，而是聚焦于智慧安防、智能车载、机器人、AI 眼镜等海量细分市场，通过极致优化的软硬件协同，提供在功耗、成本、时延与易用性上最佳平衡的解决方案。公司的核心竞争力源于深度的场景化创新能力。全自研的 ISP（图像信号处理器）、NPU（神经网络处理器）及完整 AI 工具链，这使其能在成像质量、能效比和系统集成度上持续领先。例如，在智能机器人领域，全球每生产 3 台家用扫地机器人，就有 1 台采用其主控 SoC；在智能车载市场，其面向 L2 级辅助驾驶的芯片已实现规模化前装定点；在备受关注的 AI 眼镜赛道，其推出的 SSC309QL 芯片，通过 Chiplet

等创新设计，在实现相同录像规格时，整机功耗可比市场主流方案降低约 50%，为设备实现全天候佩戴提供了关键支持。

**星宸科技正从单一的视觉处理领导者，向感知、计算和连接的一体化平台迈进。**通过战略投资与内部研发，公司不仅持续升级智能视觉和机器人主控芯片，还积极布局车载激光雷达 SPAD 芯片、具备端侧大模型能力的边缘计算芯片等前沿领域。财务数据印证了其成长性：2025 年前三季度，公司实现营业收入 21.66 亿元，同比增长 19.50%，业务呈现健康扩张态势。因此，星宸科技代表了另一类成功的国产芯片范式，即凭借在垂直领域的技术深度与市场控制力，将单一优势横向复用于多个高速增长的 AIoT 赛道，从而在端侧智能市场中持续开拓领域。

**图84：星宸科技产品及终端应用**

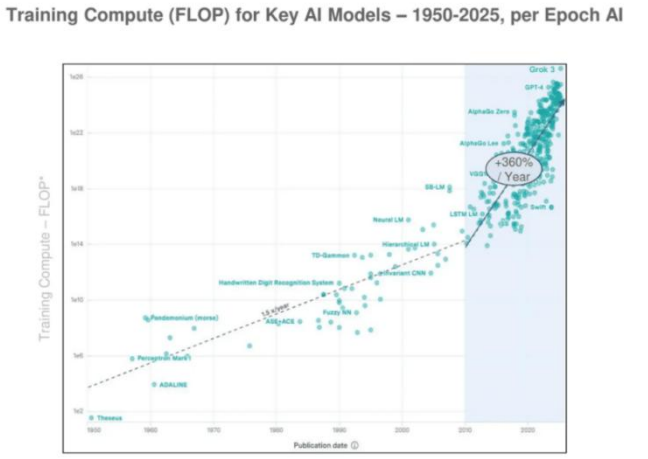

数据来源：公司官网，公司公告，爱集微，东吴证券研究所

**在泛 IoT 市场格局中，全志科技、乐鑫信息、炬芯科技、泰凌微电子、翱捷科技等公司是在特定赛道具备核心技术优势的新兴力量，凭借场景化芯片设计与成本优势覆盖多元细分需求，共同构成了国产芯片产业的多元化生态。**全志科技是国内老牌多媒体 SoC 设计商，其芯片在车载信息娱乐、智能家居、高清视频处理等领域应用广泛，车规级芯片已实现大规模前装量产。乐鑫信息是全球 Wi-Fi 与蓝牙双模 MCU 领域的领导者，其 ESP 系列芯片凭借极高的性价比和活跃的开源生态，成为全球数百万物联网开发项目的首选，在智能家居、穿戴设备中深度卡位。炬芯科技长期深耕音频 SoC，其产品在蓝牙音箱、智能手表、AR 眼镜等对音质与低功耗有高要求的设备中占据重要份额。泰凌微电子在低功耗无线连接芯片领域深耕，其蓝牙、Zigbee 等产品在键盘鼠标、遥控器、电子价签等需要超长续航的细分市场表现突出。翱捷科技作为少数具备全制式蜂窝物联网芯片研发能力的厂商，其 Cat.1 和 5G RedCap 芯片为共享经济、资产追踪、工业互联等需要广域连接的场景提供了关键解决方案。这些公司在各自擅长的领域构筑了纵深防线，共同推动着国产替代从点的突破走向面的普及。

### 4.2.4 AI 化进展的 TinyML 趋势：极低功耗 IoT 芯片的 AI 能力突破

**TinyML 是适配极低资源端侧设备的轻量 AI 技术，既具备高增长的经济价值潜力，也在技术特性上与高算力 AI 形成差异化定位。**未来 5 年 TinyML 可在全球释放超 700 亿美元的经济价值，其中物流、制造/工业自动化领域分别贡献 280 亿美元、220 亿美元，对应市场的年均复合增长率（CAGR）高达 273%。它与依赖智能设备、工业网关等硬件的 EdgeAI 差异显著，TinyML 适配 MCU、传感器节点等仅含几十 KB 至 1MB 内存的极低功耗设备，模型规模仅几千到几十万参数、体积几十 KB 至 1MB，聚焦声音唤醒、事件检测等轻量智能能力，支持端原生推理与超低功耗断网运行；而在 AI 训练算力增速高速扩张的行业背景下，TinyML 恰好填补了低资源端侧设备的智能需求空白，成为泛 IoT 芯片 AI 化的核心方向之一。

**图85：训练对算力资源要求的激增**




数据来源：BondCapital，东吴证券研究所

**图86：未来五年 TinyML 市场规模增速**


数据来源：ABIResearch，边缘 AI 报告，东吴证券研究所

**TinyML 技术已突破极低功耗 IoT 芯片（含 MCU）的 AI 部署瓶颈，实现轻量级 AI 模型在资源受限硬件上的稳定运行。**TinyML 的核心逻辑是通过模型压缩、算法优化，将传统需 GB 级内存、GPU 算力支撑的 AI 模型，适配至仅具备几 KBRAM、几 MHz 主频的微控制器（MCU）。这类芯片是泛 IoT 设备的核心硬件，广泛用于智能穿戴、工业传感器、智能家居终端，成本低至 2 美元，功耗仅为毫瓦级，相较云端 AI 的瓦级功耗降低约 1000 倍，解决了边缘设备低功耗需求与 AI 算力诉求的矛盾。传统 MCU 仅能完成基础控制功能（如家电开关、温度监测），而搭载 TinyML 的 MCU 可本地运行 AI 模型，无需依赖云端数据传输，延迟缩短至毫秒级，同时避免了网络波动对功能稳定性的影响。

图87：EdgeAI 与 TinyML 对比

| | Edge AI (如EmbeddingGemma) | TinyML |
|---|---|---|
| 硬件资源 | 智能手机、平板、PC、车载主机、工业网关<br>内存：100MB~1GB+<br>CPU/NPU/TPU等专用加速器 | MCU、传感器节点、可穿戴、嵌入式设备<br>内存：几十KB~1MB<br>极低功耗，无专用AI芯片 |
| 模型规模 | 千万~数亿参数<br>如EmbeddingGemma 3.08亿参数，<br>模型体积180MB~500MB | 几千~几十万参数<br>模型体积仅几十KB~1MB |
| 智能能力 | 复杂语义理解、嵌入生成、本地知识检索、<br>RAG、自然语言问答、多模态 | 事件检测、声音/图像/动作简单分类、<br>信号处理、状态监测、异常报警 |
| 应用场景 | 本地文档/知识库检索　智能助手<br>端侧RAG　多语种搜索<br>工业网关数据分析　隐私合规AI | 农作物病害检测　健康/心率监测<br>设备异常检测　环境/野生动物监测<br>可穿戴运动设别　声音唤醒 |
| 部署方式 | 较强终端本地推理，支持主流AI框架<br>云-端协同开发与管理 | 设备端原生推理无需操作系统，<br>支持超低功耗断网运行 |

数据来源：ABIResearch，边缘 AI 报告，东吴证券研究所

**TinyML 的典型应用已覆盖泛 IoT 多场景，轻量级 AI 模型落地消费、医学与工业领域。**在消费级场景中，人脸唤醒功能已批量应用于智能音箱、智能门锁等设备：搭载 TinyML 优化模型的 MCU，可在本地完成人脸特征提取与比对，响应时间低，误识别率低，单设备日均功耗低；在医学领域，北大人工智能研究院燕博南团队与合作者成功研制出世界首款大规模全柔性存算一体 AI 芯片，有望为可穿戴健康监测设备、柔性机器人等智能应用提供关键硬件支撑；在工业场景中，TinyML 赋能工业传感器实现故障预判，例如对电机振动数据的 AI 分析，可提前 24 小时预警设备异常，这类工业级 AIMCU 的部署成本仅为传统工业 SoC 的 1/5，且功耗降低 70%，已在扫地机器人、智能电表等泛 IoT 设备中规模化应用。2025 年高通通过收购 TinyML 领域核心企业 Edge Impulse 加速布局，强化自身在边缘 AI 与低功耗物联网场景的技术壁垒与市场竞争力。该平台已吸引超 17 万开发者入驻，支持超 45 万个嵌入式 AI 项目，客户覆盖意法半导体、恩智浦等芯片厂商及工业、医疗领域头部企业，高通通过整合其工具链可快速补足边缘 AI 开发能力，重点瞄准工业自动化、智慧医疗场景中低功耗设备的 AI 部署需求。

### 4.2.5 算力提升：芯片厂商角逐高阶 AI 的战略焦点

**面对机器人、高级辅助驾驶等需要实时多模态感知与决策的蓝海市场，提升核心算力已成为芯片厂商的战略焦点。**为支撑实时感知、融合与决策，头部厂商正全力推动 AI

算力从现有水平向更高数量级跃进，这场算力升级的本质，是为高阶智能应用铺设不可或缺的底层硬件基石。以瑞芯微、星宸科技为代表的头部厂商，正通过提升算力密度、优化芯片架构，积极向高阶端侧 AI 领域进军。

**芯片算力将从数 TOPS 向数十 TOPS 发展。**国内芯片公司不只是聚焦于泛化的芯片销售，而是与细分领域的头部终端厂商进行深度协同开发，因此对算力提出了更高的要求。在机器人领域，瑞芯微、星宸科技正与机器人整机企业合作，提供不仅算力充足，更在功耗、实时性、多传感器接口等方面深度定制的 SoC 方案。目前高通的旗舰 SoC AI 算力已突破 80 TOPS，瑞芯微的目前量产的 RK3588 芯片算力为 6 TOPS。为了抢占机器人等更高阶的市场，其下一代旗舰平台的目标算力为 30 TOPS。公司还探索 片外扩展专用 NPU”的方案，直接、高效地补齐复杂场景所需的算力缺口。这种算力的军备竞赛本质上是为了满足具身智能中多传感器融合、低延迟实时决策的刚性需求。

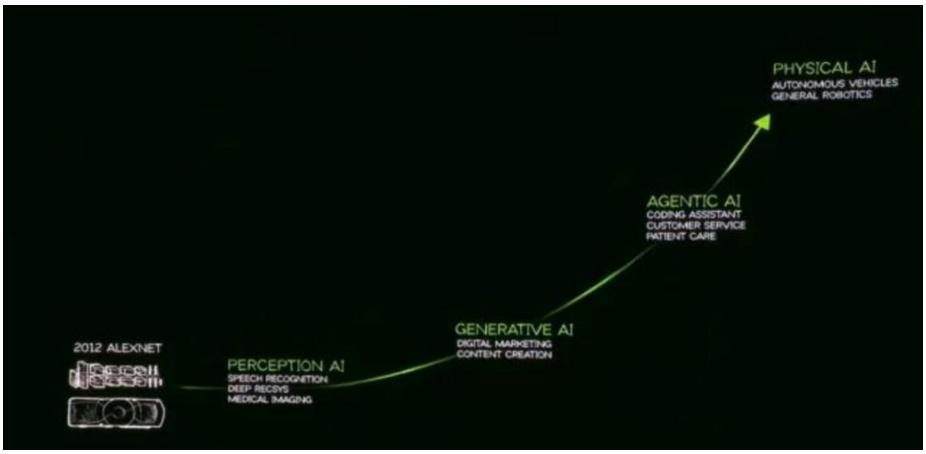
## 4.3 具身智能与物理 AI（Physical AI）：算力的新物种

摩根士丹利研报预测 2026 年中国人形机器人销量将实现翻倍增长，由 1.4 万台提升至 **2.8 万台**，产业正步入从研发原型向跨行业多场景验证转换的关键窗口期。

### 4.3.1 概念定义与架构剧变：从数字思维到物理交互的范式革命

**Physical AI 的核心在于突破传统人工智能的虚拟边界，赋予其理解物理法则、感知三维环境并安全执行复杂操作的能力，从而实现从“数字思维”到“物理交互”的范式革命。**这一定义远超出让机器人完成预设动作，而是要求 AI 系统能像人类一样，在一个动态、不确定的真实世界中进行实时推理与决策。其终极形态是成为具备长期记忆、情境理解乃至情感共鸣的“数字生命形态”，而不仅仅是执行指令的工具。

**图88：AI 正经历从 Agentic AI 到 Physical AI 的演进**




数据来源：英伟达 GTC 2025 大会，东吴证券研究所

目前，业界主要沿世界模型与视觉-语言-动作模型两条技术路径推进，前者专注于环境预测与因果推理，后者解决具体的“理解与执行”问题，两者正呈现融合趋势，共同对底层算力提出前所未有的高要求。

**Physical AI 正从学术概念走向产业落地，这对芯片的异构计算能力、实时性和多模态并发处理、IO 吞吐量和实时性提出了全新挑战，SoC 架构从追求峰值算力（TOPS）转向确定性实时响应。**Sim-to-Real（仿真到现实）成为核心开发范式。机器人 SoC 需硬件支持 NVIDIA Isaac Lab、ROS 2 等仿真环境的部署，即让算法在虚拟世界中完成大量训练与测试后，能高效、安全地在实体芯片上部署运行，实现"云端训练，端侧直行"。在仿真环境中，开发者可低成本生成数百万种极端场景数据（如物体滑落、碰撞反弹），训练完成后通过模型压缩将算法部署至端侧芯片。这要求 SoC 具备既能在云端训练时作为推理加速器，又能在端侧以低功耗实时运行的双模式能力。多模态并发处理是 Physical AI 的硬件刚需。

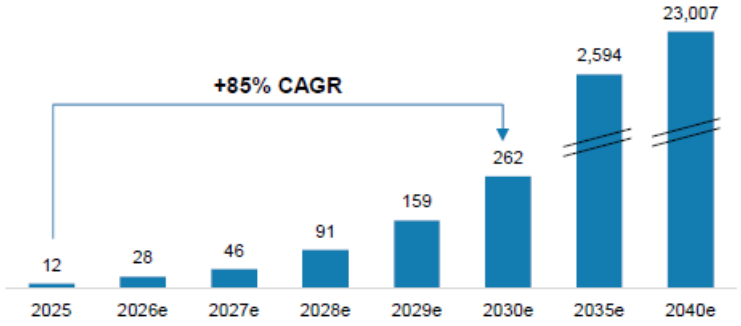
不同于手机 SoC 只需处理视觉信号，机器人 SoC 需同步处理高清视觉、激光雷达、麦克风阵列、关节编码器及高精度力/触觉传感器的海量数据流，这对芯片内部的数据搬运效率和输入输出（I/O）吞吐带宽构成了首要挑战。以人形机器人为例，从传感器数据采集到关节电机响应的端到端延迟需<10ms，而传统手机 SoC 的相机预览延迟约 50-100ms，相差一个数量级。此外，保障实时性要求芯片内部互联（NoC）具备强大的服务质量（QoS）保障和可预测的低延迟特性。最后，仿真到现实的迁移成为关键开发范式。这种架构剧变，使得机器人 SoC 的设计理念更接近高性能计算（HPC）与汽车功能安全（ASILD）的融合体。

### 4.3.2 市场空间：从实验室奇观到万亿美元产业的临界点

**受益于 2025 年人形机器人运动能力的快速提升，和 2026 年初高自由度灵巧手和机器人大脑的快速发展，2026 年具身智能有望进入规模化落地阶段。**机器人有望从替代结构化环境中重复劳动的“扫地僧”（清洁机器人），进化到能适应复杂非标任务的“蓝领工人”（人形机器人）。据 IDC 数据，2025 年全球人形机器人销售额约 4.4 亿美元，同比增长约 508%。

2026 年，行业正式从“技术验证”迈入“规模量产”关键期。根据摩根士丹利预测，2026 年中国人形机器人销量将翻倍至 2.8 万台，2030 年预估可达 26.2 万台。从厂商格局来看，中国厂商凭借完善的制造体系、快速迭代能力与成本优势，在全球市场中占据主导地位。

图89：中国具身机器人销售量预测（千台）




数据来源：Morgan Stanley Research，东吴证券研究所

**TAM 扩张的核心驱动力是成本下降，成本端的持续下降让商业化落地具备可行性。**若人形机器人整机 BOM 成本降至 2-3 万美元，其应用场景将从有限的工业检修、物流搬运，爆炸式渗透至商业服务、家庭辅助乃至医疗护理等领域，市场总规模将跃升至万亿级。当 BOM 成本降至 2-3 万美元，人形机器人的月使用成本（折旧+能源+维护）将低于蓝领工人月薪，触发大规模替代。特斯拉 Optimus 量产版通过核心部件国产化与供应链优化，成本已降至 1.8 万美元，较第二代产品下降 40%，进入工业与消费级市场可接受区间；场景端的深度渗透打开需求空间，工业领域的高精度装配、物流行业的智能搬运以及家庭场景的老年护理等刚性需求，正逐步被具身智能机器人满足。国内优必选 2025 年已斩获近 14 亿元人形机器人订单，交付超 500 台 Walker S2 工业级产品；政策端的强力支持构建良好发展环境，中国《机器人产业发展规划（2024-2027）》提出到 2027 年形成 3-5 家全球领先的机器人企业集团，核心零部件国产化率达到 70%以上，为产业发展提供政策保障。

**短期内，产业正沿着“工业落地”与“家庭渗透”两条清晰路径实现快速放量，为长期愿景积累数据、迭代技术和验证商业模式。**在工业与商用领域，2026 年 CES 显示，机器人已在物流分拣、精密装配、仓储搬运及医疗康复等场景展现出成熟的即战力。在家庭场景，行业领导者正从单一的扫地机器人，扩展到窗户清洁、园艺护理、泳池维护乃至情感陪伴的“全场景居家生态”矩阵。这两条路径的并行推进，不仅为上游芯片和零部件供应商提供了明确的短期订单，更在不断验证和拓宽物理 AI 的可行性边界，为其最终成为通用平台奠定基础。

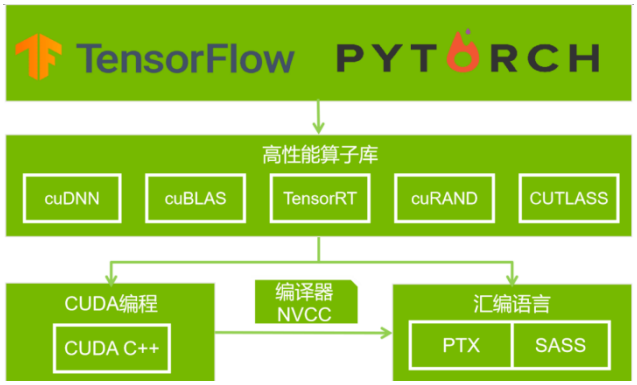
### 4.3.3 竞争格局：产业链协同与生态之争

全球 Physical AI 芯片市场呈现鲜明的梯队分化与路径竞争，核心已不是单纯的算力比拼，生态协同与垂直整合成为竞争核心。

**英伟达凭借其 CUDA 软件生态的优势垄断高端市场，以 Jetson Thor（2000 TOPS）和 Isaac 机器人平台为核心，构建了覆盖从仿真、训练到部署的全栈工具链。**作为英伟

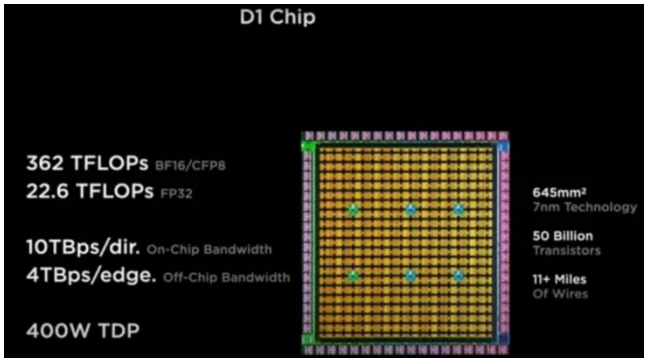
达专为物理 AI 打造的旗舰计算平台，Jetson Thor 基于最新 Blackwell 架构 GPU 构建，集成 2560 个 CUDA 核心与 96 颗第五代 Tensor 核心，实测 AI 算力达 2070 TFLOPS（FP4 稀疏计算格式），较前代产品性能提升 7.5 倍，能效比优化 3.5 倍，完美匹配人形机器人对高算力与低功耗的双重需求。其内置的专用 Transformer 引擎是核心技术突破，支持 Llama、Gemini、阿里 Qwen 等主流生成式 AI 模型在边缘端实时推理，同时兼容视觉语言动作模型（VLA）与视觉语言模型（VLM），能够将环境感知、动作规划与决策执行的端到端延迟压缩至 10 毫秒以内，为人形机器人提供接近人类的反应速度。英伟达通过 Isaac Sim 仿真平台提供数字孪生训练环境，借助 JetPack 7 SDK 实现云边端一体化开发流程，搭配 GR00T 基础模型降低开发者门槛，目前全球已有超过 500 家企业参与早期测试，客户覆盖波士顿动力、Agility Robotics、智元创新等头部机器人厂商，其中智元精灵 G2 工业级人形机器人依托 Jetson Thor 平台实现了复杂场景下的高精度作业。该平台量产模块批量采购单价 2999 美元，开发者套件起售价 3499 美元，通过规模化供应进一步巩固通用市场主导地位，2026Q1 在高端人形机器人芯片市场的市占率已达 68%。

**图90：Nvidia CUDA 生态**




数据来源：Nvidia 官网，东吴证券研究所

**图91：Tesla Dojo 芯片**




数据来源：Tesla 官网，东吴证券研究所

**Tesla 是垂直整合的代表，将其自动驾驶时代的 FSD 芯片（HW 4.0）与 Dojo 超算集群复用于 Optimus 机器人，通过芯片、算法、数据的全栈自研与深度协同，追求极致的迭代效率与成本控制，为核心自用场景服务。**特斯拉摒弃通用芯片路线，将自动驾驶领域成熟的 FSD 芯片直接复用于 Optimus 人形机器人，通过技术复用与规模效应实现极致成本控制。其最新的 FSD HW4.0 芯片延续 7nm 制程工艺，针对机器人场景优化了力控算法与运动控制单元的硬件耦合，能够通过 8 路摄像头实现三维环境识别，完美适配 Optimus 的 22 个手部自由度与复杂动作执行需求，单芯片算力达 1000 TOPS，足以支撑工业场景下的零件操作、产线补给等核心任务。垂直整合的核心优势体现在全产业链成本摊销。

FSD 芯片已实现年产能超 1000 万颗，机器人芯片与车端芯片共用生产线，大幅降低研发与制造成本；原 Dojo 超算项目积累的训练经验已迁移至 Cortex 超级集群，将单动作训练周期压缩，进一步提升迭代效率。尽管特斯拉已解散 Dojo 自研训练芯片团队，

转而采用由 6.7 万块 H100/H200 GPU 组成的 Cortex 集群，并与三星签署 165 亿美元 AI6 芯片采购协议，但车端芯片复用的核心战略并未改变，这种模式使 Optimus 量产版成本降至 1.8 万美元，较第二代产品下降 40%，其中芯片成本占比仅为 8%，远低于行业平均水平的 15-20%。

**在此格局下，国产厂商通过聚焦高确定性细分场景、依托本土供应链与快速工程化能力实现差异化突围。**地平线征程 7 芯片针对工业场景的高可靠性要求进行专项优化，通过本土供应链快速迭代，已成功适配优必选 Walker X 系列人形机器人，其 500-700 TOPS 算力能够满足工业装配、物流搬运等场景的实时决策需求，核心优势体现在极端环境下的稳定性与性价比。华为 MDC 910（1500 TOPS）借助智能驾驶领域的技术积累，实现跨场景技术复用，在车载机器人场景形成独特壁垒，其芯片-操作系统-应用生态”的完整布局能够为客户提供端到端解决方案，目前已与小鹏 Robotaxi 达成深度合作，完成从自动驾驶到机器人场景的技术迁移。

### 4.3.4 关键瓶颈与协同风险：硬件与软件的双重约束

物理 AI 产业的发展并非依赖 SoC 单一环节的算力突破，其效能释放与商业化进程，核心受制于上游核心硬件的性能边界与全产业链数据资源的稀缺性，构成当前产业发展的关键协同风险与核心瓶颈。

**从硬件端来看，感知器件与执行机构的性能约束构成机器人智能落地的核心物理限制。**尽管 SoC 算力已达到千 TOPS 级别，但感知与执行环节的技术短板直接限制算力效能的实际转化。一方面，高精度触觉与力觉传感器在灵敏度、耐久性及成本控制方面尚未实现根本性突破，导致机器人在精细操作场景中缺乏对接触力度、材质特性的精准感知，难以完成易碎品抓取、柔性装配等复杂任务。另一方面，灵巧手与关节执行器（包括无框力矩电机、空心杯电机等）仍处于技术与成本的双重高位，灵巧手成本占整机成本比例可达 25% 以上，其微型驱动、精密传动及自适应控制等核心技术门槛较高；高性能电机的转矩密度、响应速度及热管理能力等关键指标，国内供应链与国际领先水平仍存在明显差距。这一现状导致整机厂商在追求产品高性能时，对进口核心部件依赖度较高，不仅推高了 BOM 成本，更在供应链稳定性与产品迭代自主性方面形成长期潜在风险。

**从数据与算法端来看，模型训练的数据依赖特性构成智能泛化能力提升的根本性障碍，行业核心矛盾已从算力供给不足转向数据供给短缺。**与依托互联网海量文本、图像数据训练的大语言模型不同，物理 AI 模型训练所需数据为多模态高维度时空连续数据，涵盖视觉、力觉、触觉及本体感知等多通道同步信息。此类数据的获取与处理面临显著挑战：真实场景下的交互数据采集需搭建高成本复杂实验环境，且数据生成过程不可逆、试错成本较高；数据标注环节对专业知识要求较高，自动化实现难度大。因此头部企业积累的有效真实交互数据规模，与大模型训练数据量级存在显著差距，形成数据供给短缺的行业现状。更为关键的是，不同企业的数据格式缺乏统一标准，应用场景差

异较大，导致数据流通与共享难度较高，形成数据孤岛效应，进而制约通用机器人基础模型的研发进程，难以支撑模型在多元场景下的可靠应用。

**针对上述协同风险，行业正通过软硬协同路径推进突破。**硬件领域，国内供应链加速核心部件的国产化替代与技术攻关，同时积极探索柔性电子皮肤等新型传感器技术及准直驱关节等新型执行器方案。数据与算法领域，高保真仿真平台（如 NVIDIA Isaac Lab）已成为合成数据生成的核心工具，有效弥补真实场景数据供给不足的缺口；开源数据集与标准化协议的推广应用，正在推动数据流通共享，促进行业协同创新。但相关技术突破与生态建设仍需长期积累，算力、硬件与数据三者的有效协同及循环迭代，仍是物理 AI 产业发展面临的核心挑战。

# 5. 互联网大厂构建端云协同闭环硬件生态，筑牢向 AI 转型的硬件底座

26 年春节期间，阿里、字节密集预热自研芯片为缩影，互联网巨头正加速筑牢 AI 硬件底座。无论是阿里与字节的“芯片至应用”全栈布局、腾讯的 AIoT 场景赋能，还是小米的“人车家”生态联动，其核心诉求高度一致：即通过端云协同闭环锁定物理入口与核心流量，确立长期的生态卡位优势。

## 5.1 阿里巴巴全面布局“云+AI+芯片”战略

### 5.1.1 芯片层：自研芯片矩阵品类齐全，实现“云端一体”全覆盖

**阿里平头哥已构建多品类自研芯片矩阵，高端 AI 训推一体芯片真武 810E 实现规模化落地，芯片层与云、大模型协同形成全栈 AI 体系。**平头哥是阿里巴巴集团全资半导体芯片业务的主体。其官网在 2026 年 1 月正式上线了高端 AI 训推一体芯片真武 810E。真武 810E 采用自研并行计算架构与 ICN 片间互联技术，单卡搭载 96GB HBM2e 内存，支持 7 个独立 ICN 链路，片间互联带宽可达 700GB/s，可灵活配置多卡组合，且已在阿里云实现多个万卡集群部署。阿里巴巴已将“真武”PPU 大规模应用于千问大模型的训练与推理，并结合阿里云完整的 AI 软件栈进行深度优化，为客户提供一体化产品与服务. 目前，平头哥实施云端一体战略，产品从云端 AI 芯片到端侧 SoC 全覆盖，除训推一体 AI 加速芯片真武 810E 外，还涵盖 AI 推理芯片含光 800、Arm 服务器 CPU 倚天 710、高性能 SSD 主控芯片镇岳 510，以及超高频 RFID 电子标签芯片羽阵 611 和羽阵 600 等，产品线持续丰富。至此，由通义实验室、阿里云、平头哥组成的阿里巴巴 AI 黄金三角“通云哥”正式亮相。阿里巴巴正将“通云哥”打造为一台 AI 超级计算机，它不仅拥有全栈自研芯片的平头哥、亚太领先的阿里云，还具备开源模型“千问”，能够在芯片架构、云平台架构与模型架构上实现协同创新，进而确保在阿里云上训练和调用大模型时达到最高效率。

**图92：阿里平头哥芯片产品详情**

| 类别 | 具体产品 | 产品优势 | 技术规格 | 应用场景 | 合作伙伴 |
|---|---|---|---|---|---|
| 人工智能芯片 | 真武 810E | 全自研、高性能、强互联、高易用、规模验证、芯云一体 | **ICN 片间互联** 700 GB/s<br>**内存** 96GB HBM2e<br>**Host 总线** PCIe5.0 x 16 | 自动驾驶、AI 训练、AI 推理、多模态模型 | - |
| | 含光 800 | 全自研、高性能、高能效、深度优化 | **峰值算力** 820TOPS<br>**推理性能** 78563IPS<br>**能效比** 500IPS/W<br>**完整软件栈**<br>支持 TensorFlow，Caffe，MXNet 和 ONNX 等主流深度学习框架 | 电商营销、电商搜索、AI 推理 | - |
| 服务器 CPU | 倚天 710 | 高性能、高宽带、高能效、高可靠 | **CPU 架构** Arm v9<br>**核心数** 128 核<br>**主频** 2.75 GHz<br>**内存** 8 通道 DDR5<br>**I/O** 96 通道 PCIe 5.0<br>**最大功率** 250 W | AI 推理、电子商务、大数据分析、视频解编码 | - |
| SSD 主控芯片 | 镇岳 510 | 高宽带、低时延、高可靠、高能效、低成本、大容量 | **IO 处理能力** 3400KIOPS<br>**时延** 4 μs<br>**误码率** $10^{-18}$<br>**高速接口** PCIe 5.0 / DDR 5.0<br>**能效比** 420K IOPS/watt | 在线交易、分布式存储、AI 推理、AI 训练 | 亿恒创源、Biwin、得瑞领新、长江万润半导体 |
| 超高频 RFID 芯片 | 羽阵 611 | 高灵敏度、高一致性、高可靠性、强稳定性 | **协议**<br>EPCglobal G2 V2<br>ISO/IEC18000-6C<br>**灵敏度**<br>读取灵敏度-24dBm<br>写入灵敏度-20dBm<br>**ESD 防静电性能** HBM±10KV<br>**操作温度范围** -40℃ ~ +85℃<br>**芯片存储**<br>96-bit TID / 128-bit EPC | 供应链管理、智慧物流、快销品零售、鞋服零售 | ARIZON、菜鸟、信达物联、上扬无线射频科技 |
| | 羽阵 610 | 高灵敏度、强环境适应性、全方位读取、高可靠性 | **协议**<br>EPCglobal G2 V2 / ISO/IEC18000-6C<br>**灵敏度**<br>单端口读取灵敏度-21dBm<br>双端口优于-22.5dBm<br>**芯片存储** 96-bit EPC<br>**操作温度范围** -40℃ ~ +85℃<br>**ESD 防静电性能** HBM±2KV | 仓储物流、智慧零售、供应链管理 | ARIZON、菜鸟、信达物联、上扬无线射频科技 |

数据来源：平头哥官网，东吴证券研究所

### 5.1.2 模型层：多模态适配，各场景全参数覆盖，性能对标国际顶尖

**阿里云 Qwen 系列作为底座，各业务线模型实现垂直化定制。**阿里云 Qwen 系列覆盖文本生成、图像生成、语音合成、语音识别、视频合成以及推理等基础功能，成为各业务线的技术底座。与此同时，夸克、蚂蚁、淘宝、阿里大文娱等团队的 AI 模型均与自身业务场景深度绑定，助力垂直场景化定制。

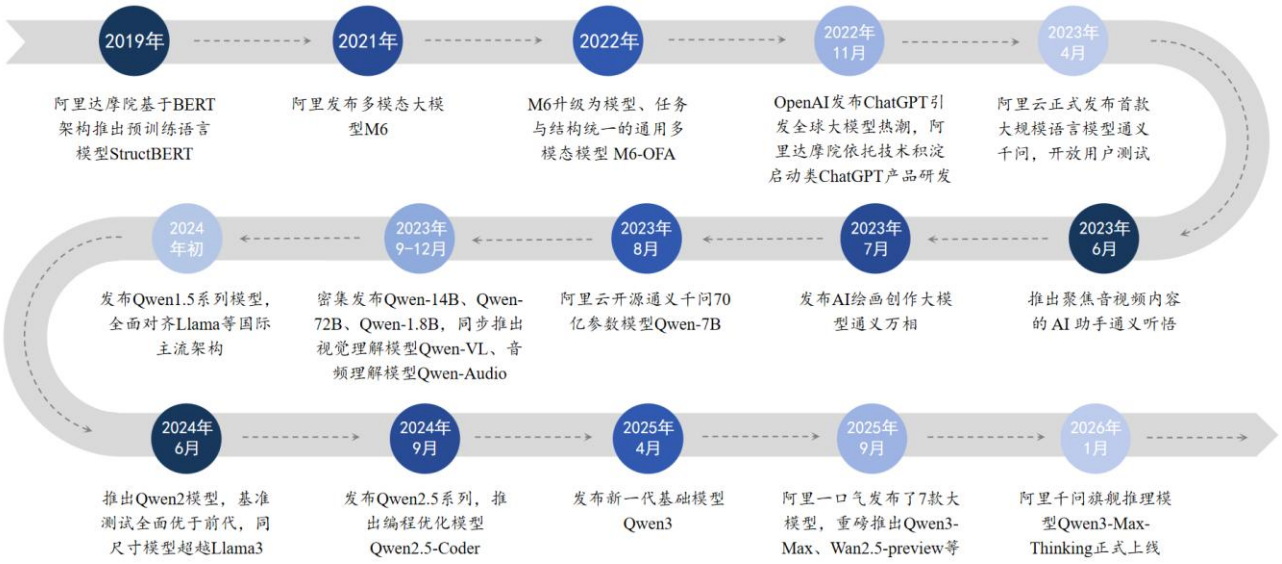
**图93：阿里 AI 基础模型总布局**

| 团队 | 类型 | 产品 |
|---|---|---|
| 阿里云 | 文本生成模型 | qwen-max/plus/turbo/Long |
| | | qwen-math（数学） |
| | | qwen-coder（代码） |
| | | qwen-vl（图像理解） |
| | | qwen-vl-ocr（OCR） |
| | | qwen-audio（音频理解） |
| | 图像生成与修改模型 | wanx 系列（通用生成/涂鸦作画/布局重绘/背景生成/动漫人物生成/虚拟模特/创意海报生成/图配文） |
| | | image 系列（画面扩展/实例分割/擦除补全） |
| | 语音合成模型 | cosvoice/sambert（文本转语音） |
| | 语音识别模型 | paraformer/sensevoice（语音转文本） |
| | 视频合成模型 | emo/liveportrait/animatemotion（人像） |
| | | video-style-transform（视频编辑与生成） |
| | 推理模型 | QwQ（推理） |
| | | QvQ（多模态推理） |
| 夸克 | 多模态大模型 | 夸克大模型（通识/搜索） |
| | | 灵知大模型（学习） |
| 蚂蚁 | 垂直场景 | 高考志愿大模型 |
| | 多模态大模型 | 百灵大模型（垂直行业） |
| 淘宝 | 多模态大模型 | 星辰大模型（电商生活服务） |
| 阿里大文娱 | 图像生成 | 神力觅裳（文生图） |
| 阿里云通义&鹿机机器人 | 具身智能大模型 | LPLM-10B |

数据来源：阿里云，通义实验室，夸克，蚂蚁，东吴证券研究所

**阿里通义大模型发展历经多年技术积淀与迭代升级，兼具全球影响力与中国市场领导地位。**通义大模型的诞生并非一蹴而就，其背后是阿里在 AI 领域的长期投入和战略演进。

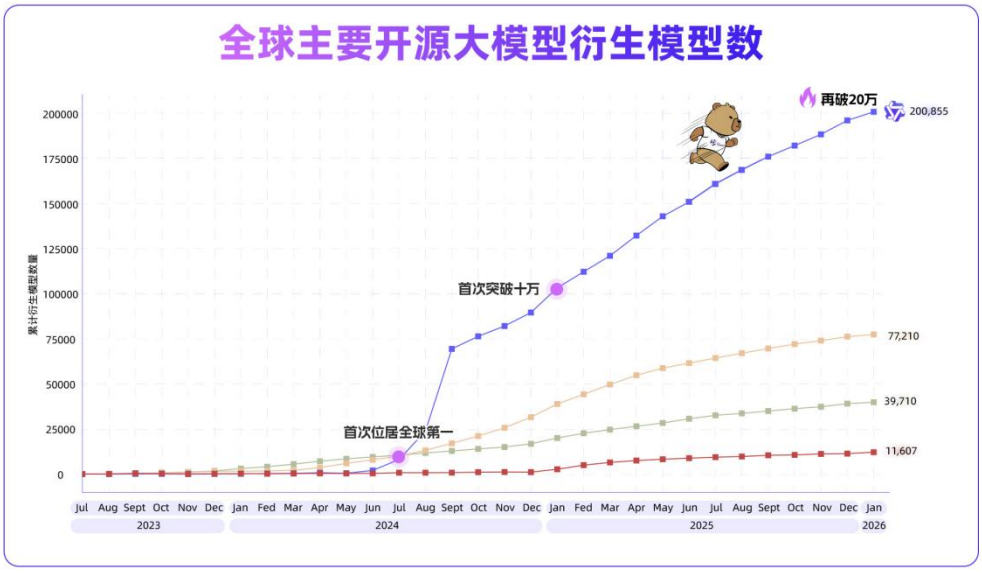
**图94：通义大模型发展历程**




数据来源：阿里云官网，财联社公众号，机器之心公众号，二进制法研社公众号，东吴证券研究所

**通义大模型在全球发展进程中斩获亮眼成绩，开源生态与企业级落地均实现行业领跑。**开源层面，通义大模型多次登顶全球开源榜单。旗下 Qwen 已于 2025 年 10 月在 HuggingFace 平台累计下载量上超越 Llama 位列全球第一；截至 2026 年 1 月底，Qwen 系列衍生模型数量超 20 万，同时，整体下载量突破 10 亿，日均开发者下载量达 110 万次；2026 年 2 月，千问模型的开源数量超 400 款，包揽开源榜单前十多数席位。企业级市场上，通义大模型拿下中国市场调用量第一，已服务全球企业级客户超 100 万。凭借全球开源的领先优势与企业级市场的龙头地位，通义大模型奠定了兼具全球影响力与中国市场领导者的双重地位。

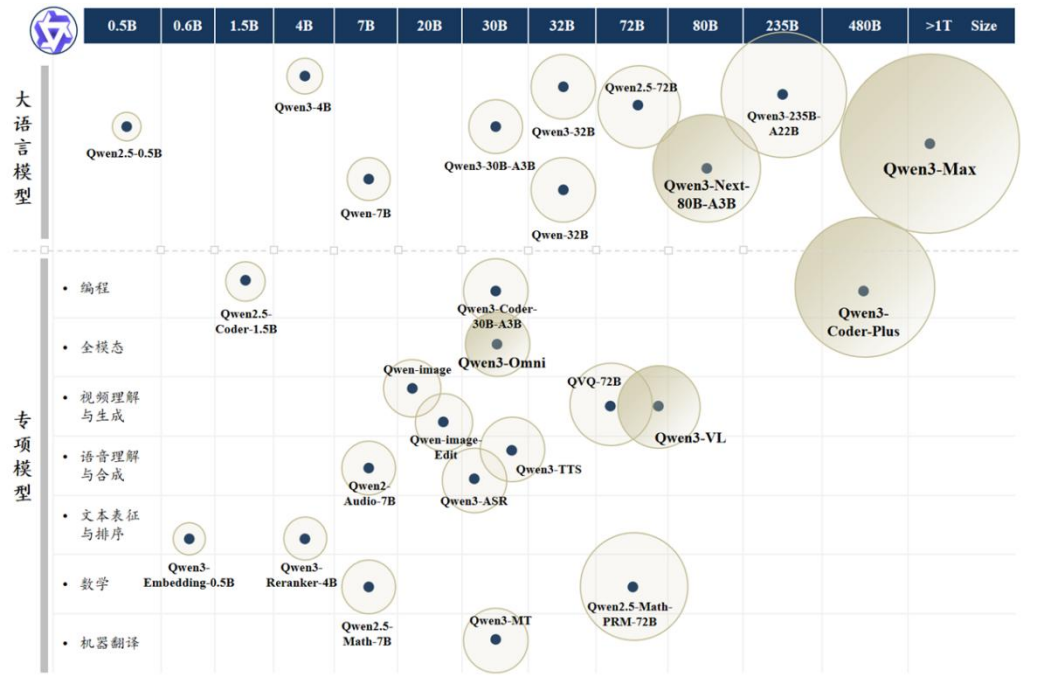
**图95：全球主要开源大模型衍生模型数情况**




数据来源：千问 Qwen 公众号，东吴证券研究所

**通义千问端侧突破万亿级，比肩国际顶级闭源模型，多任务能力突出。**通义千问大语言模型凭借超万亿参数规模的预训练基础，具备自然语言理解、文本生成、视觉理解、音频理解、工具使用、角色扮演、AI Agent互动等完备能力。阿里云为其构建了覆盖全场景的参数体系，整体规模横跨5亿到10000亿+，其中0.5B、1.8B、4B、7B、20B、30B的小尺寸模型可满足手机、PC等端侧设备的轻量化部署需求，72B、235B、480B以及万亿参数以上的大尺寸模型，则能够为企业落地与科研探索提供高性能支撑。作为通义团队迄今打造的规模最大、能力最强的旗舰型号，Qwen3-Max-Thinking总参数规模突破万亿，跻身全球最大规模AI模型行列，配套的预训练数据量高达36T Tokens且囊括海量高质量语料。该模型在多项权威基准测试中表现出众，不仅在包含事实科学知识、复杂推理、编程能力的19项核心测试中达到顶尖水准，多项指标达成或刷新全球SOTA，综合性能更可与GPT-5.2-Thinking、Claude-Opus-4.5、Gemini-3 Pro等国际顶级闭源模型抗衡乃至实现超越。

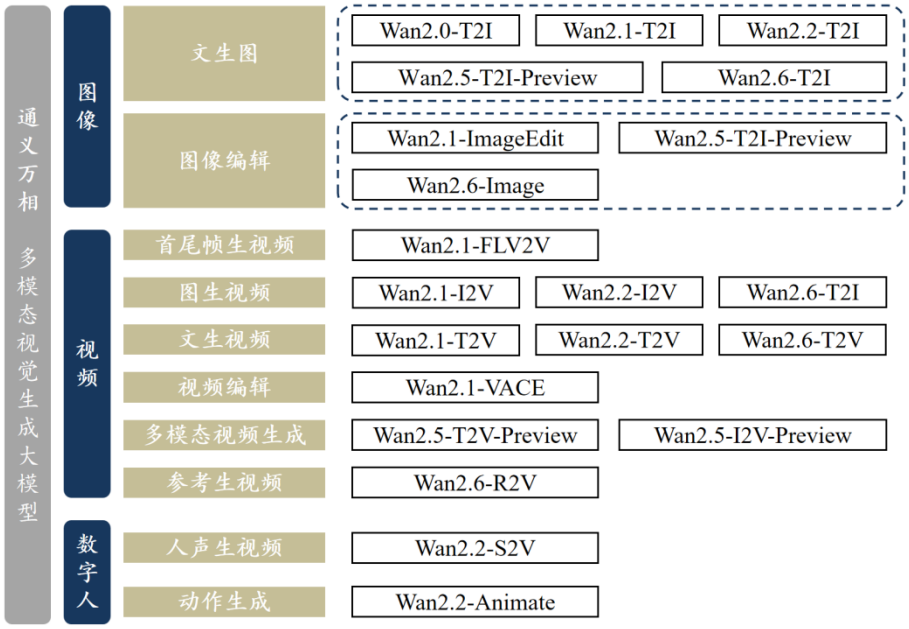
**图96：通义千问Qwen模型家族**




数据来源：阿里云，东吴证券研究所

**通义万相则实现多模态全维度一致生成突破，成为国内顶尖、全球一流的视觉生成模型家族。**通义万相视觉生成大模型基于原生多模态统一框架训练而成，支持图像、视频、声音等多模态内容生成，在画面质量、语义理解、运动幅度、物理规律遵循及艺术质感等核心维度均达到行业领先水平。作为目前全球功能覆盖最全面的视频生成模型家族之一，通义万相2.6系列涵盖文生视频、图生视频、参考生视频三大核心场景，共包含5款图像生成与文生图专项模型，更成为全球唯二、国内首个具备“参考生视频”能力的模型——用户不仅可通过“角色参考”功能固定视频中人或物的IP形象，还能提取输入视频的音色特征，实现从画面到声音的全维度复刻，轻松驾驭单人独白、双人对手戏等多元创作场景。通义实验室进一步强化万相大模型画质精细度、音效还原度与指令遵循准确性，其中单次视频生成时长达到国内最高的15秒。

图97：通义万相 Wan 模型家族




数据来源：阿里云公众号，东吴证券研究所

**通义百聆依托自研双引擎，适配多元行业场景。**通义百聆语音大模型依托自研 Fun-ASR 与 Fun-CosyVoice 两大引擎，支持多类语及方言，精准识别嘈杂环境、专业术语及混合语种，实现低延迟高准确率转写，从而提供自然流畅、情感丰富的语音识别及合成能力。其中，Fun-ASR 基于数千万小时真实语音数据训练而成，具备强大的上下文理解能力与行业适应性；Fun-CosyVoice 可提供上百种预制音色，可以用于客服、销售、直播电商、消费电子、有声书、儿童娱乐等场景。

图98：通义百聆语音大模型系列


数据来源：通义实验室官网，东吴证券研究所

### 5.1.3 应用层：筑牢 B 端壁垒，抢占 C 端先机，阿里 AI 双线齐发

**AI to C 恰合时宜，阿里深耕 C 端超级入口，依托模型和生态两大核心优势，进行**

**全方位 C 端产品布局与推广。**当前终端用户需求保持旺盛态势，后续市场增长潜力仍处于加速释放阶段。在面向消费者的 AI 赛道上，阿里向 C 端集中发力，加速构建全场景的 C 端产品矩阵。组织架构上，阿里成立智能信息事业群，现已升级为“千问 to C 事业群”，整合千问 APP、夸克、AI 硬件、UC、书旗等业务；硬件布局上，阿里发布“夸克 AI 眼镜”，探索 AI 硬件入口；核心应用上，阿里重新包装并大力推广千问 APP，将夸克学习升级为千问智学，将 AQ 升级为蚂蚁阿福。经过多年布局，阿里已搭建起覆盖电商、出行、支付、酒旅等场景的完整本地生活生态体系，深度渗透用户衣食住行等核心生活场景。在大模型性能与稳定性达标、可支撑用户规模化日常使用的基础上，阿里整合集团全域资源，实现本地生活生态与千问 App 的全面打通。目前千问 App 已可执行 400 余项核心任务，业务场景覆盖应用开发、Office 办公、学习辅导、咨询调研、数据分析、可视化报表等。

**图99：阿里巴巴 C 端产品矩阵**


数据来源：阿里巴巴官网，阿里云公众号，科技行者公众号，东吴证券研究所

**在阿里的代表性智能硬件方面，夸克 AI 眼镜 S1 作为阿里巴巴首款自研硬件，凭借极致轻量化设计、突破性显示技术与全链路生态闭环，成为千问大模型落地物理世界的理想载体，标志着阿里正式完成对下一代个人移动入口的布局。**阿里巴巴旗下首款自研硬件夸克 AI 眼镜 S1 在 2025 年 11 月正式发布。这款产品核心亮点在于极致轻量化与显示功能的兼顾：采用钛合金一体双料注塑材质搭配亲肤 PU 漆，通过与嘉联益联合首创的 7 层柔性电路板技术，将内置主板、双芯片等核心部件的镜腿压缩至全球最窄 7.55 毫米，整机重量控制在 51 克，配合 1:1 前后均衡配重、大叶仿生鼻托与超薄仿生耳弯设计，连续佩戴 2 小时无明显压痕与酸痛感。在显示技术上，其搭载 JBD 提供的 0.15 立方厘米微型 Micro-LED 光引擎，结合至格科技定制光波导，实现 4000 尼特高亮度，无

彩虹纹与漏光问题，且与康耐特合作推出一体化近视镜片，解决近视用户痛点。交互与生态方面，它支持双向视觉交互，联动支付宝、高德地图、拍立淘等阿里生态服务，借助千问大模型与自研 MasterAgent 中控，实现刷眼支付、近眼导航、所见即搜等高效功能，多意图指令处理能力出色。夸克 AI 眼镜 S1 的算力、显示、感知等系统设计让它成为千问大模型在物理世界中的理想载体。该产品构建了从云计算、AI 大模型到硬件终端与应用服务的全链路闭环，软硬件全自研模式推动了生态融合与场景变革。随着这款眼镜的推出，阿里已完成生活、办公场景及眼镜这个下一代个人移动入口的布局。

**图100：夸克 AI 眼镜 S1 功能矩阵**


数据来源：夸克 AI 眼镜公众号，东吴证券研究所

**AI to B 增长动能持续增强，阿里云+钉钉+菜鸟供应链形成极深护城河。**在 B 端市场竞争中，阿里巴巴并未依托独 APP 门户开展业务布局，而是构建起阿里云（算力底座）+钉钉（组织协同）+菜鸟（供应链履约）三位一体的全链路企业服务体系。针对传统制造、中大型企业及政务领域的数字化需求，阿里巴巴的服务范畴可深度覆盖生产设备数据接入、库存管理、物流对接等全流程产业数字化场景，在企业核心经营链路的数字化落地具备突出优势。在面向中大型传统企业、制造业与政务领域的企业效率提升场景中，阿里巴巴已构筑起深厚的竞争护城河，具备承载复杂商业逻辑、支撑企业全链路数字化转型的核心能力。

**阿里云坐拥全栈自研的云与大模型技术及协同优势，营收与 AI 相关产品均实现高速、持续增长。**阿里云已成为全球少数同时实现大模型与云计算全栈自研、技术能力行业领先的全栈人工智能提供商。大模型领域，通义模型家族收获全球开发者的广泛认可，现已成为全球规模最大的模型家族；云计算领域，阿里云凭借多年技术深耕与布局积累，已发展为全球领先、亚太地区排名第一的云计算服务商。两大核心业务板块相辅相成、深度协同，阿里云通过云计算架构的持续升级，搭建起 IaaS 层、PaaS 层、MaaS 层三层技术体系，并依托软硬件一体化的联合优化，为通义大模型的持续迭代进化提供支撑，也由此巩固了自身在全栈人工智能领域的领先地位。2026 财年第二季度阿里云季度营收同比增长 34%，主要由公共云收入的增长驱动，其中包括 AI 相关产品的采用率持续提升。AI 相关产品收入继续保持强劲增长态势，连续 9 个季度实现三位数的同比增长。

**阿里云已构建完整的 AI 基础设施，具备大规模、高性价比的优势。**AI 与云计算的深度融合对阿里云基础设施提出多维升级要求，新一代阿里云基础设施聚焦三大核心方向：具备大规模、高性价比特性，满足 AI 训练与推理核心需求；具备高性能、高可用特性，支撑业务扩张并保障连续性；具备易用性、智能化特性，提升开发与运维全流程效率。阿里云构建了全栈式 AI 基础设施层体系，底层以计算、存储、网络与 CDN、开发工具提供基础支撑，中层通过容器化技术实现算力统一编排，上层依托 PAI 智算平台完成调度、编译优化与弹性伸缩、容错、迁移等能力落地。通过全栈技术优化，阿里云 AI 基础设施连续训练有效时长达 99%，处于业界第一梯队；同时依托迁移与运维管理等环节优化，GPU 算力利用率 MFU 提升超 20%，算力转化效率与基础设施综合性能得到显著增强。

**图101：阿里云基础设施层布局**

| 层级 | 一级分类 | 二级分类 | 具体产品 |
|---|---|---|---|
| 基础设施层 | 计算 | 云服务器 | 云服务器 ECS，弹性容器实例 ECI，弹性伸缩，云盒，云虚拟主机，计算巢服务，GPU 云服务器，弹性裸金属服务器，弹性加速计算实例，专有宿主机，轻量应用服务器 |
| | | 高性能计算 | 弹性高性能计算 E-HPC |
| | | Serverless | Serverless 应用引擎 SAE |
| | | | 函数计算 FC |
| | | 边缘计算 | 边缘节点服务 ENS，边缘网络加速 ENA |
| | | 操作系统 | Alibaba Cloud Linux，龙蜥操作系统 |
| | 容器 | 容器服务 | 容器服务 Kubernetes 版，容器服务 Serverless 版 ACK Serverless，容器计算服务 ACS，分布式云管平台 ACK One，容器镜像服务，服务网格 ASM |
| | 存储 | 基础存储服务 | 对象存储 OSS，块存储 EBS，文件存储 NAS，文件存储 CPFS，表格存储 Tablestore |
| | | 存储数据服务 | 云备份，数据灾备中心 BDRC，智能媒体管理 IMM，网盘与相册服务 PDS |
| | | 混合云存储 | 混合云存储，混合云容灾服务 HDR |
| | | 数据迁移与工具 | 闪电立方 Transport，云存储网关 |
| | 网络与 CDN | 云上网络 | 负载均衡 SLB，专有网络 VPC，弹性公网 IP EIP，NAT 网关，云数据传输 CDT，私网连接 PrivateLink，共享带宽，云解析 PrivateZone，共享流量包，网络智能服务 NIS |
| | | 跨地域网络 | 全球加速 GA，云企业网 CEN，转发路由器 TR |
| | | 混合云网络 | VPN 网关，高速通道 Express Connect，智能接入网关 5AG |
| | | 内容分发网络 CDN | CDN，边缘安全加速 ESA |
| | 开发工具 | API 与工具 | 资源编排，云命令行，OpenAPI Explorer，云控制 API，Dragonwell |
| | | 开发与运维 | 云工作流 CloudFlow，云原生应用开发平台 CAP，移动研发平台 EMAS，多端低代码开发平台魔笔，云原生应用组装平台 BizWorks |
| | 迁移与运维管理 | 运维与监控 | 日志服务 SLS，系统运维管理 OOS，云监控，云网管，智能顾问，运维事件中心，应用诊断分析平台 |
| | | 云管理 | 操作审计，访问控制 RAM，配置审计，资源管理，配额中心，云治理中心，云速搭 |
| | | 备份与迁移 | 服务器迁移中心，云迁移中心 |

数据来源：阿里云公众号，东吴证券研究所

**人工智能与机器学习层智算升级，依托全栈方案与 PAI 平台服务，满足政企“云 + AI”发展需求。**阿里云人工智能与机器学习层已完成智算升级，面向政企客户推出全栈式 AI 解决方案，精准匹配其“云+AI”协同发展的核心诉求，覆盖模型训练、推理部署

到智能体开发的 AI 全生命周期流程。产品深度融合通义千问大模型、AI 原生技术与软硬一体化 AI 解决方案，涵盖基础大模型应用服务和智能语音交互、自然语言处理等开放服务。阿里云人工智能平台 PAI 面向企业用户提供涵盖数据预处理、仿真数据生成、模型训练评估、机器人强化学习、仿真测试的全链路一体化平台服务，可为具身智能、辅助驾驶等场景化 AI 应用提供全流程技术支撑，有效缩短相关应用的研发落地周期。

**图102：阿里云人工智能与机器学习层布局**

| 层级 | 一级分类 | 二级分类 | 具体产品 |
|---|---|---|---|
| 人工智能与机器学习层 | 大模型服务与应用 | AI 应用 | 通义晓蜜，通义灵码，通义听悟，通义星尘，虚拟数字人 |
| | | 行业智能 | 交通云控平台，工业大脑 |
| | | 基础大模型 | 通义千问，通义万相，通义百聆 |
| | | 模型服务平台 | 大模型服务平台百炼 |
| | 开放服务 | 智能语音交互 | 录音文件识别，实时语音识别，语音合成 |
| | | 自然语言处理 | 自然语言处理 NLP，文档智能，机器翻译 |
| | | 视觉智能 | 文字识别 OCR，视觉计算服务，视觉智能开放平台，图像搜索，全息空间 |
| | | 智能搜索与推荐 | 智能开放搜索 OpenSearch，智能推荐 AIRec |
| | 人工智能平台 | 人工智能平台 | 人工智能平台 PAI |

数据来源：阿里云公众号，东吴证券研究所

**阿里云中间层完成云原生到 AI 原生的能力升维，成为支撑企业 AI 稳定落地的神经中枢。**阿里云中间层已完成面向 AI 原生的升维进化，其函数计算 FC、EventBridge、Higress、RocketMQ 等核心产品均完成场景化能力重构，形成适配 AI 全流程的专业化支撑体系：Higress 由传统 API 网关升级为 AI 安全网关；EventBridge 从通用事件总线演进为 AI 数据管道；RocketMQ 由消息队列升级为 AI 会话管理组件。

**图103：阿里云中间层布局**

| 层级 | 一级分类 | 二级分类 | 具体产品 |
|---|---|---|---|
| 中间层 | 中间件 | 云消息队列 | 云消息队列 Kafka 版，云消息队列 RocketMQ 版，云消息队列 RabbitMQ 版，云消息队列 MQTT 版，轻量消息队列 |
| | | 云原生可观测 | 可观测可视化 Grafana 版，可观测监控 Prometheus 版，可观测链路 OpenTelemetry 版，性能测试服务，应用实时监控服务 ARMS |
| | | 应用集成 | 事件总线，API 网关 |
| | | 微服务工具与平台 | 企业级分布式应用服务 EDAS，微服务引擎 MSE，应用高可用服务 AHAS |

数据来源：阿里云公众号，东吴证券研究所

在底层能力上，阿里云中间件从三大维度构建 AI 应用落地核心保障：安全维度实现从边界防御到 AI 全链路防护的升级；性能维度突破传统网络延迟局限，聚焦推理效率优化；高可用维度以服务质量连续性替代基础在线时长，全面提升 AI 服务可靠性。

当前阿里云中间件已成为 AI 体系的神经中枢，直接决定企业 AI 应用落地质量，是企业依托 AI 打造增长引擎的关键底层支撑。

**阿里云的云服务全面覆盖域名与网站、媒体服务、数据库、大数据计算、云通信、终端用户计算以及企业云服务各场景。**阿里云服务与应用层已构建完备全栈云产品体系，可提供超 60 款云产品，全面覆盖数据库、大数据、人工智能等核心技术领域，贯通 IaaS、PaaS、MaaS 全层级云服务，能够支撑主权云客户将核心生产系统、大数据处理应用、创新 AI 业务等全品类业务统一上云，助力客户搭建集约化、一体化的统一 IT 技术底座。

**图104：阿里云服务与应用层产品矩阵**

| 层级 | 一级分类 | 二级分类 | 具体产品 |
|---|---|---|---|
| 服务与应用层 | 域名与网站 | 域名与网站 | 域名，备案服务，云解析 DNS |
| | | 知识产权服务 | 商标服务 |
| | 媒体服务 | 视频服务 | 直播，实时音视频，点播，超低延时直播 |
| | | 媒体开发服务 | 音视频终端 SDK |
| | | 媒体处理与内容生产 | 智能媒体服务，媒体处理，云端智能剪辑 |
| | 数据库 | 关系型数据库 | 云原生数据库 PolarDB，云数据库 RDS |
| | | 数据仓库 | 云原生数据仓库 AnalyticDB，云数据库 ClickHouse，云数据库 SelectDB 版 |
| | | 数据库管理工具 | 数据传输服务 DTS，数据库自治服务 DAS，数据管理 DMS |
| | | NoSQL 数据库 | 云原生多模数据库 Lindorm，表格存储 Tair，云数据库 MongoDB 版 |
| | | 数据库平台与服务 | 云数据库专属集群 |
| | 大数据计算 | 数据计算与分析 | 云原生大数据计算服务 MaxCompute，实时计算 Flink 版，实时数仓 Hologres，向量检索服务 Milvus 版，检索分析服务 Elasticsearch 版，向量检索服务 DashVector，图计算服务 GraphCompute |
| | | 数据应用与可视化 | DataV 数据可视化，智能商业分析 Quick BI，智能用户增长 Quick Audience |
| | | 数据湖 | 开源大数据平台 E-MapReduce，数据湖构建 DLF |
| | | 数据开发与服务 | 大数据开发治理平台 DataWorks，数据集成 Data Integration，数据总线 DataHub，智能数据建设与治理 Dataphin |
| | 云通信 | 云通信 | 短信服务，智能联络中心，语音服务，号码认证服务，号码百科，Chat APP 消息服务，号码隐私保护 |
| | 终端用户计算 | 无影 | 无影云电脑，无影云手机，无影云应用，无影终端 |
| | 企业云服务 | 企业云服务 | 能耗宝，机器人流程自动化 RPA，场景金融链接器，云行僧，营销引擎，企业商城 LinkedMall，Salesforce on Alibaba Cloud |

数据来源：阿里云公众号，东吴证券研究所

**阿里以钉钉为核心的 B 端 AI 协作平台，凭借钉钉的产品重构，实现工作模式升级、多模型兼容调用与高合规本地化部署，构建了完整的企业级 AI 协作落地能力。**阿里的 B 端 AI 协作平台以钉钉为典型代表，核心目标是打造高效的工作协作载体，重点提升流程可靠性、垂直行业适配能力以及与企业现有系统的融合度。钉钉 1.1 围绕三大核心维

度完成产品重构：一是首发全球首个专为 AI 打造的工作智能 OS——AgentOS，从根本上改变钉钉传统办公应用的定位，升级为可调度 AI 智能体执行操作的工作操作系统，推动 AI 从被动响应指令向主动感知并处理工作任务进阶；二是推出全新交互入口钉钉 ONE，该入口将群聊、文档、待办等分散信息整合为 AI 驱动的优先级信息流，实现工作模式从“人找事”到“事找人”的转变，同时支持企业在合规前提下，通过钉钉 AI 搜问调用 GPT-5.1、Gemini3、nano banana 等全球主流大模型；三是发布 AI 硬件 Ding Talk Real，通过本地化部署将模型、数据与应用管控在企业内网，满足金融、政务等领域的高等级数据安全合规要求，同时以开箱即用的软硬一体方案降低企业部署门槛，为企业打造安全专属的 AI 智能体运行环境。

**图105：钉钉 AgentOS 系统完整架构**


数据来源：钉钉公众号，东吴证券研究所

**图106：钉钉 ONE AI 搜索引擎界面展示**

数据来源：钉钉公众号，东吴证券研究所

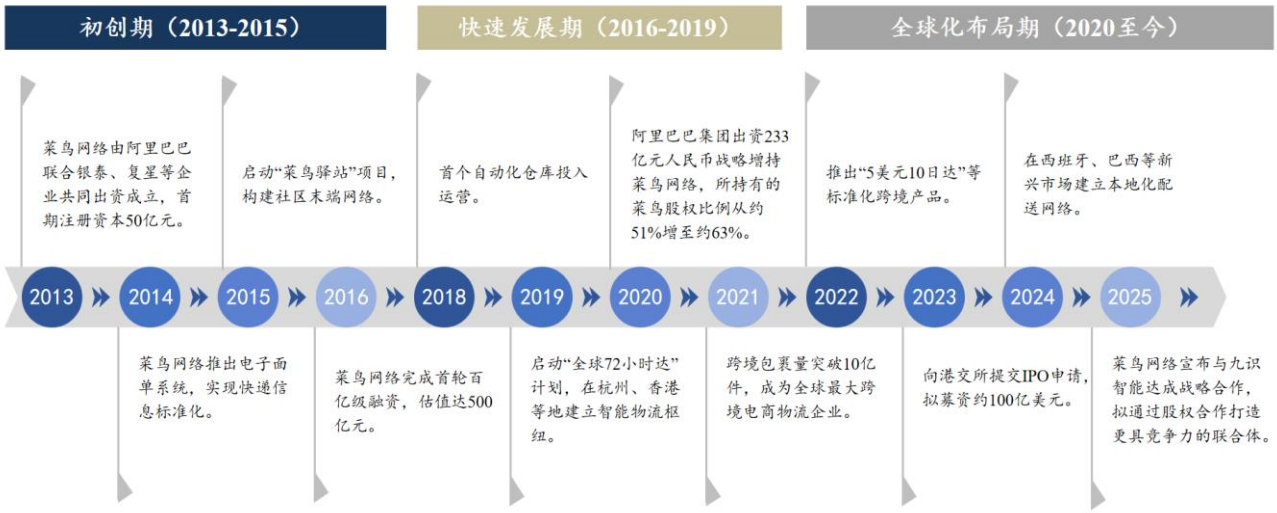
**图107：钉钉 DingTalk Real 功能展示**


数据来源：钉钉公众号，东吴证券研究所

**作为阿里 B 端“算力+组织+供应链”体系的关键一环，菜鸟网络的物流科技与全球**

**网络布局，直接承接阿里数字化的物流刚需。**菜鸟网络是阿里巴巴集团旗下核心物流科技平台企业，依托技术创新与大数据驱动，搭建起覆盖全球的智慧物流网络。作为中国领先的产业互联网公司，菜鸟专注于物流基础设施的数字化、智能化升级，打造包含国际快递、国际供应链、海外本地服务在内的综合物流解决方案，目前服务范围已遍及全球 200 多个国家和地区，日均跨境包裹处理量超 400 万件。业务布局层面，菜鸟以电商物流为核心场景，围绕该核心延伸推出国际快递、国际供应链、海外本地服务、国内快递等系列产品与服务，全方位满足电商商家及消费者的多元物流需求；在此基础上，菜鸟积极开辟业务第二增长曲线，拓展物流科技服务、供应链金融服务等创新业务，在扩大业务覆盖范围的同时，进一步提升平台综合盈利能力。

**图108：菜鸟发展历程**




数据来源：经济资信公众号，东吴证券研究所

**图109：菜鸟网络业务矩阵**

| 业务板块 | 核心服务内容 | 业务板块 | 核心服务内容 | 业务板块 | 核心服务内容 |
|---|---|---|---|---|---|
| 国际快递 | 跨境小包 | 海外本地服务 | 海外本地快递 | 供应链解决方案 | 供应链规划 |
| | 国际专线 | | 海外本地仓储 | | 供应链优化 |
| | 国际快递 | | 海外本地配送 | | 供应链管理 |
| 国际供应链 | 海外仓 | 国内快递 | 国内小包 | 物流科技服务 | 物流软件 |
| | 保税仓 | | 国内专线 | | 物流硬件 |
| | 国际货运代理 | | 国内快递 | | 物流咨询 |

数据来源：经济资信公众号，东吴证券研究所

## 5.2 字节跳动硬件链：采用“一盘棋”打法

### 5.2.1 芯片层：云端服务器芯片与专用终端芯片双路线布局

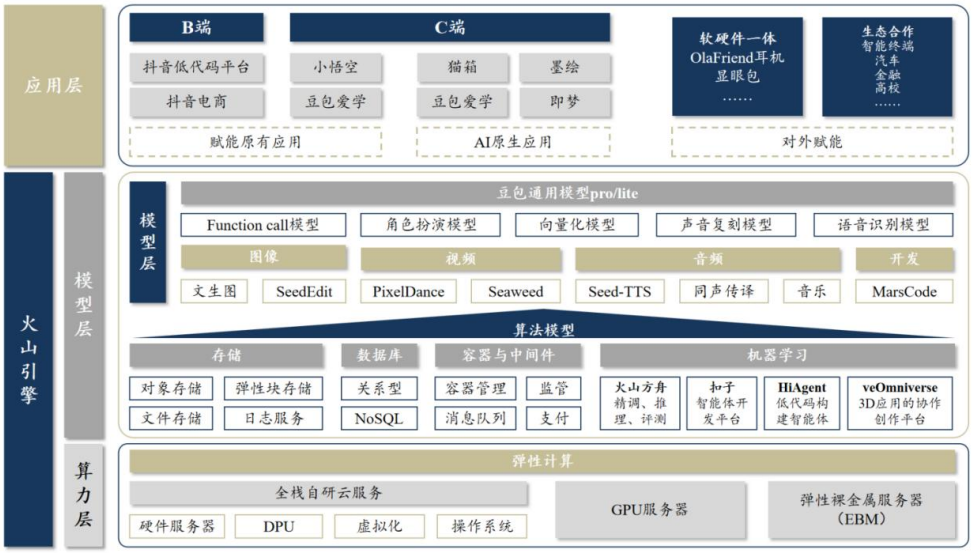
**字节跳动自研芯片聚焦双轨并行：一是面向 AI 算力的云端服务器芯片，二是 PICO**

**头显等 XR 硬件的专用终端芯片。**AI 服务器自研处理器领域，字节跳动依托 TikTok、抖音两大核心社交平台的业务基础，采用“自主研发+参股初创企业”的双规策略布局。其内部芯片部门已完成一款 AI 服务器处理器的流片，该产品性能可对标英伟达中国特供版 H20 芯片。PICO 头显芯片领域，字节跳动 2026 年面市的 PICO 新品头显将搭载全链路自研的专用芯片，该芯片已于 2024 年完成回片并步入量产阶段。这款定制化芯片破解了 MR 场景下高清高帧率视频的实时处理难题，同时满足设备在 SLAM 定位、运动补偿、逆畸变处理等场景的大规模实时计算需求，从底层优化 XR 设备普遍存在的眩晕、画面撕裂等体验痛点。芯片可兼顾高精度画面处理与低延迟表现，将系统延迟控制在 12 毫秒，同步达成算力与能效的双重高规格要求。

### 5.2.2 模型层：加速构建以豆包大模型为核心，覆盖多模态与开发领域的 AI 云原生架构

**字节跳动正迈入以 Agent 为核心技术载体的全新阶段，传统 IaaS、PaaS、SaaS 分层规划的 IT 架构不再有效，以大模型为中心的 AI 云原生架构正加速成型。**在新一代架构体系中，模型成为软件核心，MaaS 是模型调用与应用的优选方案，算力通过 Tokens 转化为智能服务；云平台与中间件围绕 Agent 的开发与运营，将 Tokens 封装为具备独立能力的智能体，并实现智能体与现有业务流程、智能体之间的互联互通。基于此技术判断，火山引擎全面升级 AI 云原生全栈服务：MaaS 层面，推出企业自有模型推理代工服务与强化学习平台；面向 Agent 开发，发布企业级 AI 智能体平台 AgentKit；针对 Agent 规模化运营，推出 HiAgent “1+N+X” 智能体工作站，全面支撑智能体产业落地。

**图110：字节跳动以大模型为中心的 AI 云原生架构**




数据来源：火山引擎官网，豆包官网，东吴证券研究所

**字节跳动在 AI 领域施行“高举高打”的激进扩张战略，依托雄厚的资金储备、规模化算力供给、丰富的应用场景与完备的技术储备，实行“全方位布局”策略。**豆包大模型是字节跳动整体 AI 生态的核心基石，也是当前公司在生成式 AI 领域具备最高市

场影响力与行业热度的核心品牌。其体系内包含基础层通用大模型，以及在此之上搭建的全品类垂直领域模型生态，全面覆盖视频生成、多模态理解生成，以及口型同步、TTS 等特色化能力。

**图111：字节跳动 AI 大模型布局**

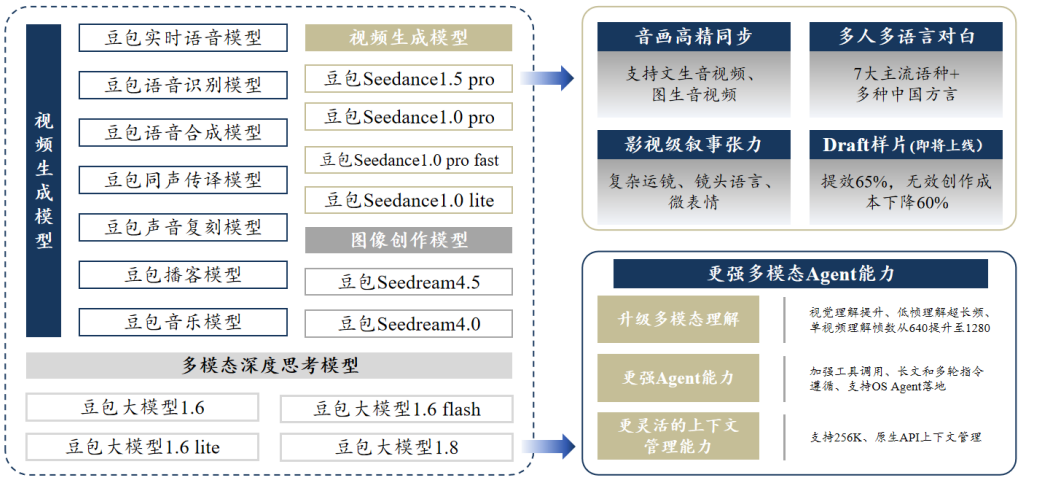
| 层级 | 研发团队 | 类型 | 产品 |
|---|---|---|---|
| 模型层 | Seed | 大语言模型 | Doubao-pro |
| | | | Doubao-lite |
| | | 语音模型 | Seed-ASR（语音识别） |
| | | | Seed-TTS（语音生成） |
| | | | Seed-ICL（声音复刻） |
| | | | Seed LiveInterpret（同声传译） |
| | | 图片模型 | SDXL-Lighting（文生图） |
| | | | Seed-Edit（图像编辑） |
| | | 音乐生成模型 | Seed-Music（音乐制作） |
| | | 视频编辑模型 | Boximator（视频编辑） |
| | | 视频生成模型 | MagicVideo-V22（文生视频） |
| | | | AnimateDiff-Lighting（文生视频） |
| | | | PixelDance（生视频、图生视频） |
| | | | Seaweed（文生视频、图生视频） |
| | | 多模态大模型 | BuboGPT |
| | | 3D 模型 | MVDream |
| | ByteDance Research | 具身智能模型 | GR-1 |
| | | | GR-2 |
| 中间层 | Flow | 智能体开发平台 | 扣子（国内，基于豆包） |
| | | | Coze（海外，基于 GPT） |

数据来源：新皮层 NewNewThing 公众号，火山引擎官网，东吴证券研究所

**经过多年的持续升级，豆包大模型家族在多模态理解和生成能力、Agent 能力上，已位于全球第一梯队。**2023 年 8 月，豆包前身“云雀”大模型成为国内首批通过官方备案的人工智能大模型。2024 年 5 月，“云雀”正式更名为“豆包”，并于同月 15 日在火山引擎原动力大会上完整发布豆包大模型家族产品矩阵。2025 年 12 月，火山引擎正式发布豆包大模型 1.8，专门面向多模态 Agent 场景进行了定向优化，其核心升级包括 Agent 能力增强、多模态理解帧数翻倍至 1280 帧、智能上下文管理。该模型在公开测评中表现优异，多项多模态任务超越全球顶尖模型，在 BrowserComp 通用智能体测评中位列全球领先。此外，豆包视频生成模型 Seedance 1.5 pro 同步发布，支持音画同步输出与多人多语言对白配音，具备影视级叙事张力，可满足影视、电商、广告等场景的高阶创作需求。更强模型、更低价格的优势驱动字节跳动豆包大模型的产业落地高速增长。

截至 2025 年 12 月，豆包大模型日均 Tokens 使用量已突破 50 万亿，居中国第一、全球第三。目前，已有超过 100 家企业在火山引擎上累计 Tokens 使用量超过一万亿。时至今日，豆包家族的产品矩阵仍在持续扩容，全新模型相继推出。

**图112：豆包大模型家族全景及模型升级情况**




数据来源：火山引擎公众号，东吴证券研究所

### 5.2.3 应用层：豆包扎根 C 端大众市场，MaaS 攻坚 B 端行业市场

**从产品矩阵架构而言，字节跳动 C 端 AI 生态以豆包为核心龙头，构建全场景海内外 AI 生态，完成全维度用户覆盖。**公司践行“产品驱动基建”的发展逻辑：依托 C 端应用实现规模化用户触达与数据积累，再通过海量真实场景数据反向赋能大模型训练迭代、交互体验优化，最终构建起闭环增长的业务飞轮。QuestMobile 监测数据显示，截至 2025 年 11 月，豆包日活用户达 5670 万，为国内用户活跃度第一的 AI 应用。字节 AI 应用核心聚焦内容生成与通用智能助手赛道，以豆包、猫箱、星绘、集梦 AI、剪映等数十款细分产品为载体，将成熟的 App 工厂规模化产品打法迁移至 AI 产业。其 To C 端 AI 产品布局具备全面性与完整性，核心对话式产品豆包有望成为国内规模领先的 AI 原生应用；垂直场景方面，公司已覆盖智能对话、教育、图像视频、情感陪伴等主流赛道；区域布局上实现海内外产品对标落地，产品形态同步覆盖独立 APP 与 Web 端。

**图113：字节跳动 AI 应用布局**

| 层级 | 研发团队 | 类型 | 产品 |
|---|---|---|---|
| 应用层 | Flow | 智能助手 | 豆包（国内） |
| | | | CiCi（海外） |
| | | 工具集 | 小悟空（国内） |
| | | | ChitChop（已停止运营） |
| | | 社交 | 猫箱（国内） |
| | | | AnyDoor（海外） |
| | | 图像 | 星绘（国内） |
| | | | PicPic（海外） |
| | 剪映 | 图片/视频生成 | 集梦 AI（国内） |
| | | | Dreamina（海外） |
| | | 视频剪辑 | 剪映（国内） |
| | | | CapCut（海外） |
| | | 教育 | 豆包爱学（国内） |
| | | | Gauth（海外） |
| | 大力教育 | 数字人 | 抖音 AI 分身（KOL 内测） |
| | | | TikTok AI 网红 |
| | 抖音/TikTok | 抖音电商内容生成 | 即创（国内） |
| | 字节跳动开发者服务团队 | 编程助手 | 豆包 MarsCode（国内） |
| | | | MarsCode（海外） |
| | | 模型分享社区 | 炉米 Lumi（国内） |
| | 其他 | 教育 | 识典古籍（国内） |
| | | 音乐生成 | 海绵音乐（国内） |
| 智能硬件 | Oladance+Flow | 智能体耳机 | Ola Friend |
| | 斐耳科技 | | FIIL GS Links |
| | 大力教育 | 智能台灯 | 大力智能学习灯 |
| | FoloToy+火山引擎 | 智能玩偶 | 显眼包 |

数据来源：新皮层 NewNewThing 公众号，东吴证券研究所

**智能硬件方面，字节跳动深耕大模型端侧应用，以耳机、玩具为豆包终端首批落地场景。**一方面，字节跳动与乐鑫科技、ToyCity 合作，推出 AI 毛绒玩具“显眼包”，集合了火山引擎的多项人工智能技术，如豆包大模型、扣子专业版、语音识别、语音合成等；另一方面，字节发布 AI 智能体耳机 Ola Friend，该产品可接入豆包大模型，并与豆包 App 深度结合。字节跳动正在研发豆包新一代 AI 耳机，将由歌尔股份承接代工生产。

**字节 B 端以火山引擎为核心载体，锚定 MaaS 为主赛道。**火山引擎是字节跳动 B 端服务战略的核心承载平台，依托字节跳动在 AI、大数据、云计算等领域长期积淀的核心技术优势，快速搭建起面向企业级市场的完整云产品体系。其业务发展始 2020 年，字节跳动正式将内部沉淀多年的技术能力面向外部客户开放输出，初期以推荐算法、数据

分析、视觉智能、多媒体技术等优势技术为核心支点，逐步拓展边界，构建起覆盖计算、存储、网络、安全等领域的全栈式云服务能力。

**图114：智能硬件之 AI 耳机 Ola Friend**

数据来源：极客公园公众号，东吴证券研究所

**图115：智能硬件之 AI 玩具“显眼包”**

数据来源：多知公众号，东吴证券研究所

在此发展阶段，火山引擎持续完善产品矩阵，以 IaaS 层 CPU 云、GPU 云与 MaaS 模型服务层为基础底座，不断丰富适配企业服务的全场景产品体系。迈入大模型时代后，火山引擎通过打造开放协同的技术生态与工具生态，持续为各行业、各应用场景的数字化转型提供深度技术赋能。

**图116：火山引擎云产品矩阵**

| 大类 | 小类 | 云产品 |
|---|---|---|
| 云基础 | 弹性计算 | 云服务器，操作系统 |
| | 数据库 | 云数据库 RDS，云数据库 Redis |
| | | 混合云 veStack |
| | | 中间件 |
| | 存储 | 弹性块存储，对象存储 |
| | 安全 | DDes，Web 应用防火墙，密码管理服务，渗透测试服务，红蓝对抗服务 |
| | 网络 | 私有网络，负载均衡，NAT 网关，公网 IP |
| | 容器 | 镜像仓库，容器服务 |
| 视频及内容分发服务 | 视频云 | 视频点播，企业直播，实时音视频，veImageX，智能处理 |
| | | 内容分发网络，边缘计算节点 |
| | 云通信 | 短信服务，语音服务 |
| | | 云手机 |
| 数据中台 | | 智能数据洞察 |
| | | 大数据集成与计算 |
| | | ByteHouse |
| | | E-MapReduce |
| | | 统一身份认证服务 |
| 开发中台 | | 低代码平台 aPaaS |
| | | veCompass |
| | | 链路服务 |
| 人工智能 | AI 开放平台 | 人像人体，图像技术，视频技术，文字识别，语音技术，自然语言处理，音频技术，机器翻译 |
| | | AI 开发平台 |

数据来源：火山引擎官网，东吴证券研究所

**火山引擎作为云计算与 AI 服务领域的后来者得以快速突围，核心在于其清晰的战略布局：专注押注 MaaS 模型即服务这一核心赛道，凭借单点突破策略，在 AI 应用迎来爆发的初期成功抢占市场先机。**火山引擎业务布局的核心便是自研 MaaS 平台火山方舟，字节跳动也始终锚定两大确定性发展方向，即以 MaaS 作为营收核心，并坚定认定 MaaS 会成为未来用户使用云服务的主流形态。其核心使命是把字节跳动在高速发展过程中沉淀的技术能力与应用工具向外部企业开放，MaaS 业务正是其对外输出 AI 能力的关键载体，且该业务底层依托的是与豆包同源的自研大模型底座，为服务能力提供了坚实的技术支撑。2025 年，火山引擎整体收入约 200 亿元，其中 MaaS 业务贡献了核心部分，且该收入结构在未来三至五年内不会改变。

**图117：字节跳动大模型产品关系图**


数据来源：火山引擎官网，东吴证券研究所

**火山引擎持续通过生态协同，联动智能终端、汽车、消费、金融等多领域行业主体，将核心技术能力转化为产业级影响力，共建零售大模型生态联盟。**终端消费电子领域，火山引擎已实现对主流手机厂商的广泛覆盖，合作方包括 OPPO、vivo、小米、荣耀、三星、传音等行业头部企业；汽车产业层面，火山引擎联动豆包大模型，围绕智能座舱、数字化营销等核心应用场景，为奔驰、宝马、奥迪、特斯拉、蔚来等主流整车企业提供全流程 AI 升级解决方案；金融服务领域，火山引擎的服务体系已覆盖华泰证券、国信证券、招商银行、浦发银行等头部券商、持牌基金公司与全国性商业银行；高等教育领域，火山引擎与超五成 985 高校达成战略合作；在消费零售领域，瑞幸咖啡借助豆包大模型开发了具备意图识别与槽位抽取能力的智能体。

图118：火山引擎携手千行百业迈向 Agent 时代


数据来源：火山引擎公众号，东吴证券研究所

## 5.3 腾讯与小米硬件链：腾讯场景深度赋能，小米“人-车-家”生态引领新范式

### 5.3.1 腾讯硬件链：模型筑基+场景落地+硬件赋能，腾讯端侧 AI 全链路成型

**腾讯端侧 AI 布局以长期主义为核心导向，完成从模型层硬核技术筑基到应用层全场景生态化落地的全链路构建，依托持续的研发投入、组织架构优化与得天独厚的生态场景优势，形成了差异化且极具落地性的端侧 AI 发展体系。**模型层，腾讯通过实质性动作完成技术与团队的双重升级，同时依托自 2018 年起超 4000 亿元的累计研发投入，完成了混元大模型的快速升级。混元 2.0 采用混合专家（MoE）架构，总参数规模达 4060 亿、激活参数 320 亿，还支持 256K 上下文窗口，通过显著改进预训练数据和强化学习策略，在复杂推理与文本生成核心场景实现国内领先；混元 3D 模型更保持全球领先水准，在开源社区的下载量超 300 万。应用层，腾讯采用多维策略，打造了覆盖 C 端个人用户与 B 端产业场景的全场景应用体系。在 C 端，腾讯构建了元宝等 AI 原生超级入口，并将元宝以分布式方式深度嵌入微信、QQ、腾讯会议、腾讯文档、视频号、公众号、腾讯新闻、腾讯地图等国民级产品中，打造了低门槛、高适配的端侧 AI 体验。在 B 端，腾讯坚持先内部规模化验证，再对外稳健输出的原则。目前混元大模型已在内部超过 900 款应用和场景中落地，其中腾讯云代码助手 CodeBuddy 成为国内首个支持插件、IDE 和 CLI 三种形态的 AI 编程工具。同时腾讯电子签、腾讯乐享、智能客服、智能办公等内部关键业务线也实现全面 AI 化，完成了自身生产方式的 AI 改造。在内部验证成熟后，腾讯精准切入医疗影像的病理分析、千万级车辆的智能座舱等落地深水区。在硬件端，腾讯以腾讯云 AIoT 2.0 产品解决方案深度赋能，依托软硬一体的技术架构，将语音智能体 TWeTalk、视觉智能体 TWeSee 及 AIoT 基础服务进行模块化封装，为用户提供一站式

的端到端服务，实现企业与开发者对 AI 能力的积木式按需组合，助力其快速推出全新的智能设备产品。TWeTalk 已深度应用于陪伴玩具、机器人、智能穿戴、耳机同传、智能点餐、智能导览、AI 面试等各类语音交互硬件场景；TWeSee 则全面服务于网络摄像机、可视门铃、智能门锁、电话手表、机器人、智慧屏幕等视觉感知硬件终端。

**图119：腾讯端侧 AI 布局**

| 层级 | 板块 | | 核心内容 |
|---|---|---|---|
| 模型层 | 模型层 | | 腾讯混元大模型+优秀开源模型 |
| | 基础设施层 | | AI Infra |
| 中间层 | ToC | | 腾讯元器 Tencent Yuanqi |
| | ToB | | 腾讯云智能体开发平台 Tencent Cloud ADP |
| | 机器人 | | 腾讯具身智能开放平台 Tencent Tairos |
| 应用层 | 顶层应用 | AI 原生应用 | 腾讯元宝，腾讯 ima，QQ 浏览器 |
| | | 腾讯特色生态 | 微信，企业微信，QQ，QQ 音乐，腾讯会议，腾讯新闻，腾讯文档，腾讯地图 |
| | 场景化 Agent | 企业服务 | 企点营销云 Agent，数据分析 Agent，语音 Agent，视觉 Agent，AI 漏洞检测 Agent，悟空代码安全 Agent |
| | | 生活场景 | 旅游规划 Agent，健康管理 Agent，学习助手 Agent，AI 高考通 Agent，智能座舱服务助手 Agent |
| | | 办公场景 | CodeBuddy Agent |
| | 行业应用 | 云 API | AI 陪伴玩具，具身机器人，智能手表，AR 眼镜，耳机同传，智能点餐、智能导览、AI 面试 |
| | | 应用端 API | 安防 IPC，可视门铃，智能门锁，机器人，智能屏幕 |

数据来源：腾讯公众号，腾讯云公众号，东吴证券研究所

### 5.3.2 小米硬件链：落实“人-车-家”全生态科技战略，开启“全品类科技高端化”新征程

**小米以玄戒 O1 高端 SoC 筑牢算力底座，凭借 MiMo 开源模型搭建技术核心，深耕“个人智能设备-家庭智能设备-智能出行”的全场景智能硬件生态，构建起一体化的端侧 AI 体系。**芯片层，玄戒 O1 于 2025 年正式发布并量产，成为小米在高端 SoC 领域的重要起点。从技术规格来看，玄戒 O1 具备 3nm 制程、10 核 CPU、旗舰 GPU 等配置，已达当前旗舰级水准，性能极具竞争力。模型层，小米重磅发布 Xiaomi MiMo-V2-Flash 开源 MoE 模型，总参数量达 309B、激活参数量 15B，在多个 Agent 测评基准中跻身全球开源模型 Top2，代码能力超越所有开源模型且比肩标杆闭源模型 Claude 4.5 Sonnet，基准测试中其性能与 DeepSeek-V3.2 基本相当。小米还同步推出在线 AI 聊天服务 Xiaomi MiMO Studio，打造模型体验、能力验证与场景落地的线上配套服务载体，构建起兼具高性能、高效率、高适配性与低成本的模型层技术底座。

**图120：小米硬件链布局**

| 层级 | 大类 | 子类别 | 代表产品 |
|---|---|---|---|
| 芯片层 | | 自研芯片 | 澎湃 S1（28 nm），玄戒 O1（3 nm） |
| 模型层 | | 开源大语言模型 | MiMo-V2-Flash，MiMo-7B |
| 中间层 | | 智能体开发平台 | Xiaomi MiMo Studio |
| 应用层 | 智能手机 | 旗舰系列 | 小米数字系列（小米 16），MIX 系列（折叠屏） |
| | | Redmi 系列 | Redmi K 系列，Note 系列，Turbo 系列 |
| | | POCO 系列（海外市场） | POCO M7，POCO X7 |
| | 可穿戴设备 | 智能手表 | Xiaomi Watch S4，Xiaomi 手环 10 |
| | | 智能手环 | Mi Band 8（健康版） |
| | | TWS 耳机 | Xiaomi Buds 4 Open（开放式设计），Redmi Buds 5 Pro |
| | | AR/VR | Xiaomi Glass 2（硅基 MicroLED 眼镜） |
| | | | 米家全景相机（360° 影像模组） |
| | 智能家居与 AIoT | 智能家电 | 米家空调，米家洗衣机，米家冰箱，小米电视 |
| | | 影音娱乐 | 米家投影仪 Ultra 2（4K 120Hz），小米音箱 |
| | | 清洁工具 | 米家扫地机器人 4，米家洗地机 3 Pro（自动集尘） |
| | | 安防系统 | 米家智能门锁，智能门铃，米家智能摄像头 |
| | | 智能网关（兼容 Zigbee / 蓝牙协议） | 多模网关，蓝牙 Mesh 网关 |
| | | 路由器（Wi-Fi 7 技术，支持 Mesh 组网） | 小米 BE7000 |
| | | 充电设备（支持多设备快充协议） | 小米 100W 氮化镓充电器，磁吸无线充电宝， |
| | | 智能插座与开关 | 小米智能插座，米家智能开关 |
| | PC / 平板与外设 | 轻薄本 | Xiaomi Book Air 13 OLED（1.3kg/14 小时续航），Redmi Book Pro 14（3.2K 90Hz） |
| | | 游戏本 | Redmi G 15（RTX 4070/2.4K 165Hz 高刷屏），小米游戏本 Pro 16（32GB 内存 / 2TB SSD） |
| | | 平板 | Redmi Pad 3，Xiaomi 米兔平板 |
| | | 周边外设 | Xiaomi 米家机械键盘，米家桌面打印机，标签打印机，小米游戏手柄 |
| | 生活消费品 | 健康护理 | 米家体脂秤，电子血压计，米家电动牙刷，吹风机 |
| | | 出行用品 | 90 分旅行箱：登机箱 Pro（TSA 密码锁），米家车载空气净化器 3 Pro（CN95 滤芯） |
| | | 厨房电器 | 米家电饭煲，米家电磁炉，智能压力锅，米家空气炸锅 |
| | | 智能灯具 | 米家净水器（云米代工），米家智能台灯，氛围灯带 |
| | 智能汽车 | 乘用车 | SU7 系列/YU7 系列（标准版 / Pro 版 / Max 版） |
| | | 生态车辆（九号公司出品） | 米家两轮车 S50（400km 续航），米家滑板车 4 Pro |
| | 互联网服务 | 操作系统 | MIUI 15（全球月活 6.5 亿），HyperOS（小米系统） |
| | | 金融科技 | 小米钱包（支付/理财/借贷），米家保险（健康/意外险） |
| | | 内容生态 | 小米视频（短视频聚合），米云同步（多端数据互通） |
| | 其他创新产品 | 机器人 | CyberOne 2.0（全尺寸人形机器人），米家服务机器人 X2（家庭配送） |
| | 子品牌 | - | 华米（智能穿戴），紫米（全能智能），润米 |

数据来源：行研大师傅公众号，东吴证券研究所

在应用层，小米以智能手机为智能硬件生态核心，通过逆周期推进高端化战略在全球 450 美元

以上高端手机市场实现份额逆势双升，依托 4000-6000 元价位段密集卡位与 AI 手机、折叠屏赛道的技术协同完成从高端市场参与者到规则定义者的角色转换，同时以印度出海经验为基础在中东、非洲、拉美等新兴市场实现出货量高速增长，通过线下渠道布局推动新兴市场份额下沉与欧洲、日本等发达市场高端机型渗透率提升，构建起兼具高端化与全球化的手机业务增长底座；同时以 IoT 和生活消费品为生态链两大支柱，采用“盟主带盟友”的产业链赋能模式打造覆盖全品类的智能硬件体系，且在可穿戴设备、音频、PC/平板与外设三大细分赛道实现高速增长并跻身全球前列，形成极具竞争力的 IoT 硬件矩阵；此外小米跨界布局智能汽车赛道，以首款 SU7 精准切入新能源乘用车 20-30 万核心价格带，通过 CarWith APP 打通手机与车机链路构建“人-车-家”智能生态闭环，规划 2026 年产能突破 30 万台以释放规模效应，最终形成以手机为核心、IoT 生态为重要支撑、智能汽车为全新增长极，各板块技术协同、生态互通，兼具高端化升级与全球化扩张的智能硬件全场景战略布局，实现从单一硬件产品到全生态智能硬件体系的深度构建。

## 6. 总结：锚定全球化竞争优势与技术积淀，见证端侧 AI 领军厂商的位阶提升与成长跨越

**在 AI 应用迭代驱动端侧硬件规格持续升级的背景下，2026 年端侧智能进入加速渗透阶段。**当前终端市场竞争格局正经历深度重塑：华为凭借麒麟系列芯片回归及鸿蒙生态的全栈整合，进一步巩固了其在高端市场的领先地位；小米、荣耀、OPPO、Vivo 等主流厂商在旗舰产品线上持续强化 AI 算力卡位，联想等 PC 龙头厂商则积极协同 AI PC 标准推进算力架构重构。紫光展锐聚焦“AI 赋能”带来的增量价值，依托 T9300 等 5G SoC 实现从入门级到中高端市场的全系列覆盖，在传音、中兴等新兴市场及主流位段得到广泛应用。在数据主权与隐私安全需求的驱动下，以绿联科技为代表的厂商通过 AI NAS 产品的快速迭代，助力边缘算力市场规模向 500 亿美元目标迈进。而豆包手机、Openclaw 部署的 Mac Mini 等终端 AI 应用增量表现场景的涌现，是国产供应链实现技术外溢与位阶提升的关键。中国有望凭借成熟的供应链体系与敏捷的生态响应力，在全球端侧 AI 产业竞争中占据核心话语权。

**展望 2026 年，国产智驾芯片赛道正由单纯的硬件参数竞赛演进为系统级生态博弈。**地平线作为行业领军者，凭借其覆盖全谱系的硬件矩阵及与大众、上汽等头部 OEM 的深度资本绑定，已构筑起坚实的市场壁垒，2025 年 1-11 月其智驾域控 SoC 安装量份额达 8.1%，在 10-20 万元主流市场渗透率更提升至 16.6%，具备显著的规模效应。与此同时，在第三方城市 NOA 市场占据 61.06%绝对份额的 Momenta 正加速推行“算法定义硬件”策略，试图以 254 TOPS 自研芯片及 4000-5000 元的极致性价比方案实现市场下探。虽然目前双方尚未爆发全面直接竞争，但 2026 年将成为重塑产业逻辑的关键窗口期：地平线的先发规模优势能

否受到实质性挑战，依赖于 Momenta 自研硬件的工程化落地表现及其对软件溢价的承接能力。在“舱驾一体”及集成化架构的必然趋势下，芯片能效比、算法兼容性及量产成本将成为定义行业竞争胜负的核心变量。

与此同时，2026 年是国产座舱芯片从“局部替代”向“全局引领”跨越的战略质变期。国内厂商正通过“舱驾一体”与“多域融合”的技术降维打击，以及“智驾平权”带来的规模化红利，深度重构全球智能汽车的算力底座。从后续发展可能性来看：芯擎科技与黑芝麻智能有望凭借在集成化架构上的先发优势，通过大幅降低整车 BOM 成本，成为车企中端及主流机型平台化部署的优先选择；瑞芯微与芯驰科技通过向 4nm 先进制程及高算力 NPU 的跨越，将驱动端侧大模型在 10 万-20 万元价位段实现从“极客配置”向“行业标配”的普及；华为海思与比亚迪将继续依托强大的终端生态粘性或极致的垂直整合规模，分别稳固其在高端溢价区与装车基数上的领跑地位。而随着地平线等跨域玩家的舱驾一体芯片步入量产周期，行业将加速向单芯片中央计算时代演进，国内厂商有望在这一进程中实现从本土配套向全球主流供应体系的实质性突破。

在 AI 应用迭代驱动端侧硬件规格持续升级的背景下，国产 IoT 芯片厂商正通过细分赛道的异构算力部署，实现从单一功能模块向边缘计算平台的价值跨越。瑞芯微（Rockchip）依托 RK3588、RK3576 等高性能算力平台，深化在工业视觉及汽车电子领域的渗透，其在 10 万元以下入门级座舱芯片市场的标配安装量份额已达 8.7%。晶晨股份在稳固全球机顶盒（市占 31.5%）与智能电视（16.8%）SoC 领军地位的同时，依托全球 250 余家具备深度渗透力的运营商渠道及知名消费电子大厂合作，展现出显著的国际供应链优势与品牌背书，正通过 6nm 工艺向全场景 AIoT 平台加速重构。此外，黑芝麻智能通过控股亿智电子切入通用 AI 推理市场；星宸科技以超 40%的视频安防份额卡位 AI 眼镜赛道；恒玄科技与乐鑫科技则分别在智能音频与 Wi-Fi MCU 领域巩固其平台化地位。机器人赛道（含具身智能）方面，根据摩根士丹利预测，2026 年中国人形机器人销量将翻倍至 2.8 万台。瑞芯微在该领域具备显著的先发优势，其 SoC 方案已广泛承担机器人“小脑”核心功能，且凭借在机器视觉与音频领域的成熟方案以及 SDK/BSP 的深度成熟度，其硬件适配生态优势已不逊色于国际大厂，具备抢占该蓝海市场高份额的确定性机遇。同时，地平线征程 7 系列已针对工业场景提供千 TOPS 级算力支持并适配优必选等头部厂商。展望 2026 年，国产芯片厂商正转型为“算力+算法+连接”的一体化平台商，有望凭借成熟的本土供应链与敏捷的生态响应力，在具身智能机器人及 AI 新终端赛道实现全球位阶的实质性提升。

国内互联网大厂正通过构建端云协同的闭合生态，加速确立以自研算力、专属模型与多形态终端为核心的全栈竞争优势。阿里巴巴依托云端一体化布局实现了从底层算力到边缘触达的深度覆盖，字节跳动通过整合底层芯片、自研模型与豆包手机等终端应用强化了业务协同，腾讯与小米则分别利用场景赋能与全

生态联动，构筑了端侧 AI 落地的产业护城河。这种闭环模式不仅能通过本地推理优化运营成本并强化隐私安全，更在范式革命中通过掌控物理入口锁定了核心流量，从而确立了长期的卡位优势。在此趋势下，成功切入大厂供应链的国产硬件厂商将深度受益于终端创新带来的订单红利与技术外溢，迎来位阶提升与份额扩张的确定性机遇。

## 7. 风险提示

**终端需求复苏不及预期：** 手机、PC 等终端是端侧 SoC 的基本盘，若宏观消费意愿偏弱或缺乏“杀手级”AI 应用带动，将拉长终端换机周期，导致下游去库存及芯片出货量承压。

**端云协同建设不及预期：** 端侧 AI 体验高度依赖端云异构算力协同，若底层互联协议标准推进迟缓或生态割裂，将导致应用体验降级，进而压制端侧 SoC 的市场扩容速度。

**国产替代及量产落地不及预期：** 高阶 AI SoC 技术壁垒高，若本土厂商在先进制程流片、良率爬坡或软硬件工具链适配上遇阻，将导致产品交付延后，错失国产替代验证窗口期。

**存储涨价压制利润风险：** 端侧 AI 对高带宽大容量存储需求严苛，若上游存储元器件价格持续居高不下，推高的整机 BOM 成本不可避免会向上游 SoC 环节传导，阶段性挤压芯片端利润。

**行业竞争加剧风险：** 端侧算力赛道正处于高烈度博弈阶段，国际巨头价格策略调整及国内跨界玩家不断涌入易引发价格战，或将大幅压缩本土芯片设计企业的毛利空间。

**国内数据安全与政策调整风险：** 端侧大模型涉及敏感数据本地化处理，若国内针对数据隐私与硬件级安全隔离出台更为严格的监管标准，将拉长芯片研发周期并推高企业合规成本。

## 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，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

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。



## 东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

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