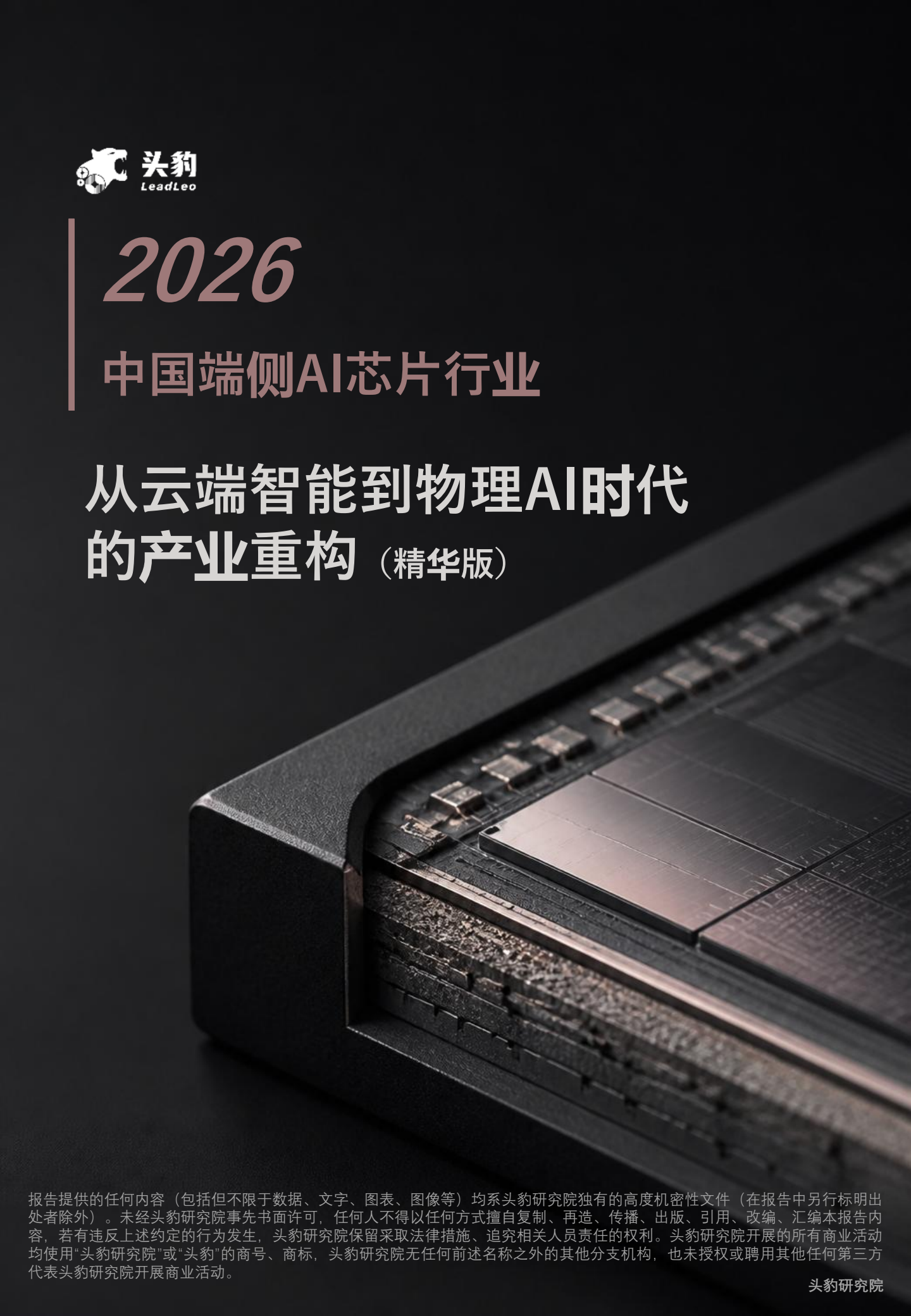
头豹
LeadLeo

2026

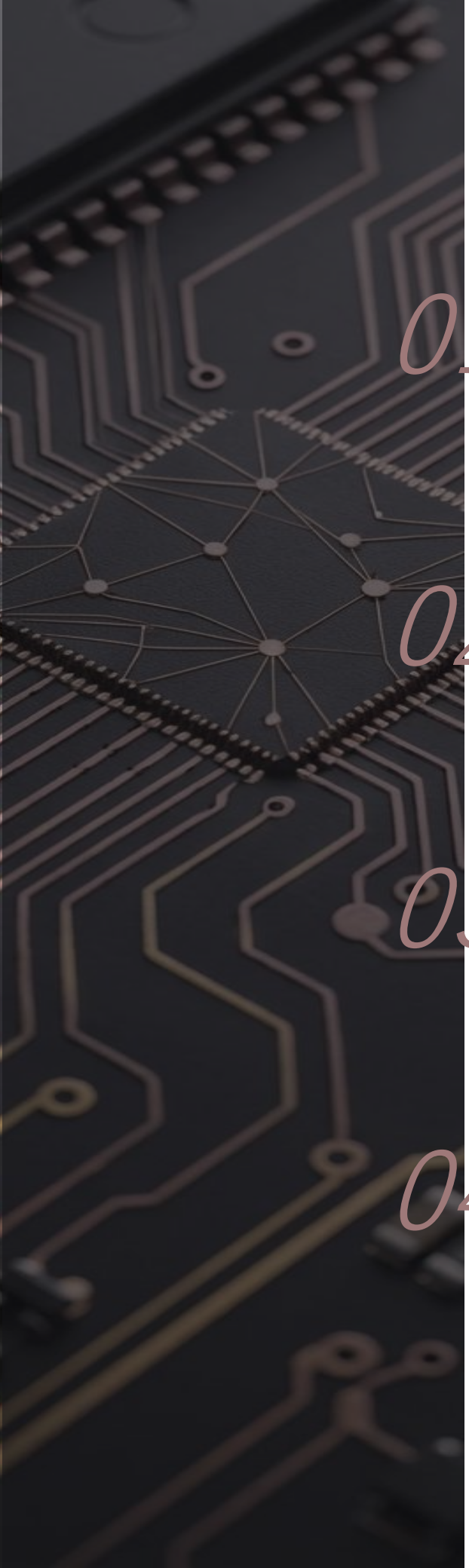
中国端侧AI芯片行业

从云端智能到物理AI时代的产业重构（精华版）



报告提供的任何内容（包括但不限于数据、文字、图表、图像等）均系头豹研究院独有的高度机密性文件（在报告中另行标明出处者除外）。未经头豹研究院事先书面许可，任何人不得以任何方式擅自复制、再造、传播、出版、引用、改编、汇编本报告内容，若有违反上述约定的行为发生，头豹研究院保留采取法律措施、追究相关人员责任的权利。头豹研究院开展的所有商业活动均使用"头豹研究院"或"头豹"的商号、商标，头豹研究院无任何前述名称之外的其他分支机构，也未授权或聘用其他任何第三方代表头豹研究院开展商业活动。

头豹研究院



核心洞察

01 端侧AI芯片正由单一功能组件升级为物理AI终端的核心计算底座

云端推理受时延、带宽成本、隐私合规及网络稳定性约束，叠加终端本地感知、决策与控制需求，以及模型轻量化、异构计算和工具链成熟，推动端侧AI芯片成为物理AI终端的核心计算底座。

02 端侧AI芯片评价体系正由峰值TOPS转向有效智能密度

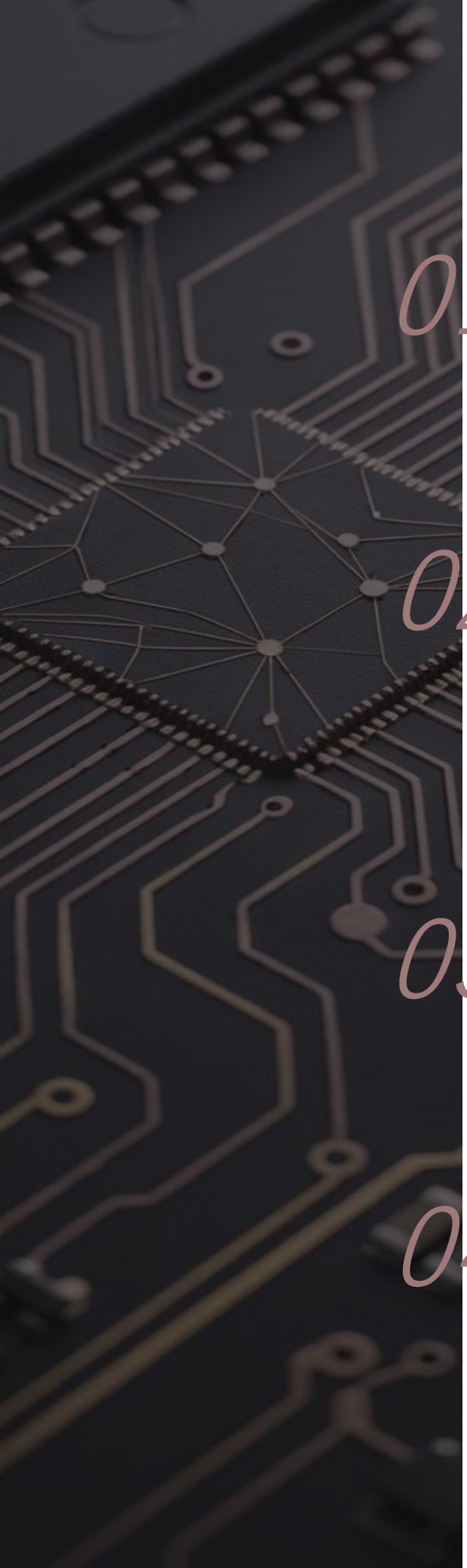
有效智能密度衡量芯片在功耗、散热和成本约束下运行真实模型并实现规模交付的综合能力，覆盖有效算力、功耗、内存带宽、异构协同、工具链、场景适配、量产能力及安全国产化。

03 机器人是端侧AI芯片产业最高压验证的终端场景

机器人需持续并行处理多模态感知、环境理解、任务决策、运动控制、实时反馈和多模型调度，对异构协同、响应周期、计算精度、可靠性及工具链提出系统性要求。

04 云厂商的下一轮机会是成为机器人和AI终端的开发基础设施

未来云厂商的产业定位由算力资源供给方延伸至开发基础设施提供方，通过整合算力、大模型、仿真训练、模型部署及设备管理，贯通开发、训练、验证、部署与更新流程，形成面向机器人和AI终端的新增长空间。



目录

01 端侧AI芯片的兴起逻辑与研究边界 5

- 为什么是现在开始研究端侧AI芯片？ 6
- 什么是端侧芯片？其与云端、边缘芯片有何差异？ 8
- 云/边/端如何分配AI任务，端侧承担哪些关键环节？ 10

02 端侧AI芯片产业链现状与竞争体系 12

- 端侧AI芯片的产业链全景 13
- 端侧AI芯片产业链的核心价值落点在哪些环节？ 14
- 不同类型端侧AI芯片厂商依靠哪些能力竞争？ 15
- 当前哪些系统性指标能真正评判端侧AI芯片竞争力？ 17

03 端侧AI芯片的场景落地与规模化约束 19

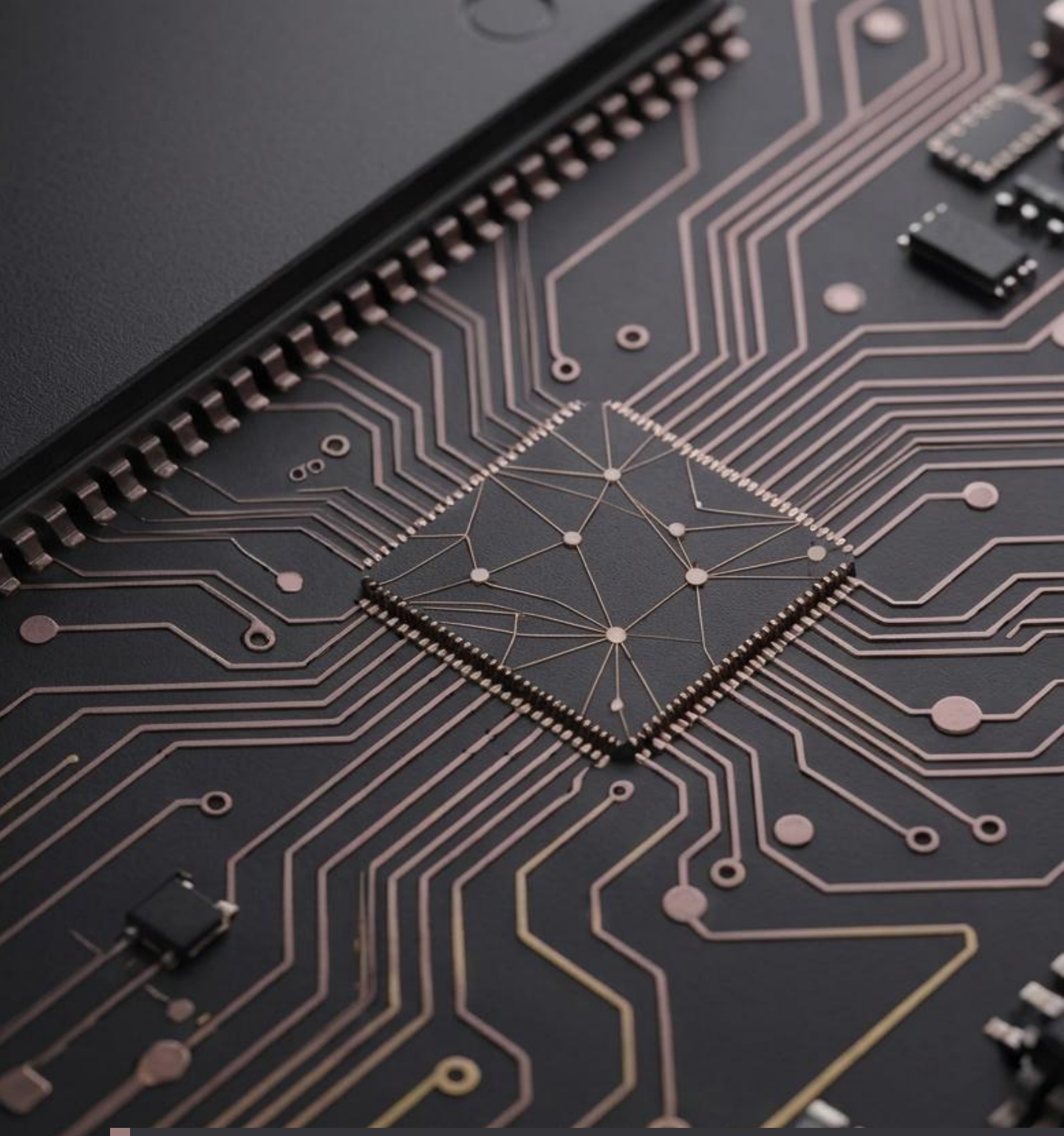
- 端侧AI芯片将率先在哪些应用场景形成规模化需求？ 20
- 哪个应用场景将构成端侧AI芯片产业最高压验证场景？ 22
- 哪些因素正在制约端侧AI芯片产业爆发？ 24

04 端侧AI芯片的产业机遇与演进方向 26

- 产业重构后，端侧AI芯片产业链的机会应如何分层？ 27
- AI机器人产业链中的新机会有哪些？ 28
- 中国端侧AI芯片产业将如何持续演进？ 29

名词解释

- **物理AI：**能够感知现实环境、理解任务、形成决策并控制物理设备执行动作的人工智能系统。
- **本地推理：**直接在终端设备上运行已训练模型并输出识别、预测、决策或控制结果的过程。
- **异构计算：**利用CPU、GPU、NPU等不同计算单元分别处理适合的任务，以提高整体计算效率。
- **异构调度：**根据任务类型、响应要求和资源状态，将计算负载分配给不同处理器并协调运行的过程。
- **多模态感知：**同时采集并融合视觉、语音、触觉、位置及其他传感器数据，以形成对环境的综合判断。
- **RTT（往返时延）：**数据从终端发送至远端并返回所需的总时间，用于衡量网络通信延迟。
- **I/O（输入输出）：**芯片或设备与传感器、存储器、执行器及其他外部组件之间传输数据的接口体系。
- **CPU（中央处理器）：**负责通用计算、系统控制、任务管理及复杂逻辑处理的核心处理器。
- **GPU（图形处理器）：**具备大规模并行计算能力，主要用于图形处理、模型训练及部分AI推理任务。
- **NPU（神经网络处理器）：**针对神经网络计算进行专门优化，用于提高AI推理速度和能源效率。
- **DSP（数字信号处理器）：**面向音频、图像、通信及传感器信号进行高速数学运算的专用处理器。
- **MCU（微控制器）：**面向低功耗、确定性控制和实时响应任务的小型控制芯片。
- **SoC（片上系统）：**将CPU、GPU、NPU、存储控制器及各类接口集成在单颗芯片中的系统级芯片。
- **AI SoC：**集成通用计算、AI推理、存储与接口能力，面向智能终端完整计算任务设计的片上系统。
- **IP（芯片知识产权核）：**可被芯片设计厂商授权使用的处理器架构、接口或功能模块设计。
- **EDA（电子设计自动化）：**用于芯片设计、仿真、验证及制造准备的软件工具体系。
- **内存容量：**芯片或终端能够存放模型参数、运行数据和中间计算结果的存储空间。
- **内存带宽：**处理器与内存之间单位时间内能够传输的数据量，直接影响模型推理和多任务并发效率。
- **BOM（物料清单）：**生产一台终端设备所需全部零部件及其成本构成的清单。
- **TOPS（每秒万亿次运算）：**衡量芯片在特定精度和频率条件下理论计算能力的峰值指标。
- **FPS（每秒帧数）：**系统每秒能够处理的图像帧数，常用于衡量视觉推理性能。
- **TTFT（首个词元生成时间）：**生成式模型接收输入后输出第一个词元所需的时间。
- **TPOT（单个输出词元时间）：**生成式模型连续输出过程中生成每个词元平均所需的时间。
- **降频：**芯片因温度、功耗或供电限制而降低运行频率，导致持续性能下降的保护机制。
- **编译器：**将模型计算图和程序代码转换为芯片可执行指令，并针对硬件架构优化计算过程的软件。
- **SDK（软件开发工具包）：**向开发者提供接口、程序库、示例和调试工具的一组开发资源。
- **运行时：**负责模型加载、内存管理、任务执行和硬件调用的软件运行环境。
- **工具链：**覆盖模型转换、编译、量化、算子适配、性能分析、调试及部署的完整软件工具体系。
- **算子：**神经网络中执行卷积、矩阵乘法、激活或归一化等具体计算功能的基本运算单元。
- **SoM（系统级模组）：**将芯片、内存、存储及基础接口预集成为可直接嵌入终端的计算模块。
- **开发板：**集成芯片、接口和基础软件，用于产品原型开发、模型测试及功能验证的硬件平台。
- **AMR（自主移动机器人）：**能够自主感知环境、规划路径并完成移动任务的机器人。
- **AGV（自动导引运输车）：**通常依靠磁条、二维码或预设路线完成物料运输的自动化车辆。



01 端侧AI芯片的兴起逻辑与研究边界

为什么是现在开始研究端侧AI芯片？（1/2）

- **当前AI正由云端内容处理迈向物理世界实时执行；供给约束集中显现、端侧技术逐步成熟及终端需求增长，共同驱动端侧AI芯片由传统功能组件向物理AI时代的核心计算底座升级**

从云端智能走向物理世界的变迁史

AI范式演进

数字内容处理时代

云端大模型

大模型崛起/云端算力竞赛

AI处理文字、图片、视频等，计算发生在云端数据中心

物理AI时代

机器人 + 端侧计算

感知-理解-决策-执行

AI开始理解现实世界，计算以端侧为主、云端辅助

2017 — 2022 — 202 — 2025

端侧芯片角色变迁

配件时代

单点推理加速器

功能单一、可替代性强、价值未显现

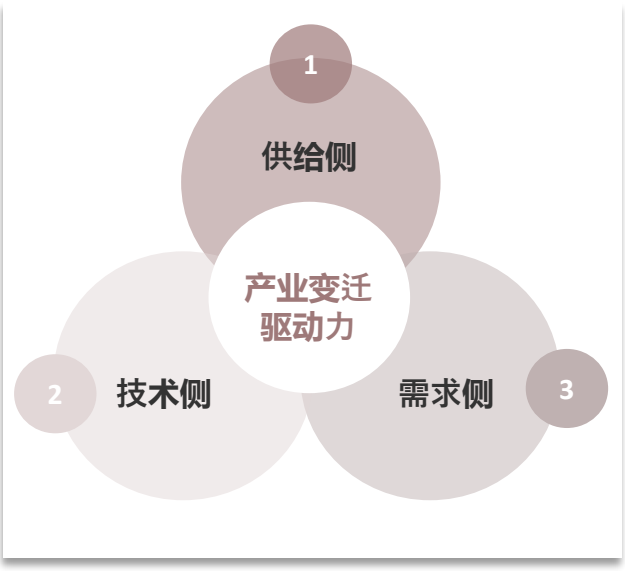
以手机NPU等为例，主要支持影像增强、语音识别等特定任务，功能价值集中于局部场景

基础设施时代

系统级计算底座

难以替代、决定AI智能上线、价值高

支撑多模态感知、本地推理、实时决策与执行控制，是AI终端的核心计算基础



- **端侧AI芯片产业的崛起由供给侧压力、技术侧成熟与需求侧拉动三重力量共同驱动**

1. **延迟的物理边界：**云端推理依赖网络往返传输，典型RTT延迟在50～200ms量级，而机器人运动控制、车载紧急制动、工业设备实时反馈等场景的响应窗口往往低于10ms。这一差距无法通过工程优化弥合——光速传播上限与网络协议栈的固有开销，共同构成云端推理无法逾越的物理架构约束。因此毫秒级、确定性控制任务通常需优先在端侧或近端完成。

（接下页）

来源：漫谈Light the Star，头豹研究院

为什么是现在开始研究端侧AI芯片？（2/2）

（接上页）

2. **带宽与成本的规模化压力：**多传感器终端（机器人、自动驾驶车辆等）在持续运行状态下，每秒可产生数百MB至GB级原始数据。当终端设备规模从百台扩展至万台乃至百万台时，全量上传云端推理的带宽与云端处理成本将随设备规模和数据量快速上升，全量云端推理在规模部署下可能面临较高成本，端侧或端云协同可降低部分传输与推理成本。

3. **隐私与合规的制度性约束：**医疗影像、金融交易记录等高敏感数据受到跨境传输、数据安全和个人信息保护等监管要求约束。GDPR及中国《数据安全法》等法规提高了数据采集、传输与跨境处理的合规门槛，推动本地处理成为降低数据外传及合规复杂度的技术路径之一，由此扩大端侧计算在高敏感场景中的部署需求。

4. **离线运行的场景需求：**矿山井下、远海作业平台及偏远工业现场等部署环境存在网络覆盖不足、连接不稳定甚至无网络可用等情况。依赖云端推理难以保障系统连续运行，端侧自主推理与实时响应能力因此成为相关场景实现可靠部署和商业化落地的必要条件。

需求侧形成实际采购牵引。人形机器人、智能汽车等下游硬件正从概念验证走向产品导入与规模量产，并加速引入本地智能。多传感器融合、实时环境感知、运动控制与持续任务执行，对芯片的算力密度、能效水平、接口兼容性及运行稳定性提出更高要求。终端竞争重点逐步由硬件参数转向实时感知、自主决策和持续执行能力，端侧AI芯片在整机成本与产品定义中的重要性同步提升，由功能组件向智能终端核心计算底座演进。

技术侧为端侧部署提供可行条件。模型量化、剪枝、蒸馏及小模型能力提升，有效降低模型对算力、内存和存储资源的占用；叠加NPU、异构计算、先进封装和端侧软件工具链发展，芯片能够在受限功耗与散热条件下承载多模型协同推理。编译器、SDK及模型转换工具逐步完善，也在降低模型迁移、硬件适配和应用开发门槛，使更多AI模型能够在端侧稳定运行并实现规模化部署。

来源：漫谈Light the Star，头豹研究院

什么是端侧AI芯片？其与云端、边缘芯片有何差异？(1/2)

- **端侧AI芯片是嵌入终端、承担本地感知处理、AI推理、交互或控制任务；区别于云端的集中计算与边缘侧的区域协同，其以低时延、低功耗、数据本地处理、离线运行与实时控制支撑终端智能闭环**

端侧芯片的定义与研究范畴界定

- **端侧AI芯片的定义：**是指部署于终端设备本地、面向实时感知、本地推理与物理执行任务的AI计算芯片。其核心特征在于：计算发生在数据产生的物理现场，可在无需持续网络连接的情况下完成任务，并可与云端协同；具备低延迟响应、低功耗运行、本地隐私保护与离线可用的能力。

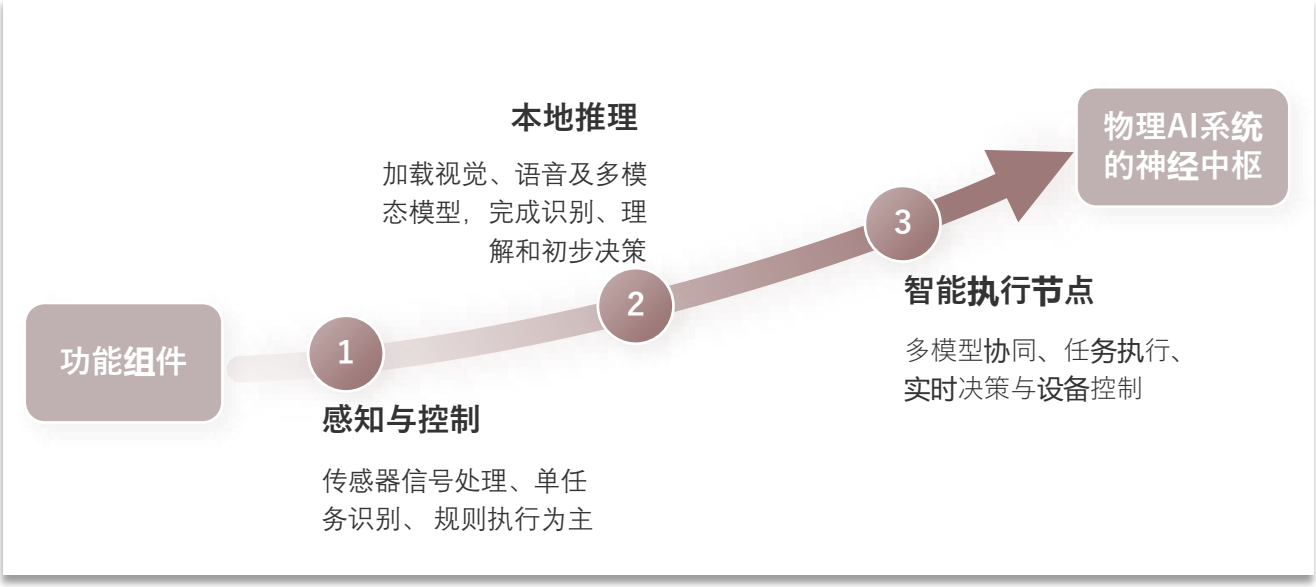
不同部署位置的芯片特征差异

对比维度	云端AI芯片	边缘AI芯片	端侧AI芯片
部署位置	集中式数据中心	基站、园区服务器、边缘网关	手机、汽车、机器人、可穿戴等智能终端内部
核心任务	模型训练、复杂推理、知识更新、全局分析	区域数据汇聚、多设备调度、模型缓存、近场推理	数据预处理、本地推理、现场决策、设备控制
响应要求	更关注吞吐量和任务规模	强调区域级低时延与多设备协同	强调确定性响应和感知执行闭环
资源条件	算力、存储和散热条件充足	资源规模介于云端与终端之间	功耗、散热、内存和成本约束最强
数据隐私	数据离开本地，合规风险较高	数据部分本地化	数据不离开设备，天然合规
时延要求	秒级～分钟级	百毫秒～秒级	微秒～毫秒级
竞争重点	算力规模、集群效率、互联带宽	任务调度、设备连接、区域协同	能效、实时性、模型适配、成本与量产能力

- **端侧AI芯片产业研究范畴界定：**为明确后续产业链与厂商分类边界，本报告采用狭义定义区分边缘计算与端侧计算。边缘侧指位于终端与云端之间的区域计算节点；端侧指嵌入最终设备、连接数据采集与物理执行的计算节点。因此，芯片是否纳入本报告研究范围，主要取决于其是否部署于终端，并承担本地AI推理或智能控制任务。

来源： ARM， 头豹研究院

什么是端侧AI芯片？其与云端、边缘芯片有何差异？(2/2)



- **端侧AI的本质是场景实时智能，即在资源与环境约束下完成感知、理解、决策和控制闭环；低时延、低功耗、数据本地处理、离线可用与实时控制，共同构成端侧芯片区别于云端及边缘芯片的核心边界**

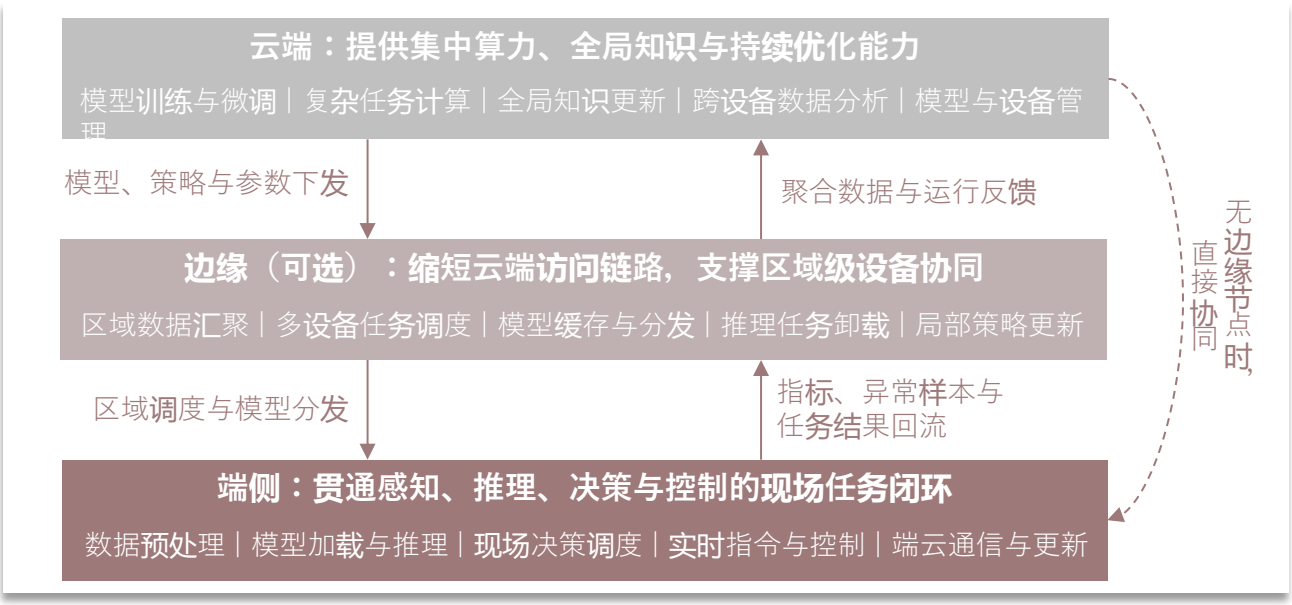
低时延：机器人运动控制、车载紧急制动及工业设备异常响应等场景通常要求毫秒级反馈，时延阈值由物理过程和安全要求决定。端侧计算可减少数据上传、网络传输与云端处理环节，缩短感知至执行链路。**低功耗：**端侧设备多依赖电池或有限供电，功耗范围覆盖毫瓦级至数十瓦级。芯片需在严格的功耗与散热约束下运行多类模型，对单位功耗推理性能、持续算力及能效管理提出更高要求。**数据本地处理：**端侧计算可在设备内部处理敏感数据，减少原始数据外传，降低传输与集中存储风险，为医疗、工业、金融等场景满足数据安全、个人信息保护及跨境传输要求提供技术支撑。**离线可用：**矿山、远海、工业现场及偏远部署点可能长期处于弱网或无网环境。端侧芯片需在缺少云端连接时维持核心推理与控制功能，保障关键业务连续运行。**实时控制：**在机器人、汽车及工业控制等场景，端侧芯片还需支持将模型输出转化为可执行控制指令，并通过CPU、NPU、MCU等计算单元协同，贯通感知、推理、决策与控制链路，实现低时延、低抖动的稳定调度。

来源：ARM，头豹研究院

云/边/端如何分配AI任务，端侧承担哪些关键环节？(1/2)

- **端云协同需依据响应时延、计算复杂度、数据敏感性和网络连续性分配任务：云端负责模型训练与全局优化，边缘侧按需承担区域协同，端侧AI芯片贯通数据处理、本地推理、现场决策与实时控制**

云边端协同任务分工示意图



- **云边端协同任务分工逻辑——由任务约束确定计算位置，使不同层级承担与其资源条件和响应要求相匹配的任务**
  - ✓ **响应时延：** 对即时性和确定性要求越高，任务越需靠近终端执行；机器人运动控制、汽车安全响应及工业设备控制等任务应优先在端侧完成。
  - ✓ **计算复杂度：** 模型训练、大规模知识更新及复杂非实时推理需要较高算力和存储资源，主要由云端承担；超出终端承载能力的推理任务可按网络条件向边缘或云端卸载。
  - ✓ **数据敏感性：** 涉及个人信息、生产数据及高敏感传感器数据的任务，应优先完成本地处理，仅回传脱敏结果、特征数据或异常样本。
  - ✓ **网络依赖：** 核心功能需要在弱网或断网环境下持续运行，端侧必须保留完成关键任务闭环所需的最低计算能力。

来源：ARM，ETSI，头豹研究院

(接下页)

云/边/端如何分配AI任务，端侧承担哪些关键环节？(2/2)

（接上页）

✓ **协同范围：** 单设备实时任务主要由端侧承担；园区、车路、机器人集群等多设备协同任务可由边缘节点处理；跨区域知识更新与全局优化由云端完成。

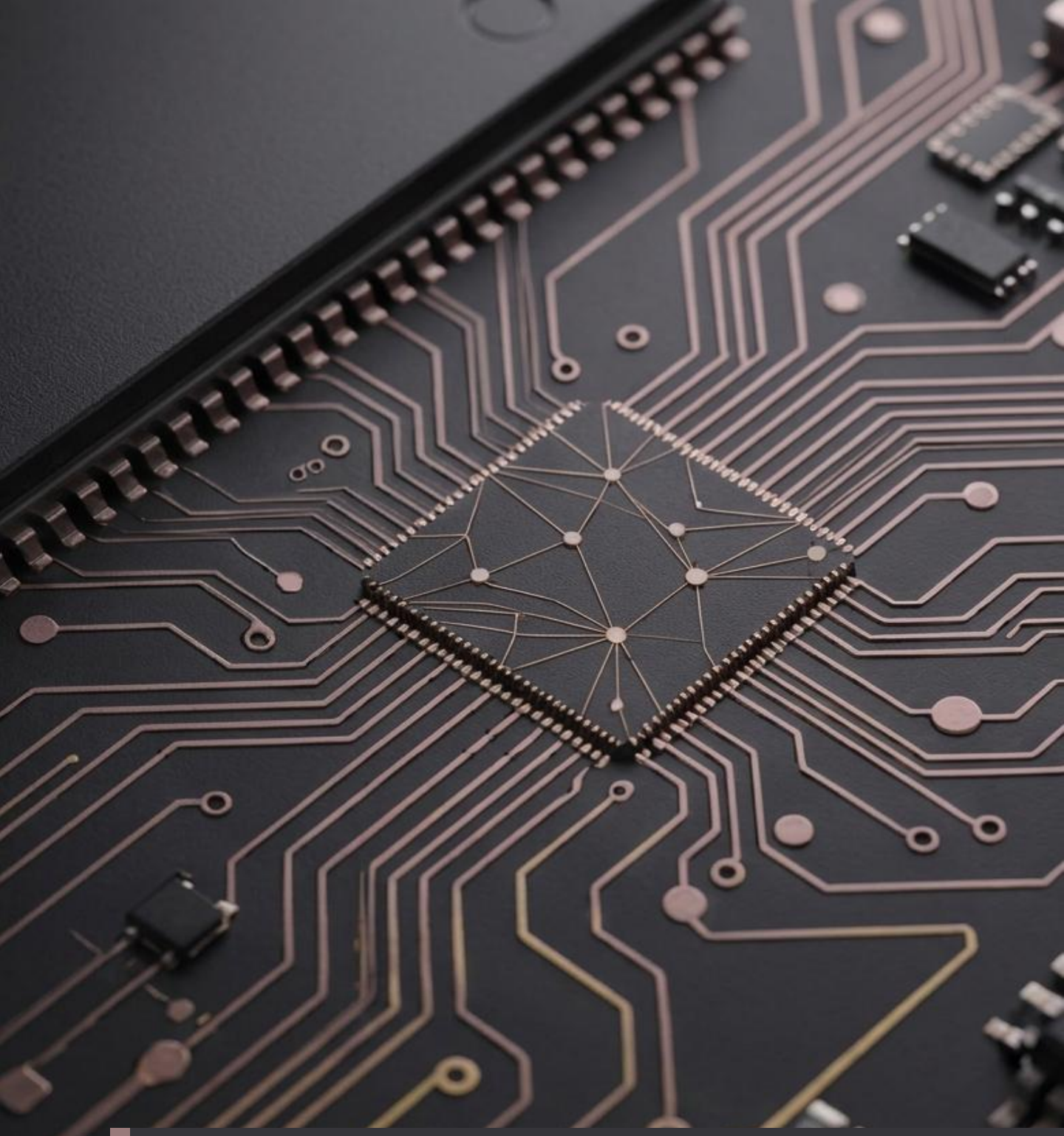
端侧AI通过本地处理减少数据传输并支持实时决策；边缘计算则将云计算能力部署至网络边缘，为区域低时延服务和多设备协同提供支撑。

- **端侧AI芯片承担的关键环节——端云任务分工进一步决定端侧芯片的功能边界；通常需优先由端侧或近端节点承载**

1. **数据预处理与多模态融合：** 接收摄像头、雷达、麦克风及各类传感器数据，完成降噪、编码、特征提取、时间同步与信息融合。
2. **模型加载与本地推理：** 调用端侧视觉、语音、语言及控制模型，完成目标识别、环境理解、状态判断和结果输出。
3. **现场决策与任务调度：** 依据推理结果进行任务规划，并协调CPU、GPU、NPU、DSP及MCU等计算单元执行不同负载。
4. **低时延指令与设备控制：** 将决策结果转化为执行器指令，控制机器人运动、车辆响应及工业设备运行，保障控制链路的实时性与稳定性。
5. **端云通信与模型更新：** 完成数据编解码、通信管理、运行状态上报和模型版本更新，维持端侧能力与云端知识体系同步。

英伟达的嵌入式AI平台也采用多类计算单元与多传感器接口，在终端侧并行运行模型并处理传感器数据，体现出端侧计算对异构协同和实时处理的需求。

来源：ARM，ETSI，英伟达，头豹研究院

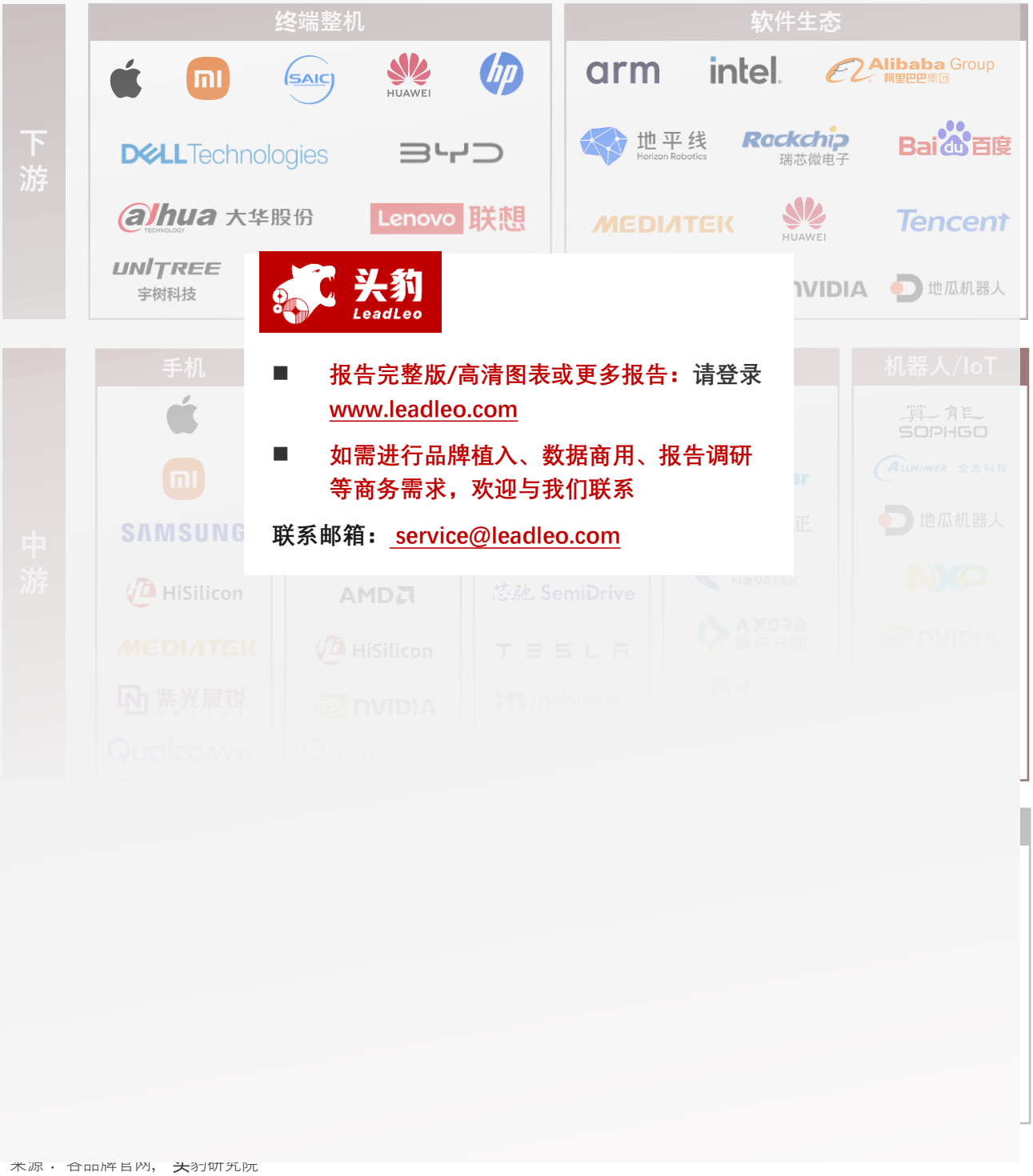


02 端侧AI芯片的产业链现状与竞争体系

端侧AI芯片的产业链全景

- 端侧AI芯片产业链涵盖上游IP与架构、EDA工具、晶圆制造及封装测试厂商，中游面向手机、AI PC、车载、视觉及机器人等场景的芯片设计与计算平台供给，下游终端整机与软件生态厂商

端侧AI芯片产业链全景图谱



头豹 LeadLeo

- 报告完整版/高清图表或更多报告：请登录 www.leadleo.com
- 如需进行品牌植入、数据商用、报告调研等商务需求，欢迎与我们联系

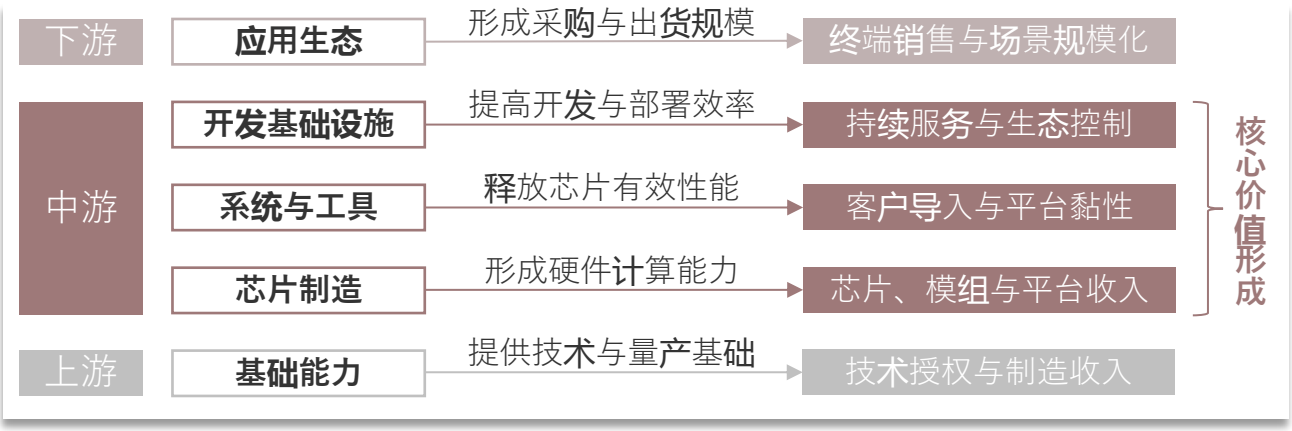
联系邮箱：service@leadleo.com

来源：各品牌官网，头豹研究院

端侧AI芯片产业链的核心价值落点在哪些环节？

- **产业的核心价值主要集中于芯片产品、系统与工具、开发基础设施三个中间环节，三者共同控制端侧算力从硬件形成、工程部署到持续使用的完整商业化过程，并直接影响厂商收入、产品差异化及客户黏性**

端侧AI芯片产业链各环节价值落点



- **端侧芯片产业链可分为基础能力、芯片产品、系统与工具、开发基础设施、应用与生态五层，其中核心价值主要集中于中间三层，分别承担算力形成、算力释放与规模部署，共同决定芯片能力能否转化为可规模落地的终端智能**

芯片产品层将CPU、GPU、NPU、存储等集成为SoC或AI芯片，在性能、功耗、成本与可靠性之间实现平衡，进而形成芯片销售收入，并可叠加IP/软件授权与定制开发收入，是商业价值实现的第一环；**系统与工具层**通过编译器、SDK、运行时、模型转换及异构调度，将理论算力转化为模型可调用能力，降低适配成本、提高算力利用率和开发效率，并形成平台黏性；**开发基础设施层**覆盖模型优化、仿真验证、部署更新、设备管理及数据回流，连接云端训练与端侧运行，使一次性硬件交付延伸至持续的软件与设备服务。中间三层共同控制端侧算力从形成、使用到规模化运行的核心过程，是国产端侧AI芯片厂商形成差异化、客户黏性与持续收入的主要价值落点。

产业链两端同样具有重要支撑价值，但直接价值获取逻辑有所不同：基础能力层提供IP、EDA、晶圆制造及封装测试等设计量产条件，关键能力和产业资源集中于少数头部供应商，国产端侧AI芯片厂商通常通过授权、采购或委外方式获得，更多体现为研发生产所需的外部投入与成本，较难在该环节形成自身持续差异化；应用与生态层负责将算力转化为终端功能、场景体验和用户价值，但其竞争壁垒更多来自场景积累、终端品牌、渠道及用户规模，价值分配较为碎片化，并依赖中间三层提供稳定的芯片、开发工具与部署能力，因此其主要商业价值更多由终端整机及应用服务厂商获取，对国产端侧AI芯片厂商的直接价值获取相对有限。

来源：ARM，高通，英伟达，头豹研究院

不同类型端侧AI芯片厂商依靠哪些能力竞争？(1/2)

- **参与厂商已分化出四类竞争路径：国际综合型以架构与生态建立基准，国内平台型以成本与灵活适配拓展增量，垂直场景型以专用架构与场景Know-How深耕需求，终端自研型以全栈协同形成系统级差异化**

端侧AI芯片4大类型厂商能力坐标

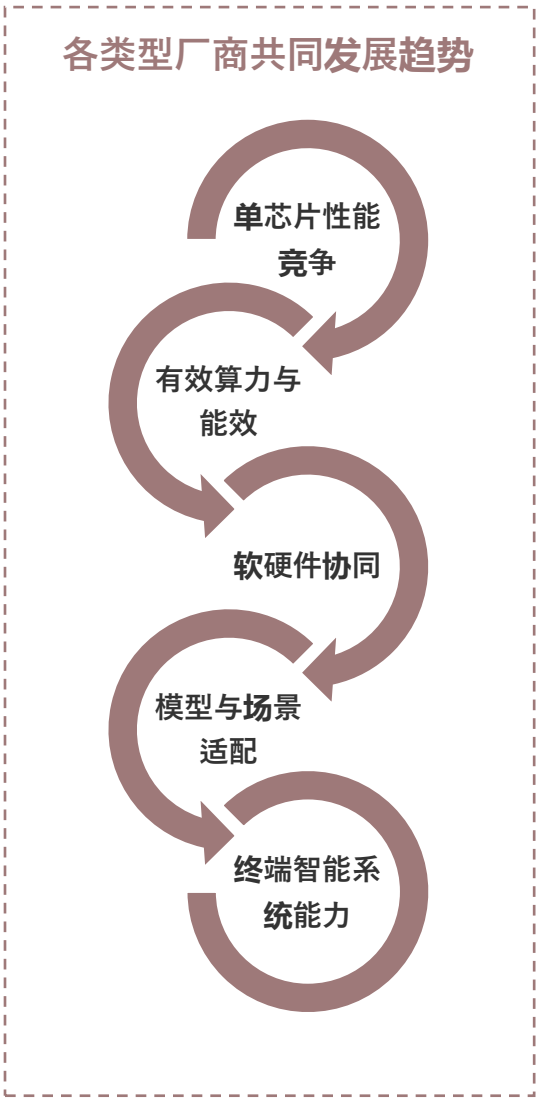


- **端侧芯片厂商可依据场景/终端控制深度和系统平台化能力分为国际综合平台型、国内平台型SoC、垂直场景型和终端自研型四类；不同类型厂商的竞争优势来源不同，并由此形成差异化发展路线**

1. **国际综合平台型厂商**以高通、NVIDIA、联发科、恩智浦等为代表，核心竞争优势来自成熟计算架构、芯片产品矩阵、软件工具链和全球开发者生态，其架构、IP或软件栈具备跨多个终端复用能力，具体覆盖范围因厂商而异。其竞争重点正由单芯片性能进一步延伸至跨终端统一架构和软硬件平台，通过模型工具、运行时和开发生态提高客户迁移成本。（接下文）

来源：高通，苹果，地平线，头豹研究院

不同类型端侧AI芯片厂商依靠哪些能力竞争？(2/2)



（接上文）

2. **国内平台型SoC厂商**以瑞芯微、紫光展锐、晶晨股份、全志科技等为代表，核心**优势**在于本土客**户**覆盖、成本与功耗控制、快速**产**品迭代和灵活**场**景适配，通**过**通用SoC平台覆盖消**费电**子、AIoT、工**业**、**车载**及机器人等多类**终**端。未来**发**展重点将由中低算力和**单**一芯片供**给**向更高算力、更强模型兼容性及完整工具**链**升**级**，通**过**芯片、算法和系**统**能力提升平台附加价**值**。

3. **垂直场景型厂商**以地平线、黑芝麻智能、爱芯元智、星宸科技、富瀚微、地瓜机器人等为代表，聚焦**车载**、**视觉**或机器人等特定**终**端，以**专**用**计**算架构、**场**景算法、工程适配及客**户联**合开**发**形成**竞**争壁**垒**。其**发**展路**线**将沿**场**景**纵**深持**续**强化，从**单颗专**用芯片**扩**展至算法、工具**链**和系**统**解决方案，并逐步覆盖更高算力和相**邻场**景。

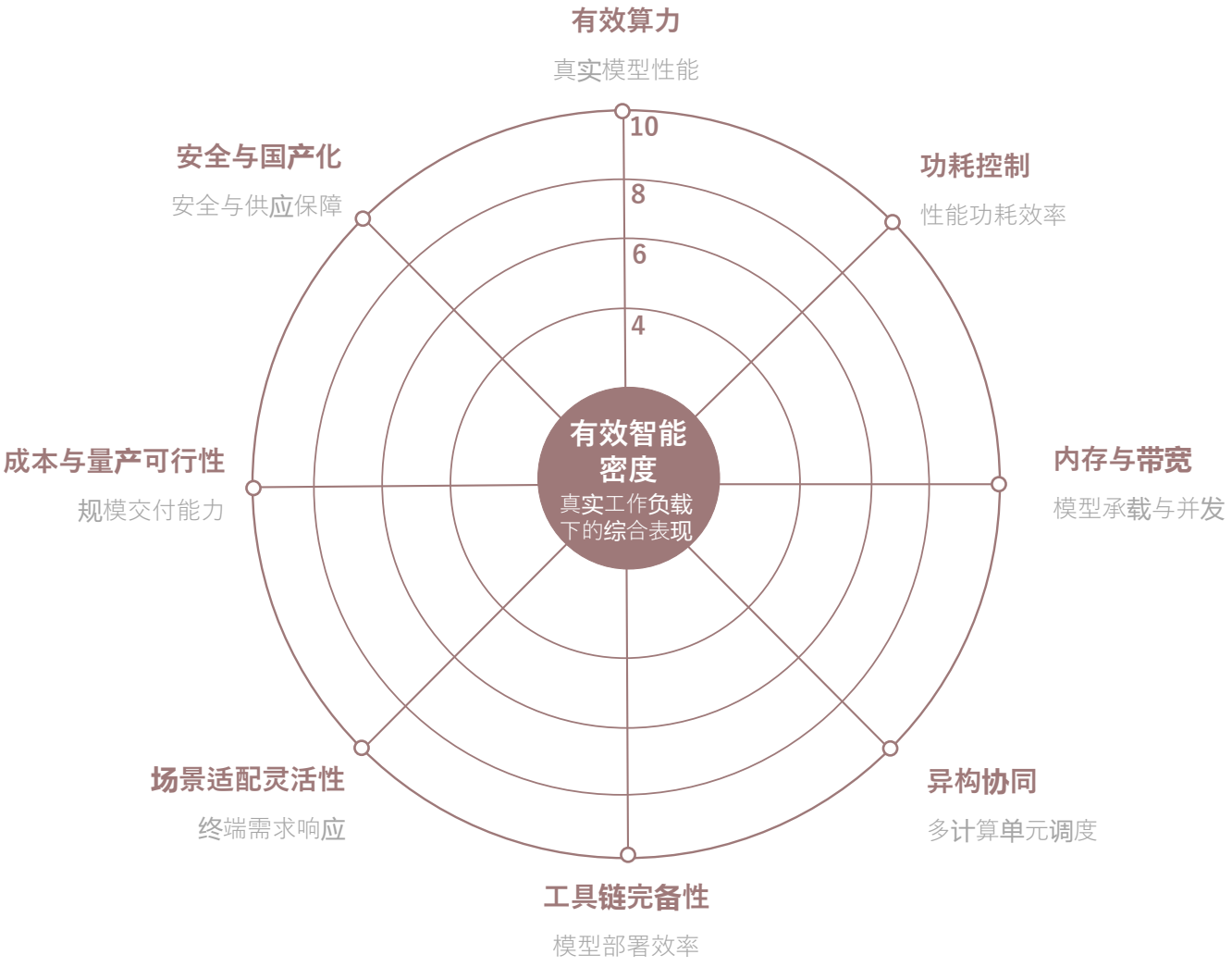
4. **终端自研型厂商**以苹果、海思、玄戒科技、特斯拉、小鹏和蔚来等为代表，**竞**争**优势**来自**终**端定**义**权、操作系**统**、模型、用**户**需求和硬件平台的一体化**协**同，可根据自身**产**品需求定制芯片，在性能、功耗和用**户**体**验**之**间进**行系**统级优**化。未来将**进**一步强化“芯片—系**统**—模型—**终**端”垂直整合，以**专**用芯片承**载**更多本地AI任**务**；其主要**约**束在于芯片能力首先服**务**自有**终**端，第三方市**场规**模和开**发**者生**态**开放程度相**对**有限。

来源：高通，苹果，地平**线**，**头**豹研究院

当前哪些系统性指标能真正评判端侧AI芯片竞争力？(1/2)

- 端侧AI芯片竞争评价体系正由单一峰值算力转向真实工作负载下的综合表现，领先者需在有效算力、功耗控制、内存与带宽、异构协同、工具链、场景适配、成本与量产、安全合规八个维度形成系统协同

端侧AI芯片“有效智能密度”评价指标雷达图



传统TOPS指标难以完整反映真实端侧AI性能

TOPS是衡量特定精度与峰值频率条件下的理论计算能力，能够反映芯片算力上限，但真实模型运行还受到内存带宽、模型精度、异构调度、功耗散热及软件优化等多重约束。因此，相近TOPS的芯片在实际任务中的推理速度、能效、稳定性及部署效果仍可能存在明显差异，评价端侧AI芯片需进一步考察真实工作负载下的系统表现。（接下文）

来源：MLCommons，高通，头豹研究院

当前哪些系统性指标能真正评判端侧AI芯片竞争力？(2/2)

（接上文）

- **“有效智能密度”评价体系由有效算力、功耗控制、内存与带宽、异构协同、工具链完备性、场景适配、成本与量产可行性以及安全与国产化八项核心指标构成**

有效算力关注真实模型在既定精度下的稳定推理能力，可观察推理延迟、吞吐量、FPS及TTFT/TPOT等表现；功耗控制衡量受限功耗与散热条件下的持续性能，重点观察性能功耗比、单次推理能耗及降频情况；内存与带宽反映模型承载、数据供给与多模型并发能力，可通过内存容量、带宽、模型占用及数据搬运效率衡量；异构协同关注CPU、GPU、NPU等计算单元的任务分工效率，重点观察调度开销、利用率及并行能力。工具链完备性衡量模型迁移、编译、量化、调试及部署效率，核心观察框架兼容性、算子覆盖与开发工具支持；场景适配灵活性反映芯片应对不同模型、终端形态及实时任务的能力，可观察接口与模型适配范围、响应能力及定制周期；成本与量产可行性关注芯片成本、系统BOM、良率、供货稳定及规模交付能力；安全与国产化则从本地数据处理、模型保护、关键IP自主性及国产软硬件适配等方面评价安全与供应保障水平。

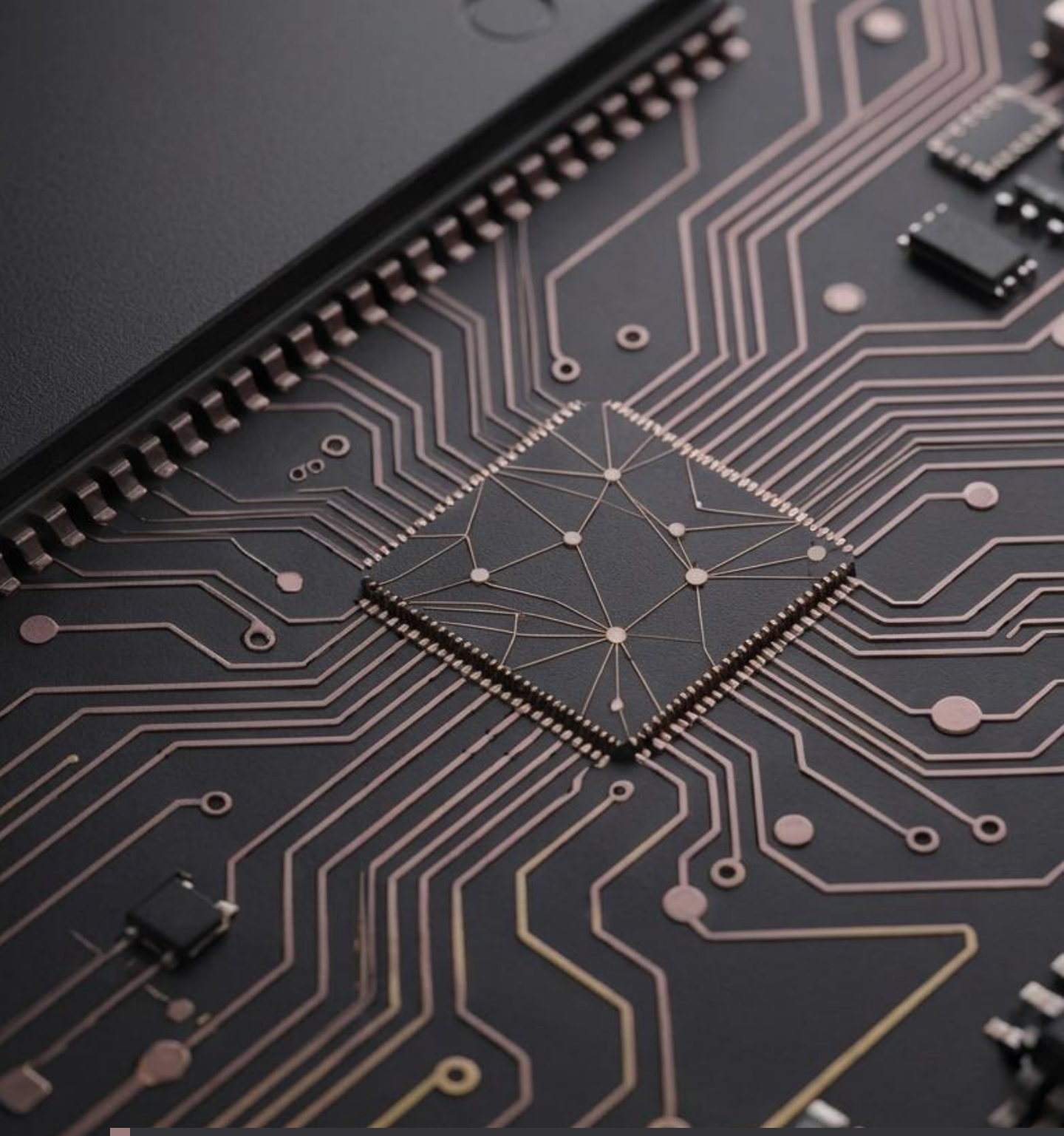
- **评价维度权重随应用场景调整**

不同终端可采用统一八维框架，但需根据任务特征和部署约束配置评价权重，例如：机器人重点关注有效算力、异构协同与场景适配，手机更强调功耗控制、内存带宽与成本，工业设备则优先考察稳定运行、安全国产化及规模交付能力。

- **“有效智能密度”评价体系将竞争力判断从单一峰值参数扩展至真实系统能力**

相比TOPS主要回答理论上能提供多少算力，八维体系进一步回答算力能否有效释放、能否持续运行、模型能否高效部署以及产品能否规模落地。端侧芯片竞争由算力参数逐步转向有效智能密度、功耗效率、工具链、场景适配及量产能力的系统竞争，领先厂商的核心特征将体现为软硬件、开发生态与终端场景之间的系统协同。

来源： MLCommons，高通，头豹研究院

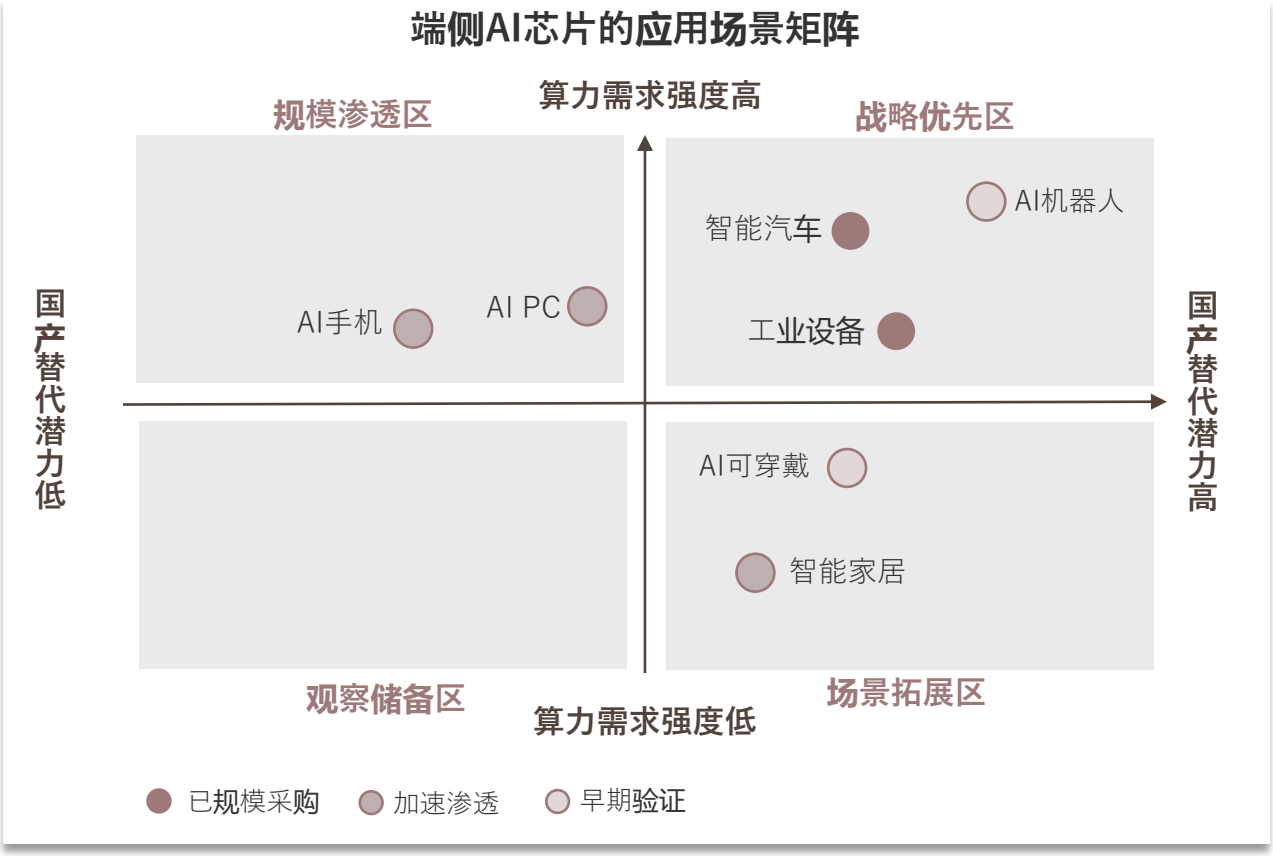


03 端侧AI芯片的场景落地与规模化约束

端侧AI芯片将率先在哪些应用场景形成规模化需求？(1/2)

- 端侧AI芯片将率先在智能汽车与工业设备形成规模需求，AI PC和AI手机依托存量终端加速渗透；通用/人形机器人短期仍以场景验证为主

端侧AI芯片的应用场景矩阵与场景拆解



- **端侧AI芯片的落地节奏主要取决于三项条件：终端是否已经形成稳定出货规模，AI功能能否创造明确使用价值，以及芯片能否满足整机成本、功耗、可靠性与量产要求**

1. **智能汽车是当前确定性最高的增量市场。** 智能驾驶与智能座舱已经形成明确的芯片采购、车型适配和量产验证体系。其中，辅助驾驶需要在本地完成环境感知、目标识别、路径判断与紧急避障，智能座舱则持续增加语音、多模态交互和个性化服务需求。汽车场景单机芯片价值较高、产品生命周期较长，但功能安全、可靠性认证和整车验证周期也构成较高进入门槛。（接下文）

来源：专家访谈，星能资产，头豹研究院

端侧AI芯片将率先在哪些应用场景形成规模化需求？(2/2)

（接上文）

2. **工业设备是国产端侧芯片较具现实落地空间的市场。** 工业视觉、视频分析、AMR、AGV及无人工厂等场景已经形成实际采购需求，核心诉求集中于低时延、离线运行、数据可控和长期稳定供货。工业客户需求高度分散，通常需要芯片、SoM模组、算法和系统集成协同交付，国产厂商可依托本地适配、成本控制和服务响应建立差异化优势。

3. **AI PC与AI手机将依托既有终端规模推动端侧AI普及。** AI PC已普遍将NPU作为关键本地AI计算单元之一，主要承载本地办公辅助、内容处理、隐私数据分析及操作系统级AI功能；手机端AI则由旗舰机型逐步向中端机型渗透，并向语音、多模态交互和生成式AI扩展。两类场景拥有成熟供应链和庞大设备基础，但需求释放仍取决于超级应用、用户付费意愿及AI功能的实际使用频率。

4. **AI可穿戴设备处于产品形态与需求同步验证阶段。** 智能眼镜、耳机等设备具有高频交互入口优势，可承载语音助手、环境感知、实时翻译和视觉识别等功能。其主要限制来自电池容量、散热空间、重量与成本，芯片需要在较低功耗下完成持续感知和轻量推理，因此对能效与集成度的要求高于单纯峰值算力。

5. **机器人具有较高算力需求和国产平台导入潜力，但短期商业化成熟度有限。** 机器人需要协同处理感知、理解、决策、控制、反馈及多模型调度，对异构计算、接口能力和工具链提出综合要求。当前通用机器人仍处于技术与场景验证期，规模采购尚未充分形成；工业机器人、AMR和AGV等任务边界较清晰的产品将更早产生需求。随着机器人架构尚未完全固化，国产芯片厂商仍具备提前进入整机平台并形成深度适配关系的机会。

6. **智能家居等低算力终端将形成广泛但分散的基础需求。** 摄像头、门锁、家电和家庭控制设备可通过本地语音、视觉识别及状态判断提升响应速度与隐私水平，但其单机价值较低、价格敏感度较高，需求更多体现为芯片出货规模，而非高端算力价值。

来源：专家访谈，星能资产，头豹研究院

哪个应用场景将构成端侧AI芯片产业最高压验证场景？(1/2)

- **机器人需在受限功耗与散热下，并行完成感知、理解、决策、控制、反馈及多模型协同六项计算任务；芯片性能直接影响动作稳定、安全响应与任务完成率，使其成为检验端侧芯片系统能力的高压场景**

机器人的六类计算任务及对端侧AI芯片的相应需求

机器人计算任务由单点推理升级为完整智能闭环

具身智能的发展推动机器人计算重心由导航和运动控制，逐步转向感知、理解、决策、控制与反馈一体化。机器人需要持续接收视觉、语音、触觉、位置及关节状态等多源数据，在动态环境中完成信息融合、空间理解、任务规划和动作生成，并根据执行结果实时修正决策。

上述六类计算任务需要持续并发运行，且在模型规模、计算精度、响应周期与可靠性方面存在显著差异。机器人端侧计算平台因此需要具备AI推理、通用计算、实时控制、数据处理及多任务调度能力，单一计算单元难以覆盖完整任务链条。

计算任务	计算内容	对端侧AI芯片的核心需求
多模态感知	处理视觉、语音、触觉、位置等传感器数据，完成图像处理、目标识别、深度估计和状态检测	• NPU/ISP • 多路I/O • 高带宽数据传输能力
环境理解	融合视觉、语音、空间位置和设备状态，识别环境关系、物体属性及人的行为意图	• 多模态模型 • 较高内存容量与带宽 • 较高模型算子兼容性
任务决策	将用户指令转化为可执行的任务序列，完成任务拆解、路径规划、动作选择及异常处理	• Transformer/VLA推理 • 异构计算
运动控制	将决策结果转化为步态、平衡、关节、力矩及灵巧操作指令	• MCU/实时CPU • 确定性低时延
实时反馈	持续读取机器人姿态、动作结果和环境变化，判断执行偏差并快速调整控制策略	• 具备高速I/O • 实时操作系统支持 • 低时延任务调度 • 故障响应能力
多模型协同	感知、定位、规划、控制和交互模型需要并发运行，并共享传感器数据与中间结果	• CPU/GPU/NPU/MCU异构调度 • 共享内存 • 工具链

来源：漫谈Light the Star， 头豹研究院

哪个应用场景将构成端侧AI芯片产业最高压验证场景？(2/2)

机器人端侧计算闭环图



端云协同成为机器人智能的基本架构

云端依托更充足的算力，承担大模型训练、复杂任务规划、知识更新、仿真学习及数据闭环；端侧承担多模态感知、本地推理、运动控制和安全响应等实时任务。端侧模型可通过云端训练、蒸馏及参数更新持续迭代，并将新版本下发至终端，形成云端能力生成、端侧实时执行、运行数据反馈、模型持续优化的协同闭环。

端侧AI芯片的核心作用，是在网络不稳定或云端响应不足的情况下，维持机器人基本感知、决策和控制能力，并将必须实时完成的计算任务稳定留在本地。

国产机器人芯片开始进入大脑与小脑协同阶段

国产计算平台已由单一视觉感知和AI推理，逐步扩展至多模态理解、任务决策与实时运动控制。部分厂商通过集成CPU、NPU与MCU，或采用大脑芯片与小脑芯片组合方案，同时覆盖高算力推理与确定性控制。

当前具身算法、机器人形态及应用场景仍在快速迭代，计算架构与开发基础设施尚未完全收敛。国产平台后续竞争重点将集中于复杂模型实际运行效率、大小脑协同、模型与算子兼容性、开发工具成熟度、长期运行稳定性及规模量产成本。

来源：漫谈Light the Star， 头豹研究院

哪些因素正在制约端侧AI芯片产业爆发？(1/2)

- **端侧AI芯片产业爆发当前受五重因素制约：硬件有效算力不足、软件工具与模型适配不成熟、超级应用及用户价值缺失、场景交付分散，以及供应链与数据闭环保障不足**

端侧AI芯片需求爆发制约因素矩阵



头豹 LeadLeo

- 报告完整版/高清图表或更多报告：请登录 www.leadleo.com
- 如需进行品牌植入、数据商用、报告调研等商务需求，欢迎与我们联系

联系邮箱：service@leadleo.com

- **端侧AI芯片部分场景已进入初期放量，但全行业尚未形成稳定、持续的规模增长，当前产业制约因素可归纳为底层硬件、软件工具、应用需求、场景交付与产业保障层面**

1. **底层硬件层面制约：**先进制程、SoC集成、内存带宽及整机能效共同限制有效算力释放。 先进制程受限会压缩芯片性能、能效与集成度的改善空间；SoC集成成熟度不足，则影响CPU、GPU、NPU、内存控制器及多媒体接口的稳定协同。模型规模和多模态数据持续增长，使内存容量与带宽逐步成为推理瓶颈，数据读取速度不足会导致计算单元利用率下降。端侧设备还受到整机功耗预算和散热能力约束，持续推理容易引发温升、降频与性能波动，使峰值算力难以稳定转化为实际运行能力。（接下文）

来源：专家访谈，头豹研究院

哪些因素正在制约端侧AI芯片产业爆发？(2/2)

（接上文）

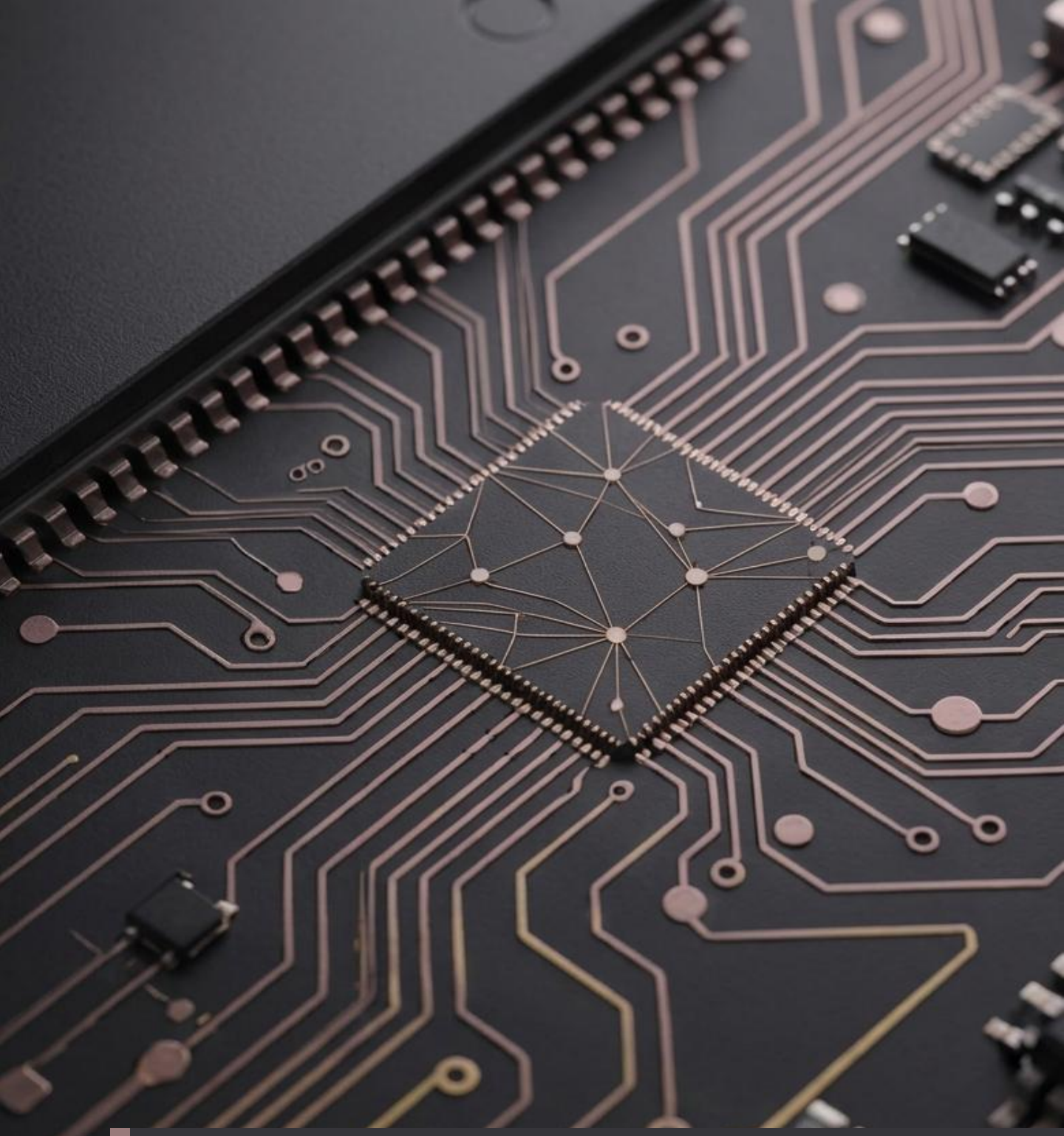
2. **软件工具层面制约：**编译器、算子、部署工具及模型适配能力尚未形成完整闭环。 编译器成熟度不足会降低模型向芯片架构的映射效率，主流算子覆盖不完整则需要客户重写算子或调整模型结构。模型转换、性能分析、调试和仿真工具相对分散，进一步增加部署流程的复杂度。同时，量化、剪枝和蒸馏可能造成模型精度损失，不同芯片与模型之间的适配差异也会影响有效算力调用，使模型部署仍依赖较多定制开发和反复调优。

3. **应用需求层面制约：**超级应用缺失与用户价值验证不足，削弱终端厂商持续增加AI配置的动力。 手机、AI PC和可穿戴设备虽然拥有较大的终端基础，但部分AI功能仍停留在硬件预装和体验验证阶段。缺少高频、稳定且难以由云端替代的超级应用，使端侧AI使用频率难以持续提升；功能带来的效率改善、体验增益和付费价值尚未充分验证，也使芯片、内存及散热成本难以转化为明确收益，终端出货规模因而无法直接形成稳定的端侧芯片增量。

4. **场景交付层面制约：**需求碎片化与客户验证周期降低芯片平台的跨场景复制效率。 汽车、工业、机器人、手机和可穿戴设备在接口、模型、功耗、可靠性和采购方式方面差异明显，同一芯片平台通常需要针对具体终端重新完成硬件设计、模型适配和系统集成。汽车与工业场景还需要经历行业认证、功能安全、长期运行及量产验证，机器人等新兴终端的产品架构与技术路线尚未完全收敛。高度定制化和较长验证周期共同推高项目交付成本，延缓芯片由单点导入转向规模复制。

5. **产业保障层面制约**：IP授权、供应链稳定性与数据闭环能力影响长期量产和持续迭代。 核心架构及关键IP授权依赖会增加产品持续开发和路线规划的不确定性，先进产能、关键设备及长期供货能力则直接影响新品导入与客户量产决策。端侧模型还需要通过设备运行数据、云端训练和模型更新实现持续优化，但隐私合规、数据回流成本、设备连接状态及跨平台接口可能阻断数据闭环。若训练、验证、更新与端侧执行无法稳定衔接，芯片平台将难以支撑大规模设备的长期运维和能力迭代。

来源： 专家访谈， 头豹研究院



04 端侧AI芯片的产业机遇与演进方向

产业重构后，端侧AI芯片产业链的机会应如何分层？

- **端侧AI芯片产业机会可分为芯片设计与IP、模组与计算平台、系统软件与开发平台、终端与场景应用四层；机会由底层算力供给逐级延伸至工程交付、开发效率与终端需求，形成相互牵引的增量结构**

端侧AI芯片产业链的机会分布图

汽车、工业、AI PC、手机、机器人、可穿戴

由本地智能持续渗透，实时控制和数据本地处理需求增长驱动

SoM、开发板、机器人计算平台、参考设计

因场景集成复杂，客户需要降低开发和量产门槛而驱动形成



编译器、SDK、模型部署、仿真、设备管理、端云协同

因跨芯片适配以及重复开发成本较高驱动形成

AI SoC、NPU、异构计算、内存互联、关键IP

因端侧算力升级与关键能力国产化驱动形成

- **端侧AI芯片产业机会贯穿芯片与IP、模组与计算平台、系统软件及终端应用，分别承接算力供给、集成适配、部署开发与采购落地**

芯片与IP：算力升级与国产化构成基础机会。多模态推理与物理控制进入终端，推动芯片向CPU、GPU、NPU、MCU、存储及接口协同的异构SoC演进，扩大AI SoC、专用NPU、低精度计算及关键IP需求。国产厂商可凭能效、场景适配和本地服务切入汽车、工业及机器人市场，但仍受制程、生态、SoC完整度和量产能力约束。**模组与计算平台：预集成降低终端导入门槛。**集成芯片、存储、接口、基础软件及模型工具的平台可减少底层开发与重复适配，缩短验证和量产周期，工业设备与机器人成为主要需求方向。**系统软件与开发平台：工具链短板释放平台机会。** 编译器、SDK、运行时、算子库及异构调度决定算力利用效率；机器人还需仿真训练、数据回流、设备管理和端云协同。平台正向算法、数据、模型与硬件一体化延伸，跨芯片标准及商业模式仍待验证。**终端应用：成熟场景率先放量，新终端提供增长弹性。**智能汽车、工业设备、AI手机及AI PC构成当前主要需求，AMR、AGV和工业视觉已形成采购；工业及商用机器人具备中期放量潜力，人形机器人、家用机器人及AI可穿戴仍需验证产品价值与规模成本。

来源：专家访谈，头豹研究院

AI机器人产业链中的新机会有哪些？

- **作为端侧AI芯片的最高压验证场，机器人产业落地依赖云厂商、芯片与计算平台、基础设施公司、本体厂商及开发者协同；其中云厂商的下一轮竞争机会是成为机器人和AI终端时代的开发基础设施**

AI机器人产业参与者协作闭环图



最高压验证场推动产业由单点供给转向系统协作

机器人需在受限功耗、散热和网络条件下，并行完成多模态感知、任务决策、运动控制与实时反馈，对有效算力、内存带宽、异构调度、工具链及可靠性提出综合要求。单一企业难以覆盖模型训练、端侧计算、软件部署、本体集成和场景开发全流程，机器人因此成为检验端侧AI系统能力的最高压场景，推动产业形成多主体协作体系。

五类参与者构建端云协同闭环

云厂商提供算力、模型、仿真、设备管理及数据回流；芯片与计算平台厂商承担传感器接入、本地推理、异构调度和实时控制；基础设施公司通过操作系统、中间件及部署工具实现能力封装；本体与场景厂商负责产品定义、系统集成、量产和交付；开发者调用模型、算法和机器人技能推动应用迭代。五类主体围绕云端训练、基础设施适配、端侧执行、数据回流和模型优化形成闭环。

云厂商向机器人及AI终端开发基础设施延伸

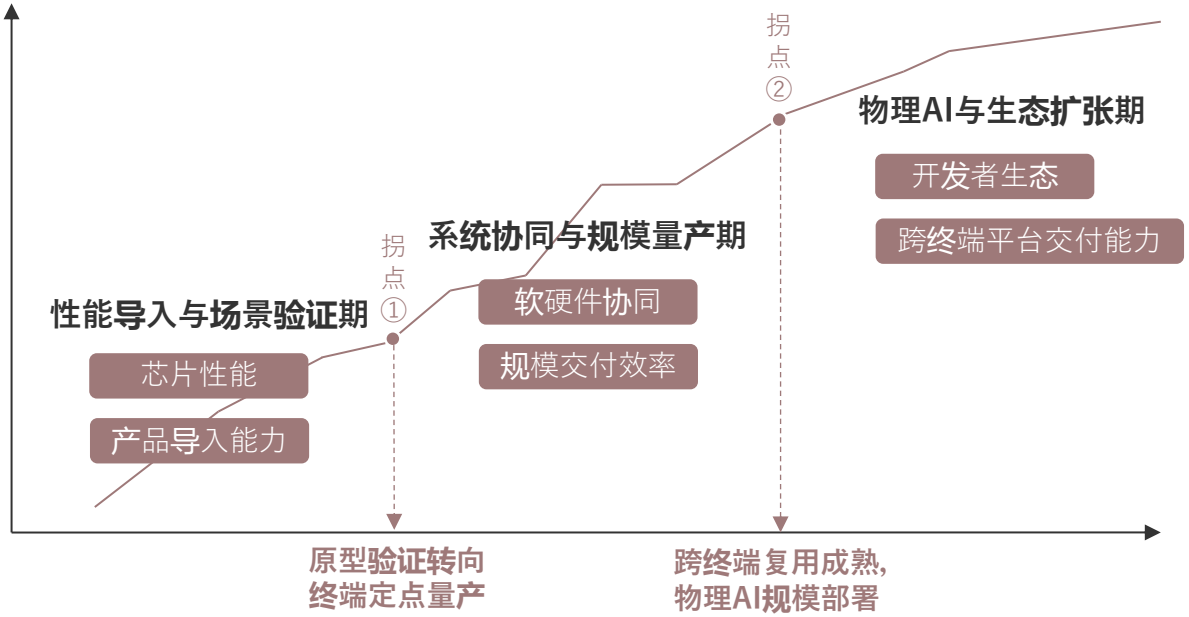
机器人开发新增云端GPU、AI云工作站、预装环境、仿真训练、模型部署和设备管理需求。预装镜像可显著缩短环境配置时间，推动云服务由资源供给向研发提效和全流程管理扩展。云厂商可整合计算、存储、网络、模型及平台能力，贯通开发、训练、部署、运行与更新，成为机器人及AI终端的开发基础设施；规模化进程仍取决于共性工具、跨平台接口、开发者生态和数据闭环的成熟度。

来源：专家访谈，头豹研究院

中国端侧AI芯片产业将如何持续演进？

- **中国端侧AI芯片产业将沿性能导入、系统协同与物理AI生态化三阶段演进，竞争焦点由芯片参数逐步转向软硬件协同与规模交付，最终延伸至跨场景平台与开发者生态能力**

端侧AI芯片产业演进路线



- **产业将沿性能导入与场景验证、系统协同与量产、物理AI生态化扩张三阶段持续演进**

1. **性能导入与场景验证期：**端侧AI需求首先推动芯片进入手机、AI PC、机器人、工业设备等智能终端进行模型适配与产品验证，产业竞争仍以芯片性能和产品导入能力为核心；厂商重点提升有效算力、功耗效率、内存带宽及模型兼容性，并通过开发板、参考设计和客户验证降低导入门槛。
2. **系统协同与规模量产期：**随着端侧模型和应用逐步成熟，竞争重点进一步转向软硬件协同与规模交付效率；芯片厂商需要完善SDK、编译器、量化工具及模型库，强化端云协同、异构调度和跨产品复用能力，并通过供应稳定、成本控制和标准化开发平台降低终端部署成本。
3. **物理AI与生态扩张期：**当机器人、自主设备等物理AI终端形成更广泛的规模应用，端侧芯片将进一步承担实时感知、推理、决策、控制及端云协同任务；竞争核心转向异构计算、实时可靠性、跨终端平台能力、模型持续迭代及开发者生态。

来源：漫谈Light the Star，专家访谈，头豹研究院

头豹
LeadLeo

头豹业务合作

全球视野 · 本土洞察 · 助力企业把握市场先机

核心业务

行业数据API

开放原创报告与研究数据接口，支持企业知识库、系统平台及AI应用高效接入和调用

KNIT解决方案

构建企业可信内容体系，提升品牌在AI搜索与问答中的可见度、准确性与转化效果

报告会员账号

可阅读全部原创报告和百万数据，提供PC及移动端，方便触达平台内容

定制报告/白皮书

对产业及细分行业进行现状梳理和趋势洞察，输出全局观深度研究报告

商业尽调

面向投资并购和商业决策，评估标的公司的商业前景、价值及风险

招股书引用

研究覆盖国民经济19+核心产业，内容可授权引用至上市文件、年报

业务咨询

客服电话：
400-072-5588

官方网站：
www.leadleo.com

商务咨询与深度合作

报告作者

袁栩聪
首席分析师

徐小雅
行业分析师

service@leadleo.com

办公地点

深圳办公室
广东省深圳市南山区粤海街道华润置地大厦E座4105室
邮编： 518057

上海办公室
上海市静安区南京西路1717号会德丰国际广场2701室
邮编： 200040

南京办公室
江苏省南京市栖霞区经济开发区兴智科技园B栋401
邮编： 210046

方法论

- 头豹研究院布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。

- 研究院依托中国活跃的经济环境，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。

- 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。

- 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。

- 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明

- 本报告著作权归头豹所有，未经书面许可，任何机构或个人不得以任何形式翻版、复刻、发表或引用。若征得头豹同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“头豹研究院”，且不得对本报告进行任何有悖原意的引用、删节或修改。

- 本报告分析师具有专业研究能力，保证报告数据均来自合法合规渠道，观点产出及数据分析基于分析师对行业的客观理解，本报告不受任何第三方授意或影响。

- 本报告所涉及的观点或信息仅供参考，不构成任何投资建议。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。在法律许可的情况下，头豹可能会为报告中提及的企业提供或争取提供投融资或咨询等相关服务。本报告所指的公司或投资标的的价值、价格及投资收入可升可跌。

- 本报告的部分信息来源于公开资料，头豹对该等信息的准确性、完整性或可靠性不做任何保证。本文所载的资料、意见及推测仅反映头豹于发布本报告当日的判断，过往报告中的描述不应作为日后的表现依据。在不同时期，头豹可发出与本文所载资料、意见及推测不一致的报告和文章。头豹不保证本报告所含信息保持在最新状态。同时，头豹对本报告所含信息可在不发出通知的情形下做出修改，读者应当自行关注相应的更新或修改。任何机构或个人应对其利用本报告的数据、分析、研究、部分或者全部内容所进行的一切活动负责并承担该等活动所导致的任何损失或伤害。

中国互联网协会人工智能工作委员会 | FROST & SULLIVAN 沙利文 | 头豹 LeadLeo | 君合 JUNHE | 嘉源律师事务所 JIA YUAN LAW OFFICES

中国端侧AI产业发展与应用实践蓝皮书（2026）

标杆案例征集



聚焦端侧AI产业链、终端场景、竞争格局与发展趋势

面向端侧AI产业链企业、投资机构及产业园区，形成一份兼具市场判断、产业链研究、标杆案例和趋势研判的专业蓝皮书。

核心服务形式

中文蓝皮书
精简、专业、适合公开传播

英文摘要版
提炼核心洞察与企业案例

企业案例专章
展示业务定位、核心能力与端侧AI实践

发布与传播
线上发布、媒体稿及定向传播

交流活动
邀请产业链企业、投资人与生态伙伴对接

合作企业获得什么

发布活动现场
蓝皮书预计2026年11月，拟于世界互联网大会乌镇峰会期间发布（具体安排以峰会组委会确认为准）

品牌表达
以第三方研究语言清晰表达企业定位、产品价值与端侧AI能力。

行业背书
进入代表性企业案例体系，强化在客户、渠道与资本市场中的专业认知。

市场洞察
获取产业链、终端场景与发展路径的系统判断，辅助战略决策。

资源连接
通过报告发布与定向活动连接产业链伙伴、客户、投资人与服务机构。

联系邮箱

案例征集与合作咨询

service@leadleo.com

欢迎来信咨询合作详情

中国互联网协会人工智能工委会、弗若斯特沙利文、头豹、君合、嘉源诚邀合作伙伴参与！