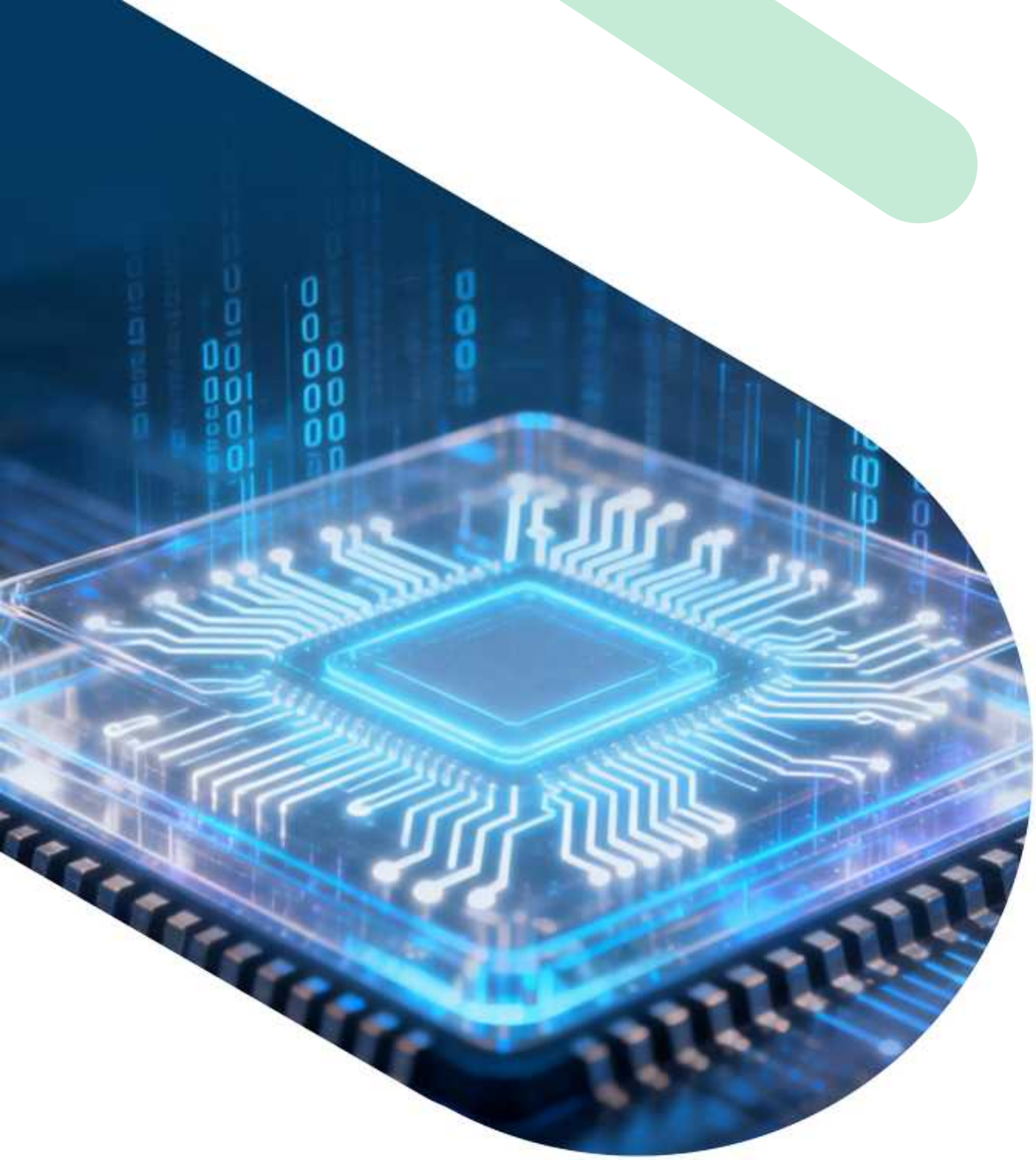


行业洞察 | 芯片

AI依旧是挖潜点，应用进一步多样化

文：魏德龄

“当当人们已经开始渐渐习惯，听听AI给出的参考建议。其背后的算力支撑也在逐步加强，像是更先进的制程，以及向更多端侧设备的拓展。另一方面，AI又一石激起千层浪，机遇与需求并行，面向未来进一步前行。

现状 | 围绕先进制程与AI迭代

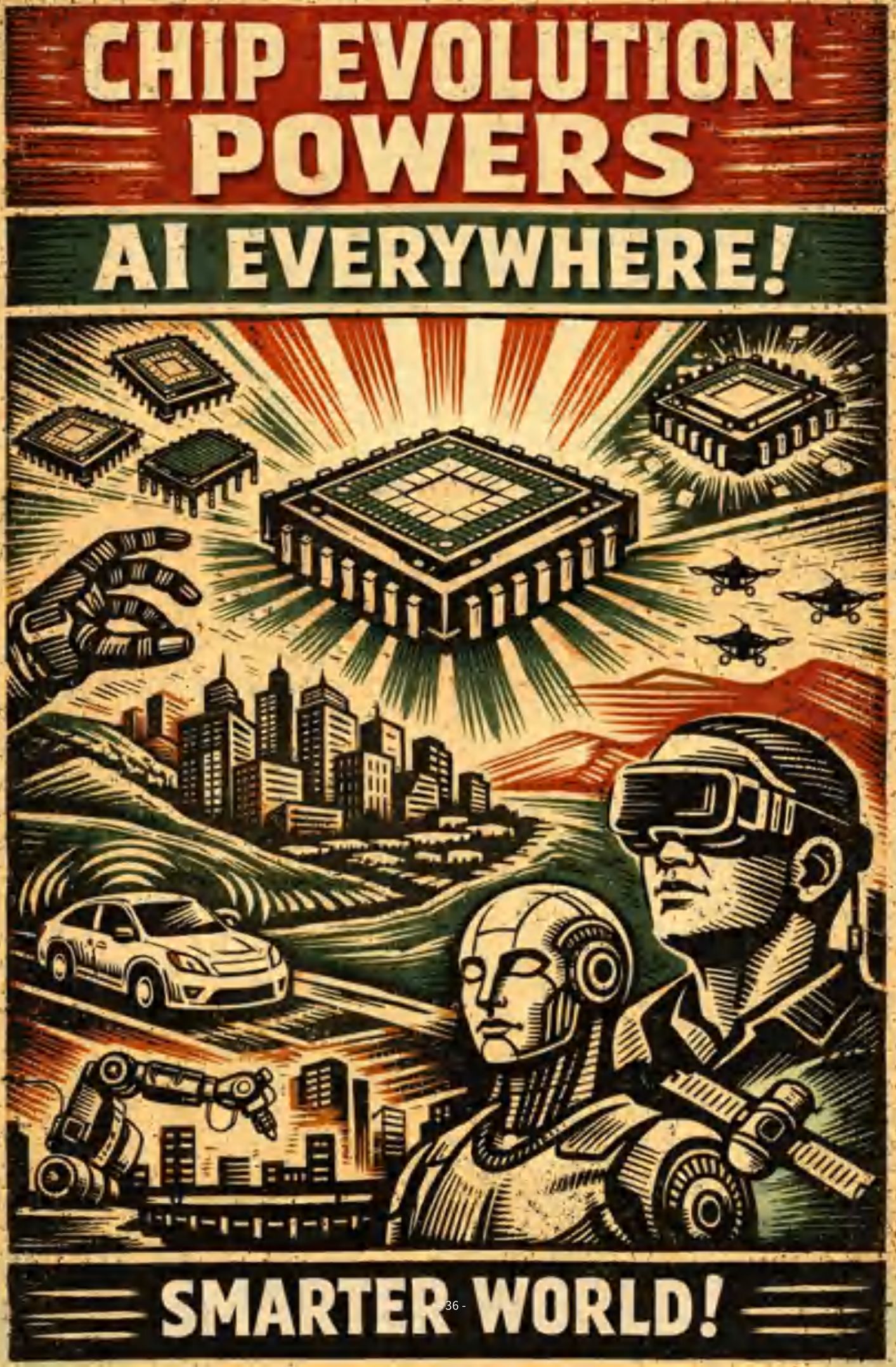
现状1：2nm时代蓄势待发

今年8月，台积电宣布其 2 纳米制程晶圆将以每片 3 万美元的固定高价供应，主攻 AI 与高性能计算等高端客户，强化“高端联盟”策略。三星则凭借更低价格和快速供货争取市场，近期成功获得价值 23 万亿韩元的特斯拉 AI 芯片订单。台积电计划在未来 34 个月内完成试产，2026 年实现 4 座工厂月产能达 6 万片的目标。初期 2 纳米工艺良率约为 60%～65%，其中 SRAM 存储单元良率更高，超过 90%。2 纳米工艺可在相同功耗下实现 10%—15% 的性能提升，或在相同性能下降功耗降低 20%—30%。

业内分析师预测，苹果将在明年秋季推出的iPhone 18系列上，正式采用台积电2nm制程工艺代工的芯片。预计iPhone 18 Pro、iPhone 18 Pro Max以及备受瞩目的折叠屏版本，将搭载全新的A20芯片。A20芯片不仅在制程工艺上有所升级，还将引入台积电最新的晶圆级多芯片封装技术（WMC-M）。这一技术将实现RAM与CPU、GPU以及神经网络引擎在同一晶圆上的高度集成，从而大幅提升内存带宽并减少延迟，为用户提供更加流畅的使用体验。据外媒报道，与采用3nm制程工艺的A19芯片相比，A20芯片的性能预计将提升15%，能效则将提高30%。

三星代工厂则以更低价格、加快供货速度吸引客户。目前三星 2 纳米良率约为40%，尚处于爬坡阶段，三星通过提供更低的价格和更灵活的供货，积极吸引新客户。自家Exynos 2600芯片有望在明年成为全球首款2nm手机芯片。随着DRAM和NAND闪存价格的持续上涨，三星正为提升良率、降低Exynos 2600的残次品数量展开相关试验。三星目标是在2025年底前将2nm GAA工艺的良率提升至70%，从而具备承接客户订单的能力。

英特尔在今年发布了代号为Panther Lake的下一代AI PC平台。这是首款基于最新Intel 18A制程工艺的客户端SoC。英特尔表示，这意味着产业将正式进入埃米时代。


CHIP EVOLUTION POWERS
AI EVERYWHERE!
SMARTER WORLD!

埃米时代的两大核心技术是RibbonFET全环绕栅极晶体管技术和PowerVia背面供电技术。英特尔是首家将这两项技术融合并实现量产的芯片厂商，将芯片的能效和密度推向了新的高度。基于Intel 18A制程推出的Panther Lake，正是工艺与架构智慧的集大成者。在相同功耗下，Panther Lake的多核性能提升了50%。Cougar Cove性能核、Darkmont能效核以及低功耗能效核的协同工作。Panther Lake的图形性能比上一代提升50%以上。它提供多达12个第三代Xe核心，每个都具备完整的矢量引擎、AI加速矩阵单元和硬件级光线追踪能力。2026年下半年，英特尔新一代至强6+处理器产品也将采用18A制程，带来更高性能与更低功耗。

AMD Zen 6架构将于2026年登场，迈入先进的2nm制程工艺时代。Zen 6架构将会新增支持多种AI数据类型，并增加更多的AI管线，这将使其在人工智能领域具备更强大的处理能力。代号为Venice（威尼斯）芯片目前已进入实验室阶段，表现极佳。相较于前一代Zen 5架构的都灵CPU，威尼斯在性能、效率与运算密度上取得了实质性的进步。AMD表示，Venice预计将在2026年如期上市，有望早于使用台积电最新节点的苹果产品。

现状2：本地AI能力支持

2025年初，DeepSeek成为国内AI领域的新宠。同时，也吸引了很多厂商通过本地运行满血或蒸馏版DeepSeek，来证明自身的AI算力表现，以及挖掘市场上的AI需求。让DeepSeek 正从传统“大模型 + 数据中心/云服务器”的运行方式，快速向“个人 PC / AI-PC /轻薄本 /笔记本 /终端设备端侧部署”转变。用户不必依赖云服务或外部服务器，就能在自己的 PC 上进行 AI 助手、文档写作、翻译、会议记录等任务，隐私性、数据安全性更高，也更方便在离线 /断网场景下使用。

英特尔发布了通过Flowy AI 助手支持本地DeepSeek R1 大模型的教程，使最新版 AI PC 助手支持 DeepSeek-R1 模型，为用户提供离线智能服务，并且针对酷睿Ultra 处理器进行了优化，专为酷睿™ Ultra 处理器的 CPU + GPU + NPU 异构架构打造，利用平台 XPU 的 AI 算力加速本地推理，让 AI PC 设备具备使用本地大模型的能力，实现翻译、会议纪要、文档撰写、设备控制等功能。

在年初的AMD AI PC创新峰会2025期间，面向前沿AI大模型，华硕通过深度整合，全面接入满血联网版DeepSeek R1 671B大模型，推出了全新的豆叮AI助手，兼顾不同场景，满足用户对AI应用的全方位使用需求。目前，从游戏娱乐到创作生产力，从高能硬件到高效软件，华硕已构建起完整的AIPC生态。

微软也在3月宣布，通过Azure AI Foundry接入DeepSeek的R1 7B和14B蒸馏模型，Copilot+ PC可实现本地运行这些模型。早在今年1月，微软就宣布要把DeepSeek-R1模型的NPU优化版本带到搭载高通骁龙X处理器的Copilot+ PC上。这些模型运行在NPU上，可以在减少对PC电池续航和散热性能影响的同时，持续获得AI计算能力，同时CPU和GPU可以执行其他任务，提高设备的整体效率。

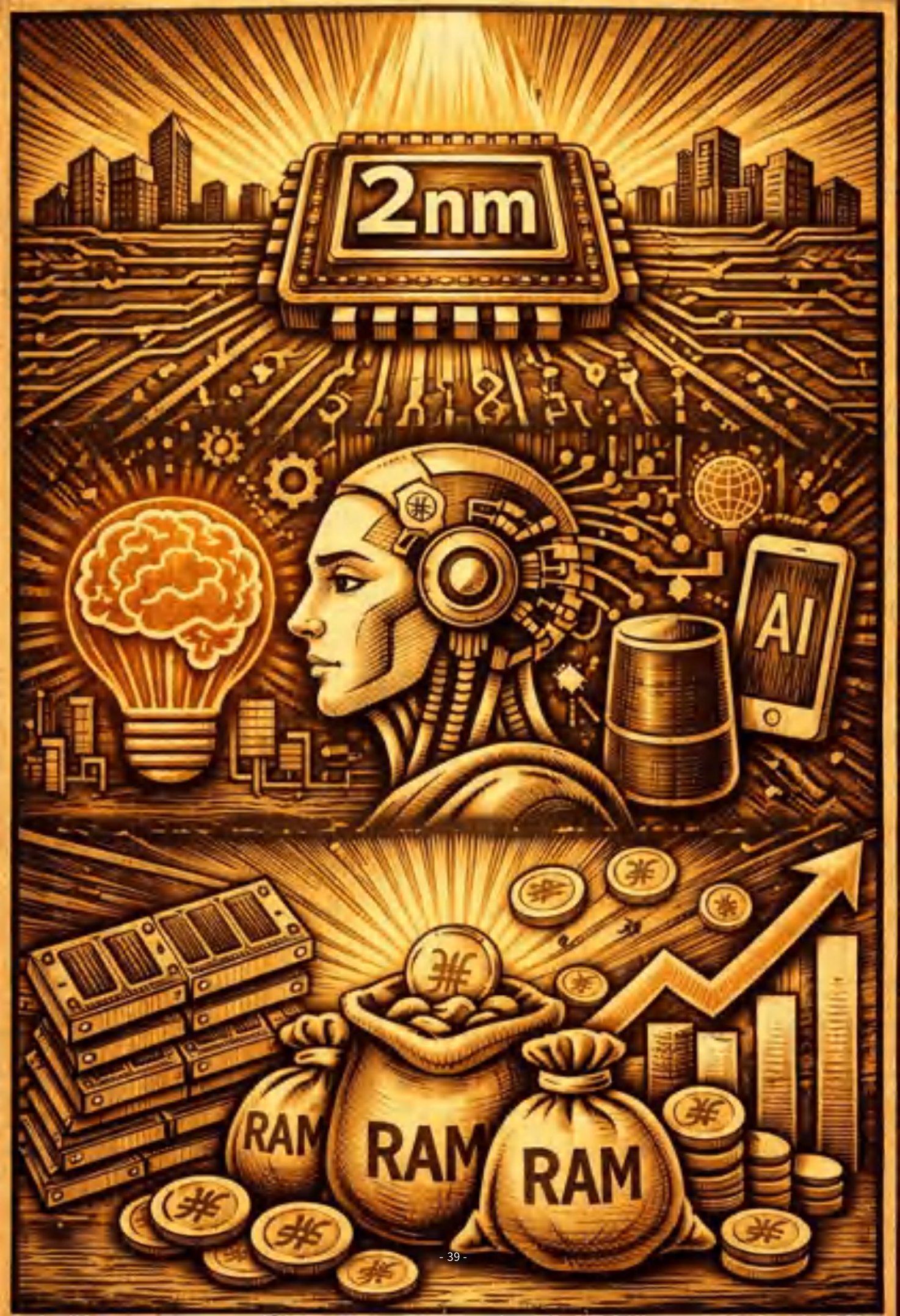
在智能手机侧，使用DeepSeek蒸馏后的Qwen-7B模型，已经能够在性能上与去年所推出的且当时最为先进的GPT-4o云端模型持平。但两个模型的参数规模却相差甚多。另外，从对比蒸馏后的Llama 700亿模型在推理、编程、数学、数据分析等方面的表现来看，同样已经超越了原始模型，只在**语言理解和指令遵循方面有待进一步优化**。而在今年很多展会现场，iQOO、努比亚、OPPO、荣耀、小米和一加等中国生态伙伴，均带来了基于骁龙平台的终端侧生成式AI和智能体AI的最新应用成果。三星也在Galaxy S25 Ultra上展示谷歌全新AI助手Gemini。Deepseek蒸馏模型涌现的背后是终端侧AI所迎来的全新机遇，使用户在本地也能获得媲美甚至超越云端的生成式AI能力，这种能力还正逐步演变为全新的交互方**式，让用户能够更加自然地与设备沟通**，引领智能终端迈向下一场变革。

现状3：内存价格飙升

从今年下半年开始，全球存储芯片市场掀起涨价潮。由于AI产业需求爆发，AI模型训练和大数据处理需要海量高带宽、低延迟的内存支撑，单台AI服务器对DRAM的需求是普通服务器的8倍。受此影响，大小容量存储芯片进入供应紧张态势。

另以4GB DDR4x为例，颗粒现货市场价格已从年初的最低7美金涨至11月中旬的30美金以上，涨幅达到4至5倍，Flash产品方面，以64GB eMMC为例，价格也从年初的3.2美金上涨至近期的8美金以上。11月份12GB LPDDR5X内存的价格已攀升至70美元，较年初翻了一倍多。

随着AI推理大模型的落地，存储产品不再是AI算力中可有可无的配件，而是成为AI基础设施集约化发展中的战略性物资。受全球AI算力需求激增影响，内存芯片价格飙升。面对内存疯狂涨价的局面，全球前两大存储巨头三星、SK海力士却拒绝扩大产量，而是以盈利考虑优先。三星和SK海力**士在疫情时期及之后经历了艰难的内存**周期，这也是目前生产线受限的原因。他们自然有自己考虑，如果内存公司大力投资扩充容量，一旦AI热潮消退，最终可能导致“供过于求”的局面。供应商预计，内存短缺可能持续到2028年，短期内状况预计不会改善，至少在未来一两个季度内，这意味着像**内存和GPU这样的产品将继续处于供应限制**之下。


2nm
AI
RAM
RAM
RAM

挑战 | 成本挑战与需求放缓

挑战1：不断攀升的成本

每年更新迭代，性能攀升的代价，就是成本的上升。以手机芯片为例，在智能手机总成本中占有重要比重，尤其随着存储（DRAM/NAND）和主控芯片（SoC）价格上涨，其成本占比压力显著增加，早期拆解报告显示处理、内存、存储等关键芯片可占到手机物料成本的30%—40%甚至更高。

对于2025年发布的旗舰级Android机型来说，仅仅是手机处理器的价格就可能高达1700-2000元人民币，相比同系列上代芯片的成本就已经上涨了27%，且行业普遍预估下一代进入2nm时代的芯片成本还将进一步上升。最终这一压力必将转嫁给OEM厂商与消费者。

据报道，台积电宣布对2nm、3nm和5nm先进制程晶圆进行全面提价，涨幅最高达8%—10%，新价格将从明年开始执行。2nm制程晶圆的价格预计将比3nm高出至少50%，台积电强调，由于2nm技术的研发和生产涉及巨额的EUV光刻设备投入、良率优化和生产线改造成本，因此不会对价格进行优惠。业界分析认为，台积电的涨价是为了应对先进制程设备投入和制造成本的持续增长，并进一步加强其在2/3/5nm制程领域的定价能力。预计到2026年，旗舰手机市场将普遍面临价格上涨和利润压力。

挑战2：电子产品需求放缓

尽管生成式人工智能、5G、物联网和边缘计算正成为推动先进半导体需求增长的核心动力。但另一边，消费电子产品的需求却在持续放缓。

GSMA全球消费者循环经济调研及其他市场研究显示，消费者的设备使用周期正在显著延长，全球平均换机周期已延长至约3.5年，在中国则延长到约3.7年。另一方面，是手机维修与二手市场的兴起，根据GSMA《移动净零排放》报告，共有73%的中国受访者表示曾维修过旧手机，远高于全球平均水平60%。还有37%的受访者会在下次购机时考虑购买翻新手机。近年来，中国翻新手机的销量呈现持续增长趋势，2024年增幅达6%。目前，在中国的智能手机行业，二手和翻新手机占比已达20%。香港是全球翻新手机市场的重要中转枢纽。

在PC市场方面，根据Omdia的预测，预计到2025年底，中国PC市场将同比增长5%，达到4150万台。增长动力来自上半年稳健的消费需求和强劲的商用采购，特别是信创领域的推动。预计这一趋势将延续至2026年，但由于消费需求进一步走弱，市场预计将小幅下降2%。平板电脑市场预计在2025年底增长12%至3500万台，这一增长主要受国内厂商激进的产品发布与定价策略带动；但在2026年，随着市场调整，出货量预计将回落9%至3200万台。

趋势 | 算力攀升与应用多样化

趋势1：AI算力的攀升

为了能够更好地运行AI，算力自然需要进一步提升。

第五代骁龙8至尊版中的Hexagon NPU被称为性能杰作的关键要素，全面加速AI特性，让应用变得更智能、更快速、更流畅，从而以最优性能带来令人愉悦的交互体验。新一代NPU实现了显著的架构升级：配备更多标量与向量加速器，速度更快的张量加速器，并采用全新的64位内存架构，NPU性能提升高达37%。这意味着用户可以获得超快响应的个性化智能体AI查询，或通过最新图像生成模型创作出高质量、高分辨率的图像。同时，AI 赋能的应用体验也将全面升级，例如实时通话翻译、流畅照片编辑等功能都将更加出色。除此之外，高通Hexagon NPU的每瓦特性能提高16%，实现了能效的全面提升。作为高通AI引擎不可或缺的一部分，高通传感器中枢如今解锁了前所未有的个性化体验，通过诸如个人知识图谱和个人记录（Personal Scribe）等新功能，那些广受用户欢迎的应用可以基于设备中的个人情境信息主动采取行动，同时确保所有数据隐私完全受控，始终由用户掌握。

面向PC设备，骁龙X2内建的Hexagon NPU具有高达80 TOPS的AI运算效能，不但高于现在已推出的AMD Ryzen AI9 HX HX375的55 TOPS，以及Intel Core Ultra 9 288V的48 TOPS，也高于Apple M4 Max的38 TOPS。高通为了改进Hexagon NPU的效能表现，分析超过300种不同AI模型运算过程中所占用的资源，并改进架构设计以达到更平均的纯量、向量、矩阵运算以及记忆体资源分配，达到提升整体效能表现的成果。为了加强在处理器执行AI应用程序的能力，从内部整合Qualcomm Matrix Engine矩阵引擎，它能够提升矩阵与AI、ML等运算负载的效能，在进行AI运算时不需要将数据传输到GPU或神经处理器NPU，具有更快的反应速度并节省搬运资料所消耗的电力，适合轻量AI应用程序。Qualcomm Matrix Engine矩阵引擎可以加速矩阵以及AI、ML运算，支援512 Bit长度向量以及FP32、FP16、BF16、INT32、INT16、INT8等资料类型。其运作时脉能与处理器脱钩，在闲置时降速或关闭，以节省电力消耗，并降低发热以避免影响处理器。

英特尔明年1月正式发布的Panther Lake整体AI算力高达180 TOPS。CPU、NPU和GPU构成的XPU，为领先的端侧AI体验奠定了坚实基础。英特尔XPU为AI提供了强大的综合算力。而更重要的是，在软件层面，我们在稀疏注意力、推测解码、KV Cache压缩等一系列关键技术上的创新，充分释放硬件潜能。

面向数据中心领域，作为目前应用最为广泛且实用性最高的云端AI，算力的支撑更是至关重要。英伟达今年GTC大会期间发布Blakwell Ultra，它由两颗台积电N4P工艺Blackwell GPU+Grace CPU+更大容量的HBM封装而来，即搭配了更先进的12层堆叠的HBM3e，显存容量提升至288GB，和上一代一样支持第五代NVLink，可实现1.8TB/s的片间互联带宽。基于存储的升级，Blackwell GPU的FP4精度算力可以达到15PetaFLOPS，基于Attention Acceleration机制的推理速度，比Hopper架构芯片提升2.5倍。

亚马逊在年度 re:Invent 大会上正式发布了新一代自研人工智能芯片Trainium3，直接向当前占据市场主导地位的英伟达（NVDA）发起挑战。亚马逊AWS首席执行官Matt Garman在发布会上表示，Trainium芯片业务已是价值数十亿美元的生意，并持续快速增长。Trainium 3是首款3nm AWS AI芯片，提供2.52 PFLOPs FP8算力，内存容量较前代增加1.5倍、带宽提升1.7倍，搭载它的Trn3 UltraServer系统能效较前代提高40%。开发中的Trainium 4将支持英伟达NVLink Fusion互联技术。

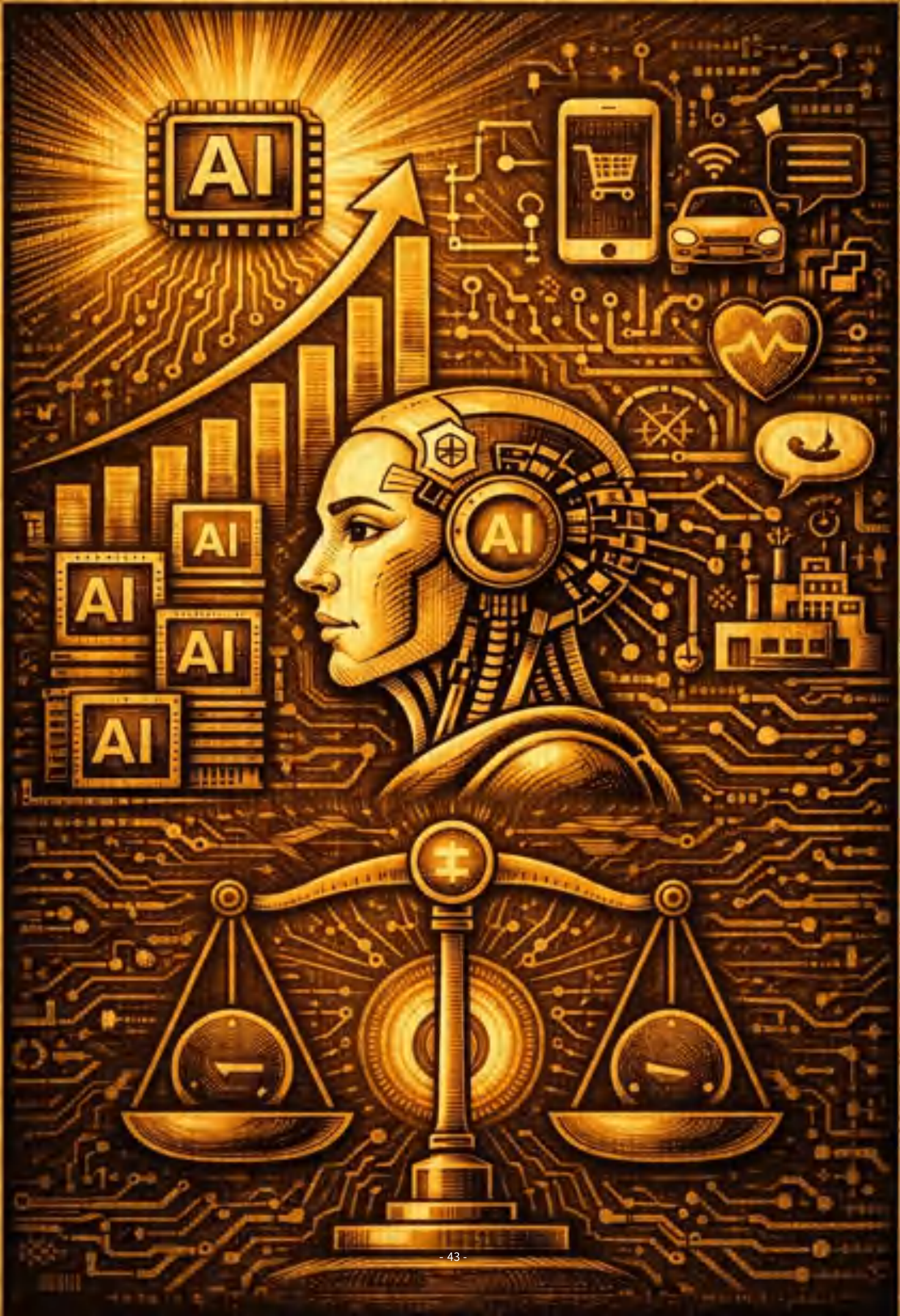
挑战2：电子产品需求放缓

随着传统智能手机市场趋于饱和，芯片行业的需求重心正快速向更多元的场景扩展。汽车电子方面，智能驾驶、车载娱乐、域控制器等需求强劲增长，使汽车成为仅次于手机之后的“第二大算力终端”，高算力SoC、传感器芯片和功率器件持续放量。物联网领域，随着 AIoT 的兴起，大量轻量级终端需要具备本地推理能力，从智能家居、工业传感到可穿戴设备，新应用不断涌现，对低功耗 MCU、连接芯片和端侧 AI 加速器提出更高要求。同时，平板、笔记本、XR头显等智能终端加速搭载 AI 本地推理能力，推动 NPU 成为芯片的新标配。这三大场景共同构成了驱动芯片产业增长的“新三角”，不断拉开需求天花板，也让行业竞争重心从单一性能比拼转向生态、能效与端侧智能的综合能力竞争。

过去五年间，汽车行业发生的一项重大变革，是从传统MCU为基础、多个ECU协同实现汽车功能的架构，迈向中央计算架构。电动平台赋予了重新思考“中央计算”的机会，通过全新的软件定义汽车架构，统一管理机电系统、功耗、各类机械或数字子系统的方方面面。汽车厂商在这一过程中，也在自身产品的设计上表现出了对“中央计算”架构的需求。例如零跑汽车通过将两个SoC集成到一个域控中，来降低成本和线束复杂性，从而提高车辆效率。另外，即便是一些仍在继续采用传统架构的传统车型，也开始配置集成多计算模块的集中式分区控制器。

高通也在过去两年左右的时间里，构建起一套高度复杂，且同时具备高度可扩展性和高度模块化的架构，能够将芯片定位为打造数字座舱的专用芯片或ADAS专用芯片，以及同时支持两者的Flex芯片。Snapdragon Ride Flex SoC作为能够满足中央计算新趋势的SoC，通过打造单芯片解决方案，免除双芯片之间的通信，从而显著降低从座舱到ADAS，以及从ADAS到座舱的传感器通信延迟，从而实现更加真实、更加出色的体验。

高通于今年2月正式发布全新品牌“高通跃龙”，覆盖工业物联网、蜂窝基础设施和工业连接解决方案。依托高通跃龙产品组合，企业能够做出更明智的决策、提高运营效率并加快产品上市，增强竞争优势并有助于在不断变化的市场中取胜。通过“高通跃龙”这一面向To B市场的全新品牌与消费者广为熟知的骁龙品牌形成互补，高通进一步完善了从终端到行业应用的布局，展现出推动产业智能化升级的战略愿景与长期承诺。具身智能则通过将AI融入机器人等物理实体，使其具备类似人类的感知、规划、决策与执行能力。基于高通跃龙QCS8550芯片平台，钛虎机器人打造了人形机器人钛虎T170“瑶光”，可实现双足行走，并能够跨越复杂地形，自重较业内其他产品轻20%～40%，动态响应速度提升了30%。在工业场景中，AI正成为重构制造体系的底层逻辑，为每一件工业品从原料分选到成品交付的全生命周期提供智能把关。移远通信旗下


AI
AI
AI
AI
AI
AI
AI

品牌宝维塔借助高通跃龙 QCS6490芯片平台的强大算力与高集成度架构推出AI分选解决方案，助力色选制造商以低成本的方式轻松打造更具国际竞争力的色选机产品。

挑战3：功耗平衡是永久课题

无论是数据中心，还是端侧计算，功耗都成了一个绕不开的话题，能效比一词，相比单一性能，更加受到行业的关注。据预测，人工智能相关能耗还将增加一倍多，从2024年的260太瓦时增至2027年的500太瓦时。以OpenAI的ChatGPT为例，单词查询耗电2.9瓦时，约为谷歌搜索耗电量的10倍。

不过在未来，AI在节能方面也会发挥更多作用。例如在通信领域，6G将面向AI时代而设计，它能够将层出不穷的新应用、新的工业用例，以及新的边缘侧AI应用场景融入通信系统中。虽然云端在数据存储和数据训练方面具有优势。

在6G网络的初期硬件采购方面，拥有成本高效的资本支出（CAPEX）非常重要，但更重要的是聚焦运营支出（OPEX），从而确保系统在运行时能够实现高效节能。AI恰好能够在这方面发挥重要作用，让网络更具预测性和响应性，使系统能够以尽可能小的功耗运行，从而实现端到端系统运行效率的全面提升。例如，通过将AI设计融入通信协议，并利用AI优化终端与网络间的信道状态信息（CSI）交换过程，可以从基于统计模型的方法转向基于数据驱动的方法，从而实现信息交换效率的优化。

相比云端，端侧人工智能在本地处理数据，无需进行高能耗的数据传输。专为端侧设备设计的人工智能芯片优先考虑能效，而不是纯粹的计算能力，因此与云端人工智能相比，每项人工智能任务的能耗可降低99%到99.9%。同时，端侧芯片因为需要更加考虑续航、发热等问题，对于功耗与性能释放间的平衡也更加关注。

结语

“面向 2026 年，芯片产业将进入“多引擎增长”阶段：汽车电子、电动化与智能驾驶全面推高车规级 SoC 与功率器件需求；IoT 设备在轻量化 AI、本地推理和低功耗连接芯片带动下迎来新一轮普及；智能终端则以端侧大模型、本地 AI 办公和混合算力体验推动换机潮。三大场景共同形成更稳定、分散且互补的下游结构，使芯片产业从单一品类驱动迈向多元需求牵引，竞争核心从单纯算力比拼转向能效、系统协同与生态整合。”