

行业评级：/
证券研究报告|行业专题
海外科技
2024年6月21日



AI进化加速端侧落地，新一轮换机潮蓄势待发

——AI终端行业专题

证券分析师
姓名：郑嘉伟
资格编号：S1350523120001
邮箱：zhengjiawei@huayuanstock.com



- **AIPC&手机是大模型落地的重要载体，颠覆传统应用模式，AI进化加速端侧落地，新一轮换机潮蓄势待发。** AIPC是算力平台+个人大模型+AI应用的新型混合体，通过整合 AI 技术，提供个性化、智能化的服务，提升了传统 PC 的功能和用户体验，成为未来计算平台的重要发展方向。生成式AI手机是利用大规模、预训练的生成式AI模型，实现多模态内容生成、情境感知，并具备不断增强的类人能力。近两年消费电子市场需求持续受压。根据IDC数据，2022与2023年连续两年全球智能手机、PC、平板电脑出货量同比下滑，随着AI的端侧落地，AI手机、AIPC的推出有望打破消费电子近年来创新不足无法打动消费者换机的局面，有望推动消费电子换机需求的到来。根据Counterpoint Research预计，2024年全球生成式AI智能手机出货量将超过1亿部，2027年将达5.22亿部，在整个智能手机出货量当中的占比达40%，年复合增长率高达83%。IDC 预计 2024 年将是PC的扩张年，全球出货量将达到 2.654 亿台，比上一年增长 2.0%，因为AIPC的推出，最终将推动市场在2028年达到2.922亿台，在 2024-2028 年预测期内的复合年增长率 (CAGR) 为 2.4%。
- **硬件端：处理器架构变化，内存价值量提升。** 1) ARM架构因其低功耗的特点份额有望逐步提升，高通+微软联手发力推动WOA，Copilot+PC的推出可能是WOA的重要转折点，根据counterpoint research预测，2027年基于ARM架构的笔记本电脑市场份额有望达到25%。2) AIPC处理器异构计算成为共性，增加NPU本质是为了进行低功耗计算，NPU专为实现以低功耗加速AI推理而全新打造，并随着新AI用例、模型和需求的发展不断演进。引入NPU对于提升计算效率、降低能耗与成本具有重要意义。3) 大模型的端侧落地，同样要求内存空间的提升，由此带来内存需求的提升，单机价值量提升，根据集邦咨询预测，DRAM于笔电的单机平均搭载容量年增率约12.4%，后续随着AI PC量产后，2025年成长幅度会更明显。
- **大模型端：轻量化小模型陆续推出，多模态大模型改变人机交互体验。** 端云混合有望成为未来AI计算的最终解决方案，AI落地终端具有必要性。据高通《混合AI是AI的未来》，AI落地终端能够在全球范围带来成本、能耗、性能、隐私、安全和个性化优势。大模型压缩技术使得复杂大模型在端侧运行成为可能，多模态大模型，如GPT-4o和Project Astra，使得人机交互变得更加自然，丰富AI应用场景。
- **AI落地端侧为产业链各环节带来投资机会，首推比亚迪电子，建议关注苹果、高通、微软、联想集团。** 从硬件层面来看，CPU是核心，高通具备先发优势，有望在intel和AMD的产品尚未满足微软AIPC要求的情况下，率先抢占部分市场；从系统软件层面来看，微软copilot渗透率迅速提升，形成正向循环，一方面可以促进其windows OEM业务增长，另一方面同样可以促进其云业务需求；从整机组装环节来看，比亚迪电子受益其海外大客户落地AI，有望推动其手机出货量，此外AI落地端侧，推动消费电子景气度复苏，公司消费电子组装和零部件业务有望受益；从品牌运营的角度来看，建议关注联想集团，公司率先布局本地大模型，AIPC有望带动公司智能设备业务恢复增长；此外建议关注苹果，自身在硬件、软件生态、终端产品具备一体化优势，苹果AI正式落地，多环节有望受益。
- **风险提示：**大模型技术发展不及预期，PC/手机等终端需求复苏不及预期

- 1、AIPC&手机是大模型落地的重要载体，颠覆传统应用模式
- 2、硬件端：处理器架构变化，内存价值量提升
- 3、大模型端：轻量化小模型陆续推出，多模态大模型改变人机交互体验
- 4、投资逻辑：AI端侧落地加速，换机潮蓄势待发，关注PC/手机厂及硬件相关投资机遇
- 5、风险提示

1.1 AI PC：AI与PC的结合，颠覆传统PC定义，成为个人AI助理

■ AI PC：AI 普惠的首选终端，算力平台 + 个人大模型 + AI 应用的新型混合体

- **AIPC是算力平台+个人大模型+AI应用的新型混合体，通过整合 AI 技术，提供个性化、智能化的服务，显著提升了传统 PC 的功能和用户体验，成为未来计算平台的重要发展方向。** AI PC不仅继续作为人们重要的生产力工具和内容消费平台，还在硬件上集成了混合AI算力单元，能够本地运行“个人大模型”。这种新型计算机可以创建个性化的本地知识库，实现自然语言交互，这将彻底改变传统PC的定义。
- **AIPC的各类应用场景为用户提供的价值是节约时间。** AI PC能够针对工作、学习、生活等场景，提供个性化创作服务、私人秘书服务、设备管家服务在内的个性化服务，共性都是为了提高用户的使用效率，节省时间。云端的公共大模型反馈速度慢会影响用户的实际体验，AIPC落地端侧，推理直接在本地运行，边缘和云端推理为辅使得AIPC能够提供即时的服务响应。
- **AIPC将是一个包含 AI 模型和应用以及硬件设备的混合体。** 根据IDC《AIPC白皮书》，AI PC 产品拥有本地部署的大模型与个性化本地知识库组合构成的个人大模型，第一交互入口为个人智能体，可实现自然语言交互，AI PC 将通过内嵌 AI 计算单元的方式提供混合 AI 算力，还可以依靠开放生态来满足不同场景的需求。在满足生产力提升的同时，通过本地数据存储和隐私及数据保护协议来保护个人隐私和数据安全。

图：AI PC在通用场景下的个性化服务

	工作	学习	生活
个性创作	- 会议材料准备 - 会议总结和纪要 - 专业 PPT/Word/Excel...	- AI 课堂笔记和记录 - 文献翻译和总结 ...	- 游戏攻略 - AI 游记 ...
秘书服务	- 个人日程表 - 同声传译 ...	- 个人课程表 - 选课和提醒 ...	- AI 旅行计划 - AI 实时游戏指导 ...
设备管家	- 主动调优 - 专业模式 ...	- 智能防护 - 学习模式 ...	- 智能互联 - 游戏模式 ...

图：AI PC 核心特征

自然语言交互的个人智能体		内嵌个人大模型	
- 多模态自然语言交互 UI - 基于本地大模型的个人智能体		- 本地为主，边缘与云为辅的大模型 - 个性化本地知识库	
标配本地混合 AI 算力	开放的 AI 应用生态	设备级个人数据 & 隐私安全保护	
- CPU&NPU&GPU 本地混合计算架构 - 个人终端和家庭主机 / 企业边缘主机协同计算	- AI 原生应用、AI 赋能应用 - 能够被智能体任务调度、适配混合 AI 算力平台等	- 本地隐私推理 & 非敏感任务调用云端大模型 - 硬件级安全芯片保护 & 个人数据加密 / 脱敏传输	

1.2 AI手机：重构手机生态，成为用户定义的个性化专属AI助理

■ 手机智能化的演进：每一轮技术创新驱动智能手机功能和用户体验的重大升级

- 1) **初始阶段（2000年代初）**：诺基亚塞班操作系统首次允许用户自行下载应用，开启了智能手机发展的先河，并促成了移动互联网生态的形成，智能手机逐渐成为人们生活中不可或缺的工具。
- 2) **iPhone 时代（2007年开始）**：iPhone 的问世颠覆了传统的手机设计理念，淘汰了物理键盘，使触控屏幕成为主流，同时苹果应用商店的建立推动了智能手机应用生态的繁荣。
- 3) **智能语音助手与多模态交互（2010年代）**：智能语音助手如Siri和Google Assistant的出现，增强了用户的交互体验，多模态交互技术如手势控制和眼球追踪提高了用户的操作效率。
- 4) **AI 技术赋能（2017年至今）**：AI 技术广泛应用于智能手机，如计算摄影、安全强化、续航优化等，安卓阵营在SoC平台中加入独立的AI计算单元，显著提升了影像处理等能力。
- 5) **生成式 AI 手机时代（2023年至未来）**：生成式 AI 技术通过大规模预训练模型，实现了智能手机的多模态内容生成和情境感知，推动从GUI（图形用户界面）到VUI（语音用户界面）的转变，并孕育出智能生命体，为用户提供个性化的内容和服务。
- 总体来看，过去手机智能化演进的过程包括操作系统和平台的创新、交互方式的改进、硬件能力的提升、软件和应用生态的丰富，而AI手机的变革有望将这四方面变化重新演绎。

图：手机智能化演进路线图

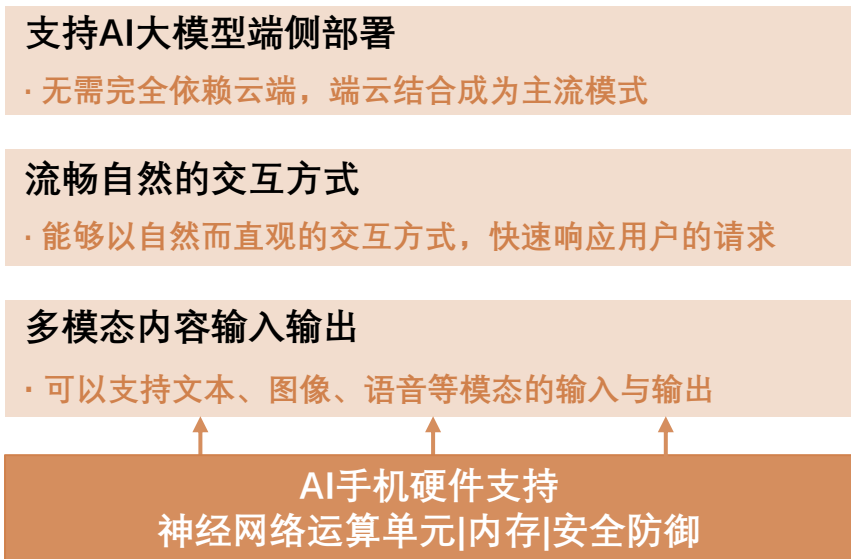


1.2 AI手机：重构手机生态，成为用户定义的个性化专属AI助理

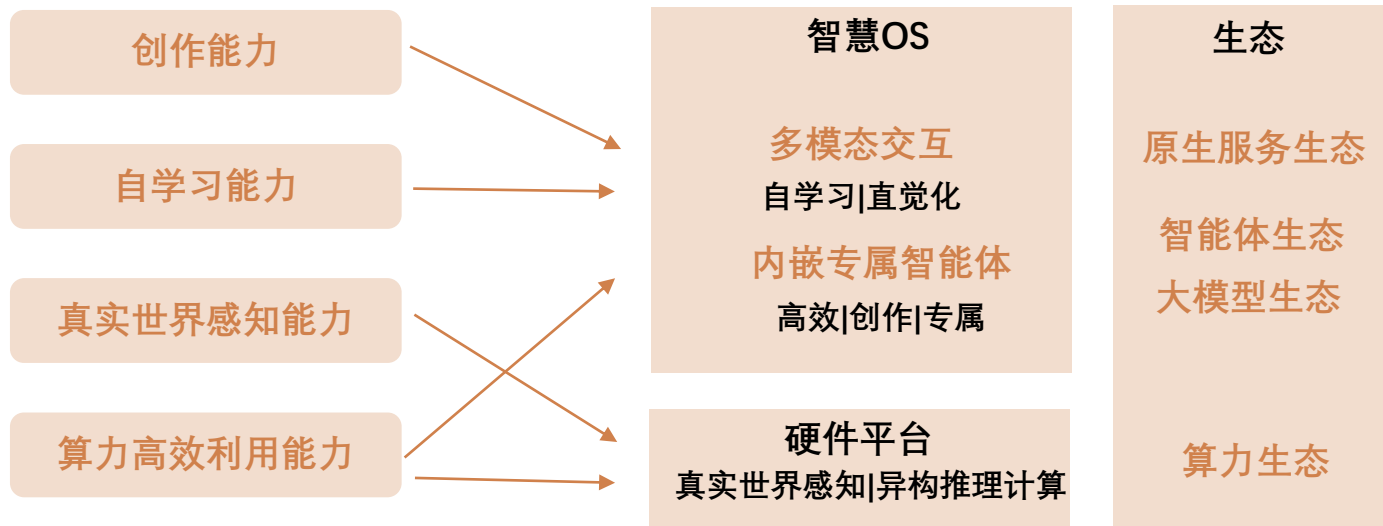
■ AI 手机：自在交互、智能随心、专属陪伴、安全可信的个人助理

- **生成式AI手机**是利用大规模、预训练的生成式AI模型，实现多模态内容生成、情境感知，并具备不断增强的类人能力。Counterpoint 基于历史每一轮手机的变革，提出了生成式AI手机的概念：生成式AI手机是利用大规模、预训练的生成式AI模型，实现多模态内容生成、情境感知，并具备不断增强的类人能力。此外生成式AI手机还将具备支持大模型的本地部署，具备多模态能力、确保流畅、无缝的用户体验的功能、以及拥有实现上述功能的硬件支持等必要特征。
- **多模态能力是实现AI手机的关键**。AI手机在原生服务生态的基础上，构建智能体生态、大模型生态和算力生态，使AI手机具备了创作能力、自学习能力、真实世界感知能力和算力高效利用能力。AI手机的智慧OS能够与用户进行多模态交互并且拥有内嵌专属智能体，使手机更加高效和富有创造力。多模态大模型还可以为智能体的成长提供更多维度和更加丰富的训练语料，使智能体自学习、直觉化，满足用户的个性化需求。

图：AI 手机核心特征



图：AI 手机全栈革新及生态重构

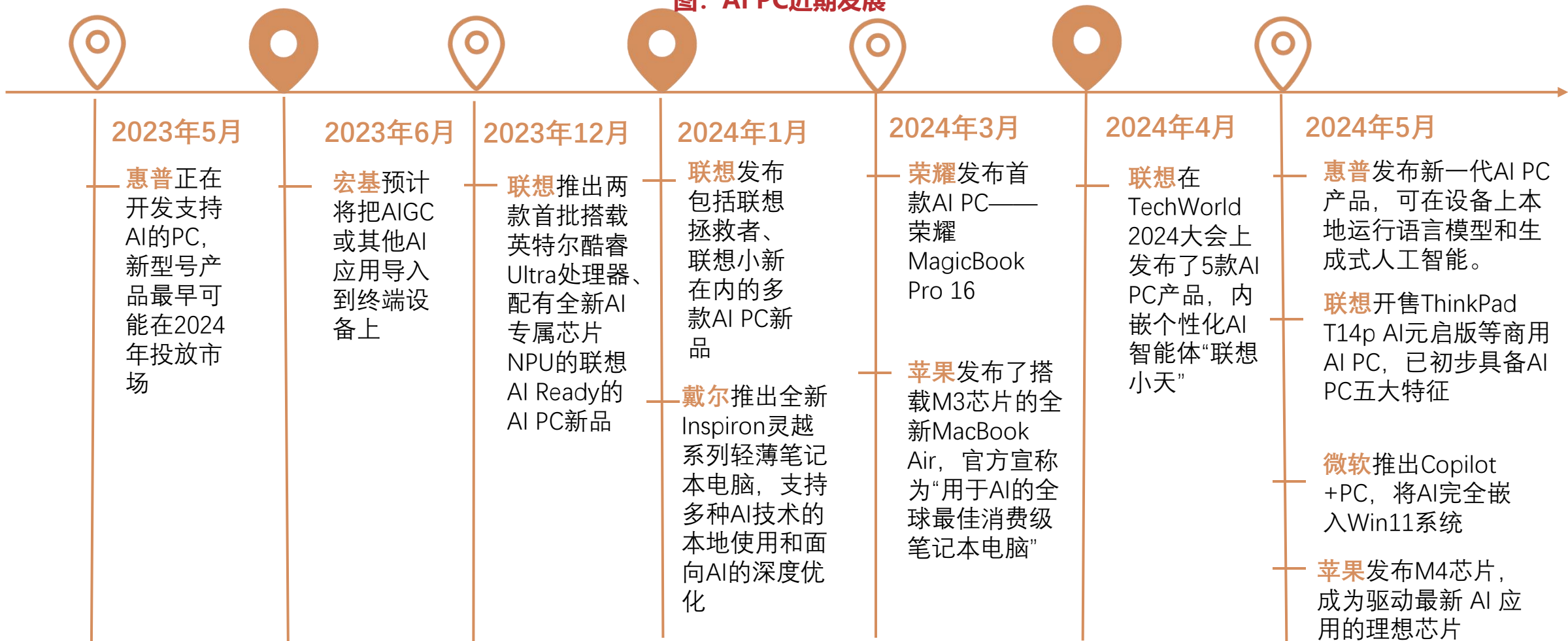


1.3 AI落地终端，各大厂商纷纷布局

■ 主流电脑品牌对AI PC的相关布局

- 2023年5月，主流电脑品牌已经开始探索AI+PC的模式，逐步推进AI与PC的融合。2024年多品牌加速布局AI PC，2024年1月，联想和戴尔发布了其新款AI PC产品；2024年3月，荣耀发布其首款AI PC产品；2024年4月，联想发布5款AI PC产品；2024年5月，微软推出Copilot+PC，多个品牌推出AI PC产品。

图：AI PC近期发展



1.3 AI落地终端，各大厂商纷纷布局

■ 主流手机厂商纷纷布局AI手机，各类产品层出不穷

- AI手机的产生与发展离不开AI大模型和手机硬件等的支持，自2023年7月至2024年5月，荣耀、小米、OPPO等主流手机制造商将目标投向AI手机领域，加大AI大模型、数据存储、芯片等的研发，推出诸多适用于AI手机的新技术，并逐渐实现AI与手机的融合。

图：AI手机近期发展



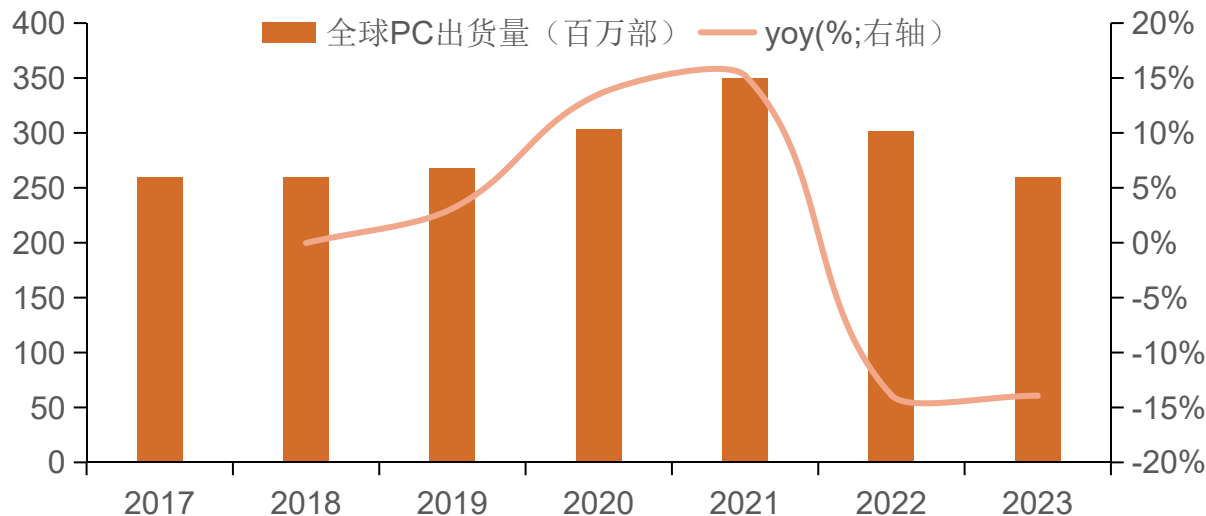


1.4 PC/手机有望受益AI推动消费电子行业复苏

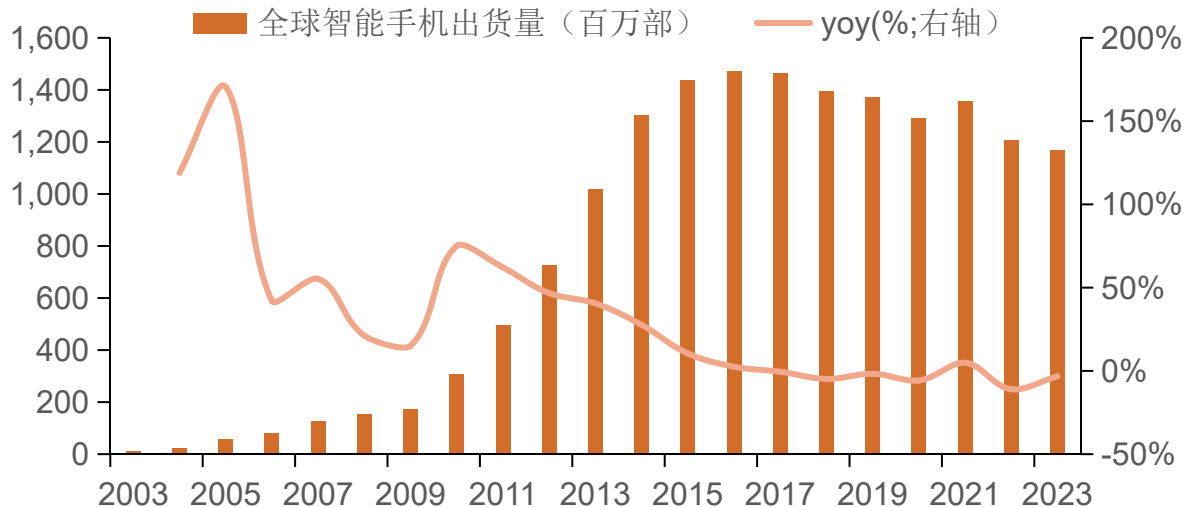
AI进化加速端侧落地，新一轮换机潮蓄势待发

- 近两年消费电子市场需求持续受压。根据IDC数据，2022与2023年连续两年全球智能手机、PC、平板电脑出货量同比下滑，其中2023年全球智能手机出货量为11.7亿部，同比下降3.2%，2016年全球智能手机出货量达到顶峰，2021年出货量小幅反弹后继续下降；2023年全球PC出货量为2.6亿台，同比下降13.9%。
- 分季度来看消费电子近期呈底部复苏的态势。根据IDC数据，3Q23全球智能手机出货开始同比增长，3Q23、4Q23、1Q24分别实现同比增长0.3%、8.6%、7.7%；1Q24全球PC出货开始同比增长，同比增长5.1%。
- 随着AI的端侧落地，AI手机、AIPC的推出有望打破消费电子近年来创新不足无法打动消费者换机的局面，有望推动消费电子换机需求的到来。根据TechInsights预计，生成式AI将是2024年智能手机市场的焦点，2024年智能手机出货量将恢复低个位数的增长，主要由新兴市场需求复苏所带动，Counterpoint Research预计，2024年全球生成式AI智能手机出货量将超过1亿部，2027年将达5.22亿部，在整个智能手机出货量当中的占比达40%，年复合增长率高达83%。IDC预计2024年将是PC的扩张年，全球出货量将达到2.654亿台，比上一年增长2.0%，因为AIPC的推出，最终将推动市场在2028年达到2.922亿台，在2024-2028年预测期内的复合年增长率(CAGR)为2.4%。

图：全球PC出货量（年度）



图：全球智能手机出货量（年度）





■ AI PC & Phone 产业链分布：上游软硬件供应商+中游代工及品牌厂商+下游应用场景

- 产业链的上游包括硬件和软件供应商。硬件供应商提供AI PC所需的各种组件，如半导体（CPU、GPU、NPU芯片）、存储设备（SSD、HDD）、内存（RAM）、主板、电源等。软件供应商则提供操作系统、AI框架、应用程序等。
- 中游主要由代工厂商和品牌厂商组成。代工厂商负责将上游提供的硬件组件组装成完整的AI PC，并进行质量控制和测试。品牌厂商则负责AI PC的设计、品牌建设、营销、销售、客户服务等。
- 产业链的下游是各类应用场景，包括教育、办公、娱乐、医疗、科研、金融、制造业、零售、政府和家庭等。

图：PC产业链全景图



- 1、AIPC&手机是大模型落地的重要载体，颠覆传统应用模式
- 2、硬件端：处理器架构变化，内存价值量提升
- 3、大模型端：轻量化小模型陆续推出，多模态大模型改变人机交互体验
- 4、投资逻辑：AI端侧落地加速，换机潮蓄势待发，关注PC/手机厂及硬件相关投资机遇
- 5、风险提示

2.1 变革一：ARM架构因其低功耗的特点份额有望逐步提升

CPU: ARM架构 VS X86架构

- ARM (Advanced RISC Machine) 架构和x86架构是两种主要的处理器架构，它们在设计理念、应用场景和性能特点等方面有显著差异。
- 从设计理念来看，ARM架构追求指令集的简化，每条指令执行速度快，能够在较低功耗下高效工作。ARM架构强调整能和高效，适用于电池供电设备。x86架构包含复杂的指令集，能够在一条指令中执行更多的操作。x86架构强调性能和灵活性，适用于需要高计算能力的设备。
- 从应用场景来看，ARM架构以其高能效和低功耗特性在移动和嵌入式市场占据重要地位，而x86架构以其高性能和丰富的软件生态系统在桌面和服务市场保持主导地位。
- ARM架构的低功耗特性是多种设计因素共同作用的结果，包括简洁的指令集、简化的硬件实现、优化的流水线设计、低功耗模式、较少的晶体管数量、专用处理单元的使用以及针对嵌入式系统和移动设备的优化。这些设计特点使得ARM处理器在需要低功耗的应用场景中表现尤为出色。
- 在AIPC场景下，ARM架构具有更高的能效比，在AIPC中能够提供高效的计算能力，同时保持较低的能耗。ARM架构允许高度定制化，芯片设计者可以根据具体的AI计算需求优化硬件架构。这种灵活性使得ARM架构能够在不同的AI应用场景中提供优化的解决方案。

表：ARM架构和x86架构对比

属性	ARM架构	x86架构
设计理念	RISC（精简指令集计算），指令集小而简单，每条指令执行时间短，提高了能效和性能。设计目标之一是低功耗，适用于需要延长电池寿命的设备。	CISC（复杂指令集计算），指令集复杂且丰富，每条指令可能执行多个操作，设计更注重功能的多样性。强大的计算能力使其适合高强度计算任务。
应用场景	主要应用于移动设备（如智能手机、平板电脑、可穿戴设备）和嵌入式系统（如物联网设备、工业控制系统、消费电子产品），因其低功耗和高能效而受欢迎。	主要应用于个人电脑（如台式机、笔记本电脑）和服务（如数据中心、高性能计算），因其强大的计算能力和广泛的兼容性而成为主流选择。
性能特点	能效比高，通常在较低的功耗下提供良好的性能。架构具有很好的扩展性，可以根据不同应用需求进行定制，适应多样化的市场需求。	计算能力强，在单线程和多线程性能上都表现出色，适合运行复杂应用程序和多任务处理。兼容性好，支持大量已有的软件和操作系统，尤其是在PC和服务市场。
市场现状与未来趋势	随着移动设备和物联网市场的快速发展，ARM处理器的市场份额不断增加。近年来，ARM也开始进军高性能计算市场，推出了适用于服务器和数据中心的高性能处理器。	在PC和服务市场中占据主导地位，特别是英特尔和AMD在高性能领域具有强大优势。但随着ARM处理器性能的提升和市场扩展，x86架构面临来自ARM和其他架构的竞争压力。
代表厂商	主要代表厂商包括ARM Holdings、Qualcomm、苹果（Apple）、三星（Samsung）、联发科（MediaTek）等。	主要代表厂商包括英特尔（Intel）和AMD。

资料来源：redhat、亿欧智库、华源证券研究

表：不同品牌跑分及功耗对比

架构	品牌	产品	发布时间	Cinebench得分		TDP (PL1/PBP)
				单核得分	多核得分	
X86	Intel	Intel® Core™ Ultra 5 processor 125H	Q4'23	1671	12707	20~65W
		Intel® Core™ Ultra 5 processor 135H	Q4'23	1763	13907	20~65W
		Intel® Core™ Ultra 7 processor 155H	Q4'23	1762	14845	20~65W
		Intel® Core™ Ultra 7 processor 165H	Q4'23	1676	15827	20~65W
		Intel® Core™ Ultra 9 processor 185H	Q4'23	1809	18545	35~65W
		Intel® Core™ i9 processor 14900KF	Q4'23	2308	40814	125W
ARM	AMD	AMD Ryzen™ 7 8840U	Q2/2024	1769	12994	15~30W
		AMD Ryzen™ 7 7840U	Q2/2023	1761	12942	15~30W
	高通	X1E-78-100	Q2/2024	1539	12182	23W
		X1P-64-100	Q3/2024	1718	12881	23W
	Apple	Apple M3 Pro	Q4/2023	1981	15209	30W
		Apple M3 Max	Q4/2023	1953	24029	40W

资料来源：Cinebench，华源证券研究

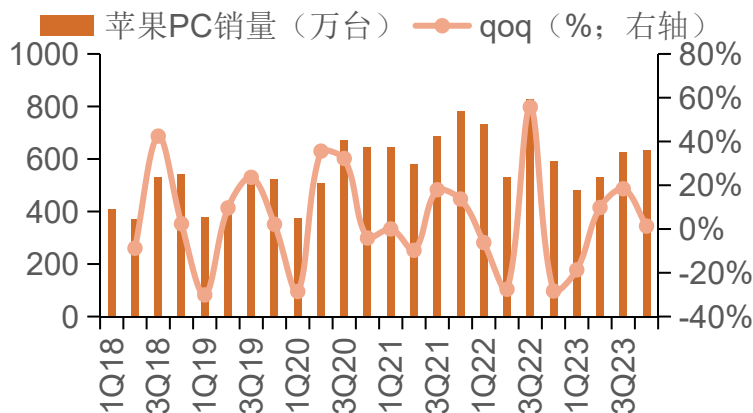


2.1 变革一：ARM架构因其低功耗的特点份额有望逐步提升

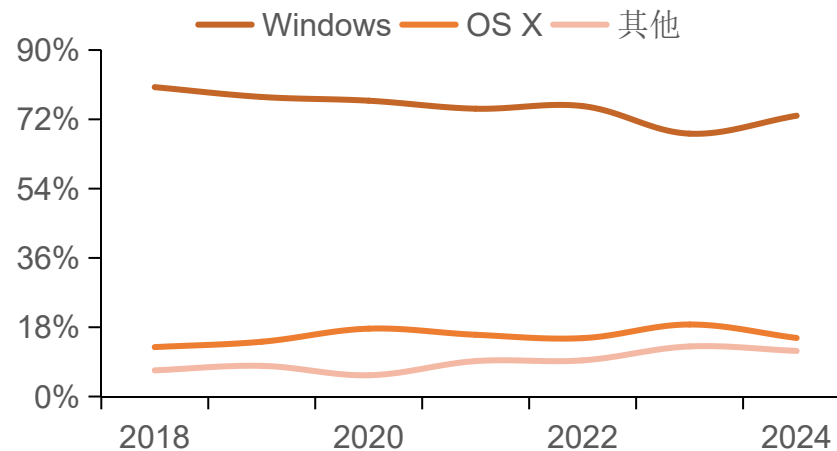
■ 苹果由x86架构转向ARM架构收获成功，是其市场份额上升的关键因素之一

- **2020年苹果由x86架构的CPU转向ARM架构CPU，推出M系列芯片。**苹果在2020年11月10日正式推出了其首款自研芯片M1，它搭载于MacBook Air (2020 年末)、Mac Mini (2020 年末)、MacBook Pro (13 英寸, 2020 年)、iMac、iPad Pro及iPad Air (第五代)上，标志着苹果从使用英特尔处理器转向使用自主研发的芯片，这是一个重要的里程碑。苹果宣称该芯片在所有低功耗中央处理器产品中性能最佳，同时具有最佳的性能功耗比。
- **苹果MacOS市场份额提升，有部分原因可能是M系列芯片高能低耗对于其mac产品销量的拉动。**根据statcounter数据，2018年MacOS在PC操作系统的占比仅为12.84%，2023年提升至18.71%，而windows操作系统市场份额则从2018年的80.36%下降至2023年的68.28%，根据苹果公司公告数据，2020年四季度推出ARM架构的M芯片后，其mac产品线销售收入在1Q21环比上涨5%，同比上涨70%，增速显著高于全球PC销量的增速。

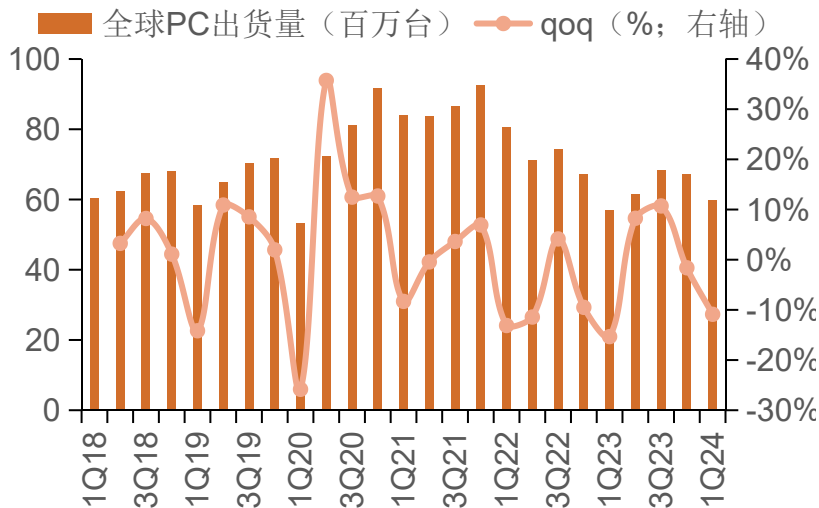
图：苹果PC销量（万台）



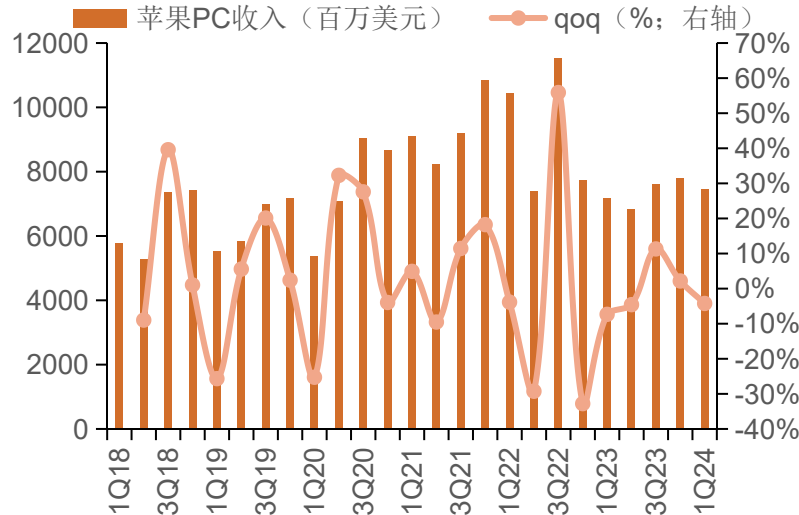
图：微软、苹果在PC操作系统的占比变化情况



图：全球PC出货情况



图：苹果PC销售收入情况



2.1 变革一：ARM架构因其低功耗的特点份额有望逐步提升

■ Rosetta 2确保苹果从x86架构转向ARM架构的过渡期平稳度过

- 苹果由intel芯片转而使用ARM自研芯片，主要原因是苹果希望根据自己的硬件想法来规划芯片的性能分配，达到高性能低功耗的效果。
- 苹果在从x86架构向ARM架构转换过程中面临的最大挑战是确保应用程序的兼容性和性能，同时需要平衡硬件设计、操作系统优化、用户体验、生态系统整合和市场竞争等多方面的因素。
- **Rosetta 2使得macx86架构转向ARM架构平稳过渡。**由于当时大量现有的Mac应用程序是为x86架构编写的，为了确保现有的x86应用能够在ARM架构的Mac上运行，苹果推出了Rosetta 2翻译层。Rosetta 2在运行时将x86指令翻译为ARM指令。这种翻译是动态进行的，意味着应用程序在启动和运行时，Rosetta 2会实时进行指令翻译，从而确保兼容性，对于用户而言，Rosetta 2的运行是透明的。用户不需要进行任何额外的操作即可运行他们现有的x86应用程序，只需像往常一样打开应用程序即可。使用Rosetta 2翻译的应用程序与原生运行在x86设备上的体验几乎没有区别。这种无缝体验是苹果致力于确保过渡期间用户不会感受到明显差异的重要组成部分。
- **ARM架构契合苹果生态系统的一致性和连贯性。**苹果的产品生态系统需要保持一致性和连贯性，新的ARM架构设备能够无缝地与现有的iPhone、iPad和其他Mac设备进行协同工作。此外开发者能够轻松地在iOS和macOS平台之间移植应用程序，使整个苹果生态系统更加紧密地集成和协同。

表：苹果从x86架构向ARM架构的关键措施

措施	具体情况
研发自有芯片	苹果开发了自己的基于ARM架构的芯片系列，称为Apple Silicon。首个用于Mac的Apple Silicon芯片是M1芯片，这些芯片专为苹果的硬件和软件进行优化，旨在提供更高的性能和更低的功耗。
Rosetta 2翻译层	为了确保现有的x86应用能够在ARM架构的Mac上运行，苹果推出了Rosetta 2翻译层。Rosetta 2是一种动态二进制翻译器，能够在后台将x86指令转换为ARM指令，从而使用户几乎感觉不到性能损失。
开发者工具和支持	苹果提供了丰富的开发者工具和文档，以帮助开发者将他们的应用程序从x86架构迁移到ARM架构。Xcode被更新以支持跨平台编译，开发者可以轻松地为M1芯片重新编译他们的应用。
Universal App格式	苹果引入了Universal App格式，允许开发者在同一个应用包中包含针对x86和ARM架构的二进制文件。这使得用户可以无缝地在任何Mac上运行相应的应用程序。
逐步转换和兼容性	苹果并没有立即淘汰所有x86架构的设备，而是逐步引入基于ARM架构的新设备，并继续支持旧设备。这样平滑的过渡策略确保了用户和开发者有足够的时间适应新的架构。
硬件与软件深度整合	苹果设计自己的芯片，使其能够更好地优化操作系统（macOS）和硬件之间的协同工作，带来更高的性能和更长的电池续航。

资料来源：亿欧智库、bphotovideo、华源证券研究

图：苹果转向ARM架构的路径及原因

第三方软件的适配工具

为了帮助开发者更好地“迁移”到新平台，苹果推出**Universal、Rosetta、虚拟化**技术等一整套解决方案将现有的Mac应用高效快捷地无缝移植到Arm架构能兼容的系统中;这套解决方案同时支持x86和Arm。

实现统一的Apple生态

由于其他苹果设备(iPhone、iPad、Apple TV、AppleWatch、Airpods)搭载的都是苹果自研的**Arm 芯片**，若Mac电脑也采用**Arm 架构**，则苹果能实现统一的**Apple 生态**，即“不同设备都有同样的架构，运行同样的程序，差别在于外形尺寸与性能”。



工具介绍

Universal 2: 开发者花几天时间重新编译App，就能分别在苹果和intel的两种处理器上运行
Rosetta 2: 对于没有被重新编译的软件，可在应用安装时将英特尔的指令自动翻译成苹果芯片可以听懂的形式
Virtualization: 提供Mac电脑运行 Linux、Docker 等开发环境

资料来源：亿欧智库、华源证券研究

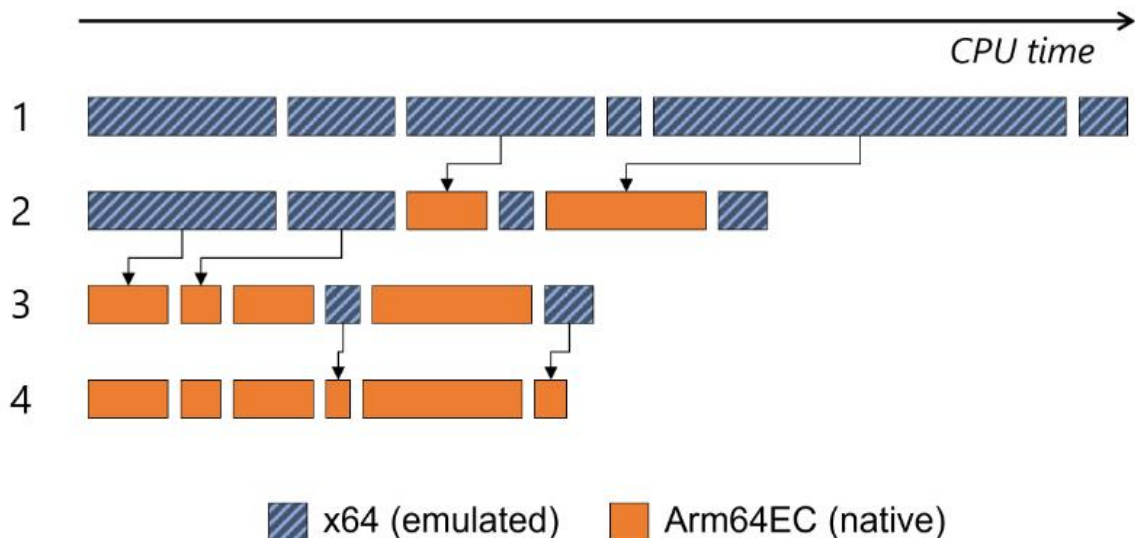


2.1 变革一：ARM架构因其低功耗的特点份额有望逐步提升

■ 微软逐步完善ARM架构的生态系统

- 2021年微软推出了ARM64EC，可以将基于X86的应用程序代码转移到基于ARM的代码上，ARM64EC可以看作是连接 x86 和 ARM 架构的重要桥梁，它通过提供兼容性和优化性能，促进了两种架构之间的互操作性和协同发展。ARM64EC是一种应用二进制接口（ABI），旨在促进在ARM系统上运行x64应用程序的兼容性。这使得x64应用程序在WOA（Windows on ARM）设备上能够更高效地运行，实现原生执行ARM代码。为了使 ARM 设备能够运行 x86 应用程序，微软引入了 ARM64EC。ARM64EC 允许开发者将现有的 x64（即 x86-64）应用程序部分或全部移植到 ARM 平台上，而无需完全重写代码。通过支持 x86 应用程序，ARM64EC 使得 ARM 设备（如基于 ARM 的 Windows 笔记本电脑和平板电脑）能够运行更多的现有软件，从而扩大了这些设备的应用范围和吸引力。
- 基于ARM架构的一系列软件开始逐步推出。目前Windows on ARM已经有87%的应用是ARM原生应用，仅剩13%需要用到兼容层转译。2023年，高通骁龙X系列发布后，高通与主要应用开发商密切合作，2024年3月谷歌也发布了 Windows on ARM 原生版完整 Chrome 浏览器，可在搭载高通骁龙的 Windows PC 上下载使用。

图：使用 ARM64EC 让现有应用在 ARM 版 Windows 11 上运行得更快



图：基于ARM架构的应用程序正在不断丰富

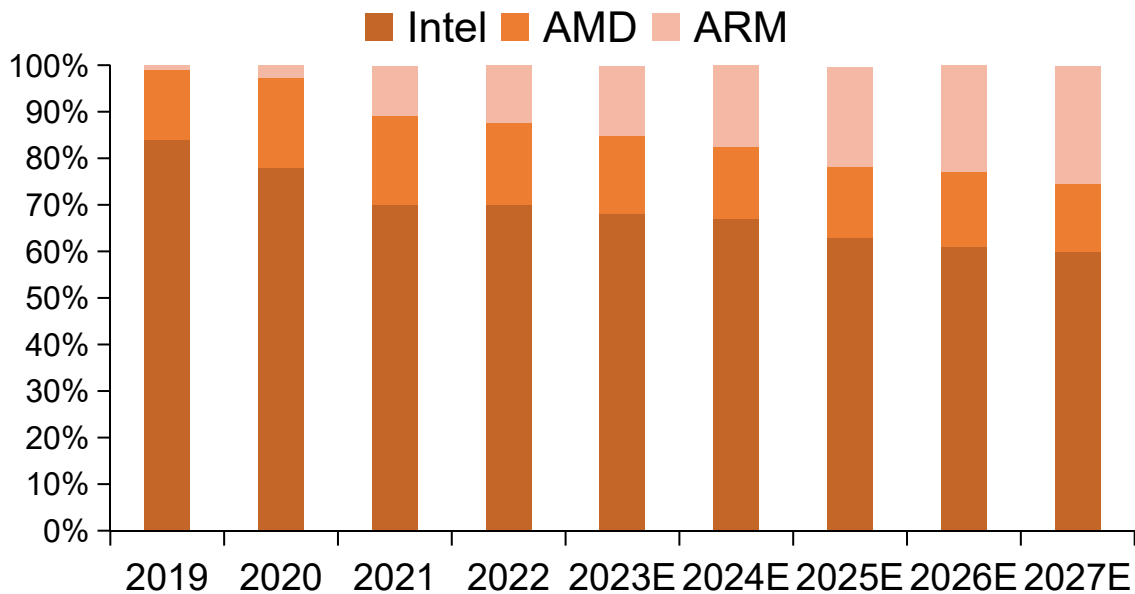


2.1 变革一：ARM架构因其低功耗的特点份额有望逐步提升

ARM架构市场份额有望逐年上涨，高通+微软联手发力WOA

- 根据counterpoint research预测，2027年基于ARM架构的笔记本电脑市场份额有望达到25%。
- 根据前文分析，当前基于ARM的笔记本电脑基本来自来自苹果，操作系统为MacOS，而微软的windows操作系统仍然是最受欢迎的操作系统。而制约过往ARM市场份额未能明显提升的主要原因是windows操作系统应用程序与ARM架构的兼容性问题，这一问题已经在逐步解决。

图：ARM架构市场份额逐步提升



图：高通+微软联手推动WOA



2.1 变革一：ARM架构因其低功耗的特点份额有望逐步提升

- 高通+微软联手发力推动WOA，Copilot+PC的推出可能是WOA的重要转折点。在Build 2024开发者大会前夕，5月20日微软发布了自家最新的“Copilot+PC”，是专为AI设计的新型Windows PC，有望开启Windows on ARM新时代。其他各大品牌同步推出Copilot+PC。微软不仅自己推出全新的Copilot+PC，同时与各大OEM厂商（宏碁、华硕、戴尔、惠普、联想、三星）合作全新Copilot+PC，于6月18日开始上市。

表：各大品牌的新款 Copilot+ 电脑于 6 月 18 日上市

品牌	具体产品
Acer	Acer 的 Swift 14 AI 2.5K 触摸屏使用户能够以更高的精度和色彩准确的图像绘制和编辑用户的视觉。只需轻触专用的 AcerSense 按钮即可启动并发现 AI 增强功能，例如 Acer PurifiedVoice 2.0 和 Purified View。
华硕	华硕 Vivobook S 15 是一款功能强大的设备，通过 Snapdragon X Elite 平台和内置 Qualcomm® AI 为生活带来 AI 体验。它拥有 40+ NPU TOPS、双风扇冷却系统和高达 1 TB 的存储空间。下一代人工智能增强功能包括 Windows Studio 特效 v2 和华硕 AiSense 摄像头，以及自适应调光和锁定的存在检测功能。它专为便携性而设计，采用超薄、轻便的全金属设计、大容量电池以及单区 RGB 背光键盘的高级造型。
戴尔	戴尔推出五款全新 Copilot+ PC，包括 XPS 13、Inspiron 14 Plus、Inspiron 14、Latitude 7455 和 Latitude 5455，提供一系列消费级和商用选项，可实现突破性的电池寿命和独特的 AI 体验。XPS 13 由 Snapdragon X Elite 处理器提供支持，具有优质的未来派设计，而 Latitude 7455 则拥有令人惊叹的 QHD+ 显示屏和具有 AI 降噪功能的四扬声器。Inspiron14 和 Inspiron 14 Plus 采用 Snapdragon X Plus 1，采用轻质低碳铝制成。
惠普	惠普的 OmniBook X AI 电脑和搭载 Snapdragon X Elite 的 HP EliteBook Ultra G1q AI 电脑采用纤薄时尚的设计，提供先进的性能和移动性，带来更加个性化的计算体验。功能包括持久的电池寿命和人工智能驱动的生产力工具，例如实时转录和会议摘要。Poly Studio 水晶般清晰的音频支持具有自动取景和眼部对焦功能的 5MP 摄像头，以增强虚拟交互。
联想	联想推出两款 AI PC：一款是专为消费者打造的 Yoga Slim 7x，另一款是专为商业打造的 ThinkPad T14s Gen 6。Yoga Slim 7x 为创意人士带来效率，配备 14.5 英寸触摸屏，支持 3K Dolby Vision，并针对 3D 渲染和视频编辑进行了优化。T14s Gen 6 为用户的工作任务带来企业级体验和 AI 性能，其功能包括网络摄像头隐私快门、Wi-Fi 7 连接和高达 64GB 的 RAM。
三星	三星新款 Galaxy Book4 Edge 超薄轻便，配备 3K 分辨率 2x AMOLED 显示屏和 Wi-Fi 7 连接。它拥有持久的电池，可提供长达 22 小时的视频播放时间，非常适合在旅途中工作或娱乐。

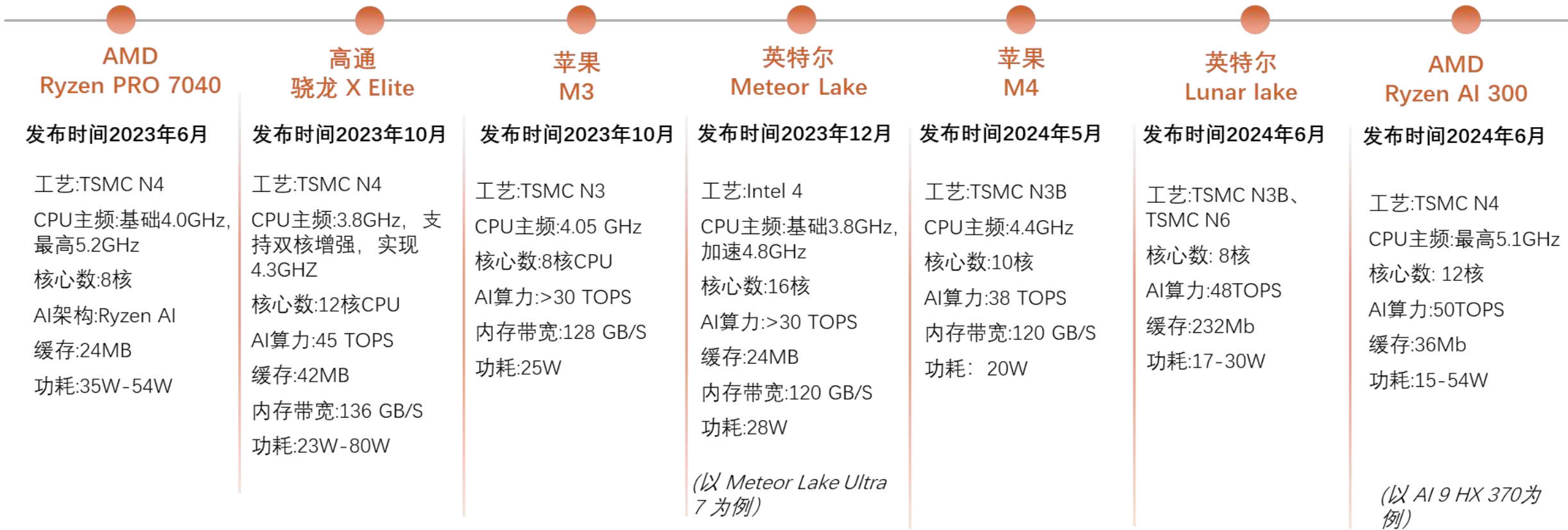
资料来源：微软官网、华源证券研究

2.2 变革二：异构计算成为AIPC新品主流，引入NPU成为关键

■ **AIPC处理器异构计算成为共性，增加NPU本质是为了进行低功耗计算**

- 引入NPU对于提升计算效率、降低能耗与成本具有重要意义。1) **计算效率与性能**：NPU专门为神经网络计算设计，能够高效处理大量并行计算任务，显著提高AI应用的性能。例如，在图像识别、自然语言处理和自动驾驶等领域，NPU可以加速训练和推理过程。2) **能耗与成本**：与通用处理器（如CPU、GPU）相比，NPU在执行特定AI任务时具有更高的能效比。这对于移动设备和嵌入式系统尤为重要，因为这些设备对功耗和电池寿命有严格的要求。3) **灵活性与可扩展性**：现代NPU通常采用可重构设计，能够适应不同的AI模型和算法。这种灵活性使得NPU可以在不同应用场景中高效运行，满足多样化的计算需求。

图：AIPC处理器发布历程



2.2 变革二：异构计算成为AIPC/手机新品主流，引入NPU成为关键

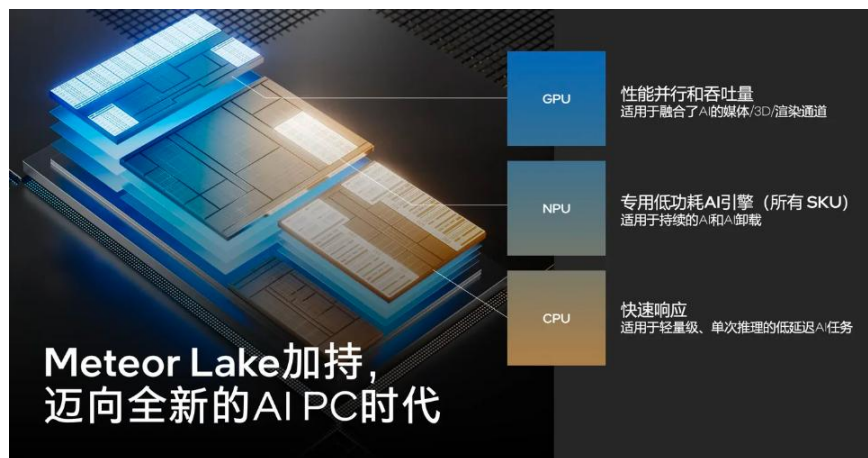
■ **AIPC处理器异构计算成为共性，增加NPU本质是为了进行低功耗计算**

- 硬件侧，“CPU+GPU+NPU”的异构架构成为英特尔、高通、AMD新一代AIPC处理器的共性。较传统CPU+GPU增加NPU，NPU专门用于处理人工神经网络计算，三种芯片结合，实现更快速、更高效率的边缘AI模型推理。
- **2023年12月英特尔推出首款面向AI PC的Meteor Lake酷睿Ultra处理器。**酷睿Ultra处理器采用分离式模块架构，由四个独立的模块组成，并通过Foveros 3D封装技术连接。其计算模块首次采用Intel 4制程工艺打造，带有基于Redwood Cove架构的P-Core和基于Crestmont架构的E-Core；核显采用了全新的Alchemist Xe-LPG设计；SOC模块里面包含了2个全新的LP E-Core，用于新型低功耗负载，进一步优化节能与性能间的平衡。同时酷睿Ultra处理器还采用了英特尔首个用于客户端的片上AI加速器“神经网络处理单元（NPU）”，将高能效AI加速提升到了新的高度，带来2.5倍于上一代产品的能效表现。**根据公司官网信息，intel计划将在 2025 年之前在超过 1 亿台 PC 上实现人工智能。**
- **2024年6月，Intel正式发布下一代面向AIPC的移动处理器Lunar Lake，综合算力达到120TOPS，NPU算力达到48TOPS，同时能耗大幅降低。**Lunar Lake处理器AI总算力达到120TOPS，其中CPU可提供5 TOPS的算力，驱动轻度AI工作；GPU提供67 TOPS算力提供游戏与创作所需的AI性能；NPU提供48 TOPS算力能够提供AI辅助与创作等功能。经过重新设计，为 x86 的能效设定了新的标准，lunar lake再供电和电源管理方面大幅改进，采用更先进的工艺节点，相较上一代meteor lake能耗降低40%。目前Lunar Lake已有来自20家OEM厂商，超过80款设计，预计第三季度开始出货。

- **2023年12月英特尔推出首款面向AI PC的Meteor Lake酷睿Ultra处理器。**酷睿Ultra处理器采用分离式模块架构，由四个独立的模块组成，并通过Foveros 3D封装技术连接。其计算模块首次采用Intel 4制程工艺打造，带有基于Redwood Cove架构的P-Core和基于Crestmont架构的E-Core；核显采用了全新的Alchemist Xe-LPG设计；SOC模块里面包含了2个全新的LP E-Core，用于新型低功耗负载，进一步优化节能与性能间的平衡。同时酷睿Ultra处理器还采用了英特尔首个用于客户端的片上AI加速器“神经网络处理单元（NPU）”，将高效AI加速提升到了新的高度，带来2.5倍于上一代产品的能效表现。**根据公司官网信息，intel计划将在 2025 年之前在超过 1 亿台 PC 上实现人工智能。**

- 2024年6月，Intel正式发布下一代面向AIPC的移动处理器Lunar Lake，综合算力达到120TOPS，NPU算力达到48TOPS，同时能耗大幅降低。Lunar Lake处理器AI总算力达到120TOPS，其中CPU可提供5 TOPS的算力，驱动轻度AI工作；GPU提供67 TOPS算力提供游戏与创作所需的AI性能；NPU提供48 TOPS算力能够提供AI辅助与创作等功能。经过重新设计，为 x86 的能效设定了新的标准，lunar lake再供电和电源管理方面大幅改进，采用更先进的工艺节点，相较上一代meteor lake能耗降低40%。目前Lunar Lake已有来自20家OEM厂商，超过80款设计，预计第三季度开始出货。

图：CPU、GPU、NPU组成多元AI算力



图：英特尔 ultra处理器架构



图：下一代lunar lake将进一步提升算力



资料来源：IT之家、intel、华源证券研究



2.2 变革二：异构计算成为AIPC/手机新品主流，引入NPU成为关键

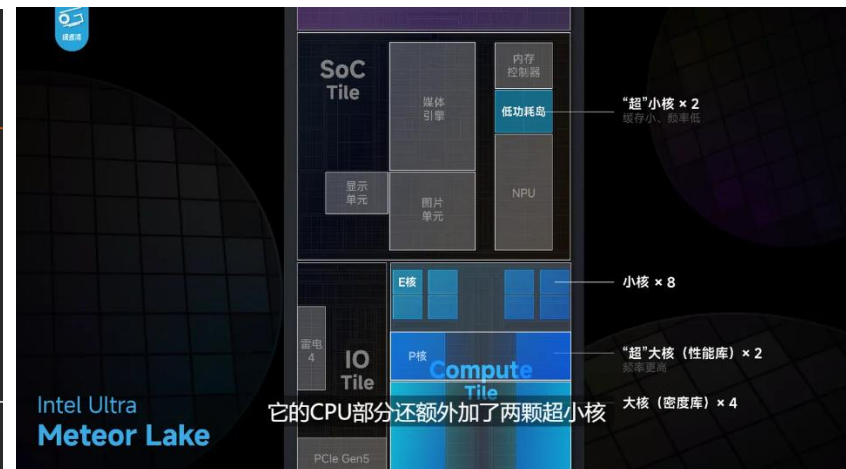
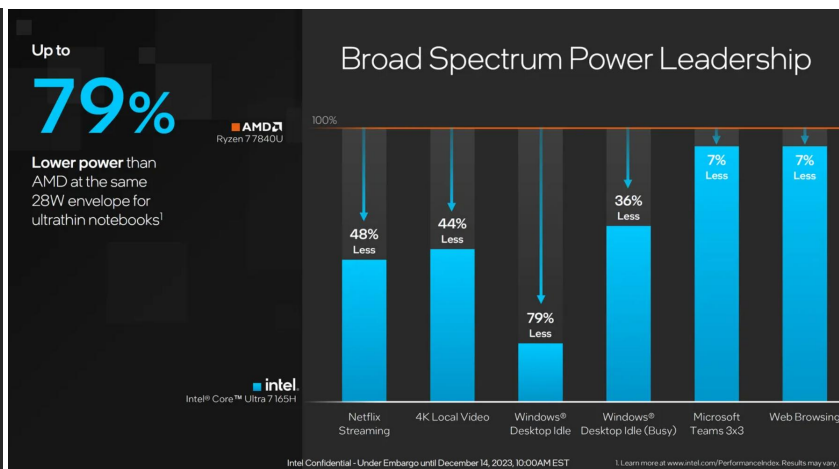
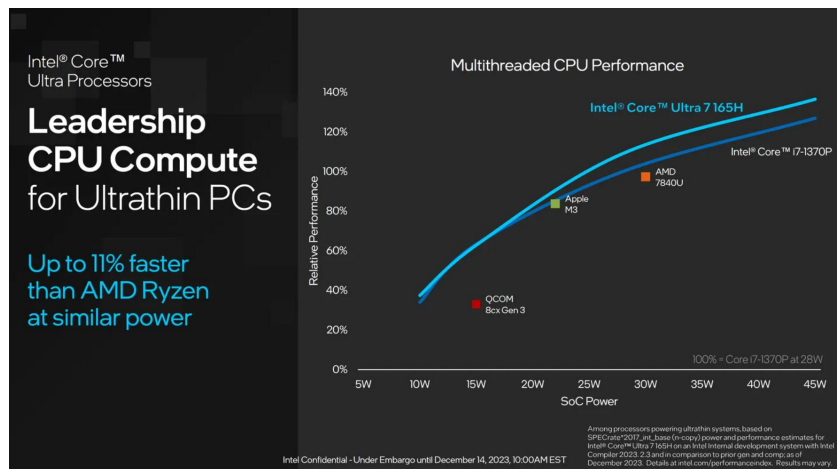
■ 硬件性能大幅提升的同时，intel Ultra依托3D高性能混合架构，功耗实现大幅下降

- 在AI应用方面，英特尔ultra处理器性能显著提升，较上一代性能提升70%，同时也领先AMD上一代产品。根据英特尔公布的对比数据显示，与英特尔Core i7-1370P相比，Core Ultra 7 165H的Generative AI（生成式AI）性能提高了70%；与AMD Ryzen 7 7840U相比，英特尔Core Ultra 7 165H在GIMP Stable Diffusion性能方面提高了5.4倍，在Stable Difference A1111性能方面提高了3.2倍，在Adobe Premier Pro（ColorGrade+场景编辑+导出）方面提高了1.7倍，在Adobe Lightroom Classic（AI照片编辑）方面提高了1.5倍，DaVinci Resolve（渲染+AI遮罩+导出）的性能提高了1.2倍，Wondershare Filmora（A1FX+预览+导出）性能提高了1.1倍。
- Ultra系列产品较上一代产品功耗显著降低。除了性能大幅提升，功耗也明显降低，与英特尔Core i7-1370P相比，Core Ultra 7 165H在Zoom视频会议中可降低38%的功耗；相较于AMD Ryzen 7 7840U，在同样的28W超薄设备上，在空闲模式下，英特尔Core Ultra 7 165H的功耗要AMD Ryzen 7 7840U低了79%；当功率突破20W后，比苹果M3更快。
- Ultra系列功耗的下降核心在于其采用Tile分离模块设计，将中高功耗、低功耗、超低功耗、GPU等负载场景和运算模块分成了不同的Tile。在低功耗场景可以让NPU或者LPE单独承载，CPU和GPU降频或者基本静置，大幅降低功耗。英特尔酷睿Ultra采用全新的3D高性能混合架构，由性能核（P-Core）、能效核（E-Core）和低功耗能效核（LP E-Core）组成。这种设计允许处理器根据不同的应用场景智能地调节核心的活动状态。

图：英特尔 ultra处理器在相同功率下的性能表现

图：英特尔 ultra处理器不同场景下的功耗下降情况

图：英特尔 ultra处理器较上一代新增两块LP E核



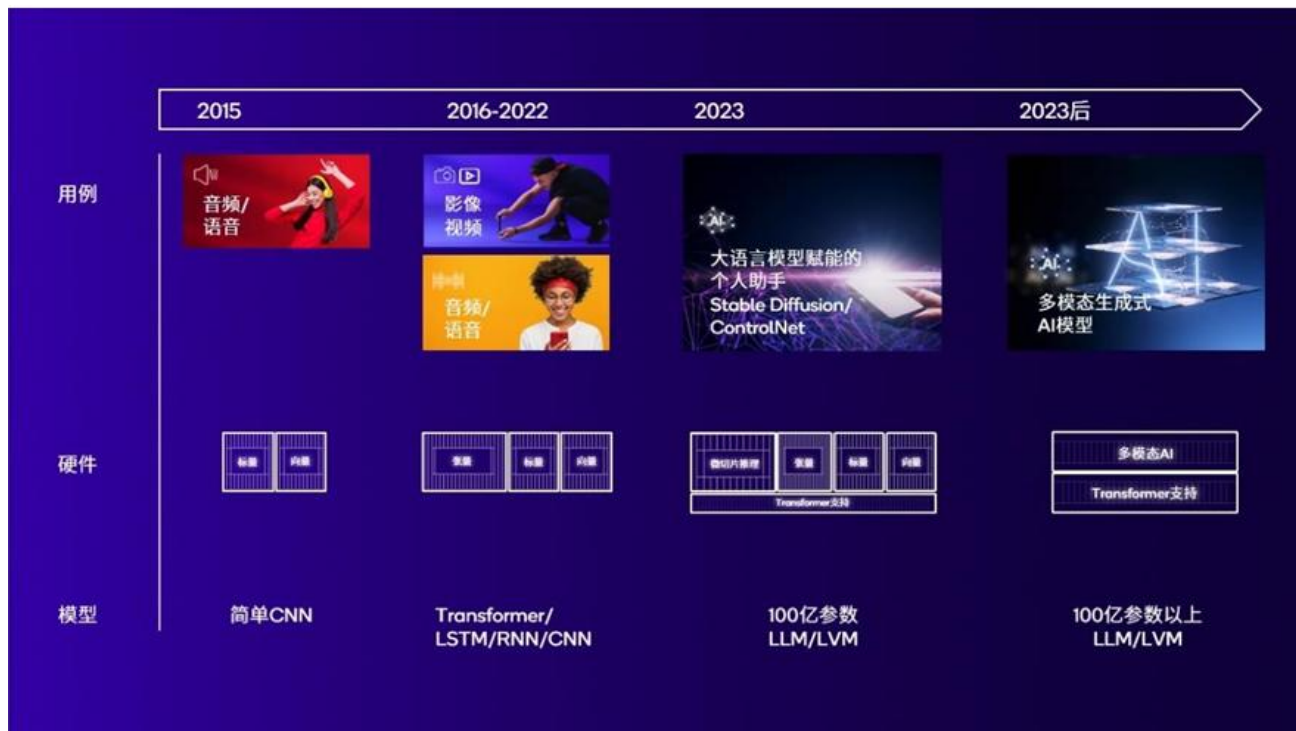


2.2 变革二：异构计算成为AIPC/手机新品主流，引入NPU成为关键

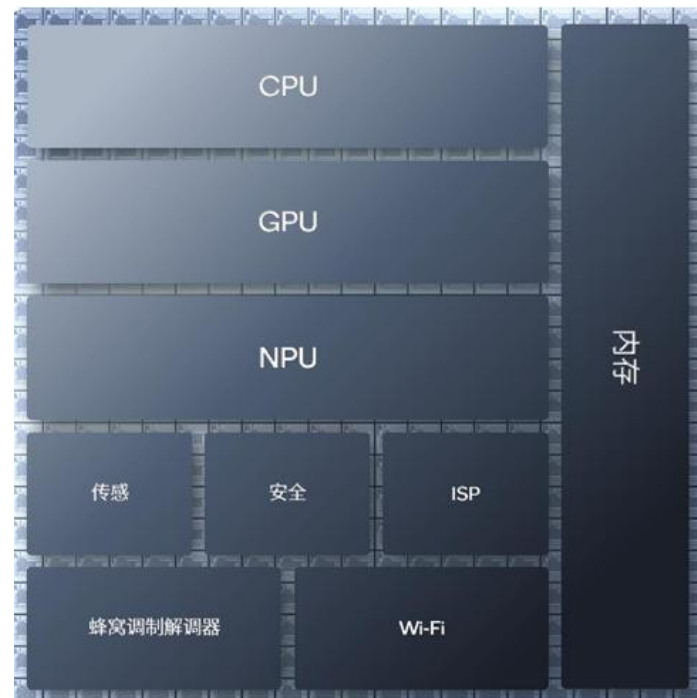
■ AIPC处理器异构计算成为共性，增加NPU本质是为了进行低功耗计算

- 随着生成式AI用例需求在有着多样化需求和计算需求的垂直领域不断增加，需要专为AI定制设计的全新计算架构。首先需要面向生成式AI全新设计的神经网络处理器（NPU），同时要利用异构处理器组合，比如中央处理器（CPU）和图形处理器（GPU）。通过结合NPU使用合适的处理器，异构计算能够实现最佳应用性能、能效和电池续航。
- NPU专为实现以低功耗加速AI推理而全新打造，并随着新AI用例、模型和需求的发展不断演进。根据高通《通过NPU和异构计算开启终端侧生成式AI》，在2015年，早期NPU面向音频和语音AI用例而设计，这些用例基于简单卷积神经网络并且主要需要标量和向量数学运算；2016年开始，拍照和视频AI用例大受欢迎，出现了基于transformer、循环神经网络、长短期记忆网络和更高维度的卷积神经网络等更复杂的全新模型，这些工作需要大量张量数学运算，因此NPU增加了张量加速器和卷积加速，大幅提升处理效率；2023年，LLM、LVM赋能的生成式AI使得典型模型的大小超过了一个数量级，因此还需要重点考虑内存和系统设计，通过减少内存数据传输以提高性能和能效。

图：NPU随着不断变化的AI用例和模型持续演进，实现高性能低功耗



图：现代SoC在单个DIE中集成多个处理器



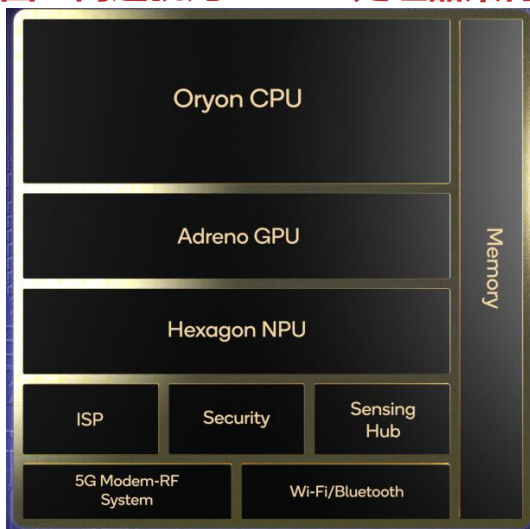


2.2 变革二：异构计算成为AIPC/手机新品主流，引入NPU成为关键

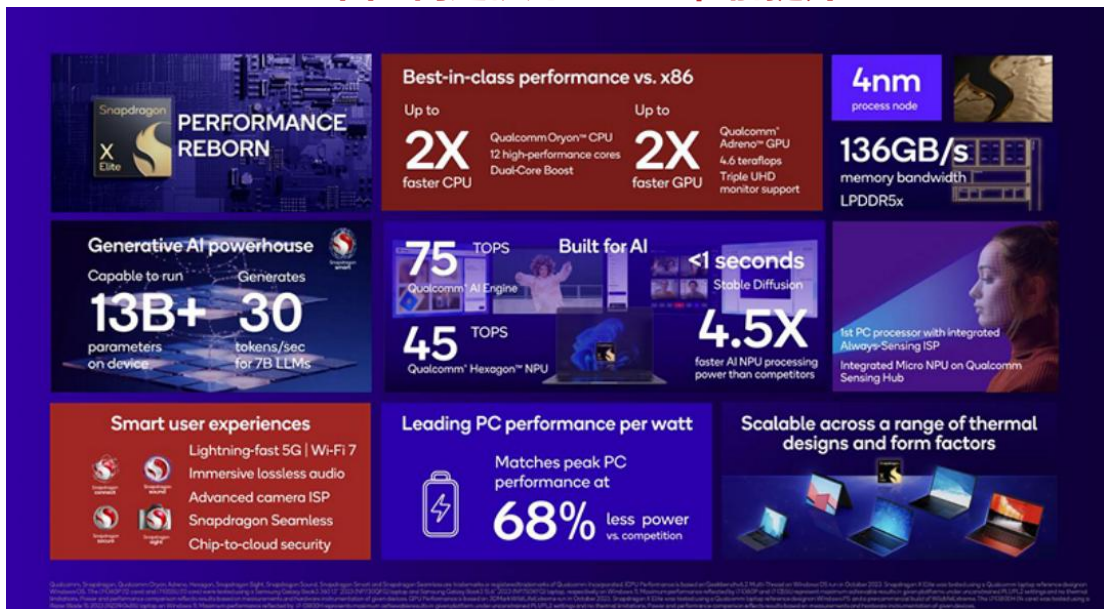
高通AIPC处理器NPU算力领先

- 2023年10月，高通推出专为AI而打造的Snapdragon X Elite，具有强大的计算和AI处理能力，有望成为未来AI PC的核心组件
- CPU:** Snapdragon X Elite 平台采用定制的集成 Qualcomm Oryon CPU（移动计算领域的新型CPU领导者），与竞争对手相比，其CPU性能快2倍，峰值性能与竞争对手相当，但功耗仅为三分之一。**GPU:** 集成了高通Adreno GPU，图形处理能力可达4.6TFLOPs，能够在图形密集型任务中表现出色；**NPU:** 内置的Hexagon NPU可以提供45 TOPS（每秒万亿次运算）的AI计算性能，能够高效处理大型生成式AI模型（如具有130亿参数的模型），大幅领先于友商最新x86架构芯片NPU的算力数值。在面向Windows的UL Procyon AI基准测试中，与其他PC竞品相比，骁龙X Elite具有领先的性能。例如，骁龙X Elite的基准测试总分分别为x86架构竞品A的3.4倍和竞品B的8.6倍。

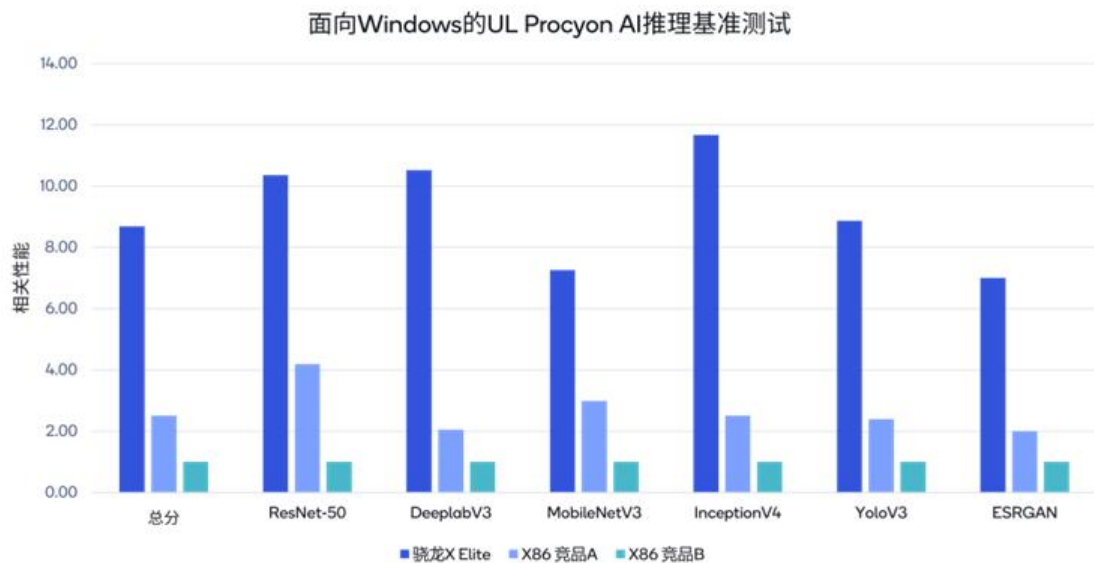
图：高通骁龙 X elite处理器架构



图：高通骁龙 X elite性能提升



图：高通骁龙 X elite在Procyon基准测试中具有领先的AI性能



资料来源：高通官网、高通《通过NPU和异构计算开启终端侧生成式AI》、华源证券研究

■ AI手机处理器同样依托NPU提供高性能算力

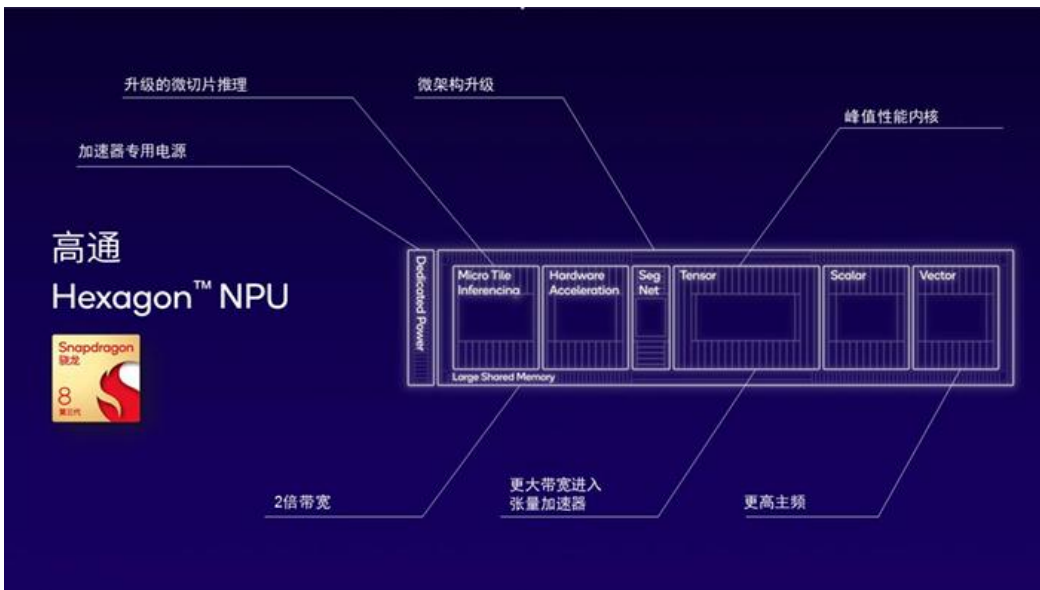
- 高通Hexagon NPU不断演进，满足快速变化的AI需求，高通NPU的核心优势在于能够以低功耗实现持续稳定的高性能表现，2023年推出的骁龙8中的Hexagon NPU为AI推理带来98%性能提升和40%能效提升。
- 高通NPU的差异化优势在于系统级解决方案、定制设计和快速创新。首先在性能方面，高通NPU结合了标量、向量和张量加速器，显著提升了AI计算性能，能够处理复杂的AI模型和任务。同时，通过先进的电源管理和优化技术，高通NPU在提供高性能的同时，保持低功耗，非常适合移动设备的需求。
- 异构计算架构是高通NPU的另一大优势。高通NPU能够与GPU、DSP等其他计算单元协同工作，实现更高效的任务调度和资源利用，进一步提升整体计算效率。此外，高通NPU还配备了专用的大共享内存，提供高带宽的数据传输，减少延迟，提高处理速度。
- 高通NPU在技术创新方面也表现出色。例如，通过量化技术降低计算精度，减少内存占用和计算量，同时保持高模型精度和性能；微切片推理技术则将神经网络分割成多个独立的微切片，最大化NPU的利用率，降低内存占用，提升推理速度。

表：高通NPU历代演进

时间	发展历程
2007年	首款Hexagon DSP在骁龙平台上正式亮相，DSP控制和架构是高通未来多代NPU的基础。
2015年	骁龙820处理器推出，集成首个高通AI引擎，支持成像、音频和传感器运算。
2018年	骁龙855中为Hexagon NPU增加了Hexagon张量加速器。
2019年	骁龙865扩展AI应用，包括AI成像、AI视频、AI语音和始终在线的感知功能。
2020年	Hexagon NPU架构更新，融合标量、向量和张量加速器，提升性能和能效，打造专用大共享内存。
2022年	第二代骁龙8中的NPU引入专用电源传输通道、微切片推理、本地4位整数运算、Transformer网络加速和其他特殊硬件。
2023年	第三代骁龙8中的NPU提升98%性能和40%能效，升级微架构，优化内存带宽，增加指令集和堆栈，带宽倍数增加。

资料来源：高通《通过NPU和异构计算开启终端侧生成式AI》、华源证券研究

图：第三代骁龙8的Hexagon NPU升级以低功耗加速实现领先的生成式AI性能



资料来源：高通《通过NPU和异构计算开启终端侧生成式AI》、华源证券研究

2.3 变革三：大模型落地端侧提升设备对于内存的需求

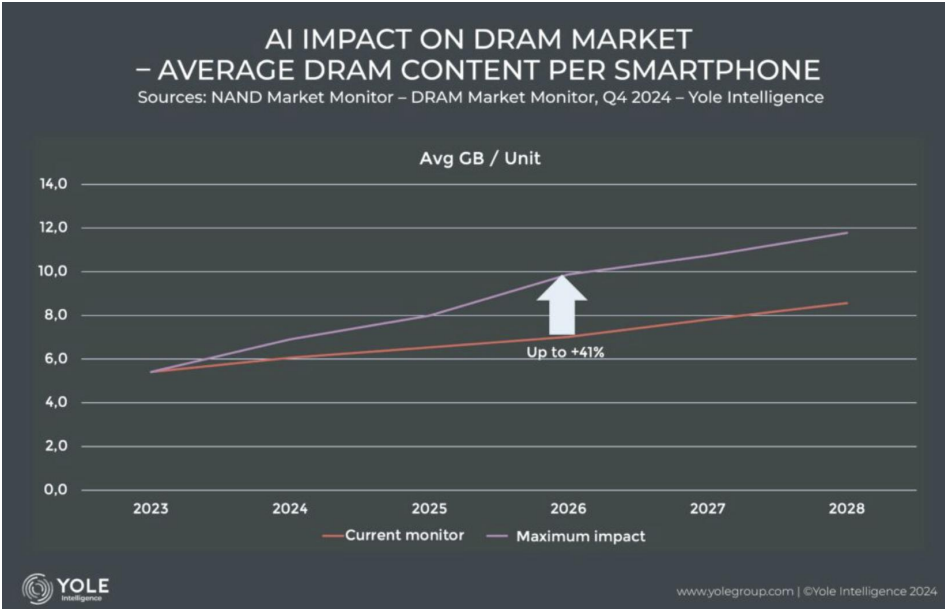
- **大模型的端侧落地，一方面要求算力的提升，另一方面也要求内存空间的提升，由此带来内存需求的提升，单机价值量提升**
- 内存是AI PC的重要组成部分，直接影响整体系统性能和AI任务的实现。随着AI模型变得更大更复杂，AI PC需要更高的内存容量。在PC市场，微软首批Copilot+PC硬件最低要求是内存16GB DDR5/LPDDR5，以及存储256GB SSD/UFS或更大的存储设备，AI PC有望带动PC平均搭载容量提升，根据集邦咨询预测，DRAM于笔电的单机平均搭载容量年增率约12.4%，后续随着AI PC量产后，2025年成长幅度会更明显。
 - 手机代际AI升级有望推动DRAM需求激增，人工智能（AI）已融入智能手机功能多年，但在高端设备中实施改良的大型语言模型（LLM）可能会导致对 DRAM 的需求上升，并加速最小 NAND 存储容量的淘汰。根据YOLE预计，2023 年，高端智能手机的平均 DRAM 容量为 9GB，随着制造商开始整合基于生成式 AI 的功能，预计到 2024 年内容将接近 10GB。

表：Surface Pro系列对比

	Surface Pro 11	Surface Pro 10
处理器	Snapdragon® X Plus (10 Core), with LCD display Snapdragon® X Elite (12 Core), with OLED display	Intel® Core™ Ultra 5 processor 135U Intel® Core™ Ultra 7 processor 165U
NPU	Qualcomm® Hexagon™ with 45 trillion operations per second	Intel® AI Boost
图形处理器	Qualcomm® Adreno™ GPU	Intel® Graphics
内存和存储	内存: 16GB or 32GB LPDDR5x RAM 存储: Removable solid-state drive (Gen 4 SSD) options: 256GB, 512GB, 1TB	内存: 8 GB, 16 GB, 32 GB, 64 GB LPDDR5x 存储: Removable solid-state drive (Gen 4 SSD): 256 GB, 512 GB, 1 TB

资料来源：Microsoft公司官网、华源证券研究

图：智能手机内存容量预计大幅增长



资料来源：YOLE、华源证券研究

- 1、AIPC&手机是大模型落地的重要载体，颠覆传统应用模式
- 2、硬件端：处理器架构变化，内存价值量提升
- 3、大模型端：轻量化小模型陆续推出，多模态大模型改变人机交互体验
- 4、投资逻辑：AI端侧落地加速，换机潮蓄势待发，关注PC/手机厂及硬件相关投资机遇
- 5、风险提示

3.1 端云混合是AI计算的最终解决方案

■ 端云混合可能是未来AI计算的最终解决方案，AI落地终端具有必要性

- 高通提出了混合AI的概念，据高通《混合AI是AI的未来》，混合AI指终端和云端协同工作，在适当的场景和时间下分配AI计算的工作负载，能够为使用者提供更好的体验同时高效利用资源。在一些场景下，计算将主要以终端为中心，在必要时向云端分流任务；而在以云为中心的场景下，终端将根据自身能力，在可能的情况下从云端分担一些AI工作负载。
- AI落地终端具有必要性，AI落地终端能够在全球范围带来成本、能耗、性能、隐私、安全和个性化优势。混合AI具有显著的五大优势。首先，它能够将部分处理从云端转移到边缘终端，从而减轻云基础设施压力并减少开支，显著降低生成式AI模型的运营成本。其次，边缘终端能够以更低的能耗运行生成式AI模型，帮助云服务提供商降低数据中心的能耗，促进环境和可持续发展目标。第三，终端侧AI处理可靠，能够在网络拥堵时仍然提供高性能和低时延，防止高峰期排队等待和拒绝服务的情况发生。第四，终端侧AI有助于保护用户隐私，查询和个人信息完全保留在终端上，减少敏感数据泄露的风险。最后，终端侧AI能够实现更加个性化的体验，基于用户行为和偏好定制响应，支持个性化生成式AI应用。这些优势使混合AI在推动生成式AI规模化扩展方面至关重要。

图：AI处理的中心正在向边缘转移



资料来源：高通《混合AI是AI的未来》、华源证券研究

表：AI终端运行在成本、能耗、性能、隐私、安全和个性化方面具有优势

优势	详细描述
成本	将部分处理从云端转移到边缘终端，减轻云基础设施压力并减少开支。生成式AI模型的终端侧运行能显著降低运营成本。
能耗	边缘终端能够以更低的能耗运行生成式AI模型，帮助云服务提供商降低数据中心的能耗，推动环境和可持续发展目标。
可靠性、性能和时延	终端侧AI处理可靠，在网络拥堵时仍能提供高性能和低时延，防止因高峰期产生的排队等待和拒绝服务情况。
隐私和安全	终端侧AI有助于保护用户隐私，查询和个人信息完全保留在终端上，减少敏感数据泄露风险。
个性化	终端侧AI能够实现更加个性化的体验，基于用户行为和偏好定制响应，支持个性化生成式AI应用。

资料来源：高通《混合AI是AI的未来》、华源证券研究

3.2 大模型轻量化技术使得复杂大模型在端侧运行成为可能

■ 模型轻量化技术使得端侧部署大模型成为可能

- 在大模型压缩方面，常规的模型压缩方法有模型稀疏化、权重矩阵分解、模型参数共享、蒸馏和量化。
- 对于超大模型精度无损的压缩，可以采用多策略组合压缩方案。通过组合使用模型稀疏化、蒸馏和参数共享等压缩技术，可以在精度无损的情况下，将模型参数量压缩至百分之一、甚至千分之一左右。

表：模型轻量化的主要方法

方法	具体介绍
模型稀疏化	通过将模型中的某些神经元、连接或层置为零，从而达到压缩模型、加速训练、减少内存消耗等目的。
权重矩阵分解	使用包括奇异值分解（SVD）等矩阵分解方法对预训练模型的Feed-Forward Network（FFN）层的权重矩阵进行分解，从而减少Attention层的参数量，提高模型的效率。
模型参数共享	部分大型模型如 ALBERT采用了权重共享的方式，特定层之间共享参数，从而减少了模型的参数量。
蒸馏	通过使用学生模型来模拟预训练教师模型的行为来减小模型大小的技术。通常情况下，学生模型由更小的神经网络或线性模型组成。蒸馏的过程是将教师模型的知识转移到学生模型，使学生模型在保持较小规模的同时，能够保持类似于教师模型的预测能力。利用蒸馏技术可以将大模型的知识 and 泛化能力迁移到小型网络，以支持轻量化的大模型部署。
量化	量化是一种将预训练模型中的权重从浮点数转换为低位数的技术。通常情况下，量化的精度可被降低到 8 位或更低。量化可以大大减少模型的存储空间和计算量，但可能会对模型的性能产生一定的影响。

资料来源：中国人工智能系列白皮书——大模型技术、华源证券研究

表：TensorFlow量化对模型的优化

Model	准确率（原始）	准确性（训练后量化）	准确性（量化感知训练）	延迟（原始）(ms)	延迟（训练后量化）(ms)	延迟（量化感知训练）(ms)	大小（原始）(MB)	大小(优化后)(MB)
Mobilenet-v1-1-224	0.709	0.657	0.7	124	112	64	16.9	4.3
Mobilenet-v2-1-224	0.719	0.637	0.709	89	98	54	14	3.6
Inception_v3	0.78	0.772	0.775	1130	845	543	95.7	23.9
Resnet_v2_101	0.77	0.768	N/A	3973	2868	N/A	178.3	44.9

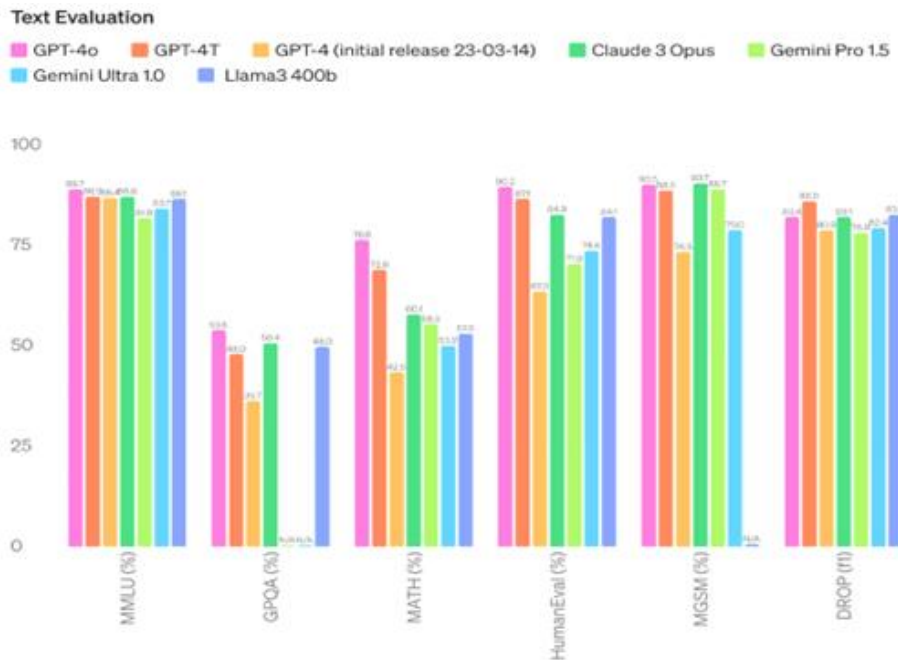
资料来源：TensorFlow官网、华源证券研究



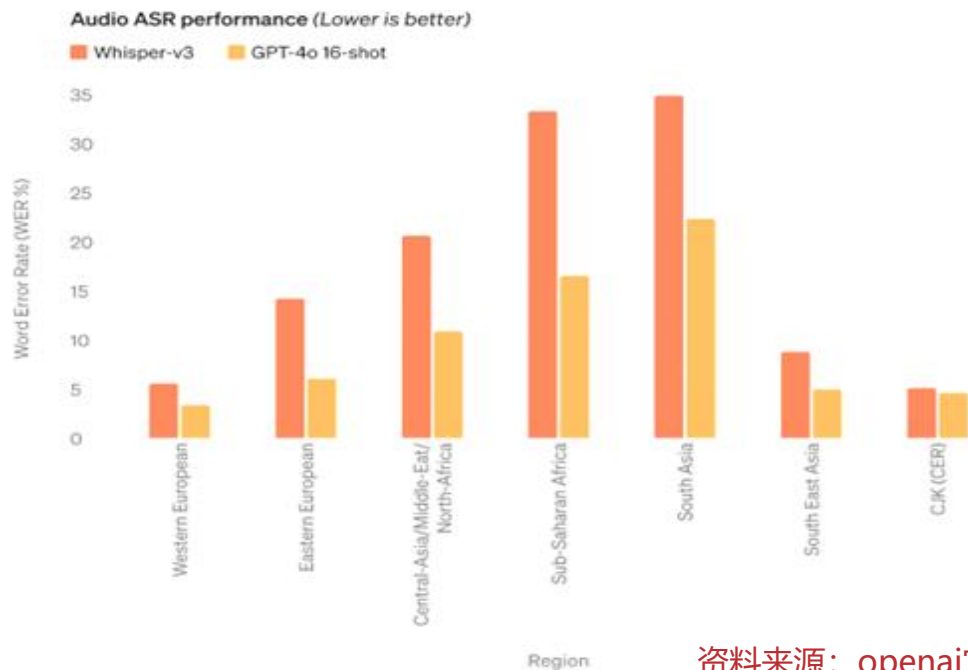
■ 多模态大模型，如GPT-4o和Project Astra，使得人机交互变得更加自然

- 2024年5月，openai发布GPT-4o，GPT-4o相较于GPT4延迟显著降低，对于语音的理解能力大幅提升，能够直接观察音调、多个说话者或背景噪音，能够输出笑声、歌唱或表达情感，实现更加自然的人机交互，响应时间的优化大幅提升实用度，使用需求有望进一步提升。**GPT-4o是迈向更自然的人机交互的一步**，它接受文本、音频和图像的任意组合作为输入，并生成文本、音频和图像的任意组合输出，整个过程由同一神经网络处理。它可以在短至 232 毫秒的时间内响应音频输入，平均为 320 毫秒。而在之前的GPT-3.5和GPT-4的语音模式中，平均延迟为2.8秒和5.4秒。GPT-4o通过端到端的训练方式处理所有的输入和输出，这意味着所有输入和输出都由同一神经网络处理。
- 在 Google 新一代 Gemini 技术的推动下，Google Astra 提供了一种更流畅、更直观的方式进行人机交互。Gemini 框架擅长分析大量图像、视频和音频数据，使 Astra 能够学习并适应周围环境的复杂性。Project Astra旨在通过结合视频、音频和文本输入，实现更加流畅的人机交互。

图：GPT-4o文本能力对比



图：GPT-4o音频ASR性能对比



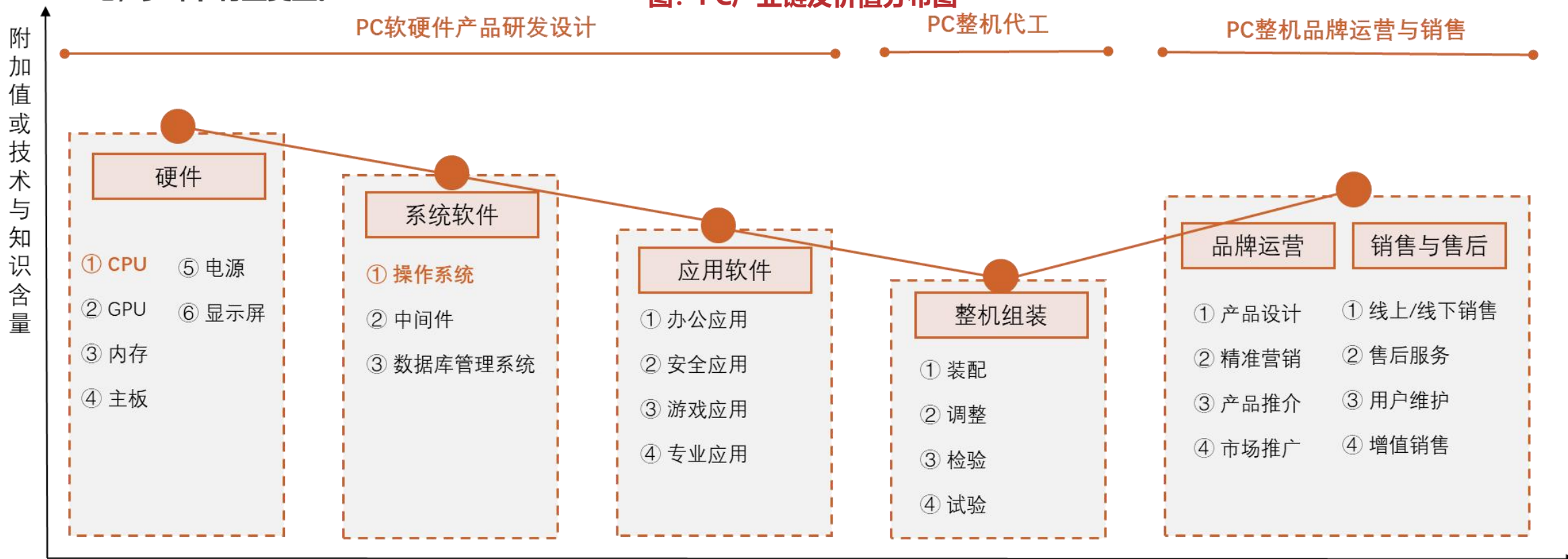
- 1、AIPC&手机是大模型落地的重要载体，颠覆传统应用模式
- 2、硬件端：处理器架构变化，内存价值量提升
- 3、大模型端：轻量化小模型陆续推出，多模态大模型改变人机交互体验
- 4、投资逻辑：AI端侧落地加速，换机潮蓄势待发，关注PC/手机厂及硬件相关
投资机遇
- 5、风险提示

4. AI落地端侧为产业链各环节带来投资机会

■ AI落地端侧为产业链各环节带来投资机会

- 从硬件层面来看，CPU是核心，高通具备先发优势，有望在intel和AMD的产品尚未满足微软AIPC要求的情况下，率先抢占部分市场；从系统软件层面来看，微软copilot渗透率迅速提升，形成正向循环，一方面可以促进其windows OEM业务增长，另一方面同样可以促进其云业务需求；从整机组装环节来看，比亚迪电子受益其海外大客户落地AI，有望推动其手机出货量，此外AI落地端侧，推动消费电子景气度复苏，公司消费电子组装和零部件业务有望受益；从品牌运营的角度来看，建议关注联想集团，公司率先布局本地大模型，AIPC有望带动公司智能设备业务恢复增长；此外建议重点关注苹果，自身在硬件、软件生态、终端产品具备一体化优势，苹果AI正式落地，多环节有望受益。

图：PC产业链及价值分布图



■ 高通在AIPC处理器领域具备先发优势

- 首批copilot+PC搭载高通骁龙X系列处理器，有望助力公司突破PC市场。**1Q24年公司实现营收93.89亿美元（yoy+1.2%，qoq-5.5%），净利润23.26亿美元（yoy+36.5%，qoq-15.9%）；稀释 EPS 2.06美元。QCT收入达 80 亿美元，反映出对 Android 智能手机的强劲需求以及汽车行业的持续增长势头，手机芯片业务仍然是公司营收的主要来源，但增速有所放缓，而汽车和IoT业务则保持着较快的增长势头；QTL收入达13亿美元。AI有望成为高通未来的重要增长点。高通PC端处理器骁龙X系列在性能和功耗上较intel和AMD具有优势，同时公司和微软共同定义了AIPC 40TOPS的门槛，公司率先满足其要求，微软的copilot+PC也率先推出搭载高通骁龙X系列处理器的版本，同时公司也在扩宽其合作伙伴，联合多家OEM厂商促进其产品落地，包括宏碁、华硕、戴尔、惠普、联想、三星等。

图：高通AI 手机领域的布局

图：高通AI PC领域的布局

First generation of flagship Android smartphones from leading OEMs featuring on-device AI



Live translation
Live transcription
Photo portrait generation
Photo AI editing
Semantic search
Text summarization
Virtual assistants
Writing assistance
Image generation

New on-device AI use cases and experiences expected in future smartphone launches

骁龙X系列平台赋能的Copilot+设备 5月20日发布

Copilot+PC



Copilot+功能的可用性因具体设备和市场而异。

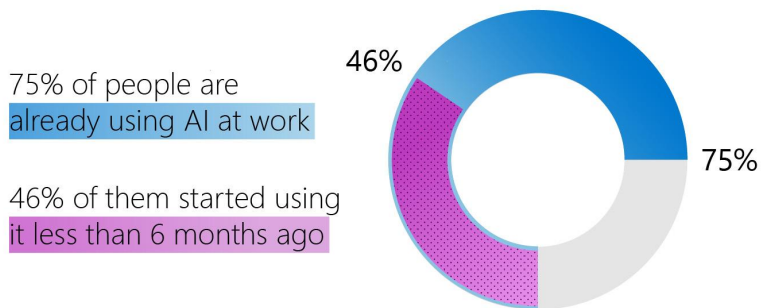
■ Copilot的融合大幅提升用户的工作效率

- 在AI技术的推动下，微软通过Copilot项目在PC端带来了显著的应用变化，特别是在工作效率和用户体验方面。Copilot集成了生成式人工智能，能够帮助用户处理各种办公任务，如会议总结、邮件处理、文档编辑和数据分析，显著提升了工作效率和创造力。最重度的Teams用户在3月使用Copilot总结了8小时的会议，相当于节省了整整一个工作日，Copilot用户在处理电子邮件时，整体上减少了11%的单个邮件阅读时间，并减少了4%的互动时间。
- 此外，Copilot在Microsoft 365应用（如Word、Excel和PowerPoint）中的集成，使用户在文档编辑和数据分析方面更加高效，用户在这些应用中编辑的文档数量增加了10%，一些公司实现了20%的增长。Copilot还能够生成高质量的文本、图像和其他多媒体内容，帮助用户在创意工作中实现更高的效率和灵感激发，用户可以通过简单的指令生成各种内容，如市场营销材料、社交媒体帖子和产品描述。
- 同时，Copilot提供了强大的数据分析和可视化功能，帮助企业更好地理解业务数据并做出明智的决策，用户可以快速创建图表和报告，实时监控业务绩效和市场趋势，通过机器学习模型对历史数据进行分析并提供未来趋势的预测，帮助企业制定战略计划。

图：2024年工作趋势指数：职场AI使用情况

Three Out of Four People Use AI at Work

Usage nearly doubled in the last six months.



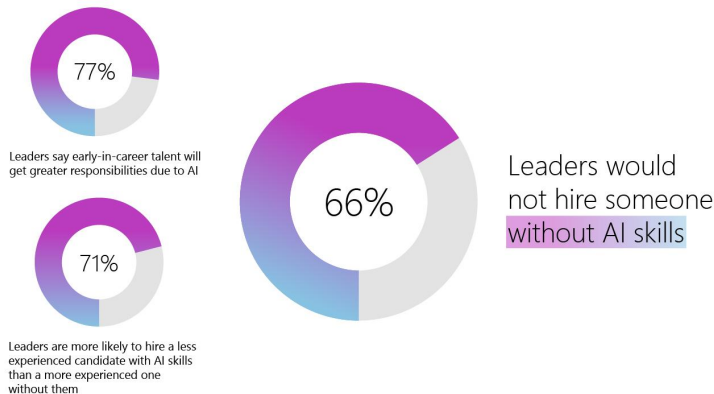
2024 Work Trend Index Annual Report
from Microsoft and LinkedIn

Survey questions:
How often do you use generative artificial intelligence (AI) for your work?
How long have you been using generative artificial intelligence (AI) at work?

图：2024年工作趋势指数：AI技能与招聘趋势

The New Hiring Imperative

AI aptitude takes center stage.



2024 Work Trend Index Annual Report
from Microsoft and LinkedIn

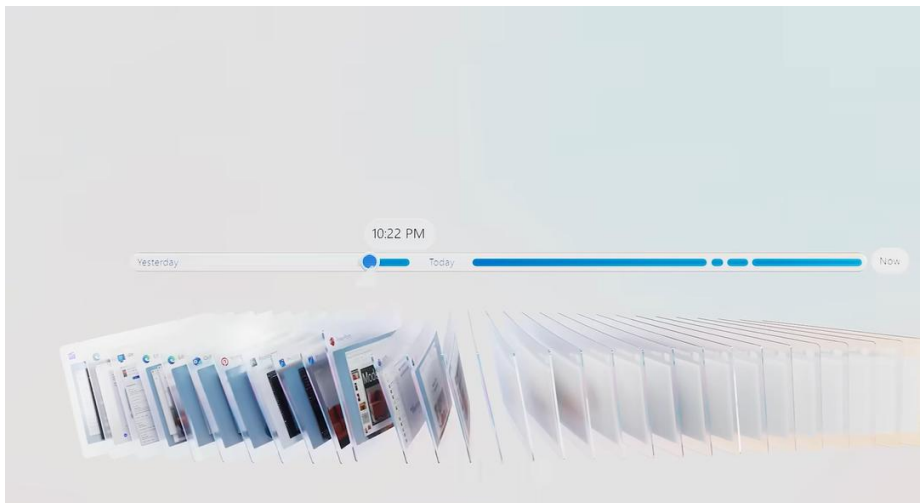
Survey question:
To what extent do you agree or disagree with the following statements about generative artificial intelligence's (AI's) impact on skills?
• I would be more likely to hire a less experienced candidate with AI skills than a more experienced candidate without AI skills
• In considering job candidates, I would not hire someone without AI skills
• Because they can delegate more work to AI, early-career talent will be given greater responsibilities

4.2 系统软件：微软copilot渗透率提升，驱动公司多项业务增长

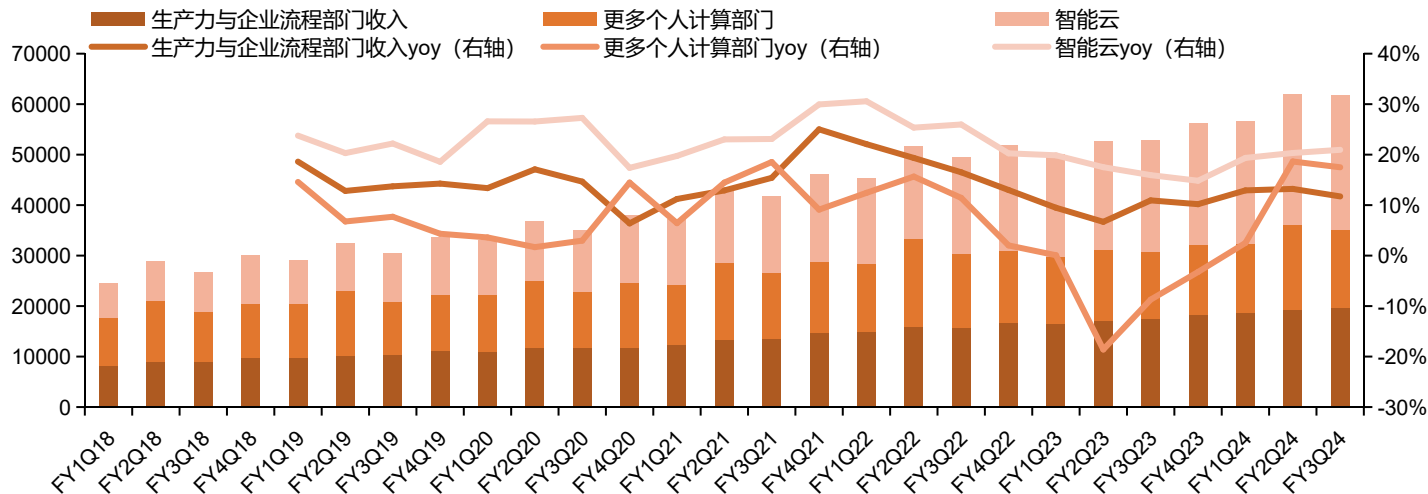
■ Copilot+PC六月正式上市，微软copilot加速渗透，形成良性循环

- 在Build 2024开发者大会前夕，5月20日微软发布了自家最新的“Copilot+PC”，是专为 AI 设计的新型 Windows PC，有望开启Windows on ARM新时代。微软不仅自己推出全新的Copilot+ PC——新一代的Surface Pro 和 Surface Laptop，同时与各大 OEM 厂商（宏碁、华硕、戴尔、惠普、联想、三星）合作全新 Copilot+PC，将于 6 月 18 日开始上市，起价为999 美元。
- 根据公司官网，Copilot+PC是迄今为止最快、最智能的 Windows PC。微软引入了全新的系统架构，将 CPU、GPU 和新的高性能神经处理单元 (NPU) 的强大功能结合在一起。凭借强大的新型芯片，NPU算力能够实现40+ TOPS、全天的电池寿命以及对最先进人工智能模型的访问，Copilot+ PC能够完成任何其他 PC上无法完成的事情。例如通过 Recall轻松查找并记住用户在PC中看到的内容，使用 Cocreator 直接在设备上近乎实时地生成和优化 AI 图像，并通过实时字幕消除语言障碍，将40 多种语言的音频翻译成英语等。此外微软还宣布与全球一些最大、最受欢迎的应用合作，利用 NPU 的强大功能提供全新的创新性 AI 体验，包括 Adobe、DaVinci Resolve Studio、capcut、cephable、liquidtext 等。
- 公司生产力与企业流程、更多个人计算部门收入增长强劲。FY3Q24公司生产力与企业流程部门实现营收195.7亿美元，同比增长12%，占总营收的32%；根据公司FY3Q24电话会，目前各行各业已有30000家组织已使用 Copilot Studio 为 Microsoft 365 定制 Copilot 或构建自己的 Copilot，环比增长 175%，近 60% 的《财富》500 强企业现在都在使用 Copilot，早期采用者的使用强度有所提高，包括 Teams 中每个用户使用 Copilot 辅助互动的次数增加了近 50%；Windows 中的 Copilot 目前已在近 2.25 亿台 Windows 10 和 Windows 11 PC 上使用，环比增长 2 倍。更多个人计算部门实现营收155.8亿美元，同比增长17%，占总营收的25%，主要得益于公司windows OEM表现好于预期。

图：recall 功能示意图



图：微软分业务收入及增速情况（百万美元）

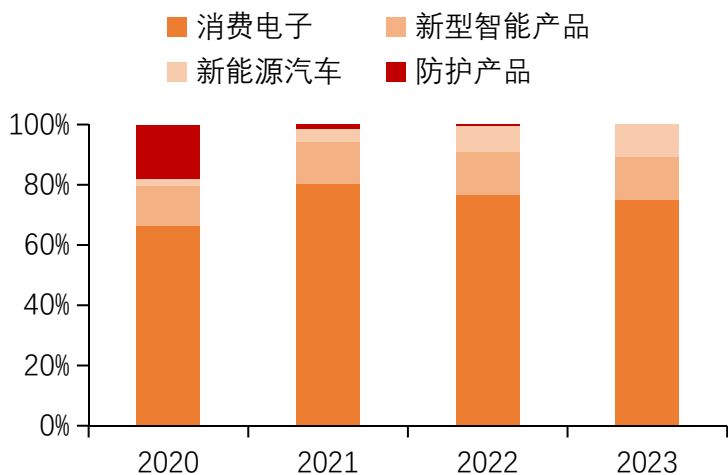


4.3 代工环节：比亚迪电子有望受益AI带动消费电子底部复苏

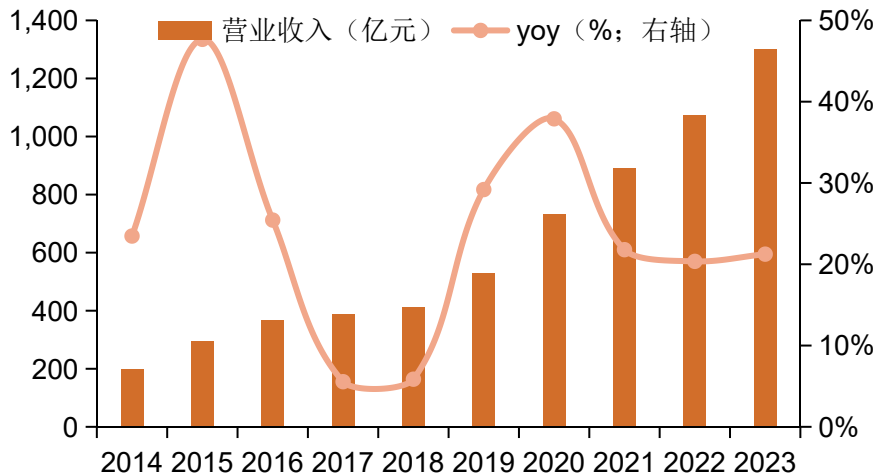
■ AI落地端侧有望推动消费电子底部复苏，比亚迪电子有望受益出货量增长以及份额提升

- 比亚迪电子是全球领先的平台型高端制造企业，为客户提供一站式服务。**比亚迪电子是全球智能手机及笔记本电脑核心供应商之一，深耕消费电子领域。凭借行业领先的材料技术、智造工艺以及“一体化”解决方案服务能力，造就其在智能手机结构件的领导地位。
- 公司消费电子板块业务由组装业务和零件业务组成。**组装业务方面，公司拥有多年的ODM、OEM制造经验，为国内外知名公司提供一站式制造服务，服务范围遍布消费电子、工业电子、汽车电子、医疗电子、智能硬件等多个领域。零件业务方面，自2006年进入结构件制造领域以来，发展成为国内最具竞争力的金属结构件供应商之一，目前已拥有30多项工艺技术，可为客户提供从初期成型到产品组装等一站式加工服务，此外公司能够为客户提供全面的塑胶结构件解决方案，也是国内最具竞争力的3D玻璃供应商之一。
- 消费电子业务受益AI端侧落地推动底部复苏。**消费电子板块2023年实现收入974亿，占公司总收入比重75%，近年来公司切入北美大客户，持续开拓核心产品线，提升市场份额，公司消费电子板块业绩逆势实现业务增长。2024年5月海外大客户发布搭载M4芯片的iPad Pro产品，硬件性能已逐步满足AI任务要求；安卓客户方面，公司持续聚焦中高端产品提供全方位服务，我们认为AI端侧落地有望带动公司消费电子业务持续健康成长。

图：公司各业务收入占比情况



图：公司营业收入情况



图：公司净利润情况



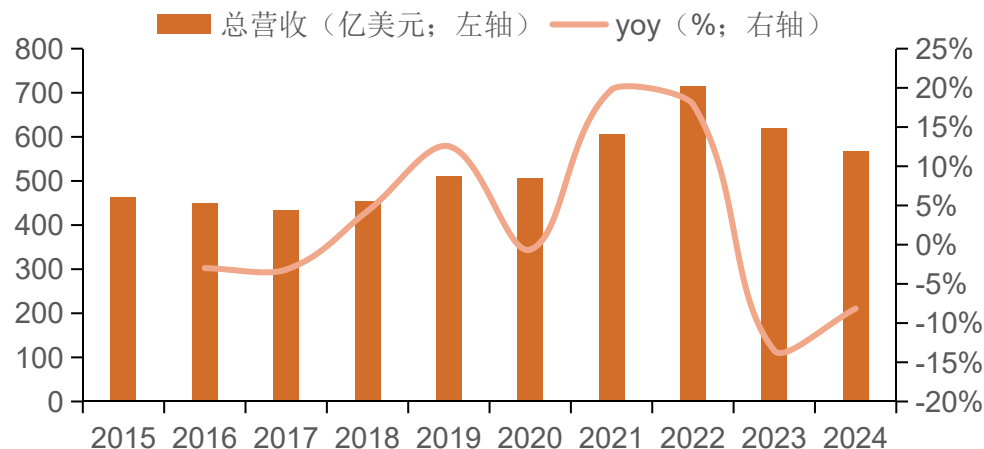


4.4 整机品牌：联想率先打造AI PC本地大模型，换机需求驱动公司业绩增长

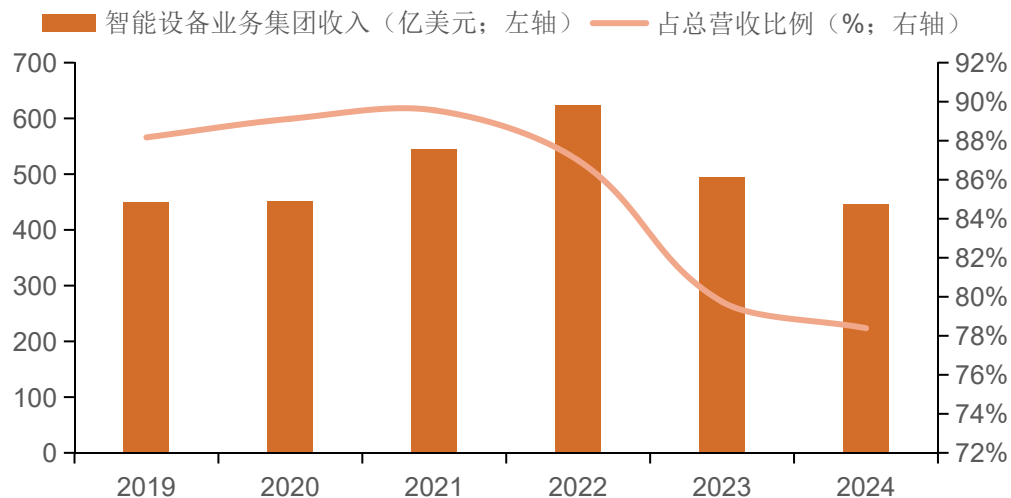
■ 联想在AI PC具有先发优势，致力于构建AI PC生态系统，开启面向AI的新十年

- 2024年4月，联想TechWorld 2024正式发布AI PC个人智能体“联想小天”。联想小天由嵌入本地的天禧大模型所驱动，能够与用户进行自然交互，并在工作、学习和生活等多个场景中应用AI功能，具备包括AI画师、AI PPT、文档总结、智会分身、知识问答、AI识图、会议纪要、设备调优、AI换机、AI服务等功能，同时联想小天具备自我学习和成长能力，通过日常互动，不断学习和理解用户的使用习惯和偏好，从而提供更加精准和人性化的服务。与此同时，个人智能体不仅有着更强的隐私保护属性，而且还可以调用本地知识库，为每一位用户提供更加个性化的服务，从而成为更懂用户的专属助理和管家。
- 产品端，2024年4月，联想TechWorld 2024发布六款AI PC新品，囊括Yoga、Think及小新系列，售价从5999元至17999元不等。
- AIPC有望拉动公司智能设备业务恢复增长，盈利改善。FY2024公司实现营业收入568.64亿美元，同比下降8.2%，实现归母净利润10.11亿美元，同比下降37.1%，其中智能设备业务（IDG）实现营收446亿美元，同比下滑10%，主要是由于上半年年初渠道库存过多，整体市场疲弱所影响，下半年需求逐步恢复，FY4Q24公司IDG业务实现营收104.63亿美元，同比上升7%，根据IDC市场数据，公司PC出货量同比增长7.8%，市场份额达到23%，同比增长1.4pct。随着联想AIPC在5月正式发售，AI PC元年开启，推动换机需求增长，公司IDG业务有望逐季回升，带动公司总体收入呈现复苏态势。

图：联想集团总营收及同比变化



图：联想集团IDG业务收入及其占比



资料来源：wind、华源证券研究

4.4 整机品牌：苹果具备一体化优势，多环节有望受益AI落地

■ 苹果AI落地，综合其一体化优势，有望带动公司收入恢复增长

- **苹果举办WWDC2024，推出Apple Intelligence，为其软硬件生态注入AI。** Apple Intelligence利用Apple芯片的强大功能来理解和创建语言和图像、跨应用程序采取行动，将AI融入智能写作工具、通知、邮件、备忘录、电话，同时还具备文生图的功能。在 Apple Intelligence 的支持下，Siri 也具备了全新的能力。凭借更丰富的语言理解能力，Siri 更加自然、更符合语境、更加个性化，能够简化和加速日常任务。Siri 具备了对个人背景的了解、在 app 内和跨 app 执行操作的能力。苹果与OpenAI合作，将ChatGPT集成到Siri和写作工具中，Siri 可以利用 ChatGPT 处理某些请求，包括有关照片或文档的问题，用户不需要 ChatGPT账户就可以免费使用GPT-4o，信息均不会被记录，苹果将ChatGPT集成在iOS18、iPadOS 18、macOS Sequoia中，同时苹果还将支持其他AI模型。
- **Apple Intelligence的核心是设备端处理，意味着用户的个人数据不会被收集或存储。** 苹果通过结合先进的硬件和软件技术，确保用户的数据仅用于实现特定的请求，并且不会被存储或访问。苹果先进的Apple芯片，包括A17 Pro和M系列芯片，给Apple Intelligence提供了坚实的算力基础，苹果走自研本地大模型加云端的策略，超出本地处理能力的问题可以借助云端大模型解决。

图：苹果发布Apple Intelligence

图：苹果将ChatGPT集成在iOS18、iPadOS 18、macOS Sequoia中

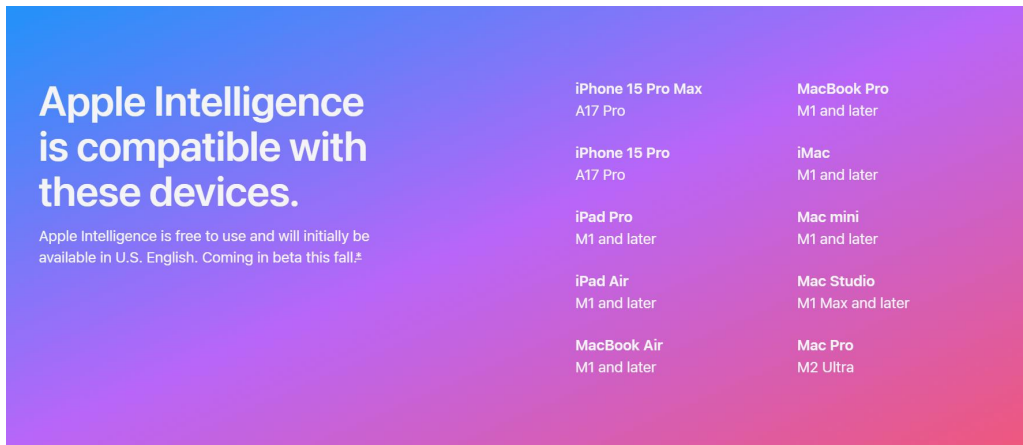


4.4 整机品牌：苹果具备一体化优势，多环节有望受益AI落地

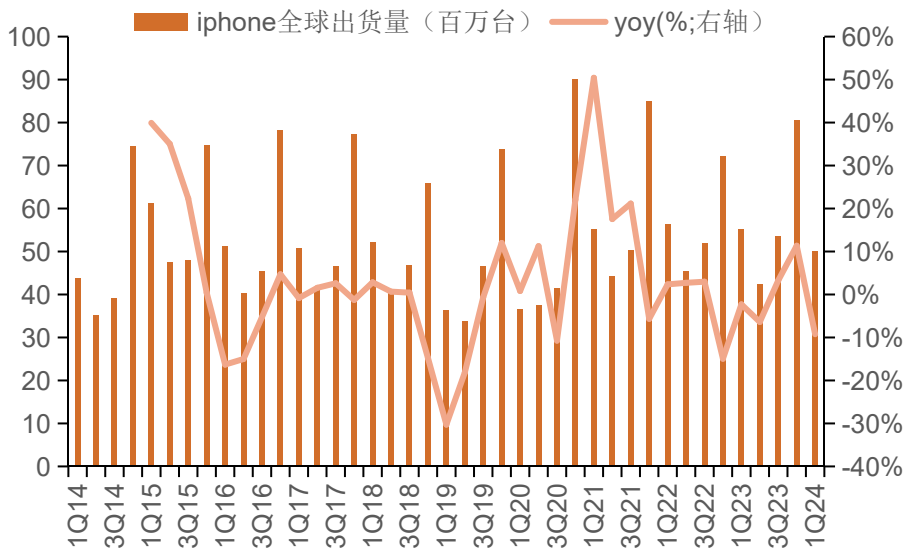
■ 苹果AI落地，综合其一体化优势，有望带动公司收入恢复增长

- Apple intelligence有望拉动公司多条产品线出货量回暖，其中手机现有产品支持本地AI功能仅有iPhone 15 Pro、iPhone 15 Pro Max，对于刺激iphone的销量回暖可能更为明显。根据公司官网，Apple intelligence将于今年秋天推出测试版，将在iPhone 15 Pro、iPhone 15 Pro Max以及搭载M1及更新机型的 iPad 和 Mac 上提供测试版（Siri 和设备语言设置为美国英语）。硬件层面，苹果已发布M4，率先用于iPad Pro，2024年5月，苹果举行春季发布会，发布搭载M2芯片的iPad Air 产品，以及搭载M4芯片的iPad Pro产品，同时首发M4芯片。

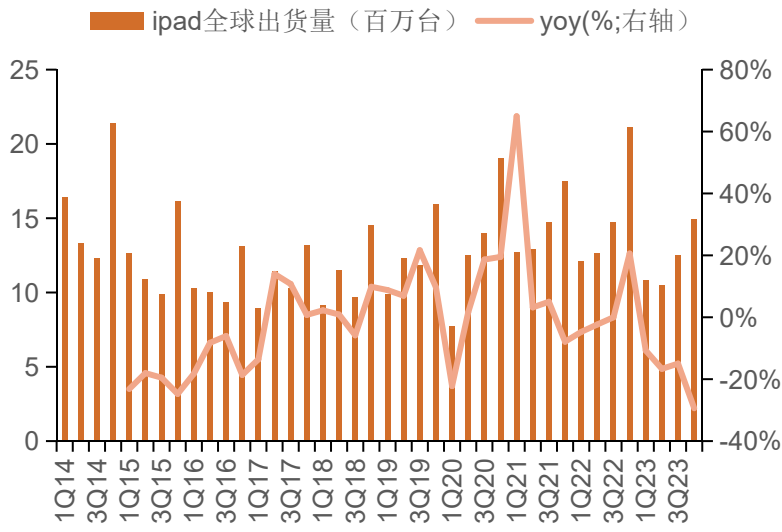
图：Apple intelligence支持的设备



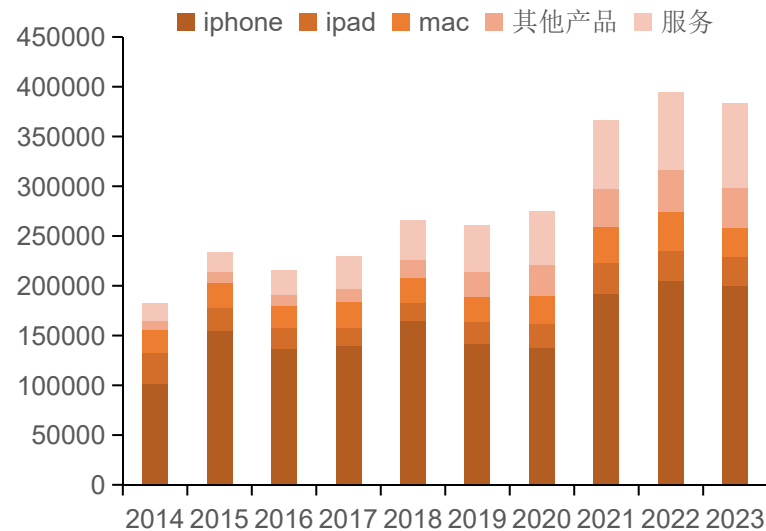
图：iphone全球出货量情况



图：ipad全球出货量情况



图：苹果营收结构情况（百万美元）



资料来源：ifind、wind、苹果官网、华源证券研究

- 1、AIPC&手机是大模型落地的重要载体，颠覆传统应用模式
- 2、硬件端：处理器架构变化，内存价值量提升
- 3、大模型端：轻量化小模型陆续推出，多模态大模型改变人机交互体验
- 4、投资逻辑：AI端侧落地加速，换机潮蓄势待发，关注PC/手机厂及硬件相关
投资机遇
- 5、风险提示



- **大模型技术发展不及预期：**目前AI技术尚未完全成熟，仍处早期快速迭代阶段，大模型技术发展不及预期可能会导致后续的应用落地存在较大不确定性。
- **PC/手机等终端需求复苏不及预期：**AI大幅提升了终端设备的使用体验，可能会使换机周期加速到来，拉动手机、电脑等终端设备销量，但对于消费者的吸引程度仍存较大不确定性，需求复苏可能不及预期。

证券分析师声明

本报告署名分析师在此声明，本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，本报告表述的所有观点均准确反映了本人对标的证券和发行人的个人看法。本人以勤勉的职业态度，专业审慎的研究方法，使用合法合规的信息，独立、客观的出具此报告，本人所得报酬的任何部分不曾与、不与，也不将会与本报告中的具体投资意见或观点有直接或间接联系。

一般声明

本报告是机密文件，仅供华源证券股份有限公司（以下简称“本公司”）的签约客户使用。本公司不会因接收人收到本报告而视其为本公司客户。本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人作出邀请。客户应对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特殊需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或使用本报告所造成的一切后果，本公司均不承担任何法律责任。

本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

本报告的版权归本公司所有，属于非公开资料。本公司对本报告保留一切权利。除非另有书面显示，否则本报告中的所有材料的版权均属本公司。未经本公司事先书面授权，本报告的任何部分均不得以任何方式修改、复制或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

本公司销售人员、交易人员以及其他专业人员可能会依据不同的假设和标准，采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论或交易观点，本公司没有就此意见及建议向报告所有接收者进行更新的义务。

华源证券股份有限公司具有中国证监会许可的证券投资咨询业务资格。

信息披露声明

在法律许可的情况下，本公司可能会持有本报告中提及公司所发行的证券并进行交易，也可能为这些公司提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。本公司将会在知晓范围内依法合规的履行信息披露义务。

销售人员信息

华东区销售代表李瑞雪lirui xue@huayuanstock.com

华北区销售代表王梓乔wangzi qiao@huayuanstock.com

华南区销售代表杨洋yang yang@huayuanstock.com

股票投资评级说明

证券的投资评级：以报告日后的6个月内，证券相对于市场基准指数的涨跌幅为标准，定义如下：

买入（Buy）：相对强于市场表现20%以上；

增持（Outperform）：相对强于市场表现5% ~ 20%；

中性(Neutral)：相对市场表现在 - 5% ~ + 5%之间波动；

减持(Underperform)：相对弱于市场表现5%以下。

行业的投资评级：以报告日后的6个月内，行业相对于市场基准指数的涨跌幅为标准，定义如下：

看好（Overweight）：行业超越整体市场表现；

中性(Neutral)：：行业与整体市场表现基本持平；

看淡(Underweight)：行业弱于整体市场表现。

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取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。

本报告采用的基准指数：沪深300指数



華源証券
HUAYUAN SECURITIES

谢 谢 聆 听