

从云到端，AI产业的新范式

——电子行业2024年中期投资策略

电子首席方竞

核心观点：AI产业的新范式

一、模型变革，云端融合为趋势。近年来生成式AI步入快速发展期，开源及垂直大模型百花齐放。而今年以来，最大的变革则是端侧模型的兴起，相较云侧模型，端侧模型在时延、隐私、个性化方面具备优势。各大厂商亦纷纷推出AI终端，**云端融合成为产业发展的下一趋势。**

二、云端变革，GB200解决互联瓶颈。当下AI产业发展和核心矛盾，是日益增长的应用+终端的需求，和算力本身增速瓶颈之间的矛盾。当前全球AI芯片竞争日趋白热化。云厂商为了降低资本开支，纷纷推出自研AI加速卡计划。而**GB200 NVL72的推出，显著拉开了英伟达与其竞争对手的差距。**英伟达通过架构的创新，解决了GPU之间互联带宽的问题，实现了最多576张卡1.8TB/s的双向互联带宽，显著领先其他竞争对手。Blackwell平台给服务器上游的光模块和PCB都带来了价值量的提升，而GB200的Rack架构的主要价值增量则在铜互连、电源等环节，尤其是铜互连的从零到一，更值得关注。

三、供应链变革：全面升级的算力产业链。自然语言的运算、图形的处理需要大量存储器协助运算及存储，并且AI服务器对数据传输速率有着更高要求，催生了存储及PCB价值量的增长。**1) 存储：**HBM是目前用于打破“内存墙”的重要技术之一。此外，DDR5渗透率提升，CXL、MCR等内容技术亦是行业发展的重心。**2) PCB：**服务器PCB要求高层数、高密度及高传输速率。AI服务器的OAM板及UBB板，最高使用到M7材料，20-30层的高阶HDI。单台AI服务器PCB价值量可达1万元。此外，交换机，电源PCB也有相应的价值量提升。为满足AI服务器的旺盛需求，上游封测及制造的工艺技术革新及产能扩张也尤为重要。**1) 封测：**算力芯片的主流封装方案为COWOS芯片+HBM显存。设备环节考量3D堆叠及测试设备的工艺能力，材料环节则看重轻薄化、高集成度以及导热性。**2) 制造：**台积电预计到2030年，半导体市场规模将达到1万亿美金，其中高性能计算（HPC）占比40%，将成为第一大细分市场。

四、端侧变革：身边的算力终端。AI云端融合趋势下，各大厂商发力AI终端，迎接下一波成长。**我们将AI终端的定义提炼为三要素：算力、模型、生态。**其中，算力和模型是AI终端的必要条件，算力是硬件基础，通过NPU执行神经网络运算，实现终端推理；而模型的减枝蒸馏，则可有效压缩模型参数，减少推理时间，降低端侧负载。上述两大要素决定了终端的硬实力。相较前者，生态需要长时间打磨，且需产业链厂商齐心协力配合，开发应用，挖掘垂直场景。生态建设代表了终端的软实力。我们认为**生态的建设和完善，会决定AI终端的成败。**而这一过程当中位居核心的是品牌厂商。苹果则得益于其多年的积累，及消费市场的号召力，更有机会打造出完备的生态。此外，端侧供应链的PCB、散热、光学等环节同样值得重视。

投资建议：标的方面，建议关注：**1) 云端：**工业富联、寒武纪、沪电股份、胜宏科技、澜起科技、沃尔核材；**2) 供应链：**中芯国际、华虹公司、ASMPT、长川科技、兴森科技；**3) 终端：**联想集团、小米集团、立讯精密、鹏鼎控股、领益智造、蓝思科技、华勤技术、水晶光电。

风险提示：下游需求不及预期；大模型等发展不及预期；晶圆厂扩产不及预期；新产品研发进展不及预期。



01

模型变革：云端融合为趋势

02

云端变革：GB200解决互联瓶颈

03

供应链变革：全面升级的算力产业链

04

端侧变革：身边的算力终端

05

投资建议

06

风险提示

CONTENTS

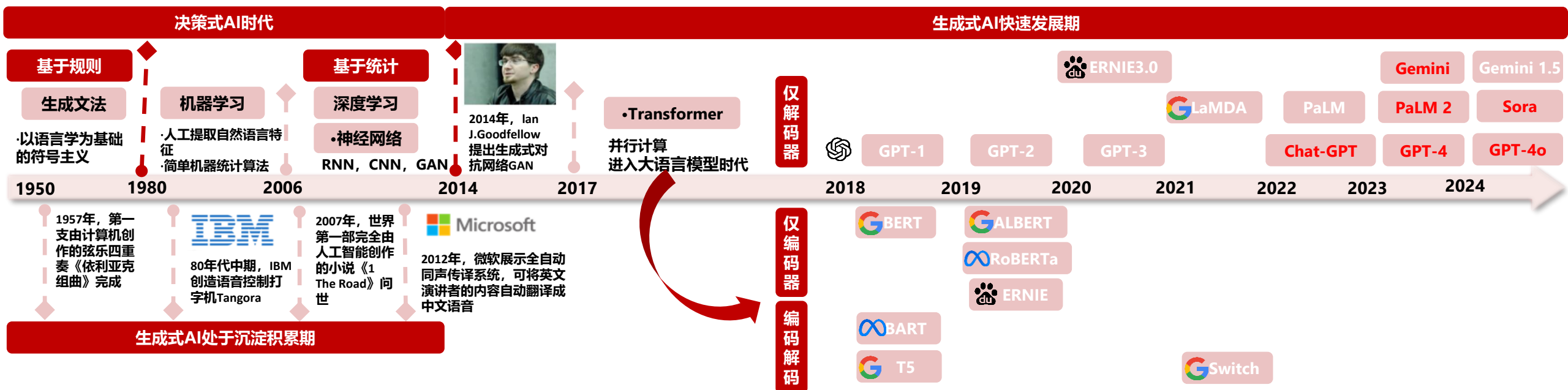
目录



01. 模型变革：云端融合为趋势

1.1 生成式人工智能：新一轮生产力革命崛起

- **近年来生成式AI步入快速发展期。**1950年开始生成式AI出现早期萌芽，此后AIGC处于漫长的沉淀积累期，决策式AI占据主流。随着2014年生成式对抗网络等深度学习算法的提出，AIGC步入快速发展期，生成内容的丰富度和精度都有了较大的提升。
- **从GPT-1到GPT-5，GPT模型的智能化程度不断提升。**ChatGPT在拥有3000亿单词的语料基础上预训练出拥有1750亿个参数的模型（GPT-2仅有15亿参数），预训练数据量从5GB增加到45TB。



1.2 大模型发展趋势：开源大模型成为重要分支

开源大模型



闭源大模型



1.3 大模型发展趋势：垂直大模型百花齐放

在AI大模型加持下给垂直领域带来革命性的新体验：

- **中科创达×PC：**中科创达的“魔方法律助手”端侧智能应用正式与联想AI PC个人智能体“联想小天”实现无缝对接。中科创达魔方法律助手是一款基于法律大模型精心打造的个人AI专属法律顾问。这款助手通过深入分析专业法律案例问答和真实案例数据库，结合法律原理和法学逻辑教材，成功构建了卓越的法言法语理解能力、法律意图识别能力、法律逻辑推理能力以及法律文本生成能力。
- **通义千问×教育：**AI教育科技公司精准学基于开源通义千问大模型训练，打造了融合文本、语音及情绪的实时多模态教育垂直模型“心流知镜”，该模型精通中小学主要教材、名师解题方法，可提供实时、低延迟的互动辅学能力，让学生学习变得轻松、高效。据悉，精准学将于6月发布旗下首个搭载该模型的AI辅学机。

图：中科创达的“魔方法律助手”



图：精准学多模态教育垂直模型“心流知镜”



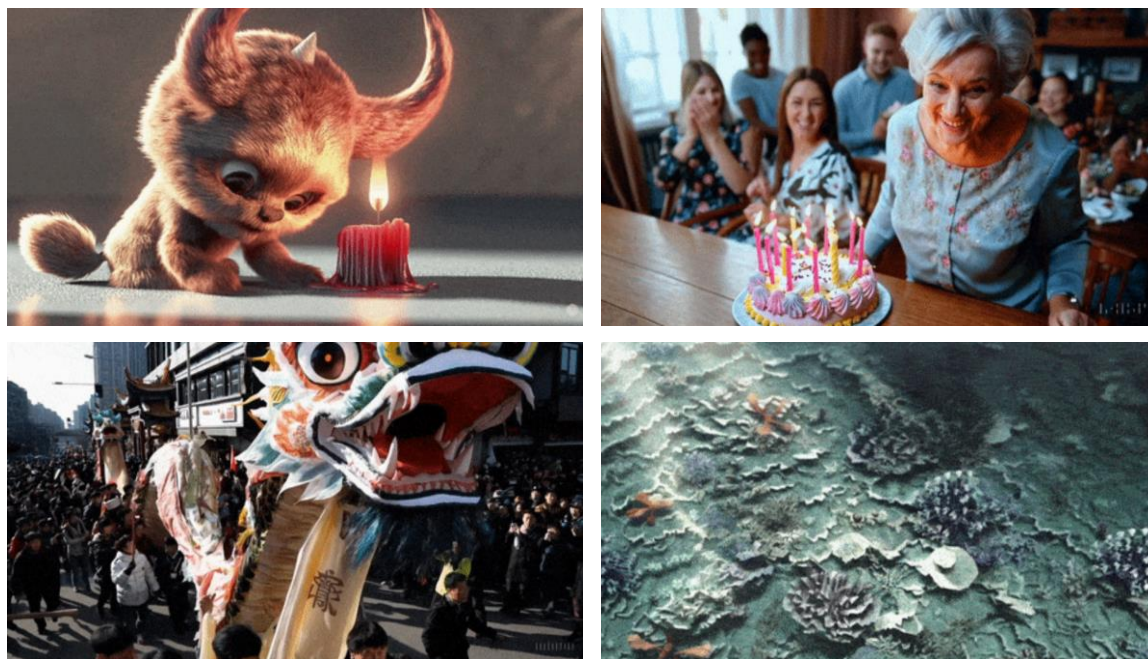
通义开源历程
已是中国最强开源大模型



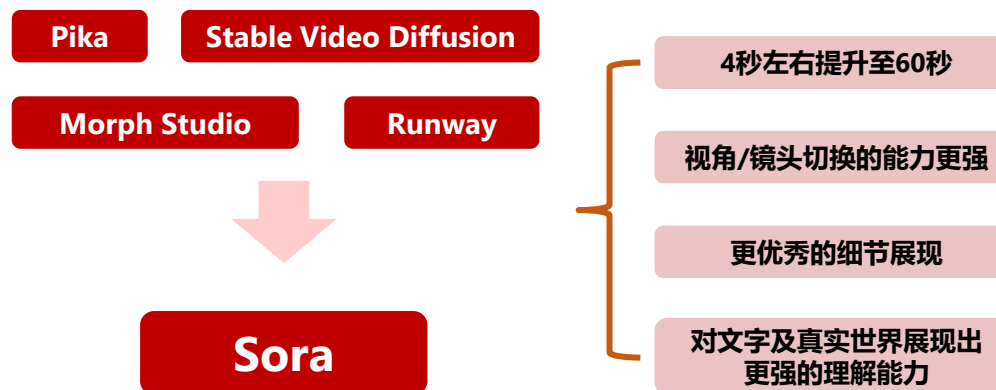
1.4 云侧：AI赋能文生视频，大幅降低视频创作成本

- **AI的视频生成能力可以为影视制作行业显著降本增效。**国内首部由文生视频创作的AI动画片《千秋诗颂》在2024年2月26日开播，AI的文生视频功能已经在B端逐步落地。
- **Sora大幅提升了AI的文生视频功能。**2024年2月16日，OpenAI推出Sora大模型，作为OpenAI的第一款文生视频大模型，其能力相较此前的文生视频模型展现出压倒性的优势。Sora强大之处表现在其能够生成最长60秒的视频，视频的视角或镜头能够更为流畅地切换，更强的细节展现能力以及对真实世界更好的理解能力。

图：文生视频案例



图：传统文生视频和Sora的能力差异



1.4 云侧：长文本分析能力打造各领域全能专家

- AI的长文本分析能力可以在短时间内阅读大量文本内容，并进行快速分析反馈。目前国内常用的长文本分析模型包括ChatGPT、同意前文、Kimi、豆包等，可以帮助使用者节省大量时间成本。
- 2024年2月，谷歌发布了Gemini 1.5模型。相较于此前Gemini 1.0Pro版本， Gemini 1.5 pro将上下文处理能提从3.2万Tokens提升至100万以上，处理能力涵盖包括1小时的视频、11小时的音频、超过3万行代码或超过70万字的代码库。

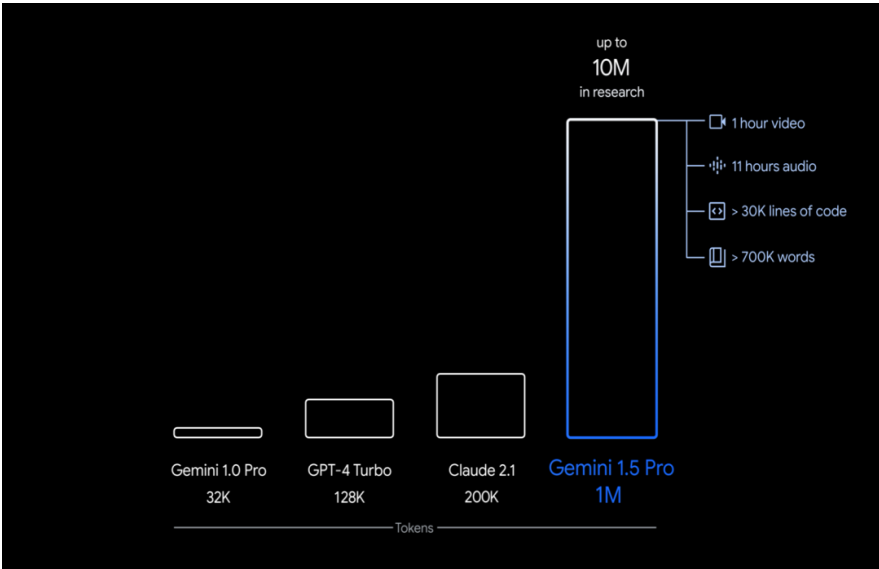
图：Kimi长文本分析案例



表：谷歌Gemini模型家族能力总结

模型	上下文窗口限制	关键功能
Gemini 1.0 Ultra	最多100万个标记	关于图像进行对话 分析、分类和总结提示中给定的大量内容- 处理高度复杂的任务,如编码、逻辑推理、遵循细微的说明以及创造性协作
Gemini 1.0 Pro	32K个标记	在长文本块中以99%的成功率找到特定信息 高水平的“上下文内学习”技能- 32K个标记的上下文窗口容量
Gemini 1.5 Pro	100万个标记以上	“上下文内学习”技能 高达100万个标记的上下文窗口容量- 分析、分类和总结提示中给定的大量内容
Gemini Nano	32K个标记	文本摘要 根据上下文的智能回复 高级校对和语法纠正 独立的无需互联网连接的功能，提高电池寿命

图： Gemini 1.5模型可以突破100万tokens上下文检索



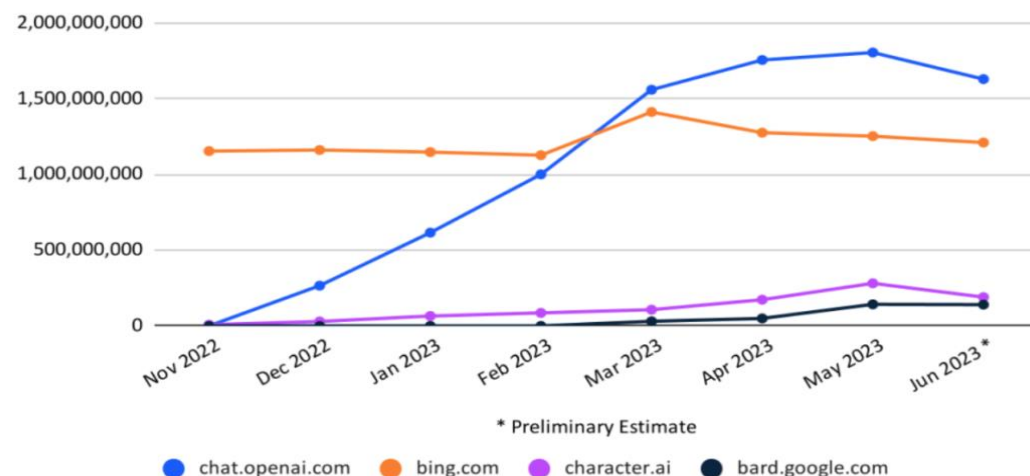
1.4 云端大模型瓶颈：云端推理受限于成本和算力瓶颈

- **云端推理成本与终端用户访问量正相关。**目前，云端是生成式 AI 模型部署的主要方式，用户可以通过网页对模型进行访问或通过 API 接口直接调用，向云端服务器发出请求，在云端收到请求后，会调用训练完成的模型对需求进行处理并返回结果。**对比训练与推理的成本，推理端成本与终端用户访问量正相关，在云端进行模型推理的成本极高，这将影响AI应用的落地及规模化扩展。**
- **云端推理受限于算力瓶颈，响应速度较慢。**随着用户数量的不断增加，ChatGPT服务器被挤爆的情况时有发生。23年11月8日晚间，ChatGPT突现重大故障，导致该应用无法使用。当时的用户页面显示“ChatGPT目前已满负荷”。此外，这一故障也影响了OpenAI的API（应用程序设计页面）服务，该服务有超过200万开发人员使用。

图：ChatGPT月访问量进展

ChatGPT and Comparisons, Worldwide

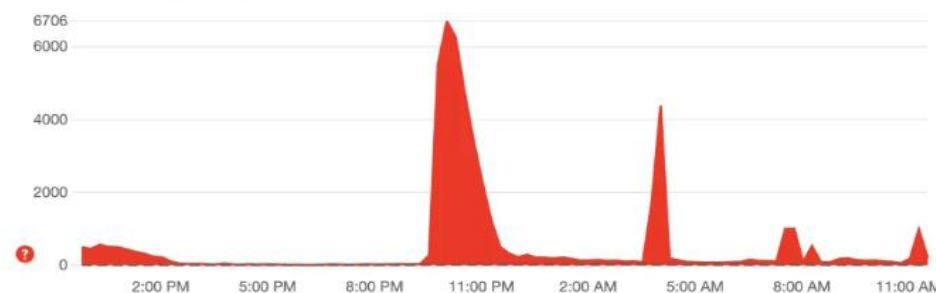
Monthly Visits Desktop & Mobile Web Worldwide



图：ChatGPT和其 API 接口出现“严重停机”

(纵轴为错误报告份数，单位：份)

OpenAI outages reported in the last 24 hours



This chart shows a view of problem reports submitted in the past 24 hours compared to the typical volume of reports by time of day. It is common for some problems to be reported throughout the day. Downdetector only reports an incident when the number of problem reports is significantly higher than the typical volume for that time of day. Visit the [Downdetector Methodology](#) page to learn more about how Downdetector collects status information and detects problems.

1.5 端侧：成本、能耗、时延、隐私安全、个性化方面具备优势

端侧模型优势

1

隐私保护

本地处理数据可以更好地保护用户隐私，数据不需要传输到云端服务器

2

定制化

利用客户个性化数据训练，定制化程度更高，贴合用户偏好

3

离线使用

本地模型允许在没有互联网连接的情况下使用AI功能，增加了应用的灵活性

4

实时响应

数据在设备上处理，可以减少到云服务器的往返延迟，提供更快响应时间



1

计算能力

云端模型通常具有更强的计算资源，能够处理更复杂的任务

2

容易更新和维护

云模型由服务提供商维护，用户无需担心软件和模型的更新问题

3

可扩展性

根据需要轻松扩展计算资源，适应不同的业务需求

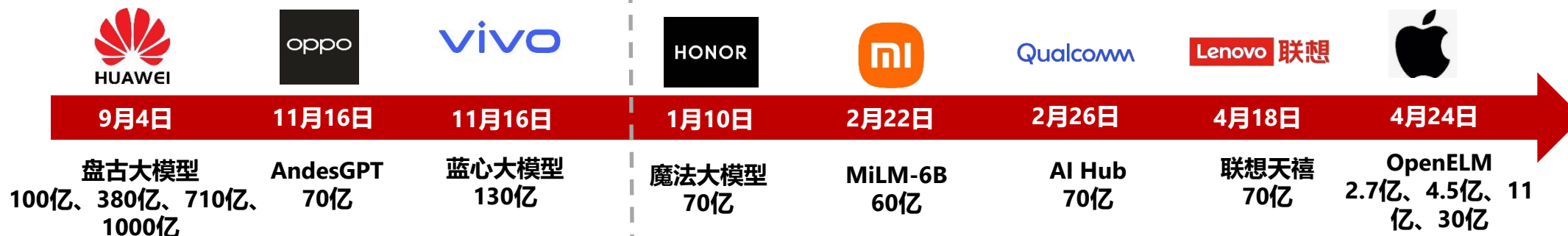
混合AI为大趋势：可结合端侧模型与云端模型的优势，在一般场景下使用端侧模型，保障数据安全和快速相应；在处理复杂问题时可灵活调用云端模型，实现端云混合

1.5 各厂商端侧AI大模型产品发布时间轴

云厂商



品牌商



2023 | 2024

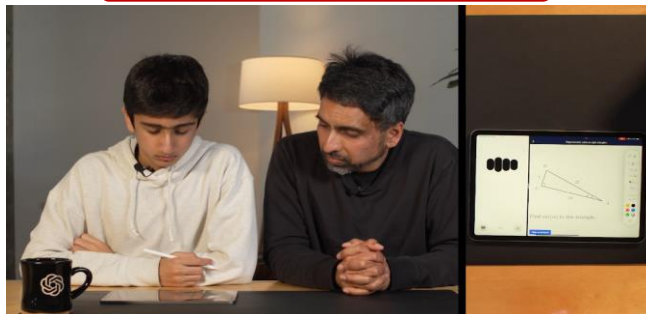
1.6 GPT-4o重磅发布，AI助手初具雏形

- 2024年5月13日，OpenAI如期举办春季发布会，推出GPT-4o。**GPT-4o的命名中的“o”代表“Omni”（全面、全能），意味着这个新模型具备强大的多模态交互能力。**GPT-4o不仅可以处理文本输入，还能接收音频和图像输入，并能生成对应的文本、音频和图像输出。GPT-4o跟此前最新版本收费不同，全球免费使用，GPT-4o API 比 GPT-4-turbo 快 2 倍，且价格便宜 50%。
- 此外，本次发布会所有的展示均通过iPhone进行，最终发布的产品也首先基于Mac OS系统，这强调了OpenAI对生态系统的重视。**我们认为，苹果作为OpenAI重要的生态合作伙伴，于2024年6月份WWDC正式宣布与OpenAI合作，智能助手Siri接入ChatGPT，AI Phone初具雏形。**后续，其他品牌也将逐步与OpenAI 合作，最后为白牌产品。

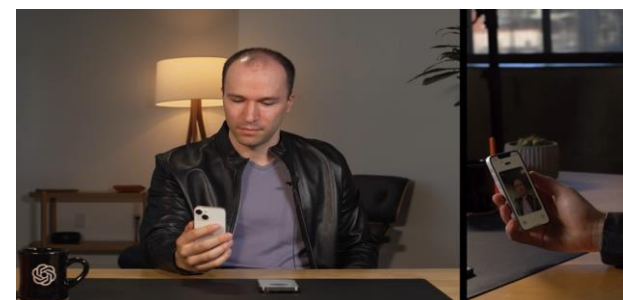
零延迟实时语音交互



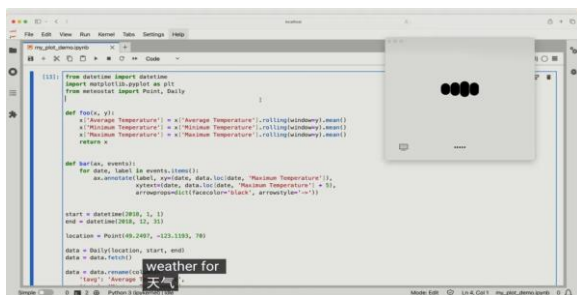
视觉传达与解析



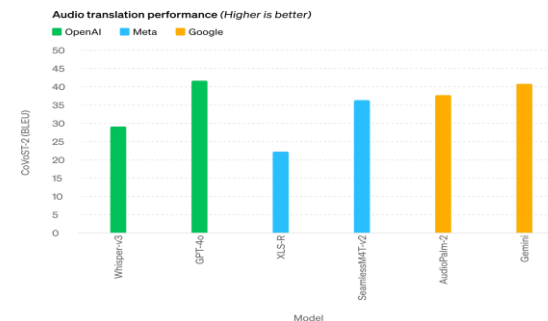
视频通话与面部情绪分析



实时编程与交互



同声传译与多语言支持

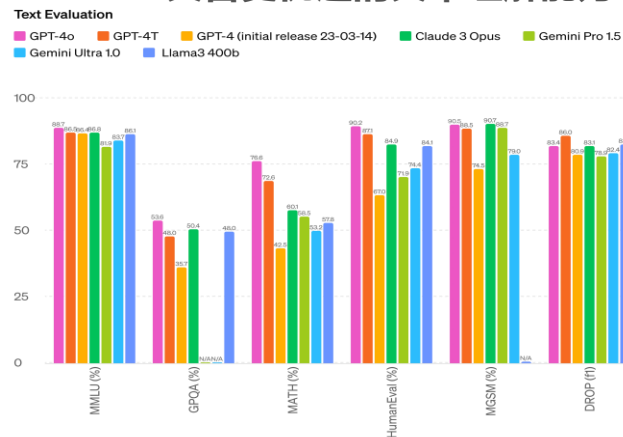


1.6

GPT-4o重磅发布，AI助手初具雏形

- GPT-4o是GPT-4模型的进化版，在基准测试中，GPT-4o在文本处理、推理和编程智能方面达到了GPT-4 Turbo级别的性能，同时在多语言、音频和视觉能力上设定了新的高标准。

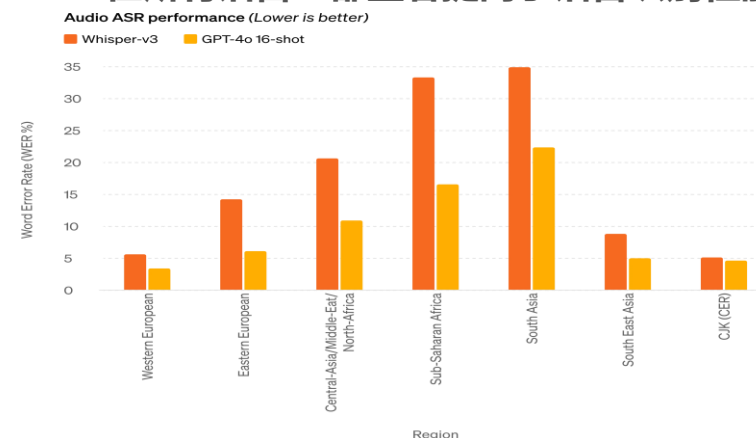
图：GPT-4o具备更优越的文本理解能力



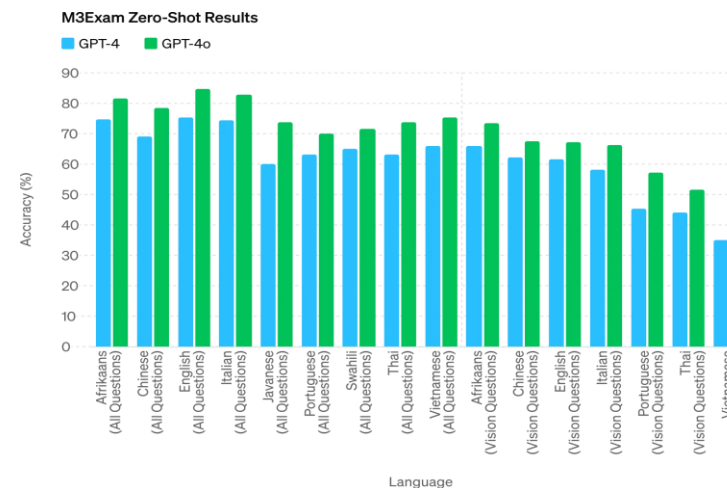
图：GPT-4o在视觉感知基准测试中具备最先进的性能

Eval Sets	GPT-4o	GPT-4T 2024-04-09	Gemini 1.0 Ultra	Gemini 1.5 Pro	Claude Opus
MMU (%) (val)	69.1	63.1	59.4	58.5	59.4
MathVista (%) (testmini)	63.8	58.1	53.0	52.1	50.5
AI2D (%) (test)	94.2	89.4	79.5	80.3	88.1
ChartQA (%) (test)	85.7	78.1	80.8	81.3	80.8
DocVQA (%) (test)	92.8	87.2	90.9	86.5	89.3
ActivityNet (%) (test)	61.9	59.5	52.2	56.7	
EgoSchema (%) (test)	72.2	63.9	61.5	63.2	

图：GPT-4o在所有语言上都显著提高了语音识别性能



图：GPT-4o在所有语言的视觉评估中比GPT-4表现更强



1.7 AI手机定位个人助手， AI PC 实现落地商用

- **必要性：**手机和PC有强大的用户基础，全球手机每年销量平均在12亿部，PC销量平均2.5-3亿台。作为个人日常生活及办公的算力终端，率先落地，具备必要性；
- **可行性：**AI 手机及PC作为目前个人强大的算力终端，应当装备能够更快、更高效地运行端侧GenAI（生成式AI）模型的处理器，NPU性能至少为30TOPS，最先进制程会率先在手机和PC落地，AI具备可行性。

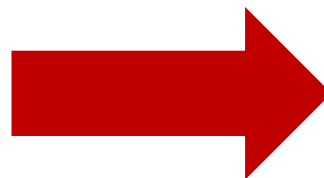
个人算力终端，率先落地



- 1) 联想发布AI PC产品，搭载端侧大模型可实现文件本地处理，宏碁、华硕、Dell等均推出AI PC新品
- 2) AMD/Intel/高通均推出嵌入专用AI模块的PC处理器
- 3) 处理器、存储、散热、电磁屏蔽、轻量化等环节迎来升级
- 4) 联想发布基于端侧模型PC个人智能体“小天”，英伟达亦发布端侧模型产品Chat with RTX



- 1) 高通推出骁龙8 Gen3，为首款专为生成式AI打造的移动平台
- 2) 以谷歌、三星、OPPO 等为代表的厂商将大模型内置于智能手机中，完成从云端 AI 向终端 AI 的转变，“AI 手机”浪潮开启
- 3) 处理器、存储、PCB、结构件等环节迎来升级
- 4) 苹果6月份WWDC2024官宣与OpenAI合作，9月有望发布首款AI手机



从IOT到汽车，AI全面赋能



IOT



MR



汽车

1.7 手机：苹果引领AI Phone产业趋势

➤ **大厂入局AI手机，端侧AI时代加速。**2023 下半年起，以谷歌、三星、OPPO 等为代表的厂商将大模型内置于智能手机中，完成从云端 AI 向终端 AI 的转变，“AI 手机”浪潮开启。2024年6月份WWDC，苹果发布个人智能化系统Apple Intelligence，同时正式宣布与OpenAI合作，智能助手Siri接入ChatGPT，苹果首款AI手机iPhone16有望于今年秋季发布。

图：全球主流端侧AI设备梳理

终端	厂商	机型	发布时间	定价	大模型	AI功能
AI手机	谷歌	Pixel 8	2023年10月4日	699美元起	Genimi Nano	照片编辑、优化影音录制、即时翻译
	三星	Galaxy S24	2024年1月18日	799美元起	云侧Gemini pro，端侧Genimi Nano	图片即圈即搜、实时双向翻译、智能图片编辑
	华为	Mate 60	2023年9月4日	5499元起	盘古大模型	AI云增强等功能
	小米	14 Ultra	2024年2月22日	6499元起	LiLM-6B	AI大模型计算摄影平台Xiaomi AISP，具有文生图、AI扩图等功能
	VIVO	X100	2023年11月16日	3999元起	蓝心大模型	AI拍照、Jovi AI助理等
	OPPO	Find X7	2024年1月8日	3999元起	小布助手	AIGC 消除、AI 通话摘要、AI 画师、小布照相馆、小布连麦等
	苹果	WWDC2024官宣与OpenAI合作，Siri接入ChatGPT，iPhone16有望成为首款AI手机				

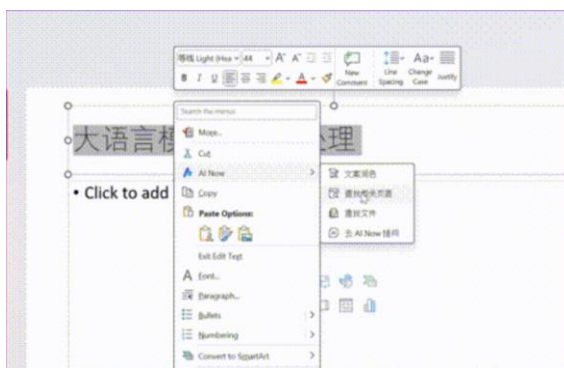
1.7 PC：微软发力模型，联想推进落地

- **24年4月18日，联想正式推出搭载端侧个人智能体“小天”的AI PC**，联想小天具备自我学习能力，可根据日常互动理解用户的使用习惯及偏好，**对于工作场景**，联想小天提供文档总结、PPT润色及生成等功能，帮助提高日常工作效率和协作能力。**对于个人生活场景**，联想小天提供AI画师等功能，使个人用户的创作更加便捷高效。
- **24年5月20日，微软发布了Windows系统最新的AI功能**，展示出对屏幕内容强大的理解和分析能力：
 - 1.通过接入GPT-4o，**可实现对屏幕内容的实时理解及处理，打造个人专属的AI Agent**
 - 2.“回溯”（Recall）功能可记忆PC上的所有操作，**该功能由40余个端侧小模型组成的模型系统同时支持运行，未来单台终端安装多个模型协同成为趋势，小模型的剪枝调优能力成为核心**

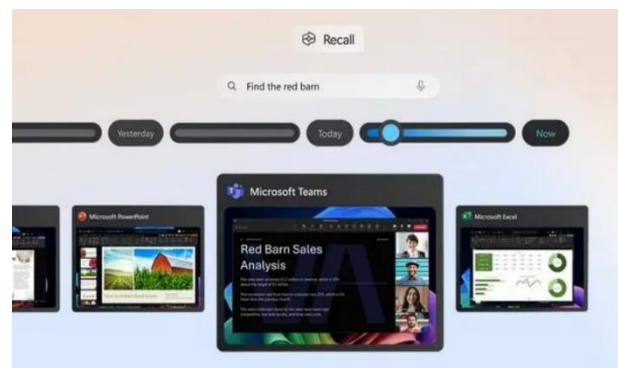
图： AI PC处理本地知识库内容



图： AI PC对PPT润色



图：“回溯功能”可记忆屏幕历史内容



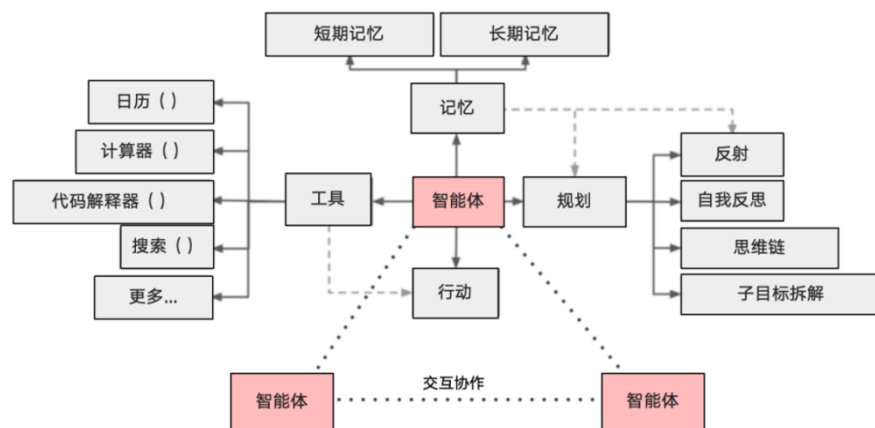
图： 通过接入GPT-4o指引玩家合成游戏道具



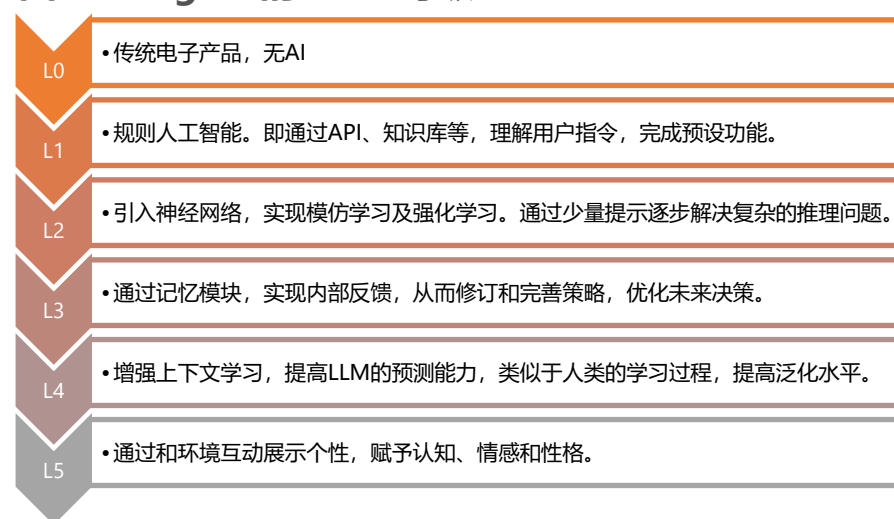
1.7 展望未来：AI终端助力Agent落地

- **Open AI为AI Agent筑基：** 2023年3月24日，ChatGPT plugins的发布，进一步扩大了ChatGPT的应用能力并催化至多场景的业务处理。2023年11月，Open AI发布GPTS，可以通过简单对话，响应用户指令，定制出满足从学习娱乐到工作分析各种需求的GPTs。距离Agent自主执行复杂任务与决策的能力大大迈进一步，**从模型层面为Agent开辟了发展路径。**
- **AI终端助力AI Agent落地：** 伴随大模型能力的日益提高，AI终端的语义理解，跨程序调用能力也持续加强。过去手机的语音助手仅可实现用户指令的简单理解和被动执行，如可询问天气信息。展望未来，AI终端将可以根据用户的使用习惯、历史行为，对用户可能需要什么服务主动进行预测，如我们在预定航班时AI Agent主动提出酒店和餐厅的推荐，并且会基于我们的价位和口味偏好。而用户在使用过程中产生的大量数据又会投喂AI进行微调训练，让其越用越懂用户。

图：基于LLM驱动的AI Agent基本框架



图：AI Agent的L0-L5等级



02. 云端变革：GB200解决 互联瓶颈

2

云端算力需求高增，GB200从架构上解决了互联瓶颈

- 前文我们讨论了模型的发展态势，而当下AI产业发展和核心矛盾，是日益增长的应用+终端的需求，和算力本身增速瓶颈之间的矛盾。
- 本章中讲探讨：
 - 1) 云厂商资本开支的展望，以及量化模型训练及推理需求。
 - AI对云厂商资本开支需求的拉动始于2023年四季度，据一季度各大云厂商的资本开支及指引，**预计2024年北美云商资本开支有望重回高速增长态势。**
 - Transformer算力需求在2年内增长750倍，平均每年以接近10倍的速度增长；英伟达平均每2年左右推出一代加速卡，从A系列到B系列的升级节奏来看，每一代产品算力提升幅度在3倍左右，价格提升幅度比算力提升幅度略低。**综合大模型参数增长和算力的单位价格来看，过去5年大模型训练对资本开支的需求持续快速提升，未来大模型参数的持续提升仍将带动算力需求激增。**
 - 2) 加速卡的竞争格局，GB200等产品给英伟达带来的优势提升，GB200服务器及产业链弹性。
 - **当前全球AI芯片仍主要被英伟达垄断，但竞争逐步白热化。**云厂商为了降低资本开支，纷纷推出自研AI加速卡计划，而AMD等芯片设计公司，也在不断推出更具性价比的产品，以求抢占英伟达的市场份额。
 - **GB200 NVL72显著拉开了英伟达与其竞争对手的差距。**英伟达通过架构的创新，解决了GPU之间互联带宽的问题，实现了最多576张卡1.8TB/s的双向互联带宽，显著领先其他竞争对手，使得GB200 NVL72的推理性能达到H100的30倍。
 - **从GB200产业链弹性来看，铜互连是核心增量。**Blackwell平台给服务器上游的光模块和PCB都带来了价值量的提升，而GB200的Rack架构的主要价值增量则在铜互连、电源等环节，尤其是铜互连的从零到一，更值得关注。

2.1



01

云商资本开支持续上修

02

AI三要素：算力、互联、存储

03

加速卡竞争格局日益激烈

04

GB200解决互联瓶颈，重视产业链弹性

05

AI云端产业链

CONTENTS

目录

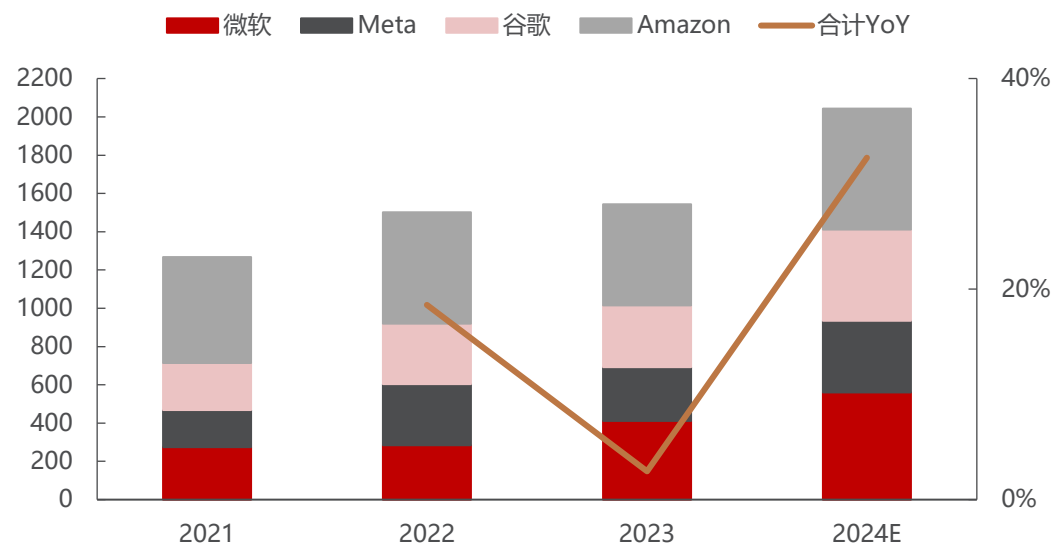


2.1

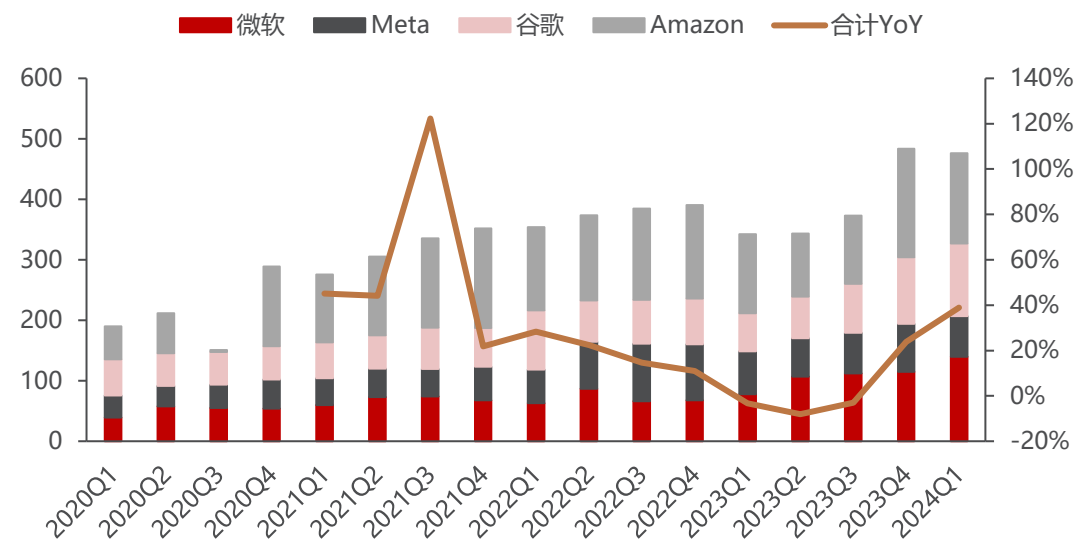
算力需求激增，云商资本开支持续增长

- 北美云商资本开支在2023年有所下滑，但伴随AI的强势增长，2024年北美云商资本开支重回高速增长态势。我们复盘了北美云厂商过去4年的Capex走势，伴随AI需求的强势增长，结合一季度各大云厂商的资本开支及指引，预计2024年北美云商资本开支有望重回高速增长态势。
- AI对云厂商资本开支需求的拉动始于2023年四季度，据Bloomberg，1Q24北美四大云厂商合计资本开支为476亿美元，同比增长39%，继4Q23云商资本开支同比转正后，同比增速进一步提升。展望2024年，多家云厂商分别上修全年资本开支指引，云厂商Capex有望重回高增。Bloomberg一致预期2024年四大北美云商资本开支合计将达到2044亿美元，同比增长32%。

图：北美四大云厂商年度资本开支及增速（亿美元，%）



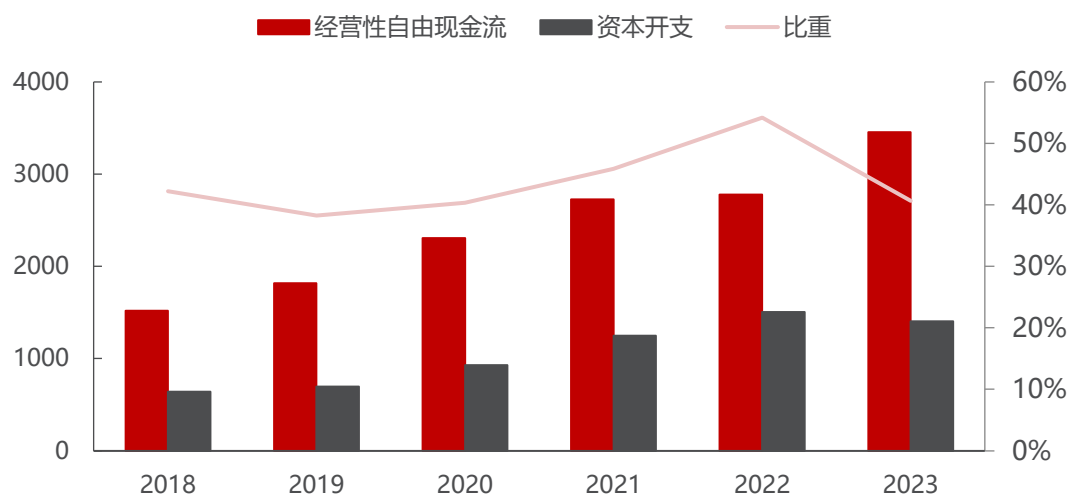
图：2020-2024年一季度北美云商资本开支（含融资租赁）（亿美元，%）



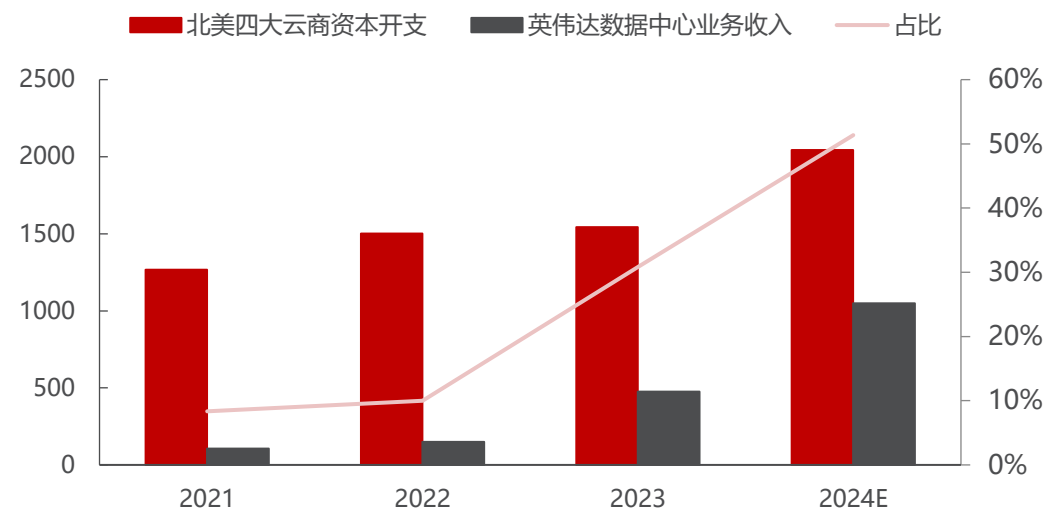
2.1 展望未来，全球AI资本开支仍有提升空间

- 全球AI资本开支的提升空间来自三个方面：1) 云商自身资本开支提升；2) AI占云商资本开支的比例提升；3) AI资本开支从云商向外扩散。
- **第一，云商自身资本开支提升：**2022年全球云厂商资本开支峰值占经营性自由现金流比例为54%，2023年仅为41%，云厂商资本开支占自由现金流的比重仍有提升空间。
- **第二，AI占云商资本开支的比例提升：**我们用英伟达的数据中心业务收入和北美四大云厂商资本开支的比例来大致测算AI占云商资本开支的比重，2021年英伟达数据中心业务收入占北美四大云商资本开支的比重为8.37%，而这一比例在2024年预计将达到51.35%。
- **第三，资本开支从云厂商向外扩散：**据Trendforce，**2022年北美四大云厂商采购的AI服务器占全球AI服务器出货量的66%**。当前教育、金融、IDC、政府等部门AI需求持续上升，服务器采购比重有望提升。2023-2024年为云商建设数据中心的主要年份，而伴随着AI应用的落地，更多场景的用户加入到采购AI服务器的群体中来，AI资本开支有望向外持续扩散。

图：云厂商资本开支占自由现金流的比例（亿元，%）



图：AI占云厂商资本开支的比例测算（亿元，%）



2.1

AI所需算力与模型参数正相关

- 在生成式AI场景下，模型训练和推理所需的算力与参数大小成正比。1) 训练所需的算力和模型参数以及训练集大小（Token）数量成正比；2) 推理所需的算力和模型参数，回答大小，以及访问量成正比关系。
- 当前大模型参数仍以较快速度增长，后续将推出的GPT-5模型，参数有望达到十万亿量级，有望引起大模型以及云厂商在算力领域的新一轮竞赛，拉动下游算力需求。

图：大模型训练成本公式

$$\text{训练成本} = \frac{\text{模型参数} \times \text{Token数} \times \text{每Token单参数所需浮点运算次数} \times \text{单位时间训练成本}}{\text{芯片算力} \times \text{算力利用率}}$$

图：大模型推理成本公式

$$\text{推理成本} = \frac{\text{推理次数} \times \text{模型参数} \times \text{回答Token数} \times \text{峰值流量倍数} \times \text{每Token单参数所需INT运算次数} \times \text{单位时间训练成本}}{\text{芯片算力} \times \text{算力利用率}}$$

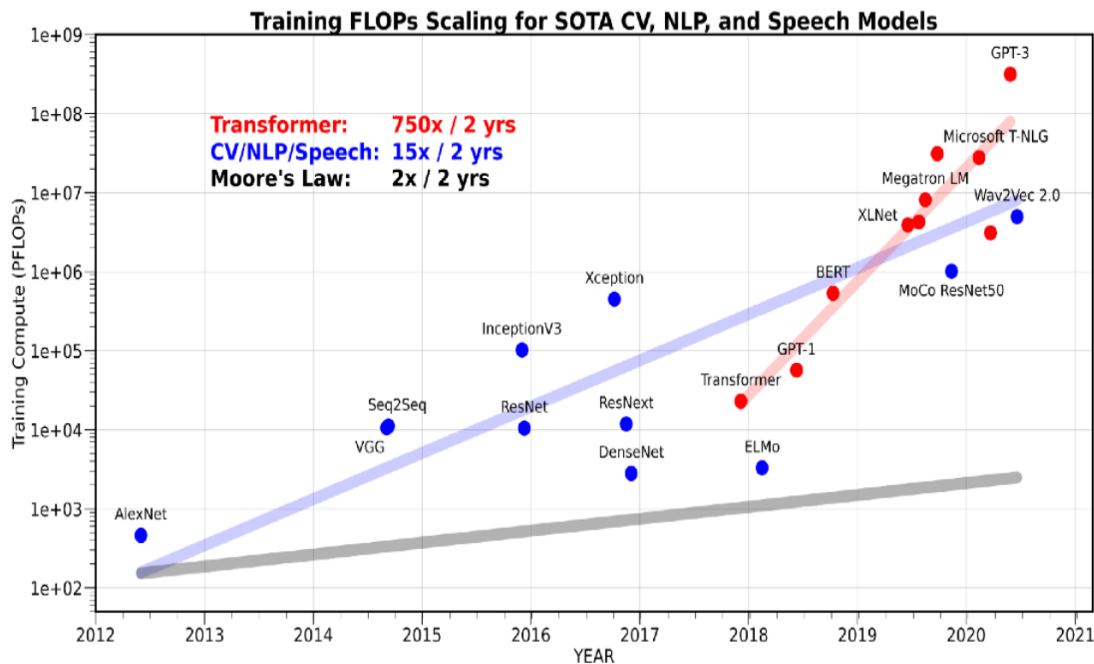
资料来源：academic accelerator，Sunyan’ s Substack，英伟达官网，Azure官网，民生证券研究院测算

2.1

模型参数指数级增长，持续带动AI算力需求

- AI大模型的算力需求在过去几年呈现快速增长的态势，Transformer算力需求在2年内增长750倍，**平均每年以接近10倍的速度增长**。以Open AI的GPT为例，GPT 1在2018年推出，参数量级为1亿个，Open AI下一代推出的GPT 5参数量级预计达到10万亿。
- 英伟达平均每2年左右推出一代加速卡，从A系列到B系列的升级节奏来看，每一代产品算力提升幅度在3倍左右，价格提升幅度比算力提升幅度略低。
- **综合大模型的参数增长和算力的单位价格来看，过去5年大模型训练对资本开支的需求持续快速提升**。当下大模型的参数仍然在不断提升的过程中，对算力的资本开支需求有望维持快速增长。

AI大模型对算力的需求超过摩尔定律



NV每两年推出一代训练GPU

型号	发布时间	峰值算力TOPS/TFLOPS					
		FP4	INT8	BF16/FP16	TF32	FP32	FP64
A100	2020	-	624	312	156	19	9
H100	2022	-	2000	1000	500	67	34
H200	2023	-	2000	1000	500	67	34
GH200	2022	-	2000	1000	500	67	34
B100	2023	7000	3500	1800	900	60	30
B200	2024	9000	4500	2250	1120	80	40
GB200	2024	20000	10000	5000	2500	190	90

注：表格中展示非稀疏算力

2.2



01

云商资本开支持续上修

02

AI三要素：算力、互联、存储

03

加速卡竞争格局日益激烈

04

GB200解决互联瓶颈，重视产业链弹性

05

AI云端产业链

CONTENTS

目录



2.2

AI三要素：算力、互联、存储决定加速卡的硬件性能

- 算力、片间互联、存储容量和带宽作为AI加速卡三要素，共同决定了加速卡的硬件性能。
- 算力是加速卡的核心性能指标。AI芯片算力根据精度有所差异，一般神经网络的标准训练精度是32位浮点数，但有时为了节省时间和资源，也可以使用16位浮点数进行训练，推理时对算力精度的要求相对较低，而对功耗、推理成本、响应速度等要求较高，通常采用INT8算力。
- 由于大模型的训练和推理需要用到多张加速卡，加速卡和加速卡之间的通信效率会成为训练和推理效率的瓶颈，片间互联带宽越大，则加速卡在训练和推理过程中的通信时延越低，计算效率越高。
- HBM显存容量和带宽决定了加速卡内存能够存储的信息量，以及GPU和系统内存之间的数据传输速率，如果显存带宽不足，则系统会出现瓶颈，GPU运算核心可能处于空闲状态。

AI算力精度和对应的下游场景

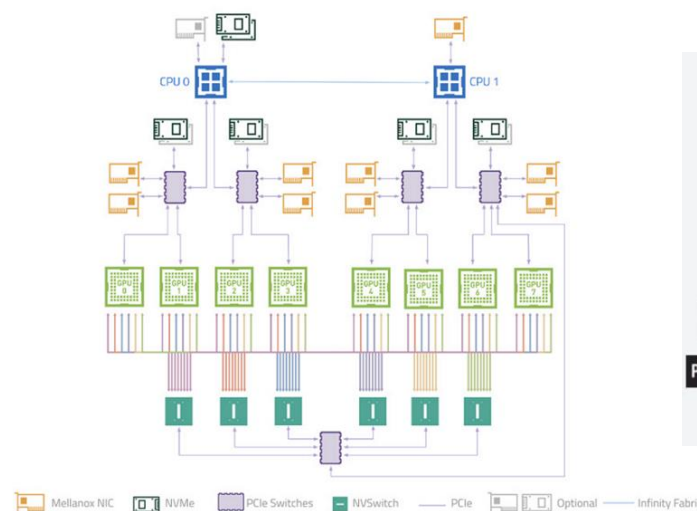
精度	定义
FP64	双精度浮点运算，精度为64位，通常用于HPC
FP32	单精度浮点数，精度为32位，是神经网络计算的标准类型
FP16	半精度浮点数，精度为16位，在对训练速度有需求的情况下也可用于训练
BF16	对FP32截断数据，8bit表示指数，7bit表示小数，和FP32精度相同但吞吐量更高
TF32	用Tensor Core运算，默认使用矩阵乘法，精度和FP32相同，但运算速度更快
TF16	用Tensor Core运算，精度为16位
INT8	用CUDA Core中的INT单元计算，8位整数，通常用于推理
FP4	英伟达B系列加速卡支持4位精度浮点数，用于推理

高性能计算

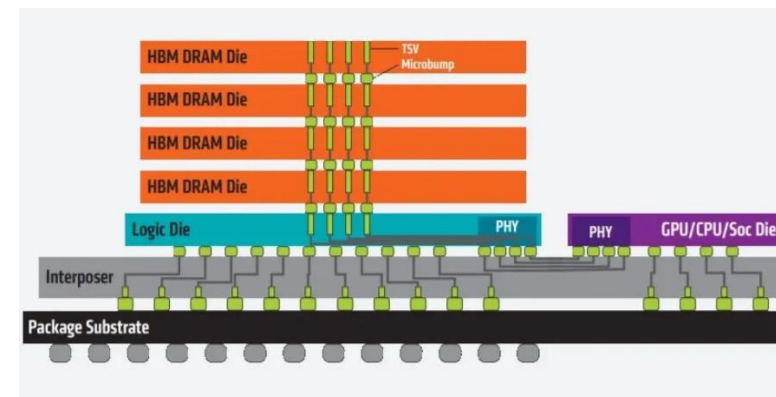
AI训练

AI推理

英伟达服务器八张GPU的片间互联方案



加速卡封装方案中采用多层HBM



2.2

算力：决定AI芯片性能的核心指标

- 算力是决定加速卡性能的核心指标。
- 英伟达的加速卡在算力方面仍处于行业领先地位，最新一代Blackwell平台加速卡不仅在8~32位推理算力中显著超过竞争对手，还额外提供了FP4算力用于低精度的推理场景。
- 其他厂商来看，AMD在算力方面与英伟达最为接近，并且在FP64算力上超越英伟达最新一代的产品，更适合用于科学计算的场景。而目前其他的互联网公司自研加速卡与英伟达仍有较大差距。

英伟达及其竞争对手的加速卡算力对比

厂商	大类	型号	峰值算力TOPS/TFLOPS					
			INT4/PF4 Dense/Sparse	INT8/FP8 Dense/Sparse	BF16/FP16 Dense/Sparse	TF32 Dense/Sparse	FP32	FP64
NVIDIA	训练	H100	-	4000/2000	2000/1000	1000/500	67	34
		H200	-	4000/2000	2000/1000	1000/500	67	34
		H20	-	296	148	74	44	1
		GH200	-	4000/2000	2000/1000	1000/500	67	34
		B100	7000/14000	3500/7000	1800/3500	900/1800	60	30
		B200	9000/18000	4500/9000	2250/4500	1120/2250	80	40
		GB200	20000/40000	10000/20000	5000/10000	2500/5000	190	90
	推理	L40S	733/1466	733/1466	362/733	183/366	92	-
		L40	724/1488	362/724	181/362	91/181	91	-
		L20	-	239	119.5	60	60	-
		L4	-	485	242	120	30	-
		L2	-	193	96.5	48	24	-
		L2	-	193	96.5	48	24	-
AMD	训练	MI300X	-	2614.9	1307.4	654	163	163
Intel	训练	Gaudi 3	-	1835	1835	-	-	-
谷歌	训练	TPUv5E	-	394	197	-	-	-
		TPUv5P	-	918	459	-	-	-
Meta	推理	MTIA v2	-	354/708	177/354	2.76	2.76	-
微软	训练	Maia 100	3200	1600	800	-	-	-
亚马逊	训练	Trainium2	-	861	431	215	215	-
TESLA	训练	D1	-	376	376	22.6	22.6	-

2.2

片间互联：加速卡间的通信速率会影响运算效率

- 英伟达在片间互联方面始终维持领先地位。英伟达的NV Link和NV Switch保持着2年一代的升级节奏，目前NV Link带宽已达到1.8TB/s的双向互联，在市面上处于领先地位。
- 竞争对手的情况来看，AMD和谷歌的片间互联带宽分别达到了896GB/s和600GB/s，与英伟达的上一代H系列产品较为接近，而其他云厂商自研加速卡大多采用PCIe通信协议，在片间互联能力方面与英伟达仍有较大的差距。

英伟达NV Link和NV Switch参数

	第二代	第三代	第四代	第五代
NVLink 总带宽	300GB/s	600GB/s	900GB/s	1,800GB/s
每个 GPU 最大链路数	6	12	18	18
NVIDIA 架构支持	NVIDIA Volta™ 架构	NVIDIA Ampere 架构	NVIDIA Hopper™ 架构	NVIDIA Blackwell 架构
	第一代	第二代	第三代	NVLink Switch
一个 NVLink 域内直连 GPU 的数量	最多 8 个	最多 8 个	最多 8 个	最多 576 个
NVSwitch GPU 之间带宽	300GB/s	600GB/s	900GB/s	1,800GB/s
聚合总带宽	2.4TB/s	4.8TB/s	7.2TB/s	1PB/s
NVIDIA 架构支持	NVIDIA Volta™ 架构	NVIDIA Ampere 架构	NVIDIA Hopper™ 架构	NVIDIA Blackwell 架构

英伟达及其竞争对手的加速卡片间互联带宽对比

厂商	大类	型号	互联带宽GB/s
NVIDIA	训练	H100	NVLink 900
		H200	NVLink 900
		H20	NVLink 900
		GH200	NVLink 1800
		B100	NVLink 1800
		B200	NVLink 1800
		GB200	NVLink 3600
	推理	L40S	PCIe 64
		L40	PCIe 64
		L20	PCIe 64
		L4	PCIe 64
		L2	PCIe 64
AMD	训练	MI300X	Infinity Fabric Links 896
Intel	训练	Gaudi 3	PCIe 128
谷歌	训练	TPUv5E	200
		TPUv5P	600
Meta	推理	MTIA v2	PCIe 128
微软	训练	Maia 100	PCIe 128
TESLA	训练	D1	PCIe 64



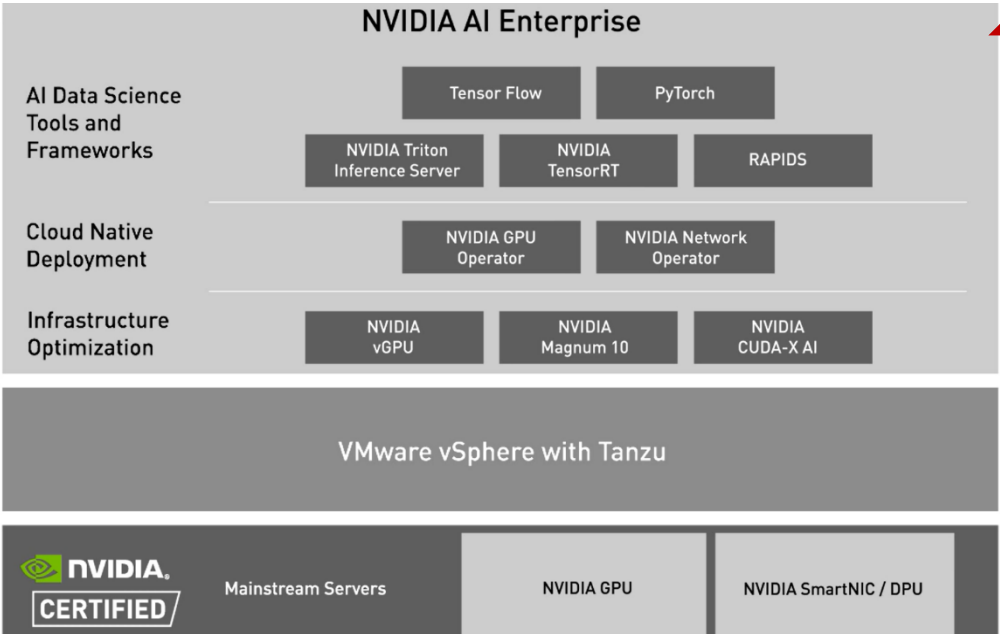
2.2

软件生态是衡量加速卡厂商能力更重要的因素

GPU的软件生态核心包括两个部分：

- **深度学习框架：**一套对深度学习算法进行模块化封装的开发工具。英伟达基于CUDA技术在2014年推出了深度学习库cuDNN，同时英伟达的GPU也支持基于TensorFlow、PyTorch的TensorRT等深度学习库。
- **计算架构：**异构编程语言，是实现异构计算单元（CPU/GPU/DSP/ASIC/FPGA等）在同一个计算框架下的关键。目前主流的GPU编程语言包括英伟达开发的CUDA、AMD开发的ROCm以及开源的OpenCL。其中CUDA和OpenCL分别主要被英伟达和AMD采用。
- 由于目前CUDA是最主流的深度学习架构，因此市面上大部分GPU相关软件都是用CUDA开发，**兼容CUDA可以大幅降低国产GPU的开发难度，提升产品可用度，增加产品稳定性，而部分厂商则选择推广自己的软件生态，如寒武纪、百度等。**

图：英伟达人工智能企业软件层



深度学习框架

K8S集群运算符

计算架构

容器平台

GPU、网卡等硬件

框架名称	维护方	支持语言
TensorFlow	Google	C++/Python/Java/R等
Keras	Google	Python/R
Caffe	BVLC	C++/Python/Matlab
PyTorch	Facebook	C/C++/Python
Theano	UdeM	Python
CNTK	Microsoft	C++/Python/C#/.NET/Java/R
MXNet	DMLC	C++/Python/R等
PaddlePaddle	百度	C++/Python
Deeplearning4j	Eclipse	Java/Scala等
ONNX	Microsoft/Facebook	Python/R

编程语言	厂商
OpenCL	苹果等
CUDA	英伟达
ROCm	AMD

2.3



01

云商资本开支持续上修

02

AI三要素：算力、互联、存储

03

加速卡竞争格局日益激烈

04

GB200解决互联瓶颈，重视产业链弹性

05

AI云端产业链

CONTENTS

目录



2.3 加速卡竞争格局：英伟达和他的挑战者们

- **当前全球AI芯片仍主要被英伟达垄断，但竞争逐步白热化。**据TechInsights，2023年英伟达占全球数据中心GPU出货量份额的98%，处于垄断地位。同时，AMD、谷歌、特斯拉等厂商相继推出MI300系列、TPU V5以及Dojo D1等产品，挑战英伟达的垄断地位，尽管2024年4月英伟达推出的B系列加速卡再一次在算力方面与竞争对手甩开差距，但当前全球加速卡市场竞争者不断增加已成定局。
- **国内市场方面**，伴随美国禁令趋严，2023年10月17日美国商务部公布算力芯片出口管制新规，A100、H100、A800、H800、L40、L40S等芯片进入管制名单，同时国内昇腾、寒武纪等龙头厂商产品能力不断追赶海外龙头，**AI芯片国产化成为大趋势。**

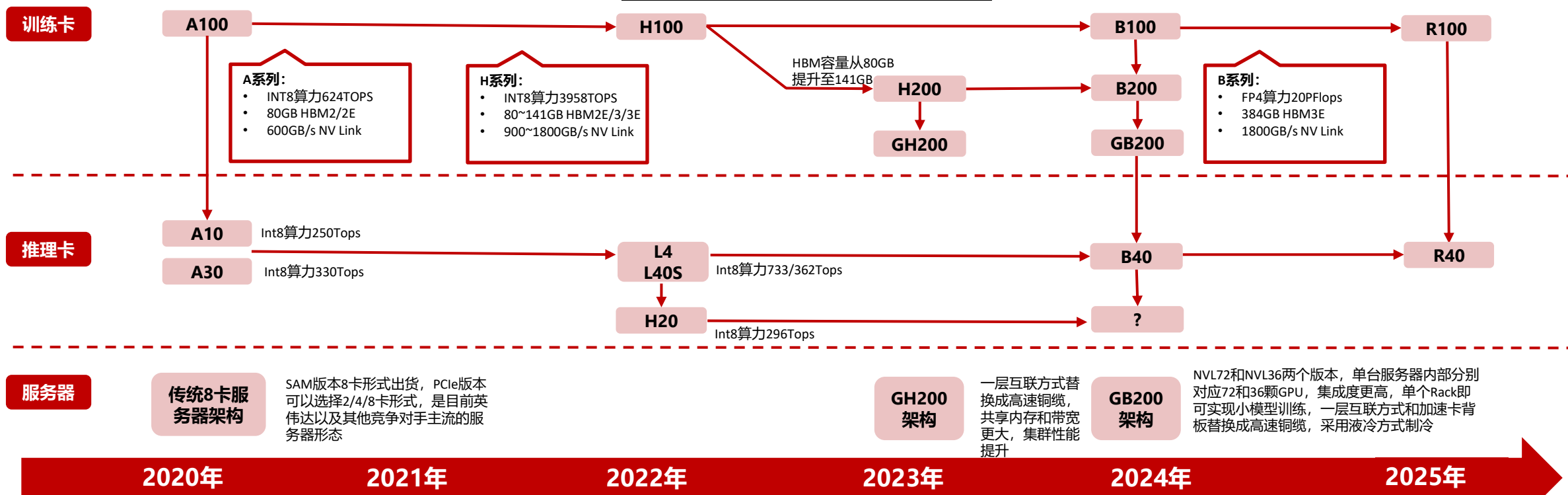
英伟达及其竞争对手的产品情况梳理

全球龙头			英伟达的追赶者			云厂商自研加速卡		
英伟达			AMD	MI300A	MI300X	Google	TPU v5E	TPU v5P
H100	H200	GH200	昇腾	910B	910C	AWS	Trainium	Graviton
B100	B200	GB200				Microsoft	Maia	
L40	L40S	L4				Meta	MITA	
L20	H20	L2	Intel	Gaudi		Tesla	Dojo	
			其他玩家	寒武纪	海光	腾讯	燧原	
					景嘉微	阿里巴巴	平头哥	
					沐曦等一级市场公司	百度	昆仑芯	

2.3 英伟达的加速卡及服务器产品线梳理

- 英伟达具有在训练侧和推理侧完善的产品布局，自2020年以来，平均每2年推出一代产品，每一代产品在算力、互联带宽等指标方面的提升幅度均在一倍左右。
- 训练卡方面，除了英伟达常规的H100、B100等加速卡外，H系列以后，英伟达还增加了H200、B200等产品，从而给客户提供更多的产品选择和更好的性价比，例如H200加速卡在HBM容量方面相较于H100均有较大提升，但价格方面提升幅度相对较少。
- 推理卡方面，2024年英伟达的产品出货以L40和L40S为主，并且推出了L20，L2，L4等产品供客户选择。

英伟达加速卡及服务器产品线梳理



2.3 英伟达的加速卡产品线梳理

- 英伟达具有在训练侧和推理侧完善的产品布局，自2020年以来，平均每2年推出一代产品，每一代产品在算力、互联带宽等指标方面的提升幅度均在一倍左右。
- 训练卡方面，除了英伟达常规的H100、B100等加速卡外，H系列以后，英伟达还增加了H200、B200等产品，从而给客户提供更多的产品选择和更好的性价比，例如H200加速卡在HBM容量方面相较于H100均有较大提升，但价格方面提升幅度相对较少。
- 推理卡方面，2024年英伟达的产品出货以L40和L40S为主，并且推出了L20，L2，L4等产品供客户选择。

英伟达主流加速卡产品参数汇总

厂商	大类	型号	发布时间	参数信息		峰值算力TOPS/TFLOPS						内存信息			互联带宽GB/s
				制程nm	功耗W	INT4/PF4 Dense/Sparse	INT8/FP8 Dense/Sparse	BF16/FP16 Dense/Sparse	TF32 Dense/Sparse	FP32	FP64	类型	容量GB	带宽GB/s	
NVIDIA	训练	H100	2022	4	700	-	4000/2000	2000/1000	1000/500	67	34	HBM3	80	3350	NVLink 900
		H200	2023	4	700	-	4000/2000	2000/1000	1000/500	67	34	HBM3E	141	4800	NVLink 900
		H20	2023	4	400	-	296	148	74	44	1	HBM3	96	4000	NVLink 900
		GH200	2023	4	700	-	4000/2000	2000/1000	1000/500	67	34	HBM3E	141	5000	NVLink 1800
		B100	2024	4	700	9000/18000	3500/7000	1800/3500	900/1800	60	30	HBM3E	192	8000	NVLink 1800
		B200	2024	4	1000	7000/14000	4500/9000	2250/4500	1120/2250	80	40	HBM3E	192	8000	NVLink 1800
		GB200	2024	4	最高2700	20000/40000	10000/20000	5000/10000	2500/5000	190	90	HBM3E	384	16000	NVLink 3600
	推理	L40S	2023	-	350	733/1466	733/1466	362/733	183/366	92	-	GDDR6	48	864	PCIe 64
		L40	2023	-	300	724/1488	362/724	181/362	91/181	91	-	GDDR6	48	864	PCIe 64
		L20	2023	5	275	-	239	119.5	60	60	-	GDDR6	48	864	PCIe 64
		L4	2023	-	72	-	485	242	120	30	-	GDDR6	24	300	PCIe 64
		L2	2023	-	-	-	193	96.5	48	24	-	GDDR6	24	300	PCIe 64

2.3 下游云厂商纷纷推出自研AI芯片

- 英伟达2024财年数据中心业务收入达到475亿美元，预计2025财年仍将以翻倍态势增长，云厂商作为英伟达的主要数据中心客户，纷纷推出自研AI加速卡计划，以求降低在数据中心领域的资本开支。
- 云厂商中，加速卡发展最快的公司为谷歌，谷歌研发TPU的时间始于2013年，采取自研加速卡为主，同时采购部分英伟达加速卡的策略。由于谷歌在加速卡领域布局早，产品完善度高，谷歌是北美四大云厂商中采购英伟达加速卡最少的厂商。2023年12月，谷歌推出面向云端的AI加速卡TPU v5p，相较于TPU V4，TPU v5p提供了二倍的浮点运算能力和三倍内存带宽提升，芯片间的互联带宽最高可以达到600GB/s。
- 其他云厂商也纷纷推出自研加速卡计划。1) Meta：2023年，Meta宣布自研MTIA v1芯片，2024年4月，Meta发布最新版本MTIA v2加速卡；2) 微软：微软Azure的企业数量已经达到25万家，微软的Maia 100在2023年推出，专为Azure云服务设计；3) 亚马逊：2023年，亚马逊推出了用于训练的Trainium2加速卡，以及用于推理的Graviton4加速卡，目前亚马逊在训练和推理卡均有布局。

英伟达客户在加速卡领域的布局情况

厂商	大类	型号	发布时间	制程nm	峰值算力TOPS/TFLOPS			内存信息		互联带宽GB/s
					INT8/FP8 Dense/Sparse	BF16/FP16 Dense/Sparse	TF32/FP32 Dense/Sparse	类型	容量GB	
谷歌	训练	TPUv5E	2023	-	394	197	-	HBM2	16	200
		TPUv5P	2023	-	918	459	-	HBM2	95	600
Meta	推理	MTIA v2	2024	5	354/708	177/354	2.76	-	128	128
微软	训练	Maia 100	2023	5	1600	800	-	HBM3	64	128
亚马逊	训练	Trainium2	2023	4	861	431	215	-	-	-
	推理	Graviton4	2023	-	-	-	-	GDDR5	-	-
TESLA	训练	D1	2021	7	376	376	22.6	-	32	64

2.3

H200和Blackwell GPU与竞争对手拉开差距

- **AMD等厂商推出的MI300等强有力的产品，对英伟达的市场份额产生了一定的争夺，但英伟达很快发布了H200、Blackwell GPU等系列的产品，与其竞争对手又拉开了一个身位。**以AMD的MI300X为例，这颗加速卡在Int8、FP16、FP32算力方面均为H100的1.3倍，互联带宽方面达到了接近于NV Link4.0的896GB/s双向互联，FP64算力和HBM容量更是达到了H100的2倍以上，一系列的堆料和价格优势使得下游云厂商考虑转用一部分AMD的产品。
- **英伟达推出的H200很大程度上是对竞争对手的回应。**相较于H100，H200将此前的HBM3提升为HBM3E，同时将HBM容量从上一代的80GB提升至141GB。在价格方面，H200相较于上一代产品体现出极强的性价比，该款加速卡发售后预计将受到下游客户的欢迎。**英伟达的下一代Blackwell GPU系列产品，在算力、内存和互联带宽的AI三要素领域与竞争对手的差距进一步拉开，巩固了英伟达的领先地位。**

B100、B200、H100、H200、MI300X参数对比

	B100	B200	H100	H200	MI300X
FP4算力 (TFlops)	7000	9000	-	-	-
FP8/INT8算力 (TFlops)	3500	4500	1979	1979	2615
FP16/BF16算力 (TFlops)	1800	2250	989	989	1307
TF32算力 (TFlops)	900	1120	495	495	654
FP64算力 (TFlops)	30	40	67	67	163
显存	192GB HBM3E	192GB HBM3E	80GB HBM3	141GB HBM3	192GB HBM3
互联带宽 (GB/s)	1800	1800	900	900	896

2.4



01

云商资本开支持续上修

02

AI三要素：算力、互联、存储

03

加速卡竞争格局日益激烈

04

GB200解决互联瓶颈，重视产业链弹性

05

AI云端产业链

CONTENTS

目录



2.4

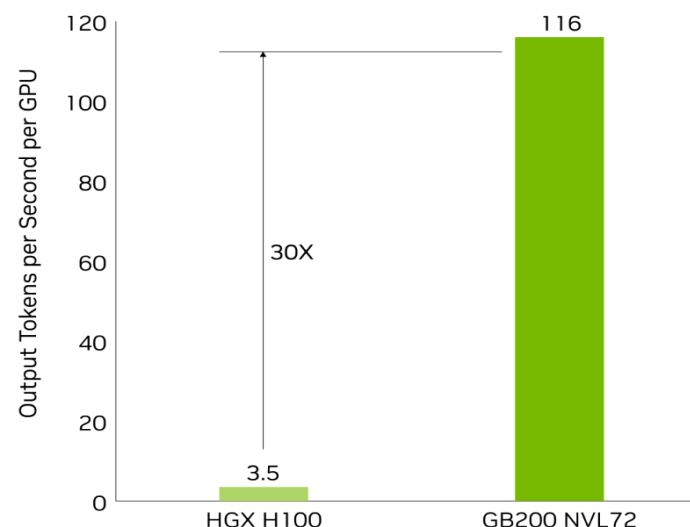
GB200 DGX NVL 72显著提升集群性价比和片间互联能力

- **GB200 DGX NVL 72给英伟达算力性能带来的提升主要来自性价比和片间互联能力方面。**1) 从性价比来看：NVL 72的集群规模增大，一方面节省了除算力芯片以外的系统成本，另一方通过提升产品性能，间接提升了下游客户购买的算力性价比。2) 从片间互联能力来看：市场上主流的AI服务器仍然是传统的8卡服务器架构，而伴随Blackwell平台推出的最多可以支持576卡互联，片间互联数量和带宽的提升极大改善了英伟达平台的推理和训练性能。
- **凭借架构优势，GB200 NVL72的推理性能达到H100的30倍，相较其他竞争对手领先优势更为显著。**按照每10亿个参数的大模型大约需要占用4GB的显存容量计算，当前GPT4已经达到万亿参数量级，运行时至少需要几十张GPU同时进行推理工作，GB200 NVL72在单个Rack内部可以实现万亿参数量级的大模型推理工作，大大降低了大模型在通信环节的算力资源占用，使得GB200 NVL72的推理性能达到H100的30倍。

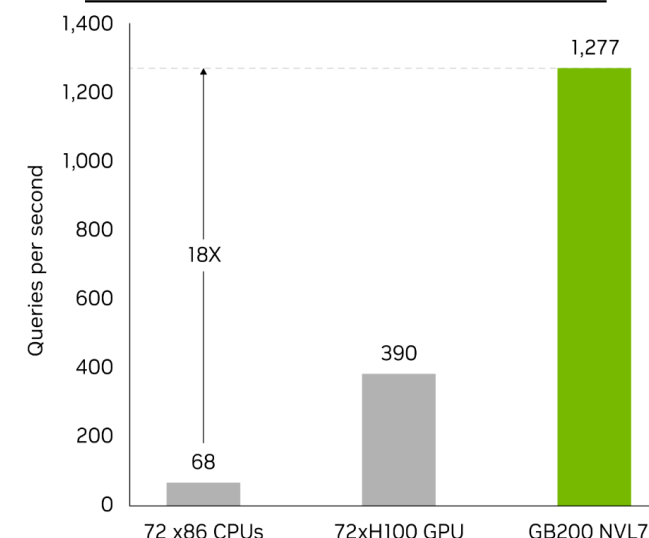
英伟达DGX NVL 72参数

	GB200 NVL72	GB200 Grace Blackwell Superchip
组成	36 Grace CPU : 72 Blackwell GPUs	1 Grace CPU : 2 Blackwell GPU
FP4 Tensor 核心 ²	1,440 PFLOPS	40 PFLOPS
FP8/FP6 Tensor 核心 ²	720 PFLOPS	20 PFLOPS
INT8 Tensor 核心 ²	720 POPS	20 POPS
FP16/BF16 Tensor 核心 ²	360 PFLOPS	10 PFLOPS
TF32 Tensor 核心	180 PFLOPS	5 PFLOPS
FP32	6,480 TFLOPS	180 TFLOPS
FP64	3,240 TFLOPS	90 TFLOPS
FP64 Tensor 核心	3,240 TFLOPS	90 TFLOPS
GPU 記憶體頻寬	最高 13.5 TB HBM3e 576 TB/s	最高 384 GB HBM3e 16 TB/s
NVLink 記憶體頻寬	130TB/s	3.6TB/s

GB200的推理实时吞吐量达到H100的30倍



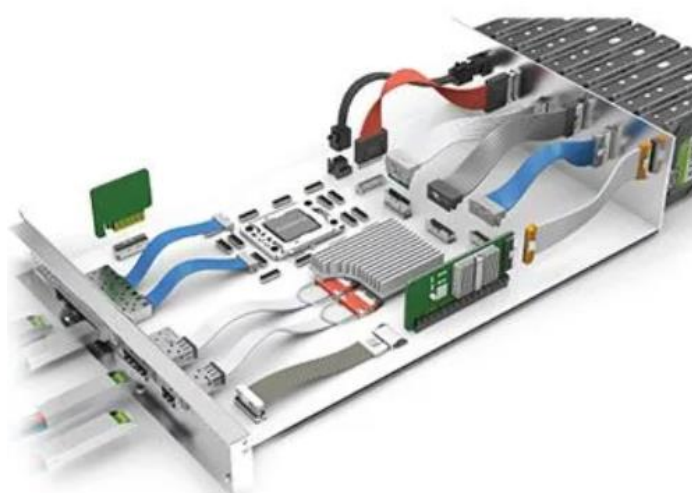
GB200 NVL72、72*H100、72*x86 CPU之间的吞吐量对比



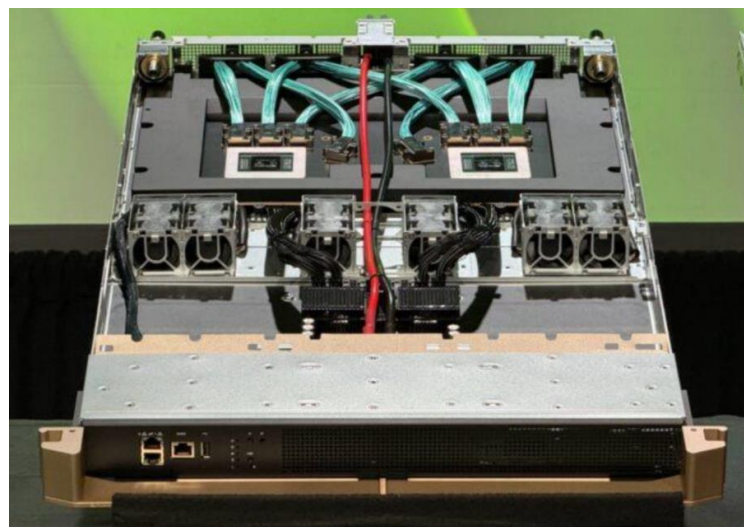
2.4 GB200——铜互连方案为全新增量

- **GB200 NVL72 Rack内部的GPU之间的互联**，英伟达采用了高速铜缆的方案，优点包括：**1) 成本低**——相较于光模块，高速铜缆在相同成本的情况下可以提供更高的互联带宽，从而提升Rack的推理以及训练效率；**2) 功耗低**——采用铜互联方案可以节省光电转换产生的能量损耗，同时也降低了散热问题；**3) 故障率低**——光模块每年有2%-5%的损坏率，而铜连接更加稳定。
- **除了GPU之间，Compute Tray内部以及Rack之间也可以采用铜互连的方案**。在英伟达的高速铜缆解决方案中，Compute Tray内部采用跳线对GPU、网卡等产品进行互联；同时多个Rack之间也可以采用铜缆的方案，GB200的Rack架构下，铜互连方案最多可以在8个Rack之间实现576卡的高速互联。

服务器内部Overpass跳线



Switch tray内部柜内线



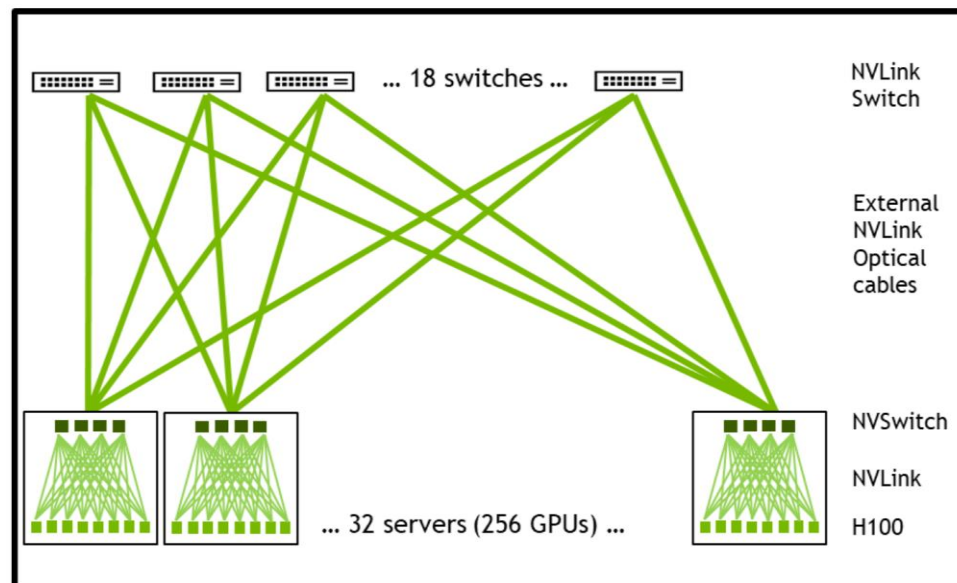
Compute Tray to Switch Tray柜内线



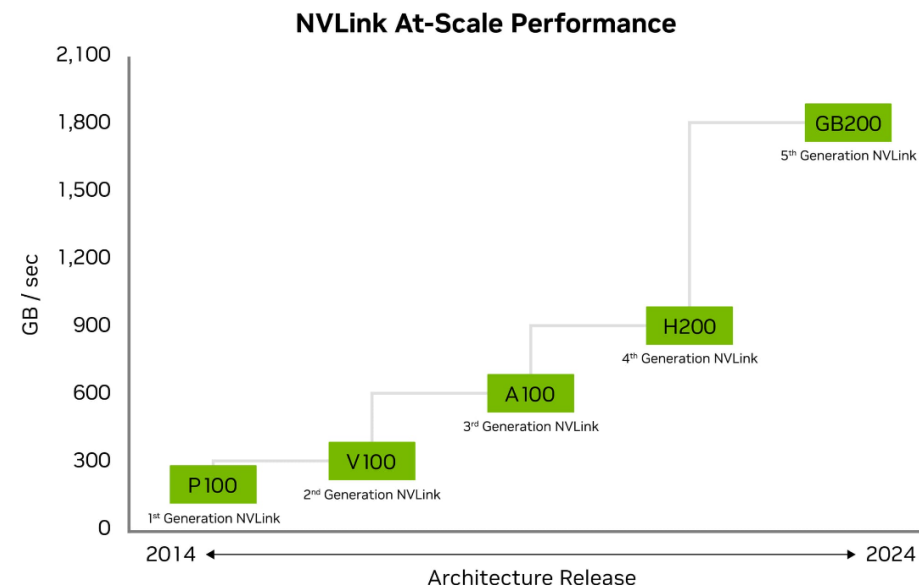
2.4 GB200——光模块带宽持续提升

- **B系列加速卡的发布，带来光模块带宽的翻倍提升。** 英伟达加速卡的数据中心通过IB网络进行互联，采用光模块方案传输数据，光模块的总传输带宽与NV Link带宽成正比例关系。Blackwell平台的推出将英伟达的NV Link带宽从900GB/s的双向互联带宽提升一倍达到了1.8TB/s，从而带来光模块带宽升级一倍，数据中心主流光模块方案也从此前的800G光模块升级为1.6T。
- GB200数据中心架构尽管在Rack内部的GPU之间才用了铜缆进行互联，但Rack之间仍主要采用光模块的方案，目前云厂商普遍构建大集群算力，AI数据中心对光模块仍有大量需求。

H100 8卡服务器数据中心架构



英伟达NV Link升级节奏及性能



2.5



01

云商资本开支持续上修

02

AI三要素：算力、互联、存储

03

加速卡竞争格局日益激烈

04

GB200解决互联瓶颈，重视产业链弹性

05

AI云端产业链

CONTENTS

目录

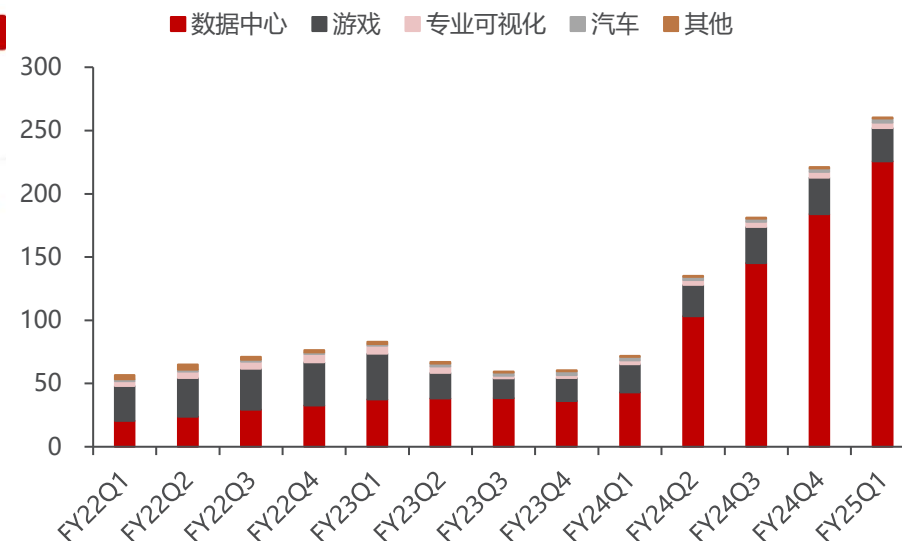


2.5 英伟达：AI时代的算力领军人

- 英伟达成立于 1993 年，公司业务布局包括数据中心、游戏、专业可视化、汽车四大板块。数据中心业务伴随着AI浪潮高速成长，1Q24实现营收225.63亿美元，同比增长426.68%。2023年公司实现营收609.22亿美元，同比+125.85%，实现净利润297.60亿美元，同比+581.32%。
- 此前市场担忧AI训练的需求会随着大模型训练告一段落而放缓，而2024年以来北美四大云商纷纷上修全年资本开支指引，1Q24四大云商合计资本开支476亿美元，同比增长38.98%，反映出客户对训练算力的需求仍在超预期。英伟达的H200、Blackwell平台以及GB200等产品持续提升公司竞争优势，业绩有望持续增长。Bloomberg一致预期2025-2026财年公司净利润为670.47/894.33亿美元，对应2024年6月17日收盘价PE分别为48/36倍。

英伟达软硬件产品线

FY2022-2025Q1公司各业务营业收入（亿美元）



2.5

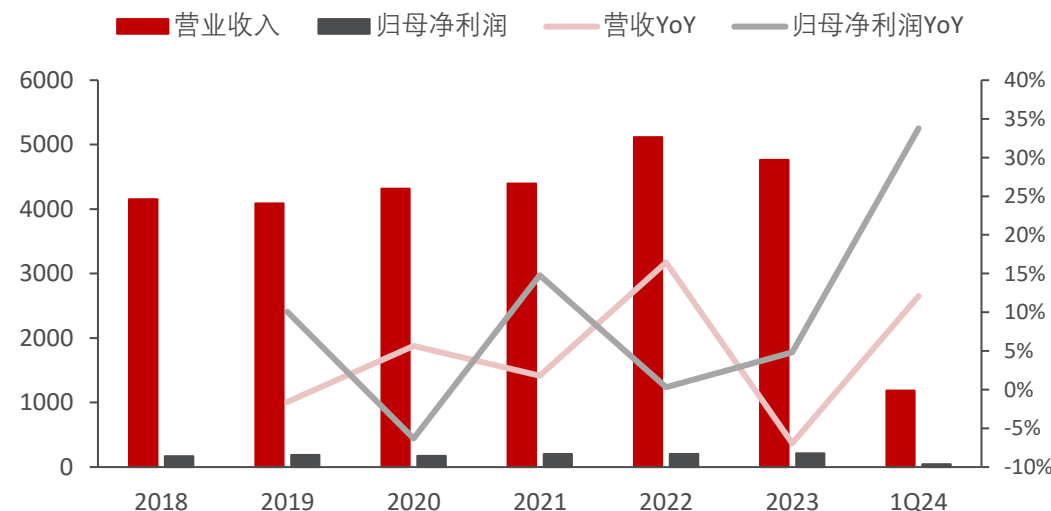
工业富联：产品结构优化调整，AI服务器带来全新成长空间

- **工业富联业务主要包括通信及网络设备、云计算、工业互联网三大板块。**2023年三大业务板块营收占比分别为58.74%/40.91%/0.35%。受益于工业富联在服务器代工领域的前瞻布局，以及AI下游的强劲需求，工业富联的云计算业务收入占比持续提升，2023年占比已超过40%。工业富联与全球各大CSP厂商深度合作，在云计算服务器出货量持续位列全球第一，同时公司具有强劲的供应链关系以及较强的成本控制能力，有望伴随AI浪潮持续成长。
- 财务情况来看，公司营业收入和利润持续增长，2024年一季度营业收入 1186.88亿元，同比增长12.09%，归母净利润41.85亿元，同比增长33.77%。我们预计2024-2026年工业富联的归母净利润分别为255.69/294.06/320.16亿元，对应2024年6月17日收盘价PE分别为21/19/17倍，维持“推荐”评级。
- **风险提示：**AI行业需求波动的风险；行业竞争加剧的风险；宏观需求不及预期的风险。

工业富联三大业务板块



2018-2024Q1工业富联财务情况 (亿元, %)



2.5 海光信息：DCU和CPU全面布局

- 海光信息主要研发应用于服务器、工作站等计算、存储设备中的高端处理器，公司的产品包括海光通用处理器（CPU）和海光协处理器（DCU），公司产品线完善，性能强劲，公司为国内领先的高端处理器企业。
- 海光信息骨干研发人员多拥有国内外知名芯片公司的就职背景，拥有成功研发x86处理器或ARM处理器的经验。CPU产品方面，公司产品按不同业务场景分成了高、中、低端的7000、5000、3000系列。DCU产品方面，海光信息具有完整的软件栈布局和强大的硬件能力，在智能计算以及人工智能市场具备竞争优势。
- 财务情况来看，海光信息业绩稳健增长。1Q24实现营收15.92亿元，同比增长37.09%；实现归母净利润2.89亿元，同比增长20.53%。我们预计海光信息2024-2025年归母净利润为16.50/21.76亿元，对应2024年6月17日收盘价PE 分别为104/79倍，维持“推荐”评级。
- 风险提示：产品研发进度不及预期的风险；政策推进不及预期的风险；供应链风险。

海光信息CPU产品梳理

系列	7000系列CPU	5000系列CPU	3000系列CPU
核心规模	16-32个核心	8-16个核心	4-9个核心
PCIe通道	128路	64路	32路
DDR4通道	8个	4个	2个
内存容量	2TB	1TB	
下游市场	数据中心、云计算中心	云计算、边缘计算、分布式存储等	入门级服务器、工作站、工业控制等市场

客户情况

inspur 浪潮

中科曙光
Sugon

Lenovo

H3C
杭州华三通信技术有限公司

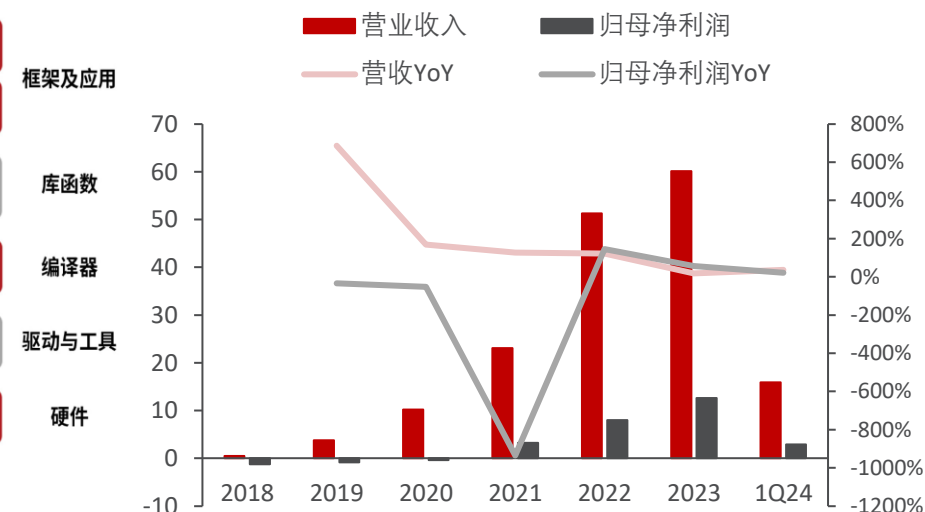
SUPERCLOUD
中国云计算服务器第一品牌

宝德
PowerLeader

海光信息深度计算处理器软件栈



2018-2024Q1海光信息财务情况（亿元，%）



2.5

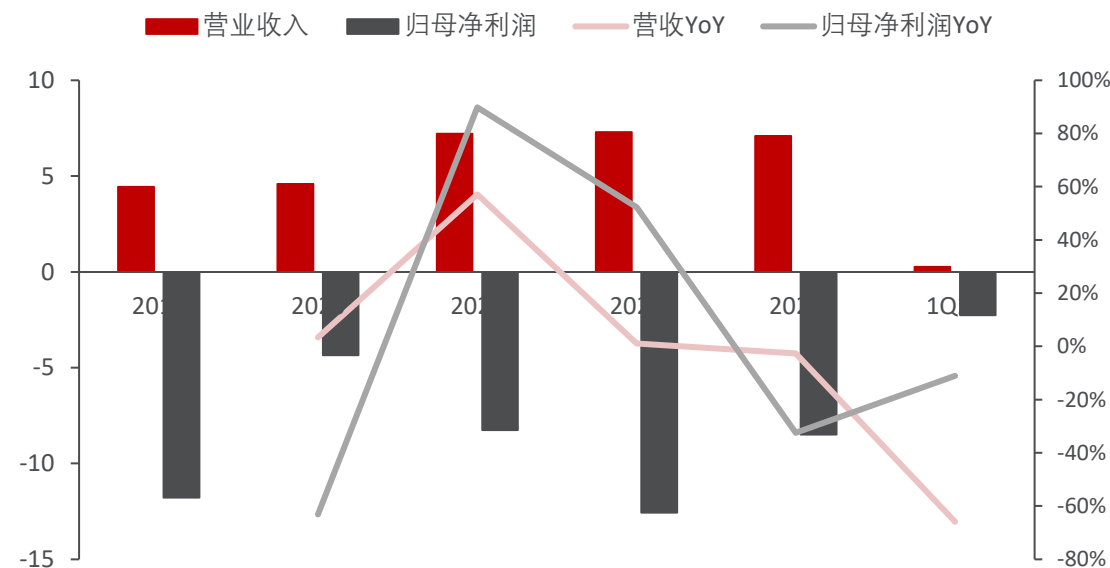
寒武纪：加速卡领域的国内领先厂商

- 公司坚持自主研发，目前公司的主要产品线包括云端产品线、边缘产品线、IP 授权及软件。主要产品为云端智能芯片及加速卡、训练整机、边缘智能芯片及加速卡、终端智能处理器 IP 以及上述产品的配套软件开发平台。公司在国内加速卡领域处于领先地位，产品算力强劲。寒武纪推出的思元590系列产品对标英伟达A100，性能领先，有望为公司业绩带来较大贡献。
- 财务情况来看，1Q24公司实现营收0.26亿元，同比下降65.91%，实现归母净利润-2.27亿元。我们预计寒武纪2024-2026年营业收入分别为15.36/26.74/33.42亿元，对应2024年6月17日收盘价PS 分别为57/33/26倍，维持“推荐”评级。
- **风险提示：**宏观需求不及预期的风险；产品研发不及预期的风险；行业竞争加剧的风险。

寒武纪主要产品布局

产品线	产品类型	寒武纪主要产品
云端产品线	云端智能芯片及加速卡	思元 100 (MLU100) 芯片及云端智能加速卡
		思元 270 (MLU270) 芯片及云端智能加速卡
		思元 290 (MLU290) 芯片及云端智能加速卡
		思元 370 (MLU370) 芯片及云端智能加速卡
		思元 590 (MLU590) 芯片及云端智能加速卡
	训练整机	玄思 1000 智能加速器
边缘产品线	边缘智能芯片及加速卡	思元 220 (MLU220) 芯片及边缘智能加速卡

2019-2024Q1寒武纪财务情况 (亿元, %)



2.5

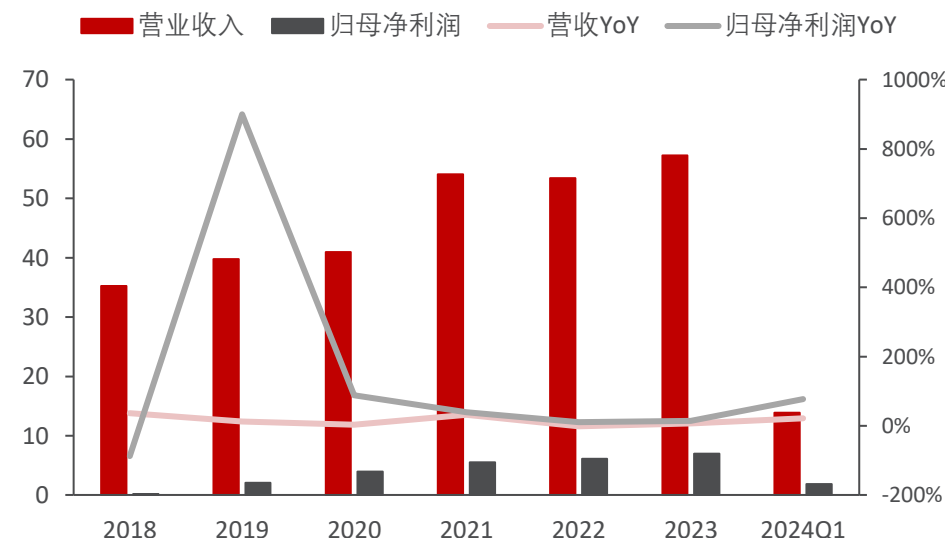
沃尔核材：子公司乐庭智联在高速线缆领域布局领先

- 沃尔核材专业从事高分子辐射改性新材料研发、制造和销售。公司产品线包括电子系列产品、电力系列产品、可分离电缆附件、线缆及新能源产品以及泡棉材料，产品包括热缩套管、热缩母排、热缩电缆附件、冷缩电缆附件等2500多种产品，广泛应用于电子、电力、冶金及航天航空等领域，已形成“沃尔核材”品牌效应。公司产品种类齐全、质量稳定、性能可靠，在电力电缆附件细分行业中的规模总量处于前列水平。
- 沃尔核材子公司乐庭智联在高速线缆领域能力领先，主要客户包括安费诺、莫仕、泰科等，有望受益于GB200对高速线缆的需求拉动。
- 财务情况来看，2024年一季度公司实现营收13.89亿元，同比增长21.63%，实现归母净利润1.84亿元，同比增长76.92%。iFinD一致预期2024-2026年沃尔核材的归母净利润分别为8.48/10.15/11.76亿元，对应2024年6月17日收盘价PE分别为20/17/15倍。

沃尔核材主要产品布局



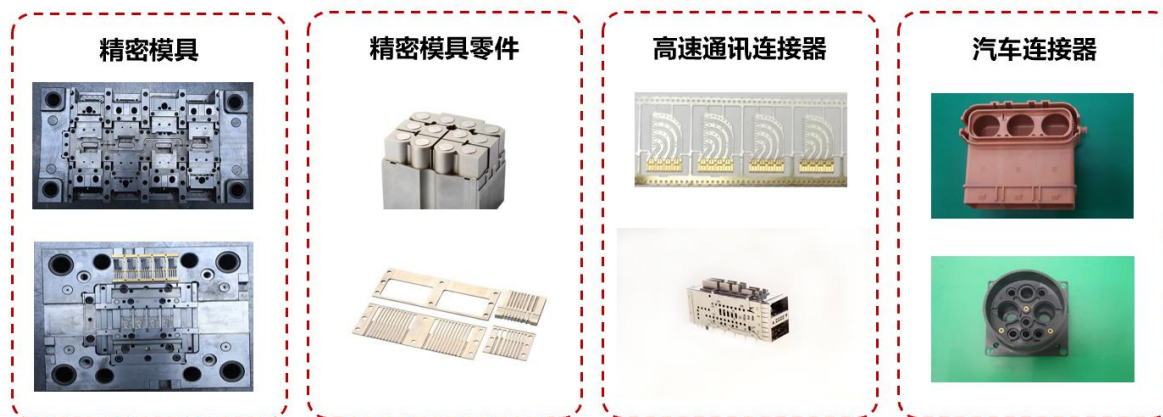
2019-2024Q1沃尔核材财务情况 (亿元, %)



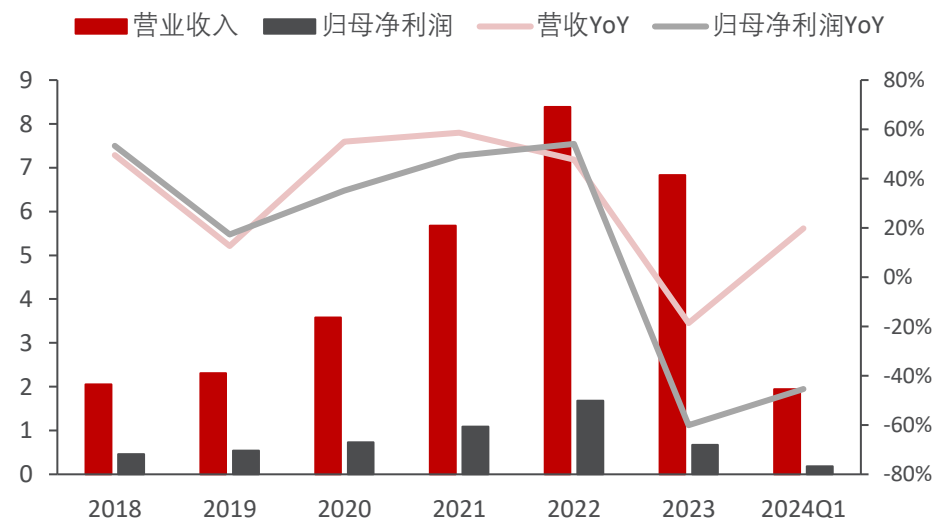
2.5 鼎通科技：连接器组件供应的优秀厂商

- 鼎通科技专注于高速通讯连接器和汽车连接器研发、生产和销售。拥有产品研发、精密零件加工、模具制造、精密冲压和注塑成型技术。公司产品布局包括精密模具、精密模具零件、高速通讯连接器和汽车连接器四个板块。公司凭借先进的技术创新能力、强大的模具制造能力和产品生产制造能力，处于连接器组件供应商的领先地位。
- 安费诺为鼎通的重要客户之一，产品包括高速背板连接器组件和I/O连接器组件，公司有望受益于安费诺在英伟达GB200供应链中的领先地位，获得业绩增长。
- 财务情况来看，2024年一季度公司实现营收1.94亿元，同比增长19.75%，实现归母净利润0.18亿元，同比下降45.45%。我们预计2024-2026年鼎通科技的归母净利润分别为1.02/1.55/2.25亿元，对应2024年6月17日收盘价PE分别为61/40/28倍，维持“推荐”评级。
- **风险提示：**公司产品海外需求不及预期的风险；新能源车销量不及预期的风险；公司产品品类拓展不及预期的风险。

鼎通科技主要产品布局



2019-2024Q1鼎通科技财务情况 (亿元, %)



03. 供应链变革：全面升级的算力产业链

3

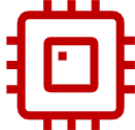
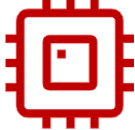
服务器供应链变革

- 通常单台传统服务器价格在1万美金以内，而搭载8张H100算力卡的DGX H100 AI服务器价值量可达40万美金（300万人民币左右）
- 自然语言的运算、图形的处理需要大量存储器协助运算及存储，并且AI服务器对数据传输速率有着更高要求，催生了存储及PCB价值量的增长。
- 下文，我们将延续上一章节，从供应链角度出发，探讨存储、PCB，以及更上游的封测、制造环节的产业投资机遇。

传统服务器



AI服务器

 整机	6千美金	 整机	40万美金	DGX H100服务器搭载8张H100，国内售价一度达40万美金
 算力芯片	无，一般采用集显	 算力芯片	24万美金	单张H100卡售价最高可达3万美金，AI服务器满载需要8张；算力芯片在AI服务器中成本占比可达80%+
 存储	2900美金	 存储	2万美金	普通服务器256G内存+16T HDD/960G SSD混合RAID，而AI服务器所需DRAM/NAND分别8倍/3倍于普通服务器，RAID卡价值量亦有所提升
 PCB	260美金	 PCB	1300美金	AI服务器中PCB面积基本不变，PCB材质提高、层数由14/16层提高至20层以上，价值量可达普通服务器5倍以上

3.1



01

存储：从HBM到CXL

02

PCB：服务器引领高端化浪潮

03

封测：先进封装解决算力瓶颈

04

制造：AI卖铲人

CONTENTS

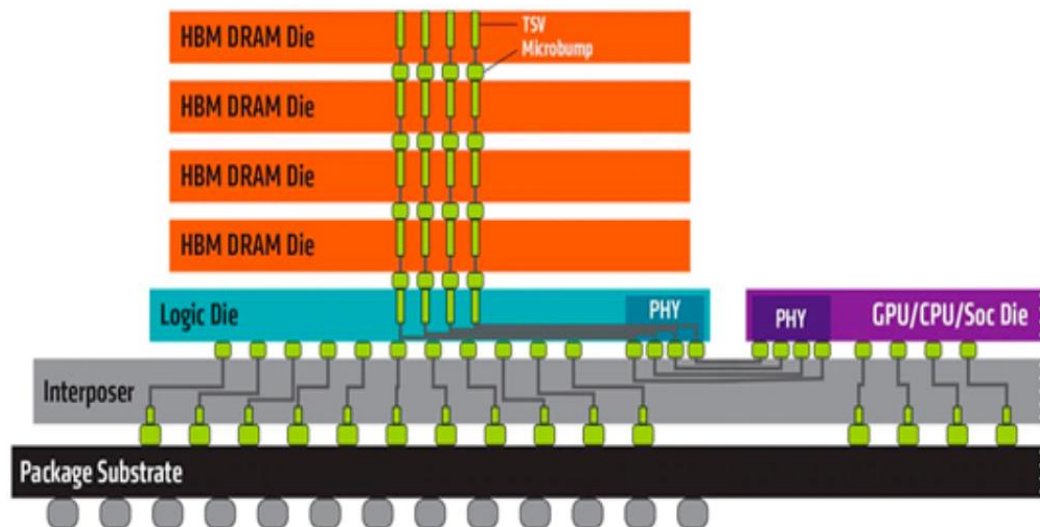
目录



3.1 服务器存储：HBM打破“内存墙”

- 如我们前文所述，AI三要素包括算力、互联和存储，从服务器供应链角度看，存储是AI时代变革重点。
- 处理器性能不断提升，“内存墙”成为计算机系统的瓶颈。计算机系统的运行受到处理器和内存的配合影响，但处理器性能因摩尔定律不断提升，而内存DRAM的传输带宽没有跟随工艺的演进而快速增长，导致访存时延迟高、效率低，严重制约处理器性能发挥，即出现“内存墙”。在AI和视觉等领域，需要大量的内存带宽来支持复杂的计算操作，若内存性能落后，会导致实际算力下降50%甚至90%。
- HBM是目前用于打破“内存墙”的重要技术之一。高带宽存储器（HBM）是一种基于3D堆栈工艺的高性能DRAM，可以提供更高的内存带宽和更低的能耗，适用于高存储器带宽需求的应用场合，如HPC、网络交换设备等。最新的HBM3E产品可提供超过1TB/s的数据带宽，具有8Gb/s的I/O速率，缓解了因内存部件延迟而阻碍算力增长的问题。

图：HBM示意图



表：HBM技术性能发展

	HBM1	HBM2	HBM2E	HBM3	HBM3E
发布时间	2013	2016	2019	2021	2023
颗粒密度	2 Gb	8 Gb	16 Gb	16 Gb	-
堆积高度	4 Hi	4/8 Hi	4/8 Hi	8/12 Hi	-
内存容量	1GB	4/8 GB	8/16 GB	16/24 GB	-
带宽	128 GB/s	307 GB/s	460 GB/s	819 GB/s	1TB/s
I/O速度	1 Gbps	2.4 Gbps	3.6 Gbps	6.4 Gbps	8Gbps

资料来源：AMD，民生证券研究院

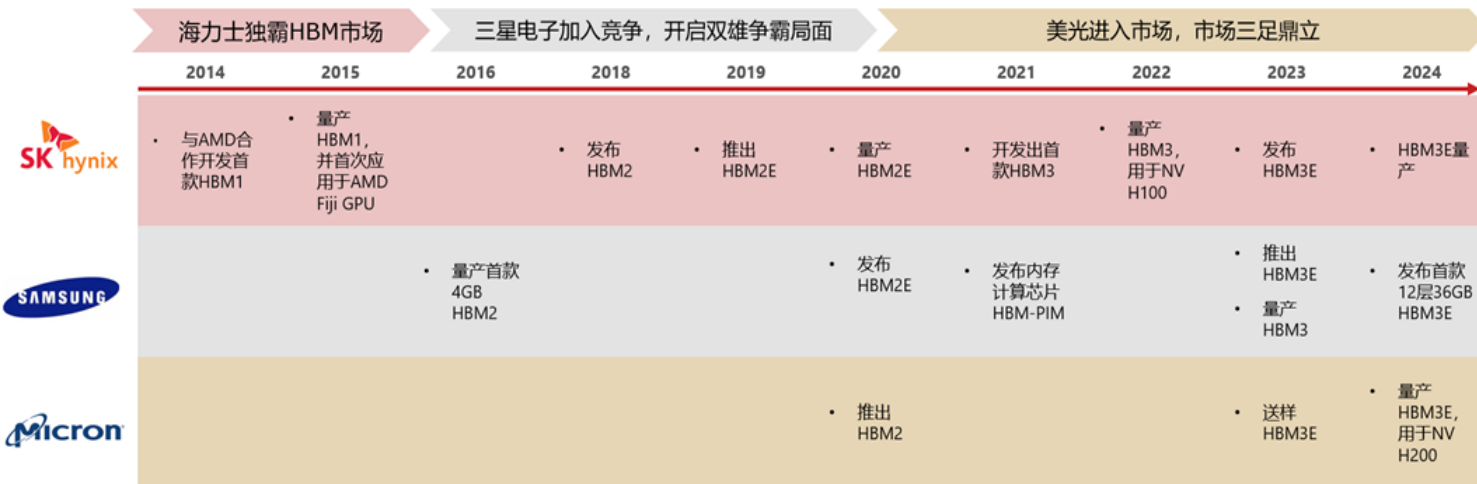
资料来源：Wikipedia，SK hynix官网，民生证券研究院

3.1

服务器存储：三大原厂争抢HBM技术高地

- **三星与海力士占领主要市场份额，美光加速追赶。**根据Yole Group统计，2023年海力士以55%的营收占比在HBM市场占据主导权，其后是三星和美光，营收占比分别为41%和3%。美光早期在堆栈式DRAM的探索中最先选择了HMC技术，在数据传输时延迟和速度方面存在劣势，但其及时转向，在2020年推出首款HBM2产品，此后跃过HBM3，直接投入研发HBM3E，并于2023年7月发布24GB 8-High HBM3E。
- **三家原厂最新的HBM3E产品对比来看，**低能耗是美光的主要竞争优势，三星拥有更小的堆叠间隙，有利于更高层数堆叠，SK海力士则得益于其开发的MR-MUF技术，散热性能优异。

图：三大存储厂商HBM进展图



资料来源：Wikipedia、海力士官网、IT之家、三星官网、CFM闪存市场、美光科技官网，民生证券研究院整理

表：三大原厂HBM3E产品对比

	美光	SK海力士	三星
发布时间	2024年2月26日	2024年3月19日	2024年2月27日
容量	24GB or 36GB	最高36GB	36GB
最大带宽	超过1.2TB/s	超过1.18TB/s	1.25TB/s
堆叠层数	8 or 12	12	12
最大数据传输速率	超过9.2Gbps	9.2Gbps	10Gbps
出货时间	2024年2月	2024年3月	2024H2
特点	功耗低，率先导入到NVIDIA H200 GPU中	散热性能提高10%	芯片堆叠间隙更小，密度更高

资料来源：各公司官网，民生证券研究院

3.1

服务器存储：AI服务器给内存技术带来的变化

➤ AI服务器需求核心在于更大带宽的存储，带来了存储技术路线变革：

- 1) **CXL** (compute express link)：全新的互联技术标准，其带来的DRAM池化技术可以大大节约数据中心的建设成本，同时带动DRAM的用量。
- 2) **MCR/MDIMM** (Multiplexer Combined Ranks)：大大提高内存带宽，**AMD已经在MemCon 2023上表达了它帮助推动JEDEC的MRDIMM开放标准** 的承诺，英特尔也与SK hynix和瑞萨合作，基于与MRDIMM类似的概念，开发了多路合并阵列 (MCR) DIMM。
- 3) **PCIe 5.0** (Peripheral Component Interconnect Express 5.0)：新一代总线技术，构建了更加高速的串行通信系统标准。

表：AI时代服务器新技术方案

技术/产品名称	主要应用场景	功能	相关互连芯片
DDR5	PC和服务内存模组	具备更高速度、更大容量与更低能耗等特点。DDR5内存的最高传输速率达6.4Gbps，比DDR4内存提升了一倍，同时减少功耗。	接口及配套芯片：用于内存模组满足服务器CPU对内存模组日益增长的高性能及大容量需求；用来缓冲来自内存控制器的地址/命令/控制信号
MCR/MRDIMM	AI服务器，多核CPU等对内存带宽要求高的场景	结合两个DDR5 DIMMs，向主机提供两倍的数据速率，大幅提升内存带宽	MRCD/MDB：结合了每个列的传输，有效地将两根DDR（双倍数据速率）模组转换为根QDR（四倍数据速率）模组。
CXL	数据中心等	CXL提供有效的资源共享/池来提高性能，最大限度地减少对复杂软件的需求，并降低系统总成本，同时也将大大带动DRAM的用量。	MXC：可用于和CPU相连接的设备提供内存共享，对于那些需要高速内存访问和大量数据处理的应用（如数据中心、云计算、AI和机器学习），使用MXC芯片可以提高内存访问效率和系统性能。
PCIe 5.0	主要是高速SSD卡，AI服务器GPU和网卡	用于连接CPU和高速外设，PCIe 5.0相较4.0速率及带宽翻倍，满足更高带宽的应用场景；	retimer：随着CPU性能越来越快，高速信号传输可能会有紊乱，因此添加一个中继PCIe Retimer来恢复信号，没有会对性能影响较大

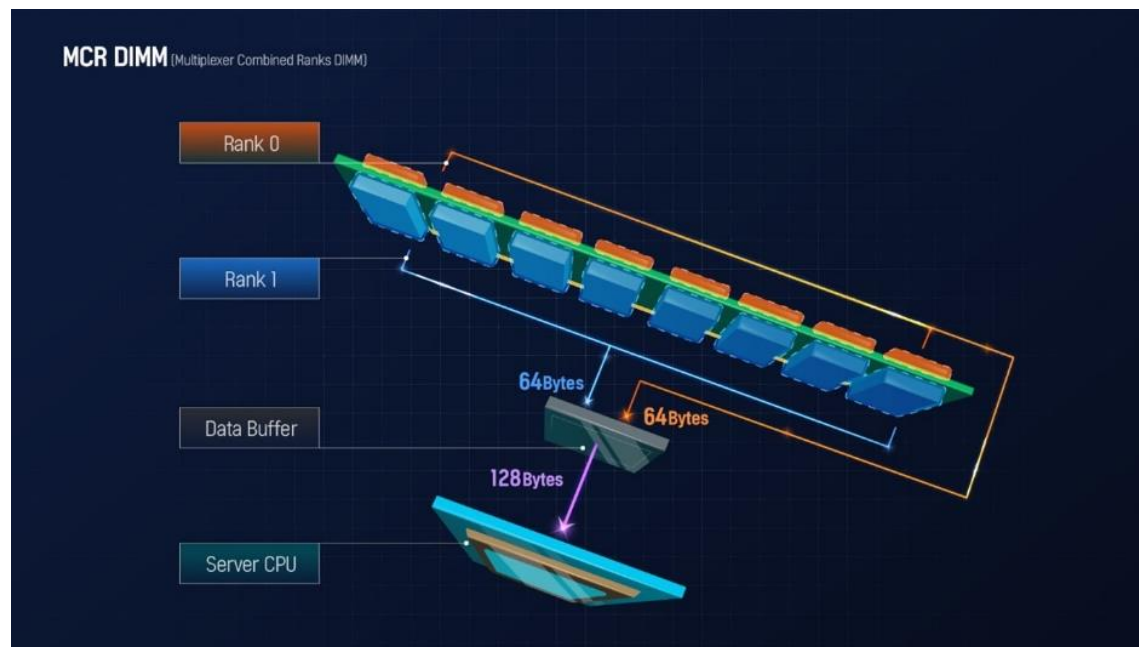
资料来源：民生证券研究院整理

3.1

服务器存储：MCR/MRDIMM内存模组带来控制芯片升级

- MCR/MRDIMM可以很好的满足AI服务器对内存带宽的高需求，该技术将多个DRAM内存模块组合在一起，通过将两个Rank形成伪多内存通道（Pseudo Channel），并使用专门的控制器（接口芯片）来管理它们之间的数据传输，并大大提高内存带宽，**理论上，MCR/MRDIMM内存的带宽是DDR5的两倍。**
- 澜起科技已于2022年完成MCR控制芯片（MRCD/MDB）研发。

图：MCR/MRDIMM架构



资料来源：海力士，民生证券研究院整理

表：各种内存条对比

封装类型	性能	价格	特点	应用
UDIMM	低	低	产品无缓冲器，容量和频率较低，但相同频率的产品延迟较低	桌面、服务器
RDIMM	较高	较高	采用单RCD芯片，数据传输更稳定，具备更高的存储容量和频率	工作站、服务器
LRDIMM	高	高	接口芯片采用RCD*1+DB*10接口芯片，其主板电压较低，可降低了内存总线负载和功耗，具备较高的内存容量	服务器
MRDIMM	极高	高	接口芯片采用MRCD*1+MDB*10，结合两个DDR5 DIMMs，向主机提供两倍的数据速率，大幅提升内存带宽。理论带宽上限为DDR5的两倍	高性能服务器

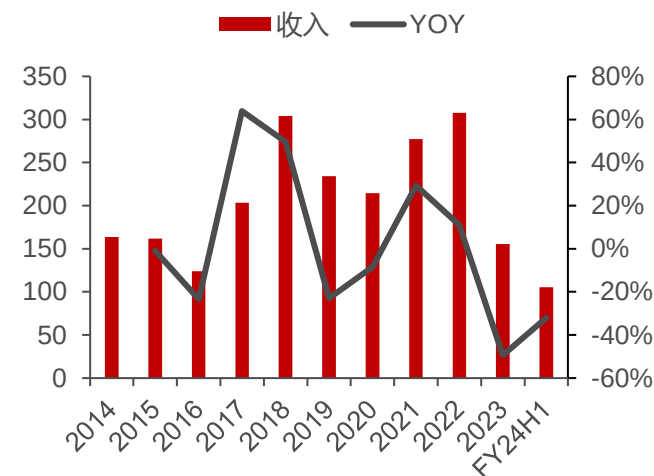
资料来源：CSDN，民生证券研究院整理

3.1

美光科技：全球存储巨头，HBM后发先至

- 美光创立于1978年，总部位于美国爱达荷州，主要产品包括DRAM、NAND和NOR。根据CFM闪存市场数据，2023年公司DRAM份额位居全球TOP3，市场占有率22.9%，仅次于三星和SK海力士；NAND市场占有率则为10.8%，同样为市场主要玩家之一。
- 美光在HBM市场起步晚于三星和SK海力士，但于2023年7月率先发布HBM3E完成新产品反超。后续产品规划上，根据美光披露的产品路线图，其预计将于2025年发布36GB 12-High HBM3E用以完善其HBM3E产品线，并于2026年推出革命性产品36GB 12-High HBM4，带宽预计将超过1.5TB/s。预计在2027年前，美光将对HBM4产品容量进一步升级，发布48GB 16-High HBM4。2028年，美光预计发布带宽升级至2 TB/s 以上的HBM4E。2023-2028年内，美光几乎年均推出一款HBM系列新产品，展示出其对高速增长HBM市场的勃勃雄心。
- 估值：据Bloomberg一致预期，公司2024-2026财年经调整净利润9.5/97.4/117.5亿美元，对应现价PE 154/16/13倍（截至2024年6月14日收盘价）。

图：美光科技营收（时间按财年划分，亿美元）



资料来源：Bloomberg，民生证券研究院

图：美光HBM内存技术路径



资料来源：DRAMeXCHANGE，民生证券研究院

3.1 澜起科技：内存接口芯片龙头，算力时代多种新品打开成长空间

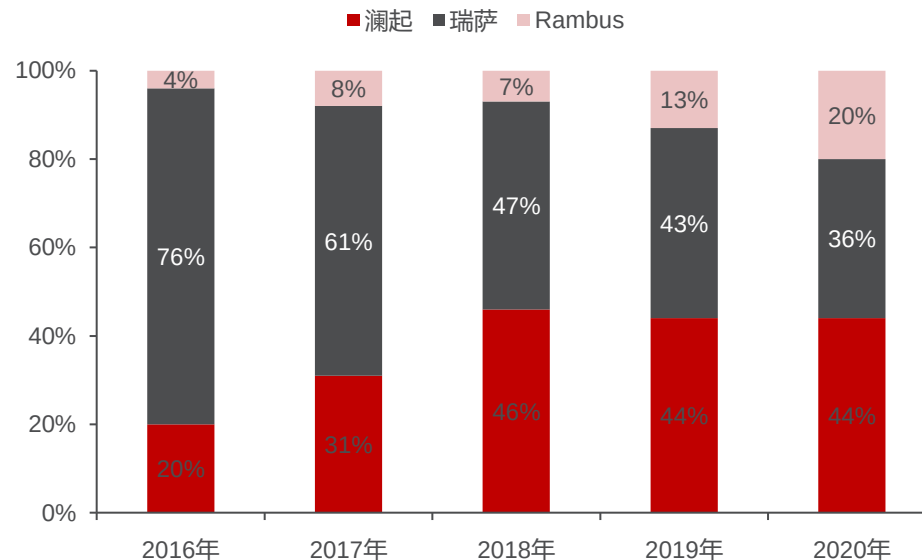
- 起步于2004年，20年经验的老牌芯片制造厂商，曾在Nasdaq上市，公司获三星/英特尔战略投资，是内存接口芯片的领导者 and 全球微电子行业标准制定机构的董事会成员之一，深度参与JEDEC相关产品的标准制定，从顺应行业发展到引领未来趋势。澜起科技拥有两大产品线，互连类芯片产品线和津逮服务器平台产品线。其中，互连类芯片产品主要包括内存接口芯片、内存模组配套芯片、PCIe Retimer芯片、MXC芯片、CKD芯片等，津逮服务器平台产品包括津逮CPU和混合安全内存模组。
- 新品序列基本在2023年完成研发及送样，未来业绩核心驱动一是DDR5芯片不断迭代保持价值量，二是新品逐步上量打开成长空间。高性能运力产品方面Retimer，MRCD/MDB已经实现出货，并在推进开展DDR5第四子代RCD、第二子代MRCD/MDB芯片和PCIe 6.0 Retimer芯片的研发。
- 估值：我们预计公司2024-2026年归母净利润14.00/20.57/25.24亿元，对应现价PE 43/29/24倍，维持“推荐”评级。
- 风险提示：DDR5渗透不及预期；新品导入进度不及预期；终端市场复苏不及预期。

图：澜起科技产品线



资料来源：澜起科技公司官网、澜起科技2022年年度报告，民生证券研究院

图：内存接口芯片竞争格局



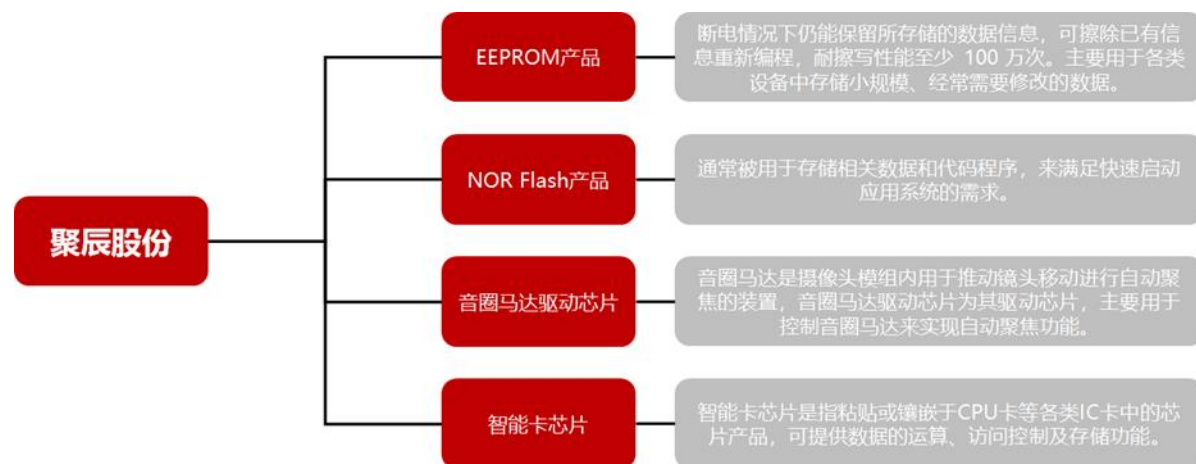
资料来源：立鼎产业研究院、华经产业研究院、澜起科技招股说明书，民生证券研究院

3.1

聚辰股份：SPD EEPROM受益DDR5加速渗透

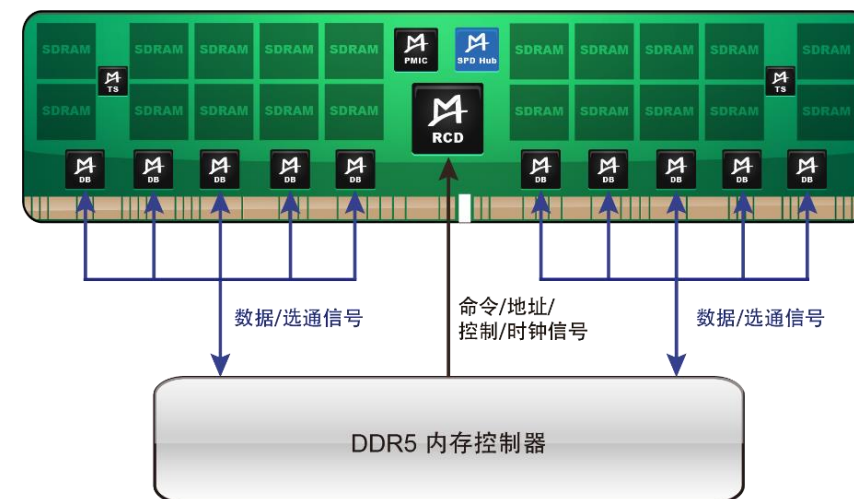
- 聚辰股份是国产EEPROM龙头厂商。产品主要包括EEPROM、音圈马达驱动芯片、智能卡芯片和NOR Flash，并广泛应用于智能手机、汽车电子、白色家电等众多领域。公司持续拓宽EEPROM产品的应用领域，与澜起科技合作开发的SPD EEPROM产品于2021年第四季度实现量产。
- 内存迈入DDR5世代，SPD EEPROM必需性突显。除内存接口芯片RCD、DB外，串行检测集线器（SPD）是内存管理系统的关键组成部分，适用于DDR5系列LRDIMM、RDIMM、UDIMM、SODIMM等内存模组。伴随DDR5内存渗透率不断提升，SPD EEPROM将迎来更广阔的市场空间。
- 估值：我们预计公司2024-2026年归母净利润3.21/4.47/5.72亿元，对应现价PE 30/21/17倍，维持“推荐”评级。
- 风险提示：新品研发销售进度不及预期；DDR5渗透进度不及预期；行业竞争加剧。

图：聚辰股份产品线



资料来源：聚辰股份2022年年度报告，民生证券研究院

图：SPD Hub在内存模组上的应用



资料来源：澜起科技官网，民生证券研究院

3.2



01

存储：从HBM到CXL

02

PCB：服务器引领高端化浪潮

03

封测：先进封装解决算力瓶颈

04

制造：AI卖铲人

CONTENTS

目录



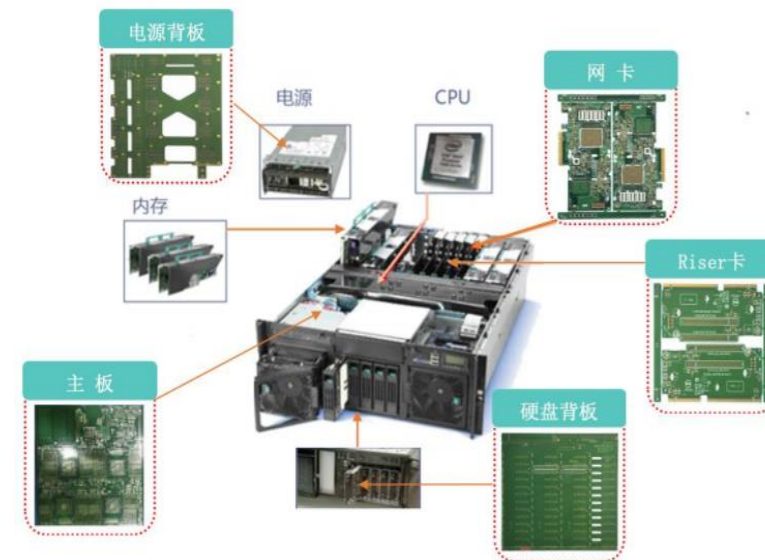
3.2 PCB——服务器引领高端化浪潮

- 服务器内部涉及PCB板的主要部件包括电源背板、网卡、主板、硬盘背板、Riser卡等，其中主板面积最大、价值量最高。服务器PCB板的特点体现在高层数、高密度及高传输速率，所用PCB板以6层、8-16层、18层以上刚性板，以及封装基板和软板为主。
- 随着芯片性能的提升，云端对数据的处理能力增强，对信号的传输速度提出更高要求。**对于传统服务器，当前PCIe 5.0渗透率正逐步提升，PCB层数、材质及价值量有相应提升。**英特尔于23年推出最新服务器平台Eagle Stream，所用PCB为16-20层，使用Very Low Loss的覆铜板材料。AMD最新的服务器平台属于ZEN 4架构，所用PCB同为是16-20层，同样采用Very Low Loss的材料。

表：PCB跟随服务器平台升级

	Platform	Purley		Whitley	Eagle Stream		Birch Stream
Intel	CPU	Skylake	Cascade Lake	Ice Lake	Sapphire Rapids	Emerald Rapids	Granite Rapids
	Nano Process	14 nm	14 nm	10 nm	Intel 7	Intel 7	Intel 3
	PCIe Gen	PCIe 3.0	PCIe 3.0	PCIe 4.0	PCIe 5.0	PCIe 5.0	PCIe 5.0
	MP Time	2017 Q3	2019 Q3	2021 Q1	2023 H1	2023 H2	2024
	CCL Material	Mid loss	Mid loss	Low loss	Very Low loss	Very Low loss	VVL/Ultra Low Loss
	Layer count	8 to 12	8 to 12	12 to 16	16 to 20	16 to 20	TBD
AMD	Architecture	ZEN	ZEN2	ZEN3	ZEN4		ZEN5
	CPU	Naples	Rome	Milan	Genoa	Bergamo	Turin
	Nano Process	14 nm (global)	7nm (TSMC)	7nm (TSMC)	5nm (TSMC)	5nm (TSMC)	4nm/3nm (TSMC)
	PCIe Gen	PCIe 3.0	PCIe 3.0	PCIe 4.0	PCIe 5.0	PCIe 5.0	PCIe 5.0
	MP Time	2017 Q3	2019 Q3	2020 Q4	2022 Q4	2023	2024
	CCL Material	Mid loss	Mid loss	Low loss	Very Low loss	Very Low loss	VVL/Ultra Low Loss
	Layer count	8 to 12	8 to 12	12 to 16	16 to 20	16 to 20	TBD

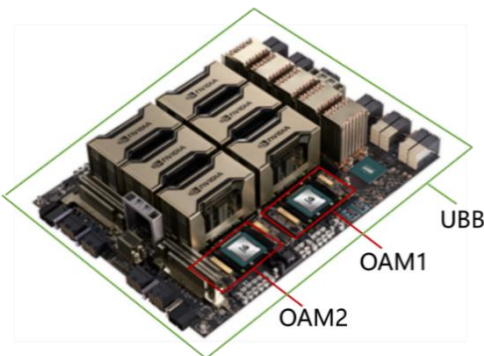
图：服务器内部PCB



3.2 AI服务器PCB量价齐增

- 相比传统服务器，AI服务器采用了CPU+GPU的异构架构，以英伟达DGX H100服务器为例，主要需要增加两种类型的PCB：**承载GPU的OAM（OCP Accelerator Module，加速卡模组）及实现GPU多卡互联的UBB（Universal Baseboard，通用基板）**。以搭载8颗H100芯片的AI训练服务器为例，每颗H100需要配备一张OAM，共计8张OAM搭载至1张UBB之上，形成8卡互联。
- AI服务器对信号传输速率要求更高，PCB的层数、材质等均有升级，如CPU主板层数需14-24层；GPU OAM需20-30层高阶HDI，GPU UBB亦需20-30层。材质方面，CPU主板及GPU OAM需Very Low Loss（M6）覆铜板材料，而GPU UBB则最高使用到Ultra Low Loss（M7）材料。**AI服务器PCB用量增多，并且材料等级高、加工难度大，单价亦有明显提高，单台AI服务器PCB价值量可达1万元。与AI服务器相匹配，400G/800G交换机升级速度加快，交换机用PCB可达30层以上，价值量提升更为显著。**
- AI服务器机柜所需的功率数倍于传统服务器，服务器电源升级以提高用电效率，保证供电稳定。服务器电源可分为AC-DC/DC-DC两类，其中AC-DC主要负责电网中交流电到服务器所需直流电的转换，DC-DC则将转换后的直流电精细化，为设备内部各部件提供更精确的直流电压，PCB亦伴随电源有所升级。**DC-DC电源在承受大电流的同时，对于电路控制更加精细，其所运用的PCB工艺相较AC-DC更为复杂，价值量更高，深南电路、威尔高、生益电子等公司在服务器电源PCB领域有相应产品布局。**

图：AI服务器中的OAM和UBB



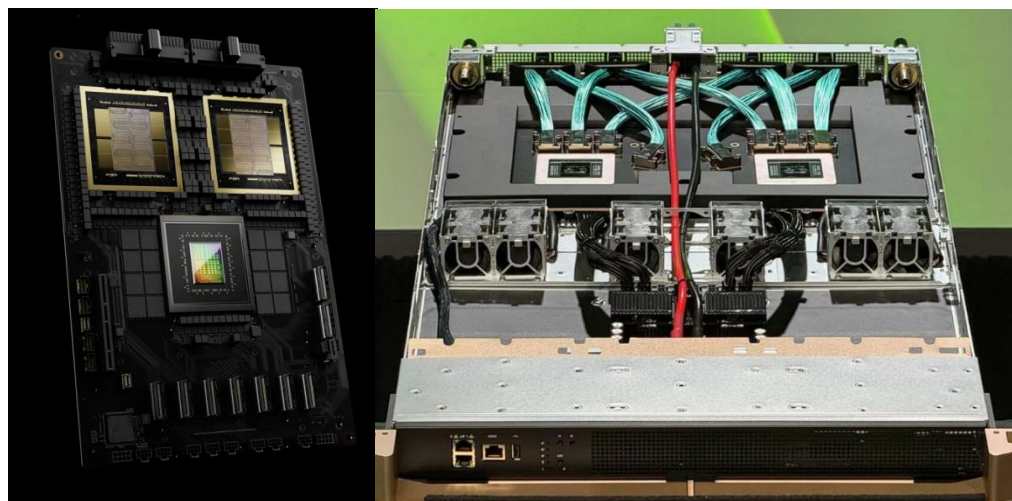
图：AI服务器中的AC-DC电源（左）及DC-DC电源（右）



3.2 GB200——AI加速PCB向高速+HDI升级

- 与DGX H100相比，GB200由于取消了单台服务器内8颗GPU互联的设计，**因此不再需要UBB板，同时Switch tray取代UBB承担了互联的功能**。DGX H100中每颗GPU芯片需要承载至一张OAM加速卡上，而CPU单独搭载至主板。GB200采用Superchip的设计，每张Superchip承载有1颗Grace CPU及2颗Blackwell GPU，CPU不再单独搭载至主板；**因此与H100相比，GB200取消了主板，并且用一张大面积的Superchip板取代之之前单颗GPU的OAM**。GB200中Blackwell GPU相较上一代Hopper GPU 的AI计算性能大幅提升，因此在更高速信号传输的需求下，PCB材料及规格或将有所提升，PCB单位面积价值量亦有所提升。

图：GB200中的Superchip板（左）和Nvlink Switch板（右）



图：GB200中PCB升级

服务器类型	类型	单 GPU 芯片用量 (张)	规格	覆铜板材料
DGX H100	CPU 主板	0.125	14-24 层	M6
	OAM	1	20-30 层 5 阶 HDI	M6
	UBB	0.125	20-30 层	M6/M7
GB200 NVL72	CPU 主板	无	/	/
	Superchip 板	0.5	或为 6 阶 HDI	或为 M7/M8
	NVlink Switch 板	0.125	或为多层板/高阶 HDI	或为 M7/M8

3.2

沪电股份：全球AI服务器及交换机PCB龙头

- **产品结构：**2023年企通板占公司整体营收66%，汽车板占公司整体营收24%。
- **AI服务器：**AI服务器主要增加UBB和OAM两类高价值量PCB，公司为AI服务器PCB核心供应商。
- **AI交换机：**AI拉动400g和800g交换机出货比例提高，沪电核心收益交换机PCB升级。
- **技术+客户双壁垒，供需两端优势显著：**
- **供给端：**沪电深耕高多层板领域，技术实力处于T1的位置，核心产品交换机PCB层数可达30层以上，加工难度大，技术壁垒高，“玩家”较少避免内卷
- **需求端：**公司海外营收占比80%+，PCB是高度定制化的行业，海外客户选定供应商后不轻易更换，每年进行正常的cost down，此外通讯类客户对供应商背景要求严格，贸易摩擦的背景下进入门槛高。因此海外占比高就意味着客户壁垒高、竞争格局好，需求端利润释放有保障。
- **2024年一季度公司实现营收25.84亿元，同比增长38.34%，实现归母净利润5.15亿元，同比增长157.03%。我们预计2024-2026年沪电股份归母净利润分别为22.55/28.84/32.04亿元，对应PE分别为31/25/22倍（股价为6月14日收盘价）。**
- **风险提示：**AI算力建设进度不及预期，800g交换机渗透进度不及预期，行业竞争加剧

表：沪电生产线情况

公司名称	主要业务	设计产能
沪利微电	高端汽车板	150 万平方米/年
黄石一厂	企业通讯市场板	75 万平方米/年
	22 层以下 PCB 产品（青淞厂转入）	
	中低阶汽车 PCB 产品（沪利微电转入）	
黄石二厂	汽车板为主的生产线	75 万平方米/年
青淞工厂	企业通讯市场板	220 万平方米/年

图：沪电股份主要客户



3.2

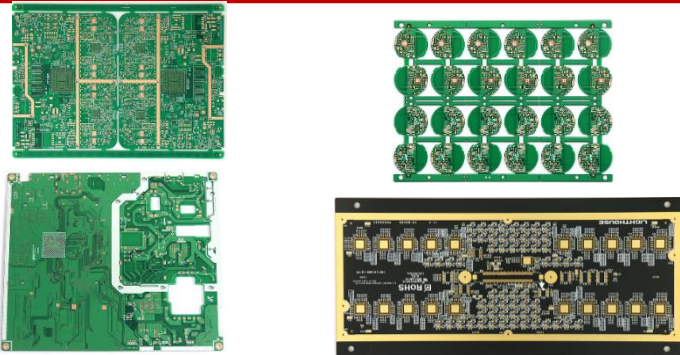
胜宏科技：UBB及OAM突破，AI服务器PCB放量可期

- 胜宏科技从事高精密多层、HDI PCB、FPC、软硬结合板的研发、生产和销售，产品广泛应用于计算机、航空航天、汽车电子（新能源）、5G新基建、大数据中心、工业互联、医疗仪器等领域。胜宏科技是CPCA副理事长单位，行业标准的制定单位之一，位居中国印制电路行业企业百强排行榜内资第4名，全球第20名。
AI服务器： OAM及UBB通过北美大客户认证，有望拿下大批量订单。
- GB200：** GB200 compute tray采用类似OAM的设计，或需使用基于M7/M8高速材料的6阶HDI工艺，技术难度极高；switch tray同样有望采用高阶HDI方案，胜宏在OAM方面积累了充分的开发与制造经验，有望受益于GB200 PCB向高阶HDI的升级。
- 消费类方面，** 公司为英伟达消费级显卡核心供应商，此外产品广泛应用于PC、存储、Mini LED等下游，AI PC催生存储需求，且PCB主板有望升级，有望推动公司相关产品量价齐；Mini LED下游需求持续旺盛，公司产品亦受益于此。
- 车载方面**公司在安波福、大陆等客户持续放量；**服务器方面，** 公司拥有浪潮、中科曙光等优质客户群体，受益于1Q24服务器行业的复苏
- 2024年一季度公司实现营收23.92亿元，同比增长36.06%，实现归母净利润2.10亿元，同比增长67.72%。Wind一致预期2024-2026公司归母净利润分别为11.61/14.91/17.99亿元，对应PE分别为25/19/16倍（股价为6月14日收盘价）。**
- 风险提示：AI算力建设进度不及预期，产品放量进度不及预期，行业竞争加剧

表：胜宏具备4-28层6阶HDI、2-42层高多层板生产能力

项目 Item	工艺能力 Technology Capability	2021	2022	2023	2024
板厚 Board Thickness	层数 Layer count	2~32L	2~36L	2~40L	2~42L
	最大成品板厚 Max. Finished board thickness	5.0mm	5.0mm	5.0mm	5.0mm
	最薄芯板厚（不含铜） Core thickness (Exclude copper)	50um	50um	50um	50um
	最薄PP厚度 Prepre thickness	40um	40um	30um	25um
	板厚公差 Board thickness tolerance	±10%	±8%	±5%	±5%
层数 Layers Counts	N/M/N	4~20L (5 Steps)	4~20L (5 Steps)	4~24L (6 Steps)	4~28L (6 Steps)
	任意层互联 Anylayer	12L	14L	14L	16L
板厚 Board Thickness	总板厚 Finished board thickness (um)	300~3200	200~3200	200~3200	200~3200
	半固化片类型和厚度 Prepreg Type (um)	PP1027/40um	PP1017/30um	PP1017/30um	PP1010/22um
	基板厚度 CCL Thickness (um)	50um	50um	50um	50um
	铜箔类型 Copper Foil Type	HTE. RTF. VLP	RTF. VLP. SLP	RTF. VLP. SLP	VLP. SLP. SLP2
	镭射孔最小孔径/pad Min laser via/pad (um)	75/175	75/160	70/150	70/140

图：胜宏部分PCB产品



3.3



01

存储：从HBM到CXL

02

PCB：服务器引领高端化浪潮

03

封测：先进封装解决算力瓶颈

04

制造：AI卖铲人

CONTENTS

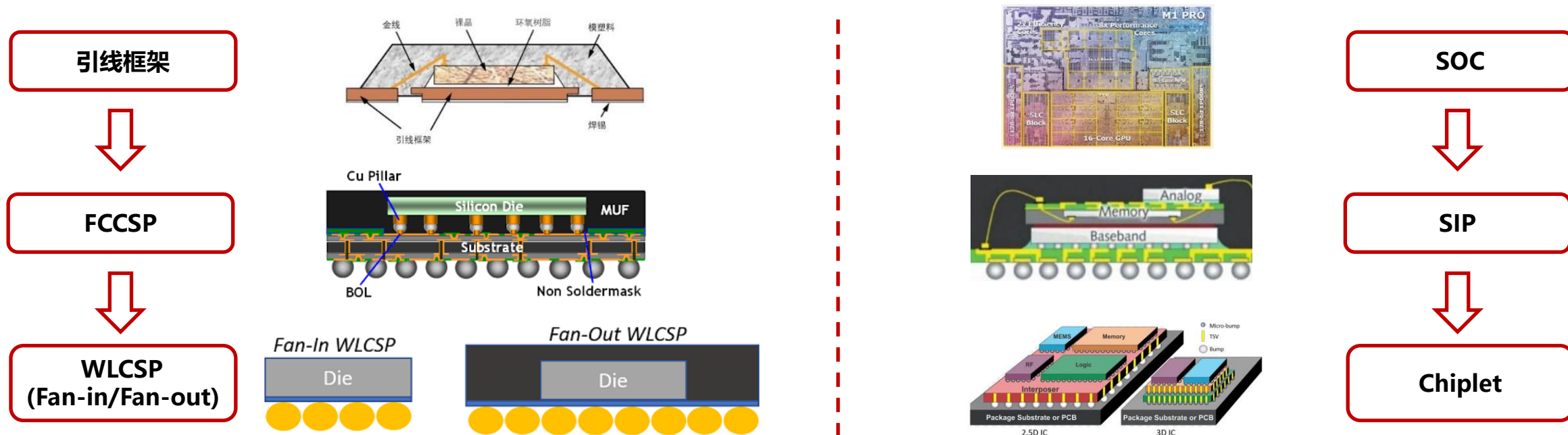
目录



3.3 先进封装两大发展方向：单芯片密度提升+多芯片集成

- AI时代，先进封装成为算力芯片的关键技术突破方向。过去的发展历程中，先进封装沿着两大方向发展：
- 1、单芯片封装体积减小，密度提升。从传统的引线框架封装，到倒装芯片封装，到更先进的WLCSP封装，单芯片的封装体积逐渐减小，封装密度提升。该技术主要用于消费类产品。
- 2、多芯片集成。从传统的单芯片封装到chiplet技术下的2.5D/3D封装，芯片封装集成度逐步提升。该技术方向主要应用于算力类产品。

图：先进封装发展两大路径



资料来源：艾邦半导体，2011 IEEE 13th Electronics Packaging Technology Conference, anysilicon, anandtech, nexpcb, semiengineering, 民生证券研究院整理

3.3 算力芯片主流技术路径：COWOS封装+HBM存储

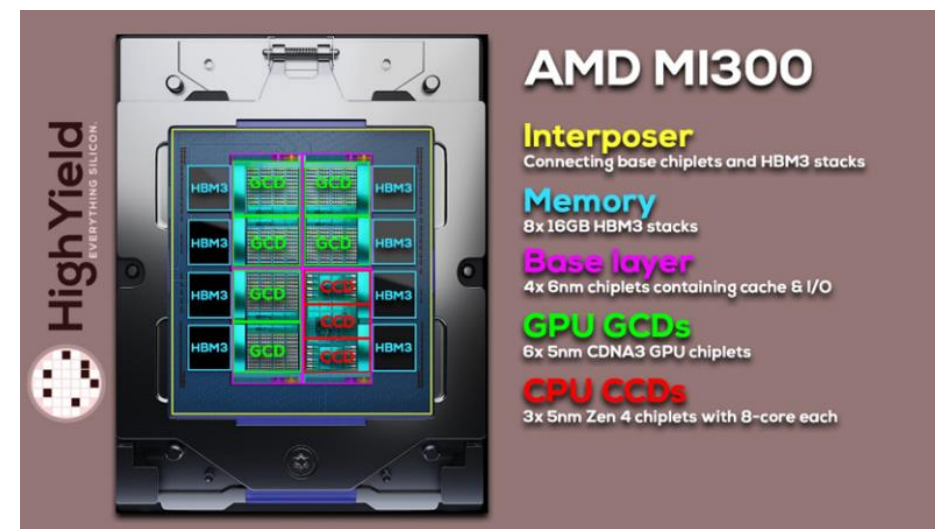
- 算力芯片的COWOS封装+HBM显存已经成为主流方案。
- **Nvidia**方面，当前主力产品H100采用**台积电CoWoS封装**，H100采用4nm制程，并配备最高80GB HBM3。
- **AMD**方面，MI300 AI加速卡采用COWOS的后继升级版SOIC方案，拥有13个小芯片，共包括9个5nm的计算核心（6个GCD+3个CCD），4个6nm的I/O die兼Infinity Cache（同时起到中介层的作用，位于计算核心和interposer之间），还配备了累计8颗共计128GB的HBM3芯片。相较Nvidia的A/H系列产品，MI300更进一步的将SOC拆分成了多颗小芯粒，并拥有更大的面积、芯粒数量、缓存颗粒数量。

Nvidia A100/H100的2.5D方案



资料来源：wccftech，民生证券研究院整理

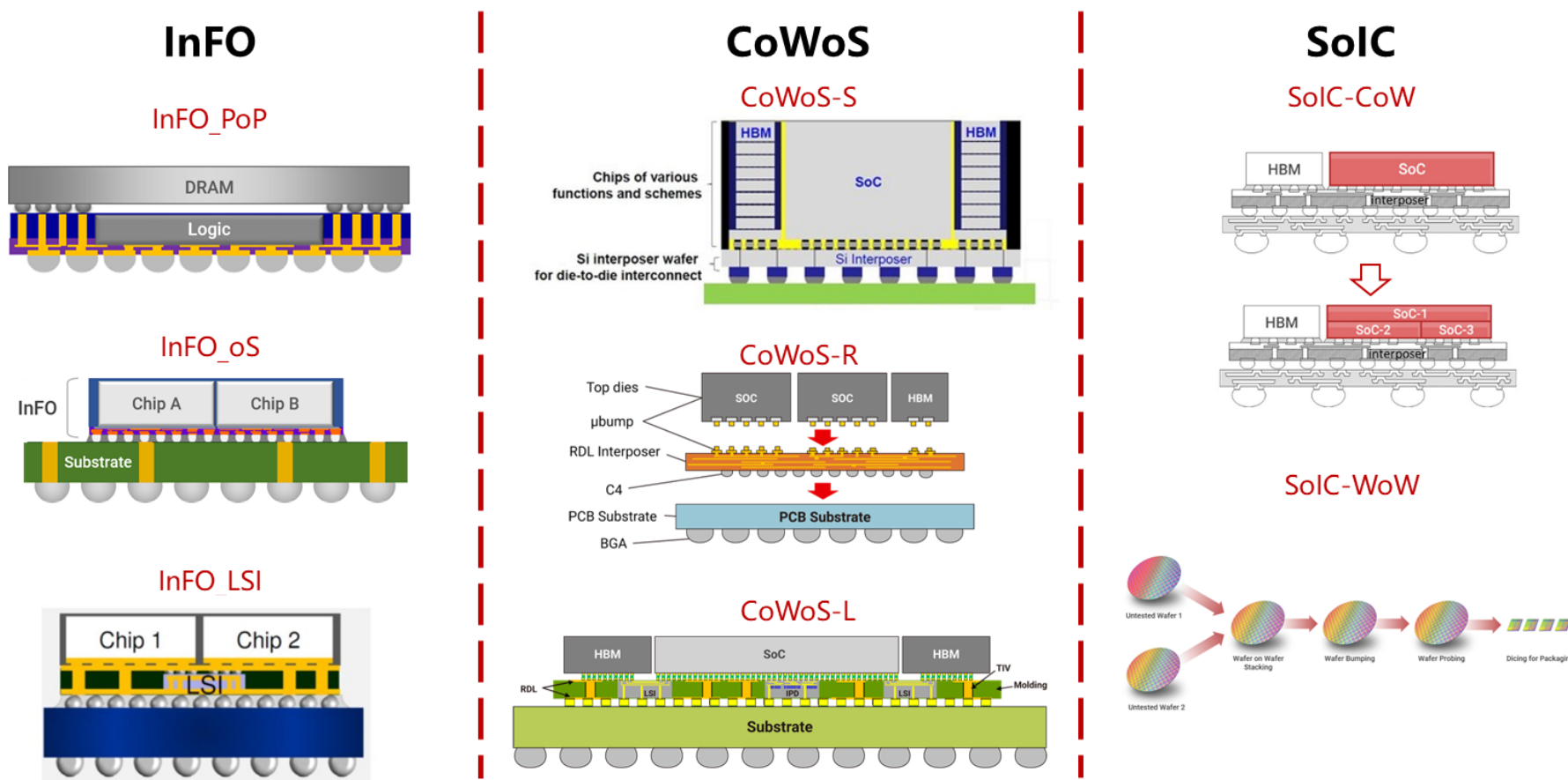
AMD MI300的3D方案



资料来源：HighYield，AMD，民生证券研究院

3.3 COWOS及其继任的封装技术路径

图：台积电用于Chiplet技术的三种封装工艺



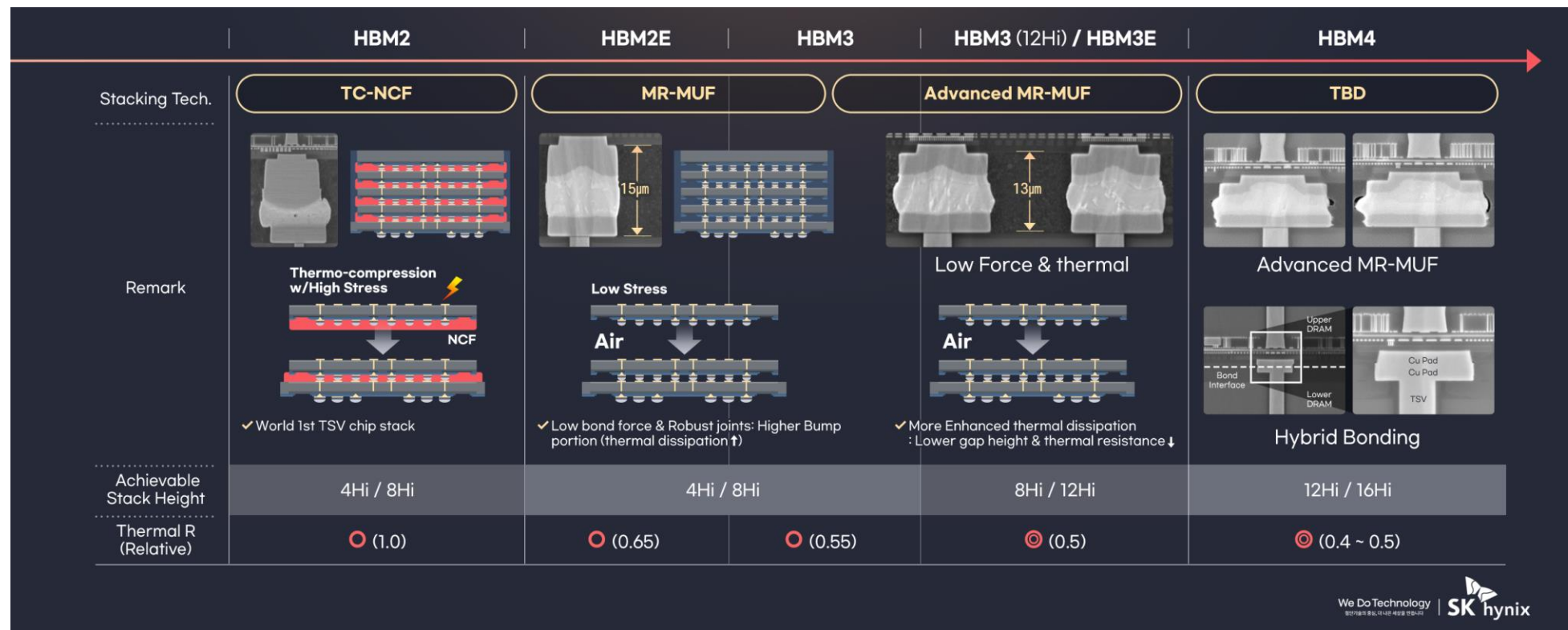
- **InFO: 无Interposer**, die直接在载板上互联。**成本低、互联速率低**。其中InFO_LSI是折中方案, 在载板中内嵌了局部硅桥。**多用于消费类、通信类**, M1 Ultra采用此种方案。
- **CoWoS: 有Interposer**, die在interposer上互联, 再封在载板上。Interposer方案也分硅基、有机两种。造价高, **多用于算力芯片 (Nvidia H100)**。
- **SoIC: 晶圆键合**, 前道堆叠, 而非后道封装环节堆叠。2021年发布, **目前用于AMD MI300系列**。

资料来源: TSMC, eetimes, 民生证券研究院整理

3.3 HBM技术路径

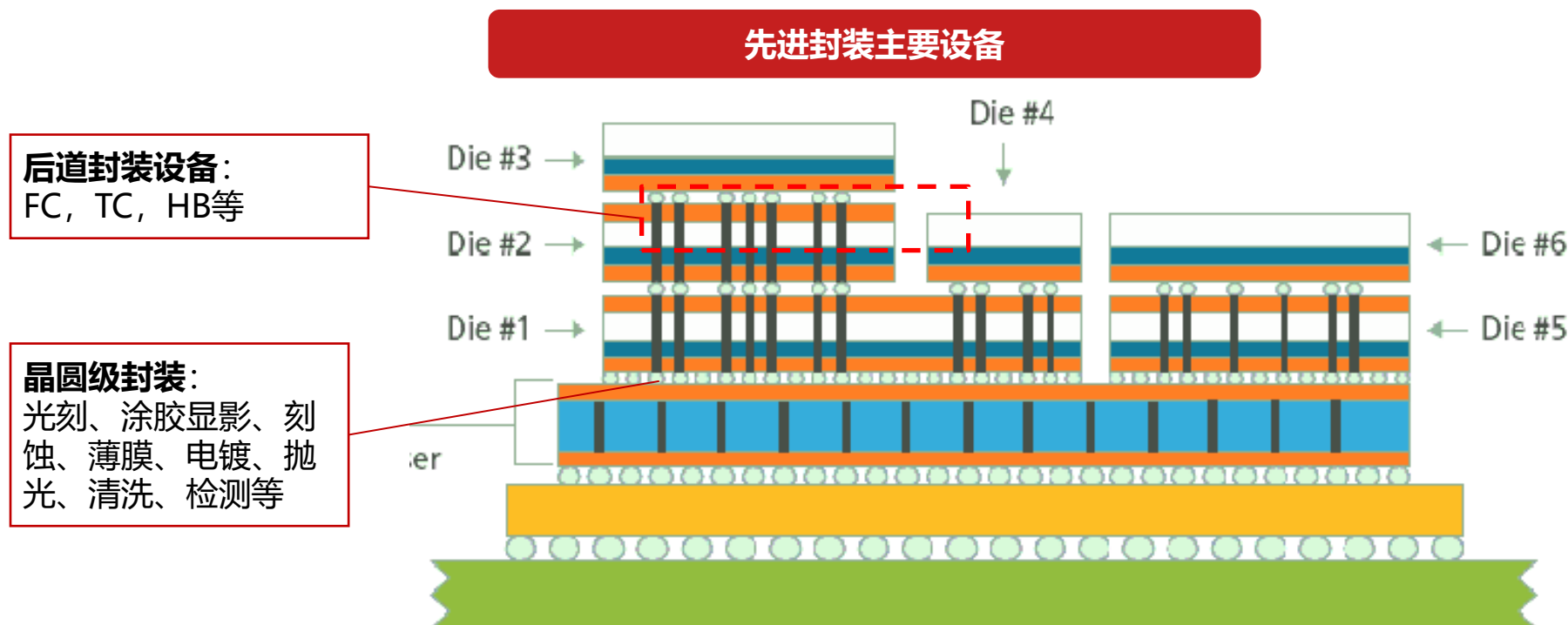
- 当前HBM的主流方案为**MR-MUF**（海力士）和**TC-NCF**（三星、美光）两种。海力士从HBM2E开始从TC-NCF切换至MR-MUF，因MUF方案有更低的热阻抗和更好的散热。
- 两种技术路径均有TCB需求。TC-NCF直接使用TCB进行堆叠，MR-MUF则是使用TC进行堆叠和pre-bonding，之后再行MUF填充和Mass Reflow。

图：海力士HBM2-HBM4的技术路径



3.3 先进封装设备：3D堆叠为核心环节

- 在HBM和COWOS封装中，区别于传统封装的设备主要有高精度FC（用于COW环节，主要厂商ASMPT）、TC（用于HBM堆叠和OS环节，主要厂商ASMPT）、HB（用于SOIC，未来可能用于HBM堆叠，主要厂商BESI、EVG）；
- 晶圆级封装环节增加了检测设备用量，国产厂商：赛腾股份，海外龙头：Camtek；
- 后道测试环节增加了测试机用量，国产厂商：长川科技，华峰测控，精智达，海外龙头爱德万、泰瑞达。
- 封装设备产业链亦受益于先进封装扩产：芯源微（先进封装涂胶显影、清洗机）、光力科技（划片机）、芯碁微装（直写光刻机）等。

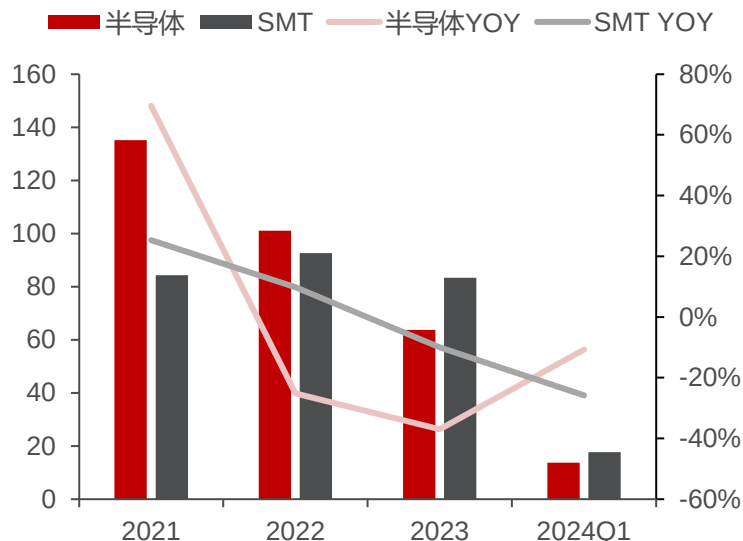


资料来源：Pancholi, A. 《3-Dimensional Integrated Circuits》，民生证券研究院整理

3.3 ASMP: 算力先进封装设备龙头

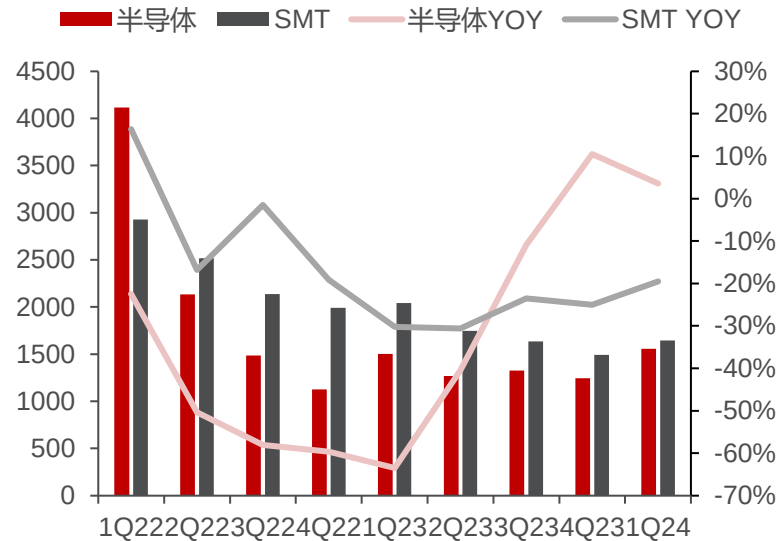
- ASMP成立于中国香港，总部位于新加坡，大股东荷兰ASMI为全球ALD设备龙头。公司两大业务板块：**半导体后道设备**（下游为半导体封装）、**SMT表面贴装设备**（下游为消费电子、汽车电子、工业终端制造等整线设备）。在半导体封装领域拥有全球领先的市场份额。
- ASMP在先进封装设备方面有广泛布局，能为先进封装流程提供一系列设备产品线。公司的**热压式固晶（TCB）**广泛应用于各种先进逻辑存储芯片；**混合键合式固晶（HB）**用于CIS和3D NAND堆叠等应用的晶圆到晶圆混合键合，以及HPC和数据中心的逻辑存储器3D堆叠。根据公司2023年报披露，公司在2023年下半年获得晶圆代工客户C2S应用的重要TCB订单，已经开始交付，预期2024年第二季度获得更多来自该客户及其供应链合作伙伴的TCB订单。此外，公司也向该晶圆代工客户交付了下一代超细间距TCB用于C2W工艺。
- 估值：据iFinD一致预期，公司2024-2026年归母净利润11.45/21.04/26.98亿港元，对应现价PE 34/18/14倍。

图：分业务营收（百万港元）



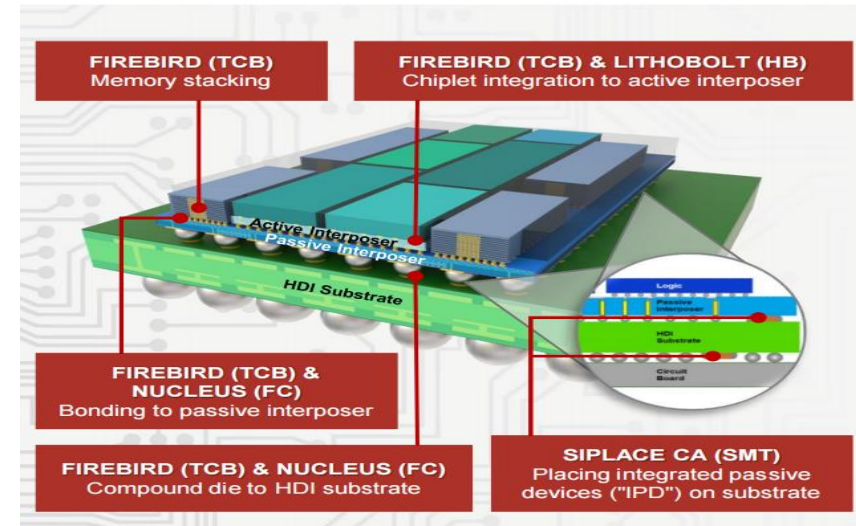
资料来源：iFinD，民生证券研究院

图：分业务订单（百万港元）



资料来源：iFinD，民生证券研究院

图：公司产品在算力芯片中的应用



资料来源：ASMP 2023Q4业绩发布PPT，民生证券研究院

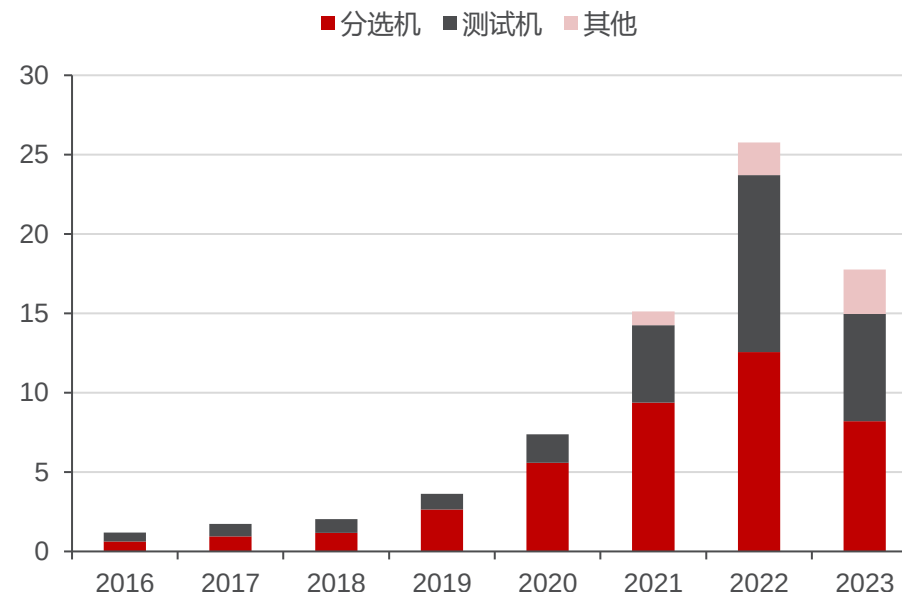
3.3 长川科技：数字测试机新品放量

- 长川主业包括分选机、测试机等，覆盖功率、数模混合、数字多种测试应用。客户覆盖TI、安靠、日月光、美光等海外龙头公司和长电、通富、华天等国内封测厂。2019年公司收购STI，拓展AOI产品线。2023年公司收购长奕科技，拓展转塔式分选机产品线，并募投实施转塔式分选机扩产。
- 多种测试机新品放量构筑公司成长性。D9000测试机可适应于高性能集成电路CP测试和FT测试，适用产品类型数字逻辑芯片、数模混合芯片、微处理器、系统级SoC及其射频类芯片，三温分选机C6800T适用于GPU、服务器CPU、AI芯片等产品测试。
- 估值：我们预计公司2024-2026年归母净利润4.40/6.73/8.56亿元，对应现价PE 45/29/23倍，维持“推荐”评级。
- 风险提示：测试机新品研发进度不及预期；下游封测行业景气波动；行业竞争加剧

图：长川科技产品线



图：长川科技分业务营收（亿元）



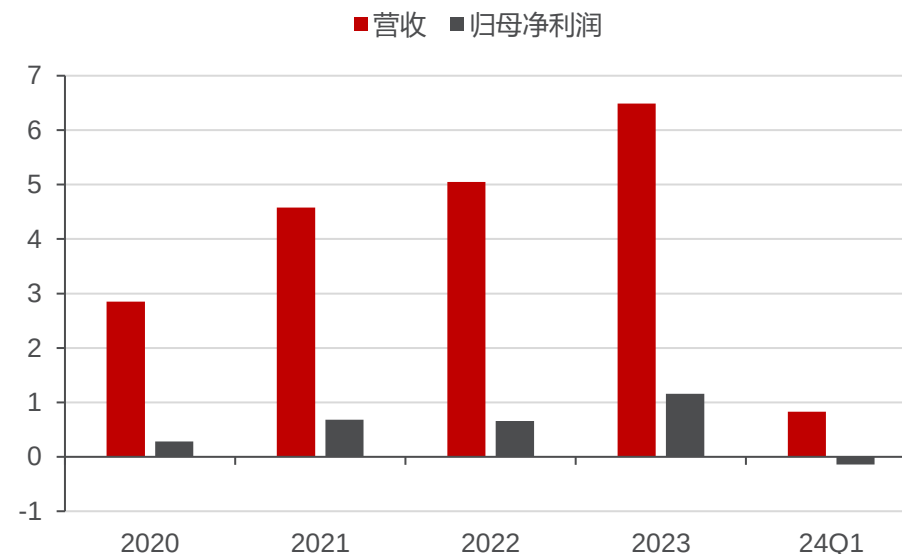
3.3 精智达：存储测试机国产化领军

- **精智达是检测设备与系统解决方案提供商。**公司传统业务主要集中在新型显示的检测、老化等设备，而新业务存储器检测设备有望成为公司新的增长点。
- 公司的存储器件测试解决方案主要包括存储器晶圆测试系统、存储器老化修复系统、存储器封装测试系统及其他测试配件等。截至2023年报披露，老化修复设备及治具板、探针卡等已经获得客户订单并在客户供应商名录上实现靠前排序，**晶圆测试机已经基本完成样机验证**。公司根据新技术和新应用在半导体存储器方面的新要求，**配合客户业务稳步推进开发针对如HBM等新一代半导体存储器测试需求的测试技术和设备。**
- **估值：我们预计公司2024-2026年归母净利润1.60/2.19/2.89亿元，对应现价PE 31/23/17倍。**
- **风险提示：测试机新品研发导入不及预期；行业竞争加剧；存储器行业需求周期性波动**

图：精智达产品线



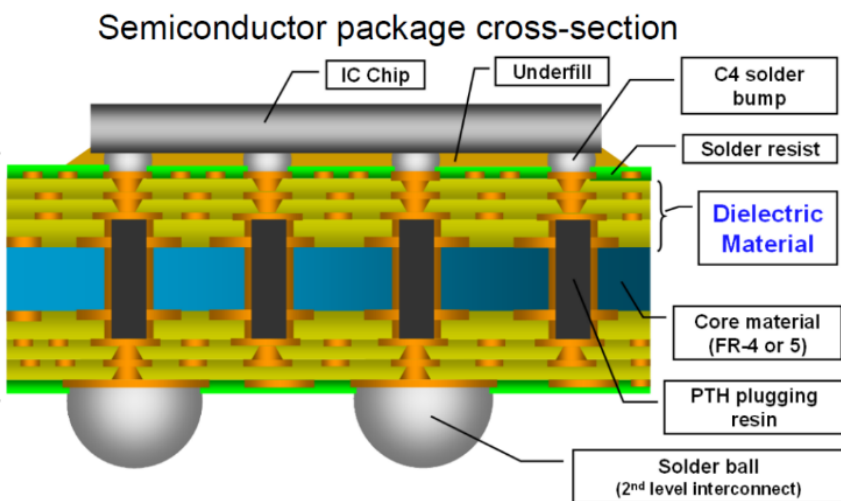
图：精智达收入利润（亿元）



3.3 先进封装材料：高端芯片国产替代先锋，材料国产化迫在眉睫

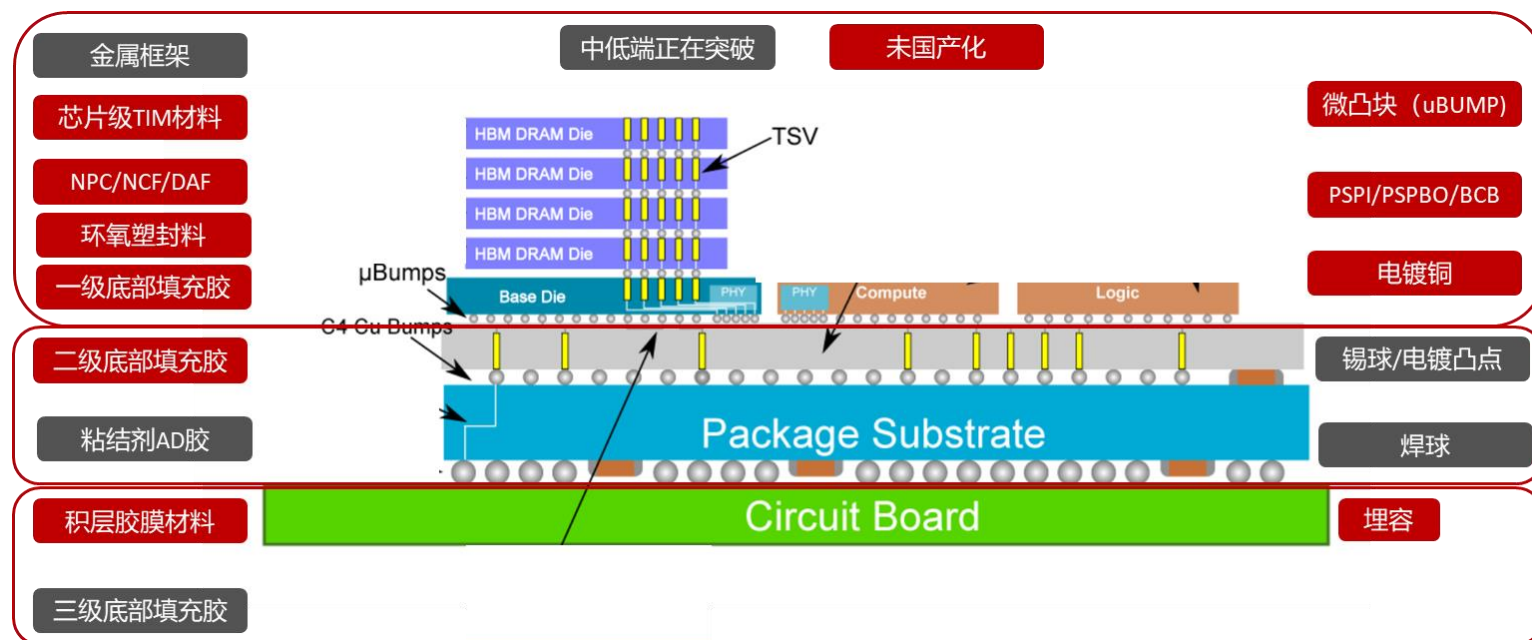
- **先进封装材料是先进封装产业链核心上游：**先进封装所呈现出轻薄化、高集成度等特点对材料提出了更高的性能要求，除了封装基板以及EMC以外，先进封装大大提高了对导热材料的性能需求，内部导热材料主要包括底填材料和TIM材料等，目前日本、美国厂商在中高端产品占有较大份额，材料国产化提速之下，具备前瞻性的技术布局的内资厂商将有望在未来的竞争中脱颖而出。
- **目前国产龙头厂商主要包括：**封装基板（兴森科技、深南电路）、环氧塑封料（华海诚科、飞凯材料、凯华材料）、电镀液（艾森股份、强力新材、天承科技、上海新阳）、Underfill（德邦科技、华海诚科）、PSPI（强力新材）、bumping厚胶（飞凯材料）、Low- α 球硅与球铝（联瑞新材、壹石通）、载板基膜（华正新材、宏昌电子）、锡膏（唯特偶）、临时键合材料（飞凯材料）等。

图：ABF载板结构



资料来源：味之素官网，elecfans，民生证券研究院

图：先进封装材料需求



资料来源：超能网，民生证券研究院

* 请务必阅读最后一页免责声明

3.3

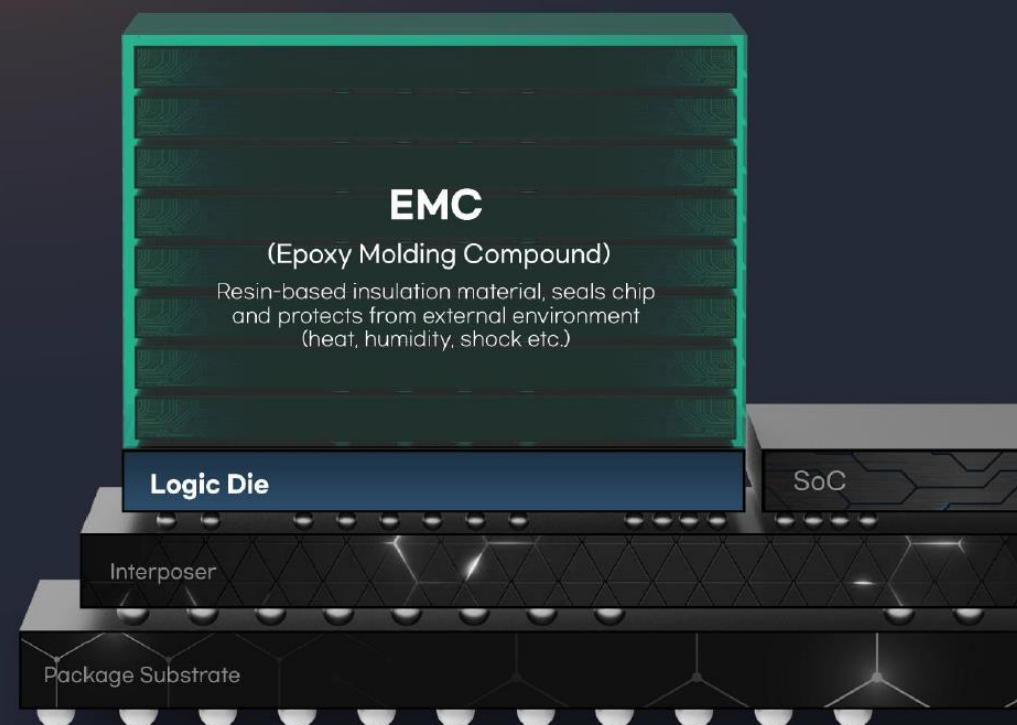
EMC: 海力士MR-MUF Process Flow核心耗材

MR-MUF* Process

Reflow & Molding done simultaneously at room temperature, achieves 3~4x thermal bumps for heat dissipation
(Low Stress 10N/Room-Temp → Reflow)

MR-MUF Molded UnderFill

- 1 Stacking of micro bump-attached chip by applying metal molding material
- 2 Melt all micro bumps at once, connecting chip and circuit
- 3 **Gapfill between chips or between substrate, insulation and molding done simultaneously**



*MR-MUF (Mass Reflow Molded Underfill)

We Do Technology
SK hynix

3.3 HBM拉动EMC价值量大幅提升

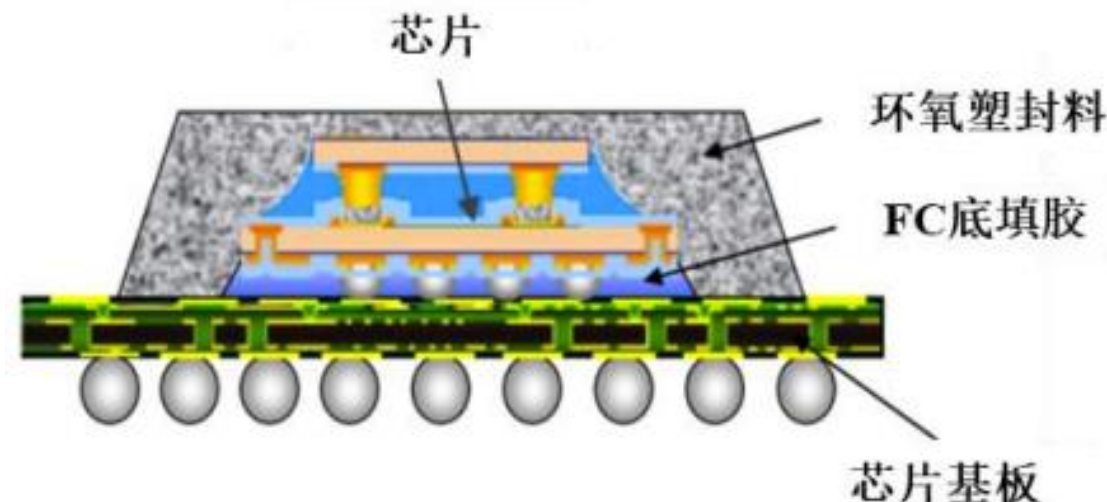
- **HBM拉动EMC和填料价值量大幅提升。**由于1颗HBM中将搭载4-8层甚至更多层的存储芯片，单颗HBM中所涉及到的数据存储和数据运输量较大，将带来发热量大的问题，散热方案的优劣将决定整颗HBM 性能，因此HBM的EMC中将开始大量使用low α 球铝和球硅来保证快速散热和控制热膨胀问题。目前华海诚科在先进封装领域GMC产品已经在多个客户考核通过，自主研发的GMC制造专用设备已经具备量产能力并持续优化；已经完成验证的芯片级底填正在做前期重复性量产准备，和最终客户协同开发的适用于“芯粒”封装的特殊性能底部填充胶正在认证考核；公司新购置一套LMC专用压缩模塑设备，进一步加快LMC产品的研发进度。
- **环氧塑封料的主要组成成分可分为聚合物、填料、添加剂三类。**聚合物主要包括环氧树脂、偶联剂、硬化剂等；填料主要由硅微粉和氧化铝；添加剂则主要包括脱模剂、染色剂、阻燃剂、应力添加剂、粘结剂等。各种成分中占比最大的两种为填料以及环氧树脂。

表：环氧塑封料主要成分及功能

	组成成分	常用化合物	主要功能	重量占比
聚合物	环氧树脂	邻甲酚型、联苯型	提供交联结构	5%-10%
	偶联剂	硅氧烷	增强结合	/
	固化剂	酚醛树脂型	提供交联结构	5%-10%
填料	硅微粉、氧化铝	二氧化硅、三氧化二铝	提高强度、降低热膨胀系数	60%-90%
添加剂	脱模剂	天然蜡	提供成形能力	<1%
	阻燃剂	/	阻燃	<10%
	催化剂	胺化物、磷化物	加速反应	<1%
	其他（染色剂、粘结剂）	/	染色、增强粘结力	<1%

资料来源：独角兽智库，民生证券研究院

图：环氧塑封料

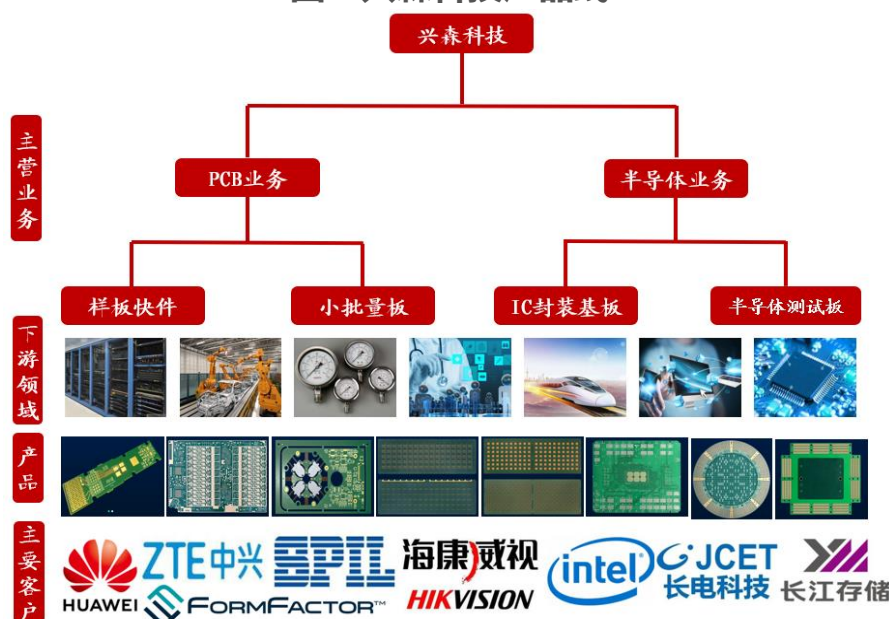


资料来源：华海诚科，民生证券研究院

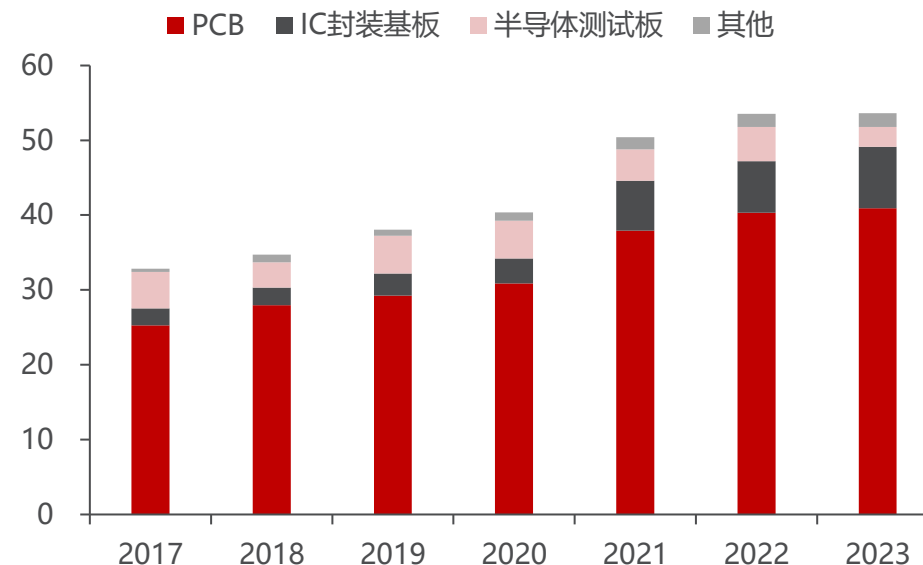
3.3 兴森科技：ABF封装基板国产替代先锋

- 兴森科技是国内PCB样板快件、批量板设计及制造领域的龙头企业，同时是国内本土 IC 封装基板行业的先行者之一，公司于2012年进入CSP 封装基板领域，并于2022年正式进军FCBGA封装基板领域。
- **公司是ABF封装基本领域国产替代先锋。**公司CSP封装基板业务在坚守存储芯片赛道的基础上，射频类产品顺利进入量产，打开了进军中高端市场的通道。公司FCBGA封装基板珠海项目已于2022年12月底建成并成功试产，预计2024年Q2开始量产；广州项目已于2023年Q4完成产线建设、进入试产阶段，目前工厂审核及产品认证工作按计划推进，预计2024年Q3开始量产。目前公司FCBGA封装基板良率迅速提升，低层板和高层板良率分别提升至90%和85%以上，预计24年底之前产品良率达到海外龙头企业同等水平。
- **估值：**预计公司24-26年归母净利润3.95/6.08/11.29亿元，对应现价PE 49/32/17倍（截至2024年6月17日收盘价），维持“推荐”评级。
- **风险提示：**下游需求不及预期，扩产进度不达预期，行业竞争加剧的风险。

图：兴森科技产品线



图：兴森科技收入结构（亿元）



资料来源：兴森科技公告，民生证券研究院整理

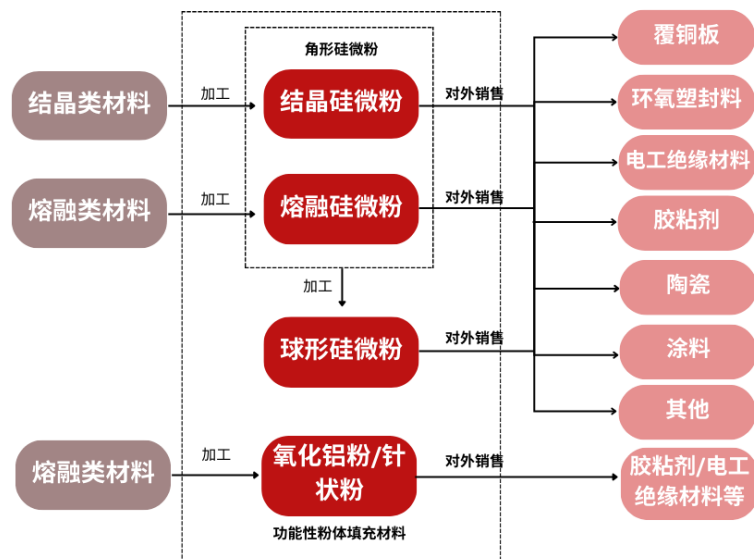
资料来源：Wind，民生证券研究院

3.3

联瑞新材：高端粉体领先企业，受益于HBM需求增长

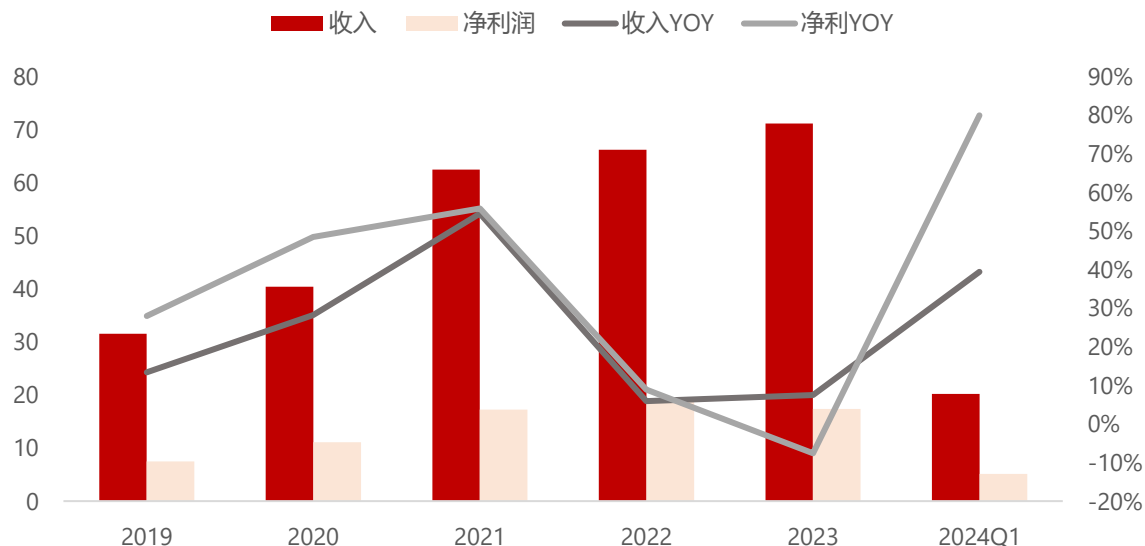
- **公司主营业务为硅微粉的研发、生产和销售。**主要产品包括结晶硅微粉、熔融硅微粉和球形硅微粉，已成为建滔集团、生益科技、住友电工、科化新材等行业知名企业的材料供应商；公司23年角形粉体、球形粉体与其他粉体材料营收占比分别为36.2%、45.9%、17.8%；24年Q1营收为2.02亿元，YoY+39.46%；归母净利润为0.52亿元，YoY+79.94%。
- **HBM上游关键材料供应商，受益于产能扩张。**HBM对封装材料要求越来越高，对粉体颗粒及性能要求也越来越高。对添加的超细粉体材料，需要用到Low α 球硅和Low α 球铝，并且伴随着HBM的技术迭代升级，对于填料的需求正不断向着更低CUT点推进。随着HBM产能扩张，公司凭借全套工艺技术和适应市场需求的产能成为受益者。
- **国内外AI需求强劲，带动硅微粉需求增长。**硅微粉在AI服务器中的应用主要集中在芯片覆铜板和封装环节，用来改善覆铜板的CTE、耐热性和可靠性，具有低损耗和低介电常数的优势。覆铜板向高频高速方向发展，市场对高性能硅微粉的需求将不断增加。
- **业绩与估值：**根据iFinD一致预期，预计2024-2026年归母净利润分别为2.3/3.0/3.6亿元人民币，对应PE分别为44/35/29倍（截至2024年6月17日收盘价）。

联瑞新材产品线



资料来源：联瑞新材招股书，民生证券研究院

联瑞新材2019-2024Q1营收及归母净利润（单位：千万元）



资料来源：Choice，民生证券研究院

3.4



01

存储：从HBM到CXL

02

PCB：服务器引领高端化浪潮

03

封测：先进封装解决算力瓶颈

04

制造：AI卖铲人

CONTENTS

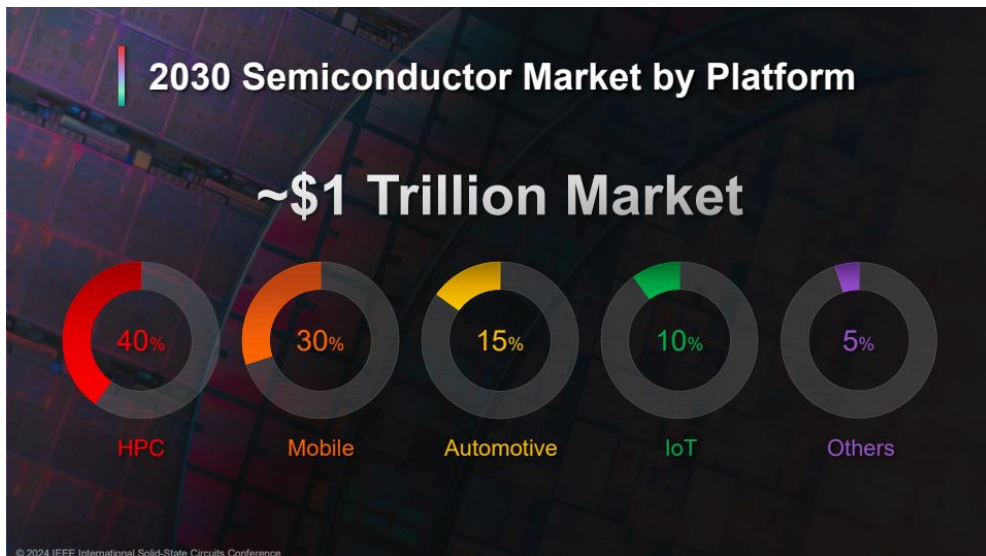
目录



3.4 晶圆代工：全球龙头看多AI需求，HPC显著拉动代工市场增长

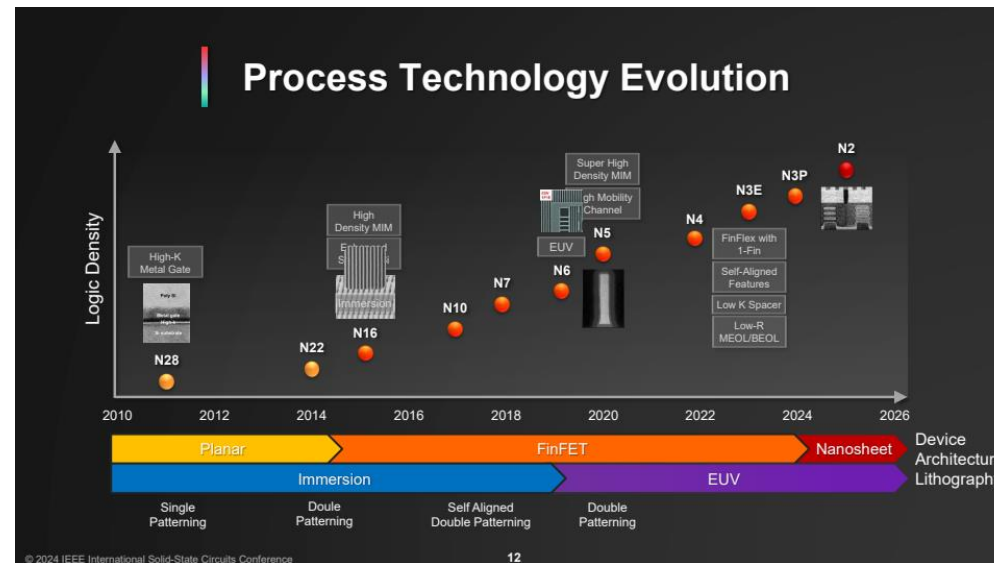
- **HPC将成为半导体下游第一大市场，2030年市场规模预计达到4000亿美金。** AI作为高性能计算应用的关键驱动力，随着生成式人工智能及语言大模型的复杂性不断增加，其计算需求将呈现指数级增长，将显著拉动半导体芯片需求。台积电预计到2030年，半导体市场将达到1万亿美金，其中高性能计算（HPC）占比40%，将成为第一大细分市场，显著拉动先进制程晶圆代工需求。
- **全球龙头对AI芯片需求展现出强烈增长预期。** 博通将AI芯片年度营收预期上调了10%；三星预测今年其AI芯片代工销售额将是去年的1.8倍，客户数量将比去年增加1倍，到2028年其AI芯片代工的客户数量将比2023年增长4倍，代工销售额也将实现9倍的增长；台积电预计到2028年，AI芯片代工业务的收入贡献将显著增加，预计占公司总收入的比重达到20%；
- **AI芯片性能提升主要依赖制程技术。** 台积电计划于2025年量产N2制程（2nm）芯片，2027~2028年量产A14（1.4nm）芯片；三星同样预计2025年量产2nm芯片，2027年量产1.4nm芯片；英特尔则预计在2024年下半年量产Intel 18A，2027年前开发出Intel 14A。

2030年全球半导体市场规模（按平台分）



资料来源：TSMC IEEE 2024，电子工程世界，民生证券研究院

台积电制程升级



资料来源：TSMC IEEE 2024，电子工程世界，民生证券研究院

3.4

台积电：AI芯片卖铲人，先进工艺遥遥领先

- **台积电作为AI芯片代工的卖铲人，先进工艺遥遥领先。**目前代工工艺主要包括N3家族 / N2 制程 / NanoFlex / A16 / 超级电轨 / CFET，先进封装包括SoW / 3DFabric / SolC (&Hybrid bonding) / CoWoS/InFo。
- **3nm制程：**N3（3nm级）工艺技术系列由多种变体组成，包括基线N3（又名N3B）、降低成本的宽松N3E、具有增强性能和芯片密度的N3P以及具有更高电压容差的N3X。N3E 去年第四季进入量产，至于今年下半年准备量产的N3P，良率表现接近N3E。由于N3P在效能、功耗、面积（PPA）表现更优异，大多数3nm产品都将采用N3P制程技术。
- **2nm制程：**N2（2nm级）技术开发步入正轨并取得良好进展。N2技术采用了该公司第一代纳米片晶体管技术，在性能和功耗方面实现了全节点的进步。目前2纳米技术进展顺利，纳米芯片转换表现达到目标90%、转换成良率也超过80%，预计2025年量产。
- 除此以外，A16 技术将使用下一代纳米片技术结合超级电轨（Super Power Rail）架构，预计2026 年下半年量产。这次会采用不同布线，台积电认为这是高效能运算（HPC）产品的最佳解决方案。相较于N2P 制程，使用超级电轨的A16 在相同Vdd（工作电压）下，运算速度增加8~10%，在相同速度下，功耗降低15~20%，芯片密度提升高达1.10X。

N3E PPA（与 N5 v1.0 对比）		
技术：	N3P(参考: N3E)	N3X(参考: N3P)
速度：	+5%（在相同漏电下）	+5%（在1.2伏电压下的最大频率）
功耗：	-5~-10%（在相同速度下）	约3.5倍漏电
芯片密度：	1.04倍	1.04倍

资料来源：台积电官网，图灵算力研究院，民生证券研究院

N2 PPA（相对于N3E V1.0）	
相同功耗下的速度提升：	10~15%
相同速度下的功耗降低：	25~30%
芯片密度：	> 1.15倍

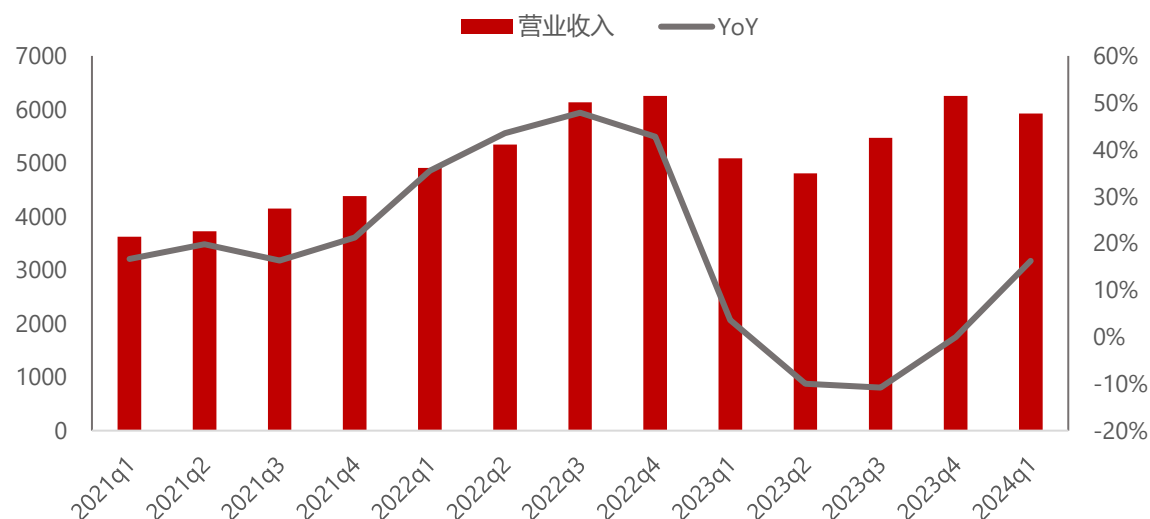
资料来源：台积电官网，图灵算力研究院，民生证券研究院

3.4

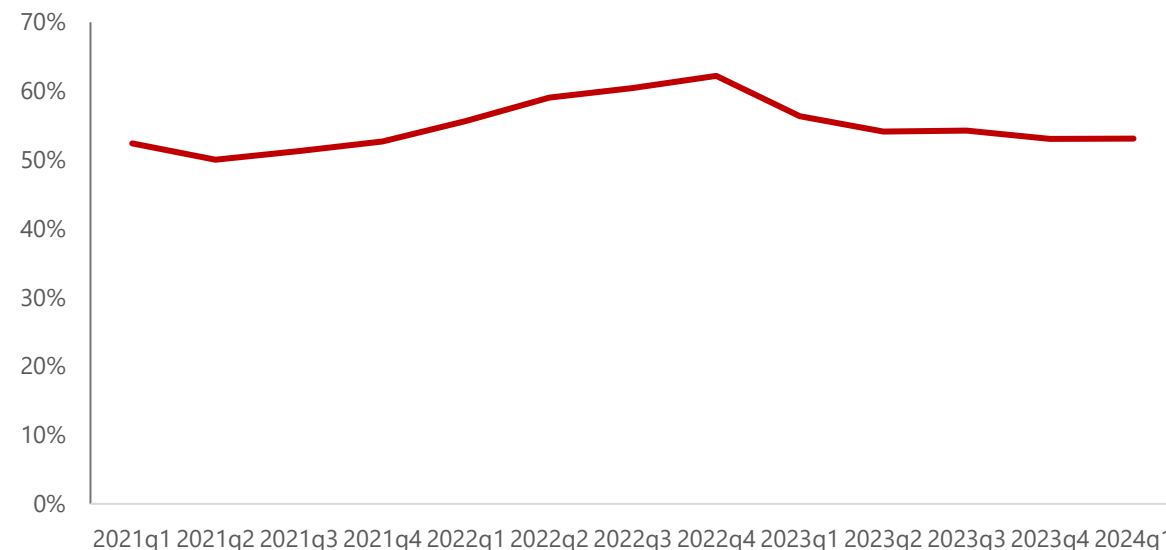
台积电：Q1业绩超预期，看好AI需求增长

- **Q1业绩同比高增，4月营收创历史次高。** Q1实现营收5926.4亿新台币（188.7亿美元），同比增长12.9%，略超指引上限；毛利率为53.1%，符合指引；4月实现营收2360.2亿新台币，月环比增长20.9%，同比增长59.6%，为同期新高、历史次高。公司1~4月累计营收达8286.7亿新台币，同比增加26.2%。
- **Q2和全年业绩指引：** Q2营收以美元计算预计增长3.9%~8%，对应区间196~204亿美元，毛利率为51%~53%；24年全年资本开支为280~320亿美元。
- **下修24年半导体行业增速，指引AI需求持续强劲。** 由于PC、智能手机需求复苏但增长缓慢等原因，台积电下修2024年全年半导体行业增速至10%（此前预计超10%），下修代工行业增速至15~20%（此前预计超20%）；受益于AI的强劲需求，AI处理器逐渐成为高性能计算平台增长的最强驱动力。公司指引AI处理器收入将同比翻倍以上增长，未来5年CAGR将达50%，至2028年收入占比将达到20%。

2021-2024季度营收（亿台币）及增速



2021~2024季度毛利率



资料来源：choice，民生证券研究院

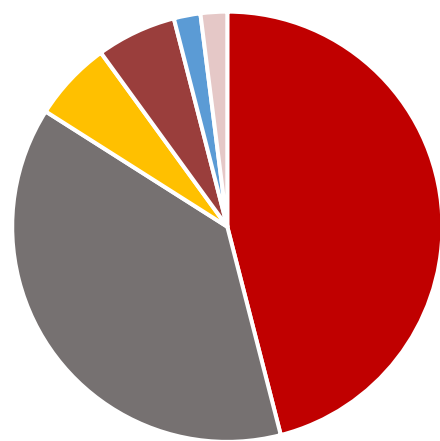
资料来源：choice，民生证券研究院

3.4

台积电：HPC业务占比近一半，3/5nm稼动率接近满载

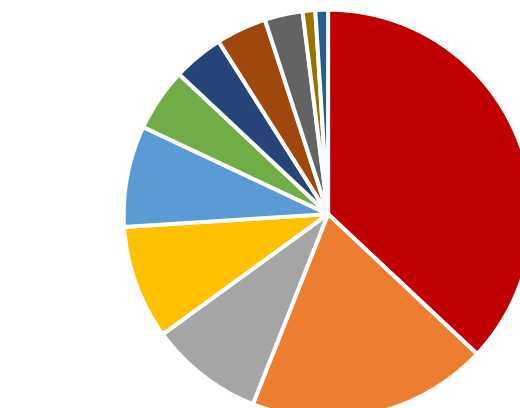
- **先进制程占比逐年攀升，HPC占比46%**：3nm制程占第一季度晶圆收入的9%，而5nm和7nm制程分别占37%和19%，7nm及以下的先进制程占晶圆收入的65%；**按照应用领域**：高性能计算（HPC）一季度收入环比23Q4增长3%，占Q1总收入的46%；智能手机受到季节性影响，一季度收入环比23Q4减少16%，占Q1总收入的38%；物联网一季度收入环比23Q4增长5%，占Q1总收入的6%。汽车收入环比持平，占Q1总收入的6%；DCE一季度收入环比23Q4增长33%。
- **稼动率满载，AI需求火热亟待增产**。据DigiTimes报道，台积电的4/5nm工艺在2023年底的产能利用率就已接近90%，2024年第一季度更是达到了满载，需求主要来自智能手机与AI芯片在内的HPC客户。同时3nm工艺的订单量也优于预期，产能利用率从2023年底的75%拉升至现在的95%。
- **科技巨头陆续导入3nm，涨价或将启动**。全球七大科技巨头（英伟达、AMD、英特尔、高通、联发科、苹果及谷歌）将陆续导入台积电3nm 制程，例如高通骁龙8Gen 4、联发科天玑9400 及苹果A18、M4 系列都将采用N3家族打造，其中基于N3E工艺的高通骁龙8Gen4已率先开始涨价，较上一代报价激增 25%，预计将超过 250 美元。
- **估值**：根据iFinD一致预期，预计24-26年归母净利润分别为319.8、387.3、450.4亿美元，对应PE分别为28、23、20倍（截至2024年6月17日收盘价）。

2024年Q1营收占比（按平台分）



■ 高性能计算 ■ 智能手机 ■ 物联网 ■ 汽车 ■ 数字消费电子 ■ 其他

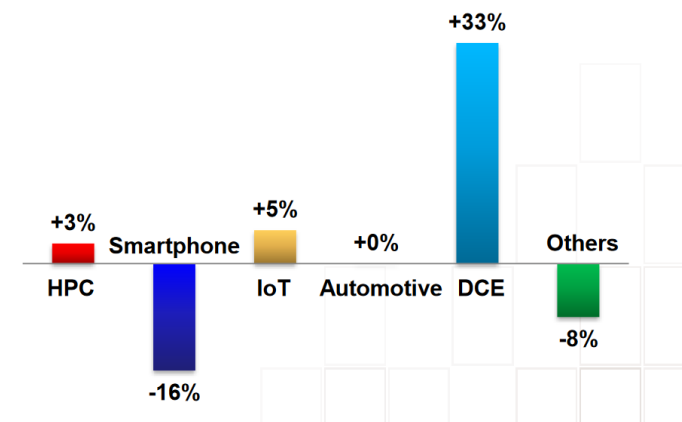
2024年Q1营收占比（按制程分）



■ 5nm ■ 7nm ■ 3nm ■ 16nm
■ 28nm ■ 40/45nm ■ 65nm ■ 0.15/0.18um

下游应用环比增长

Growth Rate by Platform (QoQ)

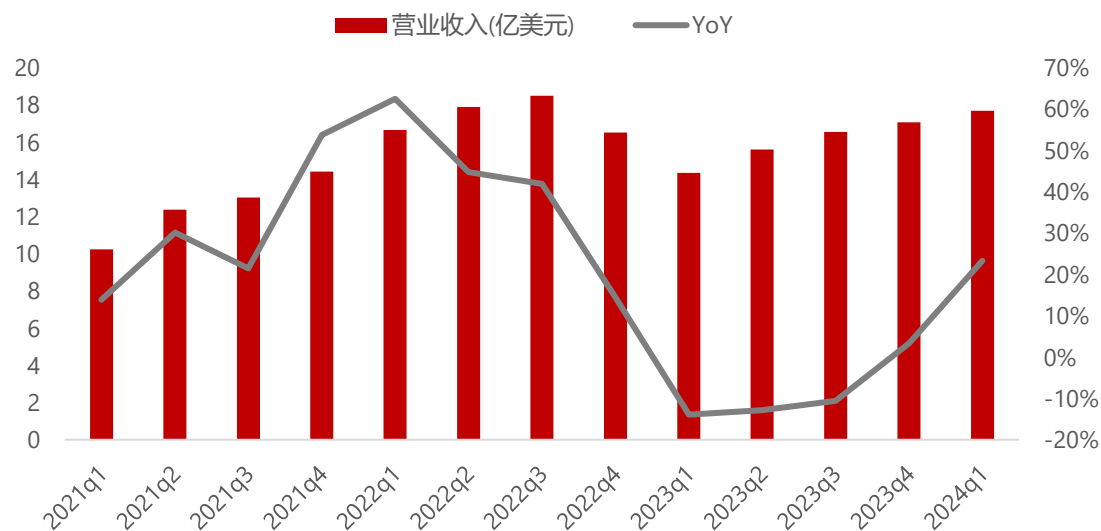


资料来源：台积电官网，民生证券研究院

3.4 中芯国际：行业触底回升，业绩逐步复苏

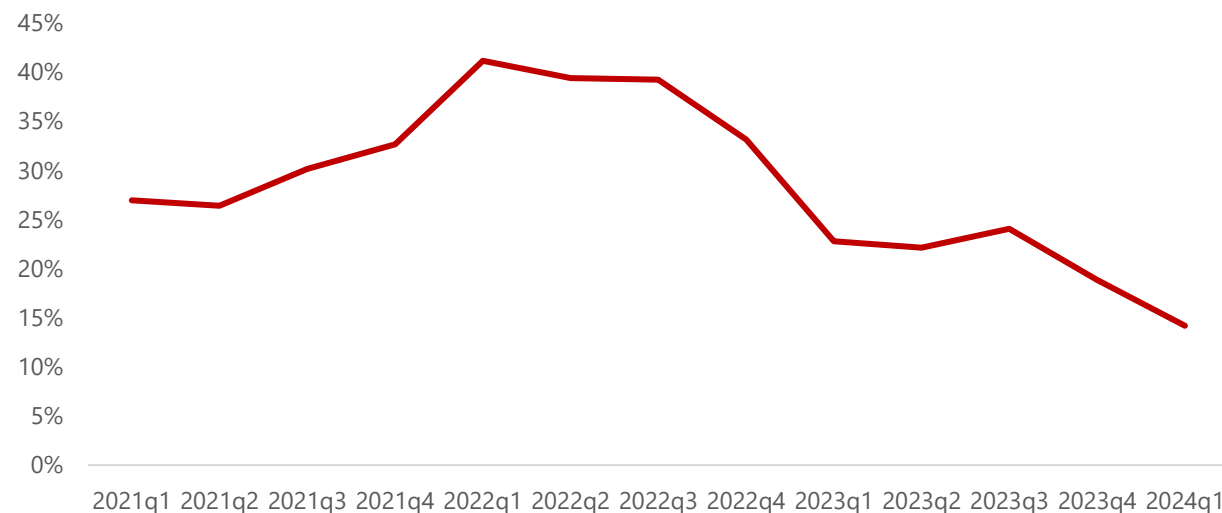
- **业绩复苏：**中芯国际24Q1营业收入为17.5亿美元，环比增长4.3%，同比增长19.7%，实现连续四个季度爬升；毛利率为13.7%，好于指引。在经历了2023年的复苏期后，中芯国际基本回到了稳步增长的状态。按24Q1收入计算，中芯国际在全球芯片代工市场占据的份额为6%，超越联电和GF首次位居全球第三。
- **资本开支：**24Q1资本开支为22.4亿美元，研发开支1.88亿美元。2023年，公司资本开支为74.7亿美元，主要用于成熟产能扩产以及新厂基建，稳步推进四个成熟制程12英寸工厂的产能建设。
- **2Q24指引：**部分客户的拉货需求还在持续，收入环比增长5%-7%，伴随产能规模扩大，折旧逐季影响下毛利率介于9%到11%之间。

2021-2024中芯国际季度收入及增速



资料来源：IFind，民生证券研究院

2021~2024季度销售毛利率

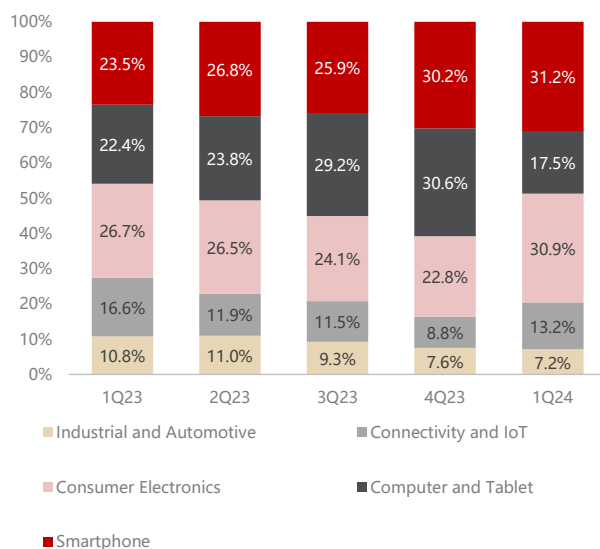


资料来源：IFind，民生证券研究院

3.4 中芯国际：行业触底回升，业绩逐步复苏

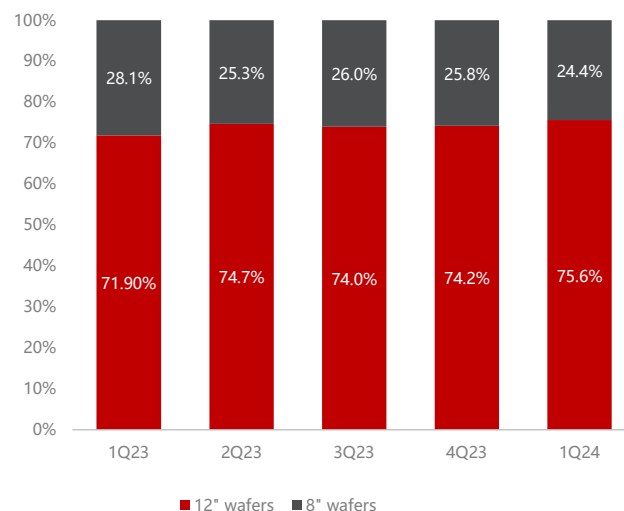
- **营收结构：**收入按尺寸拆分，12寸占比进一步提高，24Q1营收占比达到75.6%；按照下游应用领域拆分，手机业务、消费电子业务的占比分别从去年同期的23.5%、26.7%提升为31.2%、30.9%，增长明显。市场回暖主要表现在：1) 低功耗蓝牙，MCU急单增多；2) 体育年导致机顶盒，电视等销售起量；3) 中国厂商智能手机快速备货。
- **稼动率修复好于预期：**稼动率呈现环比回升趋势，24Q1稼动率约80.8%，环比增加4.0pcts，同比提高12.7pcts，下游客户需求持续回暖。
- **价格：**12寸在24年2月以来一直处于满载状态，价格较为稳定，ASP整体变化不大。8英寸价格目前处于低谷，预计25年中恢复健康水平。
- **估值：**我们预计公司24-25年营收分别为522.6/623.5亿元，归母净利润分别为60.3/74.5亿元，所对应PE估值分别为63/51倍，维持“推荐”评级（截至2024年6月17日收盘价）。
- **风险提示：**下游需求复苏不及预期；产能爬坡不及预期；先进工艺研发不及预期；地缘政治加剧。

图：中芯国际营收结构（按应用）



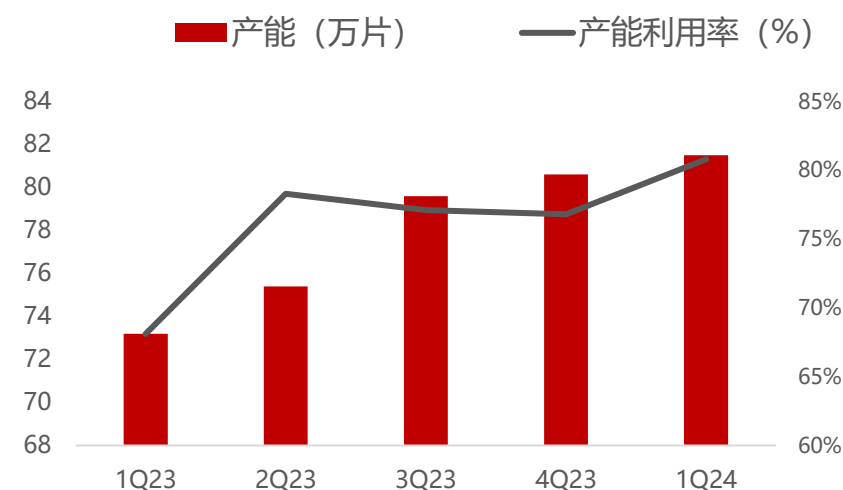
资料来源：中芯国际官网，民生证券研究院

图：中芯国际营收结构（按晶圆尺寸）



资料来源：中芯国际官网，民生证券研究院

图：中芯国际产能利用率（折合8英寸/万片）

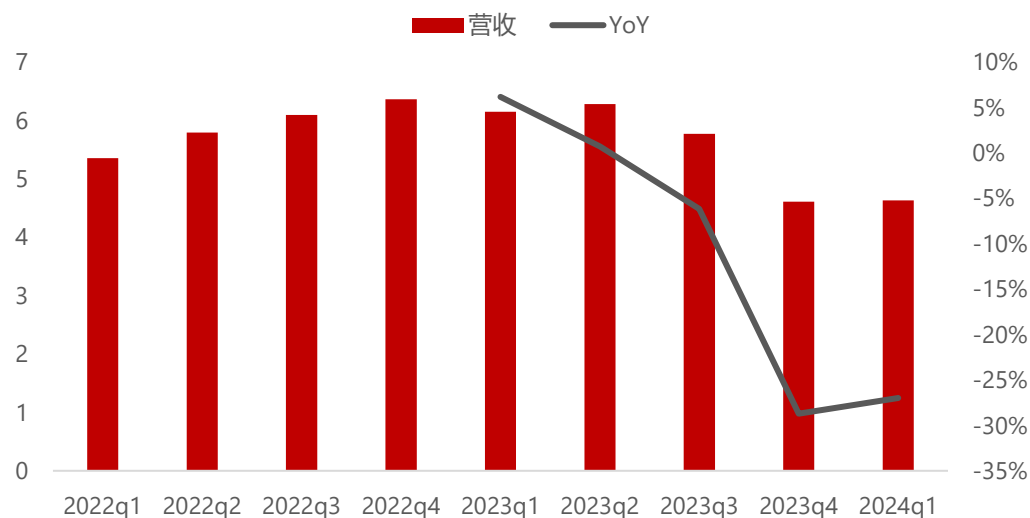


资料来源：中芯国际官网，民生证券研究院

3.4 华虹公司：业绩拐点明确，需求改善有望带动价格回升

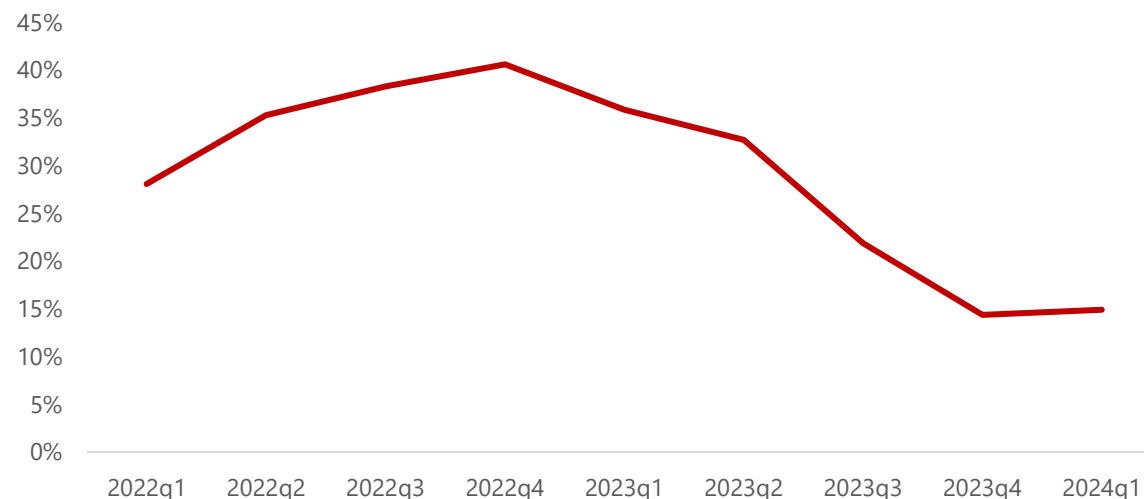
- **业绩情况：**华虹半导体一季度营收4.6亿美元，环比增长1%，符合指引；毛利率6.4%，环比改善2.4个百分点，略高于指引，主要由于产能利用率提高。
- **资本开支：**24Q1公司资本支出为3.026亿美元，其中1.994亿美元用于华虹制造，0.714亿美元用于华虹无锡，0.317亿美元用于华虹8英寸线。第一条12英寸生产线今年月产能将在9.45万片，第二条12英寸生产线也在建设中，预计三季度完成建设，四季度试生产。
- **2Q24指引：**考虑到电源管理和CIS产能满载、nor flash和低压Mos需求改善，但MCU复苏仍在早期、高压功率需求仍然疲软。公司预计销售收入约在 4.7亿美元至 5.0 亿美元之间，对应环比增长2%-9%，毛利率在6%至10%之间。

图：华虹公司2021Q1-24Q1季度收入及增长率（单位：亿美元）



资料来源：iFind，民生证券研究院

图：华虹公司2022-2024季度毛利率



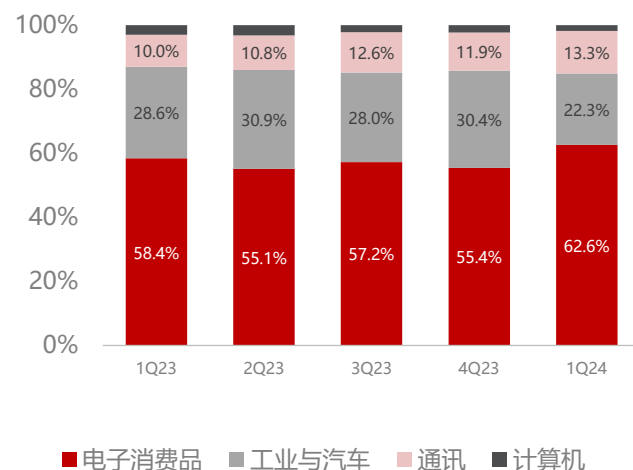
资料来源：iFind，民生证券研究院

3.4

华虹公司：业绩拐点明确，需求改善有望带动价格回升

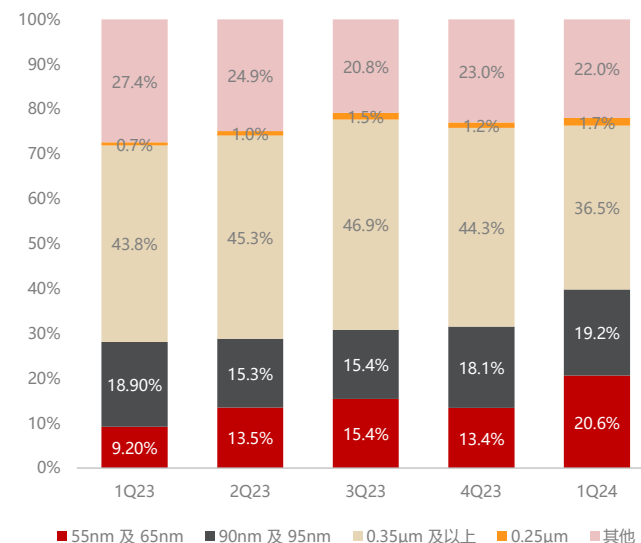
- **营收结构：**通讯业务、消费电子业务环比有较大增长，主要受益于智能手机、电源管理等领域的复苏和需求的提升。从技术节点来看，第一季度华虹半导体在55nm及65nm工艺节点实现61.5%的同比增长，在0.25μm工艺技术节点实现83.7%的同比增长，前者主要得益于CIS及其他电源管理产品的需求增加，而0.25μm工艺技术主要得益于逻辑产品的需求增加。
- **电源管理与CIS需求带动公司产能利用率提升：**Q1综合产能利用率91.7%（12寸84.2%，8寸100.3%），QoQ+7.6pcts，环比提升显著。
- **估值：**当前港股、A股PB分别为0.82、1.50；根据iFinD一致预期，24-26年净利润分别为18.4、20.9、28.3亿人民币，对应PE分别为36、31、23倍（截至2024年6月17日收盘价）。

图：华虹半导体营收结构（按应用）



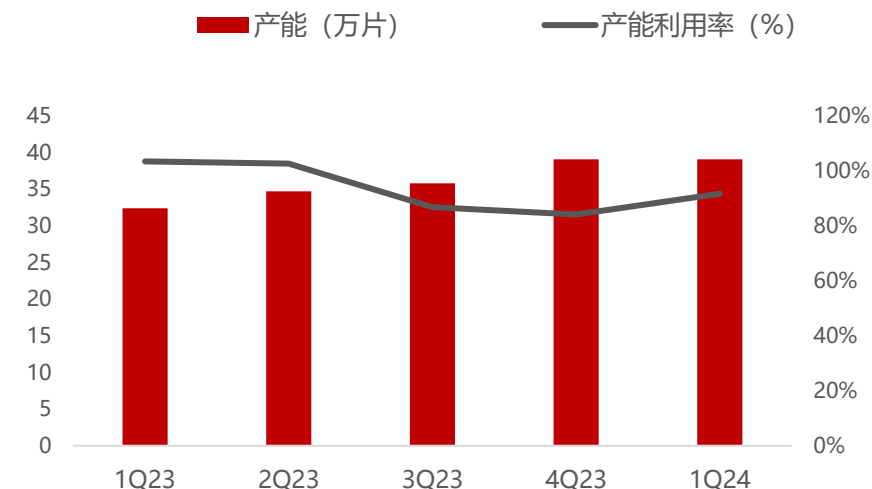
资料来源：华虹半导体，民生证券研究院

图：华虹半导体营收结构（按技术节点）



资料来源：华虹半导体，民生证券研究院

图：华虹半导体产能利用率（折合8英寸/万片）



资料来源：华虹半导体，民生证券研究院

04. 端侧变革：身边的算力终端

4.1



01

传统终端的困境与破局

02

详解苹果公司AI终端布局

03

AI 终端三要素

04

AI 终端产业链

CONTENTS

目录

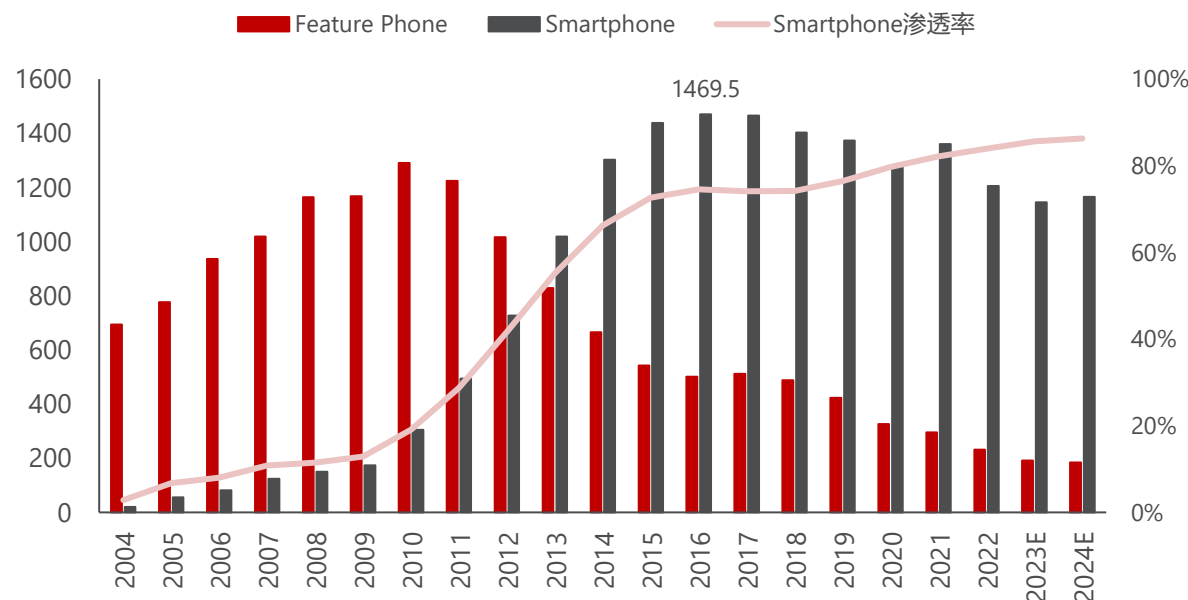


4.1

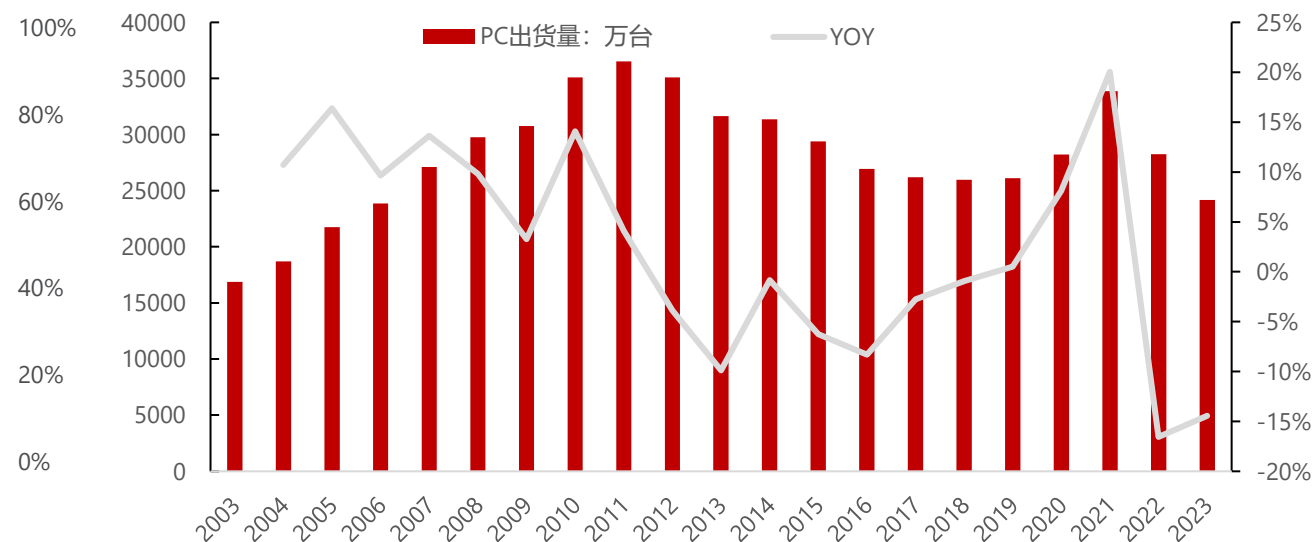
传统终端困境：步入存量市场，增长动能不足

- **手机：全球智能机销量于2016年达14.7亿部见顶，此后逐年缓降。**2023年全球智能手机出货量同比下降3.2%，降至11.7亿部，为近十年来最低。其中23Q4同比增长8.5%，出货量达到3.3亿台。据IDC预测，考虑到换机周期叠加AI大模型技术的应用以及新兴市场的增长，预计2024年全球智能机销量将回升至近12亿部水平。从市占率看，全球智能机品牌集中度逐年提升，CR5占比从14年的51.4%提升至23年的67.8%。
- **PC：**随着移动互联网时代的开启，部分工作、娱乐活动由PC切换至手机，PC销量在2011年达到3.65亿台见顶，随后进入大下行周期，直到17-19年企稳。居家办公初期明显拉动PC需求，21年PC销量达到近10年来的峰值3.4亿台，但提前购机透支了未来的需求，叠加21年供应链紊乱导致的高备货高库存，PC进入了漫长的去库阶段，22-23年销量出现大幅下跌。

图：2008-2024E全球智能机出货量及渗透率（亿部）



图：2003-2023全球PC出货量



4.1

传统终端破局：借力AI迎接下一波成长

- 当前，无论从销量或市占率角度，传统电子终端都进入了存量市场。为了破局，各大品牌厂商纷纷发力AI，智能终端时代加速。2023 下半年起，以谷歌、三星、OPPO 等为代表的厂商将大模型内置于智能手机中，完成从云端 AI 向终端 AI 的转变，“AI 手机”浪潮开启。苹果也于2024年6月份WWDC发布个人智能化系统Apple Intelligence，同时正式宣布与OpenAI合作，智能助手Siri接入ChatGPT，苹果首款AI手机iPhone16有望于今年秋季发布。
- 下文中，我们将借Apple的AI布局，解读AI终端的产业趋势。

图：全球主流端侧AI设备梳理

终端	厂商	机型	发布时间	定价	大模型	AI功能
AI手机	谷歌	Pixel 8	2023年10月4日	699美元起	Genimi Nano	照片编辑、优化影音录制、即时翻译
	三星	Galaxy S24	2024年1月18日	799美元起	云侧Gemini pro，端侧Genimi Nano	图片即圈即搜、实时双向翻译、智能图片编辑
	华为	Mate 60	2023年9月4日	5499元起	盘古大模型	AI云增强等功能
	小米	14 Ultra	2024年2月22日	6499元起	LiLM-6B	AI大模型计算摄影平台Xiaomi AISP，具有文生图、AI扩图等功能
	VIVO	X100	2023年11月16日	3999元起	蓝心大模型	AI拍照、Jovi AI助理等
	OPPO	Find X7	2024年1月8日	3999元起	小布助手	AIGC 消除、AI 通话摘要、AI 画师、小布照相馆、小布连麦等
	苹果	WWDC2024官宣与OpenAI合作，Siri接入ChatGPT，iPhone16有望成为首款AI手机				

4.2



01

传统终端的困境与破局

02

详解苹果公司AI终端布局

03

AI 终端三要素

04

AI 终端产业链

CONTENTS

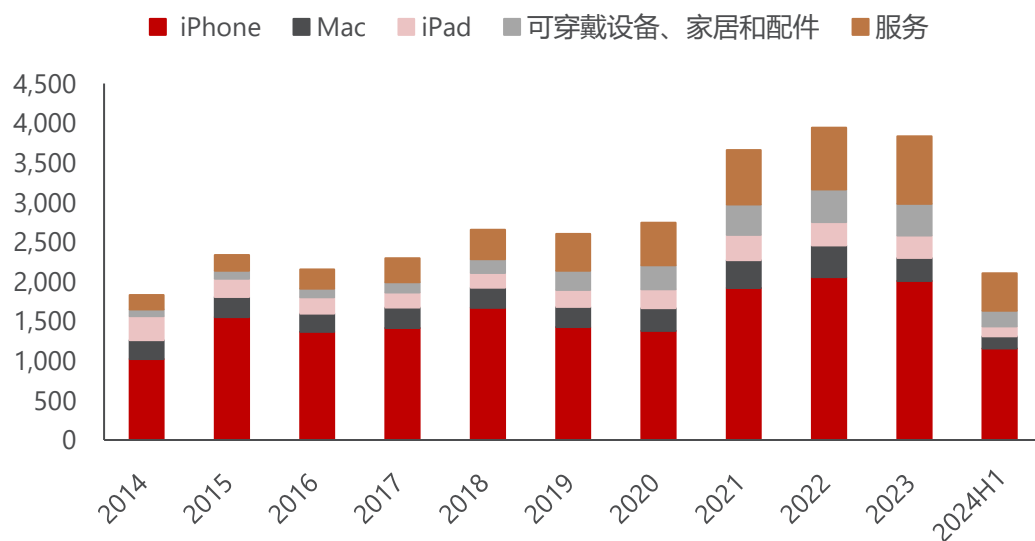
目录



4.2 苹果公司：可穿戴+服务战略卓有成效

- 苹果作为全球科技龙头，在手机、PC、可穿戴、MR等多个领域推出极具创新性的产品，23年智能手机市占率以20%位居全球首位。近年来，公司持续完善生态，从硬件公司向软硬结合的平台型公司发展。
- **业绩方面：**苹果FY24Q2实现营收908亿美元（YoY -4.31%），毛利率：46.58%（YoY+2.32pct），整体高于彭博一致预期；
- **从收入结构来看：硬件方面，**iPhone 为主要销售收入来源，基本占比五成以上，23年，iPad、Mac 和可穿戴设备的营收占比基本为 7%、7%和 10%。**软件方面，**近年来公司服务业务营收占比不断增长，24H1服务业务营收占比达22%。
- **服务业务卓有成效：**公司的服务业务涵盖了广告授权、AppleCare、云业务、数字内容和支付业务五大板块，服务业务营收占比从15年的8.5%提升至24H1的22%。

图：2014-24H1苹果分业务营收（亿美元）



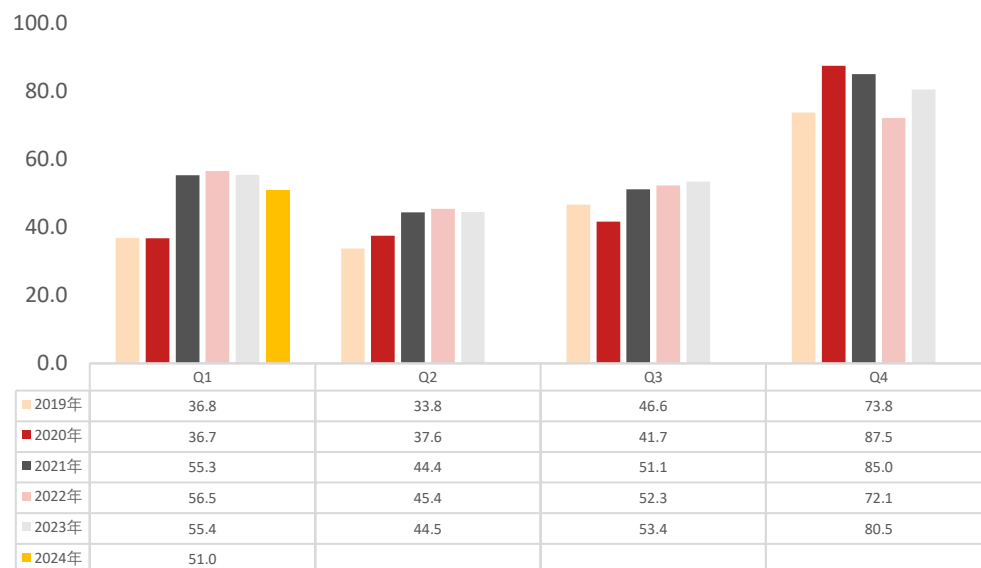
图：苹果软件服务业务板块



4.2 迎接AI晨曦，苹果开启下一十年

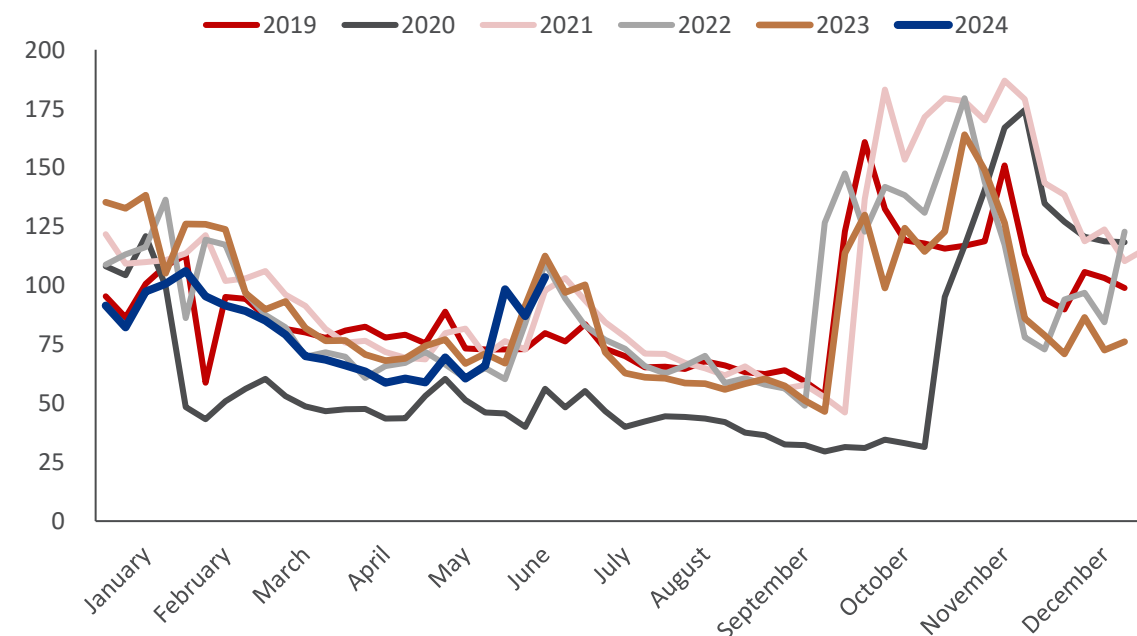
- 相较可穿戴及服务，iPhone作为公司支柱业务，在前期表现相对平淡。国内市场，iPhone受国产高端机型冲击，销量有所下滑。海外市场，三星S24爆火一定程度上也抢占了市场关注度。
- 作为应对，苹果开启了一系列“组合拳”，近期销量有明显改善。国内市场受益“618”降价促销，iPhone促销首周销量表现亮眼，跃居第一，W21（5.20-5.26）为98.3万部（同比+46.8%，环比+49.0%）。同时，公司作为全球最有底蕴的消费电子品牌，正全面拥抱AI，通过生成式人工智能再一次改变用户的终端使用习惯。

图：2019-2024年iphone分季度销量（百万台）



2019年 2020年 2021年 2022年 2023年 2024年

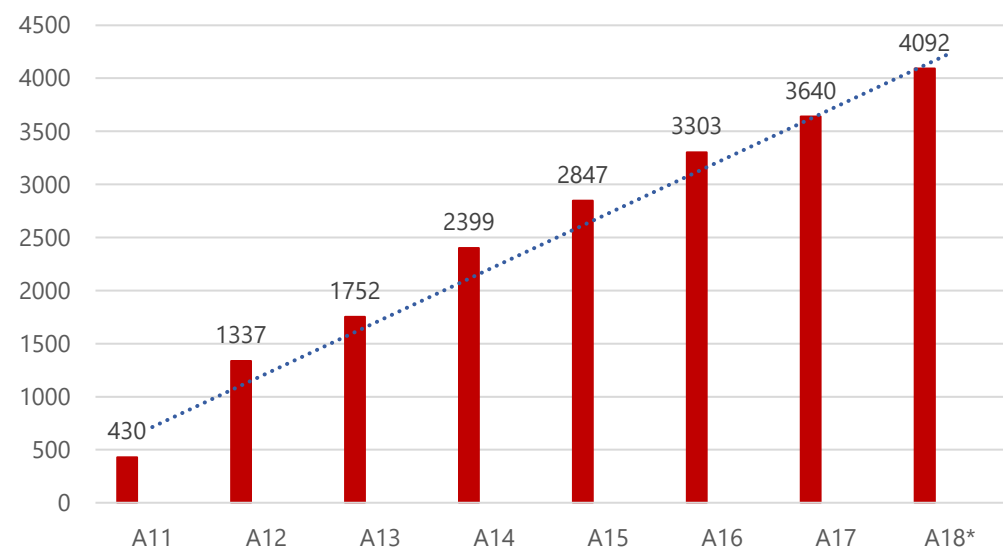
图：2019-2024年iphone周度销量（万台）



4.2 算力：苹果A+M组合拳前瞻布局

- **A系列处理器：**苹果于23年9月发布A17 Pro，作为业界首款 3nm芯片，可实现 35 TOPS。A17 Pro采用 6 核CPU，性能核心提升最快可达 10%，内部搭载 16 核AI神经引擎，神经网络引擎运行机器学习模型的速度提升最快可达上一代的 2 倍，算力可达 35TOPS，可支持iOS 17 中的自动纠错、从照片背景中提取主体、为有失语风险的用户创建个人语音等功能。
- **M系列处理器：**苹果于23年10月发布超先进个人电脑芯片 M3、M3 Pro 和 M3 Max，为首批采用业界领先 3nm工艺打造的个人电脑芯片；24年5月，M4芯片（胜任AI任务的超强芯片）发布，基于第二代3nm工艺打造，拥有全新显示引擎，CPU速度相较M2最高提升50%，并拥有动态缓存功能，搭载NPU，运算速度达每秒38万亿次。

图：手机A系列芯片GeekBench ML NPU得分



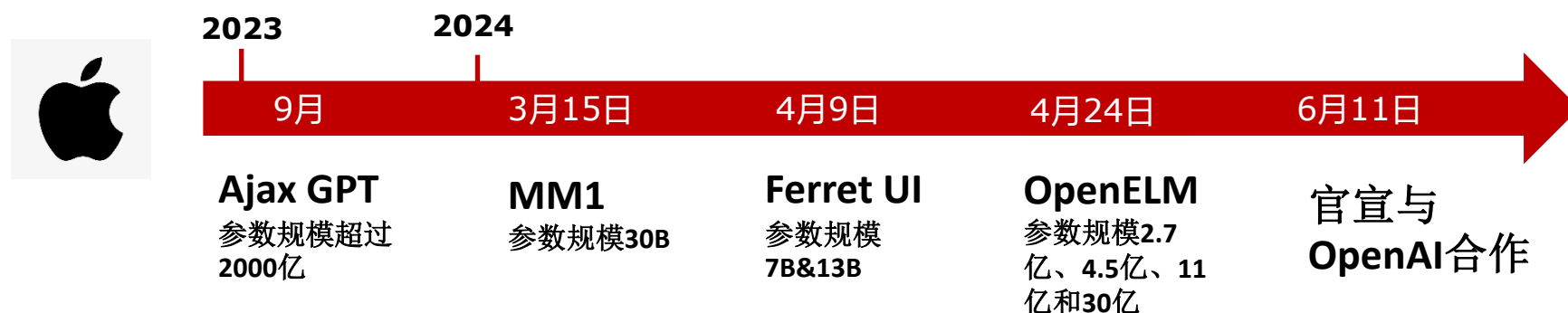
表：M系列芯片参数性能

	M1	M2	M3	M4
晶体管数量	160亿	200亿	250亿	280亿
制程工艺	台积电N5	台积电N5P	台积电N3B	台积电N3E
CPU性能核心	4核心(Firestorm) 12MB共享L2缓存 128KB数据缓存 192KB指令缓存 3.2GHz主频	4核心(Avaianche) 16MB共享L2缓存 128KB数据缓存 192KB指令缓存 3.5GHz主频	4核心(Unknown) 12MB共享L2缓存(Maybe) 128KB数据缓存 192KB指令缓存 4.1GHz主频	3/4核心(Unknown) 16MB共享L2缓存(Maybe) 128KB数据缓存 192KB指令缓存 4.42GHz主频
CPU能效核心	4核心(Icestorm) 4MB共享L2缓存 64KB数据缓存 128KB指令缓存	4核心(Bizzard) 4MB共享L2缓存 64KB数据缓存 128KB指令缓存	4核心(Unknown) 4MB共享L2缓存 64KB数据缓存 128KB指令缓存	4核心(Unknown) 4MB共享L2缓存 64KB数据缓存 128KB指令缓存
CPU单核心/多核心性能	2352/8395	2633/9747	3055/11690	3P+6E:3797/13346 4P+6E:3832/14905
GPU	7核心/8核心(2.6TFLOPS)	8核心/10核心(3.6TFLOPS)	8核心/10核心 (4.4TFLOPS(Maybe)) 全新架构/网格着色/硬件光追	10核心(4.4TFLOPS(Maybe)) 全新架构/网格着色/硬件光追
GPU OpenCL/Metal跑分	7核心GPU:20232/31996 8核心GPU:18232/29566	8核心GPU:28012/45584 10核心GPU:25257/41397	8核心GPU:30331/47214 10核心GPU:25697/41002	10核心GPU:54189(Metal)
NPU	16核心, 11TOPS	16核心, 15.8TOPS	16核心, 18TOPS	16核心, 38TOPS

4.2 模型：自主研发&三方合作齐头并进

- 苹果深耕AI项目多年，2010年花费1.5-2.5亿美元收购Siri，就被普遍视为乔布斯时代最重要的AI投资。截止2023年底，苹果收购AI初创公司达32家，高于谷歌的21家，Meta的18家及微软的17家。苹果在硬件设计、操作系统、生态系统等方面都有着深厚的积累，这为其在AI领域的探索提供了独特的优势。**目前，苹果在大语言模型、多模态模型等多领域均有进展：**
- **1、Ajax GPT：**The Information报道称，苹果的大语言模型内部代号是Ajax GPT，其参数规模超过2000亿；
- **2、MM1：**具有高达 30B 参数的多模态 LLM 系列；
- **3、Ferret UI：**Ferret UI多模态模型（参数包括7B和13B）具备强大的图片区域识别能力，可应用在手机UI之中，通过优化，使它能够更好的识别手机应用的界面。并将自然语言翻译为界面操作点；
- **4、OpenELM：**2024年4月24日，苹果开源大语言模型OpenELM，该模型有指令微调和预训练两种模型，一共有2.7亿、4.5亿、11亿和30亿4种参数，预训练数据达1.8万亿tokens，可提供生成文本、代码、翻译、总结等功能；
- **5、与OpenAI合作：**6月11日，苹果WWDC2024召开，正式宣布与OpenAI合作，智能助手Siri接入ChatGPT。

图：苹果 AI模型进展



4.2 生态：WWDC 2024揭开面纱

- **WWDC 2024召开，苹果端侧AI系统发布。**北京时间6月11日凌晨1点，苹果WWDC2024召开，发布个人智能化系统Apple Intelligence。苹果正式宣布与OpenAI合作，智能助手Siri接入ChatGPT。此外，WWDC大会延续之前惯性，发布新一代的iOS18、watchOS11、macOS15、iPadOS18、tvOS18等六大操作系统。

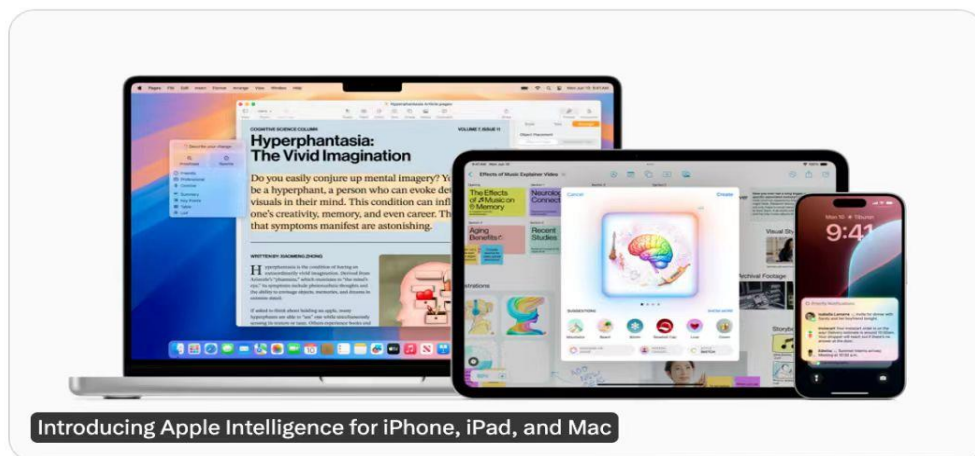
图：苹果个人智能化系统Apple Intelligence



Tim Cook
@tim_cook

It's personal, powerful, and private—and it's integrated into the apps you rely on every day.

Introducing Apple Intelligence—our next chapter in AI.



图：苹果六大操作系统



4.2 生态：WWDC 2024揭开面纱

- 苹果AI核心在于，聚焦系统级产品，构建横跨自家数据，软件，硬件的智能助理系统，而非特定功能或者服务。“Apple Intelligence”定位个人智能系统，并将大模型集成到 iPhone、iPad 和Mac等全生态。
- 具体到应用体验层面，iOS18系统升级包括：
- 1) AI+Siri：正式宣布与 OpenAI 达成合作，ChatGPT整合进 Siri；2) 主屏幕个性化：可自定义排序，加入更多控制可选项，可统一桌面色调；3) 照片：引入智能功能精选集，照片分类、查找更精准；4) iMessage：字体样式、表情自定义选项增加，可以设置稍后发送；5) 邮件：自动整理分类，提炼后供快速阅读；6) 地图：新型地形图；7) 钱包：Tap to cash 功能，手机顶端接触可快速付款；8) 手记：搜索功能从过往手记找到内容；9) 隐私：推出app应用锁。

图：ChatGPT整合进 Siri



图：iOS18的更新功能



4.2 生态：Apple Intelligence重新定义端侧AI

➤ Apple Intelligence重新定义端侧AI：

苹果端侧AI 为诸多使用场景设计了更为便利的体验，相较前期的安卓AI手机，在实用性上有较大提升。同时我们认为，更值得重视的是**苹果底层能力的突破**：

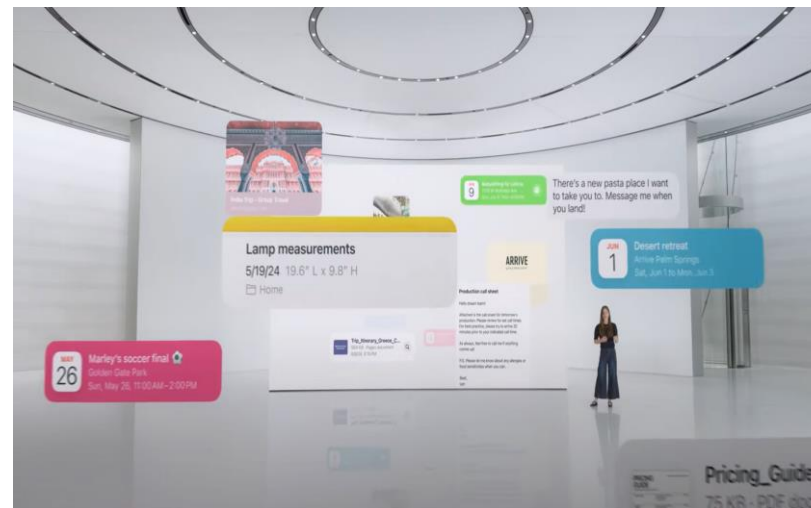
①屏幕感知：

功能层面，Apple Intelligence赋予Siri屏幕内容感知能力，同时结合用户语音指令，灵活判断给出更为精准的指令操作。

②跨APP的输入输出：

Apple Intelligence赋能Siri，可以打通多个APP的输入输出，具备AI Agent的雏形。如编辑表格使用场景中，可通过Siri控制读取相册图片中的文字，并输出至Excel。

如苹果未来将这一生态开放给第三方APP，则可实现：通过APP订机票，Siri可获取航班起落时间，并输出给打的软件，酒店软件。从而实现一键安排行程。



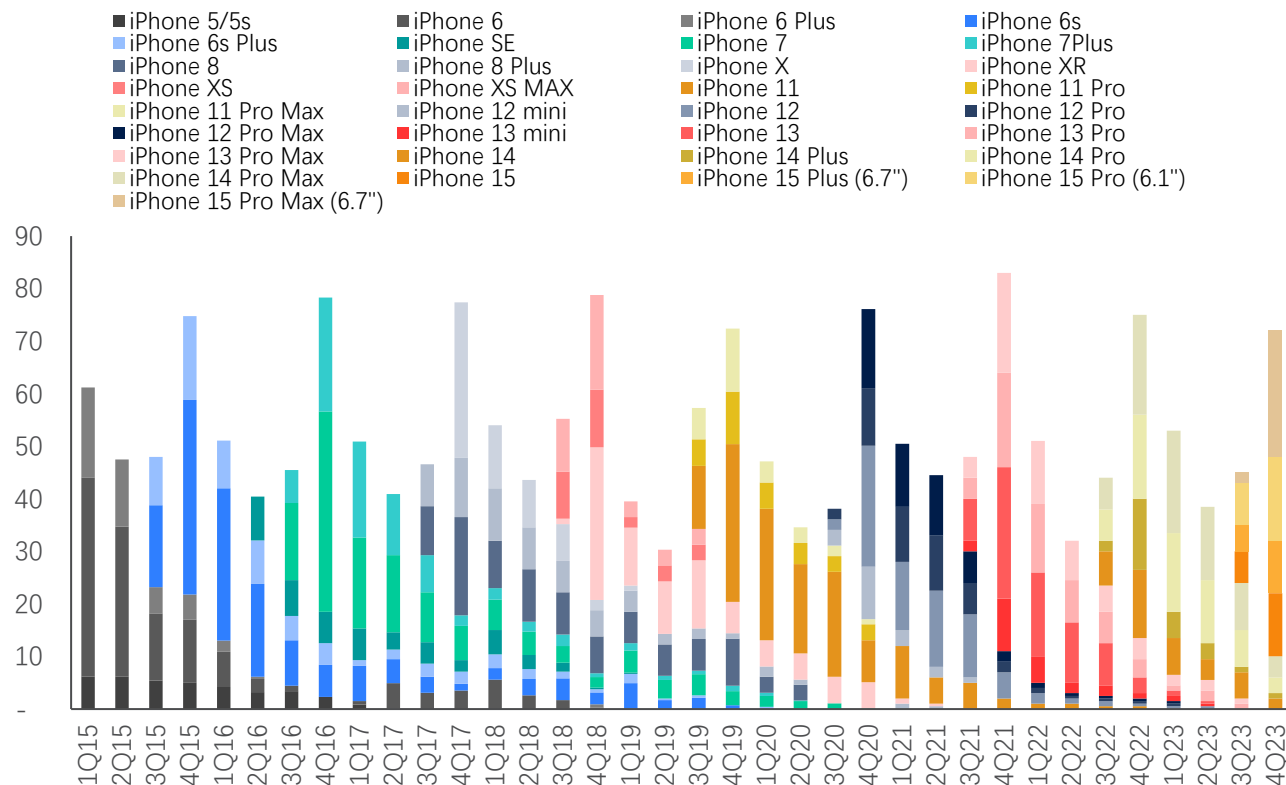
4.2 生态：Apple Intelligence重新定义端侧AI



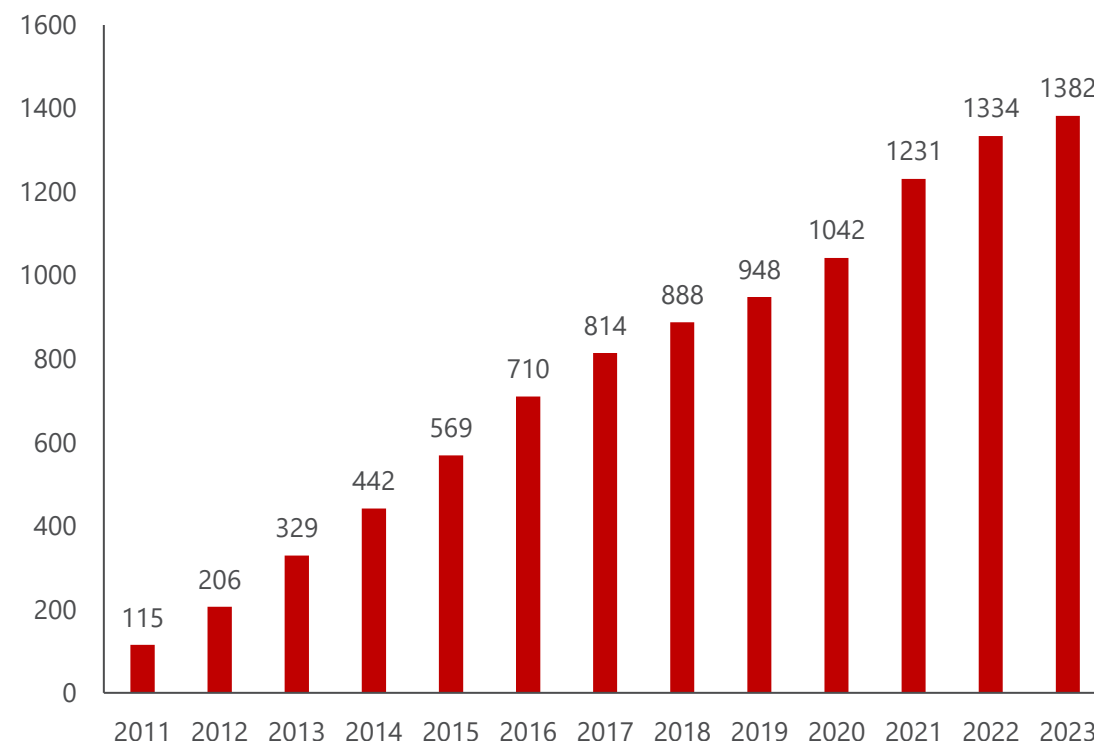
4.2 苹果开启新一轮换机周期

- 苹果于WWDC表示，仅iPhone15 pro和iPhone 15pro max以上机型才能支持AI功能，这将成为加快更换周期的关键。据IDC数据，iPhone15 pro&pro max两款机型23年销量为5000万部，占当年iPhone销量的24%，占iPhone存量活跃设备数的仅4%；作为对比，23年iPhone潜在换机基数庞大，存量活跃设备数达13.8亿部。后续在AI的带动下，iPhone换机动能强劲。

图：1Q15-4Q23 iPhone分机型销量（百万部）



图：iPhone存量活跃设备数（百万部）



4.3



01

传统终端的困境与破局

02

详解苹果公司AI终端布局

03

AI 终端三要素

04

AI 终端产业链

CONTENTS

目录



4.3

安卓算力：骁龙8系列、天玑系列持续迭代

- 前文我们讨论了苹果的AI布局，公司数年前便已前瞻布局，并于今年WWDC正式披露其AI雄心，引领了AI终端的产业发展。
- 同时，安卓及PC阵营也在积极布局。接下来我们将继续解读非苹果阵营的AI终端布局。

表：高通各代AI旗舰芯片比较

	高通骁龙8 Gen3	高通骁龙8 Gen2
工艺	台积电 4nm	台积电4nm
架构	1+2+2+3	1+2+2+3
CPU	1x3.3GHz Cortex-X4 超大核 3x3.15GHz Cortex-A720大核 2x2.96GHz Cortex-A720大核 2x2.27GHz A520 小核	1xCortex-X3 超大核3.19GHZ 2xCortex-A715 中核2.85GHz 2xCortex-A710 中核2.85GHz 3xCortex-A510小核2.00GHZ
GPU	Adreno 750 930MHz	Adreno 740
存储规格	UFS 4.0 LPDDR5X 4800MHz	UFS 4.0 LPDDR5X 4200MHz
影像	Spectra, 3x18bits, 单摄最高200MP, 双摄最高64MP+36MP, 三摄最高3x36MP, 视频拍摄最高8K@30FPS/4K@120FPS	Spectra, 3x18bits, 单摄最高200MP, 双摄最高64MP+36MP, 三摄最高3x36MP, 视频拍摄最高8K@30FPS/4K@120FPS
AI	Hexagon NPU 支持运行超100亿个参数的大模型，跑大语言模型每秒可生成超过20个Token；AI性能提升98%、能效提升40%	Hexagon DSP
连接	WiFi7、蓝牙5.4	WiFi7、蓝牙5.3
基带	骁龙X75	骁龙X70
发布时间	2023年10月25日	2022年11月16日

表：联发科各代AI旗舰芯片比较

	MTK天玑9300	MTK天玑9400
制造工艺	台积电4nm	台积电3nm
CPU配置	4+4 4 Cortex-X4 4 Cortex-A720	1+3+4 1 CortexX5 3 CortexX4 4 Cortex-A720
内存与存储性能	LPDDR5T 9600Mbps内存和UFS 4.0闪存	LPDDR5T 9600Mbps内存和UFS 4.0闪存
Meta Llama2 7B	10tokens/S	15tokens/S
Stable Diffusion V1.5	3.5秒	少于3秒
发表时间	2023年10月	2024年10月
代表机型	VIVO x100系列、OPPO Find x7、IQOO neo 9pro	VIVO 下一代旗舰机型

4.3

PC算力：NPU算力提升，低功耗高算力护航系统级AI

	Intel	AMD	高通
名称	Lunar Lake 	Ryzen AI 300 	X Elite 
制造工艺	台积电3nm	台积电4nm	台积电4nm
核心频率	8核心，基础频率2.2GHz	12 核，基础频率2.0GHz	12核，基础频率3.8GHz
NPU算力	48TOPS（前代10TOPS）	50TOPS（前代16TOPS）	45TOPS
TDP（热设计功耗）	17W	18W	23W
出货时间	24年三季度（预计）	24年7月	24年6月18日
特点	内存与芯片封装在一起的统一内存架构，更好满足AI对于大量数据即时传输的需求	业内首款支持16 FLOPS的 NPU，结合了 FP16 的精度和INT8 的性能	基于ARM架构打造，具备低功耗特点

4.3 模型：端侧AI模型加速落地

云
厂
商



品
牌
商



2023 | 2024

4.3 生态：当前AI终端功能一览

类云端AI：

- **1.文本类：**具备文字分析处理及生成功能，如撰写邀请函、文档要点提炼等；
- **2.视觉类：**利用 AIGC 可实现文生图、图片美化、图片扩展与消除、视频优化，视频智能剪辑等功能；

APP增强AI：

- **1.语音类：**主要为AI通话功能，如三星内置通话实时双向翻译和文字翻译功能，无需第三方应用并支持 13 种语言；
- **2.图片即圈即搜：**三星 Galaxy S24系列可实现即圈即搜功能，圈取屏幕上任何内容即可获得搜索结果，无需跳转APP或切换界面；

系统级AI：

- **AI Agent：**苹果Apple Intelligence具备AI Agent的雏形，通过赋能Siri，可以打通多个APP的输入输出。

图：Galaxy S24的实时翻译功能



图：Galaxy S24的“即圈即搜”功能



图：AI图片消除功能



图：AI通话摘要功能



4.3 AI 终端的三要素：算力、模型、生态

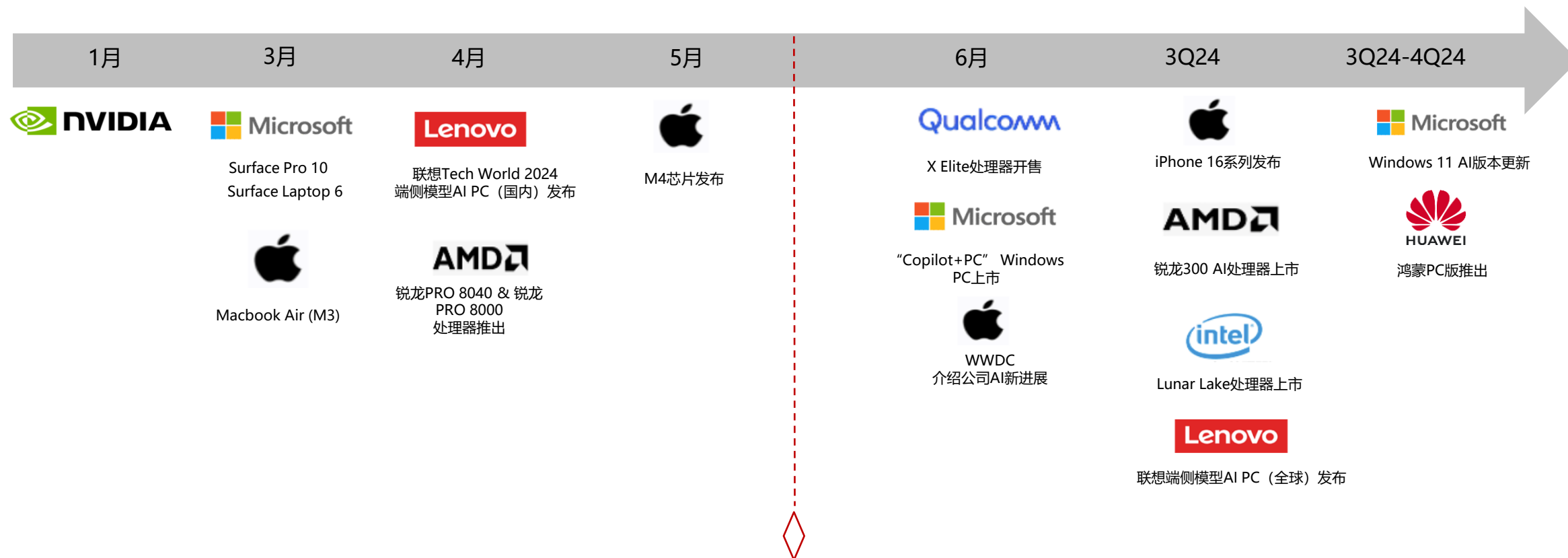
- 综上，我们将AI终端的定义，提炼为三要素：算力、模型、生态
- 其中，算力和模型是AI终端的必要条件，算力是硬件基础，通过NPU执行神经网络运算，实现终端推理；而模型的减枝蒸馏，则可有效压缩模型参数，减少推理时间，降低端侧负载。**上述两大要素决定了终端的硬实力。**
- 相较前者，生态需要长时间打磨，且需产业链厂商齐心协力配合，开发应用，挖掘垂直场景。**生态建设代表了终端的软实力。**对于用户而言，算力和模型的纸面参数固然重要，但真正决定使用体验和购买倾向的，是生态。
- 我们认为生态的建设和完善，会决定AI终端的成败。而这一过程当中位居核心的是品牌厂商。苹果则得益于其多年的积累，及消费市场的号召力，更有机会打造出完备的生态。



4.3 AI终端大事记

2024年1月-5月

NEXT



4.4



01

传统终端的困境与破局

02

详解苹果公司AI终端布局

03

AI 终端三要素

04

AI 终端产业链

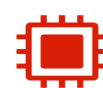
CONTENTS

目录



4.4 AI终端产业链一览

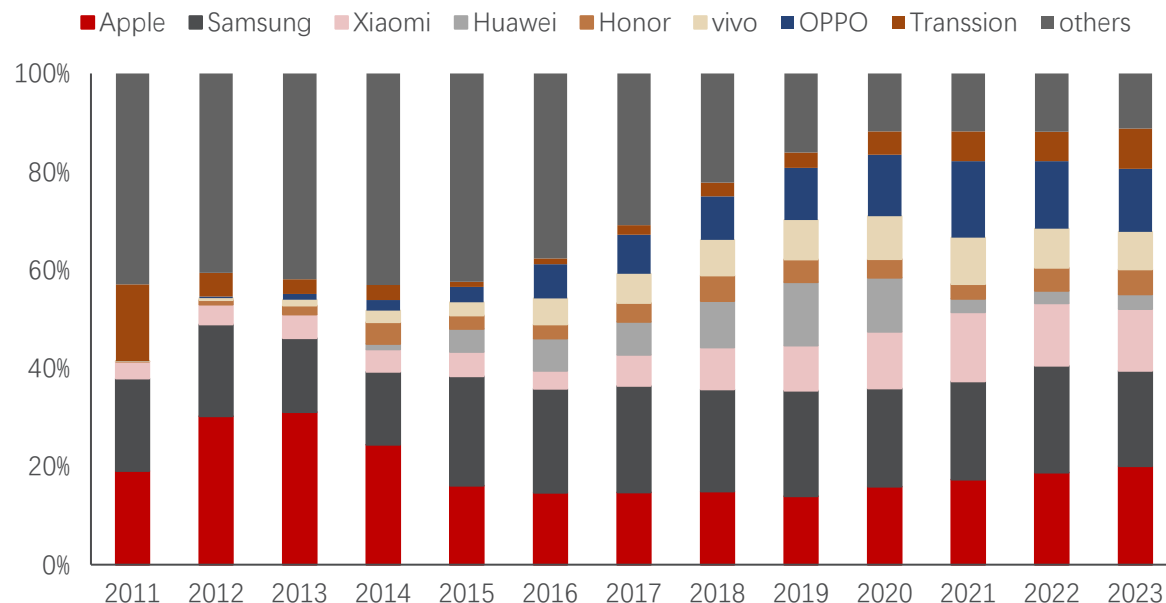
硬件价值量全面提升

	品牌	AI终端的浪潮下，首先将迎来量的变化，换机周期再度加速；随后则是质的变化，品牌商的话语权会得到提升	● PC: 联想、DELL ● 手机: 苹果、小米、传音
	算力	CPU: 加入NPU单元，满足一般AI场景 GPU: 需要高算力的AI场景，需要独立GPU配合	● PC: 高通、intel、AMD ● 终端: 高通、MTK、intel、苹果
	存储	内存: 7B端侧模型运行需8G内存，整机内存向16/32G升级，AI手机的内存和存储容量分别是普通手机的1.7倍和2倍	● 存储厂: 三星、海力士、美光 ● 模组: 德明利、江波龙、佰维、朗科 ● DDR5: 澜起科技、聚辰股份
	PCB	AI赋能手机，SLP硬板加速升级	● PCB: 鹏鼎控股、东山精密、弘信电子、燕麦科技
	散热	高算力产生高功耗，热管、石墨、硅脂、风扇等散热方案升级，价值量提升	● 散热: 瑞声科技、领益智造、思泉新材、中石科技
	结构件	结构件或将迎升级，以满足轻量化要求	● 结构件: 领益智造、长盈精密
	电池	AI终端高算力下功耗提高，电池芯能量密度升级，且或采用不锈钢电池壳	● 电池: 珠海冠宇、信维通信、豪鹏科技
	光学	光学传感器持续升级，满足AI终端的感知需求	● CIS: 韦尔、思特威、格科微 ● 光学零部件: 水晶光电、蓝特光学
	代工	AI 终端对整机设计提出更高要求，ODM厂商代工价值量提高	● 代工: 立讯精密、比亚迪电子、蓝思科技、华勤技术

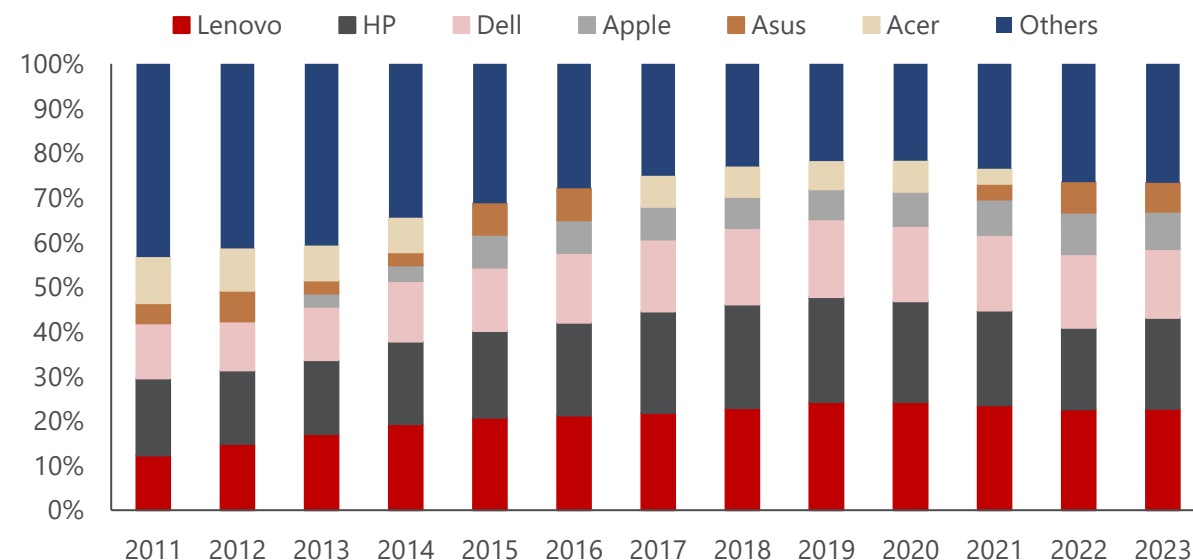
4.4 品牌商：话语权提升，价值增量显著

- 我们认为，AI终端的浪潮下，首先将迎来量的变化，换机周期再度加速；随后则是质的变化，品牌商的话语权会得到提升：
- 苹果AI生态普及后，Siri有望成为我们工作及生活的助理，其触角将延伸至方方面面，届时出行旅游消费等都会变得更为便捷，AI拥有更强的建议及选择权，品牌商的话语权也随之提升。
- 联想AI PC搭载端侧AI Agent“小乐同学”，当前小乐同学已开放端侧模型API，开发者可将AI小程序接入天禧AI生态，类似APP Store，接入端侧模型的程序将陆续上线，为联想AI PC提供更丰富的AI功能和收入来源。在传统PC硬件同质化较为严重的背景下，其自身AI大模型的嵌入AI PC的能力或将重构PC品牌商的竞争力，助力联想主导AI PC时代。

图：2011-2023年手机品牌厂商市占率



图：2011-2023年PC品牌厂商市占率



4.4 散热：AI终端能耗提升，散热成刚需

- 由于大模型的本地化运行对计算资源和存储空间的需求，处理器和存储芯片的功耗会显著增加，对AI终端的功耗管理、散热效率提出了新的挑战，推动散热材料升级。
- 手机中最常用的散热材料为**石墨散热膜**和**VC均热板**，**AI手机石墨层或需要增厚，价值量提升，亦有可能加速VC均热板方案渗透。**
- **石墨散热膜**：石墨膜具有良好的柔韧性、耐高低温性、化学稳定性及优秀的水平传热能力，消除小型化智能电子设备局部热点效果好。石墨为当前最主流的散热方式，**苹果全系列产品采用纯石墨散热方案**，其余手机品牌兼具纯石墨/石墨+VC均热板混合方案；石墨烯散热效果相比石墨有一定提升，但综合性价比来看尚未得到广泛推广。
- **VC均热板**：均热板具有更高的导热散热效率、可快速将热量均匀扩散到一个较大的面积上、能够集成解决多个高功耗器件的散热需求等。当前VC向超薄方向发展，以节省手机内部空间，**在三星、华为Mate60、OPPO等旗舰机型上大量使用，成本较石墨高，但散热效果更好。**
- **PC散热系统较为复杂**，芯片热量主要经导热硅脂—热管—散热鳍片—风扇散出，高性能PC需要用多热管增强散热，当前轻薄本亦有VC均热板升级的趋势。

图：iPhone石墨散热



图：vivo、华为VC均热板散热



图：PC散热系统

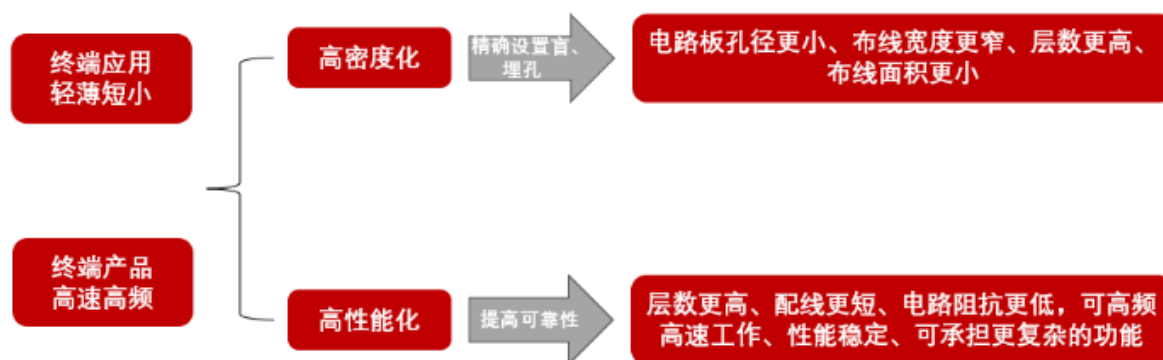


4.4 PCB：AI赋能手机，SLP硬板加速升级

- **智能机进入AI时代，主板方案有望迎来升级。**当前时点手机AI创新升级对硬件要求升级，将对主板技术路线产生影响，芯片I/O数增加导致PCB直径缩小、走线数量增加，压缩PCB的线宽线距；功能升级需要占用空间等，均需要更加高阶的主板去完成。但目前对终端要求的尺寸却不断缩小，所以对PCB导线的宽度、间距，微孔盘的直径和孔中心距离要求进一步提高。
- 从HDI工艺角度来看，SLP是采用了mSAP（改进型半加成工艺）的Any Layer技术，在一定资金的设备投入下，还利用了HDI的现有设备、技术，相对载板的制造，生产成本低，效率高。伴随AI手机升级，对手机内部空间及功能提出更高要求，SLP的线宽线距升级有望加速。

表：HDI和SLP参数对比

板材	工序	板厚	辐射孔径	孔数/每部手机主板	线宽/线距
HDI 板	120-144	0.7mm	100/200 μ m	超过 1w	40/50 μ m
SLP 板	177	0.5mm	70/140 μ m	最高超 10w	20/35 μ m



表：AI有望加速SLP的线宽线距升级

	2018年	2019~2020年	2021~2022年
层数/板厚	12层/0.65 mm	12层/0.60 mm	12层/0.55 mm
芯板厚/介电层厚度	0.065 mm/35 μ m	0.06 mm/35 μ m	0.05 mm/30 μ m
最小线宽/间距	30 μ m/30 μ m	25 μ m/25 μ m	20 μ m/20 μ m
最小激光孔径/焊盘	90 μ m/170 μ m	80 μ m/160 μ m	80 μ m/150 μ m
BGA Pitch (球栅阵列间距)	250 μ m/400 μ m	230 μ m/350 μ m	200 μ m/300 μ m
板材	H/F (高频)	H/F, LowCTE (低热膨胀系数)	H/F, LowCTE, D_k , D_f

4.4 联想集团：PC及服务器王者，AI时代继往开来

公司为全球最大的PC品牌商、全球第四大服务器制造商，以及全球第三大AI基础设施提供商，在PC及服务器领域深耕多年，技术及渠道积累丰厚。

业务结构： PC占比55%，智能手机占比18%，服务器相关占比15%，解决方案占比12%。公司分为IDG、ISG、SSG三大业务集团：

IDG： 主要包含PC及手机业务

- PC：联想全球pc市占率达22.9%，连续6年蝉联全球PC销冠，4Q23去库结束开启复苏，同比增长3.2%；市场份额达24%，同比提升1.5pct。联想商用占比超65%，且商用PC市场毛利率最高可达30%+，率先受益AI PC的落地
- 手机：23年位列全球手机销量前十，出货量位居拉美第二位，北美第三位，欧洲第四位。

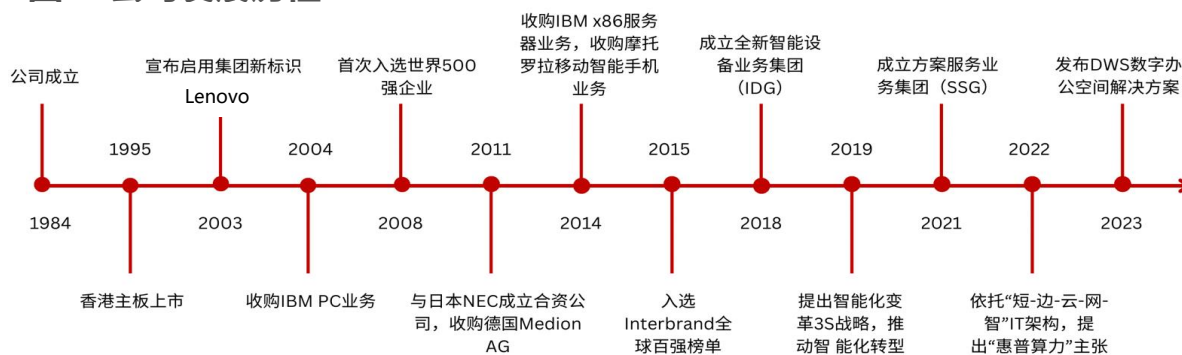
ISG： 主要包括服务器、存储和软件业务，同时为微软核心供应商，AI服务器持续爆单，占ISG营收比重超20%； ▲ 随着AI逐渐步入推理侧及中小客户，自有品牌优势体现，盈利能力显著提高。

SSG： 主要包括运维服务、项目解决方案服务、附加和支持服务，该业务非硬件类服务收入占比超50%，利润率较高且增速较快。

FY4Q23（自然年1Q24）公司实现营收138.64亿美元，同比增长9.86%，实现净利润2.48亿美元，同比增长118.05%。我们预计FY2025-2027联想集团净利润分别为12.15/15.47/19.03亿美元，对应PE分别为14/11/9倍（股价为6月14日收盘价）。

风险提示：AI PC落地不及预期，AI服务器需求不及预期，服务器行业竞争加剧

图：公司发展历程



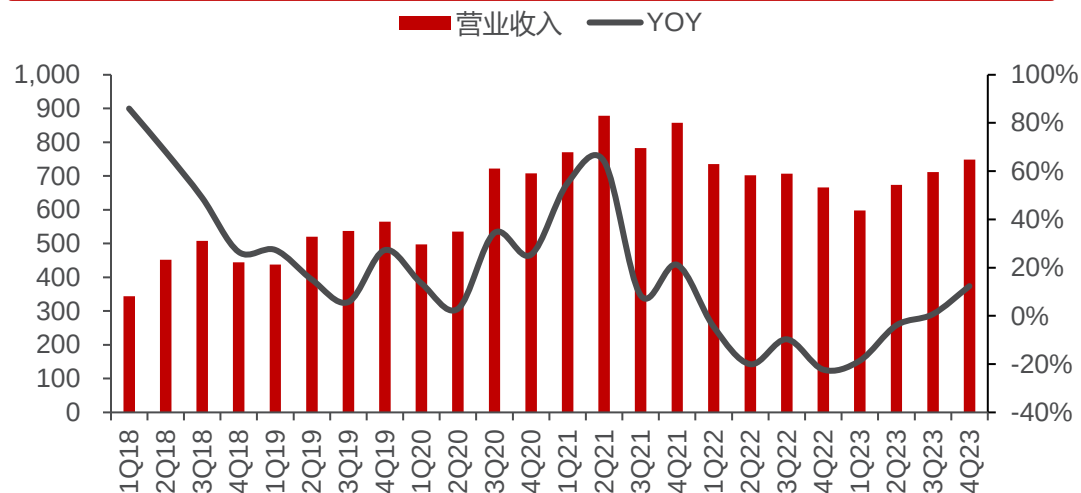
图：公司产品及客户情况



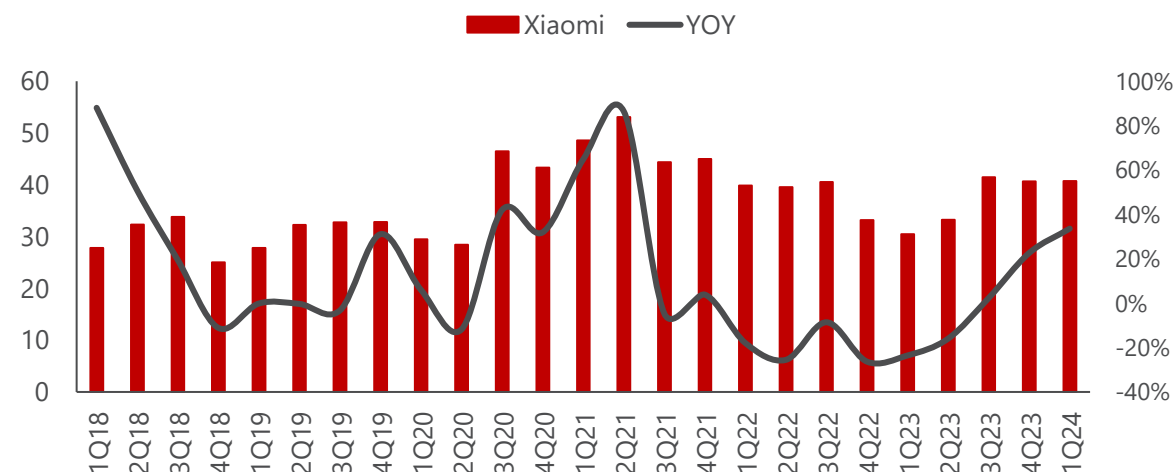
4.4 小米集团：“人车家全生态”扬帆起航

- **总体**：23年，实现营收 2,710 亿元，调整后的净利润193 亿元，YOY+126%，毛利率创历史新高为21%。在汽车等创新业务上费用约67 亿元。还原后的核心业务的利润达到 260 亿，同比增长124%。24Q1实现营业收入755亿，同比增长27%，实现经调整净利润64.9亿，同比增长100.75%。
- **手机**：24Q1全球出货量达4060万部，同比增长33.7%；智能手机ASP于中国大陆及境外市场同比分别上升；
- **IoT**：24Q1实现收入203.74亿元，同比增长21%，连续四个季度保持在200亿以上水平，毛利率19%，历史新高。AIoT平台连接的设备总数已达到 7.86亿，同比增长27%；
- **互联网**：24Q1实现收入80.48亿元，同比增长14.5%，创历史新高，同比增长14.5%，毛利率达74.2%，同比提升1.9pct；
- **汽车**：24Q1研发费用52亿元，截至2024年5月15日，小米汽车累计交付新车达1万辆。
- **估值**：我们预计公司24/25/26年归母净利润为158.8/172.8/220.5亿元对应PE为25/23/18倍，维持“推荐”评级。
- **风险提示**：市场竞争加剧；海内外政策波动；造车等新业务进展不及预期。

图：小米营业收入（亿元）



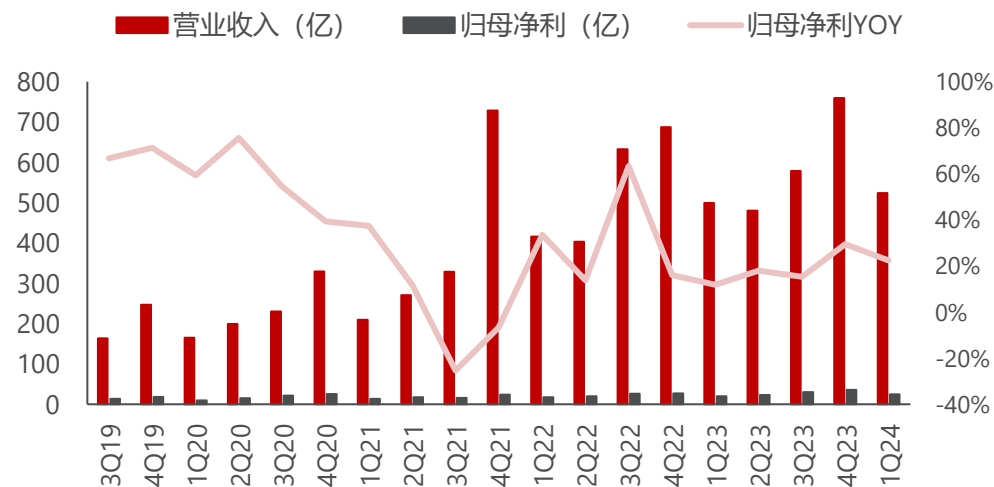
图：小米智能手机季度出货量（百万部）



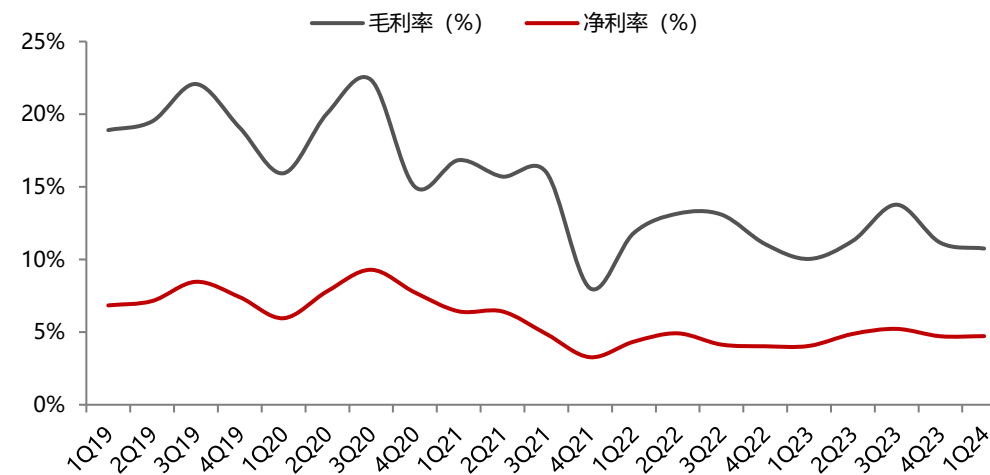
4.4 立讯精密：纵向+横向，延伸拓宽产品布局

- **公司简介：**立讯是国内精密制造龙头企业，通过近二十年发展，目前已形成消费电子、汽车、通信三大业务平台。公司积极推进产业链垂直整合，形成“零件-模组-系统”的全方位布局；同时，将精密制造能力横向复用至汽车和通信业务，实现平台扩张；外延+内生：公司通过外延并购+内生研发，不断进行产品研发与品类扩张，充分发挥垂直一体化优势，打造智能制造基地。
- **消费电子业务：**2023年消费电子业务实现收入1972亿元，同比+9.75%，公司第一大客户收入占比持续提升，后续伴随AI手机的“iphone”时刻到来，智能机有望重回成长；
- **通讯&汽车业务：**通讯业务方面，公司持续深耕电连接、光连接、风冷/液冷散热、电源管理、射频等产品。在数据中心高速互联领域，公司成功开发112G PAM4无源铜缆DAC、112GPAM4有源铜缆ACC，此外其224G新产品技术及布局处于全球领先地位。汽车业务方面，公司深耕汽车零部件领域十余年，高/低压线束、高速线束、充电枪、汽车连接器、智能座舱域控制器、液晶仪表、三合一电机系统、AR HUD、DMS等全面布局。23年公司汽车业务收入实现50%+增长。
- **估值：**我们预计公司24/25/26年归母净利润为142.2/175.8/213.8亿元对应PE为14/12/10倍，维持“推荐”评级（截至2024年6月17日收盘价）。
- **风险提示：**消费电子需求疲软；行业竞争加剧；新业务拓展不及预期。

图：立讯营收归母净利及同比增速（亿元）



图：立讯毛利率及净利率

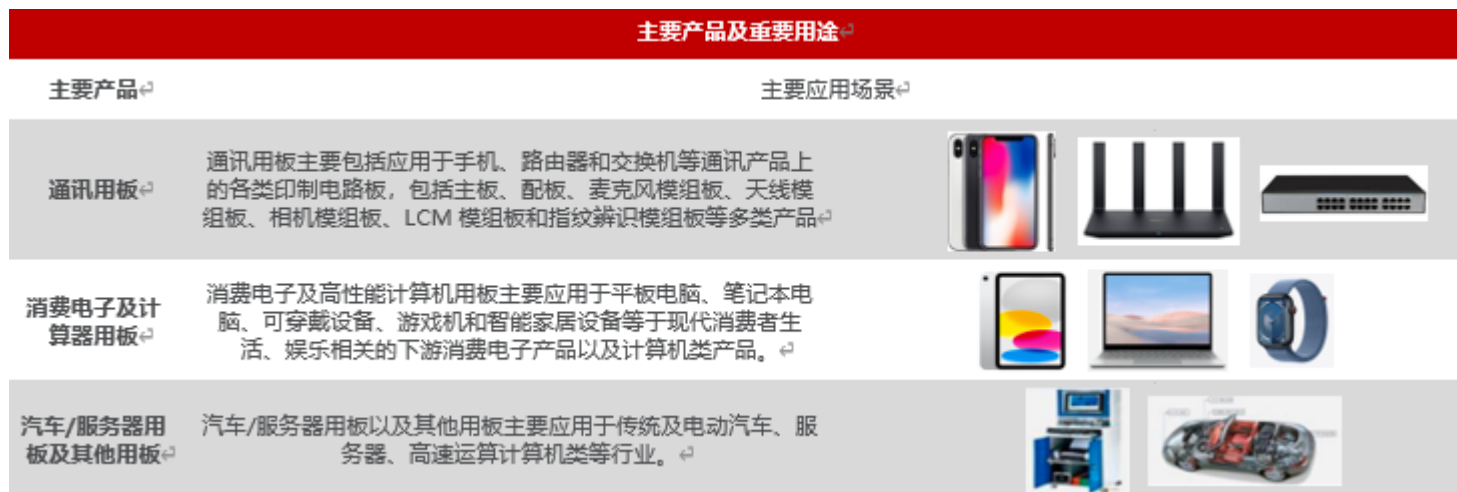


4.4

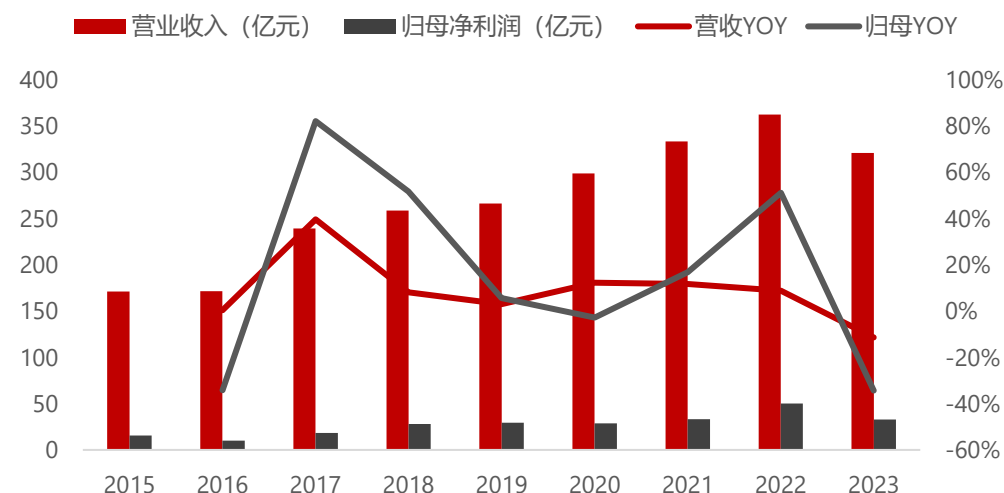
鹏鼎控股：AI赋能，FPC+SLP加速创新

- **公司简介&客户：**鹏鼎控股是一家大型PCB产品制造商，提供各类PCB产品的研发、设计、制造和销售服务。PCB产品线涵盖FPC、SMA、SLP、HDI、Mini LED、RPCB、Rigid Flex等多种产品，应用于通讯电子产品、消费电子、高性能计算机、电动汽车和人工智能服务器等领域。深耕行业数年，行业地位稳固。自公司成立以来，公司服务过的客户包括诺基亚、摩托罗拉、索尼爱立信等国际领先品牌客户，目前为苹果主要供应商。
- **营收情况：**2022年，公司实现营收总额362.1亿元，同比+8.71%；归母净利润50.12亿元，同比+51.10%。2023年受宏观环境、外汇变动以及国际贸易摩擦等因素影响，公司实现营业收入320.7亿元，同比-11.43%；归母净利润32.87亿元，同比-34.42%。1Q24公司实现营收66.86亿元，同比上升0.29%，实现归母净利润4.97亿元，同比上升18.81%。
- **估值：**据iFinD一致预期，公司2024-2026年归母净利润38.65/44.02/48.60亿元，对应现价PE 21/18/16倍（截至2024年6月14日收盘价）。

图：鹏鼎控股产品情况



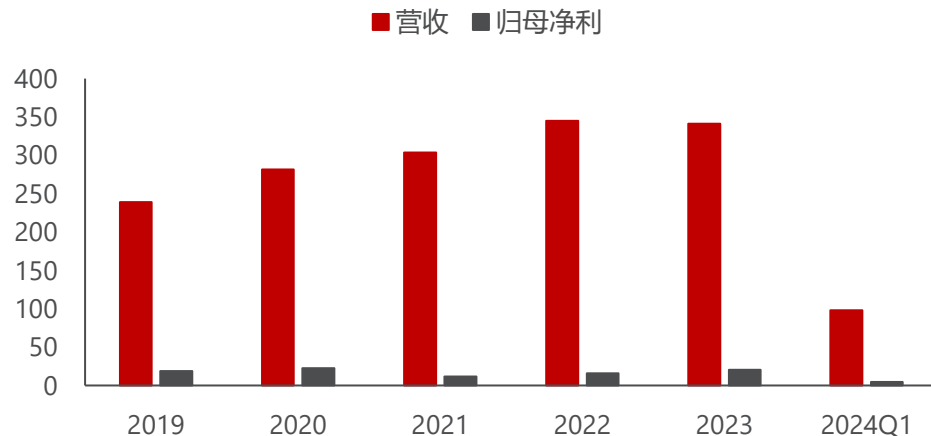
图：鹏鼎控股2015-2023营收（亿元）及归母净利润（亿元）



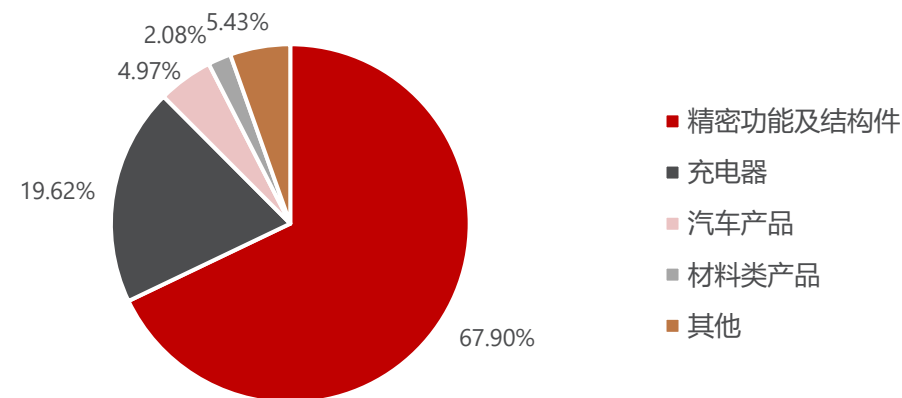
4.4 领益智造：业绩逐季稳步增长，深耕AI终端智能智造平台

- **公司简介：**作为国内精密智造领军者，公司专注精密器件市场，“材料-零部件-模组-组装”上下游垂直整合，打造一站式智能智造平台。公司拥有全球化生产基地，海外业务广泛分布于美国、法国、巴西、印度等地，为公司高速成长奠定根基。
- **营收情况：**24Q1公司实现营收97.99亿元，同比持续提升35.83%，环比增加3.4%，主要系新业务增长所致。分业务来看，2023年，公司AI终端及通讯类业务实现收入306.87亿元，毛利率21.62%。其中电源、电池、散热、显示、摄像头、中框等零组件收入占比较高，XR、折叠屏手机、服务器、机器人等终端相关零组件收入增长较快。汽车业务实现收入16.97亿，同比增长43.61%。光伏储能业务收入达15.55亿元，同比增长29.89%。
- **估值：**我们预计公司24-26年归母净利23.84/33.79/42.19亿元，对应现价PE为17/12/9倍，维持“推荐”评级（截至2024年6月17日收盘价）。
- **风险提示：**下游需求不及预期；行业竞争加剧；新业务拓展不及预期。

图：领益智造2019-24Q1营收（亿元）及归母净利润（亿元）



图：领益智造2023年产品结构



4.4 蓝思科技：玻璃盖板龙头的新成长周期

- **公司简介：**公司为消费电子玻璃盖板龙头，公司消费电子业务垂直整合，纵向以组装业务为抓手，拓展上游玻璃、金属结构件及模组产品，横向拓展新能源等领域，公司铲平渗透率和市场份额持续提升。
- **收入构成：**2023年营收按产品种类，智能手机与电脑类业务占比82%，新能源汽车类业务占比9%，智能头显与穿戴类占比6%。
- **消费电子业务：**横向拓展陶瓷、蓝宝石等新材料以及触控传感器项目，纵向拓展下游组装业务，良率已超过行业平均水平，得到客户高度认可，取得良好盈利表现，同时拓展上游材料，实现主要辅料辅具内部高比例自供与配套。2023年该业务实现营收449.01亿元，YOY+17.50%。
- **汽车业务：**公司深入拓展汽车玻璃产品及其他车载产品，并拓展充电桩、动力电池精密金属结构件等，汽车业务增长迅速。2023年，公司新能源汽车与智能座舱类业务实现营业收入49.98亿元，同比增长39.47%，毛利率15.47%；公司已经与超过30家国内外新能源及传统豪华汽车品牌合作。
- **估值：**我们预计公司24-26年实现归母净利润分别为42.27/53.80/67.12亿元，对应PE分别为18/14/11倍，维持“推荐”评级。
- **风险提示：**下游需求不及预期的风险；市场竞争加剧的风险；新产品和新市场开拓不及预期的风险。

图：蓝思科技业务布局

现有产品布局



图：蓝思科技客户



4.4 华勤技术：全产品线受益AI的ODM龙头

华勤为ODM平台型公司，产品覆盖手机、笔电、平板、智能穿戴、AIoT、服务器等多品类，全产品线受益AI赋能。

智能终端业务：包括智能手机及智能穿戴产品

根据counterpoint数据，华勤技术2023年全球手机ODM市占率达28%，位列行业第一；主要客户包括三星、小米、OPPO、VIVO、Moto等；相比笔电行业ODM的高渗透率，手机行业ODM渗透率较低，未来更多自研机型或将释放给ODM厂商，提供增长动能。

高性能计算业务：包括笔电、服务器、交换机、平板产品

➢ **笔电：**公司为内资唯一形成千万台以上PC代工规模的ODM厂商，出货量1150万台，同比增长15%，主要客户包括联想、华硕、宏碁、华为、小米、荣耀、三星等；份额提升空间大，越南、印度、墨西哥建厂有望引入北美品牌商。

➢ **服务器：**AI服务器拉动下数据中心业务高增，主要客户包括腾讯阿里字节等；400g交换机已开始出货，板卡-服务器-交换机全产业链设计能力更好配合CSP。

AIOT业务：受益于耳机等AIOT市场向手机品牌商集中的趋势，公司该业务高速增长

汽车业务：在智能座舱、智能车控、智能网联、智能驾驶四大业务模块均已实现突破，与国内外汽车主机厂达成多项合作，23年实现营收110%增长

业绩方面，1Q24公司实现营收162.29亿元，同比下滑-3.52%，实现归母净利润5.96亿元，同比增长4.28%。我们预计2024-2026华勤技术净利润分别为29.88/34.55/39.76亿元，对应PE分别为20/17/15倍（股价为6月14日收盘价），维持“推荐”评级。

风险提示：消费电子需求波动，数据中心建设进度不及预期，客户拓展不及预期

图：公司发展历程

2017-2018

- 无锡研发中心成立
- 南昌制造中心开工
- A轮融资8.7亿元（英特尔领投）
- 服务器量产发货
- 东莞研发中心开园

2019

- 两大海外制造基地完成布局
- 南昌制造中心投产开园
- B轮融资11.5亿元（高通领投）

2020-2021

- 软件与评测中心成立
- 无锡研发中心开园
- C轮融资10亿元（中国移动领投）
- 汽车电子事业部，创新业务事业群和X-lab成立

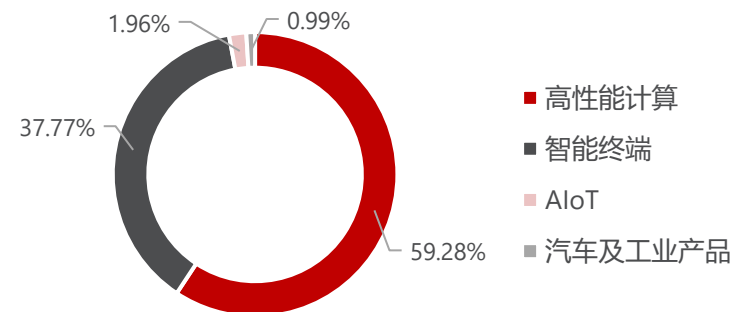
2022

- 全球研发总部动工
- 汽车电子临港研发制造基地签约
- 汽车电子产品量产出货

2023

- 上海证券交易所主板成功上市
- 西安研发中心（丝路总部）开园

图：2023年公司业务构成

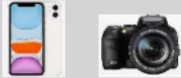
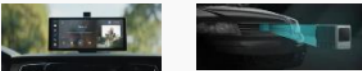


4.6

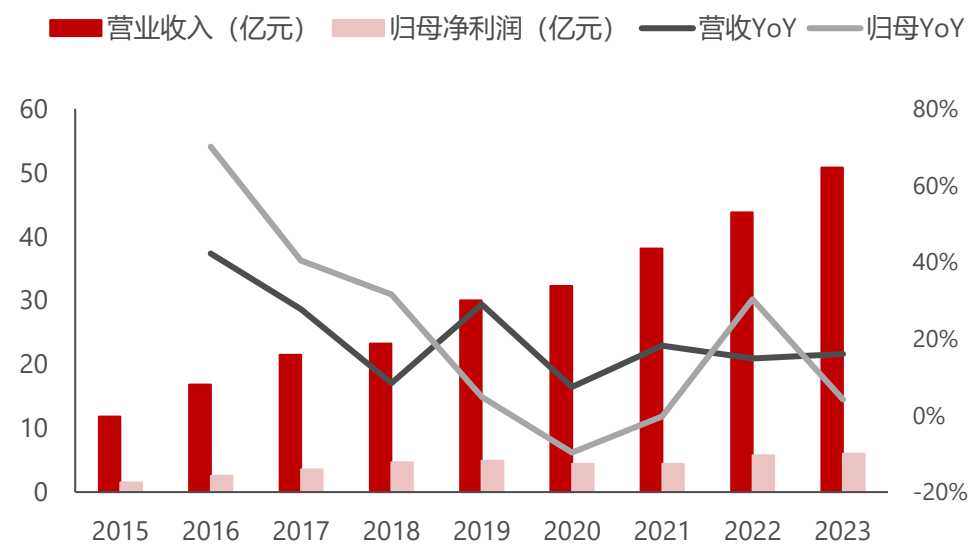
水晶光电：光学平台公司，iPhone光学创新核心受益

- **公司简介：**公司已构建光学元器件、薄膜光学面板、半导体光学、汽车电子（AR+）、反光材料五大业务板块，产品形态已由单一的光学元器件向元器件、模组及解决方案并存转型，产品已广泛应用于智能手机、相机、智能可穿戴设备、智能家居、安防监控、车载光电、元宇宙 AR/VR 等领域。深耕行业数年，行业地位稳固。公司的客户包括华米OV、吉利、东风岚图、长安马自达等品牌客户，目前为苹果的光学核心供应商。
- **营收情况：**2023年公司实现营业收入50.76亿元，同比+16.01%；归母净利润6.00亿元，同比+4.15%。1Q24公司实现营收13.45亿元，同比上升53.41%，实现归母净利润1.79亿元，同比上升92.49%。
- **估值：**据iFinD一致预期，公司2024-2026年归母净利润8.05/9.92/11.79亿元，对应现价PE 28/23/19倍（截至2024年6月14日收盘价）。

表：水晶光电产品情况

主要产品及重要用途		
主要产品	主要应用场景	
光学元器件	产品应用于手机、相机、车载等成像摄像头	
薄膜光学面板	应用于手机/pad/无人机/相机等镜头保护盖板；传统手表、智能手表	
半导体光学	接近光传感器滤光片/3D 窄带滤光片产品应用于手机接近光传感器；投影仪自动对焦；多种方案的 3D 摄像头发射及接收端模组；TWS 耳机	
汽车电子（AR+）	AR-HUDW-HUD 应用于车载电子领域；车载激光雷达视窗片应用在车规级激光雷达产品上	

图：水晶光电2015-2023营收（亿元）及归母净利润（亿元）

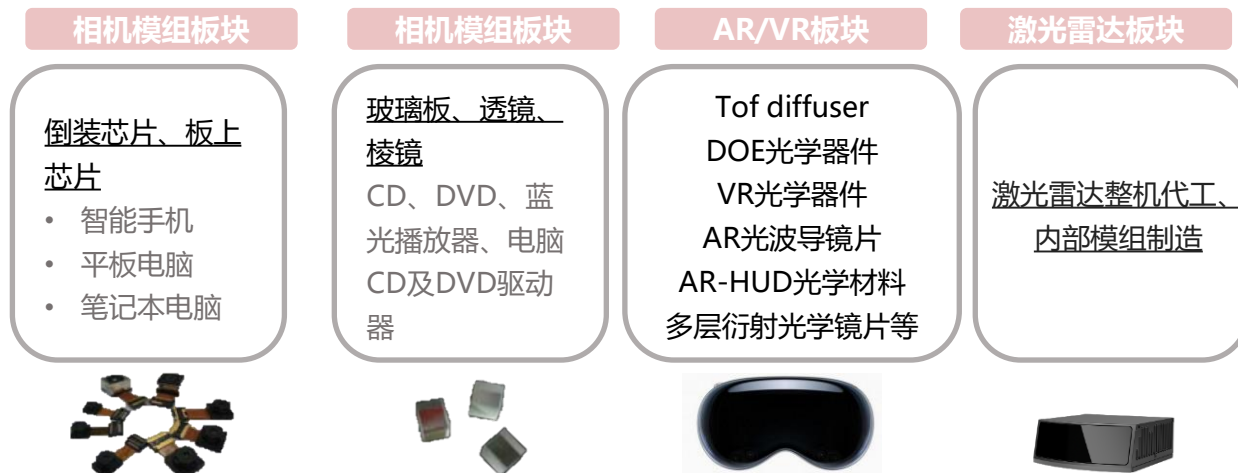


4.7 高伟电子：苹果摄像头模组核心厂商，汽车贡献第二成长极

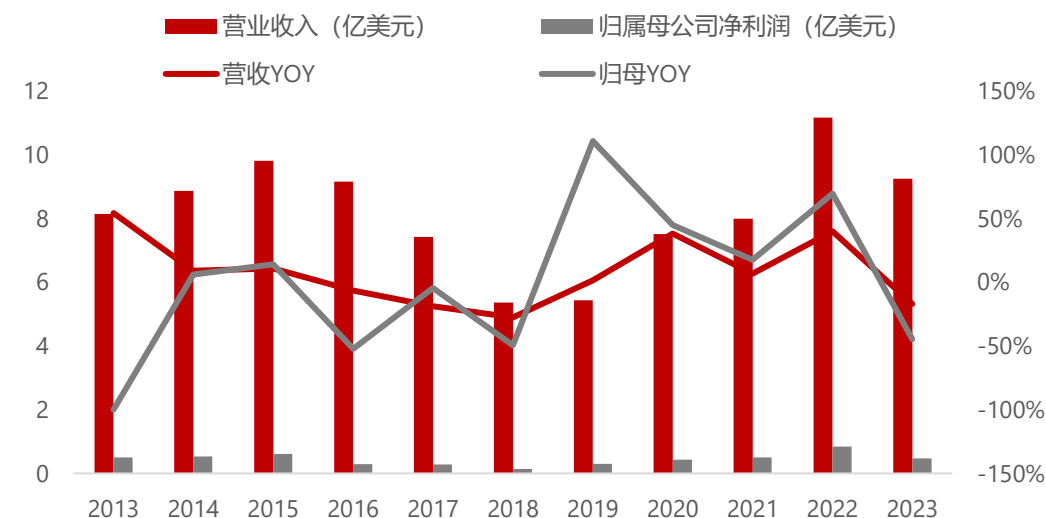
- **公司简介：**高伟电子是电子设备的精密光学模组供应商，从事设计开发、生产制造及销售各类光学部件、相机模组及系统集成产品，业务聚焦于智能手机、笔记本电脑、平板电脑等3C电子产品的摄像模组领域。公司积极横向拓展业务，开拓AR/VR、激光雷达赛道，不断扩张自身能力边界，创造中长期成长空间。
- **营收情况：**2023年实现营收9.24亿美金，同比-17.2%，归母净利润0.47亿美金，同比-44.74%。1Q24公司实现营收66.86亿元，同比上升0.29%，实现归母净利润4.97亿元，同比上升18.81%。
- **估值：**我们预计24-25年归母净利润为1.10/1.87亿美元，对应现价PE为26/15倍，公司在苹果终端的镜头模组布局多年、AR技术积淀较久，维持“推荐”评级。（截至2024年6月14日收盘价，采用汇率：1 HKD = 0.1393USD）。
- **风险提示：**下游需求不及预期，项目验证不及预期，产线爬坡不及预期。

图：高伟电子业务模块

高伟电子业务模块



图：高伟电子2013-2023营收（亿美元）及归母净利润（亿美元）



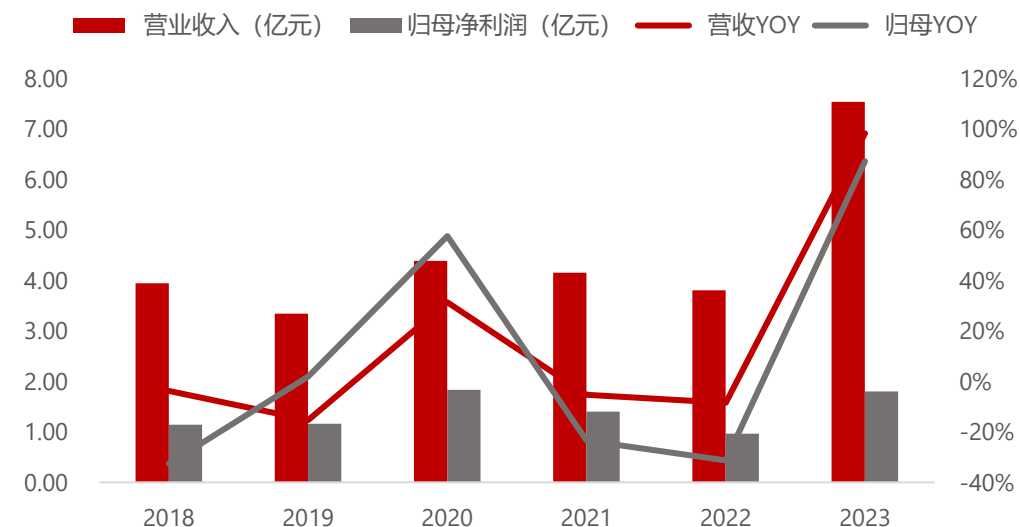
4.8 蓝特光学：iPhone微棱镜核心供应商

- **公司简介：**公司涉足消费类电子、光学仪器、车载领域及AR等在内的众多下游细分领域，已成为AMS集团、康宁集团、索尼、麦克赛尔、舜宇集团等国内外知名企业优质的合作伙伴。深耕精密光学元件领域，已形成光学棱镜、玻璃非球面透镜、玻璃晶圆三大产品系列。
- **营收情况：**2023年营收实现增长，为7.54亿元，同比上升98.35%，归母净利润为1.80亿元，同比上升87.27%。1Q24公司实现营收1.67亿元，同比上升95.47%，实现归母净利润0.29亿元，同比上升510.63%。
- **估值：**我们预计24-25年归母净利润为2.89/4.03亿元，对应现价PE为26/18倍，公司微棱镜产品的竞争力较强，玻非和晶圆产品在车载和AR技术积淀较久，维持“推荐”评级。（截至2024年6月14日收盘价）。
- **风险提示：**下游需求不及预期，客户开拓不及预期，项目验证不及预期。

图：蓝特光学主要产品



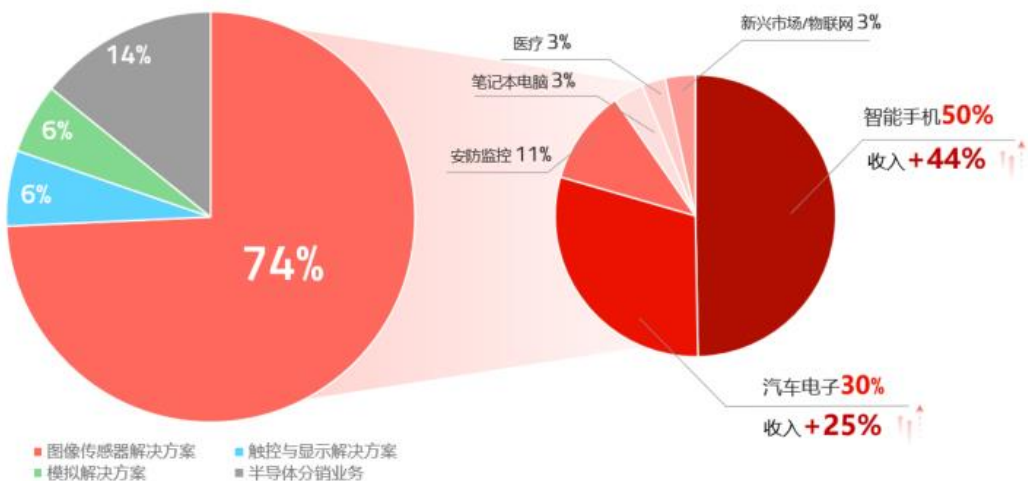
图：蓝特光学2018-2023营收（亿元）及归母净利润（亿元）



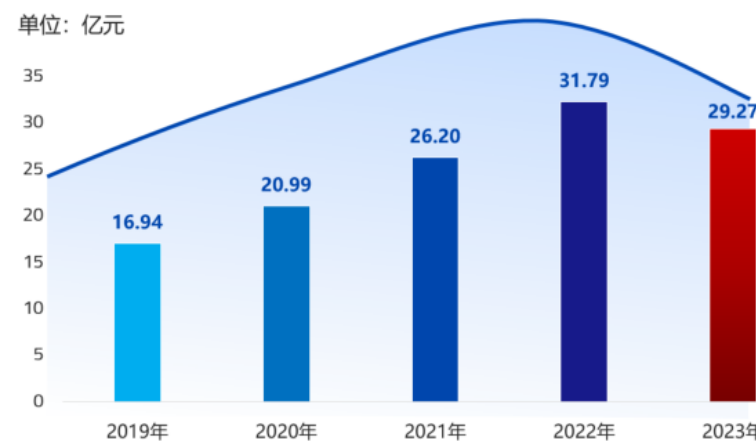
4.4 韦尔股份：国产CIS龙头厂商

- **库存去化完成，产品结构持续优化，盈利能力稳步改善。**2023年公司主营业务中图像传感器解决方案/触控与显示解决方案/模拟解决方案/半导体分销业务的收入占比分别为74%/6%/6%/14%。其中图像传感器收入中来自智能手机/汽车电子/安防监控/笔记本电脑/医疗/新兴市场&物联网的占比分别为50%/30%/11%/3%/3%/3%，来自智能手机和汽车电子的收入分别实现了同比44%和25%的较大幅度增长，智能手机CIS收入占比从22年的39%提升至50%。
- **智能手机新品销售强劲，汽车电子星辰大海。**受益于公司5000万像素及以上高端产品放量，23年公司智能手机CIS收入中5000万像素及以上产品占比突破60%。在5000万像素手机CIS产品方面，公司2023年先后推出50H、50E、50K40等产品，结合公司PureCel® Plus-S 晶片堆叠技术、DCG技术、LOFIC等先进技术，不断满足客户旗舰级后置主摄需求。汽车电子方面，公司CIS产品性能优异，市场份额快速提升；此外公司持续推进车用模拟芯片的布局，推进 CAN/LIN、SerDes、PMIC、SBC等多产品的验证导入，23年车用模拟IC在模拟解决方案收入中占比达到13%。
- **投资建议：**预计公司24-26年归母净利润分别为27.12/41.71/63.59亿元，对应当前价格的PE分别为47/30/19倍，维持“推荐”评级。（截至2024年6月17日收盘价）
- **风险提示：**下游需求不及预期；新产品研发进展不及预期；市场竞争加剧的风险。

图：韦尔股份2023年收入结构



图：韦尔2019-2023年研发投入情况

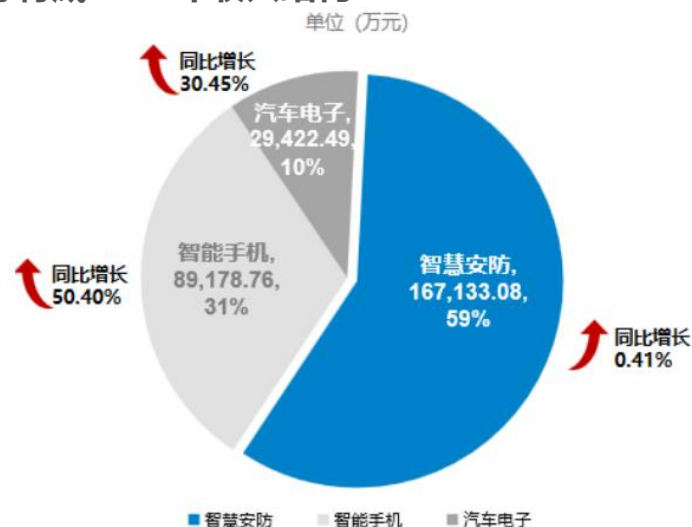


注：2023年公司研发投入较2022年略有下降，主要是受股份支付费用确认和冲回的影响。

4.4 思特威：高端手机CIS打造第二增长曲线

- **“安防+手机+汽车”三箭齐发，产品矩阵持续完善。**思特威是国内领先的CIS厂商，公司坚持“智慧安防+智能手机+汽车电子”三足鼎立的发展方向，不断优化完善产品矩阵，成功构建高阶智能手机产品第二增长曲线。2023年公司智慧安防、智能手机、汽车电子的收入分别为16.71亿元、8.92亿元、2.94亿元，占主营收入的比例分别为58.49%、31.21%、10.3%。
- **智能手机高端CIS产品顺利出货，开辟第二增长曲线。**2023年公司智能手机CIS收入实现高速增长，营收占比持续提升。目前，公司的手机CIS产品分辨率已覆盖80万到5000万像素需求，且公司推出的数颗用于旗舰机主摄、辅摄（包括长焦和广角）的XS系列高端产品已经大规模出货，成功为公司开辟第二增长曲线。除2023年推出的550XS和520XS外，2024年1月，公司推出首颗5000万像素1/1.28英寸CIS新品580XS，此款产品沿用550XS的22nm HKMG Stack工艺，单像素pixel尺寸1.22μm，具备高动态范围、低噪声、100%全像素对焦、超低功耗等性能优势，主要面向旗舰级智能手机主摄。
- **投资建议：**预计公司24-26年归母净利润分别为2.75/5.49/8.60亿元，对应当前价格的PE分别为73/37/23倍。（截至2024年6月17日收盘价）考虑到公司在安防监控领域的产品竞争力，且不断向智能手机、车载CIS等领域横向布局，打造第二增长极，维持“推荐”评级。
- **风险提示：**下游需求不及预期；市场竞争加剧；新产品进展不及预期。

图：思特威2023年收入结构



表：思特威主要手机CIS产品

XS系列	分辨率	像素尺寸	光学尺寸
550XS	50MP	1.0μm	1/1.56"
520XS	50MP	0.7μm	1/2.42"
580XS	50MP	1.22μm	1/1.28"
CS系列	分辨率	像素尺寸	光学尺寸
200/201CS	2MP	1.75μm	1/5.1"
500/501/520CS	5MP	1.12μm	1/5"
800/820CS	8MP	1.12μm	1/4"
1300/1301CS	13MP	1.12μm	1/3.06"
1600CS	16MP	1.0μm	1/3.06"

05. 投资建议

核心观点：AI产业的新范式

一、模型变革，云端融合为趋势。近年来生成式AI步入快速发展期，开源及垂直大模型百花齐放。而今年以来，最大的变革则是端侧模型的兴起，相较云侧模型，端侧模型在时延、隐私、个性化方面具备优势。各大厂商亦纷纷推出AI终端，**云端融合成为产业发展的下一趋势。**

二、云端变革，GB200解决互联瓶颈。当下AI产业发展和核心矛盾，是日益增长的应用+终端的需求，和算力本身增速瓶颈之间的矛盾。当前全球AI芯片竞争日趋白热化。云厂商为了降低资本开支，纷纷推出自研AI加速卡计划。而**GB200 NVL72**的推出，显著拉开了英伟达与其竞争对手的差距。英伟达通过架构的创新，解决了GPU之间互联带宽的问题，实现了最多576张卡1.8TB/s的双向互联带宽，显著领先其他竞争对手。Blackwell平台给服务器上游的光模块和PCB都带来了价值量的提升，而GB200的Rack架构的主要价值增量则在铜互连、电源等环节，尤其是铜互连的从零到一，更值得关注。

三、供应链变革：全面升级的算力产业链。自然语言的运算、图形的处理需要大量存储器协助运算及存储，并且AI服务器对数据传输速率有着更高要求，催生了存储及PCB价值量的增长。**1) 存储：**HBM是目前用于打破“内存墙”的重要技术之一。此外，DDR5渗透率提升，CXL、MCR等内容技术亦是行业发展的重心。**2) PCB：**服务器PCB要求高层数、高密度及高传输速率。AI服务器的OAM板及UBB板，最高使用到M7材料，20-30层的高阶HDI。单台AI服务器PCB价值量可达1万元。此外，交换机，电源PCB也有相应的价值量提升。为满足AI服务器的旺盛需求，上游封测及制造的工艺技术革新及产能扩张也尤为重要。**1) 封测：**算力芯片的主流封装方案为COWOS芯片+HBM显存。设备环节考量3D堆叠及测试设备的工艺能力，材料环节则看重轻薄化、高集成度以及导热性。**2) 制造：**台积电预计到2030年，半导体市场规模将达到1万亿美金，其中高性能计算（HPC）占比40%，将成为第一大细分市场。

四、端侧变革：身边的算力终端。AI云端融合趋势下，各大厂商发力AI终端，迎接下一波成长。**我们将AI终端的定义提炼为三要素：算力、模型、生态。**其中，算力和模型是AI终端的必要条件，算力是硬件基础，通过NPU执行神经网络运算，实现终端推理；而模型的减枝蒸馏，则可有效压缩模型参数，减少推理时间，降低端侧负载。上述两大要素决定了终端的硬实力。相较前者，生态需要长时间打磨，且需产业链厂商齐心协力配合，开发应用，挖掘垂直场景。生态建设代表了终端的软实力。我们认为**生态的建设和完善，会决定AI终端的成败。**而这一过程当中位居核心的是品牌厂商。苹果则得益于其多年的积累，及消费市场的号召力，更有机会打造出完备的生态。此外，端侧供应链的PCB、散热、光学等环节同样值得重视。

投资建议：标的方面，建议关注：**1) 云端：**工业富联、寒武纪、沪电股份、胜宏科技、澜起科技、沃尔核材；**2) 供应链：**中芯国际、华虹公司、ASMPT、长川科技、兴森科技；**3) 终端：**联想集团、小米集团、立讯精密、鹏鼎控股、领益智造、蓝思科技、华勤技术、水晶光电。

06. 风险提示

6.1

风险提示

- **下游需求不及预期：**全球经济环境、地缘政治等不确定因素下，电子行业下游需求复苏仍然存在不确定性，若需求复苏不及预期，将影响整个电子板块业绩表现。
- **大模型等发展不及预期：**AI产业的发展取决于技术的进步和迭代速度，若大模型等发展不及预期将影响AI产业的发展进程。
- **晶圆厂扩产不及预期：**晶圆厂扩产的顺利进行涉及到设备采购、工艺整合等多方面因素，若不及预期将影响上游设备材料需求。
- **新产品研发进展不及预期：**AI智能硬件浪潮初启，若新产品研发进展不及预期将影响业内公司业绩表现。国产设备、材料等供应链仍在起步阶段，较多产品处于验证环节，如果研发进度不及预期将影响相关上市公司业绩表现。

THANKS 致谢

民生电子研究团队：



分析师 方竞

执业证号：S0100521120004

邮箱：fangjing@mszq.com



分析师 李少青

执业证号：S0100522010001

邮箱：lishaoqing@mszq.com



分析师 李萌

执业证号：S0100522080001

邮箱：limeng@mszq.com



分析师 张文雨

执业证号：S0100524060002

邮箱：zhangwenyu@mszq.com



分析师 宋晓东

执业证号：S0100523110001

邮箱：songxiaodong@mszq.com

民生证券研究院：

上海：上海市浦东新区浦明路8号财富金融广场1幢5F； 200120

北京：北京市东城区建国门内大街28号民生金融中心A座19层； 100005

深圳：广东省深圳市福田区益田路6001号太平金融大厦32层05单元； 518026

分析师声明：

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并登记为注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰准确地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、不因、也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

评级说明：

投资建议评级标准		评级	说明
以报告发布日后的12个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A股以沪深300指数为基准；新三板以三板成指或三板做市指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普500指数为基准。	公司评级	推荐	相对基准指数涨幅15%以上
		谨慎推荐	相对基准指数涨幅5%～15%之间
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅5%以上
	行业评级	推荐	相对基准指数涨幅5%以上
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅5%以上

免责声明：

民生证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。

本报告仅供本公司境内客户使用。本公司不会因接收人收到本报告而视其为客户。本报告仅为参考之用，并不构成对客户的投资建议，不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，客户应当充分考虑自身特定状况，不应单纯依靠本报告所载的内容而取代个人的独立判断。在任何情况下，本公司不对任何人因使用本报告中的任何内容而导致的任何可能的损失负任何责任。

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、意见及预测仅反映本公司于发布本报告当日的判断，且预测方法及结果存在一定程度局限性。在不同时期，本公司可发出与本报告所刊载的意见、预测不一致的报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问、咨询服务等相关服务，本公司的员工可能担任本报告所提及的公司的董事。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

若本公司以外的金融机构发送本报告，则由该金融机构独自为此发送行为负责。该机构的客户应联系该机构以交易本报告提及的证券或要求获悉更详细的信息。本报告不构成本公司向发送本报告金融机构之客户提供的投资建议。本公司不会因任何机构或个人从其他机构获得本报告而将其视为本公司客户。

本报告的版权仅归本公司所有，未经书面许可，任何机构或个人不得以任何形式、任何目的进行翻版、转载、发表、篡改或引用。所有在本报告中使用的商标、服务标识及标记，除非另有说明，均为本公司的商标、服务标识及标记。本公司版权所有并保留一切权利。