民生证券
MINSHENG SECURITIES

AI投资的新范式

—— 电子行业2025年中期投资策略

电子行业首席：方竞

核心观点

我们于去年发布了AI深度报告，**提出了“CSP引领、ASIC为王”的投资观点。**而在本篇2025年中期策略报告中，我们将进一步解读算力的长期成长空间，探寻GPU和ASIC的最新动向，并挖掘国产算力和AI终端的新变化。

海外算力：近期美股AI软硬件标的纷纷创新高，**短期催化为英伟达业绩超预期，而长期驱动力则为AI赋能互联网应用，带动推理需求增长+Token消耗提升，以此为锚点，实现AI投资的ROI闭环。**算力需求高增的背景下，英伟达产品加速迭代，而CSP自研的ASIC则迎来了更快的成长。**算力的升级离不开功率和速率两条路线：1）速率方面，**我们看到了ASIC和服务器架构变化带来的PCB升级，传统光模块向CPO的演进，以及AEC的渗透率快速提升；**2）功率方面**，HVDC、超级电容为下一代服务器提供供电保障，液冷则成为芯片功耗提升下的刚需。而在英伟达的ComputeX大会上，各大厂商展示了具体的产品升级路线，也为未来几年算力行业的发展敲定了大方向。

国产算力：豆包+DeepSeek破局，国产大模型弯道超车：豆包和DeepSeek分别在多模态和轻量化两方面加速了国产大模型的发展进程。国内其他模型厂商也加速了追赶节奏，2025年以来，豆包、通义千问、百度、腾讯混元、阶跃星辰和Kimi等其他国产大模型加速了更新迭代，AI应用加速放量下推理侧需求有望提升。**算力基建加码，解决供给短板：**国内云计算厂商正加大算力储备及模型优化投入，AI计算基础设施建设布局逐步清晰，相关资本开支进入新一轮扩张周期。而短期内，国产算力基建难以满足迅速增长的需求，算力租赁成为破局之道。**向“芯”而行，国产算力破局元年：**在国产大模型密集落地背景下，芯片厂商加速适配国产算力生态。中芯国际N+1工艺已逐步成熟，N+2持续推进，构建国产算力底座；昇腾910C量产落地，920系列研发加快，性能持续逼近国际主流水平；寒武纪、海光等在AI训推方向深度布局，硬件端多点突破，生态融合加快。云端ASIC正成为算力演进主流，芯原等设计企业快速成长，与海内外头部厂商形成紧密合作，成长弹性充足。

核心观点

AI终端：手机AI功能仍待完善，但仍有光学、折叠屏、指纹识别等硬件的结构性创新。智能眼镜市场近期不断升温，销售火热。复盘AI/AR眼镜的发展历程，Meta&Rayban AI眼镜的成功证明了“先眼镜后智能“的思路，即AI眼镜替代传统眼镜，然后在AI眼镜上增加AR效果使消费者逐步提升对眼镜这类新终端的接受度。**我们认为从AI向AR演进，品牌厂商与光学方案商深度绑定，光学/显示逻辑有望得到充分演绎：**①AI眼镜交互模式和功能比较单一，AR眼镜加入显示功能，能显著提升用户体验；②光学显示模块成为AR眼镜中BOM占比最高的部分之一，且相较于AI眼镜是纯增量环节，目前主流方案是MicroLED+衍射光波导。

我们坚定看好AI产业的长期叙事，**英伟达持续强势，云厂商崛起，国产算力突破的当下，投资机遇也会更加多元化。**具体到细分赛道，算力链重点关注服务器、PCB、CPO、铜缆、电源、液冷等产业链，这也是国内企业深耕多年，具备优势的环节。而AI终端则相对预期充分，需观察热门新品的放量节奏。

建议关注：1)服务器：工业富联、华勤技术；**2)算力芯片：**芯原股份、寒武纪、海光信息；**3)PCB：**沪电股份、胜宏科技、广合科技、生益科技、景旺电子、威尔高；**4)铜/光互联：**瑞可达、博创科技、太辰光、东山精密；**5)电源及温控：**禾望电气、中恒电气、麦格米特、申菱环境、江海股份；**6)品牌及代工：**小米集团、影石创新、歌尔股份、国光电器；**7)SOC：**乐鑫科技、恒玄科技、星宸科技； **8)存储：**兆易创新、普冉股份； **8)渠道商：**博士眼镜、孩子王、明月镜片等。

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

目录

CONTENTS

01 海外算力：推理需求高增下的ROI闭环

1.1

目录

CONTENTS

1.1.1 美股AI软硬件核心标的纷纷创新高

- 4月7日关税战风波以来，海外AI在低预期下基本面稳步改善，软硬件板块持续上涨，其中龙头厂商英伟达涨幅45%，硬件板块中Coreweave涨幅最大，达到195%；软件板块中Palantir涨幅最大，为77%。
- 我们认为，当下海外AI板块龙头厂商股价纷纷创新高，国内AI公司股价处于相对低位，值得重点关注。

图：美股AI软硬件公司2025年4月7日以来涨幅

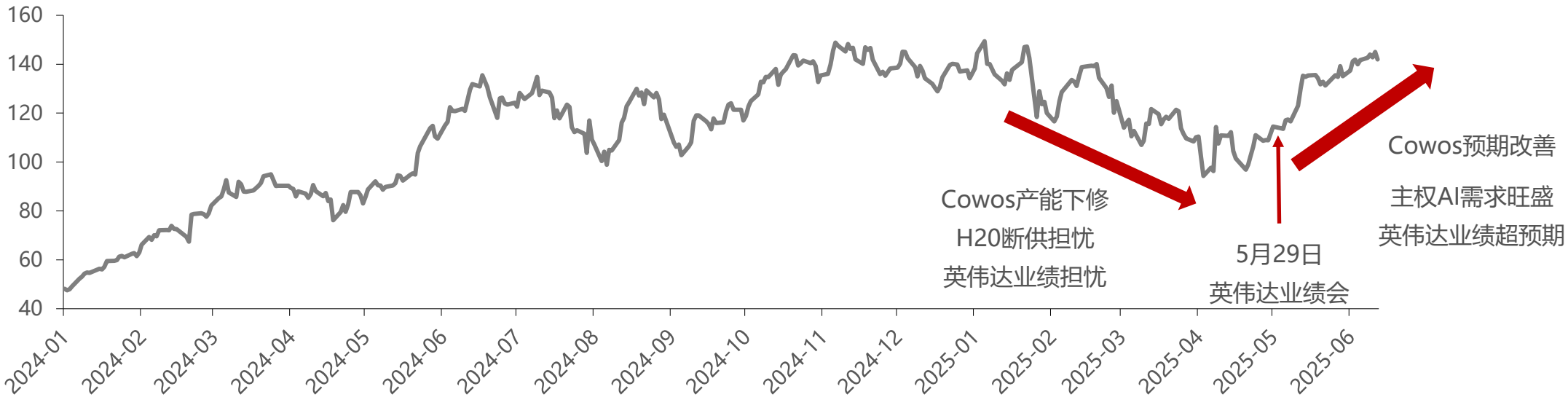
硬件					软件				
代码	公司名称	收盘价-4月7日	收盘价-6月13日	涨跌幅	证券代码	证券简称	收盘价-4月7日	收盘价-6月13日	涨跌幅
CRWV.O	CoreWeave	49.85	147.19	195%	PLTR.O	Palantir	77.84	137.40	77%
CRDO.O	Credo	35.91	73.49	105%	DUOL.O	多邻国	296.51	478.48	61%
ORCL.N	甲骨文	126.70	215.22	70%	SNOW.N	Snowflake	131.04	208.18	59%
MU.O	美光科技	68.37	115.60	69%	APP.O	AppLovin	232.22	364.49	57%
AVGO.O	博通	154.14	248.70	61%	DDOG.O	Datadog	88.20	120.45	37%
APH.N	安费诺	60.90	92.49	52%	INTU.O	Intuit	555.27	753.98	36%
DELL.N	DELL	74.52	109.56	47%	NOW.N	ServiceNow	735.50	988.66	34%
NVDA.O	英伟达	97.63	141.97	45%	MSFT.O	微软	357.20	474.96	33%
AMD.O	AMD	83.64	116.16	39%	VEEV.N	Veeva Systems	214.74	282.55	32%
MRVL.O	Marvell	50.94	67.19	32%	SAP.N	SAP	239.45	293.36	23%
SMCI.O	SMCI	33.00	41.56	26%	GOOGL.O	谷歌	146.57	174.67	19%

资料来源：Wind，民生证券研究院，注：收盘价按照前复权后价格计算

1.1.1 复盘本轮美股上涨，直接原因：英伟达业绩超预期

- **5月29日英伟达披露FY25Q3业绩，H20影响下业绩依旧超预期。**英伟达FY25Q3营收441亿美元，同比+69%，环比+11%，原指引430亿美元（±2%），Bloomberg一致预期433亿美元，超预期。英伟达表示，H20禁令对Q1、Q2收入影响分别为25、80亿美元，Q2业绩指引在H20断供影响下低于预期，但后续边际利空因素基本消除。
- **以英伟达业绩会为转折点，市场预期从悲观转向乐观。**英伟达业绩会前，市场持续担忧算力需求，2025年和2026年英伟达Cowos产能预期也在不断下修，而英伟达业绩会后，算力需求预期边际改善。6月11日英伟达巴黎GTC大会上预言，未来两年欧洲AI算力将实现十倍增长，目前正在筹建超20个“AI超级工厂”，单个AI工厂建设成本可能达到500亿美元；欧洲主权AI需求快速提升，黄仁勋预计欧洲算力需求将达到300万P，英伟达与Mistral共建“AI云平台”，搭载1.8万颗B卡。

图：英伟达2024年初至今股价走势（美元）



资料来源：英伟达，Bloomberg，Wind，富途，民生证券研究院，注：收盘价按照前复权后价格计算

1.1.1 复盘本轮美股上涨，直接原因：算力需求旺盛

- **除英伟达外，其他市场也反应出算力需求紧缺的态势，而ASIC有望成为本轮算力需求中增长最快的板块。**目前谷歌、亚马逊、微软、meta等海外CSP和字节、阿里等国内云厂商均在对接Design Serves公司匹配自研算力，2024年主流云厂商中仅有谷歌的TPU较为成熟且出货量较大，进入2025年，亚马逊的Tranium系列加速卡也将快速放量。而meta、微软、OpenAI等厂商的加速卡则有望在2026年开始大批量出货。
- **年初以来，云厂商自研ASIC出货量不断上修，**据VerifiedMarket，2024年全球ASIC市场空间为65亿美元，预计2033年将增长至152亿美元，复合增速达到12.8%。博通在6月的业绩会上表示目前已经与7个客户展开定制化ASIC业务合作，包括谷歌、Meta、OpenAI等客户；Marvell则表示当前定制化AI芯片设计市场需求旺盛，预计2028年需求将达到940亿美元，CAGR增速达到35%。
- **云厂商自研加速卡的主要优势有三点：**1）更高的单位算力性价比；2）更灵活的性能参数选择以匹配自身需求；3）更灵活的服务器产业链配套选择。当云厂商的软件生态、互联方案等逐步成熟，自研有望成为其最核心的算力来源。

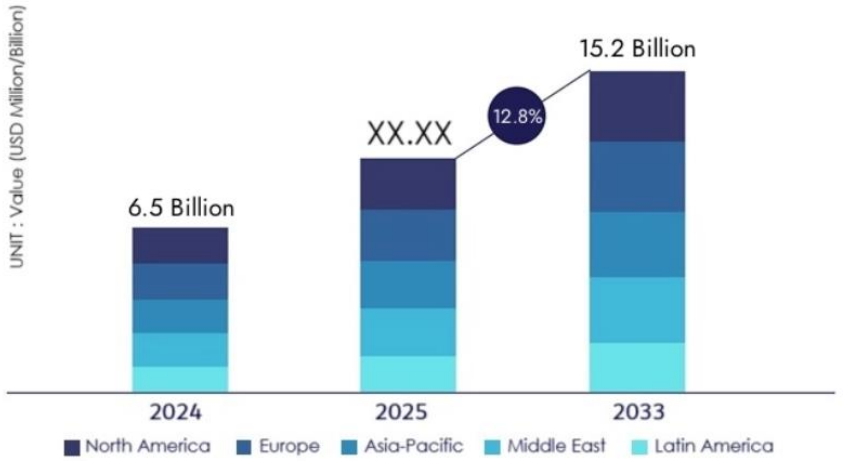
图：云厂商自研ASIC加速卡进展

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	400
		TPUv5P	2023	-	918	459	-	HBM2	95	800
Meta	推理	MTIA v2	2024	5	354/708	177/354	2.76	LPDDR5	128	-
微软	训练	Maia 100	2023	5	1600	800	-	HBM3	64	1200
亚马逊	训练	Trainium2	2023	4	861	431	215	-	96	-
	推理	Graviton4	2023	-	-	-	-	DDR5	-	-
TESLA	训练	D1	2021	7	362	362	22.6	-	32	-

资料来源：各公司官网，Semianalysis，VerifiedMarket等，民生证券研究院

注：未标注的数据为没有在公开渠道披露的信息

图：2024-2033年全球ASIC市场空间及增速

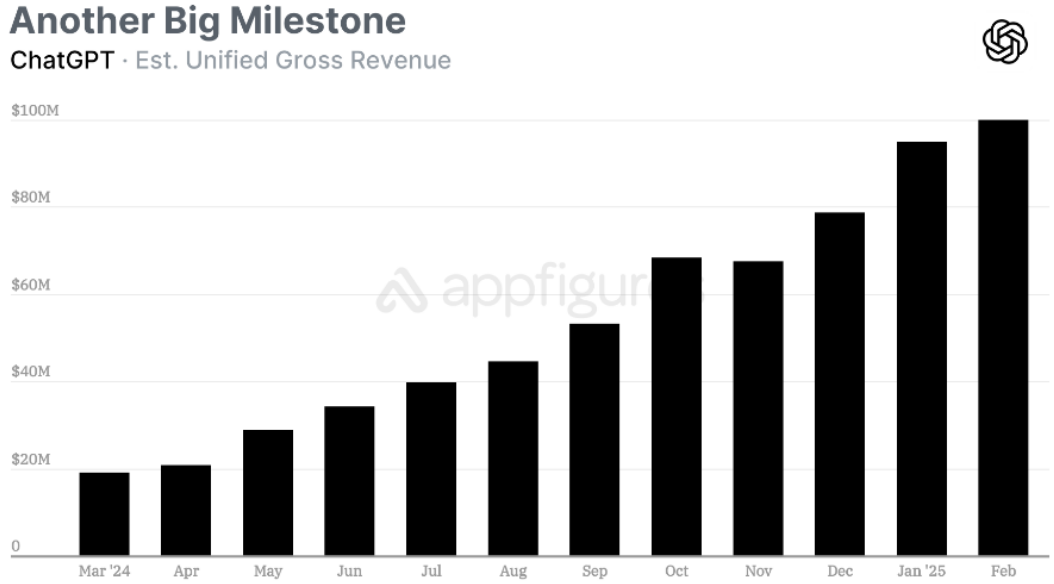


1.1.2 复盘本轮美股上涨，根本原因：大模型ROI闭环

- **目前大模型正在发生从量变到质变的过程，模型的Token消耗数量快速提升，意味着大模型的渗透率快速提升，同样也意味着应用侧的落地提速。**以谷歌的Gemini 2.5为例，在2025年I/O大会上谷歌公布了Gemini 2.5的月度Toekn消耗量变化趋势，Gemini 2.5的Token消耗量以每个月翻倍的速度快速增长，5月的月均Token数从9.7万亿激增至480万亿，同比增长50倍。
- **大模型目前已经给厂商提供了相当可观的回报，ROI逐步实现闭环。**以OpenAI为例，2025年OpenAI的ARR已经突破100亿美元，相比2024年的55亿美元同比提升80%，稳定的客户订阅收入大幅提升了大模型厂商的盈利能力。

图：谷歌Gemini 2.5月度Token消耗量

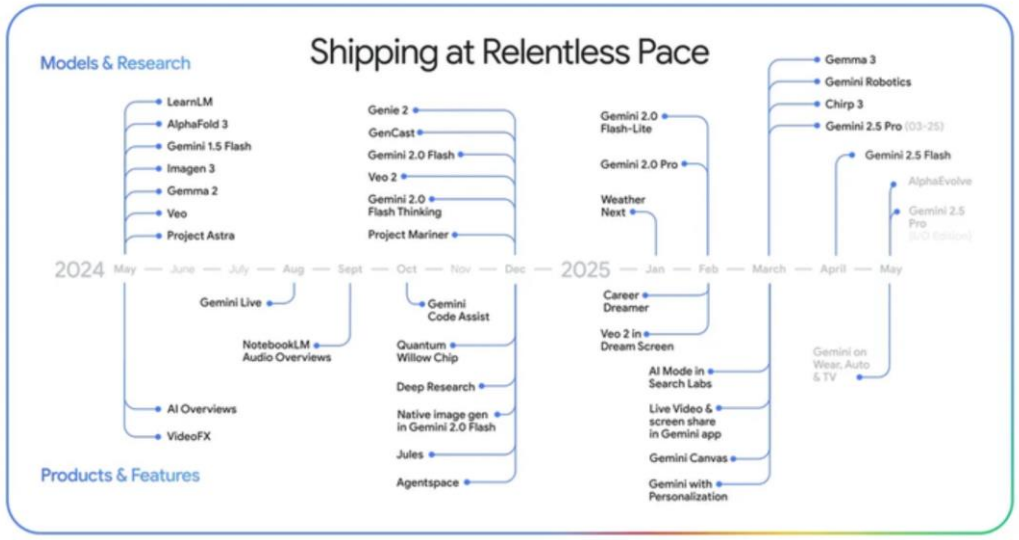
图：ChatGPT月度收入



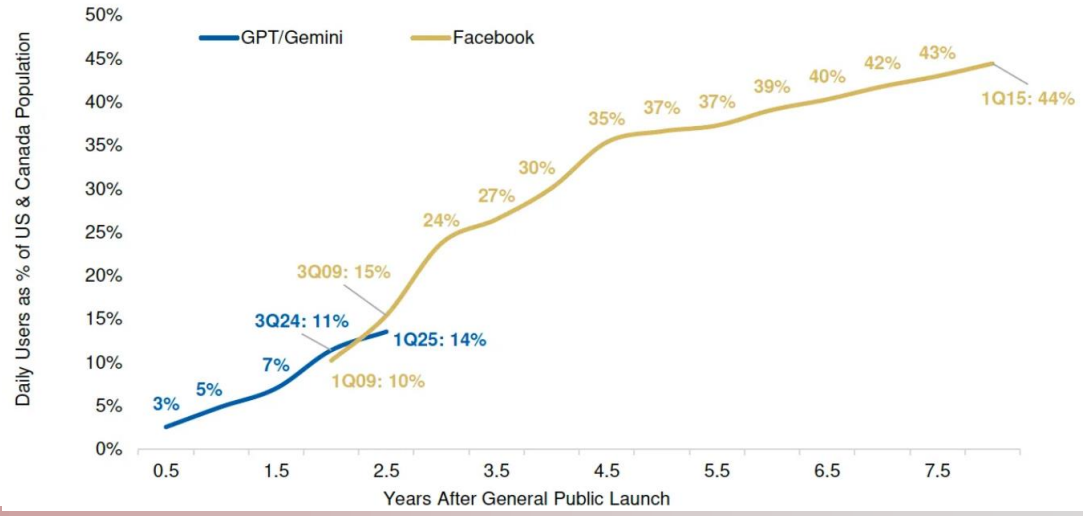
1.1.2 复盘本轮美股上涨，根本原因：大模型ROI闭环

- 前期，市场更多关注全新的热门AI应用，而我们认为**当下AI推理需求的快速增长，更多源自AI大模型赋能传统互联网应用，潜移默化的改造普通用户的行为习惯。**
- **谷歌率先利用Gemini打造AI Agent。**Gemini 2.5目前已经接管了谷歌的搜索功能，且Deep Research模式允许以用户自己上传资料并打通使用各种数据库，使得谷歌全家桶转变为AI Agent。除谷歌外，其他大模型厂商也纷纷把原生应用和AI大模型结合，如Telegram与xAI达成合作，将Grok AI聊天机器人引入其平台。
- **上下文的扩充提升了模型的认知推理能力。**Gemini 2.5 Pro在Gemini 2.0基础上，通过优化基础模型和后训练技术，将模型认知推理能力提升到了新高度，从而能更好地理解上下文，并处理复杂问题。Gemini 2.5 Pro可支持100万token的上下文窗口，相当于可在一次提示中处理约75万英文单词的文本。
- **大模型加速接入云厂商的原生应用，有望大幅加速AI应用的落地节奏，使得大模型ROI实现闭环，反向拉动训练和推理的算力需求。**据Retailgentic，截止1Q25生成式AI的渗透率仅为14%，参考前两轮的互联网革命和移动互联网革命，生成式AI有望凭借各大云厂商的传统业务导入，在未来几年实现渗透率的加速提升，从而带动推理需求高增。

图：谷歌大模型全家桶打造AI Agent



图：大模型和Facebook在发布后渗透率对比



资料来源：谷歌I/O大会，Retailgentic等，民生证券研究院

1.2

目录

CONTENTS

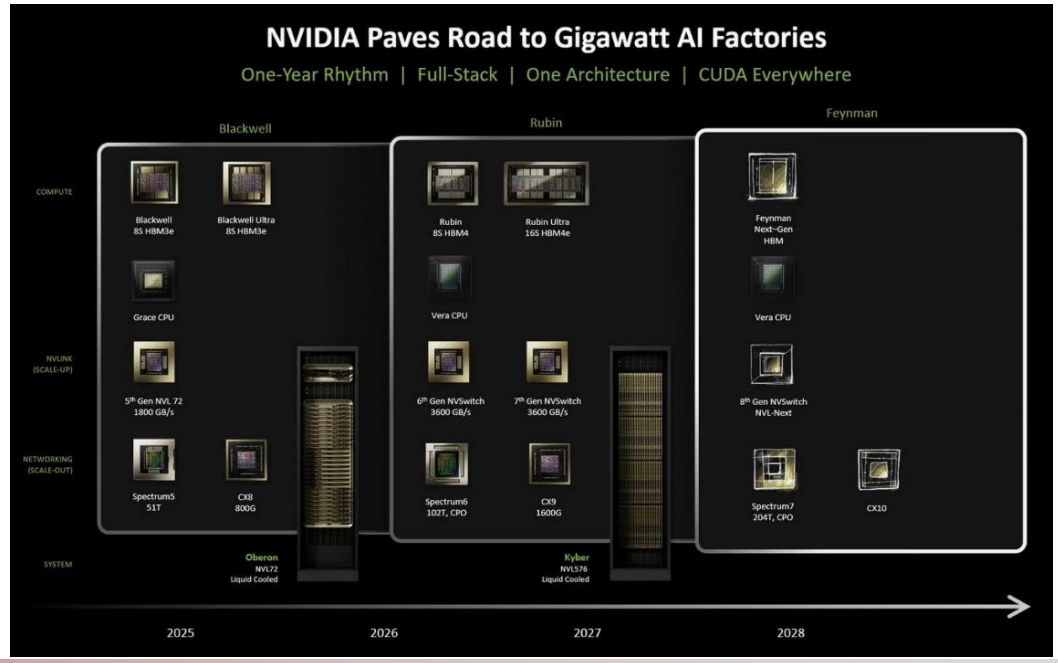
1.2.1 英伟达算力路线图：加速卡迭代周期

- **2024年英伟达推出了全新的Blackwell架构加速卡，产品性能全面提升。**2022年以后英伟达加快了产品迭代节奏，从此前的2年一代产品提升为每年迭代。英伟达最新的B300加速卡单芯片FP4稠密算力达到15PFlops，相较上一代B200提升50%，且在HBM容量上进一步提升至288GB。
- **英伟达服务器正在从8卡架构转向72卡，未来将推出144卡和576卡机柜。**2024年英伟达推出了GB200服务器，目前已经进入全面生产环节，GB300 NVL72将于2025下半年推出，相较上一代产品，处理能力提升1.5倍；2026年下半年，英伟达将推出全新的Vera Rubin NVL144机柜，并且推出全新的CPU，NVLink6.0以及HBM4.0；2027年英伟达将推出Rubin Ultra576机柜单个机柜的算力将达到15EFlops，功耗达到600kW，算力密度将再次得到大幅度提升。此外，2028年英伟达将推出更新一代的Feynman架构产品，单芯片算力进一步提升。

图：英伟达加速卡迭代节奏及性能对比

	A100	H100	H200	B200	B300
发布时间	2020	2022	2023	2024	2025
封装	CoWoS-S	CoWoS-S	CoWoS-S	CoWoS-L	CoWoS-L
HBM（GB）	80	80	141	192	288
die颗数	1	1	1	2	2
算力（TFlops）	624	3958	3958	10000	15000
功率（W）	300	700	700	1000	1200

图：英伟达产品Road Map



资料来源：英伟达GTC大会，Semianalysis，英伟达官网，民生证券研究院，
注：A100/H100/H200采用8位稀疏算力，B200/B300采用4位稠密算力

1.2.1 英伟达B30：海外算力封锁下的GPU新方案

- 英伟达正在为中国市场研发一款名为“B30”的降规版AI芯片，这款芯片将首度支持多GPU扩展，允许用户通过连接多组芯片来打造更高性能的计算集群。B30芯片预计将采用最新的Blackwell架构 （目前唯一适应美国要求的架构），使用GDDR7显存，而非高频宽内存（HBM），也并未采用台积电的先进封装技术。B30售价预计在6500美元至8000美元之间，远低于H20芯片的1万至1.2万美元。
- 由于未采用先进封装技术，且在多GPU扩展时可能增加系统整体功耗，B30在大规模集群部署时，散热和功耗管理也将面临挑战。同时受显存带宽和架构调整影响，B30单芯片在处理高精度计算任务，如FP16时，效率低于H20 。不过通过多GPU扩展，100块B30组成的集群理论性能可达H20集群的85%，这使其在大规模计算任务中仍具备一定竞争力，能够满足部分云服务提供商的中小规模计算需求以及一些对成本较为敏感的大规模模型训练场景。

英伟达出口中国GPU系列解析

GPU型号	架构	CUDA核心数	Tensor核心数	显存(HBM)	带宽	NVLink	功耗	主要用途
A100	Ampere	6912	432	40GB/80GB HBM2e	1.6/2TB/s	NVLink 600GB/s	400W	适用于深度学习训练&推理
H100	Hopper	14592	456	80GB HBM3	3.35TB /s	NVLink 900GB/s	700W	适用于大模型AI训练&超大模型推理
A800	Ampere	6912	432	40GB HBM2 /80GB HBM2e	1.6-2TB/s	NVLink 400GB/s	400W	适用于推理&训练
H800	Hopper	16896	528	80GB	3.35TB/s	NVLink 400GB/s	700W	中国市场AI训练&大规模推理
H20	Hopper（阉割）	未知	未知	96GB	受限4.0TB/s	NVLink 900GB/s	400W	中国市场中等规模训练&推理
B30	Blackwell	未知	未知	GDDR7	1.7TB/s	未知	250W	中国市场中小等规模训练&推理

英伟达B30 VS H20

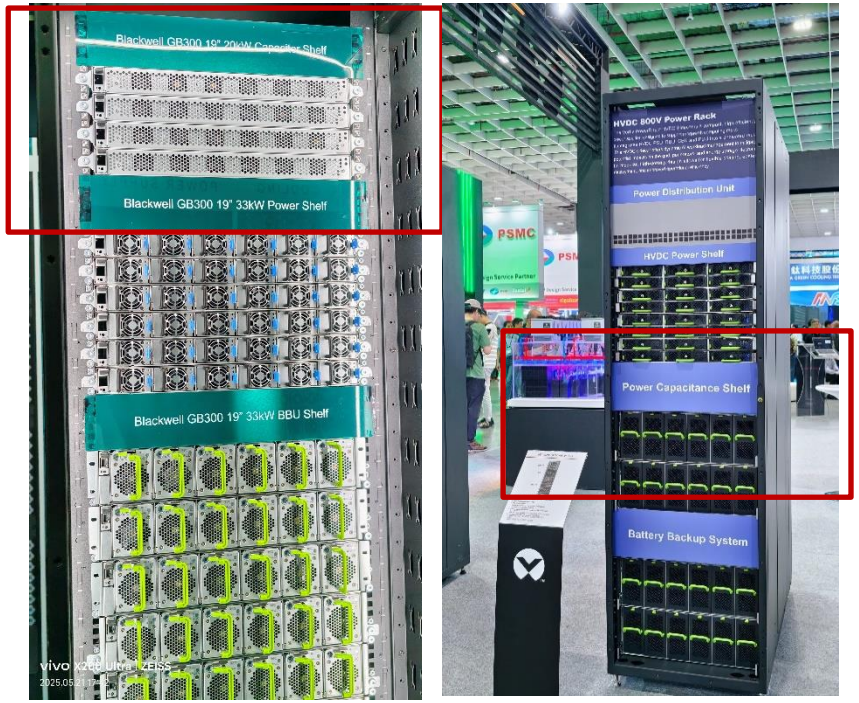
参数	B30	H20
架构	Blackwell（阉割版，台积电4N工艺）	Hopper（台积电4N工艺，CoWoS先进封装）
显存类型	GDDR7	HBM3（部分版本为141GB HBM3e）
FP32算力	200 TFLOPS (提升40% vs H20)	44 TFLOPS（阉割版）
FP16算力	80-100 TFLOPS	148 TFLOPS（稀疏计算）
单卡售价	6500-8000美元（降价40%-50% vs H20）	1.2万-1.5万美元（2024年预订价，2025年市场波动至1.3万-1.8万美元）

1.2.2 英伟达算力路线图：从ComputeX大会看服务器创新点

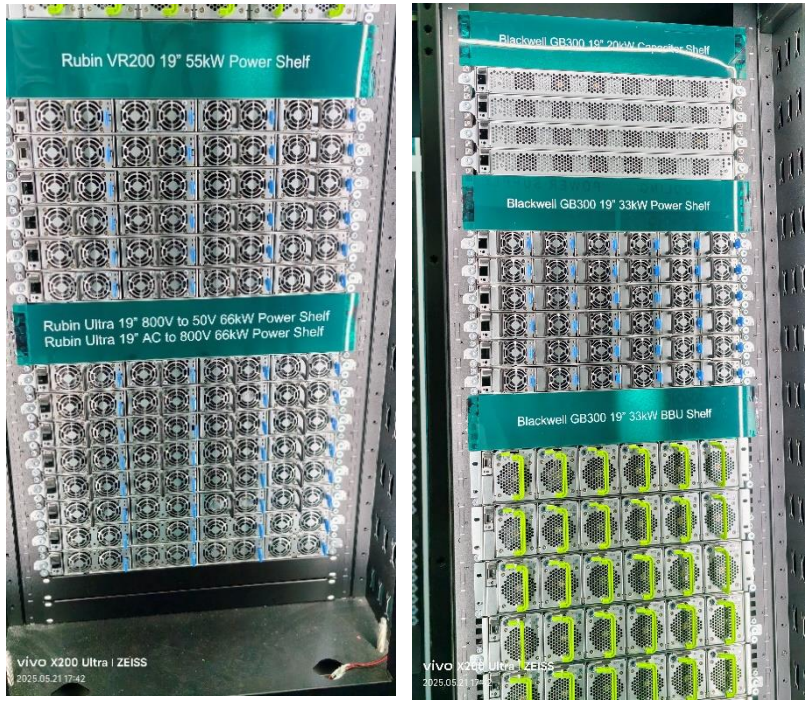
- **HVDC**：在ComputeX大会上，维谛、台达、光宝、Foxcon现场展示了800 HVDC的电源柜demo，大会还展示了HDVC 800V Sidecar电源柜，将PSU、BBU、超容进行整合，供应左右两侧2~4柜GPU Server。
- **超级电容**：维谛技术、光宝在ComputeX现场展出Power Rack的demo搭载电容模组，富士康态度则相对模糊
- **PSU**：由于GB300机柜内空间有限，在ComputeX展出的方案中，独立电源柜（融入PSU、BBU、超容等）成为趋势；

图：维谛技术 和光宝800V HVDC

图：维谛及光宝均展示了搭载超级电容的机柜demo



图：GB300和Rubin Ultra Powershelf和BBU

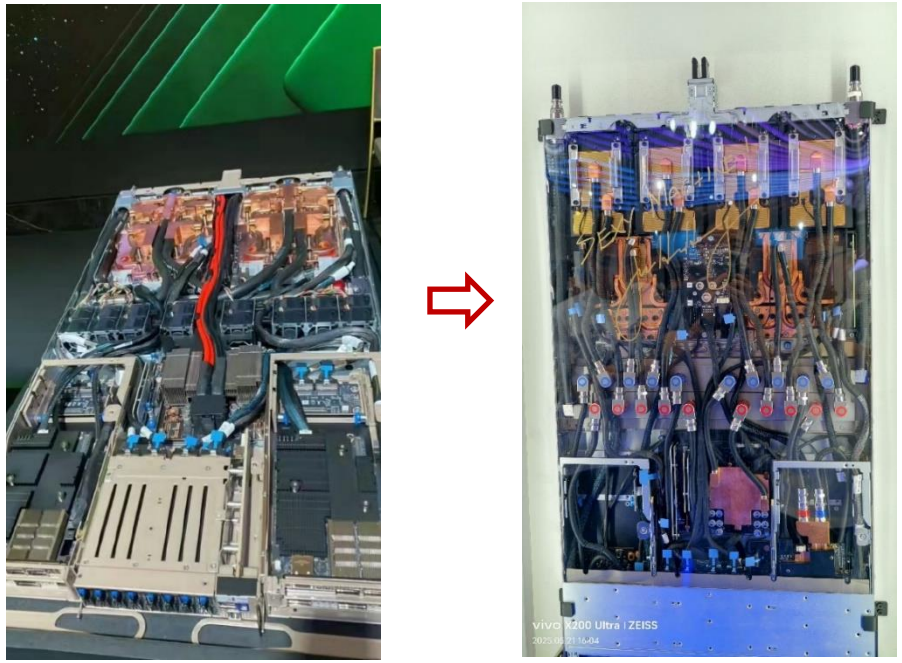


资料来源：ComputeX大会现场实拍，民生证券研究院

1.2.2 英伟达算力路线图：从ComputeX大会看服务器创新点

- 在GB200和GB300中，英伟达使用UQD和NVQD进行液冷链接，其中柜外采用UQD（已应用于GB200），柜内采用NVQD（GB300率先使用，体积相比UQD更小，链接稳定性等方面做出优化）。
- **GB300 compute tray的液冷系统包括14组液冷快接头，其中内部12组，外部2组，相比GB200（内4外2）数量翻倍以上。**
- 除了搭载在独立电源柜中的BBU方案外，英伟达还展示了搭载在Switch Tray内部的BBU方案，该方案在GB300 Switch Tray中预留了用于放置BBU的空间。

图：Compute tray GB200（左）VS GB300（右）



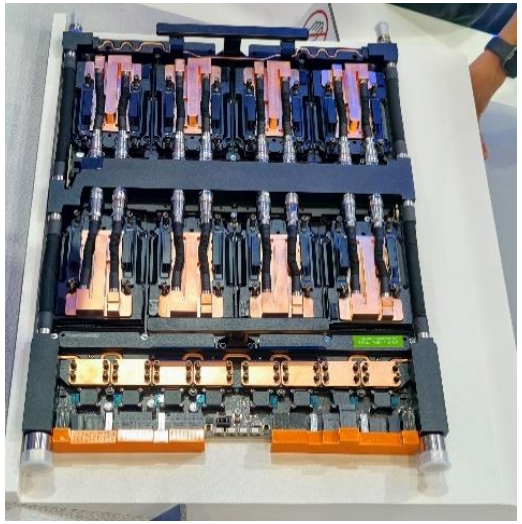
图：GB300 Switch Tray预留空间用于BBU

资料来源：ComputeX大会现场实拍，民生证券研究院

1.2.2 英伟达算力路线图：从ComputeX大会看服务器创新点

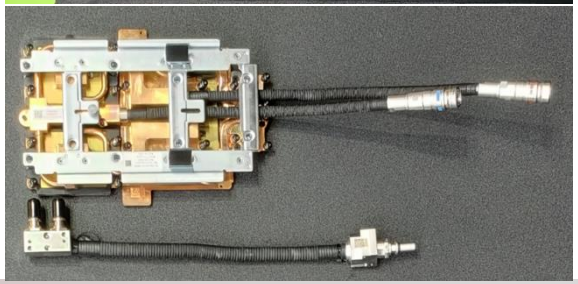
- **英伟达的B300八卡服务器有风冷和液冷两种模式供客户选择，其中B300液冷高度为2-3U，风冷则为7-8U，液冷更节省集群空间。**
- **MGX Cold Plate Inner Manifold（冷板内部歧管）供应商：**双鸿科技、奇宏科技、Boyd、酷冷至尊、CoolIT、台达电子、富士康、Lead Wealth、品达科技、立敏达；
- **MGX 44RU Manifold（分歧管）供应商：**双鸿科技、奇宏科技、品达科技、台达电子、富士康、Lead Wealth、光宝科技、酷冷至尊；
- **UQD（快接头）供应商：**双鸿科技、富世达、丹佛斯、英维克、富士康、Lead Wealth、帕克、立敏达、史陶比尔、诺通；
- **NVQD（GB300快接头）供应商：**双鸿科技、富世达、比赫、酷冷至尊、英维克、富士康、Lead Wealth、嘉泽、立敏达、泰昌、诺通、中金科工业。

图：B300 8卡液冷服务器



图：B300 8卡风冷服务器

图：MGX 44RU Manifold以及UQD、NVQD供应商

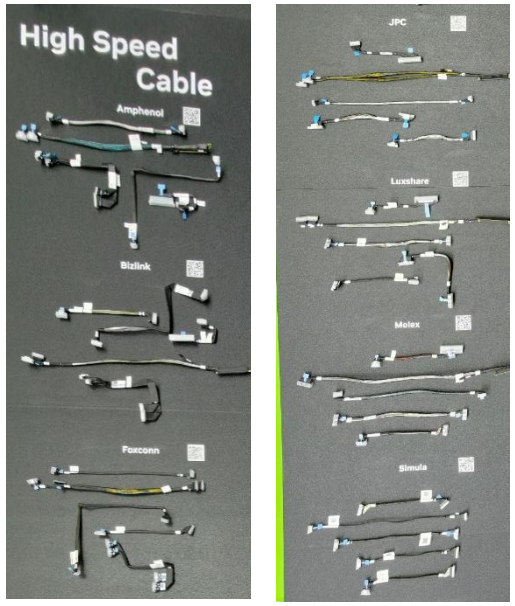


资料来源：ComputeX大会现场实拍，民生证券研究院

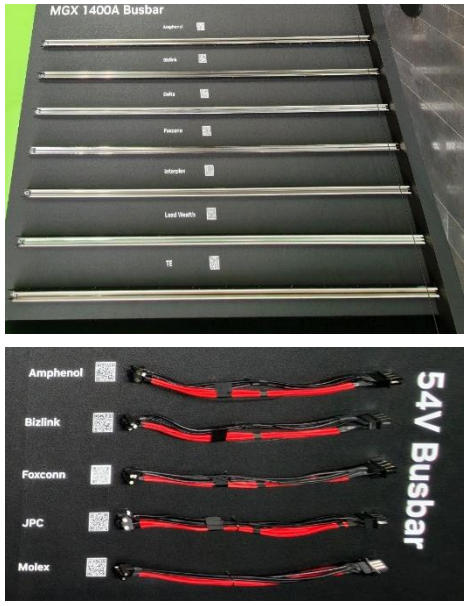
1.2.2 英伟达算力路线图：从ComputeX大会看服务器创新点

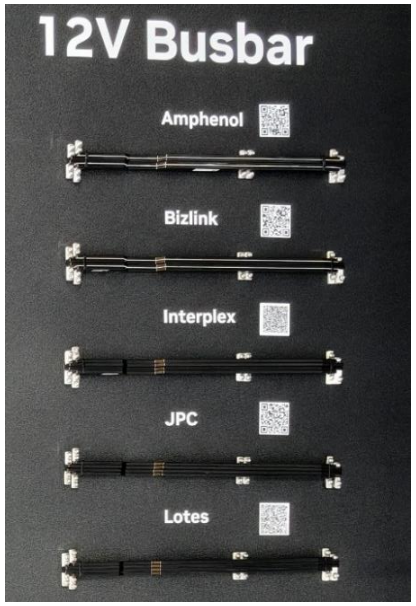
- **ComputeX大会展示了AI服务器电力电缆和高速铜缆的解决方案，主要供应商包括：**
- **High Speed Cable（高速铜缆）供应商：**Amphenol、Bizlink、富士康、佳必琪、立讯精密、Molex、矽瑪科技；
- **60A Power Whip（电力电缆）供应商：**Amphenol、Bizlink、佳必琪、瑞可达；
- **MGX 1400A Busbar（电力母线）供应商：**Amphenol、Bizlink、台达电子、富士康、Interplex、Lead Wealth、TE；
- **12V Busbar（电力母线）供应商：**Amphenol、Bizlink、Interplex、佳必琪、嘉泽；
- **54V Busbar供应商：**Amphenol、Bizlink、富士康、佳必琪、Molex。

图：High Speed Cable（高速铜缆）



图：MGX 1400A和12V Busbar（电力母线）





图：60A Power Whip（电力电缆）

资料来源：ComputeX大会现场实拍，民生证券研究院

1.2.3 ASIC成为英伟达加速卡的核心竞争者

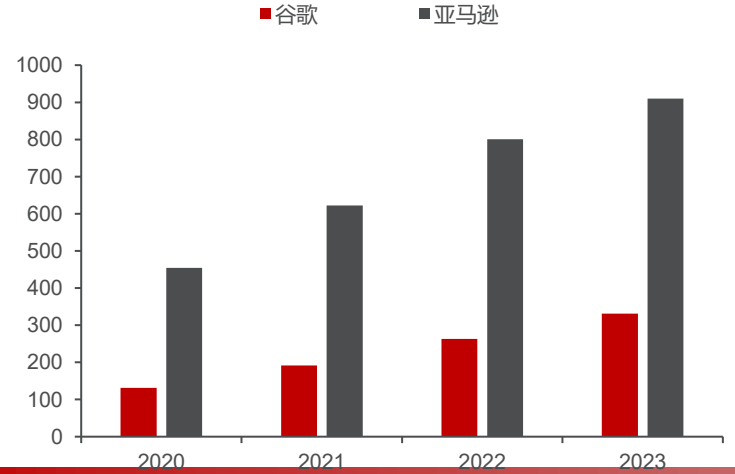
- **当前全球AI芯片仍主要被英伟达垄断，但ASIC已经成为算力芯片的主要增长极，且市场份额不断提升。**2023年以来，谷歌、AWS、meta、微软等公示相继入局ASIC赛道，挑战英伟达的垄断地位，2025年亚马逊的Trainium芯片和谷歌的TPU加速卡出货量均达到了百万颗量级，2026年以后则会有更多的厂商进入到百万颗芯片梯队，全球加速卡市场从一家独大转为多强混战的局面已成定局。
- 随着ASIC芯片的出货分额不断提升，云厂商自研加速卡供应链同样值得重视，同时ASIC服务器价值量更大的PCB、光模块、电源等环节，则有望获得更大的业绩弹性。

全球龙头			英伟达的追赶者				云厂商自研加速卡		
英伟达			AMD	MI300X	MI325X	MI355X	Google	TPU	
H100	H200	GH200					AWS	Trainium	Graviton
B200	GB200	GB300	昇腾	910B	910C		Microsoft	Athena	
L40	L40S	L4					Meta	Oculus	MITA
L20	H20	L2	Intel	Gaudi	Goya		Tesla	Dojo	FSD
							腾讯	燧原	
			其他玩家	寒武纪	海光	景嘉微	阿里巴巴	平头哥	
				沐曦等一级市场公司			百度	昆仑芯	

1.2.3 ASIC的成长态势，以谷歌和亚马逊为例

- 谷歌和亚马逊同为全球云计算业务的主要玩家。谷歌TPU加速卡最早始于2006年，2016年推出了TPUv1，大幅领先于其他云计算公司，而亚马逊的Trainium1芯片推出时间为2022年；但从云计算业务收入来看，**亚马逊2023年云计算业务收入为908亿美元，相较于谷歌的331亿美元具有显**著优势。
- **从加速看产品性能来看：**谷歌2024年推出的TPUv6e单卡BF16算力达到918TFLOPs，互联带宽则达到3584GB/s，相较于亚马逊更有优势，而亚马逊的Trainium2则在服务器架构上支持更多的加速卡高速互联，机柜内可以支持64卡互联。
- 从服务器架构来看，谷歌仍然采用传统的AI服务器架构，一个Shelf内放两个Board，每个Board放置4张TPU和1张CPU，8卡之间通过ICI Network高速互联；而**亚马逊则推出了机柜架构，最多可以支持2个机柜，64张加速卡通过AEC高速互联。**

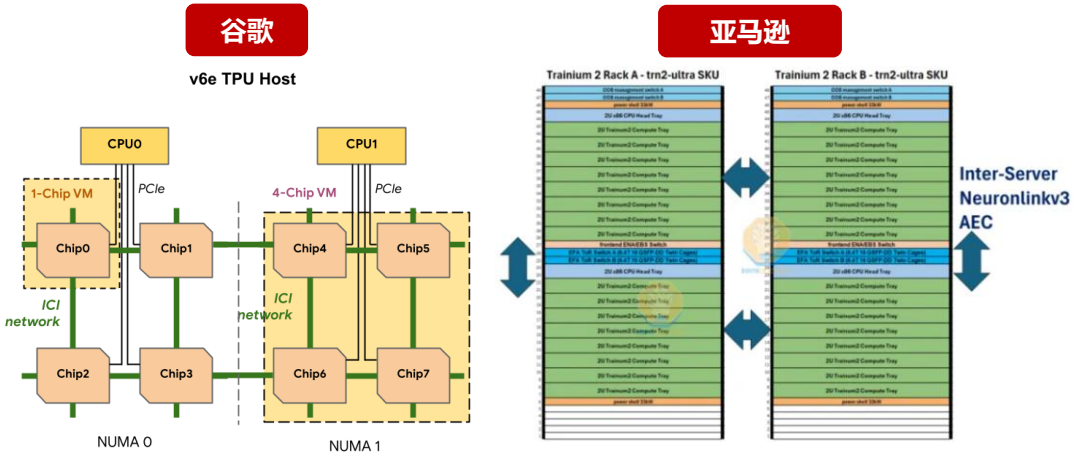
谷歌和亚马逊云业务收入（亿美元）



谷歌和亚马逊最新加速卡新能对比

厂商	谷歌	亚马逊
加速卡	TPUv6e	Trainum2
发布时间	2024	2024
int8算力（TOPs）	1836	1300
BF16算力（TFLOPs）	918	650
互联带宽（GB/s）	3584	640
互联数量	2~8	16~64

谷歌和亚马逊AI服务器架构对比



资料来源：谷歌官网、AWS官网、Semiconductor、民生证券研究院

1.3

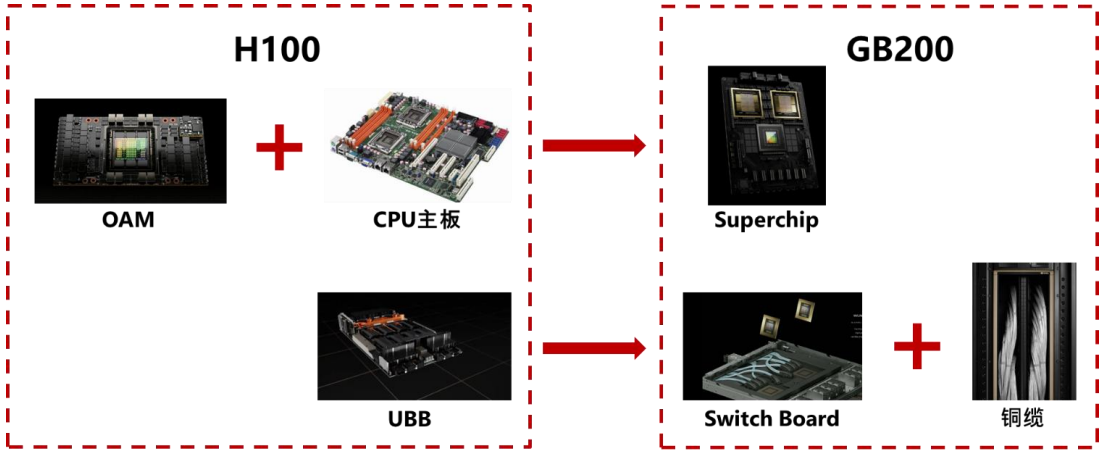
目录

CONTENTS

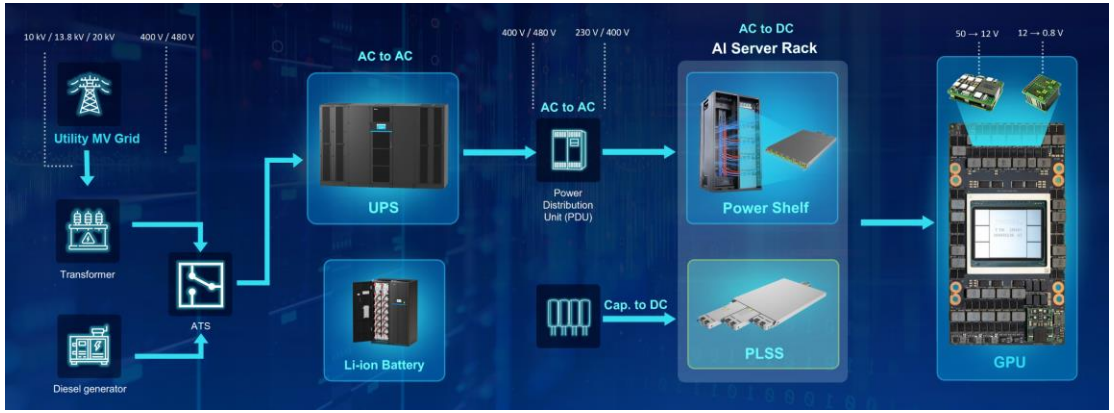
1.3 速率及功率——AI算力的两大核心矛盾

- **我们于1月深度报告中提出，2025年AI算力的核心矛盾在于“传输速率+功率密度”**，前者背后主要是PCB、CPO、AEC等，后者背后是温控液冷+电力系统等。
- **速率主要用于解决互联瓶颈。**从光模块到CPO，从DAC到AEC，以及PCB的材料和用途创新，都是业内在解决速率问题上的技术演进，将成为速率产业升级的重要组成部分，引领光电共进的产业趋势。
- **功率主要面对供电及温控难题。** 伴随摩尔定律放缓，制程升级迭代延后，功耗墙成为挡在高算力需求前面的拦路虎，对温控和电源系统提出挑战。温控方面，风冷已难以满足需求，功耗超过700W后液冷成为刚需；电源方面，高功率电源成为刚需，HVDC、超容等全新方案，保证供电系统稳定及安全性。

图：GB200服务器PCB方案升级



图：AIDC电气设备架构



1.3.1 PCB——关注下一代英伟达机柜PCB升级

1）Socket方案：

GB200：Compute tray的PCB全部采用HDI

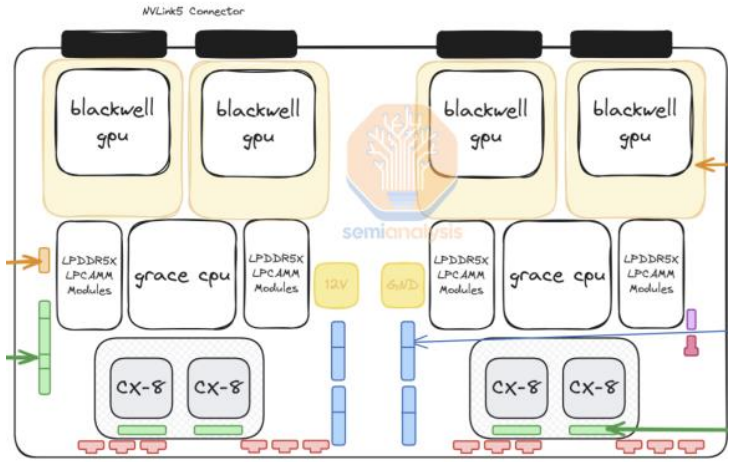
GB300：采用Socket设计，PCB回归OAM+UBB组合，GPU通过Socket搭载至OAM，CPU搭载至UBB；HDI面积减少，高多层面积增加

预计GB300 Compute tray初期仍然使用类似GB200的Bianca板；Socket有望搭载至GB300的中后期的Coedeliar板（or Rubin使用）

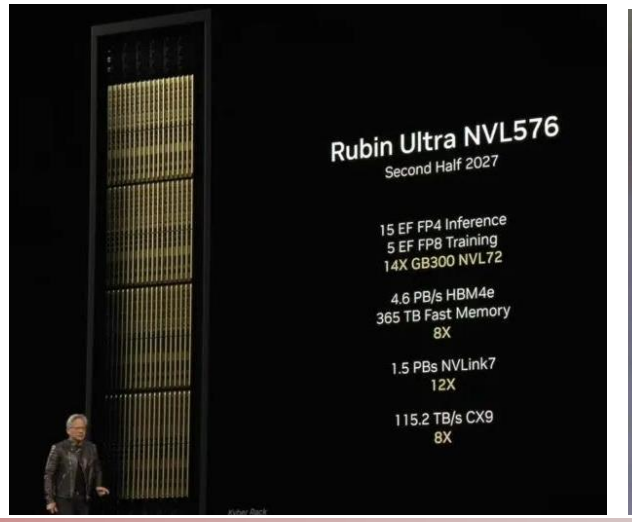
2）高速背板方案：

黄仁勋于GTC大会宣布Rubin ultra NVL576架构，该机柜由四组Canister组成，每个Canister由18个Compute tray+9个Switch tray正交放置组成，Compute及Switch之间通过高速背板实现互联，该PCB背板或将使用PTFE/传统M9覆铜板材料，预计价值量较高。黄仁勋宣布该机柜方案将于2H27正式推出。

图：搭载Socket方案的PCB示意图



图：NVL576机柜方案及高速背板（右）



1.3.1 PCB——重视云厂商自研AISC芯片PCB放量

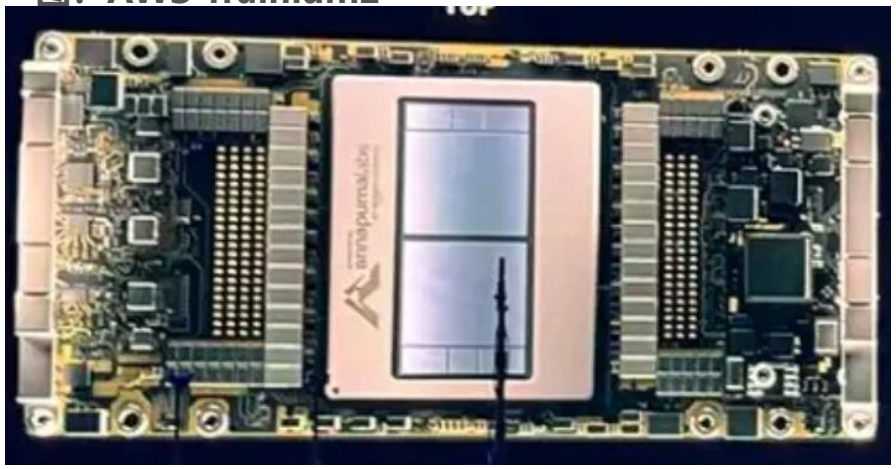
当前谷歌、微软、亚马逊、Meta四大海外云厂商中，谷歌TPU及亚马逊Trainium已形成大规模出货，微软Maia及Meta的MTIA出货量较少，但随着性能的改善自26年开始或显著放量。**云厂商自研ASIC的PCB总体呈现高多层为主，层数多，材料等级要求高的特点（多为M8），预计单芯片对应PCB价值量较高，生益电子、沪电股份等算力PCB核心供应商显著受益。**

与英伟达设计方案不同，谷歌及亚马逊的AI服务器既没有采取类似H100的八卡架构，也没有采取像类似GB200将CPU与GPU集成在一张PCB板的方案。如谷歌最新一代（第六代）的TPU芯片Trillium，将每四颗TPU搭载至一张PCB上，组成Compute Tray；CPU单独搭载至另一张PCB中，放置于另一个Tray。

亚马逊同样采取了将CPU及GPU分离的设计方案，两颗Trainium2芯片搭载至一张PCB上，组成Compute Tray，而两颗CPU单独搭载至另一张PCB上，组成Head Tray。

图：谷歌TPU

图：AWS Trainium2



资料来源：中国科学院文献情报中心，半导体行业观察，AWS,美通社，芯智讯，华尔街见闻，民生证券研究院

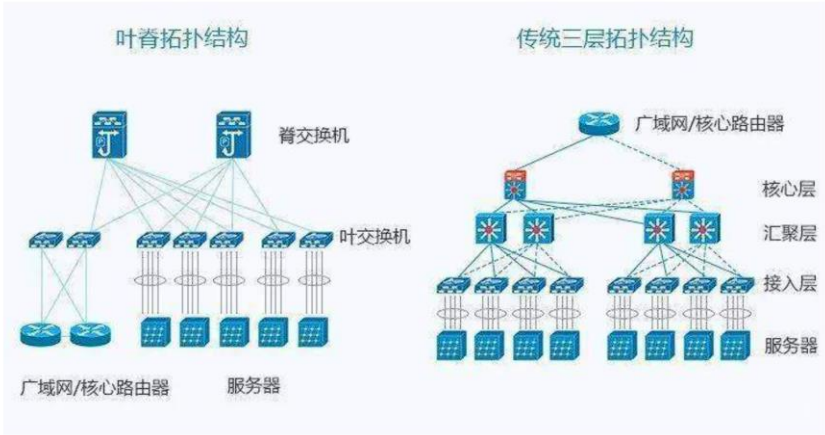
1.3.1 PCB——交换机组网升级，PCB核心受益

组网架构向5层迈进，利好交换机出货。AI大模型参数量及推理需求持续增长，带动万卡乃至十万卡集群发展，Scale out 推动组网架构从 2 层向 3 层、4 层架构拓展，当前海外推理侧已开始使用4-5层组网架构，以避免信号通道阻塞和降低延迟，相比原有2、3层架构交换机数量增多，且对交换机传输速率要求较高，带来高速交换机海量需求。

2025年6月3日，博通宣布其Tomahawk 6交换芯片开始出货，该产品为全球首款102.4T交换芯片，最多可处理64个1.6T端口，正式宣告交换机进入1.6T时代。**200G-400G-800G-1.6T交换机每次代际升级，PCB板的层数及加工难度均会显著增加，如800G交换机PCB已达到接近40层，售价也相应提高；随着AI算力需求催生高速交换机加速迭代，沪电股份等交换机PCB核心供应商将显著受益。**

图：Marvell 400G交换机PCB

图：交换机叶脊拓扑结构及传统三层拓扑结构

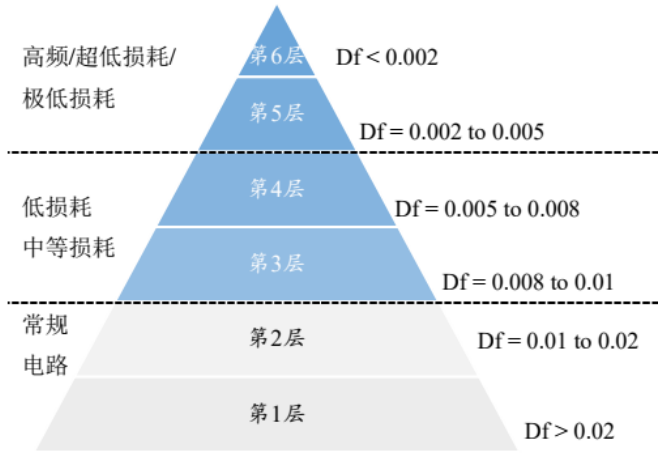


1.3.1 PCB——高速覆铜板壁垒较高，关注内资厂商突破

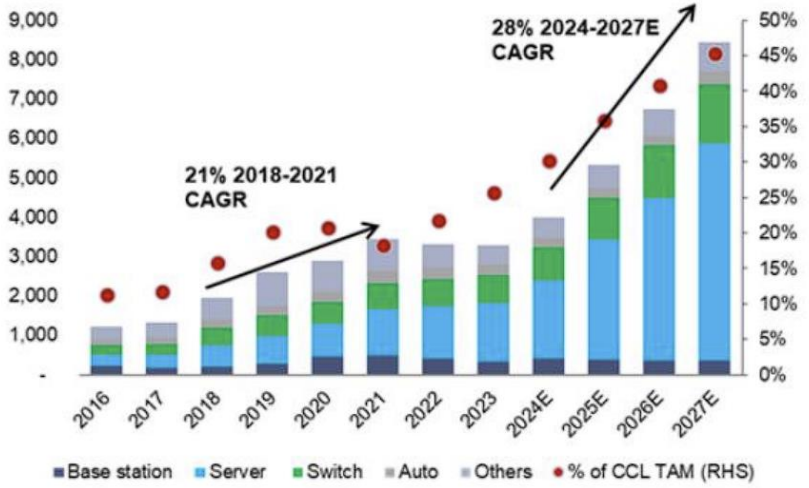
覆铜板为PCB制造中的基板材料，PCB的性能、品质、制造中的加工性、制造水平、制造成本以及长期可靠性等很大程度上取决于覆铜板。高频高速环境下，高频信号本身的衰减很严重，另一方面其在介质中的传输会受到覆铜板本身特性的影响和限制，进而造成信号失真甚至丧失。因此高频高速应用领域对于覆铜板电性能的要求非常高。

算力芯片及高速交换机对信号传输提出更高要求，高速覆铜板的需求快速增长，Ultra low loss（M7）及Super Ultra low loss（M8）等级高速覆铜板得以广泛应用，根据台光预计，高阶覆铜板市场2024-2027年复合增长率达28%。高速覆铜板作为算力相关PCB上游材料，技术难度大，壁垒较高，长期被台光、松下、斗山得外资厂商垄断，**生益科技正积极同国内外各大终端就GPU和AI展开相关项目开发合作，并已有产品在批量供应，有望成为高速覆铜板领导者。**

图：覆铜板电性能等级



图：服务器拉动覆铜板需求增长



资料来源：南亚新材招股说明书，台光材料官网，民生证券研究院

1.3.1 PCB——算力PCB相关厂商逻辑梳理

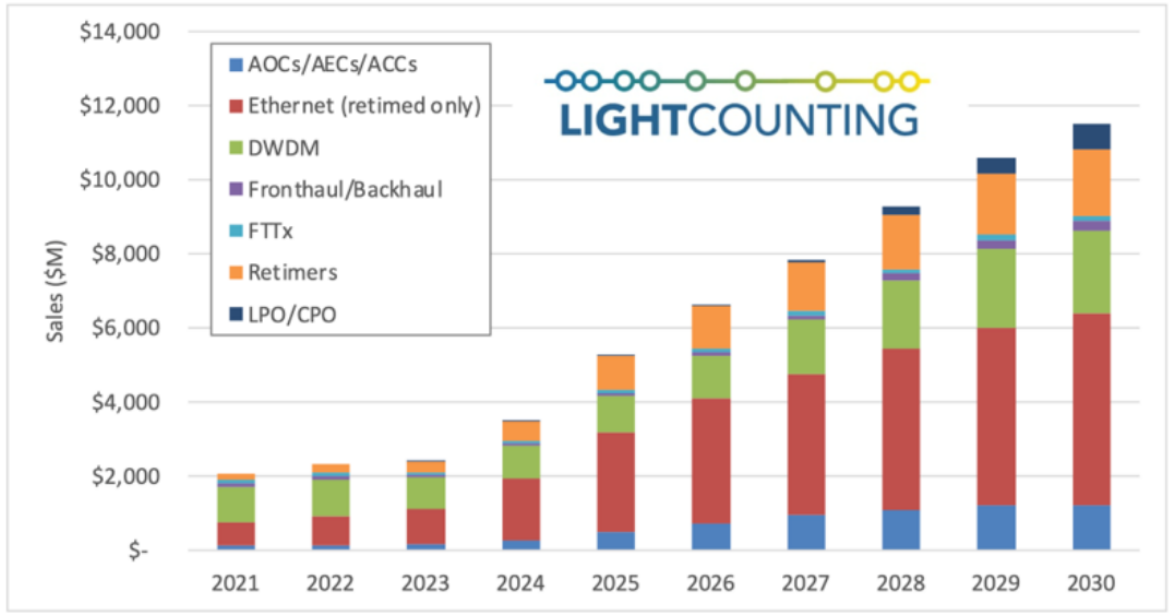
英伟达	ASIC	交换机	覆铜板
胜宏科技	**生益电子**	**沪电股份**	**生益科技**
沪电股份	**沪电股份**	深南电路	南亚新材
景旺电子	广合科技	方正科技	
方正科技			

资料来源：新浪财经，公司年报，公司官网，ifind，INBM财经视界等，民生证券研究院；
注：本页仅为逻辑梳理，不代表公司已具备相关客户

1.3.2 速率——CPO、AEC等方案伴随AI快速成长

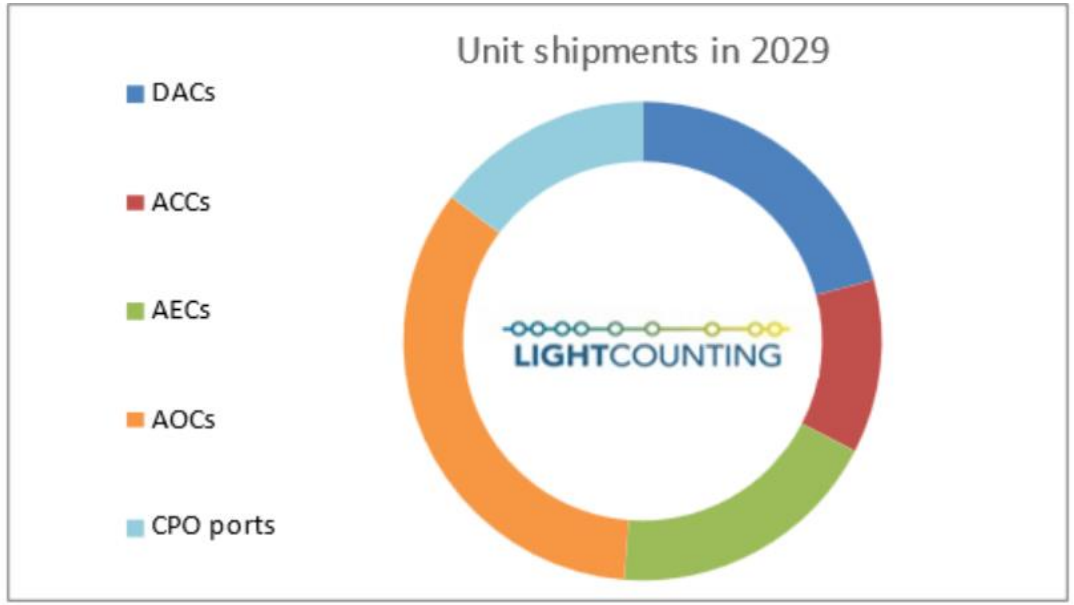
- **伴随AI需求的快速增长，CPO、AEC等技术方案成为互联方案中成长最快的环节。**据LightCounting，当前互联市场的主流技术路线仍然是以太网和DWDM，而伴随AI需求的快速增长，铜互联、LPO/CPO成为互联方案中增长最快的环节。
- **考虑光通信在低功耗、低成本、高可靠性的需求，CPO成为光通信升级的主流路线。**LightCounting表示CPO的部署将会很快开始，预计到2029年，CPO将成为1.6T互联方案的主流选择之一，2029年3.2T CPO端口的出货量将超过1000万个。而AEC作为铜互联中性能更优的解决方案，渗透率同样有望快速提升。

图：2021-2030年铜缆、以太网、CPO等互联市场空间预测



资料来源：LightCounting，民生证券研究院

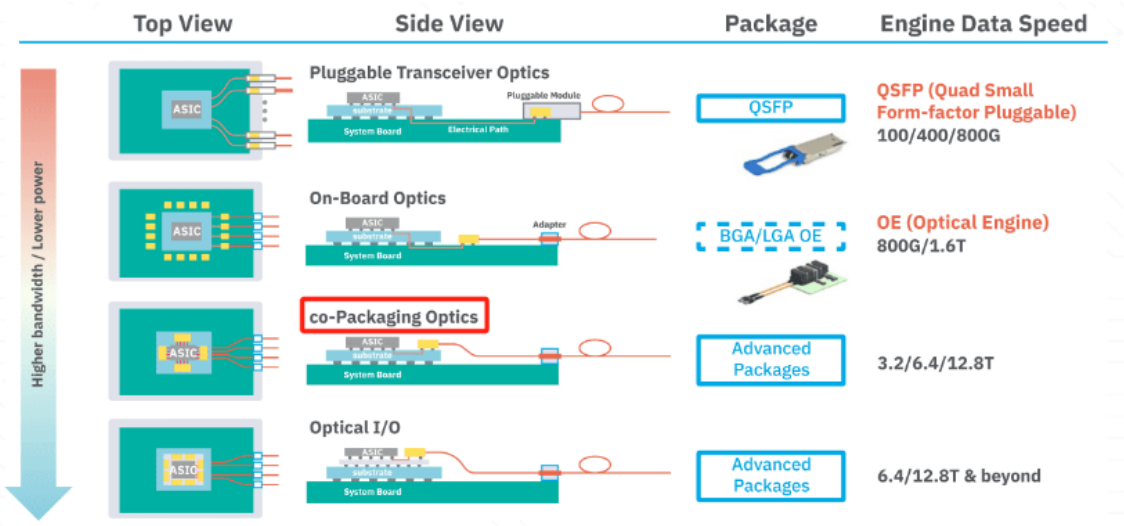
图：2029年1.6T铜缆和1.6T CPO出货量预测



1.3.2 CPO——光通信的终局方案

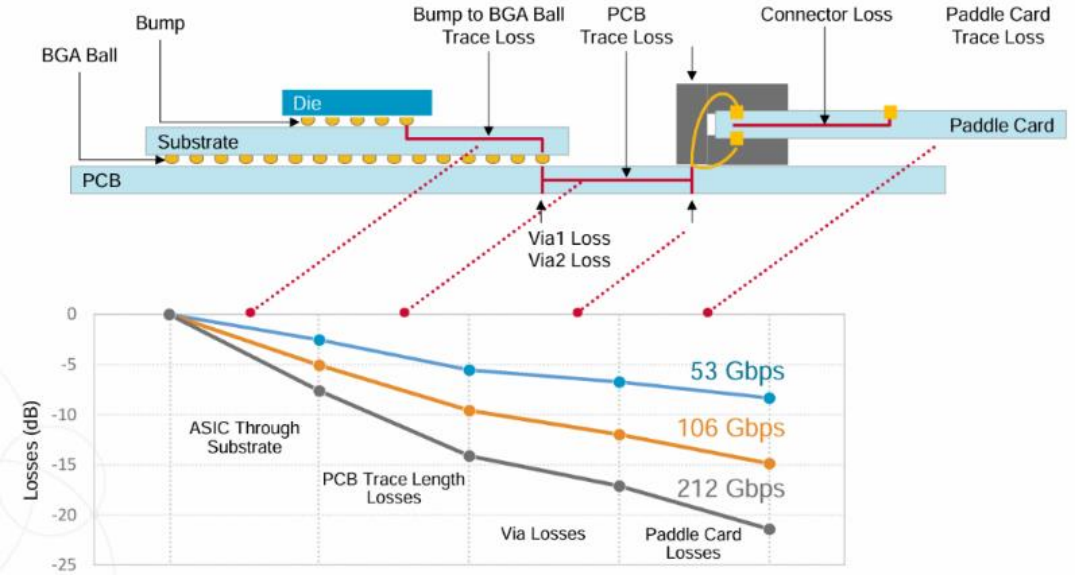
- **CPO指把光引擎和交换芯片共同封装在一起的光电共封装，没有采用可插拔光模块的形式，响应数据中心光模块降耗趋势。**在数据传输路径上，基板线路、PCB线路、通孔、光模块板卡都会产生一定损耗，且传输速率越大损耗越大。算力需求提升带动网络台积电在2024年北美技术研讨会上发布硅光路线图带宽成倍增加，数据中心能耗呈指数型增长，如何解决功耗问题成为下一代高速光互联应用的最大挑战。
- 在光模块降耗的发展趋势下，行业围绕驱动器、调制器、激光器及电接口4个方面降低功耗。在电接口方面，共封装光学（CPO）可以缩短交换芯片和光引擎之间的连接长度，实现更高密度的高速端口，提升整机的带宽密度。CPO背后的技术是硅光技术，即以光子和电子为信息载体的硅基光电子大规模集成技术。

图：CPO指把光引擎和交换芯片共同封装在一起的光电共封装



资料来源：ASE官网，博通官网，民生证券研究院

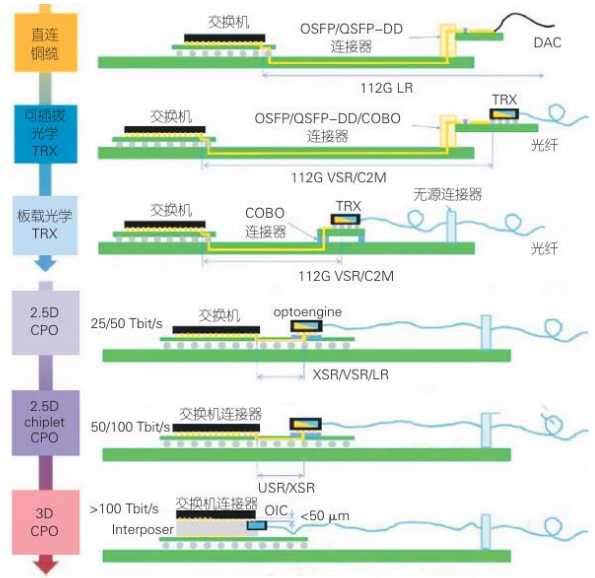
图：基板线路、PCB线路、通孔等产生一定损耗，且传输速率越大损耗越大



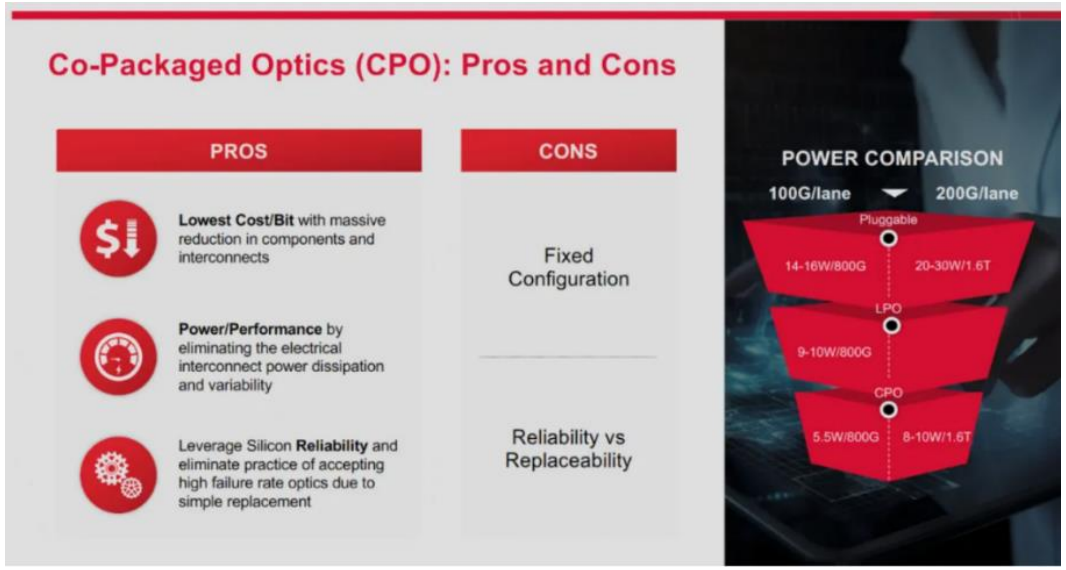
1.3.2 CPO——光通信的终局方案

- **按照物理结构，CPO可分为3种技术形态——1）2D平面CPO：**将光子集成电路PIC和集成电路并排放置在基板或PCB上，通过引线或基板布线实现互连；**2）2.5D CPO：**2.5D封装将EIC和PIC均倒装在中介层（Interposer）上；**3）3D CPO：**3D封装技术将光电芯片进行垂直互连。
- **CPO通过减少组件和互连器件数量降低成本，节省30%功耗。**CPO的主要优点之一是能够最大限度地减少所需的组件和互连器件数量，从而大幅降低单位比特成本。此外，与可插拔光模块器件相比，CPO通过消除电互连功率耗散和变异性，提供了更优越的功率和性能特性。博通官网显示，CPO能节省30%功耗，将每比特光学成本降低40%，可支持1Tbps/mm带宽密度。

图：CPO技术路线，2D平面CPO、2.5DCPO和3DCPO



图：CPO能够减少所需的组件和互连器件数量，大幅降低单位比特成本



资料来源：张平化等《数据中心光模块技术及演进》，博通，逍遥科技，民生证券研究院

1.3.2 CPO——行业龙头纷纷布局CPO

- **目前全球龙头均已在布局CPO产品，2025年有望成为CPO落地元年。**2020年以来，CPO逐渐从学术研究成果转变为市场需求产品。博通、Cisco、Marvell等行业内龙头企业均已推出多款基于CPO样品，其中英伟达计划在2025年通过CPO技术实现GPU芯片与NVSwitch芯片之间的光连接，而其他企业也在积极地布局相关产品。
- **国内CPO企业同样在加速追赶全球龙头的脚步。**中际旭创表示CPO是光模块未来的重要演进形式，公司正在投入力量对CPO进行研发；太辰光的光柔性板、超小型连接器及多芯保偏产品已可应用于CPO解决方案。

图：全球主流企业的CPO技术研究进展

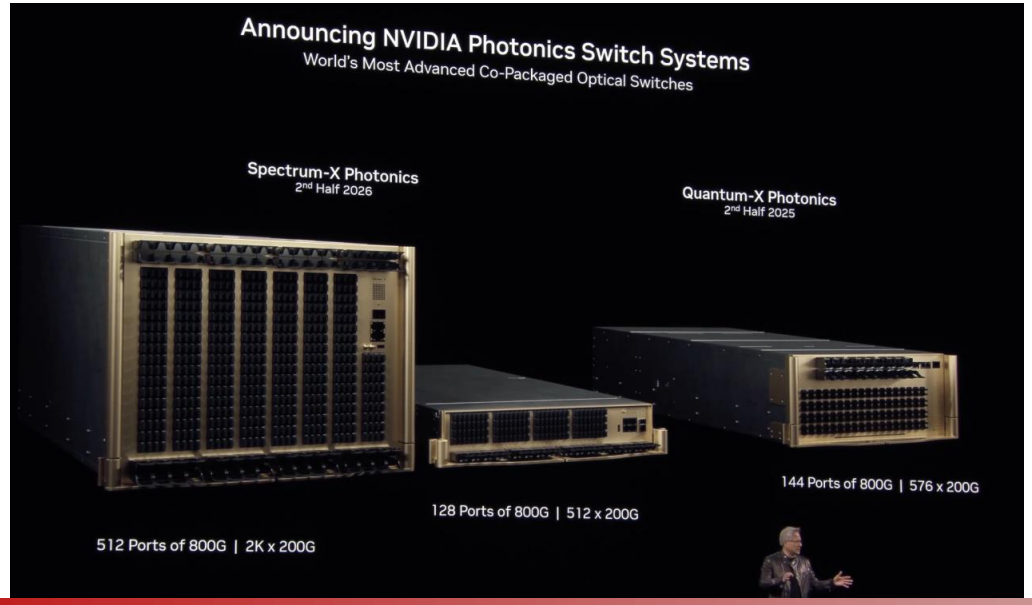
公司	进展情况
英伟达	积极推进芯片光互连策略，通过战略投资AyarLabs、TSMC构建硅光平台、芯片共封装能力，计划2025年左右通过CPO技术实现GPU芯片与NVSwich芯片之间的光连接。
台积电	在2024年北美技术研讨会上发布硅光路线图：2025年，第一代3D光学引擎(COUPE)将集成到运行速度为1.6Tbps的OSFP可插拔设备中。2026年，第二代COUPE旨在作为与交换机共同封装的光学器件集成到CoWoS封装中，功耗降低50%以上，延迟降低90%以上。第三代COUPE将传输速率提高到12.8Tbps，同时使光学连接更接近处理器本身，功耗降低90%以上，延迟降低95%。
博通	2023年10月联合MicasNetworks（云母网络）在美国OCP全球峰会上展示了基于博通的BailCPO首款51.2T(CPO)交换机。该产品的每个光引擎为3.2T(2x8PCS)，集成CMOS电芯片和MUX/DMUX，整体相对领先。
Marvell	2022年推出基于2.5DCP0技术的12.8TbitsTeralynx7交换机，2023年推出由超低延迟MarvellTeralynx1051.2Tbit/s交换芯片和川界首款PAM41.6Tbit/s光电平台MarvellNova组成的新平台。

资料来源：张平化等《数据中心光模块技术及演进》，半导体产业洞察，民生证券研究院

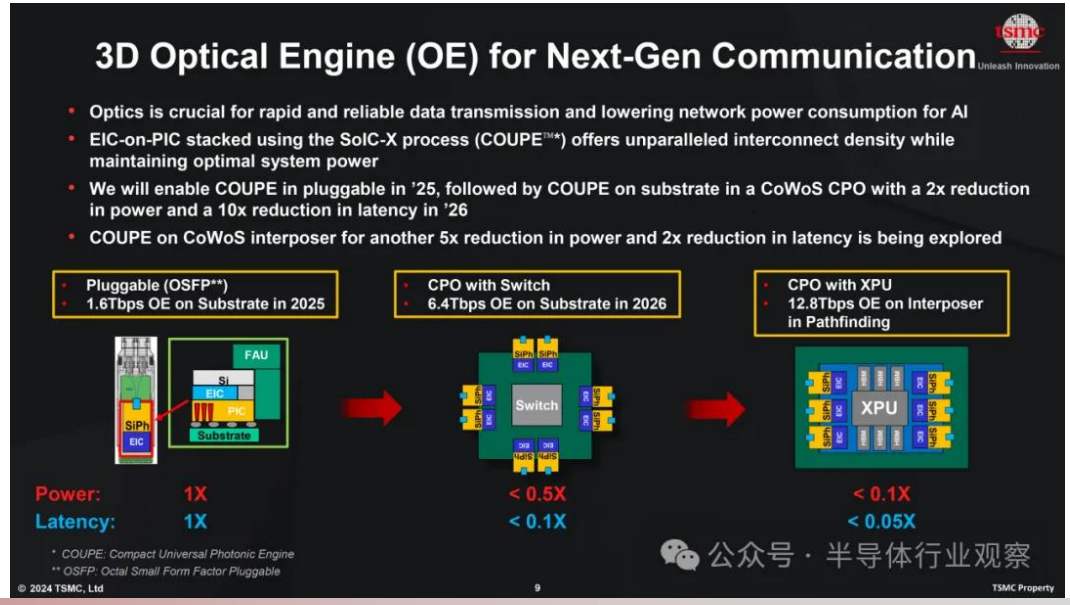
1.3.2 CPO——行业龙头纷纷布局CPO

- **英伟达并于2025年3月GTC大会推出CPO交换机新品。**英伟达的首款共封装光学(CPO)解决方案将部署在其横向扩展交换机中，可将能效提高3.5倍，网络可靠性提高10倍，部署时间缩短1.3倍。借助CPO，收发器现在被外部激光源(ELSs)取代，这些激光源与直接放置在芯片硅片旁边的光学引擎(OEs)一起促进数据通信。光纤电缆现在不再插入收发器端口，而是插入交换机上的端口，将信号直接路由到光学引擎。
- **台积电在2024年北美技术研讨会上发布硅光路线图，并在2025年表示CPO有望在1到1.5年内量产。**2025年，第一代3D光学引擎(COUPE)将集成到运行速度为1.6Tbps的OSFP可插拔设备中。2026年，第二代COUPE旨在作为与交换机共同封装的光学器件集成到CoWoS封装中，功耗降低50%以上，延迟降低90%以上。第三代COUPE（在CoWoS中介层上运行的COUPE）预计将进一步改进，将传输速率提高到12.8Tbps，同时使光学连接更接近处理器本身，功耗降低90%以上，延迟降低95%。

图：英伟达于2025GTC大会推出CPO交换机



图：台积电在2024年北美技术研讨会上发布硅光路线图



资料来源：英伟达官网，半导体行业洞察，民生证券研究院

1.3.3 AEC——性能优势更为显著的高速铜缆

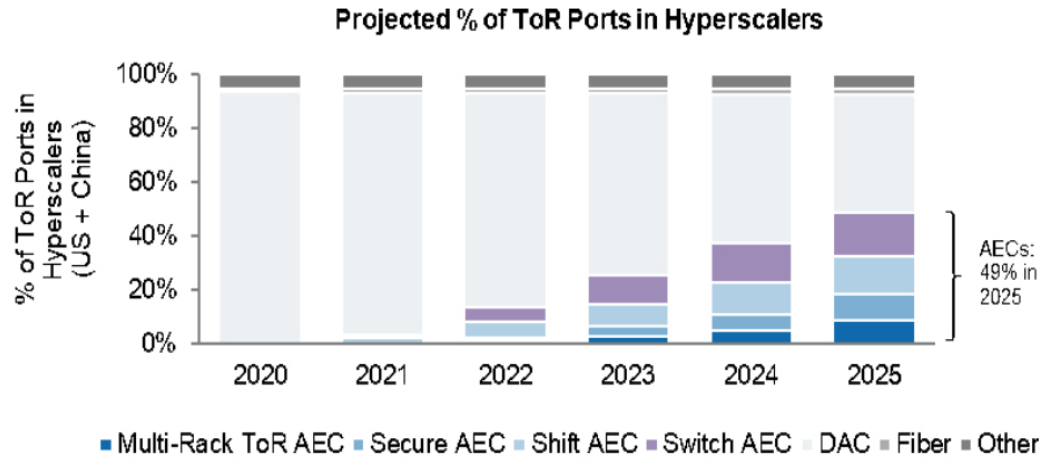
- **机柜铜互联方案始于英伟达在2024年3月推出的GB200机柜，该方案采用DAC作为机柜内互联的媒介。**高速铜缆方案主要包括DAC（无源铜缆）、ACC（有源铜缆）、AEC（有源电缆），其中DAC具有很高的成本效益，但传输距离随传输速度和带宽增加受限，而AEC在两头引入了具有时钟数据恢复功能的retimer芯片，对电信号重新定时和驱动，补偿铜缆的损耗，在7米以内的传输距离下相较于AEC具有更好的性能表现。
- **更长的传输距离，更低的损耗以及更优的传输性能使得AEC市场快速增长。**据650 Group，美国和中国超大规模企业采用的NIC ToR连接方案中，AEC的占比将从2021年的3%增长到2025年的49%。

图：DAC、AEC性能对比

	DAC	AEC
400G传输距离	< 3m	< 7m
800G传输距离	< 2m	< 2.5m
功耗	低	低
费用	低	中等
传输速度	快	快

资料来源：Asterfuison，650 Group，Credo招股书，民生证券研究院

图：2021年-2025年AEC在NIC ToR中的比重从3%提升至49%



Source: 650 Group

1.3.3 AEC——亚马逊Trainium2机柜已采用

- **亚马逊Trainium2机柜已经采用AEC的方案。**Trainium2包含16卡和64两种机柜方案，PCB负责同一个Compute Tray内部的2颗加速卡之间的互联；AEC负责其他加速卡之间的互联，以及机柜和机柜间的互联。
- **1）16卡服务器：**每个Trainium2服务器占用18个机架单元（18U），由1个2U 的CPU Head Tray和8个与之相连的2U Trainum2 Compute Tray组成。每个Compute Tray搭载两个Trainium2 芯片，两个CPU单独搭载至一个Head Tray中，通过外部PCIe 5.0 x16 DAC 无源铜缆与8个Compute Tray相连。每台Trainium2服务器整合了16颗Trainium2芯片，具备20.8 PFLOPS的算力，适合数十亿参数大语言模型的训练与部署；
- **2）64卡服务器：**Trainium2 Ultra机架方案集成 64颗Trainium2芯片，将4台Trn2服务器连接成1台巨型服务器，可提供相比当前EC2 AI服务器多达5倍的算力和10倍的内存，FP8峰值算力可达83.2PFLOPS，能够支撑万亿参数AI模型的实时推理性能。

图：亚马逊Trainium2 16卡Rack

图：亚马逊Trainium2-Ultra 64卡Rack

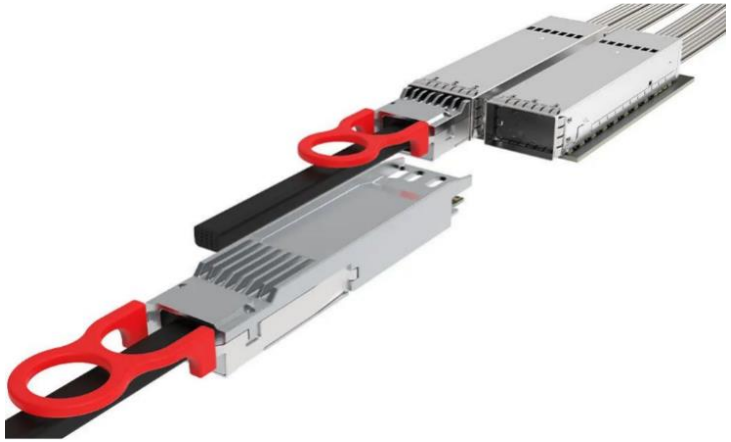
资料来源：AWS，SemiAnalysis，民生证券研究院

1.3.3 AEC——海外龙头厂商进展

- **Credo定义了AEC产品，是目前全球市场的龙头厂商，海外Molex、安费诺等也在加速追赶。**2024年10月，Credo发布HiWire AEC新品——线缆长度为7米的800G ZeroFlap（ZF）系列。此系列高性能HiWire ZF AEC旨在为人工智能（AI）后端网络提供高度可靠的互连解决方案。
- **Molex**发布的业界首个芯片到芯片之间互连的224G产品组合中，OSFP 1600解决方案包括无源电缆（DAC）和有源电缆（AEC）产品，可满足每个高速通道传输224 Gbps-PAM4信号或汇聚到每个IO连接器实现1.6T的带宽。而**安费诺**也在2023年7月发布了ExaMAX2® Gen2产品，布局AEC市场。

图：Credo 800G ZeroFlap（ZF）系列AEC新品

图：Molex的OSFP 1600解决方案



图：安费诺发布ExaMAX2® Gen2

资料来源：Credo官网，安费诺官网，讯石光通讯网，民生证券研究院

1.3.3 AEC——国内厂商加速追赶

- **AEC市场空间快速增长下，国内厂商也在配合海外龙头芯片公司，加速在AEC领域产品的追赶进度。**其中新易盛、瑞可达、博创科技均已推出400G、800G及更高速率的AEC产品，在国内外厂商处积极导入中。兆龙互联配套海外龙头厂商Credo，凭借线缆产品的优势打入AEC供应链，而沃尔核材子公司乐庭智联为国内无源铜缆的龙头供应商，在AEC快速渗透的浪潮中，公司同样有望受益。

图：国内AEC厂商产品进展

公司	进展情况
新易盛	新易盛表示公司在AEC技术领域已深入布局，与客户建立起了良好的沟通与合作关系。
瑞可达	瑞可达是能同时提供光、电、微波、高速数据、流体连接的综合解决方案的优质供应商，公司已逐步开发了应用于 AI 与数据中心领域的 SFP+、CAGE 系列，高速板对板连接器、高速 I/O 连接器，AEC 系列产品，目前相关项目正在推进中。
博创科技	博创科技在2024年大会上推出高性能400G/800G TRX/AOC/AEC系列产品，支持高达400G/800G的传输速率，覆盖OSFP、QSFP-DD和QSFP112等多种封装形式。800G AEC产品内置DSP Retimer芯片，支持互通DAC的Half Active应用，为用户提供更灵活的网络配置和管理选项。
兆龙互连	兆龙互连在2024年智能制造产业生态合作大会上推出LongTronic数据通信传输系统和LongFlex运动控制连接系统、预端接光纤布线方案产品及100G-800G AEC/ACC/DAC高速直连布线系统产品。兆龙提供的10G/25G/40G/100G/200G/400G/800Gbps的AEC/ACC/DAC高速无源/有源连接解决方案，高速率且低功耗，支持当前以太网各种速率协议应用。
沃尔核材	沃尔核材的AEC产品主要由子公司乐庭智联生产，类型包括SAS系列、PCIe系列和QSFP系列；主要客户包括美国安费诺集团（Amphenol Corporation），英国豪利士（Volex），美国莫仕（Molex），爱尔兰泰科（TE Connectivity）等国际线缆连接器龙头。

资料来源：各公司官网，民生证券研究院

1.4

目录

CONTENTS

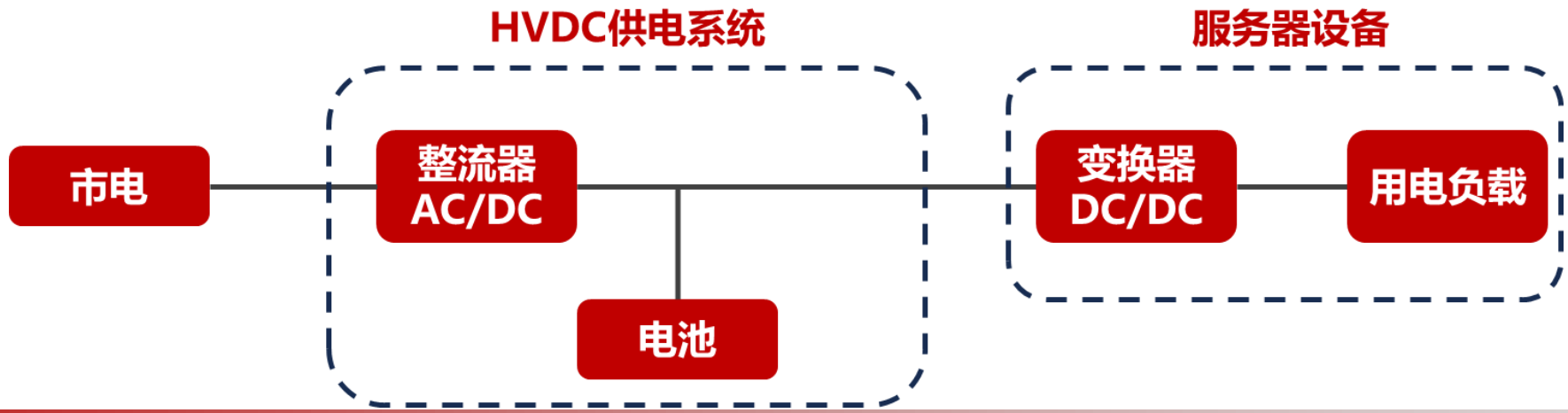
1.4.1 电源—— HVDC优势明显，数据中心供电新趋势

- **UPS即不间断电源，是传统数据中心的供电方式。**市电输入UPS后，UPS需要首先进行ACDC转换，用于给蓄电池直流充电，然后进行DCDC，把电压等级转换到合适后续设备工作的电压，再进行DCAC逆变到230V或400V交流电，用于后续的交流设备供电。

- **而HVDC则是高压直流输电，是一种新型的直流不间断供电系统。**过去主要长距离、大容量的电力传输。高压直流可以大幅减少输电损耗。并支持更高电压+功率等级。所以目前已有部分AIDC机房开始采用HVDC方案作为第一级供电。如阿里巴巴张北云计算庙滩数据中心。

- **目前数据中心呈现向HVDC（高压直流输电）升级的趋势。**国内方面，最早由中国电信推进使用HVDC，目前已成功扩展至多个领域，主要应用有阿里巴巴的巴拿马电源、腾讯的第三代数据中心T-MDC和第四代数据中心T-block等；国外方面，Intel、微软、Meta等公司也已采用400V直流供电系统。相较于UPS，HVDC取消了逆变器环节，供电线路更加简单。当市电正常时，市电经由整流器转换为直流电后直接向服务器设备供电，同时为电池充电；当市电发生故障时，由电池直接向服务器设备供电。

HVDC全链路供电架构

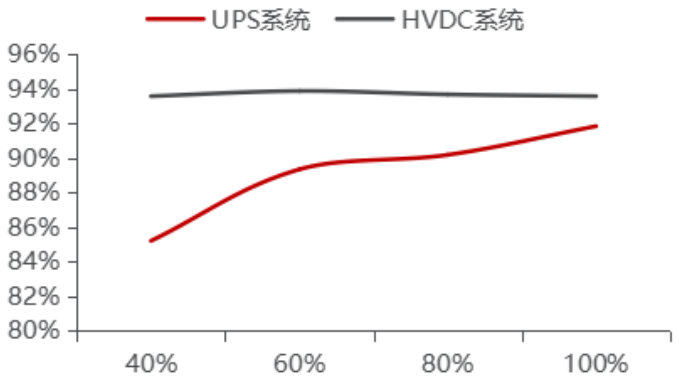


资料来源：中华人民共和国工业和信息部，中国储能网，奥默生科工程有限公司官网，环球网，搜狐，民生证券研究院

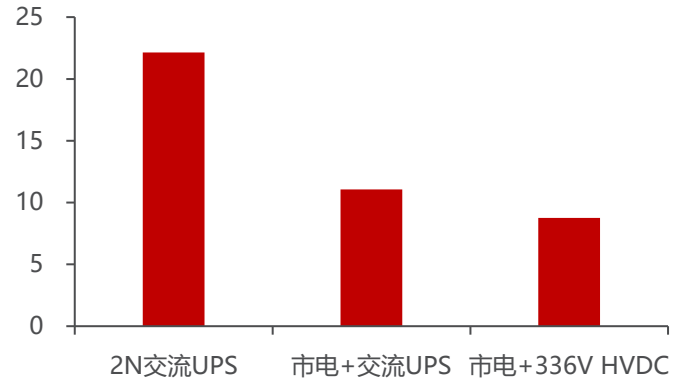
1.4.1 电源—— HVDC优势明显，数据中心供电新趋势

- **HVDC系统在多个方面表现优于UPS系统。**随着数据中心的发展，HVDC系统在效率、扩展性、可靠性等方面的优势逐渐凸显：
- **1）高效节能：**HVDC系统中取消了逆变环节，导致功率器件减少，能量损耗降低，使得用电效率得以提高，在低负载的情况下优势更加明显。
- **2）稳定可靠：**HVDC的电池可直接向设备进行供电，相比于UPS而言无需再经过逆变器环节，减少了故障点，同时由于线路运行更加简单，电池供电时只需进行降压操作，无需再考虑频率和相位同步问题，使得可靠性进一步提高。
- **3）灵活扩展：**HVDC的模块化设计使得后期扩容便利性上升，用电方式更加灵活，而UPS即使采用模块化设计也需要考虑并机问题，扩展和维护难度相对较高。
- **4）节约空间：**HVDC中功率器件减少，在提供相同功率的前提下相较于UPS可节省25%的占地面积，为服务器提供了更多宝贵空间。
- **5）降低成本：**从建设成本考虑，HVDC减少了逆变器和STS（静态旁路转换开关），使得建设成本降低25%-50%；从使用成本考虑，得益于效率提升，以市电+336V HVDC方案为例，每年系统损耗相比于市电+交流UPS方案可减少2.31万度电。

不同负载率下UPS与HVDC效率对比



不同电源系统方案每年电量损耗（万度）

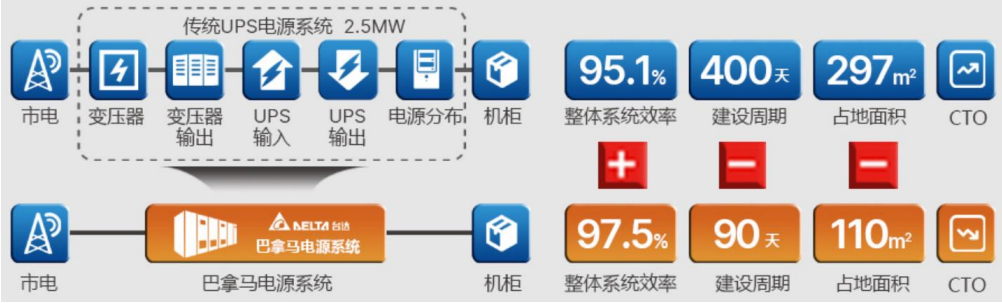


资料来源：华为官网，中国IDC圈，《数据中心的节能研究与实践》_胡鹏涛，《数据中心UPS系统与HVDC系统的效率比较》_高凯亮，民生证券研究院

1.4.1 电源—— HVDC优势明显，数据中心供电新趋势

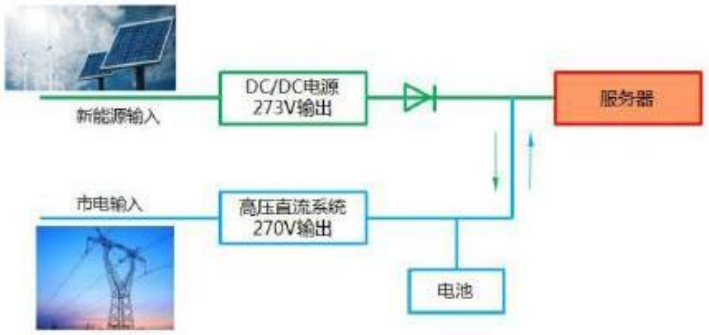
- **系统集成度更高，提升空间利用率：新一代数据中心电源将集成更多设备功能。**阿里巴巴的巴拿马电源彻底革新了传统IDC的供电架构，相较于传统UPS而言，巴拿马电源的设备数量和安装工程量减少了40%，安装空间减少了50%，总建置成本减少了40%-60%。
- **HVDC电压进一步提升，支持高算力集群：新一代数据中心电源将输出更高电压。**百度2024年推出的“瀚海”直流电源配备了750V输出柜，支持单机柜100kw，效率提升2%-4%；台达最新推出的10kV供电系统DPSST系列最高输出电压更是达到1000V，整机效率达到98%以上。
- **清洁能源加持，推动数据中心节能提效：清洁能源正逐渐成为数据中心供电新方向。**目前国内外已有多数厂商为数据中心寻求清洁能源，如腾讯、微软、亚马逊等。

巴拿马电源系统与传统UPS架构对比

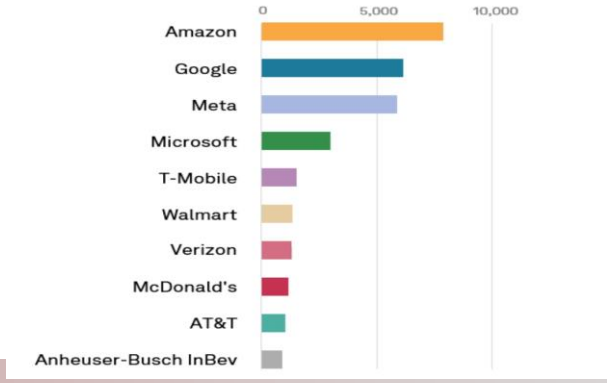


百度“瀚海”一体化直流电源

腾讯T-block“光伏+HVDC“供电系统



2022年美国企业可再生能力排名（MW）

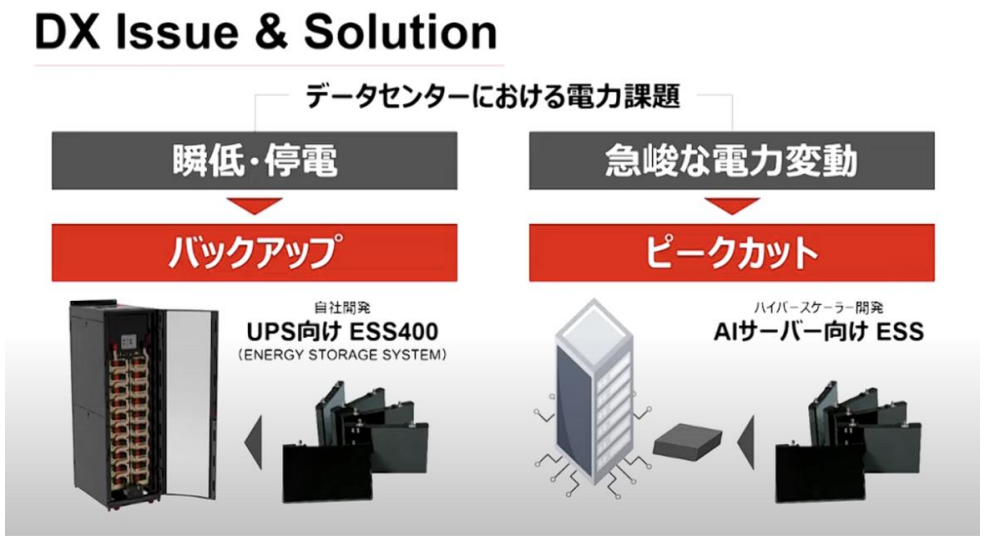


资料来源：台达官网，明阳电路官网，益企研究院，腾讯云，S&P全球，数据中心DATA，环球Tech，中国新能源网，智创狮，《利用 SiC 设计新一代数据中心冷却系统》_黄赞，民生证券研究院

1.4.2 超级电容—— GB300搭载，削峰填谷+紧急供电

- **与传统电解电容相比，超级电容具备以下有优点：** 1）高储能密度；2）充放电速度快；3）长寿命及高可靠性。**与蓄电池相比，**超级电容器更适合于短时间、高功率的能量储存和释放，而蓄电池更适合于长时间的能量储存和利用。**超容在数据中心中的作用主要为以下两点：**
- **紧急供电：**当数据中心发生停电时，**超级电容作为一级备电，会率先为服务器提供临时电源。因其响应速度极快，可在毫秒级时间内完成充电和放电。最大程度防止服务器因供电延迟而导致的数据丢失的情况。**而后则由不间断电源（UPS）的蓄电池以及柴油发电机接续供电。
- **平滑波动：**随着AI芯片功耗增大，在训练或推理的过程中功率波动会更加陡峭，对电路造成损害并影响芯片性能。**超级电容器的高功率密度快速响应特性可以较好应对电源波动，能够在电力负载突然变化时立即提供瞬时功率补偿，实现削峰填谷，平抑负载浪涌，确保电压的稳定及电路和芯片的正常运行，**不仅提供了高功率支持，而且优化了空间和能效。

图：超容的主要功能



表：锂离子混合超容与各种蓄电装置的特性比较

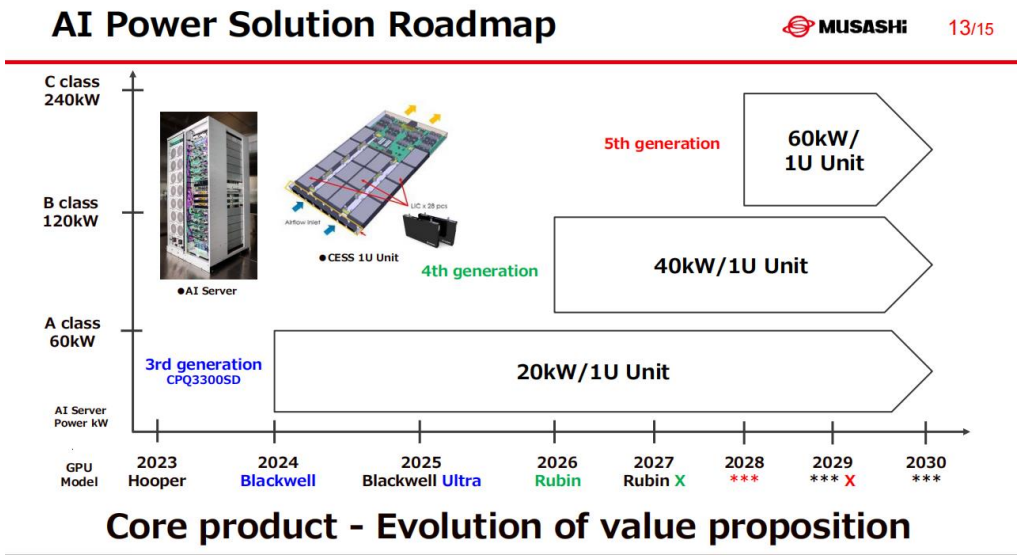
	锂离子混合超容	EDLC	锂电池	铅酸电池
能量密度	中 (在高电流时更高)	低	非常高	高
功率密度	高	高	低（不适合快速充电）	很低
快速充放电	以秒为单位	以秒为单位	以小时计（需要充电控制）	以小时计（需要充电刷新）
内部电阻	低	低	高	非常高
低温性能	好	好	非常差	差
高温性能	非常好（可达70℃）	好（可达60℃）	非常差（可达40℃）	非常差
自放电	小	大	小	大
维护	无需维护	无需维护	需要频繁替换	需要经常更换
使用寿命（浮动/循环）	长	长	相对较短	短（突然关机发生）
安全与易燃性	高，易燃	高，易燃	低，易燃（自热/点燃）	高，不易燃
应用程序	非常高的功率（中等能量）	非常高的功率（低能量）	中等功率（高能量）	低功率（高能量）

资料来源：武藏官网，汉斯期刊，观研报告网，元件网，民生证券研究院

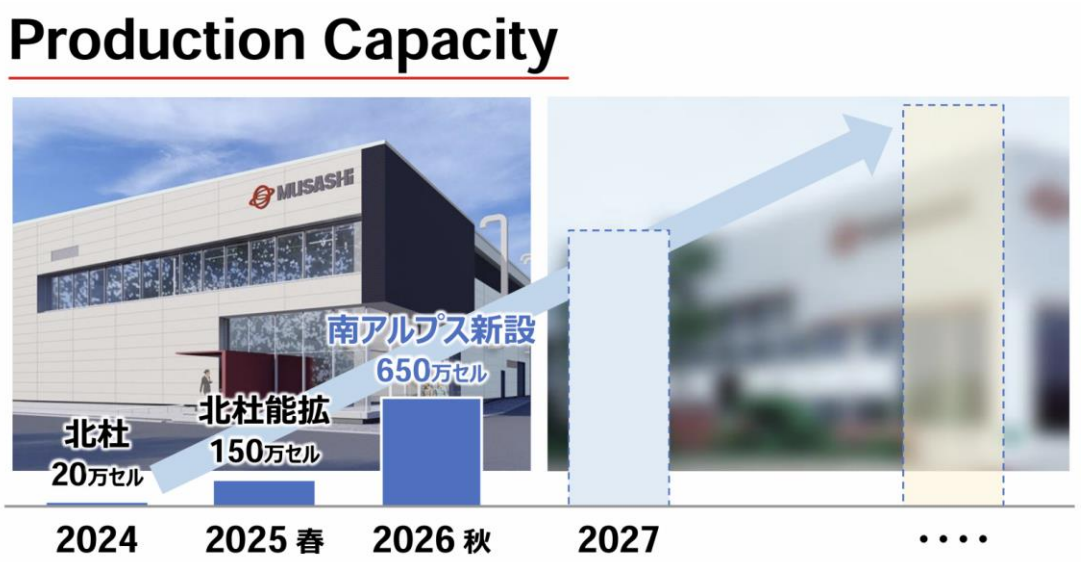
1.4.2 超级电容——详解武藏超容方案

- **超级电容成为GB300标配,当前主要供应商为日本武藏。**武藏与超容模组制造商Flex（伟创力）合作，伟创力计划在2025年上半年开始生产CESS，在生产开始后的一个季度内实现商业化。
- 武藏超级电容采用方形结构，每个超容模组包含28颗超级电容，单模组可提供20kw功率，GB300单机柜根据电力系统设计需求采用若干个模组；在武藏的产品规划中，英伟达下一代Rubin架构的超容模组单模组功率将提高至40kw，较Blackwell Ultra（GB300）翻倍。
- **武藏的扩产规划显示，AI对于超容的需求快速扩张。**当前武藏超容年产能仅20万颗；武藏在2025前半年扩建北杜工厂，并于1Q25将产量从20万颗电芯提高至150万颗；2026年9月预计开始运营山梨县新厂，产量进一步扩张至650万颗，并计划在27年实施更大的扩产计划。

图：武藏超容产品路线图



图：武藏扩产规划



资料来源：武藏官网，伟创力官网，中国工业新闻网，超级电容+，民生证券研究院

1.4.2 超级电容——卡位AI赛道，江海股份优势凸显

- **江海股份深耕超级电容器，已成为国内超容龙头。**经过多年积累，超容产品已形成多方面的竞争优势：
- **产业布局角度**，江海股份是全球在电力电子领域少数几家同时在铝电解、薄膜、超容三大类电容器进行研发、制造和销售的企业之一，各产品形成协同效应，共享客户资源。**底层技术路线角度**，江海锂离子超级电容技术源自其2013所接受的日本公司ACT的全部知识产权转让，武藏则是源自其所收购的GM Energy，二者均为锂离子超级电容，技术路线相近。2016年开始，江海股份利用8亿募投资金对原有超容技术及生产工艺进行改造，**当前公司已于相关生产商探讨技术方案，在方案上完全达到GB300的性能、功用要求，并具备产能大、扩产周期短、特别是成本低的优势；此外公司MLPC等产品也已适配AI数据中心，有望将AI打造成公司第二成长曲线。**

图：江海股份锂离子超容产品

Computex大会 维谛技术展示 800V HVDC

资料来源：深交所互动，公司公告，江海股份官网，ifind，新浪科技，民生证券研究院

1.4.2 液冷——芯片功耗提升，液冷成为刚需

- 传统数据中心以CPU云计算为主，芯片功耗较低，风冷即可满足大部分需求；
- **为什么当前液冷成为刚需?**
- **一：单芯片功耗（TDP）提高。**AI浪潮下，大功率的GPU搭载至AI服务器，**GPU功耗超过700W时，风冷一般便难以满足散热需求。**以NV为例，每一代芯片升级功耗要提高300/400W，A100及H100均采用风冷即可满足需求，**但B200（1000W）开始需要搭配液冷散热。**

- **对于传统CPU服务器，CPU功耗超过450W时，风冷一般便难以满足散热需求。**英特尔最新“至强6代”服务器CPU处理器功耗达到500w，**传统服务器芯片功耗也已达风冷极限**（既不具备经济性，又难以满足散热需求）。

图：历代英伟达GPU功耗提升

型号	发布时间	功耗（W）	芯片类型
A100	2020	**400**	推理+训练
H100	2022	**700**	推理+训练
B200	2024	**1000**	推理+训练
B300	2025	**1400**	推理+训练

图：历代英特尔CPU功耗提升

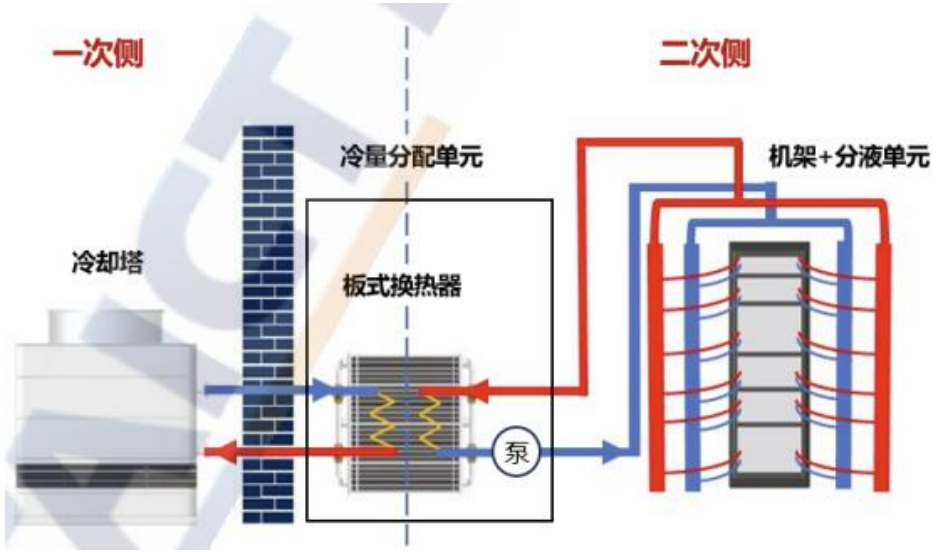
芯片型号	发布时间	最大功耗（W）
第一代至强可扩展服务器	2017	205
第二代至强可扩展服务器	2019	400
第三代至强可扩展处理器	2020	270
第四代至强可扩展处理器	2023	350
第五代至强可扩展处理器	2023	385
第六代至强性能核处理器	2024	500

资料来源：腾讯网，芯智讯，与非网，英伟达官网，CSDN，英特尔官网等，民生证券研究院

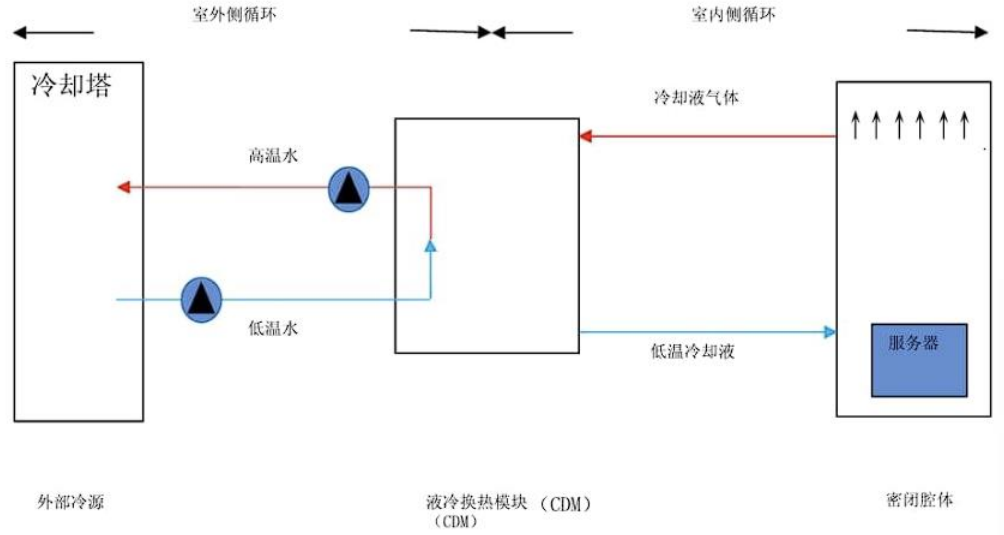
1.4.3 液冷——冷板式为主，浸没式尚未起量

- 传统数据中心中，由于以CPU云计算为主，芯片功耗较低，因此一般采取风冷散热。随着AI算力芯片功耗密度大幅提高，传统风冷散热技术已难以满足当前的高密度计算散热需求，诸如英伟达GB200 NVL72、谷歌TPU等需要采取散热效果更好的液冷辅助散热，**数据中心液冷时代已至。**
- 液冷技术主要**可分为非接触式液冷及接触式液冷两类。其中非接触式液冷主要指冷板式液冷**，冷板式液冷对于服务器整体改动较小，相较其他液冷方案，冷板式液冷在可靠性、可维护性、技术成熟度、适用性等方面具有优势，利于算力中心机房改造。**冷板式液冷成为当前绝对主导的液冷解决方案，占据国内市场95%以上份额，预计3年内冷板式仍将占据绝对主要地位**
- **接触式液冷的液体与发热源直接接触，包括浸没式液冷和喷淋式液冷两种。**相比非接触式液冷，接触式液冷可完全去除散热风扇，散热及节能效果更好，数据中心PUE值可降至1.1及以下，但相应对服务器机柜及机房配套设施的投入及改造成本更高，运维难度更大。**浸没式液冷随着未来技术标准化推进、应用部署成本降低，有望加速大规模商用进展。喷淋式液冷作为中间方案，客户兴趣不高。**

图：冷板式液冷系统示意图

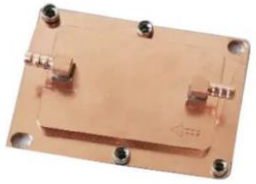


图：浸没式液冷原理示意图



1.4.2 液冷——主要供应链环节梳理

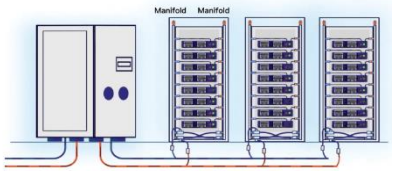
- 冷板式液冷主要可拆分一次侧系统及二次侧系统，一次侧系统主要由室外散热单元、一次侧水泵、定压补水装置和管路等部件构成。二次侧系统主要由换热冷板、热交换单元（CDU）和循环管路（Manifold）、快接头、冷源等部件构成。



液冷板：换热冷板常作为电子设备的底座或顶板，通过空气、水或其他冷却介质在通道中的强迫对流，带走服务器中的耗散热，从而有效降低算力中心的 PUE 值。
内资相关厂商：英维克、飞荣达、中航光电等

CDU ：CDU是将冷却液分配到各个需要散热设备的部件，可以看作室内机与室外机的连接点，具有流量分配、压力控制、防凝露等作用。
内资相关厂商：英维克、申菱环境、高澜股份、曙光数创等

快速接头：快速接头实现液体通路连接和断开，具有双向自密封功能，插合和断开过程中不会有液体泄漏，一般分为手动插拔和盲插拔。液冷快速接头一般由公头和母头组成。
内资相关厂商：英维克、溯联股份、川环科技、中航光电等



Manifold：循环管路是连接换热冷板、冷量分配单元和室外冷源的必要部件。按连接方式不同，可分为直连式（异程式）和环路式（同程式）两种。
内资相关厂商：英维克、溯联股份、川环科技、中航光电等

02 国产算力：算力平权，国产AI力量崛起

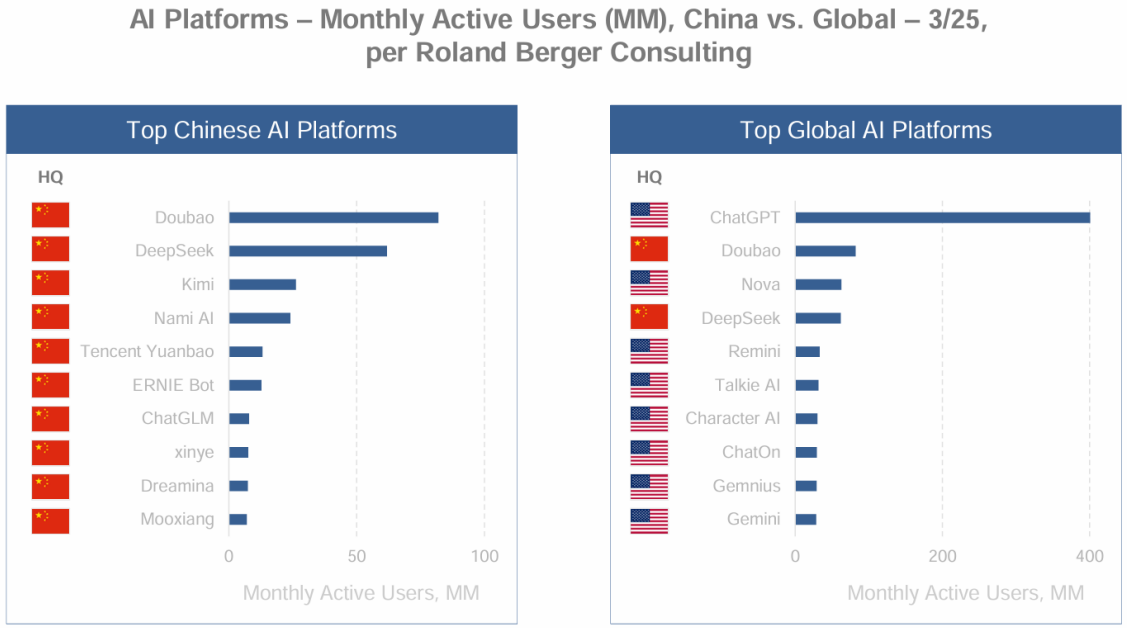
CONTENTS

目录

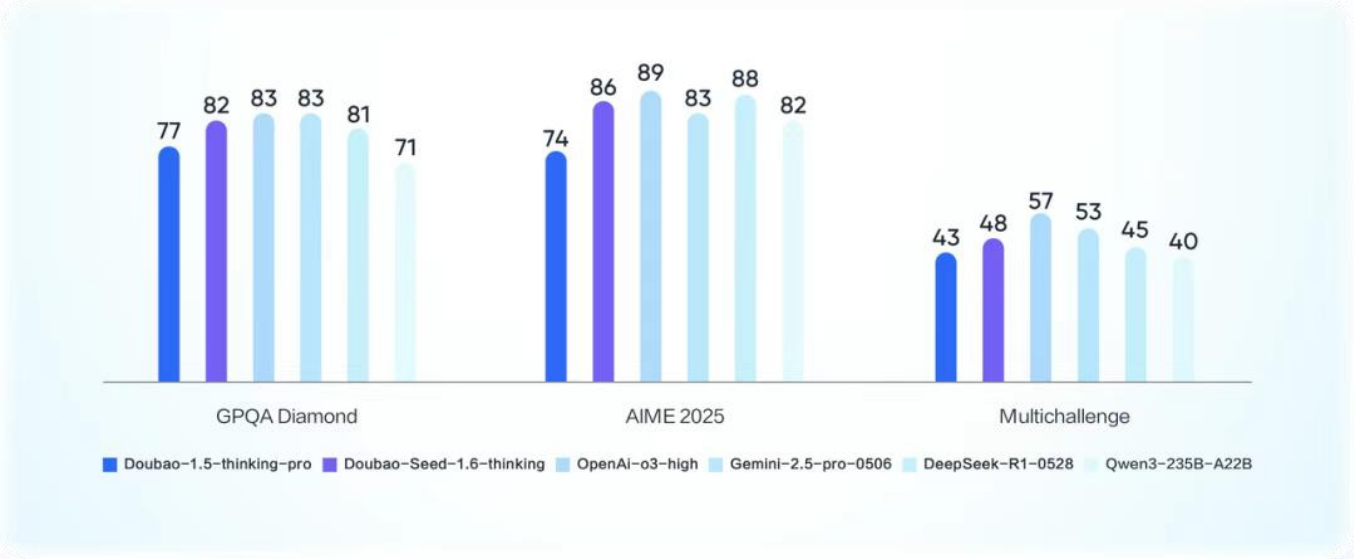
2.1.1 DeepSeek+豆包，国产”双子星“推动产业发展

- 近年来，随着模型能力持续提升和应用场景不断拓展，国产大模型正呈现蓬勃发展之势。其中，豆包与DeepSeek作为目前**国内最具代表性以及月活用户最多**的两个大模型，各自走出了差异化的发展路线。
- 作为国产大模型的“双子星”， **DeepSeek和豆包**构建起截然不同的生态模式。**DeepSeek以开源为战略核心**，采用了自主改进的DeepSeekMoE架构，支持128k上下文，在**数学、编程与通用逻辑**等测试表现上已接近其他国际顶尖模型，成为“国产开源力量”的重要代表。
- **豆包**则依托字节跳动强大的产品体系，聚焦**多模态能力**的实际应用，在最新发布的豆包1.6大模型中，不仅支持**256k超长上下文**，其深度思考强化版本的性能更是达到**国际一流水准**，与DeepSeek表现相当，而**综合成本仅为后者的1/3**。

AI平台每月活跃用户排名



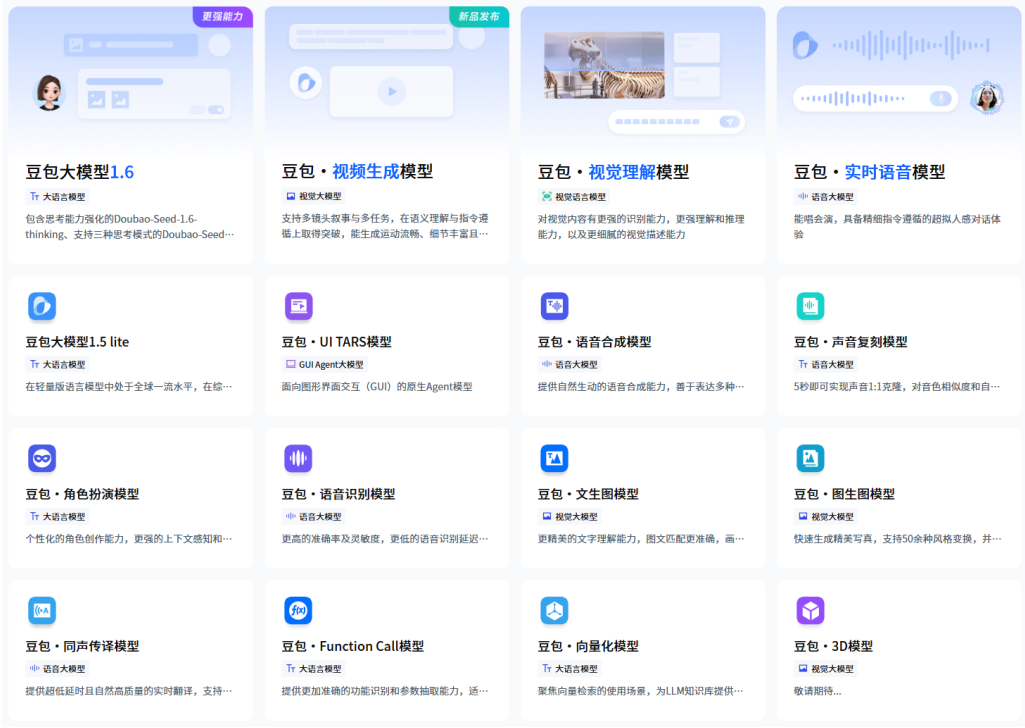
豆包和DeepSeek与其他大模型的测试对比



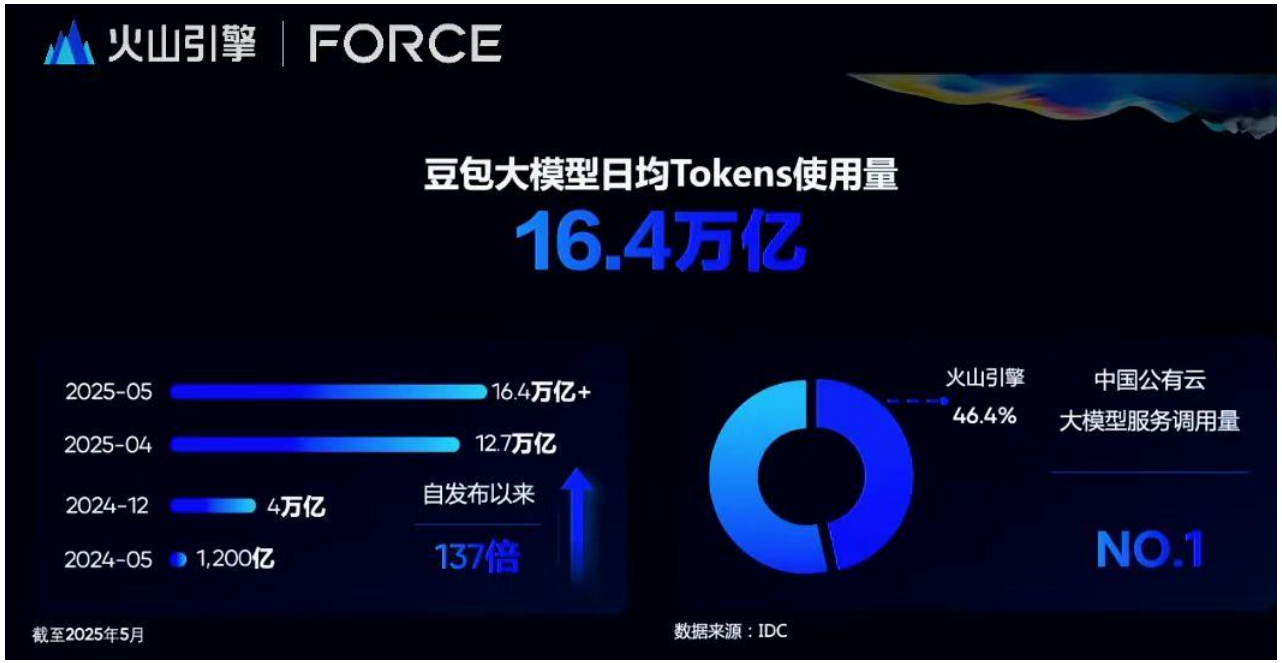
2.1.1 豆包多模态扩张，Tokens 用量飙升彰显应用热度

- 豆包大模型始于2023年8月17日字节的AI对话产品“豆包”公测，2024年，豆包逐步补全了**语音、图像、代码等能力**，并在12月18日发布豆包视觉理解模型，实现了更强的内容识别、理解和推理、以及视觉描述等能力，一举成为国内最领先的多模态大模型之一，引领国产化大模型在多模态领域的升级之路。
- 2025年6月11日，火山引擎正式发布豆包大模型1.6、豆包·视频生成模型 Seedance 1.0 pro、豆包·语音播客模型，豆包·实时语音模型在火山引擎全量上线，豆包大模型家族已成为拥有**全模态、全尺寸、高性价比**的领先模型。截至2025年5月底，豆包大模型日均 tokens 使用量超过16.4万亿，较去年5月刚发布时**增长137倍。**

豆包超全模态支持



豆包Tokens使用量数据



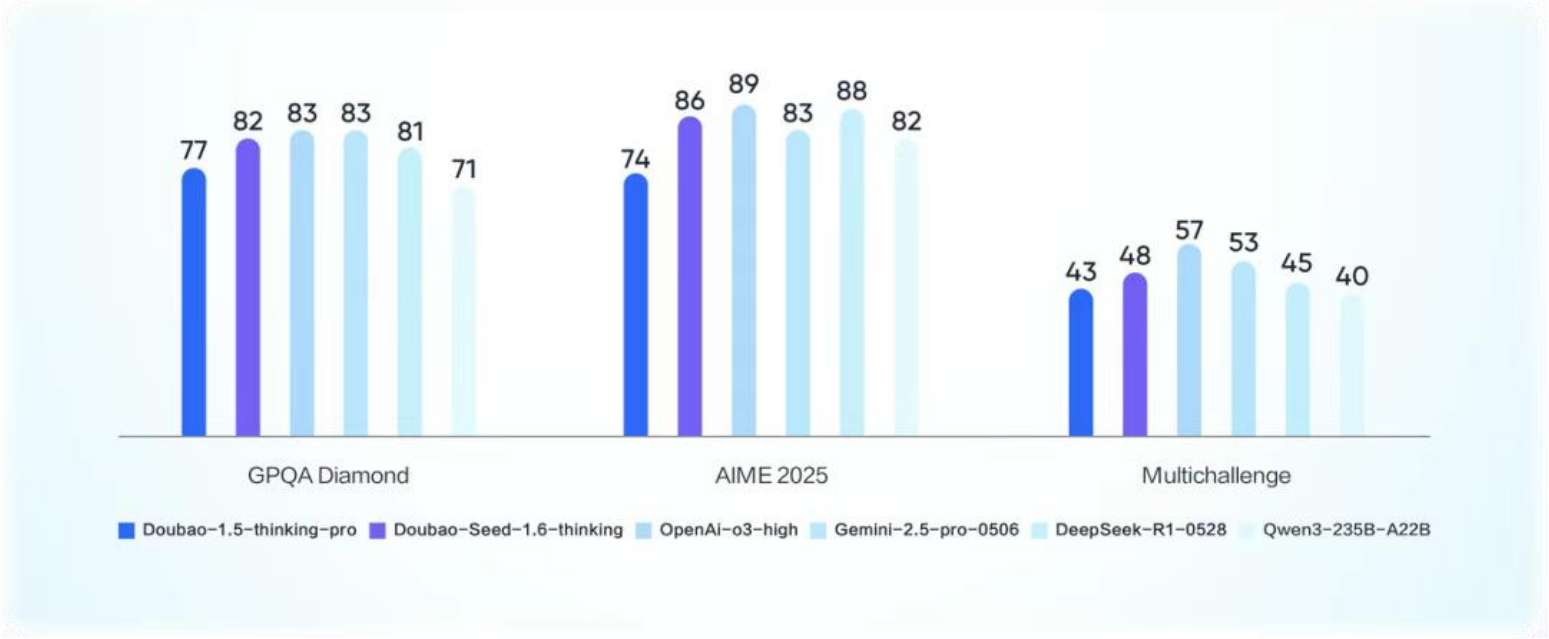
资料来源：火山引擎官网，民生证券研究院

资料来源：硬AI，民生证券研究院

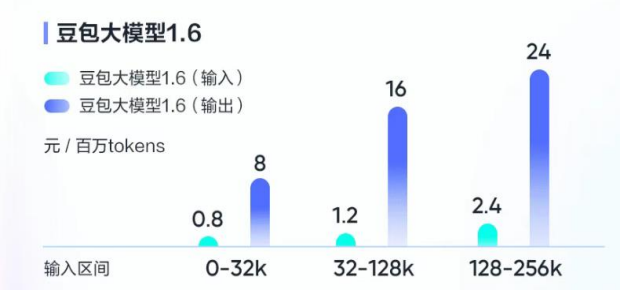
2.1.1 豆包大模型 1.6：深度思考表现亮眼，成本优势凸显

- 豆包大模型1.6系列包括：1）doubao-seed-1.6，All-in-One 的综合模型，是国内**首个支持256K 上下文**的思考模型，支持深度思考、多模态理解、图形界面操作等多项能力；2）doubao-seed-1.6-thinking：豆包大模型1.6系列在**深度思考方面的强化版本**；在代码、数学、逻辑推理等基础能力上进一步提升，支持256K 上下文；3）doubao-seed-1.6-flash：豆包大模型1.6系列的极速版本，支持深度思考、多模态理解、256K 上下文；延迟极低，TOPT 仅需10ms；**视觉理解能力比肩友商旗舰模型。**
- 豆包大模型1.6具有更强的模型效果，在众多权威测评集上，得分均属于**国际第一梯队。**在**推理能力、多模态理解能力、GUI 操作能力**上具备领先优势。
- 在另一方面，从综合成本来看，绝大部分请求输入都在32K 以内，输入输出占比在3:1，豆包大模型1.6的综合成本（2.6元）比豆包大模型1.5·深度思考模型、DeepSeek R1的综合成本（7元）**下降63%**，相当于只需原来三分之一的价格，就能使用能力更强、原生多模态的新模型。

豆包大模型1.6测试表现对比



豆包大模型1.6定价



豆包大模型1.6综合成本对比

2.1.1 Seedance 1.0 pro 视频生成登顶，火山引擎生态助力增长

- 2025年6月11日，在 FORCE 原动力大会上，火山引擎正式全新发布的豆包·视频生成模型 Seedance 1.0 pro 具备三大特性：**无缝多镜头叙事、多动作及随心运镜、稳定运动与真实美感**，在电商、影视、游戏等行业应用前景广阔。根据第三方权威榜单 Artificial Analysis 最新结果，Seedance 1.0 pro 在**文生视频、图生视频**两个维度都超越了业界诸多主流模型，**登顶全球竞技场第一。**
- 同日，火山引擎发布 AI 云原生全栈服务，帮助企业加速 Agent 落地。核心组件为，链接云服务与开发工具的**MCP 服务**，与AI自然语言对话完成开发全流程的**IDE 产品TRAE**，以及一站式**AI Agent开发工具扣子**等。更加完善的AI应用生态使得让全行业的企业都能更轻松地使用大模型、提升生产力。截至2025年初，火山引擎合作伙伴规模达到5,000多家，**助力伙伴产出增长超200%。**

Seedance 1.0 pro测试表现对比

豆包大模型1.6合作伙伴简况

Artificial Analysis Video Arena Leaderboard

Text to Video | Image to Video

Creator	Model	Arena ELO	95% CI	# Appearances
ByteDance Seed	Seedance 1.0	1299	-13/+13	4,947
Google	Veo 3 Preview	1252	-10/+10	8,033
Google	Veo 2	1128	-8/+8	12,114
Kuaishou	Kling 2.0	1114	-9/+9	9,467
Kuaishou	Kling 1.5 (Pro)	1054	-5/+5	24,211
OpenAI	OpenAI Sora	1053	-5/+5	25,251

Artificial Analysis 文生视频榜单

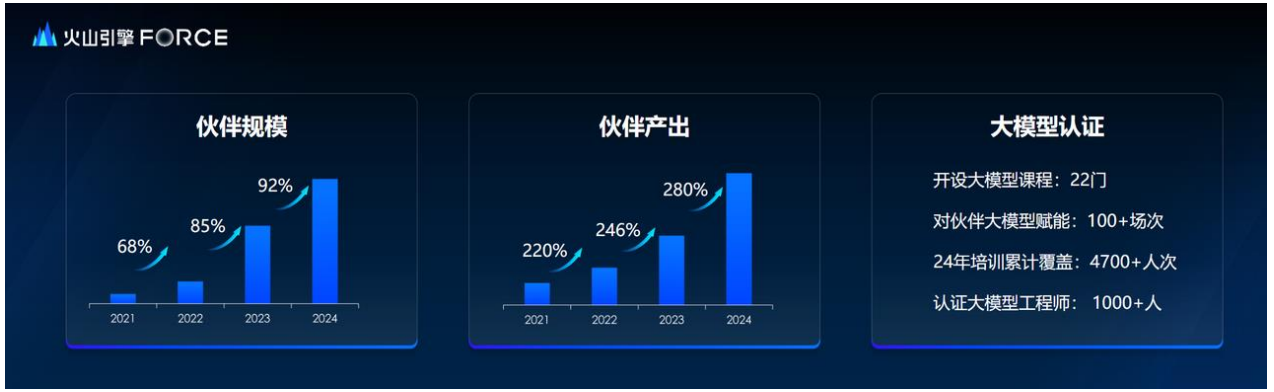
Artificial Analysis Video Arena Leaderboard

Text to Video | Image to Video

Creator	Model	Arena ELO	95% CI	# Appearances
ByteDance Seed	Seedance 1.0	1343	-13/+14	5,454
Google	Veo 3 Preview	1238	-11/+11	8,261
Kuaishou	Kling 2.0	1197	-10/+10	9,748
Kuaishou	Kling 1.6 (Pro)	1138	-9/+9	10,471
Runway	Runway Gen 4	1121	-9/+9	21,161
Google	Veo 2	1118	-9/+9	10,564

Artificial Analysis 图生视频榜单

截止2025年6月9日



2.1.2 DeepSeek最新升级，推理与精度双突破

- 2025年5月28日，Deepseek发布了版本为 DeepSeek-R1-0528的升级，虽仍然使用 2024 年 12 月所发布的 DeepSeek V3 Base 模型作为基座，但在后训练过程中投入了更多算力，显著提升了模型的思维深度与推理能力。**更新后的 R1 模型在数学、编程与通用逻辑等多个基准测评中取得了当前国内所有模型中首屈一指的优异成绩，并且在整体表现上已接近其他国际顶尖模型，如 o3 与 Gemini-2.5-Pro。**
- 相较于旧版 R1，新版模型在复杂推理任务中的表现有了显著提升。例如在 AIME 2025 测试中，新版模型准确率**由旧版的 70% 提升至 87.5%。**这一进步得益于模型在推理过程中的思维深度增强：在 AIME 2025 测试集上，旧版模型平均**每题使用 12K tokens**，而新版模型平均**每题使用 23K tokens**，表明其在解题过程中进行了更为详尽和深入的思考。同时，新版 DeepSeek R1 针对“幻觉”问题进行了优化。与旧版相比，更新后的模型在改写润色、总结摘要、阅读理解等场景中，**幻觉率降低了 45～50% 左右**，能够有效地提供更为准确、可靠的结果。

DeepSeek发展历程

DeepSeek-R1-0528测试表现

Benchmarks	DeepSeek-R1-0528	OpenAI-o3	Gemini-2.5-Pro-0506	Qwen3-235B	DeepSeek-R1
AIME 2024 数学竞赛 *pass@1*	91.4	91.6	90.8	85.7	79.8
AIME 2025 数学竞赛 *pass@1*	87.5	88.9	83.0	81.5	70.0
GPQA Diamond 科学测试 *pass@1*	81.0	83.3	83.0	71.1	71.5
LiveCodeBench 代码生成 *pass@1*	73.3	77.3	71.8	66.5	63.5
Aider 代码编辑 *pass@1*	71.6	79.6	76.9	65.0	57.0
Humanity's Last Exam 推理与百科知识 *pass@1*	17.7	20.6	18.4	11.75	8.5

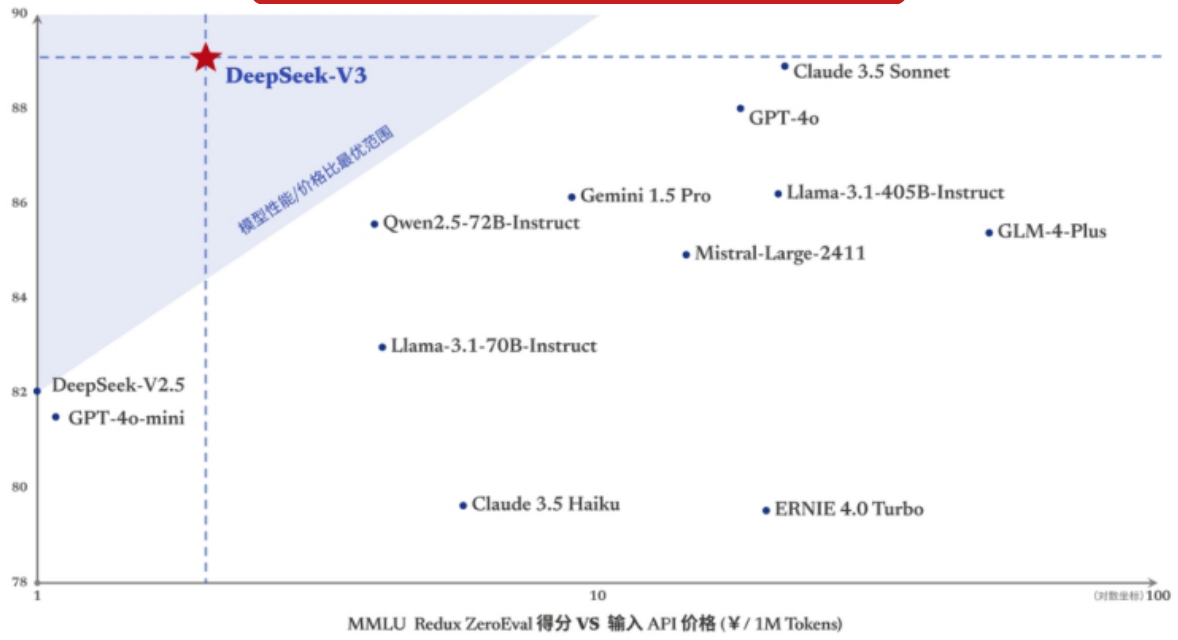
资料来源：央视网，DeepSeek API文档，未尽研究，HsuDan，民生证券研究院整理

资料来源：DeepSeek API 文档，民生证券研究院

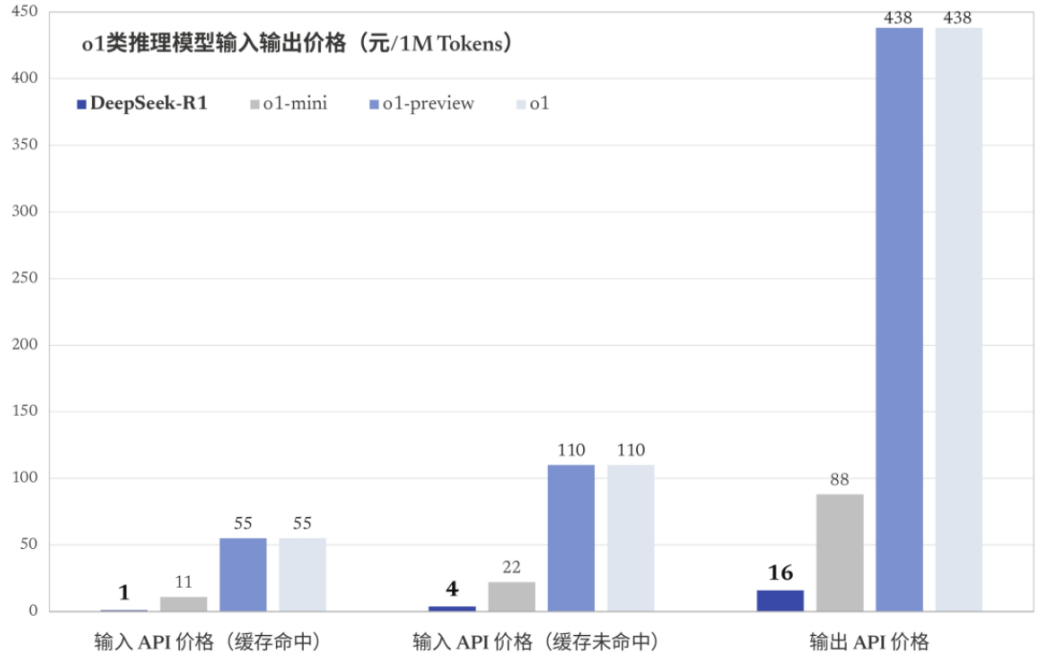
2.1.2 DeepSeek架构创新降本，解锁轻量化部署新可能

- Deepseek优越的测试表现离不开其始终专注的算法优化与创新，致力于通过前沿技术推动大模型性能和效率的突破。其中一大核心创新，**DeepSeek大模型均采用了自主改进的DeepSeekMoE架构**，融合了专家混合系统(MoE)、多头潜在注意力机制(Multi-Head Latent Attention, MLA)和RMSNorm三个核心组件。通过专家共享机制、动态路由算法和潜在变量缓存技术，该模型在保持性能水平的同时，实现了**相较传统MoE模型40%的计算开销降低**，真正意义上打开了**轻量级部署**的无限可能性。
- DeepSeek开创了低成本大模型的新范式。得益于技术路线优化，DeepSeek-V3仅使用2048块NVIDIA H800 GPU，累计训练278.8万个小时，**耗费557.6万美元**（假设单位训练成本为2美元/小时）。作为对比，**GPT-4o训练成本约为1亿美元。**
- DeepSeek-V3和DeepSeek-R1的线上系统服务器均采用NVIDIA H800 GPU，并在白天将所有节点用于推理，晚上用于训练，日均成本为8.7万美元（假设单位租赁成本为2美元/小时），对应理论日均收入为56.2万美元（假设所有token按照DeepSeek-R1 API定价），**理论成本利润率高达545%。**

DeepSeek-V3与其他大模型性能/价格对比



o1类推理模型输入输出价格（元/1M Tokens）



资料来源：DeepSeek API 文档，民生证券研究院

资料来源： DeepSeek API 文档，民生证券研究院

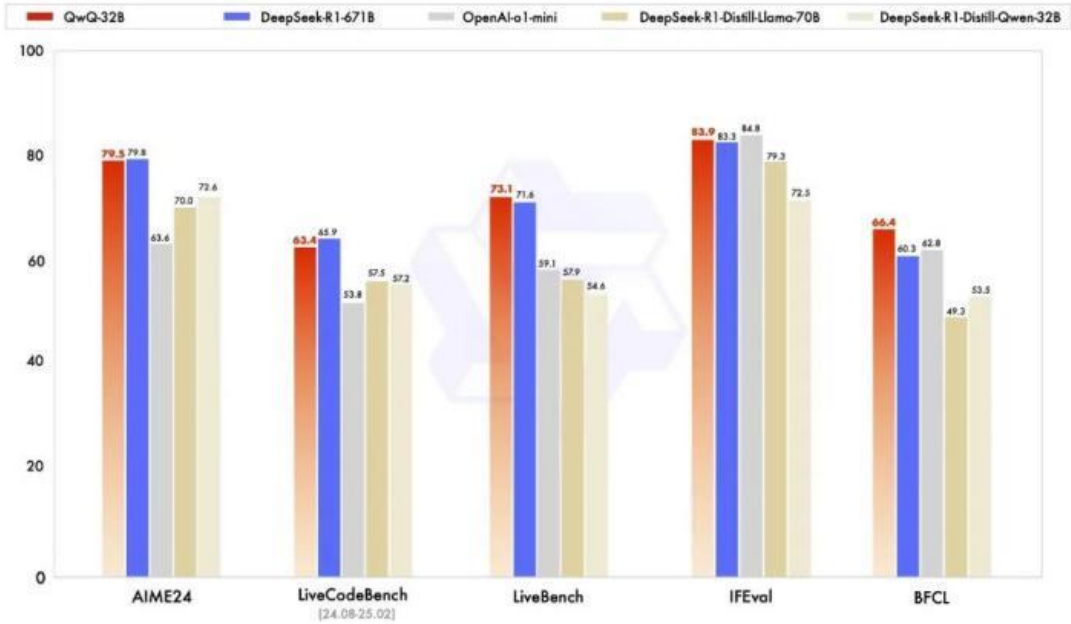
2.1.3 国产模型竞速升级，轻量化浪潮提速

- 自**DeepSeek-R1**模型发布以来短短几个月内，豆包、通义千问、百度、腾讯混元、阶跃星辰和Kimi等其他国产大模型均有重大更新，在模型轻量化方面取得了显著进展。以豆包和通义千问为例：
- **豆包：**2025年1月22日豆包发布Doubao-1.5-pro，模型使用与DeepSeek相同的 MoE 架构，通过训练与推理一体化设计，**在保持高性能的同时显著降低了推理成本。**在相同的9T tokens训练数据下，激活参数仅为稠密模型1/7的MoE模型，超过了稠密模型的性能，实现了7倍的性能杠杆提升。
- **通义千问：**2025年3月6日，阿里巴巴发布并开源全新的推理模型通义千问QwQ-32B（320亿参数），该模型通过大规模强化学习，性能可与具备 6710 亿参数（其中370亿被激活）的DeepSeek-R1媲美。并在测试数学能力的 AIME24评测集上和评估代码能力的LiveCodeBench中，表现与DeepSeek-R1相当。同时，**QwQ-32B大幅降低了部署使用成本，在消费级显卡上也能实现本地部署**，适合高效或高安全需求场景，显著优于同类轻量模型。

豆包Doubao-1.5-pro与其他大模型性能对比

		Doubao-1.5-pro	Llama3.1-405B	GPT4o-0806	Gemini-exp-1205	Claude-3.5-Sonnet-latest	Qwen2.5	DeepseekV3
Knowledge	MMLU	88.6	88.6	**88.7**	86.8	88.5	85.6	88.5
	MMLU_PRO	**80.1**	73.3	74.9	76.4	78.0	71.1	75.9
	GPQA	**65.0**	51.1	53.1	62.1	**65.0**	49.0	59.1
MATH	Math	88.6	73.8	75.9	**89.7**	78.3	83.1	87.8
	Olympiadbench	59.8	34.1	40.7	**64.7**	43.5	50.0	59.1
Code	MBPP+	78.0	72.8	78.3	78.6	76.5	76.9	**79.3**
	Mbpp	**70.2**	58.7	68.2	67.0	68.2	61.7	69.4
	Livecodebench	**65.1**	53.6	61.8	62.6	60.3	56.9	63.3
Reasoning	BBH	91.6	89.2	91.7	**92.6**	**92.6**	88.3	92.3
	DROP	**93.0**	91.2	79.8	89.7	88.3	87.4	91.6
Instruction Following	IFEval	89.5	86.0	84.7	**89.8**	89.3	84.1	86.1
	SysBench	57.6	58.9	62.2	**69.0**	**69.0**	47.2	66.3
Chinese	CMMLU	**90.9**	75.4	77.3	84.3	81.2	84.3	83.5
	C-Eval	**91.8**	72.7	76.0	85.9	80.0	86.1	86.5

通义千问QwQ-32B与其他大模型性能对比



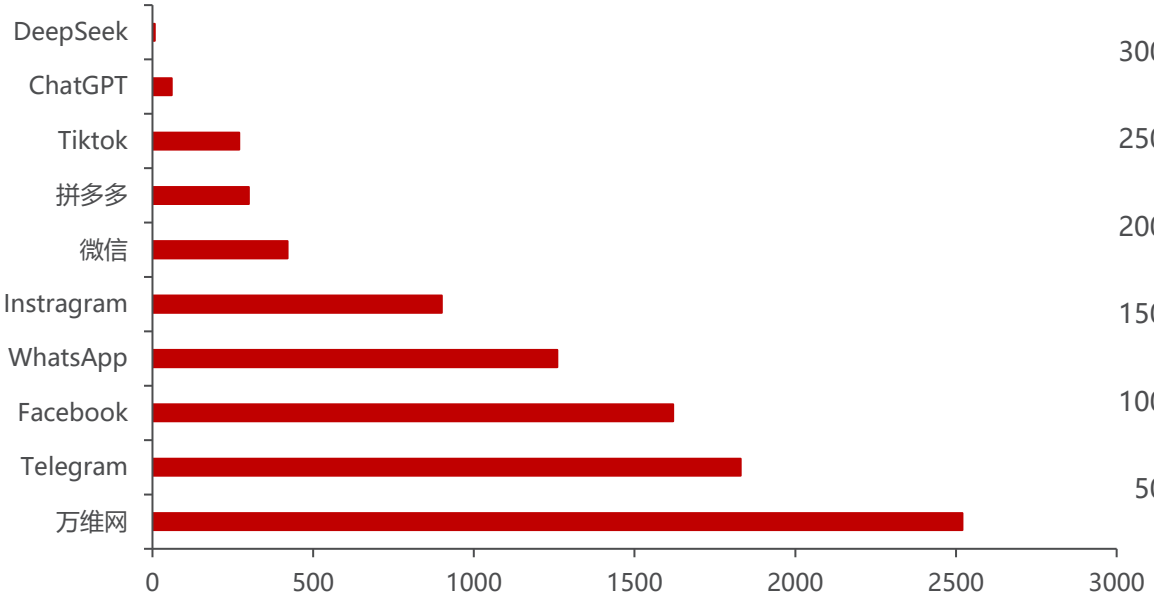
资料来源：豆包官网，民生证券研究院

资料来源：通义千问Qwen，民生证券研究院

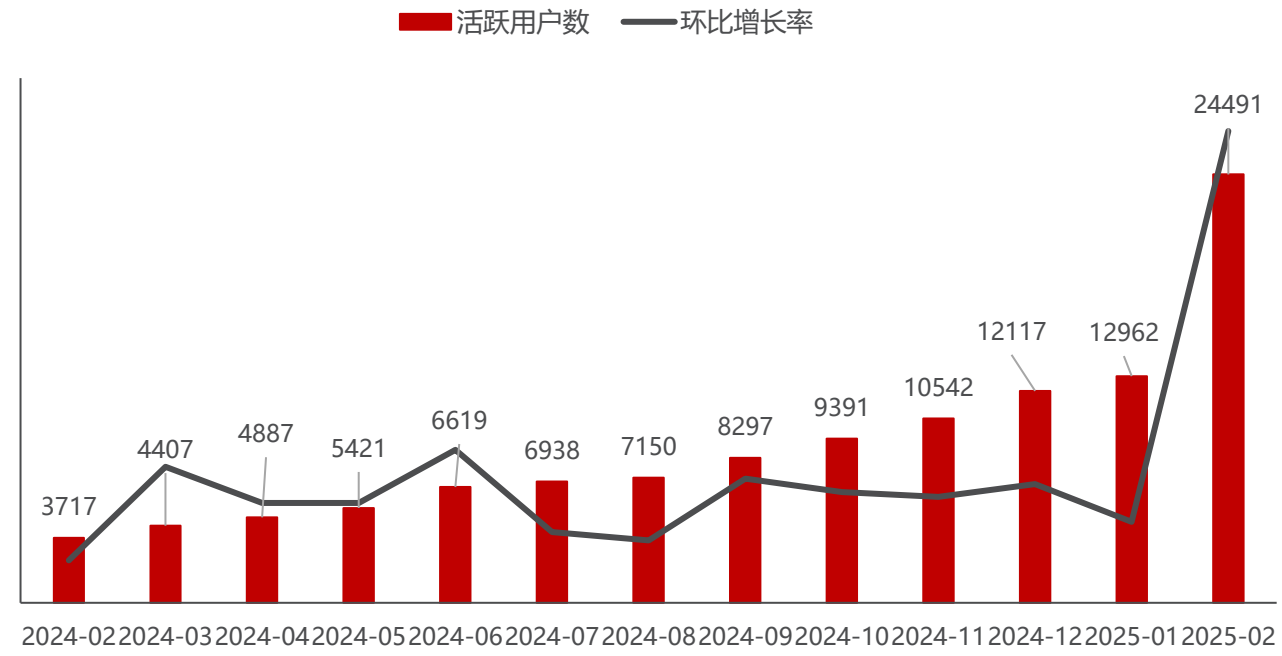
2.1.3 基于轻量化趋势，探讨推理需求激增

- **我们认为，随着大模型轻量化的趋势，更低的门槛激发出了更广泛的推理场景与需求。**轻量化降低了AI模型的使用成本，加速了AI在消费级市场和垂直行业的渗透，催生出更丰富的AI原生应用，进而推动推理算力需求持续上升。
- DeepSeek推出后实现了用户的高速增长，彰显出大众对低成本轻量化大模型的强烈需求。在无广告投入的前提下，仅用**短短7天便实现用户量突破1亿**的里程碑式增长，充分展现了DeepSeek大模型产品的吸引力与市场渗透能力。
- DeepSeek-R1模型在2025年1月份发布后迅速引爆AI大模型需求。在DeepSeek-R1发布后的一个月内，国内原生App行业规模几近翻倍，截至2025年2月底，AI原生APP用户规模达到2.4亿，**相比1月份增长88.9%。**这一增长速度体现出DeepSeek-R1在大模型市场上的助推效应，进一步印证了**轻量化、高性能的大模型具备强大的用户吸引力，推动了大模型的需求扩张。**

各产品增长1亿用户所用时间（天）



AI原生App月活跃用户规模趋势（万人）



资料来源：AI产品榜，民生证券研究院整理

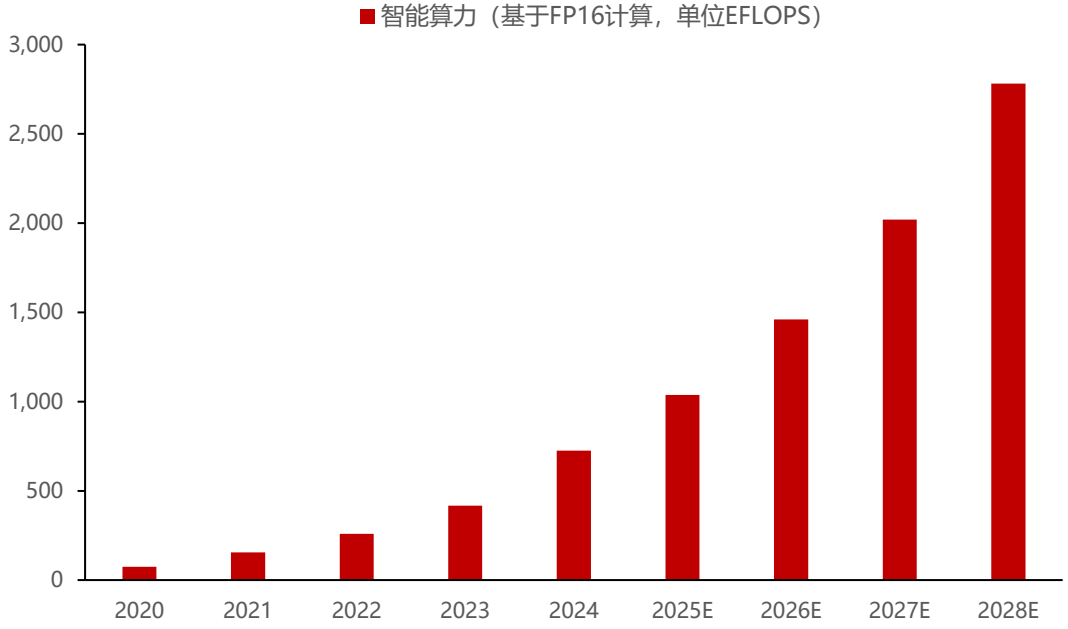
资料来源：QuestMobile，民生证券研究院整理

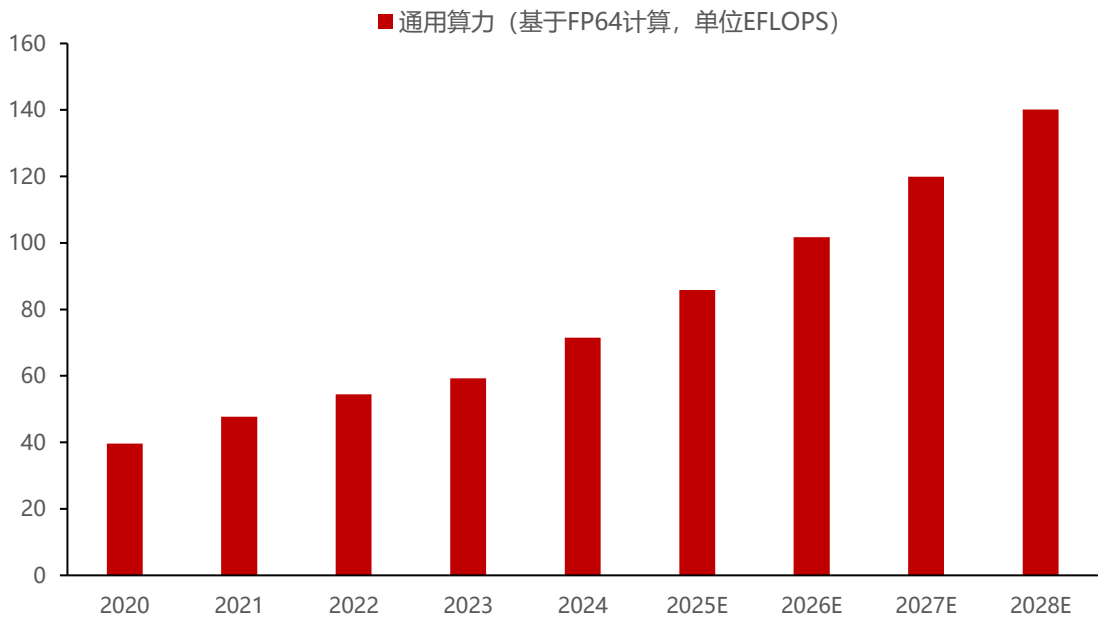
CONTENTS 目录

2.2.1 中国AI基建全面提速，BAT开启“抢算”竞速

- 根据 IDC 测算，2025年**中国智能算力规模**将达**1,037.3 EFLOPS**，预计到2028年将**突破 2,781.9 EFLOPS**，2023–2028年**年均复合增长率达46.2%**。同期，**中国通用算力规模**预计将从2025年的 **85.8 EFLOPS 增长至2028年的 140.1 EFLOPS，年均复合增长率为 18.8%**，体现出 AI 基础设施建设全面提速。
- **BAT开启“抢算”竞速。**2024年，字节跳动和腾讯各自订购了**约 23 万块**英伟达芯片（包括H20），用于满足背后对大模型和推理对算力的需求。未来，腾讯、阿里、字节跳动在算力资源领域的竞争将愈发激烈。同时，面对日益增长的算力需求与有限的资源供给，巨头们可能会通过建立算力交易平台、开展算力合作项目等方式，实现算力资源的高效利用与互利共赢。

中国智能算力和通用算力规模及预测（2020-2028）





资料来源：IDC浪潮信息，民生证券研究院整理

资料来源：IDC浪潮信息，民生证券研究院整理

2.2.1 算力投入驱动腾讯CAPEX激增

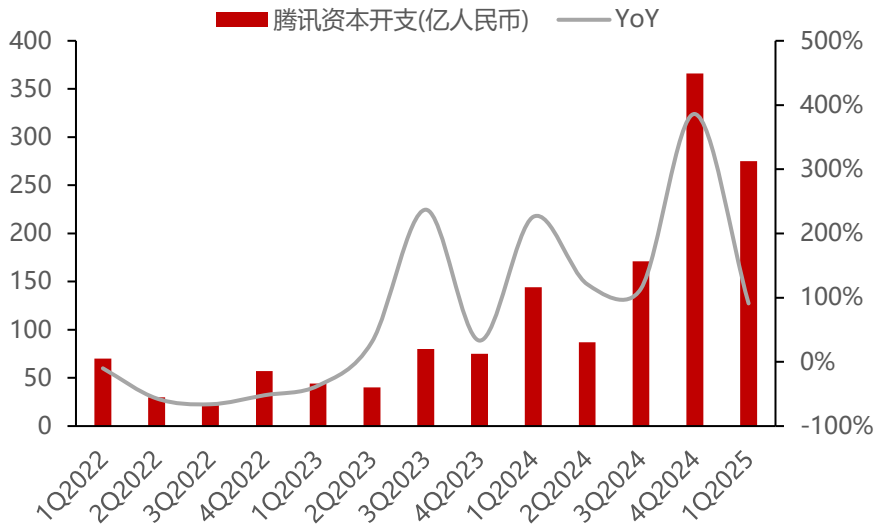
- 腾讯资本开支从2020年**340亿元增长到2024年768亿元**，年均复合增长率（CAGR）约为22.59%，体现出其在AI基础设施、数据中心及算力布局等领域持续加码的趋势。2022至2023年间资本开支增速相对放缓，主要聚焦于传统业务的稳健推进；而2024年起，**AI战略已进入重投入期**，资本开支出现大幅跃升，连续四个季度实现同比三位数增长，其中24Q4资本支出增加十分显著，**同比增长386%至366亿元**，主要因为腾讯在这一季度购买了更多GPU以满足推理需求。年度资本开支更**达到768亿元，同比增长221%**，创历史新高，展现出新一轮战略转型的力度与节奏。
- 腾讯2025年Q1资本开支**同比增长91%达到275亿元，**其中运营资本支出为264亿元人民币，**同比增长近300%，**主要是加大对GPU和服务器的投资；腾讯计划进一步加大资本开支，预计会占2025年总收入的“低两位数百分比“，这意味着**2025年腾讯的资本开支可能接近千亿元的水平。**腾讯目前已经在AI领域发起了“**从算力到人力**”的全面加码，重点投向AI原生应用研发及算力基础设施。

2025Q1腾讯季报 - 开支部分

亿元	1Q 2025	1Q 2024	YoY	4Q 2024	QoQ
运营性资本开支	26.4	6.6	+299 %	34.9	-24%
非运营性资本开支	1.1	7.8	-86%	1.7	-35%
总资本开支	27.5	14.4	+91%	36.6	-25%

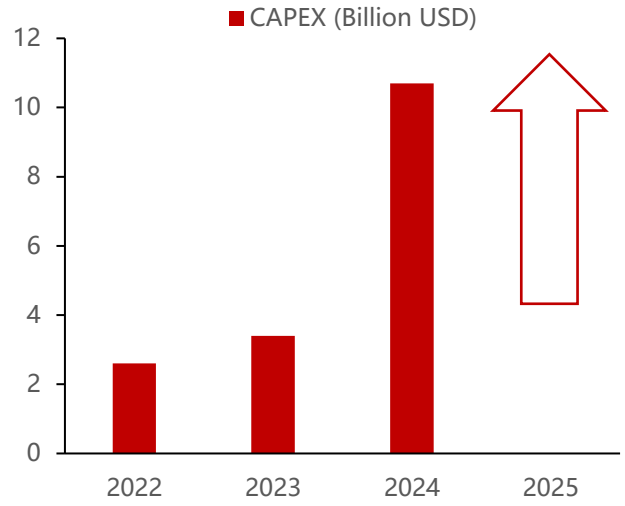
资料来源：腾讯公司官网，民生证券研究院整理

2022Q1-2025Q1腾讯资本开支



资料来源：腾讯公司官网，民生证券研究院整理

腾讯AI投资展望

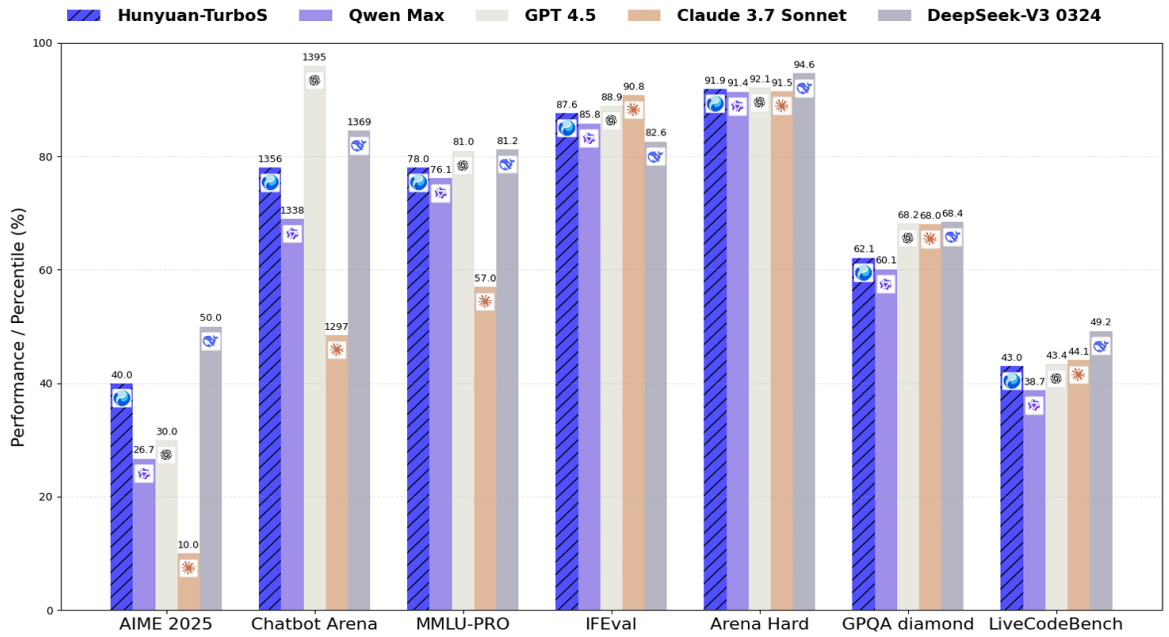


资料来源：腾讯公司官网，民生证券研究院整理

2.2.1 腾讯AI产品加速渗透，算力布局走向全球

- **腾讯采用“自研+开源”的多模型策略。**不仅相继推出深度思考模型 T1 和快思考模型 Turbo S，同时，腾讯还与国产开源模型 DeepSeek 深度合作，将其模型整合至元宝、微信 AI 搜索等核心产品。**在C端市场，AI已全面渗透腾讯用户场景，**腾讯元宝借助“混元+DeepSeek”双引擎驱动，用户规模显著增长。35天迭代30个版本，不断上线实用功能。2月至3月，腾讯元宝日活激增超20倍，成为了中国DAU排名第三的AI产品。
- **此外，腾讯云持续推进全球基础设施布局。**其国际业务2021年-2024年**连续三年保持两位数增长**，截至2024年底，基础设施已覆盖全球五大洲21个国家和地区，运营58个可用区；同时，腾讯云将在沙特建设首个中东数据中心，并计划在未来数年内在中东地区投入超1.5亿美元。

混元-TurboS 基准性能测试

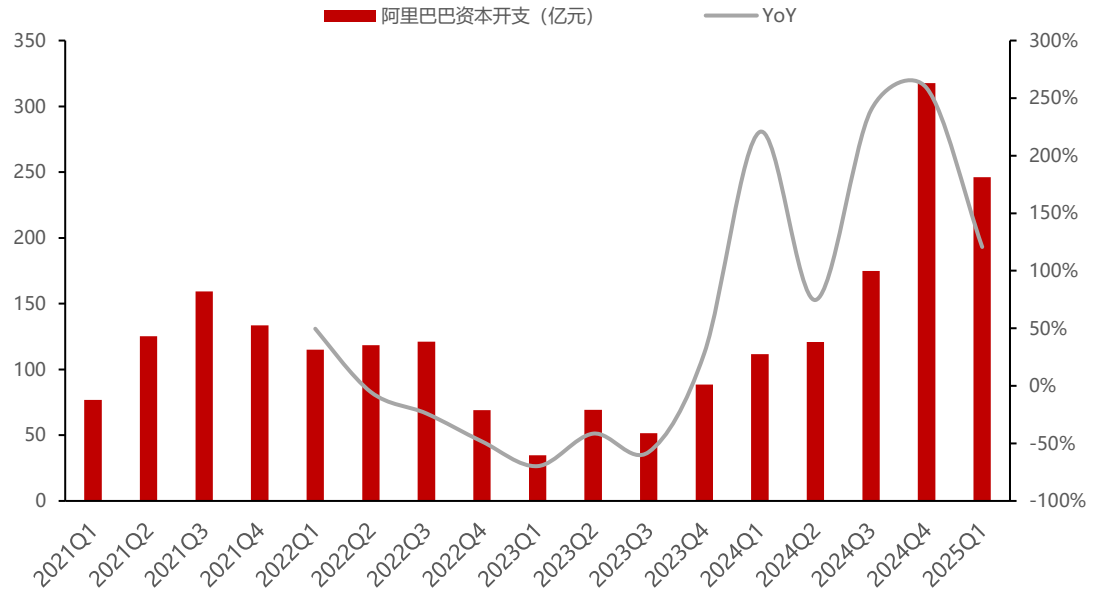


腾讯云产品全球布局

2.2.1 阿里巴巴加码CAPEX，千亿投入押注AI未来

- 阿里巴巴2021年至2024年（自然年）的资本开支**从494.95亿元跃升至725.13亿元，年均复合增长率（CAGR）约为13.58%**，展现出其在技术基础设施领域长期、持续的高强度投入态势。2021–2023年支出趋于平稳，围绕现有核心业务与技术平台进行优化整合；**24年开始大幅度增长**，全年资本开支达725.13亿元（自然年），**同比增长197.04%。**其中2024Q4 （自然年），阿里当季资本开支317.75亿元，较上个季度的174.91亿元**环比增长81.66%。**
- 阿里巴巴2025年Q1 （自然年）资本开支为**246.12亿元，同比增长120.68%。**值得注意的是，阿里云收入301.27亿元，**加速增长至18%，为三年来最快增速。**阿里将持续加码AI与云计算领域投资，围绕“AI+云计算”核心战略，聚焦AI基础设施、基础模型平台及AI原生应用三大方向。**未来三年，阿里将投入超过3800亿元用于建设云和AI硬件基础设施，总额超过过去十年总和，**这笔投资也创下了中国民营企业在云和AI硬件最大规模投资纪录。

2021Q1-2025Q1（自然年）阿里巴巴资本开支



2025Q1（自然年）阿里巴巴季报 - 开支部分

现金流(百万元)	2024Q1	2025Q1
经营活动产生的净现金流量	23340	27520
购置财产、设备及无形资产	(10174)	(23993)
买方保障基金存款的变动情况	2195	216
自由现金流	15361	3743
与资本支出有关的净现金流出	(11153)	(24612)
来自投资和收购活动的净现金（流出）流入	(328)	2008
股份回购	(34014)	(4584)

资料来源：阿里巴巴公司官网，民生证券研究院整理

资料来源：阿里巴巴公司官网，民生证券研究院整理

2.2.1 阿里巴巴模型创新领跑，云芯协同全球布局

- **阿里巴巴在Qwen系列模型上持续加码，巩固其AI领先地位。**4月29日，阿里巴巴开源新一代通义千问模型Qwen3，其将“快思考”与“慢思考”集成一体，成了全球第一个开源的混合推理模型。对比主流大模型，Qwen展现出国际领先水平。目前，通义千问Qwen衍生模型数量已突破10万，超越美国Llama模型，**通义成为全球第一AI开源模型。**
- 阿里巴巴云计算基础设施全球扩张，阿里云已在全球**29**个地域、**87**个可用区部署数据中心，形成全球服务网络。同时，阿里巴巴在AI芯片、AI云服务、AI算法、AI平台、产业AI等方面实现全线领先。旗下平头哥半导体已推出首款云原生处理器芯片——倚天710，**性能超过业界标杆20%，能效比提升50%以上。**此外，阿里还推出了高性能RISC-V处理器如玄铁910等产品，**出货超40亿颗。**

Qwen3-235B-A22B与其他主流模型对比

	Qwen3-235B-A22B MoE	Qwen3-32B Dense	OpenAI-o1 2024-12-17	Deepseek-R1	Grok 3 Beta Think	Gemini2.5-Pro	OpenAI-o3-mini Medium
ArenaHard	95.6	93.8	92.1	93.2	-	96.4	89.0
AIME'24	85.7	81.4	74.3	79.8	83.9	92.0	79.6
AIME'25	81.5	72.9	79.2	70.0	77.3	86.7	74.8
LiveCodeBench v5, 2024.10-2025.02	70.7	65.7	63.9	64.3	70.6	70.4	66.3
CodeForces Elo Rating	2056	1977	1891	2029	-	2001	2036
Aider Pass@2	61.8	50.2	61.7	56.9	53.3	72.9	53.8
LiveBench 2024-11-25	77.1	74.9	75.7	71.6	-	82.4	70.0
BFCL v3	70.8	70.3	67.8	56.9	-	62.9	64.6
MultiIF 8 Languages	71.9	73.0	48.8	67.7	-	77.8	48.4

1. AIME'24/'25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.
2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.
3. BFCL: The Qwen3 models are evaluated using the FC format, while the baseline models are assessed using the highest scores obtained from either the FC or prompt formats.

阿里云全球分布图

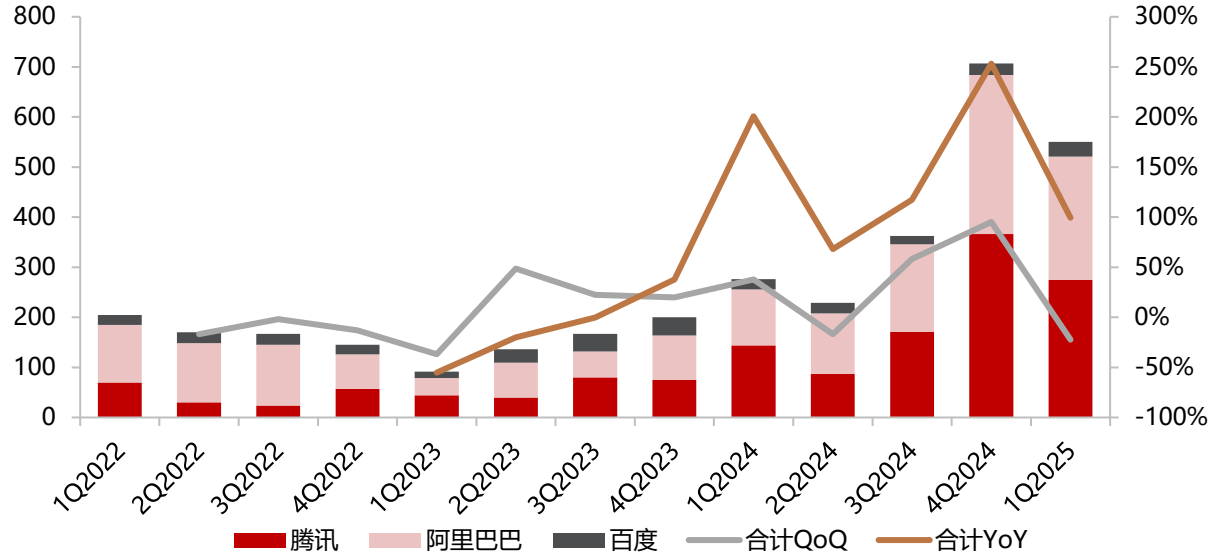
资料来源：阿里Qwen团队，民生证券研究院

资料来源：阿里云官网，民生证券研究院

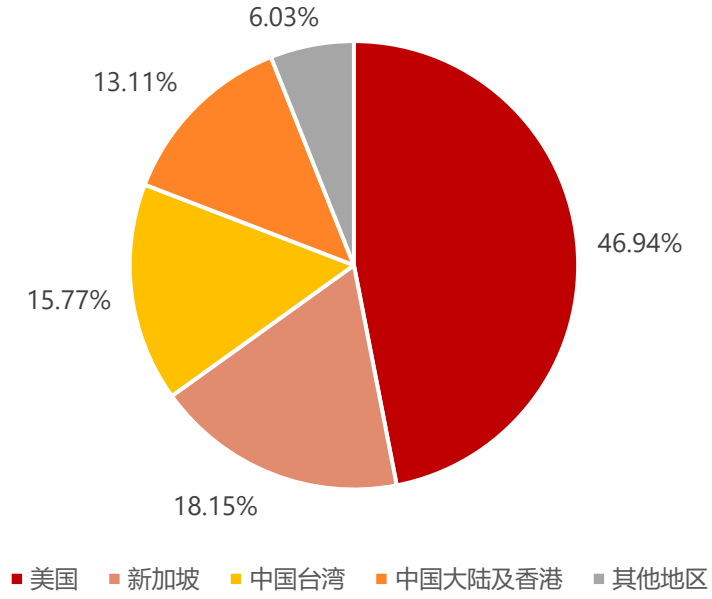
2.2.1 BAT开启算力竞速，资本开支持续高涨

- BAT（百度、阿里、腾讯）合计资本开支从**2022年687.4亿元到2024年1574.47亿元**，在两年内实现了年均复合增长率（CAGR）约为**51.34%**，展现出在AI大模型、云计算基础设施和高性能算力部署上的全面提速。从英伟达2025财年收入上看，中国大陆及香港地区销售额为170.1亿美元，占比已达到13.11%。
- **BAT 1Q2025合计资本开支同比继续高增，环比小幅下降。**1Q2025 BAT合计资本开支同比提升99.38%至550.12亿人民币，环比较4Q24的707.08亿人民币下降了22.2%。其中，腾讯，阿里巴巴和百度1Q2025资本开支分别为275，246.12和29亿人民币，同比增速分别为90.97%，120.68%，42.3%，环比分别为-24.86%，-22.54%，24.30%。
- 1Q2025国内互联网厂商BAT合计资本开支环比下降，我们认为主要原因可能包括于美国对英伟达高端GPU出口政策，其次很多企业会集中在年底确认固定资产、完成设备采购和部署合同，形成“季末集中确认”效应。但是**算力作为科技发展的核心驱动力这一事实不会改变**，BAT必将加大对AI领域的研发投入，试图抢占下一代算力技术的制高点。

BAT资本开支及增速（亿元，%）



英伟达2025财年按地区划分收入



资料来源：各公司官网，民生证券研究院整理

资料来源：各公司官网，民生证券研究院整理

2.2.2 服务器：算力之基座，全面适配国产算力

- 大模型蓬勃发展催生算力需求，云厂商资本开支高增，AI服务器作为算力之基，率先受益于国内算力基建。传统大型企业倾向于采购**品牌服务器厂商**提供的标准化服务器。2024Q1戴尔、**浪潮、联想、**惠普、超微及其他品牌厂商以强大的品牌影响力、全面的服务和支持，占据**64%**的市场份额。**白牌服务器厂商**通过提供定制化的服务器解决方案，能满足云计算厂商的特定需求。如广达、纬颖等ODM厂商，2024Q1占据了服务器市场36%的市场份额，**华勤技术则为内资白牌服务器ODM龙头，成为云厂商AI服务器核心供应商。**
- **腾讯、阿里等国内头部云厂商资本开支快速增长，带动国内算力需求蓬勃发展。**受中美政治经济摩擦影响，国内云厂商难以采购英伟达先进的H100或B系列芯片，而是主要采购英伟达中国特供版的H20芯片，**然后交由华勤、联想、浪潮等服务器厂商进行加工制造。**2025年4月，英伟达发布声明称应美国政府要求，向中国出口H20相关芯片需获得出口许可，自此H20芯片事实被禁，国内AI服务器制造商面临“缺卡”窘境。
- 英伟达即将推出B20/B40芯片，进一步削弱算力、显存等配置，以满足美国对中国出口需求；尽管芯片参数遭到进一步限制，但受制于高性能算力芯片的短缺，预计短期内英伟达B20/B40芯片仍是国内云厂商的重要算力来源之一。

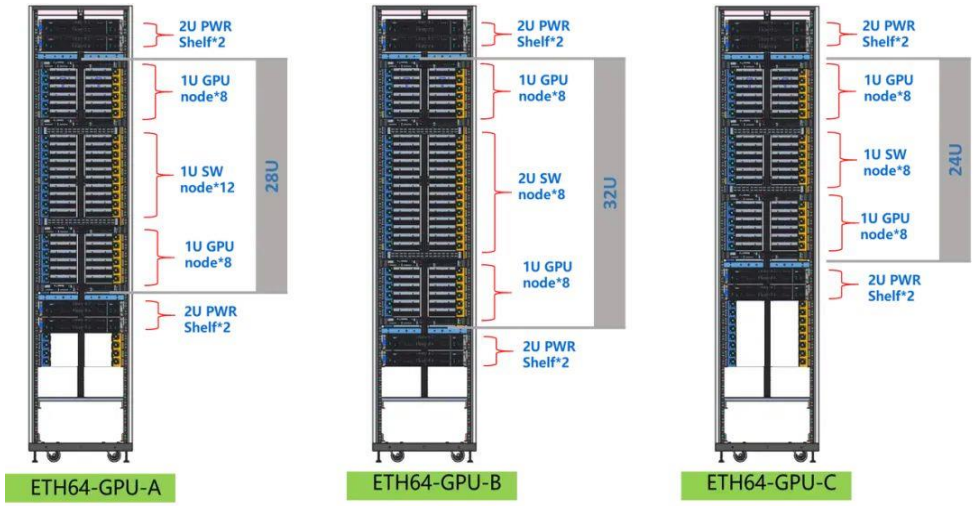
华勤技术8卡AI服务器

华勤技术400G交换机

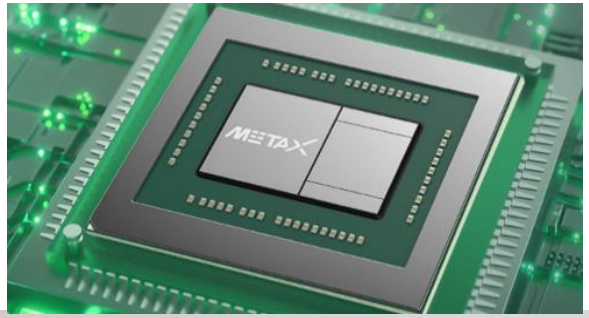
2.2.2 服务器：算力之基座，全面适配国产算力

- 远期来看，美国对英伟达AI算力卡实行出口限制，国内对于自主可控的追求必将引起国产AI算力芯片浪潮，**而国内服务器厂商已推出适配国产算力芯片的服务器，为国产算力快速增长做好准备。**由开放数据中心委员会（ODCC）主导、中国信通院与腾讯牵头设计，联合华勤技术、立讯技术等30余家产学研机构共同研发的ETH-X开放超节点项目，在华勤技术东莞智能制造基地下线。ETH-X以超大带宽、超低时延、开放互联技术为核心构建业界首个弹性超节点系统，探索在单芯片算力发展受限情况下突破算力瓶颈的新途径，**同时可以支持多种不同厂家GPU芯片、交换芯片的组合使用。**

ETH-X超节点方案



燧原、沐曦、寒武纪、摩尔线程国产算力芯片



资料来源:开放数据中心委员会，新浪财经，遂原官网，寒武纪官网，摩尔线程官网，MetaX官网

2.2.2 服务器：本地部署需求旺盛，助推AI一体机发展

大模型一体机是专为企业和政府设计的“一站式AI工具箱”，它将高性能算力硬件（如GPU芯片）、大模型软件（如DeepSeek）和安全管理功能集成在一个机柜中。**用户无需连接互联网**，即可在**本地完成数据训练、模型部署**等任务，**确保敏感数据不外泄**。其核心价值在于**降低算力使用门槛**，提升AI训练与推理效率，尤其适用于大模型训练、自动驾驶、智能制造等高算力需求场景。相比于云部署模式，采用一体机进行私有化部署具有以下优点：

1）**数据安全：政府、金融、电信**等行业客户具备大量敏感数据，本地化部署数据不脱离内网，保障数据安全。

2）**简化部署：**支撑客户**开箱即用**，无需进行二次硬件适配，从而降低了技术门槛，同时充分释放硬件性能，从而**提升模型训练和推理的效率。**

3）**成本更低：**长期使用公有云API按token付费成本较高，通过一体机私有化部署有助于**降低总体成本**并更好地掌控预算。

DeepSeek使用知识蒸馏技术，将R1大模型的复杂知识及思维链能力蒸馏至Qwen/Llama的开源小模型中，实现模型的轻量化，使用DeepSeek-R1进行蒸馏后的小模型推理能力显著提高，甚至能够超越o1-mini，表明了大模型的推理能力向小模型迁移的可能。蒸馏后的小模型参数量在1.5B-70B不等，**适合利用价格相对便宜、配置相对较低的一体机进行本地部署。**

目前，DeepSeek已赋能金融、医疗、汽车等20余个行业。这种“模型+算力+场景”的深度融合，使得AI一体机能够以开箱即用、灵活定制的特性，降低企业智能化转型门槛，加速“AI+垂类应用“的全方位落地。政务或教育、医疗等行业有不同的专属数据，利用特定数据对大模型投喂而开发的垂类应用，不仅具备隐私性、快速响应的优势，**更重要的是客户可以利用个性化的行业数据对模型进行调优，生成更符合行业特点和使用习惯的模型。**

Deepseek蒸馏小模型能力突出

	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759.0
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717.0
o1-mini	63.6	80.0	90.0	60.0	53.8	**1820.0**
QwQ-32B	44.0	60.0	90.6	54.5	41.9	1316.0
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954.0
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189.0
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481.0
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691.0
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205.0
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633.0

传统云服务和大模型一体机技术对比

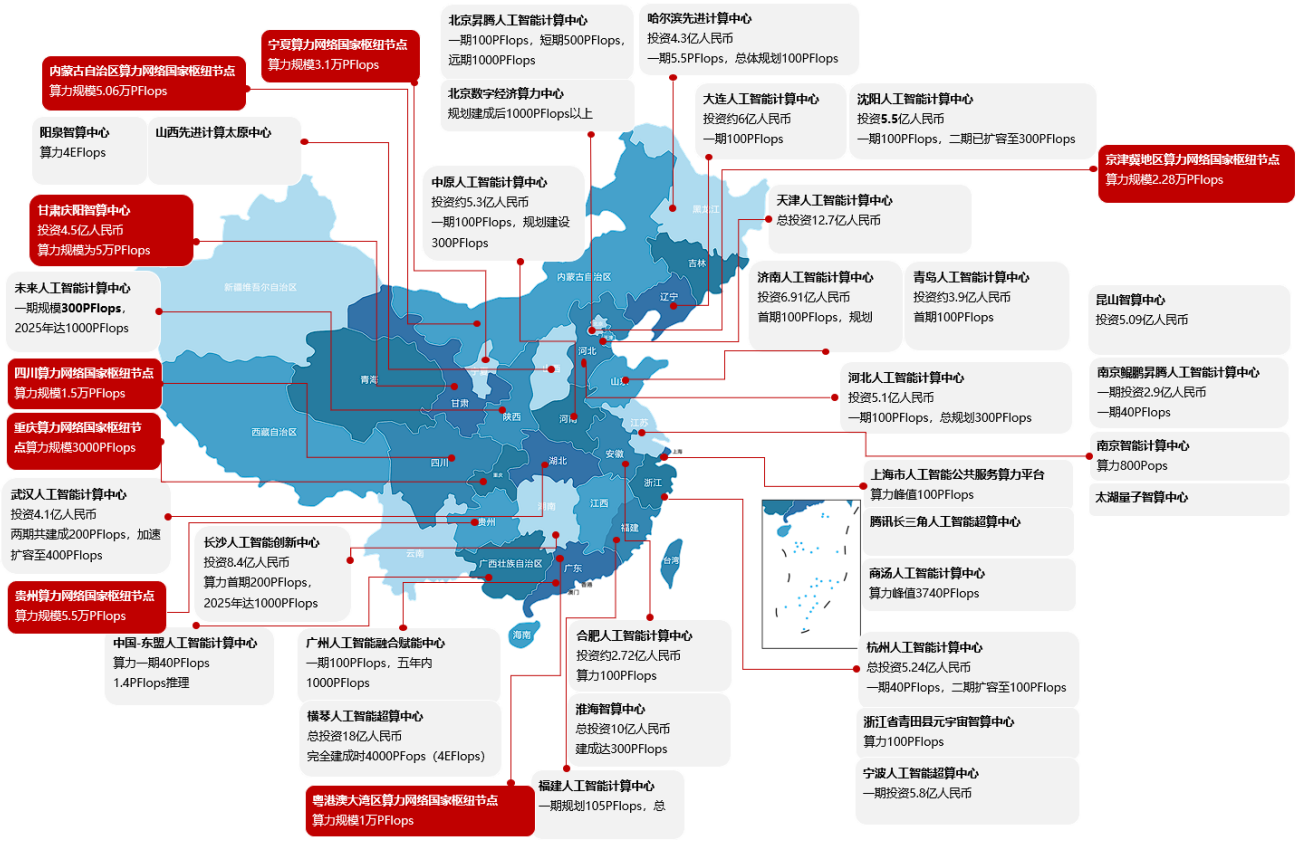
比较项目	传统云服务	大模型一体机
安全性	数据需上传至云端，存在泄露风险	数据完全本地处理，物理隔离
安装便捷程度	部署周期长达数周	3小时开机即用
成本	按流量/时长付费，成本不可控	一次性采购，5年运维成本降低40%
兼容性	依赖国外芯片（如英伟达）	支持国产芯片（华为昇腾等）

资料来源：DeepSeek官网，CSDN，中国教育和科研计算机网等，民生证券研究院

2.2.3 算力租赁：短期内算力短缺破局之道

- **算力租赁助力解决短期内算力缺口问题。**算力租赁，即算力资源方通过自建或外购算力的方式，形成一定规模的自有算力资源，然后以租赁收费的模式，为AI公司、科研院所、云厂商等提供算力支持。算力租赁厂商一般拥有AI算力卡拿卡渠道，以及有充沛的在手现金可以满足重资产的算力投资，能够为没有足够现金自建机房的中小机构提供算力支持，同时也能为云厂商提供AI算力补充。
- DeepSeek推出以来，各大AI公司、开发者甚至制造业纷纷尝试开发相关垂类应用，同时云厂商也积极接入DeepSeek，如腾讯“元宝“接入DeepSeek后APP下载量登顶排行榜第一，催生广泛算力需求，算力租赁价格随之上涨，**根据人民财讯3月5日报道，闲鱼平台近三个月高端显卡算力租赁搜索量增长300%，RTX 3090、4090等型号显卡成了“电子硬通货”。**
- **算力租赁公司成为各地智算中心的重要算力提供商。**除大模型厂商、云厂商直接对算力租赁公司下单，各地智算中心算力建设如火如荼，亦成为算力租赁公司重要客户。智算中心一般由当地政府主导规划，算力提供商（算力建设出资方）包含资金充沛的央国企、CSP以及算力租赁公司等，其中算力租赁公司为智算中心算力建设的重要一环。根据通信产业网等数据，**甘肃、宁夏、内蒙等八大算力网络国家枢纽节点2024年算力规模之和超过20万PFlops，算力建设需求旺盛，算力租赁公司深度受益。**

中国智算中心布局



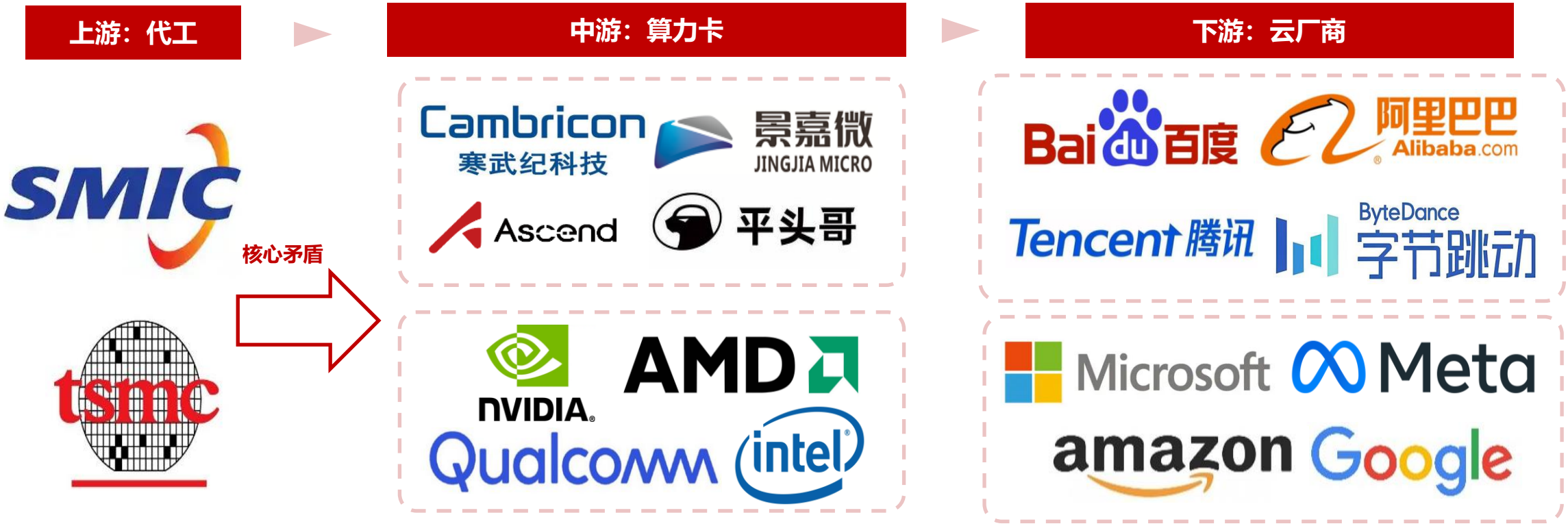
资料来源：黑龙江新闻网，金普新闻，华为官网，天津数据发展中心，济南公共资源交易中心，山东省人民政府官网等，民生证券研究院

CONTENTS

目录

2.3.1 国产大模型密集落地，芯片厂商加速适配国产算力生态

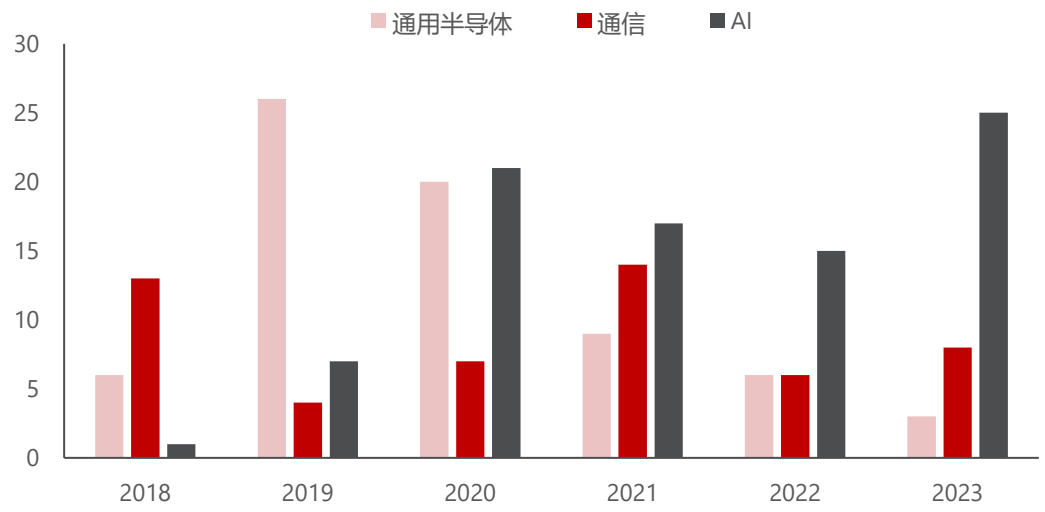
- **在国产大模型密集落地背景下，芯片厂商加速适配国产算力生态。**中芯国际N+1工艺已逐步成熟，N+2持续推进，构建国产算力底座；昇腾910C量产落地，920系列研发加快，性能持续逼近国际主流水平；寒武纪、海光等在AI训推方向深度布局，硬件端多点突破，生态融合加快。云端ASIC正成为算力演进主流，谷歌、亚马逊持续加码自研芯片体系；国内翱捷科技、芯原等设计企业快速成长，覆盖多元应用场景，并与海内外头部厂商形成紧密合作，成长弹性充足。在软件层面，适配节奏同样加快。以DeepSeek为例，发布即获得17家芯片厂商支持适配，训推效率大幅提升，助力算力生态向自主可控稳步迈进。



2.3.1 海外封锁不断升级，国产算力亟需破局之道

- **AI芯片已成为美国政府科技卡脖子的新工具。**自2018年以来，被美国列入“实体清单”的中国AI芯片企业持续增加。同时，美国聚焦于高算力芯片，限制英伟达、AMD等企业的GPU出口（AI扩散限制规则），试图全面遏制我国AI产业发展。2022年美国发布《先进计算芯片和半导体制造设备出口管制规则》，英伟达产品**A100/H100芯片**因“触线”，对华出口受阻。因此，英伟达推出专为中国市场定制的**A800/H800芯片**，得以稳住局面。2023年美国发布对华半导体出口管制最终规则，英伟达推出基于Hopper架构的中国特供版**H20芯片**。
- **封锁升级迫使国产算力加速突破。**美国最新出口管制将 “先进的国产芯片” 列为重点监管对象，试图通过技术标准割裂维持优势，英伟达还在开发专为中国市场设计的Blackwell芯片**B30芯片**，将采用GDDR7而非HBM以满足监管要求。B30 的推出本质是美国技术封锁的2.0 版本，通过 “特供”芯片维持依赖，同时挤压国产芯片生存空间。美国的限制措施激发了中国AI芯片产业的自主创新和研发，加速了国产替代的进程。中国企业面对外部压力，加大研发投入，努力构建自主可控的产业链。

2018-2023被列入“实体清单”的AI芯片企业数量



美国商务部工业和安全局对华高算力芯片出口限制

芯片名称	性能密度 PD	总计算力 TPP	是否被限制	限制依据区域
英特尔 GPU Max 1550	<1.6	1600+	否	无
英特尔 GPU Max NEXT	1.6+	1600+	否	无
AMD RX7900XTX	3.2+	1600+	是	ECCN 3A090b
AMD MI200 / MI210	3.2+	2400+	是	ECCN 3A090b
英伟达 RTX4090/RTX4090Ti	3.2+	2400+	是	ECCN 3A090b
英伟达 L40/L40S	3.2+	2400+	是	ECCN 3A090b
英伟达 A100 / A800	5.92+	2400+	是	ECCN 3A090a
AMD MI250 / MI250X	5.92+	4800+	是	ECCN 3A090a
英伟达 H100 / H800	5.92+	4800+	是	ECCN 3A090a

注：PD 超过 3.2 且 TPP 超过 2400 的都进入ECCN 3A090b（需BIS审查）；PD 超过 5.92 且 TPP 超过 4800 的被列入ECCN 3A090a（推定拒绝）

资料来源：甲子光年智库，银创智库，民生证券研究院整理

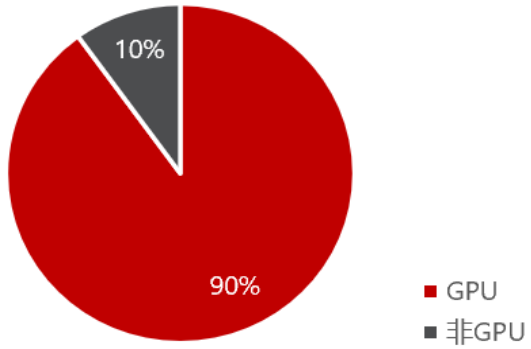
资料来源：甲子光年智库，银创智库，民生证券研究院整理

2.3.1 英伟达占据国产AI芯片主导地位，国产厂商加速跟进

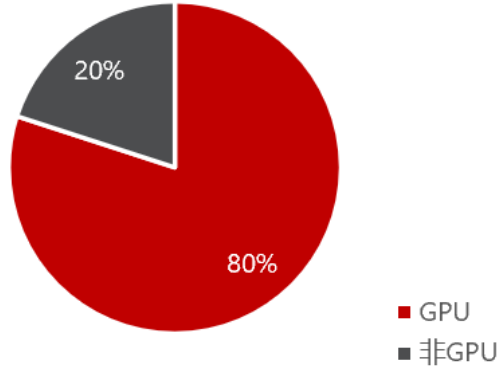
- 2023年上半年，中国加速芯片的市场规模超过50万张，GPU卡占有90%市场份额，非GPU卡占据10%市场份额。中国本土AI芯片品牌出货量近5万张，占整个市场的10%。
- 2024年上半年，中国加速芯片市场规模超过90万张。GPU卡占据80%市场份额，非GPU卡占据20%市场份额，中国本土AI芯片品牌出货量近20万张，约占整个市场的20%。

公司	型号	GPU架构	峰值INT8计算性能	峰值半精度(FP16)性能	显存容量	最大功耗	工艺制程	发布时间
NVIDIA	H100 SXM	Hopper	3958 TOPS	1979 TFLOPS	80GB	700W	4nm	2022Q1
	A100 80GB PCIe	Ampere	642 TOPS	312 TFLOPS	80GB HBM2	300W	7nm	2020Q2

2023年H1中国AI芯片市场份额



2024年H1中国AI芯片市场份额



资料来源：甲子光年智库，银创智库，民生证券研究院整理

资料来源：甲子光年智库，银创智库，民生证券研究院整理

2.3.1 国产 AI 算力卡群雄逐鹿：各主流厂商产品解析

GPU型号	910B OAM	天数 天垓150	太初 元碁T100	沐曦 曦云C500 OAM	摩尔显存 S4000 OAM	燧原 云燧T21 OAM	寒武纪 MLU370X8
单卡算力 FP16稠密	376T	190T	240T	240T	100T	128T	96T
整机8卡	约3P	1.5P	1.9P	1.9P	800T	1P	768T
显存	64G HBM2e	64G HBM2e	64GB HBM2e	64GB HBM2e	48GB GDDR6	32GB HBM2e	48GB LPDDR5
显存带宽	1.6 TB/s	1.6 TB/s?	1.2 TB/s	1.84 TB/s	768 GB/s	1.6 TB/s	614.4 GB/s
GPU互联	392 GB/s	64 GB/s	128 GB/s	896 GB/s	240 GB/s	64 GB/s	64GB/s
TDP	400W	350W	300W	350W	450W	300W	250W
使用场景	训练、微调和推理场景						微调和推理场景
千P算力8卡 机台数 FP16稠密	**342台**	**683台**	**540台**	**540台**	**1311台**	**1024台**	**1366台**

2.3.2 中芯国际：先进制程领军者，国产算力底座

- **中芯国际是中国大陆技术最先进、规模最大的专业晶圆代工企业。**可向全球提供 0.35微米到FinFET多种技术节点的晶圆代工与技术服务。根据集邦咨询发布的25年Q1全球晶圆代工业报告，中芯国际位居全球第三位、中国大陆企业第一，也是国内先进制程龙头晶圆厂。
- 中芯国际目前无实控人，持股份额分散，最大股东持股比例不超过20%。**其子公司中芯南方主要负责先进制程工艺产线，**根据企查查，中芯南方第一大股东为中芯国际，持股38.52%，第二大股东为大基金二期（23.08%），随后分别为国家集成电路产投基金（14.56%）、上海集成电路产投基金（12.31%）。

中芯国际主营业务

中芯南方股权结构（截至25年05月）

股东名称	持股比例	认缴出资额
中芯国际控股有限公司	38.52%	25.04亿美元
国家集成电路产业投资基金（二期）	23.08%	15亿美元
国家集成电路产业投资基金	14.56%	9.47亿美元
上海集成电路产业投资基金	12.31%	8亿美元
上海集成电路产业投资基金（二期）	11.54%	7.5亿美元

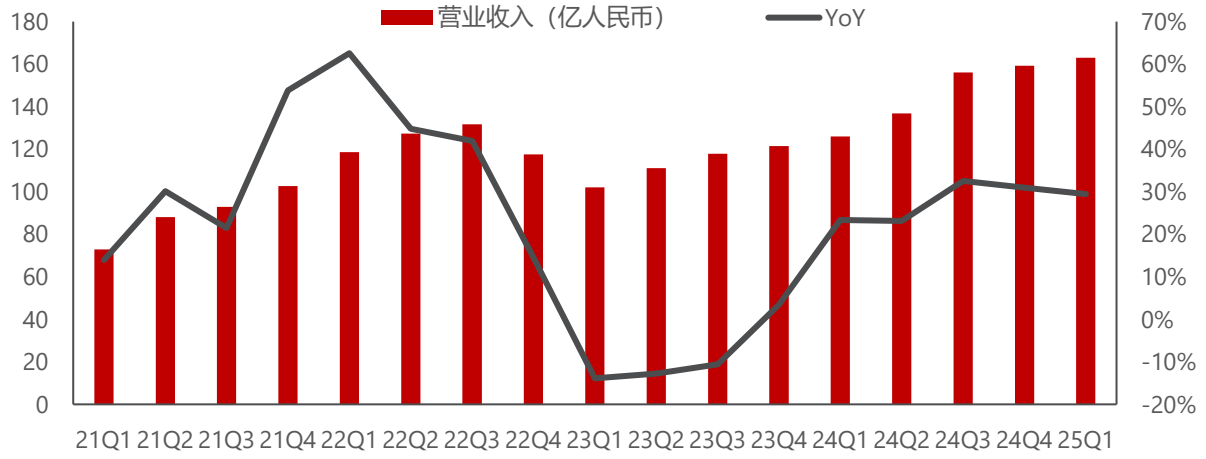
资料来源：中芯国际官网，民生证券研究院整理

资料来源：iFinD，企查查，民生证券研究院整理

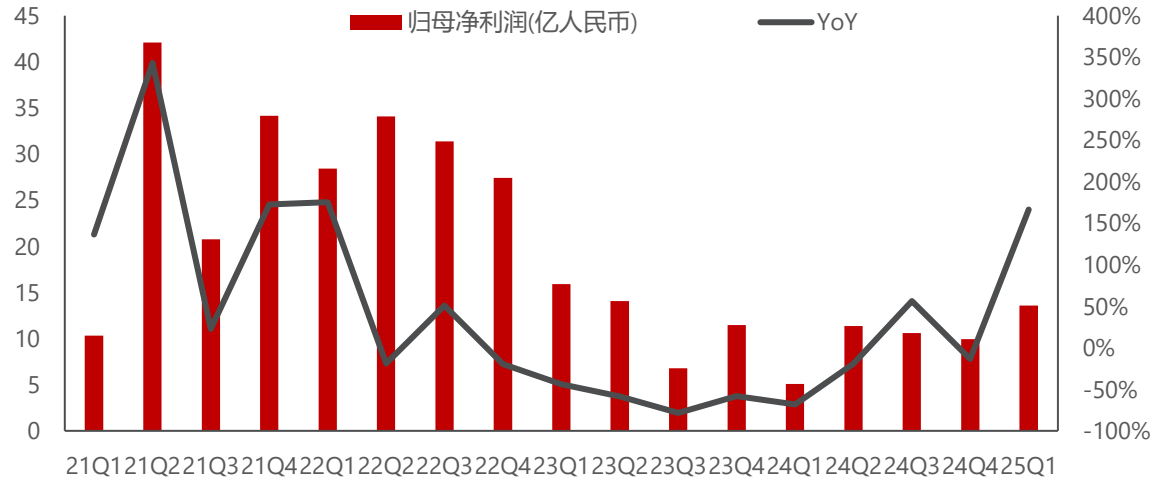
2.3.2 中芯国际：收入创历史新高，净利润显著改善

- **收入创历史新高，净利润显著改善：**2024 年中芯国际实现营收80.3 亿美元，同比+ 27%，创历史新高，实现归母净利润为4.93亿美元，同比-23.3%。营收增长主要受本土化制造需求带来的产业链的重新组合、客户市场份额的提升、产能规模的扩大、新增产能能够较快地完成验证并投入生产及国家刺激消费政策的拉动；2025年Q1中芯国际实现营收22.47亿美元，同比+29.4%，环比+1.8%，实现归母净利润为1.88亿美元，同比+166.5%，环比+36.7%。营收主要受益于国际形势变化引起的客户提拉出货，国内以旧换新、消费补贴等政策推动的大宗类产品的需求上升，以及工业与汽车产业的触底补货。
- **25年Q2指引：**展望Q2，中芯国际指引销售收入预计环比将下降 4~6%。出货数量预计相对稳健，平均销售单价预计下降；毛利率预计在18到20%之间，中芯国际通过降本增效来抵抗价格波动的压力，但Q2设备折旧继续上升，导致毛利率指引在Q1的基础上有所下降。

2021-2025中芯国际季度收入及增速



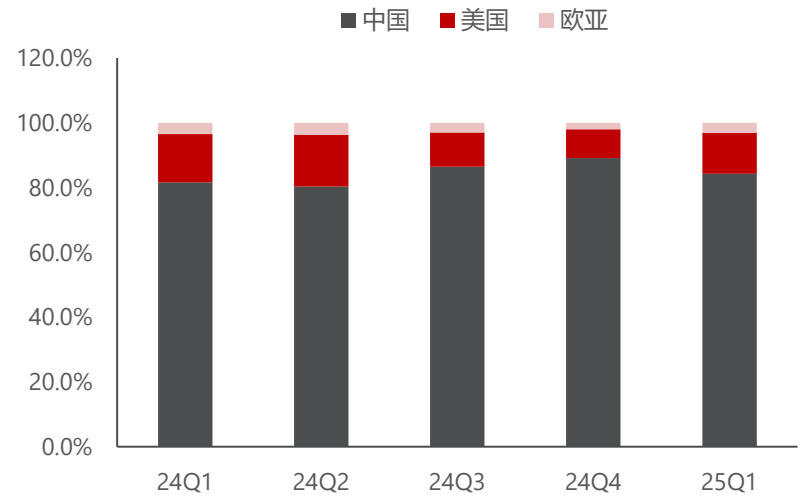
2021-2025中芯国际季度归母净利润及增速



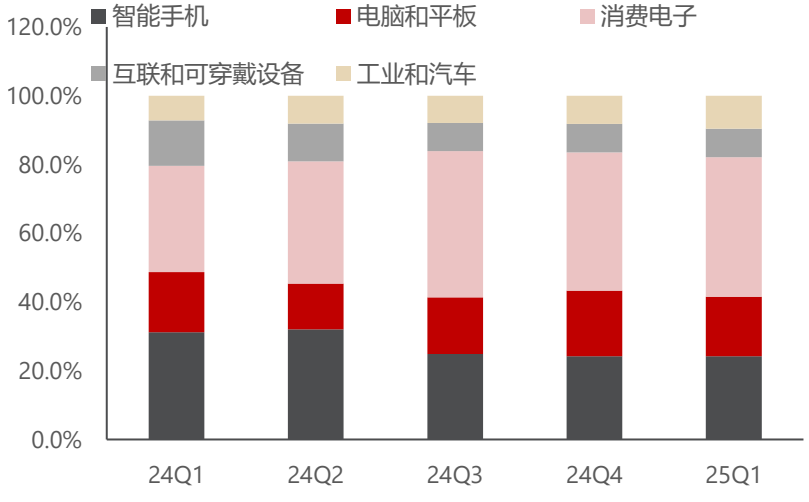
2.3.2 中芯国际：本土收入占比持续增长，下游消费类需求复苏

- **按地域拆分**：25年Q1，从金额来看，中国区营收保持稳定，占比升至84.3%，同比+2.7pct。美国与欧亚区占比分别下滑至12.6%（同比-2.3pct） 和3.1%（同比-0.4pct），主要受益于国际形势变化引起的客户提拉出货。
- **按应用领域拆分：** 25年Q1，智能手机、电脑与平板、消费电子、互联与可穿戴收入额相对稳定，占比分别为 24.2%（同比-7.0pct），17.3%（同比-0.2pct），40.6%（同比+9.7pct）和 8.3%（同比-4.9pct）。工业与汽车收入额环比增长超过两成，占比从 7.8%上升至 9.6%（同比+2.4pct）。尤其是汽车电子，得益于主要客户在汽车领域取得的进展和公司过去几年加大对汽车电子平台的投入和重点布局，在 BCD、CIS、MCU、域控制器等领域，与产业链紧密合作，公司车规产品的出货量稳步提升。
- **按尺寸占比拆分**： 25年Q1，12英寸产品占比提升至78.1%，同比+2.5pct， 8英寸产品占比21.9%，同比pct-2.5pct。8英寸、12 英寸晶圆收入环比+18%和+2%，主要受益于国内以旧换新、消费补贴等政策推动的大宗类产品的需求上升。

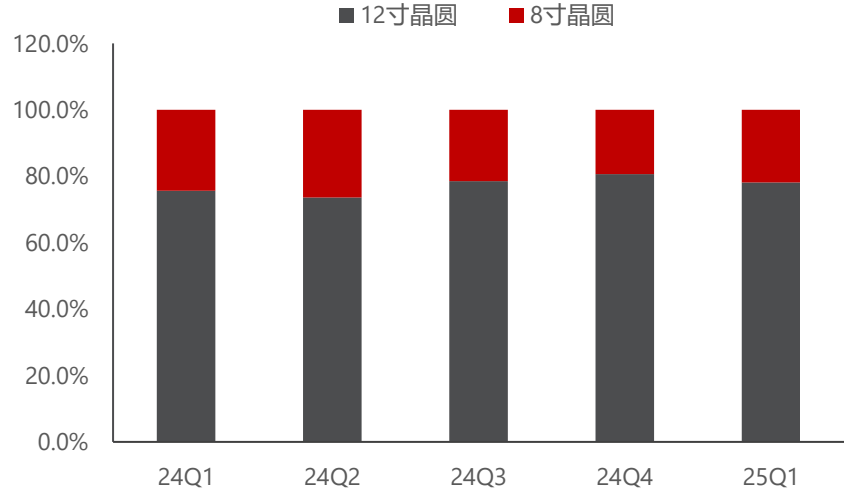
24Q1-25Q1中芯国际季度地域收入占比



24Q1-25Q1中芯国际季度应用收入占比



24Q1-25Q1中芯国际季度尺寸收入占比



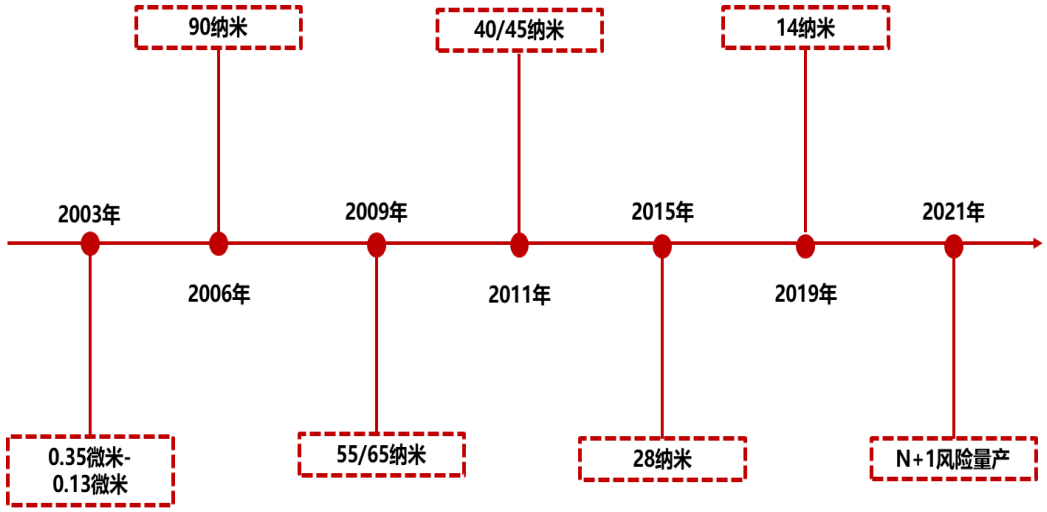
资料来源：iFinD，民生证券研究院整理

资料来源：iFinD，民生证券研究院整理
资料来源：iFinD，民生证券研究院整理

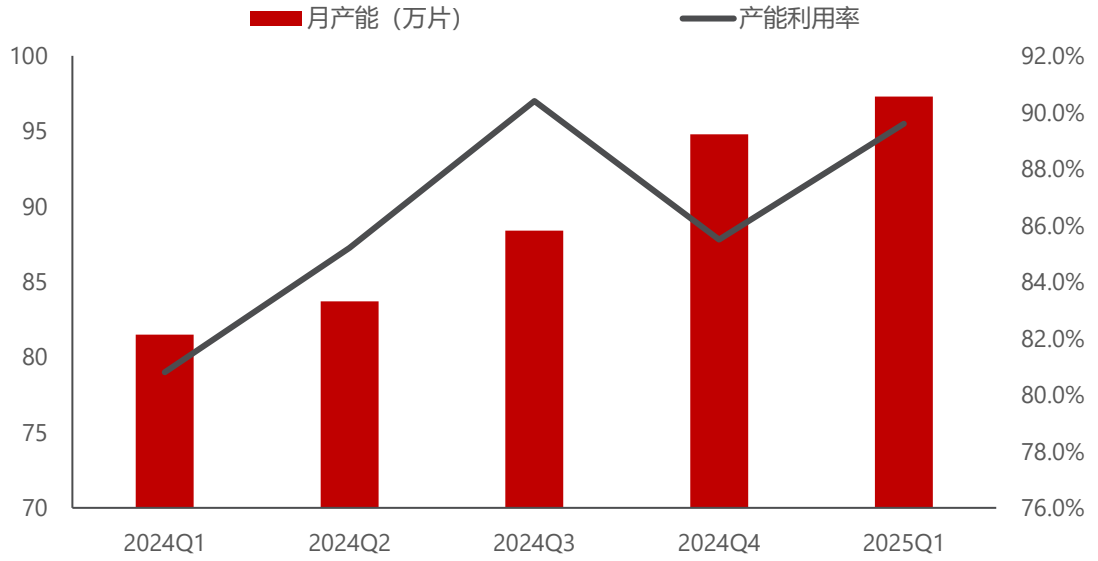
2.3.2 中芯国际：先进制程领军者，产能持续放量

- **在先进制程方面，中芯国际围绕FinFET路线持续演进，已逐步形成较为完整的代工能力体系。**2019年公司率先实现14nm FinFET的量产，并主要应用于AI与高性能低功耗计算等领域。公司进一步推出N+1工艺，于2020年完成测试及流片，第一次采用四重图形处理（SAQP）形成 Fin架构，并由自对准双重成像技术（SADP）形成 Dummy Gate。N+1工艺功耗较14nm下降57%，性能提升20%。最新的N+2制程也在加速突破中，制程有望等同于7nm工艺。尽管中芯国际在技术水平上仍与台积电存在一定差距，但通过对成熟制程的持续优化与重构，逐步推进先进工艺的国产替代。
- 截止至25年Q1，公司折合 8 英寸月产能达到 97.3万片，同比+19.4%/环比2.6%；产能利用率在新产能逐步释放的情况下达到 89.6%，同比+8.8pcts/环比+4.1pcts；25年Q1折合8英寸晶圆出货量为229.2万片，同比+27.7%/环比+15.1%。

中芯国际关键技术节点的量产时间



24Q1-25Q1中芯国际月度产能（折合8英寸，万片/月）及稼动率



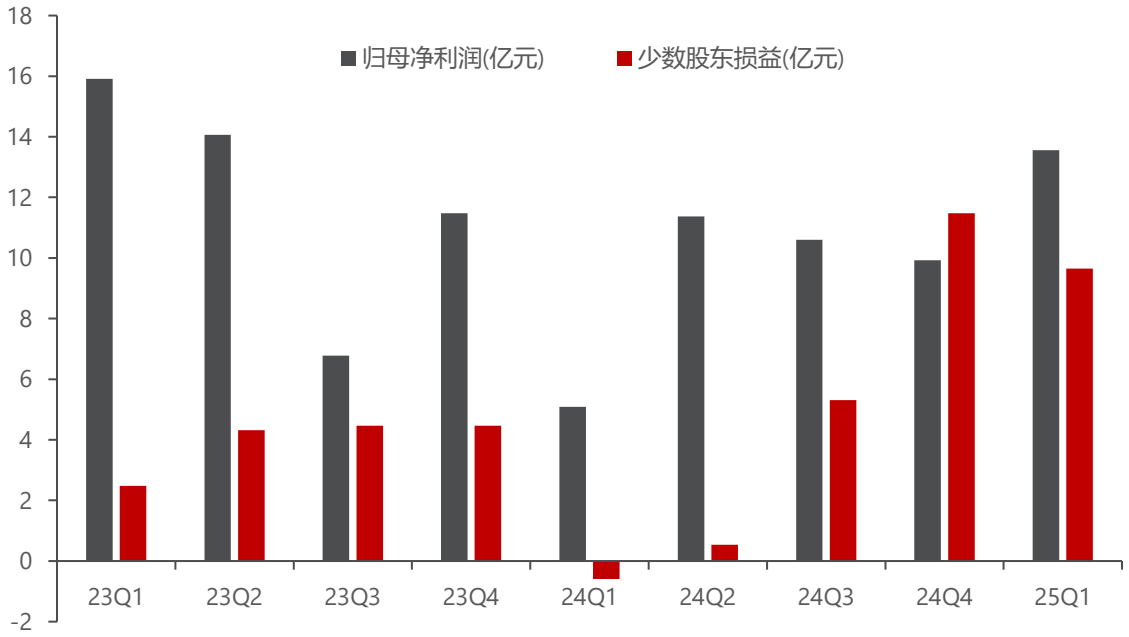
资料来源：iFinD，民生证券研究院整理

资料来源：公司公告，ifind，民生证券研究院整理

2.3.2 中芯国际：利润结构优化，归母占比持续提升

➢ **少数股东当期损益同比大幅改善，先进制程盈利逐步释放。**25年Q1中芯国际少数股东损益1.35亿美元，24年Q4 1.63亿美元，环比下滑17.12%，但较24年Q1 -0.08亿美元大幅改善并维持高位。其中，中芯国际持有中芯南方38.52%股权，剩余股份为中芯国际重要的少数股东权益,盈利释放凸显公司先进制程工艺进展明显。展望未来，在地缘政治推动算力自主可控的背景下，先进制程需求有望在Q2进一步放量，公司依托先进工艺发展核心逻辑，成长前景依旧向好。

中芯国际2023Q1-2025Q1归母净利润及少数股东损益



中芯国际少数股东权益构成

子公司名称	中芯国际持股比例（%）	少数股东持股（%）
中芯南方	38.52	61.48
中芯北方	51.00	49.00
中芯京城	54.04	45.96
中芯深圳	55.05	44.95
中芯东方	67.03	32.97

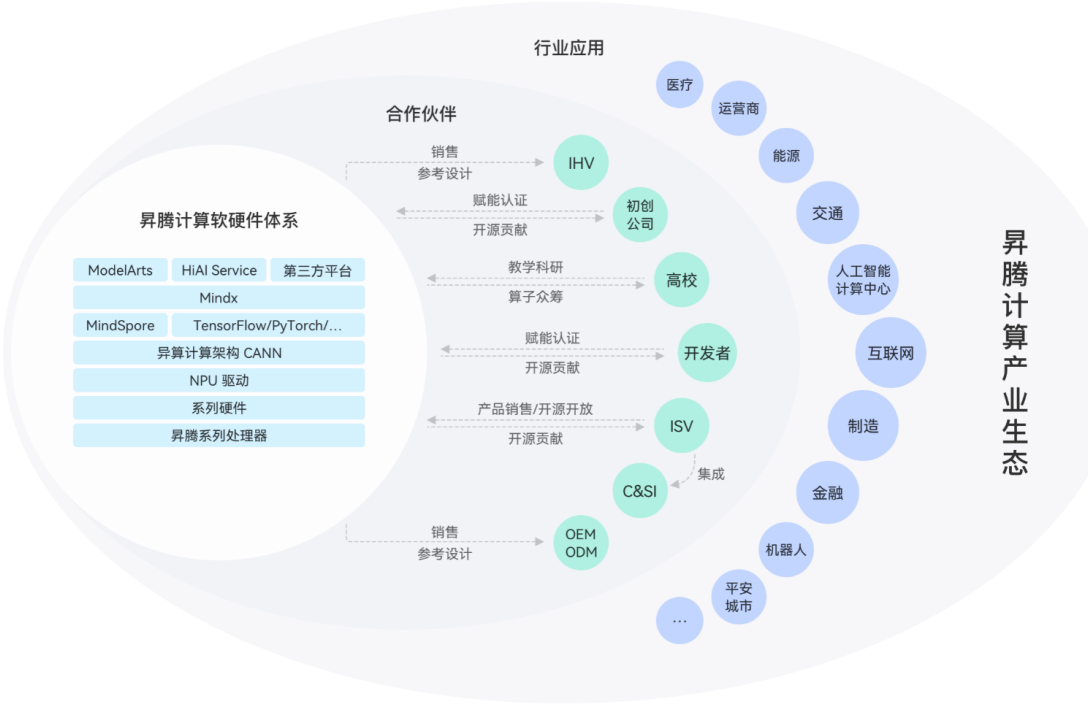
资料来源：iFinD，民生证券研究院整理

资料来源：iFinD，公司公告，民生证券研究院整理

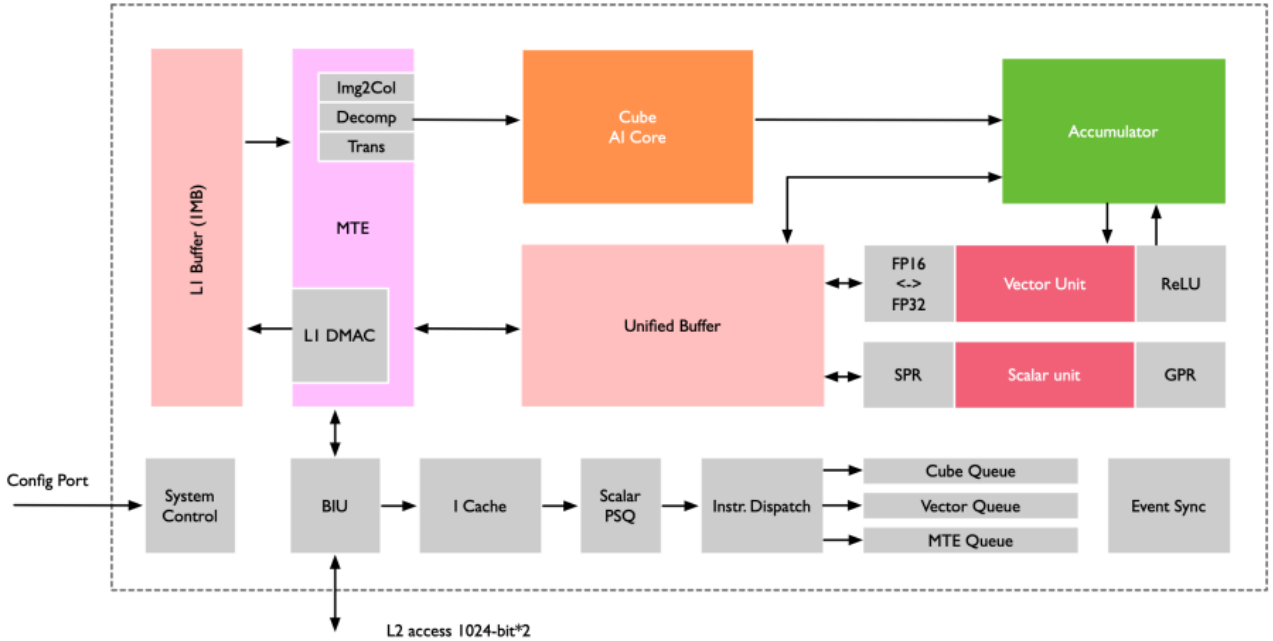
2.3.3 昇腾计算产业：以 910系列为基，构建全栈 AI 生态

- **昇腾计算产业**是基于昇腾系列（HUAWEI Ascend）处理器和基础软件构建的全栈AI计算基础设施、行业应用及服务。目前，昇腾计算产业已形成 **“昇腾芯片 + CANN 异构计算架构 + MindSpore AI 框架 + 行业应用解决方案”** 的全栈布局。作为这一产业的算力核心，昇腾 910 系列芯片具备强大的 AI 计算性能，用以**对标英伟达的GPU系列。**
- 昇腾910A于2019年发布，是华为昇腾910系列的初代产品，采用达芬奇架构，其半精度（FP16）算力**达到 256 Tera-FLOPS**，整数精度（INT8）算力达到 512 Tera-OPS；重要的是，达到规格算力**所需功耗仅 310W，明显低于设计规格的 350W。**

昇腾计算产业全景



达芬奇架构



2.3.3 昇腾910B & 910C：架构与工艺驱动的算力突围

- **昇腾910B**具备算力320 TFLOPS/FP16，搭载64GB HBM 内存，其性能表现追赶上了英伟达的A100芯片。通过自研AI推理引擎，潞晨科技实现昇腾910B与DeepSeek-R1模型的推理适配优化，在同等算力条件下，DeepSeek-R1系列模型在昇腾910B上的**推理延迟较H100 GPU仅相差5%**，吞吐量差距缩小至8%以内。
- **昇腾910C**采用7nm/N+2工艺，引入**双die封装设计**将两个910B处理器集成于单封装，显著优化计算密度并支持先进稀疏计算，**实现了算力的显著提升（800 TFLOPS/FP16）**。910C通过与DeepSeek模型的深度优化，实现**训练效率提升40%**，推理速度突破每秒3000次，进一步缩小与国际算力表现的差距。

昇腾910B与英伟达B200性能对比

	昇腾910B	英伟达B200
产品方向	加速机器学习等任务专为数据中心设计的高性能ai处理器	为需要最高级别计算性能和内存带宽的应用提供支持，如科研等
框架	DaVinci架构，面向人工智能应用的高效计算架构。	Blackwell架构，专为AI计算架构
晶体管数量	496亿	2080亿
制程工艺	7纳米制程工艺	4纳米制程工艺
算力	320Tera-FLOPS(FP16);640Tera-OPS(INT8)	40 Peta-FLOPS(FP4);4.5/9Peta-OPS(INT8);
功耗	310W	1000W
内存支持	64GBHBM内存,400GB/s内存带宽	192GB内存，达到了8TB/s内存带宽
灵活性	支持多种AI框架的灵活扩展如MindSpore	具备出色的灵活性，适用于多种高性能计算和AI应用
数据中心应用	支持高效应用，包括高吞吐量、低延迟和高可靠性场景	应用于数据中心，支持云计算、大数据处理等场景
智能网卡特性	高性能、低延迟的智能网卡特性，用于内部的数据传输	低延迟，适用于数据中心内部的数据传输

910C与910B性能对比

型号	昇腾910B	昇腾910C
制程	7nm	7nm
浮点FP16算力	320 TFLOPS	752 - 800 TFLOPS
整型Int8算力	640 TFLOPS	1504 TFLOPS
显存	64GB HBM2e	128GB HBM2e
功耗	310W	310W
GPU互联	400GB/S	392GB/S

资料来源：IT技术订阅，芯智讯，民生证券研究院整理

资料来源：IT技术订阅，半导体产业纵横，AI科技前哨战，开源大模型，民生证券研究院整理

2.3.3 昇腾910D & 920：引领下一代算力革新

- **昇腾910D，**作为华为昇腾系列的最新旗舰产品，采用达芬奇3.0架构，单芯片集成64个AI Core，**算力密度提升200%。**自研HBM3e显存通过3D堆叠技术，带宽达4TB/s，**超越H100的3.35TB/s。**采用硅光模块实现芯片间超高速互联，**延迟降低至纳秒级。**这种架构革新让昇腾910D的理论峰值算力达到1.2 PFLOP/s，**较H100的672 TFLOP/s提升78%。**更关键的是，其能效比优化至2.1 TFLOP/W，**较H100的1.7 TFLOP/W提升23%。**
- **昇腾920，**作为昇腾第三代AI处理器，昇腾920采用中芯国际6nm（N+3）工艺以及Davinci架构，晶体管密度达到120亿个，**较上一代910提升约50%。**在BERT-large模型训练场景下，千卡集群**吞吐量提升可达30%。**峰值算力达到1800TOPS（INT8）/900TFLOPS（BF16），内存带宽达4000GB/s。与英伟达H20对比，920C在BF16**算力上提升约37%，内存带宽增加33%，能效比（TOPS/W）优化约42%。**

910D参数

指标	参数	技术亮点
架构	达芬奇架构 3.0	采用 3D Cube 技术，单芯片集成 64 个 AI Core，算力密度提升 200%
生产工艺	7nm	由中芯国际制造
带宽	4TB/s	自研HBM3e显存，通过3D堆叠技术，超越H100的3.35TB/s
理论峰值算力	1.2 PFLOP/s	较 H100 的 672 TFLOP/s 提升 78%
能效比	2.1 TFLOP/W	较 H100 的 1.7 TFLOP/W 提升 23%

昇腾920与英伟达H20训练任务表现对比

模型	昇腾920（单卡）	英伟达H20（单卡）	性能提升
ResNet-50	2800 images/s	1550 images/s	80.6%
BERT-large	125 tokens/s	89 tokens/s	40.4%
GPT-3 175B	1.2 tokens/s	0.8 tokens/s	50%

2.3.3 华为384架构领衔，910C蓄势待发

- 4月10日华为云宣布推出**CloudMartix 384架构**， CloudMatrix 384 超节点基于 384 颗昇腾 910C 芯片构建。通过**五倍芯片堆叠配置**下，BF16性能达到300 PFLOPS，**约为GB200 NVL72的1.7倍**；HBM总容量达49.2TB，是GB200的3.6倍；在功耗上，CloudMatrix 384总功率上达到559.4kW，是GB200 NVL72的近四倍功耗。能耗设计上的取舍，换来了在大规模训练及推理场景下显著的性能释放。
- 据华为介绍，昇腾384超节点架构**最适合MoE AI模型**。华为公布的基准测试结果显示，超节点384在处理Meta的LLama 3等密集型AI模型时，单卡性能达到132 tokens/秒（TPS），是传统集群的2.5倍。对于Qwen以及DeepSeek等通信密集型多模态及MoE模型，华为架构的单卡性能达到**600至750 TPS，可以达到3倍以上的提升。**

CloudMatrix 384 架构示览图

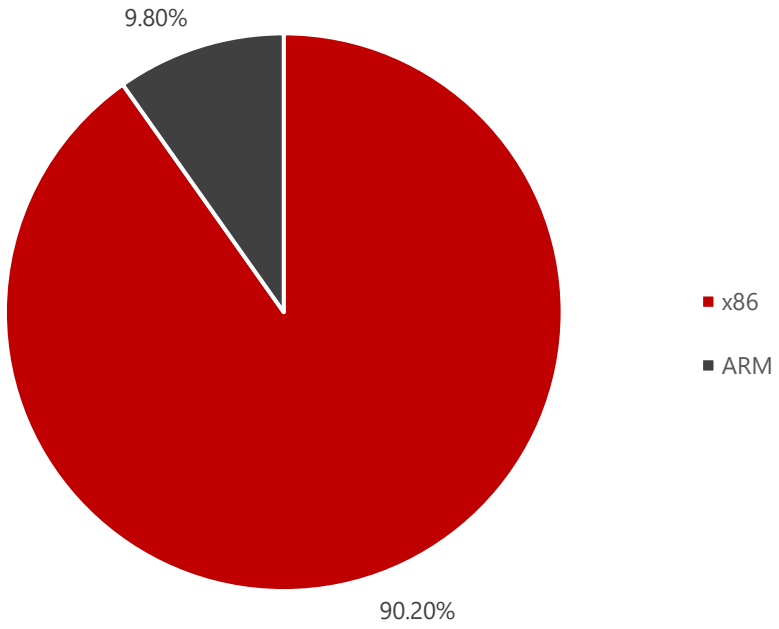
CloudMatrix 384 与 NVL 72 性能对比

标准	单位	Nvidia GB200 NVL72	CloudMatrix CM384	Huawei vs Nvidia
BF16 密集型每秒千万亿次浮点运算能力	PFLOPS	180	300	1.7x
高带宽内存容量	TB	13.8	49.2	3.6x
高带宽内存带宽	TB/s	5.76	12.29	2.1x
向上扩展带宽	Gb/s uni - di	64800	134400	2.1x
向上扩展域规模（GPU 数量）	GPUs	72	384	5.3x
向外扩展带宽	Gb/s uni - di	28800	153600	5.3x
系统总功耗	W	145000	559378	3.9x
每 FLOP BF16 运算功耗	W/TFLOP	0.81	1.87	2.3x
每 TB/s 内存带宽功耗	W per TB/s	251.7	455.2	1.8x
每 TB 内存容量功耗	kW/TB	10.5	11.4	1.1x

2.3.3 海光：DPU+CPU双引擎驱动，自研C86升级国产性能

- 海光信息是国内领先的高端计算芯片设计企业，围绕CPU和DCU产品双重布局。公司的CPU产品早年获得AMD在高端处理器的技术授权及相关技术支持，并基于X86架构研发。而x86服务器则占据服务器市场中的首要位置，据《中国算力发展报告（2024）》，2023年x86服务器在我国CPU服务器市场份额占比达到**90.20%**。在2022年国产服务器领域排名中，海光以**约53.6%的市场份额**持续领跑国产CPU产业。 而“海光三号”作为海光信息最主要产品之一，在参数上处于**国产领先地位，性能表现优秀。**
- 海光已全面掌握授权技术的整套源代码，并成功消化吸收x86相关全套技术，实现了独立的CPU设计与迭代开发。海光在微体系结构层面持续推进自主创新，基于完整的x86指令集源码**实现C86架构的国产化研发**，显著提升产品在计算性能、总线带宽、内存频率、安全性及可扩展性等方面的核心指标，持续强化产品竞争力。同时，海光自研C86系列处理器也保持着**每代至少15%-30%的性能提升。**

2023年我国CPU服务器市场份额分布



海光第三代CPU产品参数对比

	Intel	AMD	海光	兆芯	海思	飞腾	龙芯	申威
产品型号	Xeon6980P	EPYC9965	海光7390	KH-4000/32	鲲鹏920-7260	S5000C-64	3D5000	申威1621
指令集	x86	x86	x86	x86	ARM	ARM	LoongArch	SW_64
核心数	128	192	32	32	64	64	32	16
超线程	256	384	64	不支持	不支持	不支持	不支持	不支持
主频（GHz）	2.0	2.25	2.7	2.0	2.6	2.1	2.0	2.0
内存类型	DDR5	DDR5	DDR4	DDR4	DDR4	DDR5	DDR4	DDR3
内存通道数	12	12	8	8	8	8	8	8
最高内存频率（MHz）	6400	6000	3200	3200	2933	4400	3200	2133
PCIe版本	5.0	5.0	4.0	3.0	4.0	5.0	/	3.0
PCIe通道数	96	128	128	128	40	96	/	16

资料来源：Gartner，中国信息通信研究院，民生证券研究院

资料来源：各厂商官网及招股说明书，中关村在线，民生证券研究院整理

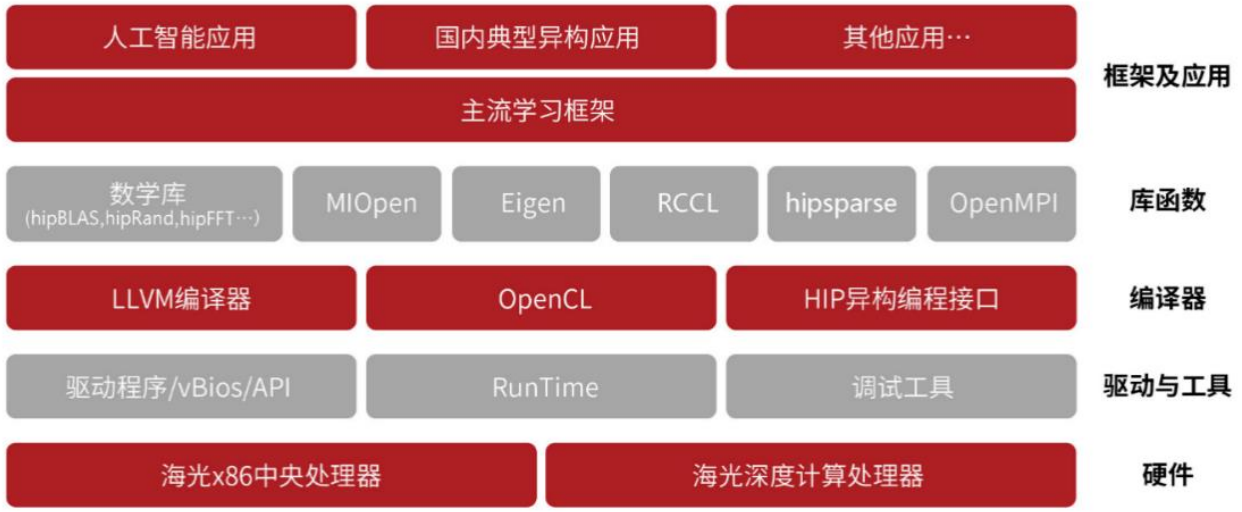
2.3.3 海光：GPGPU 架构赋能，国产化算力体系加速成型

- 海光信息的DCU产品以**GPGPU架构**为基础，兼容“类CUDA”通用环境，面向数据中心和云计算场景，随着AI算力时代的到来，DPU产品发展前景广阔。公司目前的主力产品为深算二号 K100，相较于前代产品，**整体性能提升超过100%。**针对AI领域对高算力的特殊需求，公司基于深算二号架构推出K100 AI版。在BF16/FP16（半精度）模式下峰值为**192 TFLOPS，较K100提升一倍，**而FP32通用算力**提升4倍达到98 TFLOPS，性能对标NVIDIA H20**，约为A100性能的60%。
- 自2月2日DeepSeek V3和R1模型与海光信息DCU达成适配，海光信息的硬件场景优势进一步显现。DeepSeek的MTP（多令牌预测技术）技术优化数据处理流程，使海光DCU在单位时间内**处理更大规模文本数据。**在边缘计算与物联网融合的AI算力场景下，能耗与体积约束进一步收紧，海光DCU凭借轻量化设计、优化后的**低功耗特性及强化的学习兼容能力**，在这一领域具备更强市场竞争力。

海光K100与K100 AI版性能对比

规格参数分类		K100	K100_AI
芯片计算核心		120	-
性能指标	FP64	24.5 TFLOPS	-
	FP32	24.5 TFLOPS	24.5TFLOPS/ Tensor 98 TFLOPS
	FP16	100 TFLOPS*	Tensor200 TFLOPS
	INT8	200 TOPS*	400 TOPS
显存		64GB	64GB
PCIe 接口		PCIe 4.0 x16	PCIe 5.0 x16
TDP		300W	350 - 400W
尺寸规格		全高全长双宽	全高全长双宽，采用稀疏技术

海光完整软件栈系统



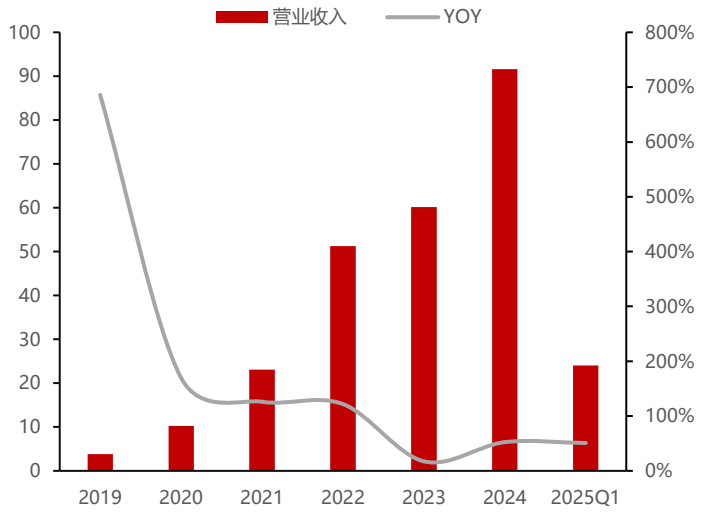
资料来源：AI时代的技术服务商，民生证券研究院整理

资料来源：海光官网，民生证券研究院整理

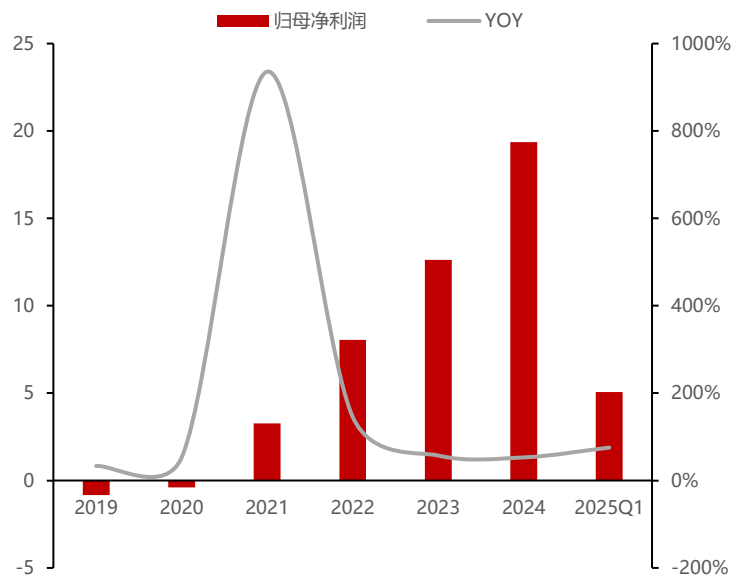
2.3.3 海光：营收利润双升，国产化与大模型驱动业绩突破

- 2025年Q1，海光实现营业收入24亿元，**同比增长50.76%**；归母净利润为5.06亿元，**同比增长75.33%。**营收高增受益于**国产化进程及大模型中的快速适配。**同时，海光Q1研发投入达7.64亿元，同比增长16.26%，**连续5年维持上涨趋势**，体现出海光在技术迭代和性能突破上的持续投入，通过构筑产品护城河夯实长期竞争力。海光25Q1合同负债32.37亿元，较去年末大增23.33亿元，**订单需求持续旺盛**，后续业绩释放具备较强增长动能。
- 海光盈利能力持续增强，**毛利率呈上升趋势**，由2019年的37.31%提升至2024年的63.72%，直到25年Q1海光销售毛利率依旧维持在61.69%。受益于国内算力基础设施投入加速及DeepSeek等国产大模型应用落地，**DCU出货有望提速。**在中美政策摩擦加剧背景下，具备自主可控属性的算力芯片需求将进一步提升。海光CPU+DCU双赛道驱动需求加速释放，叠加海光2024年高端处理器领域业绩表现亮眼，预计海光2025-2027年将**保持高端处理器业务的高增长态势。**

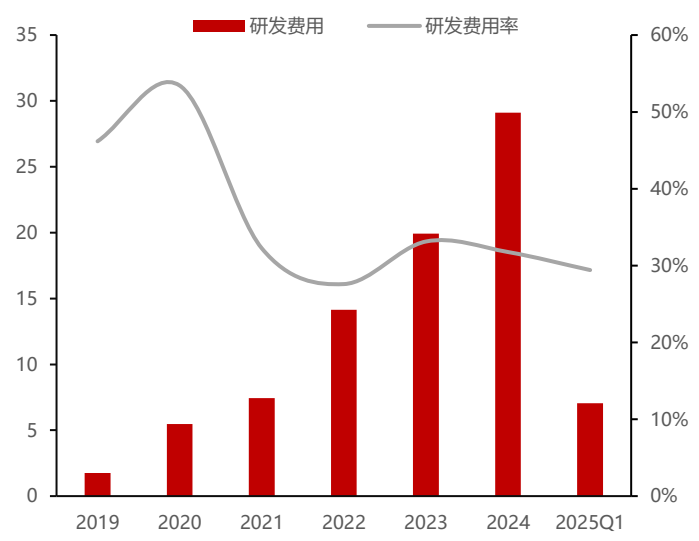
海光2019-2025Q1营收及增速



海光2019-2025Q1归母净利润及增速



海光2019-2025Q1研发费用（亿元）及研发费用率



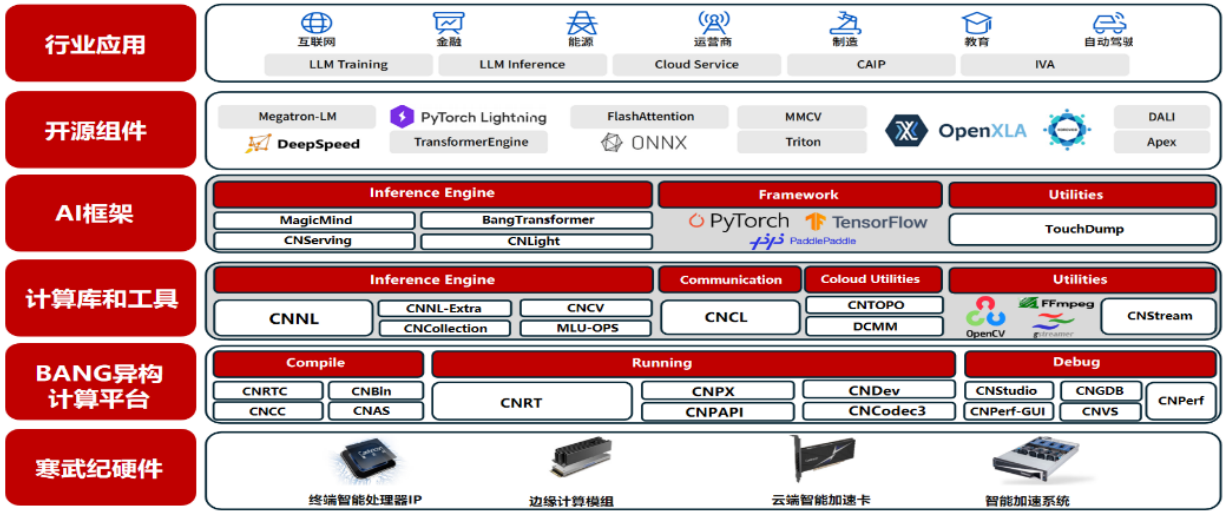
资料来源：iFinD，民生证券研究院整理
资料来源：iFinD，民生证券研究院整理

资料来源：iFinD，民生证券研究院整理

2.3.3 寒武纪：软硬协同发力，推进高性能AI芯片国产化

- **寒武纪**是全球知名的智能芯片企业，聚焦芯片技术突破与商业化落地，具备从云到边、从硬件到系统软件的完整能力。当前全新一代云端AI训练芯片为思元590，采用ASIC架构设计，具备低功耗、高效率，**单卡算力达512 TOPS，对标A100。**在存储架构方面，思元590的MTP与DDR/HBM间的带宽大幅优化，支持L2 Cache读写缓存，MTP Cluster的访存带宽**较上代提升4倍**，支持大规模数据处理。思元590已部署在智能安防、自动驾驶等多元应用场景。值得关注的是，搭载思元590的阿里云智能计算集群在Llama-3训练任务中**已实现训练成本下降40%**，进一步验证其商业应用潜力。
- 在业绩层面，公司积极把握国产算力需求所带来的市场机遇。25年Q1公司实现营收11.11亿元，**同比大增长4230.22%**；归母净利润为3.55亿元，实现扭亏为盈；25年Q1在存货端27.55亿元，**较24年末大增9.81亿，未来良好放量支撑。**细分行业方面，公司在互联网、运营商产品性能获得认可，金融行业完成大模型升级，持续推动AI算力在银行、保险、基金等业务场景中的深度落地。

寒武纪基础软件系统平台



寒武纪产品策略

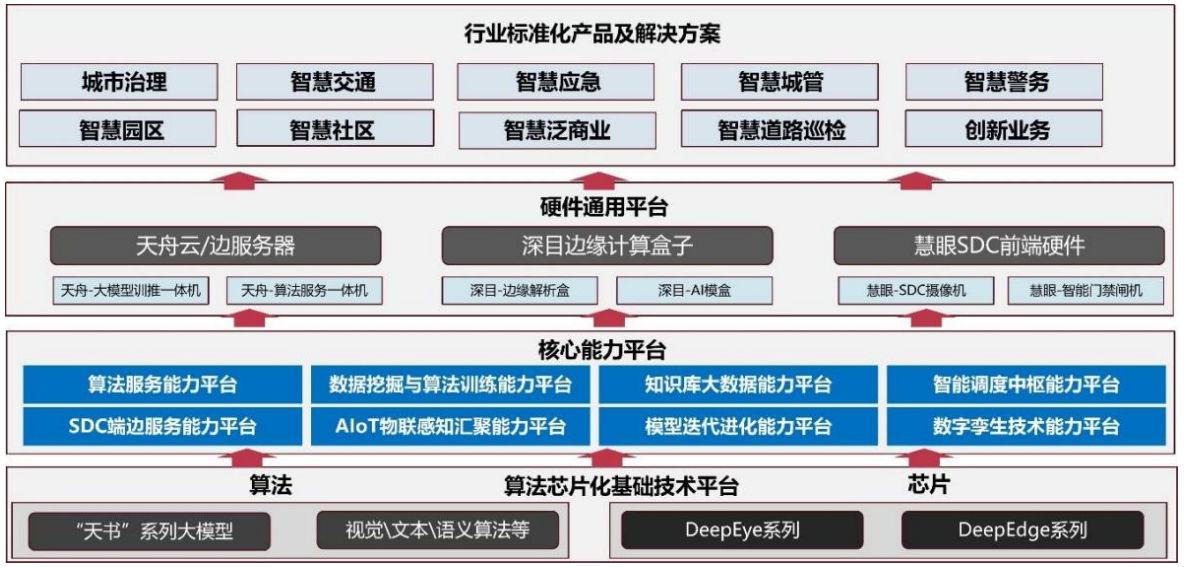
2.3.3 云天励飞：算法芯片化加速推进，边缘AI应用深化突破

- 云天励飞长期深耕算法与芯片双平台架构，围绕算法芯片化和端云协同的技术路径，已在数字城市等核心应用场景中实现落地突破。公司自ChatGPT发布以来，持续加码大模型训练及推理方向，在软硬协同能力、产品落地广度及算力平台演进方面取得积极进展。公司在首款自研AI芯片DeepEye 100推出后，陆续推出DeepEdge 10系列SoC芯片，**性能与适配能力持续提升。**最新产品DeepEdge 10 Max基于自研的NNP400T处理器打造，浮点与整数计算性能上显著提升，适配主流深度学习网络结构，面向多样化算力场景，**满足不同行业在边缘部署中的高效推理需求。**
- 财务端有望释放进一步的业绩潜力。公司2025年Q1实现营业收入2.64亿元，**同比+168.23%**；归母净利润为-0.86亿元，仍处于亏损状态，但亏损幅度**同环比均有显著改善。**随着大模型场景化需求持续释放，公司在训练与推理一体化的产品链条将逐步完善。

云天励飞DeepEdge 10系列性能参数

性能	DeepEdge 10	DeepEdge 10 C	DeepEdge 10 Max
MCU 核	单核 MCU 处理器，450 DMIPS	单核 MCU 处理器，450 DMIPS	4个单核 MCU 处理器，1.8K DMIPS；
异构计算平台	NNP400T 8 TOPS (INT16), 2 TFLOPS (FP16) 的	NNP400T 4 TOPS (INT16)/ 1.5 TFLOPS (FP16)	NNP400T; 32 TOPS (INT16), 8 TFLOPS (FP16)
图形处理器	OpenGL ES 3.1/3.0/2.0/1.1	支持 3D 图形处理单元	OpenGL ES 3.1/3.0/2.0/1.1
存储器接口	LPDDR4@3733Mbps/ LPDDR4X@3733Mbps/ DDR4@3200Mbps	LPDDR4@3200Mbps / LPDDR4X@3200Mbp	LPDDR4@3733Mbps / LPDDR4X@3733Mbps；
接口	RGMII、RMII	PCIe 3.0	PCIe 3.0
功耗	10W	-	40W
封装	FCBGA 封装: 23nm * 23nm	FCCSP 封装：15mm x 15mm	FCBGA 封装：40mm * 40mm/0.8mm间隙

云飞励天下游场景解决方案



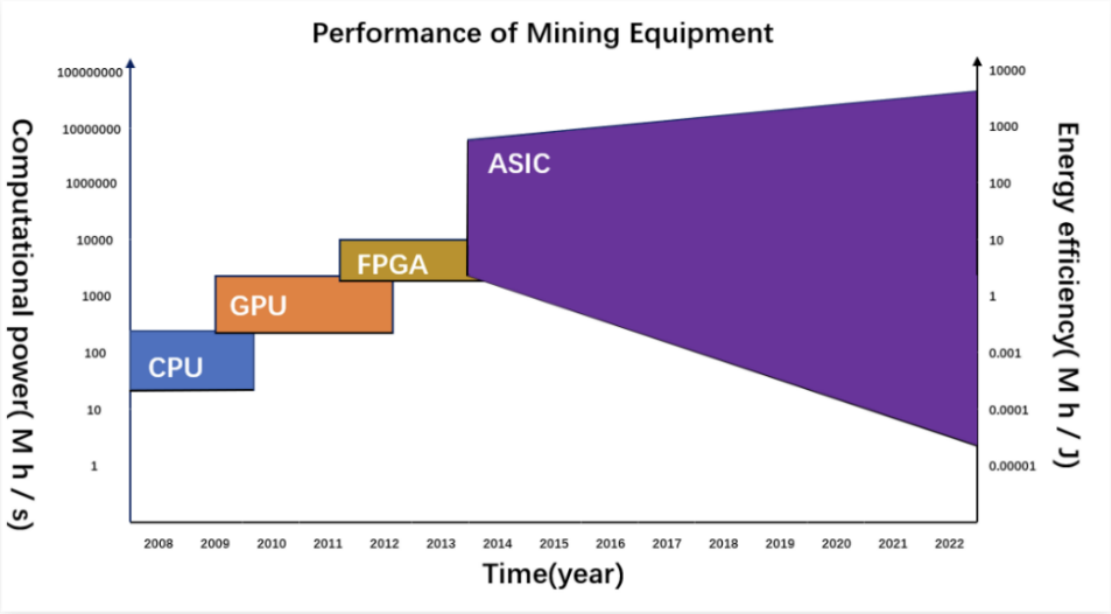
资料来源：云天励飞官网，民生证券研究院整理

资料来源：云天励飞年报，民生证券研究院

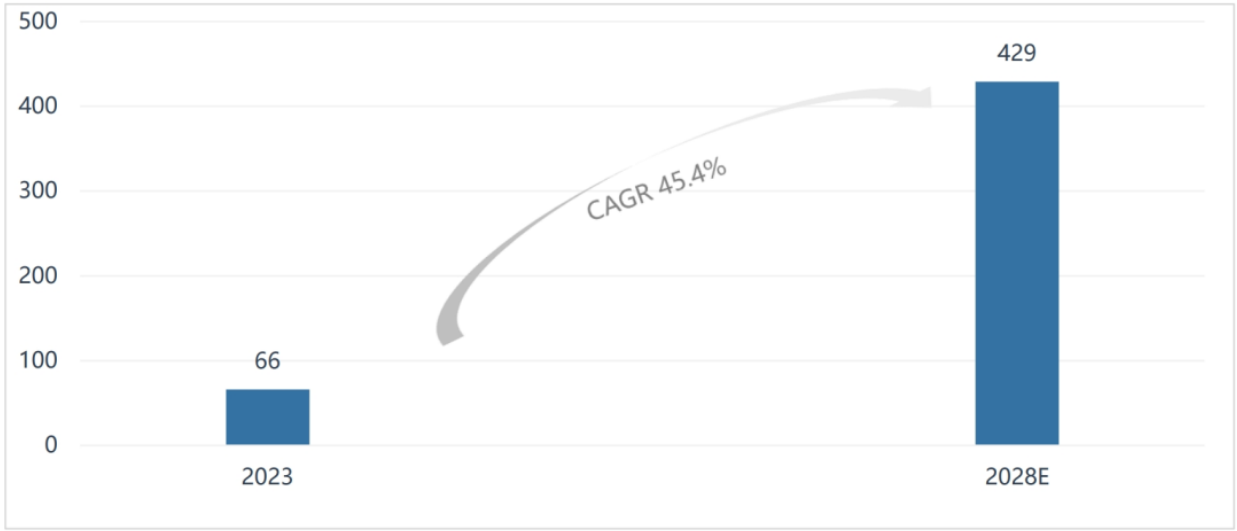
2.3.4 ASIC：云厂商的曙光，挑战GPU垄断

- 目前在AI加速卡领域英伟达仍然占据主要供应商的地位，但云商自研ASIC的比例正在逐步提升，将成为未来AI芯片增量最核心的来源。一方面，云商成本显著优于向英伟达等商业公司外采，4Q24英伟达**毛利率已达到73.0%**。另一方面，云商自研ASIC更加灵活，可以根据自身的模型训练和推理需求，进行AI芯片和服务器架构的设计。摩根士丹利预计，AI ASIC市场规模将**从2024年的120亿美元增长至2027年的300亿美元，年复合增长率达到34%。**
- TrendForce的最新研究报告指出，随着人工智能服务器需求的迅猛增长，美国主要的云计算服务提供商（CSP）正加快内部开发专用集成电路（ASIC）芯片的步伐，**平均每1至2年便推出新一代产品。**在中国，人工智能服务器市场正逐步适应美国自2025年4月起实施的新出口管制政策。据TrendForce预测，这些措施将导致2025年进口芯片（如NVIDIA和AMD产品）的市场份额**从2024年的63%下降至约42%。**与此同时，在积极推动国产人工智能处理器的政策扶持下，预计中国本土芯片制造商的市场份额将**提升至40%，与进口芯片的市场份额几乎持平。**

不同处理器的计算能效对比



2023-2028E全球ASIC市场规模（亿美元）



资料来源：电子工程世界，民生证券研究院

资料来源：Wind，Marvell，来觅数据，民生证券研究院

2.3.4 聚焦芯原与翱捷，探寻国产 ASIC 发展路径

- **芯原股份**是国内领先的半导体IP供应商，拥有包括GPUIP、NPU IP在内的多款处理器IP，具备为客户提供一站式芯片定制服务的能力。公司芯片设计能力覆盖14nm/10nm/7nm/5nm FinFET和28nm/22nmFD-SOI。2023年3月，蓝洋智能发布与芯原股份合作打造的基于Chiplet架构的高性能AI芯片，其中CC8400在提供强大算力的同时，还可优化面积和功耗；VIP9400支持Transformer模型，能够为数据中心和汽车应用提供强大的AI算力；VC8000D具有高吞吐量、多格式等特性，可用于视频内容分析。目前，内置芯原GPU IP的芯片在全球范围内出货**近20亿颗**，芯原NPU IP已被82家客户用于其142款人工智能芯片中，在全球范围内出货超过1亿颗。
- **翱捷科技**是国内少数同时完成“**5G+AI**”技术突破的企业，专注于无线通信芯片领域。公司掌握2G/3G/4G/5G全制式蜂窝基带芯片及多协议非蜂窝物联网芯片设计能力，相关客户涵盖国家大型电网、中兴通讯、美的集团、Hitachi、360、TP-Link等知名企业。翱捷科技具备提供**超大规模高速SoC芯片定制服务能力**，为人工智能、互联网、智能手机等不同行业内的头部企业提供芯片定制服务。

GPU与几种ASIC芯片架构特点对比

特性	GPU	LPU	TPU	NPU
核心类型	众核(数千小核心)	顺序处理单元	张量核心(专用矩阵运算单元)	神经元模拟电路(突触权重集成)
并行性	高并行(SIMD架构)	低(序列化处理)	中等(任务级并行)	高(数据驱动并行)
能效比	较低(功耗高)	中等(专注语言任务)	高(专为张量优化)	极高(存储计算一体化)
灵活性	高(通用并行计算)	低(仅限语言任务)	中(绑定TensorFlow生态)	低(专用神经网络)

芯原股份和翱捷科技ASIC研发实力一览

	芯原股份		翱捷科技	
	2023年	2024年	2023年	2024年
IP 授权收入（亿元）	7.65	7.36	1.23	0.35
ASIC / 芯片定制收入（亿元）	15.64	15.81	2.26	3.36
研发人员规模（人）	1662	1800	1135	1138
研发团队背景	Broadcom、Marvell 等		Marvell等	
IP 能力	拥有 GPU、NPU、VPU 等全面的芯片 IP 布局，AI 储备第一梯队，服务于多家全球知名企业		专注无线通信芯片领域，ISP、高速通信接口 IP、射频 IP 等能力领先	
ASIC 能力	多工艺节点芯片设计能力，多家客户量产经验。为新能源车企提供 5nm 的自动驾驶芯片；为蓝洋智能设计基于 Chiplet 的 AI 芯片		超大规模芯片量产经验，单颗芯片晶体管数量高达 177 亿	

资料来源：电子工程世界，民生证券研究院整理

资料来源：芯原股份官网，翱捷科技官网，上海证券报，经艺求精，民生证券研究院整理

03 AI 终端：技术浪潮下的硬件重塑进行时

3.1

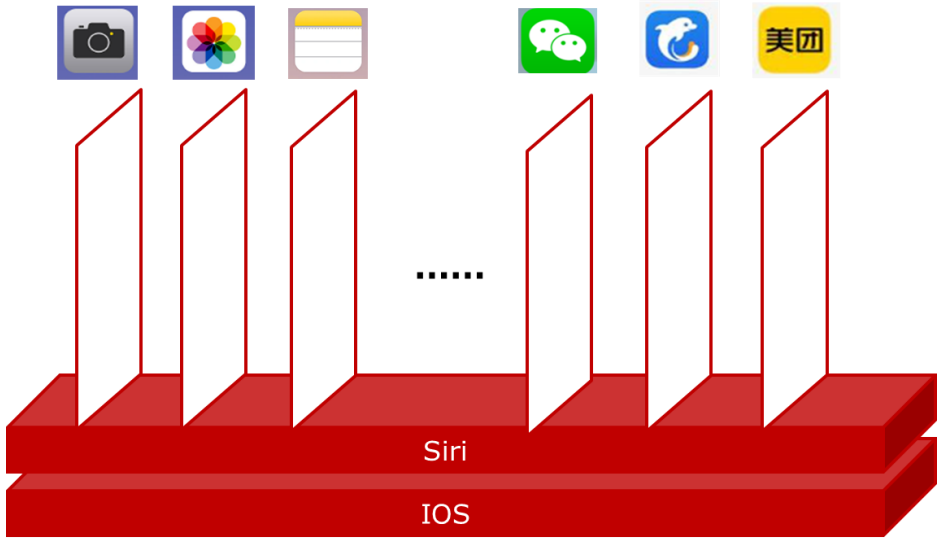
CONTENTS

目录

3.1.1 苹果AI布局：Apple Intelligence持续迭代，AI Siri和国行版再延期

- **苹果在WWDC 2024上发布Apple Intelligence：**Apple Intelligence对Siri进行了增强，让各个App向下通过Siri互通，Siri能在App中执行操作，借助App Intents和App Entities，实现互联互通。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。**截至目前，Apple Intelligence与第三方App交互的能力尚未落地，WWDC 2025仍未推出AI Siri和国行版Apple Intelligence，语言适配方面，预计将在今年年底前支持以下语言：丹麦语、荷兰语、挪威语、葡萄牙语、瑞典语、土耳其语、繁体中文和越南语。**

图：Apple Intelligence通过Siri打通和各个App的交互



图：屏幕感知功能

图：写作工具的摘要功能

图： Create a Memory Movie功能

3.1.1 苹果AI布局：Apple Intelligence持续迭代，AI Siri和国行版再延期

➢ 2024年12月6日，苹果推出iOS 18.2 Beta版，**Siri接入ChatGPT，Apple Intelligence在图像功能上进行了显著升级，同时iPhone 16用户可体验全新的视觉智能（Visual Intelligence）功能。**同年12月12日，苹果发布iOS 18.2 正式版。

表：Apple Intelligence功能简介

功能	简介	18.1 推出与否	18.2 Beta版 推出与否
Writing Tools	帮助用户对文本进行改写、校对和摘要。使用任何标准UI框架来呈现文本字段，App都将自动兼容Writing Tools功能。利用TextView委托API，用户可以对Writing Tools处于活跃状态时的App行为进行自定义，如在Apple Intelligence处理文本时暂停同步以避免冲突。	√	Writing Tools**新增“Compose”按钮**，可生成与所选主题相关的文本，并**新增“描述您的更改”选项**，从而自由修改文本的语气或内容，例如将电子邮件转换为诗歌等。
Image Playground	用户可利用其在“信息”、“备忘录”、“Keynote讲演”、“Pages文稿”等App中创作图像，开发者通过Image Playground API可将这一功能引入自己的App中，图像创作在设备端进行，无需开发和托管模型。	×	√
Genmoji	用户可利用其生成新的表情符号，Genmoji以内嵌图像的形式表示，使用标准UI框架呈现的文本字段及AttributedString均可支持Genmoji。	×	√
Siri & App Intents	苹果使用Apple Intelligence对Siri进行了增强，从而使得Siri与系统的结合更加自然、深入和个性化。**开发者可利用不同领域的预定义和预训练App Intents，让Siri能在App中执行操作**，并可让App的操作在“聚焦”、“快捷指令”App和“控制中心”等位置更容易被发现。使用SiriKit的App将自动获得Siri增强的对话能力，**借助App Entities，Siri可以了解App中的内容，并从系统任何位置为用户提供App中的信息**。	推出部分功能，如新版Siri能够更好地理解用户，拥有更强的上下文能力和新外观及唤醒动画等；但屏幕感知等功能尚未推出。	Siri集成了**ChatGPT以及Visual Intelligence（视觉智能）功能**，利用iPhone 16 的相机控制功能，帮助用户了解周围事物。

3.1.1 苹果新品及发布节奏展望

- **iPhone 17系列将于2025年秋季新品发布会推出，其创新主要包括：**几款机型的摄像头或将大幅升级（部分升级为48MP长焦镜头+24MP前置摄像头），搭载A19系列芯片，拥有更大的内存，以及全新超薄Air（Slim）型号。
- 据Apple Roadmap，苹果将于2026年推出折叠iPhone，该折叠iPhone外屏尺寸为6英寸，内屏尺寸为8英寸，还将使用低延迟、低功耗的LLW DRAM。iPhone 18 Pro机型将配备显示屏下的Face ID。此外，苹果还将推出OLED面板的10.9英寸的iPad Air平板电脑和配备8.4英寸的OLED显示屏的iPad mini，以及14和16英寸OLED面板的MacBook。2027年，苹果有望发布AR眼镜或可折叠iPad产品。

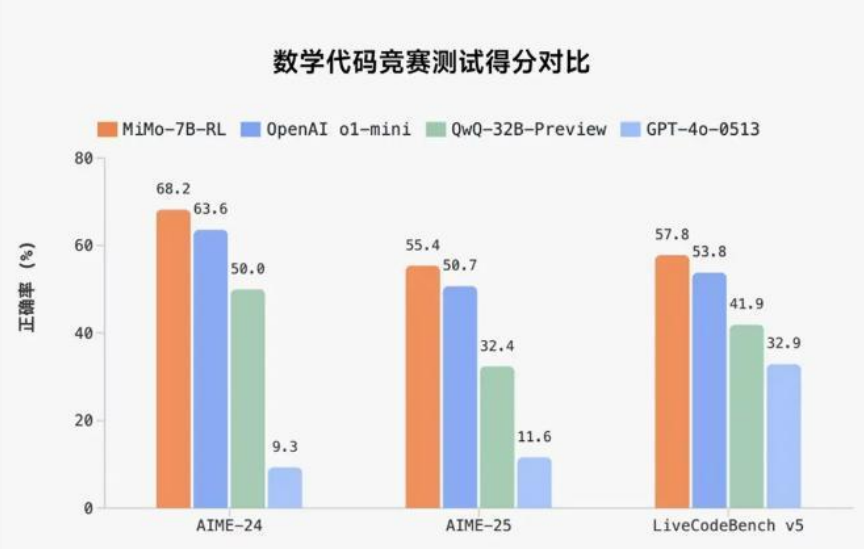
表：iPhone 16/17系列配置对比

16/17	数字系列	Plus/Slim系列	Pro系列	Pro Max系列
屏幕尺寸	6.1”/6.3”	6.7”/6.6”	6.3”/6.3”	6.9”/6.9”
处理器	A18（N3E）/A19（N3P）	A18（N3E）/A19（N3P）	A18 Pro（N3E）/A19 Pro（N3P）	A18 Pro（N3E）/A19 Pro（N3P）
基带芯片	X71M/-	X71M/自研5G BP	X71M/-	X71M/-
DRAM	8GB/8GB	8GB/8GB	8GB/12GB	8GB/12GB
后置摄像头	48MP主摄+12MP超广角/48MP主摄+12MP超广角	48MP主摄+12MP超广角/48MP主摄+12MP超广角	48MP主摄+12MP潜望式长焦+48MP超广角/48MP主摄+48MP潜望式长焦+48MP超广角	48MP主摄+12MP潜望式长焦+48MP超广角/48MP主摄+48MP潜望式长焦+48MP超广角
前置摄像头	12MP/24MP	12MP/24MP	12MP/24MP	12MP/24MP
Wi-Fi	Wi-Fi 7/-	Wi-Fi 7/-	Wi-Fi 7/Wi-Fi 7（自研）	/Wi-Fi 7（自研）
外观	铝/-	铝/-	钛/铝+玻璃	钛/铝+玻璃

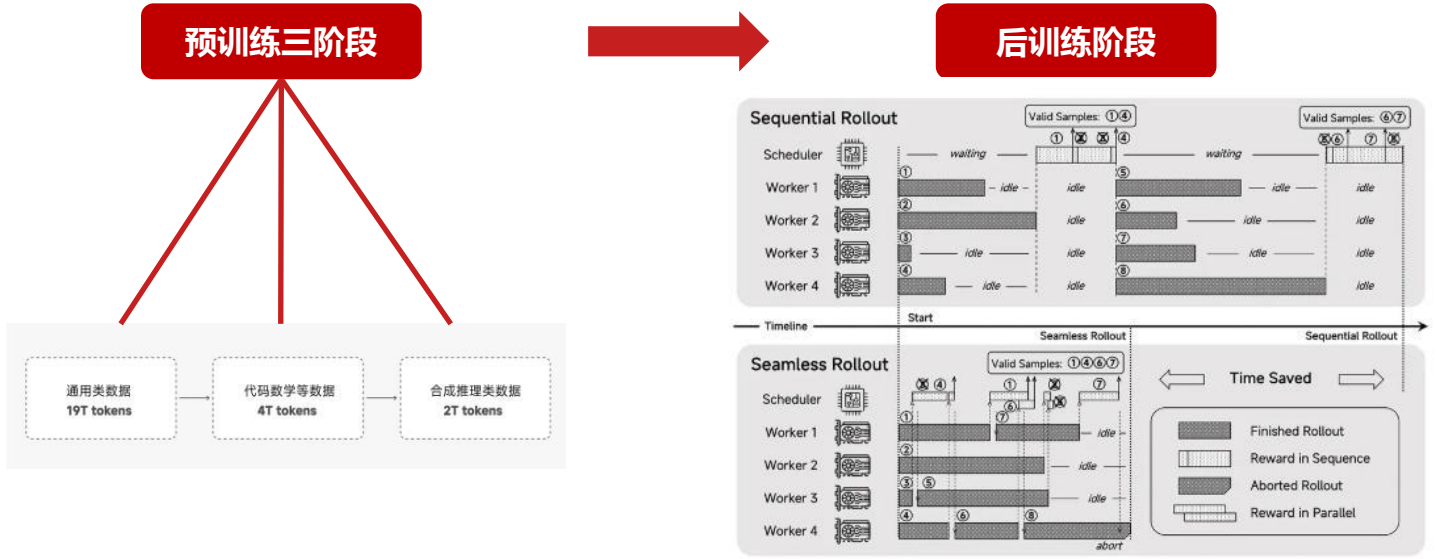
3.1.2 小米推出首个推理开源大模型MiMo，后续有望引入小爱同学赋能米家生态

- **4月30日，小米推出专注推理能力的开源大模型MiMo，仅用7B参数规模就在数学推理和代码竞赛测评中超越OpenAI的闭源模型o1-mini以及阿里32B规模的QwQ。**据小米介绍，Xiaomi MiMo诞生之初探索的核心问题就是激发模型推理潜能，这款模型联动预训练到后训练，全面提升推理能力：①预训练阶段：着重挖掘富含推理模式的语料，并合成了约200B tokens的推理数据。训练过程采用三阶段策略，逐步提升训练难度，累计训练了25T tokens，这一训练量在同等规模模型中处于领先水平。②后训练阶段：小米团队提出了“Test Difficulty Driven Reward”机制，有效解决了困难算法问题中奖励稀疏的问题。同时引入“Easy Data Re-Sampling”策略，显著提升了强化学习训练的稳定性。在框架层面，设计了“Seamless Rollout”系统，使得强化学习训练速度提升2.29倍，验证速度提升1.96倍。此外，**小米正在搭建自己的GPU万卡集群，将对AI大模型进行大力投入。小米的AI人才布局也在加速。**2024年12月20日，第一财经报道称DeepSeek开源大模型DeepSeek-V2的关键开发者之一罗福莉将加入小米，或供职于小米AI实验室，领导小米大模型团队。罗福莉是MLA（Multi-head Latent Attention）技术的核心开发者之一，该技术在降低大模型使用成本上发挥了关键作用。**我们认为随着小米对AI的全面布局，后续MiMo有望持续迭代并引入小爱同学赋能米家生态。**

图：Xiaomi MiMo测评得分对比



图：MiMo预训练&后训练阶段示意图



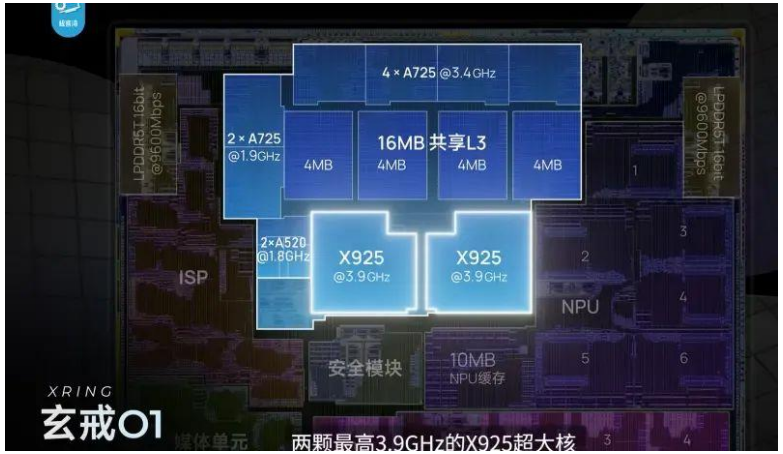
3.1.2 玄戒芯片正式落地，自研芯片+AIOS+大模型拼图完成

- **5月22日，小米召开15周年战略新品发布会，自主研发设计的芯片“玄戒XRING”正式发布。**继澎湃OS持续迭代、开源推理大模型MiMo之后，小米自研芯片也正式落地。卢伟冰曾在小米24年业绩会上将AI、OS和芯片这三项列为小米核心技术，芯片的突破标志着小米核心技术拼图的完成。**过去5年，小米在核心技术研发上投入约1020亿元，未来5年（2026-2030），将再投入2000亿元。其中，截至2025年4月底，小米玄戒芯片研发投入达135亿元。**
- **目前，“玄戒XRING”包括用于手机和平板的玄戒O1芯片，以及用于运动手表的玄戒T1芯片。（1）玄戒O1芯片：**采用全球最先进的第二代3nm工艺，拥有190亿个晶体管，芯片面积仅109mm²，采用十核四丛集CPU（多核性能跑分超过A18 Pro）+ 16核GPU，安兔兔跑分超过300万分，成功跻身旗舰芯片第一梯队。玄戒O1芯片采用Arm公版架构+自研NPU IP及ISP单元。（2）玄戒T1芯片：集成了小米自研4G基带，蜂窝通信全链路自主设计，支持4G eSIM独立通信。小米自研基带完整覆盖4G-LTE各层协议，7000+测试用例，具有海量的现网适配，覆盖100+城市。
- **据极客湾测评：**（1）玄戒O1芯片确为小米自研；（2）玄戒O1芯片CPU表现接近A18 Pro和天玑9400，GPU峰值性能与天玑9400持平，能效介于A18 Pro与骁龙8 Elite之间；（3）同模具的小米15S Pro，玄戒O1在续航上逊于骁龙8 Elite版本。

图：玄戒O1 Layout设计图

- ✓ **芯片布局与同期旗舰（天玑9400、骁龙8 Elite、A18 Pro）差异显著**，尤其是CPU核心簇和缓存结构设计独特，因此绝非公版套壳。
- ✓ **小米在标准单元库外设计了大量定制Cell**，用于优化X925超大核高频性能和A725表现。
- ✓ **公版IP后端设计独具匠心：**如同一A725架构采用两种后端设计，高频性能核（3.4GHz）与低频能效核（1.9GHz），后者面积更大且晶体管布局不同，体现自主优化能力；无SLC（系统级缓存）设计，通过增大各单元独立缓存（如NPU配备10MB缓存）规避低功耗场景风险等。

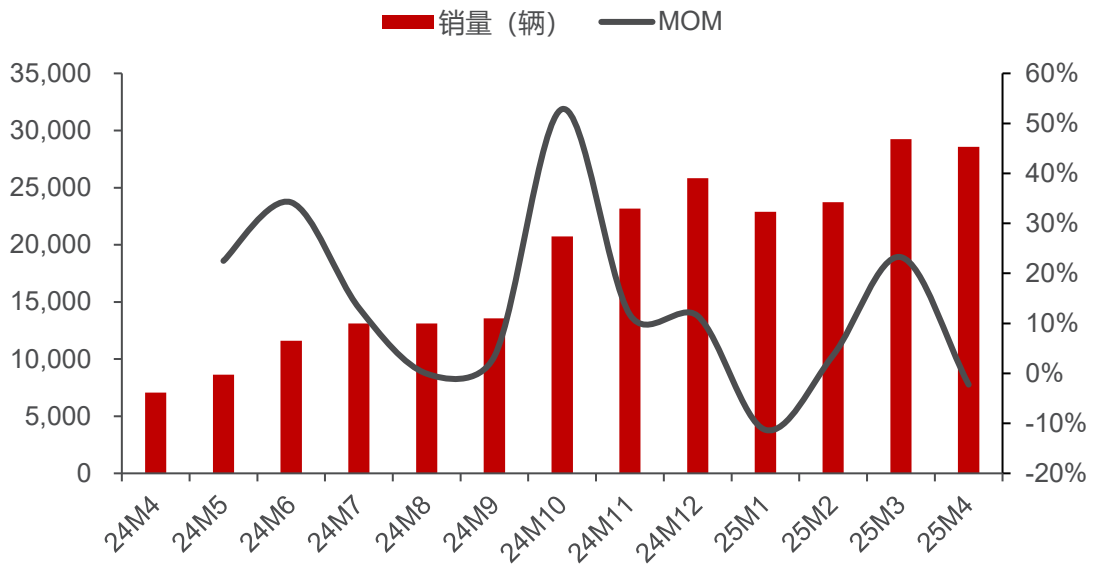
图：玄戒O1 CPU架构



3.1.2 小米人车家生态正式建成，科技家电&汽车业务协同发展

- **大家电赛道增速明显，家电业绩持续向上：**小米在家电领域坚持“和用户做朋友”的理念，用户需要什么，小米就生产什么，因此小米的产品容易戳中用户痛点，引发用户共鸣。此外，小米不断优化家电业务服务，尤其是大家电“拆送装一体式”服务，已覆盖2898个区县，平均完成时效仅需1.6天。
- **智能家电工厂奠基，科技家电新起点：**继“手机工厂”、“汽车工厂”之后，小米第3座大型智能工厂，也是小米首座“智能家电工厂”正式奠基，这意味着**小米“人车家”全生态战略在大型自建智能工厂方面完成闭环。小米智能家电工厂一期聚焦“空调”品类，计划于2025年正式投产，并于2026年大规模量产。**
- **SU7的发布标志着人车家全生态正式建成，随着YU7在小米15周年战略新品发布会亮相，小米汽车产品矩阵进一步得到扩充。**小米YU7系列定位豪华高性能SUV，包括标准版、Pro和Max三个版本。YU7暂不开放小定，将会在7月正式上市，并公布价格，展车即将到店，用户可以线下体验。小米YU7全系列拥有超长续航（YU7标准版续航达到835km，是所有中大型纯电SUV的续航第一）+ 小米天际屏全景显示 + 700 TOPS辅助驾驶算力 + 激光雷达 + 连续阻尼可变减振器。

图：小米汽车月度销量情况跟踪（单位：辆）



图：豪华高性能SUV YU7于小米15周年战略新品发布会正式亮相

3.2

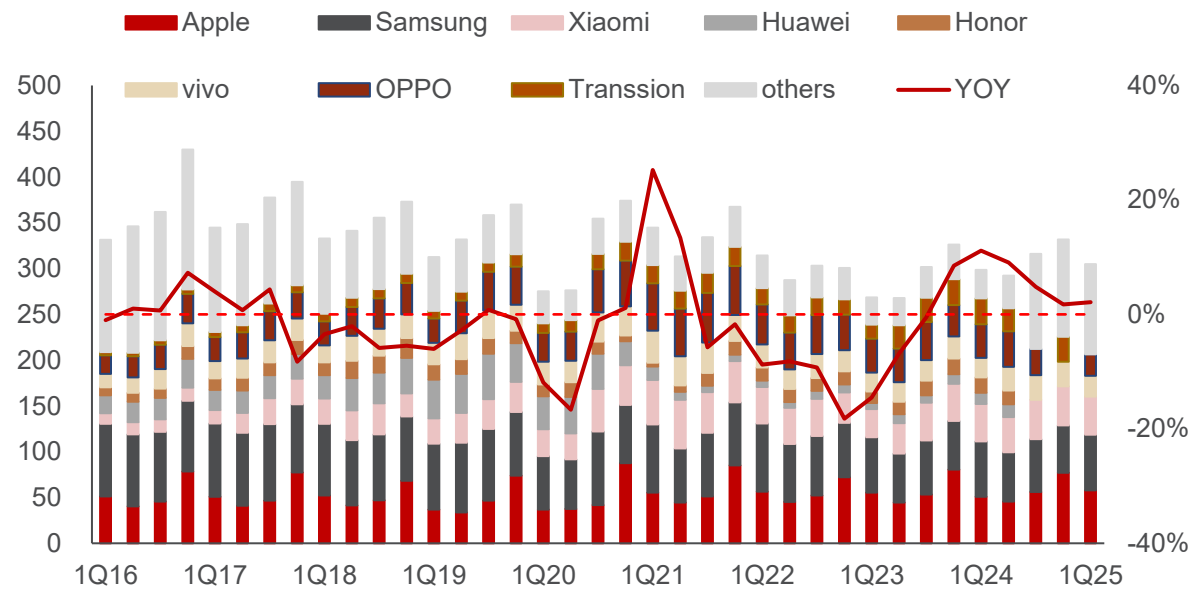
CONTENTS

目录

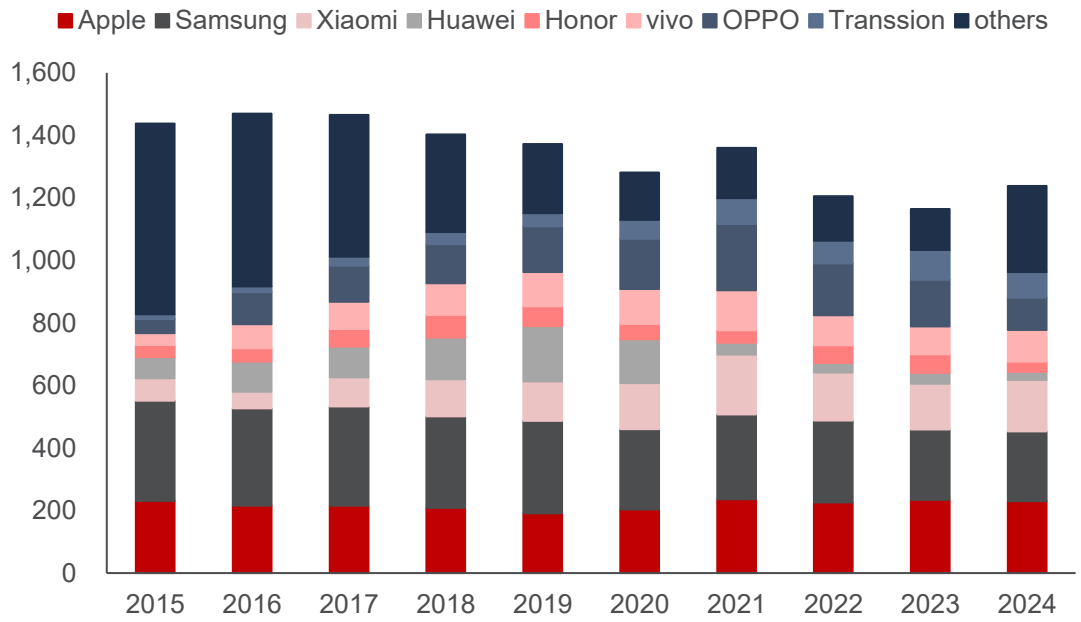
3.2 智能手机：步入存量市场，但仍有结构性创新

- 过去几年，智能手机新品硬件创新乏力，市场增速放缓，全球公共卫生事件后智能手机销量反弹，叠加GenAI或带来增量需求并推动高端化，市场对后市仍有期待。
- 24年6月，苹果在开发者大会上正式推出的系统级AI Apple Intelligence率先完成了AI手机的“打样”；2024年，消费电子市场回暖，叠加AI功能带来的换机需求，据IDC，全球智能手机出货量的同比增长6.4%，达到12.39亿部。
- 25Q1，海外智能手机销量数据表现低于预期，智能手机AI功能仍待完善，如截至目前苹果Apple Intelligence与第三方App交互的能力尚未落地，AI Siri和国行版Apple Intelligence仍未推出，尽管如此，供应链的创新依旧不止，下文我们将就核心创新点进行展开介绍。

图：全球智能手机出货量季度数据（百万部）



图：全球智能手机出货量年度数据（百万部）



资料来源：IDC，民生证券研究院整理

3.2 手机核心硬件创新一览

光学

镜头方面，玻塑混合镜头在提升进光量、降低镜头高度方面具有优势；**潜望式下沉**是重要创新之一，潜望式镜头通过微棱镜结构实现光学变焦，同时支持远景和微距拍摄；其他光学创新还包括**可变光圈、超光谱摄像头、外挂镜头；CIS方面，**超大底（4/3英寸）+大像素是升级趋势；**TOF系统，**同3D结构光技术，可用于生物识别和手势识别。

- **模组：**舜宇光学、丘钛科技、高伟电子、欧菲光
- **光学零部件：**水晶光电、蓝特光学、瑞声科技、东田微
- **CIS：**韦尔股份、思特威、格科微
- **TOF：**力芯微

折叠屏

在全球智能手机存量竞争的背景下，中国已成为全球最大折叠屏手机市场。24年9月，华为发布开创性三折屏产品Mate XT；苹果预计26年将推出折叠iPhone。影响折叠屏渗透率提升的三个关键要素分别为价格、重量和厚度。

- **折叠屏：**精研科技、东睦股份、统联精密

指纹识别

相较于光学屏下，**超声波指纹识别**方案更轻薄、更省电，对屏幕要求更低，解锁体验更好，且解锁区域设计更灵活，预计今年将渗透至安卓系中端机型。

- **指纹识别：**汇顶科技、丘钛科技、欧菲光

3.2 光学：行业创新趋势

前摄

- **超小头部**
- **自动对焦/OIS**
- **小广角（90°）**
- **超薄化：**可配合折叠屏手机在有限空间内的堆叠要求，并保留更多细节的图像
- **大光圈**

主摄

- **大像面+大光圈：**实现进光量的提升，帮助夜景拍摄获得更好效果。
- **外挂镜头：**可用于手机的类单反镜头，可带来专业的影像效果
- **玻塑混合：**玻璃镜头透光率高、能做到更薄且不易老化；而塑胶镜头可塑性强，易制成非球面的形状，方便小型化，加工成本低，但透光率低；玻塑混合镜头结合了二者的特点，在成本和透光率及厚度两个方面进行了平衡。
- **可变光圈：**镜头可以通过调节光圈大小来控制进光量从而影响曝光、调节景深以及控制图像质量
- **芯片防抖**

超广角

- **小畸变**
- **超薄化**
- **大光圈**

长焦

- **潜望式：**利用反光镜或棱镜折射光线，从而实现高倍数光学变焦。其工作原理类似于潜水艇上的潜望镜，通过改变光路的方向，使光路在机身内部发生折叠，从而在不显著增加手机厚度的情况下实现更长的焦距。**我们认为未来潜望式主要有两个升级方向：①双潜望镜头；②多群潜望。**
- **棱镜中置/后置**
- **低高度长焦**

3.2 光学：外挂镜头，全新技术路径

- **小米MWC 2025秀影像“肌肉”，磁吸式M43大底镜头震撼来袭：**在MWC 2025大会上，小米展示了一款概念产品——磁吸式可拆卸镜头，其被命名为“小米模块化光学系统”，该系统内置了1亿像素的M43传感器，并配备了F1.4大光圈、35mm焦段的定焦镜头，镜头具备可变光圈，低光环境下表现出色。其安装方式类似于MagSafe磁吸配件，轻贴即可固定在手机背面，然后在相机应用中点击图标切换至可拆卸镜头模式。镜头内置自动对焦马达，可通过屏幕点击对焦，同时也配备实体对焦环，便于手动微调。这一系统的关键在于小米的LaserLink技术：一种专有光通信模块，手机和镜头背部各有一个小型光电，数据通过近红外激光传输，速度可达10Gbps。这一高速连接不仅能保证拍摄时的流畅体验，还能与小米的AI计算摄影系统协同优化成像质量。**相较于vivo仅有光学元件的方案，小米M43镜头内置了独立传感器并采用了LaserLink技术保证拍摄体验，我们认为后续会有更多厂商跟进这一全新的技术路径。**
- **vivo外挂长焦镜头：**vivo率先推出了一款与蔡司联合研发的2.35×长焦增距镜，能够打造更远距离的拍照覆盖，适用于演唱会、音乐节等拍摄场景。

图：小米M43超级大底镜头

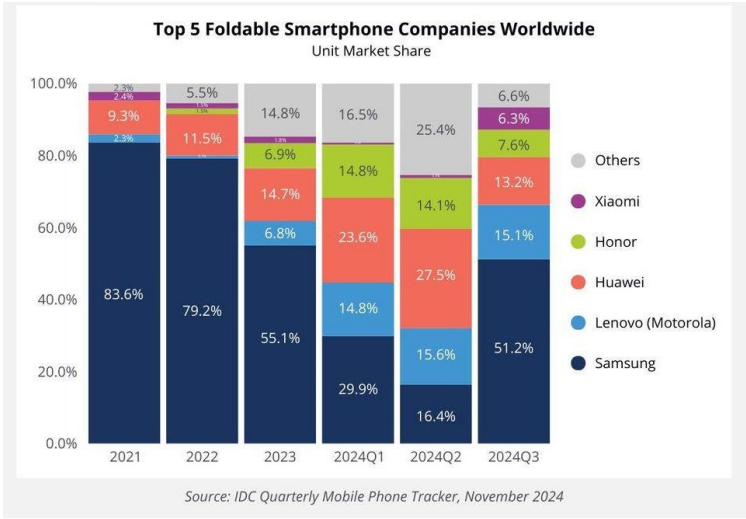
图：vivo蔡司2.35倍增距镜/外挂镜头

3.2 全球折叠屏手机市场：2024年增速放缓，2025小年，2026年苹果入局

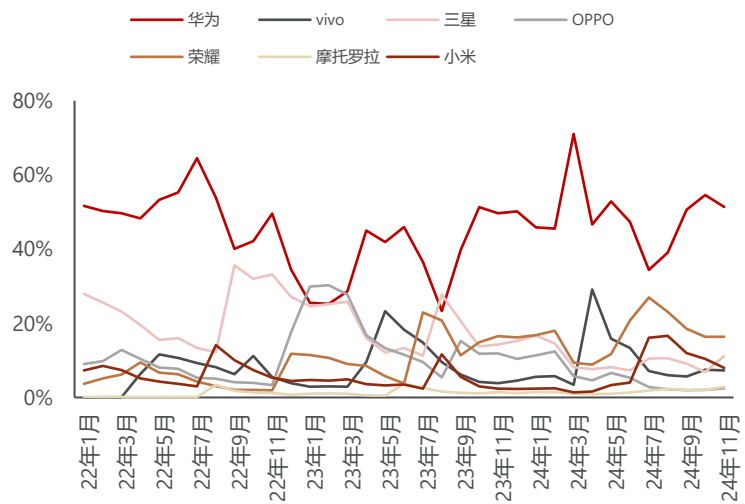
- **2024年增速放缓，2025年或将同比下滑，下一个增长高峰将由苹果在2026年发布的折叠屏手机推动：**据Canalys，2024年全年折叠屏手机出货量仅能实现13%的同比增长，约1970万台。当前该市场仍被华为和三星两大品牌主导，二者在2023年和2024年的面板采购市场中合计占据了70%的份额。2025年折叠屏手机销量或将迎来同比下滑，预计2026年苹果入局将为行业带来超过30%的增速，2027~2028年也将有超过20%的增长表现。
- **全球竞争格局：**当前折叠屏手机市场格局并不稳定，且受新机发布周期影响，但总体来看，TOP5（24Q3）玩家包括三星（51.2%）、联想（15.1%）、华为（13.2%）、荣耀（7.6%）和小米（6.3%）。
- **国内竞争格局：**中国是全球最大的折叠屏手机市场，而国内折叠屏手机市场是一个寡头市场。据IDC，2024年国内TOP5玩家包括华为（48.6%）、荣耀（20.6%）、vivo（11.1%）、小米（7.4%）和OPPO（5.3%）。

图： 2019-2028E 全球折叠屏手机出货量预测（百万部）

图：2021~2024Q3全球各品牌折叠机市占率



图：2022-2024年国内各品牌折叠机市占率



资料来源：金融界，Canalys，新浪财经，IDC，新浪科技等，民生证券研究院整理

3.2 铰链：折叠屏手机的主要增量成本

- **铰链：**铰链一般由多个金属零件组装而成，其中精密金属零件的制造工艺主要包括**MIM（金属注射成型）、锆基液态金属、3D打印，部分常规零件也可以用CNC、冲压等传统工艺生产。**以铰链制造的核心零部件工艺MIM（金属注射成型）为例，美国企业**安费诺**作为消费电子、军工铰链设计王牌，一家包揽了荣耀、小米、vivo、OPPO等企业铰链方案设计。而随着国内折叠屏产业链的羽翼丰满，行业内的头部厂商也逐渐触及到海外电子消费品牌。如国内企业**精研科技、统联精密、东睦股份等**也可为折叠屏提供MIM产品。

表：折叠屏产业链布局

上游材料	中游零部件					铰链组装
液态金属	**MIM**	**钛合金轴盖**	**转轴**	**3D打印**		**海外为主**
宜安科技	**精研科技**	铂力特	科森科技	**铂力特**		安费诺（美国）
东睦股份	**东睦股份**	金太阳	大富科技	**华曙高科**		奇鋐（中国台湾）
铝合金材料	**统联精密**	**碳纤维支架**	**CNC**	**华工科技**		兆利（中国台湾）
福蓉科技		领益智造	长盈精密			KH Vatec（韩国）
						S-Connect （韩国）

3.2 指纹识别：生物识别技术迭代路径

非全面屏 → 传统电容式指纹识别：物理按钮与指纹识别合二为一，典型如早期iPhone的Home键

不仅改变了智能手机形态，使手机拥有了更高的屏占比，还带来了生物识别技术的变迁

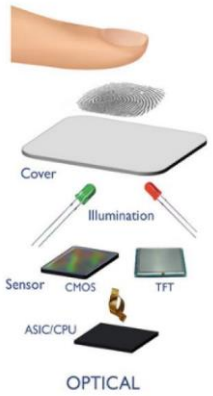
全面屏

- （3D）人脸识别：通过结构光构建人脸模型进行识别，缺点是不可避免刘海、灵动岛、三挖孔等，比指纹识别更贵，主要是苹果、华为在使用
- 屏下指纹识别：主要是安卓系厂商在使用

电容式 → 光学式 → 超声波式

LCD → OLED（AMOLED）

CAPACITIVE



OPTICAL

ULTRASONIC

3.2 指纹识别：超声波指纹识别原理及优势

➢ **超声波指纹识别技术原理**：利用声波在物体表面的反射和散射特性来识别指纹，能显著提升安全性并给予解锁区域设计更高的灵活度。

发射：芯片发射一段特定频率的超声波脉冲，声波穿过屏幕到达用户指纹

→

接收：用户指纹反射该声波信号并被传感器接收

→

处理：芯片将接收到的声波信号转换为电信号并通过测量时间延迟构建起指纹的三维图像作为提取到的指纹的特征信息

➢ **超声波 VS 光学屏下**：超声波方案更轻薄，更省电，屏幕要求更低，解锁体验更好。

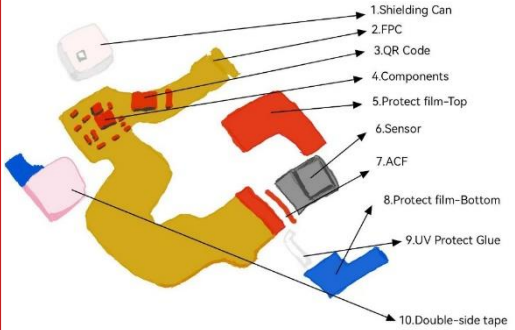
性能	超声波屏下指纹	光学屏下指纹
厚度	仅为0.1-0.2mm，可以贴附在屏幕背面，有利于元器件堆叠，为电池等核心部件挪位置	更厚
对屏幕透光率要求	无要求，可降低屏幕成本	较高
功耗	更低	为实现透光屏幕必须亮起，功耗更高
支持湿手解锁效果	在手指、屏幕有污渍或水渍的情况下，仍能够快速准确解锁	识别效果不理想

光学屏下

模组对应区域屏幕要做开孔透光处理

模组位置固定，比较占用空间

超声波

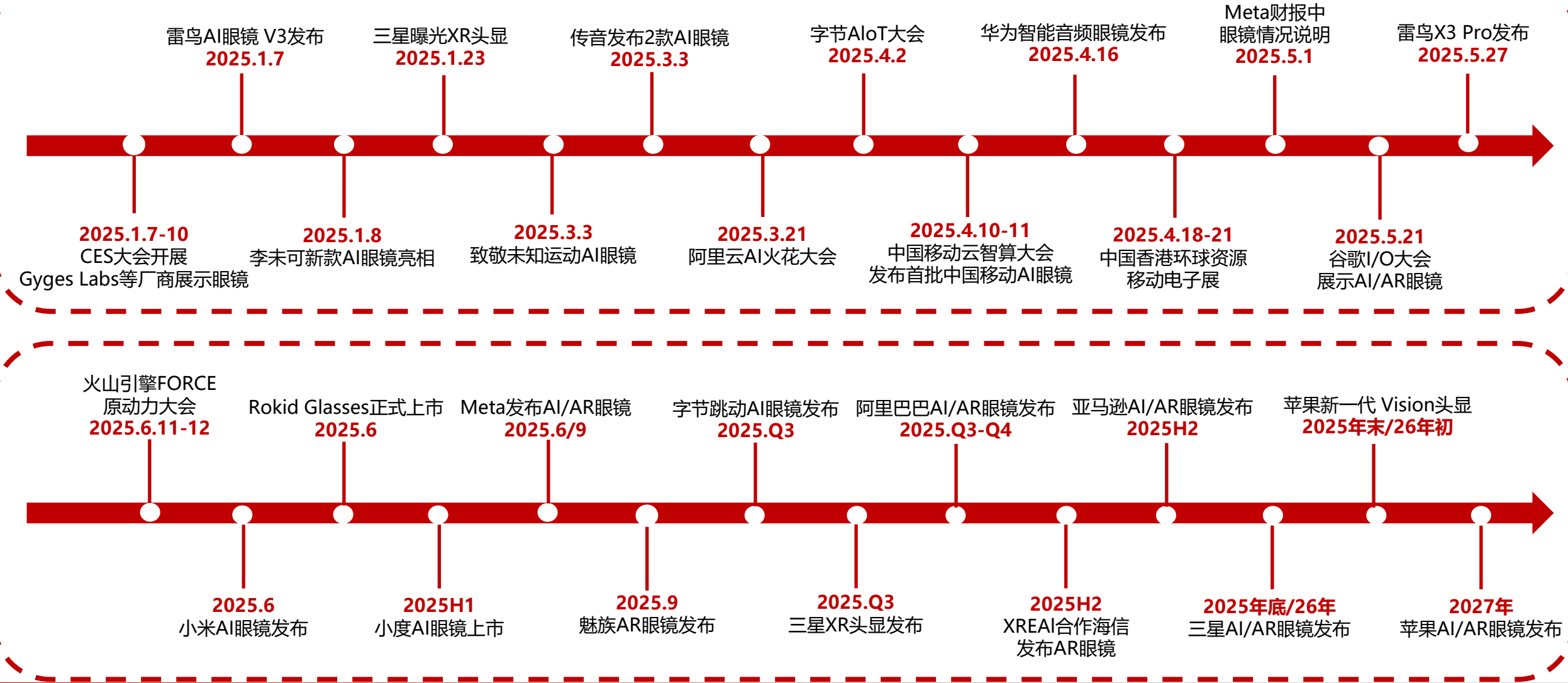


3.3

CONTENTS

目录

3.3.1 AI/AR眼镜日历表



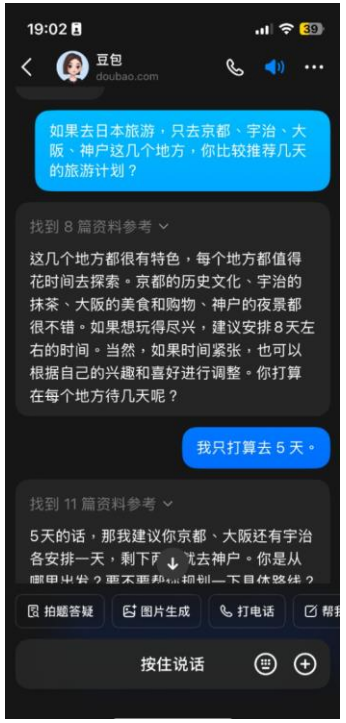
3.3.2 智能硬件的终端体验成为决胜关键

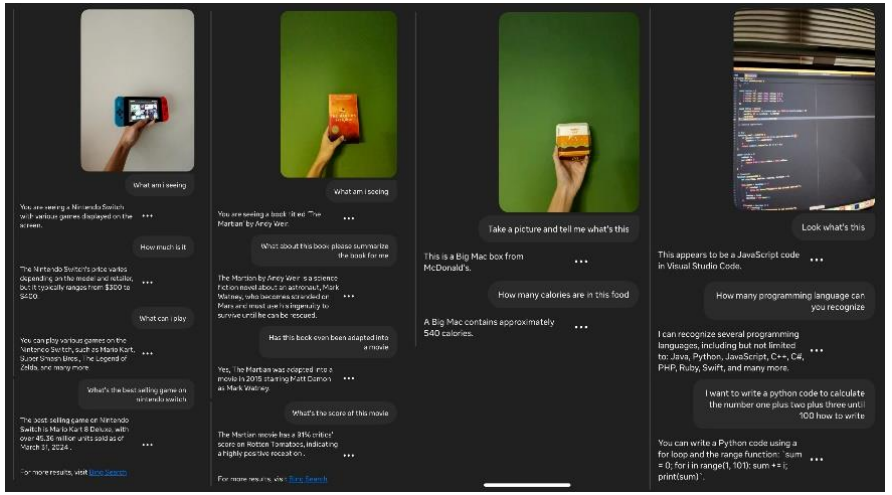
- 目前**A股的AI应用目前集中于ToC领域**，终端设备销量制约AI应用落地的进展。
- 分析Rayban-Meta眼镜和Ola豆包耳机发现，Rayban-Meta由于高度可用的音视频多模态功能成为现象级产品，而Ola豆包耳机因存在较多AI体验溢价导致销量不及预期（可被普通耳机+豆包替代）。
- 因此，我们认为最终的终端设备销量取决于用户的实际体验，**AI终端需要深耕应用场景，解决用户的刚需**，让用户愿意为设备的AI可用性付出产品溢价。

图：Rayban-Meta眼镜和豆包耳机对比

	Rayban-Meta 二代眼镜	Ola Friend 豆包耳机
外观		
售价	2200元	1199元
目标场景	太阳镜	运动耳机
功能	AI助理、多模态功能（摄像头翻译、识别等）	AI功能仅实现基础对话 无法调用系统和第三方APP
销量	截至2024年H1， 累计出货超100万台	截至2025年1月9日， 天猫旗舰店销量仅6000+

图：用户与设备终端互动展示（左：豆包耳机，右：Rayban-Meta眼镜）





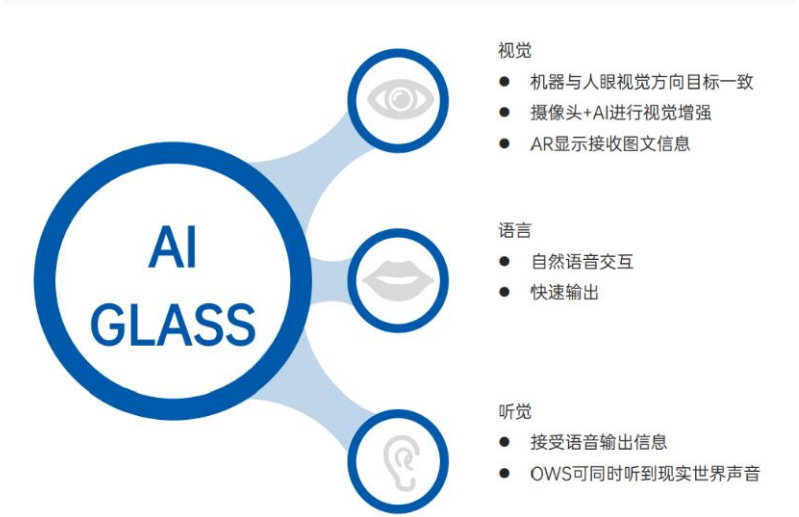
资料来源：艾邦AR/VR网，VR陀螺，淘宝天猫，民生证券研究院整理

资料来源：前方智能，雷科技，民生证券研究院整理

3.3.2 如何看待AI终端的未来

- **1）价值量提升：**与早先的消费电子创新，如TWS耳机对比，AI终端是系统级产品，需适配诸多应用+AI大模型+多元化交互方式，产品形态的升级演进需要循序渐进。但优点在于，一旦用户体验成熟，带来便利解决刚需，用户的付费意愿更强。**我们认为，AI终端的定价=硬件成本+AI体验，有更高附加值，可以打开品牌厂商的利润天花板。**
- **2）交互体验：**当下的AI终端产品，从前期发布的智谱、荣耀手机，到近期发布的Rokid眼镜，均主打语音交互，极大程度地解放双手，未来还将结合手势识别等功能，进一步革新交互体验。**我们认为，未来手机的定位将转向随身的算力终端，其交互过程可更多通过AI硬件实现，**即：1）通过AI眼镜的摄像头，实现“所见即所得”；2）通过麦克风，实现“所言即所行”。
- **3）产业链生态：我们认为当前的AI终端可类比十年前的智能机。**2012-2015年间，传统国产智能机四强“中华酷联“，遭遇OPPO、vivo、小米等新“玩家”的挑战，行业加速洗牌。而当下，AI赋能加速智能终端崛起，类比智能机产业的格局演变，**我们看好未来双线并行的市场格局，**即品牌厂商自研硬件+大模型；白牌厂商则采用公版方案配合第三方模型，各自占据目标市场。

图：AI眼镜可实现的交互方式



图：未来消费电子的交互方式将被重构

3.3.2 有何值得期待：大模型加持，AI玩具新风口袭来

- 在字节引领的新一轮AI终端浪潮中，**我们认为大模型将赋予传统玩具生命力，因此AI玩具有望作为情感陪伴载体，适合成为消费领域AI终端的优先落地场景。**AI玩具能通过AI技术及时发现并回应用户情绪，从而提供高质量的情感支持，满足了用户被理解、被倾听和被关注的需求，更容易刺激消费者的购买欲望。
- 目前的AI玩具市场的发展趋势：**1）AI功能不断优化；2）IP+玩具打造特色；3）应用场景持续拓宽**
- 同时，**相较于其他AI硬件，AI玩具的消费品属性更强。**因此，AI玩具领域的核心在于定义细分用户的玩具需求，并针对性提升AI能力、设计AI应用，做好价格与体验的平衡。我们判断，在AI玩具终端中，深耕行业需求和产品生态的专业玩具厂商更有机会脱颖而出；在供应链领域，具备开源社区和开发者社群的厂商，更有机会在AI玩具这个长尾市场中引领差异化AI应用的发展。

图：字节AI毛绒玩具“显眼包”

图：FoloToy的八爪鱼AI套件

图：AI玩具市场发展趋势

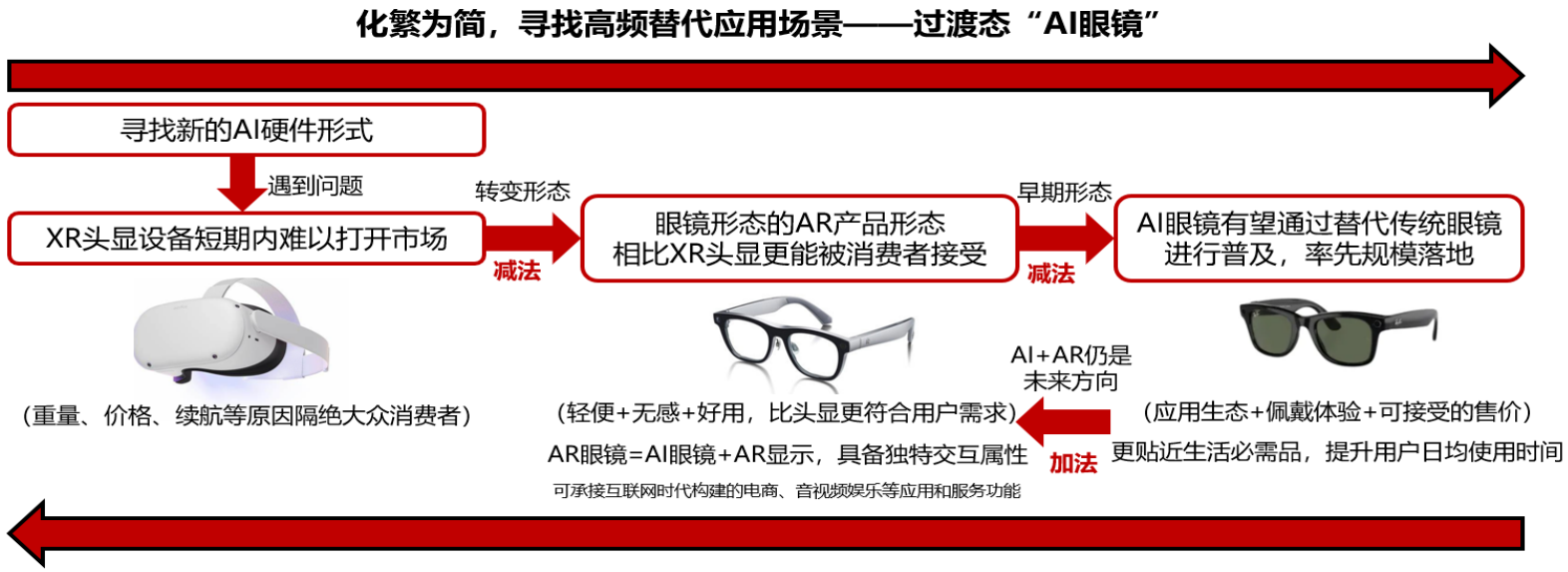
资料来源：腾讯网，民生证券研究院整理
资料来源：淘宝网，Folotoy官网，民生证券研究院整理

资料来源：芯航员，众享时刻，民生证券研究院整理

3.3.2 有何值得期待：智能眼镜AI先行，探索AR

- 总结Rayban-Meta眼镜成功的主要原因，我们认为产品的成功源于其具备优秀AI终端的特点，**可替代生活必需品（基础）+新增优秀的AI应用体验（溢价）。从产品形态来看，AI眼镜是一种符合当前市场预期和消费者认知的过渡期形态。**
- **展望后续，认为当下的AI眼镜不会是产品的最终形态。**智能眼镜在成为成熟的智能终端产品的道路上，还面临丰富的生态应用、舒适的佩戴体验、合适的售价三者平衡的考验。然而，目前AI眼镜缺少视觉输出，意味着在互联网和移动互联网时代所构建的涵盖电商、本地生活、音视频娱乐等成千上万的应用和服务，基本都难以在AI眼镜平台上重现。因此，**从生态应用的角度来看，AI眼镜很难成为一个通用的计算终端。**
- 当用户对AI眼镜的接受度逐步提高，眼镜产品由于其具备贴近用户视觉的特性，在形态上会逐步叠加光学显示模块（光波导+光机），走上AI+AR的道路。其中，AI与AR的能力是相辅相成的，AI可以提升AR交互的智能性（如手势识别、眼动跟踪等），AR则是AI合适的显示载体。因此，**后续的智能眼镜产品节奏应该是AI先行，探索AR。**

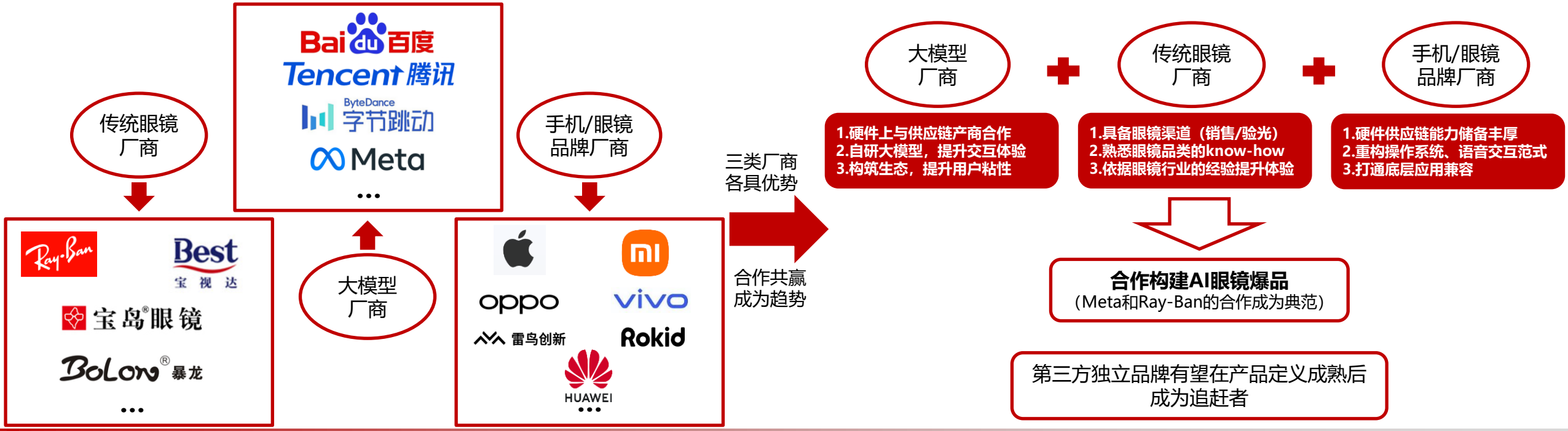
图：智能眼镜演变历史



3.3.2 有何值得期待：AI眼镜市场混战，谁会胜出？

- **从软硬件两个维度分析，在AI眼镜的硬件储备方面**，AR眼镜硬件储备相对充足，但传统眼镜渠道（销售/验光）、眼镜品类的know-how等对于各大软硬件厂商是一场全新的大考，需要有大量眼镜行业的经验托举，率先指明这条路径的正是Meta的Ray-Ban。
- **在AI眼镜的软件储备方面**，许多厂商在过去1年做好准备。纵观手机厂商们近来的发布会，基本都重构了操作系统、语音交互范式，从底层打通与多模态模型、AI应用的兼容，为AI眼镜构筑了完善的软件体系。
- 虽然当前“AI眼镜大战”有较多参与者，但最终来看，我们认为会呈现**品牌AI终端和白牌+第三方模型终端两大阵营并行的状态。**对比来看，传统硬件品牌厂商，尤其是资源密集的手机厂商，由于其掌握海量用户和流量入口，天然具有竞争优势，更有望居于AI眼镜这条产业链的核心位置。

图：三类厂商共筑AI眼镜爆品



3.3.2 AI/MR/AR 差异一览

- AI眼镜、MR和AR的代表产品分别为Meta & Ray Ban，苹果Vision Pro、Meta Quest 3和微软Hololens；**从产品形态上看**，AI主打眼镜+语音交互，AR眼镜则主打光源+光学交互。**硬件上**，SOC：除苹果外，AI眼镜及主流的AR产品均采用高通的AR1G1/XR2G1芯片；光学（AR）：目前主流方案为衍射/阵列光波导+Lcos/Micro LED；显示（MR）：目前主流方案为菲涅尔+Micro OLED。

表：apple vision pro、雷鸟X2、Xreal Air2、Ranban-Meta二代产品对比

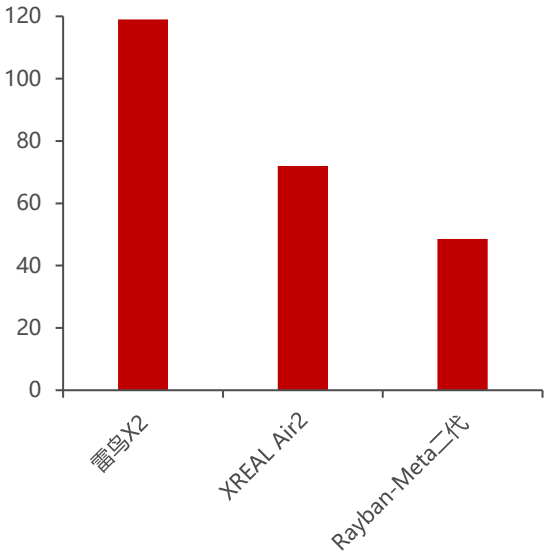
	MR头显	AR眼镜（信息提示类）	AR眼镜（观影类）	AI眼镜
代表产品	**Apple Vision pro**	**雷鸟X2**	**XREAL Air2**	**Rayban-Meta二代**
图示				
重量	600-650g	119g	72g	48.6g
价格	3499美元起	4999元	2599元	299美元
上市时间	2023年6月	2023年10月	2023年9月	2023年9月
续航/电池容量	2h-2.5h	590mAh	无	4h
光显方案	**micro-OLED**	**双目异显衍射光波导+Micro-LED显示屏**	**Micro-OLED+Birdbath**	**无**
分辨率	单眼4k	双眼640×480	AR Space：3840x1080 投屏模式：1920x1080	无
处理芯片	Apple M2 Chip、Apple R1 Chip	高通骁龙 XR2 Gen1	无	高通AR1 Gen1
接入AI模型	**无**	**Rayneo AI**	**无**	**Llama3**
交互方式	眼动跟踪、手势识别、语音	镜腿触控交互、戒指射线交互、语音交互	可直连设备或搭配Beam操作	语音、镜腿触控、按钮拍照
功能概述	1、透过现实世界背景在空间中布置窗口并利用手势对其操作。 2、拍摄并观看空间视频。 3、和其他苹果系列产品互动形成生态。	1、贴面翻译：翻译字幕紧随脸侧显示。 2、辅助对话：大模型根据翻译分析回答。 3、空间导航：悬浮于眼角的地图信息。 4、空间坐标：显现周围地标建筑信息 5、AI语音助手、息屏速拍等。	1、搭配Beam Pro，实现3DoF可悬停空间屏。 2、搭配Beam Pro空间操作系统，将移动APP生态空间化，实现空间视频即拍即看。	1、AI作为对话助手回答问题、提供信息。 2、多模态功能：调用摄像头翻译、识别物体

3.3.2 三款AI/AR眼镜差异一览

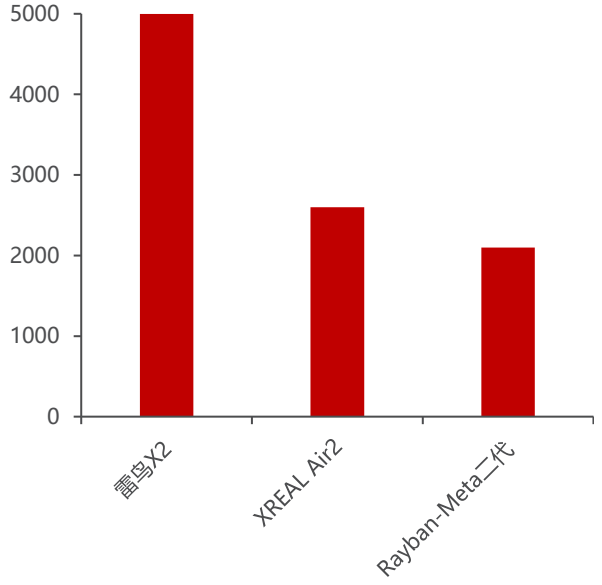
- **三款眼镜差异对比：重量方面**，Rayban-Meta因其作为户外墨镜平替的性质，重量最轻；**价格方面**，雷鸟X2因需适配日常、办公娱乐，具备丰富功能，价格最高；**重度体验时间方面**，三款眼镜重度体验时间均在2-3小时，距离满足用户长时间便捷体验还存在差距。

- **Rayban-Meta成本：**根据维深 wellsenn XR拆解，总BOM成本约164美元，核心硬件成本达75美元，合计占比约45.73%（包括SoC成本55美元、ROM+RAM成本11美元、摄像头成本9美元）

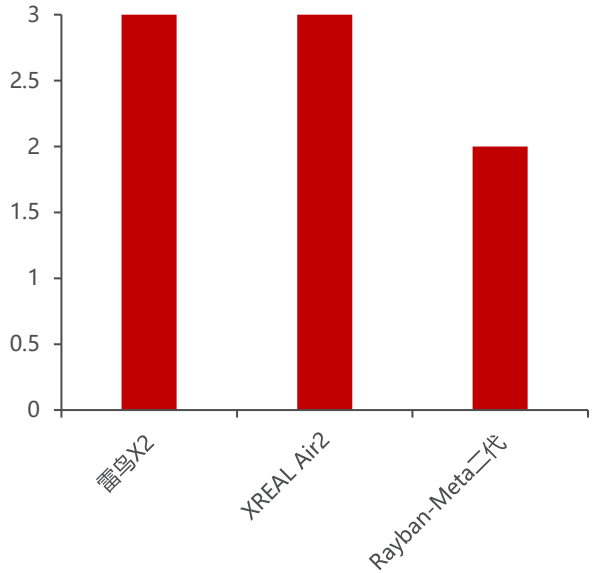
图：三款眼镜重量对比（g）



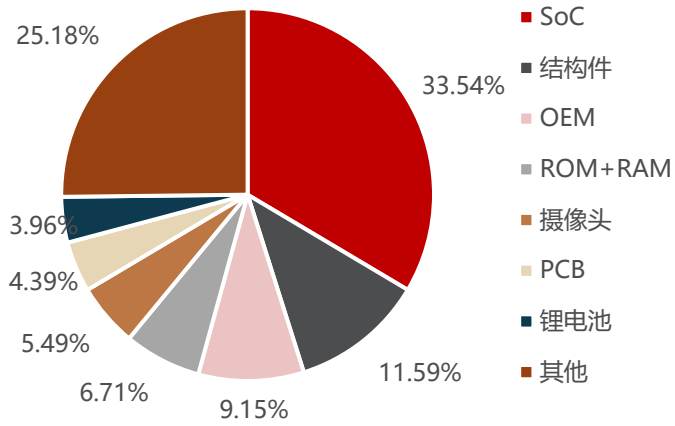
图：三款眼镜价格对比（元）



图：三款眼镜重度体验时间对比（h）



图：Rayban-Meta BOM成本拆解（按元件）



注：价格中美元：人民币=7：1，Rayban-Mate价格近似2100元

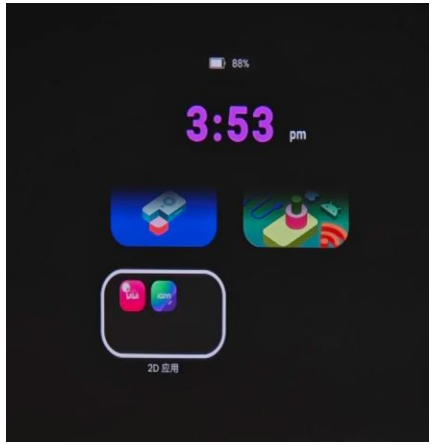
注：重度使用含义为，雷鸟X2为不间断使用场景，XREAL Aie2搭配Beam Pro持续使用，Rayban-Meta频繁调用AI或拍摄视频

3.3.3 XREAL Air2核心体验

- **观影体验：**产品与手机、游戏机等设备连接可拓展显示画面，最高等效 4 米距离观看 130 英寸屏幕的表现，提供巨幕体验；APP中自带AR空间，可在空间内进行AR运动、AR观影、AR办公等一系列AR互动操作。
- **Beam计算终端：**设备硬件方面具备方向键、确定键、音量键等一系列功能键，支持体感指针操控；设备系统内置哔哩哔哩和爱奇艺等APP；XREAL Air 2+Beam的组合可激活“悬停、云台、浮窗”三大模式，将游戏/观影体验感拉满。
- **新款Beam Pro AR计算终端：**具备基于安卓深度定制的系统，可安装各类安卓应用；设备后置双摄用于拍摄空间视频和照片。

图：XREAL Beam外观和系统

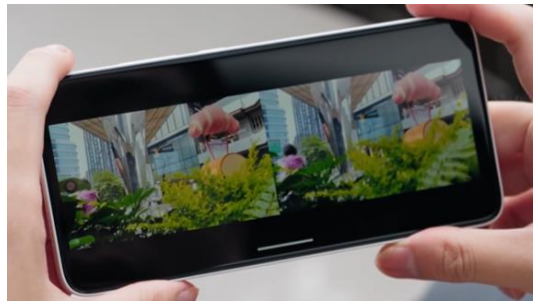
Beam外观



Beam系统

图：新款Beam Pro系统和应用

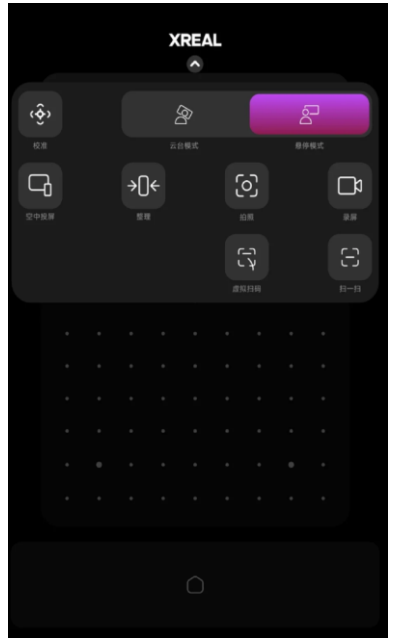
Beam Pro系统界面



Beam Pro拍摄空间视频并观看

图：AR空间效果展示

APP选择进入AR空间



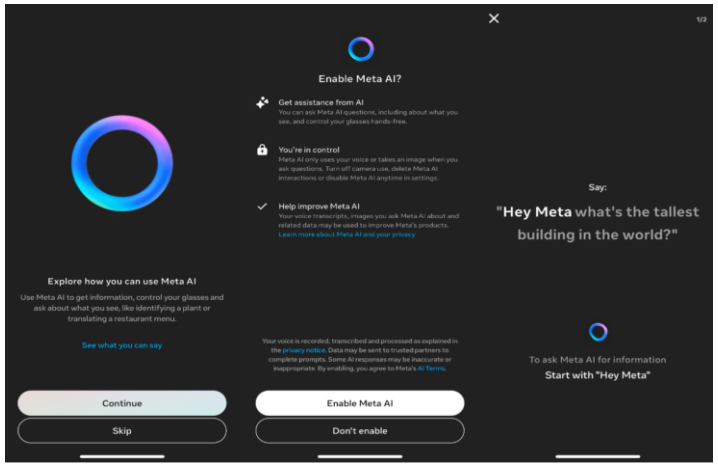
AR控制面板

AR空间效果

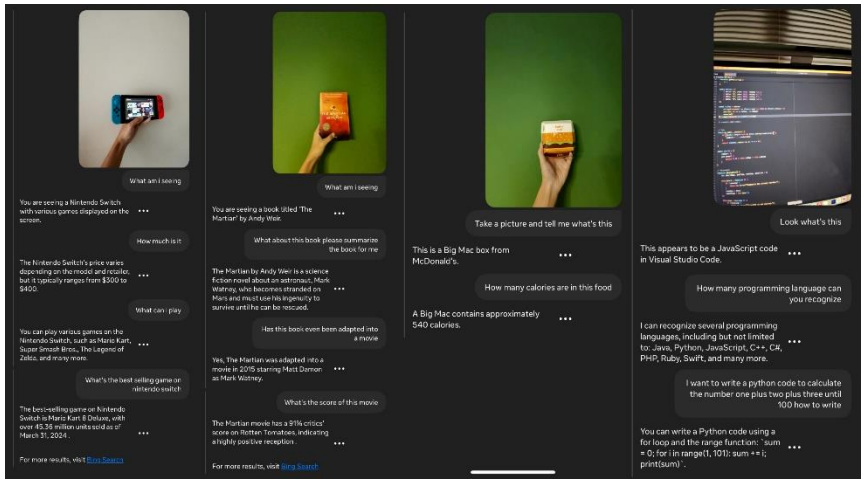
3.3.3 Rayban-Meta二代核心体验

- **Meta AI功能：**AI功能必须全程联网使用，包括语音助手、翻译、AI识物等；Meta AI可结合眼镜摄像头拍摄来识别物品，并回应问答，还可以翻译并以语音反馈拍摄到的标识和文字（目前仅支持英语、法语、德语、西班牙语和意大利语间互译）。
- **音乐与通信功能：**仅内置对 Apple Music、Spotify 和 Amazon Music 三个音乐流媒体服务的语音交互支持（不支持中文），可指定歌手、歌曲、曲风等；通信仅支持打电话和 Meta 旗下的 WhatsApp、Messenger 和 Instagram应用。
- **拍摄功能：**可语音或按键激活拍摄照片和摄像功能（单次最长可拍60秒视频），拍摄内容可导入手机；眼镜静态拍摄效果较好，但拍摄快速运动画面会出现运动模糊，不宜做运动相机。

图：Meta AI功能展示



注：AI功能须在Meta View 应用中开启，目前仅限美国和加拿大用户使用



询问AI后，AI会将获得的信息以语音的形式反馈给你，并将完整的对话内容存储在 Meta View 应用中

图：左为眼镜拍摄，右为 iPhone 15 Pro 超广角拍摄

3.3.3 Meta×依视路陆逊梯卡：AI/AR/XR眼镜新品来袭

- 作为智能眼镜的先行者，Meta的产品推出节奏在智能眼镜行业至关重要。**2025-2026年也将是Meta智能眼镜新品频出的阶段**：
- **AI眼镜方面：**Meta将与知名运动眼镜品牌Oakley联合打造AI眼镜，眼镜代号为Supernova 2，产品功能将与Ray-Ban Meta相近，同时更专注于骑行等运动的细分赛道，将融入时尚感、实用感、场景感等。Oakley联名款智能眼镜已于6月20日推出。
- **AR眼镜方面**：Meta有望推出带有小型显示器的智能眼镜Hypernova，眼镜右侧镜片下方嵌有单眼显示屏，支持多样化信息展示；另外，其摄像头性能较Ray-Ban Meta显著提升，拍摄质量媲美iPhone 13，可实现高清影像记录；此外，其配备了Meta自主研发的sEMG神经腕带技术，通过肌电信号识别实现精准手势交互，显著优化设备操作逻辑与交互体验。
- **XR头显方面：**Meta研发一款代号为“Puffin”的超轻量XR头显，通过核心组件的物理分离设计实现结构重构，在保障算力性能的同时，减轻用户穿戴负荷；应用场景方面，产品面聚焦于构建轻量化虚拟显示平台，重点布局便携式多屏协同办公与轻量级沉浸娱乐等场景。该产品预计将于2026年底前正式推向市场。

图：Meta联手Oakley打造运动智能眼镜

图：带显示的Meta新品眼镜

图：Meta超轻量XR头显

3.3.3 小米AI眼镜：6月蓄势待发，眼镜×车端联动可期

- 海外龙头厂商 Meta 率先掀起智能眼镜的形态创新浪潮，国内品牌厂商也纷纷跟进：
- **小米AI眼镜6月发布：**小米创始人雷军在微博发文预热在6月底举办的新品发布会，万众期待的小米YU7汽车将于会议发布，届时还有小米AI眼镜等诸多重磅新品同场一起发布。小米AI眼镜的产品形态预计将对标Ray-Ban Meta，预计售价 1499 元起，公司对产品预期销量30万台+。
- **产品硬件领域：**芯片方面，小米AI眼镜采用高通骁龙 AR1 Gen 1 +BES 2700的双芯片架构，针对性解决续航焦虑。其中，AR1芯片专注于影像处理，为用户带来出色的影像体验；而BES 2700 芯片作为耳机芯片，主要负责音频处理；在影像系统方面，小米AI眼镜搭载了索尼IMX681图像传感器，配备了5麦克风阵列，能够在嘈杂的环境中精准拾音，确保语音交互的准确性和流畅性。此外，还采用了电致变色镜片技术，可根据不同的光线环境，让镜片转变为墨镜，有效应对强光环境。

图：小米AI眼镜供应链拆解预测

图：小米AI眼镜多端应用联动

小米大模型提供支持

资料来源：维深wellsenn XR，民生证券研究院

资料来源：XR控，新浪新闻等，民生证券研究院整理

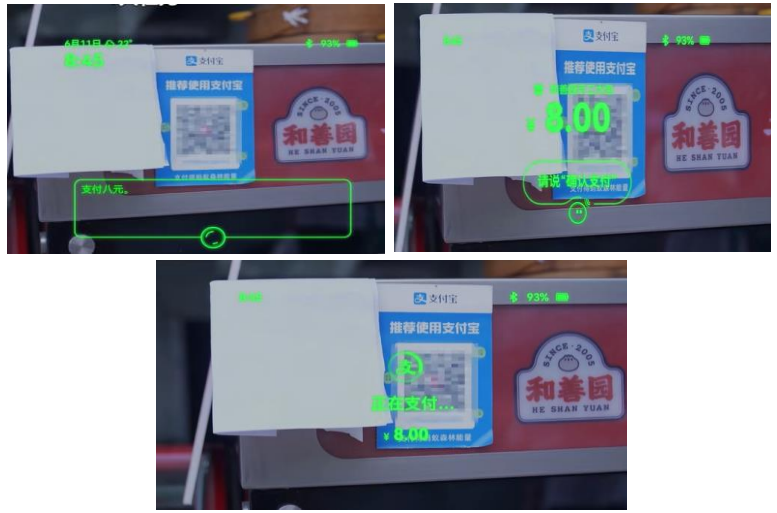
3.3.3 智能眼镜应用生态完善

- 在眼镜硬件形态持续迭代的同时，产品的应用生态也在不断完善：
- **导航功能：**高德地图智能眼镜解决方案于2025年6月全新上线，与Rokid、雷鸟创新、逸文、星纪魅族等厂商合作重新定义未来出行，具备路线规划、实时导航、建筑物识别、地点快查、景点讲解、语音问答、视觉定位、AR导航等功能，致力于构建智慧生活，满足用户的日常出行需求。
- **翻译功能**：Ray-Ban Meta、Rokid Glasses等眼镜都配备实时翻译功能，其中Rokid Glasses眼镜通过麦克风拾取对话内容，进行翻译后直接投射在虚拟屏幕上，避免了人工翻译打断发言者的情况，能让双方的对话体验更流畅。此外，做演讲时，具备字幕的Rokid Glasses也可成为眼前的提词器。
- **支付功能：**星纪魅族携手蚂蚁国际 Alipay+，实现 AR 技术与支付场景深度融合，填补智能眼镜行业支付领域长期空白。该功能已搭载于 StarV Air2 智能眼镜，能为用户带来更智能高效的支付体验。同时，Rokid Glasses眼镜合作支付宝，推出AR眼镜的支付解决方案。

图：高德地图开放平台智能眼镜解决方案

图：实时语音翻译

图：支付功能



资料来源：高德地图微信公众号，民生证券研究院

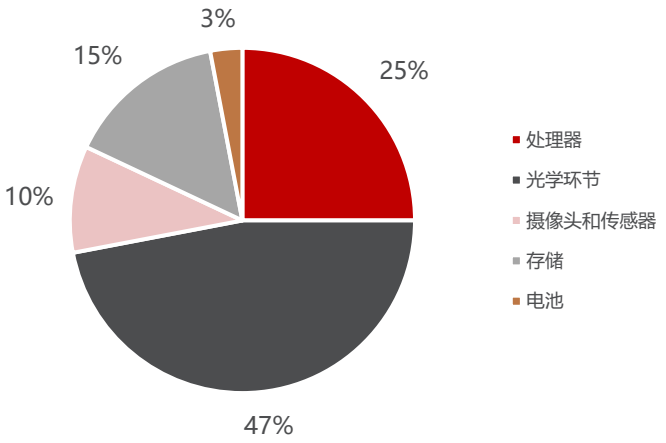
资料来源：量子位，民生证券研究院

资料来源：腾讯网，民生证券研究院

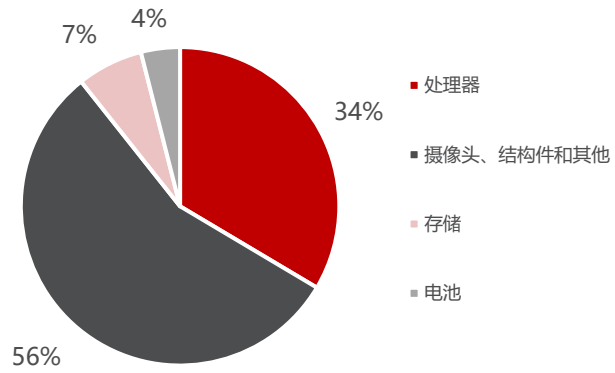
3.3.4 AI和AR眼镜硬件差异对比

- **AI和AR眼镜功能上的差异决定了二者核心硬件差异**，其中SoC是AI眼镜硬件的关键，光学则是AR眼镜硬件的核心问题。
- **AI眼镜的硬件需求主要集中在SoC处理器上。**通过在传统眼镜上配备摄像头、麦克风等传感器，将其智能化改造，然后将传感器收集到的信息传输给内置的SoC处理器进行分析和处理。根据Wellsenn XR数据，在Ray Ban Meta眼镜的成本拆解中，SoC处理器是占比最大的单一硬件结构，占比约34%。
- **相比而言，AR眼镜的硬件需求则相对较高，目前其光学显示组件无法做到满足性能需求的同时实现轻量化量产。**除了需要SoC+传感器支持复杂的图像处理和虚拟信息叠加功能外，AR眼镜还需要显示技术+光学组件来创建虚拟图像，光学组件的集成度和功耗会直接影响AR眼镜的重量和性能。根据智东西数据，以价格$1,000的Hololens开发者版本为例，光学显示模组占比最大（Lcos投影设备$180和透明全息透镜$290），达到总成本的47%；处理器约占总成本25%。由此可见，光学显示模组是AR眼镜硬件中的价值高地，而各厂家对光显方案技术的追求也一直在前进中。

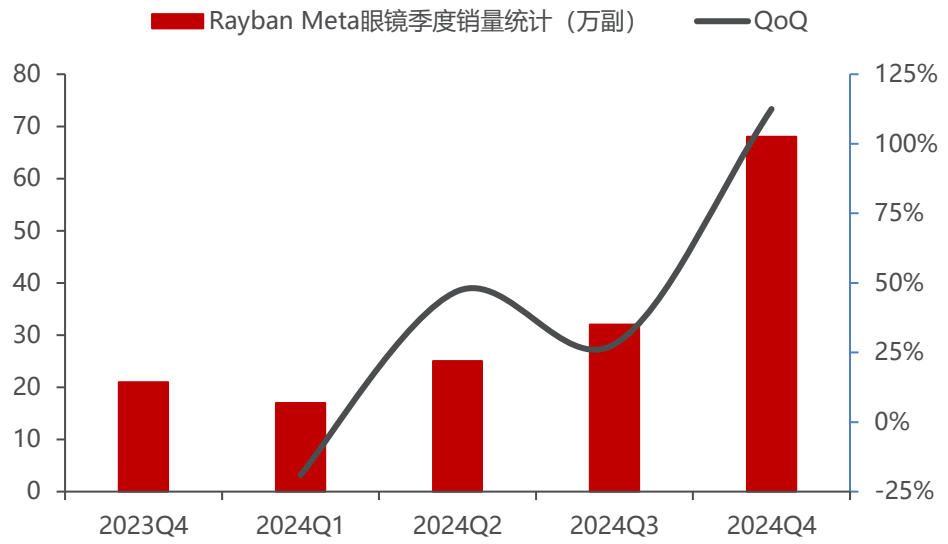
图：AR眼镜Holoens开发者版本硬件成本拆解



图：AI眼镜Ray-Ban Meta硬件成本拆解



图：Ray-Ban Meta眼镜季度销量统计（万副）



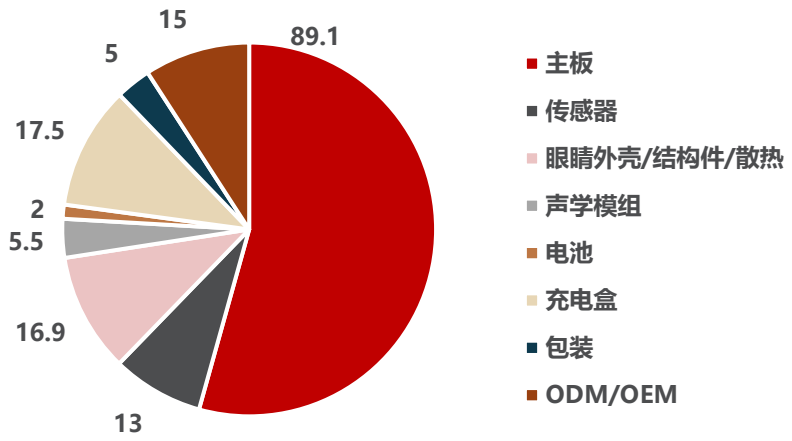
资料来源：智东西，民生证券研究院整理
资料来源：维深wellsenn XR，民生证券研究院整理

资料来源：维深wellsenn XR，民生证券研究院整理

3.3.4 AI眼镜硬件方案

- 拆解AI眼镜的硬件结构=**主板+传感器+眼睛外壳/结构件/散热件+声学模组+电池+充电和+包装+OEM/ODM**，其中主要的核心硬件集中在主板和传感器上。目前，AI眼镜仍属于轻量级产品，硬件结构较为简单，其中SoC成为影响功能体验和成本的关键因素。
- **当下AI眼镜的SoC芯片需具备三种核心能力：1）低功耗。**解决高重量+低续航的核心痛点；**2）具备连接能力。**蓝牙+WiFi集成是趋势，部分需求蜂窝能力；**3）芯片处理能力。**根据音频、图像、视频等应用场景需求渐次升级，但尚不需要运行端侧模型。
- **因此，当下主流的SOC方案有三种：**1）高通XR、AR系列芯片；2）手表方案移植；3）SOC+外挂ISP方案。当前阶段，国产SoC可依靠国内成熟的硬件供应链生态，**凭借优秀的性能+低成本方案，迅速推进国产替代的行业逻辑。**

图：Rayban-Meta BOM成本拆解（按结构）



图：AI眼镜SoC核心迭代趋势

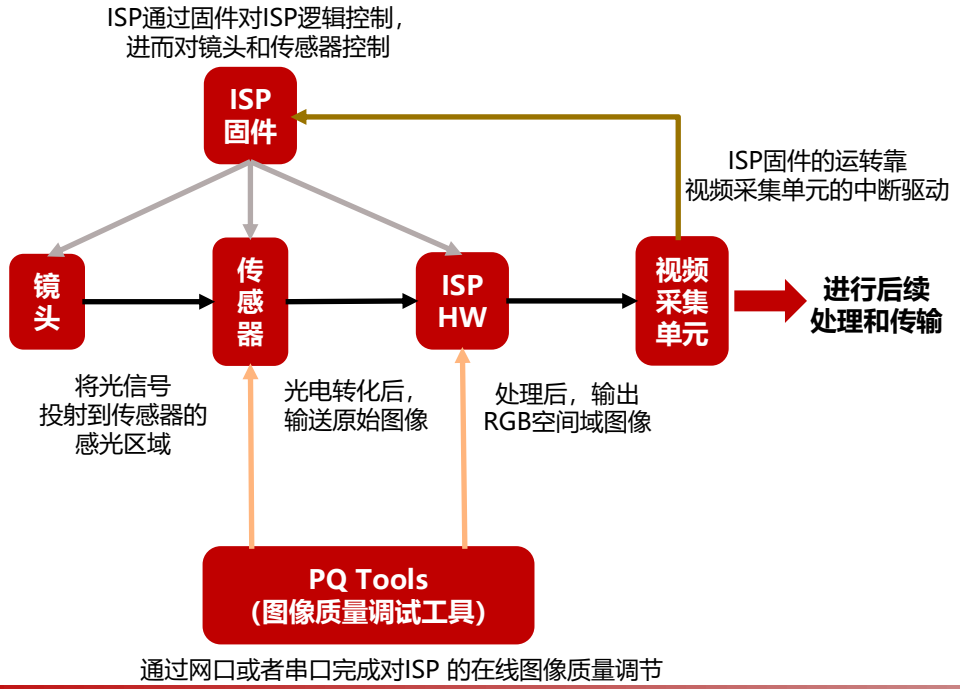
资料来源：维深wellsenn XR，民生证券研究院整理

资料来源：维深wellsenn XR，新浪新闻，恒玄科技官网等，民生证券研究院整理

3.3.4 AI眼镜硬件方案

- **全新的视觉交互方式下，AI眼镜的拍摄需求日益凸显。**连接和音频是早期耳机SoC时代的关键，但在全新的交互模型引领下，AI眼镜在视觉等交互方面的能力逐渐完善。因此，集成ISP等多模态处理能力成为AI眼镜SoC发展的重点方向。
- 以Rayban-Meta眼镜为例，产品热销可能得益于其3A的独到之处，即自动曝光、自动对焦、自动白平衡，由于AI眼镜无法像手机一样通过人手实时操控调整对焦，所以智能化的物体识别效果成为影响体验的关键。**在此过程中，ISP芯片的视觉处理能力发挥了至关重要的作用。**
- 分析未来眼镜硬件的变化趋势，我们认为ISP芯片渗透率和性能需求将会逐步提升，ISP能力将可能成为芯片厂商的决胜要素。

图：ISP处理过程



图：ISP能力成为眼镜提升交互体验的关键

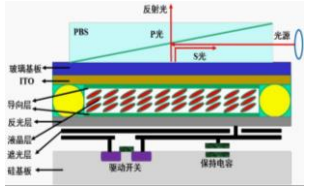
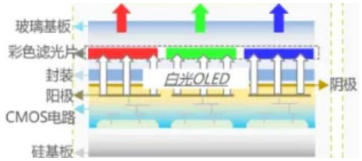
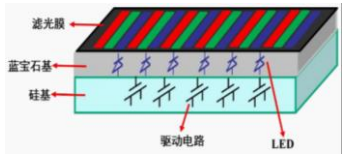
资料来源：电子工程专辑，民生证券研究院整理

资料来源：电子工程专辑，民生证券研究院整理

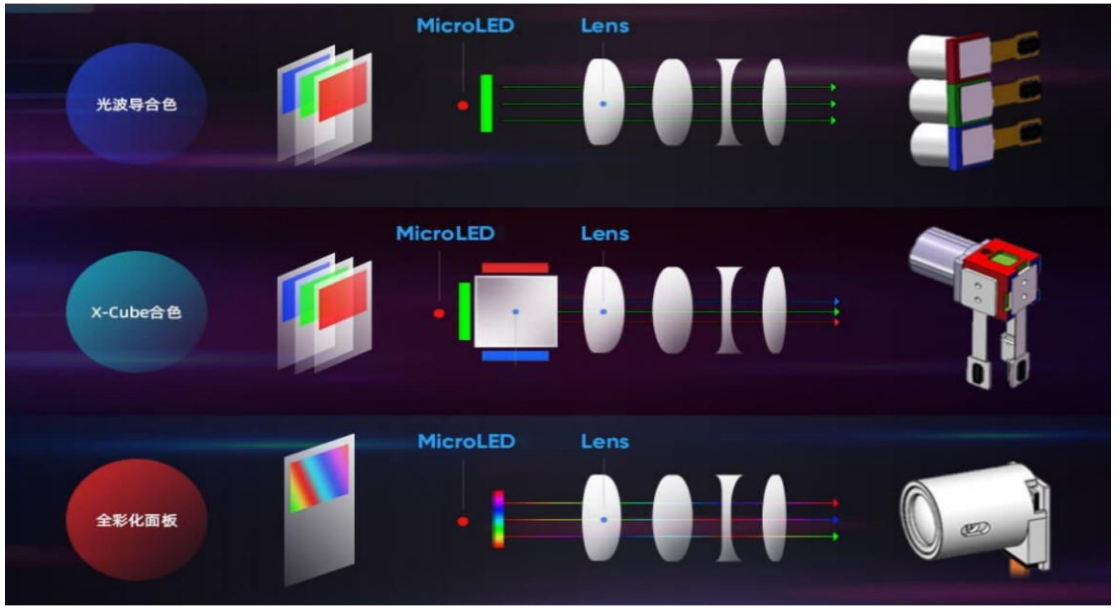
3.3.4 AR眼镜硬件方案

- **AR眼镜成本中光学显示单元和计算单元的占比最大。**AR整机的零部件可归纳为光学显示、计算、感知、存储和电池五大功能模块。根据微软Hololens BOM拆解，各模块占比为：光学显示 47%、计算 25%、感知 10%、存储 15%、电池 3% 。
- **近年来，随着AR产品创新迭代，光显模块的重要性还在逐步提升：**以华为22年发布的产品vision glass为例，其Micro OLED+BirdBath 的光学显示成本占比72%，对比16年的微软hololens开发者版本（光显占比约47%），光显成本持续显著提升。
- 光学显示=光机+镜片+显示屏，光机和镜片是光显方案中最核心的部分，其中**Micro LED有望凭借低功耗、微型化等优点成为光机的终局方案。**

表：LCoS、Micro OLED、Micro LED三类微显示技术对比

	LCoS	Micro OLED	Micro LED
结构示意图			
原理	反射式微液晶显示	有机自发光	无机自发光
反应时间	毫秒	微秒	纳秒
工作温度	0 - 50℃	-50 – 70℃	-100 - 120℃
优势	模组体积小，成本低，解析度高，色域广，高分辨率	响应速度快，功耗低，体积小，高对比度，延展性好	响应速度更快，功耗低，体积小，亮度高，寿命长
劣势	响应速度慢，功耗高，对比度低	亮度低，成本高	灵活性差，成本高，量产难
技术成熟度	实现规模量产	小规模量产	小规模试产

图：Micro LED的三类彩色化方案



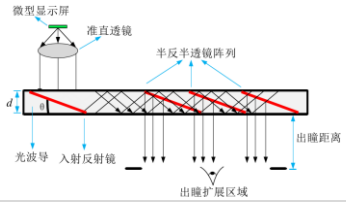
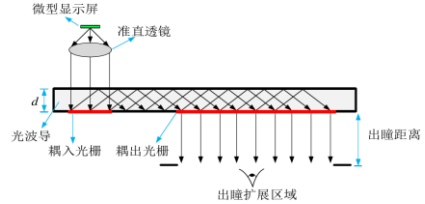
资料来源：艾邦VR产业资讯，艾瑞咨询，《硅基液晶（LCoS）微显示技术》梁宇华等，民生证券研究院整理

资料来源：艾邦AR/VR网，民生证券研究院

3.3.4 AR眼镜硬件方案

- 为了提供更舒适、沉浸式的增强现实体验，具有体积/透光率/清晰度等优势的光波导显示技术逐步被AR眼镜采用，成为主流AR光学方案。目前多种光波导方案并行，综合考虑**成本及光学模组+光机（Micro OLED）匹配效果，衍射光波导为终局方案可能性更大**，其中光波导片+纳米压印的设备供应链具备一定投资机遇：1）光学：水晶光电、蓝特光学、美迪凯、舜宇光学科技、歌尔股份；2）设备：苏大维格
- **材料方面，SiC成为衍射光波导方案的新理想基。**从特性来看，SiC材料具备高折射率、宽FOV等特点，有助于优化眼镜的显示、降低重量、散热改善等。

表： AR核心部件光波导方案对比及其供应链

	阵列光波导	表面浮雕光栅光波导	体全息光波导
原理图			
制造工艺	传统光学冷加工镀膜/贴合/切割	半导体微纳加工、纳米压印	激光全息干涉
耦入元件	反射镜	表面浮雕光栅(SRG)	体全息光栅(VHG)
耦出元件	半透半反镜面阵列	SRG	VHG
厚度	< 2mm	< 2mm	
视场角	25°-70°	40°-50°	
透光度	90%以上	80%以上	
显示器件	LCOS / Micro OLED / Micro LED	DLP / Micro LED	
光效	6%-15%	0.3%-1%	
技术优势	设计原理简单；显示性能极佳	量产性和良率更优，可实现二维扩瞳	工艺效率高；量产投资小
技术痛点	制作工艺繁琐；量产难度较大	光损严重；存在彩虹效应	量产难度大、稳定性差
技术发展	分子键合技术	纳米压印技术；多层波导技术	
代表厂商	Lumus、水晶光电、珑璟光电、灵犀微光、理湃光晶、GodView	Microsoft、Magic Leap、Vuzix、WaveOptics、驭光科技(歌尔收购)	DigiLens、Sony、三极光电、Akonia(苹果收购)、谷东科技

资料来源：Wellsenn XR，VR陀螺，艾瑞咨询，理湃光晶，《衍射光波导近眼显示方案的设计》 韦信宇等，民生证券研究院整理

3.3.4 供应链标的一览

- 2024年，中国 AR 眼镜市场全年销量达到 28.6 万台，特别是24Q4，随着更多品牌新品入市，销量创下新高。
- 以星纪魅族、Rokid 、XREAL 、雷鸟创新、INMO为代表的AR眼镜五小龙已形成国内市场的主要竞争格局，未来有望通过性价比+生态应用重塑全球市场格局。

终端品牌
小米集团
漫步者

渠道商
博士眼镜
明月镜片
孩子王

方案商
润欣科技
移远通信
广和通
美格智能

代工
歌尔股份
国光电器
天键股份
佳禾智能
亿道信息

SiC
天岳先进

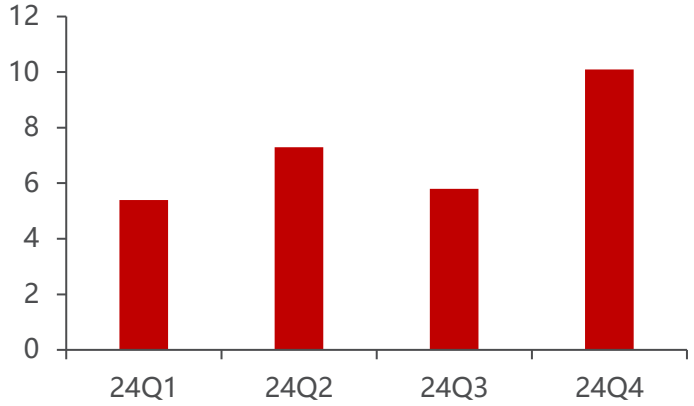
零部件
福立旺

光学
舜宇光学
蓝特光学
水晶光电

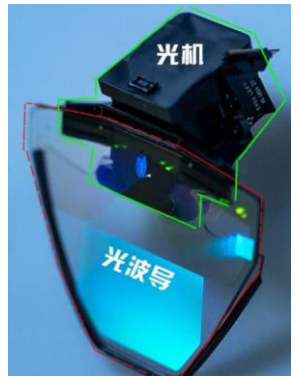
存储芯片
德明利
兆易创新
佰维存储
普冉股份
东芯股份

数字芯片
恒玄科技
星宸科技
乐鑫科技
炬芯科技
全志科技
泰凌微
中科蓝讯
翱捷科技
瑞芯微
富瀚微

图：中国AR眼镜季度销量（万台）



图：华为AR vision glass的光机



图：光波导AR适配的同光机的整体变化

2015年	2019年	2021年	2021年
LCoS	LBS	Micro OLED	Micro LED
10-30万nit	2-10万nit	3000-6000nit	10万-100万nit
单目0.4K-0.6K	单目1K	单目0.4K-0.6K	单目0.6K-1K
60-120Hz	60-90Hz	60-90Hz	60-120Hz
2000-6000ppi	1200-2000ppi(等值)	3000-5000ppi	3000-6000ppi
阵列/衍射光波导	衍射/阵列光波导	阵列光波导	衍射/阵列光波导
HoloLens 1 Magic Leap One等	HoloLens 2	INMO Air等	INMO Go 雷鸟X2等

资料来源：维深wellsenn XR，民生证券研究院整理

04 投资建议

核心观点

我们于去年发布了AI深度报告，**提出了“CSP引领、ASIC为王”的投资观点。**而在本篇2025年中期策略报告中，我们将进一步解读算力的长期成长空间，探寻GPU和ASIC的最新动向，并挖掘国产算力和AI终端的新变化。

海外算力：近期美股AI软硬件标的纷纷创新高，**短期催化为英伟达业绩超预期，而长期驱动力则为AI赋能互联网应用，带动推理需求增长+Token消耗提升，以此为锚点，实现AI投资的ROI闭环。**算力需求高增的背景下，英伟达产品加速迭代，而CSP自研的ASIC则迎来了更快的成长。**算力的升级离不开功率和速率两条路线：1）速率方面，**我们看到了ASIC和服务器架构变化带来的PCB升级，传统光模块向CPO的演进，以及AEC的渗透率快速提升；**2）功率方面**，HVDC、超级电容为下一代服务器提供供电保障，液冷则成为芯片功耗提升下的刚需。而在英伟达的ComputeX大会上，各大厂商展示了具体的产品升级路线，也为未来几年算力行业的发展敲定了大方向。

国产算力：豆包+DeepSeek破局，国产大模型弯道超车：豆包和DeepSeek分别在多模态和轻量化两方面加速了国产大模型的发展进程。国内其他模型厂商也加速了追赶节奏，2025年以来，豆包、通义千问、百度、腾讯混元、阶跃星辰和Kimi等其他国产大模型加速了更新迭代，AI应用加速放量下推理侧需求有望提升。**算力基建加码，解决供给短板：**国内云计算厂商正加大算力储备及模型优化投入，AI计算基础设施建设布局逐步清晰，相关资本开支进入新一轮扩张周期。而短期内，国产算力基建难以满足迅速增长的需求，算力租赁成为破局之道。**向“芯”而行，国产算力破局元年：**在国产大模型密集落地背景下，芯片厂商加速适配国产算力生态。中芯国际N+1工艺已逐步成熟，N+2持续推进，构建国产算力底座；昇腾910C量产落地，920系列研发加快，性能持续逼近国际主流水平；寒武纪、海光等在AI训推方向深度布局，硬件端多点突破，生态融合加快。云端ASIC正成为算力演进主流，芯原等设计企业快速成长，与海内外头部厂商形成紧密合作，成长弹性充足。

核心观点

AI终端：手机AI功能仍待完善，但仍有光学、折叠屏、指纹识别等硬件的结构性创新。智能眼镜市场近期不断升温，销售火热。复盘AI/AR眼镜的发展历程，Meta&Rayban AI眼镜的成功证明了“先眼镜后智能“的思路，即AI眼镜替代传统眼镜，然后在AI眼镜上增加AR效果使消费者逐步提升对眼镜这类新终端的接受度。**我们认为从AI向AR演进，品牌厂商与光学方案商深度绑定，光学/显示逻辑有望得到充分演绎：**①AI眼镜交互模式和功能比较单一，AR眼镜加入显示功能，能显著提升用户体验；②光学显示模块成为AR眼镜中BOM占比最高的部分之一，且相较于AI眼镜是纯增量环节，目前主流方案是MicroLED+衍射光波导。

我们坚定看好AI产业的长期叙事，**英伟达持续强势，云厂商崛起，国产算力突破的当下，投资机遇也会更加多元化。**具体到细分赛道，算力链重点关注服务器、PCB、CPO、铜缆、电源、液冷等产业链，这也是国内企业深耕多年，具备优势的环节。而AI终端则相对预期充分，需观察热门新品的放量节奏。

建议关注：1)服务器：工业富联、华勤技术；**2)算力芯片：**芯原股份、寒武纪、海光信息；**3)PCB：**沪电股份、胜宏科技、广合科技、生益科技、景旺电子、威尔高；**4)铜/光互联：**瑞可达、博创科技、太辰光、东山精密；**5)电源及温控：**禾望电气、中恒电气、麦格米特、申菱环境、江海股份；**6)品牌及代工：**小米集团、影石创新、歌尔股份、国光电器；**7)SOC：**乐鑫科技、恒玄科技、星宸科技； **8)存储：**兆易创新、普冉股份； **8)渠道商：**博士眼镜、孩子王、明月镜片等。

05 风险提示

5.1 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 宋晓东
执业证号：S0100523110001
邮箱：songxiaodong@mszq.com

分析师 卢瑞琪
执业证号：S0100524090002
邮箱：luruiqi@mszq.com

民生证券研究院：

上海：上海市虹口区杨树浦路188号星立方大厦7层； 200082
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：深圳市福田区中心四路1号嘉里建设广场1座10层 01室； 518048

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：