电子行业2026年年度投资策略

从星星之火到全面燎原的本土硬科技收获之年

行业研究 · 行业投资策略

电子

投资评级：优于大市（维持）

证券分析师：胡剑	证券分析师：胡慧	证券分析师：叶子	证券分析师：张大为	证券分析师：詹浏洋	证券分析师：李书颖
021-60893306	021-60871321	0755-81982153	021-61761072	010-88005307	0755-81982362
hujian1@guosen.com.cn	huhui2@guosen.com.cn	yezi3@guosen.com.cn	zhangdawei1@guosen.com.cn	zhanliuyang@guosen.com.cn	lishuying@guosen.com.cn
S0980521080001	S0980521080002	S0980522100003	S0980524100002	S0980524060001	S0980524090005

从星星之火到全面燎原的本土硬科技收获之年

国信证券 GUOSEN SECURITIES

- **2025年AI产业链在业绩趋势中从分歧走向共识，2026年有望成为本土硬科技收获之年。**如我们在2023年年度策略报告《在春寒料峭中枕戈待旦》所述，电子行业的景气周期自2021年下行近2年，于2023年下半年筑底回升，以华为Mate系列回归为标志性事件，如今仍处在由AI创新所拉动的温和上行过程中；如我们在2025年年度策略报告《AI革新人机交互，智能终端百舸争流，行业迈入估值扩张大年》所述，行业在“宏观政策周期、产业库存周期、AI创新周期”共振上行的过程中、在被动基金快速扩容的助力下，呈现出显著的估值扩张趋势。在经历了2Q25由Deepseek兴起所引致的“算力通缩”的叙事逻辑冲击以及由美国所发起的关税战冲击之后，3Q25以来，行情在AI产业链亮眼的业绩趋势中从分歧走向共识，截至12月16日，电子上涨40.22%，位居全行业第三位。展望2026年，我们认为，AI大模型的推理能力仍在持续迭代，大模型与端侧应用的闭环正在形成，算力+存力硬件层面供不应求的态势仍将延续，而国内先进制程的扩张进度和自主可控的推进速度仍有较大预期差。我们认为，在2020年全面开启的5G创新周期中已有冒尖趋势的中国科技产业，在国内工程师红利的支撑下，在直面美国“制裁政策”、经历了逾5年的“人财物”快速累积之后，正在新一轮AI创新周期中体现出更强的全球竞争力，也正在迎来世界范围的重新认知，2026年有望成为“从星星之火到全面燎原的本土硬科技收获之年”。

- **AI大模型群雄逐鹿，英伟达引领算力迭代，PCB、服务器产业链延续高增长。**得益于大模型在架构上的创新：采用混合专家架构通过稀疏化实现更高效的推理，采用创新的注意力机制降低计算复杂度与内存需求，深度思考模式下多轮推演以减少幻觉等，国内外大模型在多模态理解、推理及AI应用层面均实现持续进阶。受益于CSP、主权云等算力需求扩张、以及AI推理应用的蓬勃发展，TrendForce预计2026年全球八大CSP合计资本支出将增长40%达到6000亿美元+，全球AI服务器出货量将增长20.9%。与此同时，继2025年由GB200向GB300的升级之后，2026年英伟达新一代Rubin架构AI服务器将为分离式推理带来革命性变化，英伟达预计在26年底前，Blackwell和Rubin系列GPU总出货将达2000万颗，合计订单将达5000亿美元。基于算力军备竞赛的市场规模扩容以及算力产品迭代所带来的ASP提升，伴随Scale Up与Scale Out所带来的智算集群扩展，我们认为，2026年深度参与全球产业链分工的PCB、服务器产业链，包括相关的上游材料及液冷散热等环节都将迎来量价齐升的高速成长期，关注：工业富联、胜宏科技、生益科技、华勤技术、沪电股份、蓝思科技、立讯精密、鹏鼎控股、东山精密、景旺电子。

- **算力+存力：国产算力通用芯片与ASIC方案齐发力，存力缺货涨价有望贯穿全年。**算力方面，当前国产芯片积极更新迭代，华为计划1Q26推出昇腾950PR，4Q26上市超节点Atlas 950 SuperPoD；寒武纪、沐曦、壁仞、摩尔线程等国产卡顺利导入智算中心。同时，受限于美国BIS多次制裁国产芯片，CSP大厂的合规ASIC项目将同步迎来发展机遇，其中非一线云厂的自研项目有望为国内ASIC厂商带来可观增量，而存储方案也将成为ASIC差异化竞争的关键点。存力方面，AI时代的DRAM逐步从“附属角色”转变为“性能瓶颈突破口”，在供给侧结构调整叠加AI需求拉动下将加速成长，预计2026年DRAM位元需求量有望同比增加26%；与此同时，随着AI推理兴起，传统的HDD在读写速度、响应延迟及能效方面的局限性加速了SSD渗透，NAND缺货态势从局部应用蔓延至全盘，价格指数自25年9月至12月已上涨超40%，我们认为，26年DRAM及NAND仍将呈现较严重的供不应求，价格有望延续涨势。算力及存力相关产业链关注：寒武纪、翱捷科技、芯原股份、德明利、江波龙、兆易创新、北京君正、伟测科技。

从星星之火到全面燎原的本土硬科技收获之年

国信证券 GUOSEN SECURITIES

- **运力+电力：运力成为算力提升的重要突破口，算力增长推动电源架构同步升级。**运力环节既要解决数据进出内存的问题，又要实现服务器内部、机架之间以及集群之间的顺畅通信，在国内高端算力芯片流片受限的背景下，运力环节的优化成为重要突破口，预计2024-2030年全球高速互连芯片市场规模CAGR为21.2%，中国市场的占比将由25%提高至30%，为MRDIMM内存、PCIe互连芯片、CXL互连芯片、硅光芯片、OCS等产业链创造广阔增量市场，关注澜起科技、蓝特光学。与此同时，随着数据中心芯片功耗的增大，机架处理功率水涨船高，英飞凌预测单GPU的功耗将呈指数级增长，到2030年达到约2000W，机架的峰值功耗将达到300kW以上。一方面，电能转换效率的提升对降低数据中心的运营成本至关重要；另一方面，机架侧大幅、快速波动的功率曲线对公共电网的稳定性构成挑战，因此要求供电方案向HVDC方向发展，SST、DrMos及SiC、GaN器件将成为AI电源的核心方向，关注：天岳先进、新洁能、江海股份、斯达半导、芯联集成、华润微、扬杰科技、杰华特、晶丰明源、顺络电子。

- **AI端侧：AI Agent重塑交互范式，大厂争先布局端侧入口，消费电子创新大年开启。**随着大模型在多模态理解、通用推理与任务执行能力上的持续演进，AI正由工具型能力升级为能够理解用户意图并自主执行任务的AI Agent，端侧消费电子产品是AI商业化闭环的关键承载层，有望系统性重构人机交互范式。在此背景下，手机、眼镜、耳机以及家庭机器人等多种终端形态，有望围绕AI Agent构建协同网络，推动AI从单点功能升级迈向跨场景、跨终端的系统级体验。语音、视觉及环境感知等多模态输入的重要性提升，对端侧算力、感知能力与连接能力提出更高要求。我们认为，当前端侧相关技术与产业基础已趋于成熟，商业模式的关键突破有望形成“非线性放大效应”。展望2026年，从年初的CES到年中的WWDC，亦或字节、苹果等头部厂商的持续探索，叠加潜在模型厂商及互联网厂商入局硬件的预期，均可能成为引爆市场情绪与产业投资共识的关键催化，关注：恒玄科技、晶晨股份、蓝思科技、立讯精密、蓝特光学、歌尔股份、翱捷科技、乐鑫科技、华勤技术、传音控股、小米集团、顺络电子。

- **半导体：推荐进程有望超预期的自主可控产业链，以及在景气复苏阶段加速国产替代的模拟芯片。**据SIA数据，2024年中国占全球半导体销售额的28%，但本土供应比例仅4.5%，自给率仍偏低，且由于增量主要来自GPU、HBM等云侧增量，自给率较2023年有所降低。但A股半导体公司的财务表现持续改善，据我们统计的146家公司的经营数据，季度收入最高值落在2025年的占比54%，3Q25 SW半导体板块整体毛利率处于1Q21和2Q21之间，净利率与4Q20、1Q21水平相当。从行业周期来看，全球半导体销售额已连续八个季度同比增长，25年12月WSTS再次上修了对2025和2026年的预测值，预计2024-2026年全球半导体将实现连续3年两位数增长。从国内半导体产业来看，除了AI增量外，国内芯片设计企业崛起和在地化制造需求为自主制造链提供增量，重点关注晶圆代工、先进封装和上游半导体设备材料环节；另外，模拟芯片在半导体产品品类中周期靠后，国际大厂TI、ADI 2025年收入开始同比转正，标志着行业进入复苏阶段，国内企业近几年推出的新品有望进入规模放量阶段，长期来看AI数据中心以及自动驾驶、人形机器人等AI应用均为其带来广泛增量，同时模拟芯片也是国产化空间较大的细分，将持续受益国产化率提高。关注：中芯国际、华虹半导体、杰华特、思瑞浦、北方华创、中微公司、拓荆科技、圣邦股份、南芯科技、伟测科技、通富微电、长电科技、鼎龙股份、骄成超声。

- **重点关注组合：**中芯国际、工业富联、中微公司、翱捷科技、华虹半导体、寒武纪、北方华创、澜起科技、蓝思科技、蓝特光学、立讯精密、胜宏科技、华勤技术、恒玄科技、顺络电子、沪电股份、伟测科技、晶晨股份、鹏鼎控股、思瑞浦、杰华特、圣邦股份、德明利、江波龙、佰维存储、歌尔股份、兆易创新、鼎龙股份、京东方A、长电科技、水晶光电、传音控股、豪威集团、骄成超声、海康威视、小米集团、天岳先进、扬杰科技、乐鑫科技、江海股份、电连技术、晶丰明源、光弘科技、南芯科技、世华科技、北京君正、洁美科技、国芯科技、通富微电、唯特偶、福立旺。

- **风险提示：国产替代进程不及预期；下游需求不及预期；行业竞争加剧的风险；国际关系发生不利变化的风险；行业周期性波动风险；生产设备及原材料供应风险。**

目录

【1】2025年行情回顾：AI在业绩趋势中从分歧走向共识

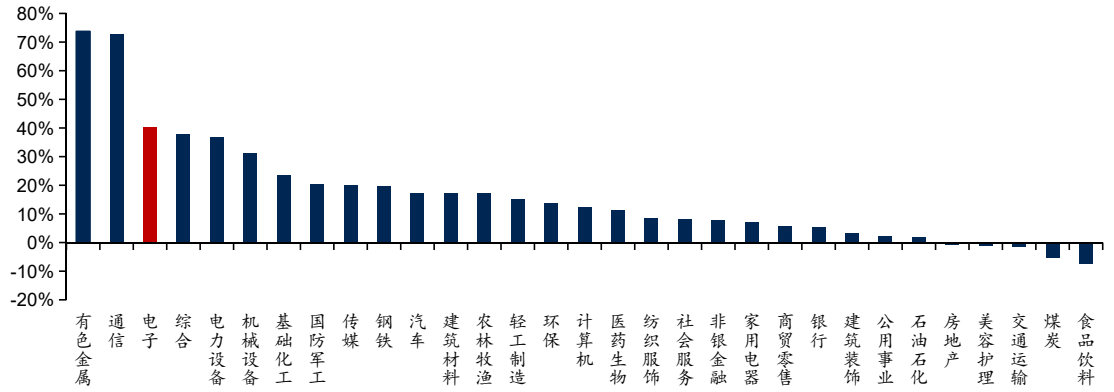
2025年行情回顾：AI在业绩趋势中从分歧走向共识

国信证券 GUOSEN SECURITIES

- 2025年初至12月16日，上证指数、深证成指、沪深300分别上涨14.11%、24.01%、14.30%。电子行业整体上涨40.22%，涨跌幅位居全行业第三位，其中元件、电子化学品、其他电子、消费电子、半导体、光学光电子分别上涨93.19%、46.88%、40.63%、39.40%、38.37%、7.66%。恒生科技指数、费城半导体指数、台湾资讯科技指数分别上涨20.91%、39.73%、27.89%。

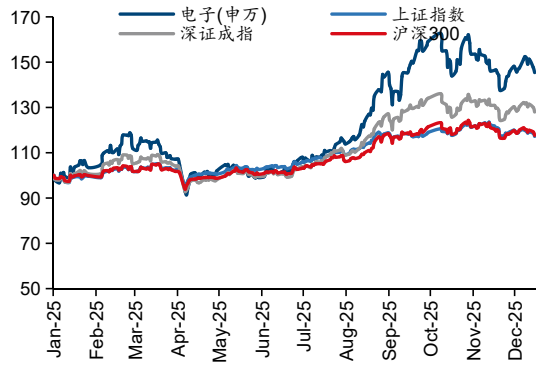
- **2025年1-2月，**延续2024年末“字节火山引擎大会”所引燃的AI端侧创新预期，端侧SoC作为半导体中盈利预期改善力度居前的细分方向，带动半导体及消费电子上涨。春节期间，Deepseek大模型受到全球范围内的认可与应用，中国科技企业在世界AI创新进程的参与度和贡献度再度受到广泛关注与认知，本土半导体、尤其是算力自主可控相关方向估值扩张明显，期间电子上涨8.02%，其中半导体、消费电子、元件上涨10.52%、8.60%、5.95%。
- **2025年3-5月，**美国大幅加征关税对全球经贸结构及资本市场造成较大冲击，全球贸易保护主义兴起，资本市场波动加大。一方面受中美关税博弈反复，美国加大对国内半导体技术限制的影响；一方面受Deepseek兴起后“算力通缩”的叙事逻辑冲击，市场对于AI硬件投入的持续性产生担忧；叠加基金行业新规所引致的超配行业减持压力等影响，期间电子下跌11.55%，其中消费电子、半导体、光学光电子分别下跌18.97%、11.02%、10.47%。
- **2025年6-10月，**在AI算力需求拉动下，全球逻辑类芯片月销售额同比增速大幅扩张，台积电及多家咨询机构先后上修全年及26年全球半导体增长预期，“算力通缩”预期被证伪。与此同时，AI基建相关的光模块、PCB、服务器环节业绩全面超预期，进而传导到存储环节的全面紧缺及大规模涨价，“存储超级周期”的预期强化；此外，叠加中美经贸谈判态势向好，与AI端侧创新关联的消费电子迎来底部反弹。在此期间，行业高景气及估值扩张弹性延续，电子上涨60.68%，其中元件、消费电子、半导体上涨99.15%、80.54%、58.51%。
- **2025年11月以来，**由于2Q25“中美关税战”抢出口透支部分订单需求、3Q25 3C消费补贴退坡以及存储缺货涨价等因素干扰，3Q25财报季反映出电子板块、尤其上游IC设计环节在高预期下的增势均有所回落，叠加3Q25机构持仓在TMT方向仍呈现较高的拥挤度，受基金考核新规下调仓预期的影响，板块市场热度有所降温，自10月至12月16日电子下跌8.66%，其中消费电子、半导体分别下跌12.26%、11.23%。

图：截至12月16日各行业涨跌幅



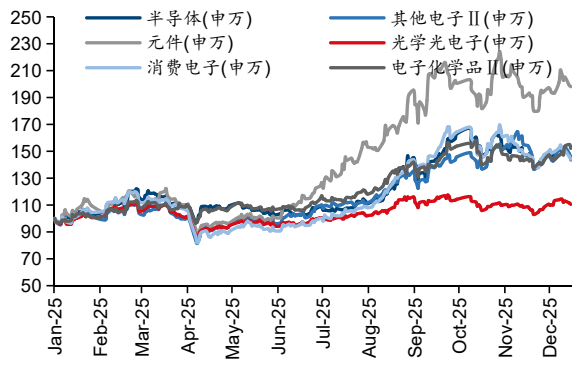
资料来源：Wind，国信证券经济研究所整理

图：2025年至今电子行业股价走势



资料来源：Wind，国信证券经济研究所整理

图：2025年至今电子各细分行业股价走势



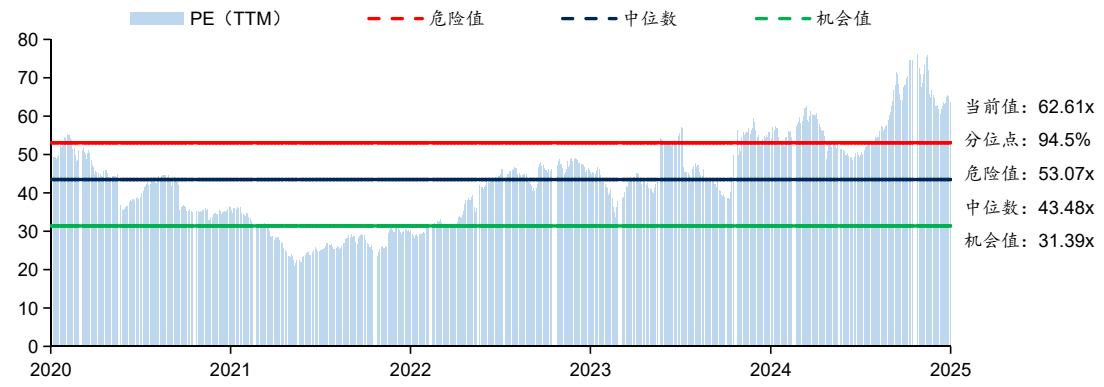
资料来源：Wind，国信证券经济研究所整理

2025年行情回顾：AI在业绩趋势中从分歧走向共识

国信证券 GUOSEN SECURITIES

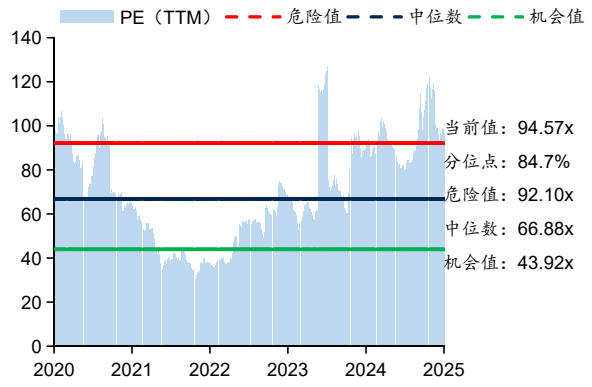
- 估值方面，截至2025年12月16日电子行业整体TTM PE（62.61x），处于近五年的94.5%分位；其中：半导体板块、消费电子、元件、光学光电子、电子化学品、其他电子TTM PE分别为94.57x、36.86x、53.78x、50.67x、70.50x、75.31x，处于近五年的84.7%、86.8%、94.6%、57.5%、96.0%、94.6%分位。

图：近五年电子(申万)PE(TTM)



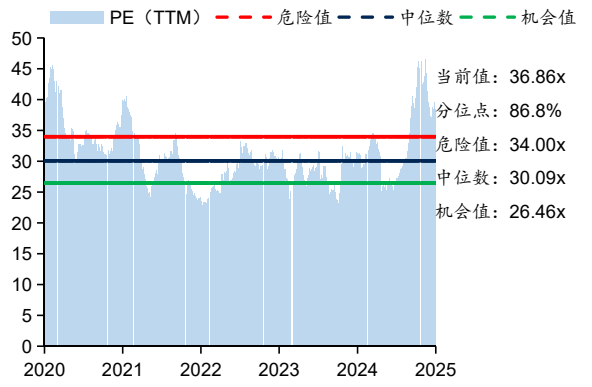
资料来源：Wind，国信证券经济研究所整理

图：近五年半导体(申万)PE(TTM)



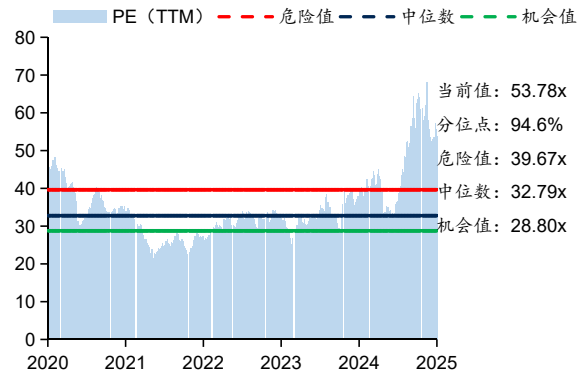
资料来源：Wind，国信证券经济研究所整理

图：近五年消费电子(申万)PE(TTM)



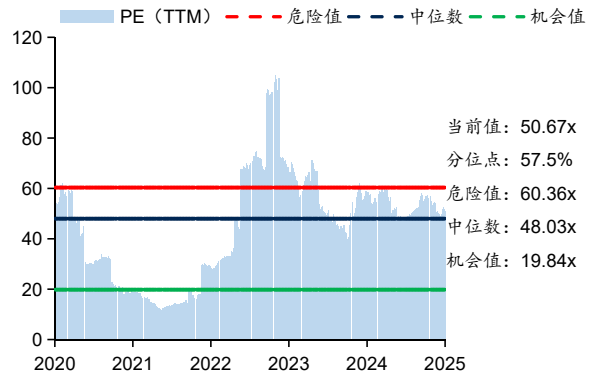
资料来源：Wind，国信证券经济研究所整理

图：近五年元件(申万)PE(TTM)



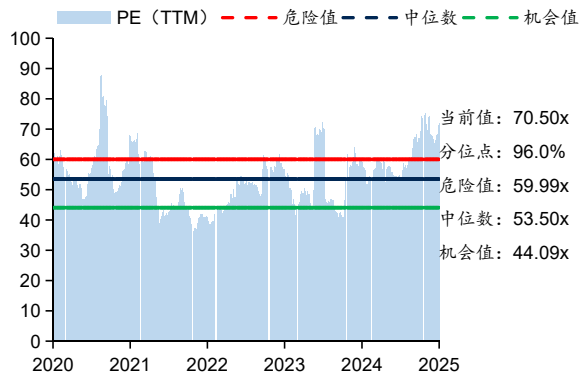
资料来源：Wind，国信证券经济研究所整理

图：近五年光学光电子(申万)PE(TTM)



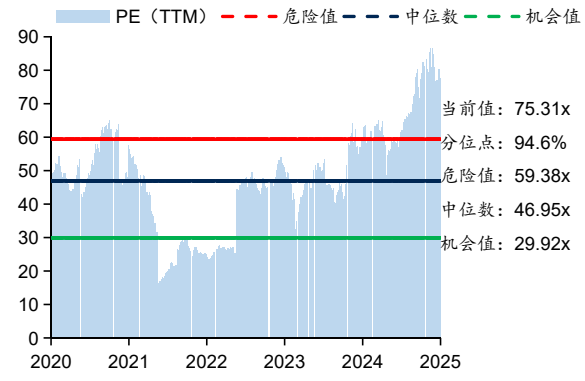
资料来源：Wind，国信证券经济研究所整理

图：近五年电子化学品(申万)PE(TTM)



资料来源：Wind，国信证券经济研究所整理

图：近五年其他电子(申万)PE(TTM)



资料来源：Wind，国信证券经济研究所整理

2025年行情回顾：AI在业绩趋势中从分歧走向共识

国信证券 GUOSEN SECURITIES

● 截至三季度末，公募基金电子板块重点持仓市值排行前五的公司分别是寒武纪-U、中芯国际、海光信息、澜起科技、立讯精密。从重仓持股的公募基金数目来看，三季度工业富联、寒武纪-U、立讯精密获得较多公募基金增持；小米集团-W、京东方A、TCL科技遭到较多公募基金减持。

图：2025年第三季度公募基金电子板块重仓持股TOP20

排名	公司代码	公司名称	持仓市值(百万元)			重仓基金数(个)			持股占流通股比(%)		
			3Q25	2Q25	变化	3Q25	2Q25	变化	3Q25	2Q25	变化(pct)
1	688256.SH	寒武纪-U	71298	37897	33401	922	397	525	12.86	15.09	-2.23
2	688981.SH	中芯国际	55468	41743	13724	592	433	159	19.80	23.81	-4.01
3	688041.SH	海光信息	52368	36297	16071	568	405	163	8.92	28.98	-20.06
4	688008.SH	澜起科技	45186	23969	21217	357	270	87	25.49	25.53	-0.04
5	002475.SZ	立讯精密	43563	23556	20007	931	569	362	9.27	9.38	-0.11
6	601138.SH	工业富联	42578	3431	39147	659	126	533	3.25	0.81	2.44
7	0981.HK	中芯国际	37615	18941	18674	629	474	155	6.47	5.82	0.66
8	688012.SH	中微公司	36683	24054	12629	377	239	138	19.61	21.07	-1.47
9	1810.HK	小米集团-W	35897	38131	-2233	225	442	-217	3.38	3.25	0.13
10	300476.SZ	胜宏科技	27506	19732	7773	446	453	-7	11.38	17.23	-5.85
11	002371.SZ	北方华创	26911	24061	2850	463	488	-25	8.22	10.19	-1.97
12	603986.SH	兆易创新	25721	15019	10702	508	311	197	18.07	17.88	0.19
13	002463.SZ	沪电股份	24291	12269	12022	355	333	22	17.20	14.99	2.20
14	688521.SH	芯原股份	16699	9404	7295	263	176	87	18.43	19.62	-1.19
15	002384.SZ	东山精密	15577	5920	9657	222	162	60	15.71	11.31	4.41
16	002916.SZ	深南电路	13832	3136	10696	161	86	75	9.60	4.37	5.23
17	603501.SH	豪威集团	10770	11407	-637	146	198	-52	5.91	7.34	-1.44
18	600183.SH	生益科技	9640	2428	7212	256	141	115	7.45	3.40	4.06
19	688347.SH	华虹公司	9303	2468	6834	169	78	91	19.92	11.48	8.43
20	688099.SH	晶晨股份	8452	3811	4641	119	39	80	18.05	12.77	5.28

资料来源：Wind，国信证券经济研究所整理

2025年行情回顾：AI在业绩趋势中从分歧走向共识

国信证券 GUOSEN SECURITIES

- 截至2025年12月16日，沪（深）股通电子板块持仓市值排行前五的公司分别是北方华创、立讯精密、工业富联、豪威集团、澜起科技；2025年至今净流入金额排行前五的公司分别是北方华创、澜起科技、中微公司、沪电股份、安集科技；2025年至今净流出金额排行前五的公司分别是工业富联、生益科技、中芯国际、海光信息、三环集团。

表：2025年至今电子板块沪（深）股通持仓变化

排名	公司代码	公司名称	净流入金额（百万人民币）	沪（深）股通持股市值（百万人民币）			沪（深）股通持股占流通股比例（%）		
				2025/12/16	2024/12/31	变化(%)	2025/12/16	2024/12/31	变化(pct)
1	002371.SZ	北方华创	24,577	40,961	12,453	228.9%	24.8%	11.7%	13.0
2	002475.SZ	立讯精密	-1,225	24,511	18,517	32.4%	9.4%	10.1%	-0.7
3	601138.SH	工业富联	-9,685	21,723	13,008	67.0%	11.4%	18.1%	-6.7
4	603501.SH	豪威集团	318	17,724	14,576	21.6%	17.8%	17.3%	0.5
5	688008.SH	澜起科技	7,122	16,466	4,950	232.7%	14.6%	7.5%	7.1
6	688256.SH	寒武纪-U	1,635	15,291	6,779	125.6%	5.9%	5.1%	0.8
7	688012.SH	中微公司	4,754	15,232	6,659	128.7%	12.9%	8.3%	4.6
8	000725.SZ	京东方A	158	11,198	11,705	-4.3%	8.5%	8.3%	0.2
9	688041.SH	海光信息	-2,793	10,431	10,216	2.1%	6.0%	7.7%	-1.7
10	002463.SZ	沪电股份	2,528	9,614	3,532	172.2%	10.7%	6.9%	3.8
11	300476.SZ	胜宏科技	3	7,334	1,075	582.5%	4.3%	4.5%	-0.2
12	600183.SH	生益科技	-4,972	7,238	5,762	25.6%	9.9%	20.6%	-10.7
13	603986.SH	兆易创新	-2,354	5,994	4,828	24.1%	5.0%	7.5%	-2.5
14	002384.SZ	东山精密	1,329	5,372	1,193	350.4%	5.6%	3.3%	2.3
15	300866.SZ	安克创新	-485	4,722	4,670	1.1%	20.8%	23.6%	-2.7
16	002916.SZ	深南电路	1,619	4,391	1,355	224.2%	9.1%	5.9%	3.2
17	000100.SZ	TCL科技	229	4,310	4,498	-4.2%	5.4%	5.1%	0.3
18	002938.SZ	鹏鼎控股	1,677	4,008	1,633	145.4%	13.1%	7.1%	6.0
19	688019.SH	安集科技	2,425	3,759	686	447.7%	16.1%	5.5%	10.6
20	300433.SZ	蓝思科技	-334	3,724	3,110	19.7%	6.9%	7.6%	-0.7

表：2025年至今电子板块港股通持仓变化

公司代码	公司名称	净流入金额（百万港元）	港股通持股市值(百万港元)			港股通持股占流通股比例(%)		
			2025/12/16	2024/12/31	变化(%)	2025/12/16	2024/12/31	变化(pct)
1810.HK	小米集团-W	25,253	184,456	137,587	34.1%	17.3%	19.4%	-2.1
0981.HK	中芯国际	26,972	149,659	59,023	153.6%	29.5%	23.3%	6.2
1347.HK	华虹半导体	4,083	24,791	6,257	296.2%	21.5%	16.8%	4.7
2382.HK	舜宇光学科技	-2,210	12,308	14,984	-17.9%	17.1%	19.9%	-2.8
0285.HK	比亚迪电子	27	8,406	10,722	-21.6%	11.3%	11.3%	0.0
6969.HK	思摩尔国际	1,719	8,034	7,338	9.5%	10.7%	8.9%	1.7
1385.HK	上海复旦	645	4,337	1,290	236.3%	37.7%	30.0%	7.7
1415.HK	高伟电子	-66	3,936	4,044	-2.7%	16.2%	16.6%	-0.4
2018.HK	瑞声科技	927	3,757	2,822	33.2%	8.1%	6.3%	1.8
2577.HK	英诺赛科	2,907	2,930	-	-	7.8%	-	7.8
1888.HK	建滔积层板	1,632	2,150	301	615.2%	6.0%	1.3%	4.7
6613.HK	蓝思科技	1,439	1,449	-	-	19.3%	-	19.3
0522.HK	ASMPT	204	1,357	1,140	19.1%	4.3%	3.7%	0.7
0148.HK	建滔集团	62	1,244	841	48.0%	4.3%	4.1%	0.2
0303.HK	伟易达	657	1,005	258	289.2%	6.4%	1.9%	4.5
2038.HK	富智康集团	-1,420	959	282	239.5%	6.4%	3.9%	2.5
1478.HK	丘钛科技	-631	760	998	-23.9%	7.5%	13.0%	-5.5
2631.HK	天岳先进	692	750	-	-	23.5%	-	23.5
1304.HK	峰岹科技	558	425	-	-	16.1%	-	16.1

资料来源：Wind，国信证券经济研究所整理

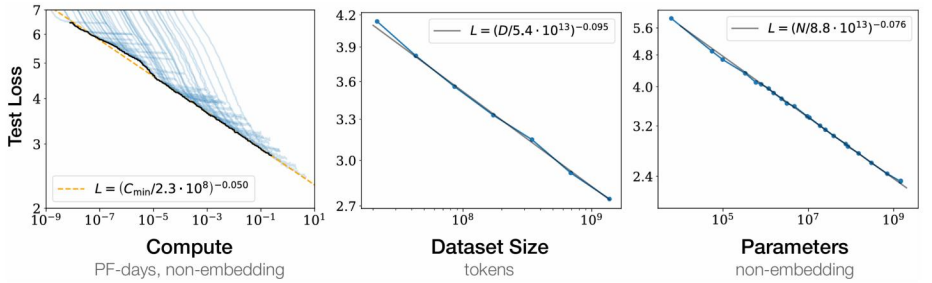
【2】AI大模型群雄逐鹿，英伟达引领算力迭代，PCB、服务器产业链延续高增长

2.1 大模型的演进：规模法则下的架构创新

国信证券 GUOSEN SECURITIES

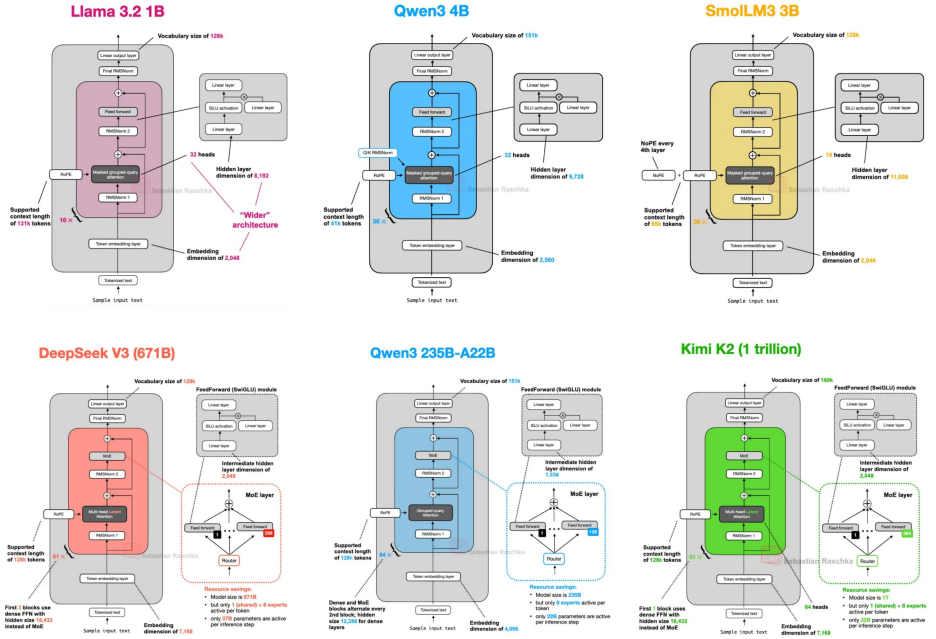
- OpenAI于2020年提出的**规模法则（Scaling Law）**指出，模型效能和模型参数规模、数据量、算力存在正相关的幂律关系，即模型的效果随模型规模指数增加而线性提高。随着模型规模的扩大，大模型将产生质变，从而获得以往小规模的模型不具备的深度思考、跨任务泛化的能力，此为**“智能涌现”效应**。
  - 近年来，伴随模型参数规模超越千亿级，人工智能技术得以“涌现”出更加强大的理解、推理、联想能力。典型大模型如2023年OpenAI发布的GPT-4，其基础架构仍为Transformer架构，通过增加参数量展现出卓越的性能。
- 提升模型的智能水平可以通过持续扩大模型规模和数据量来实现，而由于密集型Transformer模型的计算和内存开销是核心痛点，2024年至今对效率的迫切需求要求对模型架构进行改进升级。在此过程中，**混合专家（Mixture of Experts，MoE）架构**通过稀疏激活来降低计算量是用于提升效率的核心策略之一；**注意力机制（Attention）**的创新，降低了计算的复杂度。
  - 以DeepSeek-V3为例，其基本架构仍采用Transformer框架，但又创新性地融入多头潜在注意力（MLA）和DeepSeekMoE架构。这一设计在维持模型高性能的同时，极大地提升了训练与推理的效率。
- 当模型效率得到优化，模型便有条件实现计算成本更高的推理过程，此时模型有了进一步提升可解释性的需求，在进行推理任务场景时能够在给出最终答案前进行一系列复杂的思考步骤，便产生了**思维链（Chain of Thoughts，CoT）**。而对于模型AI Agent能力的开发，成为了应用推理能力的自然延伸。

图：模型规模的指数提升线性提高模型性能



资料来源：Jared等著-《Scaling Laws for Neural Language Models》-Arxiv（2020）-P3，国信证券经济研究所整理

图：部分重要模型架构示意图

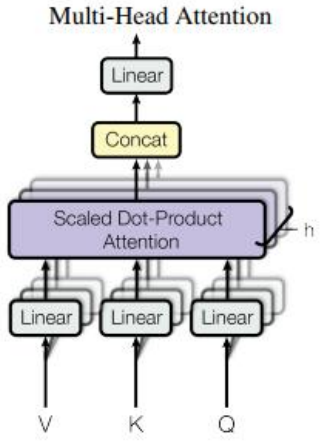


资料来源：Sebastian Raschka《The Big LLM Architecture Comparison》，国信证券经济研究所整理

2.1 注意力机制创新：降低计算复杂度与内存需求

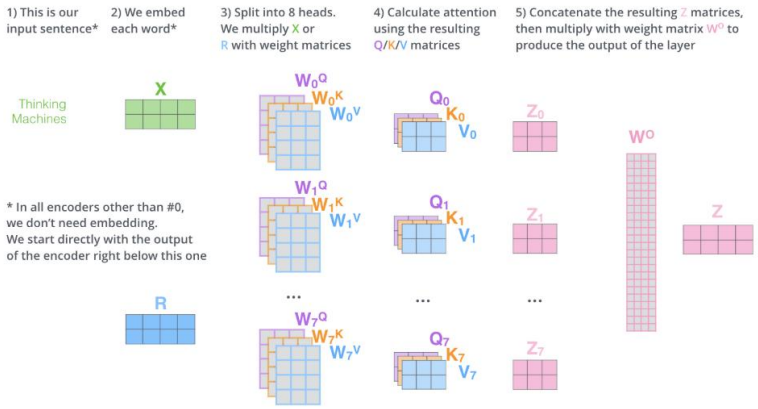
- **大模型的注意力机制经历了多头注意力机制（MHA）、分组查询注意力机制（GQA）、多头潜在注意力机制（MLA）的变化和创新，以期降低计算复杂度及内存需求。**
  - **多头注意力机制（Multi-head Attention，MHA）：**作为Transformer模型的基本结构，采用多组不同的线性变换对Q、K、V矩阵进行映射并分别计算Attention，再将不同的Attention结果拼接起来进行线性变换。MHA的本质是在参数总量保持不变的情况下，将Q、K、V映射到高维空间的不同子空间进行Attention计算，防止过拟合。
  - **分组查询注意力机制（Grouped-query Attention，GQA）：**将多个查询头进行分组，从而共享相同的键（Key）和值（Value）投影；这种计算方式减少了键和值计算的总数，从而降低了内存使用并提高了效率。近年来，GQA已成为MHA的新标准替代品，从而提供了更高的计算和参数效率。
  - **多头潜在注意力机制（Multi-head Latent Attention，MLA）：**在传统的注意力机制中，推理期间的键值缓存往往占用大量资源。而MLA通过低秩联合压缩技术，大幅削减了注意力键和值的存储空间。在生成过程中，仅需缓存压缩后的潜在向量，这一举措显著降低了内存需求。
- DeepSeek的MLA通过将长序列的Key和Value向量（即KV缓存）压缩成一个单一的、低秩的潜在向量（latent vector）来解决KV缓存瓶颈。这极大地减少了存储历史信息所需的内存，使它在支持128K上下文长度的同时，KV缓存相较于前代模型减少了93.3%。

图：MHA原理示意图



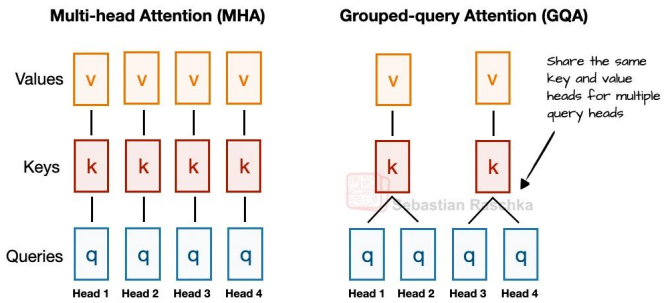
资料来源：《Attention Is All You Need》，国信证券经济研究所整理

图：MHA计算方法与流程



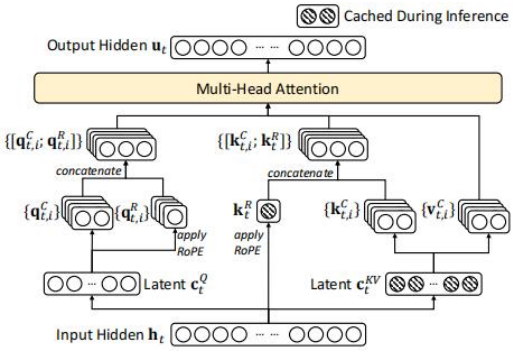
资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

图：MHA与GQA架构对比



资料来源：Sebastian Raschka《The Big LLM Architecture Comparison》，国信证券经济研究所整理

图：DeepSeek-V3的MLA架构



资料来源：DeepSeek，国信证券经济研究所整理

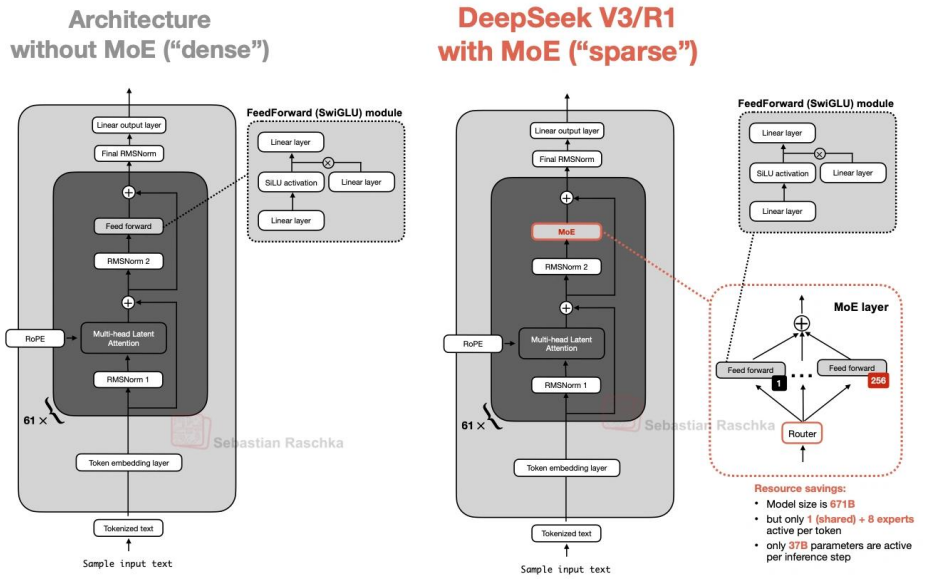
2.1 混合专家架构：稀疏化实现更高效的推理

国信证券 GUOSEN SECURITIES

- **混合专家（MoE）架构：**其基本思想是通过有选择性地仅调用最重要的专家模型，在不增加计算成本的前提下实现更高的智能水平和适应性。对于输入序列中的每个token，模型会动态地选择一小部分专家进行处理，即使模型的总参数量急剧增加，但每次前向传播（即推理）时实际激活的参数量和计算量仅占一小部分，从而实现经济的训练和高效的推理。

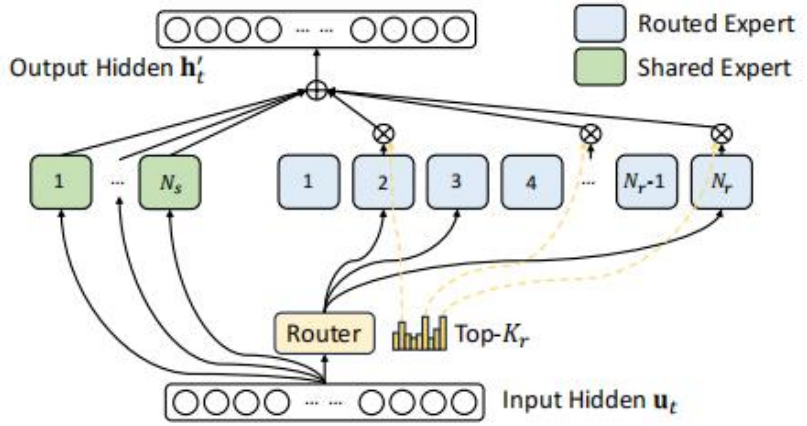
- DeepSeek的DeepSeekMoE架构采用了更为精细粒度的专家设置，还特别将部分专家设定为共享专家。在每一个MoE层中，都由共享专家和路由专家协同构成。其中，共享专家负责处理所有token的输入信息，为模型提供基础的处理支撑；而路由专家则依据每个token与专家之间的亲和度分数来决定是否被激活。这种独特的设计，使得模型在处理不同类型的输入时，能够更加灵活且高效地调配资源，进一步提升了整体的运行效率和表现。

图：MoE架构有（右）与无（左）的对比示意图



资料来源：Sebastian Raschka《The Big LLM Architecture Comparison》，国信证券经济研究所整理

图：DeepSeek-V3的DeepSeekMoE架构



资料来源：DeepSeek，国信证券经济研究所整理

2.1 深度思考：多轮推演减少幻觉，推理思考链走向前台

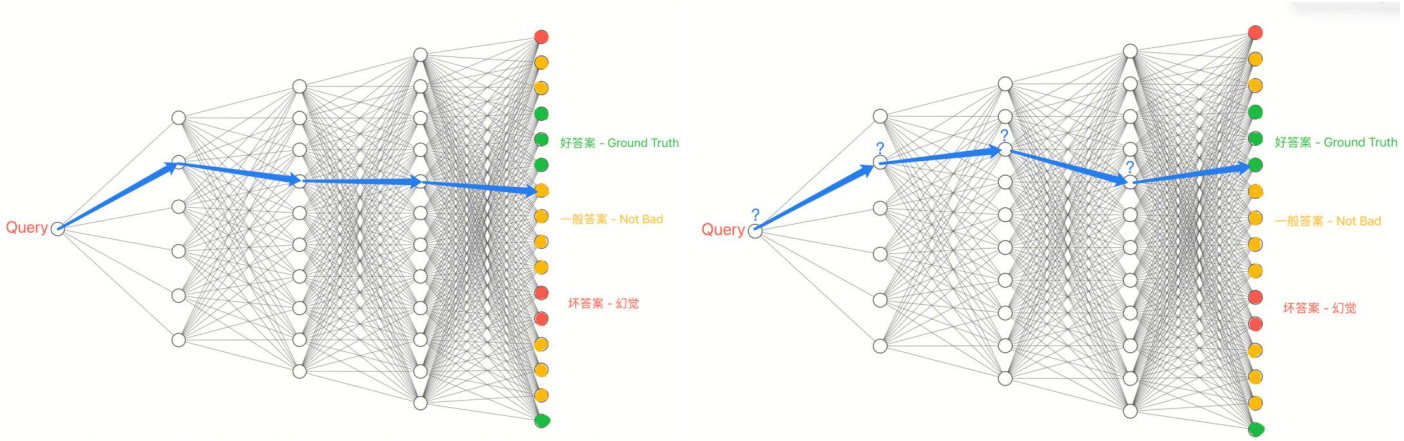
国信证券 GUOSEN SECURITIES

- 2024年9月，OpenAI推出专注于复杂推理任务的人工智能模型o1，该系列基于思维链强化学习技术开发，在数学、物理、生物和化学等学科领域展现博士级问题解决能力。该系列模型会花更多时间思考问题，然后再做出反应，就像人一样。同时通过训练，模型能够学会如何完善自己的思考过程、尝试不同的策略并认识到自己的错误。**OpenAI o1是“深度思考”模型成为发展主流的起点，同时将模型内蕴的推理能力“显化”。**

- **思考链（CoT）是模型在生成最终答案之前，投入额外的计算资源生成的一段内部思考过程，**能够使得模型在逻辑、数学和规划等复杂任务上实现性能的大幅提升。对于大模型而言，“深度思考”是“多轮推演”的过程，其最大的价值在于减少大模型的幻觉。相比于“直接输出”（即模型只经过一次思考，给出一条最佳路径），“深度思考”能够让模型在每一轮思考和决策过程中根据已知信息做更深层次的推演，更多地利用模型之前未激活的权重知识，从而得到准确率更高的答案。而思考链也标志着模型从静态的知识检索向动态的问题解决能力进行转变。

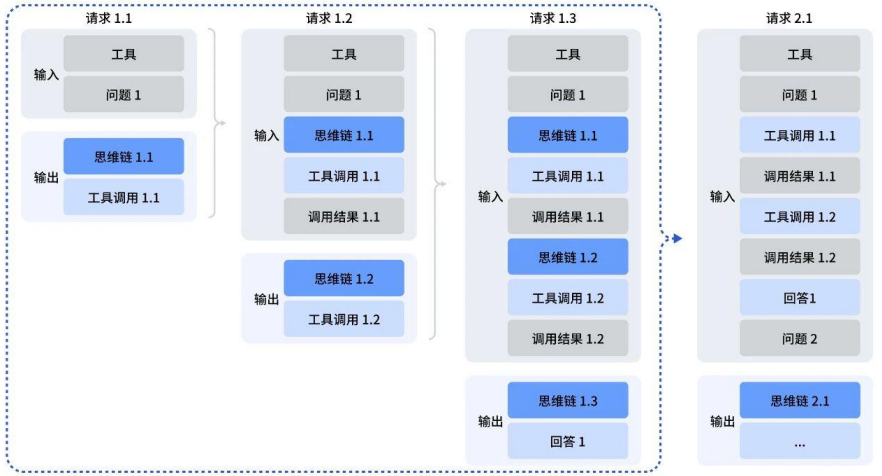
- 2025年12月，DeepSeek发布DeepSeek V3.2正式版，强化了Agent的能力，并融入思考推理。其中，DeepSeek V3.2支持思考模式下的工具调用能力，即在思考模式下，模型能够经过多轮的思考与工具调用，最终给出更详尽准确的回答。在这个过程中，用户需回传思维链内容给应用程序接口（API），以让模型继续思考。

图：模型“直接输出”（左）与“深度思考”（右）步骤示意图及对比



资料来源：腾讯技术工程，国信证券经济研究所整理

图：DeepSeek V3.2思考模式下的工具调用



资料来源：DeepSeek，国信证券经济研究所整理

2.1 最新大模型进展：群雄逐鹿，不断突破性能上限

国信证券 GUOSEN SECURITIES

- **Google Gemini 3系列：**2025年11月发布，整合进Gemini应用、谷歌的AI搜索产品AI Mode和AI Overviews，以及其企业级产品。Gemini 3 Pro在几乎所有主流AI基准测试中均显著超越了前代，展示了博士级的推理能力，重新定义了多模态推理的上限。Gemini 3 Pro高分登顶LMArena Leaderboard，在Humanity's Last Exam和GPQA Diamond上获得最高分。Gemini 3 Deep Think模式进一步拓展了智能的边界，在推理和多模态理解能力上带来重大进步，能够解决更复杂的问题，在前述测试中均优于Gemini 3 Pro。

- **Anthropic Claude Opus 4.5：**2025年11月发布，定位为旗舰级通用大模型，主打“更强推理、顶级编程、智能Agent/电脑操作”，能够自主处理模糊场景、权衡复杂决策，无需人工引导。Opus 4.5核心聚焦编程能力，SWE-bench Verified（业界公认的编程能力测试标准）达到80.9%，首次突破80%，超越Gemini 3 Pro的76.2%。

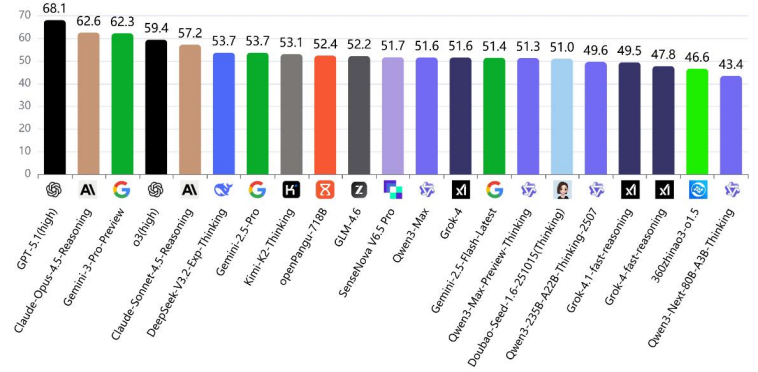
- **OpenAI GPT-5.2系列：**2025年12月发布，相较GPT-5.1在通用智能、长上下文理解、智能体工具调用及视觉方面均有显著提升。GPT-5.2在众多基准测试中超过了Gemini 3 Pro，重回SOTA（State of the Art，表示目前能够实现的最佳结果）。GPT-5.2分为三个版本，Instant用于日常工作和学习；Thinking用于更深入的工作，以更高的完成度处理复杂任务；Pro用于应对高难度问题，给出高质量答案。

- **DeepSeek V3.2系列：**2025年12月发布，分为V3.2和V3.2-Speciale版本，前者目标是平衡推理能力与输出长度，适合日常使用，例如问答场景和通用Agent任务场景；后者目标是将开源模型的推理能力推向极致，探索模型能力的边界，在主流推理基准测试上的性能表现优异。在公开的推理类Benchmark测试中，V3.2达到了GPT-5的水平，仅略低于Gemini-3.0-Pro。V3.2-Speciale模型成功斩获IMO 2025（国际数学奥林匹克）、CMO 2025（中国数学奥林匹克）、ICPC World Finals 2025（国际大学生程序设计竞赛全球总决赛）及IOI 2025（国际信息学奥林匹克）金牌。

- 据SuperCLUE数据，在2025年11月针对27个国内外大模型在中文大模型基准测评中，**GPT-5.1(high)综合得分最高**。在开源模型中，DeepSeek-V3.2-Exp-Thinking综合得分最高，同时也是表现最佳的国内模型。

  - SuperCLUE是独立领先的通用大模型的综合性测评基准，是CLUE（The Chinese Language Understanding Evaluation）基准的发展和延续。测评包括五大任务：数学推理、科学推理、代码生成（含web开发）、幻觉控制、精确指令遵循，SuperCLUE智能指数即SuperCLUE通用测评总分，直观展示各个模型的综合表现。

图：SuperCLUE智能指数（2025年11月）



资料来源：SuperCLUE，国信证券经济研究所整理

图：SuperCLUE测评基准2025年11月总体表现

排名	模型名称	机构	开/闭源	总分	数学推理	幻觉控制	科学推理	精确指令遵循	代码生成	属地	使用方式
-	GPT-5.1(high)	OpenAI	闭源	***68.11***	74.07	88.80	53.98	47.42	76.30	海外	API
-	Claude-Opus-4.5-Reasoning	Anthropic	闭源	***62.57***	58.33	90.33	44.25	50.32	69.64	海外	API
-	Gemini-3-Pro-Preview	Google	闭源	***62.26***	67.59	87.78	43.36	45.16	67.39	海外	API
-	o3(high)	OpenAI	闭源	***59.43***	66.67	83.81	34.51	48.39	63.76	海外	API
-	Claude-Sonnet-4.5-Reasoning	Anthropic	闭源	***57.23***	50.00	82.11	38.94	46.77	68.32	海外	API
🥇	DeepSeek-V3.2-Exp-Thinking	深度求索	开源	***53.69***	49.07	82.30	***36.28***	37.10	***63.70***	国内	API
-	Gemini-2.5-Pro	Google	闭源	***53.68***	60.19	90.38	32.74	33.87	51.22	海外	API
🥇	Kimi-K2-Thinking	月之暗面	开源	***53.07***	***63.89***	77.88	***36.28***	30.65	56.63	国内	API
🥉	openPangu-718B	华为	开源	***52.38***	46.30	***88.34***	27.43	***39.35***	***60.46***	国内	API
🥉	GLM-4.6	智谱AI	开源	***52.22***	***58.33***	84.73	***36.28***	23.87	57.89	国内	API
🥉	SenseNova V6.5 Pro	商汤	闭源	***51.67***	49.07	77.57	30.97	***42.90***	57.82	国内	API
🥉	Qwen3-Max	阿里巴巴	闭源	***51.56***	***55.56***	78.39	***46.02***	13.55	***64.29***	国内	API
-	Grok-4	X.AI	闭源	***51.56***	50.00	88.39	26.55	23.23	69.64	海外	API
-	Gemini-2.5-Flash-Latest	Google	闭源	***51.41***	52.78	88.18	27.43	27.10	61.58	海外	API
🥉	Qwen3-Max-Preview-Thinking	阿里巴巴	闭源	***51.31***	***55.56***	76.21	***38.94***	33.55	52.28	国内	API
🥉	Doubao-Seed-1.6-251015(Thinking)	字节跳动	闭源	***51.01***	42.06	***89.17***	31.86	36.77	55.18	国内	API
4	Qwen3-235B-A22B-Thinking-2507	阿里巴巴	开源	***49.60***	52.78	77.12	26.55	32.58	58.94	国内	API
-	Grok-4.1-fast-reasoning	X.AI	闭源	***49.48***	56.48	78.22	33.63	23.55	55.51	海外	API
-	Grok-4-fast-reasoning	X.AI	闭源	***47.78***	53.70	76.06	44.25	14.19	50.69	海外	API
5	360zhinao3-o1.5	360	闭源	***46.60***	32.41	***86.25***	24.78	***49.68***	39.87	国内	API
6	Qwen3-Next-80B-A3B-Thinking	阿里巴巴	开源	***43.41***	47.22	68.52	21.24	31.94	48.12	国内	API
-	gpt-oss-120b	OpenAI	开源	***40.88***	60.65	57.93	25.66	26.45	33.73	海外	API
7	Hunyuan-T1-20250822	腾讯	闭源	***40.74***	36.11	81.82	22.12	27.42	36.24	国内	API
7	MiniMax-M2	稀宇科技	开源	***40.02***	19.44	64.17	32.74	26.77	56.96	国内	API
8	LongCat-Flash-Thinking	美团	开源	***36.63***	23.15	77.79	17.70	26.45	38.09	国内	API
9	Spark-X1.5	科大讯飞	闭源	***31.56***	26.85	69.66	5.31	8.71	47.26	国内	API
-	Llama-4-Maverick-17B-128E-Instru	Meta	开源	***22.33***	8.33	64.60	6.19	8.71	23.83	海外	API

注：数据来源SuperCLUE，2025年11月28日；为减少波动影响，本次测评将相差1分内的模型视为并列。海外模型仅作对比参考，不参与排名。标红为国内前三。

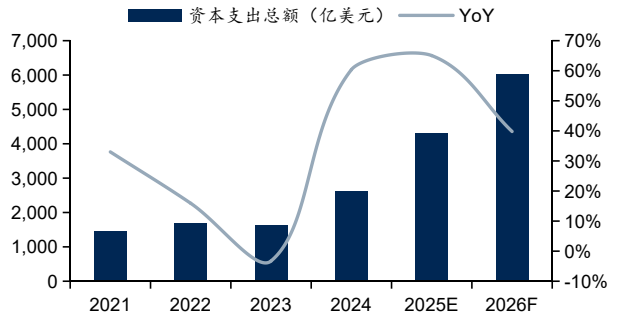
资料来源：SuperCLUE，国信证券经济研究所整理

2.1 云服务商资本开支快速增长，大力推进AI基础设施建设

- **大型云服务厂商近年资本开支快速增长，算力“军备竞赛”愈演愈烈。**国外四大CSP厂商今年前三季度资本开支均已接近或超过500亿美元，亚马逊更是超过900亿美元。国外四大CSP厂商亚马逊、微软、谷歌、Meta在2025年第三季度资本开支分别达到351亿、194亿、240亿、188亿美元，同比分别增长55.2%、30.0%、83.4%、128.0%；2025年前三季度累计资本开支分别达923亿、532亿、636亿、483亿美元，同比分别增长67.3%、33.9%、66.2%、111.6%。

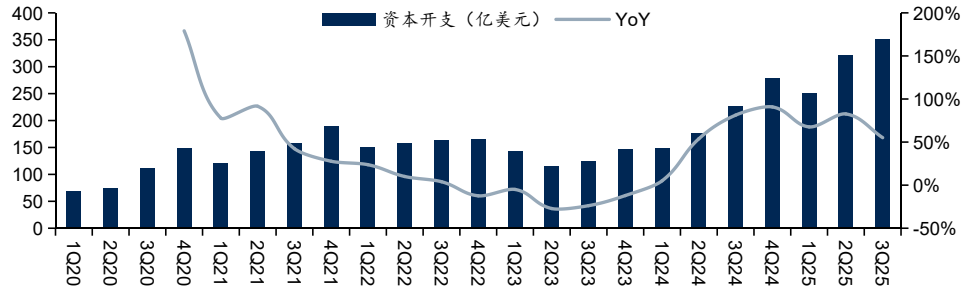
- 据TrendForce数据，基于北美CSP最新财报指引，其已将全球八大主要CSP（Google、AWS、Meta、Microsoft、Oracle、腾讯、阿里巴巴、百度）2025年资本支出总额年增长率由此前的61%上修至65%。**预计2026年八大CSP将维持积极的投资节奏，合计资本支出将进一步增长40%达到6000亿美元以上，**展现出AI基础建设的长期成长潜能。

图：全球八大CSP资本支出总额及预测

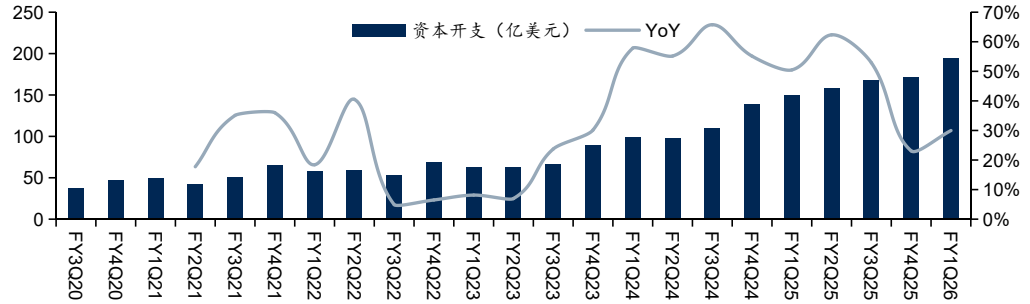


资料来源：TrendForce，国信证券经济研究所整理

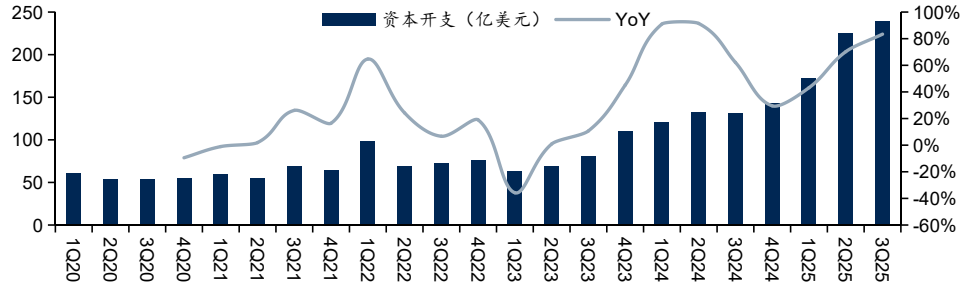
图：亚马逊季度资本开支



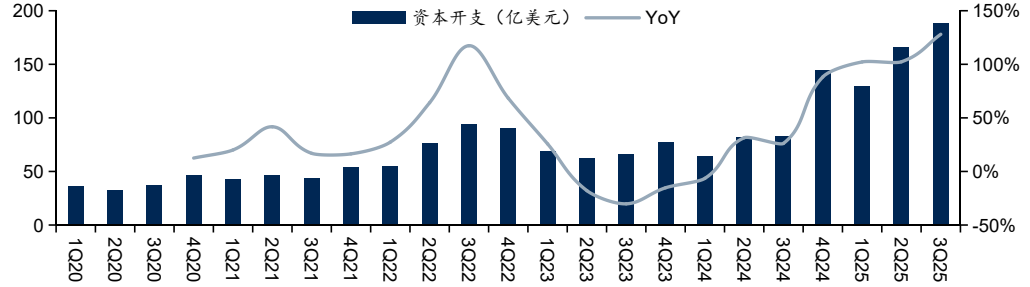
图：微软季度资本开支



图：谷歌季度资本开支



图：Meta季度资本开支



资料来源：Wind，国信证券经济研究所整理

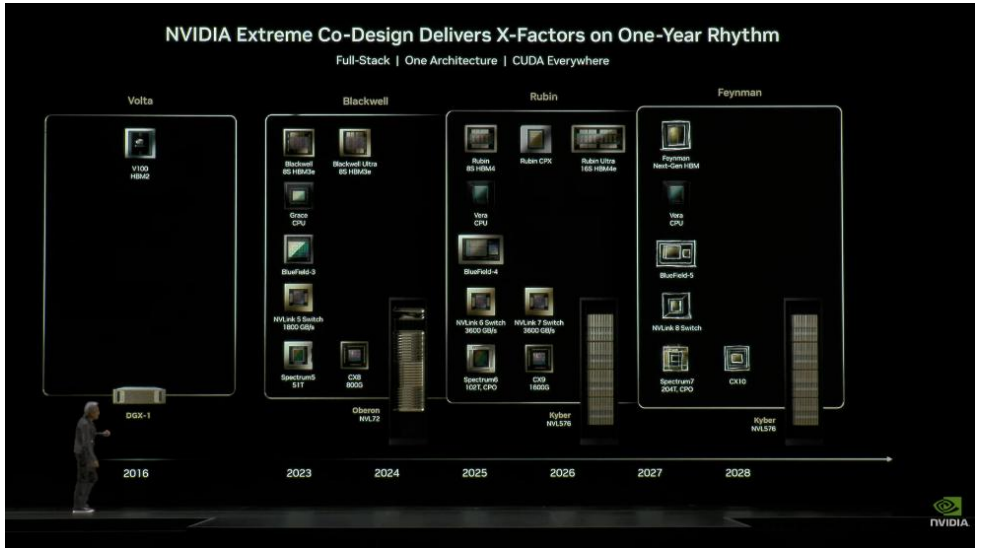
2.1 英伟达技术路线：架构持续快速更新迭代，算力成倍增长

国信证券 GUOSEN SECURITIES

- **英伟达GPU架构快速更新迭代奠定其领先地位。**英伟达GPU加速计算发展始于2008年推出的Tesla架构，其成为第一代真正开始用于并行运算的GPU架构。2010年Fermi架构推出，是第一个支持DirectX 11的GPU计算架构。2012年Kepler架构推出，作为Fermi的升级版整体架构保持一致。2014年Maxwell架构推出，通过优化架构提供了可观的能耗比提升。2016年Pascal架构推出，是首个为深度学习而设计的GPU架构。2017年Volta架构推出，专注于提高深度学习的性能。2018年Turing架构推出，是全球首款支持实时光线追踪的GPU架构。2020年Ampere架构推出，统一了AI训练和推理，并在光线追踪和DLSS（深度学习超级采样）方面有显著的改进。2022年Hopper架构推出，主要面向AI及数据中心等构建。2024年Blackwell架构推出，使用了二代Transformer、Secure AI、5代NVLink等最新技术。

- **2025年英伟达GB200系列AI服务器以整机柜形式批量出货，性能与算力成倍增长，不断满足当前井喷的算力需求。**GB200是由两个Blackwell B200 GPU和一个Grace CPU组成的AI加速平台，每个B200 GPU含有2080亿个晶体管。相较于H100，GB200的算力提升了6倍；在处理多模态特定领域任务时，算力达到H100的30倍。GB200 NVL72是一套多节点机架级扩展系统，适用于高度计算密集型的工作负载，它将36个Grace Blackwell超级芯片组合在一起，其中包含通过第五代NVLink相互连接的72个Blackwell GPU和36个Grace CPU。此外，升级版B300芯片及GB300系列服务器亦陆续出货。

图：英伟达技术路线图



资料来源：英伟达，科创日报，国信证券经济研究所整理

图：英伟达不同技术路线架构参数对比

Nvidia Roadmap							
	2022	2023	2024	2025	2026	2026	2027
Chip and Package Level							
	Hopper		Blackwell		Rubin		
Accelerator	H100 (SXM)	H200	B200/ GB200	GB300 (Ultra)	VR200	CPX	VR300 (Ultra)
GPU TDP (W)	700	700	700/1200	1,400	2,300	800	4,000+
Foundry Node	4N		4NP		N3P (3NP)	N3P (3NP)	N3P (3NP)
Logic Die Configuration	1 x Reticle Sized GPU		2 x Reticle Sized GPU		2 x Reticle Sized GPU, 2x I/O chiplet	1x Reticle Sized GPU	4 x Reticle Sized GPU, 4x I/O chiplet
FP4 PFLOPs - Dense (per Package)	4*		10	15	33.3	20.0	66.7
Memory	80GB HBM3	141GB HBM3E		288GB HBM3E	288GB HBM4	128GB GDDR7	1024GB HBM4E
HBM Stacks	5	6	8		8	N/A	16
Memory Bandwidth	3.35TB/s	4.8TB/s	8TB/s		20.5TB/s	2TB/s	53TB/s
Packaging	CoWoS-S		CoWoS-L		CoWoS-L	FC-BGA	CoWoS-L
SerDes speed (Gb/s uni-di)	112G		224G		224G	64G (PCIe Gen6)	224G
Nvidia CPU	Grace				Vera	Vera	Vera
System Form Factor							
Maximum system density	NVL8		NVL72 144 compute chiplets 72 Logical GPUs		NVL144 144 compute chiplets 72 Logical GPUs	144 CPX Chips	NVL576 576 compute chiplets 144 Logical GPUs
Form Factor Supported	HGX		HGX,Oberon		HGX,Oberon	VR CPX	Kyber
# of GPU Packages	8		72	72	72	144	144
# of GPU dies	8		144	144	144	144	576
Scale up links	UBB (PCB)		Copper Backplane		Copper Backplane	None	PCB Backplane
Aggregate FP4 PFLOPs (Dense)	32*		720	1,080	2,398	2,877	9,605
Aggregate Memory capacity	14TB	14TB	14TB	21TB	21TB	18TB	147TB
Aggregate Memory bandwidth	27TB/s	38TB/s	576TB/s	576TB/s	1,476TB/s	288TB/s	7,668TB/s

资料来源：SemiAnalysis，英伟达，国信证券经济研究所整理

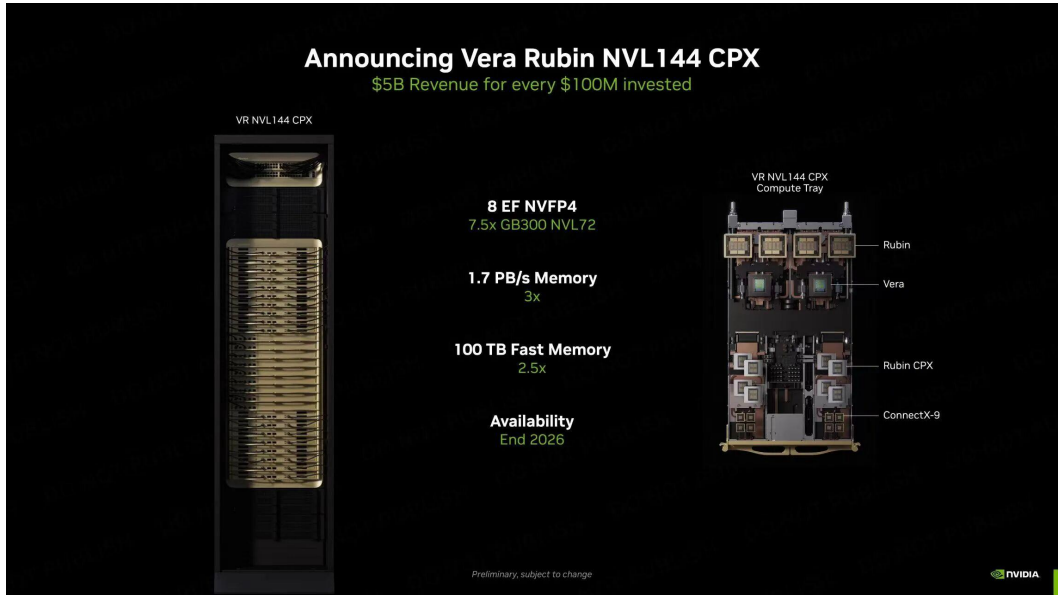
2.1 Rubin服务器：英伟达下一代芯片架构，为“推理时代”而来

国信证券 GUOSEN SECURITIES

- **随着AI大模型“推理时代”的全面到来，英伟达最新推出Rubin CPX GPU，专为长语境推理设计，或将彻底改变推理领域。**该芯片通过专门优化预填充阶段，强调计算FLOPS而非内存带宽，为分离式推理服务带来革命性变化。与此同时，英伟达推出三种Vera Rubin机架配置：VR200 NVL144（仅Rubin）、VR200 NVL144 CPX（Rubin+Rubin CPX混合）、以及Vera Rubin CPX双机架方案，其无电缆设计、全液冷方案、灵活扩展等特点解决了前两代机柜的诸多痛点。其中，英伟达Vera Rubin NVL144 CPX机架将Rubin GPU与Rubin CPX GPU整合，集成了36个 Vera CPU、144块Rubin GPU和144块Rubin CPX GPU，这种异构配置使得系统能同时高效处理推理的两个阶段。此外，双机架解决方案提供了更大的灵活性，允许客户根据自身工作负载需求，单独部署VR NVL144（纯Rubin GPU）机架和VR CPX（纯Rubin CPX GPU）机架，以精确调整预填充与解码的比例。预计2028年英伟达将推出更新一代的Feynman架构产品。

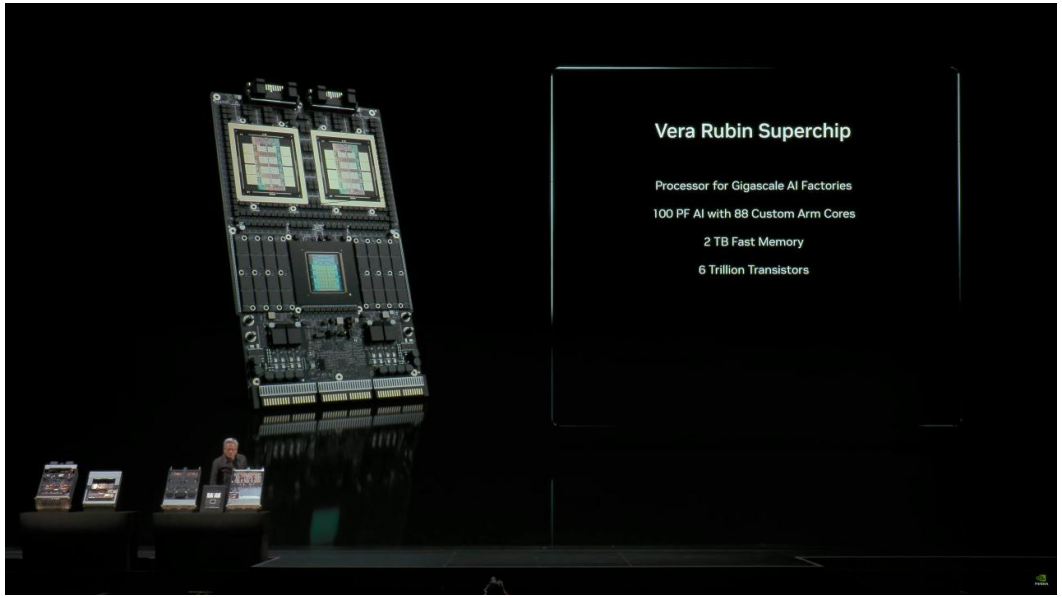
- **10月28日，下一代Vera Rubin超级芯片在英伟达华盛顿GTC大会上首亮相。**它搭载了Vera CPU和两颗强大的Rubin GPU。该主板还搭载了大量LPDDR系统内存（共32个），将与Rubin GPU上的HBM4显存配合使用。Vera CPU将配备88个定制ARM核心，共计176个线程。预计Rubin GPU将在明年10月或更早进入量产阶段。

图：VR NVL144 CPX重要组件拆解示意图



资料来源：英伟达，国信证券经济研究所整理

图：Vera Rubin超级芯片



资料来源：英伟达，国信证券经济研究所整理

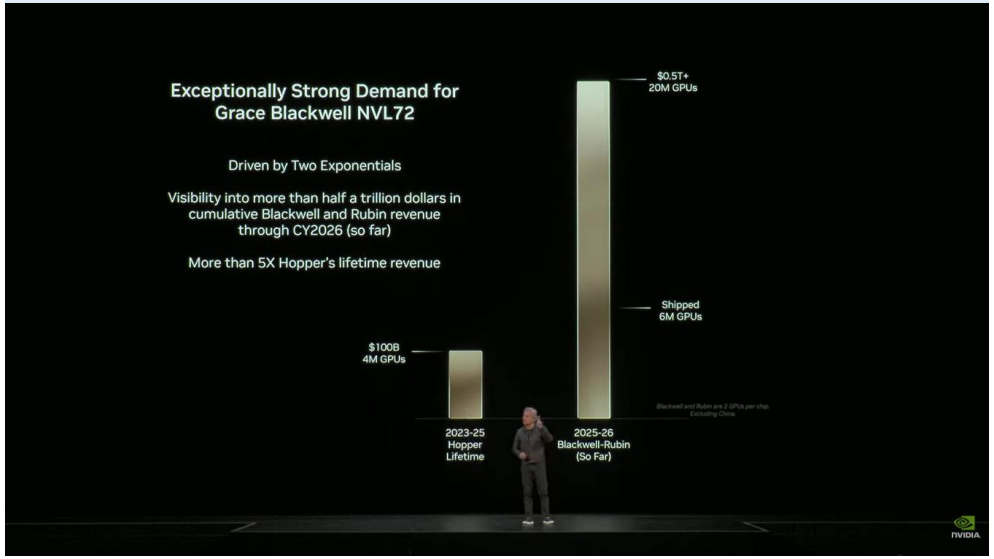
2.1 英伟达AI芯片出货量屡创新高，AI服务器出货量及占比持续增长

国信证券 GUOSEN SECURITIES

- **英伟达Blackwell和Rubin系列GPU出货有望创新高，英伟达仍占据AI芯片供应主导地位。**根据2025年10月英伟达GTC大会上黄仁勋最新披露，Blackwell GPU已经量产，过去四个季度出货600万颗；展望未来5个季度，Blackwell和Rubin系列GPU预计总出货量将达2000万颗，相较于Hopper增长5倍；2026年前Blackwell与Rubin芯片合计订单有望达5000亿美元收入。此外，据TrendForce预测，2025年英伟达仍占据70%左右AI芯片市场，仍处于绝对的主导地位。

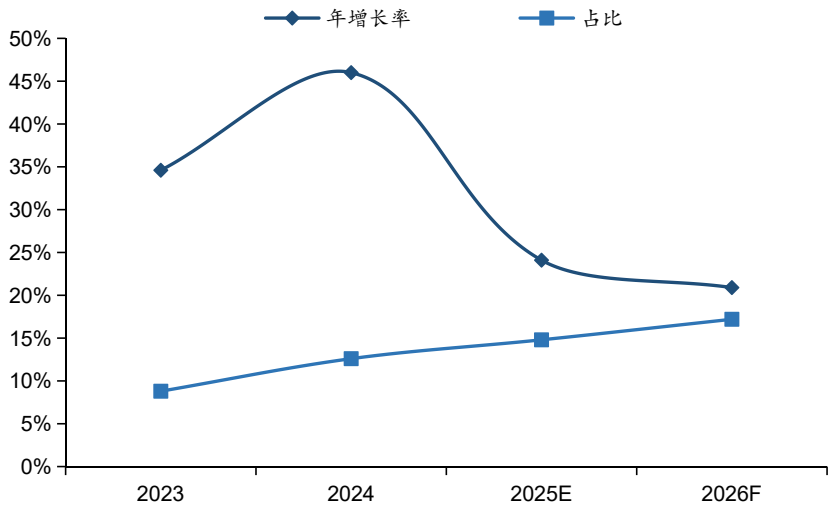
- **受益于智能算力市场的推动，全球AI服务器出货量及占比持续增长。**据TrendForce预测，预计2025年AI服务器出货量同比增长24.1%，整体服务器占比提升至14.8%，产值方面受惠于Blackwell新方案、GB200/GB300机柜等较高价值的整合型AI方案，预计将有近48%的增长。2026年受益于CSP、主权云等算力需求持续稳健，叠加AI推理应用蓬勃发展，TrendForce预计全球AI服务器出货量将进一步增长20.9%，占整体服务器比例上升至17.2%，产值方面受惠于整机型AI服务器方案的进一步渗透，有望同比增长30%以上，营收占整体服务器比重将达70%。

图：英伟达最新GPU出货量及预测



资料来源：英伟达，国信证券经济研究所整理

图：AI服务器出货年增长率及占整体服务器比例



资料来源：Trendforce，国信证券经济研究所整理

2.1 服务器架构趋势一：超节点成为新型算力基础设施架构

国信证券 GUOSEN SECURITIES

- **超节点（Hyper Node）是自2024年起逐步兴起的一种新型AI算力基础设施架构，**其核心目标是应对大模型训练与推理对极致通信效率和高密度算力协同的迫切需求。例如，华为CloudMatrix384首创将384颗昇腾芯和192颗鲲鹏芯，通过超节点高速网络MatrixLink全对等互联，形成一台超级“AI服务器”，算力从单台服务器的6.4 Pflops提升到超节点服务器的300 Pfops，算力提升了50倍。
- 超节点通过高速互连技术将数十至数百颗AI加速芯片（如GPU、NPU或TPU）在一个物理机柜或一组紧密耦合的计算单元内进行高密度集成，构建出一个高带宽域（High-Bandwidth Domain, HBD）。在此域内，所有加速器之间的通信延迟极低、带宽极高，近似于单机内部互联的性能水平。
- 该架构旨在突破传统服务器受限于PCIe总线带宽，以及跨节点依赖标准以太网或InfiniBand所带来的通信瓶颈。通过将大量计算单元整合为一个逻辑上的“超级服务器”，超节点能够高效支撑张量并行（Tensor Parallelism）、专家并行（Expert Parallelism）等对内部通信带宽和延迟极为敏感的大模型并行训练策略。

图：华为CloudMatrix384超节点

资料来源：华为开发者大会，国信证券经济研究所整理

图：英伟达DGX SuperPOD架构示意图

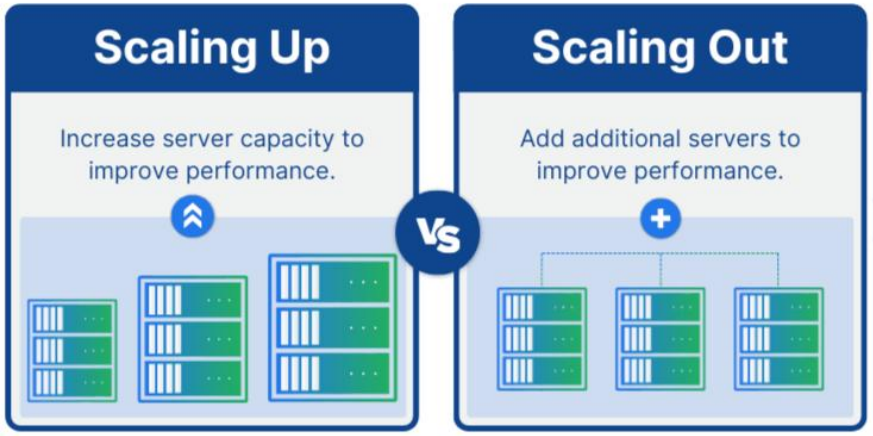
资料来源：英伟达，国信证券经济研究所整理

2.1 服务器架构趋势二：Scale Up与Scale Out带来智算集群的扩展

国信证券 GUOSEN SECURITIES

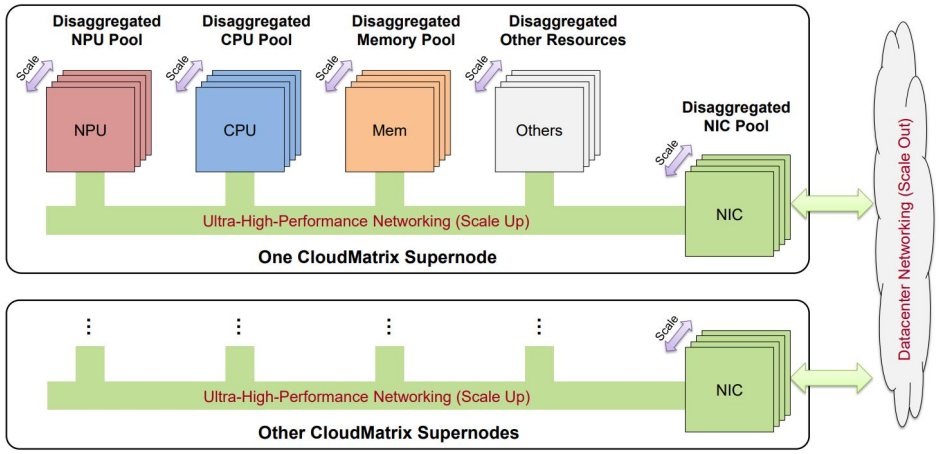
- 大模型训练与智算中心的整体网络架构通常划分为多个逻辑平面，包括参数面、样本面、业务面、存储面和管理面。这些网络平面可根据实际业务需求灵活共享资源，以高效、弹性地支撑大模型训练及各类智能计算任务。**AI智算集群的扩展主要通过两种模式实现：Scale Up（纵向扩展）和Scale Out（横向扩展），**二者共同推动智算集群在算力规模与运行效率上的持续突破。
  - Scale Up（纵向扩展）指在单个节点内部增加GPU/NPU等加速卡的数量，构建高性能“超节点”。其核心目标是优化节点内各组件（如GPU-GPU、GPU-CPU）之间的通信，通过极致的低延迟与高带宽互联，使多个加速器协同工作，如同一个统一而强大的计算单元。这种模式适用于对单机性能要求极高的场景，系统能力的提升依赖于更强的硬件配置，属于典型的垂直扩展路径。
  - Scale Out（横向扩展）则通过增加节点数量来扩大集群整体规模，将成百上千个超节点互联，形成大规模分布式训练系统。它主要解决多节点之间的通信效率问题，其网络性能直接决定了集群的可扩展性和整体训练效率。该模式适用于单一节点无法承载的超大规模模型训练任务，属于水平扩展架构，具备良好的弹性和可伸缩性。
- **在实际部署中，Scale Up与 Scale Out往往协同使用：**前者提升单点算力密度，后者拓展系统整体容量，二者相辅相成，共同构建高性能、高效率、高灵活性的AI智算基础设施。

图：Scale Up和Scale Out示意图



资料来源：IT之家，国信证券经济研究所整理

图：华为CloudMatrix超节点的Scale Up和Scale Out示意图

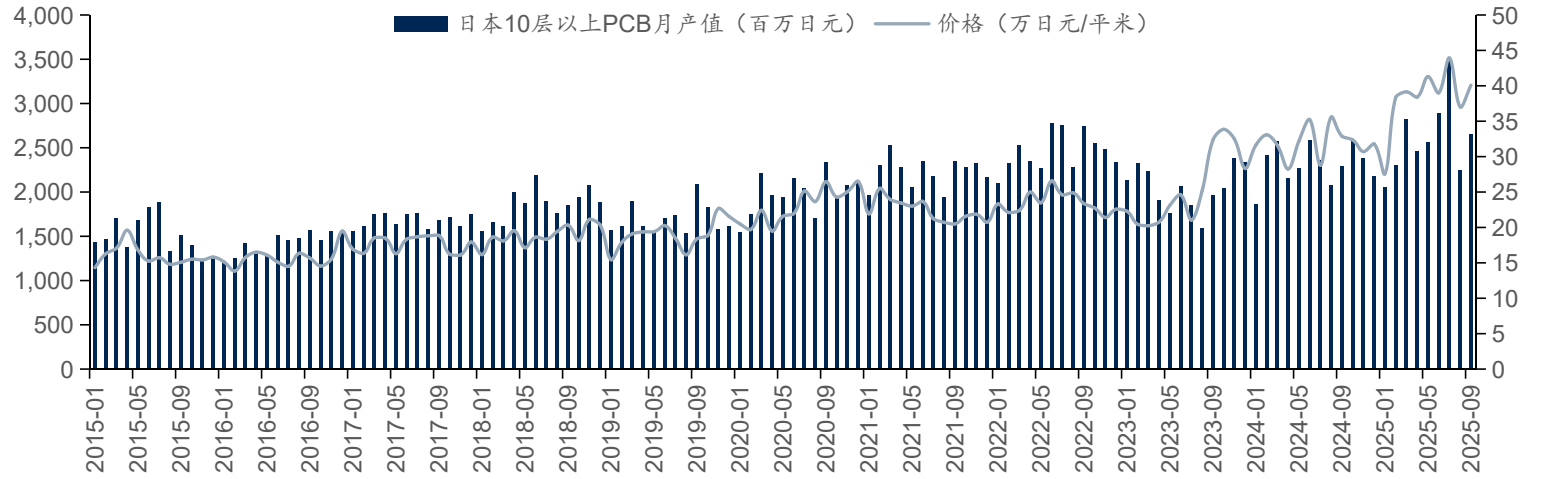


资料来源：华为，国信证券经济研究所整理

2.2 PCB自3Q23进入周期上行

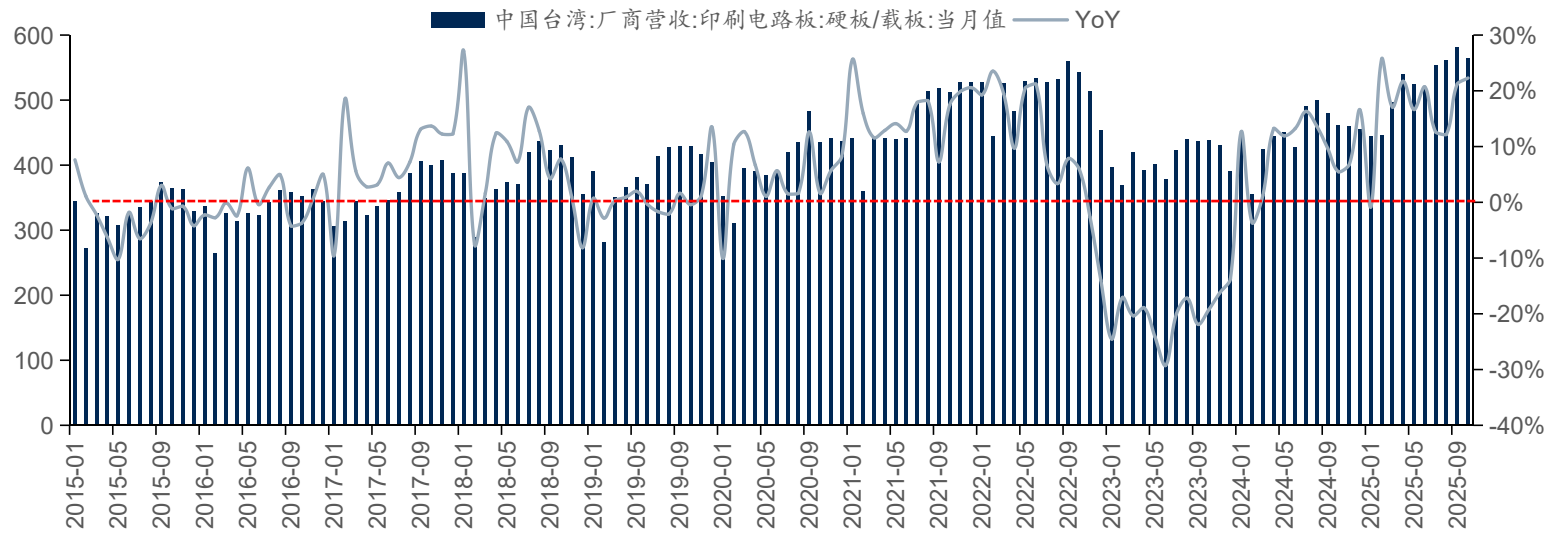
- **PCB行业在经历了2H22及1H23周期下行，2H23进入行业修复阶段，并自2024年年初进入景气上行阶段，且随着2025年AI基建加速，行业出现供不应求的状况，并预计将会持续到2027年。**
- 根据Prismark统计的日本10层以上PCB产值和价格月度情况，2015年至2025年十年间，23年9月相关产值达到15.87亿日元的十年产值低位，此后伴随单位面积价格快速增长，产值回升，25年7月达到34.90亿日元的十年高位，同比23年9月翻了一倍。
- 价格方面，2015至1H23，日本10层以上PCB均价最低价14.33万日元/平米，最高价26.57万日元/平米，8年涨幅85%。但进入2H23，价格快速提升，23年7月价格为21.04万日元/平米，25年9月价格达到40.08万日元/平米，涨幅90%。
- PCB周期上行主要由3个原因驱动，按重要性排序分别为：①AI推动的交换机、服务器等算力基建爆发式增长；②汽车智能化落地带来的量价齐升；③智能手机、PC的新一轮AI创新周期。

图：日本10层以上PCB月产值和价格情况



资料来源: Prismark，国信证券经济研究所整理

图：中国台湾省PCB厂商月度营收

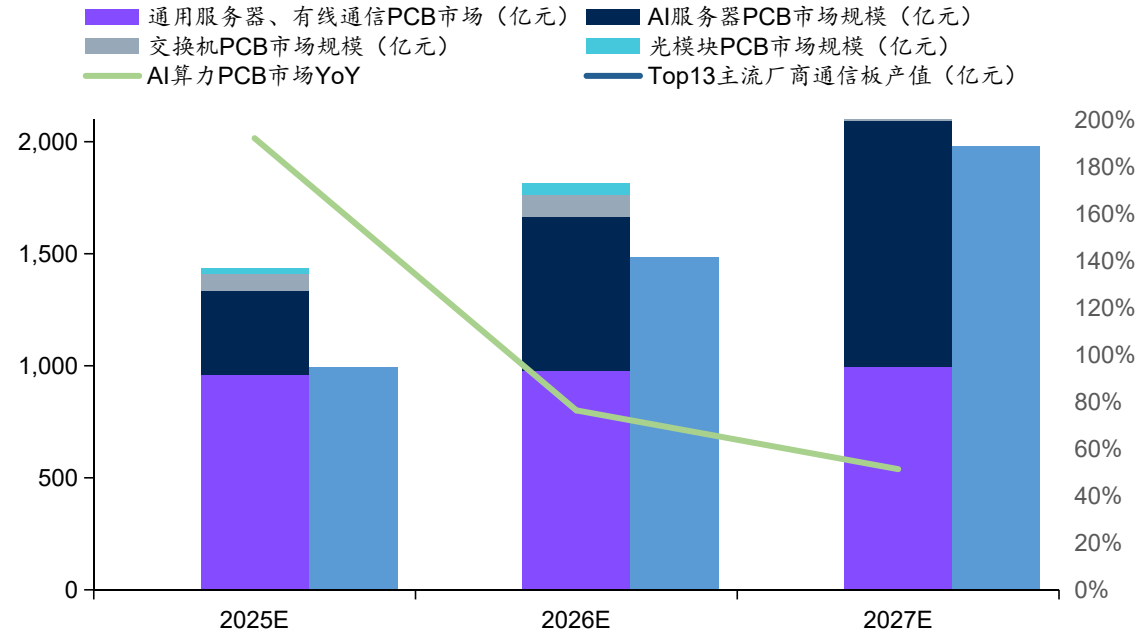


资料来源：Wind，国信证券经济研究所整理

2.2 2026年PCB供需缺口持续存在，有望在27年收窄

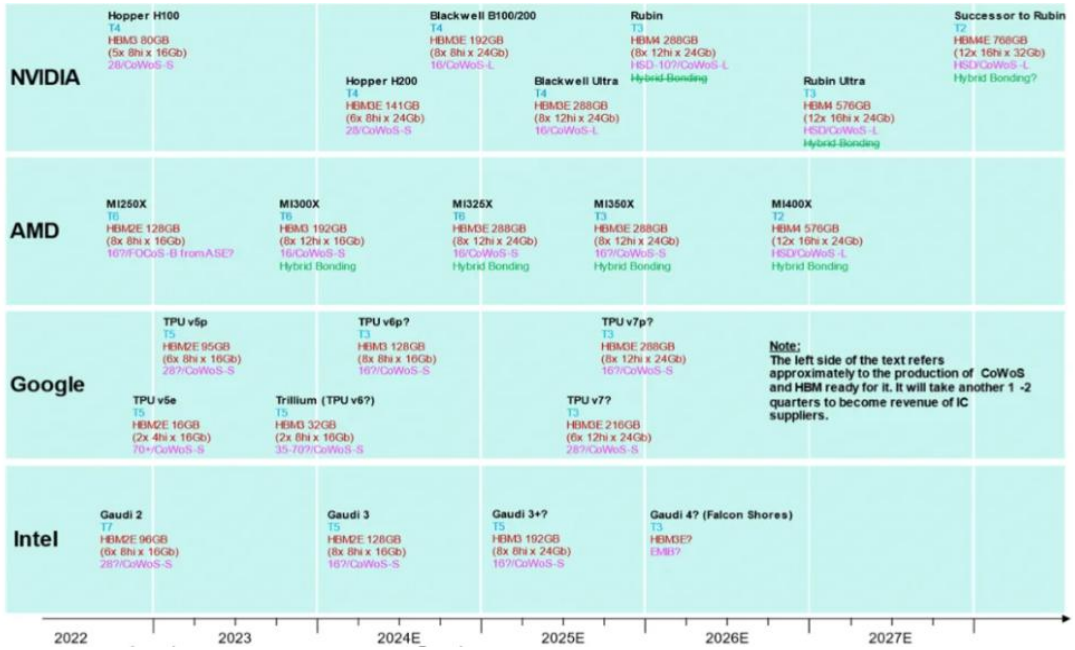
- **需求方面：**AI算力基础建设将会带动AI服务器、高速交换机、光模块等类别PCB需求大增，并间接带动通用服务器领域温和增长。根据产业链跟踪，我们预计英伟达GB300机柜中，单GPU对应PCB价值量为420美金，同代ASIC芯片对应GPU价值量为700美金，2026年，随着Rubin系列出货占比提升/ASIC芯片配套升级，对应单芯片PCB价值量分别提升，对应AI服务器PCB市场2025-2026年达到379/689亿元，考虑交换机PCB市场73/95亿元，光模块PCB市场23/53亿元，通用服务器市场958/977亿元，合计有线通信类PCB市场达到1433/1815亿元。
- **供给方面：**基于IDC、Prismark等第三方机构预测，及各公司已公开披露的产能规划，以投入产出比1:1.2-1.5范围计算，我们预计2025-2026年，Top13全球算力类（服务器+有线通信）PCB产值将达到780/1320亿元。我们的测算基于数个重要假设：我们统计了各公司2026年年底的投产预期，并将其假设为2027年的平均产能，2025至2027年线性爬产，期间稼动率始终保持满产。但目前由于海外工厂缺少成熟工人和工程师，多个公司海外工厂投产速度不及预期，国内工厂也可能面临相关设备缺货的情况。
- **综上，根据我们的测算，2026年全球算力类PCB市场需求将达到1815亿元，而全球Top13的PCB厂商相关产值约为1320亿元，考虑其他厂商，预计将有近200亿的供需缺口。**展望2027年，就目前的CSP厂资本开支预期和PCB厂商的扩产预期假设，供需缺口将大幅收窄。

图：算力PCB供需测算



资料来源: IDC，Prismark，各公司公告，国信证券经济研究所整理

图：AI Accelerator roadmap



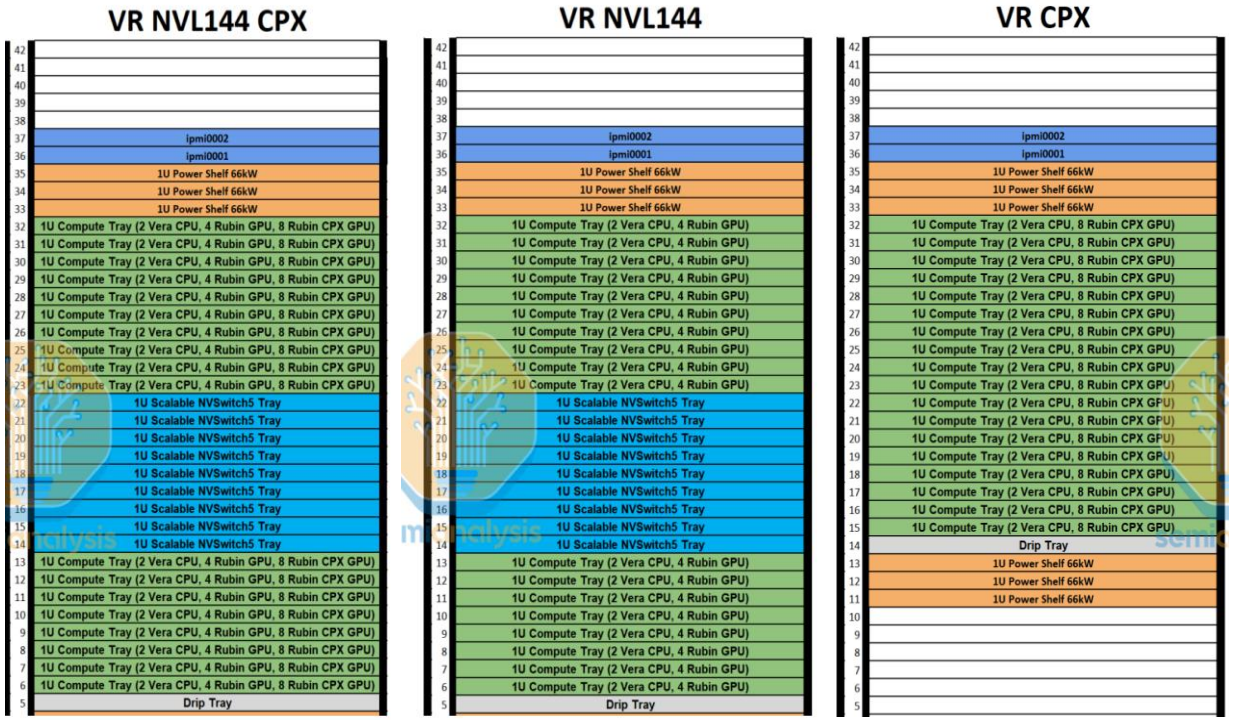
资料来源: semianalysis，国信证券经济研究所整理

注：产值参考公司各自口径，公司间可能存在口径误差。部分海外PCB厂商产能数据缺失，产值预测可能略低于实际产值

2.2 CPX机架芯片密度再提升，PCB价值量大幅增长

- **在2025年9月的AI Infra Summit上，英伟达发布了Rubin CPX解决方案，是一款专为处理超长上下文、AI智能体和大规模推理场景设计的GPU新品。**采用了创新的解耦式推理架构，单芯片Rubin CPX更侧重计算性能而非内存带宽，预计将于2026年年底上市。由于推理过程中的预填充阶段往往大量消耗计算能力而仅轻度占用内存带宽，若采用内存带宽精简但计算能力充沛的芯片可实现更高效的资源配置，Rubin CPX GPU通过针对推理过程中截然不同的预填充和解码阶段进行硬件专项优化，使分布式服务更充分发挥其潜力。
- **VR200将包含三种配置的机架级服务器：**
  1）VR200 NVL144：包含18个compute tray，每个compute tray搭载4颗Rubin GPU和2颗CPU，不含CPX；
  2）VR200 NVL144 CPX：在VR200 NVL144基础上，每个compute tray额外搭载8颗Rubin CPX GPU；
  3）Vera Rubin CPX only：包含18个compute tray，每个tray搭载8个Rubin CPX GPU+2个Vera CPU，可与VR NVL144配合使用。

图：Rubin的三种架构



Nvidia Rack Scale Servers						
	Units	GB200 NVL72	GB300 NVL 72	VR200 NVL144	VR200 NVL144 CPX	Vera Rubin CPX Only
Compute and Memory						
Compute Trays	#	18× GB200 NVL72	18× GB300 NVL72	18x VR NVL144	18x VR CPX 18x VR NVL144	18x VR CPX
GPU	Type	B200	B300	R200	-	-
CPU	Type	Grace	Grace	Vera	Vera	Vera
CPX GPU	Type	-	-	-	Rubin CPX	Rubin CPX
FP4 Dense FLOPS	PFLOPs	720.0	1,080.0	2,397.6	5,277.6	2,880.0
HBM Memory Capacity	TB	13.8	20.7	20.7	20.7	-
GDDR7 Memory Capacity	TB	-	-	-	18.4	18.4
HBM Memory Bandwidth	TB/s	576	576	1,476	1,476	-
GDDR7 Memory Bandwidth	TB/s	-	-	-	288	288
Rack-Level Content						
CPUs	#	36	36	36	36	36
GPU Packages	#	72	72	72	72	-
Rubin CPX GPUs	#	-	-	-	144	144
Total NICs	#	72	72	144	144	144
Total Compute and Networking Chi	#	**180**	**180**	**252**	**396**	**324**
Networking						
Scale-Up World Size	#	72	72	72	72	-
Number of NVSwitches	#	18	18	36?	36?	-
NVLink Scale-Up Bandwidth (uni-di)	Tbit/s	518	518	1,037	1,037	-
Scale-out NIC	Type	CX-7	CX-8	CX-9 800G	CX-9 800G	CX-9 800G
Scale-out NIC per Compute Tray	#	4	4	8	8	8
Scale-out Bandwidth (uni-di)	Tbit/s	28.8	57.6	115.2	115.2	115.2
Front-end NIC	Type	Bluefield-3	Bluefield-3	Bluefield-4	Bluefield-4	Bluefield-4
System Design						
Compute Tray Connectivity	Type	Cable + PCB	Cable + PCB	PCB	PCB	PCB
Cooling	Type	Liquid(85%) + Air(15%)	Liquid(85%) + Air(15%)	Liquid (100%)	Liquid (100%)	Liquid (100%)
Power Budget	kW	~140	~180	~225	~370	~190

资料来源: SemiAnalysis，Nvidia。国信证券经济研究所整理

2.2 CPX机架芯片密度再提升，PCB价值量大幅增长

- **由于新增CPX GPU，computer tray布局发生较大变化：**

  1）新增了一块链接bianca板以及下方CX-9网卡的midplane；

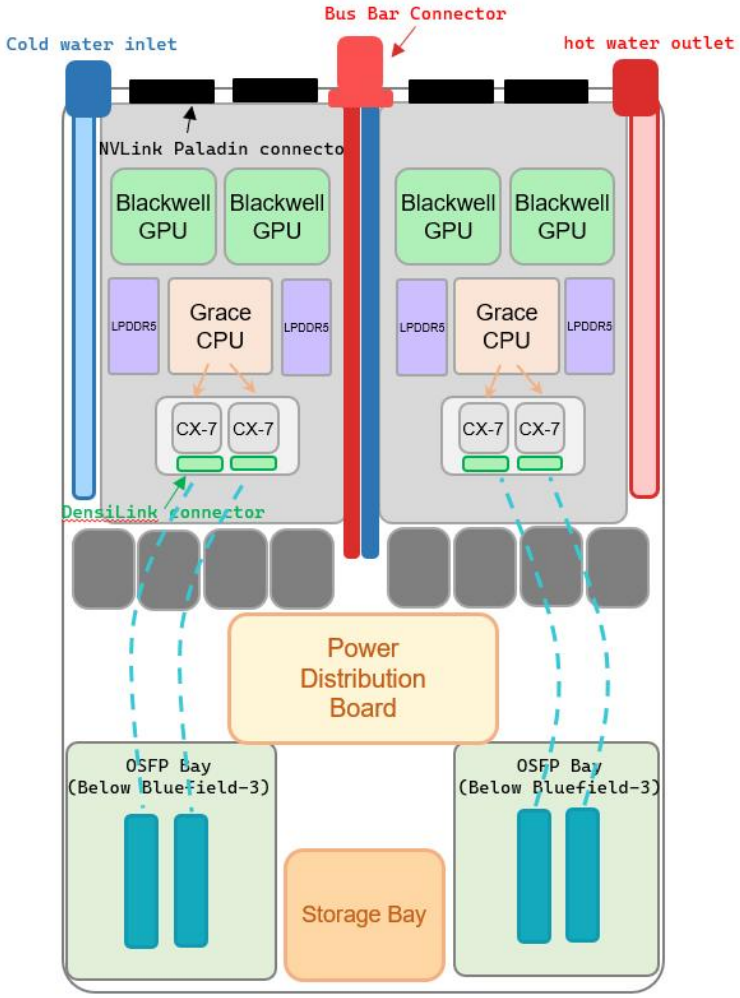
  2）Rubin CPX GPU主板，在compute tray内新增8颗Rubin CPX GPU，位于CX-9网卡上方，放置在四张子卡上。

- **具体来看，Vera Rubin NVL72通过增加中板（midplane）和改变布局，减少铜缆的使用，以应对空间和可靠性挑战。**为了应对GB200/GB300组装中Overpass布线空间受限、可靠性低等挑战，VR NVL144 CPX采用无线缆设计。overpass被替换为Paladin板对板连接器，与位于机箱中间的midplane相连。除overpass外，连接OSFP cage与connectX网卡的线缆、连接PCIe至前端Bluefield DPU及本地NVMe存储的线缆、其他侧边带线缆均被取消。由于PCB承载了更多的信号传输，推动PCB材料升级，叠加新PCB增加，VR系列机柜PCB价值量大幅提升。

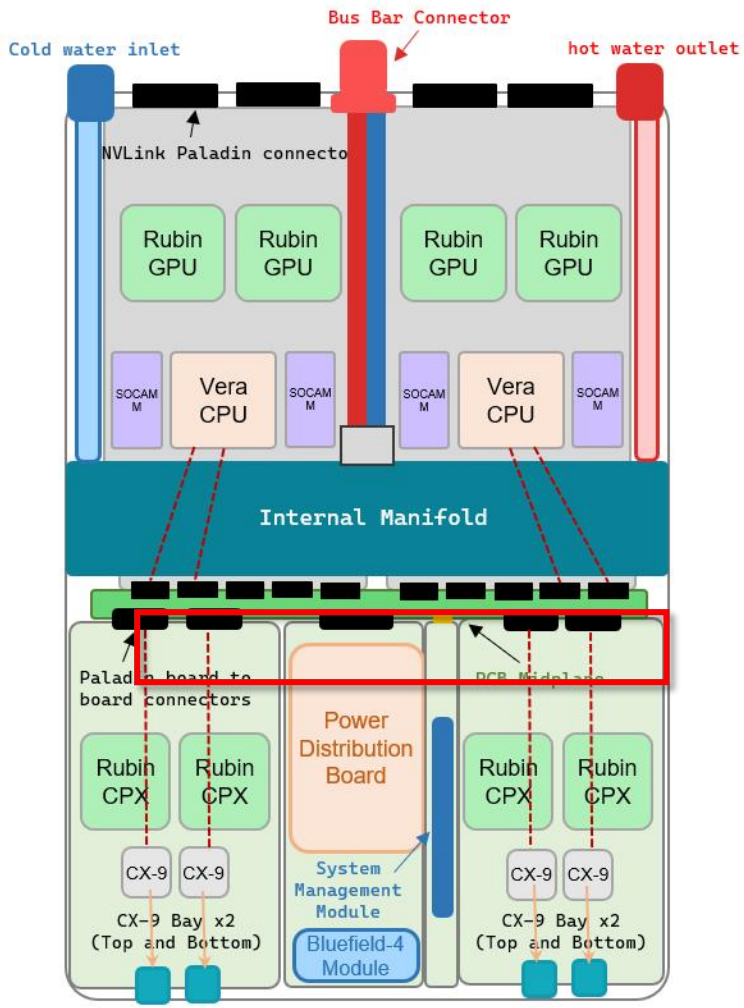
- **Midplane和CPX板的增加，带动单托盘PCB价值量快速提升，将成26年NV链PCB主要增量。**我们预计GB200 NVL72单机柜PCB价值量约23-25万元，单GPU对应PCB价值量为3000+元，而在满配的VR NVL 144 CPX机柜中，单GPU对应价值量达到8000余元（仍以单机柜72卡计算，不考虑CPX芯片）。

图：Compute Tray布局变化和scale out信号传输路径

GB200/300



VR NVL144



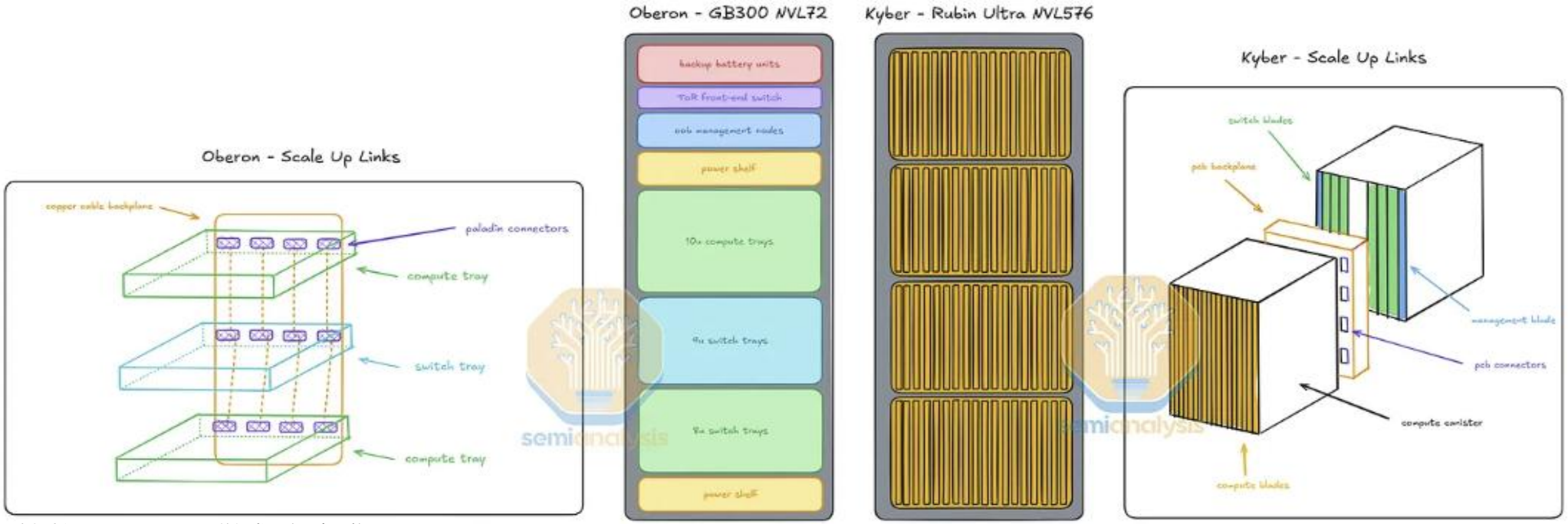
资料来源: SemiAnalysis，Nvidia，国信证券经济研究所整理

2.2 PCB重塑高密度链接格局，NV有望引入正交背板

- **展望2026年，正交背板是AI服务器领域PCB潜在的重要催化。**GTC 2025上，一个关键的更新是Kyber机架架构。Nvidia通过将计算托盘旋转90度安装，来提高机柜密度，推出NVL576（144个GPU封装）配置。Kyber机架架构与之前Oberon架构（GB300 NVL72）相比，主要区别包括：

1. 计算托盘旋转90度：为了实现更高的机架密度，计算托盘被旋转90度，采用墨盒形式。
2. 每个机架包含4个罐体：每个罐体有18个计算托盘，每个托盘中包含两个Rubin Ultra GPU和两个Vera CPU，因此每个罐体中有36个GPU封装（144个die），这使得机架内四个罐体总计达到576颗die。
3. PCB背板替代铜缆背板：作为GPU与机架内NVSwitch之间的扩展链路，PCB背板取代了铜缆背板。这一转变主要是由于在更小的空间内布置电缆的难度过高，无法继续使用铜线对每个刀片进行连接。
4. 机架背面的NVSwitch托盘通过PCB背板的背面连接到计算托盘。

图：Kyber机架与Oberon架构的对比



资料来源: Semianalysis，国信证券经济研究所整理

【3】算力+存力：国产算力通用芯片与ASIC方案齐发力，存力缺货涨价有望贯穿全年

3.1 国产算力芯片竞相涌现，把握自主可控历史机遇

国信证券 GUOSEN SECURITIES

- **海外算力芯片存在后门风险，国产算力芯片有望趁势崛起。**2025年7月，美国议员呼吁要求美出口的先进芯片必须配备“追踪定位”功能，同时美人工智能领域专家透露，英伟达算力芯片“追踪定位”“远程关闭”技术已成熟。国家网信办于7月31日约谈英伟达，要求英伟达就对华销售的H20算力芯片漏洞后门安全风险问题进行说明并提交相关证明材料。8月10日央视旗下媒体玉渊谭天发文指出，从硬件和软件的角度看，H20均存在后门风险，且H20既不先进，也不环保。

- 当前海外英伟达等算力芯片仍处于领先地位，但由于地缘政治、算力新规、网络安全等因素限制，发展国产算力芯片的重要性进一步提升。同时，国产算力芯片积极更新迭代，适配国产大语言模型等，有望加速实现国产算力芯片的自主可控。

表：国产主流算力芯片性能参数对比

厂商	产品	推出时间	工艺nm	功耗W	FP/BF16	INT8	显存GB	带宽GB/s	HBM	互联带宽
寒武纪	MLU370-X8	2022	7nm	250	96	256	48	614.4	2E	MLU Link 512 GB/s
	MLU590	2023	7nm	550	314.6		96	2700	2E	
	MLU690	2026	7nm		813		160	3000	3	
海光信息	深算三号	2024	7nm	350	350		64	1024	2E	448 GB/s
摩尔线程	MTT S4000	2023			98	196	48	768	GDDR6	MT Link 240 GB/s
	MTT S5000	2024					80			
沐曦	C500	2023	6nm	350	240	480	64	1800	2E	MetaXLink 896 GB/s
	C600	2025	12nm		优于A100		144	4000	3E	
壁仞	BR100/166	2022	7nm	550	1024	2048	64	1638	2E	Blink 448 GB/s
	BR200	2026	7nm							
燧原	S60	2021	12nm	300	128	256	32	1600	2E	GCU-LARE 300 GB/s
	L600	2025	7nm		优于H20		144	3600	3E	800 GB/s
昆仑芯	P800	2024		400	350		96	2300		
平头哥	PPU	2025	12nm				96		2E	700GB/s
华为	昇腾910	2019	7nm	350	320	640		1200	2	HCCS 784 GB/s
	昇腾910B	2023	7nm	500	376		64	1200	2E	
	昇腾910C	2025	7nm	900	790		128	1200	2E	

资料来源：寒武纪、海光信息、摩尔线程、沐曦、壁仞、燧原、昆仑芯、平头哥、华为等，国信证券经济研究所整理

3.1 国产算力芯片竞相涌现，把握自主可控历史机遇

国信证券 GUOSEN SECURITIES

- **蚂蚁部署万卡国产算力集群，DeepSeek UE8M0 FP8针对国产芯片设计。**在2025年世界互联网大会乌镇峰会前沿人工智能模型论坛上，蚂蚁集团平台技术事业群总裁骆骥表示，蚂蚁已部署万卡规模的国产算力集群，适配自研与各主流开源模型，训练任务稳定性超过98%，训练与推理性能可媲美国际算力集群，并全面应用于安全风控领域的大型训练与推理服务。同时，根据蚂蚁集团百灵团队技术成果论文《Every Flop Counts》，蚂蚁针对国产卡进行调优，其3000亿参数的MoE大模型可在国产GPU上完成训练，性能与使用海外芯片效果相当。此外，DeepSeek正式上线V3.1模型后，DeepSeek官方评论，UE8M0 FP8是针对即将发布的下一代国产芯片设计，UE8M0中U代表无符号优化，E代表指数位数，M代表尾数位数。

- 我们认为，当前国产算力卡仍在持续迭代追赶，虽然较海外英伟达有差距，但伴随国产大模型逐步接受和适配国产算力芯片，国产算力芯片的更新迭代有望迈入正循环，建议关注：寒武纪、摩尔线程、沐曦股份、翱捷科技、芯原股份、灿芯股份、澜起科技、龙芯中科等。

图：五类AI算力加速卡对比

Table 1: Characteristics of different AI accelerators (listed in descending order of availability).

Device	Peak FLOPS (T)	Memory (GB)	Fair Cost per Hour (RMB)	Support FP8
A	370	64	7	×
B	120	96	4.5	×
C	312	80	10	×
D	989	80	27.5	√
E	147	96	5.64	√

资料来源：《Every Flop Counts：SCALING A 300B MIXTURE-OF-EXPERTS LING LLM WITHOUT PREMIUM GPUS》，国信证券经济研究所整理

图：DeepSeek-V3.1搜索智能体测评

Benchmarks	DeepSeek-V3.1	DeepSeek-R1-0528
Browsecomp	30.0	8.9
Browsecomp_zh	49.2	35.7
HLE	29.8	24.8
xbench-DeepSearch	71.2	55.0
Frames	83.7	82.0
SimpleQA	93.4	92.3
Seal0	42.6	29.7

资料来源：DeepSeek官网，国信证券经济研究所整理

3.1 国产算力芯片竞相涌现，把握自主可控历史机遇

国信证券 GUOSEN SECURITIES

- **摩尔、沐曦陆续上市，政策支持助力算力突围**。科创板为硬科技企业开辟上市绿色通道，摩尔线程从2025年6月30日IPO申请获受理至9月26日过会，全程仅耗时88天，创下科创板审核速度新纪录，募资约80亿元用于新一代训推一体、图形芯片及AI SoC研发。沐曦股份于2025年10月24日科创板成功过会，募资约42亿元用于高性能通用GPU及人工智能推理GPU等研发和产业化。

➢ 摩尔线程：作为国产全功能GPU第一股，依托自主研发的MUSA架构，覆盖AI智算、专业图形加速、桌面级图形加速和智能SoC等多累产品。产品线涵盖政务与企业级智能计算、数据中心及消费级终端市场，能够满足政府、企业和个人消费者等在不同市场中的差异化需求。

➢ 沐曦股份：全面覆盖人工智能计算、通用计算和图形渲染三大领域，先后推出智算推理曦思N系列、训推一体和通用计算曦云C系列，以及图形渲染曦彩G系列。此外，沐曦打造了自主开放、高度兼容国际主流GPU生态（CUDA）的软件生态体系，具备易用性和可扩展性。

图：摩尔线程主要产品分类

分类			芯片	板卡/模组	一体机	集群设备
服务器级	AI 智算	企业级	第四代 GPU“平湖”	S5000	D800 X1/X	KUAE2
			第三代 GPU“曲院”	S4000		KUAE1
	专业图形加速	企业级	第二代 GPU“春晓”	S3000	D200/D400/D800	MCCX
			第一代 GPU“苏堤”	S1000/S2000	-	-
桌面级图形加速		消费级	第二代 GPU“春晓”	S70/S80		
		企业级	第二代 GPU“春晓”	X300		
			第一代 GPU“苏堤”	S10/S30/S50/X100		
智能 SoC 类		企业级	第一代 SoC“长江”	AI 模组-E300	AI 算力本-A140	
		消费级				
示意图						

资料来源：摩尔线程招股书，国信证券经济研究所整理

图：沐曦股份主要产品分类

产品类型	型号	产品特征	应用场景
训推一体 GPU	曦云 C500 系列	公司曦云 C 系列产品拥有多精度混合算力，内置大量运算核心，具有较强的并行计算能力和较高的能效比，适用于向量计算和矩阵计算等计算密集型应用，可广泛应用于智算训练与推理、通用计算、AI for Science 等场景	云端智算（训推一体）、通用计算、AI for Science 等
	曦云 C600 系列		
智算推理 GPU	曦思 N100 系列	公司曦思 N100 产品系面向传统人工智能场景，内置性能强劲的视频处理器和运算核心，可广泛应用于智慧城市、智慧交通、智慧教育、智能视频处理等场景	云端及边端推理、视频转码
	曦思 N260 系列	公司曦思 N 系列后续迭代产品系面向生成式人工智能场景，拥有多精度混合算力、大容量显存和较高的能效比，可广泛应用于大模型推理、生成式应用等场景	云端推理、一体机及工作站
	曦思 N300 系列		
图形渲染 GPU	曦彩 G100 系列	公司曦彩 G 系列产品系面向图形处理场景，内置性能强大的图形处理器，可广泛应用于云游戏、数字孪生、云渲染、影视动画和专业制图等场景	云端及边端图形处理

资料来源：沐曦股份招股书，国信证券经济研究所整理

3.1 国产算力芯片竞相涌现，把握自主可控历史机遇

- **央视播出智算中心项目成效，国产算力卡对齐英伟达H20。**9月16日，央视《新闻联播》播出了“中国联通三江源绿电智算中心项目建设成效”，介绍了阿里平头哥、沐曦股份、摩尔线程、壁仞科技、中昊芯英、太初元碁、燧原科技等多个国产AI芯片的已签约或拟签约情况，已签约项目共计1747台设备、22832张算力卡，总算力达3479P。

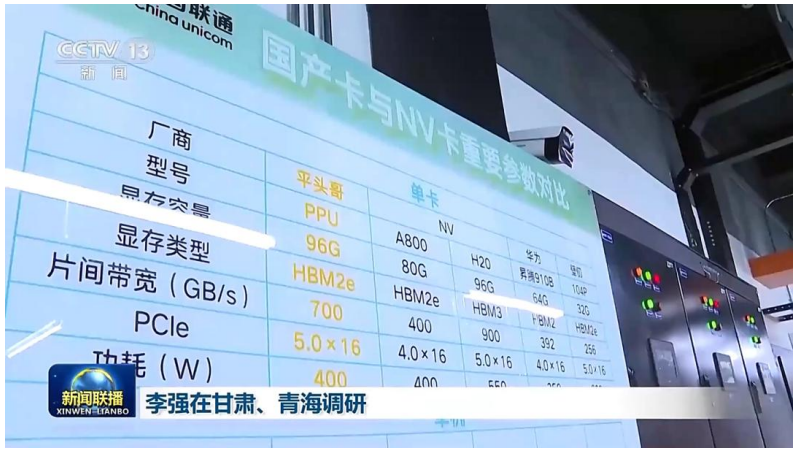
➢ 平头哥：平头哥PPU采用HBM2e显存，显存容量达96GB，片间带宽为700GB/s，功耗为400W，接口方面支持PCIe 5.0×16，多项配置规格超过A800、接近H20。

➢ 昆仑芯：百度旗下昆仑芯P800同时在被内部和外部客户采用，在中国移动2025年至2026年人工智能通用计算设备（推理型）集中采购项目中，基于昆仑芯的AI服务器产品，在3个标包中，分别拿下70%/70%/100%的份额，均排名第一，中标订单规模达十亿级。

图：中国联通三江源绿电智算中心项目

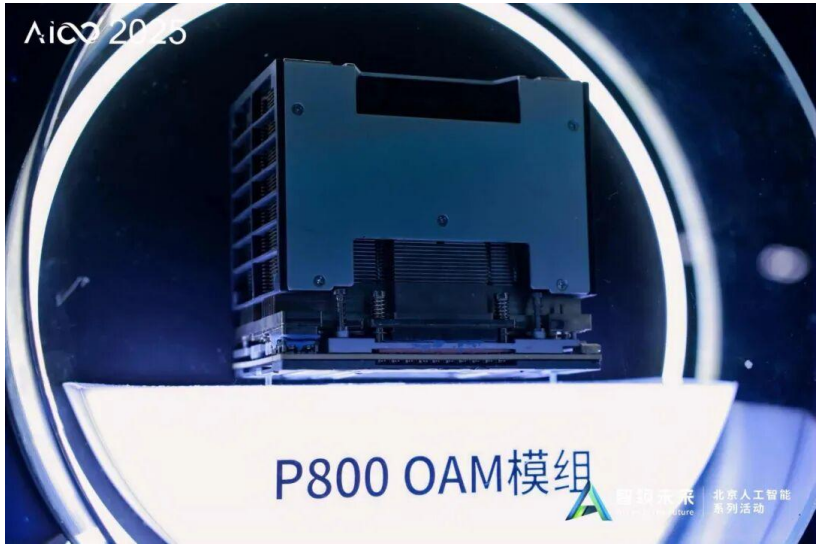
资料来源：新闻联播，国信证券经济研究所整理

图：国产卡与NV卡重要参数对比



资料来源：新闻联播，国信证券经济研究所整理

图：昆仑芯P800 OAM模组支持构建万卡级集群



资料来源：昆仑芯科技官网，国信证券经济研究所整理

3.1 国产算力芯片竞相涌现，把握自主可控历史机遇

国信证券 GUOSEN SECURITIES

- **华为计划1Q26推出昇腾950PR，4Q26上市超节点Atlas 950 SuperPoD。**2025年9月华为全连接大会上，华为首次公布2025-2028年昇腾芯片演进路线，预计2026年一季度推出昇腾950PR，互联带宽达到784GB/s，算力可在FP8下，达到1PFLOPs，内存为128GB，1.6TB/s。此外，华为发布了最新超节点产品Atlas 950 SuperPoD和Atlas 960 SuperPoD超节点，分别支持8192及15488张昇腾卡。同时基于超节点，华为发布了全球最强超节点集群Atlas 950 SuperCluster和Atlas 960 SuperCluster，算力规模分别超过50万卡和达到百万卡。基于昇腾950DT的超节点，可实现8 EFLOPs FP8算力，并实现跨柜全光互联16.3PB/s带宽；而基于昇腾950DT的超节点集群则可实现524 EFLOPs算力。目前，华为CloudMatrix 384超节点累计部署300+套，服务于互联网、金融、运营商、电力、制造等行业的20多个客户；未来，超节点将成为AI基础设施建设新常态。

图：2025-2028年昇腾芯片演进路线

资料来源：华为全连接大会，国信证券经济研究所整理

图：华为Atlas 950 SuperPoD

资料来源：华为全连接大会，国信证券经济研究所整理

3.1 谷歌TPUv7比肩英伟达GPU，ASIC对外销售开拓新市场

国信证券 GUOSEN SECURITIES

- **谷歌Ironwood性能比肩英伟达GB200，自研ASIC对外销售有望开拓新市场。**谷歌发布第七代TPU Ironwood，搭载了192GB的显存，是第六代TPU Trillium的6倍；显存带宽方面提升到7578 GB/s，是Trillium的4.5倍；双向带宽增加到1229 GB/s，是Trillium的1.5倍，单芯片峰值算力达4614 TFLOPs。此外，根据11月财联社等媒体报道，谷歌正寻求通过向客户提供本地部署选项，来拓展其TPU业务。其中，Meta Platforms正考虑2026年从谷歌云租用谷歌芯片，同时计划2027年斥资数十亿美元购买Google TPU。当前Ironwood性能比肩GB200，显示出推理端GPGPU并非唯一方案，自研ASIC已是明确的产业趋势。若未来谷歌将TPU对外销售，ASIC方案在云端推理领域市场规模有望进一步提升。

厂商	ASIC/GPU原厂	名称	发布时间	算力	连接		存力		
				算力FP16/矩阵稠密（TFLOPs/TOPs）	互联技术	互连速率(GB/s)	HBM/显存类型	显存带宽（GB/s）	显存容量（GB）
英伟达	英伟达	V100	2017	125	NVLink2	300	HBM2	900	32
		A100	2020	312	NVLink3	600	HBM2e	2039	80
		A800	2022	312	NVLink	400	HBM2e	2039	80
		H100	2022	989	NVLink4	900	HBM3	3430	80
		H800	2023	989	NVLink	400	HBM3	3430	80
		H200	2023	989	NVLink4	900	HBM3e	4915	141
		B200	2024	2500	NVLink5	1800	HBM3e	8192	192
		GB200	2024	5000	NVLink5	2*1800	HBM3e	16384	384
		B300	2025	3750	NVLink5	1800	HBM3e	8192	288
		Rubin	2025	6250	NVLink6	3600	HBM4	13312	288
谷歌	博通	TPU v4i	2020	69	ICI Links	200	-	300	8
		TPU v4	2021	275	ICI Links	672	HBM2	1228	32
		TPU v5e	2023	197	ICI Links	400	HBM2	819	16
		TPU v5p	2023	459	ICI Links	1200	HBM3	2765	95
		TPU v6e	2024	926	ICI Links	800	HBM3	1640	32
		TPU v7	**2025**	**4614**	**ICI Links**	**1229**	**HBM3e**	**7578**	**192**
亚马逊	Marvell	Trainium2	2023	667	NeuronLink v3 (PCIe Gen5)	1280	HBM3	2900	96
	Alchip	Trainium3	2024	1334	?	?	HBM3e	?	144
Meta	博通	MTIA T-V1	4Q25-1Q26	?	?	?	HBM3e	?	216
		MTIA T-V2	2027	?	?	?	HBM4	?	?

资料来源：Nvdia、Broadcom、Marvell等，国信证券经济研究所整理

3.1 谷歌TPUv7比肩英伟达GPU，ASIC对外销售开拓新市场

国信证券 GUOSEN SECURITIES

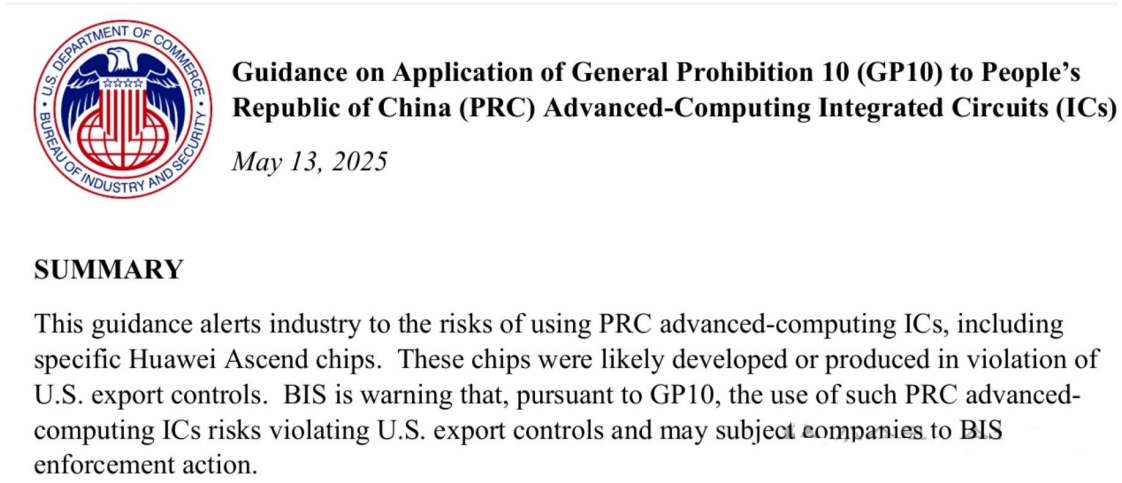
- **美国BIS多次限制国产算力芯片，合规ASIC项目成为互联网云厂优选项。**1月15日，美国BIS发布新规加强对先进计算半导体的限制，对14/16nm及以下逻辑芯片实施全球管控，该类芯片在2025/27/29年分别满足晶体管数量少于300/350/400亿个，并由可信赖的OSAT厂封装。5月13日，美国BIS发布公告，撤销拜登政府的《人工智能扩散规则》，同时宣布采取额外的措施加强全球半导体出口管制。新管制包括：1）指出在全球任何地方使用华为昇腾芯片均违反美国出口管制；2）警告公众将美国人工智能芯片用于训练和推理中国人工智能模型的潜在后果；3）向美国公司发布指导如何保护供应链免受转移策略的影响。

- 海外算力芯片采购受限且存在后门风险，推动国产算力芯片及合规自研ASIC发展。我们认为，国产AI算力芯片将坚定不移地走自主可控道路，而国内大模型及互联网厂商也有望进一步提升国产算力芯片的使用比例。但由于当前国产先进制程产能有限，云厂商自研合规ASIC项目成为短期最优选项之一。

图：2025年1月15日BIS加强先进计算半导体限制

资料来源：美国BIS官网，国信证券经济研究所整理

图：2025年5月13日BIS采取额外措施加强半导体出口管制



资料来源：美国BIS官网，国信证券经济研究所整理

3.1 谷歌TPUv7比肩英伟达GPU，ASIC对外销售开拓新市场

国信证券 GUOSEN SECURITIES

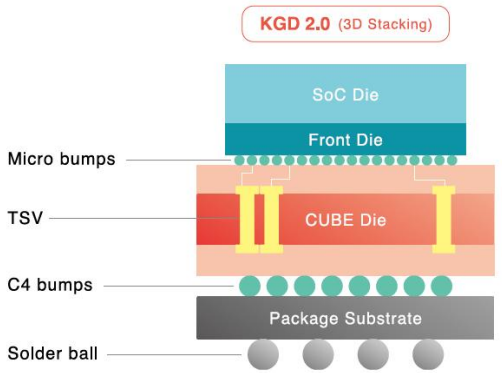
- **受限于算力芯片面积与晶体管数量，存储方案成为ASIC差异化竞争关键点。**当前海外先进制程的芯片代工受到芯片面积需小于300平方毫米，且晶体管数量需少于300亿个等限制，同时要求在指定OSAT厂封装。在此限制下，ASIC芯片项目虽然能够满足国内推理需求，但推理芯片本身难以形成差异化，而相关配套的存储方案成为关键点。国内厂商有望通过LPDDR、CUBE、3D DRAM堆叠等各类创新型存储方案，在显存容量、带宽、算力、功耗等方面，达到对标海外主流推理芯片近似效果的高性价比方案，建议关注国内ASIC公司：翱捷科技、芯原股份、灿芯股份等。

- 当前海外推理芯片以高通为例，其计划推出面向数据中心的人工智推理芯片AI200（2026 商用）和AI250（2027 商用）。AI200是专用机架级AI推理解决方案，为大型语言和多模态模型（LLM、LMM）推理及其他AI工作负载提供低总拥有成本（TCO）和优化的性能，每卡支持768GB LPDDR。AI250将首次采用近内存计算的创新内存架构，提供超10倍的有效内存带宽和更低的功耗。

图：高通计划发布推理芯片AI200和AI250

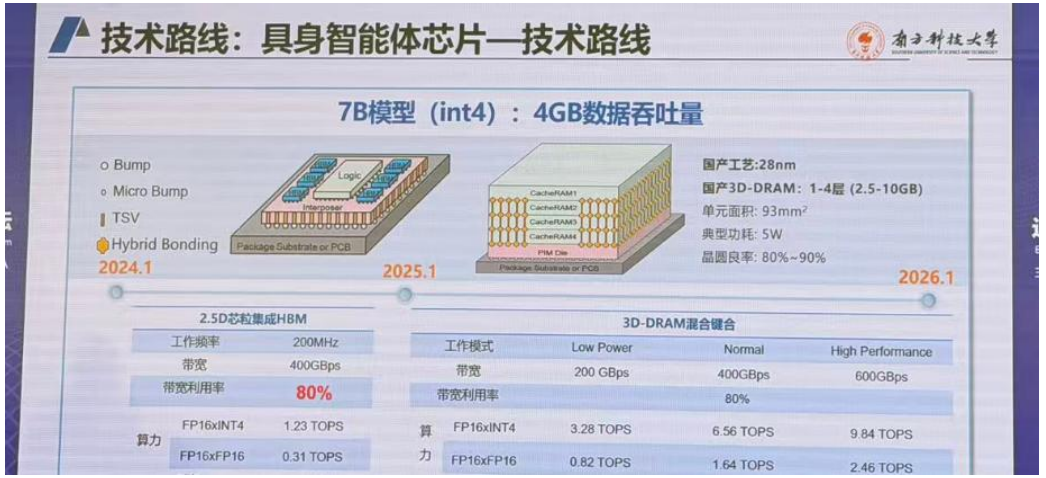
资料来源：高通官网，国信证券经济研究所整理

图：华邦CUBE存储方案



资料来源：华邦官网，国信证券经济研究所整理

图：南方科技大学3D-DRAM混合键合方案



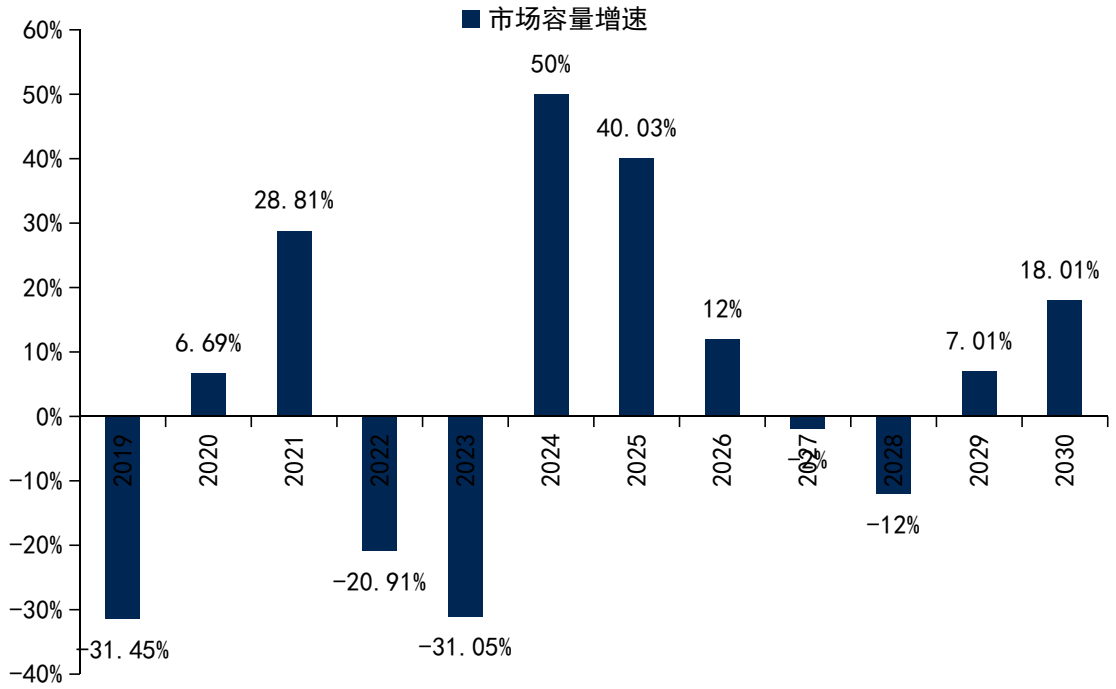
资料来源：南方科技大学、湾芯展论坛，国信证券经济研究所整理

3.2 存储产业持续增长，AI成为主要增长驱动力

国信证券 GUOSEN SECURITIES

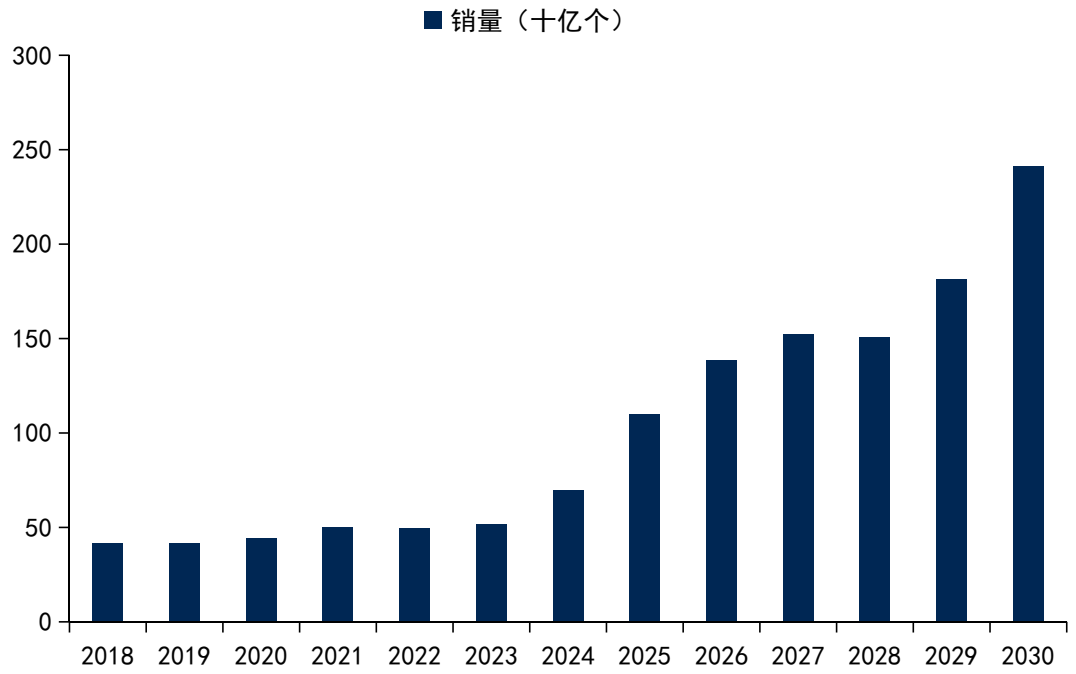
● **预计25年存储市场同比增长40%，产业从简单供需博弈向结构性增长转变。**根据Statista数据，预计存储市场2025年总规模有望同比增长 40%，出货量有望达1098亿颗，同比增长57%。经历2022至2023年周期性调整后，行业重新步入增长通道，一方面HBM等AI高价产品带动整体量价齐升，另一方面原厂在供给端保持收缩策略，整体存储价格回暖，市场容量保持增长。从需求结构看，AI需求为主要驱动力，消费电子等领域则保持平稳，需求结构化带来原厂产品结构优化，成熟产品库存去化、控产，AI服务器及数据中心存储则加速扩产升级，存储周期也有供需博弈向基于需求结构的供给侧调整转变。

图：2019-2030年全球存储市场容量增速预测（%）



资料来源：Statista，国信证券经济研究所整理

图：2019-2030年全球存储市场出货量情况预测（单位：十亿个）

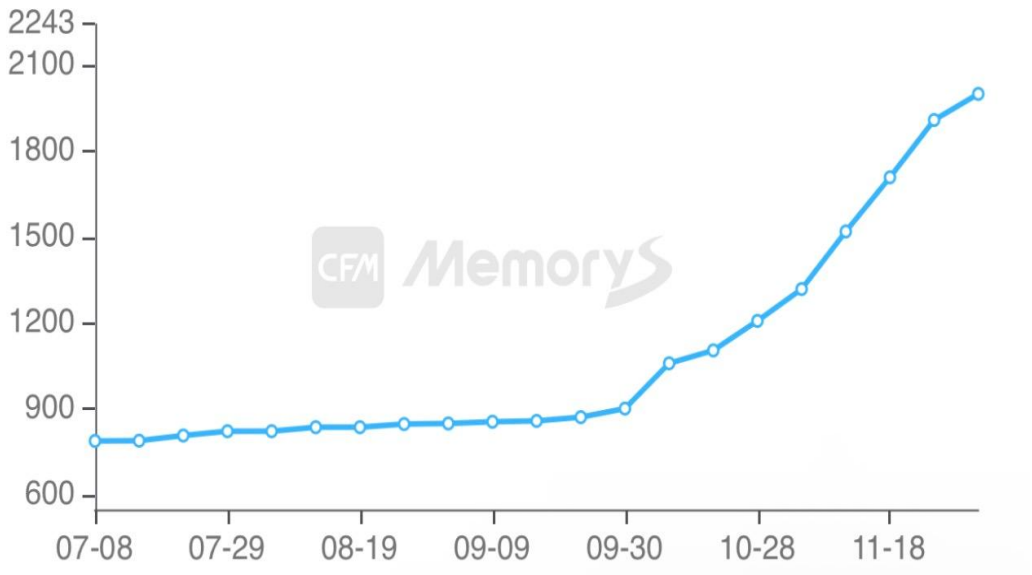


资料来源：Satista，国信证券经济研究所整理

3.2 DRAM：供给端优化，价格持续上涨

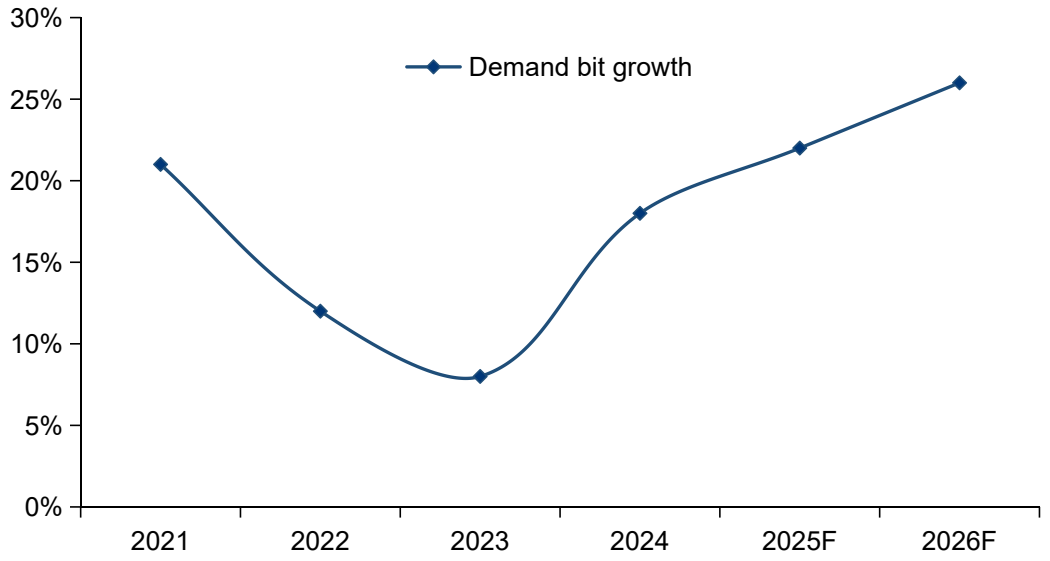
● **DRAM供给结构优化，25年价格持续上涨。**在AI拉动的结构性需求推动下，原厂将部分利润率偏低的传统DRAM产能转至DDR5、HBM等更高利润的产品。25年4月海外原厂接连宣布将停产DDR4、LPDDR4X等旧制程DRAM产品，DDR4、LPDDR4X价格开启涨价潮并延续至25年底；DDR5、LPDDR5X则受HBM产能挤占叠加新旧制程切换影响，供应趋于紧张，价格加速上涨。根据trendforce数据，DRAM在供给侧结构调整叠加AI需求拉动下仍将保持加速成长，预计26年DRAM位元需求量有望同比增加26%。

图：2025年下半年DRAM价格指数变化情况



资料来源：CFM闪存市场，国信证券经济研究所整理

图：2021-2026F DRAM 位元需求增速（%）



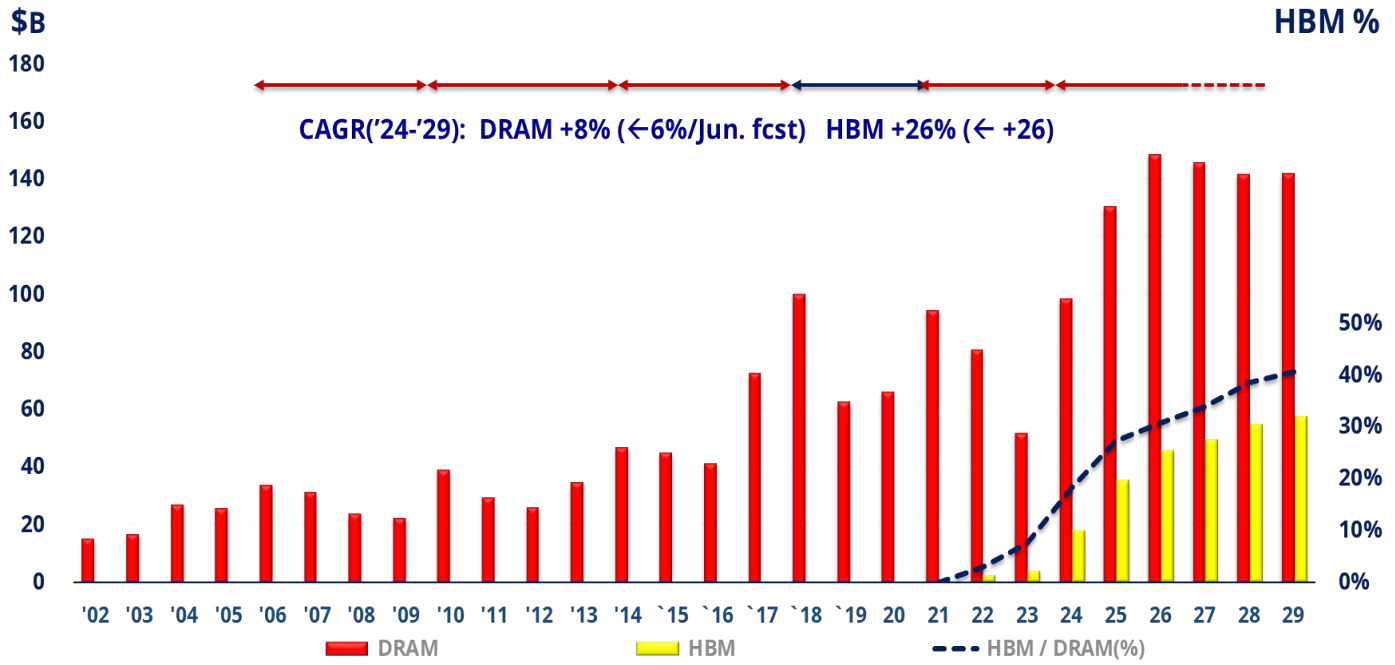
资料来源：Trendforce，国信证券经济研究所整理

3.2 DRAM：HBM进入高速发展期

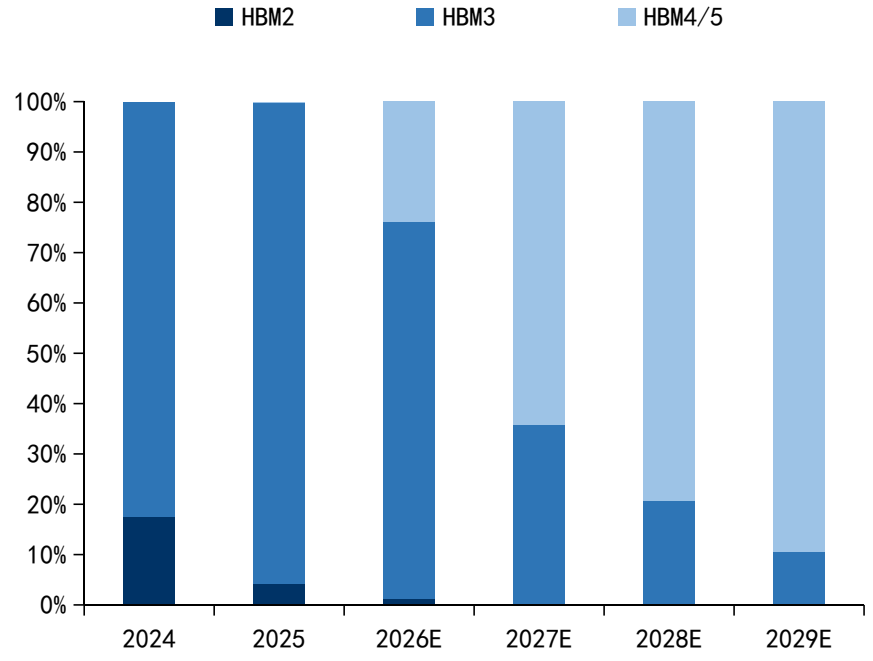
国信证券 GUOSEN SECURITIES

● 目前HBM通过3D堆栈与TSV技术，有效缩短处理器与存储器之间的距离，在即将量产的HBM4中，可实现更高通道密度与更宽I/O带宽以支撑AI GPU与加速器的超大规模运算。目前HBM4市场主要由SK海力士、三星和美光主导；SK海力士HBM份额最大，于2025年3月出货全球首款12层HBM4样品，计划在26年实现12层HBM4产品的量产。由于需求强劲，IDC预计2026年HBM市场市场规模有望超过460亿美元，HBM4的市占率则随着供应商持续放量而逐季提高，预计于26-27年将超过HBM3e产品成为市场主流。

图：HBM市场规模预测（十亿美元）



图：2024-2029F HBM结构变化



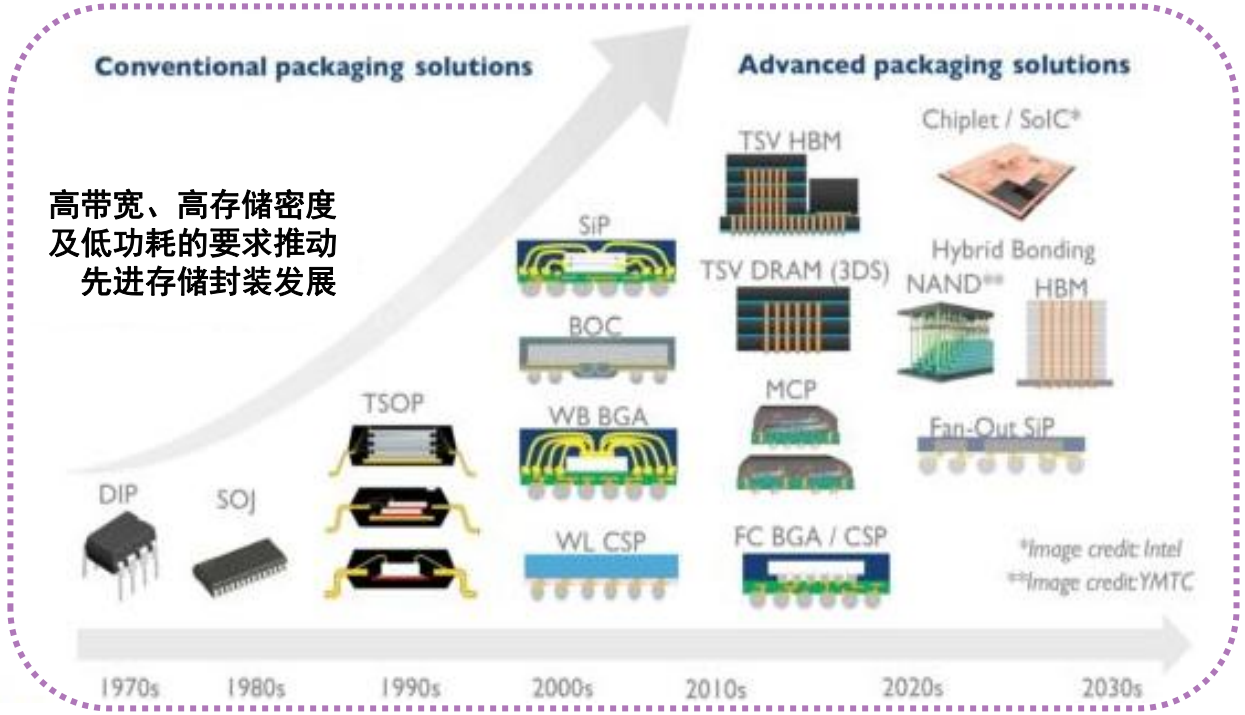
资料来源：IDC，国信证券经济研究所整理

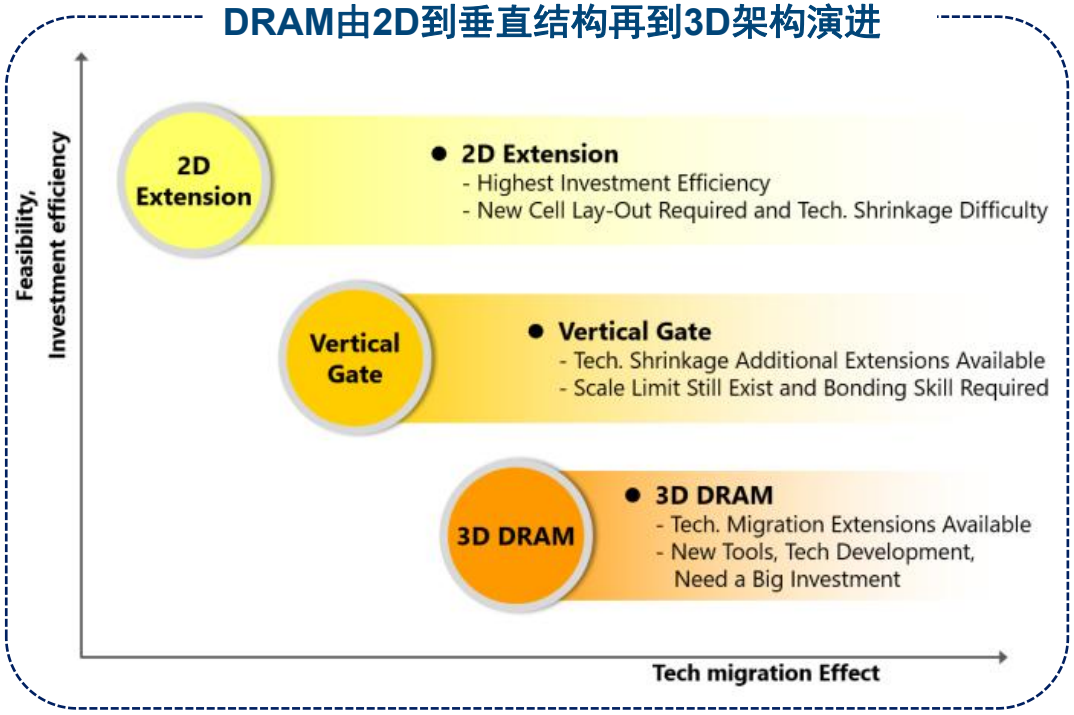
3.2 DRAM：存储介质迭代与先进封装打开多维增长空间

国信证券 GUOSEN SECURITIES

● **DRAM介质中长期向3D方向发展，短期先进封装为解决“内存墙”的主流路径。**DRAM的基本单位是存储单元（Cell），单元面积越小，有限空间内集成的单元就越多，电信号传输距离也越短，低功耗效率和处理速度得以提升。DRAM技术从最早的150nm不断缩小至10nm级别，然而随着尺寸进一步微缩，工艺的裕量存在极限，为减小单元尺寸并提升DRAM性能，垂直栅极（Vertical Gate）和三维DRAM（3D DRAM）被提出以解决目前的技术瓶颈。此外，DRAM自身结构演进仍需时间；以HBM为代表的先进封装方案成为短期解决内存和处理单元之间数据传输带宽受限即受到“内存墙”阻碍的主流形式；内存芯片逐步从“附属角色”转变为“性能瓶颈突破口”。

图：先进封装与存储介质技术同步发展





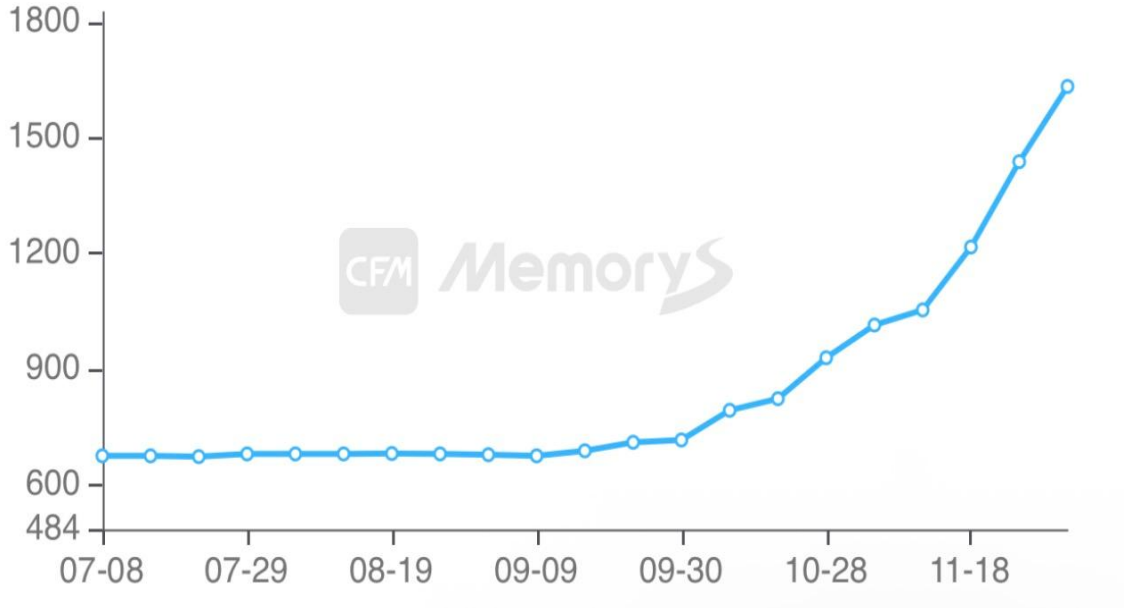
资料来源：Yole；Cha, Seon Yong. "Driving Innovation in DRAM Technology—Towards a Sustainable Future." 2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2025. 国信证券经济研究所整理

3.2 NAND：AI推理催化下迎来新增长

国信证券 GUOSEN SECURITIES

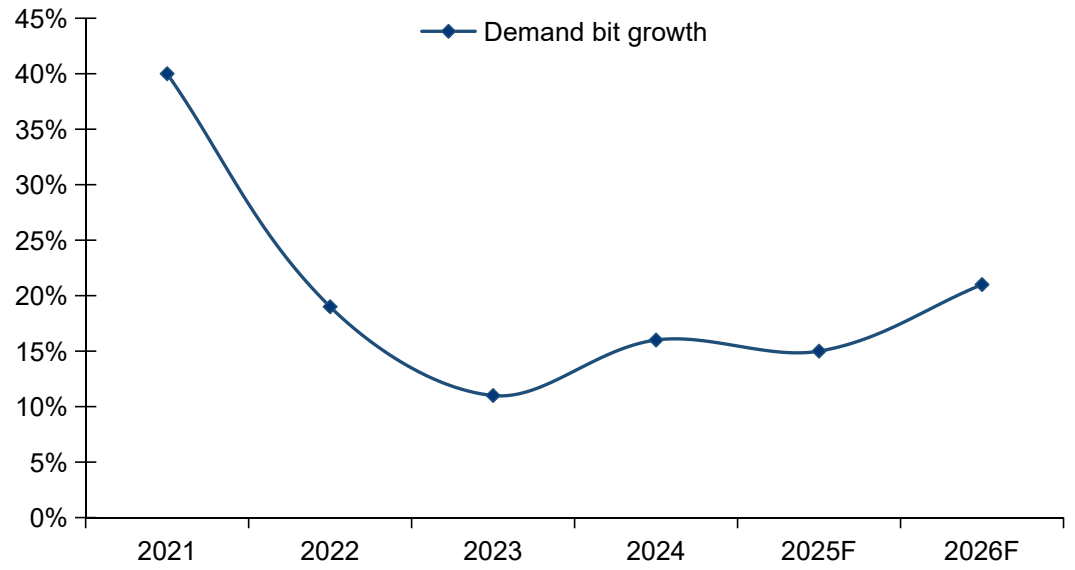
● **随着AI推理需求增加，NAND价格于25年9月进入加速上涨。**3Q25由于CSP厂商持续扩建AI基础建设，企业级SSD需求拉升，推动NAND价格持续上涨。NAND通过上半年减产，供需好转，叠加企业级SSD销售占比提高，NAND价格指数自9月初至今涨幅超40%。本次价格传导由企业级到行业大客户再到消费类渠道NAND，一方面，供给端原厂产能有限，另一方面CSP厂商已锁定26年产能，因此NAND供给紧张有望延续，根据Trendforce数据，预计26年NAND位元需求量有望同比增加21%。

图：2025年下半年NAND价格指数变化情况



资料来源：CFM闪存市场，国信证券经济研究所整理

图：2021-2026F NAND位元需求增速（%）



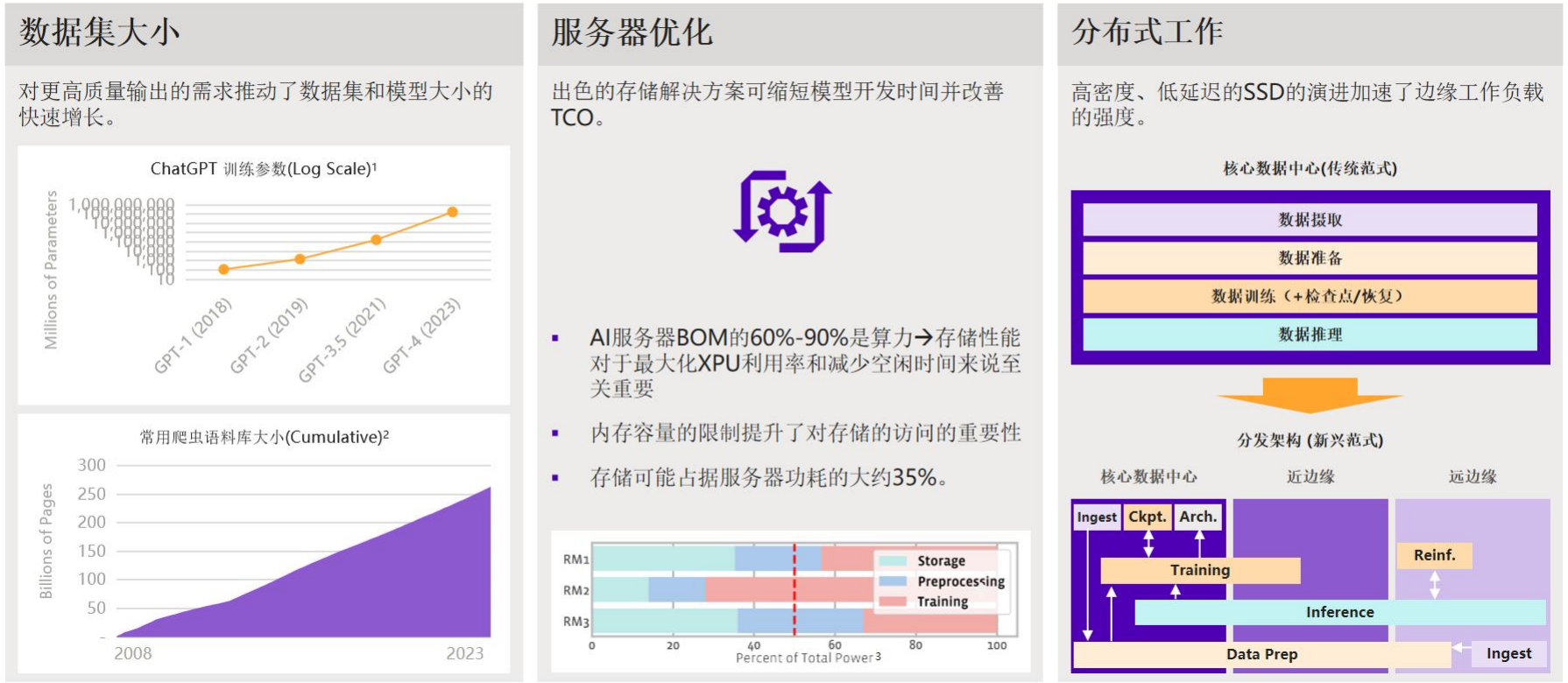
资料来源：Trendforce，国信证券经济研究所整理

3.2 NAND：AI推理带来NAND结构性增长，Nearline SSD需求提升

国信证券 GUOSEN SECURITIES

● **存储对AI的重要性体现在成本、功耗和空间的优化上。**数据集的增长速度呈现对数级态势，存力需求增加；AI服务器60-90%的物料成本用于各种计算资源（CPU、GPU、NPU等），但XPU需要高性能的存储来高效地提供数据，并在整个过程中保持高利用率，对于训练过程的顺利进行至关重要。如果存储性能不足，XPU可能会长时间处于空闲状态。在功耗方面，某些特定应用场景中存储消耗了整个服务器功耗的35%。如果能通过采用更高密度的存储和其他优化措施来降低这一比例，能够节省大量的电力和资金，企业级SSD是最优选择。

图：SSD在AI中的价值



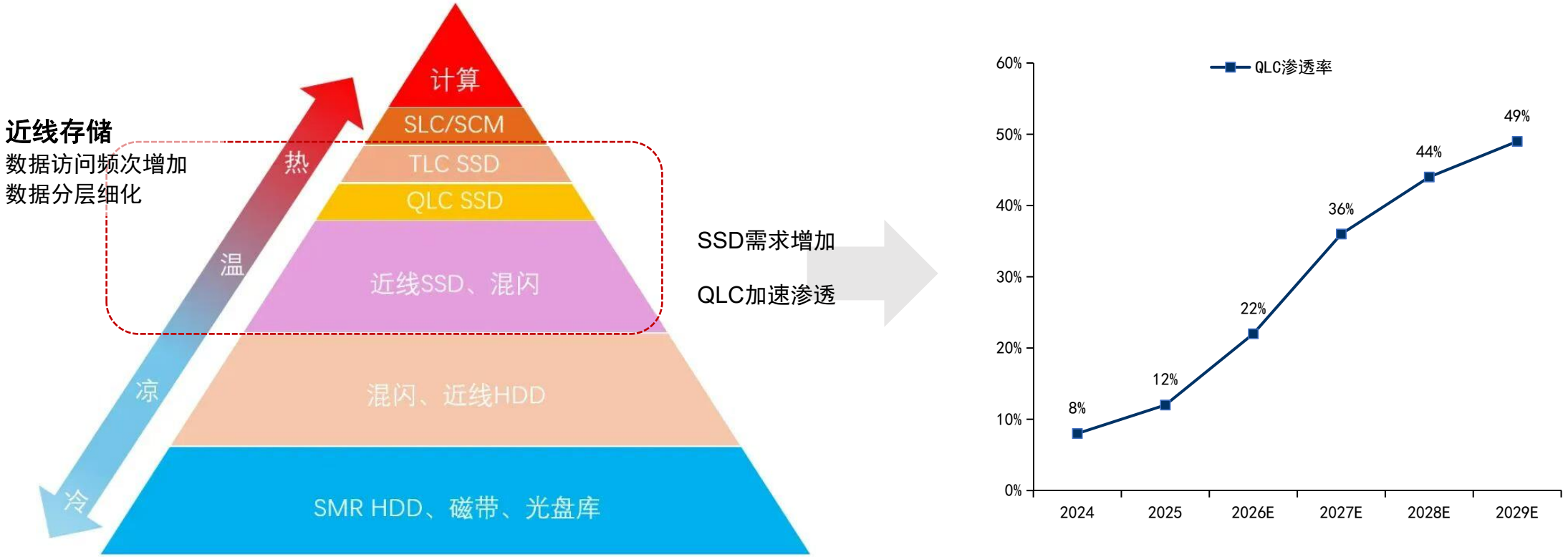
资料来源：Counterpoint，国信证券经济研究所整理

3.2 NAND：AI推理带来NAND结构性增长，Nearline SSD需求提升

国信证券 GUOSEN SECURITIES

● **随着AI技术在各行各业的广泛应用，AI推理服务成为连接算法与实际应用的关键桥梁。**这些服务需要处理海量的数据，并进行实时或近实时的分析，对储存系统的容量、速度和效率提出了高要求。传统的HDD在读写速度、延迟及能效方面的局限性推动了SSD渗透。传统近线（Nearline SSD）存储为访问量并不大的数据存储，而随着AI推理带来数据的访问频次提升，传统近线存储遇到性能瓶颈，例如硬盘阵列相对较慢的数据传输率会限制训练时GPU获取数据的效率。叠加传统机械硬盘产能有限，推动固态硬盘（SSD）需求加速攀升。目前各大NAND Flash供应商正加速Nearline QLC NAND Flash产品的验证与导入，根据IDC数据，预计QLC NAND渗透率26年有望达22%。

图：近线存储推动SSD需求提升

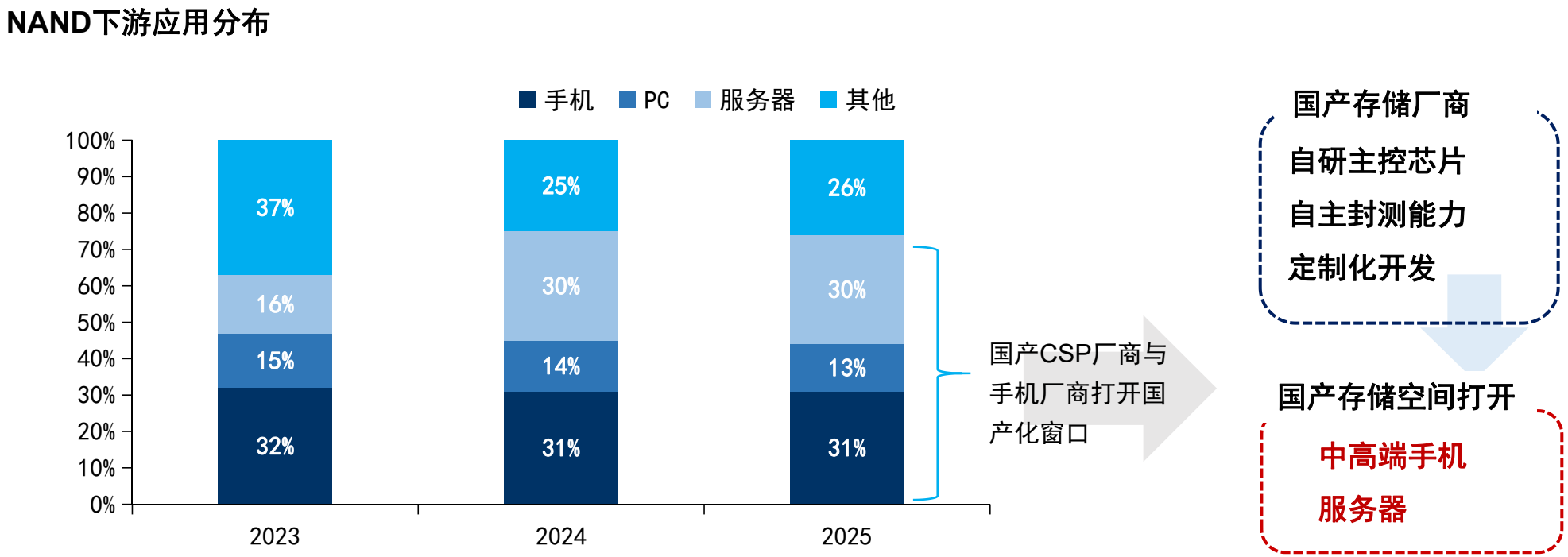


资料来源：Counterpoint，国信证券经济研究所整理

3.2 国产存储厂商迎国产化增长机遇

● **国产化进程打开存储模组产业升级机遇。**根据CFM数据，NAND Flash下游需求中手机约占31%，服务器占近30%，此前该类应用以海外原厂为主，国产厂商较少。随着AI应用出现，海外原厂将产能重心向高附加值的AI应用倾斜，退出部分存量市场。此外，国内模组厂商封测、主控等能力逐步提升，在此背景下，下游终端用户打开国产化窗口，国产存储逐步进入服务器、手机等中高端应用：在企业级市场，下游CSP厂商资本开支增加，德明利、江波龙代表的国产模组企业级产品加速放量；在手机端，标准化产品已进入国产放量期，针对中高端手机应用中，国产模组厂如江波龙通过开发5nm制程主控芯片、佰维通过晶圆级封装迎来导入机遇。

图：国产化窗口带来国产厂商增量机遇



资料来源：CFM闪存市场，国信证券经济研究所整理

【4】运力+电力：AI运力已成为AI系统功能的基石，AI算力增长推动电源架构同步升级

4.1 运力：为AI时代高带宽、低延迟的数据传输提供保障

国信证券 GUOSEN SECURITIES

● **AI技术及应用的快速发展不仅推动算力、存力需求激增，运力的重要性和需求也在增加。**算力、存力、运力三者作为AI基础设施的三大支柱，其中运力已成为系统功能的基石，决定计算与存储之间及其内部的数据传输效率。互连需要全面的解决方案，不仅要解决数据进出内存的问题，还要实现服务器内部、机架级服务器之间以及集群之间整个系统基础架构的顺畅通信。一个高性能的AI系统，需同时具备强算力支撑数据处理、大容量存储保障数据供给，以及高性能运力来实现高带宽、低延迟的数据传输，三者协同才能全面提升系统整体效率。

图：运力是AI基础设施的基石

资料来源：澜起科技公告，国信证券经济研究所整理

4.1 运力：为AI时代高带宽、低延迟的数据传输提供保障

国信证券 GUOSEN SECURITIES

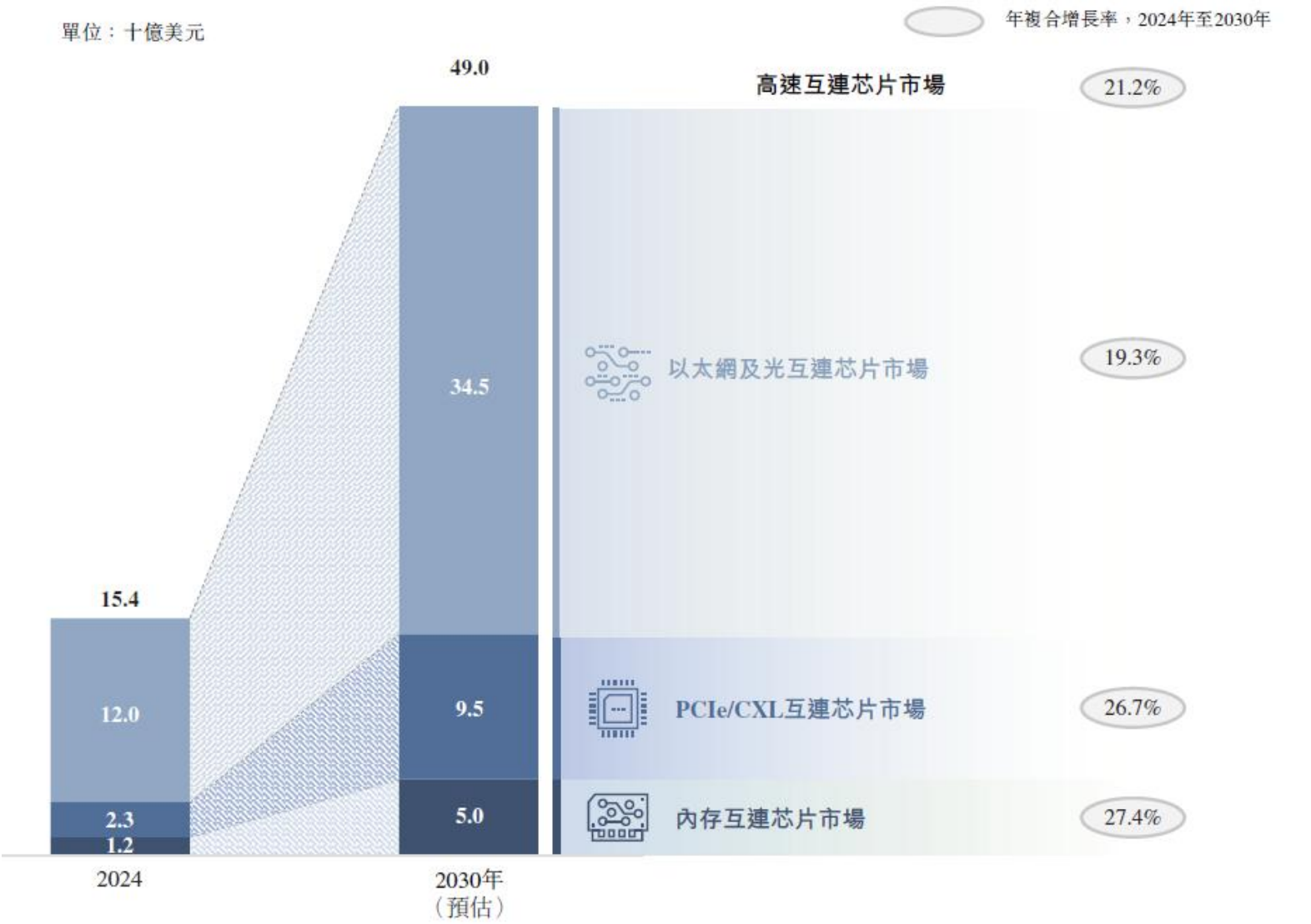
- **预计2024-2030年全球高速互连芯片市场规模CAGR为21.2%，中国市场占比将提高。**根据弗若斯特沙利文的预测，2024年全球高速互连芯片市场规模为154亿美元，预计2030年将增长至490亿美元，CAGR达21.2%，中国市场的占比将由25%提高至30%。

- **内存互连芯片**市场规模将由12亿美元增至50亿美元，CAGR为27.4%。

- **PCIe/CXL互连芯片**市场规模将由23亿美元增至95亿美元，CAGR为26.7%；

- **以太网及光互连芯片**市场规模将由120亿美元增至345亿美元，CAGR为19.3%。

图：全球高速互连芯片市场规模



资料来源：澜起科技公告，弗若斯特沙利文，国信证券经济研究所整理

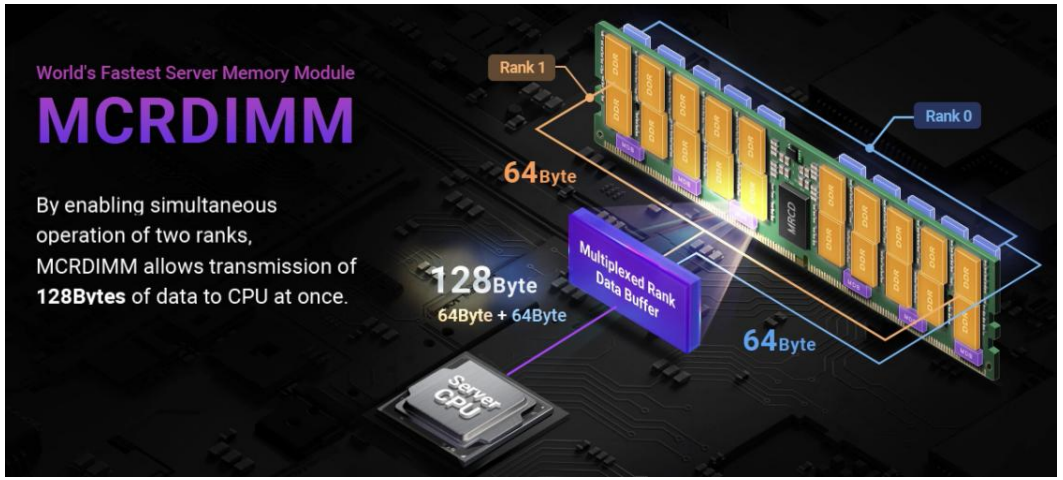
4.1 运力：为AI时代高带宽、低延迟的数据传输提供保障

国信证券 GUOSEN SECURITIES

● **MRDIMM内存模组标准为满足AI需求而生，可支持更高的速率。**为了满足AI对更高带宽、更高容量内存模组的需求，JEDEC组织制定了服务器MRDIMM内存模组相关技术标准。DDR5 MRDIMM提供创新、高效的新模块设计，有效地增加带宽而无需额外的物理连接并提供无缝带宽升级，使应用程序能够超过DDR5 RDIMM数据速率，第一子代MRDIMM支持8800MT/s速率，第二子代MRDIMM支持12800MT/s，第三子代MRDIMM的数据传输速率预计超过14000MT/s。

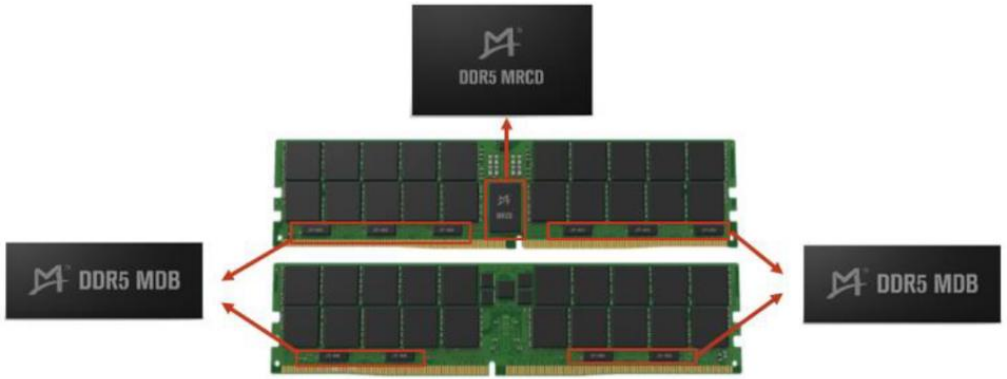
● **MRDIMM需要搭配1颗MRCD和10颗MDB芯片。**MRDIMM内存模组采用了LRDIMM“1+10”的基础架构，需要搭配的内存接口芯片为1颗MRCD芯片和10颗MDB芯片。其中MRCD芯片负责缓冲和中继来自内存控制器的地址、命令、时钟和控制信号，MDB芯片则负责缓冲和中继来自内存控制器或DRAM内存颗粒的数据信号。与普通的RCD芯片、DB芯片相比，设计更为复杂、速率更高，价值量也将有所提升。

图：MRDIMM/MCRDIMM示意图



资料来源：SK海力士官网，国信证券经济研究所整理

图：MRDIMM需要搭配1颗MRCD和10颗MDB芯片



资料来源：澜起科技公告，国信证券经济研究所整理

4.1 运力：为AI时代高带宽、低延迟的数据传输提供保障

国信证券 GUOSEN SECURITIES

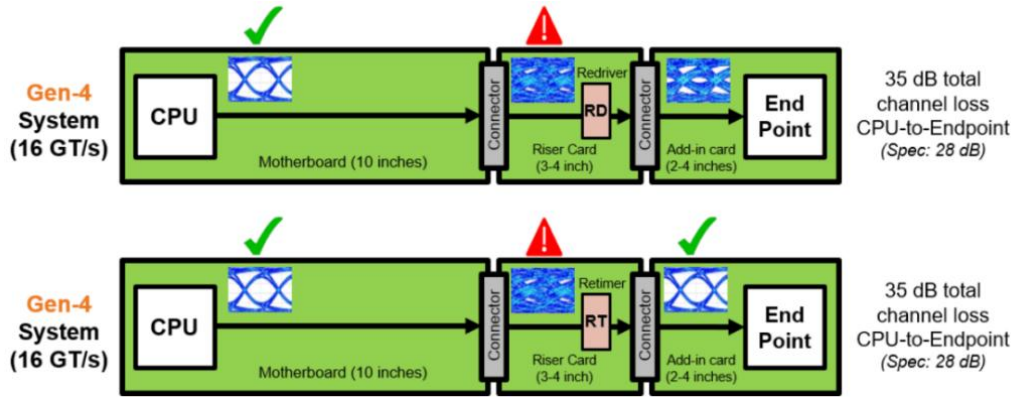
- **PCIe协议是一种高速串行计算机扩展总线标准，已成为主流通用互连接口。**PCIe（Peripheral Component Interconnect Express）是在PCI的基础上，为解决总线带宽问题发展而来，用于连接计算机的主板和各种外围设备，如GPU、固态硬盘（SSD）、网卡等。目前PCIe已成为主流互连接口，全面覆盖了包括PC、服务器、存储系统、手持计算等各种计算平台。

- **PCIe互连芯片是数据中心和服务器高速数据互连的核心组件，包括PCIe Retimer芯片和 PCIe Switch芯片。**预计市场规模将由2024年的22.89亿美元增长至2030年的77.61亿美元。

- **PCIe Retimer芯片：**适用于PCIe协议的超高速信号调理芯片，主要用于解决数据在高速、远距离传输场景中时序不齐、损耗严重、完整性差等问题，在CPU与高速外设（如GPU、AI加速卡、SSD卡及网卡等）的互连中发挥重要作用。一台主流的8卡GPU服务器通常需配套8-16颗PCIe Retimer芯片，预计全球市场规模将由2024年的3.95亿美元增长至2030年的18.90亿美元。

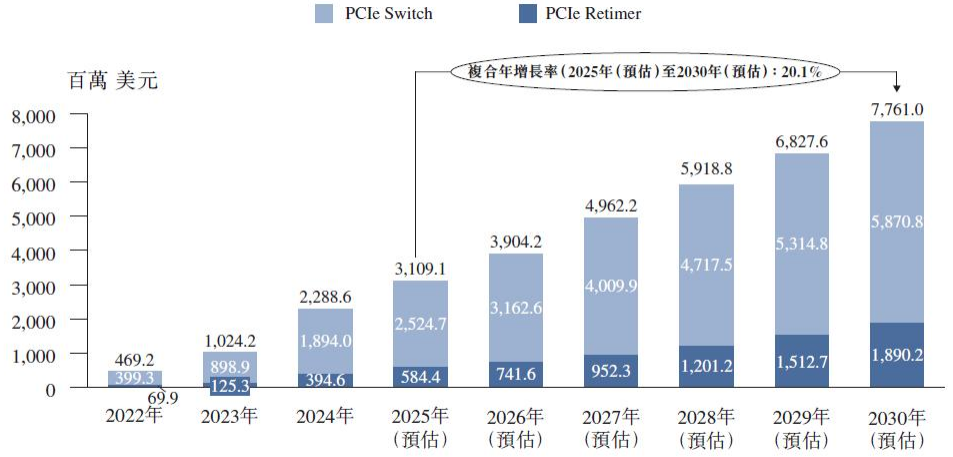
- **PCIe Switch芯片：**用于扩展接口数量，通过内部交换架构实现多设备的高速数据转发。一台主流的8卡GPU服务器通常需配套2-4颗PCIe Switch芯片，预计全球市场规模将由2024年的18.94亿美元增长至2030年的58.71亿美元。

图：PCIe Retimers可解决PCIe高速数据传输过程中的信号质量问题



资料来源：AsteraLabs官网，国信证券经济研究所整理

图：PCIe互连芯片市场规模



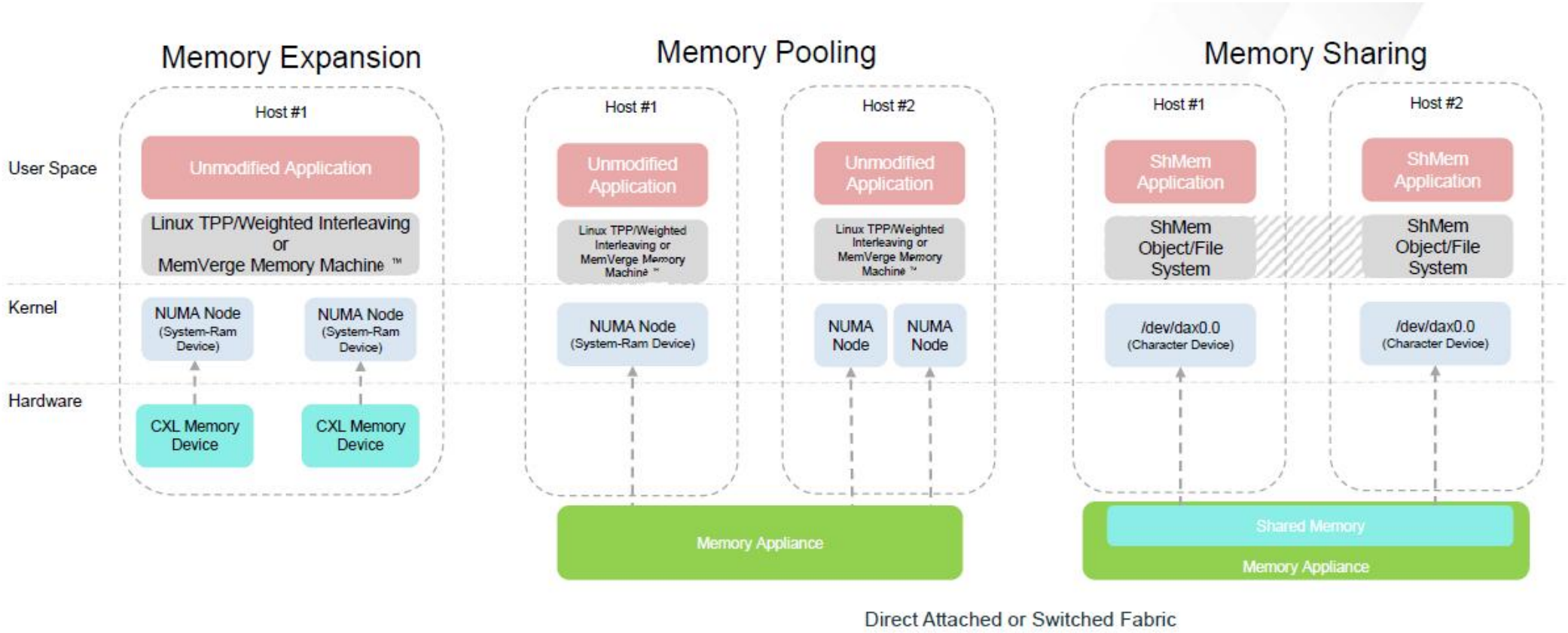
资料来源：澜起科技公告，弗若斯特沙利文， 国信证券经济研究所整理

4.1 运力：为AI时代高带宽、低延迟的数据传输提供保障

国信证券 GUOSEN SECURITIES

● **CXL可用于解决计算系统的内存墙问题。**CXL（Compute Express Link）是一种建立在PCIe基础上开放式高速互连技术，旨在通过统一内存地址空间提高数据中心和高性能计算中CPU、内存及加速器之间通信效率，同时保持低延迟和高带宽，满足高性能异构计算与存储的需求。在数据密集的计算系统中，CXL可用于内存扩展、内存池化和内存共享，达到扩展存储容量、高带宽低延迟获取内存池、根据工作负载动态分配内存、减少数据移动等目的。

图：CXL在大规模数据中的应用（内存扩展、内存池化、内存共享）



资料来源：CXL官网，国信证券经济研究所整理

4.1 运力：为AI时代高带宽、低延迟的数据传输提供保障

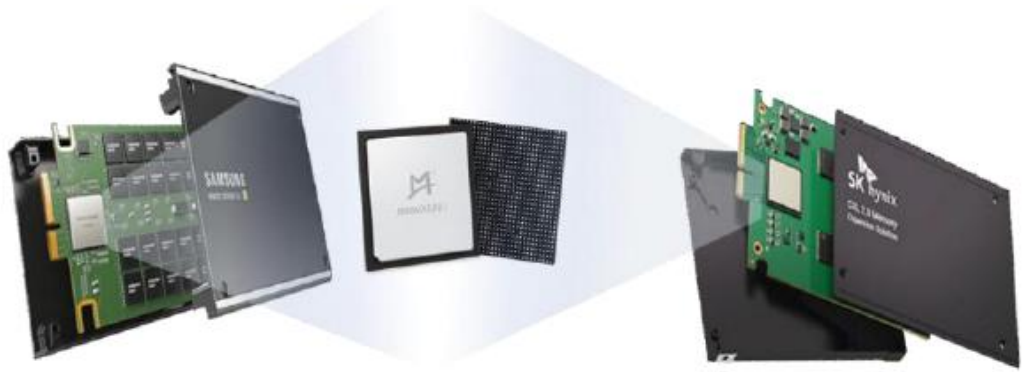
国信证券 GUOSEN SECURITIES

● **CXL互连芯片包括CXL MXC芯片和CXL Switch芯片。**CXL互连芯片是基于CXL协议构建的高速互连核心器件，可实现CPU、GPU、内存间高速低延迟的数据交互，包括CXL MXC芯片和CXL Switch芯片。预计市场规模将由2024年的430万美元增长至2030年的17.03亿美元。

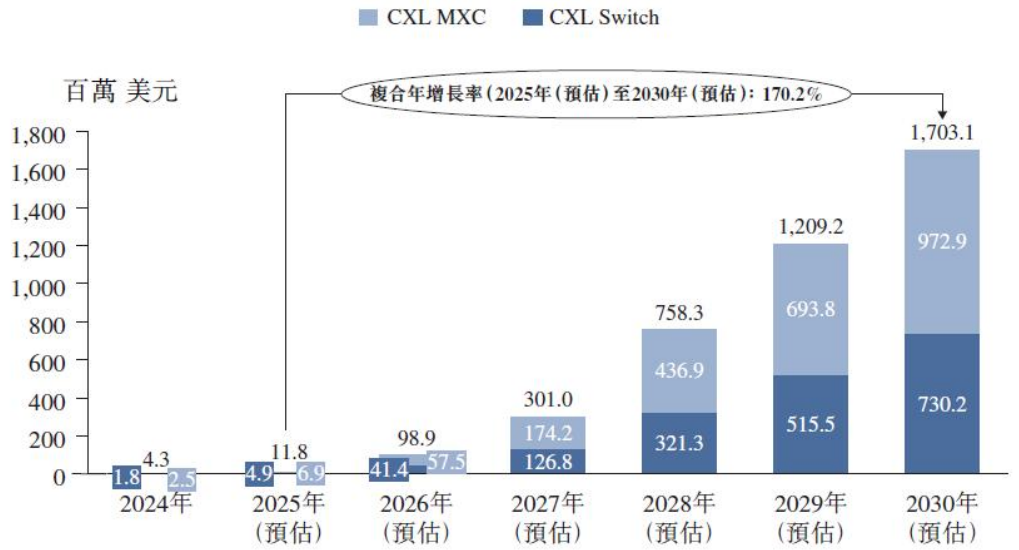
□ **CXL MXC芯片：**负责完成协议转换、内存访问调度及一致性控制，是构建内存扩展和内存池化架构的关键控制器，能够有效提升内存容量和带宽，单个内存池通常需要配置16-32颗CXL MXC芯片。预计全球市场规模将由2024年的250万美元增长至2030年的9.73亿美元。

□ **CXL Swithc芯片：**用于实现多个CXL主机及设备之间的互连与资源管理，支持单个或者多个CPU连接多个CXL内存或加速设备，提升系统扩展能力和资源利用效率，单个内存池通常需要配置2-4颗CXL Switch芯片。预计全球市场规模将由2024年的180万美元增长至2030年的7.30亿美元。

图：MXC在EDSFF模组中的应用



图：CXL互连芯片市场规模



资料来源：澜起科技公告，国信证券经济研究所整理

资料来源：澜起科技公告，弗若斯特沙利文，国信证券经济研究所整理

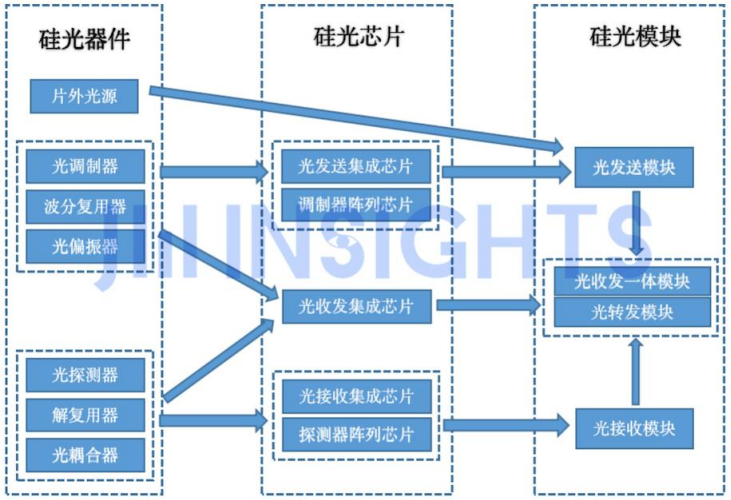
4.1 硅光技术：高性能数据传输的技术源头与基础底座

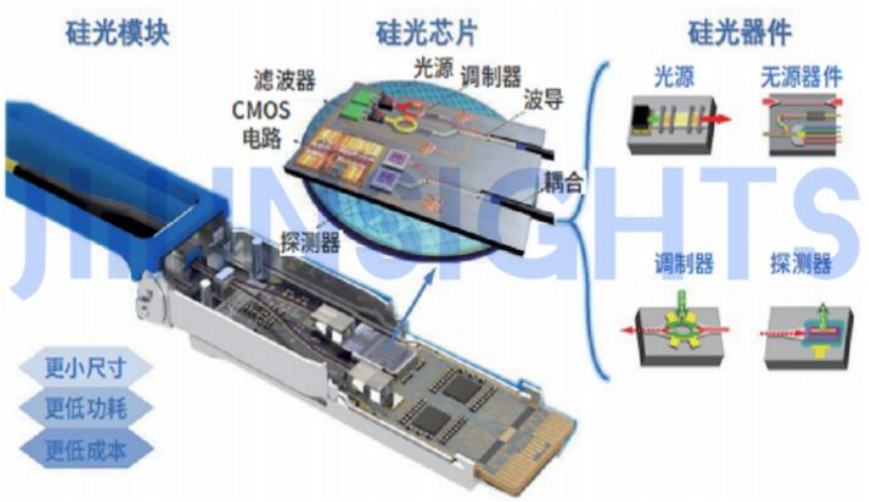
国信证券 GUOSEN SECURITIES

- 随着AI大模型对算力需求的爆发，传统的“电连接”已经无法满足芯片之间、服务器之间高速、低功耗的数据传输需求，必须用“光”来接力。**硅光是一种利用硅基工艺制造光学器件的底层技术，**其利用现有的成熟半导体工艺（CMOS工艺），在硅基衬底上制造光学器件（如调制器、波导、探测器）。其核心优势在于把“光”的器件像“电”的器件一样集成在芯片上，实现了高集成度、低成本、低功耗，是实现后面所有高级封装（如CPO）和高性能模块的必要技术基础。

- **硅光产品可以分为三个层级，分别是硅光器件、硅光芯片和硅光模块。**硅光模块主要由硅光器件、驱动电路和光接口组成。相较传统光模块，硅光模块具有传输速率大、集成度高、传输损耗低等优势，在通信互联系统中发挥着重要作用，随着大数据时代的到来，硅光模块市场前景广阔。

图：硅光产品的三个层级及结构示意图





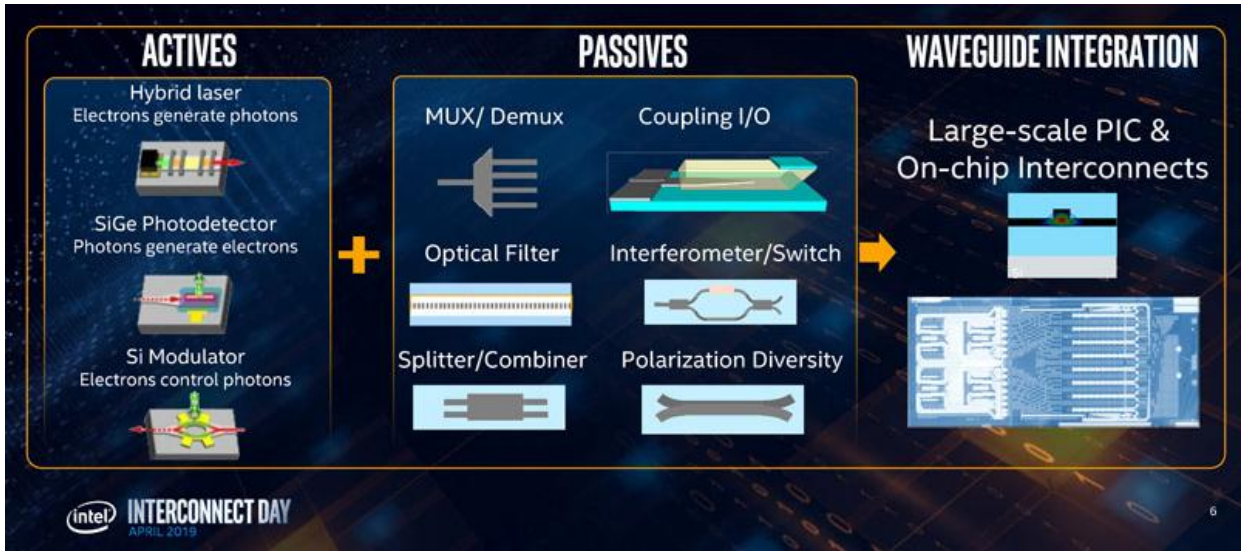
资料来源：集微咨询，国信证券经济研究所整理

4.1 硅光芯片：集成了光路的硅基芯片，硅光模块的“心脏”

国信证券 GUOSEN SECURITIES

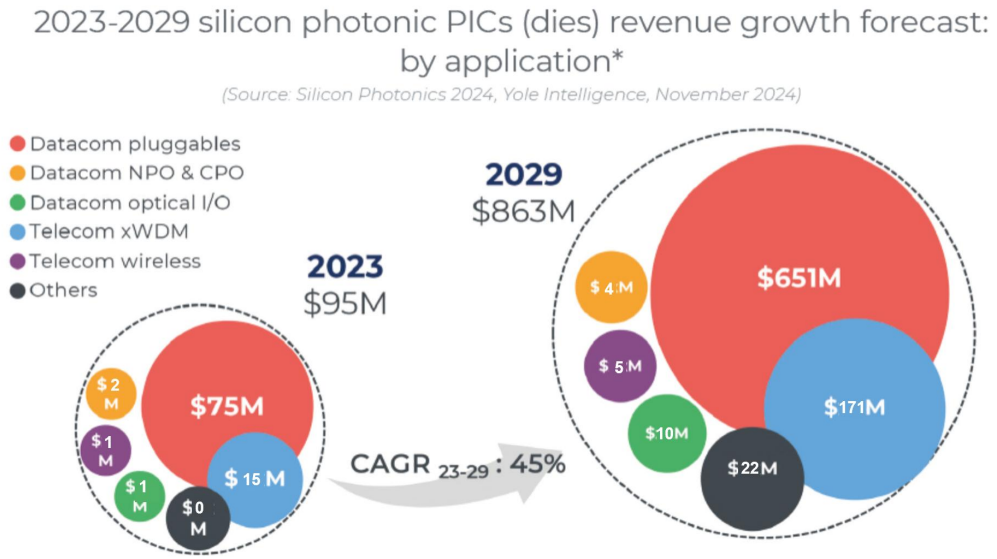
- **硅光芯片是硅光技术的核心引擎。**硅光芯片集成了激光器、调制器、探测器等光路，其作用为负责最核心的“光电转换”工作，即将电信号转换成光信号进行发送，或将接收到的光信号转换为电信号。
- **硅光芯片可分为有源和无源两大类。**有源光芯片包括激光器、调制器、探测器等。其中，激光器是将电信号转化为光信号的关键器件；调制器将输入的电信号通过物理效应转换为光信号，精确调控光波的强度、相位或频率等参数，以实现信息在光纤中的高效传输；探测器通过光电效应将接收到的光信号转换为电流信号。无源光芯片包括光波导、耦合器、复用/解复用器件等，分别用于光路由、光信号与硅光芯片的耦合以及波分复用/解复用。
- **预计到2029年硅光芯片销售额将达到8.63亿美元，2023-2029年CAGR达45%。**据Yole数据，硅光芯片销售额有望从2023年的0.95亿美元快速增长至2029年的8.63亿美元。其中，到2029年数据中心可插拔光模块规模增长至6.51亿美元，占2029年硅光芯片销售额的75%；用于电信波分复用领域硅光芯片的规模增长至1.71亿美元，占比达20%。

图：硅光芯片的组成



资料来源：Intel官网，国信证券经济研究所整理

图：硅光芯片市场规模及预测



资料来源：Yole，国信证券经济研究所整理

4.1 硅光模块：基于硅光芯片的封装产品，硅光的核心组件

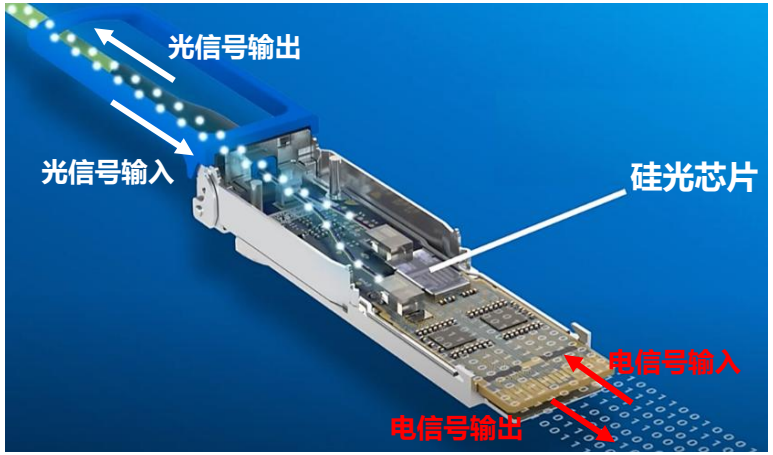
国信证券 GUOSEN SECURITIES

- **硅光模块是硅光芯片加上驱动电路、DSP等封装在一起的完整硬件产品。**作为基于硅光子技术的新一代光通信器件，其应用场景跟传统光模块类似。相比传统的光模块，硅光模块体积更小、传输速度更快、功耗更低。目前市面上硅光模块主流形态仍为可插拔的硅光模块（如800G、1.6T模块）。

- **高集成度及兼容CMOS工艺是硅光模块的核心优势。**传统光模块中各器件分立，需要连接与封装；硅光模块以其材料特性以及CMOS工艺的先天优势，能够很好的满足数据中心对更低成本、更高集成、更低功耗、更高互联密度等要求。

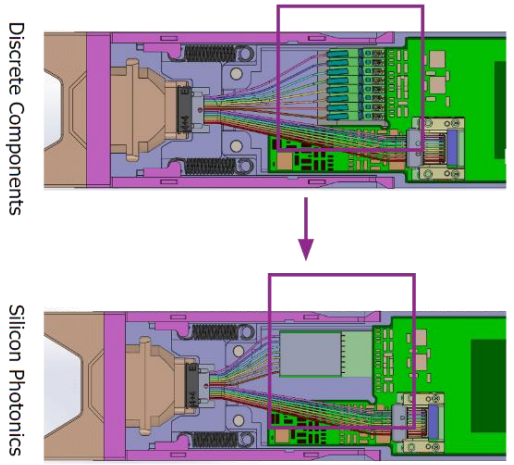
- **预计到2029年硅光模块销售额将达到103亿美元，2023-2029年CAGR达45%。**据Yole数据，硅光模块销售额有望从2023年的14亿美元快速增长至2029年的103亿美元。其中，到2029年数据中心可插拔硅光模块规模增长至53亿美元，占2029年硅光模块销售额的52%；用于电信波分复用领域硅光模块规模46亿美元，占比达45%。此外，据Yole数据，2024年硅光模块全球销量近600万只，预计2025年硅光模块销量超过1250万只，到2029年销量接近1800万只。

图：硅光模块结构示意图



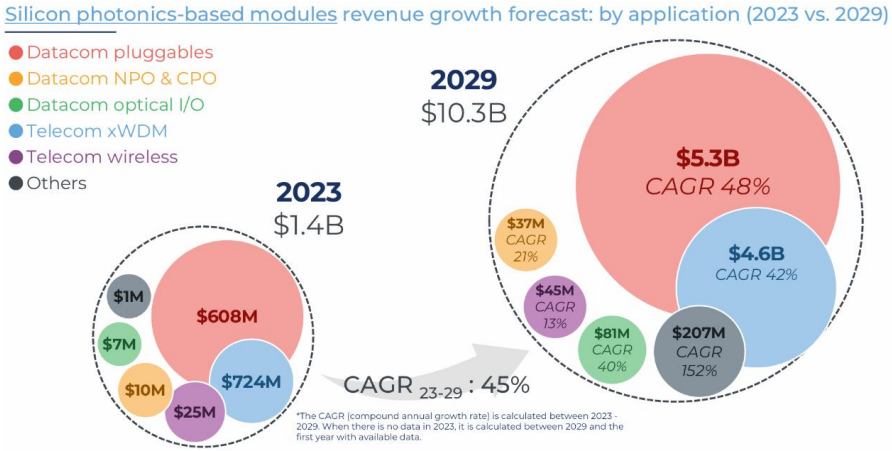
资料来源：Intel官网，国信证券经济研究所整理

图：传统光模块与硅光模块结构对比



资料来源：ITRI官网，国信证券经济研究所整理

图：硅光模块市场规模及预测

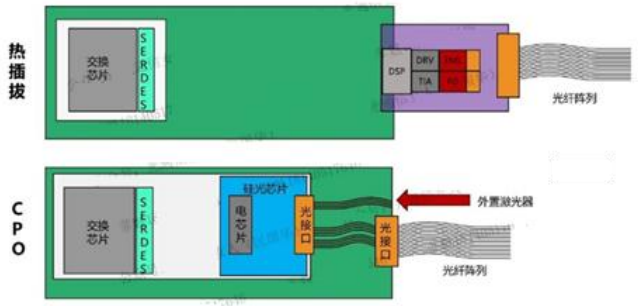


资料来源：Yole，国信证券经济研究所整理

4.1 光电共封装(CPO)：硅光模块的“进化形态”之一

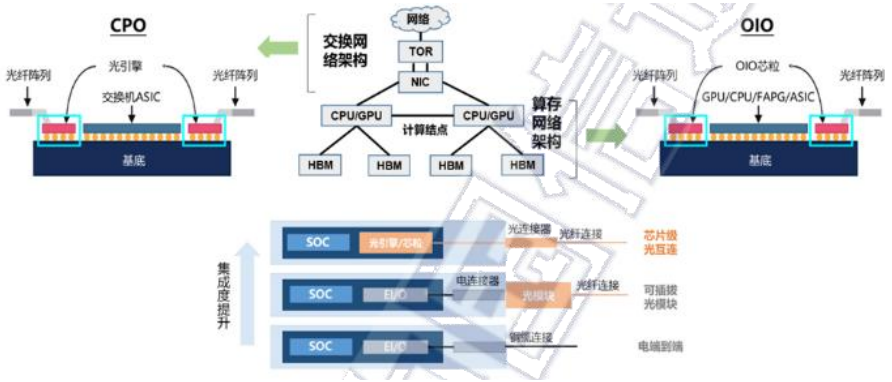
- **光电共封装（CPO）是一种在数据中心光互连领域应用的光电共封装方案。**当速率要求进一步提升（例如到3.2T甚至更高），传统的“可插拔”方式会产生太大的功耗和信号衰减，此时需要更进一步的架构变革。CPO不再把光模块做成独立的、可插拔的零件，而是把光引擎（硅光芯片）直接和交换芯片封装在同一个基板上。其优势在于消除了可插拔接口带来的损耗，功耗极低，适合AI集群内部极短距离的超高密度互联。
- **CPO技术目前主要用在交换机接口中。**CPO的核心是将光模块不断向交换芯片靠近，缩短芯片和模块之间的走线距离，并逐步替代可插拔光模块，最终把交换芯片（或XPU）ASIC和光/电引擎（光收发器）共同封装在同一基板上，最大程度地减少高速电通道损耗和阻抗不连续性，从而可以使用速度更快、功耗更低的片外I/O驱动器。
- **CPO应用在算力芯片上，向光OIO（输入输出）发展。**为了解决计算芯片CPU/GPU/XPU等之间的互联问题，OIO利用光互连低功耗、高带宽、低延迟的优势，取代传统的electrical IO方案，芯片输入输出的IO变为光信号，进而构建分布式计算网络。CPO不仅减少了高速电通损耗，其能效优势也非常显著。
- **CPO渗透率持续上升，端口出货量未来5年快速增长。**随着数据中心对高速互联需求的增长，CPO在满足大规模、高宽带连接方面市场潜力突出。据lightcounting数据，预计到2029年，1.6T CPO端口出货量显著增长，而3.2T CPO端口出货量预计将超过1000万个。

图：CPO技术与传统光模块对比



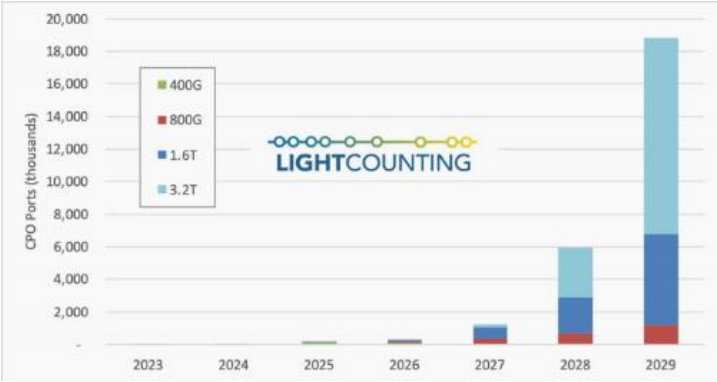
资料来源：菲魅光通信，国信证券经济研究所整理

图：CPO与OIO应用场景



资料来源：信通院，国信证券经济研究所整理

图：按速率划分的CPO端口出货量及预测



资料来源：LightCounting，国信证券经济研究所整理

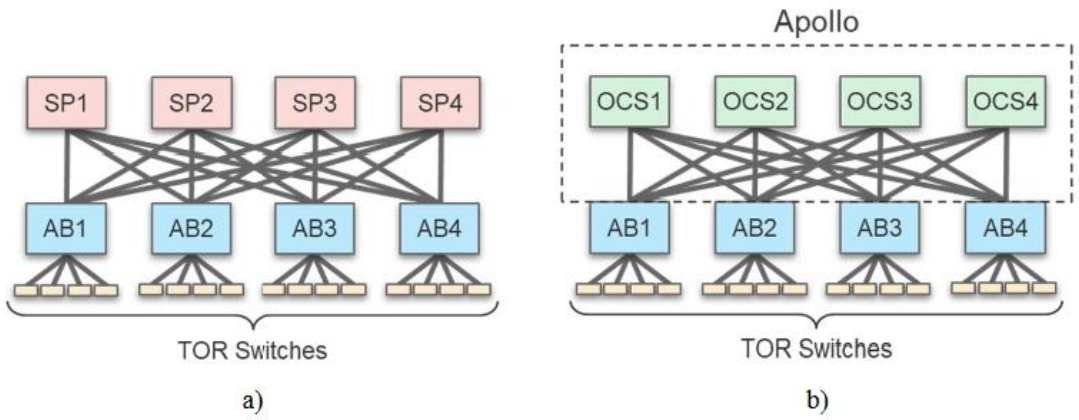
4.1 光电路交换(OCS)：网络架构层面的“去电化”思路

- **传统数据中心用“电交换机”来调度流量。**传统数据中心网络采用脊叶（Spine-leaf）结构，其中SP（Spine，脊）层主要是电网络交换机（Electronic Packet Switch，EPS）。SP层与每一个AB（Aggregation Block，汇聚）层相连，AB层与TOR（Top of Rack）交换机相连。由于传统架构中信号经过SP层进行多次电信号和光信号的转换，因此会产生较大的功耗，同时增加数据的延迟。

- **光电路交换（Optical Circuit Switch，OCS）**则是采用光交换机，其特点为利用微镜阵列（MEMS）或液晶等技术，直接在光域切换光路，不需要经过“光-电-光”的转换。OCS通常用于数据中心的核心骨干层（Spine层）或跨数据中心互联，处理大规模、稳定的数据流。在AI大模型快速发展的背景下，OCS交换机可以动态配置计算芯片间的连接关系，构建更好的大型算力网络对AI大模型的发展具有重大的意义。

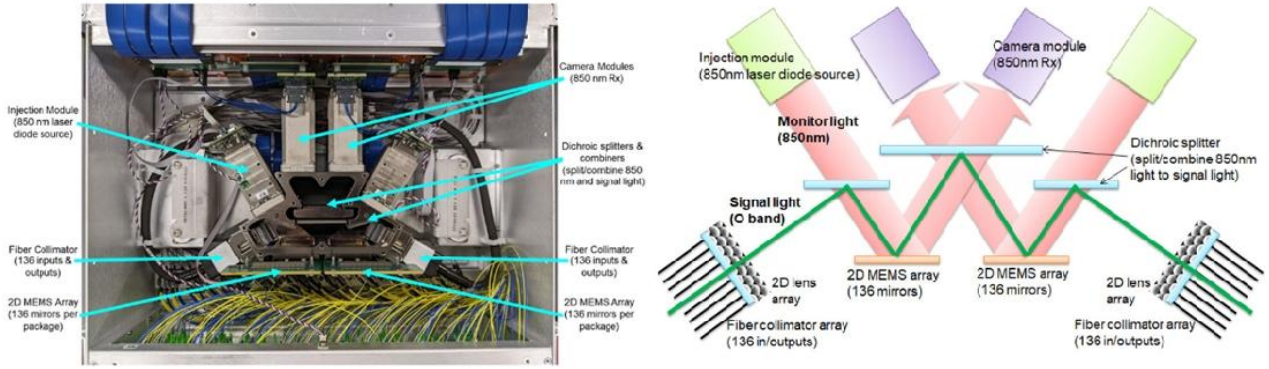
- 2023年3月，在OFC 2023（2023年美国光纤通讯博览会）上，谷歌详细介绍了其内部项目Apollo；该项目在其数据中心大范围部署OCS交换机，带来数据中心网络架构的重大变革。谷歌OCS交换机名为Palomar，输入输出端口是两个光纤准直器阵列（Fiber collimator array），包括光纤阵列和微透镜阵列，输入输出均为136个通道。当光通过光纤进入OCS交换机后，会先后经过两个2D MEMS阵列，每个阵列含有136个平面镜，用于精确调节光的传播方向。此外，系统中还包括两组监控通道，使用850nm波长的光，经过MEMS阵列反射后进入到监控相机处，通过图像处理来反馈控制MEMS阵列，从而优化链路插损。

图：传统数据中心网络与采用OCS交换机的Apollo项目结构示意图



资料来源：OFC 2023，国信证券经济研究所整理

图：谷歌Palomar OCS交换机实物图及原理示意图



资料来源：OFC 2023，国信证券经济研究所整理

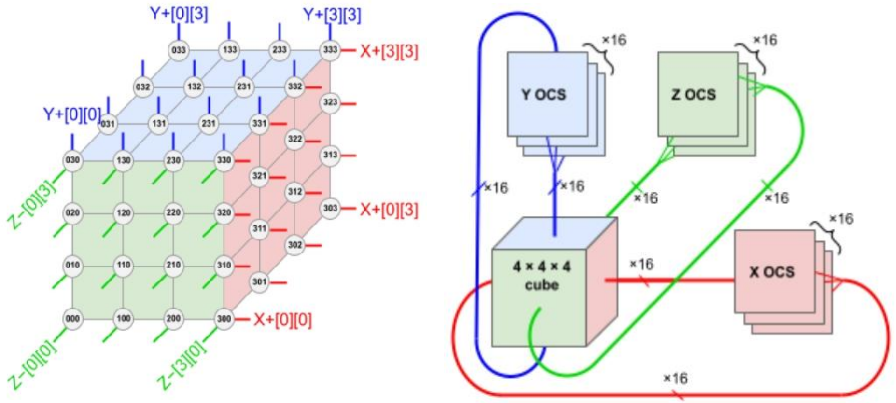
4.1 超算系统：谷歌的实践进一步强化了OCS交换机的应用

国信证券 GUOSEN SECURITIES

- 2023年4月，谷歌发布论文《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》（《TPU v4：具有嵌入式硬件支持的机器学习光学可重构超级计算机》），详细介绍了通过OCS交换机能够让超级计算机可以轻松地动态重新配置芯片之间的连接，有助于避免出现问题并实时调整以提高性能。

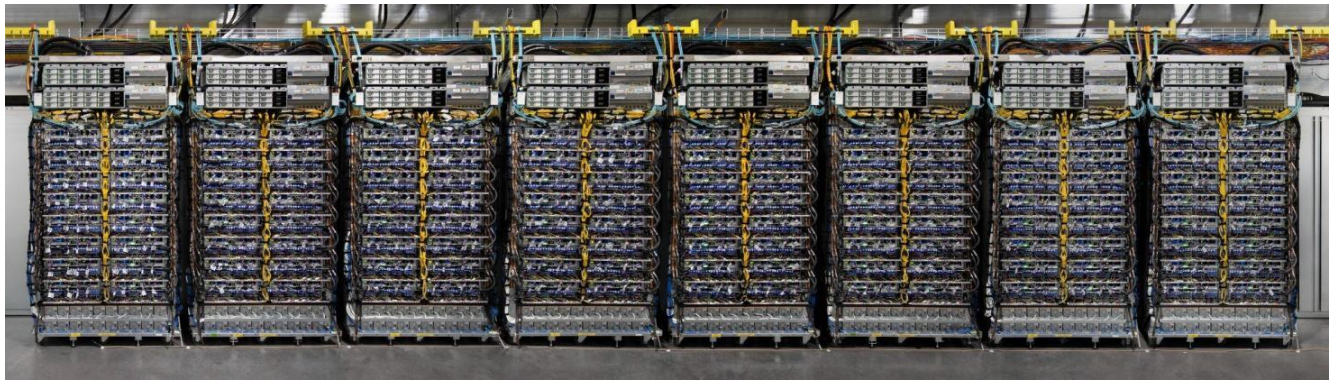
- TPU v4超级计算机的基本构造由4x4x4的TPU v4 Cube（立方体）组成。64个TPU芯片形成一个Cube，内部的TPU之间通过电缆链接，最外侧的6个面上的TPU与OCS交换机相连，每个面有16条链路，每个Cube共有96条光链路连接到OCS交换机上。为了提供三维环面的环绕链接，相对两侧的链接必须连接到相同的OCS交换机上，因此每个Cube连接到48个OCS交换机上。由于谷歌Palomar OCS交换机为136x136端口（128个端口加上8个用于链路测试和修复的备用端口），因此48个OCS交换机能够链接来自64个4x4x4 Cube的48对线缆，总共并联4096个TPU v4芯片，形成一个大型超算系统。

图：64个TPU芯片形成4x4x4 Cube及与OCS交换机的连接方式示意图



资料来源：TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings，国信证券经济研究所整理

图：由4096个TPU芯片组成的超级计算机



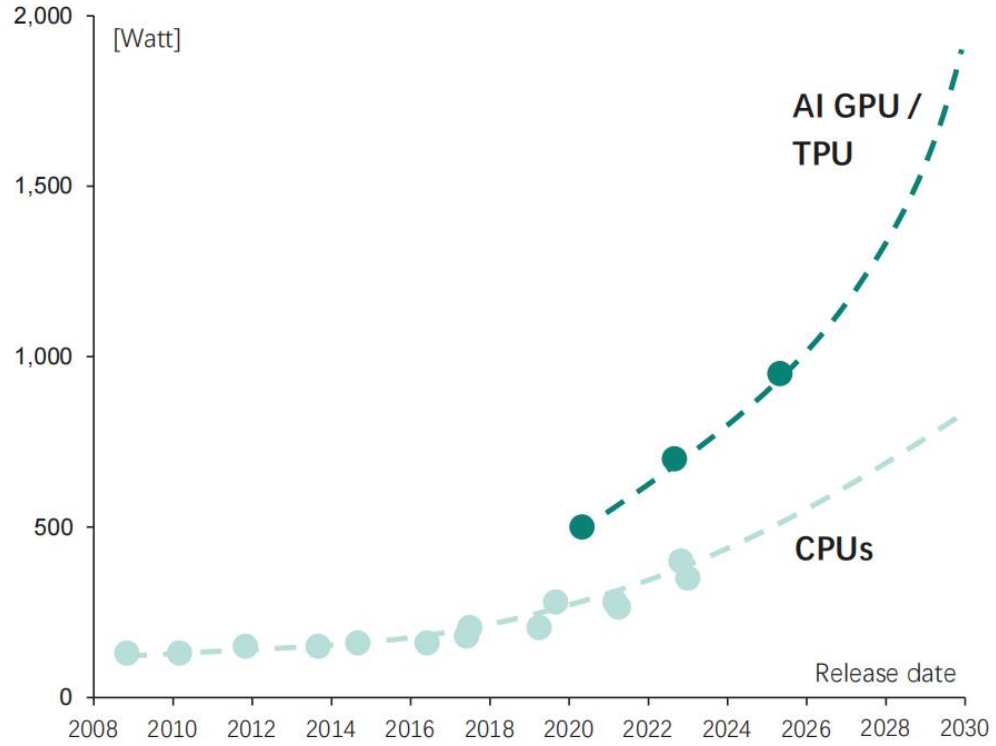
资料来源：TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings，国信证券经济研究所整理

4.2 AI算力增长推动数据中心用电量提升，电源架构同步升级

国信证券 GUOSEN SECURITIES

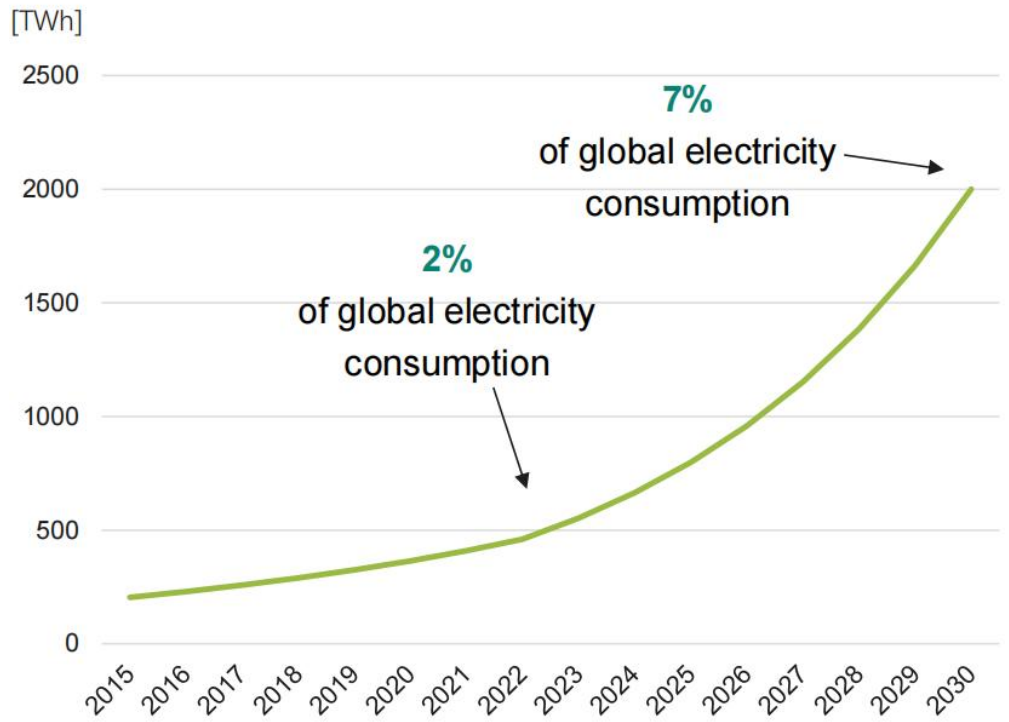
● **AI算力爆发推动数据中心用电激增，数据中心耗电量占全球电力消耗有望从22年的2%提升至7%。**随着算力应用渗透，人工智能数据中心芯片功耗的增大迫使机架处理功率越来越高，英飞凌预测单个GPU的功耗将呈指数级增长，到2030年将达到约2000W，而AI服务器机架的峰值功耗将达到300kW以上。相应地，全球数据中心耗电量占全球电力消耗有望从22年的2%提升至7%。随着数据中心能耗快速提升，数据中心电源架构随算力提升同步升级。

图：x86和基于Arm的服务器CPU与GPU和TPU的电力需求对比



资料来源：英飞凌，IEA，国信证券经济研究所整理

图：数据中心的电力消耗在全球电力需求的占比情况



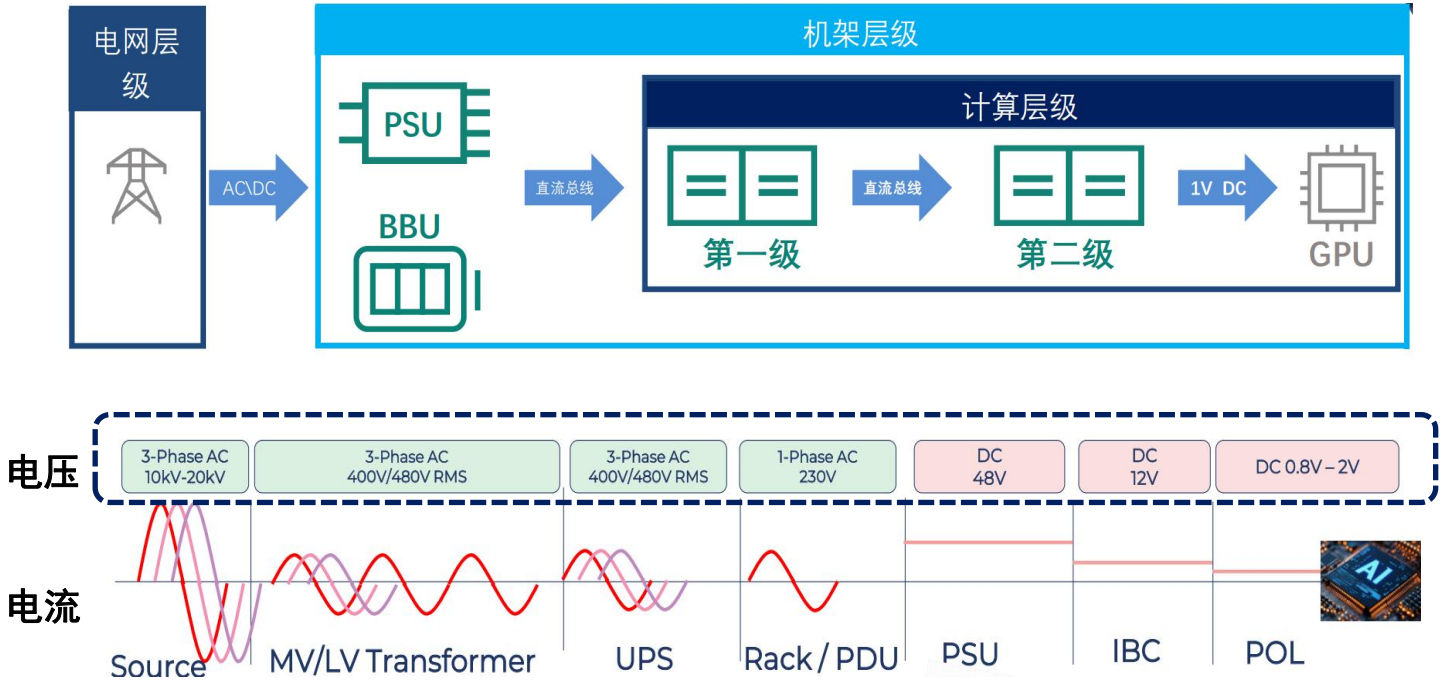
资料来源：英飞凌，IEA，国信证券经济研究所整理

4.2 AI算力增长推动数据中心用电量提升，电源架构同步升级

国信证券 GUOSEN SECURITIES

● 数据中心的供电系统沿电流方向依次从**电网层级**到**服务器机架层级**再到**计算层级**至计算芯片供电，**实现电流从交流到直流、电压从高压到低压的转换**；其对应电网的13.8 kV交流电经多个层级转换最后以1V左右的直流电对GPU进行供电。AI算力爆发推动数据中心用电激增，数据中心电力损耗、用铜量、机房区占比等均面临技术升级挑战，电源系统的能量转换效率要求大幅提升，传统供电系统很难满足需求，电源向小空间、高效率方向演进。根据测算，仅在芯片环节，50万个GPU（电费为0.10美元/千瓦时）对应年度电费总计1.19亿美元；50万个AI加速器对应年电费总计6833万美元；**电能转换效率的提升对降低数据中心的运营成本至关重要**。

图：数据中心电源架构



资料来源：英飞凌，Yole，国信证券经济研究所整理

图：GPU及AI服务器耗电水平

Power Reduction with proteanTecs	CPU	GPU	AI
Quantity	500,000	500,000	500,000
Power consumption (kWh)	0.15	0.35	0.2
Utilization rate in data center	60%	60%	60%
Power Usage Effectiveness (PUE)	1.3	1.3	1.3
Annual power costs per device	$102.5	$239.1	$136.7
Total annual cost savings with proteanTecs	**$5,637,060**	**$11,957,400**	**$8,199,360**
Performance increase due to device power reduction	Transaction per Second (TPS) increase 1,853,932,584	Frames per Second (FPS) increase 4,000,000	Inferences per Second (Inf/s) increase 12,272,727,273

资料来源：Semiconductor Engineering，国信证券经济研究所整理

4.2 AI算力增长推动数据中心用电量提升，电源架构同步升级

国信证券 GUOSEN SECURITIES

● **AI工作负载的功率需求持续攀升推动电源架构升级。**随着GPU互连数量增加，以NVIDIA Hopper架构到NVIDIA Blackwell架构为例，当扩展至72个GPU的系统时，机架的功率密度提升了3.4倍，使得单机架的功耗从数十千瓦攀升至超过100千瓦，未来甚至将达到1兆瓦。传统低压环境（例如54 VDC）所需电流庞大，**不仅会引发显著的电阻损耗，使用大量铜缆成本大幅提升**。此外，AI工作负载还带来波动性挑战。相比传统数据中心运行数千个互不相关的任务不同，在训练大语言模型时，数千个GPU会同时以近乎一致的节奏执行高强度计算，形成大幅且快速负载波动特征的功率曲线。**机架的功耗可能在几毫秒内从约30%的“空闲”状态迅速上升至100%**，随后又快速回落，对公共电网的稳定性构成挑战。

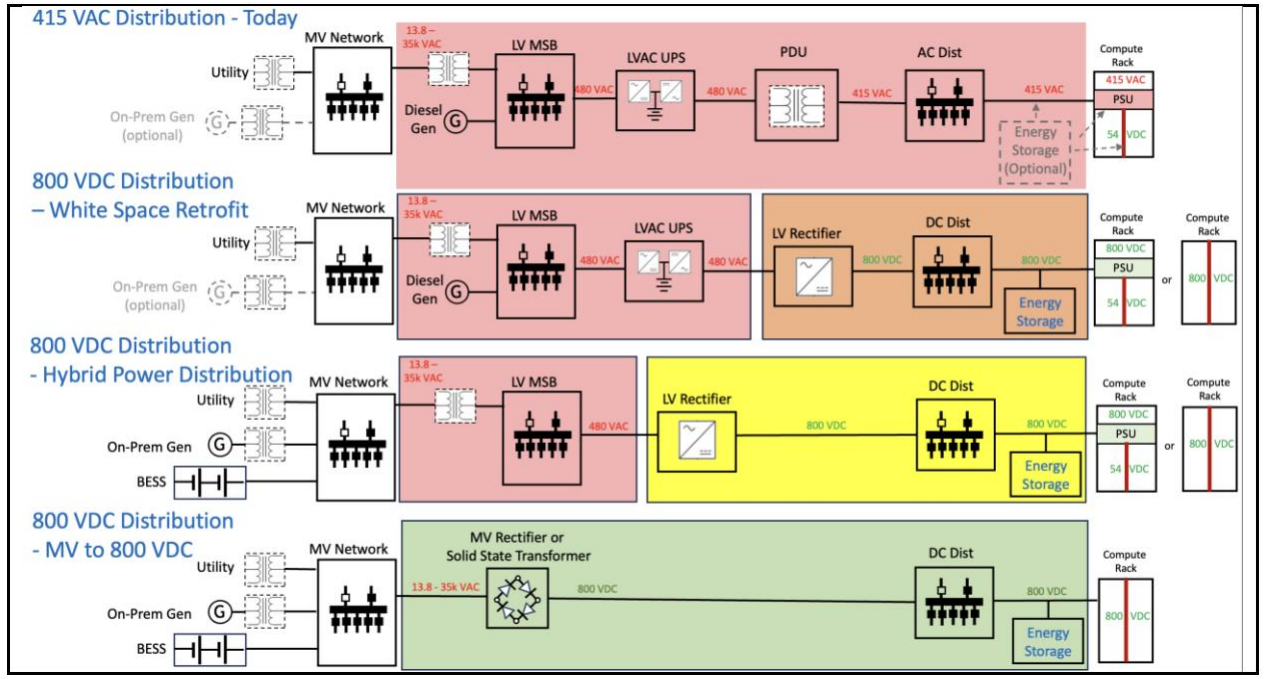
● **电源向高压（800伏直流（VDC）配电系统）方向发展。**从传统的415V或480V交流三相系统转向800V直流架构，相同线规可传输的功率比415 VAC高出15.7%，从而减少铜缆的使用并降低成本；减少了导体数量和连接器尺寸，降低了材料与安装成本；提高了整体能效。

➢ **现有架构**包含多个功率转换环节：电网的中压电能（例如35 kVAC）被降压至低压水平（例如415 VAC）随后经由交流不间断电源（UPS）调节后，通过配电单元（PDU）和母线槽系统传输至各个计算机柜。在每个机柜内部，多个电源单元（PSU）将415 VAC转换为54 VDC，并将直流电输送至计算托架，再通过板级DC-DC转换器完成最终的电压调节。

➢ **未来**在设施层面集中完成所有交流到直流的转换：中压交流电通过大型高容量电源转换系统直接转变为800 VDC，随后将该800 VDC配电至数据中心内的各个机架。这一设计通过去除交流开关设备、转换器和PDU层级，简化结构。

图：电源架构变化

从架构到器件：简化转换层级 提升能量转换效率



资料来源：英伟达，国信证券经济研究所整理

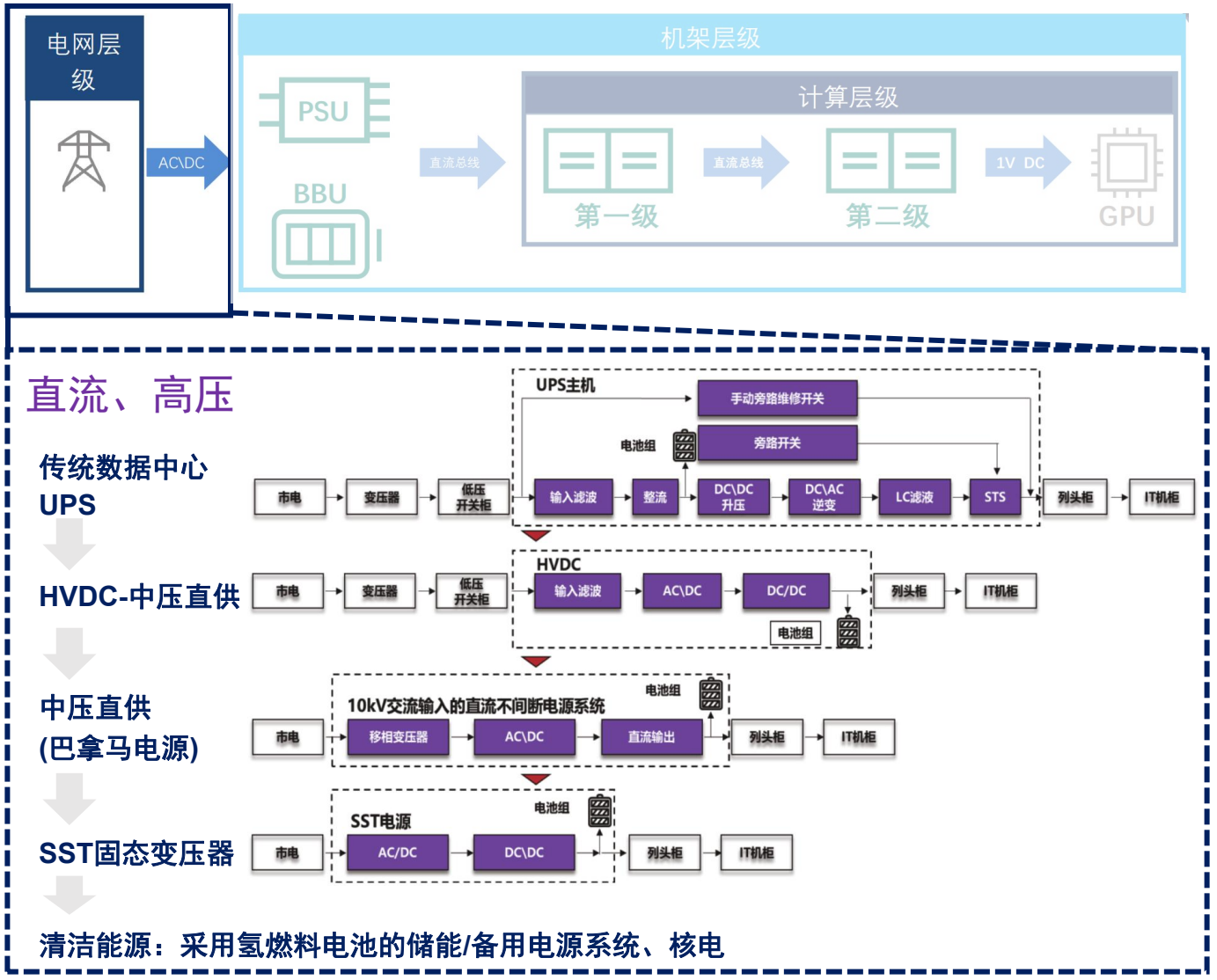
4.2 电网层级：电源向直流、高压方向发展以提升能效

国信证券 GUOSEN SECURITIES

- **电网层级主要完成10KV配电到10KV/380V的变电再到低压配电的环节，架构向直流、高压方向演进。**传统数据中心主要使用交流UPS（不间断电源）作为供电系统的核心设备；UPS系统包含从10kV变压器经过低压配电柜、UPS、精密列头柜以及PDU等配电设备，将能量层层传递至服务器电源；然而该供电架构节点多，占地面积较大且供电效率较低。

- **随着IT机柜功率密提升，数据中心供电电压进一步提升至800V或±DC400V，SST（固态变压器）应用出现。**SST系统作为10kV交流输入的直流不间断电源系统的进阶版取代传统的变压器设备进行调压和整流，并具有高功率因数、低电流谐波的输入特性。其系统链路更短，效率更高，体积更小，重量更轻，控制更方便，并具有很大的成本下降潜力。未来随着直流架构进一步成熟，清洁能源将引入直流架构作为供电来源之一。

图：GPU及AI服务器耗电水平



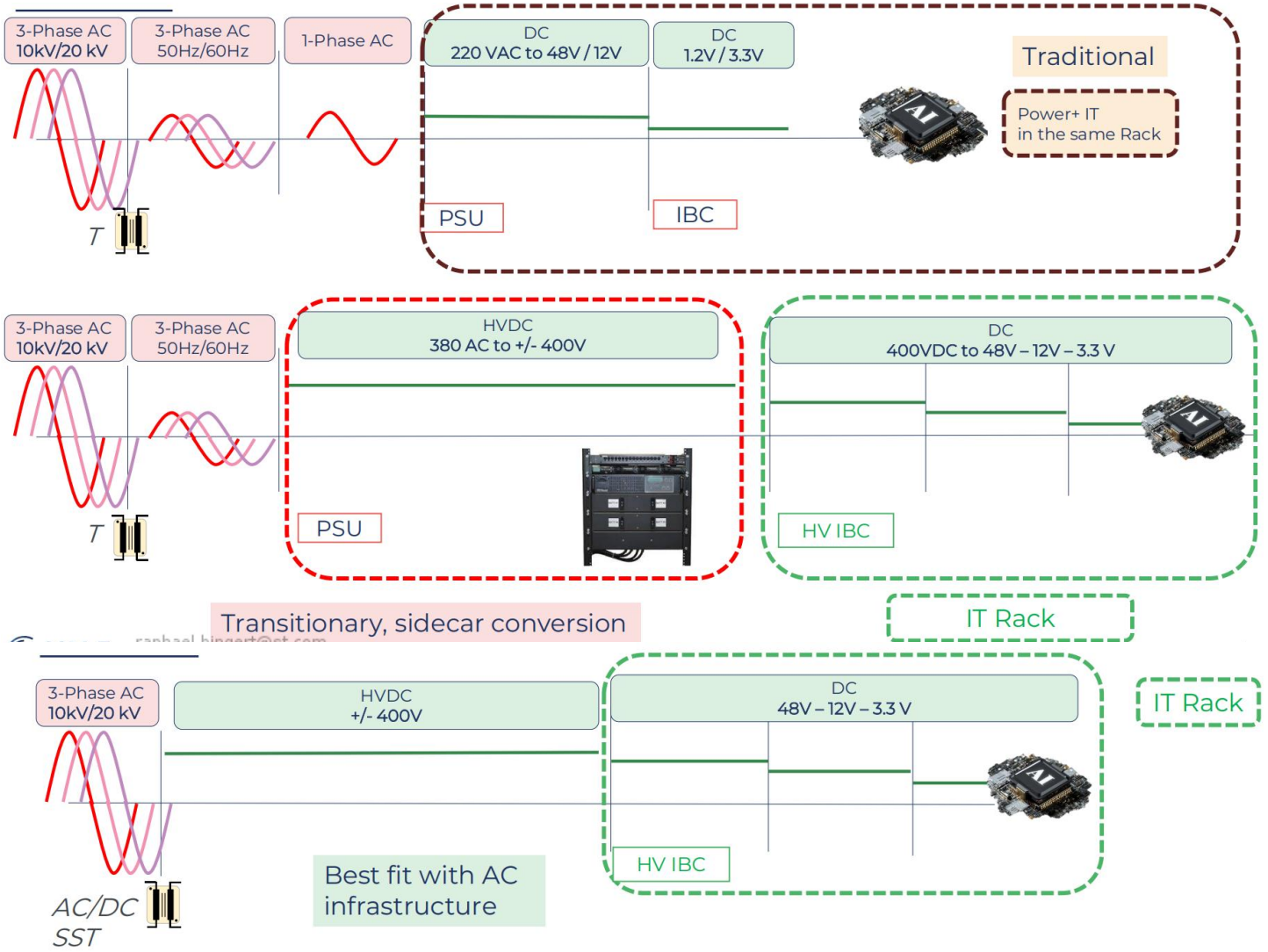
资料来源：英飞凌，数据中心800V直流供电技术白皮书，国信证券经济研究所整理

4.2 IT机柜层级：电源分级减少以提升能效

在IT机柜侧，电源结构向分级简化方向演进：

- 传统解决方案是直流电转换在每台服务器上进行，在服务器附近安装多个电源单元（PSU），导致损耗（约占服务器总功耗的5%），功率上限约为250 kW。
- SideCar过渡方案将整流阶段分离，以隔离损耗并释放机架内更多的IT空间，基于三相系统每个机架的供电能力扩展至600kW，但每个IT机架仍然需进行AC/DC转换。在中期将转换阶段移至数据中心前端。通过将电压升至800V（将配电损耗降低四倍，配合SST）。
- 终极方案使用中压直流（MVDC）输入代替交流电，即直流微电网。

图：GPU及AI服务器耗电水平

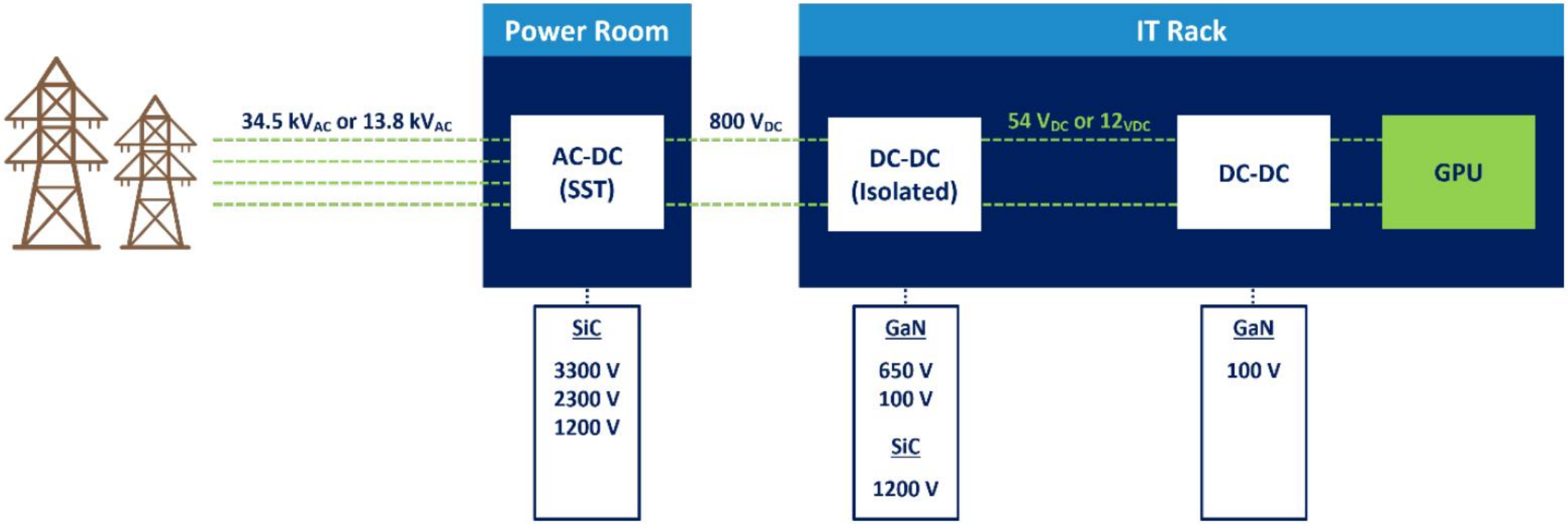


资料来源：Yole，国信证券经济研究所整理

4.2 碳化硅和氮化镓将成为AI电源的主流功率器件

- 随着电源功率密度大幅提升，硅基器件性能已接近极限，在下一代高压电源架构下，碳化硅和氮化镓将成为主要的功率器件以完成电能的高效转换。由于宽禁带半导体（例如GaN和SiC）能够承受更高的电场，因此它们可以承受更高的电压并提供更低的单位面积电阻，在实现更高的功率密度和效率的同时还可以在更高的开关频率下工作。最终，通过碳化硅与氮化镓的应用，电源转换效率提升，可使用更小的外部元件，电源体积以及系统成本均可下降。

图：碳化硅和氮化镓将成为数据中心的核心功率器件

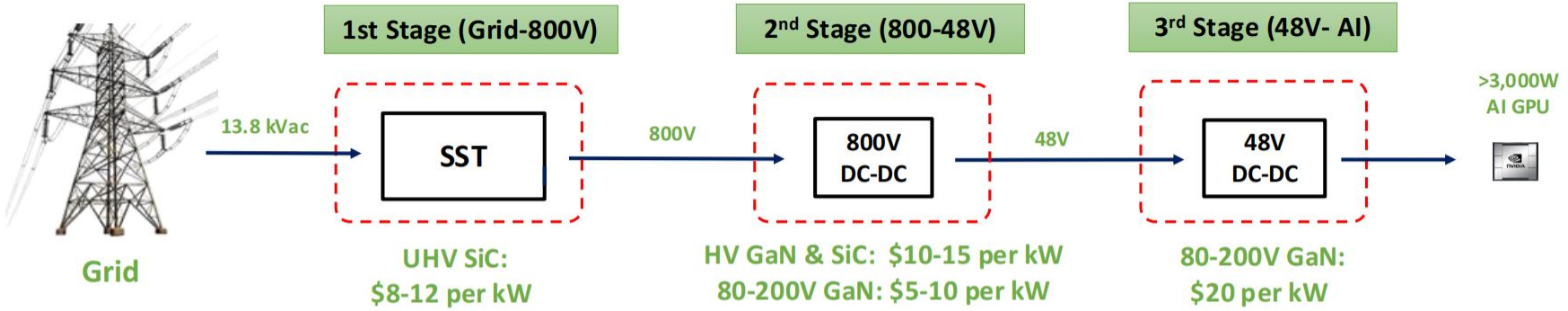


资料来源：Yole，国信证券经济研究所整理

4.2 碳化硅和氮化镓将成为AI电源的主流功率器件

- **碳化硅与氮化镓在800V数据中心市场有望2030年达26亿美元。**碳化硅主要用于电网到800V、800V到48V阶段；为固态变压器以及直流转换环节的核心器件，通过碳化硅的应用单机功率密度加速提升，预计SST中碳化硅单位价值量对应8-12美元/KW；氮化镓主要用于直流降压和三次电源部分，由于高频特性好，可以进一步缩小整机的体积，预计三次侧电源氮化镓单位价值量对应20美元/kW；根据Navitas数据，假设800V数据中心渗透率2030年提升至80%，对应碳化硅和氮化镓的市场有望达26亿美元。

图：碳化硅和氮化镓将成为数据中心的核心功率器件



800V Data Centers	2025	2026	2027	2028	2029	2030
GPU Shipments (Mu, total AI)	9	11	12	13	16	20
Power per GPU (W)	2000	2500	3000	3150	3300	3500
Total GPW Power (GW)	18	28	35	41	52	71
800V Adoption (as % of total AI)	0%	2%	9%	23%	50%	80%
Power Semi $ per kW (GPU power)	$0	$55	$52	$50	$47	$45
Power Semi TAM ($M)	**$0**	**$30**	**$164**	**$458**	**$1,229**	**$2,564**

GaN + SiC can be adopted in significant % of this $2.6B/yr 800V Data Center opportunity

资料来源：Navitas，国信证券经济研究所整理

【5】AI端侧：AI Agent重塑交互范式，大厂争先布局端侧入口，消费电子创新大年开启

5.1 存储涨价强预期下，高、低端手机预计出现分化走势

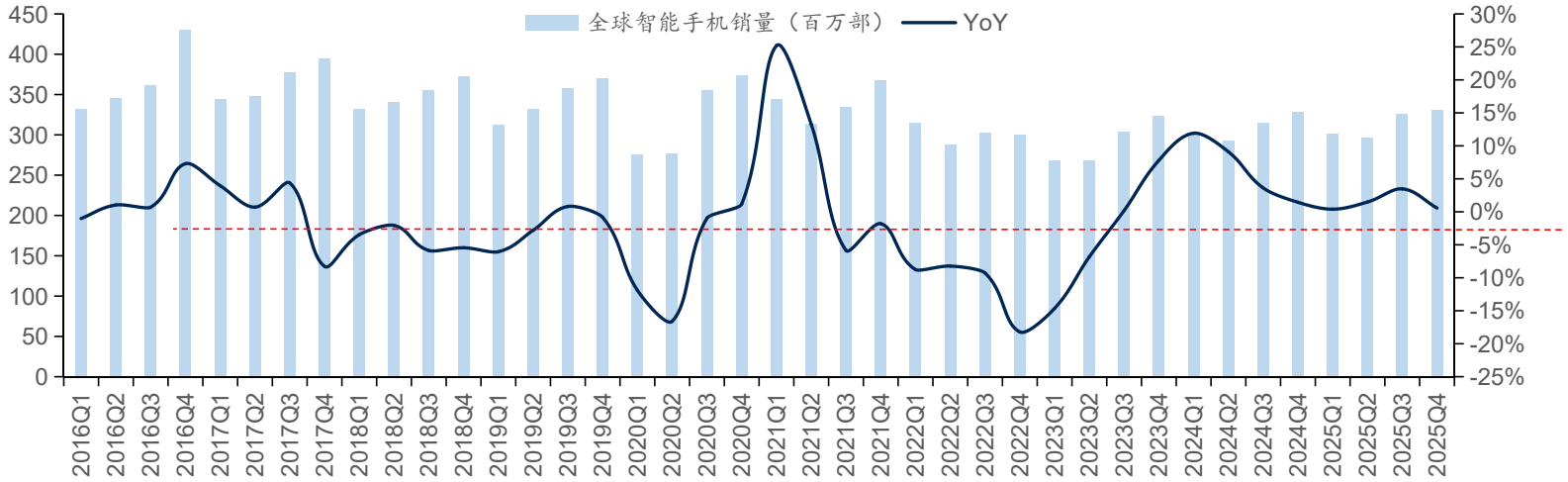
- **3Q23-3Q25，全球手机季度销量同比增速为正，在长达7个季度的同比下滑后迎来了连续5个季度的同比增长。**据IDC数据，全球智能手机出货从2007年的1.25亿部快速增至2016年高点的14.69亿部，随后连续4年同比下滑，2021年因疫情宅家经济推动出货量同比回升6%至13.6亿部，此后下滑至2023年11.64亿部。

- 根据Counterpoint Research发布的《智能手机市场展望追踪》，预计2025年全球智能手机出货量将同比增长3.3%。苹果的智能手机出货量在2025年整体表现强劲，全球出货份额将达19.4%，使其自2011年以来首次成为全球第一大智能手机品牌。三星出货量预计也将同比增长4.6%，全球份额达18.7%，但仍将让出其十余年来的榜首位置。

- 展望2026年，由于存储涨价，以及国补政策退出，转为地方接力，我们预计传统智能机中，高端手机在成本增长比例和客户价格敏感度两个方面更占优势，销量维持稳定。而低端手机受影响更大，可能出现销量下滑。

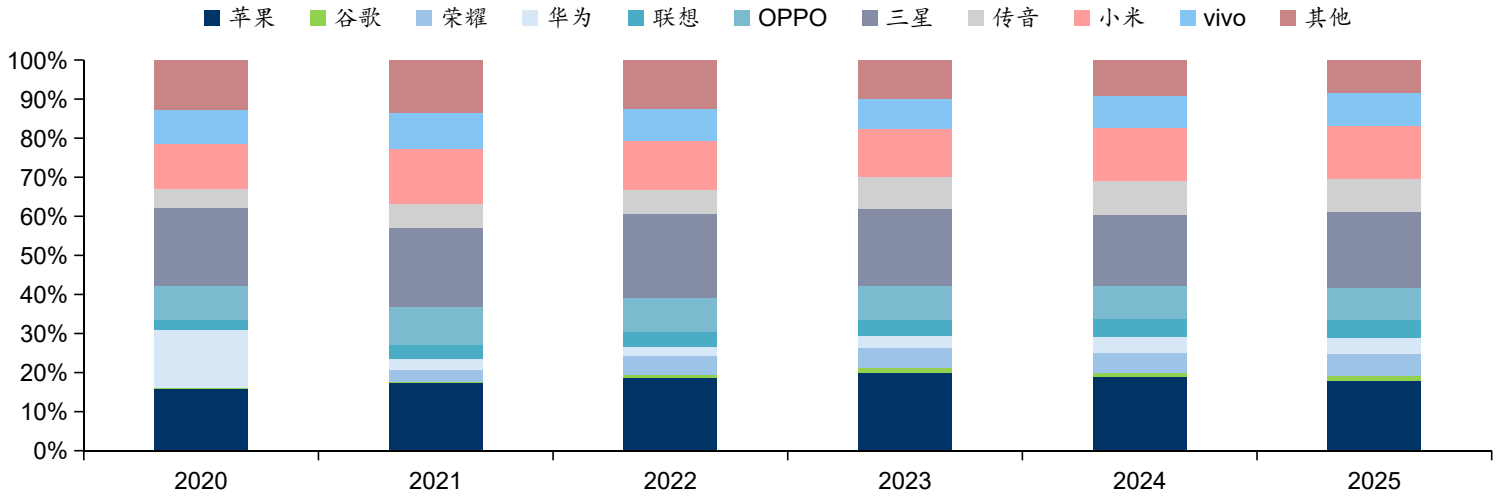
- 但随着iPhone折叠机型发布，以及AI Agent逐步落地，带来手机交互颠覆式创新，有望加速全球换机周期，带来手机领域的新一轮景气上行。

图：全球智能手机季度出货量（百万部）



资料来源：IDC，国信证券经济研究所整理

图：中国智能手机月度出货量



资料来源：IDC，国信证券经济研究所整理

5.1 折叠屏是旗舰机市场的重要增量

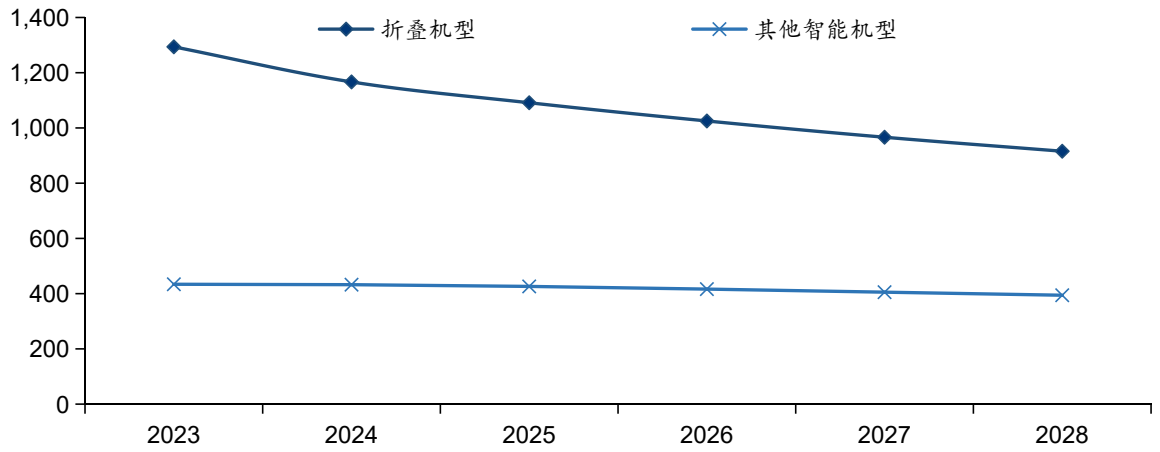
- 折叠机高昂的成本使得其定位局限于高端机型，目前头部手机厂商除了苹果，几乎按每年更新的频率推出横折叠、竖折叠两款机型。根据IDC统计预测，2025年全球智能机出货量预计12.3亿部，其中折叠屏手机出货量有望达到3千万部，占比2.6%。预计2028年全球折叠屏手机出货量将达到5千万部（23-28年CAGR为20%），占全球智能机销量的3.5%。
- 随着折叠屏关键技术规模效应提升，成本逐步下探，2024年全球折叠机型平均售价在1167美元，预计到2028年价格将下探至916美金。

图：头部品牌折叠手机迭代进程

品牌	机型	2020	2021	2022	2023	2024
华为	大折叠	Mate Xs	Mate X2	Mate Xs2	Mate X3/5	
	小折叠		P50 Pocket			Pocket2
	三折叠					Mate XT
三星	小折叠	Galaxy Z Flp	Z Flip 2/3	Z Flip4	Z Flip5	Z Flip6
	大折叠	Galaxy Z Fold2	Z Fold3	Z Fold4	Z Fold5	Z Fold6
OPPO	小折叠			Find N2 Flip	Find N3 Flip	
	大折叠		Find N	Find N2	Find N3	
VIVO	小折叠				X Flip	
	大折叠			X Fold	X Fold2	
荣耀	小折叠					Magic V Flip
	大折叠			Magic V/Vs	Magic V2 /Vs2/V Purse	Magic V3/Vs3
小米	小折叠					MIX Flip
	大折叠		MIX Fold	MIX Fold2	MIX Fold3	MIX Fold4

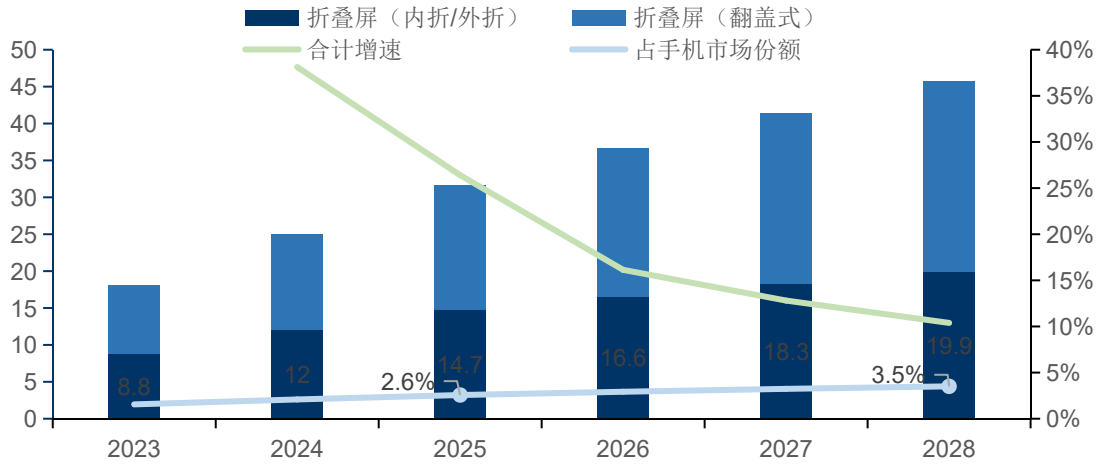
数据来源：各公司官网，国信证券经济研究所整理

图：全球折叠屏手机平均出货单价



资料来源：IDC，国信证券经济研究所整理

图：全球折叠屏手机出货量及预测（按年度）



数据来源：IDC，国信证券经济研究所整理

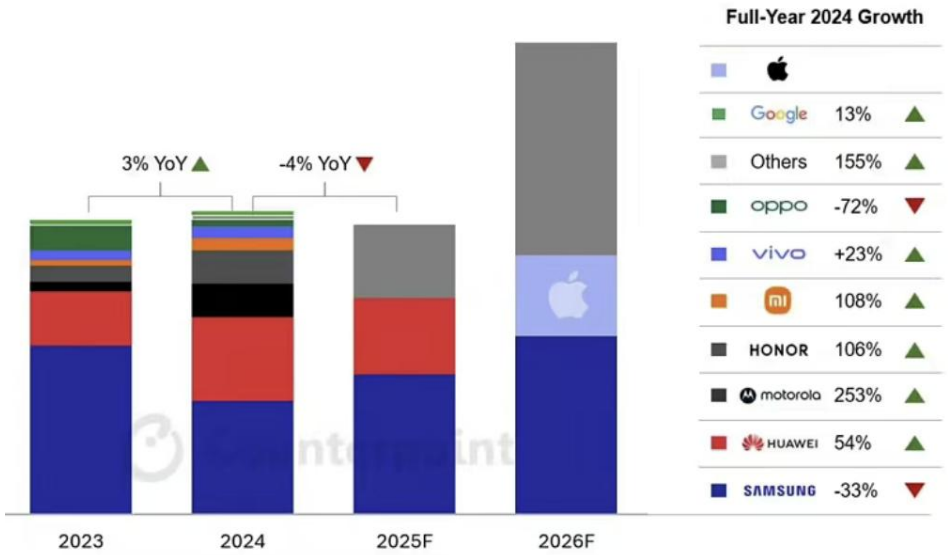
5.1 折叠iPhone有望于2026年发布

- 根据苹果供应链透露，2025年6月已开始折叠iPhone的P1，预计2025年底有机会走完Prototype的开发流程，再进入EVT(Engineering Verification Test)，按照此流程，2H26年折叠iPhone有望上市。原本苹果规划新产品当中，除了iPhone，亦包括较大尺寸的折iPad。如今确定，将率先推出折叠iPhone，折叠iPad暂时不发，推测暂缓原因包括面板等零组件生产更困难，且售价太高，市场接受度较低等。
- 苹果iPhone通常会在前一年展开P1-P3阶段，来年开始进入EVT、DVT及MP，以赶上秋季发布会，比如预计2025年秋季要发表的iPhone17，已经在第2季初完成EVT。在P1到P3阶段，将由供应链开始小量试产，再交由iPhone主力组装厂鸿海、和硕等进行组装，中检视生产与产品良率，若无问题，就进入下个阶段。上述P1到P3，每一阶段约2个月。苹果开发新机时间远超过同业，且即使已经开模，只要产品开发或市场状况不如预期，也存在临时暂停，过去在多款产品都有先例。
- 三星2024年折式手机出货量为800万台，在2011年曾达到1100万~1200万台。因此，我们预计苹果折叠iPhone有望在首年达到百万级销量，全生命周期有望达到千万级销量。

图：苹果折叠iPhone效果渲染图

数据来源：AppleInsider，国信证券经济研究所整理

图：苹果入局可能颠覆折叠机市场



数据来源：Counterpoint，国信证券经济研究所整理

5.1 潜望式摄像头：大幅提升智能手机长焦拍摄能力

国信证券 GUOSEN SECURITIES

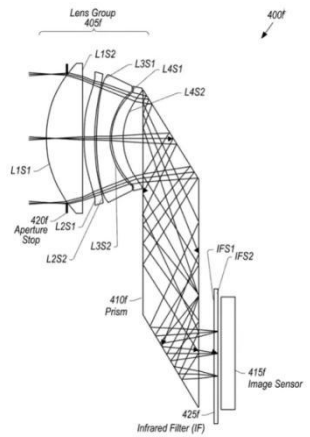
- 在硬件层面，摄像头及对应的影像升级仍是智能手机的创新核心。随着下游消费者对于手机影像功能需求的日益多样化，智能手机厂商和摄像头供应商努力推进行业变革，摄像头规格持续升级。除了智能手机单机搭载摄像头数量持续增长，超高像素、大光圈、多透镜设计、光学防抖、光学变焦、潜望式镜头模组等成为智能手机摄像头升级的演进方向。

- 潜望式镜头又称“内变焦”镜头，其光学变焦在机身内部完成。一类潜望式镜头采用折叠光变方案，通过增加棱镜折射的方式，实现光路90°转变，使得镜头横向放置以获得更大的变焦倍率而无需考虑因为镜头高度限制无法获得更大空间，能够在不增加手机厚度的同时，实现高倍数的光学变焦，大幅度提高手机拍摄性能，从而达到手机拍照更长远、更清晰的效果。

- 2023年，苹果在iPhone15 Pro Max上首次搭载潜望式摄像头，采用四重反射棱镜方案。根据苹果官网和专利显示，有别于传统的单次反射转折90°方案，苹果的潜望式镜头采用四次反射棱镜架构，通过一颗平行四边形的光学棱镜实现，这相当于搭载等效120mm f/2.8、5倍变焦的望远镜头。当光线在经过望远镜组之后，可以在棱镜内部反射四次，再到达感光元件。四重反射棱镜架构解决了潜望长焦模组的空间占用问题。与安卓主流的三角棱镜潜望式镜头设计相比，采用四重反射棱镜的潜望式镜头光学元件简单，不需要复杂的望远镜组设计，理论上有更好的光学表现。在降低成本的同时，光圈也容易设计得较大。由于棱镜中间没有其他光学元件，因此较好地解决了潜望长焦模组的空间占用问题。

- 苹果搭载潜望式摄像头的机型销量占比持续攀升，带动四重反射棱镜产品需求持续增加。据IDC数据，iPhone15 Pro Max发布后的四个季度销量超过4700万部；在4Q23-2Q24期间，iPhone15 Pro Max单款机型占到当季度苹果手机销量约1/4。2024年9月，iPhone16系列手机发布，新增iPhone16 Pro及iPhone16 Pro Max两款高端机型搭载潜望式摄像头。据IDC数据，iPhone16系列发布后的四个季度中，搭载潜望式摄像头的机型销量超过1.1亿部；在4Q24-2Q25期间，搭载潜望式摄像头的机型占到当季度苹果手机销量稳居45%以上。

图：四重反射棱镜设计原理图

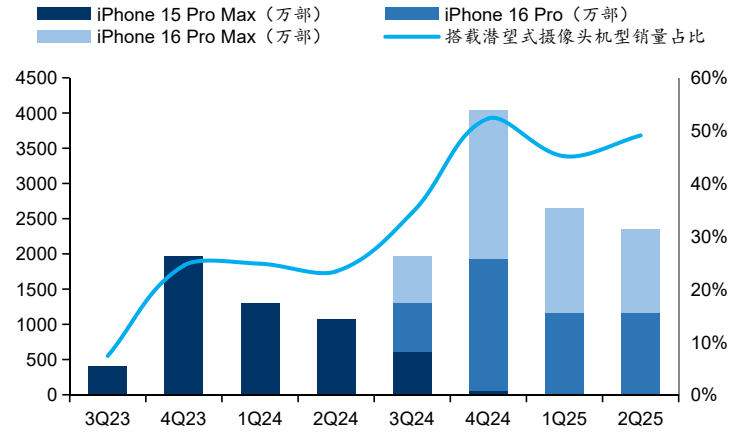


资料来源：苹果官网，国信证券经济研究所整理

图：采用四重反射棱镜的潜望式摄像头光路示意图

资料来源：苹果官网，国信证券经济研究所整理

图：苹果搭载潜望式摄像头机型销量及占比



资料来源：IDC，国信证券经济研究所整理

5.1 玻塑混合与可变光圈：智能手机光学镜头的升级方向

国信证券 GUOSEN SECURITIES

- **玻塑混合镜头融合了玻璃镜头与塑料镜头的优势，正成为智能手机镜头升级的重要新方向。**塑料镜头因体积小、成本低、易于大规模量产，广泛应用于对空间和成本敏感的移动影像系统；而玻璃镜头则凭借优异的光学性能和环境稳定性，多用于对成像质量要求较高的场景。玻塑混合方案巧妙结合两者特点，在实现小型化与轻量化的同时，兼顾高光学性能与可量产性，应用前景更为广阔，但制造工艺也更为复杂。得益于玻璃镜片的高折射率和热稳定性，此类镜头可支持更大光圈、更高解析力、更薄模组厚度以及更低的温度漂移，有效降低暗光环境下的图像噪点，改善画面边缘与近焦区域的成像质量，并扩大有效视场角。目前，该技术已在高端旗舰机型中加速落地，例如小米17系列主摄采用1G+6P（1片玻璃+6片塑料）玻塑混合镜组，凭借超低反射镀膜技术，显著抑制镜片间内部反射，大幅减少鬼影与眩光，从而呈现更纯净、通透的画面表现。

- **光圈是指相机镜头内部由多片可变叶片组成的组件，通过调节其开合大小来控制进光量，从而影响画面曝光和景深。**简单来说，光圈越大，背景虚化效果越明显，适合人像特写等突出主体的场景；光圈越小，画面整体越清晰，更适合拍摄合影、建筑等需要前后景都锐利的题材。传统手机摄像头受限于空间和成本，通常采用固定最大光圈设计，无法物理缩小光圈，背景虚化效果主要依赖算法模拟。**近年来，部分厂商开始引入物理可变光圈技术：**小米在13 Ultra和14 Ultra上搭载了该功能，三星曾在Galaxy S9、S10系列中率先尝试，华为则在最新的Mate和Pura系列旗舰机型中持续应用并优化这一技术。**配备物理可变光圈后，用户既能通过大光圈实现更自然、光学级的背景虚化，也能通过缩小光圈提升整体清晰度，从而显著降低对计算摄影算法的依赖。**2025年10月15日消息，据韩媒etnews报道，苹果计划在明年秋季发布的iPhone18Pro系列手机中首次引入可变光圈镜头，可根据拍摄环境灵活调整真实景深，实现更多样化的照片风格。

图：小米17采用1G+6P玻塑混合镜头

资料来源：小米官网，国信证券经济研究所整理

图：可变光圈结构示意图

资料来源：IT之家，国信证券经济研究所整理

5.2 端侧景气向上，关注AI生态价值

国信证券 GUOSEN SECURITIES

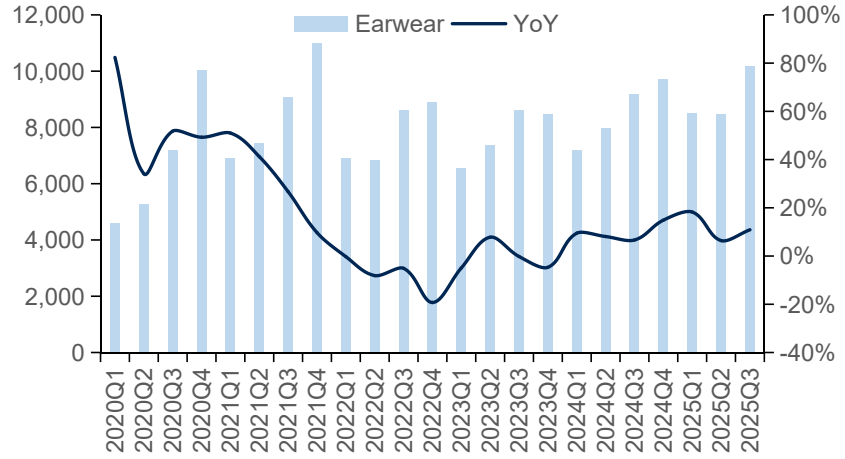
- **2025年全球耳机市场延续温和回升趋势。**根据IDC数据，2024年全球TWS耳机出货量3.4亿部(YoY +9.8%)。3Q25全球耳机出货1亿部，其中TWS耳机出货量8295万部(YoY +11%，QoQ +20%)，占比81%，与近3年比例持平。分品牌来看，苹果前三季度销量占比22%，略低于2024年全年的25%，但4Q25苹果发布Airpods Pro3，功能升级明确，销量超预期，带动供应链加单。

- **智能眼镜继续维持高增速，3Q25全球出货299万副(YoY +187.5%)，全年出货近千万部。**2025年，夸克S1、小米眼镜、Meta Rayban Display等重要产品密集落地，智能眼镜技术路线逐渐清晰。我们判断，智能眼镜是AI Agent的理想入口，未来两年有望延续高速渗透，成为消费电子周期中最具成长性的赛道。

- **智能手表温和复苏，**1-3Q25全球出货1.2亿部(YoY +7.3%)。目前，智能手表具备姿态识别、心率、血氧、睡眠等功能已日益完善。短期看，无创血糖监测仍是关键技术瓶颈，该功能若实现商用，将打开健康监测的第二增长曲线。

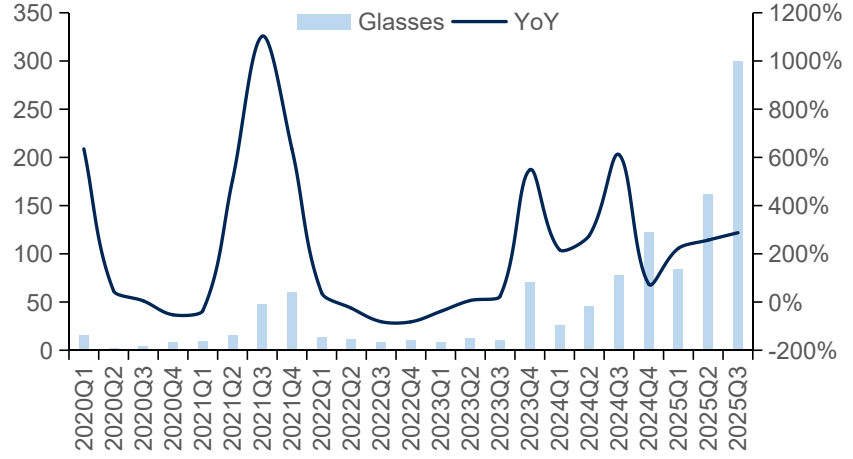
- **智能戒指崛起为新兴细分品类。**1-3Q25出货235万部(YoY +104%)。在无屏、轻佩戴等特性驱动下，正成为低存在感健康监测与AI交互的补充载体。随着智能眼镜渗透提升，戒指可通过微手势/指向识别成为关键输入设备，有望与眼镜形成复合终端组合。

图：耳机全球销量（万副）



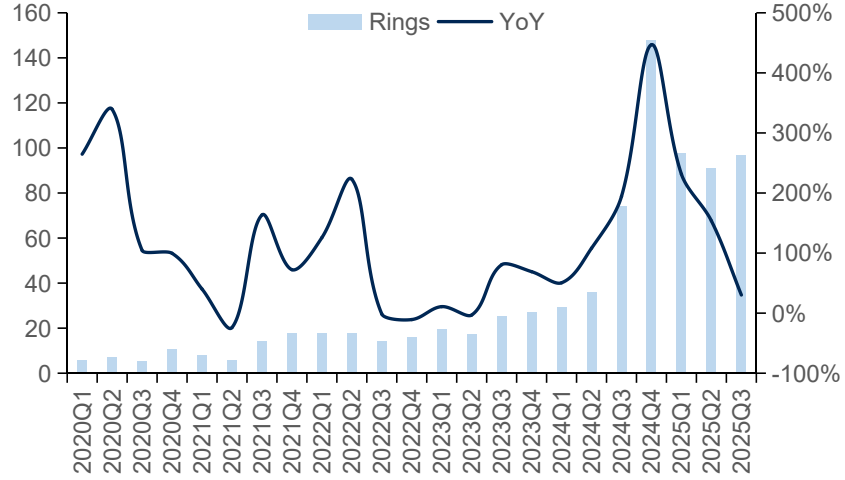
资料来源：IDC，国信证券经济研究所整理

图：智能眼镜全球销量（万副）



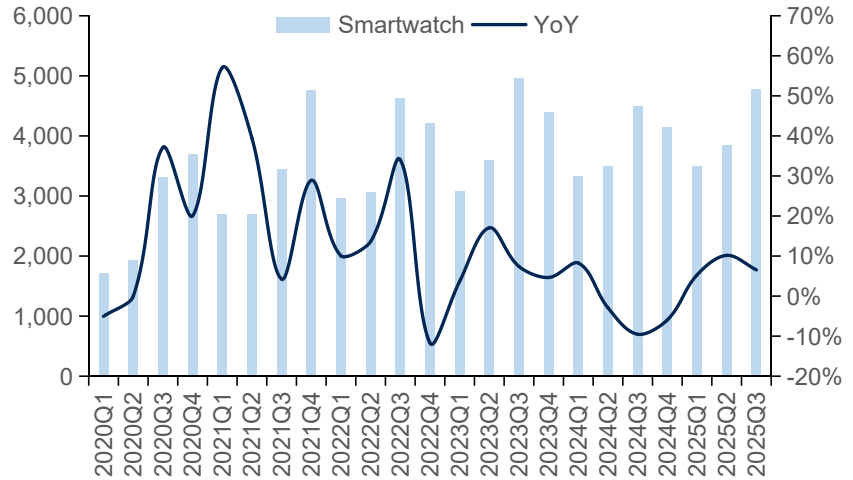
资料来源：IDC，国信证券经济研究所整理

图：智能戒指全球销量（万副）



资料来源：IDC，国信证券经济研究所整理

图：智能手表全球销量（万副）



资料来源：IDC，国信证券经济研究所整理

5.2 Apple Intelligence构建AI手机的终局蓝图

- **2024年WWDC上，苹果首次系统化阐述了Apple Intelligence的战略蓝图。**Apple Intelligence是个人智能系统，它将强大的生成模型放在iPhone、iPad和Mac的底层系统中，内置的大型语言模型具有深度自然语言理解能力，能够理解和创建文本和图像，能够代表用户执行任务，开发者工具也进行了全面提升。苹果构建了一个非常理想的AI赋能手机的终局，但实际落地来看难度超过此前预期。

- 6月11日，苹果公司软件工程高级副总裁Craig Federighi和全球营销高级副总裁Greg Joswiak在《华尔街日报》谈及其人工智能战略时，明确表示，开发传统的AI聊天机器人并不是他们的目标。相反，苹果的AI战略聚焦于系统集成，以提升用户的日常操作体验，创造一种无缝的智能交互环境。苹果布局AI非常早，Face ID解锁、相册人物识别、Apple Pencil防误触等功能，背后都是AI模型运行的结果。这延续了苹果一贯的"隐形AI"哲学，AI是工具，它必须是无感的、高效的，并且绝对忠诚于用户隐私，核心目的仍是提升用户的使用体验。 苹果追求的AI是一种环境，而不仅仅是一种功能，在终端算力和隐私保护双重限制下，技术演进速度虽暂未达市场预期，却为消费电子领域的AI集成树立了颇具前瞻性的范式标准。

- **Apple Intelligence国行版上市预期模糊。**据macrumors报道，苹果公司25年11月在官网上线“Apple Intelligence 反馈表单”，要求填写+86手机号，首次明确将中国大陆用户纳入内测范围。收集网页不久就又关闭了。目前，Apple Intelligence 所有模块在中国大陆仍呈灰色不可点状态。按照国内监管要求，苹果必须引入本土 AI 提供商完成内容审核与数据托管。苹果官方未透露确切上线日期，上线预期一再推迟。

图：Apple Intelligence功能概览

数据来源：Apple，国信证券经济研究所整理

图：Siri可以外接ChatGPT进行问答

数据来源：Apple，国信证券经济研究所整理

图：苹果AI的五大原则

数据来源：Apple，国信证券经济研究所整理

5.2 跨应用调度冲击现有商业格局，长期演进路径充满博弈

国信证券 GUOSEN SECURITIES

- **豆包手机助手的推出首次真实验证了系统级AI Agent形态，但商业与监管阻力将导致短期落地受限、长期演进路径充满博弈。** 2025年12月，字节跳动联合中兴发布豆包手机助手预览版，通过与手机系统深度集成，让AI具备跨应用执行能力，用户只需语音表达需求，即可完成原本需要多次点击的复杂操作，被视为AI时代新交互范式的首次规模化落地探索。该模式本质上以系统AI化取代App智能化：由AI Agent承担任务链的拆解与调度，实现从“点式交互”向“目标直达”的范式升级。这意味着操作入口有望从App转向AI，重塑移动互联网流量与服务分发的权力结构。这对手机厂商而言，AI Agent入口决定未来生态主导权；而对超级App而言，跨应用调度将削弱其对用户的控制，广告曝光与浏览停留等商业基础面临弱化风险，存在被“管道化”的强烈担忧。因此，尽管技术上并无高壁垒，但头部厂商长期保持克制。

- 监管层高度关注跨应用AI助手业务，重点包括：**①反不正当竞争**（是否存在屏蔽或自优先调用）；**②消费者权益保护**（授权范围、风险告知与责任边界）；**③数据安全与黑灰产风险**（跨应用任务执行与高危权限使用）。豆包助手此前获得模拟操作权限（如INJECT_EVENTS），引发外界合规疑虑，后续微信等核心应用已短期限制其相关能力。我们判断，豆包的探索反映出AI生态演进的结构性矛盾：技术已经具备颠覆交互的能力，但商业利益与监管规则尚未完成重构。短期看，豆包将继续面临权限收紧、应用对抗与体验不稳定等多重压力；但长期看，用户对交互效率提升的需求具有不可逆性，系统级AI Agent仍是下一代终极形态。模型、操作系统与服务生态的博弈，将决定这一变革节奏。谁能率先建立可信的系统级AI能力边界与利益分配机制，谁将掌握下一轮智能手机市场主导权。

图：豆包手机迅速售罄

图：豆包手机微信登陆异常

关于微信登录异常的情况

12月2日晚，多位用户反馈，在nubia M153上使用豆包手机助手操作手机功能时，如果涉及操作微信，会出现微信异常退出，甚至无法登录的情况。我们后续下线了手机助手操作微信的能力，目前，nubia M153上被禁止登录的微信账号正陆续解封，请大家等待一段时间并尝试重新登录。

图：Nubia M153 豆包手机助手 技术预览版

数据来源：豆包，国信证券经济研究所整理

5.2 意图框架方案厚积薄发，看好系统、硬件一体化厂商

- 目前AI系统要调用App，主要有两种技术路径：**调用官方接口（意图框架方案）**和**模拟用户操作（纯视觉方案）**。
- **调用官方接口方案：**应用端主动将关键功能或服务注册为「可被系统或语音助手调用的动作（action）/ 接口（intent）」，然后由AI Agent在得到用户授权后直接触发。比如，Deep Link、Intent等机制，以及Android的App Actions或iOS的App Intents、Shortcuts等开发框架。这样，AI无需模拟点击、滑动或输入，而是调用“干净”的能力接口，后台执行服务请求——速度快、稳定性高、对UI兼容性依赖极低、所需算力少、数据安全性高，是主流生态推荐的标准化做法。
- **模拟用户操作方案：通过UI自动化/系统无障碍权限(Accessibility Service)来“模拟真实用户操作”。**这种方式本质上由 AI“扮成人”操作：观察界面（文本、按钮、输入框）、决定点击或输入位置、发送触摸 / 输入事件，从而在多个应用之间跨界面完成任务。这种技术被广泛用于抢红包、自动回复、一键获取权限等应用中。该路径因不依赖应用提前适配接口，对现有生态覆盖广，但稳定性与安全性常受挑战，极易因App UI更新或权限限制而失效。此外，开启无障碍服务之后，因为需要实时监控手机，会引起手机的卡顿、耗电及严重的隐私问题。
- **因此，我们认为调用官方接口方案虽推行不易，但具备长期可发展性，系统与硬件一体化厂商在新一轮AI手机竞争中的胜出概率将进一步强化。**对于第三方App来说，意图框架方案可以避免数据泄露，降低遭到爬虫等恶意攻击的风险，最大程度保护第三方app的自主可控，因此配合意愿将显著高于模拟用户操作方案。不过，意图框架需要第三方App的适配，因此对于自有系统厂商来说更易推进。建议重点关注Apple（iOS系统+苹果硬件生态）、Google（安卓系统+Pixel手机等硬件）、华为（鸿蒙系统+华为硬件生态）产业链。

表：系统级AI Agent的两种实现路径

	调用官方能力 / 系统接口	模拟用户前台操作	注释
执行层实现路径	通过App的API接口	通过模拟用户操作	模拟用户操作方案可能导致第三方app数据泄露，安全可控性受到威胁
开放度	低	高	模拟用户操作方案可以在第三方app零适配前提下自主执行
隐私性	高	低	模拟用户操作方案需要获取无障碍权限
竞争格局	利好具有生态粘性的终端厂商	均可	模拟用户操作方案可以以第三方App形式接入，而意图框架方案以系统厂商主导，与三方app协作，仅ios、鸿蒙、Google等少数厂商具备相关能力
被禁止的风险	低	高	模拟用户操作方案容易被app检测为爬虫等程序而被禁止
算力需求	低	高	意图框架任务执行阶段以调用API接口为主，算力要求低。
稳定性	高	低	如App的UI发生更改，模拟用户操作方案可能识别错误。
参与厂商	苹果、华为、Google	荣耀、OV、豆包	

资料来源：Wind，国信证券经济研究所整理

5.2 OpenAI新形态终端+Google生态链，端侧生态一触即发

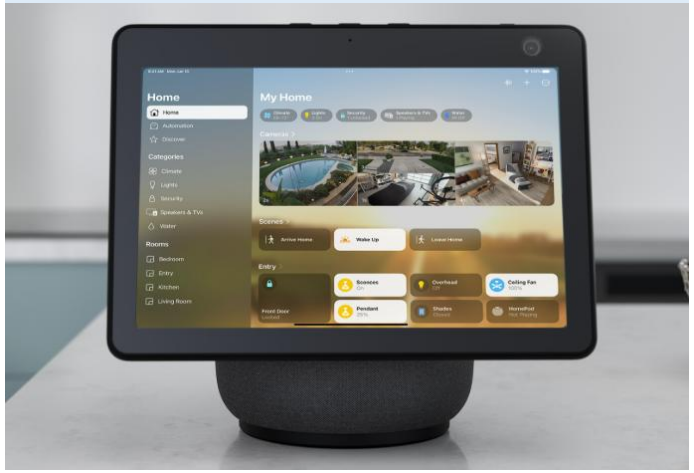
- **OpenAI联合乔尼·艾夫完成AI硬件原型，目标2026年推出“比手机更平静”的智能设备，重塑人机交互范式。**25年5月，OpenAI以65亿美元的价格收购了名为iyO的硬件公司，其负责人为苹果前设计总监，曾设计了Shuffle等苹果的优秀硬件产品。奥特曼在访谈中透露，OpenAI即将要推出的新设备比Humane的AI Pin略大，外形却与iPod Shuffle一样紧凑优雅。尽管目前关于外形的猜测较为多样，但可以明确的是，产品将拥有和苹果系列产品一脉相承的设计语言，追求简洁的外观和非凡的性能，且新产品将在2年内发布。在奥特曼的预想中，新终端可以成为被用户充分信赖的智能AI，它能长期替用户处理事务、过滤噪音。它具有情境感知能力，知道何时该保持静默，何时该呈现信息。用户不再是被动接收信息，而是拥有了选择宁静的权力。

- 10月15日消息，据彭博社报道，马克·古尔曼透露苹果正在越南扩大生产规模，并计划推出多款全新家居设备。知情人士透露，苹果计划在2026年推出一款室内安防摄像头和一款能控制家电、充当家庭中枢的显示屏设备。此外，苹果还计划在2027年推出一款能自主移动的桌面机器人，该机器人配备电机与传感器，外形类似中控显示屏，安装在一根可移动的机械臂上。

图：桌面机器人渲染图

数据来源：Bloomberg，国信证券经济研究所整理

图：带屏幕的homepod渲染图



图：OpenAI新形态终端渲染图

数据来源：AING 硬迹，国信证券经济研究所整理

图：Pixel

5.3 Meta发布量产AR眼镜反馈强烈，巨头跨界布局AI+AR生态

国信证券 GUOSEN SECURITIES

- **Meta发布量产版AR眼镜Meta Display，光学方案采用阵列光波导技术。**9月18日，Meta在Connect 2025大会上发布Meta Ray-Ban Display量产版AR眼镜，采用右眼全彩单目显示屏设计，600×600像素显示屏可提供20度视场角（FOV）内容显示能力，显示屏/内容刷新率90Hz/30Hz，峰值亮度5000尼特。该产品起售价799美元。据AR陀螺报道，该产品光波导方面采用阵列光波导技术，由Lumus授权、Schott（肖特）制造。

- **Meta Display智能眼镜开售两天即售罄，成为AR眼镜量产领域里程碑事件。**2025年9月30日，Meta该新款AR眼镜Meta Ray-Ban Display在部分实体门店开售，仅两天便几乎全面宣告售罄，且许多门店预约已经排到10月底，甚至到11月或12月。Meta表示补货工作即将开始，公司计划将销售该款眼镜的商店数量增加一倍。Meta新款AR眼镜的畅销亦彰显带显示AR眼镜市场热度及消费者认可和推崇。

- **夸克AI眼镜S1发布，阿里在AI+XR生态的布局落地。**11月27日，夸克发布其首款自研旗舰级双显AI眼镜S1，采用双芯片架构，搭载“千问”对话助手；光学方案上，夸克S1采用“二维双目衍射光波导+MicroLED”的组合，在显示效果与全天候佩戴之间取得平衡；配合入眼亮度高达4000nit的绿色光引擎，基于高折射率玻璃基底及光栅的位置设计，大幅抑制了“彩虹纹”效应。基于“近眼显示”能力，#阿里庞大的服务生态通过AI多模态形式延展至眼镜端，同时在物理层面将整机重量控制在51g左右。

图：Meta Display智能眼镜及显示效果示意

资料来源：Meta官网，国信证券经济研究所整理

图：夸克眼镜S1及亮点使用场景

资料来源：夸克官网，国信证券经济研究所整理

图：夸克眼镜S1主要参数

芯片	主芯片	高通骁龙AR1
	协处理器	恒玄BES2800
镜框信息	框型	威灵顿/波士顿
	颜色	曜黑/玳瑁
显示	光引擎	2个Micro-LED单绿光引擎
	光波导	2个1.8高折玻璃二维光栅衍射波导
	入眼亮度	4000nits
	分辨率	640*290
	FOV	26.5度
	显示位置	远近调节+高低调节
镜片信息	近视镜片	1.6、1.67、1.74折射率树脂镜片
摄像头	像素	1200万
	拍照	4032*3024
	录像	1080P@30FPS 3K@30FPS

资料来源：夸克官网，国信证券经济研究所整理

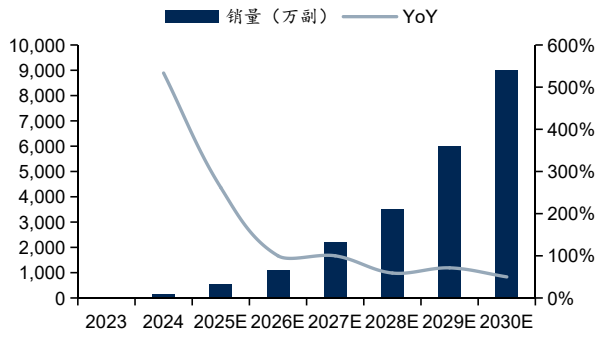
5.3 AI眼镜销量持续提高，AR眼镜市场规模空间广阔

- **全球AI智能眼镜销量持续提高，预计到2030年出货量将达9000万副。**据Wellsenn XR数据，2023-2024年全球AI智能眼镜销量分别为24/152万副。2025年上半年销量达147万副，同比增长220%；随着下半年Meta、阿里、Rokid、三星、理想以及其他品牌AI眼镜新品陆续上市和交付，以及音频眼镜逐步上线AI大模型，Wellsenn XR预计全年AI智能眼镜销量有望超550万副；同时预测到2030年全球AI智能眼镜销量将达9000万副。

- **AR设备虽出货量还处于低量级阶段，但其出货量呈现逐年增长态势。**据Wellsenn XR数据，2024年全球AR出货量为50万台。预计2025年全球AR出货量将达85万台，同比增长70%；主要由于AR眼镜受益于AI眼镜的热度、销量持续增长，同时下半年开始会有阿里、Meta等AR+AI眼镜新产品持续发布和上市，为全年高增长贡献新增量；运动类AR眼镜预计也有超预期的销量贡献。Wellsenn XR预计到2027年，全球AR出货量将达到300万台。

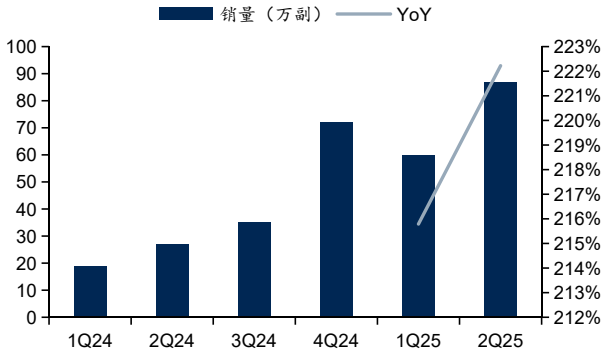
- **AR眼镜市场规模不断提高，未来具有广阔空间。**近年来随着AR芯片、AR光学等技术不断发展，AR眼镜功能日益丰富，成本不断降低，因此逐渐拓展至消费市场。据Statista数据预测，中性条件下2023年全球AR眼镜硬件+软件市场收入额达到22.11亿美元，并将在未来5年内保持稳定增长，中性条件下预计2028年收入额将达到158.46亿美元，期间年复合增长率为48.27%。

图：全球AI智能眼镜销量及预测



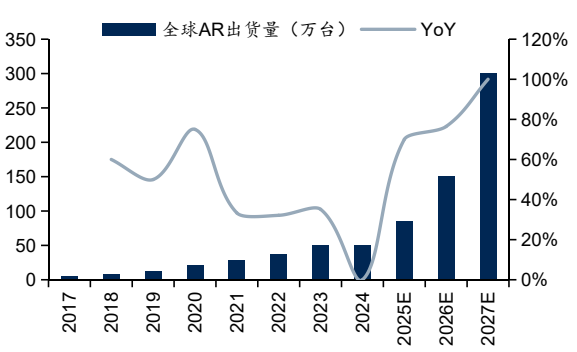
资料来源：Wellsenn XR，国信证券经济研究所整理
注：sell out口径统计，不含未接入AI大模型的音频、拍照以及AR眼镜

图：全球AI智能眼镜季度销量



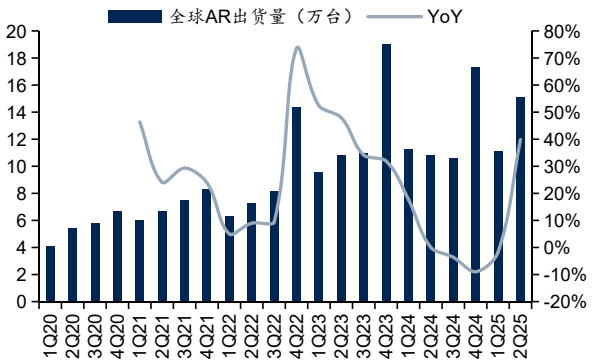
资料来源：Wellsenn XR，国信证券经济研究所整理
注：sell out口径统计，不含未接入AI大模型的音频、拍照以及AR眼镜

图：全球AR设备出货量及预测



资料来源：Wellsenn XR，国信证券经济研究所整理
注：sell out口径统计，不含无屏AR。

图：全球AR设备季度出货量



资料来源：Wellsenn XR，国信证券经济研究所整理
注：sell out口径统计，不含无屏AR。

5.3 光波导方案有望成为AR眼镜所收敛的方案

国信证券 GUOSEN SECURITIES

- **光波导方案主要分为反射光波导和衍射光波导，或将成为未来AR眼镜光学方案的必然趋势。**光波导方案利用在波导结构中光线反射或衍射前进起到传输图像的作用。光波导的这种特性可以将显示屏移至额头，减少对外界光线的阻挡，改善用户的佩戴体验。光波导方案解决了体积和视场角（FOV）与动眼框的矛盾，其轻薄和高穿透性被认为是未来AR眼镜光学方案的必然趋势。

- **衍射光波导具备轻薄和高视场角优势，量产性和良率较易提升。**衍射光波导根据光栅类型的不同可分为浮雕光栅波导与全息光栅波导，其原理利用了光栅的衍射特性，让光能够在设计好的路径上传播，将微投影系统发出的光导入人眼。衍射光波导的优点在于经镀膜后可直接加工，并且以半导体工艺为主，量产性与良率较易提升；同时，衍射光波导保留了轻薄、高视场角的光波导方案优势。而由于物理原理限制，衍射光波导方案可能会导致色散和隐私泄露等问题。

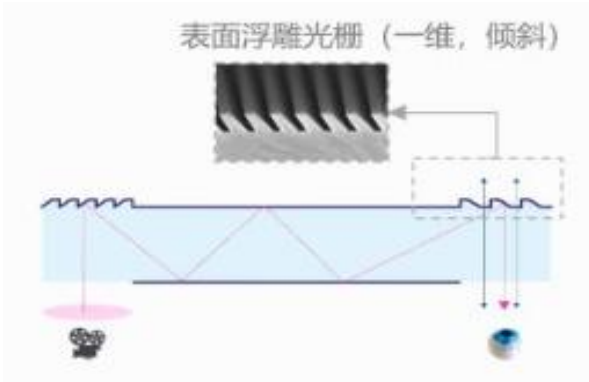
- **反射光波导同样具备轻薄和高视场角优势，成像质量较高。**反射光波导又称阵列光波导，该方案通过阵列反射镜堆叠实现图像的输出，图像光线在阵列内的每一次反射都会经过反射波导进入人眼，增大了动眼眶范围。其优点在于设计原理简单，在减小体积的同时有效增加视场角。同时成像质量、色彩和对比度能达到较高水平。而不足之处在于该方案生产对阵列贴合和切割工艺的一致性要求较高，且自动化能力较弱，因此存在量产难度大，单片价格高的问题；同时存在固有的明暗条纹问题。

表：典型AR光学方案对比

光学方案	棱镜	自由曲面	BirdBath	光波导
形态	棱镜块	楔形	眼镜	眼镜
视场角FOV°	10-20°	20-40°	40-60°	20-60°
透光率	40%-50%	40%-70%	25%-30%	80%-95%
光学效率	20%-30%	20%-40%	15%-20%	0.1%-3%
厚度	＞10mm	＞10mm	20-30mm	1-2mm
优势	量产技术成熟，成本低	成像质量高，色彩饱和度高且光学效率好，量产技术较成熟	结构简单，视场角大，量产技术较成熟	解决体积和视场角的矛盾，厚度重量接近普通眼镜，视场角大，分辨率和透光率高
劣势	重量厚度大，亮度低，视场角小，图像质量差且有畸变	重量和厚度高于普通眼镜，局部图像存在畸变	重量和厚度高于普通眼镜，透光率低，眼动范围受限，图像存在畸变	光学效率低，部分技术图像质量较差，量产能力较低

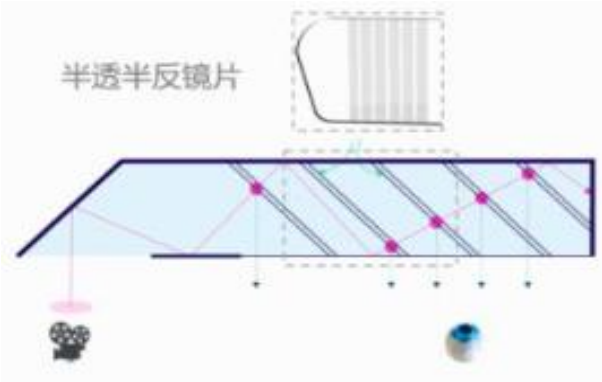
资料来源：VR陀螺，易观分析，国信证券经济研究所整理

图：衍射光波导方案原理示意图



资料来源：艾瑞咨询，国信证券经济研究所整理

图：反射光波导方案原理示意图



资料来源：艾瑞咨询，国信证券经济研究所整理

5.3 AI眼镜新品密集发布，国产SoC渗透率提升

国信证券 GUOSEN SECURITIES

● **Meta Rayban拉开智能眼镜序幕，国内品牌厂商集中投入布局。**2023年9月，Meta-Rayban发布了支持拍摄功能的AI眼镜后，2024年销量迅速超过百万副。此后，国内AR/VR品牌厂商、手机厂商、互联网厂商等，纷纷下场布局AI眼镜产品。2025年，小米AI眼镜、阿里夸克S1、理想AI眼镜等国产品牌的AI眼镜陆续发布。当前不带显示的AI眼镜品牌价格已在2000元以内，而带显示功能的AI眼镜价格也已在4000元以内。而国庆节前发布的Meta Rayban Display售价约799美元，海外媒体预测其生命周期为2年，出货量预计在15至20万部。

表：主流AI智能眼镜数据对比（部分）

序号	智能眼镜	厂商	发布时间	是否支持拍摄	是否带显示	重量	接入大模型	价格	SoC芯片
1	Meta-Rayban二代	Meta-Rayban联合	2023年9月	支持，12MP摄像头	不带	48.6克	Meta AI	299美元	高通AR1
2	魅族StarVR Air 2	星纪魅族	2024年10月	不支持	双目单绿色显示	44克	通义	2799元	恒玄2700
3	Rokid Glasses显示版	Rokid(杭州灵伴科技)	2024年11月18日	支持，1200万像素	双目单绿色显示	49克	通义千问、文心一言等	3149元	高通AR1
4	雷鸟X3 Pro	雷鸟创新	2025年5月27日	支持，1200万像素	支持显示（全彩）	76克	通义千问（多模态）	8999元起	高通AR1
5	小米AI眼镜	小米	2025年6月26日	支持，1200万像素	不带	40克	超级小爱	1999元起	高通AR1+BES2700
6	Meta Ray-Ban Display	Meta-Rayban联合	2025年9月30日	支持，1200万像素	支持显示（HUD）	69克	Meta AI	799美元起	高通AR1
7	阿里夸克AI眼镜S1	阿里巴巴	2025年11月27日	支持，1200万像素	双目单绿色显示	51克	阿里千问	3999元起	高通AR1+BES2800
8	理想AI眼镜Livis	理想汽车	2025年12月3日	支持，1200万像素	不带	36克	理想自研AI	1999元起	BES2800+SA62105X

资料来源：XR Vision Pro，国信证券经济研究所整理

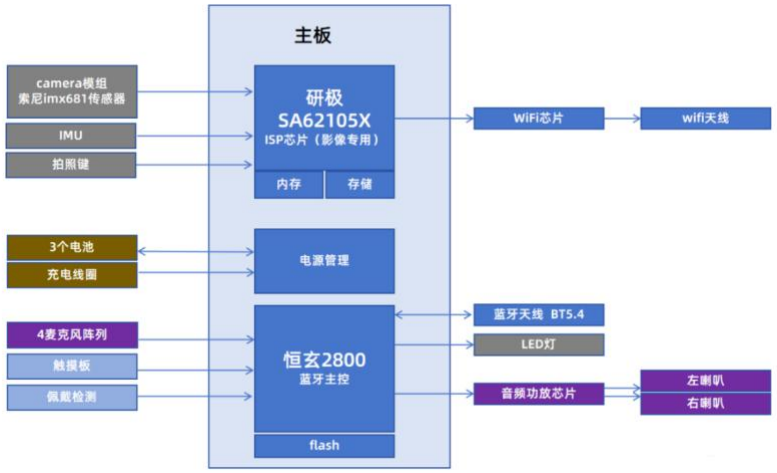
5.3 AI眼镜新品密集发布，国产SoC渗透率提升

国信证券 GUOSEN SECURITIES

- **理想Livis首款采用国产BES2800作为主控，芯片约占硬件BOM的17.5%。**2025年12月3日，理想汽车发布旗下AI眼镜Livis，携手蔡司达成全球战略合作，全系标配蔡司高品质镜片，并提供屈光、感光变色、墨镜等多种版本，售价1999元起。理想Livis是首款以国产芯片恒玄科技BES2800作为主控芯片方案的AI眼镜，可以通过理想同学语音指令，流畅实现空调开启、方向盘加热、尾门控制等操作。根据XR Vision测算，Livis的硬件BOM成本约在800元左右，主控芯片BES2800及研极微ISP合计约占总BOM的近两成。

- **阿里夸克S1采用国产BES2800作为协处理器，芯片约占硬件BOM的16.7%。**2025年11月27日，阿里巴巴旗下夸克发布双显AI眼镜S1，标准套装售价3999，整机重量控制在51g左右。S1光学方案采用“二维双目衍射光波导+MicroLED”组合，芯片方案采用高通AR1+恒玄2800，兼顾4k拍照&3k视频，实现综合续航7h/唤醒待机25h/录像60min+，双芯片方案约占总硬件BOM的不到两成。

图：理想AI眼镜Livis架构图



资料来源：XR Vision，国信证券经济研究所整理

表：理想AI眼镜Livis主要供应商及硬件BOM

模块	料号	预估价格
主控芯片	恒玄BES2800（6nm）	约70元
ISP芯片	SA62105X	约70元
结构件	TR90壳料，鼻托，转轴，散热等	约150元
ODM组装费		约150元
摄像头模组	索尼IMX 681，1200万像素镜头	约50元
PCB和其它电子元件	触摸芯片，电源管理芯片等	约100元
电池	240mAh 锂电池	约30元
麦克风	4阵列麦克风	约10元
扬声器	2个对称双磁路三明治扬声器	约25元
内存及存储	32GB	约45元
充电盒	支持无线充电及TypeC，1700mAh	约100元
总计成本		约800元

资料来源：XR Vision，国信证券经济研究所整理

表：夸克AI眼镜S1主要供应商及硬件BOM

模块	料号	预估价格
芯片	高通AR1	约400元
芯片	恒玄BES2800	约100元
显示	2个JBD蜂鸟MINI 2 MicroLED	约900元
显示	2个1.8高折玻璃二维光栅衍射波导	约600元
影像	索尼IMX681，1200万像素，FOV109°	约70元
结构件		约200元
扬声器	2颗旗舰双音圈超线性扬声器（定制）	约10元
麦克风	5麦克风（2阵列+1骨传mic）	约10元
电池	双电池（280mAh（主）+40mAh（副））	约30元
OEM		约200元
内存+存储	3GB+32GB	约100元
充电盒		约200元
PCB及其他		约200元

资料来源：XR Vision，国信证券经济研究所整理

5.3 AI眼镜新品密集发布，国产SoC渗透率提升

国信证券 GUOSEN SECURITIES

- **高通AR1为当前AI眼镜主流解决方案，第二代有望升级3nm工艺平台。**2023年9月，高通发布了第一代AR1平台，与上一代AR2不同，该平台专为时尚智能眼镜打造，可以不借助手机完成操作，主打高质量拍摄、显示效果，并强化AI功能。AR1采用14-bit双ISP，支持高达1200万像素照片拍摄和600万像素视频拍摄体验，同时还结合了智能手机上的自动曝光、自动人脸检测、计算HDR和人像模式。 同时，高通AR1增强了AI能力，能够帮助增强照片和视频的拍摄质量、通过降噪实现更清晰的通话，并通过计算机视觉实现更清晰的视频拍摄。
- **高通推出新一代AR1+ Gen1，降低尺寸并提升能效。**2025年6月在美国增强现实世界博览会上，高通新发布了第一代骁龙AR1+平台，其相较于前代平台尺寸缩小26%，同时增强了图像质量、提升了能效，并具备运行小语言模型（SLM）的能力。

图：高通骁龙AR1 Gen1

资料来源：高通官网，国信证券经济研究所整理

表：高通AR1与W5等对比

	Wear 4100+	W5/W5+ Gen1	AR1 Gen 1
工艺制程	12nm	4nm	4nm
CPU	4×A53，最高可达2.0GHz	4×A53，1.7GHz	4×A55，1.9+GHz
内存	LPDDR3，750MHz	1×16 LPDDR4，2133MHz	1×16 LPDDR4X，2.1GHz，LLC
GPU	Adreno 504@320MHz，支持OpenGL ES3.1	Adreno 702@1GHz	Adreno 621，支持OpenGL ES3.2和Vulkan 1.1
显示	1080p 30fps	1080p 60fps	1280x1280 60fps(Dual)
ISP	Dual ISP 16MP+16MP	Dual ISP 16MP+16MP EIS(电子防抖)3.0,FNR(多帧降噪算法) Pseudo ZSL(伪零快门延迟) 2x CSI 4lane DPHY/CPHY	2×12MP，IFE和IFE-lite(图像前端处理单元)， HW JPEG Enc(JPEG编码器)，Pseudo ZSL(伪零快门延迟)，CSI4×4-lane(4×2-lane+4×1-lane)CPHY1.2/DPHY1.2
DSP	Dual Qualcomm® Hexagon™ QDSP6 v56 用于调制解调器和GPS的专用MDSP 用于开放式传感器执行环境和音频的专用ADSP	Dual Hexagon QDSP V66K HiFi 5 DSP	Hexagon DSP，1.2GHz，2MB LPI,eNPU, Sensor Hub，Voice UI
接口	USB2.0	USB2.0	SPI-NOR，12 SE，1xUSB3.1 Gen1， 2x PCIe Gen3 1-lane
蓝牙	蓝牙5.0	蓝牙5.3	蓝牙5.3
WiFi	WiFi 4（2.4GHz）	WiFi 4（2.4GHz/5GHz）	WiFi 7

资料来源：高通官网，Wellsenn XR，国信证券经济研究所整理

5.3 AI眼镜新品密集发布，国产SoC渗透率提升

国信证券 GUOSEN SECURITIES

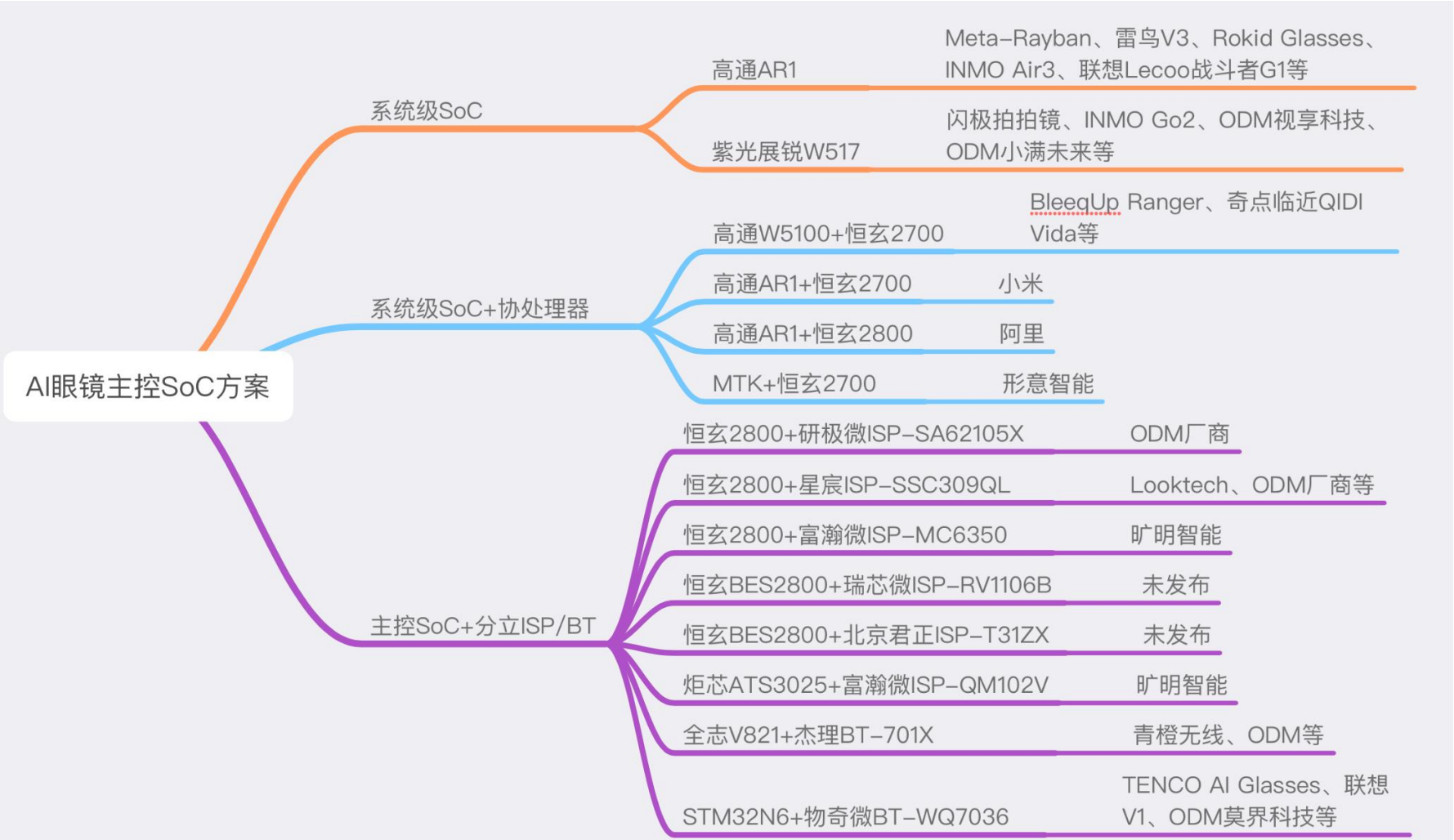
- 当前AI眼镜主流SoC方案主要包含以下三大类：

（1）系统级SoC单芯片方案：集成了CPU、GPU、ISP、DSP、WiFi、蓝牙等模块，综合性能强，如高通AR1 Gen1，展锐W517等。

（2）系统级SoC+协处理器双芯片方案：主控SoC负责拍摄、视频等复杂功能，协处理器负责待机、音频等低功耗场景，如高通AR1+恒玄BES2800等。

（3）主控SoC+ISP双芯片方案：集成度及性能弱于系统级SoC，但续航等领域具备优势，如BES2800+SSC309QL。

图：AI智能眼镜三类主控SoC方案概况



资料来源：XR Vision Pro，国信证券经济研究所整理

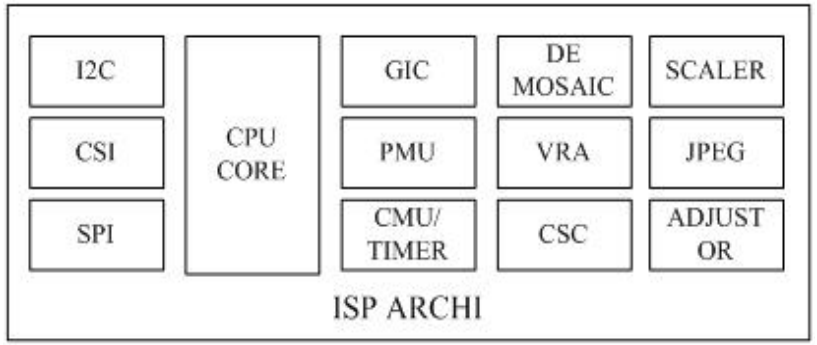
5.3 AI眼镜新品密集发布，国产SoC渗透率提升

国信证券 GUOSEN SECURITIES

- **ISP决定眼镜成像质量，动态防抖和低功耗尤为关键。**对AI眼镜来说，高性能的ISP决定AI眼镜成像质量的关键环节，AI眼镜因佩戴者的使用环境常为动态场景，故动态防抖要求较高。此外，ISP芯片需满足低功耗要求以延长续航，满足长时间佩戴需求。

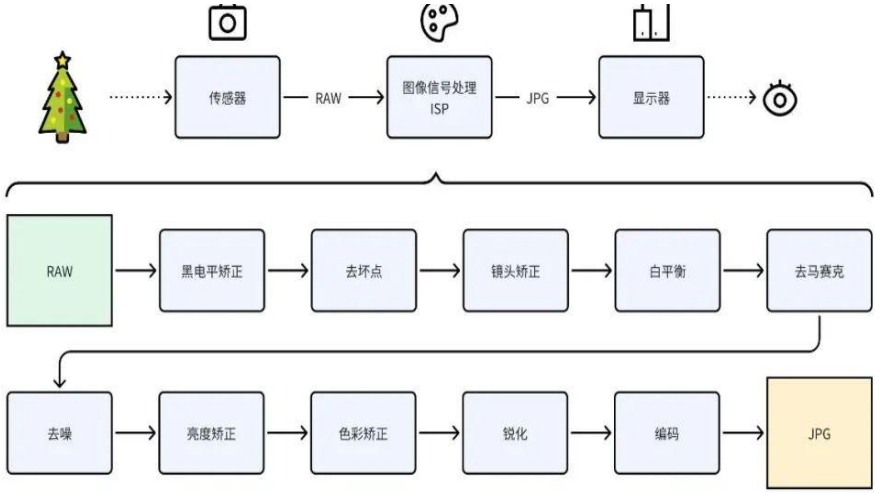
- ISP（Image Signal Processor）即图像信号处理器，主要用来对前端图像传感器输出信号处理的单元，以匹配不同厂商的图象传感器。ISP内部包含CPU、SUP IP等，可以认为ISP是一颗完整的SoC，能够运行各种算法程序，实时处理图像信号。ISP一般处理图像传感器的输出数据，如AEC（自动曝光控制）、AGC（自动增益控制）、AWB（自动白平衡）等功能。ISP处理图像的主要流程包括，黑电平矫正（Black Level Correction）、去坏点（Defect Pixel Correction）、镜头阴影校正（Lens Shading Correction）、解马赛克（Demosaicing）、白平衡校正（White Balance）、降噪（Noise Reduction）、Gamma、色彩校正矩阵（CCM）、锐化（Sharpening）等。

图：ISP内部构成示意图



资料来源：新机器视觉，国信证券经济研究所整理

图：ISP处理流程



资料来源：维深Wellsenn XR，国信证券经济研究所整理

图：黑电平校正前后对比

矫正前　　矫正前

资料来源：Wellsenn XR，国信证券经济研究所整理

5.4 AI重塑智能家居生态，Gemini焕新赋能AIoT硬件

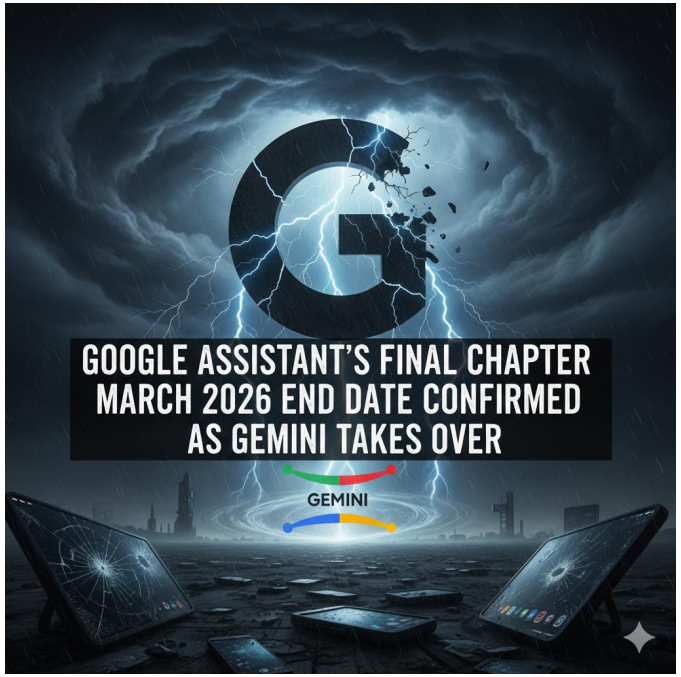
国信证券 GUOSEN SECURITIES

- **Matter引导跨品牌设备联动，Gemini AI驱动全新硬件设备。**2019年谷歌、苹果、亚马逊等巨头联合推动Matter，目标打破不同品牌智能家居设备的兼容壁，谷歌作为Matter的核心推动者，将Matter与Google Home应用深度融合。此外，谷歌推出了由Gemini AI驱动的全新Google Home和Nest设备阵容，目标Gemini普及到其现有的超过8亿台设备生态系统中。谷歌计划在特定领域打造旗舰硬件以展示创新，同时允许第三方制造商和企业集成Gemini。

① **推进Gemini替代Google Assistant并优化交互：**谷歌宣布将在2026年3月31日停用Google Assistant，开启近十年助手时代的收官，由Gemini全面接管智能家居生态核心系统，覆盖横跨Android Phone、Android Auto、Wear OS、Google TV及其他谷歌生态设备。

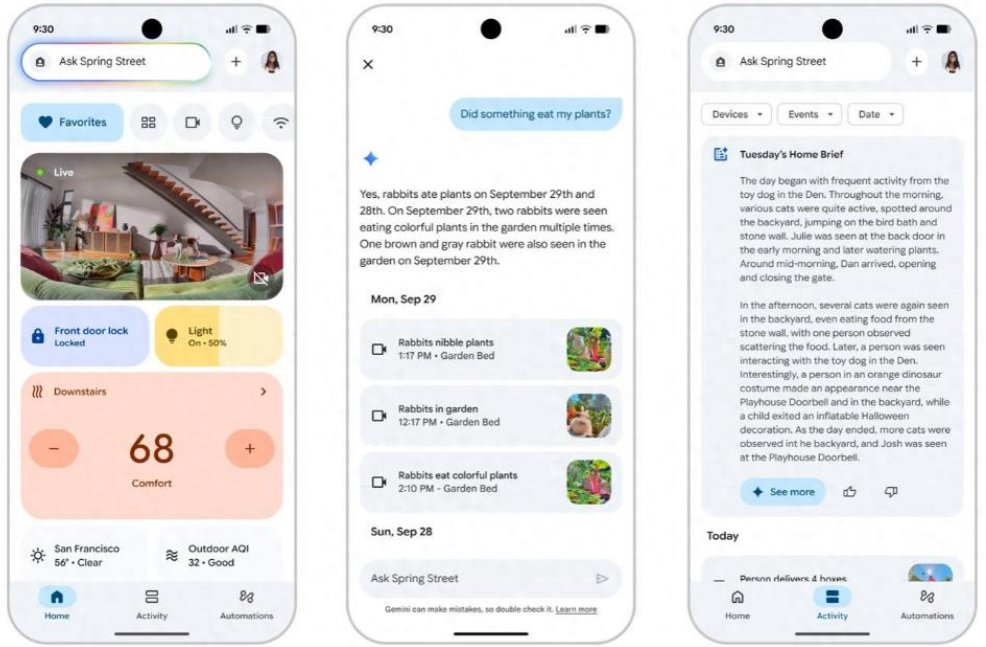
② **全面重构Google Home应用：**谷歌自10月起对该应用完成重大更新，界面简化为Home、Activity和Automations三大板块，崩溃率降低80%。新增“Ask Home”功能，支持自然语言交互执行复杂操作，同时提升Nest相机视频加载速度，还加入基于描述检索片段的功能，操作更便捷。

图：谷歌计划2026年3月停用Google Assistant



资料来源：AI Evangelist，国信证券经济研究所整理

图：谷歌全面重构Google Home应用



资料来源：Google Home官网，国信证券经济研究所整理

5.4 AI重塑智能家居生态，Gemini焕新赋能AIoT硬件

国信证券 GUOSEN SECURITIES

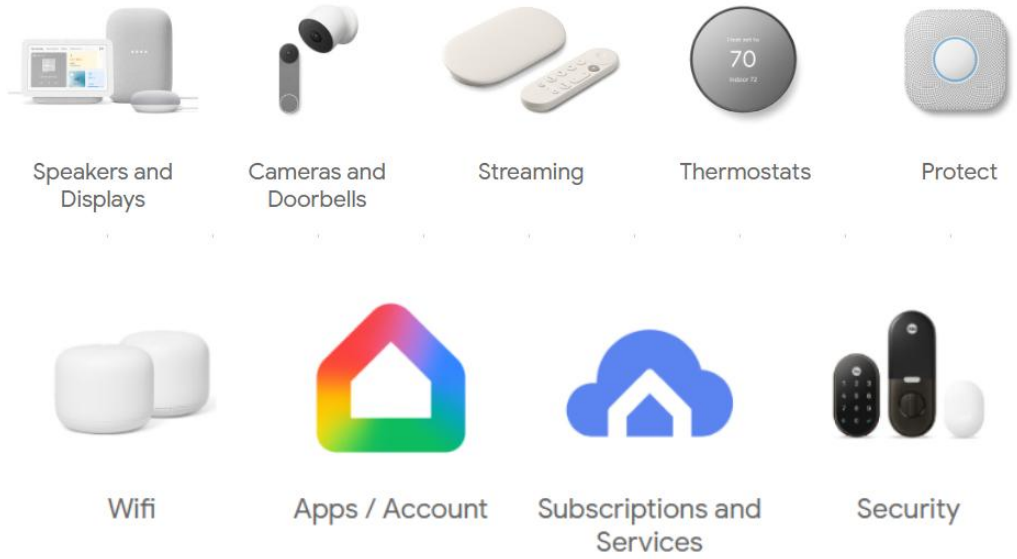
③ 推出多款Gemini驱动的Nest硬件新品：10月谷歌发布第三代Nest Cam Indoor、第二代Nest Cam Outdoor和第三代Nest Doorbell门铃，售价分别约100美元、150美元和180美元。新品支持2K HDR录制，视角拓宽且弱光下能长时间保持全彩模式，还可借助Gemini AI生成精准告警描述，比如识别“狗从玩具围栏跳出来”这类具体场景。同时还与沃尔玛合作推出Onn低成本摄像头和门铃，让Gemini生态覆盖更多价位的硬件。此外，谷歌还计划2026年春季推出新款Google Home智能音箱。

④ 推出分级订阅服务与合作计划：推出Google Home Advanced订阅服务，每月10美元或每年100美元的方案含30天事件历史等基础功能，每月20美元或每年200美元的方案可解锁60天事件片段等全部新特性。

图：Google发布新一代Nest硬件产品

资料来源：Google Nest官网，国信证券经济研究所整理

图：Google Nest硬件产品概况



资料来源：Google Nest官网，国信证券经济研究所整理

【6】半导体：自主可控进程有望超预期，受益景气复苏的模拟芯片加速国产替代

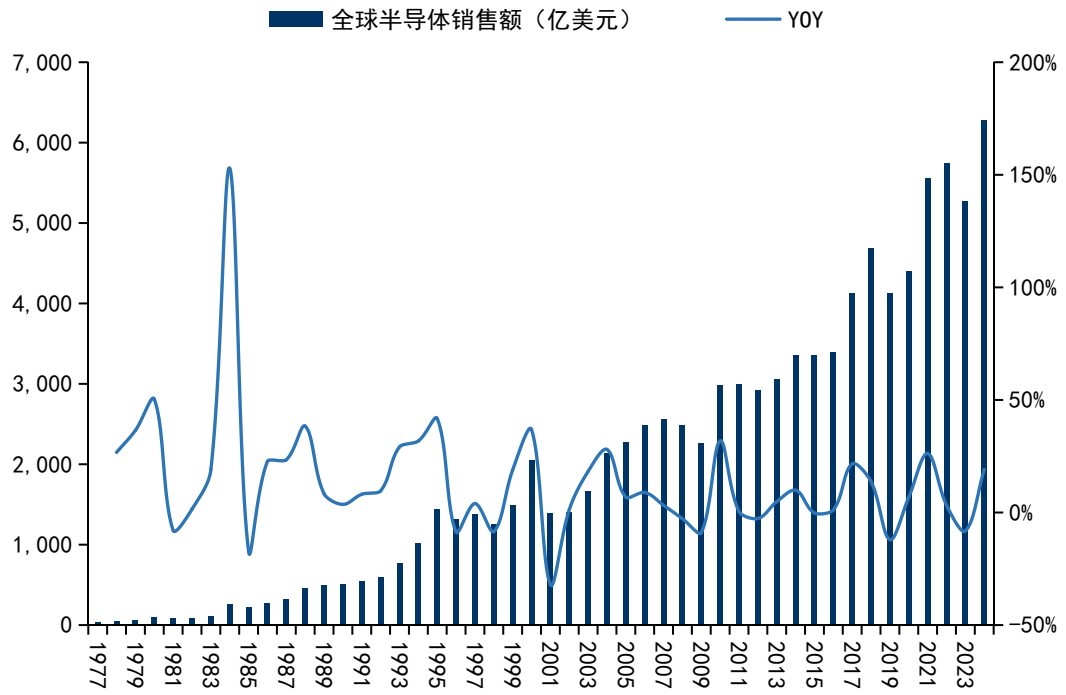
6.1 全球半导体销售额近十年CAGR为6.5%，中国占比28%

国信证券 GUOSEN SECURITIES

- **2014-2024年全球半导体市场规模的CAGR为6.5%，2024年同比增长19%至6276亿美元。**根据SIA的数据，全球半导体销售额从1977年的35.5亿美元增长到2024年的6276亿美元，近十年的年均复合增速为6.5%。

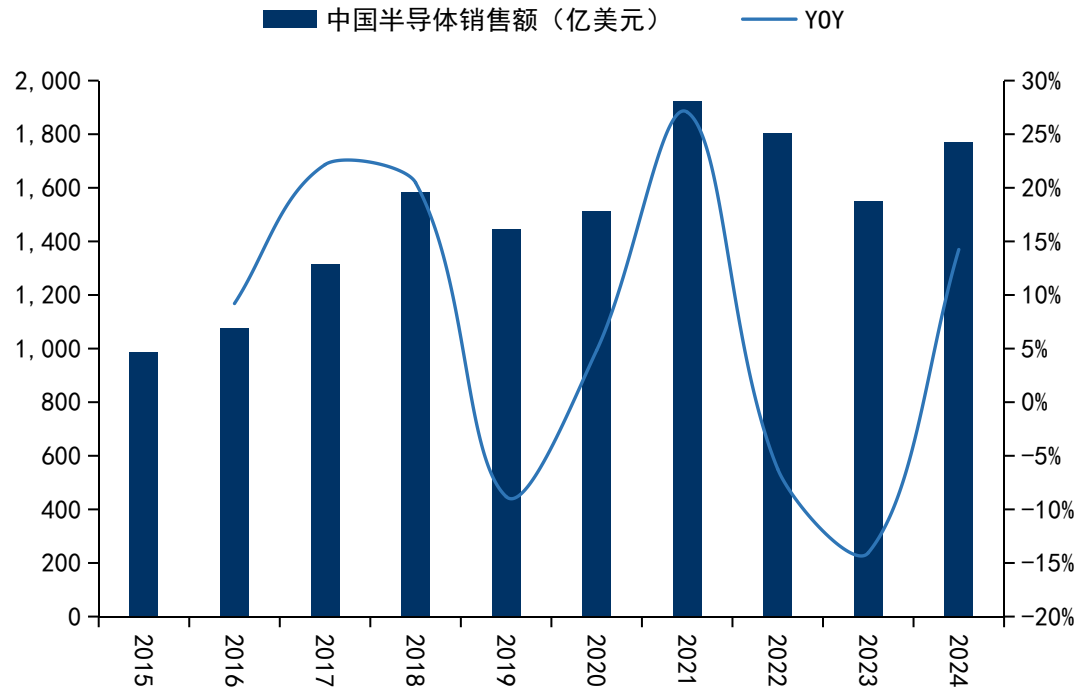
- **2024年中国占全球半导体销售额的28%。**中国是半导体销售的重要市场，2019年占比超过35%，之后几年因中美贸易摩擦占比呈下降趋势，2024年占比为28%。

图：全球半导体销售额及增速



资料来源：SIA，国信证券经济研究所整理

图：中国半导体销售额及增速



资料来源：SIA，国信证券经济研究所整理

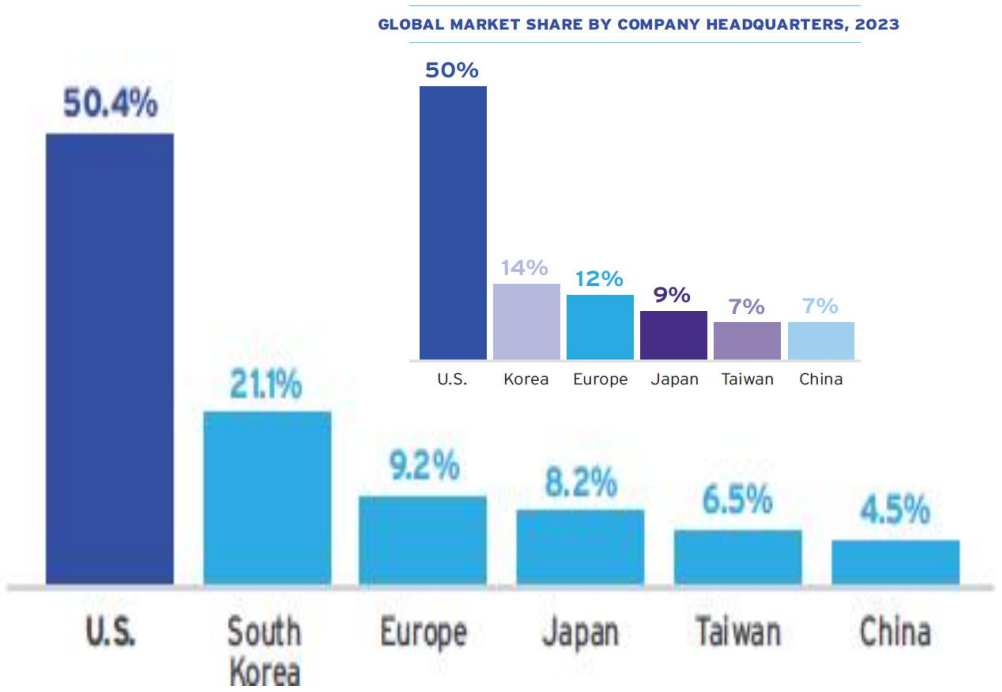
6.1 中国半导体企业的供应比例远低于中国的销售额比例

国信证券 GUOSEN SECURITIES

- **2024年中国半导体企业的供应比例约4.5%。**从半导体企业总部所在地来看，美国仍是半导体的主要供应国，2024年供应比例约50%，中国的供应比例约4.5%，远低于销售额占比。2023年中国企业供应比例为7%，比2020年的供应比例5%有所提高。

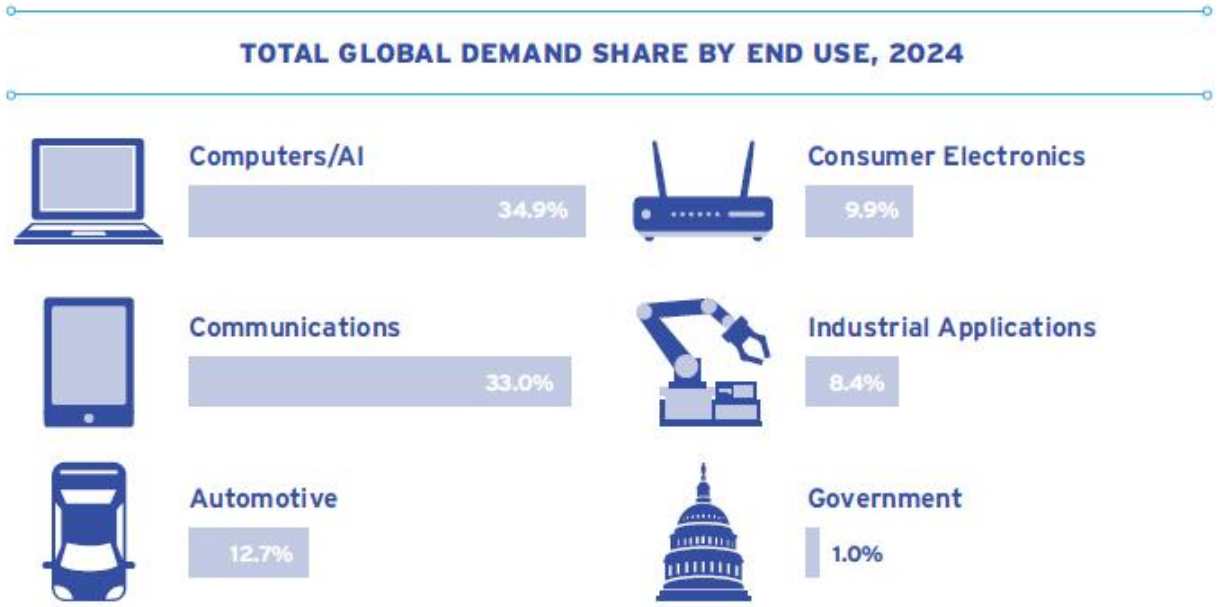
- **2024年半导体下游中计算机/AI占比最高。**从下游领域来看，根据SIA的数据，2024年计算机/AI占比34.9%，通讯占比33%，汽车占比12.7%，消费电子占比9.9%，工业占比8.4%，政府占比1.0%。

图：2024年中国半导体企业供给率仅4.5%



资料来源：SIA，国信证券经济研究所整理

图：2024年全球半导体下游构成



资料来源：SIA，国信证券经济研究所整理

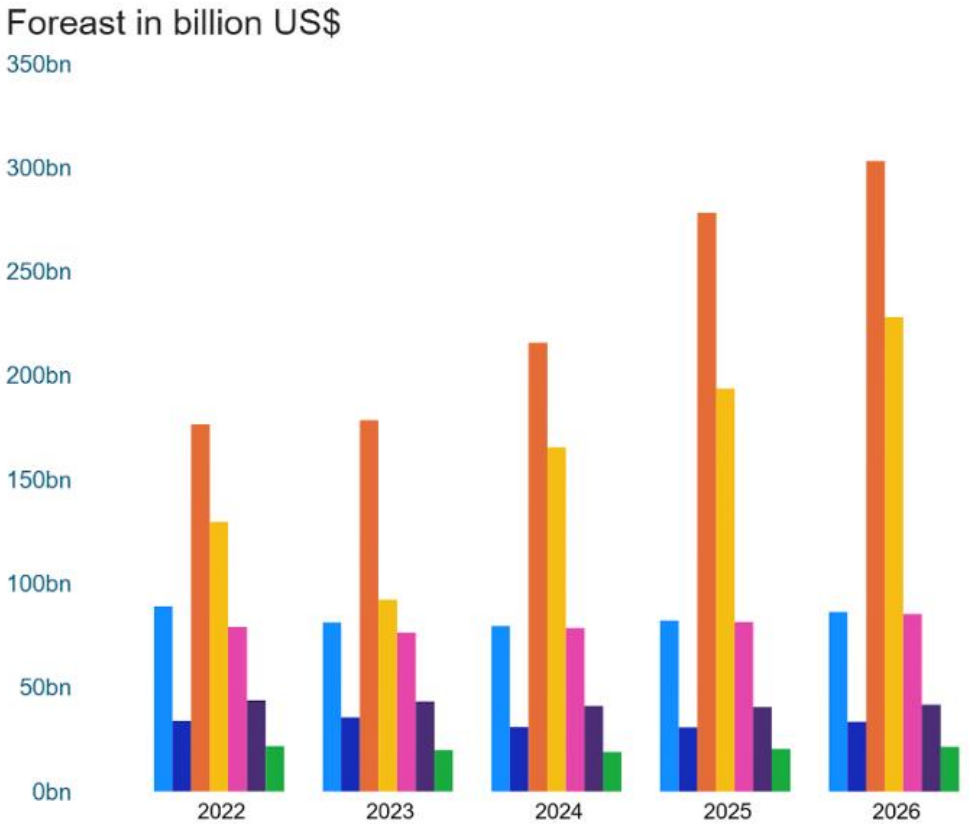
6.1 WSTS提高全球半导体销售额预测值

国信证券 GUOSEN SECURITIES

- 2025年8月，WSTS将2025年全球半导体销售额同比增速由前次预测值11.2%提高至15.4%，达到7280亿美元；将2026年预计增速由8.5%提高至9.9%，达到8000亿美元。**2024-2026年全球半导体将实现连续三年增长。**

- 从细分产品类别来看，2024年的增长主要来自存储和逻辑（AI带动GPU、HBM的需求），**2026年所有细分产品均将进入正增长**，我们认为这主要是由于各下游均去库完成，且将有更多AI端侧产品落地，带动各类辅助芯片需求增长。

图：全球半导体销售额预测



Segment Growth Y/Y	2023	2024	2025	2026
⊞ Discretes	4.5%	-12.7%	-0.6%	8.9%
⊟ IC	-9.7%	25.9%	17.9%	10.6%
⊞ Analog	-8.7%	-2.0%	3.3%	5.1%
⊞ Logic	1.1%	20.8%	29.0%	9.0%
⊞ Memory	-28.9%	79.3%	17.1%	17.8%
⊞ Micro	-3.5%	3.0%	3.9%	4.6%
⊟ Opto & Sensors	-4.2%	-4.6%	1.8%	3.6%
⊞ Optoelectronics	-1.6%	-4.8%	-1.1%	2.8%
⊞ Sensors	-9.4%	-4.1%	7.9%	5.3%
Total	-8.2%	19.7%	15.4%	9.9%

WSTS

资料来源：WSTS，国信证券经济研究所整理

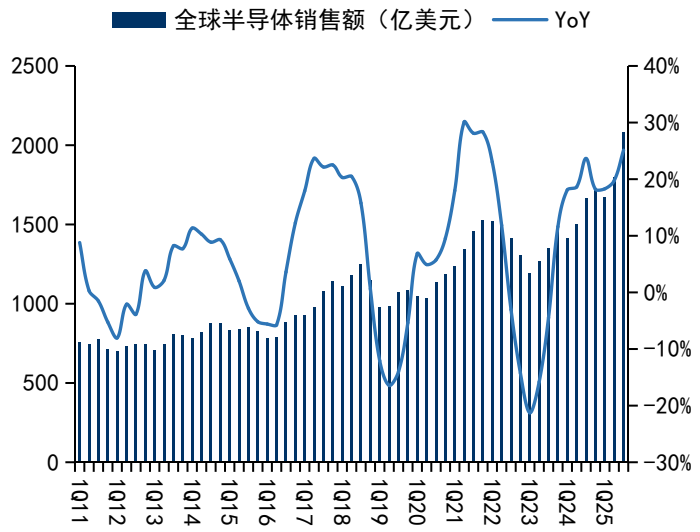
6.1 全球半导体销售额连续八个季度同比增长

国信证券 GUOSEN SECURITIES

- **3Q25全球半导体销售额同比增长25.1%至2084亿美元，连续八个季度同比增长。**根据SIA的数据，3Q25全球半导体销售额为2084亿美元，同比增长25.1%，环比增长15.8%，连续八个季度同比增长；中国半导体销售额为561亿美元，占全球的26.9%，同比增长15%。从季度同比增速来看，目前处于相对高位。

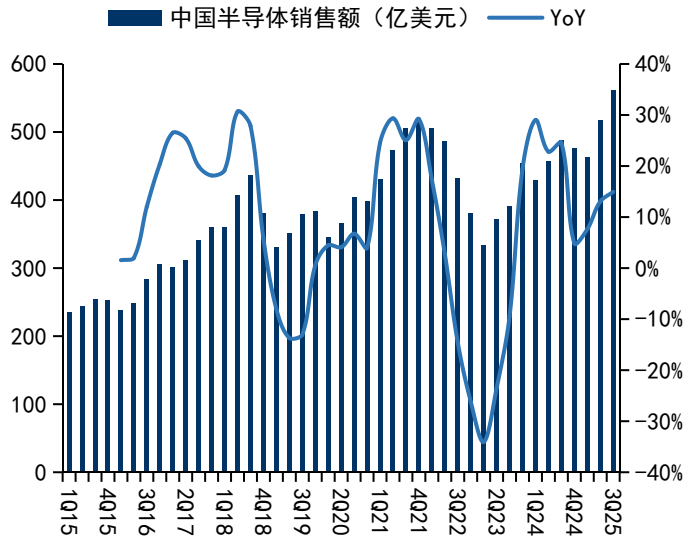
- **中芯国际和华虹半导体产能利用率持续回升。**根据中芯国际的公告，3Q25产能利用率为95.8%，环比提高3.3pct，同比提高5.4pct。根据华虹半导体的公告，3Q25产能利用率为109.5%，环比提高1.2pct，同比提高4.2pct。

图：全球半导体季度销售额



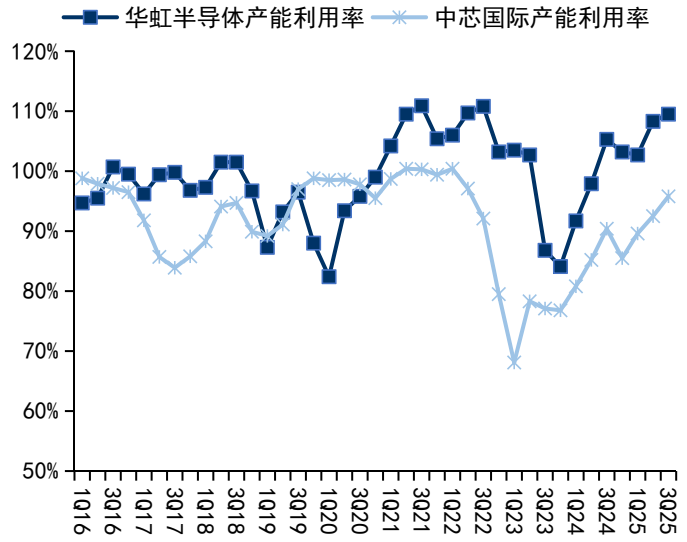
资料来源：SIA，国信证券经济研究所整理

图：中国半导体季度销售额



资料来源：SIA，国信证券经济研究所整理

图：中芯国际和华虹半导体的产能利用率



资料来源：各公司公告，国信证券经济研究所整理

6.1 半导体板块毛利率继续回升，多家公司2025年收入创季度新高

国信证券 GUOSEN SECURITIES

● **半导体板块毛利率、净利率。**A股半导体板块整体毛利率在1Q24触底后回升，3Q25毛利率为29.3%，处于1Q21和2Q21之间。净利率在4Q24触底，3Q25约11%，与4Q20、1Q21水平相当。

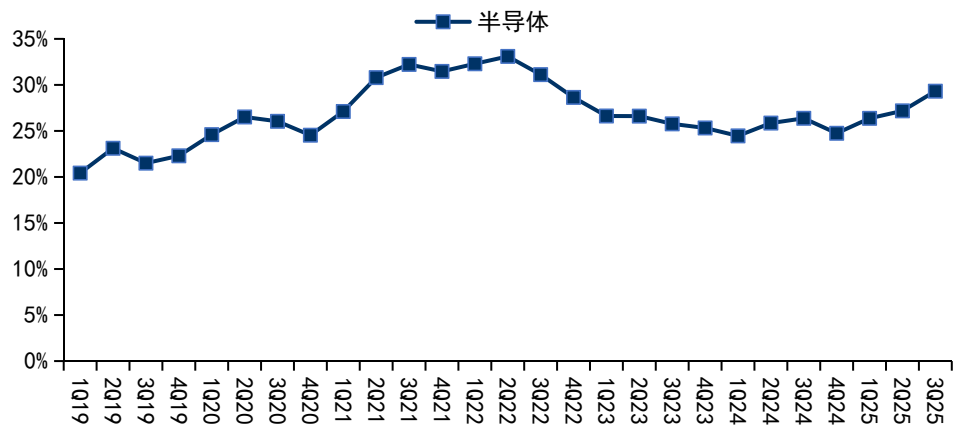
● 从公司分布来看，我们统计了从1Q22开始有季度经营数据的146家半导体上市公司，其中**季度收入最高落在2025年的有79家，占比达54%，**我们认为，在行业去库存结束后，芯片国产化、高端化以及AI带动的增量是国内半导体企业收入增长的主要动力；季度毛利率最高主要在2021、2022年缺芯涨价期间；季度毛利率最低主要在2023/2024年，**2025年季度毛利率最低的有22家公司，占比15%，说明大部分半导体公司毛利率已从低点回升。**

图：半导体上市公司经营情况好转

	季度收入最高的公司数量	占比	季度毛利率最高的公司数量	占比	季度毛利率最低的公司数量	占比
2019	2	1%	17	12%	14	10%
2020	5	3%	15	10%	14	10%
2021	18	12%	**52**	**36%**	11	8%
2022	16	11%	**35**	**24%**	10	7%
2023	11	8%	10	7%	**30**	**21%**
2024	15	10%	6	4%	**45**	**31%**
2025	**79**	**54%**	11	8%	22	15%
合计	146		146		146	

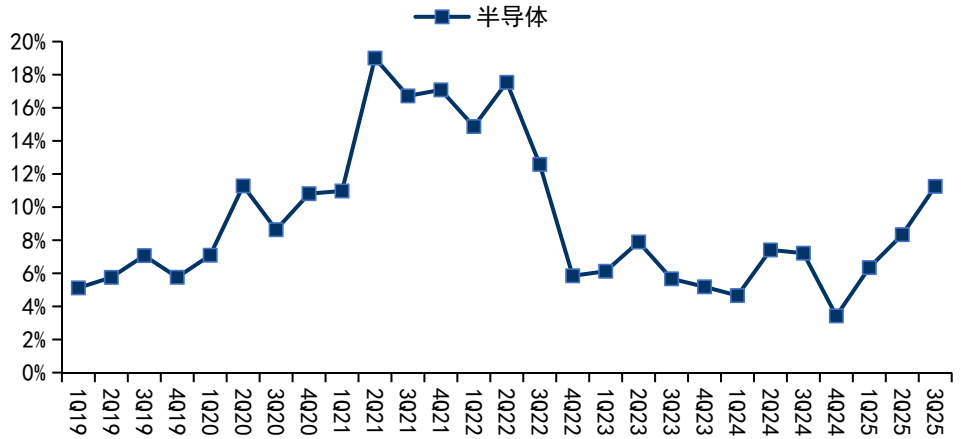
资料来源：Wind，国信证券经济研究所整理

图：SW半导体板块季度毛利率



资料来源：Wind，国信证券经济研究所整理

图： SW半导体板块季度净利率

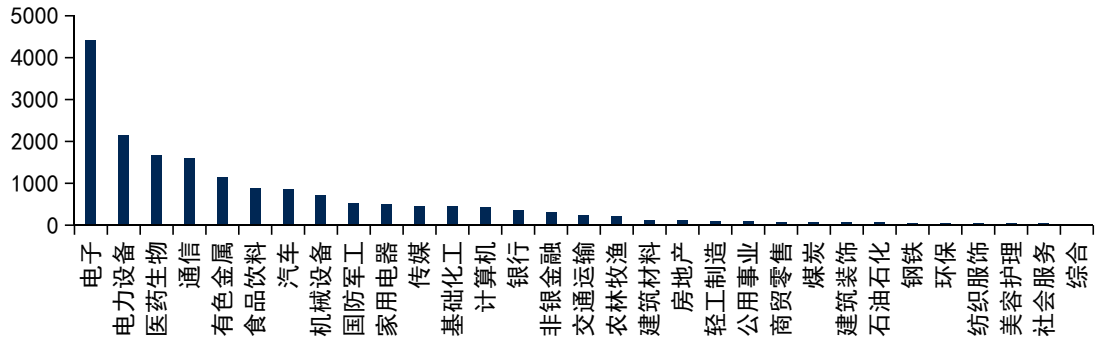


资料来源：Wind，国信证券经济研究所整理

6.1 3Q25主动基金半导体重仓持股比例为12.56%

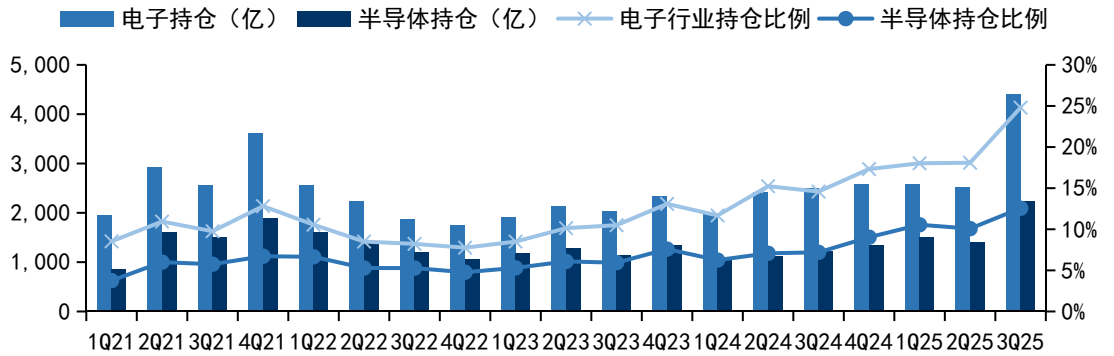
- 3Q25主动基金重仓持股中电子公司市值为4413亿元，持股比例为18.11%；半导体公司市值为2235亿元，持股比例为12.56%，环比提高2.5pct。**相比于半导体流通市值占比5.89%超配了6.7pct。**
- 3Q25前五大半导体重仓持股占比由2Q25的37%上升至41%，第一大占比为12%。

图：各行业基金重仓市值



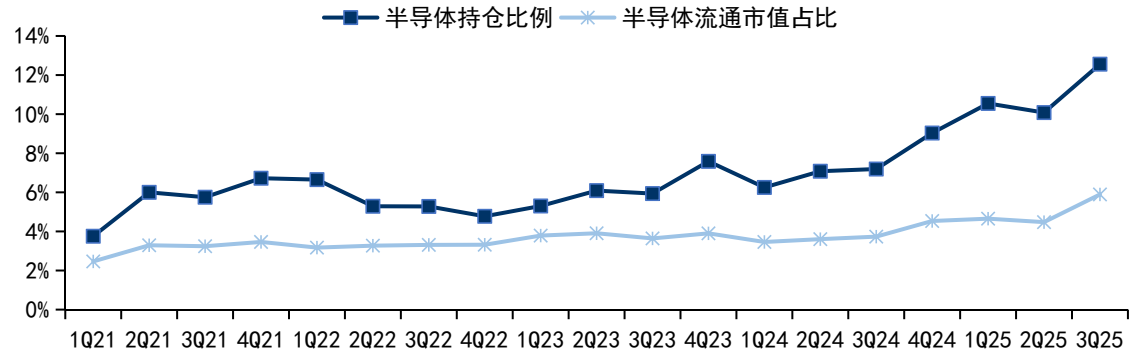
资料来源：Wind，国信证券经济研究所整理

图：半导体重仓持股市值及比例



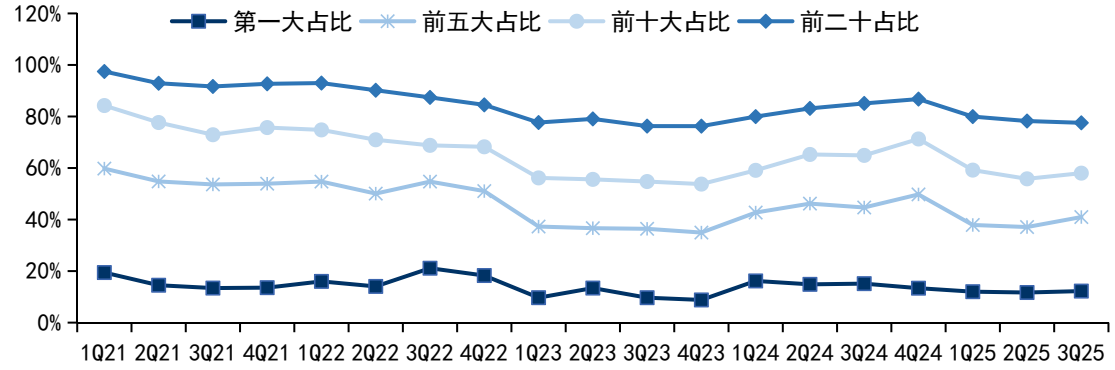
资料来源：Wind，国信证券经济研究所整理

图：半导体持仓占比及流通市值占比



资料来源：Wind，国信证券经济研究所整理

图：3Q25半导体前五大重仓持股占比为52.8%



资料来源：Wind，国信证券经济研究所整理

6.1 代表性公司主被动基金持仓情况

国信证券 GUOSEN SECURITIES

● 从统计的20家半导体细分方向龙头公司来看，2025年中被动基金持股比例超过主动基金持股比例的有15家，主动基金持股比例更高的有5家。合计来看，主动基金持股比例呈上下波动趋势，被动基金持股比例从2021年中开始一直呈上升趋势，自2022年底一直高于主动基金。

	合计		中芯国际		华虹公司		海光信息		寒武纪		北方华创		中微公司	
	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比
2021年底	3.90%	6.99%	8.10%	3.58%					2.79%	1.01%	2.28%	15.06%	3.48%	7.75%
2022年中	5.45%	6.50%	12.16%	3.58%					4.30%	0.90%	2.70%	14.21%	7.34%	6.22%
2022年底	6.88%	5.87%	17.77%	2.33%			0.81%	2.88%	6.91%	1.79%	3.96%	11.42%	12.40%	9.68%
2023年中	8.97%	7.67%	21.58%	9.81%			1.32%	2.13%	10.06%	10.20%	4.46%	13.04%	15.05%	17.22%
2023年底	11.69%	8.65%	28.98%	8.81%	0.84%	3.68%	10.45%	3.19%	12.71%	7.96%	5.11%	11.77%	16.78%	18.70%
2024年中	12.42%	6.67%	29.03%	7.26%	2.13%	5.28%	11.05%	3.12%	13.15%	10.20%	6.54%	12.37%	19.19%	16.77%
2024年底	13.79%	4.55%	27.69%	8.16%	7.72%	1.55%	12.27%	4.24%	15.37%	7.72%	7.67%	10.03%	21.88%	10.90%
2025年中	14.14%	4.40%	27.76%	5.79%	9.32%	12.36%	12.11%	3.31%	16.10%	5.03%	7.44%	8.51%	21.86%	8.43%

	豪威集团		卓胜微		澜起科技		长电科技		通富微电		华润微		士兰微	
	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比
2021年底	3.82%	12.65%	5.19%	12.06%	3.47%	9.40%	3.62%	0.43%	4.06%	0.66%	2.00%	0.44%	2.25%	11.72%
2022年中	4.44%	9.13%	5.42%	11.78%	5.33%	8.23%	4.85%	0.59%	5.05%	0.17%	4.99%	0.19%	3.11%	10.46%
2022年底	5.27%	3.98%	6.42%	14.36%	12.73%	12.94%	5.62%	0.64%	4.23%	1.66%	7.04%	0.07%	4.09%	5.16%
2023年中	6.82%	5.32%	8.92%	11.65%	15.61%	9.85%	6.53%	10.46%	5.72%	11.48%	8.80%	0.05%	4.81%	2.25%
2023年底	7.30%	6.78%	10.61%	13.98%	18.38%	15.82%	7.48%	12.55%	5.77%	22.91%	10.44%	0.05%	4.70%	0.22%
2024年中	9.40%	4.19%	11.25%	7.02%	20.16%	14.70%	8.00%	13.07%	5.64%	3.99%	10.30%	0.18%	6.03%	0.63%
2024年底	10.84%	2.51%	12.62%	2.42%	26.92%	3.39%	9.31%	5.34%	5.94%	1.26%	11.32%	0.07%	7.68%	0.14%
2025年中	10.86%	4.24%	11.96%	2.04%	31.95%	5.63%	9.75%	1.09%	6.22%	1.33%	11.28%	0.07%	7.99%	0.35%

	沪硅产业		天岳先进		兆易创新		恒玄科技		紫光国微		圣邦股份		翱捷科技	
	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比	被动基金持股占比	主动基金持股占比
2021年底	2.53%	4.99%			5.90%	16.46%	0.90%	8.40%	5.47%	24.00%	4.66%	28.38%		
2022年中	4.68%	5.58%	0.04%	2.09%	7.42%	18.78%	5.17%	3.85%	4.15%	22.53%	6.65%	27.74%	0.06%	1.13%
2022年底	7.02%	5.44%	0.83%	2.99%	9.94%	13.98%	3.16%	9.16%	6.14%	23.27%	7.57%	28.92%	0.79%	0.73%
2023年中	10.97%	5.34%	3.73%	1.92%	10.42%	13.13%	3.82%	12.09%	7.28%	9.12%	9.13%	22.64%	7.59%	8.56%
2023年底	14.78%	2.66%	8.51%	6.08%	11.29%	12.76%	6.30%	16.46%	7.88%	8.72%	9.53%	24.57%	9.75%	11.68%
2024年中	15.51%	2.22%	8.68%	2.27%	12.68%	17.29%	5.13%	20.98%	8.08%	4.33%	9.93%	25.84%	10.01%	7.09%
2024年底	17.11%	1.75%	8.85%	0.76%	12.42%	16.73%	6.56%	25.94%	9.11%	2.73%	11.47%	22.31%	10.08%	4.76%
2025年中	17.47%	1.63%	8.18%	5.29%	12.30%	19.39%	17.19%	16.91%	9.08%	4.23%	10.95%	16.93%	6.53%	11.43%

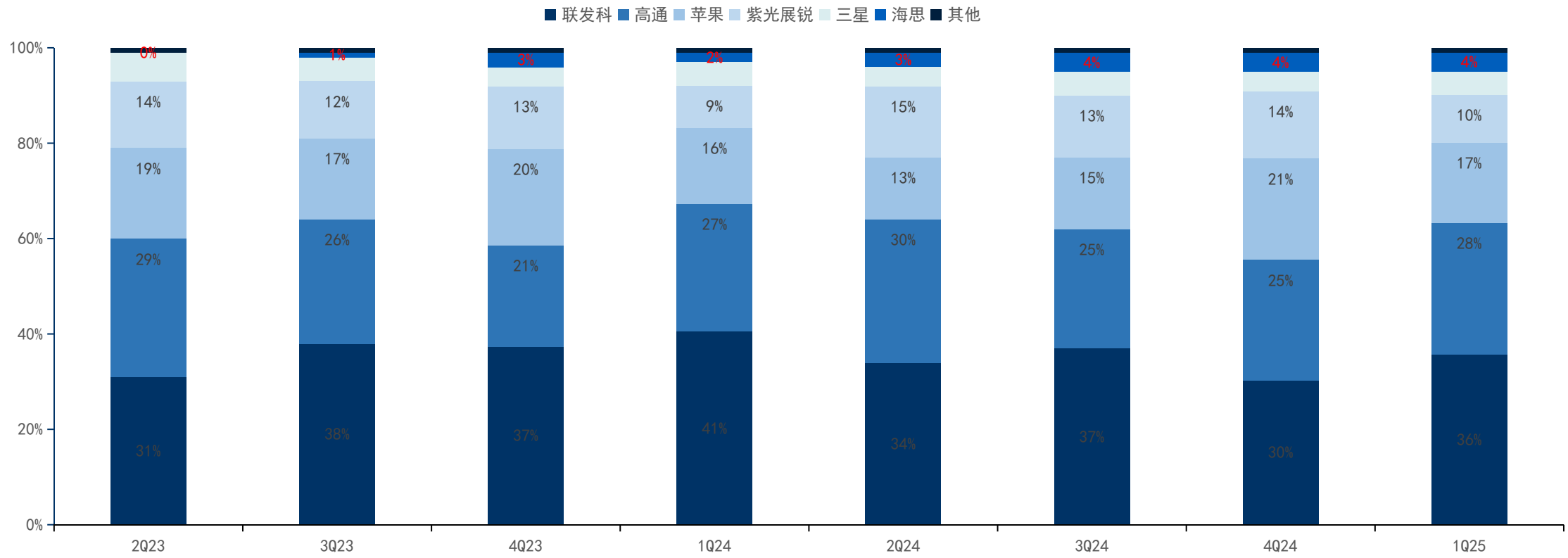
资料来源：Wind，国信证券经济研究所整理（被动基金：被动指数型基金、增强指数型基金，主动基金：偏股混合型基金、普通股票混合型基金、灵活配置型基金、平衡混合型基金）

6.2 晶圆代工：受益国产芯片设计企业崛起和在地化制造趋势

国信证券 GUOSEN SECURITIES

● **海思手机处理器全球市占率回升，华为公布昇腾芯片迭代路线图。**根据Counterpoint Research的数据，2Q23海思在全球手机处理器的市占率几乎为0。2023年8月29日，华为终端宣布“HUAWEI Mate 60 Pro先锋计划”，时隔3年再次采用麒麟旗舰芯片，3Q23海思手机处理器市占率开始逐渐恢复，1Q25回升至4%。在2025年9月4日召开的折叠屏手机Mate XTs发布会上，华为现场官宣麒麟9020系列芯片，这是自2021年以来，华为首次公开新款麒麟芯片的有关消息。另外，在华为全联接大会2025上，华为首次对外公布昇腾芯片未来三年的产品迭代路线图，预计2026年第一季度推出昇腾950PR，四季度推出昇腾950DT，2027年四季度推出昇腾960，2028年四季度推出昇腾970。

图：全球手机处理器季度市占率



资料来源：Counterpoint，国信证券经济研究所整理

6.2 晶圆代工：受益国产芯片设计企业崛起和在地化制造趋势

国信证券 GUOSEN SECURITIES

● **中芯国际稳居全球晶圆代工第三，华虹稳居第六。**根据TrendForce的数据，公司在1Q24首次超越格芯和联电，成为全球第三大晶圆代工企业，仅次于台积电和三星；之后市占率一直保持在全球第三，2Q25市占率为5.1%。除中芯国际外，华虹稳居全球晶圆代工第第六，合肥晶合也进入全球前十。

表：全球前十大晶圆代工企业市占率排名

	4Q23		1Q24		2Q24		3Q24		4Q24		1Q25		2Q25	
1	台积电	61.2%	台积电	61.7%	台积电	62.3%	台积电	64.7%	台积电	67.1%	台积电	67.6%	台积电	70.2%
2	三星	11.3%	三星	11.0%	三星	11.5%	三星	9.1%	三星	8.1%	三星	7.7%	三星	7.3%
3	格芯	5.8%	中芯国际	5.7%	中芯国际	5.7%	中芯国际	6.0%	中芯国际	5.5%	中芯国际	6.0%	中芯国际	5.1%
4	联电	5.4%	联电	5.7%	联电	5.3%	联电	5.1%	联电	4.7%	联电	4.7%	联电	4.4%
5	中芯国际	5.2%	格芯	5.1%	格芯	4.9%	格芯	4.8%	格芯	4.6%	格芯	4.2%	格芯	3.9%
6	华虹集团	2.0%	华虹集团	2.2%	华虹集团	2.1%	华虹集团	2.7%	华虹集团	2.6%	华虹集团	2.7%	华虹集团	2.5%
7	高塔半导体	1.1%	高塔半导体	1.1%	高塔半导体	1.1%	高塔半导体	1.0%	高塔半导体	1.0%	世界先进	1.0%	世界先进	0.9%
8	力积电	1.0%	力积电	1.0%	世界先进	1.0%	世界先进	1.0%	世界先进	0.9%	高塔半导体	0.9%	高塔半导体	0.9%
9	合肥晶合	1.0%	合肥晶合	1.0%	力积电	1.0%	力积电	0.9%	合肥晶合	0.9%	合肥晶合	0.9%	合肥晶合	0.8%
10	世界先进	1.0%	世界先进	1.0%	合肥晶合	0.9%	合肥晶合	0.9%	力积电	0.8%	力积电	0.9%	力积电	0.8%

资料来源：TrendForce，国信证券经济研究所整理

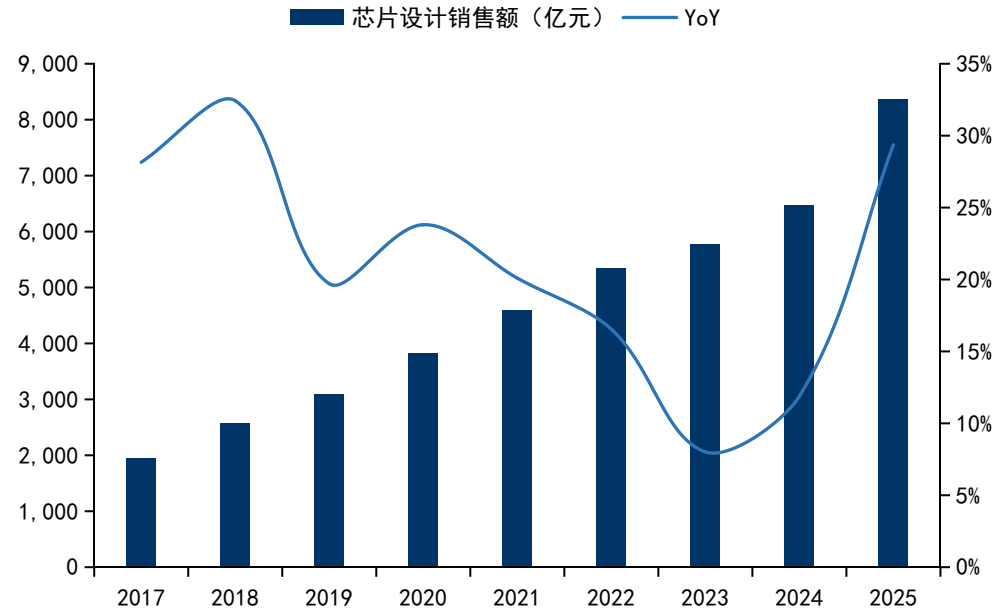
6.2 晶圆代工：受益国产芯片设计企业崛起和在地化制造趋势

国信证券 GUOSEN SECURITIES

● **2017-2025年中国芯片设计企业数量和销售额均以两位数CAGR增长。**中国芯片设计企业数量由2017年的1380家增长至2025年的3901家，CAGR 14%，其中销售额过亿的企业数量由2017年的191家增长至2025年的831家，CAGR 20%。从销售额来看，2017年为1946亿元，2024年增至6460亿元，CAGR 19%，高于全球半导体销售额同期6%的CAGR。

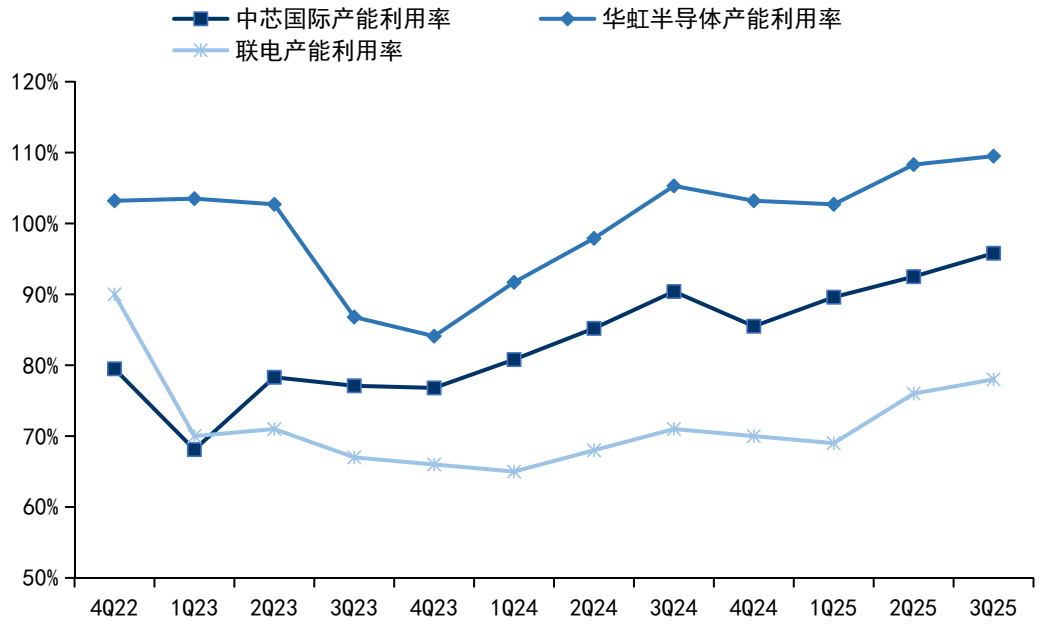
● **中芯国际和华虹半导体产能利用率高于联电。**2022年半导体行业周期下行，中芯国际、华虹半导体、联电等晶圆代工厂的产能利用率均下降，但中芯国际和华虹半导体的产能利用率早于联电触底回升，且长期高于联电。我们认为，这主要得益于中国芯片设计企业的崛起和制造本土化趋势。

图：中国芯片设计销售额及增速



资料来源：ICCAD，国信证券经济研究所整理

图：中芯国际、华虹半导体、联电季度产能利用率



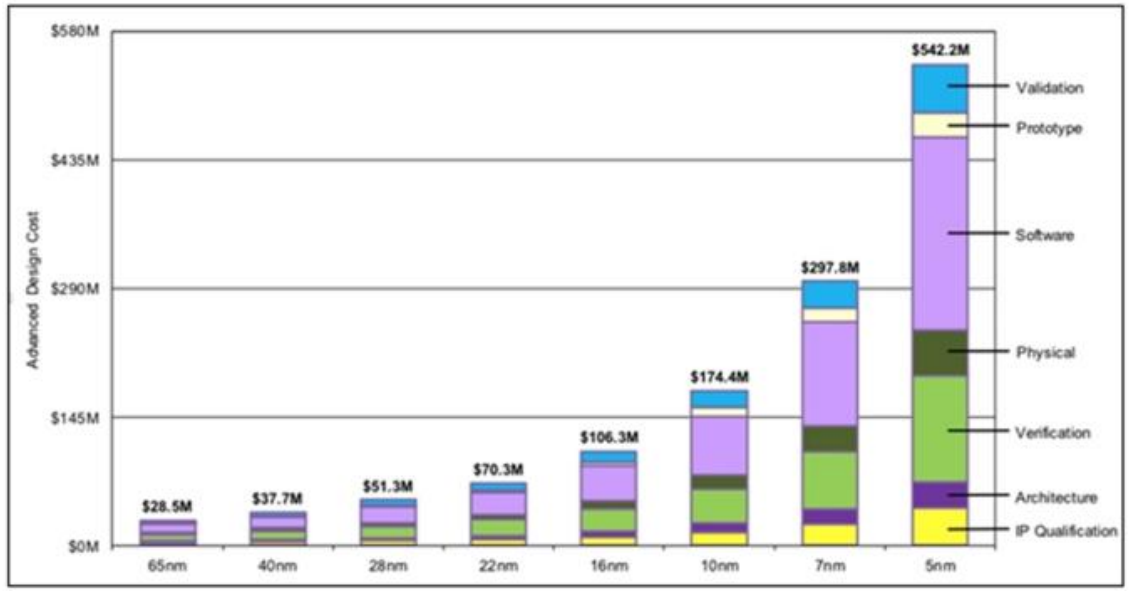
资料来源：各公司公告，国信证券经济研究所整理

6.3 先进封装：AI加速的核心驱动力

- **先进制程的成本快速提升且接近物理极限，先进封装获重视。**随着工艺制程进入10nm以下，芯片设计成本快速提高。根据International Business Strategies （IBS）的数据，16nm工艺的芯片设计成本为1.06亿美元，5nm增至5.42亿美元。同时，由于先进制程越来越接近物理极限，摩尔定律明显放缓，侧重封装技术的More than Moore路径越来越被重视。

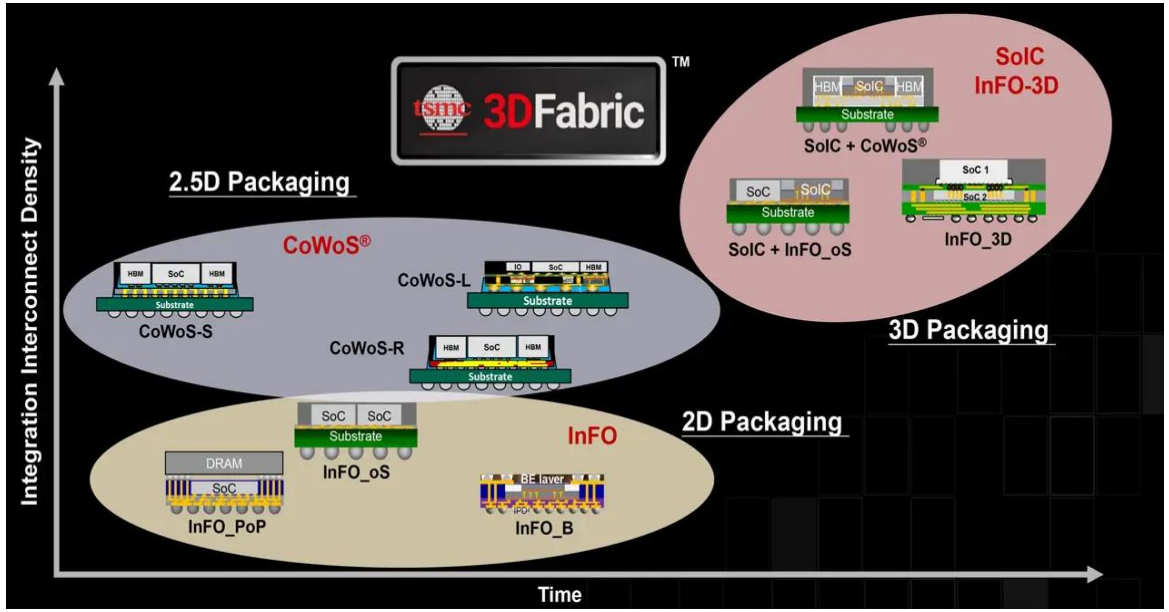
- **台积电早已入局先进封装，近年约10%资本开支主要用于先进封装。**台积电在追求先进制程的同时，早在2008年便成立集成互连与封装技术整合部门入局先进封装，目前已形成CoWoS、InFO、SoIC技术阵列。近年来，台积电每年资本开支中约10%投入先进封装、测试、光罩等。

图：芯片设计成本随着先进制程快速提升



资料来源：IBS，国信证券经济研究所整理

图：台积电先进封装技术



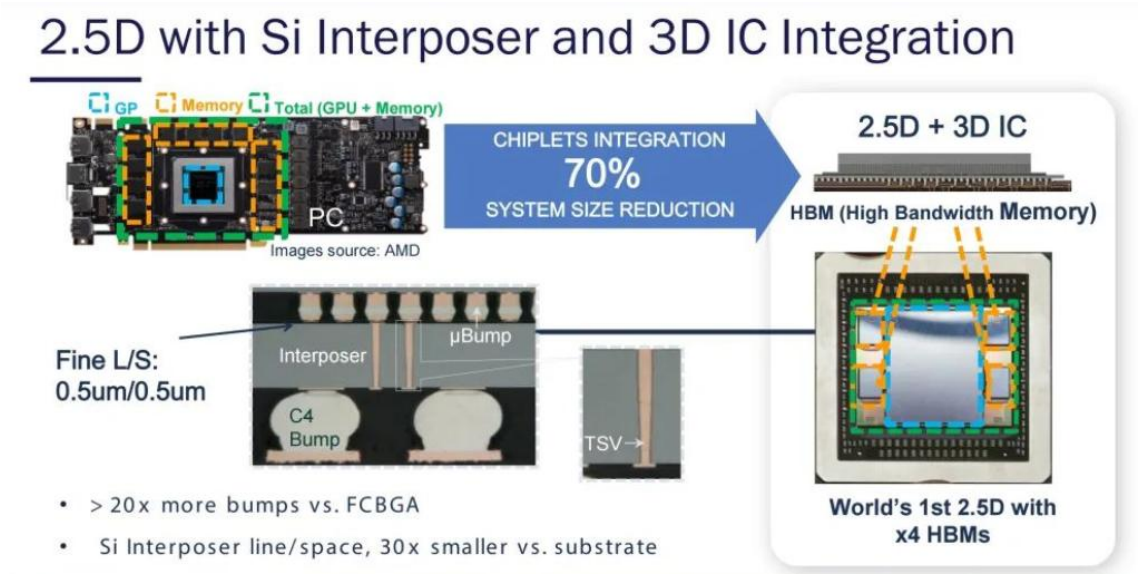
资料来源：台积电，国信证券经济研究所整理

6.3 先进封装：AI加速的核心驱动力

- **2.5D（含硅中介层）与3D IC集成技术带来了内存与计算的深度融合。**该技术通过微凸点（μBump）、硅中介层（Interposer）和C4凸点实现连接，其精细线宽/线距（L/S）可达0.5um/0.5um。与传统的倒装芯片球栅阵列（FCBGA）相比，凸点数量增加20倍以上；硅中介层的线宽/线距比基板小30倍，能实现系统尺寸减少70%，大幅提升集成密度与性能。

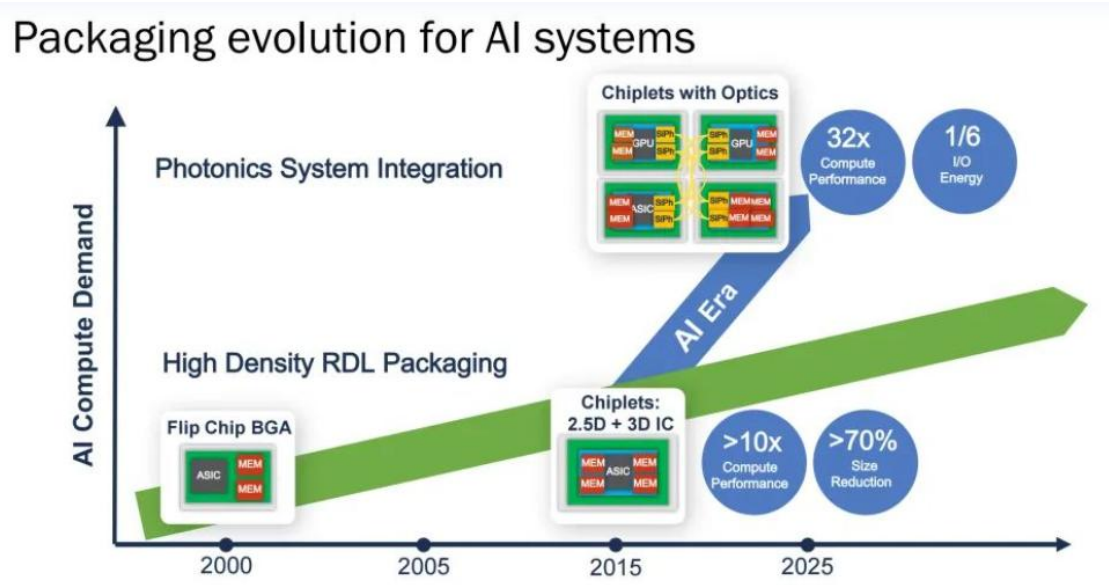
- **异构集成助力AI系统尺寸缩减和性能提高。**相比于单芯片封装和SoC，多芯片的异构集成封装可以更好的优化性能、成本和效率。2000年至2025年，AI系统封装经历了从传统芯片级到2.5D+3D IC集成的演进，系统尺寸减少70%以上且计算性能提升10倍以上。未来，随着光子chiplet的出现，AI系统封装将朝着32倍计算性能提升及1/6 IO能耗的方向演进。

图：2.5D和3D IC集成技术



资料来源：ASE，国信证券经济研究所整理

图：AI系统的封装演进

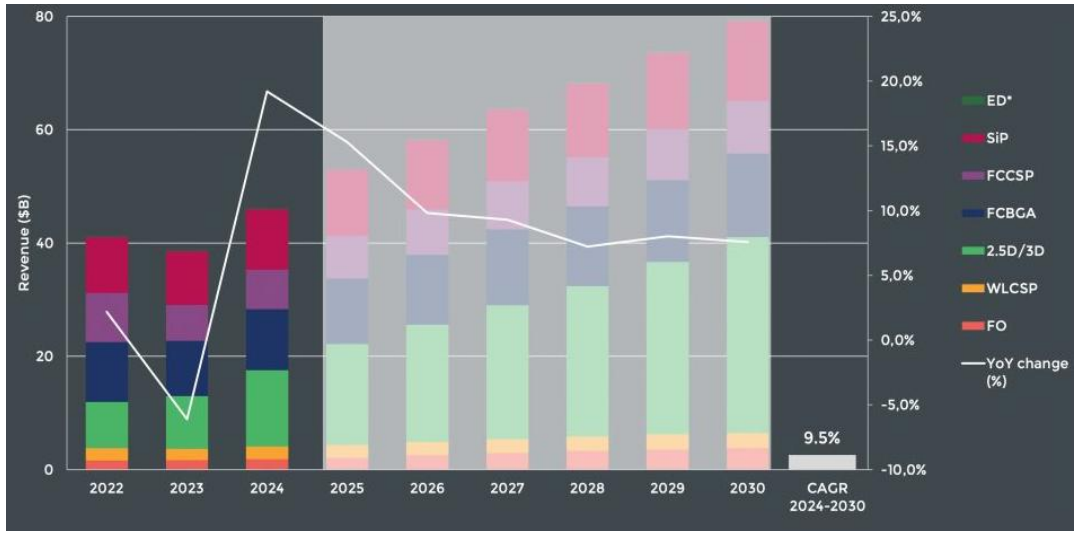


资料来源：ASE，国信证券经济研究所整理

6.3 先进封装：AI加速的核心驱动力

- **预计2030年先进封装市场规模达794亿美元，2024-2030年CAGR为9.5%。**根据Yole的预测，2024-2030年全球先进封装市场规模将以9.4%的CAGR持续增长，到2030年达到约800亿美元，主要由AI和高性能需求推动。其中通信和基础设施是增长最快的市场，CAGR达14.9%，主要由AI加速器、GPU、芯粒驱动。
- **全球前十大先进封装企业中两家来自中国第三方封测企业。**根据Yole的数据，2024年全球前十大先进封装玩家主要由IDM厂商（英特尔、索尼、三星、SK海力士、长江存储）、OSAT（安靠、日月光、长电科技、通富微电）以及晶圆代工厂（台积电）构成。
- **中国封测企业具有国际竞争力，随着更多本土高端芯片设计企业的崛起，先进封装测试的需求和能力将进一步提高。**

图：全球先进封装市场规模预测



资料来源：Yole，国信证券经济研究所整理

图：2024年全球前十大先进封装玩家

资料来源：Yole，国信证券经济研究所整理

6.4 长存长鑫上市有望扩产提速，海外限制加强国产替代需求

国信证券 GUOSEN SECURITIES

- BIS对华半导体出口管制规则三次调整

➢ 2022年10月7日（第一轮半导体制裁）

美国BIS出台“1007”出口管制规则，管制措施适用于将美国设备或零部件出口到中国国内的特定先进逻辑或存储芯片晶圆厂，主要是16/14nm以下节点的逻辑集成电路、128层以上的NAND存储器集成电路、18nm及以下的DRAM集成电路。

同时美国商务部在UVL清单中添加了更多公司，包括31家中国实体。此外，美国商务部还修改了UVL规则，如果不配合现场核实，将列入实体清单。其中，北方华创、长江存储、先导先进、佛山华国光学器材、中国科学院化学研究所等相关公司在列。

➢ 2023年10月17日（第二轮半导体制裁）

美国BIS发布“1017”出口管制新规，对“1007”规则进行调整、细化和补充，在芯片制造设备管制中增加具体性能参数阈值，并通过直接产品规则的适用，使得非美国制造的相关设备可能同样受到EAR管控。

➢ 2024年12月2日（第三轮半导体制裁）

美国BIS公布的最新一轮对华半导体出口管制新规，具体包含1）对24种半导体制造设备、3种半导体软件工具的管制；2）对HBM的新管制；3）合规和转移问题的新规；4）新增140个实体清单和14项修改等。

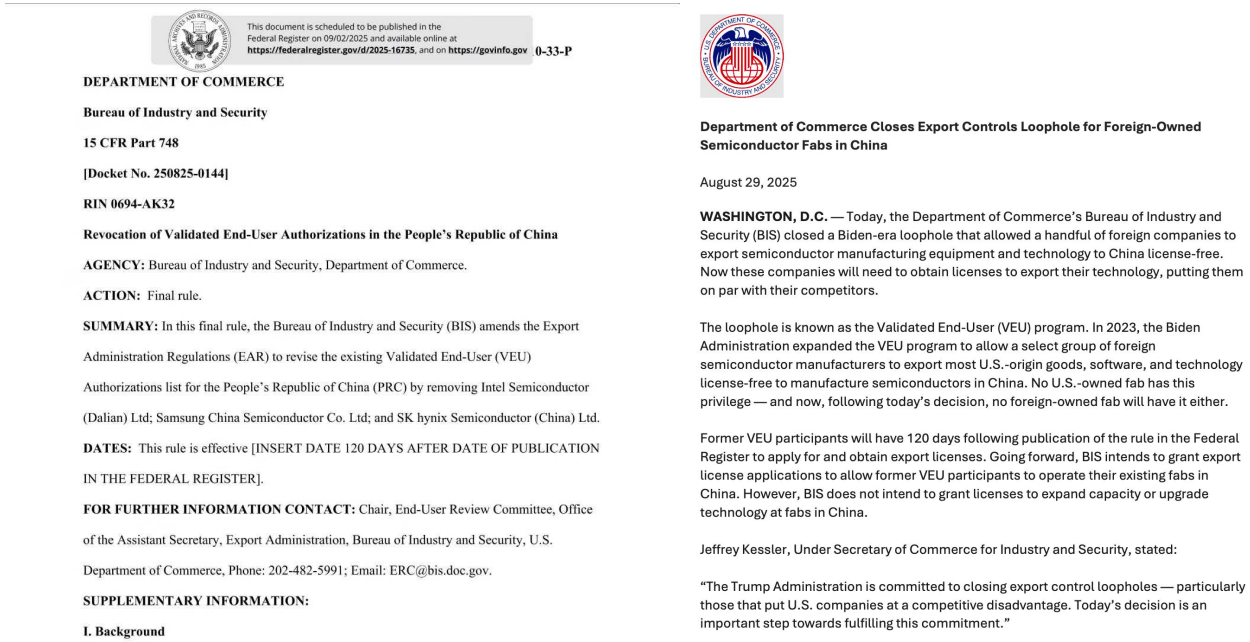
新增进入实体清单的公司包括设备厂（北方华创、盛美上海、中科飞测、新凯来、凯世通、华峰测控、北京烁科、华海清科、芯源微等）、晶圆厂（青岛芯恩、鹏芯旭、昇维旭、闻泰科技等）、零部件材料公司（至纯科技、南大光电等）、投资公司（江淮资本、智路资本、建广资产等）、张江实验室、以及EDA公司华大九天等。

6.4 长存长鑫上市有望扩产提速，海外限制加强国产替代需求

国信证券 GUOSEN SECURITIES

- 20250829：美国BIS修订EAR以修改VEU授权清单，删除英特尔半导体(大连)、三星中国半导体、SK海力士半导体(中国)。
- 20250902：美国终止台积电南京工厂的最终用户(VEU)地位，撤销台积电授权将基本设备免费运送到其中国芯片制造基地。
- 20251007：美国众议院“中美战略竞争特别委员会”调查指出，美国及其盟国的5大设备公司2024年为中国提供了价值380亿美元的半导体制造设备，提出九项政策建议，包括1）统一美国及盟友的出口管制。2）加强对华全面出口管制。3）扩大“实体清单”范围。4）防止设备转用与规避。5）限制全球 fab 使用中国设备。6）限制向中国出口对半导体制造设备生产至关重要的零部件等。我们认为海外半导体设备&材料限制进一步趋严，为国产供应链的份额提升提供机遇，建议关注：北方华创、中微公司、拓荆科技、神工股份、鼎龙股份、江丰电子等。

图：美国BIS修改VEU授权清单



This document is scheduled to be published in the Federal Register on 09/02/2025 and available online at https://federalregister.gov/d/2025-16735, and on https://govinfo.gov

0-33-P

DEPARTMENT OF COMMERCE

Bureau of Industry and Security

15 CFR Part 748

[Docket No. 250825-0144]

RIN 0694-AK32

Revocation of Validated End-User Authorizations in the People's Republic of China

AGENCY: Bureau of Industry and Security, Department of Commerce.

ACTION: Final rule.

SUMMARY: In this final rule, the Bureau of Industry and Security (BIS) amends the Export Administration Regulations (EAR) to revise the existing Validated End-User (VEU) Authorizations list for the People's Republic of China (PRC) by removing Intel Semiconductor (Dalian) Ltd; Samsung China Semiconductor Co. Ltd; and SK hynix Semiconductor (China) Ltd.

DATES: This rule is effective [INSERT DATE 120 DAYS AFTER DATE OF PUBLICATION IN THE FEDERAL REGISTER].

FOR FURTHER INFORMATION CONTACT: Chair, End-User Review Committee, Office of the Assistant Secretary, Export Administration, Bureau of Industry and Security, U.S. Department of Commerce, Phone: 202-482-5991; Email: ERC@bis.doc.gov.

SUPPLEMENTARY INFORMATION:

I. Background

Department of Commerce Closes Export Controls Loophole for Foreign-Owned Semiconductor Fabs in China

August 29, 2025

WASHINGTON, D.C. — Today, the Department of Commerce's Bureau of Industry and Security (BIS) closed a Biden-era loophole that allowed a handful of foreign companies to export semiconductor manufacturing equipment and technology to China license-free. Now these companies will need to obtain licenses to export their technology, putting them on par with their competitors.

The loophole is known as the Validated End-User (VEU) program. In 2023, the Biden Administration expanded the VEU program to allow a select group of foreign semiconductor manufacturers to export most U.S.-origin goods, software, and technology license-free to manufacture semiconductors in China. No U.S.-owned fab has this privilege — and now, following today's decision, no foreign-owned fab will have it either.

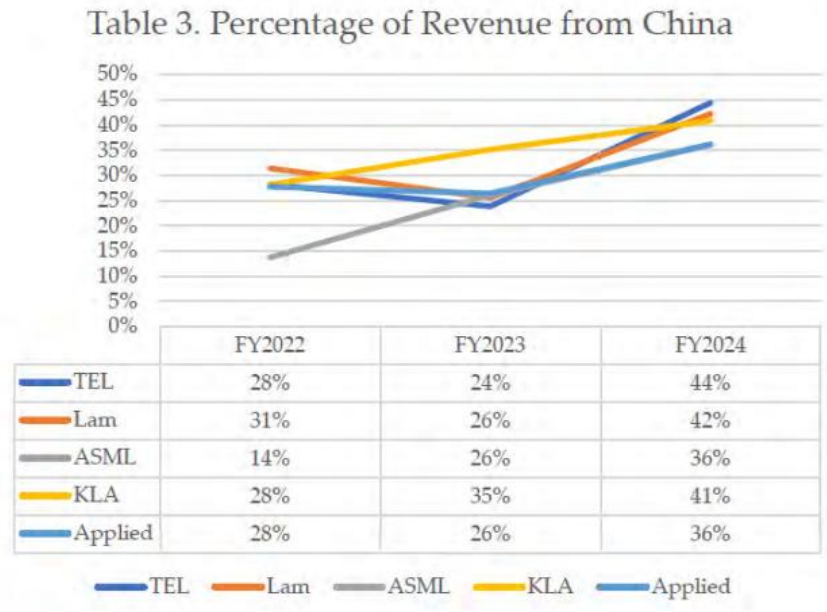
Former VEU participants will have 120 days following publication of the rule in the Federal Register to apply for and obtain export licenses. Going forward, BIS intends to grant export license applications to allow former VEU participants to operate their existing fabs in China. However, BIS does not intend to grant licenses to expand capacity or upgrade technology at fabs in China.

Jeffrey Kessler, Under Secretary of Commerce for Industry and Security, stated:

"The Trump Administration is committed to closing export control loopholes — particularly those that put U.S. companies at a competitive disadvantage. Today's decision is an important step towards fulfilling this commitment."

资料来源：BIS官网，国信证券经济研究所整理

图：美国及盟国5大半导体设备公司中国收入占比



	FY2022	FY2023	FY2024
TEL	28%	24%	44%
Lam	31%	26%	42%
ASML	14%	26%	36%
KLA	28%	35%	41%
Applied	28%	26%	36%

资料来源：美国国会官网，国信证券经济研究所整理

6.4 长存长鑫上市有望扩产提速，海外限制加强国产替代需求

国信证券 GUOSEN SECURITIES

- **长鑫上市辅导完成，长存三期成立，两长扩产有望提速**。7月7日，根据证监会披露报告，长鑫科技首次公开发行股票并上市辅导工作完成。长存方面，9月5日，长存三期（武汉）集成电路有限责任公司成立，注册资本207.2亿元，其中长江存储持股50.19%，湖北长晟三期投资持股49.81%。9月25日，长江存储科技控股有限责任公司（长存集团）召开股份公司成立大会并选举首届董事会，标志其股份制改革全面完成。伴随长存集团股份制改革全面完成，长存三期成立扩产蓄势待发，而长鑫科技加速上市也将助力其后续扩产。当前长存在半导体设备方面已经实现了较高的国产化率，而长鑫国产化率仍有较大提升空间，后续两长扩产及国产化率提升有望推动国产半导体设备、材料公司订单进一步提增长，建议关注：北方华创、中微公司、拓荆科技、神工股份、鼎龙股份、江丰电子等。

图：长鑫科技完成上市辅导工作

资料来源：中国证券监督管理委员会网上办事服务平台，国信证券经济研究所整理

图：长存三期集成电路有限责任公司成立

统一社会信用代码	91420100MAEU35CE09	企业名称	长存三期（武汉）集成电路有限责任公司		
法定代表人	陈南翔	经营状态	在营	成立日期	2025-09-05
注册资本	2072000万元人民币	实缴资本	-	参保人数	-
组织机构代码	MAEU35CE0	工商注册号	420199110037989	纳税人识别号	91420100MAEU35CE09
企业类型	其他有限责任公司	营业期限	2025-09-05 至 长期	核准日期	2025-09-05
所属地区	湖北省武汉市洪山区	所属行业	-		
登记机关	武汉东湖新技术开发区市场监督管理局	英文名	-		
注册地址	湖北省武汉市东湖新技术开发区科技五路261号未来馆三楼				
通信地址	-				
经营业务范围	一般项目：集成电路制造；集成电路销售；集成电路设计；集成电路芯片及产品销售；集成电路芯片及产品制造；技术服务、技术开发、技术咨询、技术交流、技术转让、技术推广；货物进出口；技术进出口；进出口代理。（除许可业务外，可自主依法经营法律法规非禁止或限制的项目）				

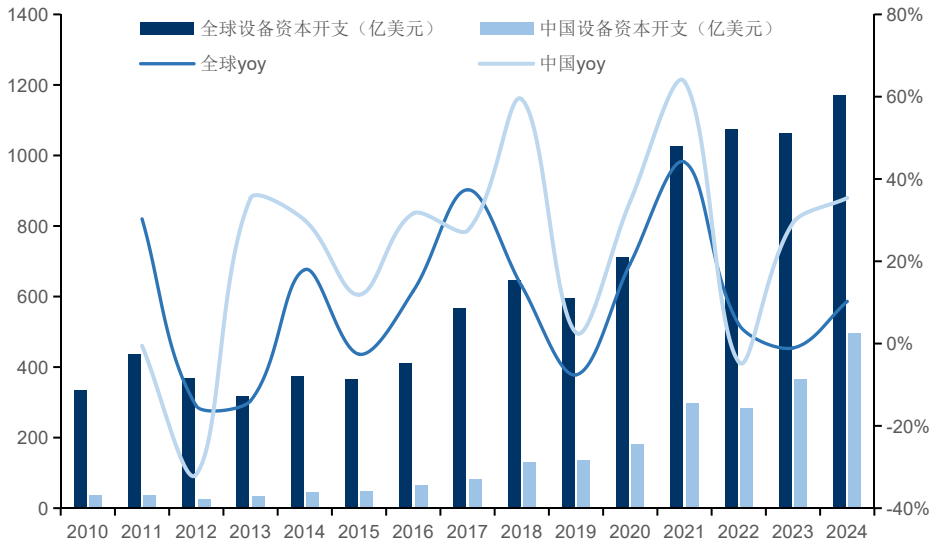
资料来源：风鸟，国信证券经济研究所整理

6.4 长存长鑫上市有望扩产提速，海外限制加强国产替代需求

国信证券 GUOSEN SECURITIES

- **中国半导体设备资本开支高增，光刻机与量检测设备国产化空间广阔。**根据Semi报告数据，2024年全球半导体设备市场规模约为1171.4亿美元（同比增长10%），其中，中国（除港澳台地区）半导体设备市场规模约495.5亿美元（同比增长35%）。半导体设备结构方面，薄膜、刻蚀和光刻设备分别占据了23%、22%、17%的市场，是十大类设备中占比最高的三类。目前，以GTA的12英寸40nm逻辑工艺产线为例，去胶设备、湿法、干法刻蚀设备均已达到较高国产化率，分别为100%、90%、88%。而光刻、量检测设备国产化率最低，分别为0%和11%，具有较大的国产替代空间。

图：全球及中国半导体设备资本开支对比



资料来源：Semi，国信证券经济研究所整理

表：以12寸40nm逻辑工艺半导体设备国产化情况

国产设备应用状态（按12寸40nm逻辑工艺要求）		工艺种类	A（量产）	B（样机）	C（研发）	国产化占比
图形生成	曝光机	4	0	3	1	0%
	涂胶显影	4	3	1		75%
减材设备	干法刻蚀	8	7	1		88%
	湿法刻蚀	10	9	1		90%
	研磨抛光	4	3	1		75%
	去胶设备	1	1			100%
增材设备	氧化沉积炉	9	5	1	3	56%
	化学气相沉积（CVD）	15	7	8		47%
	物理气相沉积（PVD）	6	2	4		33%
改性设备	离子注入	3	1	2		33%
	热处理	3	2	1		67%
量测设备	测量检测	9	1	5	3	11%

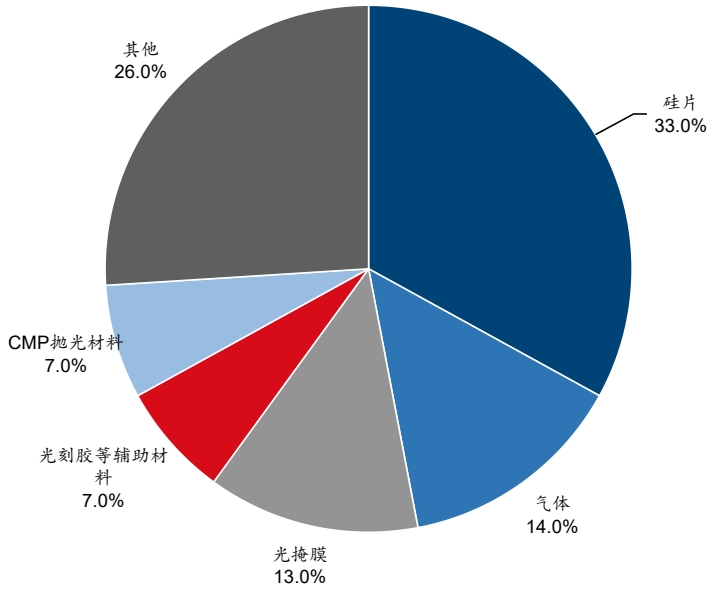
资料来源：GTA，国信证券经济研究所整理

6.4 长存长鑫上市有望扩产提速，海外限制加强国产替代需求

国信证券 GUOSEN SECURITIES

● **半导体材料国产化率普遍较低，光刻胶与湿化学品等为主要替代方向。**根据观研报告网数据，2023年全球半导体材料销售额为667亿美元，同比下降约8.2%。中国除港澳台地区为2022-2023年全球半导体材料销售额唯一正增长地区，同比增长约0.9%。当前国内半导体整体以进口为主，自2022年以来，引线框架、键合丝等半导体封装材料国产化率有较明显的提升，而前道晶圆制造的光刻胶、湿化学品、12英寸硅片等国产化率仍处于10%的较低水平。

图：2022年全球半导体材料市场结构



资料来源：观研天下数据中心，国信证券经济研究所整理

表：2022-2024年半导体材料领域国产化率变化情况

材料名称	2022 年国产化率	2024 年国产化率
晶片	9%	55%（8 英寸）、10%（12 英寸）
光掩模	30%	晶圆厂商自产为主
光刻胶	<5%	10%
电子气体	<5%	15%
湿电子化学品	3%	10%（G3 及以上）
溅射靶材	20%	30%
抛光材料	20%	30%（抛光液）、20%（抛光垫）
引线框架	<30%	40%
封装基板	<20%	<20%
环氧塑封料	-	30%
键合丝	<20%	30%

资料来源：观研天下数据中心，国信证券经济研究所整理

6.4 长存长鑫上市有望扩产提速，海外限制加强国产替代需求

国信证券 GUOSEN SECURITIES

- **芯上微装第500台步进光刻机成功交付，国产光刻机供应链有望持续突破。**2025年8月8日，上海芯上微装举办了第500台步进光刻机交付仪式，此次发运的步进光刻机将交付给盛合晶微。AMIES的先进封装光刻机能够满足Flip-chip、Fan-in、Fan-out WLP/PLP、2.5D/3D等先进封装技术的要求，目前全球市占率达到35%，国内市占率达到90%。11月25日，芯上微装自主研发的首台350nm步进光刻机（AST6200 ）正式完成出厂调试与验收。AST6200采用I-line光源（365nm），搭载大数值孔径投影物镜，实现350nm高分辨率，专为功率、射频、光电子及Micro LED等先进制造场景量身定制。我们认为，伴随AMIES的500台光刻机成功交付，公司在先进封装之外的IC前道等领域也有望积极延伸和拓展，建议关注光刻机供应链相关公司：茂莱光学、波长光电、福晶科技、永新光学、蓝特光学等。

图：芯上微装第500台步进光刻机交付

资料来源：芯上微装官网，国信证券经济研究所整理

图：芯上微装首台350nm步进光刻机发运

资料来源：芯上微装官网，国信证券经济研究所整理

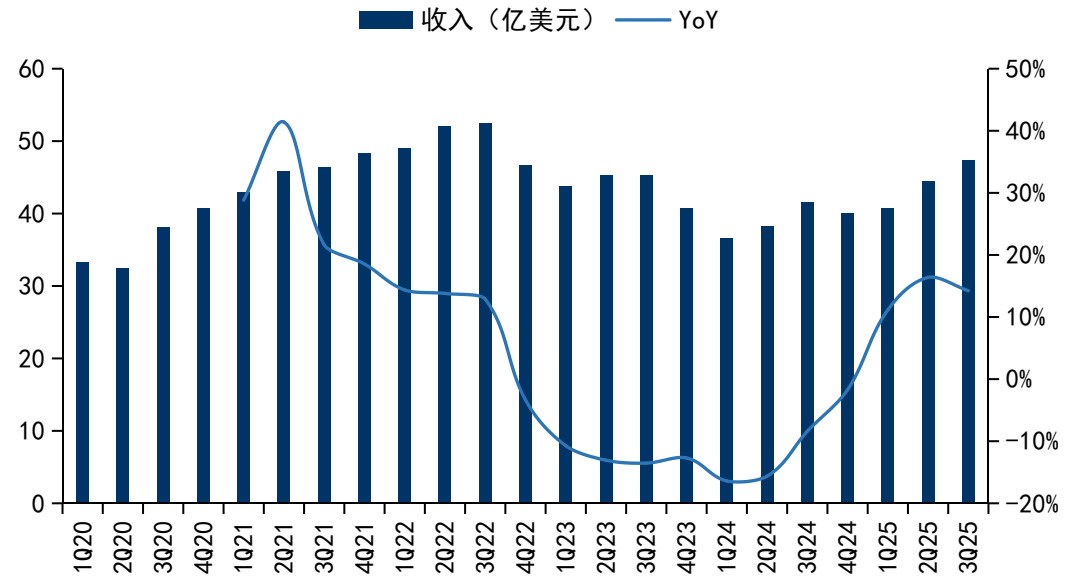
6.5 模拟芯片：去库结束后的新料号上量可期

国信证券 GUOSEN SECURITIES

- **短期去库存结束，周期向上：**TI 1Q25营收在连续九个季度同比下降后首季同比转正，2Q25、3Q25继续同比正增长，其表示当前复苏非常温和，市场正向趋势线回升，客户库存保持在低位。ADI 2QFY25（截至5月3日）营收在连续七个季度同比减少后首季同比转正，已连续三个季度收入同比增长超过20%，其对2026财年及以后的增长充满信心，预计2026财年将看到广泛的增长，所有终端市场都会增长。我们认为模拟芯片行业正处于周期向上阶段，国内企业近几年推出的新产品有望进入规模放量阶段。

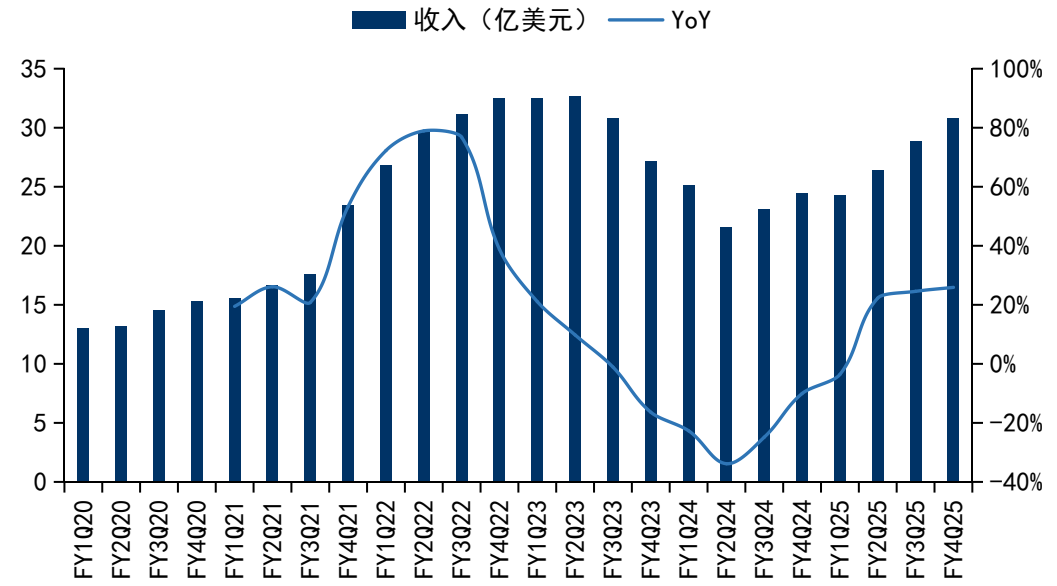
- **长期AI带来增量：**模拟芯片作为基础器件，在电子产品中广泛使用，AI数据中心以及自动驾驶、人形机器人等AI应用均为其带来广泛增量。比如机器人，根据ADI的法说会，机器人技术从固定臂机器人向自主、移动乃至人形机器人发展，将使ADI提供的价值量从数百美元增加到数千美元。

图：TI季度收入



资料来源：Wind，国信证券经济研究所整理

图：ADI季度收入



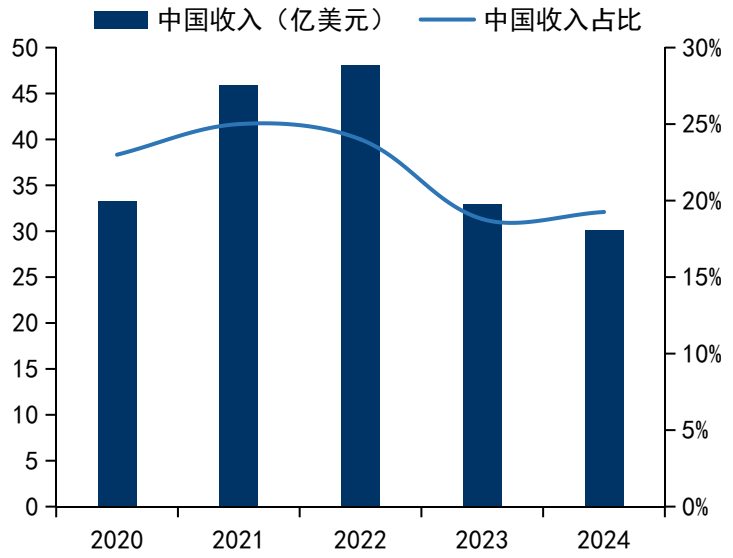
资料来源：Wind，国信证券经济研究所整理

6.5 模拟芯片：去库结束后的新料号上量可期

国信证券 GUOSEN SECURITIES

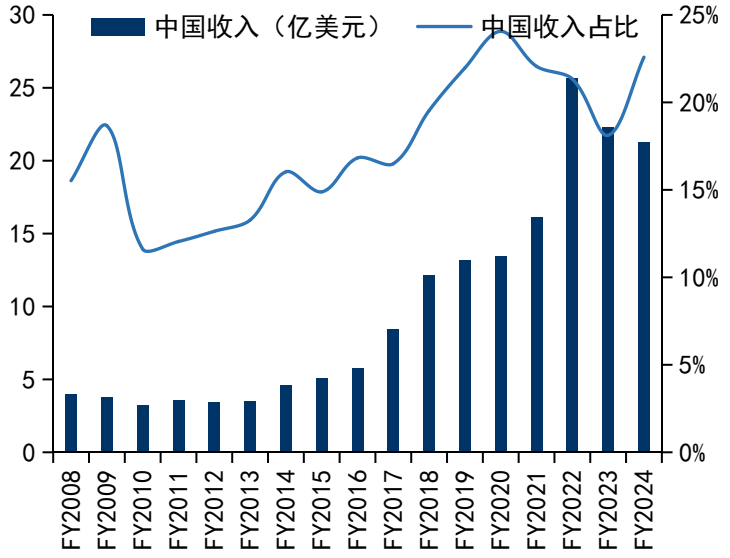
- **我国占全球模拟芯片市场规模的35%，国产化空间较大。**根据WSTS和弗若斯特沙利文的数据，2024年我国占全球模拟芯片市场的35%左右。虽然我国模拟芯片自给率在近年有所提升，但仍然偏低。从竞争格局来看，第一梯队仍然是以TI、ADI等为代表的欧美企业，中国是其收入主要来源地之一。
  - TI 2024年来自中国的收入约**30亿美元**，同比减少9%，收入占比19%（按照终端客户的总部所在地统计）。
  - ADI FY2024年来自中国的收入约**21亿美元**，同比减少5%，收入占比23%（按采购公司产品的分销商或OEM厂商所在地统计）。
  - MPS 2024年来自中国的收入约**12亿美元**，同比增长26%，收入占比53%（按发货地统计）。

图：TI来自中国的收入



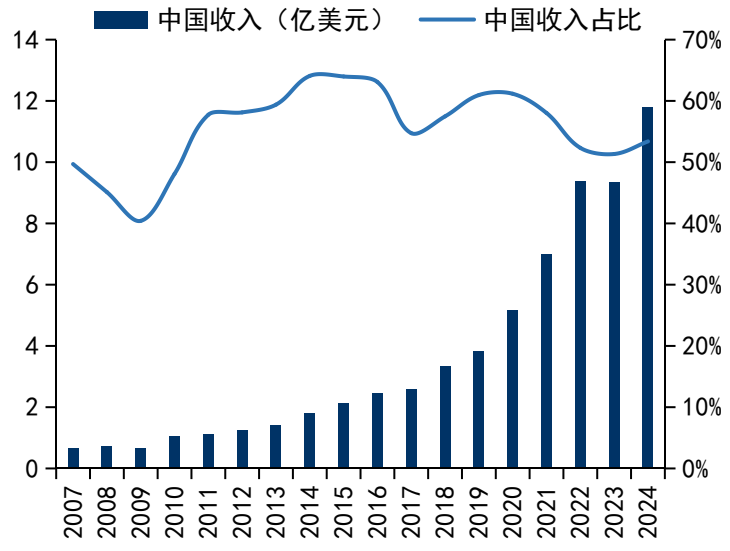
资料来源：TI公告，国信证券经济研究所整理

图：ADI来自中国的收入



资料来源：ADI公告，国信证券经济研究所整理

图：MPS来自中国的收入



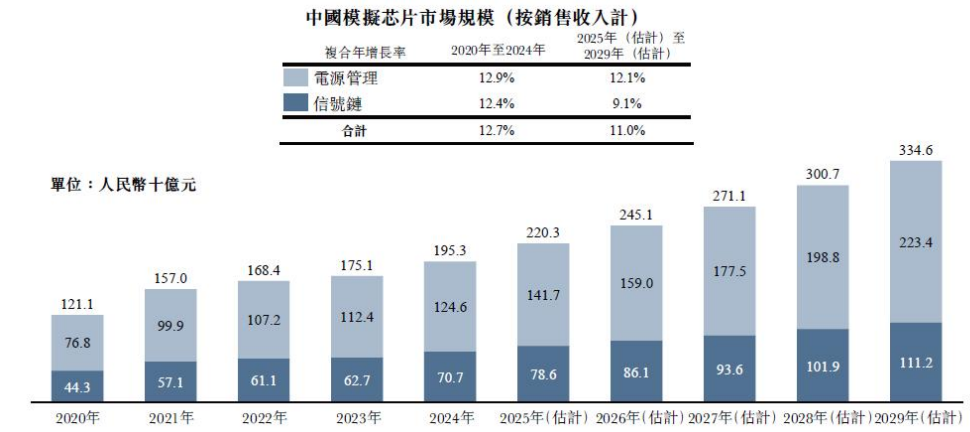
资料来源：MPS公告，国信证券经济研究所整理

6.5 模拟芯片：去库结束后的新料号上量可期

国信证券 GUOSEN SECURITIES

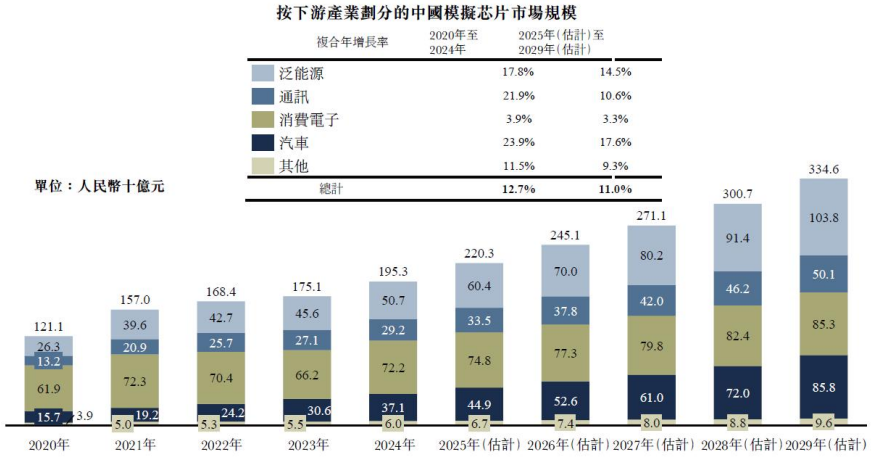
- **国内模拟芯片企业的核心推荐逻辑是【去库周期结束+AI增量+国产化率提高+份额提升+盈利能力改善】。**
  - **工业：**本轮去库周期完成时间靠后的领域，根据TI、ADI法说会，目前工业客户去库存已结束，呈广泛复苏。我们认为，下游去完库存后将恢复正常采购和新产品导入，国内企业近几年大额研发推出的新产品有望规模放量。同时，由于工业领域对模拟芯片的需求呈现小量多种类的特点，且竞争格局相对较好，工业领域的毛利率一般优于消费电子，工业领域的新产品放量，有望提高企业盈利能力。
  - **AI：**AI除了带动模拟芯片整体需求外，AI产业链的国产化也是重点，以多相电源（多相控制器+DrMOS）为代表的核心电源管理芯片既是增量市场，也是国产化的重点和难点。
  - **汽车：**汽车的电动化和智能化为模拟芯片提供增量，国内车厂在电动化和智能化方面布局积极，为国内模拟芯片公司突破汽车领域国产化提供了机会。我们认为，汽车模拟芯片仍是增量市场，且国产化处于起步阶段，前几年相关企业仍是以产品研发和导入为主，目前正进入集中放量阶段。
  - **消费电子：**模拟芯片公司早期多以聚焦个别产品系列为主，随着产品开发能力的提高以及客户关系的加深，围绕手机等应用终端丰富产品系列，为客户提供一站式解决方案成为必然选择，马太效应将愈加明显。

图：中国模拟芯片市场规模



资料来源：弗若斯特沙利文报告，纳芯微公告，国信证券经济研究所整理

图：中国模拟芯片下游结构



资料来源：弗若斯特沙利文报告，纳芯微公告，国信证券经济研究所整理

【7】面板及被动件：面板进入稳定盈利新阶段，AI需求助推被动元件涨价预期蔓延

7.1 被动件行业景气上行，涨价预期从细分单品向全盘蔓延

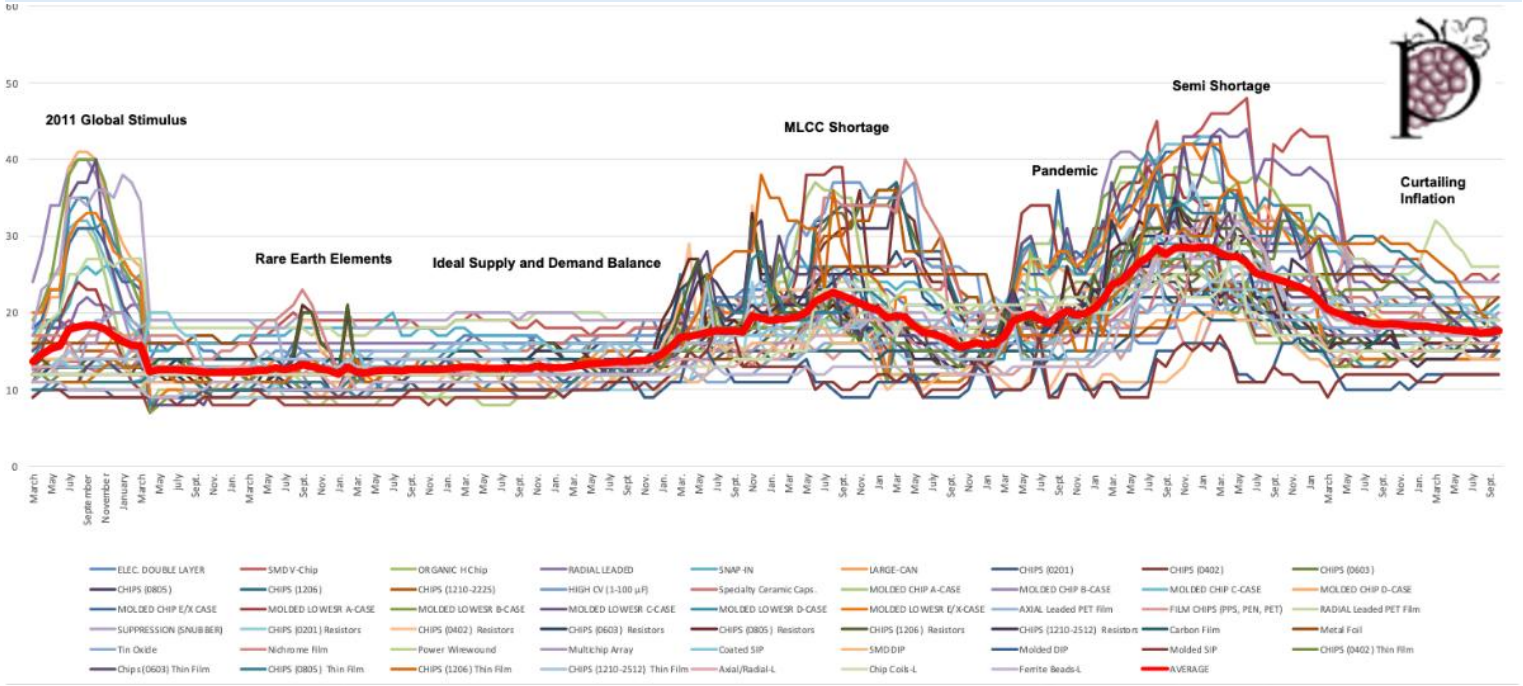
- **回顾前三次涨价原因，MLCC 涨价往往出现在终端结构升级叠加上游产能阶段性错配的窗口期，具有明显的“需求结构驱动 + 供给调整放大”的周期特征，且原材料价格波动在各轮周期中都起到放大波动作用。**

➢ ① 2011-2012年：智能机全球渗透加速，年出货量翻倍增长，同期平板电脑也同比大幅增长，2012年平板电脑同比增长53% + 3·11东日本大地震冲击上游材料与部分被动件产能。

➢ ② 2017-2018年：智能手机单机MLCC用量持续上升，例如iPhoneX整机MLCC用量较前代提升10% + 汽车电动化初始，带动高容量、高耐压MLCC需求大幅增长 + 2018年，Murata、太阳诱电等日系大厂主动将产能从通用型MLCC转向车规、高容、小尺寸高端产品，导致通用品产能被压缩+2016年中以来，被动元件相关原材料价格整体呈上行趋势。

➢ ③ 2020-2021年：疫情导致生产与物流受挫 + 5G 终端与汽车电子需求恢复超预期 + 华新科工厂火灾等个别厂商事故影响阶段性产能 + 重要成本：钯，大幅涨价。

图：53类被动元件交期历史（2012.3－2024.11周度数据）

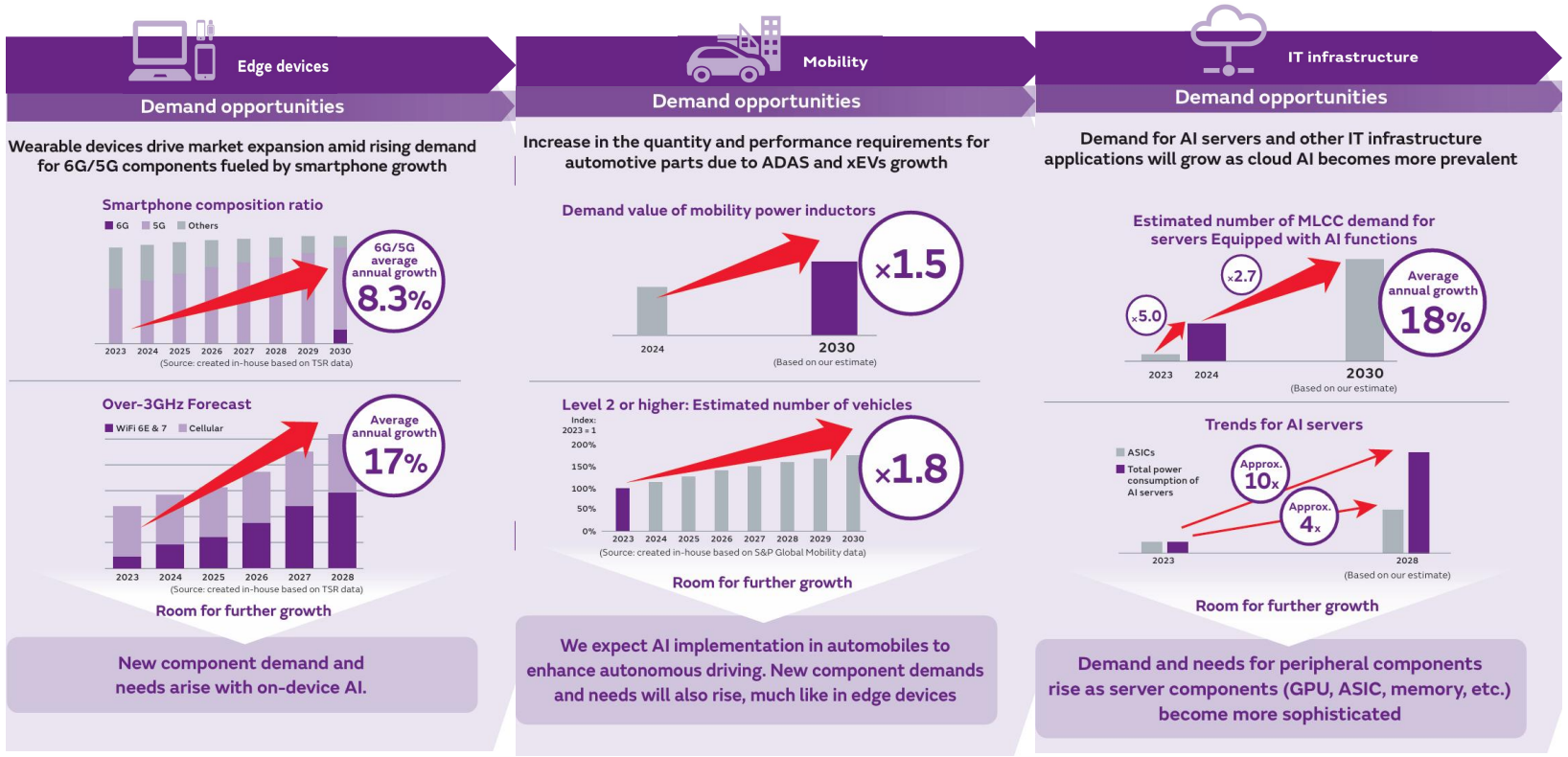


资料来源：欧洲被动元件协会，国信证券经济研究所整理

7.1 创新：汽车、AI算力和端侧是核心驱动力

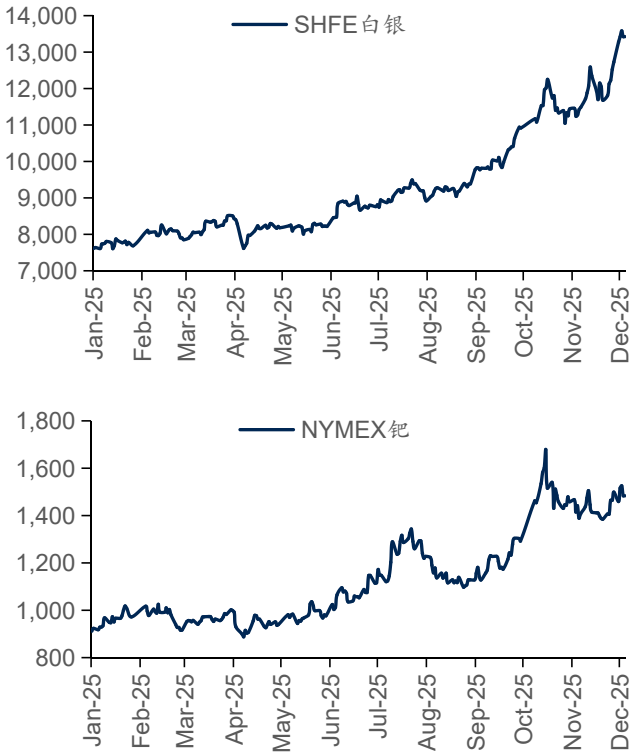
- **我们预计，2026年全球被动元件行业将从“补库存+原材料涨价带动局部涨价”走向“结构性成长与分化”，品类间景气度可能存在差异。**
- 需求方面，AI算力基建和汽车智能化，推动高端产品需求增长，产品向高容、小型化、高可靠（车规/工规）方向升级。叠加上游原材料涨价背景，海外企业有动力将产能向高端品倾斜；
- 供给方面，经历2018与2021两轮大周期后，全球龙头在通用品MLCC、电阻等环节扩产明显趋于谨慎，更聚焦车用、工控与高附加值规格，新一轮供给弹性较过去显著收敛，使得2026年在需求正常偏强的假设下，行业整体更易维持“温和紧平衡”。
- 但部分产品如钽电容、牛角型铝电解电容、超级电容等，需求弹性较大，供应格局清晰，有望出现结构性机会。

图：被动元件的三大市场驱动



资料来源：村田，国信证券经济研究所整理

图：白银、钯期货价格年初至今已翻倍



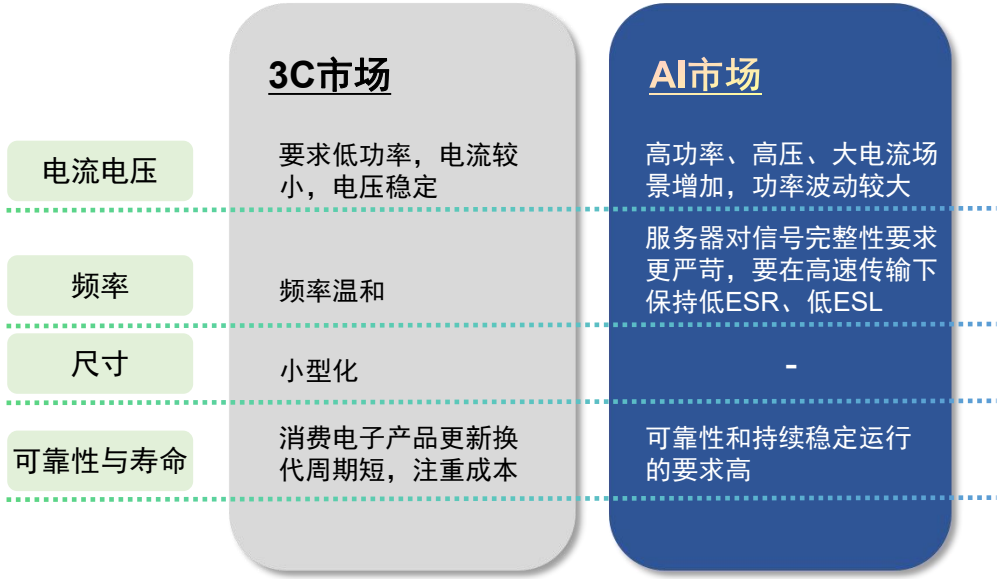
资料来源：Wind，国信证券经济研究所整理

7.1 新应用催生被动元件新料号增量，国内龙头企业抢占窗口期

国信证券 GUOSEN SECURITIES

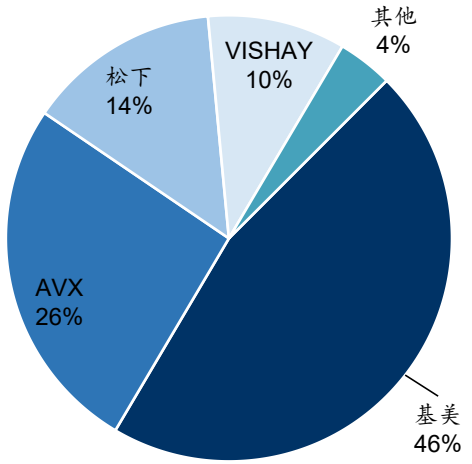
- **AI升级需求与传统3C对被动元件的要求有所变化，从而引入了新的料号或材料。**过去十年，被动元件的主要增量来自于3C类产品，为了保证终端的续航，要求系统整体功耗低，但在服务器中，功耗波动大、峰值高，从而对被动元件提出了新的要求。例如，英伟达在GB200中大量采用了Vishay的vPolyTan聚合物钽电容，导致该产品在2025年的供应可能面临短缺，并预计将在GB300中增加超级电容作为峰值功率的补充电源。
- **钽电容：**因AI伺服器处于高速运算的高温环境，将采用更多的更耐高温的高阶MLCC及钽质电容。目前全球钽电容产量主要被KEMET（占46%）、AVX（占26%）、PANASONIC（占14%）、VISHAY（占10%）这四家巨头垄断。钽电容过去常用于军工、高端消费电子等领域，整体市场维持低速增长。根据TrendForce报告，Yageo内部钽电容器约占收入的23%，推算2024年钽电容收入约为60亿元，市场占比40%，则全市场为150亿人民币。而AI服务器中GPU板和CPU板上都大量使用了钽电容，预计2025年英伟达服务器钽电容采购额将达到30亿元，此外，企业级存储为实现电保护功能，通常会配备大容量的电容，企业级ssd及数据中心级ssd存储量快速攀升，预计2026年相关钽电容需求将达到40亿元。在AI服务器和eSSD推动下，我们预计26年钽电容市场将达到217亿，较2024年增长45%。供给方面，龙头企业过去三年并未大幅扩产，我们预计2026年钽电容将出现供需缺口，国产替代窗口期打开。

图： AI增量市场对被动元件提出新的要求



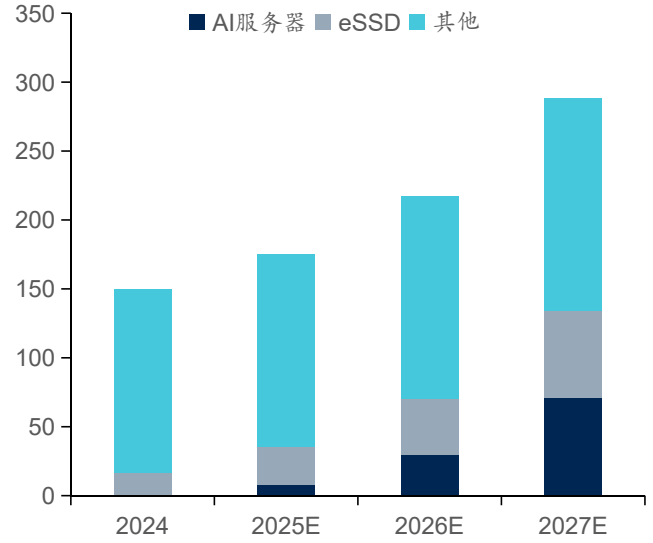
资料来源：国信证券经济研究所整理

图：钽电容市场格局



资料来源：国巨，国信证券经济研究所整理

图：钽电容市场规模测算



资料来源：TrendForce，国信证券经济研究所测算

7.1 超级电容已成AI服务器电源标配

国信证券 GUOSEN SECURITIES

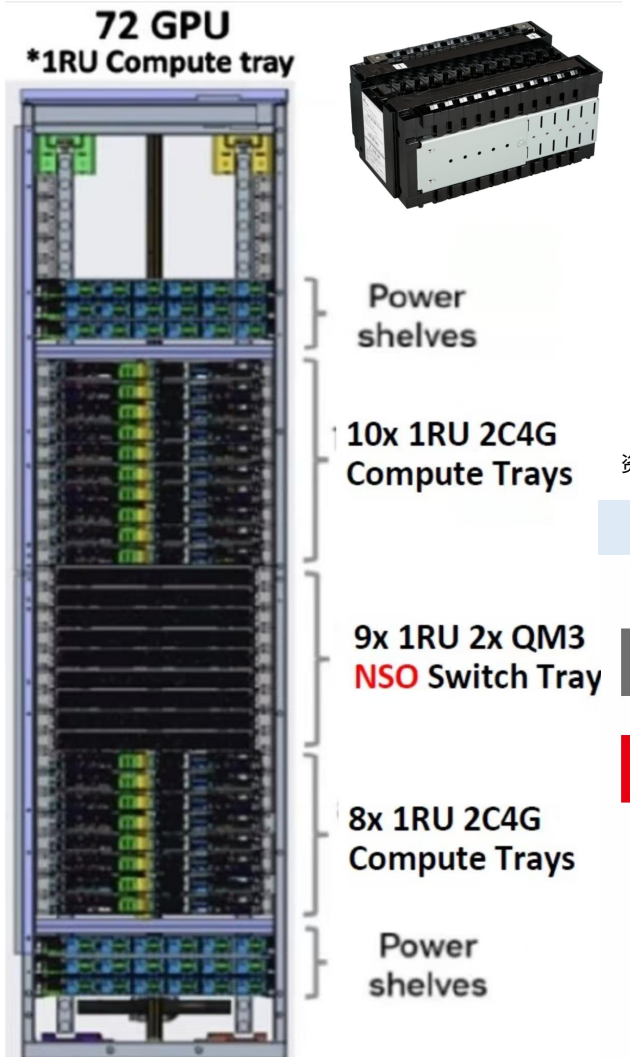
- **超级电容：**超级电容器是一种新型储能装置，以其快速充电、长寿命和高比功率等特性，在储能技术中占据独特地位。在电网调频、AI电源、交运等领域都有所应用，随着超容能量密度提升、成本下降，有望打开更广阔的市场。

- **超级电容是GB300标准电源解决方案的重要组成部分。**超级电容的最大作用是吸收瞬间的电力波动，维持稳定的电源供应。通过实施超级电容器解决方案，AI工作负载的功率波动能够被机架吸收，从而将对数据中心基础设施和电网的影响降到最低。

甲骨文目前正在探索采用Musashi的混合超级电容器解决方案的方法，以帮助甲骨文在全球部署GB200，将现有的65,536个GPU规模提升到131,072个GPU。

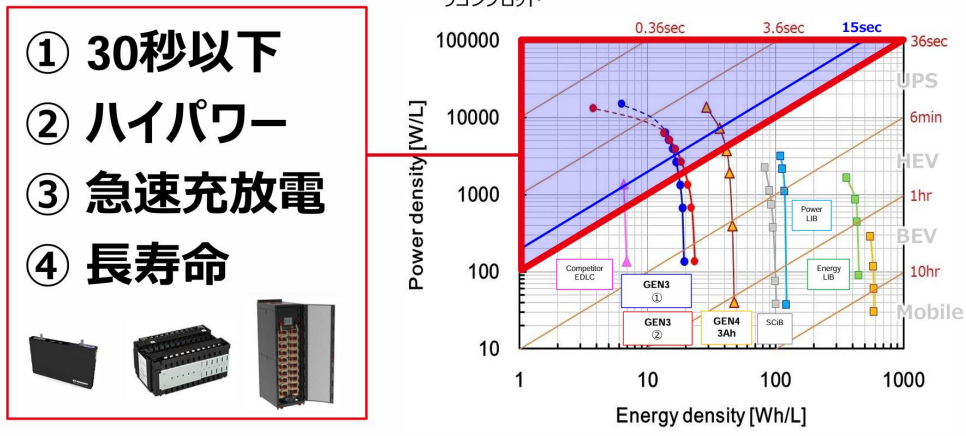
由于锂电池充放电倍率低，难以平抑负载浪涌，且高频次充放电导致使用寿命大大缩短，维护成本过高。而超级电容可以提供瞬时功率补偿及回收，实现毫秒级削峰填谷，平抑负载浪涌。此外，超级电容循环寿命超百万次，使用寿命长，降低生命周期维护成本。我们预计，AI服务器用超级电容，将随着超级电容扩产和技术成熟度提升，实现快速渗透，预计2026年相关市场将达到30亿以上。

图：武藏锂离子超容电芯模组及GB200中的布局



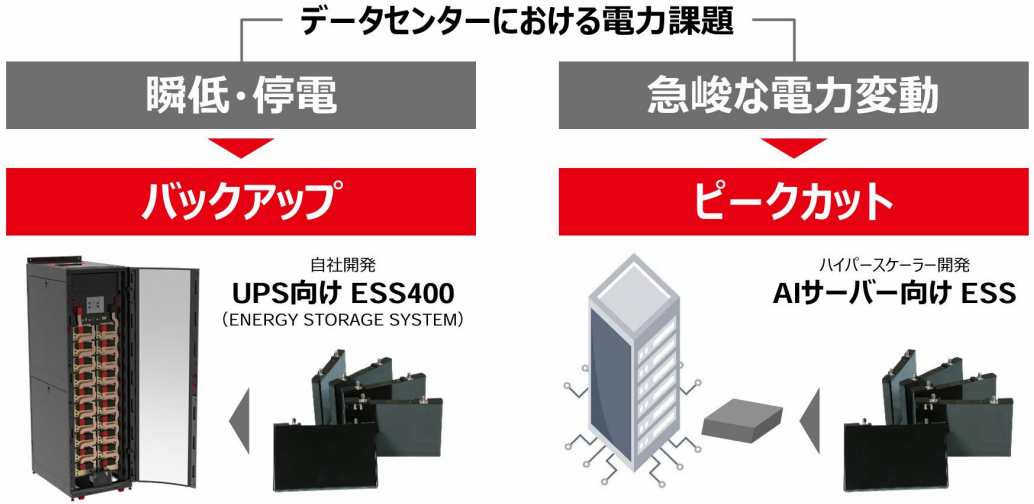
资料来源：日本武藏，国信证券经济研究所整理

图：超级电容目标市场



资料来源：日本武藏，国信证券经济研究所整理

图：数据中心的电力问题



资料来源：日本武藏，国信证券经济研究所整理

7.2 LCD价格：11月各尺寸TV面板价格环比持平

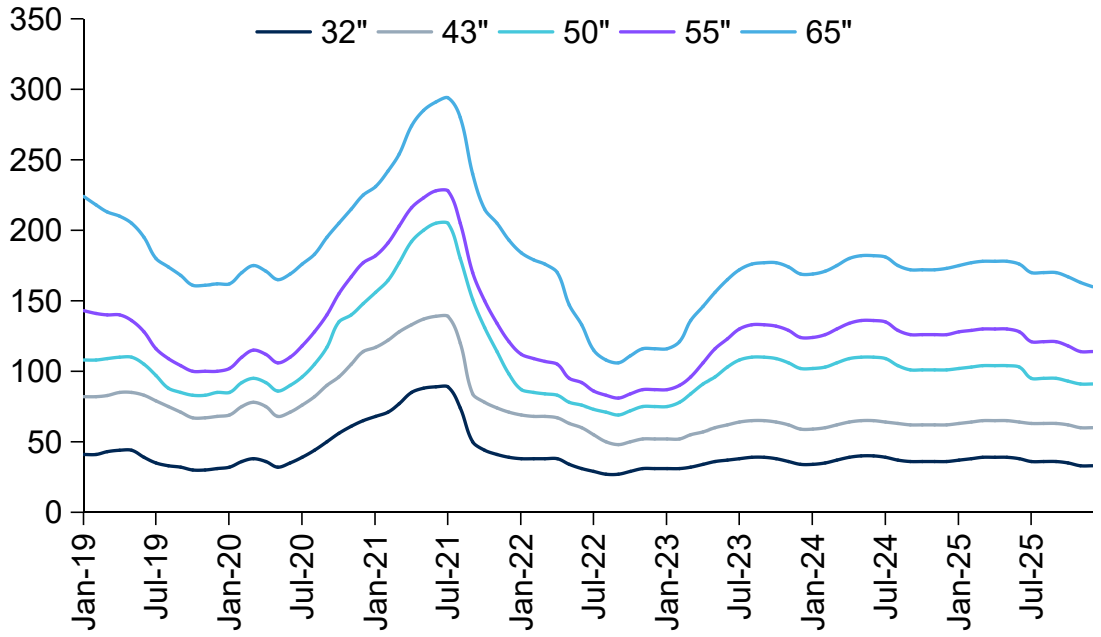
国信证券 GUOSEN SECURITIES

● **我们认为，LCD行业产能逐步稳定，伴随海外厂商产线关停以及海外厂商产线出售，行业份额有望进一步集中，供给端在份额集中的优势下可以较好的调控LCD TV面板价格，行业周期属性逐步淡化，成长属性显现，LCD面板企业的盈利稳定性有望逐步强化。**

● 2025年11月32、43、50、55、65英寸LCD TV面板价格为33、60、91、114、163美元/片，各尺寸LCD TV面板价格较10月分别环比下滑5.7%、3.2%、2.2%、3.4%、2.4%；Omdia预计12月32、43、50、55、65英寸LCD TV面板价格为33、60、91、114、160美元/片，除65英寸环比下滑1.8%外其余尺寸环比保持持平。据TrendForce，四季度是电视面板需求传统淡季，但部分品牌客户仍积极冲刺年底目标，部分品牌客户开始提前备货，带动第四季度面板需求提升。面板厂在持续有订单状态下积极满足客户需求，采用专案价格方式鼓励客户增加拉货。

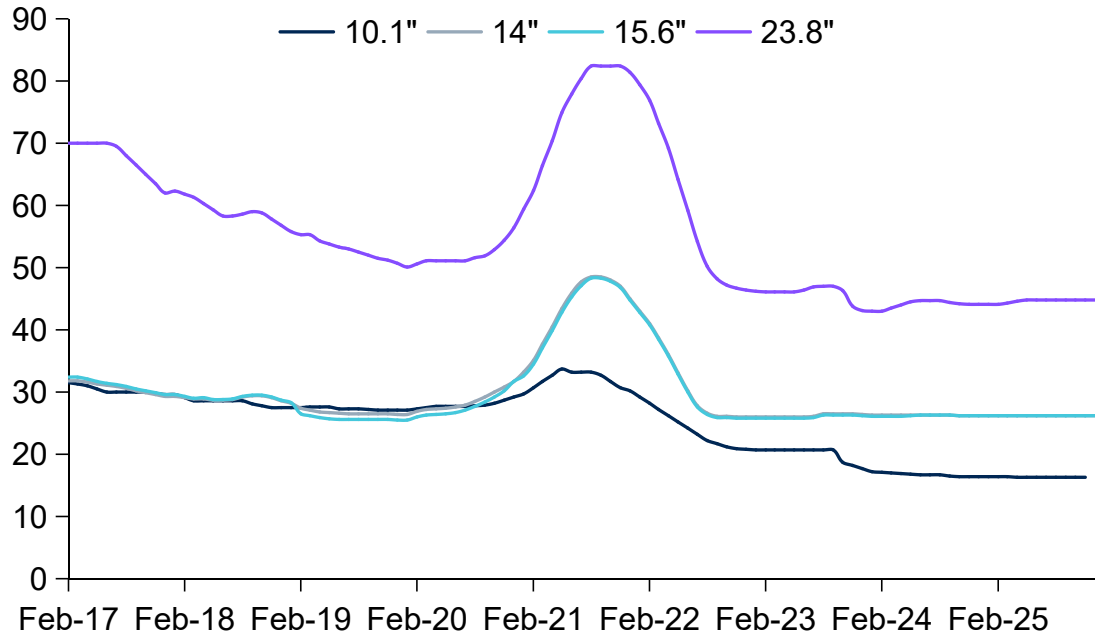
● 2025年11月10.1英寸（平板电脑）、14英寸（笔记本电脑）、23.8英寸（显示器）LCD IT面板价格16.3、26.2、44.8美元，环比持平；Omdia预计12月14英寸、23.8英寸LCD IT面板价格分别为26.2、44.8美元/片，环比持平。

图：32、43、50、55、65英寸LCD TV面板价格走势（单位：美元）



资料来源：Omdia，国信证券经济研究所整理

图：10.1、14、15.6、23.8英寸LCD IT面板价格走势（单位：美元）



资料来源：Omdia，国信证券经济研究所整理

7.2 供给：预计25年全球大尺寸LCD产能同比增长0.21%

国信证券 GUOSEN SECURITIES

- 基于退出产线和新增产线的梳理，我们对全球大尺寸LCD总产能面积进行定量测算。假设产能的新增和减少进度都按照线性完成，全球大尺寸LCD总产能面积情况测算如右图。
- 我们预计2025年全球大尺寸LCD产能面积较2024年增长0.21%，其中1Q25、2Q25、3Q25、4Q25产能分别环比变动-0.02%、-0.02%、0.00%、+0.16%。

表：2023-2025年全球大尺寸LCD供给测算

产线	1Q23	2Q23	3Q23	4Q23	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25
新增（万平米）												
京东方福州8.5代线（B10）	6.19	6.19										
京东方武汉10.5代线（B17）	22.29	22.29	22.29									
华星光电深圳11代线（T6）			7.43	7.43	7.43	7.43						
华星光电深圳11代线（T7）		22.29	22.29	22.29	22.29							
华星光电广州8.6代线（T9）	43.88	43.88	43.88				35.10	35.10	35.10	35.10		
夏普广州10.5代线（SIO）	14.86	14.86	14.86	22.29	22.29	22.29	22.29					
友达台中8.5代线（L8B）		8.25	8.25	8.25	8.25							
群创7代线（ILX Fab 7）	1.65	1.65										
深天马福建8.6代线（TM19）							17.55	17.55	17.55	17.55		13.16
退出（万平米）												
LG坡州7.5代线（P7）	24.68	24.68	24.68	24.68								
三星汤井8.5代线（L8-2）	35.48											
夏普堺市10代线（Sakai）							54.09	54.09	54.09	54.09		
全球大尺寸LCD单季产出（万平米）	8121.34	8216.06	8310.39	8345.97	8406.24	8435.96	8456.82	8455.38	8453.95	8452.51	8452.51	8465.67
环比增长（%）	0.35%	1.17%	1.15%	0.43%	0.72%	0.35%	0.25%	-0.02%	-0.02%	-0.02%	0.00%	0.16%

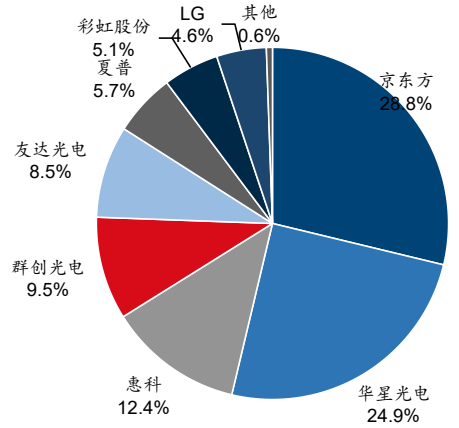
资料来源：Omdia，WitsView，国信证券经济研究所整理及预测

7.2 竞争格局：京东方、TCL华星全球LCD龙头地位稳固

国信证券 GUOSEN SECURITIES

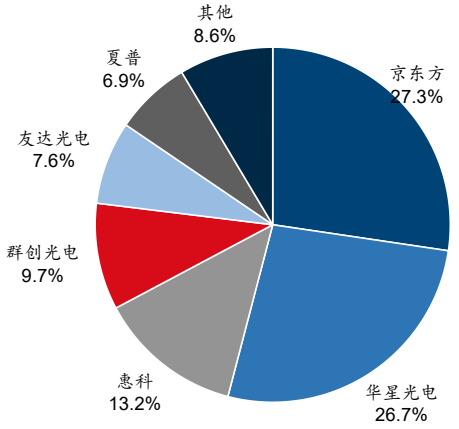
- 根据IDC数据，2025年1-9月京东方、华星光电、惠科、群创光电、友达光电分别以28.8%、24.9%、12.4%、9.5%、8.5%的市场份额（按出货面积）位居全球大尺寸LCD面板市场的前五位，其中：
- 电视：京东方、华星光电、惠科分别以27.3%、26.7%、13.2%的市占率位居前三位
- 显示器：京东方、华星光电、LG分别以30.1%、22.6%、16.8%的市占率位居前三位
- 笔记本：京东方、友达、群创分别以36.7%、17.5%、17.1%的市占率位居前三位
- 平板电脑：京东方、华星光电、LG分别以48.0%、15.9%、12.7%的市占率位居前三位

图：2025年1-9月全球大尺寸LCD面板市场份额（按出货面积）



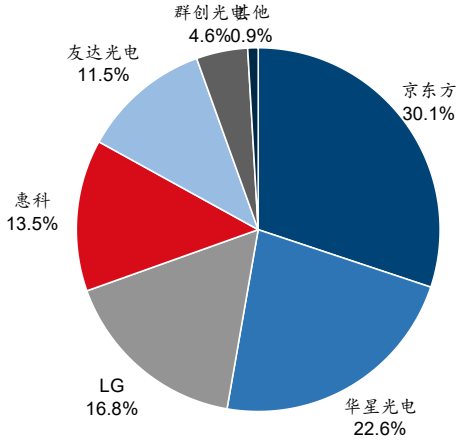
资料来源：IDC，国信证券经济研究所整理

图：2025年1-9月全球LCD电视面板市场份额（按出货面积）



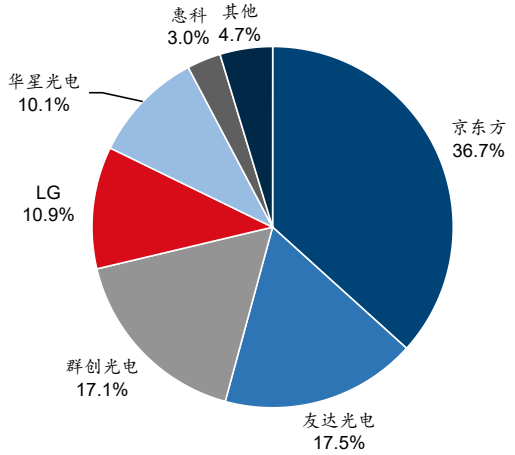
资料来源：IDC，国信证券经济研究所整理

图：2025年1-9月全球LCD显示器面板市场份额（按出货面积）



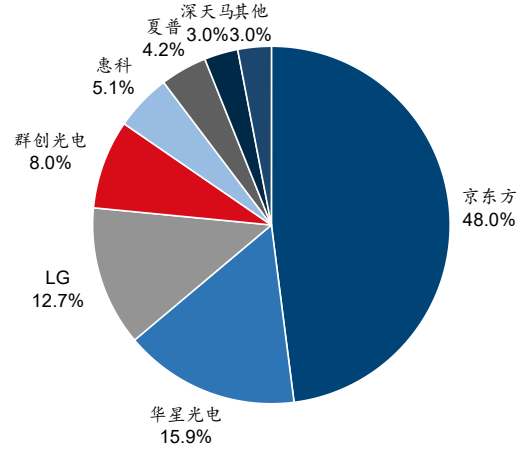
资料来源：IDC，国信证券经济研究所整理

图：2025年1-9月全球笔记本电脑面板市场份额（按出货面积）



资料来源：IDC，国信证券经济研究所整理

图：2025年1-9月全球平板电脑面板市场份额（按出货面积）



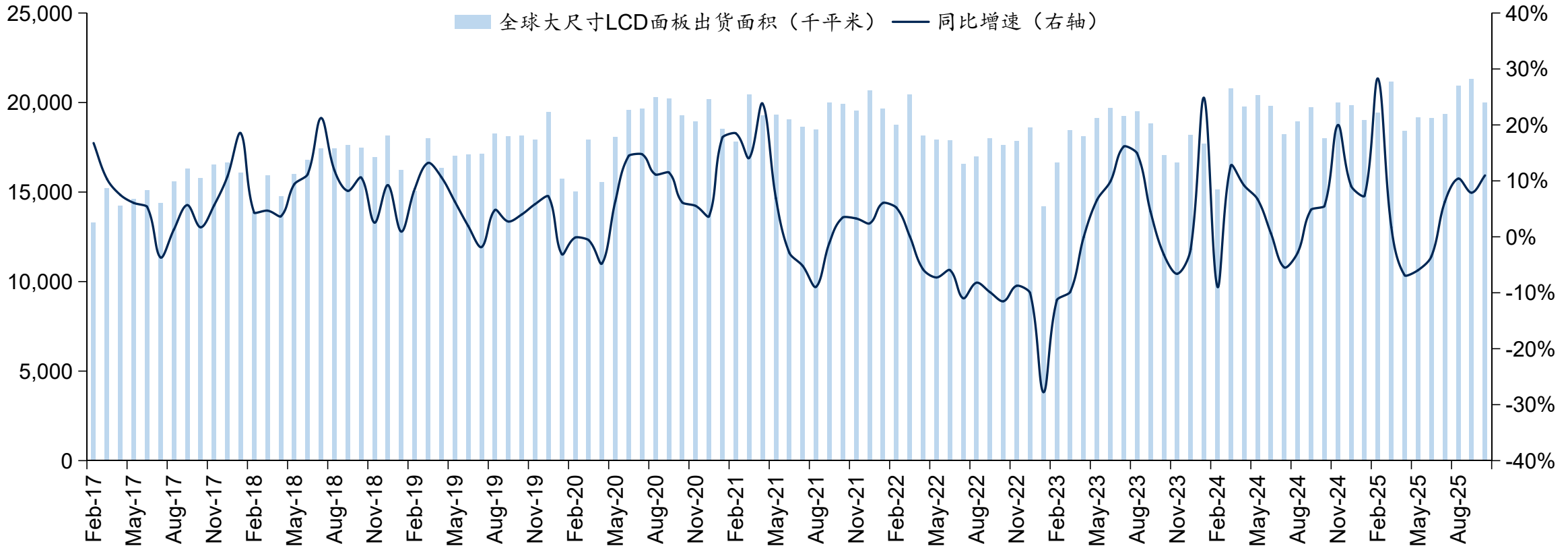
资料来源：IDC，国信证券经济研究所整理

7.2 需求：1-10月全球大尺寸LCD面板出货面积同比增长4.94%

国信证券 GUOSEN SECURITIES

- 根据WitsView数据，2025年10月全球大尺寸LCD面板（电视、显示器、笔记本电脑、平板电脑）出货量7506.0万片，同比增长15.86%，环比下滑10.45%；全球大尺寸LCD面板出货面积1999.4万平米，同比增长10.97%，环比下滑6.14%；2024年1-10月全球大尺寸LCD面板出货量7.73亿片，同比增长7.67%；全球大尺寸LCD面板出货面积1.98亿平米，同比增长4.94%。

图：全球大尺寸LCD面板出货面积



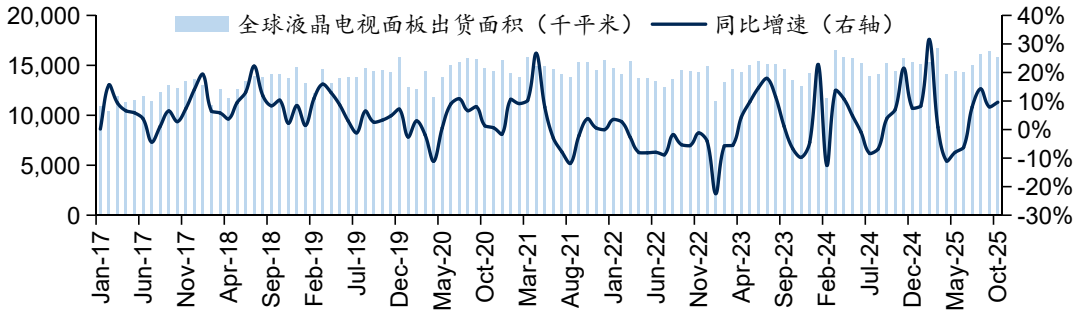
资料来源：WitsView，国信证券经济研究所整理

7.2 需求：1-10月全球大尺寸LCD面板出货面积同比增长4.94%

国信证券 GUOSEN SECURITIES

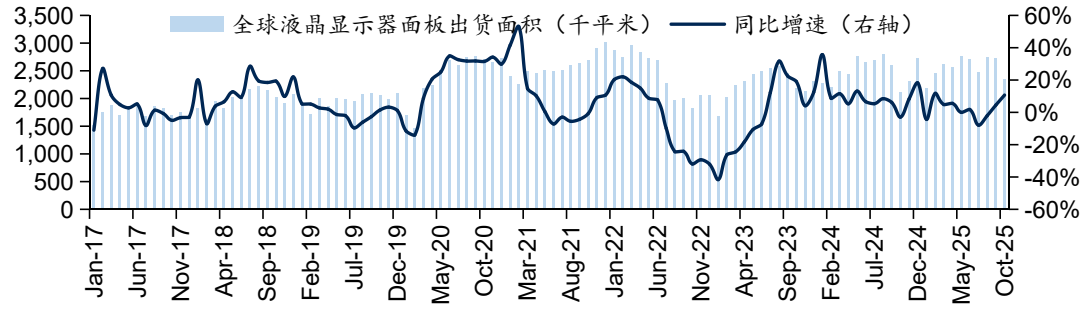
- 根据WitsView数据，2025年1-10月全球LCD电视面板出货量同比增长3.71%至2.08亿片，全球LCD电视面板出货面积同比增长4.58%至1.53亿平米；全球LCD显示器面板出货量同比增长0.50%至1.33亿片，全球LCD显示器面板出货面积同比增长2.13%至2559.7万平米；2025年1-10月全球笔记本电脑面板出货量同比增长10.18%至1.84亿片，出货面积同比增长11.13%至1189.2万平米；2025年1-10月全球平板电脑面板出货量同比增长13.73%至2.48亿片，出货面积同比增长14.19%至718.0万平米。

图：全球液晶电视面板出货面积



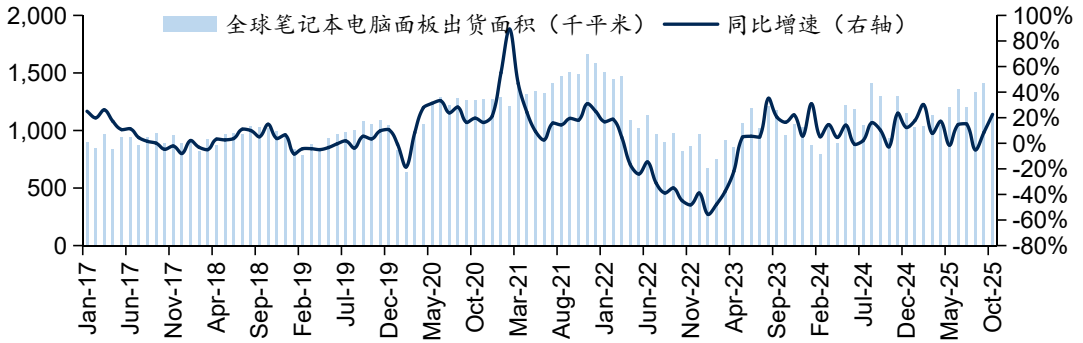
资料来源：WitsView，国信证券经济研究所整理

图：全球液晶显示器面板出货面积



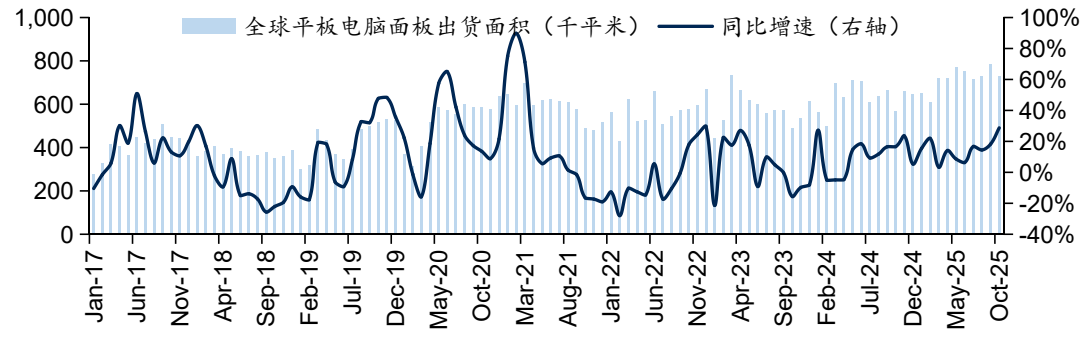
资料来源：WitsView，国信证券经济研究所整理

图：全球液晶笔记本电脑面板出货面积



资料来源：WitsView，国信证券经济研究所整理

图：全球液晶平板电脑面板出货面积



资料来源：WitsView，国信证券经济研究所整理

风险提示

风险提示

国信证券 GUOSEN SECURITIES

一、国产替代进程不及预期。国内半导体企业相比海外半导体大厂起步较晚，在技术和人才等方面存在差距，在国产替代过程中产品研发和客户导入进程可能不及预期。

二、下游需求不及预期。在地缘政治和全球经济疲软背景下，全球电子产品等终端需求可能不及预期，从而导致对半导体产品需求量减少。

三、行业竞争加剧的风险。在政策和资本支持下，国内半导体企业数量较多，在部分细分市场可能出现竞争加剧的风险，从而影响企业盈利能力。

四、国际关系发生不利变化的风险。我国半导体产业链在部分环节需要依赖海外厂商，若未来国际关系发生不利变化，可能对半导体产业链运营产生重大影响。

五、行业周期性波动风险。半导体行业渗透于国民经济的各个领域，行业整体波动性与宏观经济形势具有一定的关联性。工业控制及电源、新能源、变频白色家电等行业，如果宏观经济波动较大或长期处于低谷，上述行业的整体盈利能力会受到不同程度的影响。

六、新能源市场波动风险。新能源市场作为一个新兴的市场，可能存在较大市场波动的风险。若产业政策变化、供应链器件配套、相关设施建设和推广速度以及客户认可度等因素影响，导致新能源市场需求出现较大波动。

七、全球供应链不确定性。海外的采购与销售业务，通常以欧元、瑞士法郎、美元等外币定价并结算，外汇市场汇率的波动会影响公司所持货币资金的价值，从而影响公司的资产价值。同时全球半导体供应链受贸易限制等多方面影响将会带来供应链成本的上升。

八、显示器件需求不及预期的风险。显示器件主要应用于智能手机、平板电脑、电视以及其他新兴应用场景，消费电子产品需求弹性较大，宏观经济波动将通过改变消费者的收入预期和购买能力影响消费电子产品市场的需求，并传导至显示行业。2020年以来，全球经济受新冠肺炎影响，整体呈现复杂多变态势，贸易保护主义、单边主义抬头，世界经济运行风险和不确定性显著上升。若未来宏观经济形势持续下行，将抑制显示器件的市场需求，进而对产业链相关公司的盈利能力造成不利影响。

九、显示器件价格波动的风险。显示行业历史上具有较强的周期性，在“液晶周期”的作用下，产品价格随着技术创新对供需关系带来的冲击呈周期性波动。若未来供需关系格局发生重大变化，导致显示器件产品价格发生剧烈波动，产业链相关公司的经营业绩将受到相应影响。

十、生产设备及原材料供应风险。显示行业上游领域的技术壁垒和行业集中度较高，部分核心生产设备和原材料仍然依赖少数几家国外供应商提供，下游企业在采购该等设备和原材料时可供选择的范围较小，存在设备和原材料临时断供、价格波动较大的风险，进而对产业链公司的日常生产运营造成不利影响。

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032