

电子

2024 年中期策略报告

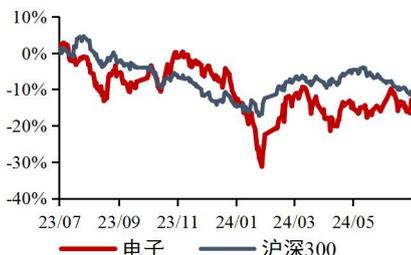
领先大市-A(维持)

AI 供需两旺铸就科技新趋势

2024 年 7 月 10 日

行业研究/行业中期策略

电子板块近一年市场表现



资料来源：最闻

相关报告：

【山证电子】山西证券电子行业周跟踪：
Q3 服务器出货动能增强，H2 晶圆厂产能利用率有望提升 2024.7.8

分析师：

高宇洋

执业登记编码：S0760523050002

邮箱：gaoyuyang@sxzq.com

傅盛盛

执业登记编码：S0760523110003

邮箱：fushengsheng@sxzq.com

张天

执业登记编码：S0760523120001

邮箱：zhangtian@sxzq.com

徐怡然

执业登记编码：S0760522050001

邮箱：xuyiran@sxzq.com

赵天宇

执业登记编码：S0760524060001

邮箱：zhaotianyu@sxzq.com

投资要点：

➢ **AI 终端：**2024 年伴随多品牌先后推出 AI PC、AI 手机，AI 端侧加速落地，带动换机潮的同时催生终端全面升级。预计到 2027 年生成式 AI 手机的市场渗透率将达到 43%，生成式 AI 手机的总量将从 2023 年的百万级增长到 2027 年的 12.3 亿部，同时 AI 算力驱动 AI 手机 PCB、散热、电池等硬件配套升级。预计 2025 年 AI PC 渗透率将达到 37%，成为市场主流，2027 年 60% 的 PC 将具备端侧 AI 功能。同时，AI PC 中的关键零部件将与处理器同步进行升级，包括内存、闪存、电池、散热、传感器等环节将有明显提升。

➢ **AI 算力：**英伟达最新旗舰产品 GB200 NVL72 算力较 DGX H100 提升 45 倍，并公布了未来 3 年“一年一代”的迭代节奏，英伟达产业链持续呈现供销两旺情形。与此同时，顶层设计下国内算力网络高速发展，运营商、地方智算中心、互联网企业不断加大算力投资，在外部政策限制和国内政策支持下，国产算力比例将获得快速提升，我们或在大芯片、存储、先进封装、液冷板块挖掘到更多预期差。

➢ **AI 网络：**展望 2024 下半年和 2025 年，我们看到 AI 网络三大主要趋势：一是以太网占比将显著提升，带动以太网交换机和光模块市场规模快速提升；二是 scaleup 能力持续迭代，铜连接成为 GB200 配套最大增量，同时以 AMD 为引领的 UALink 和以华为昇腾为引领的国产算力具备想象空间；三是 scaleout 能力从千卡、万卡到百万卡规模演进，短期看光模块配比有望再增加，未来 800G 向 1.6T、3.2T 演进价值将显著提升。

➢ **投资建议：关注 AI 终端+AI 算力+AI 网络三条主线：**

(1) AI 终端：AI 引领硬件创新升级，建议关注：**鹏鼎控股、立讯精密、东山精密、中石科技、思泉新材、珠海冠宇、洁美科技、顺络电子、三环集团。**

(2) AI 算力：业绩主线，把握海外配套和国产替代机遇，建议关注：国产大芯片**寒武纪、海光信息、景嘉微、龙芯中科**；液冷**英维克**；AI 存储**澜起科技、聚辰股份**；先进封装**长电科技、通富微电、颀中科技、汇成股份、北方华创、中微公司、拓荆科技、芯源微、鼎龙股份、华海诚科。**

(3) AI 网络：光铜共进迎接 scaleup 和 scaleout 能力升级，建议关注：光模块**新易盛、中际旭创、天孚通信、源杰科技、华工科技、铭普光磁**；铜互联**沃尔核材、神宇股份、鼎通科技。**

风险提示：AI 发展不及预期风险，外部制裁升级风险，新技术发展不及预期风险，宏观经济增长乏力风险。



请务必阅读最后一页股票评级说明和免责声明

1

目录

1. AI 终端：AI 端侧加速落地，催生终端全面升级.....	5
1.1 AI 手机：算力驱动手机 PCB、散热、电池等硬件配套升级.....	5
1.2 AI PC：关键零部件与处理器同步进行升级.....	7
2. AI 算力：业绩主线，把握海外配套和国产替代机遇.....	10
2.1 “一超多强”引领，国产大芯片产业链深度受益.....	11
2.2 AI 算力呼唤液冷全链条崛起.....	15
2.3 算力需求爆发，AI 存储大有可为.....	18
2.3.1 AI 存储达成高性能和低功耗的双解.....	18
2.3.2 HBM、DDR5 渗透率持续上升.....	20
2.4 算力时代，先进封装有望迎来加速发展.....	22
2.4.1 封装环节：Foundry 与 OSAT 各有侧重，内资厂商积极布局先进封装.....	23
2.4.2 设备环节：国产设备持续突破，进口替代进程加速.....	24
2.4.3 材料环节：关键材料性能要求升级，高端品类国产化空间巨大.....	25
3. AI 网络：光铜共进迎接 scaleup 和 scaleout 能力升级.....	26
3.1 光模块：海外需求持续性无忧，国内市场有望快速启动.....	26
3.2 铜互联：英伟达引领短距高密度互联新趋势，市场空间广阔.....	29
4. 风险提示.....	32

图表目录

图 1：各品牌 AI 手机发布加速 AI 端侧落地.....	5
图 2：生成式 AI 手机总规模预测.....	6
图 3：全球 AI PC 渗透率曲线.....	8
图 4：具有 AI 功能的处理器线路图.....	9

图 5: 关键零部件与处理器同步进行升级.....	10
图 6: 英伟达在 computex2024 公布三代数据中心芯片路线图.....	11
图 7: 我国智能算力规模及预测.....	11
图 8: 中美拥有最多的 AI 芯片公司, 国内以华为为代表并产生众多互联网自研和创业芯片公司.....	12
图 9: AI Infra 软硬件栈由 5 层架构组成.....	13
图 10: CUDA-X 数据处理平台成为英伟达在垂直行业广泛应用的重要护城河.....	13
图 11: 华为 Atlas900 服务器机柜和 Atlas800 训练服务器设计.....	14
图 12: 寒武纪思元 290 核心优势.....	15
图 13: 主力 GPU 厂商芯片功耗快速上升.....	16
图 14: 当机柜密度大于 20kw 时适合采用液冷散热.....	16
图 15: 液冷不同交付模式示意图.....	18
图 16: 美光 AI 存储产品组合.....	19
图 17: 三星存算一体化芯片模型.....	19
图 18: HBM 内部结构图.....	20
图 19: 采用 CoWoS 封装的英伟达 A100.....	22
图 20: 英伟达 A100 CoWoS 封装切面图.....	22
图 21: 全球先进封装市场规模, 亿美元.....	23
图 22: 通信领域 2022-2028 CAGR 17%, 增速最高.....	23
图 23: 2022 年先进封装市场格局集中.....	23
图 24: OSAT 与 IDM、晶圆代工厂技术布局有差异.....	23
图 25: 先进封装设备体系.....	24
图 26: 环氧塑封料是先进封装主要保护材料.....	25
图 27: 垂直互联与异质集成提升 CMP 耗材需求.....	25



图 28: coherent 认为未来几年数通光模块市场将快速增长.....	27
图 29: YOLE 最新预测全球数通光模块将在 2029 年达到 224 亿美金.....	27
图 30: 英伟达将在 2025 全面量产支持 1.6T 端口的网卡和交换机.....	28
图 31: 谷歌 OCS 兼容不同速率可快速适配 TPUv6 对更高速率端口要求.....	28
图 32: 以太网光模块单位带宽 (Gbps) 价格 (美元) 变化.....	28
图 33: LPO 出货量预计在 2025 年明显增长.....	28
图 34: NVL72 Switch tray 使用的芯片直出跳线.....	30
图 35: NVL72 使用的高速背板 cartridge.....	30
图 36: 服务器使用到的各种高速通信线类型.....	30
图 37: UALINK 拓展通用 scaleup 协议.....	31
图 38: 华为“天成”机柜级算力平台产品.....	31
表 1: 多款 AI 手机处理器的性能对比.....	6
表 2: 微软“Copilot+PC”的 AI 功能.....	7
表 3: 全球数据中心服务器冷板液冷市场规模测算.....	17
表 4: HBM 已经成为 AI 服务器的搭载标配.....	21
表 5: 内存迭代带来内存接口芯片、配套芯片量的增长.....	21

1. AI 终端：AI 端侧加速落地，催生终端全面升级

1.1 AI 手机：算力驱动手机 PCB、散热、电池等硬件配套升级

各品牌 AI 手机陆续发布，驱动新一轮手机换机潮，并加速 AI 端侧落地。2023 年 11 月 13 日，VIVO 率先发布 AI 手机，进入 2024 年，各手机品牌加速发布搭载 AI 功能的智能手机，截至 6 月底，华为、三星、小米、OPPO、荣耀、魅族等各品牌 AI 手机均已发布。与此同时，今年 9 月苹果 AI 功能将正式开放。考虑到 2022、2023 年全球智能手机的平均换机周期被拉长，和 AI 手机陆续发布带来的催化，我们认为 2024 年有望迎来一轮 AI 手机换机潮，并加速 AI 端侧落地。

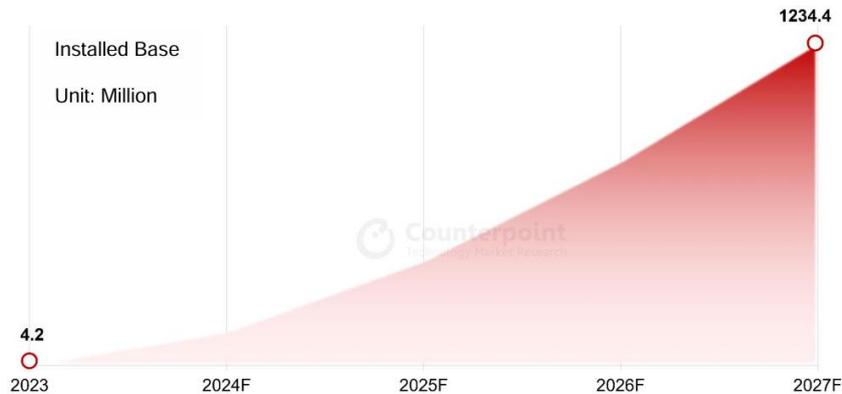
图 1：各品牌 AI 手机发布加速 AI 端侧落地



资料来源：VIVO、OPPO、荣耀、三星、小米、魅族、一加、华为官网，山西证券研究所

2024 年将成为生成式 AI 手机的元年，预计 2027 年生成式 AI 手机的市场渗透率将达到 43%。根据 Counterpoint 数据，2023 年生成式 AI 手机规模仅有 420 万台，在全年 11.7 亿部手机出货中占比不足 1%。但进入 2024 年，随着各大手机厂商将生成式 AI 功能作为智能手机升级重点，这将加速生成式 AI 手机的普及，预计 2027 年生成式 AI 手机的市场渗透率将达到 43%，生成式 AI 手机的存量规模也将从 2023 年的百万级增至 2027 年的 12.3 亿部。

图 2：生成式 AI 手机总规模预测



资料来源：Counterpoint 《生成式 AI 手机产业白皮书》，山西证券研究所

AI 算力升级驱动手机 SoC 性能显著提升。对比 AI 手机和传统手机的处理器能看到几大明显变化：（1）AI 手机搭载 NPU，并不断提升 NPU 算力以满足大模型端侧运行。（2）GPU 性能，浮点运算能力实现从传统手机的 GFLOPS 到 AI 手机的 TFLOPS 飞跃。（3）内存容量明显提升，AI 算力带来海量数据，手机内存实现从 4/6/8GB 到 16/20/24GB 的显著升级，预计存储价值量翻 2-3 倍。

表 1：多款 AI 手机处理器的性能对比

		Mediatek Dimensity 9300	Qualcomm Snapdragon 8 Gen 3	Apple A17 Pro	Samsung Exynos 2400	Qualcomm Snapdragon 8S Gen 3	Google Tensor G3
CPU	CORES	8	8	6	10	8	9
GPU	型号	Mali-G720 Immortalis MP12	Adreno 750	Apple A17 GPU	Samsung Xclipse 940	Adreno 735	Mali-G715 MP7
	FLOPS	5990.4 Gigaflops	4435.2 Gigaflops	2147.2 Gigaflops	3407 Gigaflops	3379.2 Gigaflops	2415.8 Gigaflops
存储	最大内存	24 GB	24GB	8GB	24GB	24GB	16 GB
	最大带宽	76.8 Gbit/s	77 Gbit/s	51.2 Gbit/s	68.2 Gbit/s	64 Gbit/s	64 Gbit/s
封装制程	nm 级	4	4	3	4	4	4
AI 性能	Neural processor (NPU)	MediaTek APU 790	Hexagon	Apple Neural Engine	yes	Hexagon	yes
TDP	热功耗	7W	6.3W	8W	6W	6W	6W

资料来源：NanoReview，山西证券研究所

关注 PCB、散热、电池等硬件配套升级。(1) PCB: 随着 AI 加入和 NPU 算力的不断升级, 驱动手机 FPC 层数增加、线宽线距更小, 进而提升单机 FPC 价值量, 同时手机 SLP 板的也需要面积变大、线宽线距更小以容纳更多元器件, 加上材料升级, 预计单机 SLP 价值量也将提升。(2) 散热: AI 手机算力和功耗的急速增长, 使得散热成为确保 AI 手机稳定运行的关键, 预计不锈钢 VC 均热板代替铜 VC 均热板、石墨烯材料散热片代替人工石墨片散热的散热方案将在 AI 手机中加速渗透。(3) 电池: 为支持大模型运行, 不仅手机电池容量要加大, 电芯也会更密, 电池结构设计用不锈钢外壳代替软包外壳更好散热。

1.2 AI PC: 关键零部件与处理器同步进行升级

微软推出专为 AI 设计的新型 Windows PC, 称为“Copilot+PC”, 开启 AI PC 新时代。与传统 PC 不同, Copilot+PC 具有每秒 45 万亿次操作 (TOPS) 的神经处理单元 (NPU), 专门用于加速人工智能和机器学习任务。微软首批 Copilot+PC 搭载高通骁龙 X Elite 和骁龙 X Plus 处理器, 较 Surface Pro 9 的性能提高了 90%, 并采用定制的高通 Oryon CPU, 同时还具有 Copilot 体验、Recall、实时字幕、Surface Studio 相机和 AI 增强等 AI 功能。作为行业领导者, Copilot+PC 的发布将开启 AI PC 新时代。

表 2: 微软“Copilot+PC”的 AI 功能

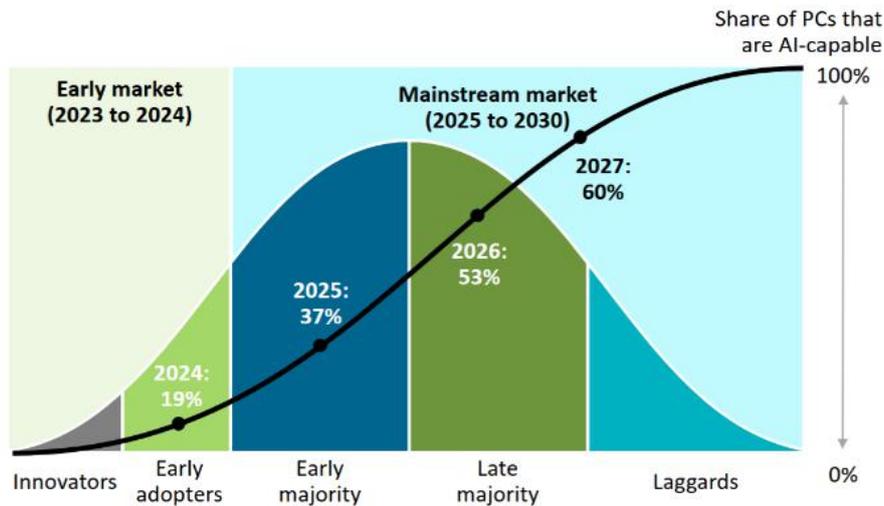
AI 功能	功能介绍
Recall	这项功能利用 NPU 芯片, 记录用户在 PC 上的所有操作, 包括打开的文件、访问的网站、使用的应用程序等。用户可以通过语音或文本查询, 轻松找到特定文件、电子邮件或网页, 就如同拥有“过目不忘”的超强记忆力。
实时字幕	支持使用实时字幕提高在视频通话和观看电影时的理解能力, 可以在屏幕上实时将口语转换为英文字幕, 并支持 40 多种语言翻译。
Copilot 体验	内置键盘包括一个 Copilot 键, 可在笔记本电脑上启动 AI 驱动的助手。
Surface Studio 相机	这些相机将使用 Surface Pro 第 11 版上的 Windows Studios 效果和 Surface 笔记本电脑第 7 版上的 Windows Studio 效果, 通过 AI 驱动的效果来增强你的相机源。
AI 增强	如 Live Cocreator 让用户能够在设备上实时生成和编辑 AI 图像, 用户只需用自然语言描述他们想要创建的图像, Cocreator 就会根据描述生成相应的图像, 并允许用户进行实时修改和优化, 释放无限创意。

资料来源: 微软官网, 山西证券研究所

2025 年 AI PC 将成为主流。Canalys 预计 2024 年出货的 PC 中, 19% 将具备 AI 功能。这些电脑将包括所有 M 系列 Mac, 以及搭载 AMD、英特尔和高通等公司 AI 加速芯片组的首款

Windows PC。Windows 10 服务将于 2025 年 10 月结束，随着客户在此日期之前更新设备，预计具备 AI 功能的 PC 在 PC 总出货量中所占的份额将从 2025 年开始飙升。Canalys 预计 2025 年 AI PC 渗透率将达到 37%，成为市场主流，2027 年 60% 的 PC 将具备端侧 AI 功能。

图 3：全球 AI PC 渗透率曲线



资料来源：Canalys 《AI PC 的现在与未来》，山西证券研究所

高算力 NPU 是决战 AI PC 性能的关键因素。目前，为了顺应 PC 领域的 AI 应用和趋势，x86 和 ARM 阵营的处理器厂商都已对其产品规格进行了战略性升级。与以往机型相比，一个关键的不同是集成 NPU。传统 PC 处理器由 CPU+GPU 组成，其中 CPU 处理综合计算需求，GPU 处理图像计算需求，而新一代大模型计算需要专业神经网络计算能力，因此引入 NPU 已经成为行业趋势，因此新一代 PC 处理器已经形成 CPU+GPU+NPU 的三核计算架构。虽然没有 NPU 并不妨碍 PC 运行 AI 功能，但集成 NPU 可显著提高计算效率、缩短响应时间并降低 CPU 功耗。在当前的 NPU 技术领域，高通公司的 X Elite 以 45TOPS 算力在所有处理器制造商产品中速度最快，一跃成为行业标杆。

图 4：具有 AI 功能的处理器线路图



资料来源：Canalys 《AI PC 的现在与未来》，山西证券研究所

AI PC 中的关键零部件将与处理器同步进行升级，其中内存、闪存、电池、散热、传感器等环节将有明显提升。除了处理器将有明显升级，而与处理器配套的芯片及零组件环节也将有直接提升，增量明显的 PC 供应链环节包括：

(1) **内存：容量至少翻倍。**云端模型一般在千亿参数级别，而边缘端模型进行参数蒸馏后参数量级也在百亿级，PC 运行 AI 模型应用，将对密集数据传输和处理有较高的需求，因此内存容量会有显著提升，我们认为未来 AI PC 入门级标配一定是 32GB 内存，当前 16GB 内存会被淘汰，明年则是 64GB PC 开始出货。

(2) **闪存：对固态硬盘性能和容量提出高要求。**AI 大模型预训练算法存储容量较大，同时边缘端模型对用户个性化需求需要进行预训练，训练数据及个性化模型参数同样占据一定容量，因此 AI PC 将对 SSD 性能和容量提出高要求。

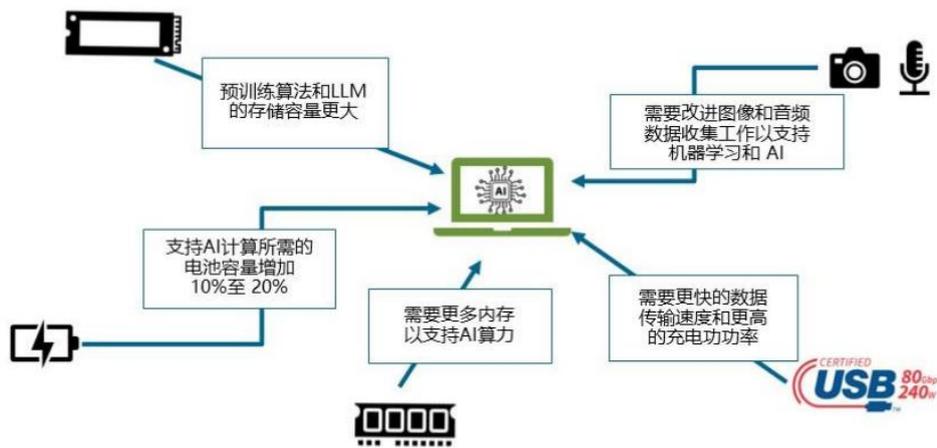
(3) **电池及充电环节：容量、功率均有所增长。**处理器算力升级提高 PC 整体功耗，电能消耗显著提升，因此为适配芯片能耗，电池容量进一步增大，同时充电功率也将随之提升。

(4) **散热：需求提升，单机散热件价值量提升。**由于电池及芯片功耗提升带来的热量提升，散热需求明显提升，传统散热解决方案散热效率有限，在 PC 整体功耗提升的背景下，散热材料的用量增长以及散热方式的多元化已成为趋势，考虑到 PC 散热由多个散热部件组成，随着散热材料用量增长和升级，我们认为单机散热件价值量有望提升。

(5) **传感器：精度升级，数量和种类提升。**AI PC 强调更高维度的用户与 PC 直接交互，而传感器则是输入及输出的窗口，随着 AI PC 对实时环境和用户反馈的需求不断增长，PC 与

外界交互的方式将更加多元化，同时对交互的准确性和及时性也有更高要求。声学、光学以及其他传感方式将持续优化，例如摄像头、麦克风、扬声器等传统光学及声学传感器的数量及准确度将持续提升，新型传感器例如触觉、惯性等也有望导入 PC 等终端，提升终端对外界的感知维度以及信息的输出方式。

图 5：关键零部件与处理器同步进行升级



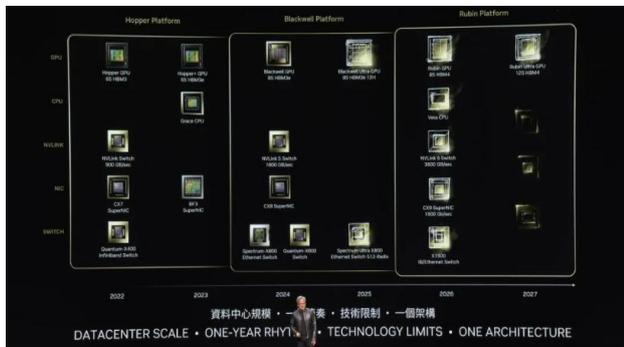
资料来源：Canalys 《AI PC 的现在与未来》，山西证券研究所

AI 端侧加速落地，催生终端全面升级，建议关注：鹏鼎控股、立讯精密、东山精密、中石科技、思泉新材、珠海冠宇、洁美科技、顺络电子、三环集团等。

2. AI 算力：业绩主线，把握海外配套和国产替代机遇

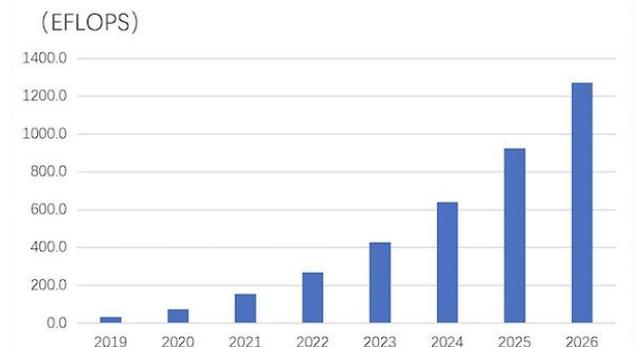
AI 算力需求继续指数级增长。英伟达最新旗舰产品 GB200 NVL72 算力较 DGX H100 提升 45 倍，随着未来 3 年“一年一代”的迭代节奏，英伟达产业链将持续呈现供销两旺情形。与此同时，顶层设计下国内算力网络高速发展，运营商、地方智算中心、互联网企业不断加大算力投资，在外部政策限制和国内政策支持下，国产算力比例将获得快速提升。展望 2024 下半年，我们仍然看好 NV 产业链配套上游的光模块、PCB、服务器代工等环节，但聚焦国产算力，我们或在大芯片、存储、封装、液冷板块挖掘到更多预期差，同时随着国产算力占比提升，我们将看到千亿国产算力大市场带来更多成长股的投资机会。

图 6：英伟达在 computex2024 公布三代数据中心芯片路线图



资料来源：新智元，山西证券研究所

图 7：我国智能算力规模及预测



资料来源：新华三&中国信通院《2023 智能算力发展白皮书》援引 IDC，山西证券研究所

2.1 “一超多强”引领，国产大芯片产业链深度受益

AI 大芯片是国产替代最亟需突破的领域，当前已形成以华为昇腾为引领的“一超多强”局面。AI 大芯片是半导体先进制程、先进封装、计算架构设计、通信接口设计集大成者。英伟达最新芯片 Blackwell B200 FP8 算力达 9PFlops，与国产芯片拉开了较大差距。美国 BIS 在 2023 年 10 月新修订的 AI 芯片出口管制令中将 AI 芯片总处理性能（TPP）大于 4800 或 TPP 大于 1600 的同时性能密度（PD）大于 5.92 作为出口禁令最新标准，几乎涉及到 B100、H100、A100、L40s 以及 AMD、Intel 的所有主力数据中心产品。与此同时，英伟达推出的国内版本 H20 综合性能大幅下降，并不能满足国内用户对于性价比的需求。我们认为，国产大芯片将迎来历史发展机遇，我们将通过 chiplet、加速国产存力运力配套以及集群部署方案来弥补单芯片在性能上的差距。根据 YOLE 最新统计，当前国产 AI 大芯片以华为昇腾系列为引领，其最新产品 Ascend 910B 目前供不应求，而以百度昆仑芯、阿里含光等互联网自研芯片以及燧原、寒武纪、登临、沐曦、墨芯等为代表的新兴势力未来也将成为市场的有机组成部分。

图 8：中美拥有最多的 AI 芯片公司，国内以华为为代表并产生众多互联网自研和创业芯片公司



资料来源：YOLE，山西证券研究所

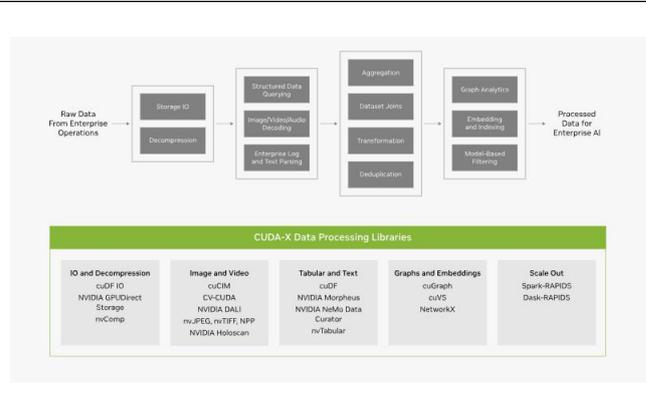
国产大芯片是算力、存力、运力、工具库、软件生态的综合竞争。AI 大芯片的高性能稳定工作需要一系列软件、工具库、平台的支持，我们称之为 AI Infra，根据浪潮信息 OGAI 大模型智算软件栈的介绍，AI Infra 包括基础设施（主要是硬件和运维支撑平台）、系统环境（包括操作系统、Driver、算子库、环境配置、系统监控）、调度平台（包括断点续训、故障定位、算力调度）、模型工具（包括数据治理、预训练、微调的开发工具）以及多模纳管层（实现以应用为目的的可视化、低代码使用），需要巨大的研发投入。此外另一个难点在于“类 CUDA”应用生态的打造，CUDA 内置的丰富应用算子库、函数大大减轻了垂直行业用户开发工作量，形成了巨大使用“粘性”，AI 芯片应用生态非长期研发和迭代更新不能解决。我们认为，国产大芯片如果从“能用”进入“好用”仍需投入大量研发成本和时间，目前丰富 AI infra 的方式一是全方位自主研发并创造开源环境促应用（类似昇腾）；二是搭建“类 CUDA”GPGPU 算子库（类似海光、沐曦、摩尔线程）；三是尽量统一标准接口与第三方协同开发（类似中国移动发起的“全调度以太网”在调度平台层做的工作）。

图 9：AI Infra 软硬件栈由 5 层架构组成



资料来源：浪潮信息官网，山西证券研究所

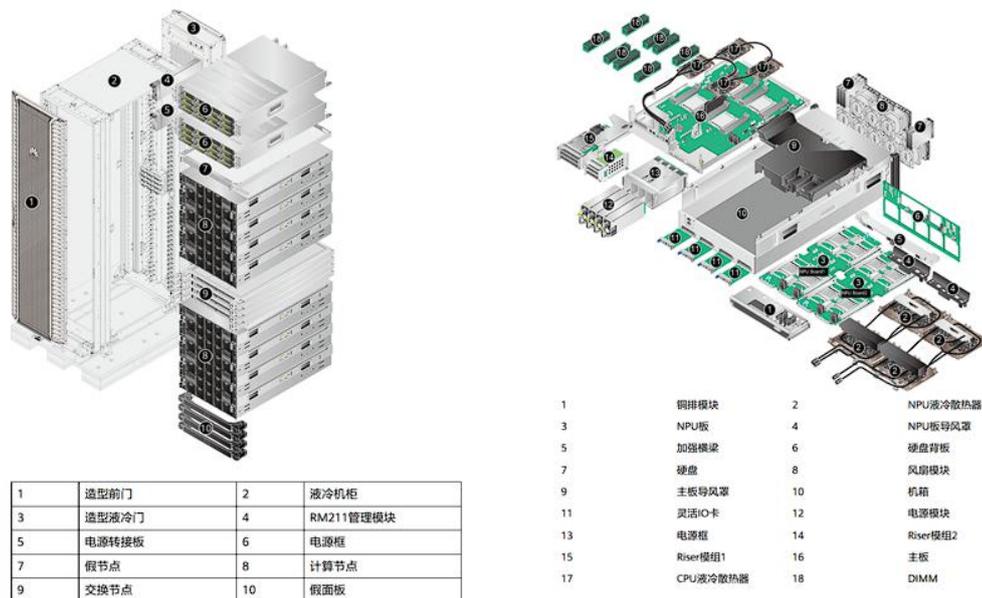
图 10：CUDA-X 数据处理平台成为英伟达在垂直行业广泛应用的重要护城河



资料来源：英伟达官网，山西证券研究所

华为昇腾引领国产替代，随着产能提升市场规模有望快速增长，供应链各环节受益。华为近年来在昇腾 AI 系列芯片迭代、自主可控产能提升方面取得了长足进展。根据《南华早报》报道，Ascend 910B 芯片在性能上已超越英伟达 A100 GPU，成为运营商、智算中心、互联网企业国产替代的主流选择。2024 年 5 月，中国移动发布的 2024 至 2025 新型智算中心 190 亿元集采中，标包 1 中标厂商均为国产化服务器鲲鹏昇腾产业链合作伙伴。根据《2023 智能算力发展白皮书》，假设到 2025 年我国智能算力规模至少要达到 900EFlops，其中昇腾市占率达到 30%，昇腾 910B 算力相当于 A100 (FP16 约 300Tflops)，平均价格为 12 万元，则保守估计昇腾 910B 累计市场空间为 1080 亿元。华为昇腾产业链配套环节包括服务器组装、电源、散热、高速 IO 连接器、背板连接模组、PCB、铜缆、光模块、DRAM、SSD、RAID 控制器、电源管理等 IC 等，一是配套昇腾出货产值快速增加，二是产品迭代升级加速、价值量或提升，三是对标海外赛道龙头长期成长空间广阔。

图 11：华为 Atlas900 服务器机柜和 Atlas800 训练服务器设计



资料来源：《atlas800 训练服务器用户指南》，《Atlas900 用户指南》，山西证券研究所

寒武纪新款产品性能提升，未来有望在智算中心、运营商、互联网多点突破。寒武纪至今在大算力 AI 芯片已发布了四代产品，2018 年的思元 100、2019 年的思元 270、2020 年的思元 290 以及 2021 年的思元 370 系列。根据官网描述，寒武纪首款云端训练芯片思元 290 实现了 INT8 算力 512TOPS，而在研的第五代智能处理器微架构对推荐系统和大语言模型训练推理场景进行了重点优化，在编程灵活性、能效、功耗、面积等方面大幅提升了产品竞争力。应用落地方面，寒武纪携手国内头部算法公司推动大模型应用，在视觉大模型领域与智象未来达成大模型深度适配，在语言大模型领域与百川智能等头部客户进行了大模型适配。2023 年，公司智能计算集群系统业务实现营收 6.05 亿元，在沈阳、台州成功落地了智能计算集群项目，未来将有更多项目落地。与此同时，互联网和电信客户基于供应链安全和不同工作负载的需要也有部署多家 AI 芯片的需求，我们认为公司的客户群体未来会更加多元化。

图 12：寒武纪思元 290 核心优势



资料来源：寒武纪官网，山西证券研究所

未来端侧和云端推理芯片也将更加多元化。随着国内外大模型快速降价、AI agent、数字人、AI 客服等应用快速落地，我们认为 AI 推理工作负载的比重将快速上升。推理芯片与训练芯片的侧重点不同（边缘侧低延时低功耗、针对视觉处理任务的优化、针对不同模型的深度优化），我们看到美国科技大厂（微软、谷歌、亚马逊、苹果）均有量产或自研芯片计划，OPENAI 也表达了自研芯片意向。短期来看英伟达的竞争优势依然明显，但未来云端推理和边缘侧应用或给更多公司带来机会。海光 DCU 在市场上具备三大竞争优势，一是具备全精度各种数据格式算力，二是高速并行数据处理能力，三是兼容“类 CUDA”环境，公司正通过参与开源软件项目加快产品推广速度。景嘉微定增募投项目：“高性能通用 GPU 芯片研发及产业化项目”由图形处理领域拓展至高性能计算领域，“通用 GPU 先进架构研发中心建设项目”主要面向高性能计算和数据处理需求，彰显了公司拓展 AI 大芯片市场的决心。

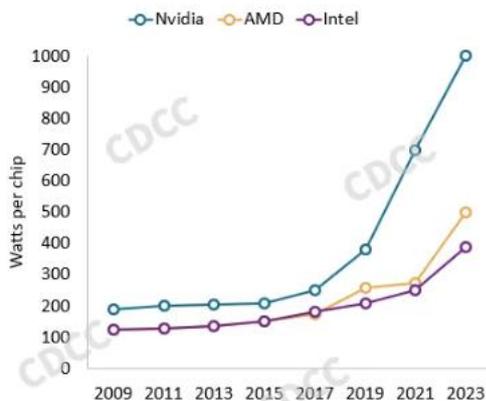
国产大芯片建议关注：寒武纪、海光信息、景嘉微、龙芯中科等。

2.2 AI 算力呼唤液冷全链条崛起

主流大芯片功耗快速提升，液冷从选配逐渐成为标配。全球 AI 算力高增，各大芯片厂商的 TDP 功率也在飞速提升，根据 S&P Global 数据，2017 年主流芯片厂商 TDP 仅为 200w 左右，到 2023 年，一般 GPU 的 TDP 已经达到 350-500W，NVIDIA 新一代 B100 功耗甚至超过 1000W。

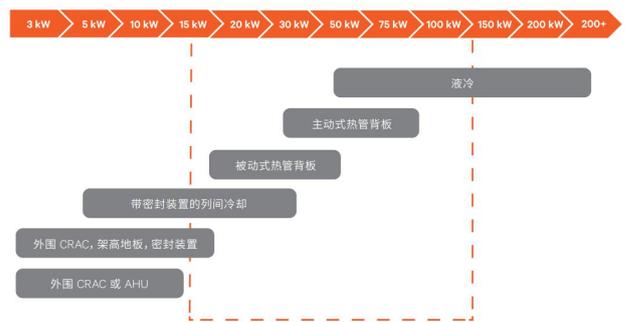
随着芯片 TDP 和机柜容量持续提升，液冷从选配逐渐成为标配。根据 S&P Global 数据，当 TDP 达 200-300W 时，风冷散热系统需要进行特殊升级；当 TDP 达 350-400W 时，风冷逐渐接近极限，液冷将成为必然选择。NVIDIA 新一代 B100 功耗可能超过 1000W，意味 AI 算力所需机柜功率更高，根据《2021-2022 年度中国数据中心基础设施产品市场总报告》和立鼎产业研究网数据，我国约 90% 以上的机柜功率密度在 15kW 以下，但未来主流机柜功率预计会达 25kW-50kW/柜，甚至更高。而液冷技术以其低 PUE（冷板式液冷 PUE<1.2、浸没式液冷技术 PUE<1.1，节能 20%-30% 以上）、满足高密度部署（降低占地和建设成本）、服务器运行更加安全可靠（CPU 温度低至 65℃ 以下）、全年全地域适用等优势，有望解决能耗和散热发展瓶颈。

图 13：主力 GPU 厂商芯片功耗快速上升



资料来源：CDCC，S&P，山西证券研究所

图 14：当机柜密度大于 20kw 时适合采用液冷散热



资料来源：《维谛技术数据中心液冷解决方案白皮书》，山西证券研究所

液冷分为服务器内冷板、机房二次侧（室内）、机房一次侧（室外）等环节，整体市场空间广阔。目前主流液冷应用方式为冷板式液冷和单相浸没式液冷，其中，冷板式液冷需要定制，成本较高，但在可维护性、空间利用性、兼容性等方面较为优秀，由于国内冷板式液冷起步较早，具有先发优势，实际应用场景广泛，因此我们认为冷板式液冷将成为未来主流路线。我们预计，2028 年冷板液冷数据中心市场空间有望超 640 亿元，冷板&管路市场空间有望超 460 亿元。我们预计 2028 年全球服务器 CPU 出货量将超 3600 万片，对应单芯片服务器平均功耗约 290w，冷板液冷渗透率将增长至 25%；全球服务器 GPU 出货量将接近 1500 万片，对应单芯片服务器平均功耗约 1800w，冷板液冷渗透率将增长至 55%；冷板液冷单 kW 投资额下降至 3400 元左右，冷板&管路价值量约 2500 元/Kw；则 2028 年 CPU 服务器冷板液冷市场空间超

130 亿元，CPU 冷板&管路市场规模超 90 亿元；AI 服务器冷板液冷市场空间超 500 亿元，AI 冷板&管路市场规模超 360 亿元。

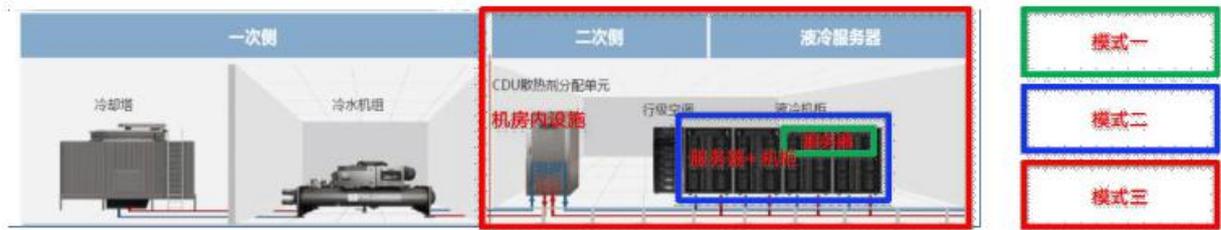
表 3：全球数据中心服务器冷板液冷市场规模测算

	2023	2024e	2025e	2026e	2027e	2028e
全球 CPU 出货量 (万片)	2811	2964	3123	3279	3443	3615
CPU 平均功率 (w)	180	198	218	240	264	290
CPU 服务器对应单芯片平均功率 (w)	270	297	327	359	395	435
冷板液冷渗透率	10%	13%	16%	19%	22%	25%
冷板液冷单 KW 投资 (元/kW)	4000	3880	3764	3651	3541	3435
CPU 冷板液冷市场规模 (亿元)	30	44	61	82	106	135
CPU 冷板&管路市场规模 (亿元)	19	29	41	56	75	98
全球 GPU/AIDSA 出货量 (万片)	549	685	914	1214	1351	1486
GPU/AIDSA 平均功率 (w)	700	752	807	867	931	1000
GPU 服务器单芯片平均功率 (w)	1260	1353	1453	1561	1676	1800
冷板液冷渗透率	25%	30%	40%	45%	50%	55%
冷板液冷单 kw 投资 (元/kW)	4000	3880	3764	3651	3541	3435
AI 冷板液冷市场规模 (亿元)	69	108	200	311	401	506
AI 冷板&管路市场规模 (亿元)	43	70	133	213	283	368
冷板液冷市场规模 (亿元)	100	152	261	393	507	641
其中：一次侧市场规模 (亿元)	65	99	170	256	330	416
二次侧市场规模 (亿元)	35	53	91	138	178	224
冷板&管路市场规模 (亿元)	62	98	174	269	358	466

资料来源：IDC，Intel，SemiAnalysis，热设计网，半导体行业观察，199IT，IT 之家，TechInsights，海光信息招股说明书，山西证券研究所

当前液冷数据中心标准化程度低，服务器冷板 ODM 以及二次侧整体交付是当前的主流模式。液冷系统涉及环节较多，根据 IT 设备侧与机房配套侧各自的交付内容，液冷服务器有三种交付模式，分别为：模式一液冷服务器；模式二液冷服务器+液冷机柜；模式三液冷服务器+液冷机柜+CDU+二次侧管路。由于目前液冷技术标准化程度较弱，各厂家之间缺少兼容性对接，当前运营商等大客户主要采用产品成熟度较好、配套对接内容少、交付及后期运维风险较小的模式二进行交付，英维克是典型的代表公司。我们认为如果液冷系统接口未来实现解耦标准化，模式一可能成为主流，客户会分别采购一次侧和二次侧部件。

图 15：液冷不同交付模式示意图



资料来源：《冷板级液冷服务器设计白皮书》，开放数据中心委员会 ODCC，山西证券研究所

服务器冷板投资：随着 AI 服务器市场占比增加、英伟达 B200 系列液冷的大幅增加，以及国内新建数据中心对 PUE 的要求。冷板式液冷服务器相较于浸没式液冷有望更快渗透，出货量快速增加。服务器厂商的冷板、管路通常采用 ODM 代工方式，定制性强，与主流服务器厂商有合作经验的厂商有望拿到更高比例，相关供应商订单有望快速增长，直接受益于液冷渗透率提升的线性增长。目前英伟达、昇腾等厂商倾向于冷板 ODM、二次侧、一次侧分别解耦的交付的解耦模式，NV 链主要配套商为台系、美系厂商，国内昇腾冷板 ODM 配套商建议关注飞荣达、科创新源、英维克。

液冷二次侧投资：模式二为非解耦整体解决方案，是液冷数据中心渗透率早期的交付主流模式，有望受益于数据中心改造、新建数据中心等，主要产品涉及服务器冷板、管路以及 Manifold、接头等，主要以项目制交付。国内曙光、中兴、新华三等倾向于该模式，建议关注英维克、高澜股份、曙光数创。

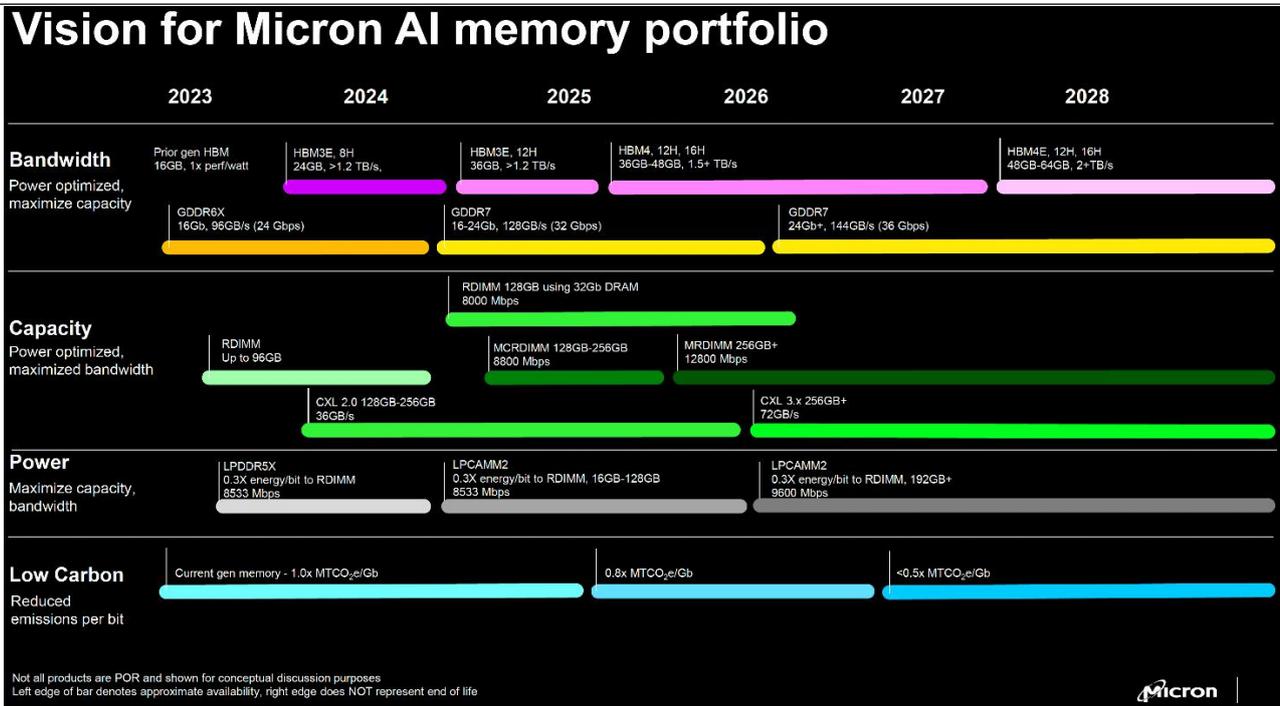
液冷建议关注：英维克等。

2.3 算力需求爆发，AI 存储大有可为

2.3.1 AI 存储达成高性能和低功耗的双解

随着算力需求不断攀升，AI 大模型对存储的带宽、容量和功耗也提出了更高要求。AI 大模型的兴起催生了海量算力需求，而数据处理量和传输速率大幅提升使得 AI 服务器对芯片内存容量和传输带宽提出更高要求。从美光发布的 AI 存储产品组合来看，AI 存储产品在向低能耗、高带宽、高容量持续迭代，以高性能牵引 AI 技术进一步革新。

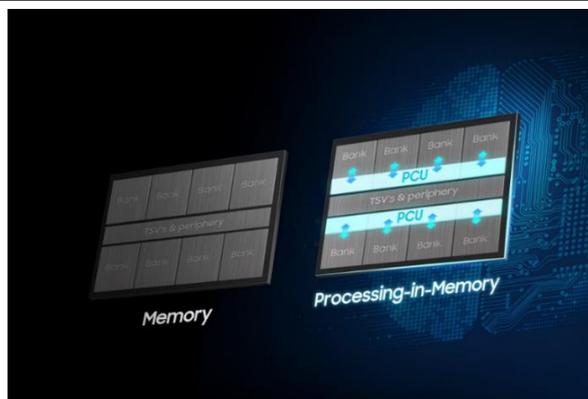
图 16: 美光 AI 存储产品组合



资料来源: Micron, Tom’s Hardware, 山西证券研究所

存算一体是 AI 存储的未来发展方向。存算一体化通过在内存内核中集成一个称为可编程计算单元的 AI 引擎来处理一些逻辑功能，可编程计算单元和 CPU 中的多核处理一样，支持内存中的并行处理，从而提高性能。与现有内存解决方案相比，三星的存算一体化可以通过可编程计算单元将性能提高达 4 倍。同时较之现有 HBM，存算一体化的系统能耗降低了 70%。

图 17: 三星存算一体化芯片模型

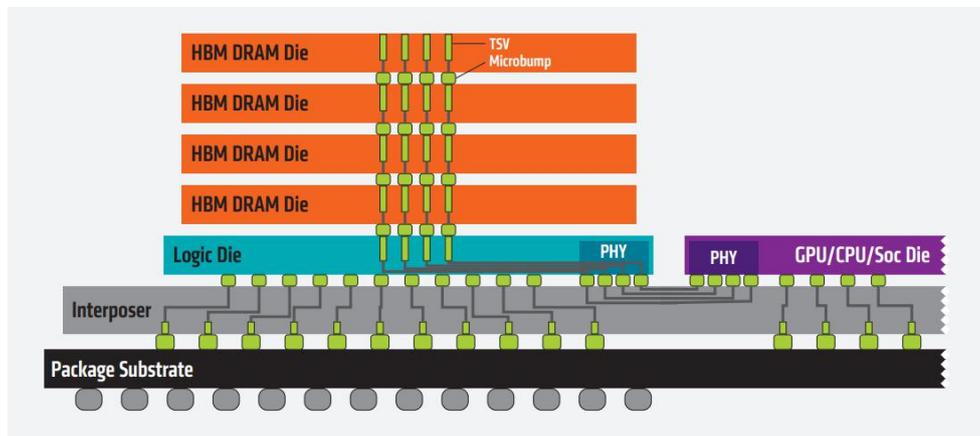


资料来源: Samsung, 山西证券研究所

2.3.2 HBM、DDR5 渗透率持续上升

HBM 作为基于 3D 堆栈工艺的高性能 DRAM，打破内存带宽及功耗瓶颈。HBM（High Bandwidth Memory）即高带宽存储器，通过使用先进封装（如 TSV 硅通孔、微凸块）将多个 DRAM 芯片进行堆叠，并与 GPU 一同进行封装，形成大容量、高带宽的 DDR 组合阵列。HBM 通过与处理器相同的“Interposer”中间介质层与计算芯片实现紧凑连接，一方面既节省了芯片面积，另一方面又显著减少了数据传输时间；此外 HBM 采用 TSV 工艺进行 3D 堆叠，不仅显著提升了带宽，同时降低了功耗，实现了更高的集成度。

图 18：HBM 内部结构图



资料来源：AMD 官网，山西证券研究所

在高性能 GPU 需求推动下，HBM 成为当前 AI 服务器的搭载标配。HBM 凭借高带宽、大容量、低延时和低功耗优势，目前已逐步成为 AI 服务器中 GPU 的搭载标配。英伟达推出的多款用于 AI 训练的芯片 A100、H100、H200 和 B200，都采用了 HBM 显存。其中，A100 和 H100 芯片分别搭载了 HBM2e 和 HBM3，H200 和 B200 芯片搭载了速率更快、容量更高的 HBM3e。AMD 的 MI300 系列也都采用了 HBM3 技术，最新的 MI325X 芯片则搭载了 HBM3e。MI250X 的容量与前一代相同为 128GB，而更高端的 MI300X 则将容量提升至 192GB，增长了 50%，相当于 H100 容量的 2.4 倍。最新的 MI325X 芯片则将容量提升到 288GB，超过了英伟达的 B200 芯片。

表 4：HBM 已经成为 AI 服务器的搭载标配

	NVIDIA				AMD		
	A100	H100	H200	B200	MI250X	MI300X	MI325X
GPU	GA100	GH100	GH100	Blackwell GPU	CDNA 2 GCD	CDNA 3 XCD	CDNA 3 XCD
Manufacturing Process	TSMC 7nm	TSMC 4nm	TSMC 4nm	4nm	6nm	5/6nm	5/6nm
Memory Capacity	80GB HBM2e	80GB HBM3	141GB HBM3e	192GB HBM3e	128GB HBM2e	192GB HBM3	288GB HBM3e
IC Package Technology	CoWoS	CoWoS	CoWoS	CoWoS	CoWoS	CoWoS	/
Memory Bus Width	5120-bit	5120-bit	6144-bit	2x 4096-bit	8192-bit	8192-bit	8192-bit
Memory Bandwidth	2TB/sec	3.35TB/sec	4.8TB/sec	8TB/sec	3.2TB/sec	5.3TB/sec	6TB/sec
Architecture	Ampere	Hopper	Hopper	Blackwell	CDNA2	CDNA3	CDNA3

资料来源：英伟达官网，AMD 官网，AnandTech，Tom's Hardware，观察者网，山西证券研究所

服务器 CPU 升级支持 DDR5，DDR5 渗透率提升，并带来内存接口芯片、配套芯片需求增长。JEDEC 的 DDR5 标准带来 DB 芯片和配套芯片数量增加，服务器需配备 SPD、PMIC 各一颗和 TS 两颗，个人电脑需配备 SPD、PMIC 甚至 CKD 各一颗。内存接口芯片上则由原先的“1+9”架构升级为“1+10”架构，带来 DB 增量需求。2022 年 11 月 AMD 发布首次支持 DDR5 的 EPYC9004 系列，随后英特尔相继发布 Sapphire Rapids、Emeralds Rapids 两代处理器，支持 PCIe5.0 和 DDR5 内存技术。新一代 CPU 出货上量可加速 DDR5 渗透率提升，预计 2024 年年中 DDR5 渗透率将超 50%，DDR5 渗透率提升有望带来内存接口芯片、配套芯片需求增长。

表 5：内存迭代带来内存接口芯片、配套芯片量的增长

应用	产品	内存接口芯片		内存模组配套芯片	
		DDR4	DDR5	DDR4	DDR5
服务器	RDIMM	RCD	RCD	-	SPD+PMIC+2*TS
	LRDIMM	RCD+9*DB	RCD+10*DB		
台式机、笔记本	UDIMM	-	-	-	SPD+PMIC+（当速率达 6400MT/S 及以上时需配备一颗）CKD
	SODIMM	-	-		

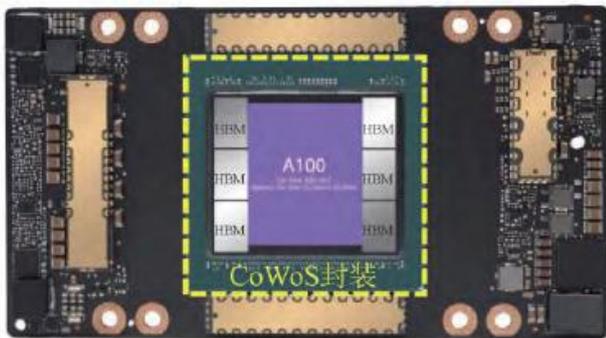
资料来源：澜起科技 2023 年半年报，山西证券研究所

AI 存储建议关注：澜起科技、聚辰股份。

2.4 算力时代，先进封装有望迎来加速发展

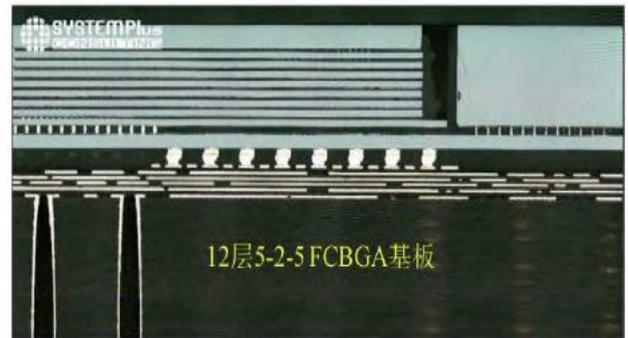
先进封装可以显著提升芯片性能。先进封装可以突破带宽瓶颈，提升芯片性能。存储器的“存储墙”限制了计算芯片性能的发挥，GDDR5 的带宽极限为 32GB/s。由逻辑芯片和多层 DRAM 堆叠而成的 HBM 技术可以突破带宽瓶颈，HBM1 和 HBM2 的带宽分别为 128GB/s 和 256GB/s，HBM3 可突破 1.075TB/s。通过先进封装，如台积电 CoWoS 技术，将 HBM 和处理器集成，可以显著提升芯片性能。英伟达从 2020 年开始采用台积电 CoWoS 技术封装其 A100 GPU 系列产品。

图 19：采用 CoWoS 封装的英伟达 A100



资料来源：《先进封装技术的发展与机遇》，山西证券研究所

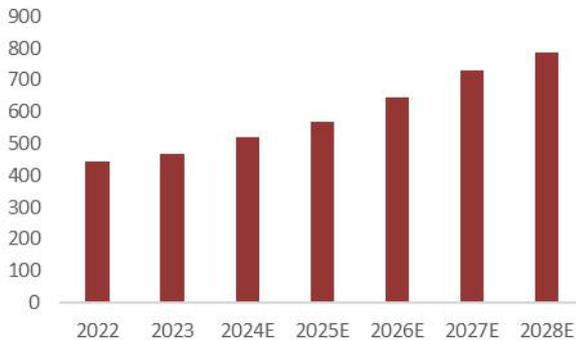
图 20：英伟达 A100 CoWoS 封装切面图



资料来源：《先进封装技术的发展与机遇》，山西证券研究所

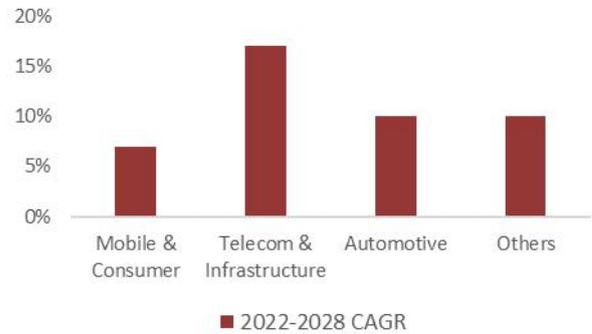
2028 年全球市场规模有望达到 785.5 亿美元。随着 AI 及 HPC 等高算力芯片对先进封装技术的需求日益提升，先进封装行业有望迎来加速发展。Yole 预计，全球先进封装市场规模有望从 2023 年的 468.3 亿美元增长到 2028 年的 785.5 亿美元。下游手机消费市场占比最大，通信基础设施领域增长最快。Yole 认为，未来手机消费领域仍是先进封装最大的市场，2028 年预计占比 61%，相较 2022 年的 70% 有所下降。从增速看，先进封装在高性能计算领域的广泛应用，使得通信基础设施成为先进封装增长最快的领域，2022-2028 年预计实现 17% 的复合增长。

图 21：全球先进封装市场规模，亿美元



资料来源：Yole 《Advanced packaging market and technology trend 2023》，山西证券研究所

图 22：通信领域 2022-2028 CAGR 17%，增速最高



资料来源：Yole 《Advanced packaging market and technology trend 2023》，山西证券研究所

2.4.1 封装环节：Foundry 与 OSAT 各有侧重，内资厂商积极布局先进封装

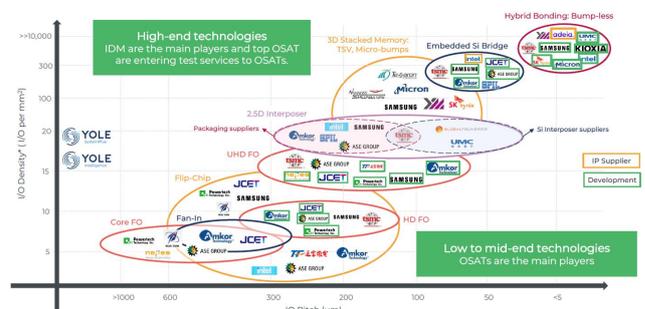
Top 10 厂商先进封装份额 89%，封测厂与晶圆代工厂各有侧重。当前，OSAT、Foundry、IDM 厂商都在大力发展先进封装。Yole 数据显示，2022 年日月光、安靠和台积电分别以 25.0%、12.4%、12.3% 份额位居先进封装市场前三；Top 5 厂商市占率 67.9%，Top 10 厂商份额达到 89%，行业高度集中。IDM、Foundry 由于在前道环节经验更丰富，能更快掌握需要蚀刻等前道步骤的 TSV 技术，在 2.5D/3D 封装、混合键合等技术方面较为领先；OSAT 更熟悉后道环节、异质异构集成，因此在 SiP、WLP 等技术相对有优势。

图 23：2022 年先进封装市场格局集中



资料来源：Yole 《Advanced packaging market and technology trend 2023》，山西证券研究所

图 24：OSAT 与 IDM、晶圆代工厂技术布局有差异



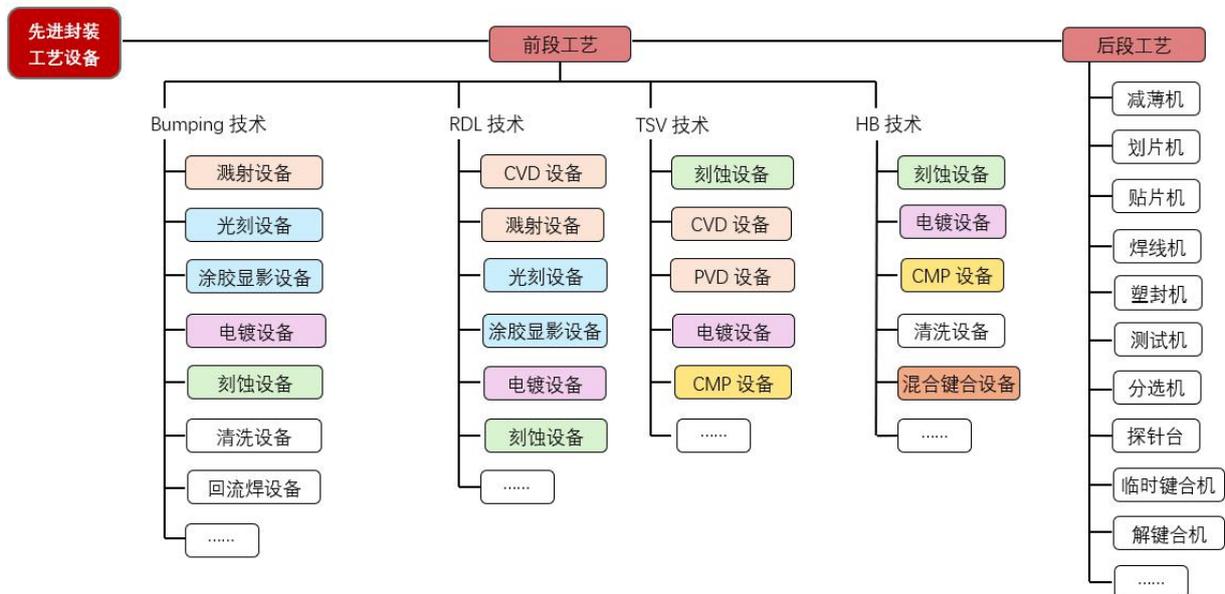
资料来源：Yole 《Advanced packaging market and technology trend 2023》，山西证券研究所

内资 OSAT 厂商积极布局先进封装。中国大陆 OSAT 厂商中，长电科技、通富微电在 2.5D、3D 等先进封装领域相对领先。长电科技推出的 XDFOI® Chiplet 高密度多维异构集成系列工艺已按计划进入稳定量产阶段，该技术涵盖 2D、2.5D、3D 集成技术。通富微电超大尺寸 2D+ 封装技术、3 维堆叠封装技术、大尺寸多芯片 chip last 封装技术已通过客户验证。此外，华天科技、晶方科技、甬矽电子等公司亦积极布局先进封装。华天科技已经具备 3D NAND Flash 32 层超薄芯片堆叠封装能力。晶方科技在 WLCSP 上技术领先，同时拥有领先的硅通孔（TSV）、WLP、Fanout 等封装技术。甬矽电子拥有成熟的 RDL 及凸点加工能力。

2.4.2 设备环节：国产设备持续突破，进口替代进程加速

先进封装核心技术提升前道设备需求。从流程看，先进封装后段工艺与传统封装相同，区别在于前段工艺引入了 Bumping、RDL、TSV 及 HB 等技术以实现连接优化，通常涉及光刻、刻蚀、电镀、CMP 等前道制造工序。其中以制作各类金属与介质薄膜的沉积设备、制作深孔的刻蚀设备、填充金属导线的电镀设备、实现电路图形的光刻工序设备、对晶圆进行减薄抛光和去除多余杂质的 CMP 设备，以及键合工序设备最为核心。

图 25：先进封装设备体系



资料来源：公开资料整理，山西证券研究所

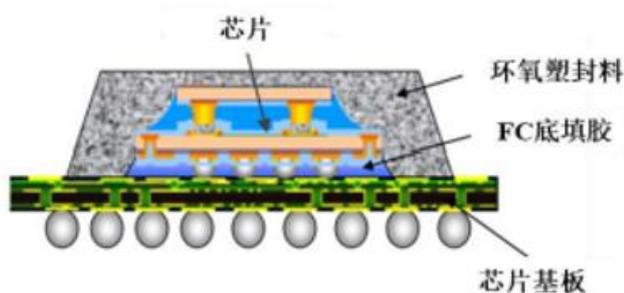
国产设备技术持续突破，进口替代进程加速。相较于先进制造，先进封装对制程节点要求

不高，随着国内前道设备厂技术持续突破，国产设备在先进封装产线的竞争力增强，进口替代进程加速。北方华创在薄膜沉积与刻蚀领域实力雄厚，TSV 刻蚀设备已广泛应用于国内先进封装厂；中微公司等离子体刻蚀设备技术领先，ICP 设备在晶圆级先进封装、2.5D 封装生产线持续获得重复订单；拓荆科技立足薄膜沉积优势积极布局混合键合市场，PECVD/ALD 在先进封装领域已实现产业化应用；芯源微先进封装用涂胶显影设备、单片式湿法设备已应用于一线大厂，并布局临时键合、解键合等新品；盛美上海在先进封装领域主要产品包括清洗设备、电镀设备和湿法设备；华海清科 CMP 设备、减薄设备是芯片堆叠与先进封装技术关键；芯碁微装直写光刻设备已经获得大陆头部先进封装客户连续重复订单。

2.4.3 材料环节：关键材料性能要求升级，高端品类国产化空间巨大

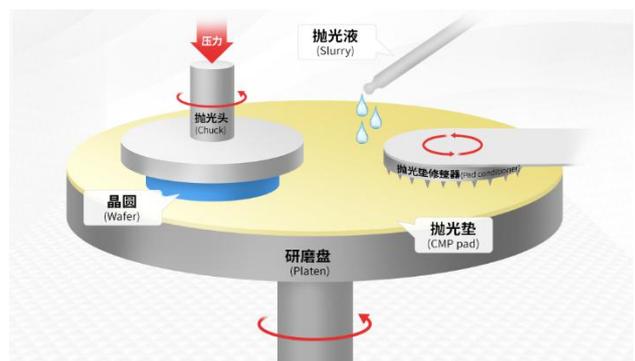
高端材料是先进封装实现性能提升的关键支撑。先进封装凸点制备与金属布线带来电镀液增量需求，铜互连是目前电镀材料最大的细分市场；三维堆叠立体结构对包封材料保护性和底部填充料接合可靠性提出更高要求，环氧塑封料是先进封装保护材料主要选择，硅微粉则是其重要原材料；垂直互联与异质集成提升 CMP 耗材需求，铜暴露及晶圆表面平坦化直接影响电学连接性能；前道工艺引入和 TGV 对 TSV 的低成本替代路线则分别带来光刻材料和玻璃基板的需求增量。

图 26：环氧塑封料是先进封装主要保护材料



资料来源：华海诚科招股书，山西证券研究所

图 27：垂直互联与异质集成提升 CMP 耗材需求



资料来源：SK hynix，山西证券研究所

海外企业长期垄断，国产化空间巨大。先进封装用高端材料长期由美国、日本、德国等企业占据市场主导地位，国产材料已部分实现技术覆盖或导入应用，替代潜力巨大。电镀液环节，上海新阳、艾森股份已实现供应并积极导入新品；抛光材料环节，安集科技 TSV 和 HB 工艺

用多款抛光液和清洗液已作为首选供应进入客户产线，**鼎龙股份**抛光垫产品已取得封装厂客户订单或进入验证测试阶段；光刻材料环节，**艾森股份**先进封装用 g/i 线负性光刻胶已实现批量供应，**鼎龙股份**封装 PI 拟在 2024 年内完成验证并开始导入；保护材料环节，**华海诚科**布局先进封装用环氧塑封料和底填胶，**联瑞新材**打破海外对电子级球型硅微粉技术垄断；玻璃基板环节，**沃格光电** TGV 载板已通过客户验证。

先进封装建议关注：长电科技、通富微电、顾中科技、汇成股份、北方华创、中微公司、拓荆科技、芯源微、鼎龙股份、华海诚科等。

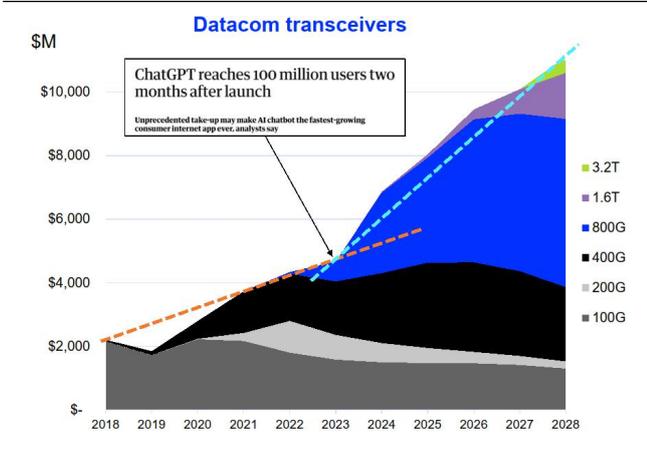
3. AI 网络：光铜共进迎接 scaleup 和 scaleout 能力升级

展望 2024 下半年和 2025 年，我们看到 AI 网络三大主要趋势：一是以太网占比将显著提升，带动以太网交换机和光模块市场规模快速提升；二是 scaleup 能力持续迭代，铜连接成为 GB200 配套最大增量，同时以 AMD 为引领的 UALink 和以华为昇腾为引领的国产算力具备想象空间；三是 scaleout 能力从千卡、万卡到百万卡规模演进，短期看光模块配比有望再增加，未来 800G 向 1.6T、3.2T 演进价值将显著提升。

3.1 光模块：海外需求持续性无忧，国内市场有望快速启动

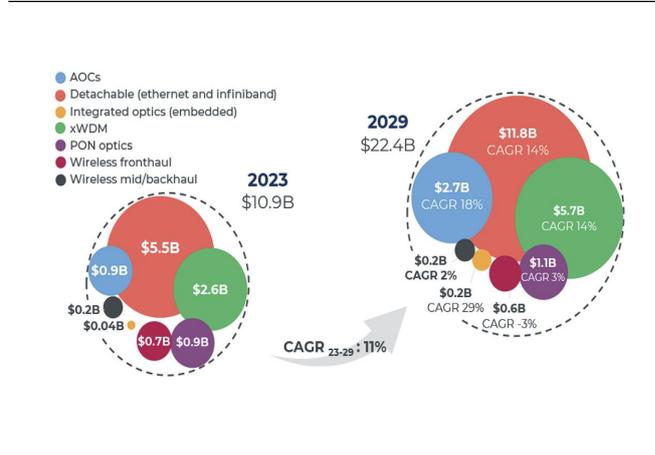
AI 浪潮继续，2025 需求无忧，同时光模块迭代周期缩短，强者将恒强。英伟达 FY25Q1 业绩超预期，并认为所有多模态应用、大量的 AI 创业公司、主权人工智能需求仍然“难以置信”，公司的订单展望乐观。Marketscreen 对于英伟达 FY2026、FY2027 的收入一致预期分别为 1574 亿（yoy +31%）、1823 亿（yoy +16%）美金，表明海外投资者对人工智能浪潮持续时间的乐观预期。Marvell 在今年 4 月的 AI Day 表明了对 AI 光互联市场的看好：光模块端口 double 节奏由 4 年缩短为 2 年，同时光模块增速或超过 XPU，这主要是因为 AI 集群规模增大带来更多层数的光模块配比要求（10 万 XPU 需要 50 万光模块，而百万 XPU 需要千万）。Coherent、YOLO 的最新观点均对数据中心光模块市场增长展望乐观，YOLO 预测全球数通光模块将在 2029 年达到 224 亿美金。展望未来，英伟达 tik-tok 的芯片节奏将光模块带宽翻倍时间缩短到 2 年，高速光模块研发成本愈来愈高，与客户的联合研发或更早进行，客户更换供应商的动机减弱，高端产品“保值期”延长，光模块龙头厂商有望保持较好的盈利能力。

图 28: coherent 认为未来几年数通光模块市场将快速增长



资料来源: coherent《Analyst Briefing at OFC 2024》, 山西证券研究所

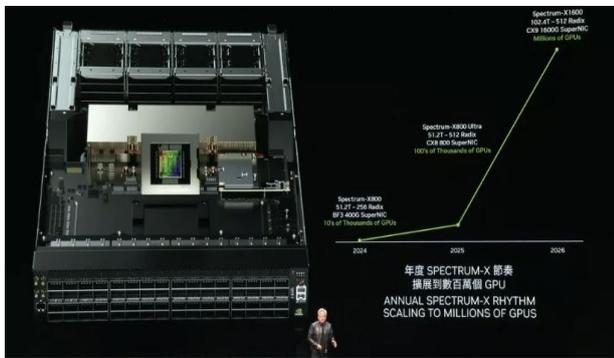
图 29: YOLE 最新预测全球数通光模块将在 2029 年达到 224 亿美金



资料来源: YOLE, 山西证券研究所

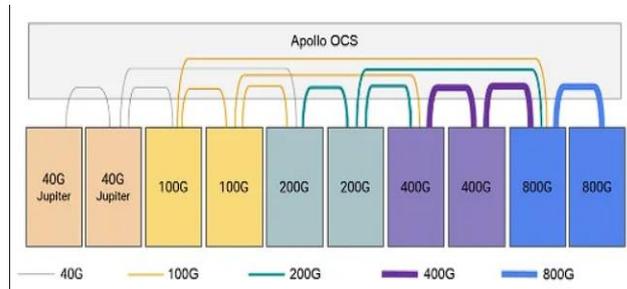
关注 1.6T 进展。英伟达在 computex2024 上宣布 2025 年将宣布量产支持 1.6T 端口的 CX8 网卡以及 X800 交换机系列以支持十万卡级 AI 集群的组网需求。谷歌在 5 月的 I/O 2024 开发者大会上宣布了第六代 TPU-Trillium, 芯片算力较上代提升 4.7 倍, 其早些在 AI 网络中部署的 OCS 光交换机因可兼容不同速率光模块, 使用寿命较电交换机或大幅延长, 同时也有望在 2025 年带来 1.6T 光模块的配套需求。我们认为 2025 年 1.6T 光模块将迎来快速上量, 主要供应商或集中于研发和产能储备领先的头部公司。中际旭创表示 1.6T 海外重点客户已进入送样测试环节, 预计下半年完成认证并下单; 新易盛表示已在 OFC2024 推出了基于单波 200G 的 EML/硅光/薄膜铌酸锂 1.6T 光模块。

图 30：英伟达将在 2025 全面量产支持 1.6T 端口的网卡和交换机



资料来源：新智元，山西证券研究所

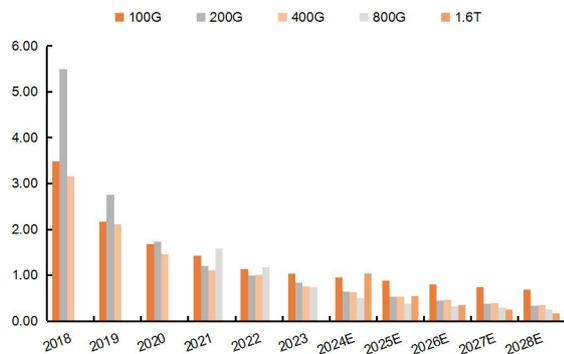
图 31：谷歌 OCS 兼容不同速率可快速适配 TPUv6 对更高速率端口要求



资料来源：Semianalysis，山西证券研究所

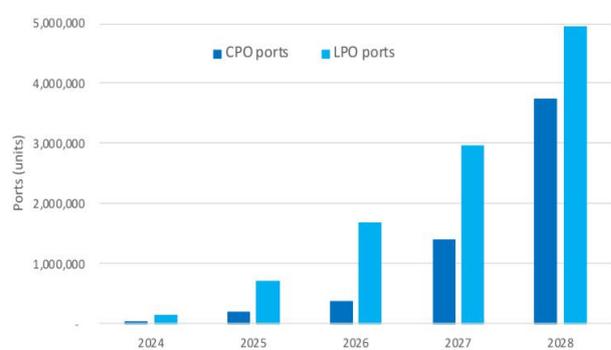
关注以太网 AI 光模块的变化：成本优化、客户多元化、LPO&LRO 渗透率提升。英伟达在 computex2024 上表示未来将同步推广 Quantum IB 交换机和 SpectrumX 以太网交换机系列，希望以更多元化的产品组合满足不同客户需求。我们认为以太网在 AI 网络中占比将逐年提升，其主要优势除了更开放的生态和多元供应商带来的成本优势还有对云客户多租户环境的更友好适配。LightCounting 预计至 2025 年数据中心 800G 光模块的总和单位带宽成本将下探 0.38 美金/Gbps, 而 2027 年 1.6T 的单位带宽成本又将较 800G 有优势。此外, 受益于以太网 LPO&LRO 产品在简化供应链、降低时延方面的优势, LPO 出货量或将在 2025 显著增长。

图 32：以太网光模块单位带宽 (Gbps) 价格 (美元) 变化



资料来源：LightCounting，山西证券研究所

图 33：LPO 出货量预计在 2025 年明显增长



资料来源：LightCounting，山西证券研究所

关注数据中心网络扩容和新建投资：高密度 MPO 布线和跳线的需求。 AI 网络对高密度布线提出了要求，以英伟达为例，其 OSFP 800G 光模块采用 2*MPO12 接口，爆发的 400G/800G 出货量正快速消耗数据中心存量预制布线端口。由于存量上架率饱和以及电力容量限制，以北美为代表正涌现新建 AI 数据中心需求，光纤预置布线将受益于 AI 资本开支的增长。北美光纤布线龙头康宁表示，GPU 机柜的平均光纤数较 CPU 机柜增加了 8 倍，价值量为每 GPU 数百美金。数据中心光网络布线的上游主要为光纤、跳线、机架以及陶瓷插芯，目前北美集成商的跳线供应商主要来自国内代工，而陶瓷插芯主要来自 UScontec、senko 等日美巨头。

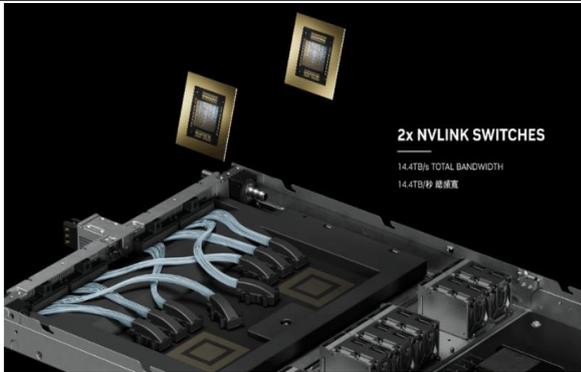
关注国产算力光模块配套厂商。 目前英伟达在 AI 交换机和光模块配置上具有领先一代的优势，国产算力卡组网主要采用以太网交换机和商用网卡，目前以 200G、400G 为主。随着国产算力集群由千卡往万卡集群发展，光模块配比数量或提升。根据星融元对国内万卡 GPU 集群互联的统计，国内传统万卡集群光模块与 GPU 比例达到 1:6。除了光模块配比提升，我们认为国产大芯片市占率的提升也将带动配套光模块需求增加。

光模块建议关注：新易盛、中际旭创、天孚通信、源杰科技、华工科技、铭普光磁等。

3.2 铜互联：英伟达引领短距高密度互联新趋势，市场空间广阔

英伟达 NVL72 高密度铜背板互联成亮点，铜连接或成为在百卡层级 HB domain 互联领域性价比最优选择。 英伟达 Blackwell 发布将机柜级产品 NVL72&NVL36 作为了旗舰主推系列，其最有吸引力的卖点之一在于将 72 个 Blackwell 芯片进行了内存统一互联，整体可作为一颗超级芯片进行编程应用。根据英伟达在 GTC2024 的介绍，NVL72 创新性的使用了高密度铜跳线将芯片 IO 引出到背板，再通过高速铜线背板将 18 个计算板互联到 9 个交换板上，其使用的电缆长度累计接近 2 英里，共有约 5000 条独立电缆。根据 the next platform，与 PCB 相比，使用 Cable（铜线）的单位功耗（J/b）相当但距离是 PCB 的 10 倍（可达 5 米），有效解决了 NVL72 机柜范围通信难题；同时与光模块 AOC 相比，Cable 的功耗仅为一般，同时成本是 AOC 的十分之一。因此面临 PCB 上高速信号提升的损耗、有限的传输距离以及机柜整体互联的需求，铜线互联成为当前百卡级别互联性价比优选。进一步推演，未来若 scaleup 全互联的需求延展至百卡以上至千卡规模，我们认为 CPO 或也将登上历史舞台，使用何种通信方式关键看当期性能和成本的平衡。

图 34: NVL72 Switch tray 使用的芯片直出跳线



资料来源: 新智元, 山西证券研究所

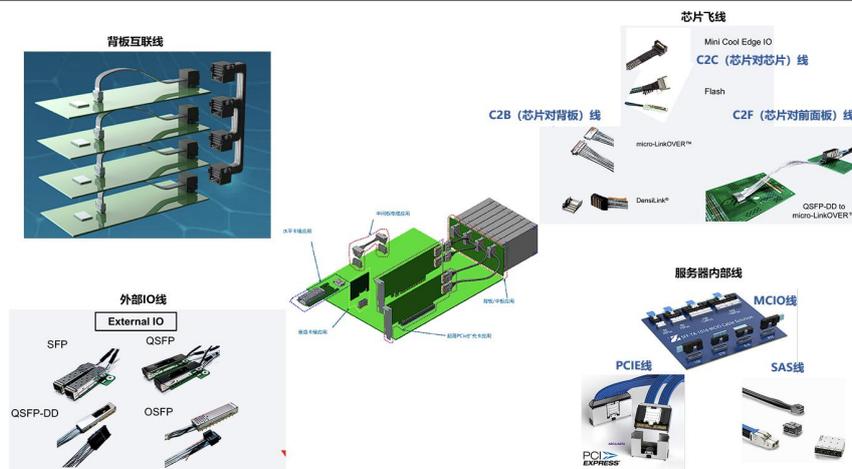
图 35: NVL72 使用的高速背板 cartridge



资料来源: 英伟达, 山西证券研究所

铜连接主要分为高速外部 IO、背板互联、芯片直出跳线、服务器内部线等，随 AI 服务器市市场均快速增长。随着市场对 NVL72 铜连接方案重视度提升，市场研究重点逐渐扩散到服务器内部所有铜连接解决方案。以安费诺、TE、Molex 为代表的国际连接器巨头已建立了完整的铜连接产品系列，我们梳理后看到，服务器使用的高速铜线通信主要在背板互联线、芯片飞线、外部 IO 线、服务器内部线四大场景，NVL72 主要使用的是芯片飞线和背板互联线，解决了 224G SERDES 信号在 PCB 上传输距离缩短、功耗增加、长距离背板互联难题。同时配套 Blackwell 的 CX8 网卡和 X800 系列交换机，外部 IO 线也有望升级至 1.6T DAC/ACC，服务器内部线升级为 PCIE6，价值量或均有提升。

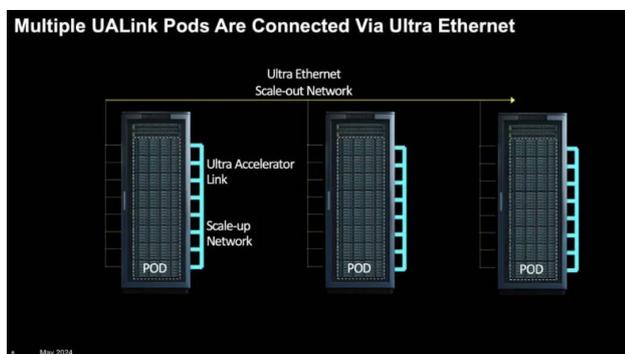
图 36: 服务器使用到的各种高速通信线类型



资料来源: 安费诺, TE, samtec, 兆龙互连, 山西证券研究所

高速铜连接市场有望从英伟达引领扩散到海外 UALINK 和国产配套。根据立讯精密立讯对 NVL72 整体解决方案的描述，NVL72 单机柜铜光连接价值量超 200 万元，包括电连接、光连接、电源管理、散热等，预计总市场规模达到千亿元。铜连接作为 Blackwell 最显著的增量产品有望“复刻”光模块行情，2025 年市场需求或爆发增长。NVL72 更大的意义在于引领 scaleup 通信技术发展，海外 UALINK 以及国内智算集群均有望跟进。5 月底，英特尔、AMD、博通、思科、谷歌、HPE、Meta 和微软宣布建立 UALink 推广工作组，以指导数据中心 AI 加速器芯片之间连接组件的发展，希望未来可以取代 NVLink 接口。UALink 1.0 规范将支持多达 1024 个加速器内存统一互联，虽具体实现方式仍未知，参考 NVL72 的铜连接架构设计仍不失为成熟方案。国内方面，中国移动编制的《面向超万卡集群的新型智算技术白皮书》倡议加速推进超越 8 卡的超节点形态服务器，优化 GPU 卡间互联协议实现通信效率跃升，可以期待国内 AI 大芯片在 scaleup 互联技术也在酝酿更大的动作。以华为为例，其 2022 年底即推出的“天成”多样算力平台旨在设计更高的算力密度，超节点形态服务器设计将是下一步工作重点。

图 37: UALINK 拓展通用 scaleup 协议



资料来源：云头条，山西证券研究所

图 38: 华为“天成”机柜级算力平台产品



资料来源：华为，山西证券研究所

海外连接器巨头配套：高速裸线、CAGE 代工将受益于产能扩张和价值量提升。由于与英伟达的联合研发以及对于核心专利的掌握，GB200 高速铜连接前期价值量或将主要集中于以安费诺为代表的国际连接器巨头厂商。以国内为例，集中了安费诺最大的信息通信产品线配套产能，其对于 224G 高速线、cage 结构件等高速产品配套需求或增加，同时对于中低端产品线的产能外包需求也将外溢。安费诺国内合作伙伴包括乐庭智联（沃尔核材）、神宇股份、鼎通科技、奕东科技等，以沃尔核材为例，根据其公告，高速通信线订单需求在不断增长，已下单采购了 70 多台绕包机和多台芯线机以进一步满足产能需求，可预见未来由产能提升和产品

价值量提升带来的收入增长。

国产替代：关注华为产业链为引领的高速产品突破。华为具有 AI 大芯片、交换机、高速 SERDES 等多重技术储备，且具备集群产品设计制造能力，或率先在 scaleup 超节点互联技术取得突破，引领国内高速背板连接器、铜互联，IO 连接器等配套厂商发展。其中，在高速背板连接器领域，国内已形成了以华丰科技、庆虹电子、中航光电为主的格局，具有技术含量高、投入大、制造难度大、竞争格局好的特点。而在高速 IO 连接器领域，随着端口由 200G、400G 向 800G 升级，单位价值量有望显著提升，意华股份、中航光电、立讯精密、方向电子等厂商均有望受益。

铜互联建议关注：沃尔核材、神宇股份、鼎通科技等。

4. 风险提示

AI 发展不及预期风险：目前大模型 AI 算力仍在早期阶段，但如果后续大模型代际演进放缓或者应用落地效果不及预期，AI 算力需求可能降低，并打破算力投资增长的线性推演。

外部制裁升级风险：AI 手机、AIPC 等端侧应用依赖高性能的 AI 芯片，如果美国等国家加大对 AI 芯片的制裁，那可能会限制国内 AI 手机、AIPC 等端侧应用发展和创新。

新技术发展不及预期风险：液冷、HBM、先进封装等技术领域面临研发难度大、技术突破缓慢等问题，若新技术研发、验证、推广进度不及预期，则可能对相关企业收入、盈利造成影响。

宏观经济增长乏力风险：宏观经济增长乏力导致国内政府、运营商、地方智算中心以及互联网企业算力投资不达预期，同时经济环境的不确定性也将影响算力投资的实际回报，对国产算力发展产生负面影响。

分析师承诺：

本人已在中国证券业协会登记为证券分析师，本人承诺，以勤勉的职业态度，独立、客观地出具本报告。本人对证券研究报告的内容和观点负责，保证信息来源合法合规，研究方法专业审慎，分析结论具有合理依据。本报告清晰地反映本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点直接或间接接受到任何形式的补偿。本人承诺不利用自己的身份、地位或执业过程中所掌握的信息为自己或他人谋取私利。

投资评级的说明：

以报告发布日后的 6--12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。

无评级：因无法获取必要的资料，或者公司面临无法预见的结果的重大不确定事件，或者其他原因，致使无法给出明确的投资评级。（新股覆盖、新三板覆盖报告及转债报告默认无评级）

评级体系：

——公司评级

- 买入： 预计涨幅领先相对基准指数 15%以上；
- 增持： 预计涨幅领先相对基准指数介于 5%-15%之间；
- 中性： 预计涨幅领先相对基准指数介于-5%-5%之间；
- 减持： 预计涨幅落后相对基准指数介于-5%- -15%之间；
- 卖出： 预计涨幅落后相对基准指数-15%以上。

——行业评级

- 领先大市： 预计涨幅超越相对基准指数 10%以上；
- 同步大市： 预计涨幅相对基准指数介于-10%-10%之间；
- 落后大市： 预计涨幅落后相对基准指数-10%以上。

——风险评级

- A： 预计波动率小于等于相对基准指数；
- B： 预计波动率大于相对基准指数。

免责声明：

山西证券股份有限公司(以下简称“公司”)具备证券投资咨询业务资格。本报告是基于公司认为可靠的已公开信息，但公司不保证该等信息的准确性和完整性。入市有风险，投资需谨慎。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，公司不对任何人因使用本报告中的任何内容引致的损失负任何责任。本报告所载的资料、意见及推测仅反映发布当日的判断。在不同时期，公司可发出与本报告所载资料、意见及推测不一致的报告。公司或其关联机构在法律许可的情况下可能持有或交易本报告中提到的上市公司发行的证券或投资标的，还可能为或争取为这些公司提供投资银行或财务顾问服务。客户应当考虑到公司可能存在可能影响本报告客观性的利益冲突。公司在知晓范围内履行披露义务。本报告版权归公司所有。公司对本报告保留一切权利。未经公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯公司版权的其他方式使用。否则，公司将保留随时追究其法律责任的权利。

依据《发布证券研究报告执业规范》规定特此声明，禁止公司员工将公司证券研究报告私自提供给未经公司授权的任何媒体或机构；禁止任何媒体或机构未经授权私自刊载或转发公司证券研究报告。刊载或转发公司证券研究报告的授权必须通过签署协议约定，且明确由被授权机构承担相关刊载或者转发责任。

依据《发布证券研究报告执业规范》规定特此提示公司证券研究业务客户不得将公司证券研究报告转发给他人，提示公司证券研究业务客户及公众投资者慎重使用公众媒体刊载的证券研究报告。

依据《证券期货经营机构及其工作人员廉洁从业规定》和《证券经营机构及其工作人员廉洁从业实施细则》规定特此告知公司证券研究业务客户遵守廉洁从业规定。

山西证券研究所：

上海

上海市浦东新区滨江大道 5159 号陆家嘴滨江中心 N5 座 3 楼

太原

太原市府西街 69 号国贸中心 A 座 28 层
电话：0351-8686981
<http://www.i618.com.cn>

深圳

广东省深圳市福田区林创路新一代产业园 5 栋 17 层

北京

北京市丰台区金泽西路 2 号院 1 号楼丽泽平安金融中心 A 座 25 层

