2026 年 04 月 09 日

计算机行业研究

买入（维持评级）

行业专题研究报告

证券研究报告

计算机组
分析师：刘高畅（执业 S1130525120005） liugaochang@gjzq.com.cn
分析师：陈芷婧（执业 S1130525120008） chenzhijing@gjzq.com.cn
分析师：鲍淑娴（执业 S1130526020002） baoshuxian@gjzq.com.cn

国产机柜时间到来

本周观点

- **超节点产业进入规模落地期。**1）大模型训练与推理对 Scale-up 域刚性需求显现。大模型万亿参数与 MoE 架构确立新常态，专家并行（EP）带来的高频 All-to-All 通信对互联带宽提出极致要求。传统 Scale-out 集群面临通信、功耗与复杂度的物理硬约束，在此背景下，具备大带宽与内存统一编址能力的 Scale-up 超节点正式成为新一代智算基建的核心底座。2）供给端：海内外厂商超节点产品密集发布。产业供给端已步入密集兑现期，海外英伟达 GB200 NVL72 整机架处于满负荷运行状态，并向 NVL576 加速演进；国内阵营全面起势，华为 CM384 进入工程化交付阶段，中科曙光（scaleX640/40）、百度（天池 256/512）、阿里（磐久 128）等头部厂商的机柜级超节点产品相继发布或商用。供需双侧强逻辑共振，印证超节点产业已跨越技术验证期，全面开启规模化落地。
- **集群化交付推动 ODM 厂商毛利率结构性抬升。**交付单元从单台白盒服务器跃迁为整机柜乃至 Pod 级系统交付，整机柜在供电架构、散热、互联三个维度均形成显著工程壁垒，使超节点的研发与交付成为对厂商技术、运维、工程能力的综合性挑战。当交付复杂度提升、合格供应商稀缺时，头部 ODM 具备向客户转移工程溢价的实际能力，毛利率有望向更高区间迁移。
- **超节点整机柜的设计需解决高密度 GPU 协同工作的挑战，多环节价值量有望显著提升。**核心设备厂商正依托深厚的技术卡位，迈入业绩兑现周期：1）浪潮信息双线布局超节点矩阵，先后发布元脑 SD200 与 CRS6000S 服务器，实现单机柜 32/64 张本土 AI 芯片的超高密部署，其构建的 32 卡 Scale-up 高速互联域将卡间通信带宽大幅提升 8 倍，推动千亿参数大模型训练周期由月级实质性压缩至周级。此外，公司牵头组建超节点创新联合体打造“北京方案”，联合产业链攻坚互连协议与应用落地，实质性打通超节点规模化部署的“最后一公里”；2）华勤技术凭借计算、网络与液冷散热的全栈设计壁垒，其超节点项目预计于今年第二季度开始发货并于下半年步入规模化交付，2026 全年相关收入指引超百亿元，加速确立行业领先身位。

相关标的

AI 机柜：浪潮信息、华勤技术、中科曙光、紫光股份等；
交换芯片：盛科通信、锐捷网络等；
高速连接器：华丰科技等；
国内算力：寒武纪、海光信息、东阳光、利通电子、协创数据、网宿科技、优刻得、豫能控股、润泽科技、亿田智能、华丰科技、神州数码、云天励飞、大位科技、润建股份、科华数据、中芯国际、华虹半导体、禾盛新材、奥飞数据、首都在线、云赛智联、瑞晟智能、潍柴重机、欧陆通等。

风险提示

- 行业竞争加剧的风险；技术研发进度不及预期的风险；特定行业下游资本开支周期性波动的风险。

内容目录

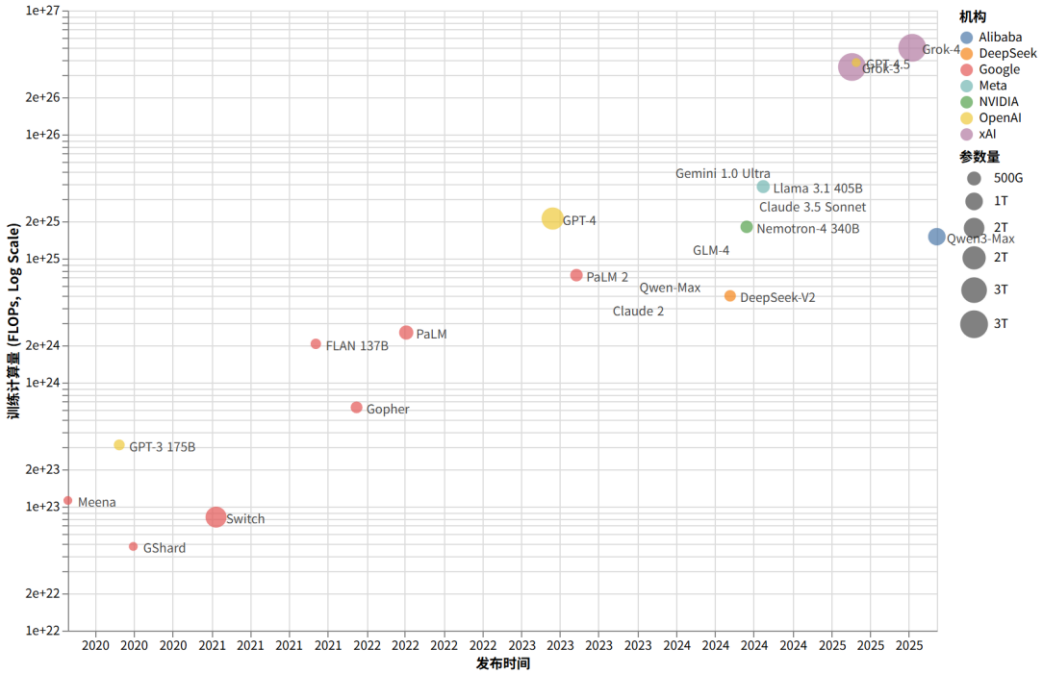
图表目录

一、超节点产业进入规模落地期

1.1 需求端：大模型训练与推理对 Scale-up 域刚性需求显现

万亿参数成为标配，大模型持续沿 Scaling Law 演进。1）参数规模方面：2020 年 GPT-3 参数规模为 1750 亿，2023 年 GPT-4 参数规模跃迁至约 1.8 万亿，2025 年 Llama-4、Kimi K2、xAI Grok4 等模型将万亿级参数与万卡级集群规模确立为新常态，Claude Mythos 5 参数量进一步突破 10 万亿级。2）数据量方面：Qwen 2.5-Max 训练数据量超过 20 万亿 token，DeepSeek-V3 训练数据量约 14.8 万亿，AI 模型大模型面向的任务场景日趋丰富，数据量加码下泛化能力得以提升。参数规模与数据量双向增长下，2020-2025 年大模型训练算力需求激增近 1000 倍，模型参数量的增长带动训练训练成本倍数级增长，前沿模型训练成本每年约 2－3 倍增长，至 2027 年或超 10 亿美元，对基础设施端提出挑战。

图表1：顶级生成式 AI 模型训练算力需求演进趋势（2020-2025）



来源：《超节点技术体系白皮书》，国金证券研究所

MoE 架构全面普及，推高高频 All-to-All 通信需求。2024-2025 年以 DeepSeek-V2/V3 为代表的 MoE 架构兴起，目前 GLM-5.1、Kimi K2.5、Gemini 3.1 Pro、Llama 4 等前沿大模型均基于 MoE 架构。MoE 架构引入的专家并行（EP）要求高频 All-to-All 通信，即每个 token 需要被路由到少量专家，且路由结果高度动态，流量模式呈现不可预测的稀疏多播特征，推高算力需求的同时，对互联带宽、延迟和动态调度能力提出了远超传统架构的要求。

图表2：MoE 架构对互联带宽、延迟和动态调度能力的要求提升

维度	Dense/TP 时代	MoE/EP 时代
主要通信模式	AlReduce, AllGather	A1l-to-A1(动态路由)
流量可预测性	高(同步、对称)	低(稀疏、动态、非对称)
带宽敏感性	高(梯度同步)	极高(每层每 token 均需路由)
延迟敏感性	中(可隐藏)	极高(EP 路由在关键路径上)
尾延迟容忍度	可接受	极低(一个慢专家拖慢全局)
对交换芯片的要求	聚合带宽	聚合带宽+动态负载均衡+低排队时延
维度	Dense/TP 时代	MoE/EP 时代
主要通信模式	AlReduce, AllGather	A1l-to-A1(动态路由)
流量可预测性	高(同步、对称)	低(稀疏、动态、非对称)

来源：超节点技术体系白皮书，国金证券研究所

传统服务器集群面临通信墙、功耗墙、复杂度墙三大瓶颈：千亿级模型一次梯度同步即 TB 级数据，传统以太网难以承受；高密度 AI 芯片推高功耗，传统风冷无法满足散热需求；万级处理器带来故障常态化，放大传统集群运维复杂度。传统服务器集群基于 Scale-out 架构，通过通用以太网连接大量标准化服务器，张量并行（TP）、流水并行（PP）

和序列并行（CP）产生大量跨节点网络通信，跨服务器的带宽与时延成为根本瓶颈，制约模型训练效率提升。

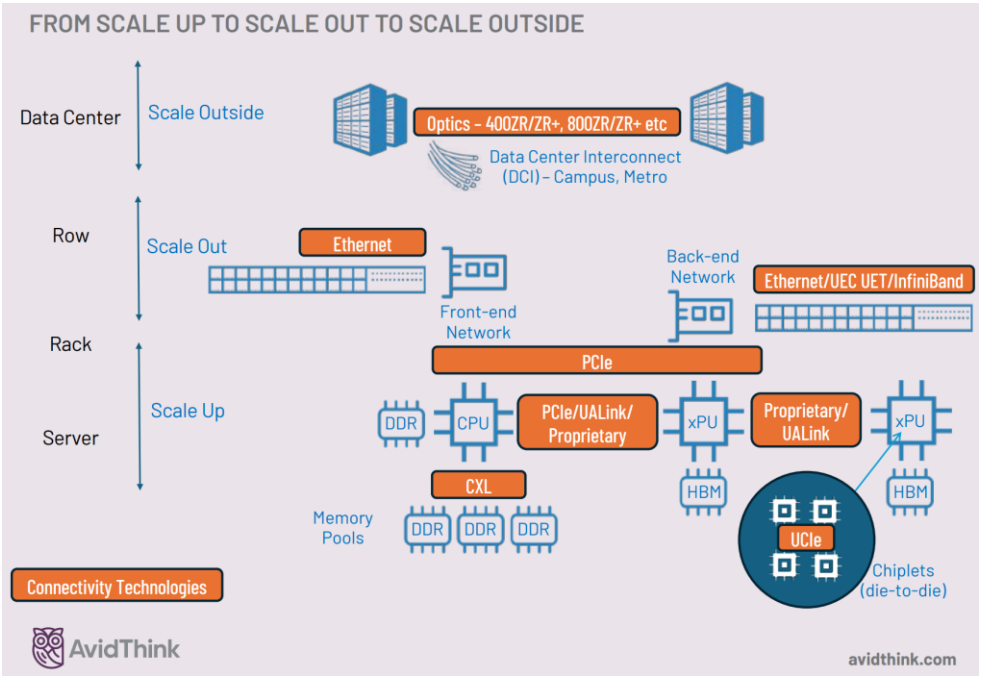
Scale-up 超节点突破 Scale-out 计算瓶颈，成为新一代算力基础单元。超节点基于 Scale-up 理念，将多颗 AI 处理器通过大带宽总线互联形成一个逻辑上的超节点，实现计算资源的紧密耦合和高效协同，天然适配 MoE 通信密集型模型架构，且基于 Scale UP 总线的超节点，大模型训练速度可提升 20%以上。超节点提供大带宽、低时延的互联能力，内存统一编址可避免传统架构中跨节点内存访问的性能骤降问题，同时标配液冷、集中供电、RMC 管理，使得数十万亿参数训练成为可能。

图表3：超节点解决传统服务器集群面临的“三堵墙”问题

维度	传统服务器集群	超节点
通信墙	千亿级模型一次梯度同步即 TB 级数据，传统以太网难以承受	超节点能够提供大带宽、低时延的互联能力，支持更大规模 AI 处理器的高效协同，实现更大范围、更高流量的数据传输
功耗与散热墙	为破通信墙而提升密度，促使液冷、48V 供电成为标配	部署液冷散热技术，精确控制核心器件工作温度，降低因热应力导致的故障风险
复杂度墙	万级处理器带来故障常态化，任一组件或一次光电转换失败都会放大为全局可用度/利用率问题	采用逻辑切分技术，赋予系统精准定位和隔离故障的能力，大幅降低故障扩散风险

来源：《超节点发展报告》，国金证券研究所

图表4：数据中心网络从 Scale-out 阶段的“服务器堆叠”升级至 Scale-up 阶段的“超节点”



来源：《Prospecting for Performance: Data Center Networking in 2025》，国金证券研究所

1.2 供给端：海内外厂商超节点产品密集发布

NVIDIA 为服务器到机柜级计算机的先行者。回顾从 Volta 到 Rubin 系列的演进，NVIDIA 的技术战略非常清晰：通过算力、互联、存储和封装等多个维度的协同创新，实现系统层面的指数级性能增长。2020 年，NVIDIA 在其推出的 HGX-A100 系统中，通过第二代 NVSwitch 将两个八卡 A100 以背板方式连接，构成一个 16 卡系统；2022 年，随 Hopper 架构推出的第三代 NVSwitch 支持更灵活的组网方式，能够实现 32 颗 GH200（32xGPU）的互联（NVL32）；2024 年 Blackwell 发布时，第四代 NVSwitch 能够实现 36 个 GB200 超级芯片（共 72 颗 GPU）的互联（NVL72），并形成 DGX GB200 SuperPod 等机柜级产品形态；未来的 VeraRubin 系列将进一步实现更大规模的互联。

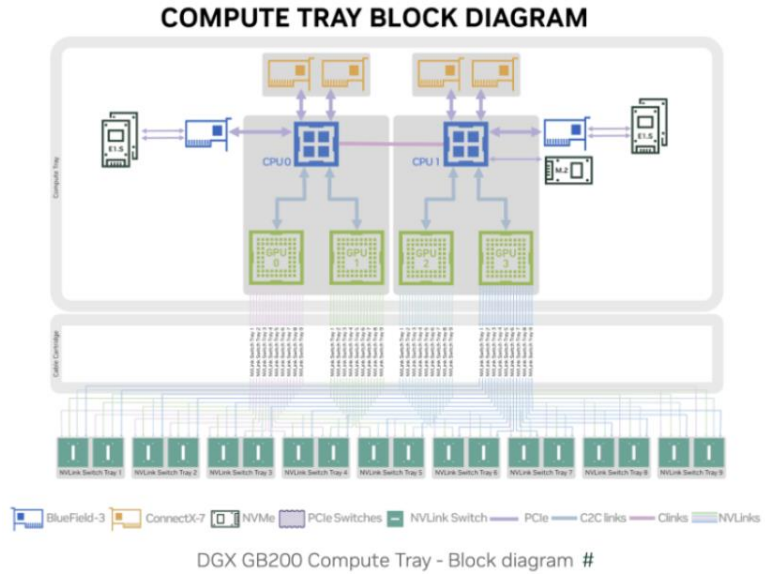
图表5：Hopper 与 Blackwell 两代 GPU 所对应的机柜级 Scale-Up 形态

参数	NVL32	GH200 SuperPod	NVL72	GB200 SuperPod
架构	Hopper	Hopper	Blackwell	Blackwell
HBM 大小	32 x 144GB = 4.6 TB	256 x 96GB = 24.5 TB	36 x 384GB = 13.8 TB	288 x 384GB = 110 TB
LPDDR5X 大小	32 x 480GB = 15.4 TB	256 x 480GB = 123 TB	36 x 480GB = 17.3 TB	288 x 480GB = 138 TB
HBM 带宽	3.35 TB/s	4.8 TB/s	8 TB/s	8 TB/s
FP16 (FLOPS)	32 PetaFLOPS	256 PetaFLOPS	180 PetaFLOPS	1440 PetaFLOPS
INT8 (OPS)	64 PetaOPS	64 PetaOPS	360 PetaOPS	2880 PetaOPS
FP8 (FLOPS)	64 PetaFLOPS	64 PetaFLOPS	360 PetaFLOPS	2880 PetaFLOPS
FP6 (FLOPS)	N/A	N/A	360 PetaFLOPS	2880 PetaFLOPS
FP4 (FLOPS)	N/A	N/A	720 PetaFLOPS	5760 PetaFLOPS
GPU-GPU 带宽	0.9 TB/s	0.9 TB/s	1.8 TB/s	1.8 TB/s
NVSwitch	Gen3 64 Port	Gen3 64 Port	Gen4 72 Port	Gen4 72 Port
NVLink 带宽	36 x 0.9 TB/s = 32 TB/s	256 x 0.9 TB/s = 230 TB/s	72 x 1.8 TB/s = 130 TB/s	576 x 1.8 TB/s = 1 PB/s
Ethernet 带宽	16 x 200 Gb/s	256 x 200 Gb/s	18 x 400 Gb/s	576 x 400 Gb/s
IB 带宽	32 x 400 Gb/s	256 x 400 Gb/s	72 x 800 Gb/s	576 x 800 Gb/s
GPUs Power	32 x 1 kW = 32 kW	256 x 1 kW = 256 kW	36 x 2.7 kW = 97.2 kW	Not provided

来源：《超节点技术体系白皮书》，国金证券研究所

GB200 引发范式转移，标志互连结构从单服务器内部进一步扩展到整机架一级域。从 GB200 NVL72 的架构来看，在交换平面一侧，NVL72 机架包含 9 个 NVLink switch trays，每个 switch tray 含 2 个 NVLink switch chips，因此整机架共有 18 个 switch chips。每个 B200 GPU 具有 18 条 NVL5 links，并采用“每个 GPU 对每个 switch chip 1 条链路”的连接方式。换言之，单个 GPU 会连接机架中的全部 18 个 switch chips；相应地，单个 switch chip 会连接整机架全部 72 个 GPU，各 1 条链路。GB200 NVL72 及后续将扩展的 576GPU，标志着互联结构从单节点 NVSwitch 走向机架级 NVLink 域。

图表6：GB200 NVL72 连接结构示意

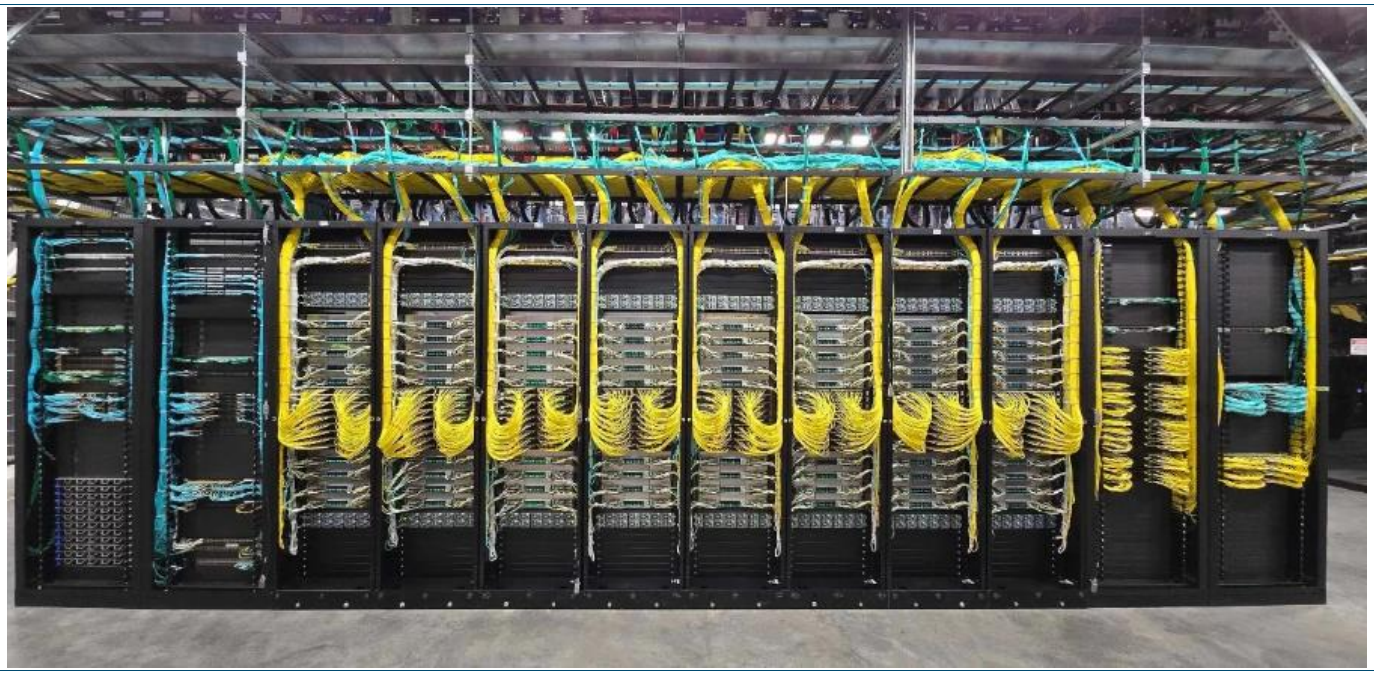


来源：《超节点技术体系白皮书》，国金证券研究所

英伟达 GB200 NVL72 放量，整机架市场需求高涨。2025 年 11 月，在英伟达第三季度财报电话会上，公司表示第三季度公司的业绩增长由 Blackwell 架构的持续增长势头推动，且 GB300 的收入占 Blackwell 总收入的比例超过了 GB200，达到约三分之二，GB300 已向大多数主要云服务提供商、超大规模云计算厂商和 GPU 云平台批量出货，此外公司预计从 2025 年年初到 2026 年 12 月，Blackwell 和 Rubin 平台的收入将达到 5000 亿美元。而根据 2026 年 2 月英伟达第四季度财报电话会，目前各大云计算服务商、超大规模云计算企业、AI 模型研发商和企业客户部署的 Blackwell 架构基础设施，算力已达 90 亿瓦，且处于满负荷运行状态。

Rubin Ultra NVL576 路线图明确。NVIDIA Vera Rubin Ultra 引入了新的两层 all-to-all NVLink 拓扑，使开发者能够扩展到 576 个 GPU。Vera Rubin Ultra NVL576 将 8 个独立的 MGX NVL 机架（每个机架配有 72 个 Rubin Ultra GPU）组合在一起，并通过铜缆和直接光纤连接在一个 576-GPU NVLink 域中。

图表7：NVIDIA Polyphe 原型，基于 GB200 的多机架 NVL576 纵向扩展系统



来源：NVIDIA 官网，国金证券研究所

华为是国内较早将 AI 集群从八卡服务器推进到超节点形态的厂商之一。其产品演进大致经历了三个阶段：第一阶段是以 Atlas 800 为代表的服务器级训练系统，核心特征是基于昇腾 910A 和 HCCS 的节点内高速互联；第二阶段是 Atlas 800T A2 等面向昇腾 910B 的 8 卡系统，继续沿用节点级高带宽互联，并逐步完善软件栈与交付体系；第三阶段则是 CloudMatrix 384，将互联范围从单节点扩展到多柜一体化超节点，把 Scale-Up 域从传统的 8 卡或 16 卡级别提升到 384 NPU 级别。

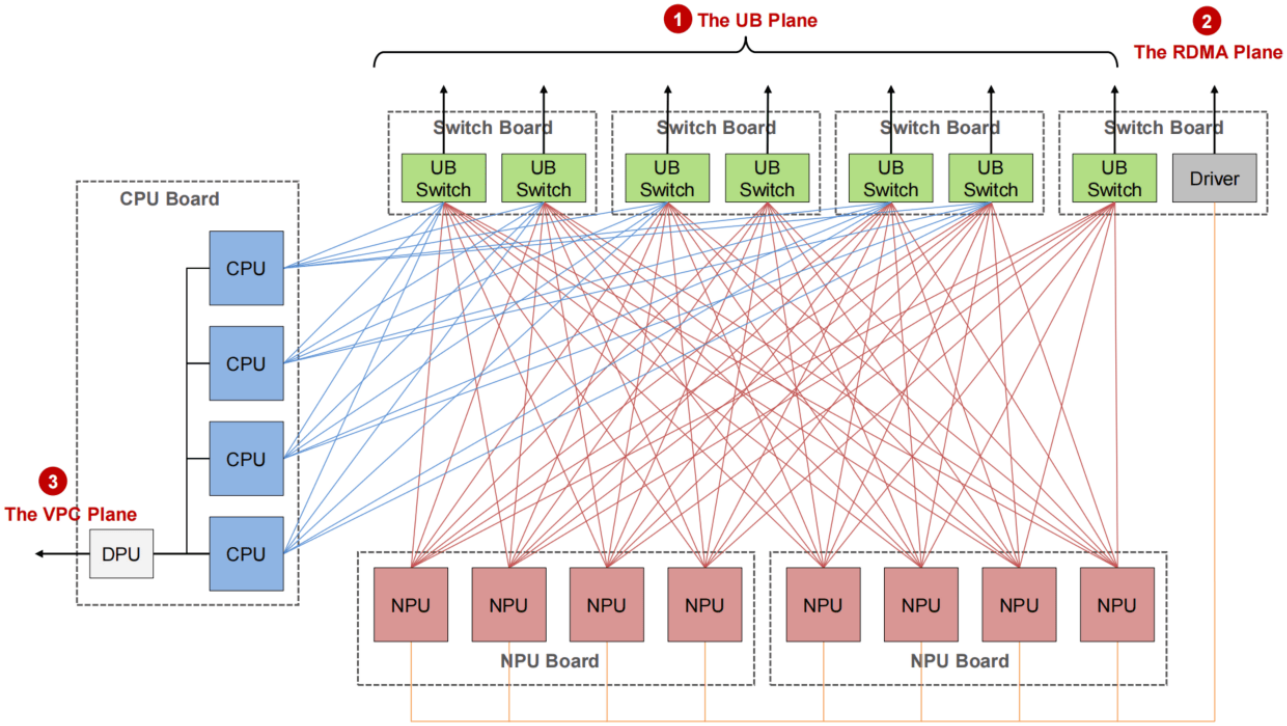
CM384 的基本思路是把芯片、节点和超节点三个层级组织为同一个连续的高带宽域：芯片层负责提供计算与本地存储能力，节点层负责将 8 个 NPU 与 CPU、交换芯片组织为单层 UB 平面，超节点层再通过多柜互联将 48 个计算节点组合为一个二层 UB 域。

图表8：华为超节点路线图

产品	芯片	Scale-Up 规模	互联协议	形态	时间
Atlas 800（训练集群）	昇腾 910A	8 卡 / 节点	HCCS v1	服务器级	2020
Atlas 800T A2	昇腾 910B	8 卡 / 节点	HCCS v2	服务器级	2023
CloudMatrix 384	昇腾 910C	384 NPU	UB 灵衢 1.0	超节点级	2024 曝光，2025 商用落地
下一代 CloudMatrix	昇腾 950（预期）	千卡到万卡级	UB 灵衢 2.0	超节点 / 集群级	2025+

来源：《超节点技术体系白皮书》，国金证券研究所

图表9：华为 CloudMatrix 384 结构示意图



来源：《超节点技术体系白皮书》，国金证券研究所

CM384 超节点进入工程化交付阶段，后续超节点产品演进路线清晰。截至 2025 年，CloudMatrix 384 已在多个智算中心公开落地或披露部署，包括芜湖、贵安、乌兰察布以及中国电信粤港澳大湾区（韶关）等站点，这表明该系统已不再停留于展示型样机，而是进入了工程化交付阶段；而根据 2025 年 9 月华为全联接大会，Atlas 900 A3 SuperPoD 超节点累计部署 300 多套，服务于互联网、金融、运营商、电力、制造等行业的 20 多个客户。在 2025 年 9 月的华为全联接大会上，结合已推出或正在研发的昇腾芯片，华为在大会上发布了多款超节点和集群产品，包括 Atlas 950 超节点（基于 Ascend 950DT 打造，支持 8192 张基于 Ascend 950DT 的昇腾卡，预计将于 2026 年四季度上市）、Atlas 960 超节点（基于 Ascend 960 打造，最大可支持 15488 卡，预计将于 2027 年四季度上市）。

图表10：Atlas 950 超节点

来源：2025 华为全联接大会，国金证券研究所

图表11：Atlas 960 超节点

来源：2025 华为全联接大会，国金证券研究所

中科曙光：实现产品化供给，打造普惠化高端算力基础设施。2025 年 11 月，中科曙光推出全球首个单机柜级 640 卡超节点 scaleX640，采用“一拖二”高密架构设计，单机柜算力密度提升 20 倍，基于 AI 计算开放架构，在硬件层面支持多品牌加速卡，实现 MoE 万亿参数大模型训练推理场景 30%-40%的性能提升。2026 年 3 月，中科曙光推出世界首个无线缆箱式超节点 scaleX40，采用正交无线缆一级互连架构，单节点集成 40 张 GPU，总算力超过 28PFLOPS，与传统 8 卡机方案相比价格持平，并且训练性能最大可提高 120%，推理性能最大提升 330%；精准适配中小规模训练与推理场景，配套的 SothisAI 平台支持一键部署、自动断点续训、故障智能隔离，支持用户独立完成全栈应用落地。

图表12：2025 世界互联网大会乌镇峰会期间中科曙光 scaleX640 备受关注

来源：中科曙光微信公众号，国金证券研究所

图表13：中科曙光 scaleX40 具备低门槛部署、高稳定运行和开箱即可用的系统创新优势

来源：中科曙光微信公众号，国金证券研究所

百度智能云：单节点实现万亿参数训练，向千卡级演进。2025 年 11 月，百度同步推出天池 256 超节点与天池 512 超节点。天池 256 将 256 张 P800 放到同一个节点内，相比 2025 年 4 月发布的超节点，单实例的卡间互联总带宽提高 4 倍、性能提高 50%以上；对比同等卡数的 P800 集群，主流大模型的推理任务上单卡吞吐提升了超过 3.5 倍，预计 2026H1 上市。天池 512 在天池 256 基础上卡数翻倍、卡间互联总带宽翻倍，单个天池 512 超节点即可万亿参数模型的训练，预计 2026H2 上市，未来百度智能云还将陆续推出相应的千卡、4000 卡超节点。

图表14：百度天池 256 超节点单卡吞吐提升 3.5 倍

来源：百度智能云微信公众号，国金证券研究所

图表15：天池 512 超节点单节点完成万亿参数训练

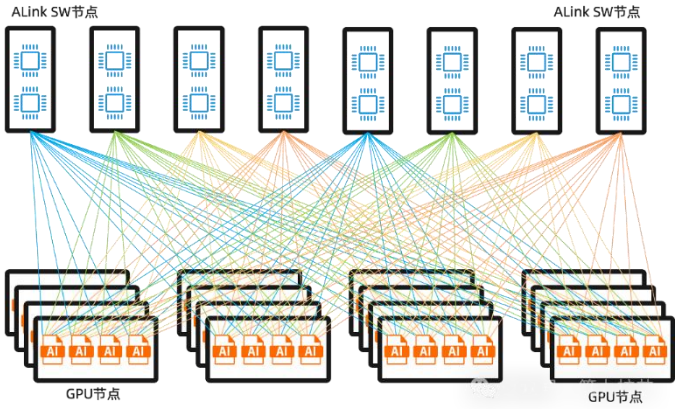
来源：百度智能云微信公众号，国金证券研究所

阿里云：软硬件深度耦合度，算力密度领先。2025 年 9 月阿里云发布全新一代磐久 128 超节点 AI 服务器，单柜支持 128 个 AI 计算芯片，集成阿里自研 CIPU 2.0 芯片和 EIC/MOC 高性能网卡，采用开放架构，扩展能力极强，可实现高达 Pb/s 级别 Scale-Up 带宽和百 ns 极低延迟，相对于传统架构，同等 AI 算力下推理性能还可提升 50%；与阿里云自研的 HPC 网络、CPFS 存储系统、人工智能平台 PAI 深度集成，使通义千问大模型训练加速 3 倍以上。基础版峰值算力与 NVIDIA H20 持平，专注于 AI 推理任务；高级版算力约为 NVIDIA H100 的 50%左右，支持 AI 训练任务，分层设计使磐久 128 可灵活适应不同场景的算力需求。

图表16：阿里磐久 AL128 超节点

来源：算力核芯微信公众号，国金证券研究所

图表17：磐久超节点 ScaleUp 互连拓扑图



来源：算力核芯微信公众号，国金证券研究所

二、超节点驱动产业链价值量重估的核心原因

2.1 价值量跃迁：AI 服务器相对通用服务器整机价值量约 25 倍，增量在环节间呈现显著梯度分化

AI 服务器整机价值量较通用服务器跃升一个数量级，传统"服务器出货台数"框架失效。根据 SemiAnalysis 对典型配置的 BOM 拆解，一台双路 Intel Sapphire Rapids 通用服务器整机成本约 1.05 万美元，而一台 Nvidia DGX H100 AI 服务器整机成本达 26.9 万美元，二者比值约 24.7 倍。这一数据背后的产业含义在于：市场以"服务器出货台数"衡量产业链景气度的传统口径已难以刻画 AI 周期的真实弹性，AI 服务器渗透率每提升一个百分点，对应的产业链收入拉动远超通用服务器周期。我们认为，评估 AI 服务器产业链标的业绩弹性，核心在于把握 25 倍价值量增量在各环节间的梯度分布结构。

图表18：Semianlysis 对 2x Intel Sapphire Rapids Server 与 Nvidia DGX H100 的成本进行了比较，AI 服务器比标准 CPU 服务器的成本多出约 25 倍

组件	成本				价值量成长性
	2x Intel Sapphire Rapids Server		Nvidia DGX H100		（3）/（1）-1
	(1)	(2)	(3)	(4)	
CPU	1,850	17.70%	5,200	1.90%	1.81
8 GPU+4 NVSwitch Baseboard	0	0.00%	195,000	72.50%	-
内存	3,930	37.50%	7,860	2.90%	1.00
存储器 硬盘	1,536	14.70%	3,456	1.30%	1.25
网卡 SmartNIC	654	6.20%	10,908	4.10%	15.68
机箱（外壳、背板、电缆）	395	3.80%	563	0.20%	0.43
主板	350	3.30%	875	0.30%	1.50
散热（散热器+风扇）	275	2.60%	463	0.20%	0.68
电源	300	2.90%	1,200	0.40%	3.00
组装测试	495	4.70%	1,485	0.60%	2.00
Markup	689	6.60%	42,000	15.60%	59.96
总成本	10,474		269,010		24.68
内存 DRAM BOM	37.50%		2.90%		
存储 NAND BOM	14.70%		1.30%		
Memory BOM	52.20%		4.20%		

来源：SemiAnalysis，国金证券研究所

GPU+NVSwitch Baseboard 是价值量跃迁的绝对主导环节，占 AI 服务器整机 BOM 比例超 70%。在通用服务器中，GPU 与 NVSwitch Baseboard 环节的价值量为零；而在 DGX H100 中，该环节单项价值量达 19.5 万美元，

占整机 BOM 比例高达 72.5%。换言之，AI 服务器本质上是一个 GPU 子系统叠加一副服务器外壳，这一结构性特征决定了 Nvidia 在全球 AI 产业链中的绝对主导地位，也解释了为何 AI 服务器周期中 GPU 环节具备最强的定价权与盈利能力。

通信是除 GPU 外增幅最陡、确定性最高的价值量提升方向。网卡环节价值量从通用服务器的 654 美元跃升至 DGX H100 的 10,908 美元，绝对值增幅达 15.7 倍，位列所有环节第二。其技术成因在于 AI 训练与推理对 Scale-out 后端网络带宽的刚性需求——每张 GPU 需单独配置一张高速 NIC（H100 时代对应 ConnectX-7 400G，B200/B300 时代升级至 ConnectX-8 800G），而通用服务器整机通常仅配置 1-2 张普通以太网卡。我们认为，通信是 AI 周期中除 GPU 外最具确定性的受益环节。

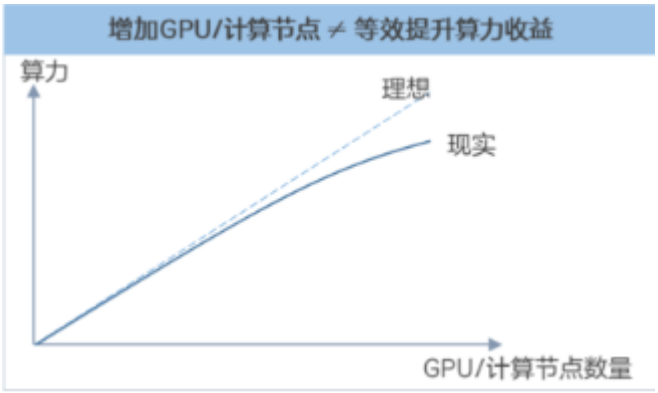
超节点架构的核心跃迁在于将互联从“服务器内部总线”提升为“系统级交换资源”。以英伟达 NVL72 为例，其单柜内部通过 NVLink Switch + 铜背板实现 72 颗 GPU 的全互联。根据 SemiAnalysis 的 BOM 模型估算，互联相关组件（交换芯片、铜背板、高速连接器、Retimer 等）在单柜 BOM 中的占比从传统 8 卡服务器（如 DGX H100）的约 3% 提升至 15%-20%。这一变化并非简单的线缆数量增加，而是交换芯片成为算力扩展的“第二颗核心”——每增加一颗 GPU 进入同一内存域，交换端口的带宽与连接数呈超线性增长。随着 Rubin Ultra NVL576 等跨柜方案引入光互联，交换芯片与光模块的价值量占比有望进一步向更高数值迈进。互联环节因此成为超节点时代价值量弹性最大、技术壁垒最高的方向。

图表19：传统 8 卡服务器和超节点互联组件占比

架构	互联组件占比	说明
DGX H100	**~3%**	PCIe + NVSwitch（4 颗）+铜缆
GB200 NVL72	**15%-20%**	18 颗 NVSwitch5 + 铜背板 + 高速线缆

来源：SemiAnalysis，国金证券研究所

图表20：GPU 集群规模扩大，内部通信数据量呈超线性增长



来源：中兴官网，国金证券研究所

组装测试环节价值量翻倍，是我们强调"集群化交付推动 ODM 毛利率结构性抬升"的 BOM 层面核心佐证。组装测试价值量从 495 美元跃升至 1,485 美元，增幅 2.0 倍。AI 服务器的装配、老化、烧机与互联联调工艺复杂度相较通用服务器显著提升，良率控制与现场交付能力成为 ODM 环节议价权转移的关键变量。进入超节点时代，交付单元从"台"升级为"柜"乃至"Pod"，组装测试环节的价值量弹性将进一步放大。

Markup 环节增幅高达 60 倍，整机集成本身即为高附加值环节。Markup（整机留存加价）从通用服务器的 689 美元跃升至 AI 服务器的 4.2 万美元，增幅达 60 倍，占整机 BOM 比例 15.6%。这一数据清晰揭示：AI 服务器整机集成环节本身即为高附加值环节，头部品牌商与 ODM 在 BOM 之上能够获取显著的议价空间。

价值量排序与国产链映射。综合上述梯度分化，AI 服务器相对通用服务器的价值量增幅排序为：GPU>SmartNIC（15.7 倍）>Markup（60 倍）>电源（3.0 倍）>组装测试（2.0 倍）>CPU（1.8 倍）>主板（1.5 倍）>存储 NAND（1.3 倍）>DRAM（1.0 倍）>散热（0.7 倍）>机箱（0.4 倍）。将这一框架映射至昇腾服务器与国产超节点生态，价值量较高、弹性最强的环节依次为 AI 芯片、通信、电源、整机集成与 Markup、CPU、存储与内存。

2.2 集群化交付推动 ODM 厂商毛利率结构性抬升

传统单台服务器代工模式下，ODM 厂商议价能力长期受限，毛利率被压制在极低水平。造成这一现象的根本原因，在于传统代工模式高度标准化：中国服务器行业向白牌生产模式转变后，众多企业将服务器模块化、标准化，进一步压缩了服务器厂商的利润空间，服务器行业整体毛利率因此持续承压。

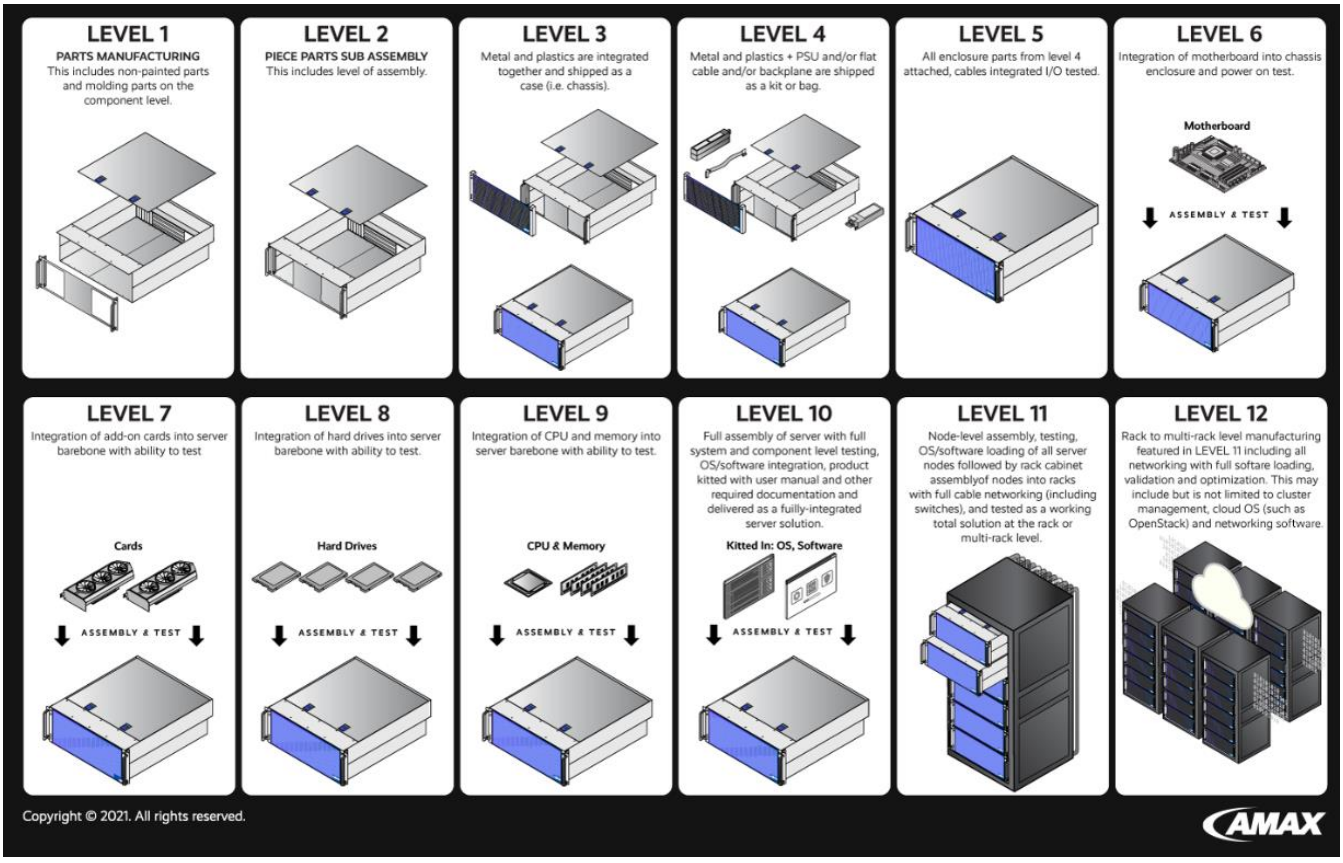
然而，超节点时代的到来正在改变这一格局，主要通过两个维度重塑 ODM 的利润结构：其一，交付单元从单台服务器跃迁为整机柜乃至 Pod 级系统交付，工程复杂度与议价权同步抬升；其二，客户结构高度集中，能够通过头部云厂商严苛验证的 ODM 极为稀缺，供给侧竞争格局优化。两重变化叠加，头部 ODM 在超节点数通业务上的毛利率有望向更高区间迁移。

2.2.1 交付单元从"台"变成"柜"乃至"Pod"

① 交付形态跃迁：从 L6 零部件到 L11 整机柜系统集成

服务器代工生产主要有 12 个层级，从最低的 Level 1（零部件制造）到最高的 Level 11-12（服务器解决方案），层级越往上难度越高，附加值随之提升。 Tencent News

图表21：服务器代工生产主要有 12 个层级



来源：AMAX，国金证券研究所

以英伟达为例，据电子产品世界报道，过去英伟达的 AI 服务器机柜，可粗分为两段，第一段为做到 L6 的主板，第二段是从 L6，再加关键零组件到 L10，然在 2026 年下半将开始量产的 Vera Rubin 架构，NVIDIA 有意指定三家 ODM 做到 L10，交货给 NVIDIA 后，再交给要做 L11 的厂商。未来英伟达计划推行的新商业模式，是由指定供应商直接做到 L10 后出货给英伟达，再由 ODM 做 L11 组装。这意味着 ODM 需承接的不再是单台白盒服务器，而是涵盖整机柜集成、Scale-up 组网、供电与结构件、上架联调在内的完整系统级交付。

② 系统级交付的工程复杂度提升，ODM 议价权上升

超节点整机柜的交付复杂度，远超传统单台服务器。据超聚变数字技术有限公司产品规划专家在亚洲数据中心峰会（Data Center Asia 2025）上的分析：超节点涉及非常多的高速线缆的连接，需要很高的部署密度，不可能用一些离散的设备来做超节点，所以必须走向整机柜，进而会带来一系列挑战。

整机柜架构的变化会提出很多工程方面的要求。首先就是供电架构。传统机房采用的是水平供电，但根据超聚变进行的模拟分析，当芯片功率超过 1 500 瓦，水平供电会面临可靠性的问题。因为铜线缆有电阻，一定有压降，会导致不同位置的芯片接收到的电压不一样。因此，在单芯片 1500 瓦以上就可能要考虑垂直供电。垂直供电对整机柜的影响主要是增加了节点空间高度，架构上带来 U 位和节点间距的变革。

其次是母线排（Busbar）。整机柜的功率如果在 250kW，用 54V 直流供电就可以了。但是，250 kW 时母线排上的电流会达到 2 500A，表面温度已经高到烫手的程度。因此，更高功率的机柜需要考虑高压直流，如 400V、800V、±400V 等方案。相应的，供电柜的架构也会发生变化。智算中心不是数据中心的简单升级，而是基于 AI 的业务复杂特性而做的重新设计。

单柜的功率越来越高，但传统风冷机房总的散热能力和供电能力是有上限的。单机柜功耗持续增加，风液混合冷板式液冷的风冷部分可能会达到风冷机房的极限，需要走向低温冷板或全液冷路线。供电能力的闲置会导致机房在改造的时候产生大量白地板的浪费。因为液冷和整机柜的方案虽然可以提高上架率，但供电和散热的上限会制约机柜的数量。在机房改造中，每一个客户、每一个行业都应该基于自己的业务特征和要求来选择一个适合自己的方案。

从构建大规模 AI 集群的角度去看整机柜，用户重点关注三点：第一，可靠性，高可靠性才能保证训练、推理的效率。第二，线性度，随着加速卡的增加，性能能否也可以线性地增长。第三，快速恢复，保证训练、推理在发生故障时能够尽快让业务保持运行，这个其实是一个更庞大的系统工程。整机柜，尤其是超节点的研发与交付，是对企业技术能力、运维能力、工程能力的综合性挑战。

我们认为，当交付复杂度提升、合格供应商稀缺时，ODM 具备向客户转移工程溢价的实际能力。

2.2.2 客户结构集中化

当前 AI 超节点服务器的需求呈现显著的“头部集中”特征。这种客户结构集中化，使得订单加速流向少数能够通过严苛认证的 ODM 厂商。超节点并非简单的服务器堆叠，而是通过高速互联协议将数十至数百颗 GPU 整合为逻辑统一的协同计算系统，在信号完整性、热设计、系统架构等方面的要求远超传统服务器。能够同时满足 PCIe 5.0/6.0 及 112G SerDes 背板设计（误码率控制在 1e-12 以下）、单柜 100kW+液冷散热方案、以及 99%以上量产直通良率的 ODM 厂商，全球范围内不超过个位数，从而形成了极高的准入壁垒。

具体的，在 400G/800G 高速互联时代，超节点单通道数据速率已攀升至 112G PAM4，信号对插入损耗、反射和串扰极度敏感，任何微小的阻抗不连续都会在眼图或误码率性能中被无限放大；热设计方面，单柜功耗突破 100kW，液冷已成为“必选配置”而非“可选技术”。这种复杂度导致具备全栈交付能力（自研交换机+AI 服务器）的 ODM 厂商极为稀缺，行业内甚至出现了“发布即巅峰，交付无日期”的乱象，部分厂商停留在“手搓样机”阶段，而真正具备工业化量产能力的 ODM 凤毛麟角。

三、相关标的

超节点整机柜的设计需解决高密度 GPU 协同工作的挑战，多环节价值量有望显著提升：

- AI 机柜：浪潮信息、华勤技术、中科曙光、紫光股份等；
- 交换芯片：盛科通信、锐捷网络等；
- 高速连接器：华丰科技等；
- 国内算力：寒武纪、海光信息、东阳光、利通电子、协创数据、网宿科技、优刻得、豫能控股、润泽科技、亿田智能、华丰科技、神州数码、云天励飞、大位科技、润建股份、科华数据、中芯国际、华虹半导体、禾盛新材、奥飞数据、首都在线、云赛智联、瑞晟智能、潍柴重机、欧陆通等

3.1 浪潮信息：超节点机柜布局已久，牵头组建创新联合体

25 年 8 月、26 年 1 月分别发布重磅超节点服务器产品。1）元脑 SD200：2025 年 8 月 7 日，2025 开放计算技术大会（OCTS25）在京举行，浪潮信息作为 OCTC 创始成员和 OCP 核心成员受邀参会，重磅发布超节点 AI 服务器“元脑 SD200”，单机即可运行超万亿参数大模型，并在多个全参模型实测中，实现 64 卡整机推理性能的超线性扩展；同时面对未来高功率、高密度算力场景的散热难题，浪潮信息推出 MW 级泵驱两相液冷 AI 整机柜方案，单芯片解热突破 3000W，解热能力高达每平方厘米 250w 以上。2）超节点服务器 CRS6000S：2026 年 1 月 9 日：浪潮计算机超节点 CRS6000S 通过架构创新，实现单机柜 32/64 张本土 AI 芯片的超高密度部署，兼容下一代高性能 AI 芯片的扩展设计更保障了资产长期价值。超节点 CRS6000S 在互联架构上实现革命性提升，通过 4 台 Switch Tray 构建的算力全互联架构，打造 32 卡 Scale-up 高速互联域，实现内存统一寻址与算力资源池化管理。柜间支持 IB/RoCE 高速通信协议，跨柜数据传输实现高带宽、低延迟特性，可轻松构建万卡以上规模的智算集群，相较传统方案，卡间通信带宽提升 8 倍，直接推动单节点 MoE 大模型训练性能和单卡推理效率大幅提升，这意味着千亿参数大模型的训练周期可从月级压缩至周级，短视频生成、智能客服等推理场景的响应延迟将降低至毫秒级。

26 年 3 月牵头组建创新联合体，共同攻关超节点。超节点创新联合体在北京市科委中关村管委会组织下，由浪潮信息等多家单位共同成立，已在超节点互连协议、系统研制、标准制定、应用部署等方面取得显著进展。2025 年以来，联合体重点联合芯片、系统、大模型开发、应用创新等领域的企业和机构，打造超节点智算应用“北京方案”，融合领先的多元模型算法，面向科研、具身智能、医疗、智造、教育等不同行业、不同应用场景开发智能体方案，实现技术与应用的良性互动，打通超节点产业落地的“最后一公里”。

图表22：浪潮超节点服务器 CRS6000S

来源：浪潮计算机、国金证券研究所

图表23：浪潮信息牵头组建创新联合体

来源：元脑服务器、国金证券研究所

3.2 华勤技术：超节点下半年量产交付，26 全年收入预计超百亿

计算、网络、液冷散热全栈布局。1）在超节点领域，公司布局前瞻、起步较早，长期与下游核心客户开展深度协同与联合预研，技术领先优势持续夯实巩固。公司内部研发体系保持“量产一代、开发一代、预研一代”的明确研发技术投入和规划；是行业内极少数同时拥有计算节点、网络节点和液冷散热全栈设计能力的厂家。2）在计算与网络技术方向，公司始终紧跟行业前沿趋势，实现产品与技术快速迭代升级；围绕整机架构设计、系统互联互通、供电方案设计、高效散热设计及差异化场景性能调优等核心技术节点，已构建起突出的技术壁垒与竞争优势，在客户联合研发环节占据行业领先身位。3）生产制造环节，公司面向数据中心类产品打造了完全自主可控的生产园区与高品质制造产能，可稳定保障产品良率，高效支撑客户规模化量产与快速交付需求。

公司预计今年超节点项目会在二季度开始发货，下半年规模交付，2026 全年预计超过百亿收入，占据行业领先地位。

图表24：华勤 AI 超节点产品

来源：电子发烧友网 Elecfans、国金证券研究所

风险提示

■ 行业竞争加剧的风险：

在信创等政策持续加码支持计算机行业发展的背景下，众多新兴玩家参与到市场竞争之中，若市场竞争进一步加剧，竞争优势偏弱的企业或面临出清，某些中低端品类的毛利率或受到一定程度影响。

■ 技术研发进度不及预期的风险：

计算机行业技术开发需投入大量资源，如果相关厂商新品研发进程不及预期，表观层面将呈现出投入产出在较长时期的滞后特征。

■ 特定行业下游资本开支周期性波动的风险：

部分计算机公司系顺周期行业，下游资本开支波动与行业周期性相关性较强，或在个别年份对于上游软件厂商的营收表现产生扰动。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。

上海	北京	深圳
电话：021-80234211	电话：010-85950438	电话：0755-86695353
邮箱：researchsh@gjzq.com.cn	邮箱：researchbj@gjzq.com.cn	邮箱：researchsz@gjzq.com.cn
邮编：201204	邮编：100005	邮编：518000
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼	地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧	地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究