



# 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

## 再谈超节点

### 投资逻辑：

我们在 4 月 9 日发布的《国产机柜时间到来》中提出，超节点产业进入规模落地期，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，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。本篇以 WAIC 2026 超节点集中展示为切入口，再谈超节点的投资机会。

**WAIC 大会国内超节点大检阅，需求端同步迈入超节点时代。** 本届 WAIC 大会上国内超节点产品集中大检阅，华为、中兴通讯、沐曦、摩尔线程等厂商集中展示超节点产品，产品形态覆盖 64 卡、128 卡、256 卡、384 卡乃至 1024 卡高带宽通信域；其中，华为昇腾 950 超节点 1024 卡真机首次公开亮相，昇腾 384 超节点已实现规模化商用，中兴 OEX、沐曦曦景 S600、摩尔线程 MTT C256、燧原 ESL64-0/C 及壁仞 NPO 光互连方案等亦集中亮相。与此同时，需求侧也在同步进入超节点时代，Kimi K3 官方部署建议明确提出 64 张及以上加速卡组成的超节点配置。我们认为，超节点逐步成为国产 AI 基础设施的共同技术路线，国内算力竞争的评价框架正在由单颗芯片峰值算力，转向一个高带宽域内可协同的加速卡规模、通信效率、软件成熟度和系统交付能力。

**超节点核心增量环节在于交换机及交换芯片、服务器及机柜、连接器。** 1) 交换机及交换芯片：超节点将原本服务器外部的 Scale-out 网络进一步延伸至机柜内和柜间 Scale-up 互联，推动交换需求从传统以太网交换机扩展至高带宽、低时延的专用交换芯片与交换系统。随着单卡算力和互联带宽持续提升，交换环节的芯片数量、端口速率和单机价值量均有望上行。2) 服务器及机柜：交付单元由单台白盒服务器升级至整机柜乃至 Pod 级系统，供电、散热、互联及系统调优难度显著提升，头部 ODM 工程壁垒与盈利能力有望同步提升。3) 连接器：计算板、交换板、电源及背板之间高速连接数量增加，SerDes 速率持续提升，同时正交、Cable-tray 和 NPO 等多种连接架构并行演进，推动高速背板、板端连接器、高速铜缆及光互连产品扩容升级。

**供需共振向上，继续重中国内算力黄金年代。** 需求端，国产模型能力正在跨过可用临界点，Coding、Agent、长上下文及 MoE 模型持续拉动 Token 调用量，高带宽通信域已经从工程优化项变成模型部署要求；供给端，国产 GPU 与国产 CPU 供给改善，叠加大厂资本开支持续上修，有望推动国产算力由芯片导入进入系统级放量阶段。我们认为，随着国产芯片供应释放，需求将依次向交换网络、服务器与整机柜、算力租赁、AIDC 上架和云服务传导，需求扩张、芯片供给突破与系统架构升级正在形成共振，继续重中国内算力黄金年代。

### 投资建议

交换芯片/交换机：盛科通信、紫光股份、锐捷网络等；

服务器/机柜：浪潮信息、华勤技术、中科曙光、中兴通讯等；

连接器：华丰科技、胜蓝股份、中航光电、航天电器等；

国内算力：寒武纪、海光信息、东阳光、中芯国际、华虹半导体、禾盛新材、摩尔线程、天数智芯、沐曦股份、壁仞科技、协创数据、华丰科技、杰华特、京基智农、华达科技、中科曙光、利通电子、浪潮信息、利扬芯片、胜蓝股份、华勤技术、国科微、中国长城、晶科科技、罗曼股份、盈峰环境、芯原股份、晶科科技、亿田智能、豫能控股、星环科技、鸿日达、盛视科技、神州数码、润泽科技、大位科技、润建股份、奥飞数据、瑞晟智能、科华数据、潍柴重机、欧陆通、杰创智能、奥尼电子。

### 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

|                                                 |    |
|-------------------------------------------------|----|
| 一、WAIC 大会国内超节点大检阅 .....                         | 3  |
| 二、需求端同步迈入超节点时代 .....                            | 7  |
| 2.1 Kimi K3: 官方建议 64 张及以上加速卡，高带宽通信域进入部署要求 ..... | 7  |
| 2.2 多个国产超节点平台披露与国产大模型完成适配与部署 .....              | 7  |
| 三、超节点增量环节：交换、整机柜与高速连接 .....                     | 7  |
| 3.1 交换机及交换芯片：端口、容量同步升级 .....                    | 8  |
| 3.2 集群化交付推动 ODM 厂商毛利率结构性抬升 .....                | 10 |
| 3.3 高速连接器：正交、Cable-tray 与 NPO 共同扩展产品边界 .....    | 12 |
| 四、供需共振向上，重中国内算力黄金年代 .....                       | 13 |
| 4.1 国内算力风眼：国产模型能力跃升、调用量持续攀升 .....               | 13 |
| 4.2 GPU：需求端芯模适配提速，供给侧改善 .....                   | 13 |
| 4.3 CPU：推理时代是 CPU 需求大起点 .....                   | 13 |
| 4.4 AIDC：期待算力供给丰富后同步迈入涨价周期 .....                | 13 |
| 4.5 算力租赁：好用的算力仍是稀缺资产 .....                      | 13 |
| 五、相关标的 .....                                    | 13 |
| 风险提示 .....                                      | 14 |

## 图表目录

|                                                                |    |
|----------------------------------------------------------------|----|
| 图表 1: Atlas 950 SuperPoD 在 WAIC 2026 首次公开真机 .....              | 3  |
| 图表 2: 中兴通讯 OEX 超节点计算柜与交换柜 .....                                | 4  |
| 图表 3: 沐曦新一代 AI 超节点新品曦景 S600 .....                              | 4  |
| 图表 4: 摩尔线程 WAIC 展台的计算节点与交换节点 .....                             | 5  |
| 图表 5: 燧原“云燧 ESL64-0”超节点 .....                                  | 6  |
| 图表 6: 新华三 WAIC 现场超节点及智算方案展示 .....                              | 7  |
| 图表 7: NVIDIA GB300 NVL72 中，NVIDIA NVLink Switch Tray 架构图 ..... | 8  |
| 图表 8: Vera Rubin NVL72 NVLink 全连接拓扑 .....                      | 8  |
| 图表 9: 昇腾未来三年路线迭代图 .....                                        | 9  |
| 图表 10: 昇腾计算柜逻辑组网（384 超节点） .....                                | 9  |
| 图表 11: 新华三 WAIC 现场智算网络及超节点方案展示 .....                           | 10 |
| 图表 12: 服务器代工生产主要有 12 个层级 .....                                 | 11 |
| 图表 13: 正交零线缆超节点对板端与正交高速连接提出更高要求 .....                          | 12 |



## 一、WAIC 大会国内超节点大检阅

超节点已经成为国内 AI 基础设施的共同技术路线。本届 WAIC，国内主要算力厂商集中发布及展示超节点、超集群新品，国内芯片、服务器、网络与云厂商集中展示超节点和超集群真机，产品规格覆盖 64 卡、128 卡、256 卡、384 卡和 1024 卡，展示内容也由单一硬件参数扩展到互联协议、模型适配、Token 吞吐、长稳运行及批量交付。如华为昇腾 950 超节点真机首次公开亮相，支持 1024 卡互联；中兴通讯发布兼容多元国产 GPU 的 OEX 超节点；沐曦发布单柜 64 卡曦景 S600；摩尔线程展示 MTT C256；中科曙光展示全国产十万卡 AI 超集群“曙光 8000（登峰）”，壁仞科技、燧原科技、新华三等亦展示相关产品或方案。我们认为，超节点逐步成为国产 AI 基础设施的共同技术路线，国内算力竞争的评价框架正在由单颗芯片峰值算力，转向一个高带宽域内可协同的加速卡规模、通信效率、软件成熟度和系统交付能力。

### 1.1 华为：昇腾 950 超节点

昇腾 950 超节点把单一高速互联域扩展至 1024 卡。华为官网 7 月 17 日披露，Atlas 950 SuperPoD 在 WAIC 首次公开真机，系统提供 1 EFLOPS FP8、2 EFLOPS FP4 算力，拥有 256TB 全局统一内存编址空间，并采用 TB 级 NPU 互联带宽与 3μs RTT，面向万亿参数大模型训练和高并发推理。1024 卡规模要求计算、交换、内存语义和调度软件共同构成连续高带宽域，产品竞争不再局限于单卡参数。

昇腾超节点已经形成规模化交付。昇腾 384 超节点已商用落地 750 多套，覆盖互联网、运营商、金融、教育、医疗、交通和制造等行业；现场还展示 Atlas 850E 风冷超节点，单个超节点支持扩展至 96 卡。750 多套交付为系统稳定性、模型适配和持续供货提供工程验证，客户采购指标也进一步转向有效吞吐与单位 Token 成本。

图表1：Atlas 950 SuperPoD 在 WAIC 2026 首次公开真机



来源：华为官网，国金证券研究所

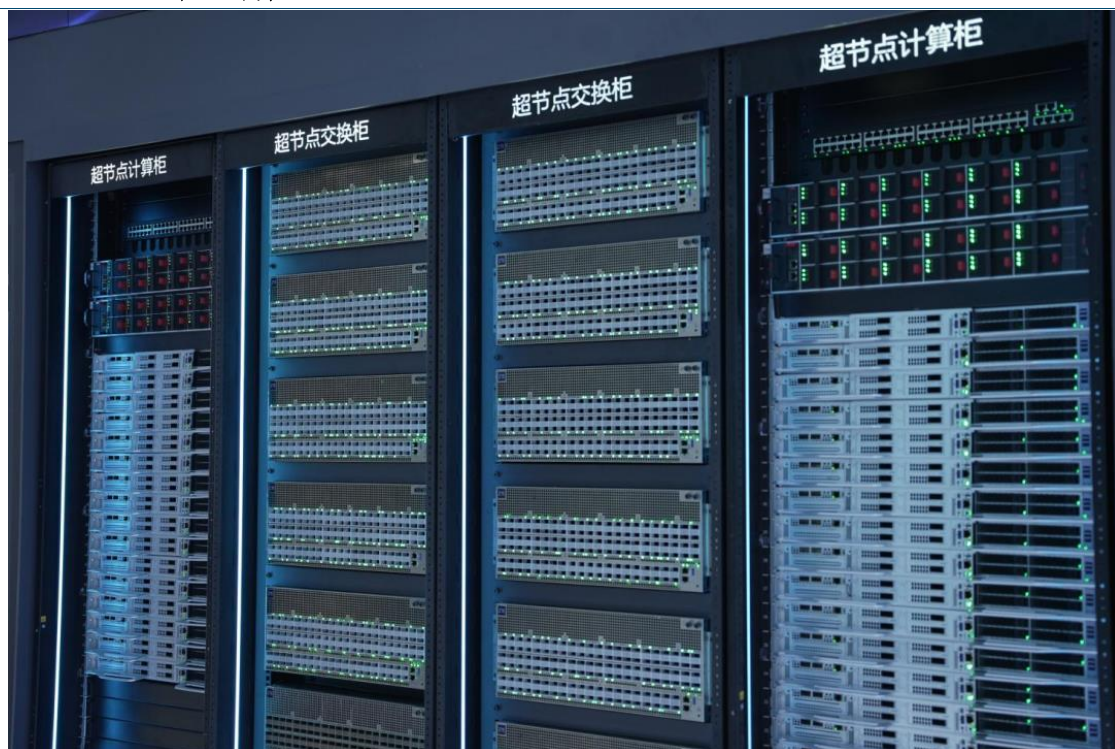
### 1.2 中兴通讯：OEX 超节点

OEX 超节点以统一接口连接多元国产 GPU。中兴通讯官网 7 月 18 日披露，公司在 WAIC 正式发布 OEX 超节点，通过底层协议标准化与接口统一兼容国内多元 GPU 生态，并支持客户结合业务特征选择芯片组合；系统采用正交无背板和“三零线缆”设计，通过高速互联总线协议与大容量交换芯片，实现 GPU 机内高速互联与机间无损交换。该方案把多芯片适配、Scale-up 交换和整柜工程纳入统一交付。

OEX 把优化目标延伸至全生命周期 Token 成本。中兴通讯官网称，公司对芯片、算法、整机、集群和软件进行全栈拉通，并通过“OEX 算力集装箱”和前置式开发压缩产品上市周期。面对快速迭代的国产 GPU 供给，标准化接口与模块化系统有助于降低客户更换芯片、扩容和运维的复杂度，同时提升交换系统、联合调优和工程交付在整机价值中的占比。



图表2：中兴通讯 OEX 超节点计算柜与交换柜



来源：澎湃新闻，国金证券研究所

### 1.3 沐曦：曦景 S600

曦景 S600 选择 64 卡作为单柜全互连规模。曦景 S600 是面向大模型训练、推理及智算中心建设打造的系统级算力产品，可实现单机柜 64 卡高密度部署，通过柜内全互连架构，大幅降低数据交互时延，提升大模型并行计算效率，并灵活支持 EP、TP 等多维并行策略，全面覆盖大模型训练和推理场景。扩展能力方面，曦景 S600 支持柜内高速互连和柜间横向扩展，集群规模可灵活扩展至万卡级别，满足大型智算中心建设需求，为 AI 基础设施提供更具效率的国产算力底座。整机柜一体化交付模式有效缩短部署周期，计算节点与交换节点采用 0 线缆直连设计，进一步降低建设和运维成本。

图表3：沐曦新一代 AI 超节点新品曦景 S600



来源：IT 之家，沐曦官方公众号，国金证券研究所

### 1.4 摩尔线程：MTT C256



MTT C256 将 256 张 GPU 组织在一层 Scale-up 网络内。IT 之家 7 月 20 日报道，摩尔线程在 WAIC 首次公开展示 MTT C256，采用计算与交换一体化高密设计，目标是减少传统多层网络引入的带宽损耗与转发时延；公司同时展示模型训练工厂、词元生产工厂和智能体生产工厂，将硬件系统与训练、推理及 Agent 负载对应。一层网络覆盖规模扩大后，交换芯片端口密度、拓扑设计和故障隔离成为决定系统效率的关键变量。

图表4：摩尔线程 WAIC 展台的计算节点与交换节点



来源：IT 之家，国金证券研究所

### 1.5 中科曙光：曙光 8000（登峰）

曙光 8000（登峰）把国产算力系统扩展到十万卡级。科创板日报 7 月 17 日报道，该系统在 WAIC 迎来真机首展，已接入国家超算互联网并纳入全国一体化算力网，可承载科学计算、大模型训练与推理及工业仿真等任务；系统已完成 300 余项重点应用优化，覆盖 20 余个科研和产业领域，超过 70 项应用实现万卡规模扩展。其重点是通过计算、网络、存储、调度和应用优化，把不同计算单元组织成可运营的大规模基础设施。

超智融合扩大同一基础设施能够承载的任务范围。科创板日报同篇报道显示，曙光 8000 支持 FP64 至 INT8 多精度算力，兼顾高精度科学计算和低精度 AI 计算，并以国家超算互联网承担跨区域算力调度。在国产算力仍具异构特征的阶段，科研、训练和推理负载共平台能够提高集群利用率，系统软件、资源调度与应用迁移服务的重要性随之提升。

### 1.6 壁仞科技：三级超节点产品矩阵

壁仞科技正式推出下一代 NPO 光互连、分布式解耦架构超节点方案，支持单个超节点 1024 卡 Scale-up 扩展。壁仞科技首创 Chiplet（芯粒）架构大算力芯片，新一代 BR2xx 系列 GPU 沿用这一技术路线。BR2xx 支持 FP8/FP4 低精度高算力计算，拥有更大显存带宽并原生集成超节点互连能力。其自研的 BLink2.0 超节点互连协议，可让最多 1024 张 GPU 共享同一个内存空间。此外，基于 BR2xx 和 BLink2.0，壁仞科技构建了三级超节点产品矩阵：16 卡标准服务器超节点（电互连）、128 卡高密整机柜超节点（电互连）、以及 1024 卡分布式解耦架构超节点（NPO 光互连）。

### 1.7 燧原科技：两类 64 卡方案并行，并探索 512 卡以上 NPO 互连

燧原同时展示正交架构和 Cable-tray 两类 64 卡超节点。燧原科技 7 月 17 日新闻稿披露，与中兴通讯联合打造的云燧 ESL64-O 采用正交架构和零线缆高密集集成，ESL64-C 则采用成熟 Cable-tray 方案，两类产品面向不同机房条件和部署偏好；公司还展示 NPO 光互连原型系统，突破传统铜互连距离限制，支持 512 卡以上超节点部署及超大规模集群弹性扩展。两类整机柜方案并行体现超节点结构仍处于快速工程化阶段。正交无背板减少计算节点与交换节点之间的高速线缆，但提高板端连接器精度、信号完整性和整柜装配要求；Cable-tray 路线供应链和维护方式更成熟；NPO



则为跨柜扩大高带宽域提供后续路径。芯片厂商与系统厂商联合开发，可在量产前共同验证接口、交换拓扑、结构件与连接器。

图表5：燧原“云燧 ESL64-0”超节点



来源：燧原科技，国金证券研究所

### 1.8 新华三：UniPoD S80000 超节点

新华三以全栈产品矩阵覆盖不同规模超节点。新华三官网披露，UniPoD S80000 覆盖 32 卡至 1024 卡，最高可弹性扩展至 16384 卡，训练性能提升 70%、推理性能提升 3 倍；WAIC 现场同步展示全系列单芯片 102.4T 智算交换机和新一代智算广域网产品。计算、网络、存储、云、安全和运维协同交付，使超节点内部 Scale-up 与跨节点网络可以纳入统一优化。

全栈协同的目标落在模型端到端效率。新华三官网称，方案通过光互连、算网协同、高密架构和拓扑优化降低全链路 TCO，并在政务云训推平台、模型训练智算中心和医疗科研 AI 场景落地。客户验收重点由单机峰值算力转向训练效率、推理吞吐、稳定性和可运维性，交换设备与系统软件的价值量同步提高。



图表6：新华三 WAIC 现场超节点及智算方案展示



来源：新华三官网，国金证券研究所

## 二、需求端同步迈入超节点时代

### 2.1 Kimi K3：官方建议 64 张及以上加速卡，高带宽通信域进入部署要求

Kimi K3 把超节点规模写入官方部署建议。Kimi 7 月 17 日技术博客显示，K3 为 2.8 万亿参数模型，采用 Stable LatentMoE，每次有效激活 16/896 个专家；官方指出推理效率同样受益于更大的高带宽通信域，因此建议在 64 张及以上加速卡组成的超节点配置中部署。

超大 MoE 模型放大专家并行与内存访问压力。Kimi 同篇技术博客披露，K3 采用量化感知训练，并围绕大规模专家并行使用静态形状和关键路径无主机同步等优化；Mooncake 分离式推理架构与前缀缓存则继续降低服务成本。随着参数量、专家数量和上下文长度增长，单卡算力难以独立解决通信和显存约束，更大的 Scale-up 域有助于提升专家路由、KV 缓存与张量并行效率。

### 2.2 多个国产超节点平台披露与国产大模型完成适配与部署

DeepSeek 已经成为国产超节点软件成熟度的重要验证负载。中国证券报·中证网 7 月 19 日报道，百度天池 256 已适配 DeepSeek、文心、GLM 和 MiniMax，推理效率提升 50%；上海证券报·中国证券网 7 月 18 日报道，阿里磐久 AL128 已通过百炼支持 Qwen、DeepSeek 和 Kimi。两家云厂商均把多模型适配与整机柜互联结合，超节点由硬件发布进入模型服务交付。

昇腾超节点也已完成 DeepSeek 系列模型适配。中国证券报·中证网 4 月 24 日转引华为官方微信称，昇腾超节点全系列产品支持 DeepSeek V4，昇腾 A3 超节点系列全面适配并提供训练参考实现，昇腾 950 通过融合 kernel 和多流并行降低 Attention 计算与访存开销。

## 三、超节点增量环节：交换、整机柜与高速连接

我们认为，超节点核心增量环节在于：1) 交换机及交换芯片：计算单元从传统 8 卡扩展至 64 卡、128 卡、384 卡乃至 1024 卡，卡间通信端口、单端口速率及交换容量同步提升，Scale-up 交换芯片、无损网络及整机交换系统价值量显著增长。2) 服务器及机柜：交付单元由单台白盒服务器升级至整机柜乃至 Pod 级系统，供电、散热、互联及系统调优难度显著提升，头部 ODM 工程壁垒与盈利能力有望同步提升。3) 连接器：计算板、交换板、电源及背板之间高速连接数量增加，SerDes 速率持续提升，同时正交、Cable-tray 和 NPO 等多种连接架构并行演进，推动高速背板、板端连接器、高速铜缆及光互连产品扩容升级。

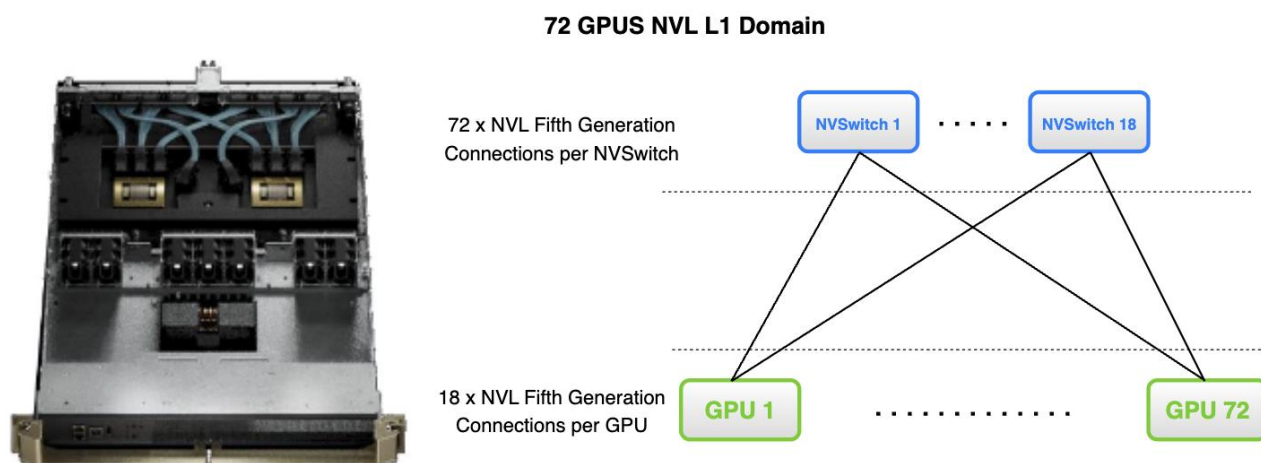


### 3.1 交换机及交换芯片：端口、容量同步升级

计算单元由传统 8 卡服务器扩展至超节点后，需要通过更多高速交换端口将 GPU 纳入统一 Scale-up 通信域，单端口速率和交换芯片容量随之提升。同时，训练中的 All-Reduce、MoE 模型中的 All-to-All 及专家并行流量，对拥塞控制、无损通信、链路遥测和故障隔离提出更高要求。因此，超节点对交换环节的增量不仅体现在设备和端口数量增长，也体现在交换芯片规格升级及网络软件价值提升。

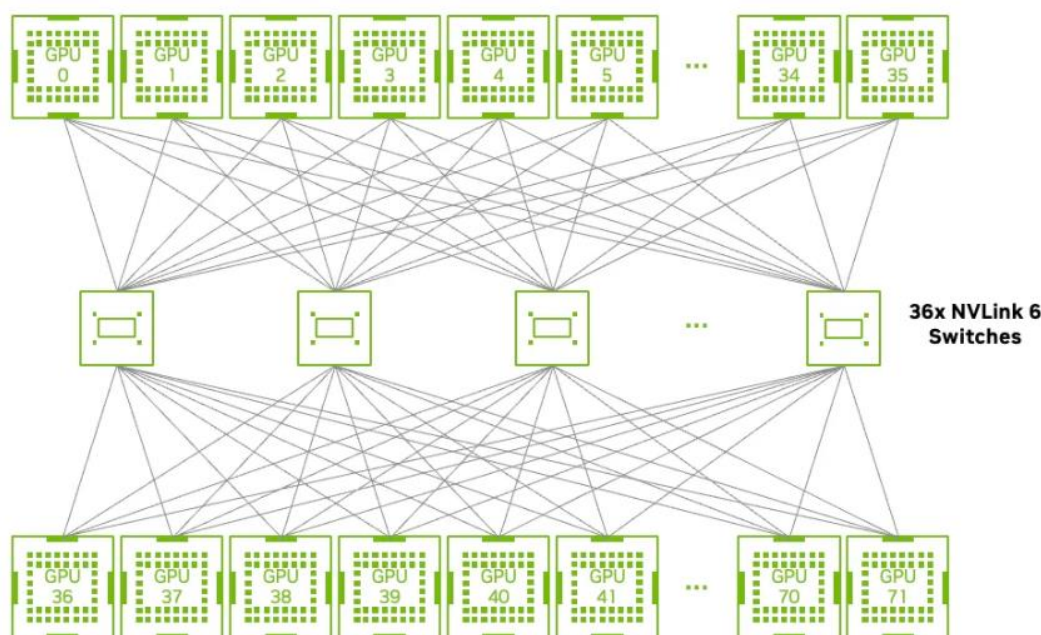
以 NV 为例，从 GB 迭代至 Vera Rubin：单卡互联带宽翻倍，单柜 NVSwitch ASIC 数量同步翻倍。1) GB200/GB300 NVL72 由 72 颗 Blackwell GPU 构成单一 NVLink 域，每颗 GPU 通过 18 条第五代 NVLink 分别接入 18 颗 NVSwitch，单 GPU 聚合互联带宽达到 1.8TB/s；整柜配置 9 个 NVLink 交换托盘，每个托盘集成 2 颗 NVSwitch，合计使用 18 颗交换芯片，GPU 与交换芯片数量之比为 4:1，整柜 GPU 侧聚合互联带宽约 130TB/s。2) Vera Rubin NVL72 仍维持 72 颗 GPU 和 9 个 NVLink 交换托盘，但单 GPU 双向互联带宽提升至 3.6TB/s，整柜聚合带宽增至约 260TB/s；与此同时，每个交换托盘集成的 NVLink 6 Switch 由 2 颗增至 4 颗，单柜交换 ASIC 总量由 18 颗增至 36 颗，GPU 与交换芯片配比由 4:1 提升至 2:1。Scale-up 交换需求不仅取决于超节点的卡数，还取决于单卡需要注入交换网络的带宽：在卡数不变的情况下，单卡互联带宽翻倍，同样可以通过增加并行交换平面，推动单柜交换芯片用量和交换系统价值量提升。

图表7: NVIDIA GB300 NVL72 中, NVIDIA NVLink Switch Tray 架构图



来源: Nvidia 官网, 国金证券研究所

图表8: Vera Rubin NVL72 NVLink 全连接拓扑



来源: Nvidia 官网, 国金证券研究所



以昇腾为例：互联带宽由 784GB/s 提升至 2TB/s，950 进一步转向灵活、多级组网。1) 昇腾芯片的互联能力也在快速迭代，Atlas 900 A3 SuperPoD 采用 384 颗昇腾 910C 构成统一 Scale-up 域，单 NPU 双向互联带宽为 784GB/s；2) 昇腾 950 系列将单芯片互联带宽提升至约 2TB/s，较 910C 提升约 2.5 倍。3) 根据华为为路线图，昇腾 960 将在 950 基础上进一步增加互联端口数，昇腾 970 的互联带宽规划提升至 4TB/s，整体呈现接近一年一代、互联带宽持续升级的趋势。

图表9：昇腾未来三年路线迭代图

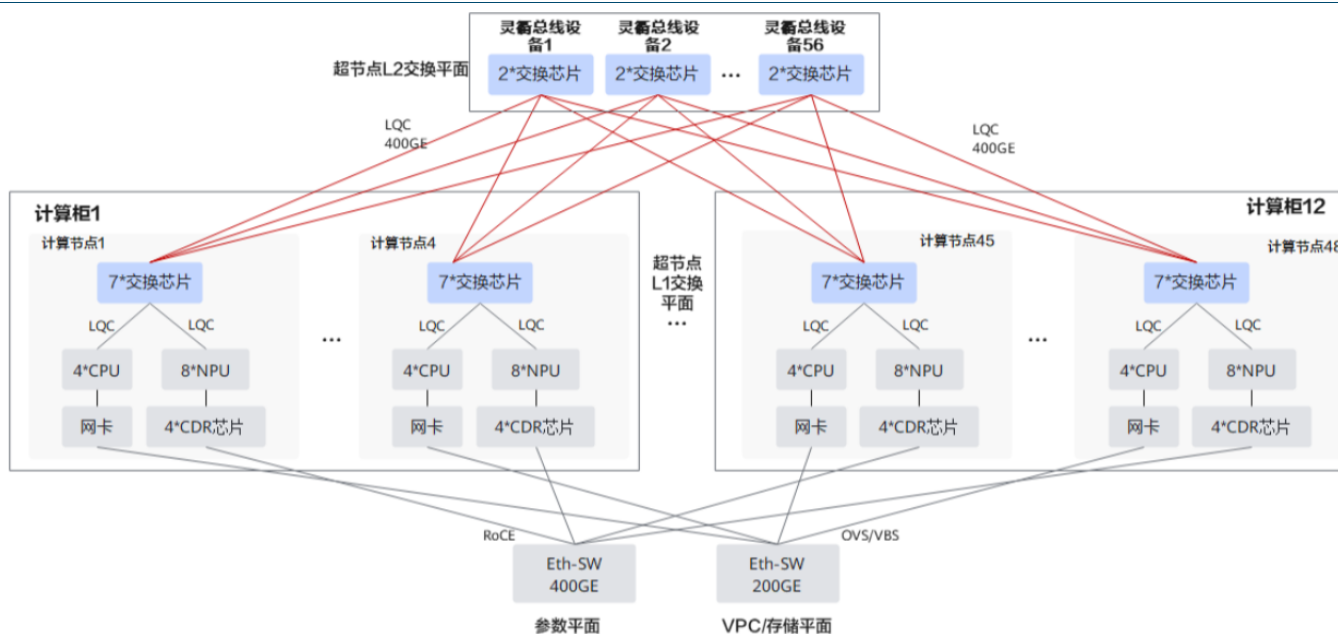


来源：华为 2025 全联接大会，国金证券研究所

从具体交换架构看，CM384 采用节点内 L1 交换+节点间 L2 交换的两级灵衢网络。系统包含 48 个 8 卡计算节点，每个节点配置 7 颗 LQC 交换芯片，L1 层合计 336 颗；上层通过 56 台灵衢总线设备连接不同计算节点，按每台设备 2 颗交换芯片计算，L2 层约 112 颗。因此，CM384 合计对应约 448 颗 Scale-up 交换芯片，384 颗 NPU 与交换芯片的配比约为 1:1.17，其中仅计算节点内 L1 层的配比即为 8 颗 NPU 对应 7 颗交换芯片。

昇腾 950 的思路则发生进一步变化。单颗 950 芯片集成 72 Lane HiLink SerDes，划分为 18 个 X4 端口，单芯片 Unified Bus 双向带宽约 2016GB/s；同时在 IO Die 内部集成 UB On Chip Switch，使部分流量可以直接在芯片端口之间完成转发，无需进入计算 Die 或占用片上内存带宽。

图表10：昇腾计算柜逻辑组网（384 超节点）



来源：华为官网《Atlas 900 A3 SuperPoD 超节点 技术白皮书》，国金证券研究所

超节点放量叠加互联带宽升级，国产交换芯片及交换机需求有望实现量级跃升。超节点将原本服务器外部的 Scale-out 网络进一步延伸至机柜内和柜间 Scale-up 互联，推动交换需求从传统以太网交换机扩展至高带宽、低时延的专用交换芯片与交换系统。随着单卡算力和互联带宽持续提升，交换环节的芯片数量、端口速率和单机价值量均有望上



行。当前国内超节点方案处于多技术路线并行阶段，而 Scale-up 环节需要高带宽、低时延的专用交换芯片与交换系统，并要求交换芯片、交换机、互联协议和算力芯片进行系统级适配，为国产交换机及交换芯片厂商参与联合研发、架构适配和设备供货提供机会。

从各家进展看，1) 盛科通信：12.8T、25.6T 交换芯片已进入客户市场推广和逐步应用阶段，并在规划面向大型数据中心和云服务的下一代高端交换芯片。2) 中兴通讯：公司 OEX 超节点在网络侧采用柜内 57.6T 自研芯片电交换与柜间 dOCS 光交换相结合的架构，并通过 OLink 协议兼容多元国产 GPU。3) 新华三 UniPoD 超节点同步展示全系列单芯片 102.4T 智算交换机，通过光电并行、算网调优和广域无损覆盖不同层级算力互联需求。4) 锐捷网络在 WAIC 将交换设备纳入超节点、AIDC 及 DCI 整体方案，关注重点由单一交换设备转向系统级智算网络，公司投资者关系材料显示，2025 年以来数据中心 400G、800G 高速交换机收入占比持续提升。

图表11：新华三 WAIC 现场智算网络及超节点方案展示



来源：新华三官网，国金证券研究所

### 3.2 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。

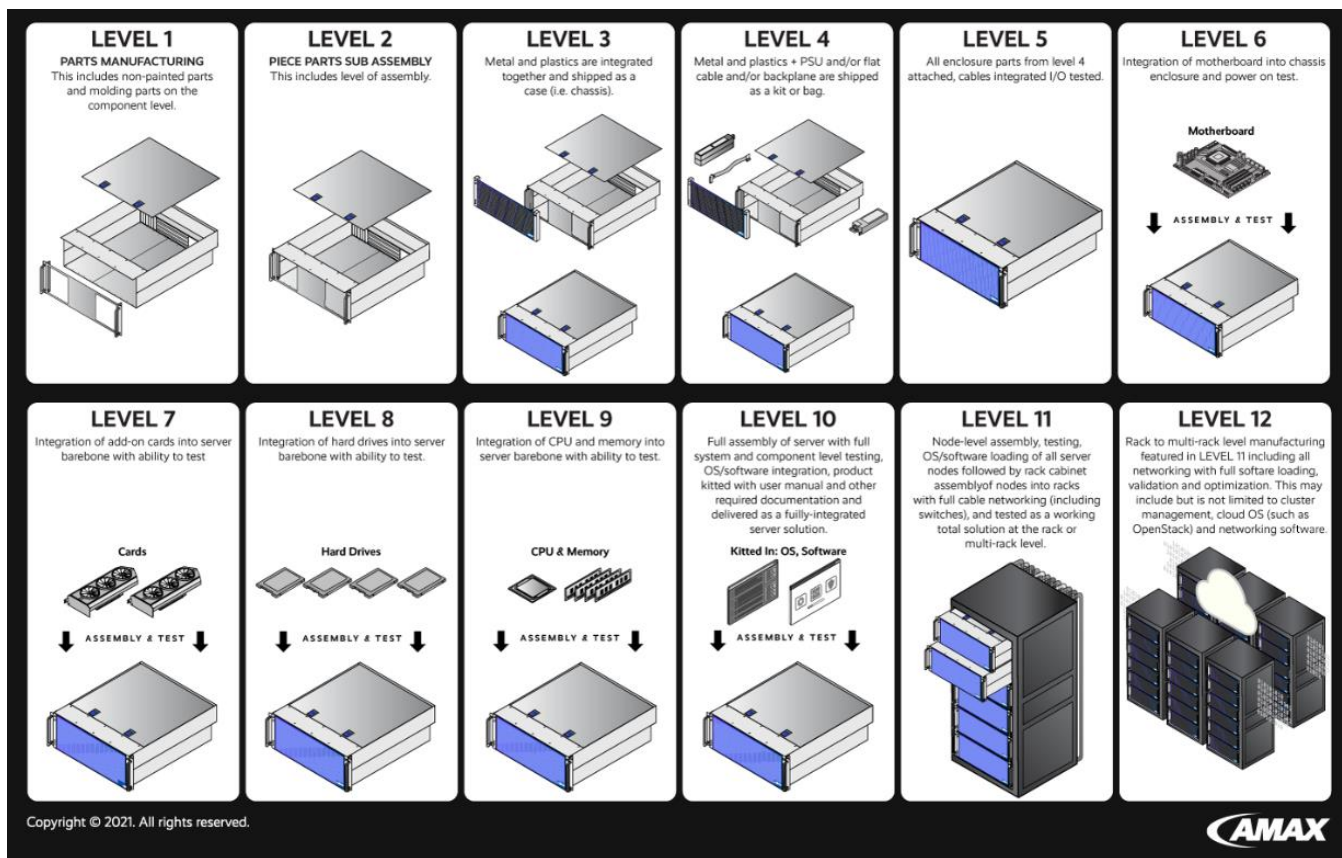
#### 3.2.1 交付单元从“台”变成“柜”乃至“Pod”

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12：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，然后在 2026 年下半年将开始量产的 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 1500 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 2500A，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。



### 3.2.2 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。

### 3.3 高速连接器：正交、Cable-tray 与 NPO 共同扩展产品边界

超节点架构带来高速连接器增量，核心在于机柜内互联密度和速率显著提升。传统服务器跨节点主要依赖外部交换网络，而超节点需要在单机柜或多机柜范围内将更多 GPU/NPU 组成低时延、高带宽的统一计算单元，连接方式从普通板级连接和外部网络互联，升级为芯片、模组、背板、线缆与光电模块之间的高密度系统级互联。随着 112G/224G 高速通道持续迭代，单柜连接点数量、线缆复杂度和连接器价值量同步提升，正交背板连接器用于提升板卡与背板之间的高速信号密度，Cable-tray 方案用于解决柜内高速线缆布线和维护问题，NPO/CPO 则进一步打开光电互联接口增量。由此，高速连接器从传统服务器配套件升级为超节点系统性能和交付能力的重要组成部分。

WAIC 大会上，真实超节点架构已经直接呈现连接方式变化。燧原科技 7 月 17 日新闻稿显示，云燧 ESL64-O 采用正交架构、零线缆和高密集成，云燧 ESL64-C 采用 Cable-tray 方案，NPO 原型系统则支持 512 卡以上部署；中兴 OEX 同样采用正交无背板和“三零线缆”设计。正交方案减少大量柜内高速线缆，却提高板端与正交连接器的密度、精度和信号完整性要求；更大跨柜通信域则推动铜互连与光互连按距离分工。

图表13：正交零线缆超节点对板端与正交高速连接提出更高要求



来源：燧原科技新闻稿（美通社发布），国金证券研究所

从目前核心连接器厂商产品进展来看：1）华丰科技：正交架构是高速背板连接器的重要演进方向，可提高信号传输密度并降低传输损耗；112G 正交高速背板已完成开发并批量生产，224G 高速连接器达到样品试制合格，相关产品适用于数据中心高端服务器和交换机。112G 量产与 224G 样品验证形成连续代际，超节点放量将增加高速通道数量并



提高单通道性能要求。2) 胜蓝股份：从消费电子/汽车连接器向 AI 数据中心高速连接器延伸，在高速连接器领域持续进行技术研发和投入，目前 224G 高速背板连接器相关产品的样品在制作当中，并配合部分客户开展测试验证工作。3) 中航光电：数据中心产品包括光传输器件及组件、高速连接器与模组等，高速铜缆组件和综合布线产品已在行业头部客户实现规模化应用；公司披露的 CPO 全光互连链路方案还涉及光 shuffle、光背板、高密度面板连接器、外置光源与光引擎。柜内短距高密连接和跨柜光互连并行，为连接器厂商拓宽可服务市场。

## 四、供需共振向上，重中国内算力黄金年代

### 4.1 国内算力风眼：国产模型能力跃升、调用量持续攀升

国产模型调用量领先，验证算力需求旺盛。据 OpenRouter，6 月 15 日至 19 日，全球 AI 大模型总调用量 47 万亿 Token，环比增长超 7%，连续九周上行，中国 AI 大模型周调用量达到近 13 万亿 Token，连续八周超过美国居全球首位；同时，全球调用量前五名中，前四款均为国产模型，其中，DeepSeek-V4-Flash 连续五周位居全球榜首，周调用量超 3 万亿 Token。

### 4.2 GPU：需求端芯模适配提速，供给侧改善

国产芯模进入协同阶段，适配持续提速。5 月 22 日，国家发改委明确表示，将指导国产大模型加大力度适配国产算力芯片，标志着国产 AI 生态建设上升至产业政策层面。国产芯片 Day0 适配逐步成为头部模型厂商的标配，DeepSeek 于 4 月 24 日发布 V4 系列模型当日，华为昇腾、寒武纪、海光等 8 家国产 AI 芯片厂商同步宣布完成全链路适配。6 月 17 日智谱 GLM-5.2 上线，已在 Day 0 完成与华为昇腾、平头哥、摩尔线程、寒武纪、昆仑芯、沐曦、海光、壁仞等国产算力平台的推理适配，预计下半年昇腾 950 超节点上市后，将成为 GLM-5.2 强劲的算力底座。

国内大厂加码采购国产 AI 芯片，供给弹性逐步释放。据彭博社 5 月 28 日报道，字节跳动正讨论将 2026 年总资本开支上限提高至 700 亿美元，较 2025 年约 250 亿美元的支出规模实现近三倍跃升，资金将主要投向数据中心和 AI 基础设施建设。据第一财经 6 月 17 日消息，字节跳动正与天数智芯讨论采购至少 5 万颗 AI 芯片，主要用于大模型推理负载，对应天数智芯智铠系列云端推理 GPU；大规模模型训练侧，字节采用华为昇腾、寒武纪高端训练卡。另外，腾讯在 2026Q1 业绩会上表示，2026 年公司资本支出将会增加，H2 将有更多国产芯片投入使用。我们认为，国内大厂持续上修资本开支，新增 CapEx 有望向国产算力基础设施倾斜，随国产芯片供应加速释放，国产算力迎来大规模导入期。

### 4.3 CPU：推理时代是 CPU 需求大起点

CPU 供给端持续承压，Intel、AMD 服务器 CPU 库存趋紧、交期延长，并分别于 3 月和 4 月起上调全系列 CPU 价格，平均涨幅达 10-15%，CPU 产业已进入新一轮景气周期。1) CPU/GPU 部署比例抬升：Intel 及 AMD 公开演讲再上调乐观配比至 4:1，NV 对 CPU 业务贡献指引乐观，预计今年独立 Vera CPU 业务（不考虑 Blackwell 和 Rubin 系统中的 CPU）营收将达近 200 亿美元，为公司开辟 2000 亿美元的全新纯增量潜在市场。2) TAM 显著扩容：AMD 与 Arm 均大幅上修服务器 CPU 市场空间，预计 2030 年全球服务器 CPU TAM 将超过 1000 亿美元。3) 所有 CPU 架构均受益，ARM 架构低功耗、高核心密度的特性更契合 Agent 工作负载，且其开放授权生态亦高度契合云厂商自主构建 AI 基础设施的需求。

### 4.4 AIDC：期待算力供给丰富后同步迈入涨价周期

AIDC：资本开支体量显著跃升，期待算力供给丰富后同步迈入涨价周期。1) 海内外大厂 CapEx 持续高增，硅谷四大科技巨头 2026 年 CapEx 将高达 6500 亿美元，AI 军备竞赛进一步加剧，具体看：亚马逊成为四家中投入规模最大的企业，将 2026 年资本支出目标定在 2000 亿美元；Alphabet 的资本支出计划高达 1750 亿美元-1850 亿美元，同比接近翻倍；Meta 预计全年资本支出将增至 1350 亿美元，同比增幅或达 87%；微软同期公布其第二季度资本支出同比增长 66%，预计其截至 6 月的财年资本支出将逼近 1050 亿美元。2) 智算中心持续扩容，国产替代加速。根据 IDC 数据，2020 年中国智能算力规模为 75.0EFLOPS，到 2028 年预计将达到 2,781.9EFLOPS，预计 2020-2028 年复合增长率达到 57.1%。在多维度数据与产业动态的交叉印证下，AI 算力基础设施投建力度维持高位，AIDC 环节呈现持续高景气扩张态势。3) 随芯片卡供应缓解、推理需求释放、国产卡性能提升，IDC 招投标有望进一步加速，并在供需挤压下同步迈入涨价周期。

### 4.5 算力租赁：好用的算力仍是稀缺资产

为什么这次算租不一样：1) 卡的迭代周期很长，供给瓶颈驱动价格持续上行，算力仍是当前的稀缺、核心资产。据 SemiAnalysis 数据，自 2025 年 8 月有统计以来，B200 现货及合约综合租赁价格由约 4.0 美元/GPU·小时升至 2026 年 7 月的约 5.0-5.2 美元/GPU·小时，累计涨幅约 25%-30%；2) 北美头部算力租赁厂商 ROI 逐步兑现、关键融资顺利；3) 国内认真做事公司已经有卡有利润，算租公司开始由订单和设备交付迈入利润释放阶段：行云科技、利通电子、智微智能等发布中报业绩预告，在智算业务推动下利润均实现加速释放。4) 此外，Token 工厂建设加速落地，软通动力与头部大模型厂商签署智算服务协议，将依托北京壹号词元工厂提供 Token 推理服务，覆盖大模型推理加速、高性能算力集群适配及行业 AI 应用落地。

## 五、相关标的

交换芯片/交换机：盛科通信、紫光股份、锐捷网络等；



服务器/机柜：浪潮信息、华勤技术、中科曙光、中兴通讯等；

连接器：华丰科技、胜蓝股份、中航光电、航天电器等；

国内算力：寒武纪、海光信息、东阳光、中芯国际、华虹半导体、禾盛新材、摩尔线程、天数智芯、沐曦股份、壁仞科技、协创数据、华丰科技、杰华特、京基智农、华达科技、中科曙光、利通电子、浪潮信息、利扬芯片、胜蓝股份、华勤技术、国科微、中国长城、晶科科技、罗曼股份、盈峰环境、芯原股份、晶科科技、亿田智能、豫能控股、星环科技、鸿日达、盛视科技、神州数码、润泽科技、大位科技、润建股份、奥飞数据、瑞晟智能、科华数据、潍柴重机、欧陆通、杰创智能、奥尼电子。

## 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，表面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。

形式的复制、转发、转载、引用、修改、仿制、刊发，或以任何侵犯本公司版权的其他方式使用。经过书面授权的引用、刊发，需注明出处为“国金证券股份有限公司”，且不得对本报告进行任何有悖原意的删节和修改。

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<br>紫竹国际大厦 5 楼 | 地址：北京市东城区建国内大街 26 号<br>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<br>18 楼 1806 |



【小程序】  
国金证券研究服务



【公众号】  
国金证券研究