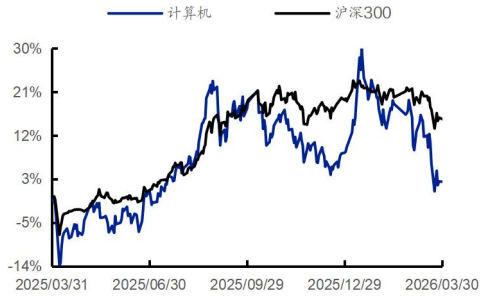
2026 年 03 月 30 日 行业研究 评级：推荐(维持)

研究所：
证券分析师：刘熹 S0350523040001
liux10@ghzq.com.cn
联系人：谢婧茹 S0350125070015
xiejr01@ghzq.com.cn

超节点 OEM：被低估的中国 AI 核心资产

——计算机行业动态研究

最近一年走势



行业相对表现			2026/03/30
表现	1M	3M	12M
计算机	-13.7%	-5.5%	2.7%
沪深 300	-4.6%	-3.4%	14.7%

相关报告

《计算机事件点评：英伟达 GTC 前瞻：或将推出 LPU 核心 AI 推理系统（推荐）*计算机*刘熹》——2026-03-08

《计算机事件点评：AI 驱动“云、边、端”算力通胀时代（推荐）*计算机*刘熹》——2026-01-29

《计算机行业专题研究：千问 APP 正式接入阿里生态，流量/模型/AI 应用产业格局有望重构（推荐）*计算机*刘熹》——2026-01-25

《计算机事件点评：Neuralink 量产在即，国产脑机链有望充分受益（推荐）*计算机*刘熹》——2026-01-07

《Gemini 3.0/Nano 密集发布，从谷歌 AI 体系看应用叙事 ——计算机行业大模型及 AI 应用专题 (2)（推荐）*计算机*刘熹》——2025-12-29

事件：

2026 年 3 月 26 日，中科曙光发布超节点新品“曙光 scaleX40”：该产品为世界首个无线缆箱式超节点，采用标准 19 英寸箱式设计，兼容主流机柜；单机 16U，部署密度是 8 卡机的 2.5 倍；最高支持 40 张 GPU 卡，FP8 算力大于 28PFLOPS，HBM 总显存大于 5TB、访存总带宽大于 80TB/S，scale-up 聚合带宽大于 17TB/S。同时，曙光 scaleX40 内置软件栈，兼容 CUDA，适配优化 800 多种大模型。

投资要点：

■ **超节点为构建大规模算力集群技术架构，已成 AI 基础设施常态**

超节点（SuperPod/SuperNode）是一种为构建大规模 AI 算力集群而设计的新型技术架构，最早由英伟达提出。其核心是通过高速互联协议将数十至数百个 GPU 或 AI 计算芯片紧密整合，形成逻辑上统一编址、高带宽、低延迟的协同计算系统。其并非简单的硬件堆砌，而是将多个计算设备通过专门的高速互联技术整合为拥有更大内存空间、能协同工作的“逻辑共同体”，让大规模算力能够“像一台计算机一样工作、学习、思考、推理”。

目前，超节点已成为 AI 算力基础设施的新常态，以英伟达、AMD、华为、中科曙光、谷歌、阿里巴巴等为代表的头部中美企业正持续推出相关产品。**2025 年被视为超节点产品“元年”，我们认为 2026 年中国国产 AI 超节点或将进入规模化应用阶段。**

■ **国内 CSP AI 资本开支展望乐观，晶圆厂/算力租赁公司产能、订单等景气度均走高**

2025 年国内整体算力资本开支规处于追赶阶段。**但 2026 年，国内 CSP 厂商资本开支展望均较为乐观。**从需求侧看，中国在 AI 推理时代具备独特优势.OpenRouter 数据显示，截至 2026 年 3 月 22 日的连续三周内，国产大模型调用量保持对美国模型的反超；3 月 1 日至 3 月 30 日，调用量前十模型中，中国模型总量占比超 50%。

从供给侧看，国内晶圆厂、英伟达、算力租赁公司产能、订单等景气度均走高。中芯国际产能利用率维持高位、华虹半导体新建产能预期乐观；英伟达 H200 已经拿到许可并获客户订单，目前正在生产过程中；宏景科技拟于 2026 年申请不超过 600 亿元授信额度同时披露 13.5 亿元定增方案，计划采购高性能算力服务器后为客户提供算力服务；协创数

据披露 2025 年至 2026Q1 客户采购额达 400 亿元+。

■ **超节点方案复杂度较高，驱动 OEM 厂商盈利能力提升**

相较传统的 GPU 集群， 超节点在通信时延、算力密度、成本和效率，机房空间利用和 TCO 总成本等方面具有优势，也具备更高的技术壁垒。在整机的系统架构设计、信号完整性、供电、散热、深度耦合上要求更高，产品规格处于非常快速的迭代中。例如，NVIDIA Vera Rubin 机架共包含 130 万个独立组件，近 1300 个芯片，整体重量接近 2 吨等。

正是这种极高的技术复杂性和快速迭代，为具备系统级能力的超节点 OEM 厂商构筑了宽阔的护城河，并直接驱动其盈利能力的提升。我们认为超节点时代，价值重心或已从标准化、可替代的硬件组装，上移至复杂的定制化系统设计、深度调优和全栈集成服务，**因而超节点 OEM 厂商凭借其在架构设计、热管理、供应链整合与大规模交付上的核心能力有望获得显著的产品溢价和更高的客户粘性。**

■ **行业评级及投资策略：**超节点已成为 AI 基础设施新常态，其具备技术复杂性和快速迭代性，为具备系统级能力的超节点 OEM 厂商构筑了宽阔的护城河，并驱动其盈利能力的提升。在国产 Token 出海，国内 CSP 资本开支展望乐观的背景下，超节点 OEM 厂商将核心受益。基于此，维持对计算机行业“推荐”评级。

■ **相关公司：1）服务器/超节点 OEM：**中科曙光、浪潮信息、华勤技术、紫光股份、工业富联、软通动力、神州数码、中兴通讯；**2）AI 芯片：**海光信息、寒武纪、芯原股份、沐曦股份、摩尔线程、壁仞科技、天数智芯；**3）CPU：**海光信息、中国长城（飞腾信息）、龙芯中科；**4）连接：**澜起科技、盛科通信、锐捷网络、华工科技、华丰科技、鼎通科技；**5）云计算：**协创数据、宏景科技、智微智能、首都在线、金山云、优刻得、云赛智联、网宿科技、深信服；**6）模型：**智谱、MINIMAX、科大讯飞；**7）IDC：**世纪互联、大位科技、润泽科技、东阳光、光环新网、数据港、奥飞数据等。

■ **风险提示：**宏观经济影响下游需求、AI 模型发展不及预期，市场竞争加剧，中美博弈加剧，AI 推理需求不及预期。

事件：

2026 年 3 月 26 日，中科曙光发布超节点新品“曙光 scaleX40”：该产品为世界首个无线缆箱式超节点，采用标准 19 英寸箱式设计，兼容主流机柜；单机 16U，部署密度是 8 卡机的 2.5 倍；最高支持 40 张 GPU 卡，FP8 算力大于 28PFLOPS，HBM 总显存大于 5TB、访存总带宽大于 80TB/S，scale-up 聚合带宽大于 17TB/S。同时，曙光 scaleX40 内置软件栈，兼容 CUDA，适配优化 800 多种大模型。

评论：

1、超节点是什么？

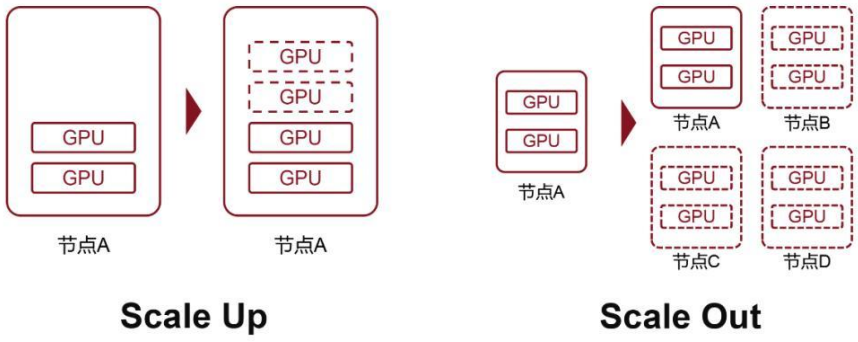
1.1、超节点为构建大规模 AI 算力集群的技术架构

超节点（SuperPod/SuperNode）是一种为构建大规模 AI 算力集群而设计的新型技术架构，最早由英伟达提出。其核心是通过高速互联协议将数十至数百个 GPU 或 AI 计算芯片紧密整合，形成逻辑上统一编址、高带宽、低延迟的协同计算系统。超节点并非简单的硬件堆砌，而是将多个计算设备通过专门的高速互联技术整合为拥有更大内存空间、能协同工作的“逻辑共同体”，让大规模算力真正能够“像一台计算机一样工作、学习、思考、推理”。

图 1：超节点示意图

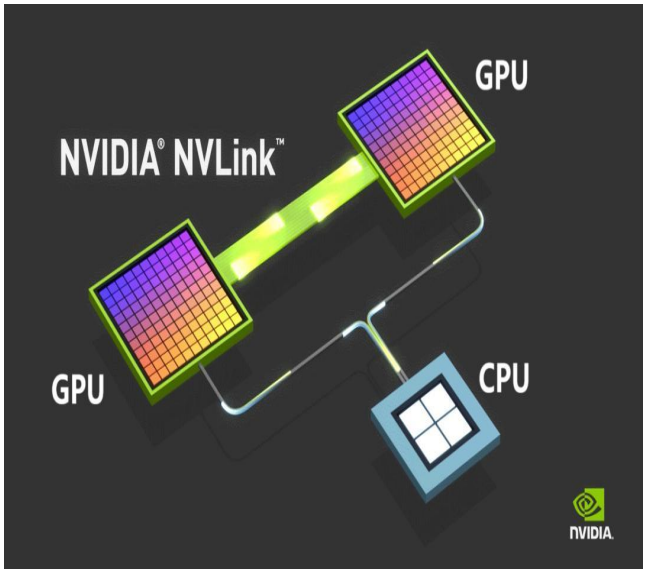
资料来源：DoNews

图 2：英伟达提出通过 Scale Up 和 Scale Out 构建 GPU 集群



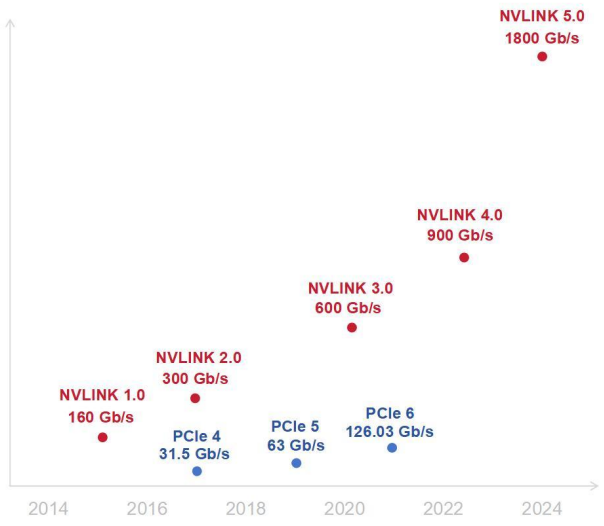
资料来源：鲜枣课堂公众号

图 3：NVLINK 总线协议解决 GPU 之间点对点通信



资料来源：鲜枣课堂公众号

图 4：NVLINK 迭代时间线



资料来源：鲜枣课堂公众号

超节点具备三大关键特征：超大带宽、超低时延和内存统一编址。超大带宽就像高速公路的宽度，确保数据传输通道顺畅；超低时延则像高速公路上没有堵车，避免计算设备等待数据；内存统一编址是核心，它像图书馆的书籍编目系统，所有内存资源提前编号并整合为“共享资源池”，实现快速调取和按需使用。**这些特征共同打破了传统集群架构中的“通信墙”瓶颈——在超高参数大模型训练中，计算单元约 40%的时间都在等待通信。**

超节点的优势主要体现在以下几个方面：

1）超节点能显著提升计算效率，相较于传统集群可以达到 3 倍以上训练性能的提升，通过资源高效调度在一定程度上弥补芯片工艺代差。

2）其次，超节点支持更大规模 AI 处理器的高效协同，为超高参数大模型训练提供基础设施支撑。

3）超节点代表了从“单点算力”到“系统算力”的范式转变，推动算力竞争从峰值性能比拼转向系统效率竞争。

1.2、超节点成 AI 基础设施常态，头部企业持续推出相关产品

目前，超节点已成为 AI 算力基础设施的新常态，以英伟达、AMD、华为、中科曙光、谷歌、阿里巴巴等为代表的头部中美企业正持续推出相关产品。**2025 年被视为超节点产品“元年”，我们认为 2026 年中国国产 AI 超节点或将进入规模化应用阶段。**

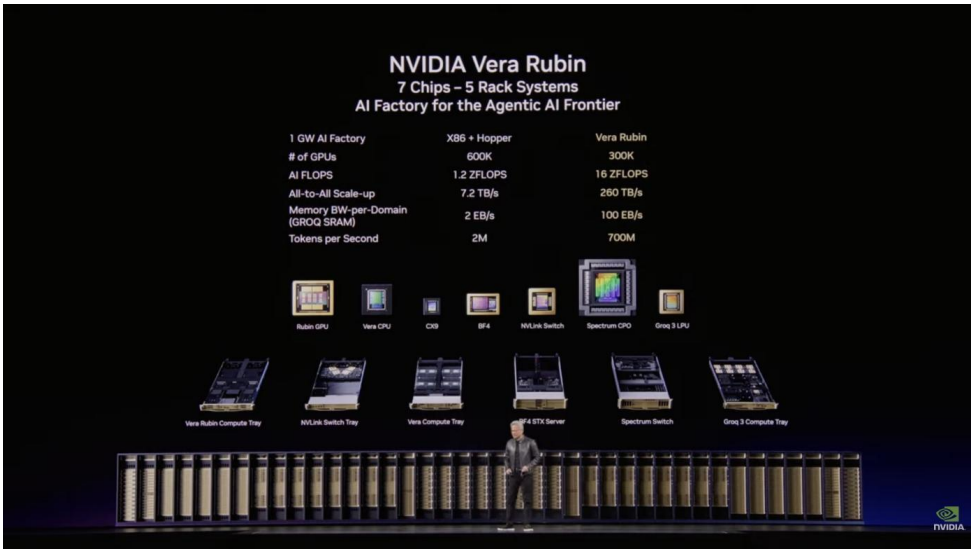
2026 年 3 月，英伟达在 GTC 大会上发布了基于 Vera Rubin 平台的最新超节点产品 Vera Rubin NVL72，内部集成 7 款核心芯片：新一代 Vera CPU、Rubin GPU、NVLink 6 交换机、ConnectX-9 超级网卡、BlueField-4 DPU、Spectrum-6 以太网交换机以及 Groq 3 LPU 推理加速器。

Vera Rubin NVL72 机架将 72 颗 Rubin GPU 与 36 颗 Vera CPU 通过 NVLink 6 高速互联整合在机柜内，形成逻辑统一的协同计算单元。相较前代，**Vera Rubin NVL72 训练大型混合专家模型所需 GPU 数量减少四分之三，推理吞吐量每瓦特提升高达 10 倍，单 token 成本降至十分之一。**

图 5： Vera Rubin NVL72 机架

资料来源：腾讯科技

图 6：内部集成 7 款核心芯片



资料来源：腾讯科技

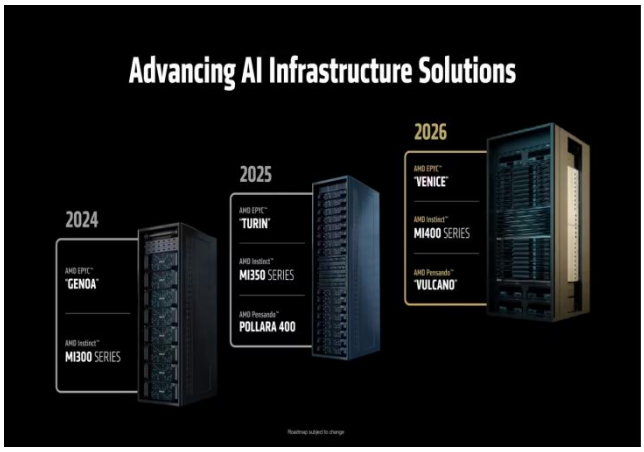
2025 年 6 月，AMD 在 Advancing AI 开发者大会上正式发布了 Helios 机架架构，其内部集成了 72 颗 Instinct MI400 系列 AI 加速器；同时 Helios 采用双宽结构，使得机架内可部署冗余电源模块和高效液冷系统。机架内部采用 18 个计算托盘与 6 个交换托盘交错排列的布局，空间利用率比传统机架提升 40%，为未来升级至 144GPU 配置预留了物理扩展空间。

性能表现方面，内部 Instinct MI400 基于 2nm 制程，配备高达 432GB 的 HBM4 内存，内存带宽达到 19.6TB/s，FP4 算力高达 40PFLOPs。整体看，Helios 机架级 AI 解决方案将整合 18 个 Zen 6 架构的 EPYC “Venice” CPU 和配备了 72 个 Instinct MI455X 系列加速器，拥有总计 31TB 的 HBM4 内存，总内存带宽为 1.4 PB/s，**预计 AI 推理时可实现最高 2.9 FP4 exaFLOPS 和 1.4 FP8 exaFLOPS 用于 AI 训练。**

图 7：AMD Helios 超节点机架

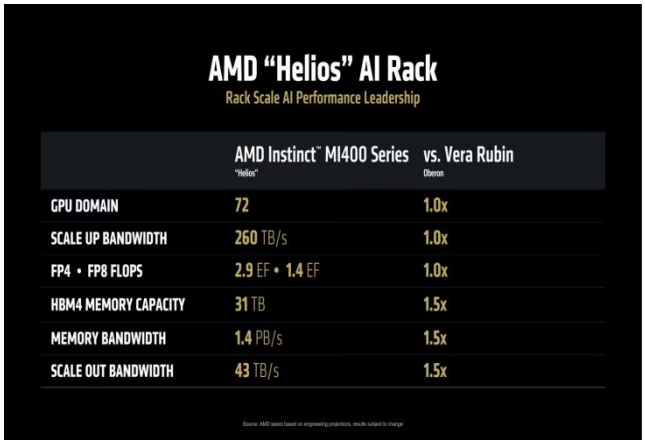
资料来源：IT 之家

图 8：历代 AMD 超节点机架



资料来源：IT 之家

图 9：Helios 机架性能表现

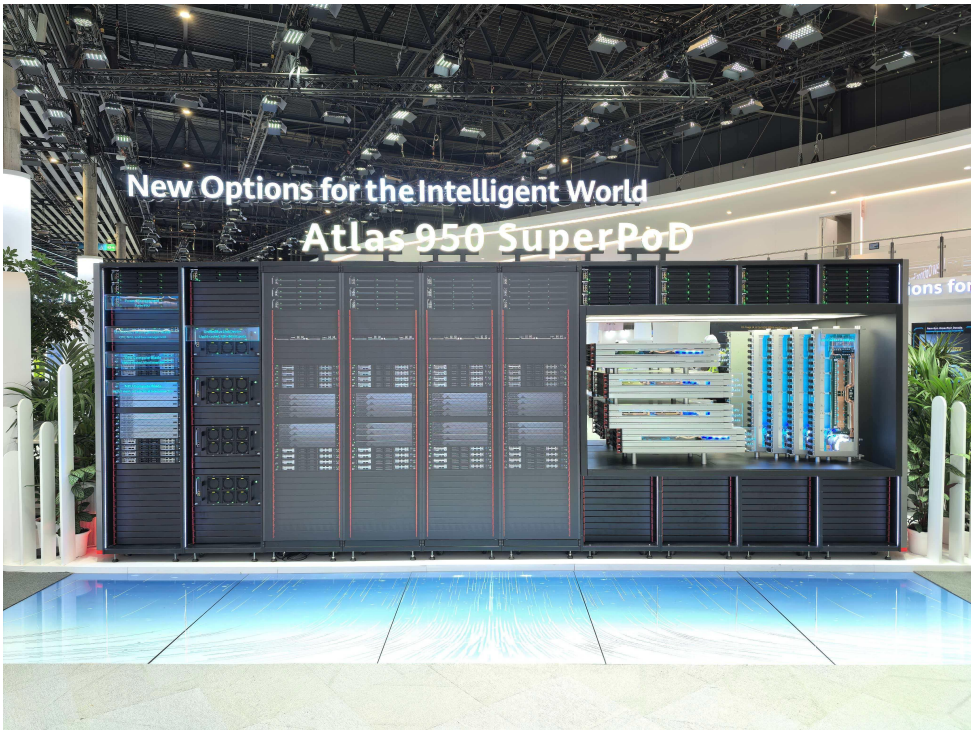


资料来源：IT 之家

2026 年 2 月，华为在 MWC 期间展示了其最新超节点产品矩阵：智算 Atlas 950 SuperPoD、Atlas 850E；通算 TaiShan 950 SuperPoD、TaiShan 500、TaiShan 200 等。其中最新智算 Atlas 950 SuperPoD 基于灵衢互联协议，**以单柜 64 卡为基本单元，最大可支持 8192 张昇腾 NPU 卡高速互联，形成逻辑上统一的超级计算实体。其 FP8 算力达到 8EFLOPS，FP4 算力达 16EFLOPS，分别达到业界水平的 6.7 倍，**并拥有 1152TB 的共享内存池，通过内存统一编址解决大规模并行计算中的数据搬运瓶颈。

在通算领域，TaiShan 950 SuperPoD 作为业界首款通算超节点，具备百纳秒级超低时延、TB 级超大带宽和 48TB 内存池化能力，为数据库、虚机热迁移、大数据处理等传统通算场景开辟了性能提升的全新路径。配合 TaiShan 500、TaiShan 200 等系列服务器，形成了高、中、低全梯度的通算产品体系。

图 10：华为 Atlas 950 SuperPoD 超节点



资料来源：环球网

2026 年 3 月 26 日，中科曙光发布超节点新品“曙光 scaleX40”：该产品为世界首个无线缆箱式超节点，采用标准 19 英寸箱式设计，兼容主流机柜；单机 16U，部署密度是 8 卡机的 2.5 倍；最高支持 40 张 GPU 卡，FP8 算力大于 28PFLOPS，HBM 总显存大于 5TB、访存总带宽大于 80TB/S，scale-up 聚合带宽大于 17TB/S。同时，曙光 scaleX40 内置软件栈，兼容 CUDA，适配优化 800 多种大模型。

而 2025 年 11 月 6 日，中科曙光在乌镇发布了“全球首个单机柜级 640 卡”的 scaleX640。**scaleX640 超节点采用“一拖二”高密架构组成 1280 卡计算单元，总算力规模达到了 630 PFlops，同时采用层次化高速互连网络，HBM 总带宽突破至 2304 TB/s，片间互连总带宽 573 TB/s。**散热方面，曙光 scaleX640 则更进一步通过先进液冷技术，将 PUE 降至 1.04。

图 11：中科曙光 scaleX40

资料来源：智能纪元 AGI 公众号、中关村论坛

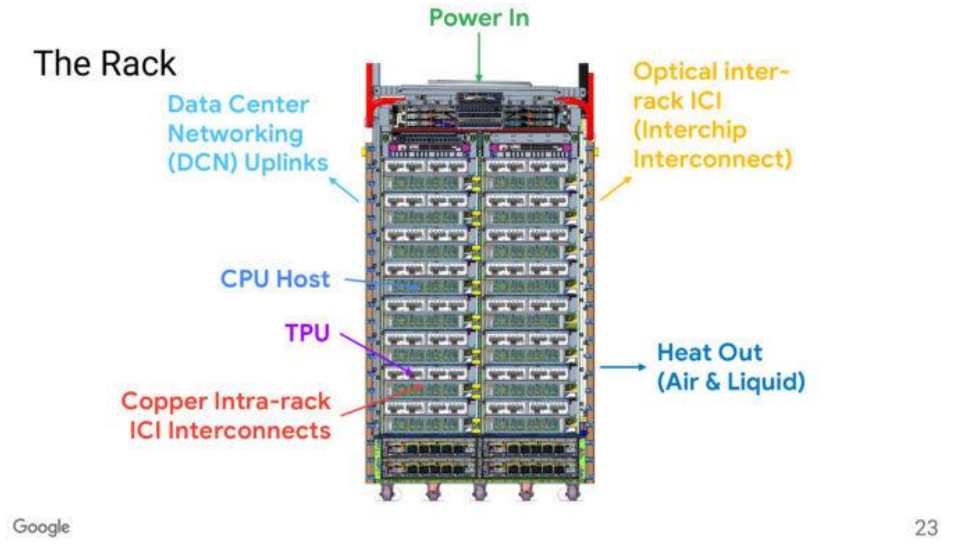
图 12：中科曙光 scaleX640

资料来源：掌上长沙

谷歌最新超节点产品为基于第七代 TPU "Ironwood"构建的 SuperPod 集群，Ironwood 采用了先进的互联架构以应对数据密集型 AI 模型的需求，单个 POD（AI 超级计算机单元）可连接多达 9216 颗 Ironwood 芯片，并通过每秒 9.6 Tb 的突破性 ICI（芯片间互联网络）连接成一个“超级 pod”，访问高达 1.77 PB 的共享高带宽内存 HBM，从而消除数据瓶颈。

在性能表现上，Ironwood SuperPod 内部单芯片 FP8 算力达 4.6 PFLOPS，配备 192GB HBM3e 内存，内存带宽 7.4TB/s，整个集群 FP8 峰值性能超过 42.5 exaFLOPS。该超节点专为大规模 AI 推理优化，现已在谷歌云数据中心规模部署。

图 13： 谷歌 Ironwood 推理超节点



资料来源：IT 之家

在 2025 云栖大会上，阿里发布全新一代磐久 128 超节点 AI 服务器。整柜采取定制双宽机柜方式，**支持 128~144 颗 GPU 芯片、高达 350kw 供电能力和 500kw 散热能力**，采用 BusBar 柜内集中供电。**该超节点服务器还集成了阿里自研 CIPU2.0 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 带宽和百纳秒极低延迟，**相对于传统架构，同等 AI 算力下推理性能还可提升 50%。**

图 14： 磐久 128 超节点 AI 推理超节点服务器

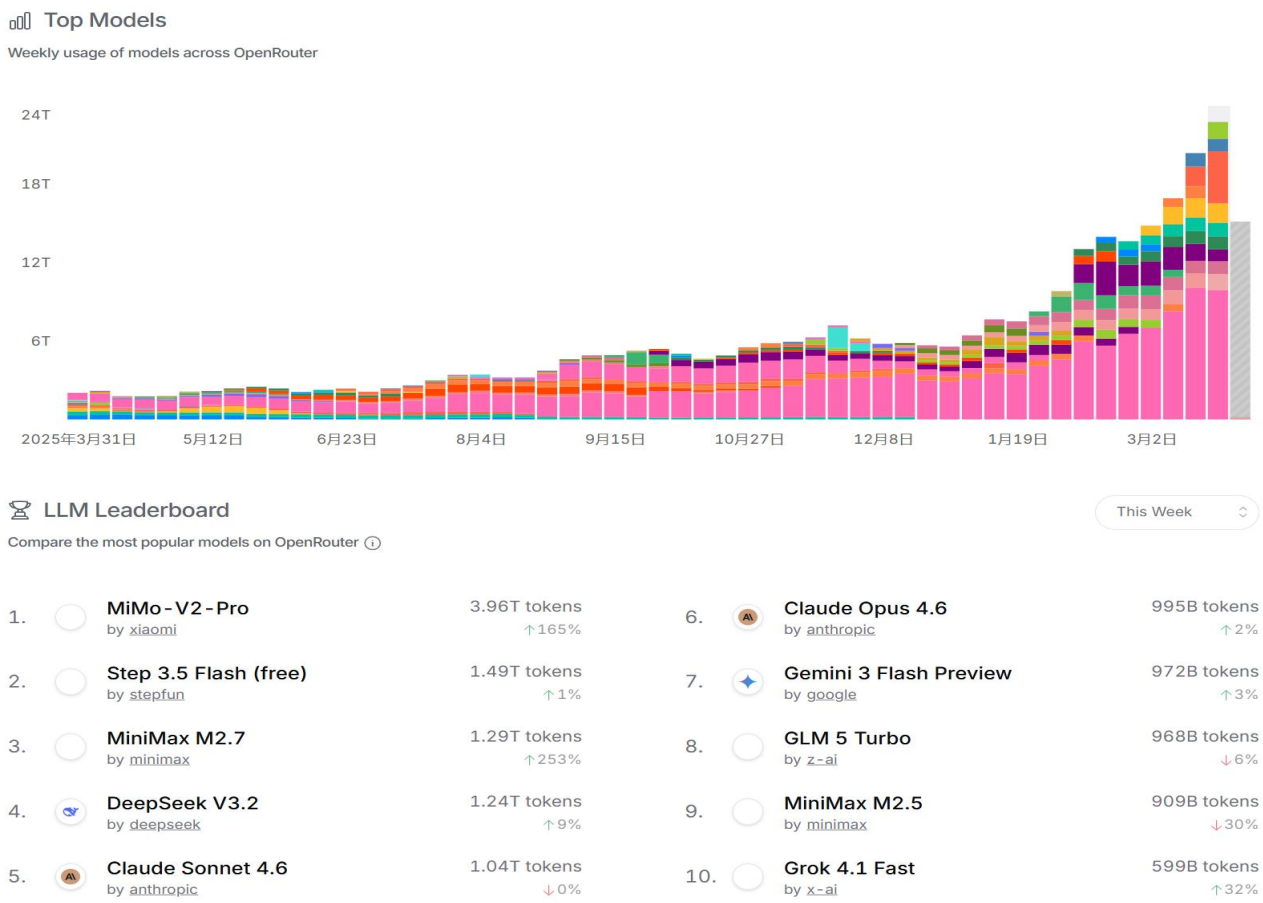
资料来源：阿里云

2、国内 CSP 厂商 AI 资本开支展望乐观

2025 年国内整体算力资本开支规处于追赶阶段。但 2026 年，国内 CSP 厂商资本开支展望均较为乐观。

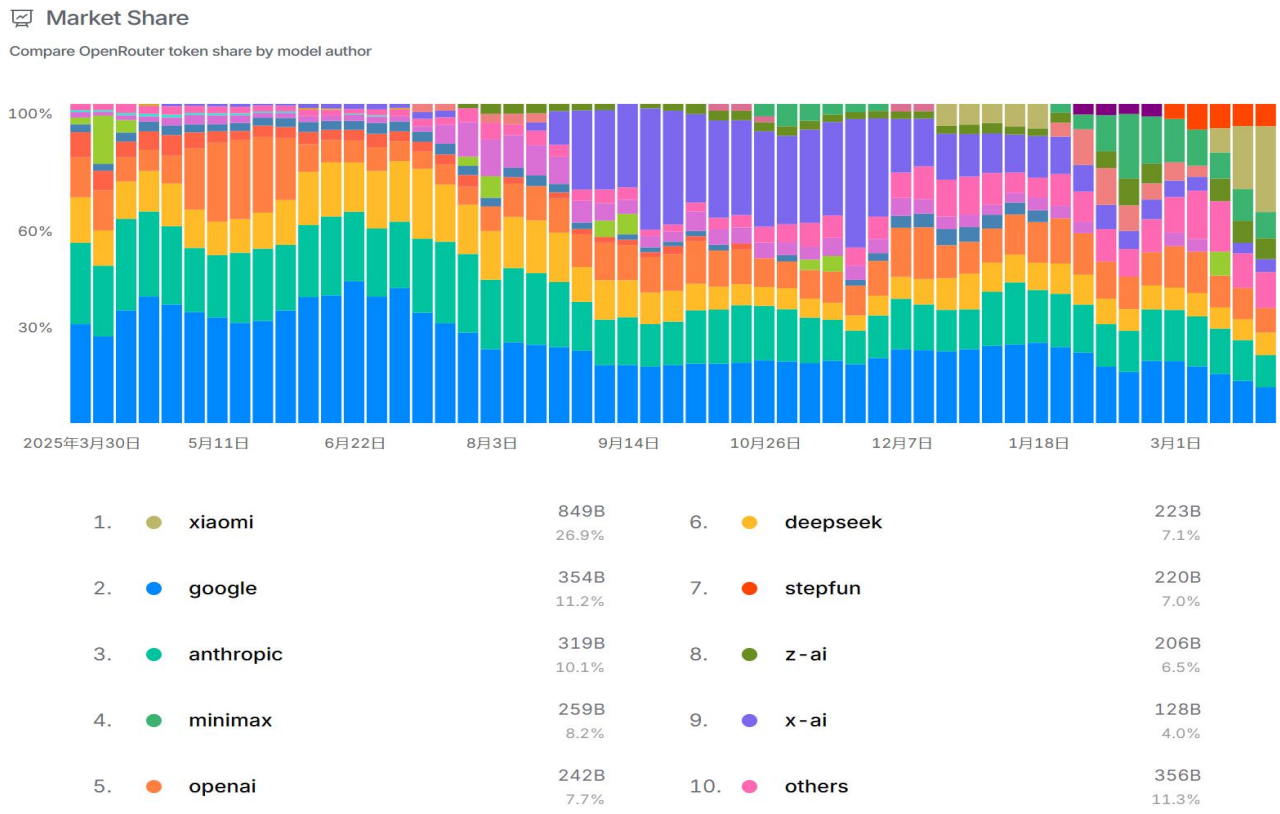
从需求侧看，中国在 AI 推理时代具备独特优势，庞大的用户基数和丰富的应用场景为模型调用提供了广阔土壤。OpenRouter 数据显示，截至 2026 年 3 月 22 日的连续三周内，国产大模型调用量保持对美国模型的反超。上周（3.16-3.22）全球模型调用量排名榜中，国产模型占五席，分别为小米集团 MiMo-V2-Pro、阶跃星辰 Step 3.5 Flash、MiniMax M2.5、DeepSeek V3.2、智谱 GLM5 Turbo，总调用量为 7.359T，较此前一周的 4.69T 增长 56.9%，增幅进一步扩大；3 月 1 日至 3 月 30 日，调用量前十模型中，中国模型总量占比超 50%。

图 15：Openrouter3.16-3.22 模型排行榜



资料来源：Openrouter

图 16：Openrouter3.16-3.22 模型厂商份额



资料来源：Openrouter

从供给侧看，国内晶圆厂、英伟达、算力租赁公司产能、订单等景气度均走高：1）**中芯国际：**中芯国际月产能由 2025Q3 的 102.275 万片折合 8 英寸标准逻辑增加至 2025Q4 的 105.875 万片，第四季度销售晶圆达 251.4 万片折合 8 英寸标准逻辑，产能利用率与 Q3 基本持平为 95.7%。**2）华虹半导体 2026-2027 持续扩产：**据无锡市公共资源交易平台，2026 年 1 月 6 日，华虹 FAB9B 项目工程总承包开启招标，总投资额达 38 亿元，计划建设一条月产能达到 5.5 万片的 12 英寸特色工艺生产线。**3）英伟达 H200 获生产许可：**英伟达在 GTC2026 表示，H200 已经拿到许可并获客户订单，目前正在生产过程中。**4）算力租赁：**宏景科技拟于 2026 年申请不超过 600 亿元授信额度同时披露 13.5 亿元定增方案，计划采购高性能算力服务器后为客户提供算力服务。协创数据披露 2025 年至 2026Q1 对外算力服务器采购额达 400 亿元+。

3、超节点方案复杂度较高，驱动 OEM 厂商盈利能力提升

相较传统的 GPU 集群， 超节点在通信时延、算力密度、成本和效率，机房空间利用和 TCO 总成本等方面具有优势，也具备更高的技术壁垒。在整机的系统架构设计、信号完整性、供电、散热、深度耦合上要求更高，产品规格处于非常快速的迭代中。例如，NVIDIA Vera Rubin 机架共包含 130 万个独立组件，近 1300 个芯片，整体重量约为 2 吨等。

正是这种极高的技术复杂性和快速迭代，为具备系统级能力的超节点 OEM 厂商构筑了宽阔的护城河，并直接驱动其盈利能力的提升。我们认为超节点时代，价值重心或已从标准化、可替代的硬件组装，上移至复杂的定制化系统设计、深度调优和全栈集成服务，**因而超节点 OEM 厂商凭借其在架构设计、热管理、供应链整合与大规模交付上的核心能力有望获得显著的产品溢价和更高的客户粘性。**

图 17：NVIDIA Vera Rubin2026 内部架构

Comparativa técnica: Evolución de los sistemas Nvidia

Especificación	Grace Blackwell (2025)	Vera Rubin (2026)	Mejora / Impacto
GPU por Rack	72 Blackwell	72 Rubin	Mayor densidad de cómputo
Eficiencia Energética	1x (Base)	10x por vatio	Reducción drástica de calor
Componentes Totales	~600,000	1,300,000	Complejidad de ingeniería
Proveedores Globales	~50 países	80 proveedores / 20 países	Optimización de suministro

资料来源：PConline 太平洋科技

4、投资建议

超节点已成为 AI 基础设施新常态，其具备技术复杂性和快速迭代性，为具备系统级能力的超节点 OEM 厂商构筑了宽阔的护城河，并驱动其盈利能力的提升；在国产 Token 出海，国内 CSP 资本开支展望乐观的背景下，超节点 OEM 厂商将核心受益。基于此，维持对计算机行业“推荐”评级。

相关公司：

1）服务器/超节点 OEM：中科曙光、浪潮信息、华勤技术、紫光股份、工业富联、软通动力、神州数码、中兴通讯；**2）AI 芯片：**海光信息、寒武纪、芯原股份、沐曦股份、摩尔线程、壁仞科技、天数智芯；**3）CPU：**海光信息、中国长城（飞腾信息）、龙芯中科；**4）连接：**澜起科技、盛科通信、锐捷网络、华工科技、华丰科技、鼎通科技；**5）云计算：**协创数据、宏景科技、智微智能、首都在线、金山云、优刻得、云赛智联、网宿科技、深信服；**6）模型：**智谱、MINIMAX、科大讯飞；**7）IDC：**世纪互联、大位科技、润泽科技、东阳光、光环新网、数据港、奥飞数据等。

5、风险提示

宏观经济影响下游需求、AI 模型发展不及预期，市场竞争加剧，中美博弈加剧，AI 推理需求不及预期。

【计算机小组介绍】

刘熹，计算机行业首席分析师，上海交通大学硕士，多年计算机行业研究经验，善于把握由技术、政策驱动的科技产业新趋势，致力于进行前瞻重磅推荐。2024 年 Wind 金牌分析师，2024 年同花顺最受欢迎分析师。
唐锦珂，计算机行业研究助理，中山大学岭南学院硕士，主要覆盖服务器、液冷、算力租赁、IDC 等方向。
谢婧茹，计算机行业研究助理，上海财经大学经济学院硕士，主要覆盖半导体、云计算、教育 IT 等方向。
朱谓晟，计算机行业研究助理，复旦大学金融硕士，主要覆盖财税 IT、工业软件、基础软件等方向。

【分析师承诺】

刘熹，本报告中的分析师均具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立，客观的出具本报告。本报告清晰准确的反映了分析师本人的研究观点。分析师本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收取到任何形式的补偿。

【国海证券投资评级标准】

行业投资评级

推荐：行业基本面向好，行业指数领先沪深 300 指数；
中性：行业基本面稳定，行业指数跟随沪深 300 指数；
回避：行业基本面向淡，行业指数落后沪深 300 指数。

股票投资评级

买入：相对沪深 300 指数涨幅 20%以上；
增持：相对沪深 300 指数涨幅介于 10%～20%之间；
中性：相对沪深 300 指数涨幅介于-10%～10%之间；
卖出：相对沪深 300 指数跌幅 10%以上。

【免责声明】

本报告的风险等级定级为 R3，仅供符合国海证券股份有限公司（简称“本公司”）投资者适当性管理要求的客户（简称“客户”）使用。本公司不会因接收人收到本报告而视其为客户。客户及/或投资者应当认识到有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司的完整报告为准，本公司接受客户的后续问询。

本公司具有中国证监会许可的证券投资咨询业务资格。本报告中的信息均来源于公开资料及合法获得的相关内部外部报告资料，本公司对这些信息的准确性及完整性不作任何保证，不保证其中的信息已做最新变更，也不保证相关的建议不会发生任何变更。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。报告中的内容和意见仅供参考，在任何情况下，本报告中所表达的意见并不构成对所述证券买卖的出价和征价。本公司及其本公司员工对使用本报告及其内容所引发的任何直接或间接损失概不负责。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或者金融产品等服务。本公司在知晓范围内依法合规地履行披露义务。

【风险提示】

市场有风险，投资需谨慎。投资者不应将本报告视为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向本公司或其他专业人士咨询并谨慎决策。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，请公众投资者慎重使用未经授权刊载或者转发的本公司的证券研究报告。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息。本报告不构成本公司向该机构之客户提供的投资建议。

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。本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

【郑重声明】