2026 年 05 月 02 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

国内算力黄金年代

本周观点

- 自 2026 年 1 月 11 日发布《国内算力斜率陡峭》起，我们连续多次发表报告观点/公开电话会，坚定看好国内算力产业链的高景气行情，国产机柜、算力租赁、CPU 等国内算力链均发布专题报告进行重点推荐。
- **本轮国内算力行情的风眼——国产模型能力跨过临界点。**1）本轮国内算力链的本质，一方面在于国产大模型能力突破关键水平，DeepSeek-V4、GLM-5 等国内模型 Agent 能力快速进阶，可以支撑不复杂的 Coding 和 Agent 应用，进而真实堆高国内算力需求；2）另一方面，国内大模型具备算法优化工程优势，整体推理算力成本低于海外，同等能力下中国头部模型的 API 价格仅为海外顶尖模型的极小比例，高性价比使得国产大模型成功争夺海外上一代大模型的市场份额，充分享受 Token 爆发的红利。
- **国产模型 ARR 和 Token 量指数级增长，真实推高国内算力全链景气度。**目前，国内模型厂 Token 与 ARR 均实现跃进式增长，Minimax ARR 已超过 1.5 亿美元；智谱 API ARR 已突破 2.5 亿美元，三度涨价仍供不应求，2026Q1 token 调用量增长 400%，真实广阔的商业化需求正转化为极速释放的算力消耗。在供需双侧强逻辑的挤压下，2026 年算力产业链已实质性进入全链通胀阶段，算力短缺程度持续升温，行业景气度正从核心芯片向 AIDC、云与算力服务等环节全面外溢：CPU/GPU 面临产能瓶颈与租赁价格持续飙升，阿里、百度等大厂的云端算力资源及配套服务亦开启多轮提价。多环节的持续涨价，从侧面强力印证了真实算力需求的爆发与供给的极度紧缺。
- **国内算力黄金年代开启，供给紧缺下好用的算力本身就是最核心的资产。**今年以来，CPU、云、算力租赁均开启涨价周期，算力链全面通胀，验证需求爆发下算力紧缺程度之深。1）国产芯片：大厂加速适配，需求爆炸+供给改善。Deepseek-V4 发布后，昇腾、寒武纪 Day 0 首发适配，后续海光信息、摩尔线程、沐曦股份、百度昆仑芯等 10 家国内芯片厂商陆续官宣支持 Deepseek-V4，国内模型厂推理侧对多家国内芯片厂商的广泛适配；从寒武纪一季报看，预付及合同负债双高增，印证国产芯片迎来需求爆炸+供给改善信号。2）国产机柜：超节点出货带来 ODM 组装厂商格局/利润率双提升。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，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。3）算租/云/IDC：①云：模型调用带来云计算需求膨胀，Token 调用爆发驱动大厂云优先涨价，第三方云跟随；②算力租赁：好用的算力本身就是最核心的稀缺资产，在需求爆发与业绩加速双重印证下，算租厂商的商业化 ROI 正加速兑现；③AIDC：海内外大厂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。

相关标的

国内算力：寒武纪、海光信息、东阳光、利通电子、协创数据、杰华特、利扬芯片、华勤技术、浪潮信息、禾盛新材、中国长城、罗曼股份、网宿科技、盈峰环境、芯原股份、华丰科技、晶科科技、亿田智能、豫能控股、星环科技、鸿日达、盛视科技、首都在线、神州数码、百度集团、中芯国际、华虹半导体、中科曙光、润泽科技、大位科技、润建股份、奥飞数据、云赛智联、瑞晟智能、科华数据、潍柴重机、金山云、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

图表目录

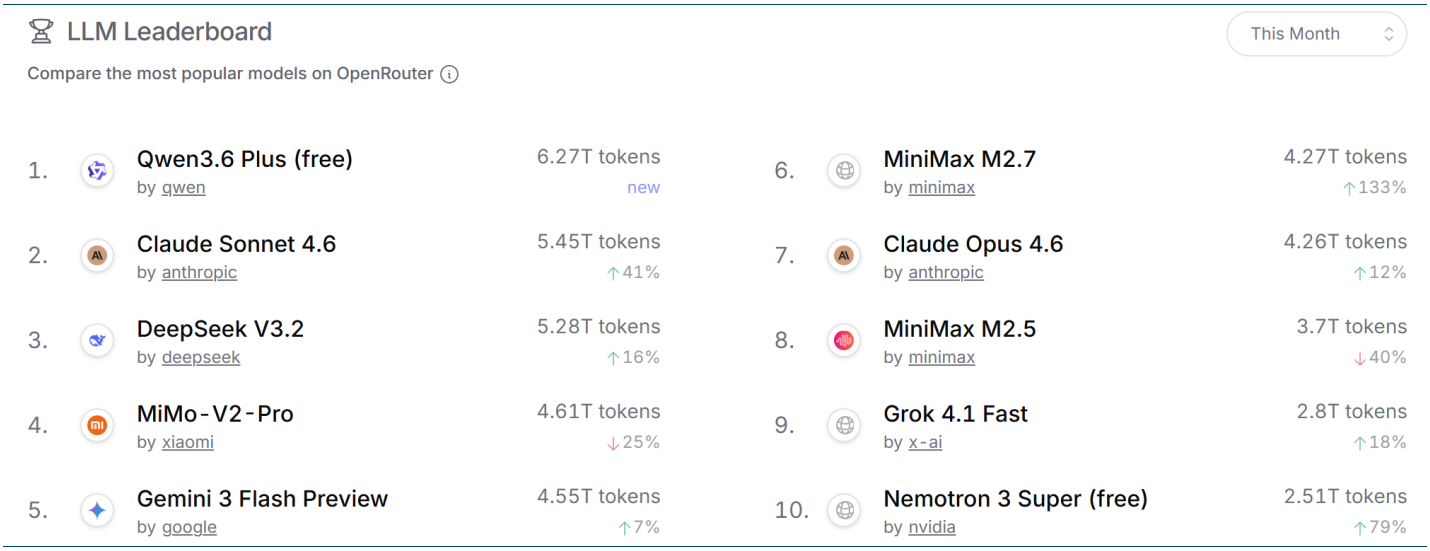
一、本轮国内算力行情的风眼——国产模型能力跨过临界点

自 2026 年 1 月 11 日发布《国内算力斜率陡峭》行业专题起，我们连续多次发表报告观点，坚定看好国内算力产业链的高景气行情，国产机柜、算力租赁、CPU 等国内算力链均发布专题报告进行重点推荐。我们认为本轮国内算力链的本质，一方面在于国产大模型能力突破关键水平，可以支撑不复杂的 Coding 和 Agent 应用，进而真实堆高国内算力需求；另一方面，国内大模型具备算法优化工程优势，整体推理算力成本低于海外，高性价比使得国产大模型成功争夺海外上一代大模型的市场份额，从而享受到 token 爆发的红利。

1.1 国产大模型能力持续进阶，从“抽卡”跨越至“稳定输出可用结果”

国产模型调用量持续居于高位，登上世界 AI 舞台。据 OpenRouter 数据，3 月 30 日-4 月 5 日，中国 AI 大模型周调用量达 12.96 万亿 token，环比增长 31.48%，调用量约为美国的 4.3 倍，连续五周超越美国。4 月，模型竞赛持续升温，国产模型围绕代码、Agent 集群、长程任务处理等能力进行升级，发布后调用量迅速攀升。阿里 Qwen3.6-Plus 于 4 月 2 日发布后仅 1 天，便以 1.4 万亿 token 调用量登顶 OpenRouter 日榜，并打破该平台的单日单模型调用量全球纪录；Kimi K2.6 发布后首周（4 月 20 日-26 日）以 1.58 万亿 token 的周调用量位居 OpenRouter 平台榜首；DeepSeek-V4 发布次日（4 月 25 日），DeepSeek-V4-Flash 调用量达到 502 亿 Token，较前日增长 85.9%；DeepSeek-V4-Pro 的调用量达到 136 亿 Token，较前日增长近四倍。3 月 30 日-4 月 30 日，全球 AI 大模型总调用量前十阵营中，中国 AI 大模型占据五席，跻身全球 AI 第一梯队。

图表1：3 月 30 日-4 月 30 日全球 AI 大模型总调用量前十阵营中，中国 AI 大模型占据五席

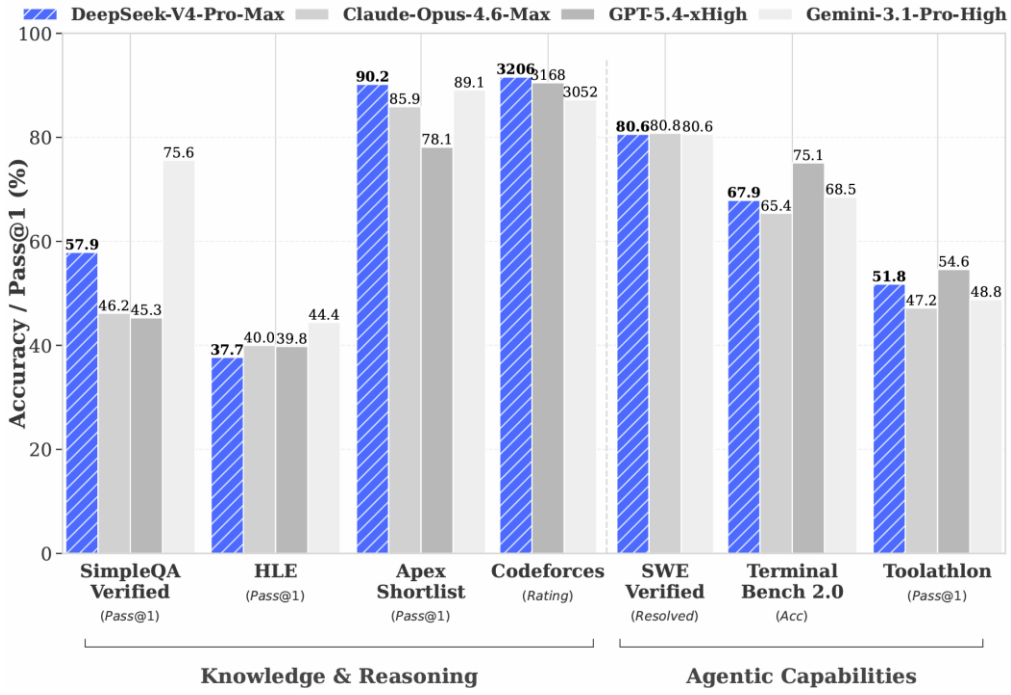


来源：OpenRouter，国金证券研究所

DeepSeek-V4 标配 1M 上下文，Agent 能力全面升级。4 月 24 日，DeepSeek 发布 DeepSeek-V4-Pro 与 DeepSeek-V4-Flash 两款基座模型，首次原生支持 1M 上下文窗口，在世界知识和推理性能上取得重要突破。

- 能力紧逼海外顶尖闭源模型。V4-Pro 在知识维度测试 SimpleQA-Verified 中得分 57.9%，领先所有开源模型 20 个百分点；长文本关键信息检索 MRCR 1M 测试准确率为 83.5%，优于 Gemin-3.1-Pro（76.3%），稍弱于 Opus-4.6（92.9%）；编程竞赛基准 Codeforces 获得 3206 Rating 评分，超越 GPT-5.4（3168）和 Gemini-3.1-Pro（3052）；在 Agentic 测评中，V4-Pro-Max 在 SWE Verified（80.6%）中表现逼近 Opus-4.6（80.8%）、现实经济任务 GDPval-AA（1554）评测中表现不俗 Opus-4.6（1619）和 GPT-5.4（1674），展现出极强的工具调用与端到端执行能力。

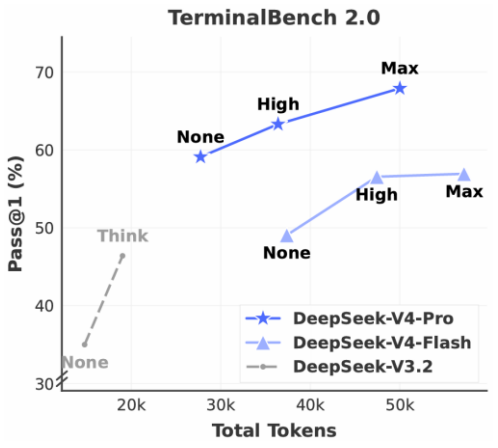
图表2：DeepSeek-V4 的代理、推理、代码能力逼近或超越海外顶尖闭源模型



来源：DeepSeek 技术报告，国金证券研究所

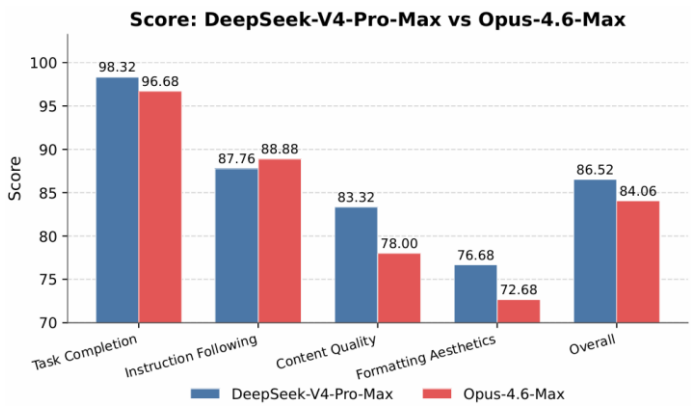
- Agent 能力深度优化。在 Agentic Coding 评测中，V4-Pro 达到当前开源模型最佳水平。DeepSeek-V4 针对 Claude Code、OpenClaw、OpenCode、CodeBuddy 等主流的 Agent 产品进行了适配优化，代码任务、文档生成任务等方面表现均有提升；OpenClaw 2026.4.24 版本更新后正式接入 DeepSeek-V4 两款大模型，并将 DeepSeek-V4-Flash 设定为默认模型。另外，V4 在 Agent 方向进行了专项优化：1）后训练阶段把 Agent 作为与数学、代码并列的独立专家方向单独训练；2）工具调用格式从 JSON 换成带特殊 token 的 XML 结构，用来降低转义错误；3）跨轮次推理痕迹在工具调用场景下完整保留，不再像 V3.2 那样每轮清空；此外，DeepSeek 自建 DSec 的沙箱平台，单集群可并发管理数十万个沙箱实例，用来支撑 Agent 强化学习训练和评测。

图表3：在 TerminalBench 2.0 测试中，DeepSeek-V4 系列模型表现显著优于 DeepSeek-V3.2



来源：DeepSeek 技术报告，国金证券研究所

图表4：白领任务中，DeepSeek-V4-Pro-Max 在完成度、内容质量、格式美观度等方面优与 Opus-4.6-Max



来源：DeepSeek 技术报告，国金证券研究所

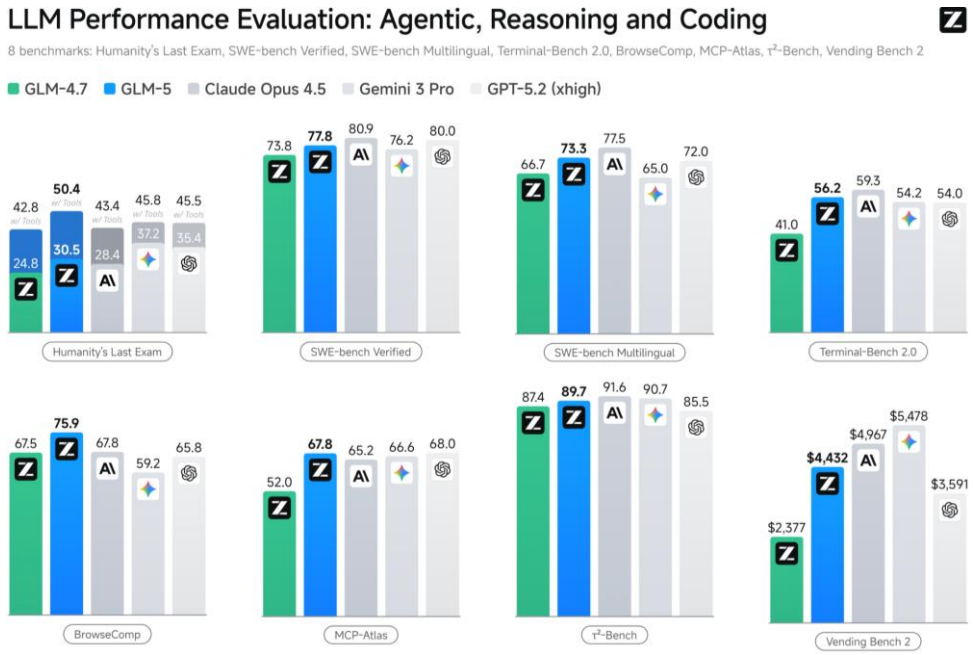
注：TerminalBench 2.0 为 Agent 全链路工程能力测试

智谱 GLM-5 在 Coding 与 Agent 能力上取得开源 SOTA。2 月 12 日，智谱上线并开源 GLM-5，其在真实编程场景的使用体感逼近 Claude Opus 4.5，擅长复杂系统工程与长程 Agent 任务。在全球权威的 Artificial Analysis 榜单中，GLM-5 位居全球第四、开源第一。

- 基座能力全面演进：1）参数规模扩展：从 355B（激活 32B）扩展至 744B（激活 40B），预训练数据从 23T 提升至 28.5T，更大规模的预训练算力显著提升了模型的通用智能水平；2）异步强化学习：构建全新的“Slime”框架，支持更大模型规模及更复杂的强化学习任务，提升强化学习后训练流程效率；提出异步智能体强化学习算法，

使模型能够持续从长程交互中学习，充分激发预训练模型的潜力；3）稀疏注意力机制：首次集成 DeepSeek Sparse Attention，在维持长文本效果无损的同时，大幅降低模型部署成本，提升 Token Efficiency。

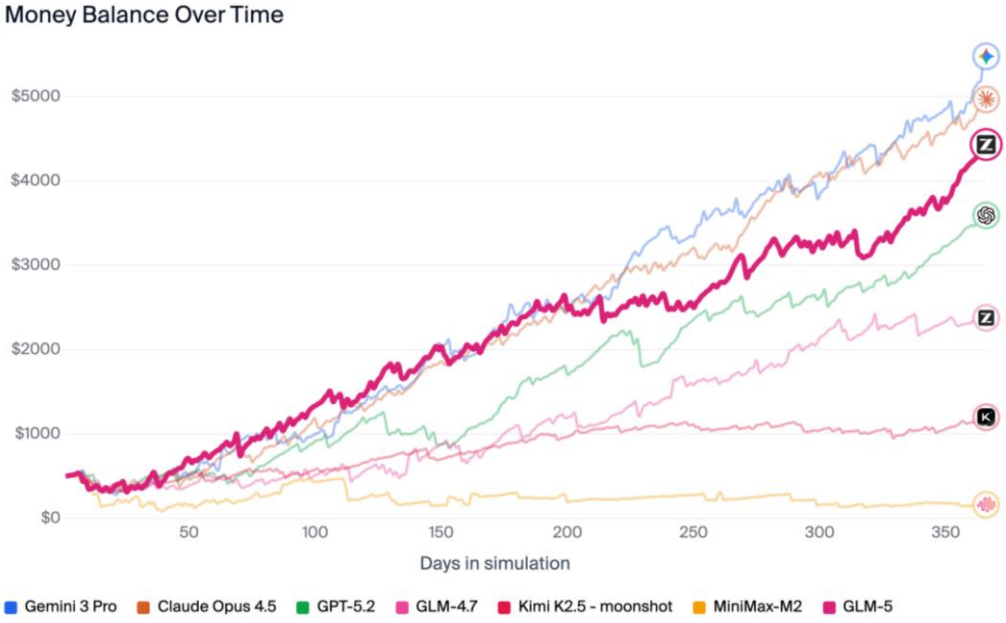
图表5：GLM-5 的代理、推理、代码能力出色



来源：智谱微信公众号，国金证券研究所

- Agentic 工程能力大幅提升。1）代码能力：GLM-5 在 SWE-bench-Verified 和 Terminal Bench 2.0 中分别获得 77.8 和 56.2 的开源模型 SOTA 分数，性能超过 Gemini 3 Pro，具备稳定交付生产结果的能力。在内部 Claude Code 评估集合中，GLM-5 在前端、后端、长程任务等编程开发任务上显著超越 GLM-4.7（平均增幅超过 20%），能够以极少的人工干预自主完成 Agentic 长程规划与执行、后端重构和深度调试等系统工程任务，使用体感逼近 Opus 4.5。2）Agent 能力：GLM-5 在 BrowseComp（联网检索与信息理解）、MCP-Atlas（工具调用和多步骤任务执行）等 Agent 能力评测基准中取得开源第一；在衡量模型经营长期能力的 Vending Bench 2 中展现了出色的长期规划和资源管理能力，表现接近 Claude Opus 4.5。GLM-5 能够在长程任务中保持目标一致性、进行资源管理、处理多步骤依赖关系，推动编程范式从 Vibe Coding（氛围编程）转向 Agentic Engineering（智能体工程）。

图表6：在要求模型在一年期内经营一个模拟自动售货机业务的 Vending Bench 2 测试中，GLM-5 最终账户余额达到 4432 美元，经营表现接近 Claude Opus 4.5



来源：智谱微信公众号，国金证券研究所

1.2 极致的算法优化下，推理性价比优势明显

极致的算法优化下，国产大模型推理算力成本优势凸显。国内大模型普遍采用 MoE 架构，可大幅降低显存占用、提升推理吞吐量，压缩推理成本。此种从底层工程驱动的结构性降本增效，使得同等能力下，中国头部模型的 API 价格仅为海外顶尖模型的极小比例，性价比优势尽显。

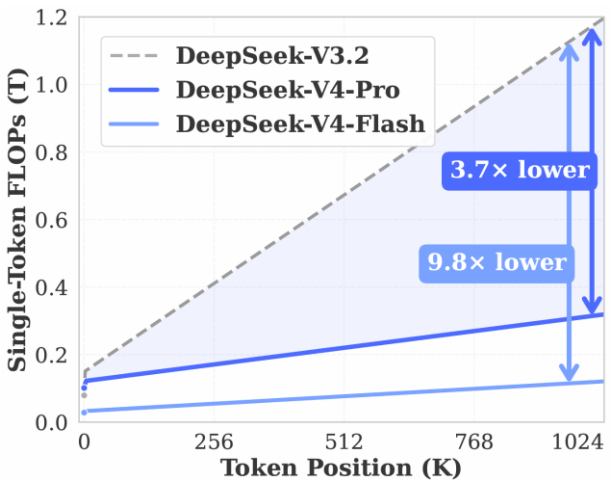
图表7：国产模型价格优势显著

模型	输入价格（每百万 token ）	输出价格（每百万 token ）
GPT-5.5	0.5（缓存输入）/5 美元	30 美元
Gemini 3.1 Pro	2（<=20 万 token）/4（>20 token）美元	12（<=20 万 token）/18（>20 token）美元
Claude Opus 4.7	5 美元	25 美元
Claude Sonnet 4.6	3 美元	15 美元
DeepSeek-V4-Flash	0.2（缓存命中）/1（未命中）元	2 元
DeepSeek-V4-Pro	1（缓存命中）/12（未命中）元	24 元
GLM-5.1	6（输入长度＜32K）/8（输入长度≥>32K）元	24（输入长度＜32K）/28（输入长度≥>32K）元
Qwen3.6-Plus	2 元	12 元
KIMI K2.6	1.1（缓存命中）/6.5（未命中）元	27 元
MiniMax-M2.7	2.1 元	8.4 元

来源：各公司官网，国金证券研究所

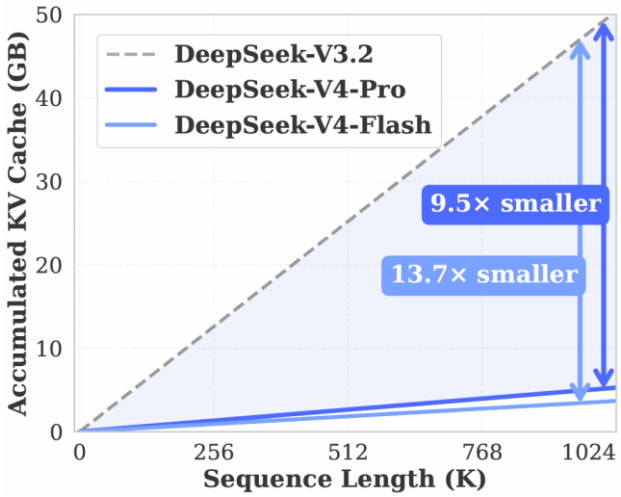
DeepSeek 通过架构创新大幅降低了计算和内存成本。DeepSeek-V4 采用 CSA+HCA 混合注意力、流形约束超连接等核心技术，大幅降低了显存占用与计算开销。V4-Pro 与 V4-Flash 最大上下文长度为 1M，技术报告数据显示，在 100 万 Token 场景下，相比 V3.2，V4-Pro 单 Token 推理 FLOPs 相比 V3.2 降低 3.7 倍，KV Cache 降低 9.5 倍；V4-Flash 进一步降低至 FLOPs 的 1/9.8、KV Cache 的 1/13.7。这意味着处理同等长度上下文的硬件成本大幅下降，使百万 Token 推理在商业环境中具备实际可行性。

图表8：DeepSeek-V4 和 DeepSeek-V3.2 的计算量随上下文长度的变化



来源：DeepSeek 技术报告，国金证券研究所

图表9：DeepSeek-V4 和 DeepSeek-V3.2 的显存容量随上下文长度的变化



来源：DeepSeek 技术报告，国金证券研究所

DeepSeek-V4 为高性能高性价比模型，开启百万上下文普惠时代。DeepSeek-V4-Pro 于 4 月 24 日上线后，百万 token 输入成本为 1 元（缓存命中）/12（缓存未命中），百万 token 输出成本为 24 元。4 月 25 日，DeepSeek 宣布对 V4-Pro 模型 API 开启限时 2.5 折价格优惠；4 月 26 日，DeepSeek 又宣布输入缓存命中的价格降至首发价格的 1/10。最新调价后，DeepSeek-V4-Pro 百万 token 输入成本为 0.025 元（缓存命中）/3（缓存未命中），百万 token 输出成本为 12 元。另外，DeepSeek 官方表示，受限于高端算力，目前 Pro 的服务吞吐十分有限，预计下半年昇腾 950 超节点批量上市后，Pro 的价格会大幅下调。

二、国产模型 ARR 和 Token 量指数级增长，真实推高国内算力全链景气度。

2.1 ARR 和 Token 迈入非线性增长通道

国内模型厂 ARR 跃进式增长，真实且广阔的商业化需求正持续转化。截至 2026 年 2 月，Minimax ARR 已超过 1.5 亿美元，是 2025 年全年营收近 2 倍，2026 年 2 月新注册用户数已经达到 2025 年 12 月的 4 倍以上。截至 2026 年 3 月，智谱 API ARR 已突破 2.5 亿美元，同比增长 60 倍，年初以来增长 6.4 倍。智谱今年以三度涨价，2 月 12 日 GLM Coding Plan 整体涨幅 30%起、3 月 16 日 GLM-5-Turbo API 涨价 20%、4 月 8 日 GLM 全系列模型再度提价 10%；多次涨价下市场仍然供不应求，截至 3 月底，智谱 API 调用价格相比去年底提升 83%、调用量增长 400%。

今年以来，随着 OpenClaw、Hermes 等智能体应用及多 Agent 协同架构加速落地，token 使用量较去年底已实现约 2-3 倍增长，需求侧快速放量。国产大模型在 Agent 能力上持续迭代，交付质量与稳定性显著提升，从“可用”向“好用”全面升级；同时，依托算法优化与工程化能力积累，国内厂商在算力利用率与系统效率方面具备优势，使得推理成本低于海外同类模型。模型能力升级与成本优势共振驱动下，国产大模型性价比优势进一步放大，竞争力持续提升，有望充分受益于 Agent 驱动的 token 需求爆发机遇。

2.2 算力链全面通胀，多环节涨价侧面印证算力需求

在供需双侧强逻辑的挤压下， 2026 年算力产业链将进入“全链通胀”周期，行业景气度将从核心芯片向 AIDC、云与算力服务、配套电力设备及服务器等环节全面外溢。

CPU 短缺加剧，涨价周期来临。推理时代已至，Agentic AI 多元化场景加速落地，CPU 作为调度中心的战略价值显著上升，需求爆发下，多家 x86 厂商的服务器 CPU 库存售罄、价格上涨、平均交货周期大幅延长。据日经亚洲 3 月 25 日报道，英特尔与 AMD 已各自通知客户，将分别于 3 月和 4 月起上调全系列 CPU 价格，平均涨幅达 10-15%，部分产品涨幅更高；同时，交货周期将从之前的 1-2 周大幅延长至 8-12 周，个别情况下甚至将长达 6 个月。AI 算力需求爆炸式增长，AI 芯片巨头占用大量原材料与产能，英特尔与 AMD 面临产能扩张瓶颈，叠加原材料价格上涨，CPU 供给端持续承压，供需错配加剧，CPU 价格进入上行通道。

图表10：算力全链共振，步入涨价周期

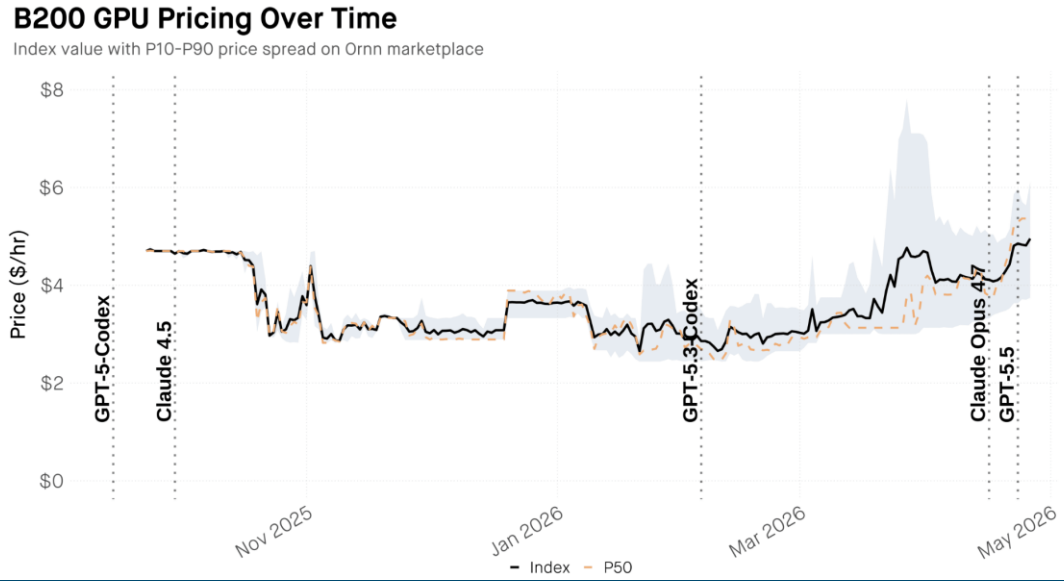
厂商	公告时间	具体内容
AWS	1 月 4 日	EC2 机器学习容量块价格上调约 15%，包括由 NVIDIA GPU 驱动的 P5en、P5e、P5、P4d，以及使用 AWS Trainium 的 Trn2 和 Trn1 实例，P5e.48xlarge 实例每小时费用由 34.61 美元涨至 39.80 美元，美国西部地区 P5e 实例价格从 43.26 美元/小时涨至 49.75 美元/小时。
三星电子	1 月 5 日	2026Q1 将 DRAM 价格较 2025Q4 提升 60%至 70%（包括向服务器、PC 及智能手机领域）。
	1 月 25 日	2026Q1 NAND 闪存供应价格上调 100%以上。
SK 海力士	1 月 5 日	2026Q1 将服务器 DRAM 价格较 2025Q4 提升 60%至 70%（向服务器、PC 及智能手机领域）。
	3 月 2 日	上调 2025Q2 DRAM 价格，DDR5 颗粒统一涨价 40%，部分产品涨幅高达 100%。
谷歌云	1 月 27 日	对 Google Cloud、CDN Interconnect、Peering 以及 AI 计算基础设施服务进行价格调整；北美地区数据传输价格从原来的 0.04 美元/GiB 上涨至 0.08 美元/GiB，涨幅达 100%；欧洲地区从 0.05 美元/GiB 上涨至 0.08 美元/GiB，涨幅 60%；亚洲地区从 0.06 美元/GiB 上涨至 0.085 美元/GiB，涨幅约 42%。
网宿科技	2 月 4 日	CDN 产品标准服务组流量上调 35%，CDN 产品快速回源通道流量上调 40%，对象存储产品存储空间上调 40%。
优刻得	2 月 11 日	对续签及新签用户的全线产品与服务进行价格上浮调整。
铠侠	2 月 14 日	2026Q1 NAND 闪存供应价格翻倍。
Hetzner	2 月 23 日	调高全线产品及服务报价，包括云服务、专用服务器、存储及负载均衡器等。云服务涨价最为显著，德国及芬兰的云服务价格涨幅在 30%到 38%之间，美国地区的 CCX 专用 vCPU 云服务器价格普遍涨幅在 30%水平。
腾讯云	3 月 11 日	GLM 5、MiniMax 2.5、Kimi 2.5 结束免费公测，转为按量计费。混元系列模型 Tencent HY2.0 Instruct 与 Tencent HY2.0 Think 的价格上调，其中，HY2.0 Instruct 输入价格从每千 tokens 0.0008 元调整为 0.004505 元，输出价格从 0.002 元调整为 0.01113 元；HY2.0 Think 输入价格从 0.001 元调整为 0.0053 元，输出价格从 0.004 元调整为 0.0212 元。
阿里云	3 月 18 日	平头哥真武 810E 等算力卡产品上涨 5%-34%，文件存储产品 CPFS（智算版）上涨 30%。
	4 月 13 日	取消 DataWorks 标准版、专业版用户每日调用 API 的数量限制，调用 API 的免费额度调整分别为 10 万次/月和 50 万次/月，超出部分采用 OpenAPI 按量付费的方式。
	4 月 15 日	上调部分 MU 模型单元的服务价格，涨幅达 2%-5%。对 DDoS 原生防护 2.0（包年包月）、DDoS 高防（中国内地）以及 DDoS 高防（非中国内地）商品的弹性 95 功能进行价格调整，DDoS 高防（中国内地）的弹性 95 由 100

		元/Mbps/月调整为 150 元/Mbps/月。
百度智能云	3 月 18 日	AI 算力相关产品服务上调约 5%-30%，并行文件存储等上调约 30%。
英特尔	3 月 19 日	计划 3 月起将全线 CPU 产品价格统一上调约 10%。
AMD	3 月 25 日	计划 4 月起将全系 CPU 价格上调约 10-15%。

来源：InfoQ，腾讯新闻，21 经济网，网宿科技官网，优刻得官网，Hetzner 官网，36 氪，腾讯云官网，阿里云官网，百度智能云官网等，国金证券研究所

英伟达全系列 GPU 租赁价格上涨，算力短缺程度持续升温。据纽约数据提供商 ORNN，近几个月来，英伟达全系列 GPU 在云端数据中心的现货租赁价格均大幅上涨，Blackwell 系列芯片单小时租金已达 4.08 美元，较两个月前的 2.75 美元上涨 48%。英伟达 B200 本周租赁价格已达到 4.95 美元/小时，三月初其价格仅为 2.31 美元/小时，6 周内价格高涨 114%。算力供不应求，旧款芯片的租赁价格也在走高，据 SemiAnalysis，英伟达 H100 的一年期租赁合约价格已由 2025 年 10 月的低点约 1.70 美元/小时/GPU，上涨至 2026 年 3 月的 2.35 美元/小时/GPU，涨幅近 40%。同时，该机构调研显示，受访 GPU 供应商中半数表示英伟达 H 系列芯片无货，2026 年 8-9 月前上线的新一代 Blackwell 系列产能已被全部预订。芯片价格持续走强、交付周期拉长，算力供给瓶颈仍未缓解。

图表11：2026 年 3 月至 4 月，英伟达 B200 租赁价格六周内实现翻倍



来源：Tomasz Tunguz 博客，国金证券研究所

三、国内算力黄金年代开启，供给紧缺下好用的算力本身就是最核心的资产

3.1 国产芯片：大厂加速适配，需求爆炸+供给改善

昇腾、寒武纪 Day 0 首发适配 DeepSeek-V4。DeepSeek-V4 发布当日，寒武纪已基于 vLLM 推理框架完成对 DeepSeek-V4 Flash&Pro 双版本的 Day 0 适配；此前，寒武纪已对 DeepSeek 系列模型开展深入的软硬件协同性能优化，达成业界领先的算力利用率水平。寒武纪在推理框架与硬件特性方面进行深度优化，软硬一体全面释放 DeepSeek-V4 推理潜能，同时降低大模型部署中的算力成本。华为昇腾也于在模型开源当日即完成快速适配，昇腾官方宣布通过与 DeepSeek 双方芯模技术紧密协同，实现昇腾超节点全系列产品支持 DeepSeek V4 系列模型；昇腾 950 通过融合 kernel 和多流并行技术降低 Attention 计算和访存开销，大幅提升推理性能，结合多种量化算法，实现了高吞吐、低时延的 DeepSeek V4 模型推理部署；昇腾 A3 超节点系列产品也全面适配，同时为便于用户快速微调，提供了基于昇腾 A3 集群的训练参考实现。

图表12：华为昇腾 Day 0 支持 DeepSeek-V4

来源：芯东西微信公众号，国金证券研究所

10 家国产芯片厂商路线完成对 DeepSeek-V4 的适配，国产化替代迎来加速放量机遇。截至 2026 年 4 月 30 日，寒武纪、华为昇腾、海光信息、摩尔线程、沐曦股份、百度昆仑芯、阿里平头哥真武、天数智芯、清微智能、曦望均已适配 DeepSeek-V4。据 Bernstein，2026 年中国 AI 芯片市场规模将达 2400 亿美元，国产份额将由 2025 年的 58%进一步上升至 79%，国产替代空间巨大。我们认为，国产芯片技术进步迅速，从“替代可用”向“自主好用”加速升级，DeepSeek-V4 广泛适配国产算力平台，为国内算力厂商直接带来规模化放量的巨大机遇。

图表13：10 家国产芯片厂完成对 DeepSeek-V4 的适配

公司	适配情况
寒武纪	基于 vLLM 推理框架完成 DeepSeek-V4 适配，适配代码开源
华为	昇腾 A2、A3 及 950 全系列产品适配 DeepSeekV4
海光信息	海光 DCU 完成对 DeepSeek-V4 适配
摩尔线程	旗舰级 AI 训推一体全功能 GPU MTTS5000 适配 DeepSeek-V4-Flash
沐曦股份	沐曦股份 AI 芯片完成 DeepSeek-V4-Flash 全量适配与推理部署
昆仑芯	昆仑芯 AI 芯片上完成 DeepSeek-V4-Flash 全量适配与推理部署
平头哥	平头哥真武芯片完成 DeepSeek-V4-Flash 全量适配与推理部署
天数智芯	天数智芯 AI 芯片完成 DeepSeek-V4-Flash 全量适配与推理部署
清微智能	RPU 可重构计算芯片通过 FlagOS 生态完成 DeepSeek-V4-Flash 适配
曦望	曦望 S2 芯片完成 DeepSeek-V4-Flash 适配

来源：芯东西微信公众号，智源 FlagOpen 微信公众号，国金证券研究所

寒武纪 2026 年一季报强力印证算力需求爆炸+供给改善。寒武纪一季报预付及合同负债双高增：1）预付款项，由年初 7.45 亿元激增至 18.97 亿元，环比大幅提升，公司主动锁定上游芯片、供应链资源，供给端瓶颈持续缓解，保障后续产能与交付；2）合同负债，由年初 610.62 万元飙升至 3.96 亿元，订单与预收款项暴增，下游 AI 算力需求呈爆炸式增长，高景气度前置确认。

3.2 国产机柜：超节点出货带来 ODM 组装厂商格局/利润率双提升

国内厂商超节点产品密集发布，国产机柜时间到来。华为（Atlas 950）、中科曙光（ScaleX640/40）、阿里（磐久 128）等厂商纷纷推出超节点产品，并在算力密度、互联带宽、计算架构等方面实现突破，满足大模型训练与推理对高带宽、低时延的刚性需求。

- 华为：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14：Atlas 950 超节点

来源：2025 华为全联接大会，国金证券研究所

图表15：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16：2025 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17：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

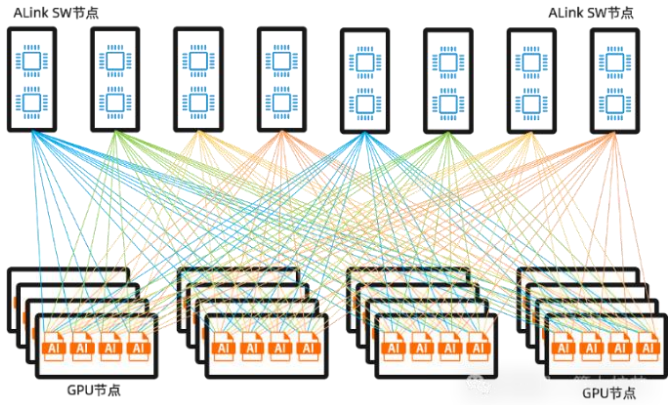
来源：中科曙光微信公众号，国金证券研究所

- 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18：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19：磐久超节点 ScaleUp 互连拓扑图

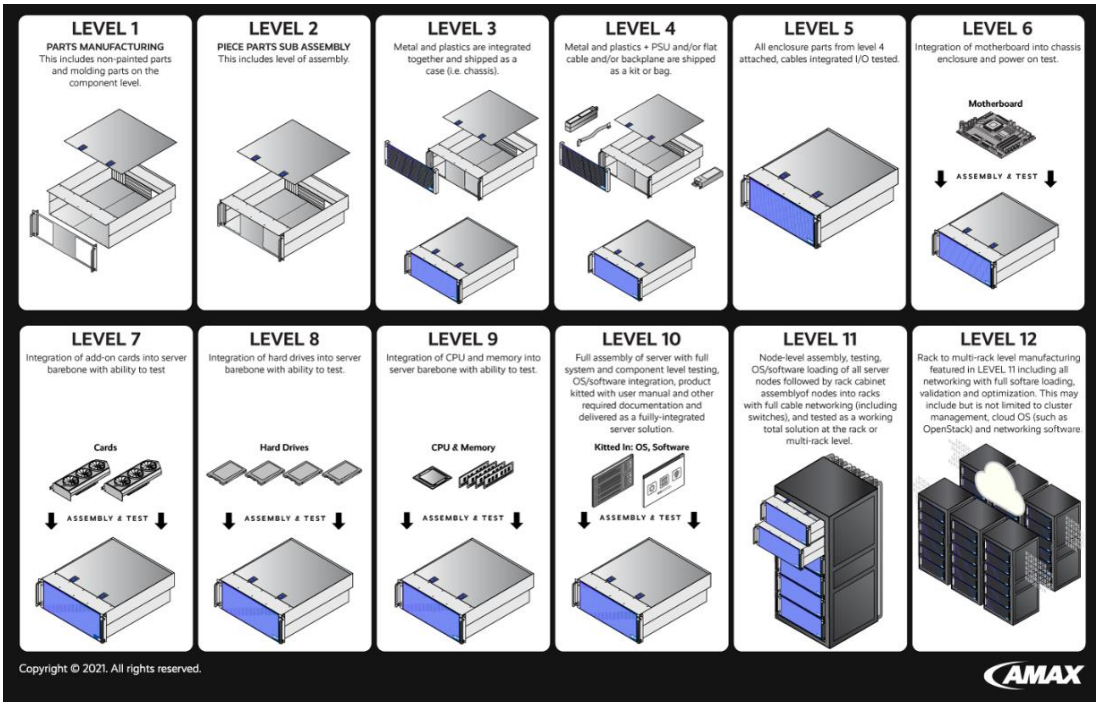


来源：算力核芯微信公众号，国金证券研究所

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。

图表20：服务器代工生产主要有 12 个层级



来源：AMAX，国金证券研究所

3.3 算租/云/IDC：模型调用带来云计算需求膨胀

云：模型调用带来云计算需求膨胀，Token 调用量呈指数级爆发，驱动云资源进入涨价通道。Token 需求的急剧膨胀，直接推高了云端的算力消耗，导致云厂资源趋于紧张并开启多轮涨价潮。2026 年 1 月，海外云厂率先出现涨价信号，亚马逊 AWS 于 1 月 4 日将其 EC2 机器学习容量块服务价格上调约 15%，谷歌云随后于 1 月 27 日宣布上调全球数据传输服务价格。2026 年 3 月起，阿里云接连发布涨价公告，3 月 18 日 AI 算力与存储等产品最高涨价 34%，、平头哥真武 810E 等算力卡产品上涨 5%-34%；4 月 13 日 DataWorks 标准版、专业版免费额度调整为 10 万次/月和 50 万次/月，超出部分按量付费；4 月 15 日部分 MU 模型单元的服务价格涨幅达 2%-5%、DDoS 防护产品价格同步上涨。国内其他云厂，如百度云、网宿科技、优刻得等，均对其 AI 算力相关产品及服务进行提价，全面步入通胀通道。

图表21：阿里云上调部分 DDoS 防护产品价格

商品	模块	调价前	调价后
DDoS原生防护2.0（包年包月）	弹性带宽	月95：82元/Mbps/月 日95：12元/Mbps/日	月95：98.5元/Mbps/月 日95：6元/Mbps/日
DDoS高防（中国内地）	弹性带宽	月95：100元/Mbps/月 日95：6元/Mbps/日	月95：150元/Mbps/月 日95：8元/Mbps/日
DDoS高防（非中国内地）	保险防护-弹性带宽	月95：105元/Mbps/月 日95：6.3元/Mbps/日	月95：150元/Mbps/月 日95：8元/Mbps/日
	加速线路/安全加速线路-弹性带宽	月95：1000元/Mbps/月 日95：60元/Mbps/日	月95：1500元/Mbps/月 日95：80元/Mbps/日
	弹性QPS	月95：12元/Mbps/月	月95：15元/Mbps/月

来源：阿里云官网，国金证券研究所

图表22：阿里云百炼平台部分 MU 价格上涨 2%-5%

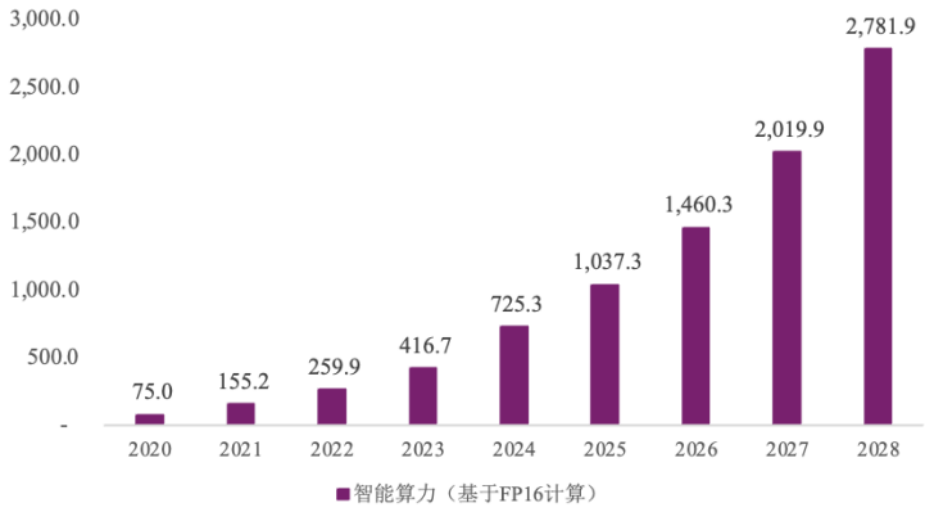
计费模式	产品型号	上架规格	售卖单位	原单价	调整后单价	调整幅度
包年包月	III 型模型单元	MU3	元/个/月	63,000	65,969	↑约5%
	IV 型模型单元	MU4	元/个/月	19,000	19,914	↑约5%
	V 型模型单元	MU5	元/个/月	9,500	10,139	↑约7%
按量计费	III 型模型单元	MU3	元/个/小时	131	137	↑约5%
	IV 型模型单元	MU4	元/个/小时	40	41	↑约2%
	V 型模型单元	MU5	元/个/小时	20	21	↑约5%

来源：阿里云官网，国金证券研究所

算力租赁：需求爆炸与业绩加速双重印证，稀缺算力资产的商业化（ROI）正逐步兑现。在算力供给紧缺的当下，好用的算力本身就是当前最核心、最稀缺的资产。从全球视角看，北美头部算租厂商的 ROI 已实质性兑现；国内来看，认真做事的智算公司已经实现"有卡有利润"，业绩进入加速兑现期。1）北美头部算力租赁厂商 ROI 逐步兑现、关键融资顺利：Oracle、CoreWeave 等 25Q4 起 GPU 云业务收入爆发；CoreWeave 完成全球首笔投资级 GPU 基础设施融资，Oracle 计划大规模融资近 500 亿美元。2）芯片迭代周期长，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：英伟达 H100 的一年期租赁合约价格已由 2025 年 10 月的低点约 1.70 美元/小时/GPU，上涨至 2026 年 3 月的 2.35 美元/小时/GPU，涨幅近 40%。3）国内相关公司已实现"有卡、有利润"：利通电子 2025 年全年实现利润 3 亿、26Q1 即实现利润 2.7 亿；协创数据 2026Q1 实现利润 7 亿（同比增长 300%+）；盈峰环境 2026Q1 实现利润 2.1 亿；智微智能 26Q1 实现利润 1.1 亿（同比+160%）。

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。

图表23：2020-2028 年中国智能算力规模及预测（EFLOPS，基于 FP16 计算）



来源：沐曦招股说明书，国金证券研究所

四、相关标的

国内算力：寒武纪、海光信息、东阳光、利通电子、协创数据、杰华特、利扬芯片、华勤技术、浪潮信息、禾盛新材、中国长城、罗曼股份、网宿科技、盈峰环境、芯原股份、华丰科技、晶科科技、亿田智能、豫能控股、星环科技、鸿日达、盛视科技、首都在线、神州数码、百度集团、中芯国际、华虹半导体、中科曙光、润泽科技、大位科技、润建股份、奥飞数据、云赛智联、瑞晟智能、科华数据、潍柴重机、金山云、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