下载日志已记录，仅供内部参考

计算机

主权 AI 时代来临：国产模型追赶至 SOTA，算力突围与生态重构

模型维度：后训练时代模式切换，模型竞争从“规模”走向“效率”。大模型能力演进正在转向预训练、后训练和推理时计算共同驱动。三类 Scaling 共同作用，使模型能力、用户需求和 Token 消耗形成正反馈。预训练阶段的壁垒是高端 GPU、集群互联和优质语料，国内厂商受算力约束较大。

进入后训练阶段后，单位算力的优化空间显著扩大，国产模型追赶海外前沿水平的路径更加清晰；GLM-5.3 在 GLM-5.2 同架构下通过重新后训练，指标显著提升，验证了后训练质量对模型能力释放的重要性。

国产模型加速迭代，开源生态强化全球可用性；定价分化，优质模型进入能力与成本再平衡阶段，国产模型性价比突出。Kimi K3 开放权重后，进一步打破“前沿能力只能由高价闭源模型提供”的认知。

算力维度：Token 经济供给不足，超节点新技术及生态构建全面突围。AI 服务器成本集中于芯片，供给约束直接推升部署成本。**模型突破后供需不匹配进一步显性化，推理工作负载持续扩大，**高端 GPU 受出口管制影响，B300、H200 等资源稀缺推动采购与租赁价格上行，万亿参数模型对显存容量、互联效率和调度能力提出更高要求。

“六张网”建设持续推进，政策端逐步明确算力网发展方向。2026 年 8 月国务院常务会议及国家发改委进一步细化部署，建立算力网、新型电网、新一代通信网“2+3+N”协调机制。

超节点成为重要方向，P/D 分离通过拆分环节提升资源利用效率，华为向万卡 SuperPoD 大规模集群演进。

国产 AI 加速卡份额提升，软硬件全栈自主生态加速闭环。2025 年国内 AI 加速卡出货量约 400 万张，国产芯片厂商合计出货约 165 万张，市占率约 41%。同时，DeepSeek V4、GLM-5.2 等国产大模型与多家国产芯片厂商深度适配，Day0 首发适配已成常态。

芯片供应链难题突破：制造、封装与存储国产化多线推进。

先进制程演进放缓背景下，系统级优化成为半导体性能提升的重要方向。随着传统摩尔定律驱动下的几何缩微逐渐面临物理极限，华为提出的“韬定律”以“时间缩微”替代“几何缩微”，强调器件、电路、芯片和系统多层级协同优化。

国产晶圆厂产能扩张，制造底座价值提升。26Q2 中芯国际月产能达 109.7 万片八英寸等值晶圆，产能利用率 93.7%。26Q2 中芯国际 12 英寸晶圆收入占比达 78.2%；华虹 12 英寸收入占比提升至 62.1%。

国产存储产线利用率提升，供给安全价值凸显。长鑫存储在 DRAM 领域产能利用率持续提升，2023/24/25 年分别达到 87.06%/92.46%/95.73%，并在合肥、北京两地拥有 3 座 12 英寸 DRAM 晶圆厂。长江存储在 NAND Flash 领域产能利用率同样维持高位，2023/24/25/26Q1 分别为 94.42%/95.43%/97.91%/98.02。

推荐配置主线：

1）GPU 与超节点：寒武纪、海光信息、中科曙光、盛科通信；

2）存储：长鑫科技、国科微；

3）设备、材料与 Fab：中科飞测、炬光科技、长川科技、中船特气、蜀道装备、华虹宏力、中芯国际；

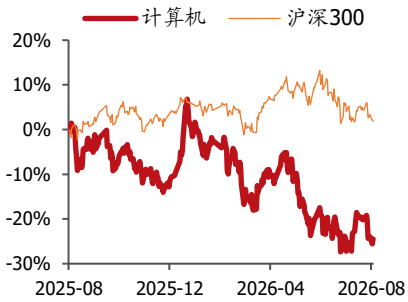
4）AIDC：东阳光、协创数据、集智股份；

5）模型与垂类应用：智谱、海致科技、金山办公、金蝶国际、中控技术、晶泰控股、深信服、容知日新。

风险提示：宏观经济不及预期；AI 应用需求释放不及预期；国产替代进程不及预期。

增持（维持）

行业走势



作者

分析师 杨然

执业证书编号：S0680518050002

邮箱：yangran@gszq.com

相关研究

1、《计算机：AI4S：科研第五范式，万亿市场拐点已现》2026-08-15

2、《计算机：上海发布五年规划：大规模集群建设扩容，国芯国模落地加速》 2026-08-13

3、《计算机：FDE：软件服务的范式革命》 2026-08-08

下载日志已记录，仅供内部参考

内容目录

1、模型：后训练时代新竞争范式，国产模型追至 SOTA 3
1.1 范式切换：后训练时代成为分水岭，国产模型逆势突围 3
1.2 开源生态扩张推动主权 AI，国产模型成为推理侧重要供给 4
2、算力基础设施：供需错配加剧，国产卡技术创新加速突围 6
2.1 高端算力供需错配，采购与租赁价格同步上行 6
2.2 超节点架构加速落地，系统级优化应运而生 8
2.3 国产 AI 芯片适配加快，自主可控趋势明确 11
3、芯片供应链：制造、封装与存储环节加速补短板 12
3.1 韬（T）定律：提供系统级追赶路径，弱化单一制程依赖 12
3.2 国产晶圆厂产能扩张，制造底座价值提升 14
3.3 存储：国产存储产线利用率提升，供给安全价值凸显 14
4、投资建议 16
4.1 GPU 与超节点：关注国产算力替代和系统级架构升级 16
4.2 存储：关注 AI 需求驱动下的 DRAM 与存储控制国产化 16
4.3 半导体设备、材料与 Fab：关注国产算力供应链补短板 16
4.4 AIDC：关注算力网建设和资源运营能力 17
4.5 模型与垂类应用：关注国产模型能力提升与行业 AI 落地 18
风险提示 19

图表目录

图表 1：三大 Scaling 路径推动模型能力与商业化进入正循环 3
图表 2：国产模型在 code 榜单中表现提升 4
图表 3：全球数据中心 AI 推理侧负载需求增长 4
图表 4：OpenRouter 周度 Token 处理量持续上行 5
图表 5：OpenRouter 月度模型调用量排名中国产模型占位提升 5
图表 6：主流模型定价 6
图表 7：算力芯片在服务器成本中占比最高 6
图表 8：英伟达和昇腾芯片指标对比 7
图表 9：国内 GPU 服务器租赁价格 7
图表 10：“六张网”政策提出时间线 8
图表 11：2021-2028 年中国智能算力规模及预测（EFLOPS，基于 FP16 计算） 8
图表 12：国内算力租赁代表企业融资进展梳理 8
图表 13：PD 分离架构示意图 9
图表 14：华为走向万卡级超节点和 52 万+卡集群 9
图表 15：华为 CM384 超节点集成计算、内存访问和通信平面 10
图表 16：CM384 为 P/D 分离和 MoE 推理做的系统优化 10
图表 17：通信带宽增长较慢，系统需要协同 11
图表 18：2025 年我国 AI 芯片市场竞争格局 11
图表 19：国产大模型加速适配国产芯片，软硬协同生态持续完善 12
图表 20：韬缩放 13
图表 21：逻辑折叠 13
图表 22：韬(T)定律和摩尔定律对比 13
图表 23：中芯国际月度产能（折合 8 英寸，万片/月）及稼动率 14
图表 24：长鑫存储募投项目 15
图表 25：全球 NAND Flash/DRAM 市场规模季度变化（单位：亿美元） 15
图表 26：2025 年 DRAM 市占率 15

下载日志已记录，仅供内部参考

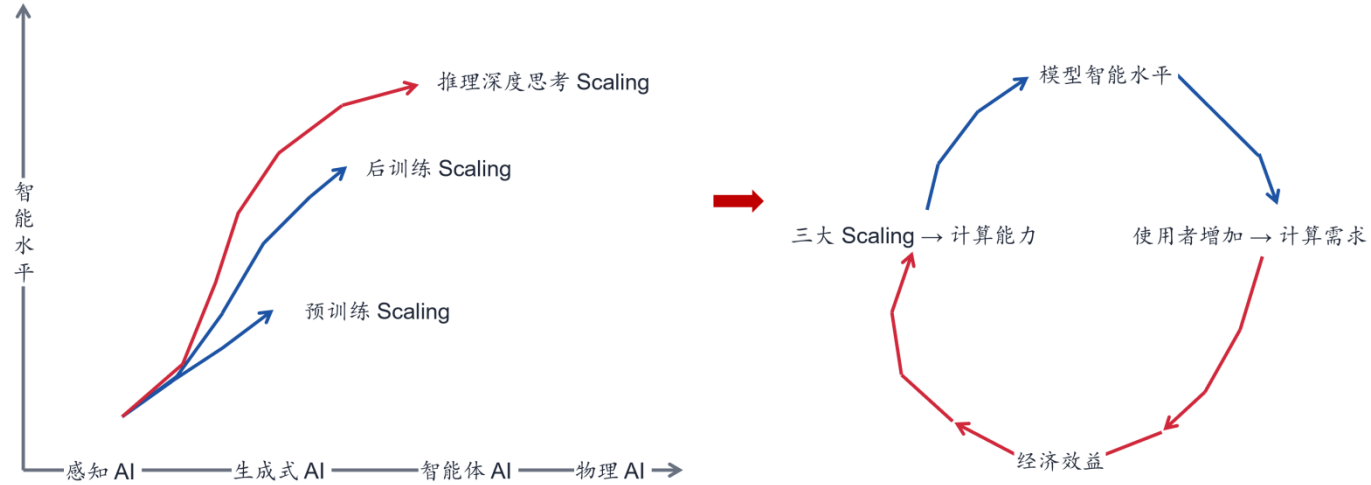
1、模型：后训练时代新竞争范式，国产模型追至 SOTA

1.1 范式切换：后训练时代成为分水岭，国产模型逆势突围

后训练时代模式切换，模型竞争从“规模”走向“效率”。大模型能力演进正在由单一预训练扩展，转向预训练、后训练和推理时计算共同驱动。模型在预训练阶段学习其固有知识，在后训练中通过微调与增强学习如何进行思考，在推理时计算过程中通过生成更多 token 实现深入推理。三类 Scaling 共同作用，使模型能力、用户需求和 Token 消耗形成正反馈，推动行业竞争从参数规模转向真实任务效率。预训练阶段的核心壁垒是高端 GPU 规模、集群互联和优质语料，国内厂商受算力约束较大；进入后训练阶段后，单位算力的优化空间显著扩大，国产模型追赶海外前沿水平的路径更加清晰。

后训练价值凸显，Agent 能力不再单纯依赖参数规模。GLM-5.3 在 GLM-5.2 同架构下通过重新后训练，在 CyberGym 漏洞推理评测中，GLM-5.3 模型从白盒源代码出发，通过触发程序故障来识别和验证漏洞。GLM-5.3 得分为 84.5%，高于 GLM-5.2 的 77.2%，也略高于 Mythos 5 的 83.8%和 GPT-5.6 Sol 的 83.6%。该案例说明，在 Agent 和代码等可验证任务中，后训练质量、奖励设计、任务数据和工程调优已成为决定模型效果的重要变量。对于国产模型而言，后训练阶段对高端 GPU 的边际依赖相对低于预训练阶段，因此更有利于在有限算力条件下实现能力追赶。

图表1：三大 Scaling 路径推动模型能力与商业化进入正循环



资料来源：英伟达官网，国盛证券研究所

代码和 Agent 任务具备结果可验证、反馈周期短、工程优化空间大的特点，因此更能体现后训练和推理工程能力。国产模型在相关榜单中持续取得前列位置，说明国内厂商已在部分高价值场景接近全球前沿水平。随着开源生态、低成本 API 和真实调用数据持续积累，国产模型有望从能力追赶进一步走向生态扩张。

国产模型加速迭代，开源生态强化全球可用性。国产模型阵营正在加速扩展。Kimi K3 具有强大的长期编码性能，智谱 GLM-5.3 面向 GLM Coding Plan 全量用户开放，承接企业级代码场景需求；DeepSeek-V4-Flash 正式版 API 上线并适配 Responses API；字节 Seedance 2.5 强化视频生成质量和多轮延长能力。整体看，国产模型不再仅停留于文本对话或单点能力展示，而是逐步覆盖代码、视频、Agent、多模态和企业级 API 等场景。随着开源模型、低成本 API 和真实用户调用数据形成反馈闭环，国产模型正在从“可用”向“好用”演进。

下载日志已记录，仅供内部参考

图表2：国产模型在 code 榜单中表现提升

Code Arena | WebDev Overall

View overall rankings across AI models on front-end web development tasks, including agentic coding workflows that require multi-step reasoning and tool use.

Aug 22, 2026 603,789 votes 118 models

Hide Filters Models Labs

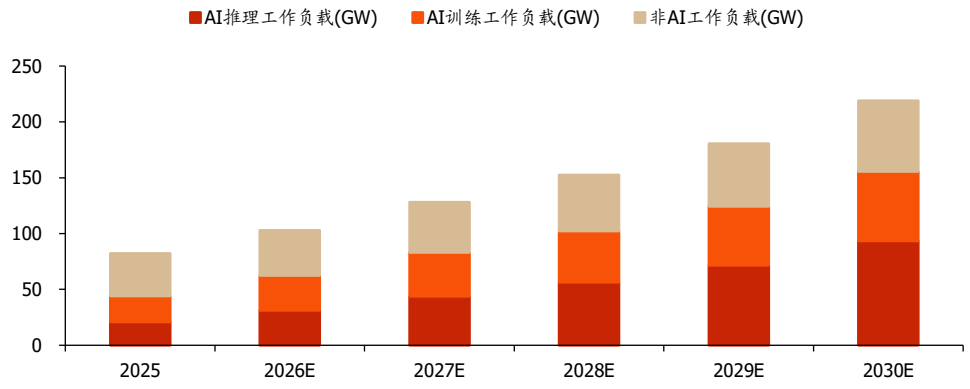
Rank	Rank Spread	Model	Score	Votes	Price $/M	Context
1	1 → 2	claude-opus-5-max Anthropic · Proprietary	1691 +9/-9	8,116	$5 / $25	1M
2	1 → 4	kimi-k3-max Moonshot · Kimi K3 license	1674 +11/-11	4,546	$3 / $15	1M
3	2 → 4	qwen3.8-max Alibaba · Proprietary	1669 +13/-13 Preliminary	3,212	$2 / $6	1M
4	2 → 4	claude-opus-5-high Anthropic · Proprietary	1663 +8/-8	8,659	$5 / $25	1M
5	5 → 8	grok-4.6-high SpaceXAI · Proprietary	1629 +17/-17 Preliminary	1,523	$2 / $6	500K
6	5 → 7	claude-fable-5 Anthropic · Proprietary	1626 +8/-8	8,382	$10 / $50	1M
7	5 → 8	gpt-5.6-sol-xhigh (codex-harness) OpenAI · Proprietary	1619 +8/-8	9,499	$4 / $20	1.1M
8	6 → 13	glm-5.3-max Z.ai · MIT	1599 +15/-15	1,916	$1.40 / $4.40	1M
9	8 → 13	qwen3.8-27b Alibaba · Apache 2.0	1595 +13/-13	2,464	$0.40 / $3	N/A
10	8 → 13	gemini-3.7-flash-high Google · Proprietary	1587 +13/-13 Preliminary	2,552	$0.75 / $3.57	1M

资料来源：Arena，国盛证券研究所
注：数据截至 2026 年 8 月 22 日

1.2 开源生态扩张推动主权 AI，国产模型成为推理侧重要供给

推理负载占比提升，Token 消耗进入持续放量阶段。随着 AI 应用从训练转向推理、从对话转向 Agent，企业和国家对于模型可控性、成本可控性和部署灵活性的要求明显提升。开源模型能够提供更高透明度、更强可定制能力和更低边际调用成本，因此在 Agent、代码生成和企业私有化部署中具备重要价值。

图表3：全球数据中心 AI 推理侧负载需求增长

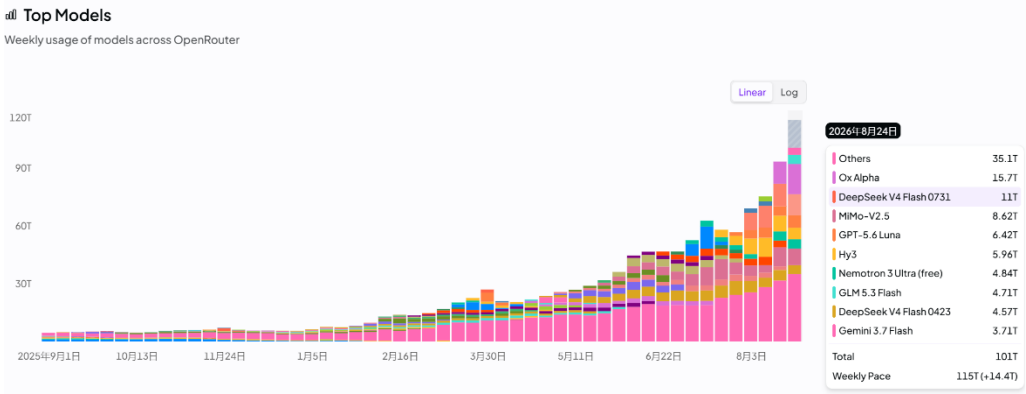


资料来源：McKinsey，国盛证券研究所

开发者调用需求扩张，OpenRouter 验证推理侧景气。Token 调用量是观察推理需求的重要指标。AI 工作重心从训练转向推理后，模型调用不再局限于简单问答，而是进入长程任务处理阶段。单次 Agent 任务往往包含多轮思考、工具调用和结果校验，其 Token 消耗显著高于传统对话。因而，Token 需求增长不仅代表用户数量增加，更反映 AI 应用正在从轻量交互走向高频生产流程。

下载日志已记录，仅供内部参考

图表4：OpenRouter 周度 Token 处理量持续上行



资料来源：OpenRouter，国盛证券研究所
注：数据截至 2026 年 8 月 29 日

国产模型全球调用占位提升，性价比优势逐步兑现。开源模型与闭源模型正形成较稳定的双轨结构：闭源模型继续定义性能和可靠性上限，开源模型则凭借成本、透明度和部署灵活性承接大规模推理负载。国产模型在开发者和中小企业场景中的调用份额提升，说明其已具备承担代码生成、自动化 Agent 和业务流程处理等真实工作负载的能力。

图表5：OpenRouter 月度模型调用量排名中国产模型占位提升

1.	Ox Alpha by stealth	20.7T tokens ↑216%	11.	Hy4 preview by tencent	1.9T tokens new
2.	DeepSeek V4 Flash 0731 by deepseek	12.3T tokens ↑5%	12.	DeepSeek V4 Pro 0423 by deepseek	1.84T tokens ↓2%
3.	MiMo-V2.5 by xiaomi	9.98T tokens ↑9%	13.	MiniMax M3 (free) by minimax	1.82T tokens new
4.	GPT-5.6 Luna by openai	6.75T tokens ↑31%	14.	GPT-5.6 Sol by openai	1.72T tokens ↑40%
5.	Hy3 by tencent	6.62T tokens ↓23%	15.	Kimi K3 by moonshotai	1.5T tokens ↑15%
6.	Nemotron 3 Ultra (free) by nvidia	5.62T tokens ↑27%	16.	Claude Opus 5 by anthropic	1.5T tokens ↓28%
7.	DeepSeek V4 Flash 0423 by deepseek	5.4T tokens ↑2%	17.	MiniMax M3 by minimax	1.49T tokens ↓14%
8.	GLM 5.3 Flash by z-ai	4.62T tokens new	18.	Laguna S 2.1 (free) by poolside	1.36T tokens ↓17%
9.	Gemini 3.7 Flash by google	4.05T tokens ↑156%	19.	GLM 5.3 by z-ai	1.24T tokens ↑222%
10.	GLM 5.2 by z-ai	3.11T tokens ↓13%	20.	Claude Sonnet 5 by anthropic	1.2T tokens ↑12%

资料来源：OpenRouter，国盛证券研究所
注：数据截至 2026 年 8 月 29 日

定价分化，优质模型进入能力与成本再平衡阶段，国产模型性价比突出。Kimi K3 开放权重后，进一步打破“前沿能力只能由高价闭源模型提供”的认知；OpenAI 下调部分前沿模型价格，有助于推动企业级 AI 应用从试用走向规模化部署；DeepSeek 近期上调 API 定价，在模型能力持续提升的背景下，进一步验证优质国产模型的商业化定价能力，同时其相较海外前沿模型仍具显著性价比优势。

下载日志已记录，仅供内部参考

图表6：主流模型定价

模型	缓存输入($/MTok)	普通输入($/MTok)	输出($/MTok)	100K 输入+25K 输出成本($)
Claude Fable 5	1.00	10.00	50.00	2.25
GPT-5.6 Sol	0.50	5.00	30.00	1.25
Claude Opus 4.8	0.50	5.00 00		1.13
Kimi K3	**0.30**	**2.99**	**14.93**	**0.67**
Gemini 3 Pro (≤200K tokens)	0.20	2.00	12.00	0.50
GPT-5.6 Terra	0.20	2.00	12.00	0.50
Claude Sonnet 5 标准价	0.20	2.00	10.00	0.45
Qwen3.7-Max	**0.33**	**1.65**	**4.95**	**0.29**
GLM-5.2	0.26	1.40	4.40	0.25
DeepSeek V4-Pro 高峰	**0.04**	**1.32**	**3.96**	**0.23**
DeepSeek V4-Pro 空闲	**0.02**	**0.66**	**1.98**	**0.12**

资料来源：各公司官网、Wind，国盛证券研究所
注：Kimi K3 模型价格按照 2026 年 8 月 20 日美元兑人民币汇率 6.7 进行换算。

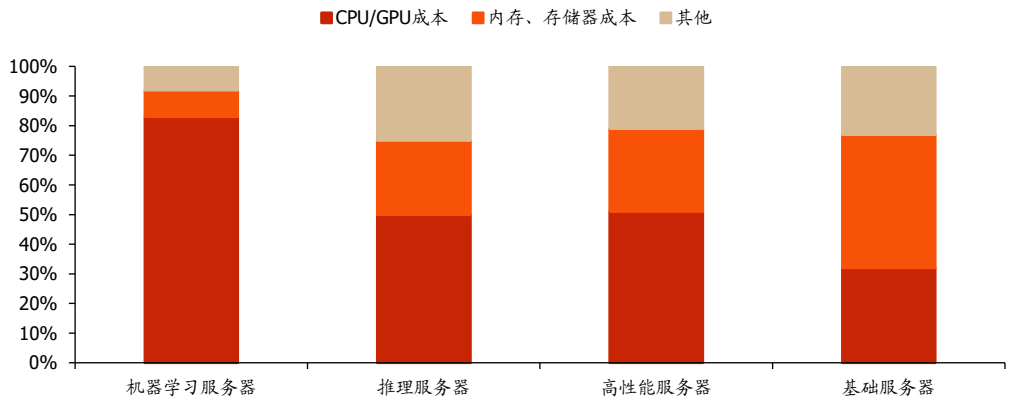
2、算力基础设施：供需错配加剧，国产卡技术创新加速突围

2.1 高端算力供需错配，采购与租赁价格同步上行

随着 AI 模型应用从训练阶段逐步向推理阶段延伸，数据中心算力需求结构正在发生变化。相比模型训练，推理任务具有持续运行、高并发和低延迟等特点，对 GPU 算力、显存容量、内存带宽、网络互联及数据中心基础设施提出更高要求。随着 Agent 等复杂 AI 应用发展，数据中心建设将进一步向计算、存储、网络和系统调度协同的方向演进。

AI 服务器成本集中于芯片，供给约束直接推升部署成本。AI 服务器成本主要集中于 GPU 及相关加速芯片，高端芯片供给变化会直接影响整机采购价格和算力租赁成本。在海外高端 GPU 进入国内存在不确定性的背景下，芯片供给紧张容易进一步传导至云端 API 价格和企业部署成本。

图表7：算力芯片在服务器成本中占比最高



资料来源：甲子光年，国盛证券研究所

下载日志已记录，仅供内部参考

单卡性能仍有差距，国产芯片需依赖系统级追赶。从规格迭代看，国产 AI 芯片与海外前沿 GPU 在单卡峰值算力、显存容量和内存带宽方面仍存在差距，但国内产品演进节奏正在加快。短期内，国产芯片较难完全依靠单卡性能对标英伟达 B 系列和 Rubin 系列；中长期看，通过超节点架构、模型适配和系统调度优化，国产算力仍可在国内训练与推理场景中提升有效供给能力。

图表8：英伟达和昇腾芯片指标对比

时间	品牌	芯片型号	FP8 算力	FP4 算力	FP16 算力	内存容量	内存带宽
2025 Q1	Ascend	910C	—	—	800 TFLOPS	128GB	3.2TB/s
2024 Q4—2025	NVIDIA	B200	4.5 PFLOPS	9 PFLOPS	2.25 PFLOPS	192GB HBM3e	8TB/s
2026 Q1	Ascend	950PR	1 PFLOPS	2 PFLOPS	—	144GB	4TB/s
2025 H2—2026	NVIDIA	B300	7 PFLOPS	15 PFLOPS	3.5 PFLOPS	288GB HBM3e	8TB/s
2026 Q4	Ascend	950DT	1 PFLOPS	2 PFLOPS	—	128GB	1.6TB/s
2026 H2	NVIDIA	Rubin R100	16 PFLOPS	50 PFLOPS	8 PFLOPS	288GB HBM4	up to 22TB/s
2027 Q4	Ascend	960	2 PFLOPS	4 PFLOPS	—	288GB	9.6TB/s
2028 Q4	Ascend	970	4 PFLOPS	8 PFLOPS	—	288GB	14.4TB/s

资料来源：英伟达官网、芯智讯、Spheron，国盛证券研究所

高端 GPU 价格上行，算力租赁景气度持续验证。国内高端 GPU 稀缺已在价格端充分体现。从租赁价格看，国内高端 GPU 供给紧张已在价格端充分体现。据 Apetops，H200 8 卡服务器月租价格达到 13.50 万元，H100 8 卡服务器为 7.80 万元，明显高于 A800、L40S 等上一代或消费级替代资源；国产高端算力中，昇腾 910C 8 卡服务器月租价格达到 5.85 万元，也显著高于昇腾 910B2、W7900D 等产品。当前价格分化反映出 AI 推理需求持续扩张下，高性能 GPU、显存容量和生态适配能力仍是租赁市场定价核心。

图表9：国内 GPU 服务器租赁价格

名称	月租价格	名称	月租价格
H200 8 卡服务器	13.50 万元/月	昇腾 910C 8 卡服务器	5.85 万元/月
H100 8 卡服务器	7.80 万元/月	天数智芯天 150 BI-V150 64G8 卡服务器	3.00 万元/月
H20 141G 8 卡服务器	5.30 万元/月	昆仑 3 P800 8 卡服务器	2.90 万元/月
NVIDIA RTX PRO 6000 96G 8 卡服务器	5.25 万元/月	昇腾 910B2 8 卡服务器	1.68-1.85 万元/月
H20 96G 8 卡服务器	4.50 万元/月	5090 8 卡服务器	1.45-1.65 万元/月
A100 80G 8 卡服务器	3.70-4.25 万元/月	A100 40G 8 卡服务器	1.05 万元/月
RTX PRO 6000D 84G 8 卡服务器	3.80-4.00 万元/月	4090 8 卡服务器	0.85 万元/月
A800 80G 8 卡服务器	2.80-3.60 万元/月	W7900D 8 卡服务器	0.68 万元/月
L40S PCIe 8 卡服务器	2.30 万元/月	Atlas 300V Pro 6 卡服务器	0.65 万元/月

资料来源：Apetops，国盛证券研究所
注：数据截至 2026 年 8 月 20 日

“六张网”建设持续推进，政策端逐步明确算力网发展方向。2026 年 3 月，国家发改委在全国两会上提出推进“六张网”建设；4 月中央政治局会议进一步明确加强水网、新型电网、算力网、新一代通信网、城市地下管网、物流网等规划建设；7 月政治局会议要求扎实推进“六张网”规划建设，推动工作重心由规划向落地执行转变；8 月国家发改委进一步细化部署，研究建立算力网、新型电网、新一代通信网“2+3+N”协调机制。政策端持续强化算力基础设施建设，有望带动算力租赁等需求增长。

下载日志已记录，仅供内部参考

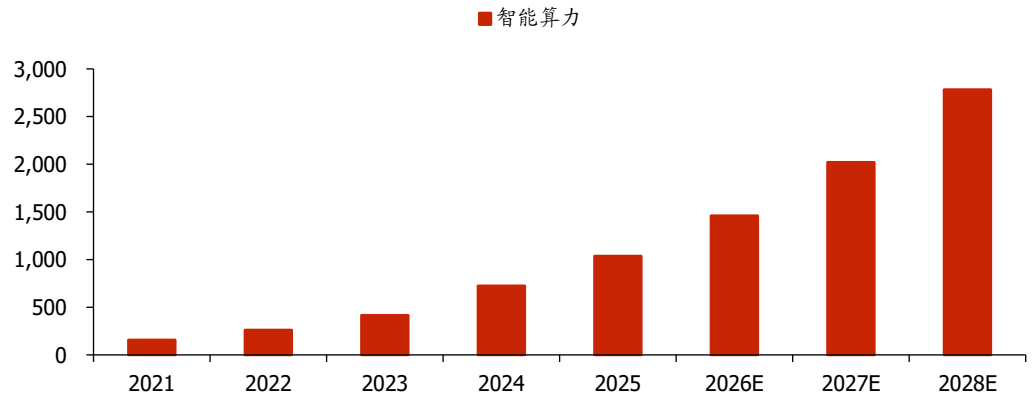
图表10：“六张网”政策提出时间线

时间	阶段	事件
2026.3	提出	国家发改委在全国两会经济主题记者会上提出推进“六张网”建设。
2026.4.28	中央定调	中共中央政治局会议明确提出“加强水网、新型电网、算力网、新一代通信网、城市地下管网、物流网等规划建设”。
2026.7.30	再部署	中共中央政治局会议要求“扎实推进‘六张网’规划建设”，工作重心从规划转向落地执行。
2026.8.19	发改委落实	国家发改委召开“六张网”重大项目协调调度机制会，研究建立算力网、新型电网、新一代通信网“2+3+N”协调机制（2 家电网+3 家电信运营商+N 家算力相关企业）。

资料来源：国家发改委、中国政府网，国盛证券研究所

融资扩张支撑供给释放，算力租赁进入资本密集阶段。算力租赁企业通过授信、定增、债券和上市融资等方式加快资源储备，反映行业已进入资本密集型扩张阶段。相关公司融资用途多指向智能算力集群建设、服务器采购和算力运营项目，说明下游需求景气正在带动供给侧扩产。后续具备融资能力、运维能力和客户资源的企业更有望在算力租赁市场中获得份额。

图表11：2021-2028 年中国智能算力规模及预测（EFLOPS，基于 FP16 计算）



资料来源：IDC，国盛证券研究所

图表12：国内算力租赁代表企业融资进展梳理

公司	融资进展
协创数据	2025 年度累计申请授信额度不超过 515 亿元，2026 年已累计申请授信额度不超过 500 亿元。
宏景科技	2025 年度累计申请综合授信额度不超过 200 亿元，2026 年已累计拟申请综合授信额度不超过 600 亿元。拟定增 12.90 亿元，其中 9.90 亿元拟用于智能算力集群建设及运营项目。
利通电子	拟申请注册发行不超过 5 亿元（含）超短期融资券及不超过 5 亿元（含）中期票据。已成功发行 2 亿元科技创新债，募集资金专项用于算力设备及备件采购。
盈峰环境	拟申请注册发行不超过 6 亿元（含）超短期融资券及不超过 24 亿元（含）中期票据。

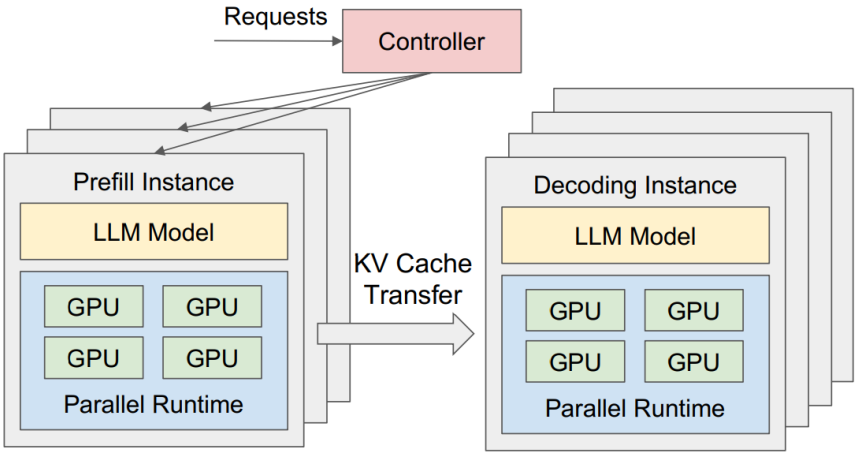
资料来源：各公司公告、协创数据投资者问答，国盛证券研究所

2.2 超节点架构加速落地，系统级优化应运而生

P/D 分离适配推理负载，资源调度效率成为关键。大语言模型推理过程可拆分为 Prefill 和 Decode 两个阶段，其中 Prefill 阶段并行处理输入 token 并生成 KV Cache，更依赖 GPU 计算能力；Decode 阶段采用自回归方式逐 token 生成，需要持续读取 KV Cache，性能更受显存带宽和内存访问效率约束。P/D 分离通过将不同阶段部署在更匹配的资源池中，有助于降低推理成本、提升吞吐效率，并成为大规模推理系统的重要演进方向。

下载日志已记录，仅供内部参考

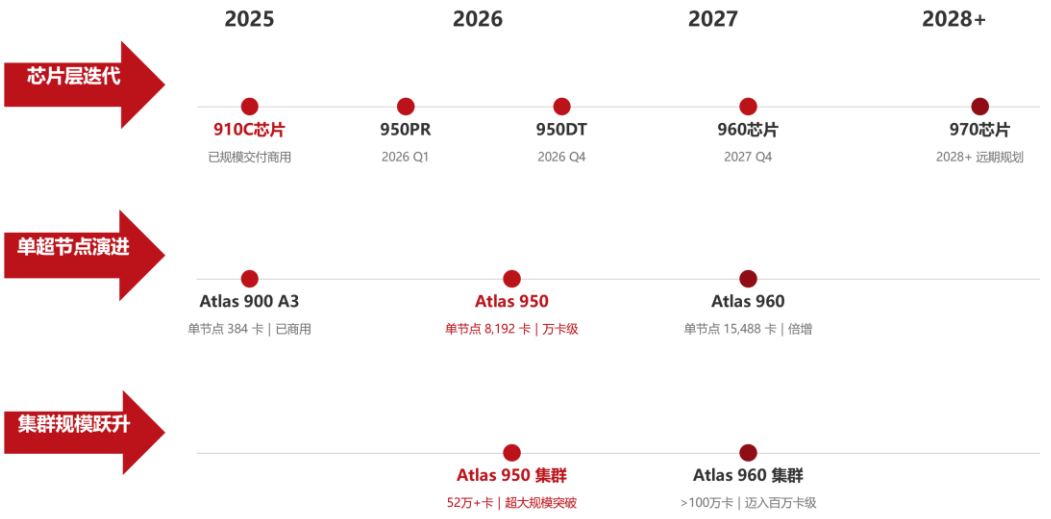
图表13：PD 分离架构示意图



资料来源：Semianalysis，国盛证券研究所

从八卡服务器到万卡集群，超节点打开 Scale-up 空间。华为 AI 集群从单机 8 卡逐步扩展至 384 卡超节点、8192 卡 SuperPoD，并规划迈向更大规模集群，体现国内算力建设从单机服务器向系统级集群演进。由于国内单芯片性能受制于先进制程和供应链约束，通过扩大 Scale-up 域提升有效算力，是更具现实可行性的路径。超节点的核心价值在于系统吞吐和资源调度，而非单纯堆叠芯片数量。

图表14：华为走向万卡级超节点和 52 万+卡集群

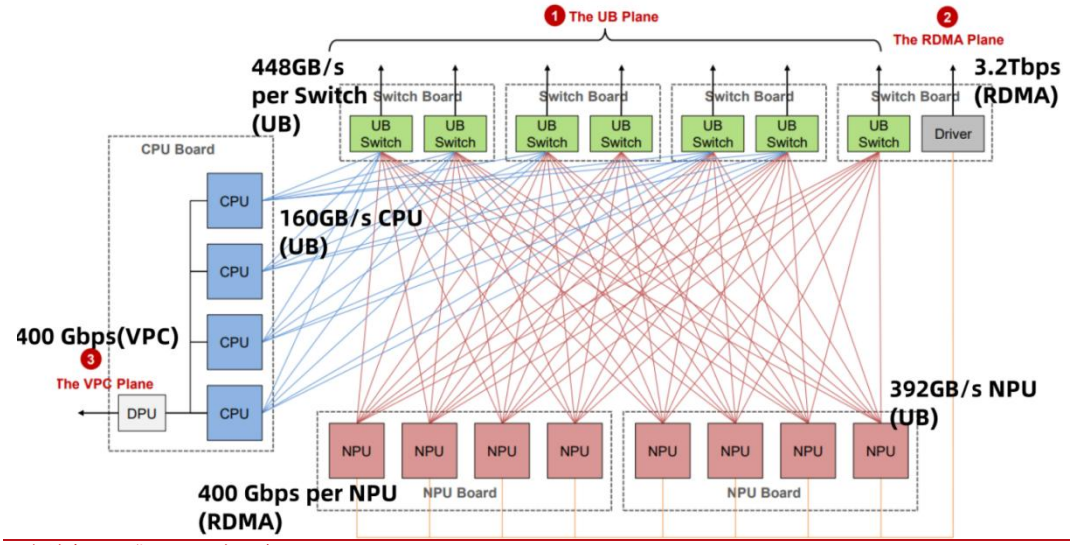


资料来源：华为官网，国盛证券研究所

计算、内存与通信一体化，提升 MoE 推理有效吞吐。CM384 超节点将计算、内存访问和通信平面进行一体化设计，适配大模型推理中的高频数据交换需求。对于 MoE 模型而言，瓶颈往往来自专家路由、KV Cache 访问、跨卡通信和尾延迟控制，而非单卡峰值算力。通过将更多强耦合通信保留在高带宽域内，超节点能够提升推理系统的稳定性和有效吞吐。

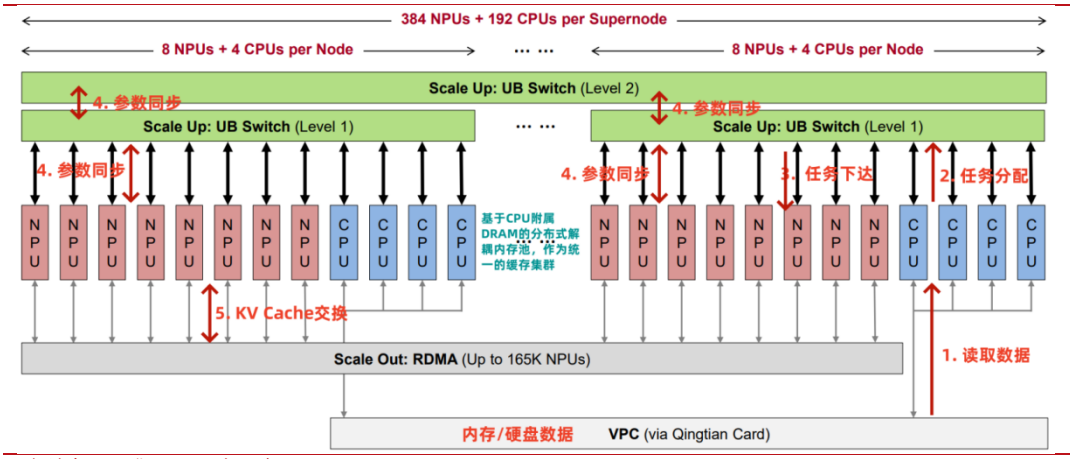
下载日志已记录，仅供内部参考

图表15：华为 CM384 超节点集成计算、内存访问和通信平面



资料来源：华为，国盛证券研究所

图表16：CM384 为 P/D 分离和 MoE 推理做的系统优化

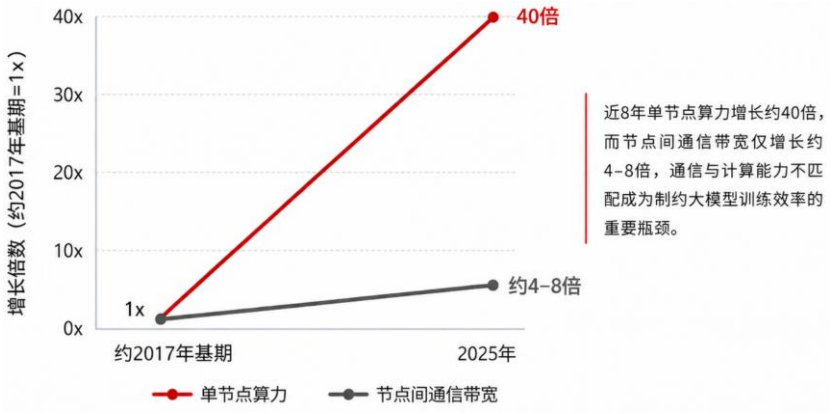


资料来源：华为，国盛证券研究所

互联带宽持续提升，AI 系统从单卡竞争转向集群协同。互联能力正在成为 AI 系统有效算力的重要决定因素。过去 PCIe 带宽明显低于 NVLink，限制了跨卡通信效率；当前封闭互联和开放互联方案均在加速演进，支撑更大规模 Scale-up 域落地。随着模型推理进入高频通信和复杂调度阶段，芯片、网络、内存和软件系统的协同效率将直接影响最终 Token 产出成本。

下载日志已记录，仅供内部参考

图表17：通信带宽增长较慢，系统需要协同



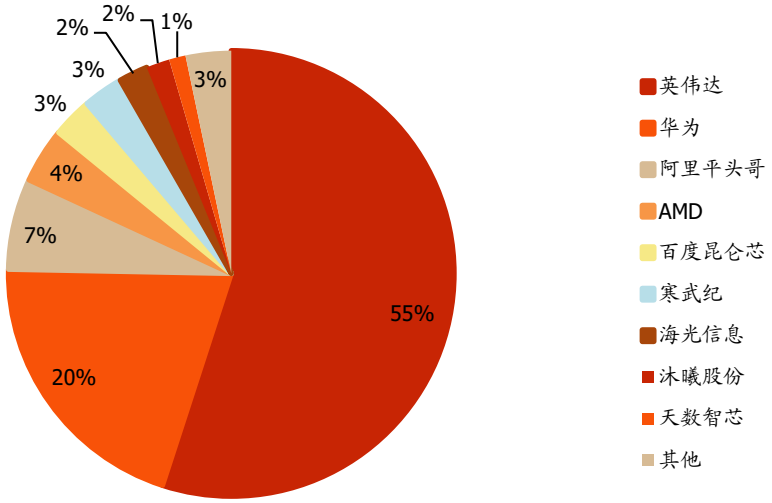
资料来源：华为《基于灵衢®的超节点参考架构白皮书》（2025），国盛证券研究所

2.3 国产 AI 芯片适配加快，自主可控趋势明确

国产 AI 加速卡份额提升，替代空间仍然明确。国产 AI 芯片进入自主可控加速期，核心驱动来自需求增长、外部约束和模型适配三方面。需求端，国内 AI 加速卡市场仍在快速扩张；供给端，海外高端 GPU 进入国内存在持续不确定性；生态端，国产大模型开始主动适配国产芯片，推动硬件从“可替代”走向“可使用、可优化”。

据 IDC，2025 年国内 AI 加速卡出货量约 400 万张，国产芯片厂商合计出货约 165 万张，市占率约 41%。这一数据说明国产 AI 芯片已具备一定规模基础，但仍存在较大的替代空间。考虑到国内 AI 应用需求持续增长、高端 GPU 供给不确定和国产模型适配加快，国产 AI 芯片未来出货量仍具备提升潜力。

图表18：2025 年我国 AI 芯片市场竞争格局



资料来源：TrendForce，国盛证券研究所

国产 AI 芯片的一大挑战是软件生态，尤其是长期以来对 NVIDIA CUDA 生态的依赖。全球顶级通用大模型训练长期绑定 NVIDIA GPU、CUDA 算子库和相关开发工具，国产芯片过去更多停留在推理侧适配，难以进入核心训练环节。但随着国产大模型持续迭代，模型厂商开始主动推进对国产算力平台的适配，也为国产芯片补齐软件生态短板提供了新的驱动力。

下载日志已记录，仅供内部参考

国产模型与国产芯片双向适配加速，自主软硬件生态逐步成型。DeepSeek V4、智谱 GLM-5.2 等模型发布后，昇腾、寒武纪、海光、摩尔线程、沐曦、昆仑芯等国产算力平台陆续完成适配，并持续拓展至 Kimi K3 等主流模型。随着模型侧与芯片侧协同优化持续深化，国产算力有望由“单点适配”迈向“多模型兼容”，并进一步向训练及复杂推理场景渗透。

图表19：国产大模型加速适配国产芯片，软硬协同生态持续完善

国产芯片厂商	DeepSeek	智谱 GLM	Kimi
华为昇腾	V4	5.2	K3
海光信息	V4	5.2	K3
寒武纪	V4	5.2	-
摩尔线程	V4	5.2	K3
沐曦股份	V4	5.2	-
昆仑芯	V4	5.2	-
平头哥	V4	5.2	K3
天数智芯	V4	-	-
壁仞科技	-	5.2	-

资料来源：芯东西、界面新闻、证券时报，国盛证券研究所

3、芯片供应链：制造、封装与存储环节加速补短板

芯片供应链补短板是国产 AI 产业长期发展的关键基础。在先进制程、EDA、光刻设备和高端存储仍受外部限制的背景下，国产算力不能仅依赖单点工艺追赶，而需要通过晶圆制造扩产、系统级架构优化、先进封装、国产存储和设备材料国产化共同提升供给能力。

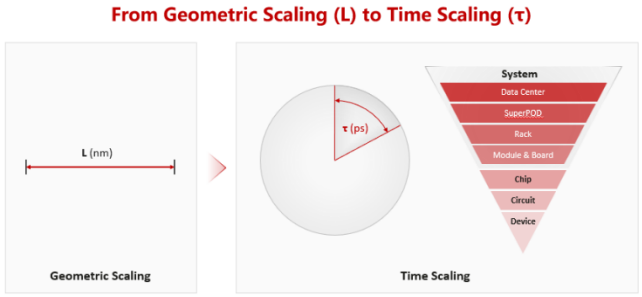
3.1 韬（T）定律：提供系统级追赶路径，弱化单一制程依赖

先进制程演进放缓背景下，系统级优化成为半导体性能提升的重要方向。随着传统摩尔定律驱动下的几何缩微逐渐面临物理极限和经济成本挑战，产业需要探索新的性能提升路径。华为提出的“韬（τ）定律”提出以“时间缩微”替代“几何缩微”，实现半导体与电子系统持续演进。

基于韬定律，华为已在智能手机与 AI 计算领域开展系统级优化实践。过去六年，公司依托器件、电路、芯片及系统等多层级协同设计，已成功设计并量产 381 款芯片，覆盖千行百业应用需求。其中，计划于 2026 年秋季面世的麒麟芯片将率先采用逻辑折叠技术，通过缩短关键路径进一步提升芯片性能；预计到 2031 年，基于韬（T）定律设计的高端芯片晶体管密度有望达到 1.4 纳米制程的同等水平，体现出系统级协同优化对延续芯片性能演进的重要作用。

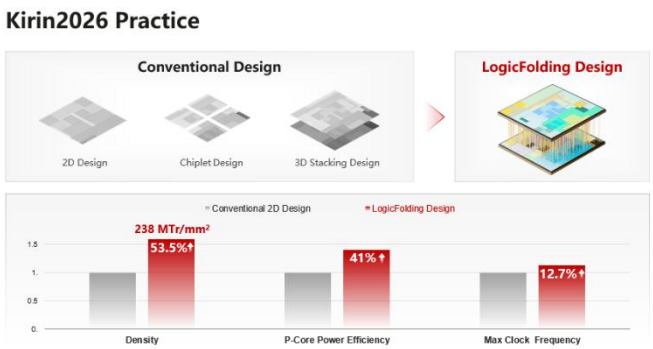
下载日志已记录，仅供内部参考

图表20：韬缩放



资料来源：华为、ISCAS、SECC，国盛证券研究所

图表21：逻辑折叠



资料来源：华为、ISCAS、SECC，国盛证券研究所

具体来看，韬定律并非单一芯片技术，而是一套覆盖器件、电路、芯片到系统的多层级协同优化框架。在器件层，通过优化晶体管结构以及互连电阻、寄生电容，降低器件级时间常数 τ；在电路层，通过逻辑折叠突破传统二维布局限制，缩短关键路径走线长度，降低信号传播延迟，提高晶体管密度和电路性能；在芯片层，通过“软件—架构—芯片”的全栈软硬协同设计，基于实际工作负载优化指令流和数据流，提高系统级并行度和执行效率；在系统层，通过灵衢总线重构计算系统互联协议，实现统一内存编址和原生内存语义，降低系统通信时延。

图表22：韬(τ)定律和摩尔定律对比

对比维度	摩尔定律	韬(T)定律
提出时间	1965 年	2026 年
提出人	戈登·摩尔（时任仙童半导体工程师）	何庭波（华为董事、半导体业务总总裁，依托华为半导体数千名工程师六年技术积累）
提出方式	在 *Electronics*《电子学》期刊发表论文《将更多元件塞进集成电路》（摩尔定律由之后加州理工教授首次提出命名）	国际电路与系统研讨会（ISCAS 2026）提出，在中科院科技论文预发布平台发表论文《A Time Scaling Theory for Multi-Layer Electronic Systems（多层电子系统的时间标度理论）》
核心表述	当价格不变时，集成电路上可容纳的晶体管数目约每 18–24 个月翻一番，性能提升一倍、成本下降一半	以“时间缩微”替代“几何缩微”，系统性降低电子系统的时间常数，持续压缩信号传播时延，实现性能、能效与密度的长期提升
衡量指标	晶体管几何尺寸（纳米）	特征时间常数 τ（秒）
优化方向	**空间维度**：缩小晶体管面积，提高单位面积晶体管密度	**时间维度**：压缩各层级信号传播与处理耗时
技术路径	单一依赖光刻工艺进步，辅以晶体管架构创新（FinFET、GAA）	全栈协同优化：器件、电路、芯片、系统、软件多层级协同
产业依赖	高度依赖顶尖光刻设备（EUV、High-NA EUV）	依赖先进封装、混合键合、光互连、EDA 工具等多环节协同
成本结构	研发与制造成本随制程节点呈指数级增长	成本分布更均衡，先进封装占比提升，光刻成本占比下降
适用场景	所有电子系统，但在 7 纳米以下节点经济效益显著下降	覆盖从移动端到数据中心的全场景，尤其适合 AI 等数据密集型应用
迭代周期	每 2–3 年一个制程节点	可实现更灵活的迭代节奏，不受光刻设备更新周期限制
理论基础	半导体物理中的尺寸缩放效应	计算系统的时间响应特性与层级优化理论
性能提升来源	晶体管数量增加+主频提升	布线长度缩短+寄生损耗降低+架构优化+系统协同
未来潜力	预计在 1 纳米左右节点接近物理极限	理论上可延续至 2035 年及以后，潜力尚未完全释放

资料来源：通信产业网，国盛证券研究所

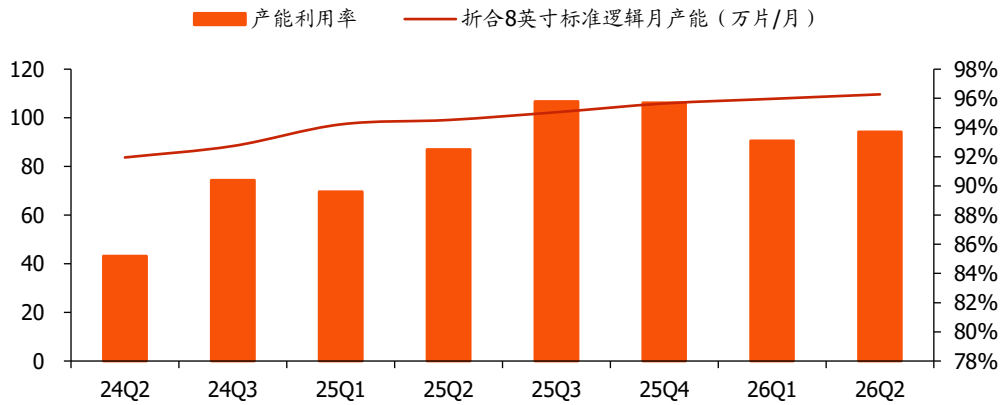
下载日志已记录，仅供内部参考

3.2 国产晶圆厂产能扩张，制造底座价值提升

中芯国际产能稳步扩张，承接国产算力制造需求。国产晶圆制造环节正在从周期复苏走向 AI 算力制造底座。中芯国际业绩保持稳健增长，2026Q2 实现营业收入 30.06 亿美元，同比增长 36.1%；归母净利润 4.79 亿美元，同比增长 261.7%。

中芯国际 26Q2 月产能达 109.7 万片八英寸等值晶圆，产能利用率 93.7%，维持高位。公司 12 英寸晶圆收入占比达 78.2%，中国区收入占比达 90.2%，产品结构持续向 12 英寸平台集中，境内客户需求保持强劲。随着国产 AI 芯片、控制芯片和算力配套芯片需求提升，中芯国际的制造底座价值有望进一步凸显。

图表23：中芯国际月度产能（折合 8 英寸，万片/月）及稼动率



资料来源：中芯国际公告，国盛证券研究所

华虹 12 英寸占比提升，特色工艺平台放量提速。华虹公司体现特色工艺平台的放量趋势。2026Q2，公司实现销售收入 7.18 亿美元，同比增长 26.8%；毛利率 16.5%，同比提升 5.6 个百分点；归母净利润 3,864 万美元，同比增长 385.9%；12 英寸收入占比提升至 62.1%。来自于中国的收入中，MCU、独立式闪存、通用 MOSFET 等平台增长明显。特色工艺广泛应用于电源管理、汽车电子、工业控制和边缘 AI 等领域，其产能扩张将与国产 AI 硬件生态形成互补。

3.3 存储：国产存储产线利用率提升，供给安全价值凸显

出口限制强化补链需求，国产存储仍需跨越工艺瓶颈。存储是 AI 算力系统中不可或缺的基础环节。大模型训练和推理对 HBM、DRAM、NAND 和 SSD 均提出更高要求，其中显存容量、内存带宽和数据读写能力直接影响模型运行效率。MoE 架构通过每个 Token 仅激活部分专家显著降低单 Token 计算量，但模型权重、KV Cache、激活值及通信缓冲区仍占用大量存储资源。因此，国产存储扩产不仅服务传统消费电子周期，也服务 AI 算力长期需求。

长鑫科技 IPO 过会，是国产存储产业资本化进程中的关键节点。2026 年 5 月 27 日，长鑫科技科创板 IPO 获上市委审议通过，拟募资 295 亿元，主要投向存储器晶圆制造量产线技术升级改造、DRAM 存储器技术升级及前瞻技术研究与开发三类项目。从募投结构看，公司资金用途不仅指向现有产线的工艺切换和量产能力提升，也覆盖下一代 DRAM 工艺平台建设及前瞻技术储备，反映国产 DRAM 龙头正由产能爬坡阶段进一步迈向技术迭代与供应链完善并重阶段。

下载日志已记录，仅供内部参考

图表24：长鑫存储募投项目

项目	投资总额	建设周期	主要内容
存储器晶圆制造量产线技术升级改造项目	75 亿	3 年（2025 年三季度前至 2028 年上半年）	对生产线进行技术改造，实现工艺技术切换，同时导入本土及新型设备、材料、零部件，促进供应链多元化
DRAM 存储器技术升级项目	180 亿	3 年（2025 年三季度前至 2028 年上半年）	示范线基础上进行技术改造，提升至更新一代工艺技术平台
动态随机存取存储器前瞻技术研究与开发项目	90 亿	3 年（2026 年 12 月前至 2028 年 12 月）	未来先进 DRAM 所需的前瞻技术研究，保持技术领先性

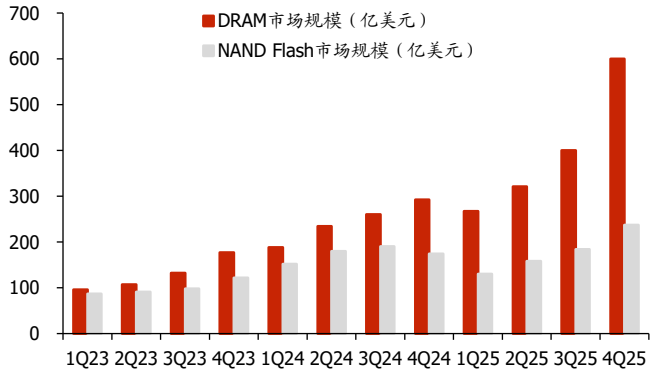
资料来源：长鑫招股书，国盛证券研究所

两长产线运行效率提升，国产存储供给能力持续夯实。长鑫存储在 DRAM 领域产能利用率持续提升，2023/2024/2025 年分别达到 87.06%/92.46%/95.73%，并在合肥、北京两地拥有 3 座 12 英寸 DRAM 晶圆厂。长江存储在 NAND Flash 领域产能利用率同样维持高位，2023/2024/2025/2026Q1 分别为 94.42%/95.43%/97.91%/98.02%，显示下游需求支撑较强、产线运行效率持续提升。DRAM 与 NAND 两条主线同步扩产，有望带动设备、材料及零部件需求释放，推动国产存储进入规模化供给阶段。

存储行业市场规模大、集中度高，国产替代空间明确。从市场规模看，据 CFM，DRAM 与 NAND Flash 在 2025 年下半年均呈现加速扩张趋势，其中 DRAM 市场规模由 1Q25 的 267 亿美元提升至 4Q25 的 600 亿美元，NAND Flash 市场规模由 1Q25 的 130 亿美元提升至 4Q25 的 237 亿美元，反映 AI 服务器、端侧 AI 和数据中心需求正在持续拉动存储容量与带宽升级。

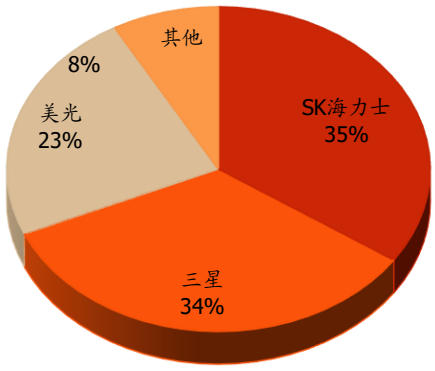
DRAM 市场集中度尤其高，根据长鑫科技招股说明书及 Omdia 数据，2025 年 SK 海力士、三星电子、美光、分别为 34.48%、33.96%、23.41%，前三大厂商合计占据绝大部分市场份额。**截至 25Q4，长鑫科技的全球市场份额已经达到 7.67%。高集中度使存储行业具备较强周期属性，也意味着国产厂商在供给安全和价格波动对冲方面具备战略价值。**

图表25：全球 NAND Flash/DRAM 市场规模季度变化（单位：亿美元）



资料来源：CFM 闪存市场，国盛证券研究所

图表26：2025 年 DRAM 市占率



资料来源：长鑫科技招股说明书，Omdia，国盛证券研究所

下载日志已记录，仅供内部参考

4、投资建议

4.1 GPU 与超节点：关注国产算力替代和系统级架构升级

寒武纪：国产 AI 芯片盈利里程碑显现，需求传导持续增强。公司 2026 年上半年营收和利润实现高增长，预付款、存货和收入之间形成较清晰的需求传导链条。对于芯片设计公司而言，客户需求确认后进入投片、存货形成、产品交付和收入确认环节，预付款与存货增长往往反映后续收入基础。我们认为，寒武纪是国产 AI 芯片的重要代表，若后续产品交付、生态适配和客户拓展持续推进，有望受益于国产算力替代和国内 AI 推理需求放量。

海光信息：CPU+DCU 双轮驱动，国产算力生态持续扩展。海光信息专注于高端处理器研发设计，形成海光 CPU 和海光 DCU 两大产品体系，覆盖服务器、工作站及人工智能计算等核心场景。公司 CPU 兼容 x86 生态，DCU 基于 GPGPU 架构适配类 CUDA 环境，凭借软硬件生态优势持续推进国产算力替代，有望受益于 AI 基础设施建设和信创产业发展。

中科曙光：验证十万卡集群能力，超节点与算力运营优势凸显。全国产十万卡 AI 超集群曙光 8000 落成，并接入全国一体化算力网络。十万卡集群对互联延迟、算力协同、负载均衡和系统可靠性提出高要求，能够验证公司在超节点产品和大规模工程建设方面的能力。我们认为，中科曙光具备服务器、超算、算力运营和国产生态协同优势，有望受益于全国一体化算力网和国产 AI 基础设施建设。

盛科通信：超节点交换芯片龙头，受益高速互联需求提升。盛科作为国内领先的以太网交换芯片设计厂商，2.4T 芯片已量产，12.8T/25.6T 产品已在客户处进入市场推广和逐步应用阶段。AI 超节点和大规模集群建设对低延迟、高带宽、高可靠互联提出更高要求，交换芯片成为算力系统有效吞吐的重要环节。我们认为，盛科有望受益于国产 AI 集群、超节点架构和高速互联网络建设，属于国产算力从单卡竞争转向系统协同过程中的关键环节。

4.2 存储：关注 AI 需求驱动下的 DRAM 与存储控制国产化

长鑫科技：国内 DRAM 龙头，深度受益存储上行周期。长鑫是中国第一、全球第四的 DRAM 厂商，已实现 DDR4、LPDDR4X、DDR5、LPDDR5/5X 等产品覆盖和迭代升级。AI 推理需求高增带动数据中心对 HBM、大容量 DDR5 和企业级 SSD 需求提升，DRAM 供需紧张有望持续。我们认为，长鑫是国产存储替代的核心标的，兼具 AI 需求拉动和供应链安全价值。

国科微：国产 AI SoC 芯片龙头，多领域布局打开成长空间。国科微围绕超高清智能显示、智慧视觉、车载电子及人工智能等领域，形成覆盖视频编解码、ISP、NPU 及 AI SoC 的芯片产品体系。公司持续推进 AI 芯片研发，布局 8TOPS-128TOPS 端侧算力芯片，有望受益于超高清产业升级、智能终端 AI 化及国产算力生态建设。

4.3 半导体设备、材料与 Fab：关注国产算力供应链补短板

中科飞测：先进制程量检测设备验证进展顺利，国产替代空间明确。中科飞测 2025 年营收实现高增长，净利润扭亏为盈，公司检测设备实现收入 13.64 亿元，同比增长 38.52%，量测设备实现收入 6.23 亿元，同比增长 72.71%。量检测设备是晶圆制造良率

下载日志已记录，仅供内部参考

提升、先进制程导入和产线稳定运行的重要支撑，国产晶圆厂扩产和先进制程验证将带动本土设备需求。我们认为，中科飞测在量检测环节具备国产替代逻辑，后续应关注先进制程客户验证、订单转化和盈利能力改善。

频准激光：精准激光技术领先企业，量子与半导体双轮驱动成长。公司聚焦量子科技与半导体领域，推出窄线宽、低噪声、高稳频的精准激光器，覆盖 177nm–5000nm 宽波段，核心技术源于“种子源 + 光纤放大 + 非线性频率变换 + 稳频”自主技术路线。2025 年毛利率超 69%，显著高于行业均值，源于高技术壁垒与定制化服务优势，客户涵盖中科院、清华等顶尖科研平台与半导体企业，形成强技术粘性与市场稀缺性。

长川科技：半导体测试设备龙头，封测检测全品类布局构筑壁垒。长川科技专注于集成电路封装测试及晶圆检测设备研发，产品覆盖测试机、分选机、探针台及 AOI 设备，广泛应用于消费电子、5G 通信、AI 计算等领域。公司通过自主研发与海外并购持续完善全球化布局，凭借丰富产品矩阵和技术积累，有望受益于国产半导体设备替代及先进封装需求增长。

中船特气：电子特气国产化推进，业绩保持稳健增长。中船特气 2025 年实现营收 22.60 亿元，同比增长 16%；归母净利润 3.46 亿元，同比增长 12.43%。电子特种气体是集成电路制造的关键材料，公司电子特气成熟产品、在建及在研产品品类已近百种，覆盖清洗、刻蚀、沉积、掺杂等半导体核心工艺环节。我们认为，中船特气具备半导体材料国产替代属性，有望受益于 Fab 扩产和关键材料自主可控。

蜀道装备：深冷装备与气体运营打开成长空间。蜀道装备是 LNG 及空分装备领域企业。我们认为，蜀道装备可从两方面受益：一是半导体、电子制造和先进材料环节对高纯气体及配套装备需求提升；二是算力基础设施扩张带来的能源和工业气体配套需求。

华虹宏力：12 英寸平台放量，特色工艺景气改善。2026Q2，公司实现销售收入 7.18 亿美元，同比增长 26.8%；毛利率 16.5%，同比提升 5.6 个百分点；归母净利润 3,864 万美元，同比增长 385.9%。12 英寸收入占比提升至 62.1%，MCU、独立式闪存、通用 MOSFET 等平台增长明显。特色工艺广泛应用于电源管理、汽车电子、工业控制和边缘 AI 等领域，其产能扩张将与国产 AI 硬件生态形成互补。

中芯国际：产能稳步扩张，承接国产算力制造需求。中芯国际 26Q2 月产能达 109.7 万片八英寸等值晶圆，产能利用率 93.7%，维持高位。26Q2 实现营业收入 30.06 亿美元，同比增长 36.1%；归母净利润 4.79 亿美元，同比增长 261.7%。公司 12 英寸晶圆收入占比达 78.2%，中国区收入占比达 90.2%，产品结构持续向 12 英寸平台集中，境内客户需求保持强劲。

4.4 AIDC：关注算力网建设和资源运营能力

东阳光：液冷+算力服务协同布局，打造 AI 算力基础设施闭环。公司依托高端铝箔核心产品研发生产能力及氟氯化工技术优势，延伸布局液冷科技业务，自主研发氟化冷却液，掌握 CDU、冷板等核心部件制造能力，建成一体化算力液冷配套体系。公司正推进重大资产重组收购秦淮数据 100%股权，并与其共同开拓 AIDC 算力基建、运维及算力服务业务，推动 AI 算力硬件、算力基建与运维、算力服务全链条贯通。

协创数据：智能算力服务商，服务器交付与算力运营协同发展。公司围绕智能算力产品与算力服务两大方向布局，已实现多型号算力服务器及算力集群解决方案交付，其中 X86 图形渲染加速服务器已应用于视频创作、电商设计等领域，ARM 服务器已导入头部游戏客户。公司依托 NVIDIA CLOUD PARTNER 资质，持续拓展云算力租赁、全球 DICT

下载日志已记录，仅供内部参考

综合服务及算力集群运维业务，为模型训练推理、自动驾驶、生物医药等领域客户提供算力服务。

集智股份：关注公司在算力基础设施相关方向的延展。集智股份原有业务具备设备与工业自动化属性，公司校直设备和校扭设备在 AI 液冷服务器中可以发挥关键作用。若后续围绕 AIDC 形成更清晰布局，有望打开新的成长空间。

4.5 模型与垂类应用：关注国产模型能力提升与行业 AI 落地

智谱：国产通用大模型核心厂商，GLM 系列持续迭代。智谱 GLM-5.2 面向长程任务场景进行优化，实现 Solid 1M 无损上下文，可支持大规模代码工程、自动化研究、性能优化等复杂任务处理。模型针对长程 Coding Agent 场景进行了强化训练，在超长上下文下保持稳定性能，并通过深度思考能力提升复杂任务执行效果。我们认为，智谱具备国产开源模型、Coding Agent 和企业级 API 场景的综合布局能力，后续有望受益于国产模型从“可用”走向“好用”的趋势。

海致科技：AI“除幻”第一股，图模融合产业智能体提供商。Atlas 图谱解决方案是公司于往绩记录期间的核心产品之一。Atlas 图谱解决方案基于公司行业领先的图计算与数据分析能力构建，包含 DMC 数据智能平台、Atlas 知识图谱平台及 AtlasGraph 图数据库。Atlas 智能体基于公司的图模融合技术开发，通过融合大语言模型的推理能力与知识图谱的准确性及可解释性，赋能企业实现智能自动化、场景化推理及自然语言交互。我们认为，海致更适合承担高可靠、强逻辑、重交付的产业 AI 场景，有望成为主权 AI 和企业智能体落地的重要标的。

金山办公：国产办公软件龙头，AI 赋能推动智能办公升级。金山办公深耕办公软件领域超过 35 年，形成覆盖个人、企业及政务场景的产品体系，包括 WPS 个人版、WPS365 及 WPS AI 等核心产品。公司依托自主研发的文档处理、在线协同、安全文档及 AI 办公技术，持续完善“云+协作+AI”智能办公生态，有望受益于办公智能化升级与国产软件替代。

金蝶国际：AI 企业级 SaaS 龙头，AI 智能体驱动企业数智化升级。金蝶围绕“AI 优先、订阅优先、全球化”战略推进业务，以灵基（Lingee）企业智能体操作系统结合 SaaS 产品，将 AI 能力融入财务、供应链、制造等核心业务场景，推动企业从流程驱动向智能驱动转型。公司持续深耕企业云服务市场，在 ERP、财务及企业应用软件领域保持领先地位，有望受益于企业数字化升级与 AI 应用落地。

中控技术：工业 AI 领军企业，AOP 推动流程工业智能化升级。中控技术专注于流程工业自动化、数字化与智能化，推出自主运行工厂 AOP，以时间序列大模型 TPT 和通用控制系统 UCS 为核心，构建“识别-评估-决策-执行”智能闭环。公司通过 AI 与工业控制系统深度融合，推动工厂由自动控制向智能决策、自主运行演进，助力流程工业实现安全、提质、降耗与低碳发展。我们认为，中控的“控制入口+工业数据+工艺 Know-how”构成差异化壁垒，是工业 AI 落地的重要标的。

晶泰控股：AI for Science 平台先行者，AI 驱动新药研发商业化加速。晶泰科技专注于 AI 与生命科学融合，围绕药物研发打造 AI for Science 平台，已覆盖全球前 20 大药企中的 17 家，并推动多款创新管线进入临床或 IND-enabling 阶段。公司持续拓展行业 AI 模型及智能体应用，推动 AI 在分子设计、实验研发等环节形成闭环，成为 AI for Science 领域首家实现盈利的港股上市公司。

深信服：云计算与 AI 安全服务商，推动企业智能化升级。深信服围绕通算与智算基础设施，构建覆盖超融合 HCI、桌面云 aDesk、统一存储及 AI 算力网关的产品体系，

下载日志已记录，仅供内部参考

为企业提供托管云、私有云等云数据中心解决方案。公司持续推进 AI 基础设施布局，通过 AI 算力网关、FastGPT 商业版及 AI 编程平台等产品，支持企业算力调度、模型治理及 Agent 应用开发。

容知日新：人工智能驱动的工业 AI 服务企业，实施“1+N+X”场景化智能运维解决方案体系，打造出 PHMGPT 故障诊断智能体，并在重点行业深度落地。报告期内 AI 诊断及服务收入同比增长约 17%。同时积极推广“软件平台 + 全无线智能监测硬件”的 AI 订阅制模式，加快国内业务闭环及海外标杆项目落地。

风险提示

宏观经济不及预期。若国内外宏观经济修复节奏弱于预期，企业 IT 支出、云服务采购、智能算力基础设施建设及终端电子消费需求可能受到影响，进而压制 AI 应用落地、算力租赁、半导体制造和存储等环节的景气传导。同时，若资本开支意愿下降或融资环境收紧，相关企业在产能建设、技术研发和客户拓展方面的投入节奏也可能放缓。

AI 应用需求释放不及预期。若 Agent、代码生成、企业级大模型调用、行业智能体等场景商业化进展慢于预期，Token 调用量增长可能放缓，进而影响模型厂商 API 收入、算力租赁需求及相关基础设施投入节奏。若下游客户对 AI 投入的 ROI 验证周期拉长，也可能导致模型应用、垂类软件和 AIDC 需求释放不及预期。

国产替代进程不及预期。国产 GPU、超节点、互联架构、半导体设备材料、晶圆制造及存储等环节仍处于持续验证和爬坡阶段，若芯片性能、集群稳定性、开发工具链、主流模型适配、先进制程良率或关键材料设备导入进度低于预期，可能影响国产 AI 产业链在训练、推理和供应链配套场景中的渗透速度。

下载日志已记录，仅供内部参考

免责声明

国盛证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，但本公司及其研究人员对该等信息的准确性及完整性不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，可能会随时调整。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。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司力求报告内容客观、公正，但本报告所载的资料、工具、意见、信息及推测只提供给客户作参考之用，不构成任何投资、法律、会计或税务的最终操作建议，本公司不就报告中的内容对最终操作建议做出任何担保。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。

本报告版权归“国盛证券股份有限公司”所有。未经事先本公司书面授权，任何机构或个人不得对本报告进行任何形式的发布、复制。任何机构或个人如引用、刊发本报告，需注明出处为“国盛证券研究所”，且不得对本报告进行有悖原意的删节或修改。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的任何观点均精准地反映了我们对标的证券和发行人的个人看法，结论不受任何第三方的授意或影响。我们所得报酬的任何部分无论是在过去、现在及将来均不会与本报告中的具体投资建议或观点有直接或间接联系。

投资评级说明

投资建议的评级标准		评级	说明
评级标准为报告发布日后的 6 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以摩根士丹利中国指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准。	股票评级	买入	相对同期基准指数涨幅在 15%以上
		增持	相对同期基准指数涨幅在 5%~15%之间
		持有	相对同期基准指数涨幅在-5%~+5%之间
		减持	相对同期基准指数跌幅在 5%以上
	行业评级	增持	相对同期基准指数涨幅在 10%以上
		中性	相对同期基准指数涨幅在-10%~+10%之间
		减持	相对同期基准指数跌幅在 10%以上

国盛证券研究所

北京
地址：北京市丰台区丽泽路 22 号院新华社国家金融信息大厦西塔 B 座 38 层
邮编：100077
邮箱：gsresearch@gszq.com

上海
地址：上海市浦东新区南洋泾路 555 号陆家嘴金融街区 22 栋
邮编：200120
电话：021-38124100
邮箱：gsresearch@gszq.com

南昌
地址：南昌市红谷滩新区凤凰中大道 1115 号北京银行大厦
邮编：330038
传真：0791-86281485
邮箱：gsresearch@gszq.com

深圳
地址：深圳市福田区福华三路 100 号鼎和大厦 24 楼
邮编：518033
邮箱：gsresearch@gszq.com