算力芯片行业报告

大模型驱动算力变革，国产算力迎增量机遇

行业研究 · 行业专题

电子 · 半导体

投资评级：优于大市（维持）

证券分析师：叶子
0755-81982153
yezi3@guosen.com.cn
S0980522100003

证券分析师：胡慧
021-60871321
huhui2@guosen.com.cn
S0980521080002

证券分析师：张大为
021-61761072
zhangdawei1@guosen.com.cn
S0980524100002

证券分析师：詹浏洋
010-88005307
zhanliuyang@guosen.com.cn
S0980524060001

证券分析师：连欣然
010-88005482
lianxinran@guosen.com.cn
S0980525080004

大模型驱动算力变革，国产算力迎增量机遇

国信证券 GUOSEN SECURITIES

- 当前算力需求正从前期的“模型训练”加速向规模化落地的“应用推理”侧外溢。随着摩尔定律边际效应减弱，算力竞争的核心已从传统的“单芯片峰值性能提升”全面转向“芯片、软件生态与系统级集群的综合效率优化” 。在海外高端芯片销售受限的背景下，国内信创需求与大模型迭代共振，推动本土AI芯片厂商加速适配并放量，国产算力全栈生态迎来增量机遇 。

- AI计算异构化与系统级协同，芯片竞争从峰值性能转向综合效率：海外大模型（如OpenAI、Google等）保持每半年一代的高频迭代，追求智能化升级 ；国产大模型在经历了技术蓄力后，自2025年起以DeepSeek-R1、智谱GLM等为代表的产品迭代显著提速，中美已成为全球大模型供给的两大核心。随着AI应用规模化落地，针对推理基础设施的投资规模从2024年开始超越训练侧，推理侧更强调高吞吐、大并发以及成本性能的平衡。AI算力产业正从“单芯片性能提升”转向由芯片、先进封装、高带宽存储（HBM）、编译框架、液冷及大规模集群构成的系统级协同优化。AI系统本质是异构计算体系：CPU负责通用调度，GPU承担大规模并行通用加速，而TPU/NPU等ASIC芯片则在特定模型和推理降本中发挥效率优势，形成百花齐放的长期共存格局。

- 海外芯片龙头从单芯片竞争走向平台化交付：英伟达依托GPU、CUDA生态、NVLink和Blackwell整柜系统，将单芯片竞争扩展为“芯片+网络+软件+系统”的平台竞争；谷歌以TPU为核心服务自有模型（如Gemini）和云客户；AWS则通过Inferentia和Trainium两条ASIC产品线将云端AI成本拆解，降低单位训练与推理成本。

- 国产算力适配与信创共振：国内信创市场正从传统的通用算力国产替代（CPU、操作系统等）转向智能算力基础设施升级。2026年5月，国家首次在安全可靠测评中设立专门AI芯片品类，华为海思、平头哥、海光信息、壁仞科技、摩尔线程等9款国产芯片获评安全可靠等级Ⅰ级，正式纳入信创体系。国产算力的焦点不只是单卡峰值，而是“芯片 + HBM + 互联 + 服务器 + 编译器/算子库 + 推理引擎 + 模型适配”的全栈效率。未来随下游国产云计算厂商、运营商等需求打开，国产算力芯片有望持续持续增长。

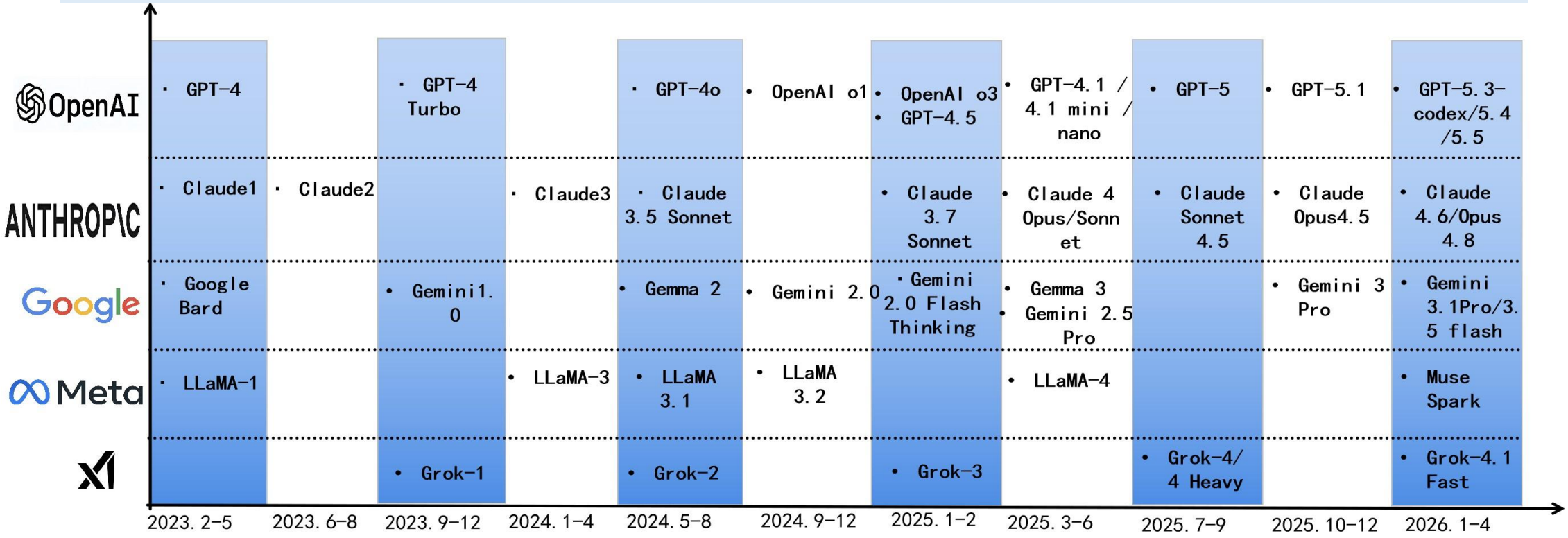
目录

大模型发展趋势——更智能、更快捷、更便宜

国信证券 GUOSEN SECURITIES

- 自2022年11月ChatGPT发布以来，生成式AI逐渐从专业领域走向大众视野。随着GPT-4多模态功能的推出和英伟达H系列芯片的起量，使得2023年被视为人工智能产业的重要转折点。
- 海外大模型追求智能化升级。自2023年以来，OpenAI、Anthropic、Google、Meta、xAI等海外大模型厂商保持高频迭代，部分厂商保持每半年一代的迭代速度，通过算力扩充和算法优化来持续推动产品智能化升级和丰富度提升。

图：主要海外大模型迭代进度

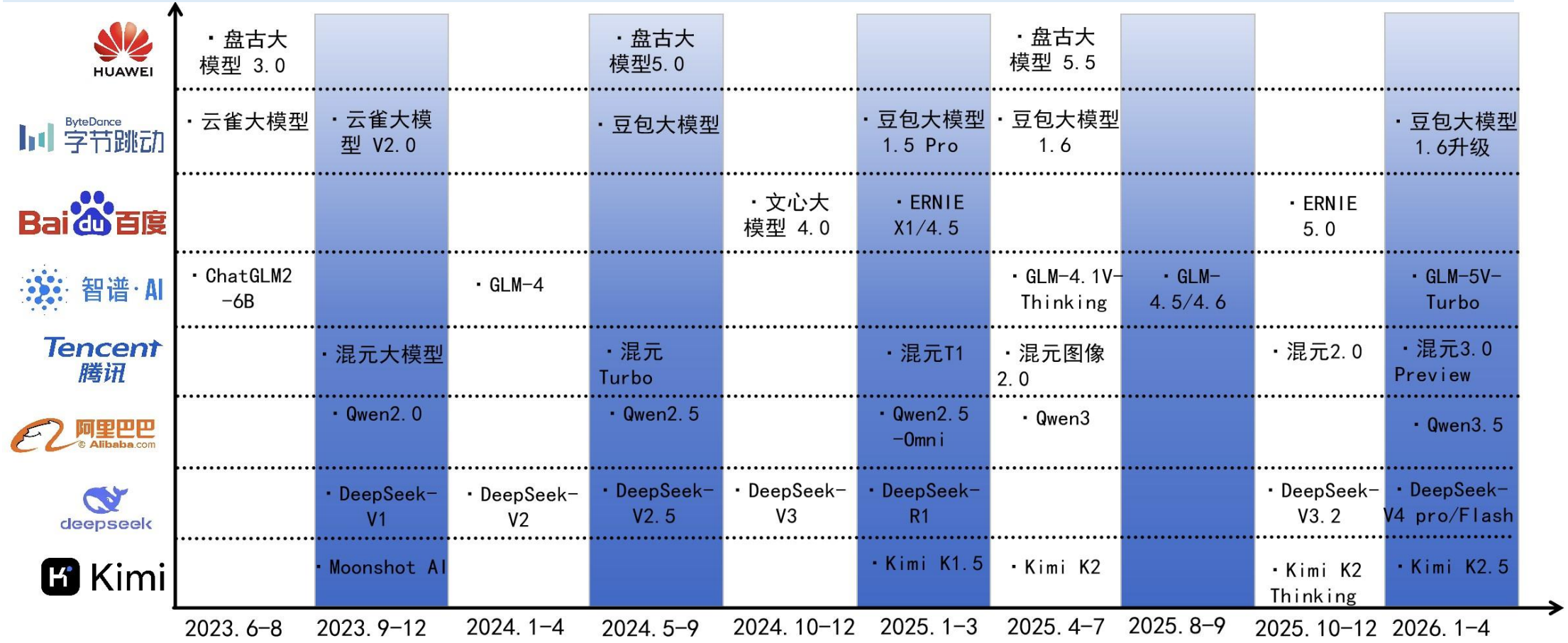


资料来源：各公司官网，国信证券经济研究所整理

大模型发展趋势——更智能、更快捷、更便宜

- DeepSeek推动国产大模型崛起。受高端芯片供给约束，2023-2024年的国产大模型迭代速度放缓。但随着DeepSeek-R1的横空出世，2025年开始国产大模型迭代提速，产品丰富度提升。

图：主要国产大模型迭代节奏



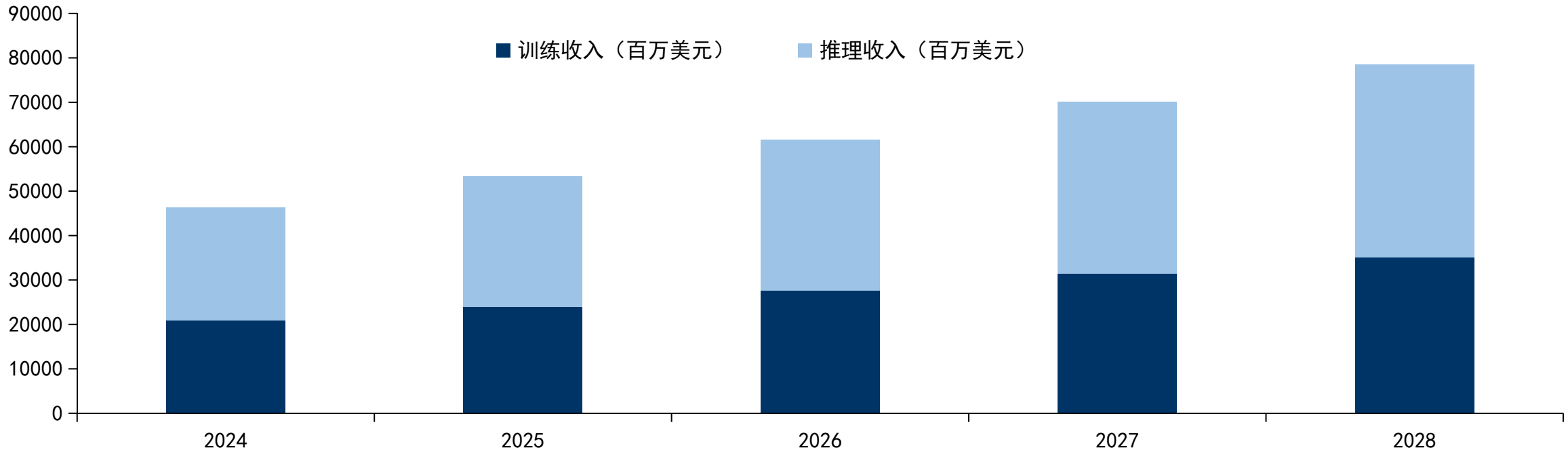
资料来源：各公司官网，国信证券经济研究所整理

大模型发展趋势——算力需求从训练侧外溢至推理侧

国信证券 GUOSEN SECURITIES

- 算力需求从训练侧外溢至推理侧。大模型发展之初，行业聚焦于大模型训练以占据市场领先地位。2022-2024年AI基础设施投资更多聚焦模型训练侧，基础设施技术栈重点在于打造千卡万卡级算力集群，确保训练过程稳定性，优化从硬件到AI开发框架到模型侧技术栈提高算力效率。随着模型应用规模化落地，AI算力需求由训练侧向推理侧外溢。据IDC，针对推理基础设施投资规模2024年开始超越训练侧。推理侧更强调高吞吐、大并发以及成本性能平衡。推理芯片需求增速更快，预计推理收入2024-2028年CAGR=14.3%，训练收入2024-2028年CAGR=13.8%。

图：全球AI训练与推理算力市场规模预测

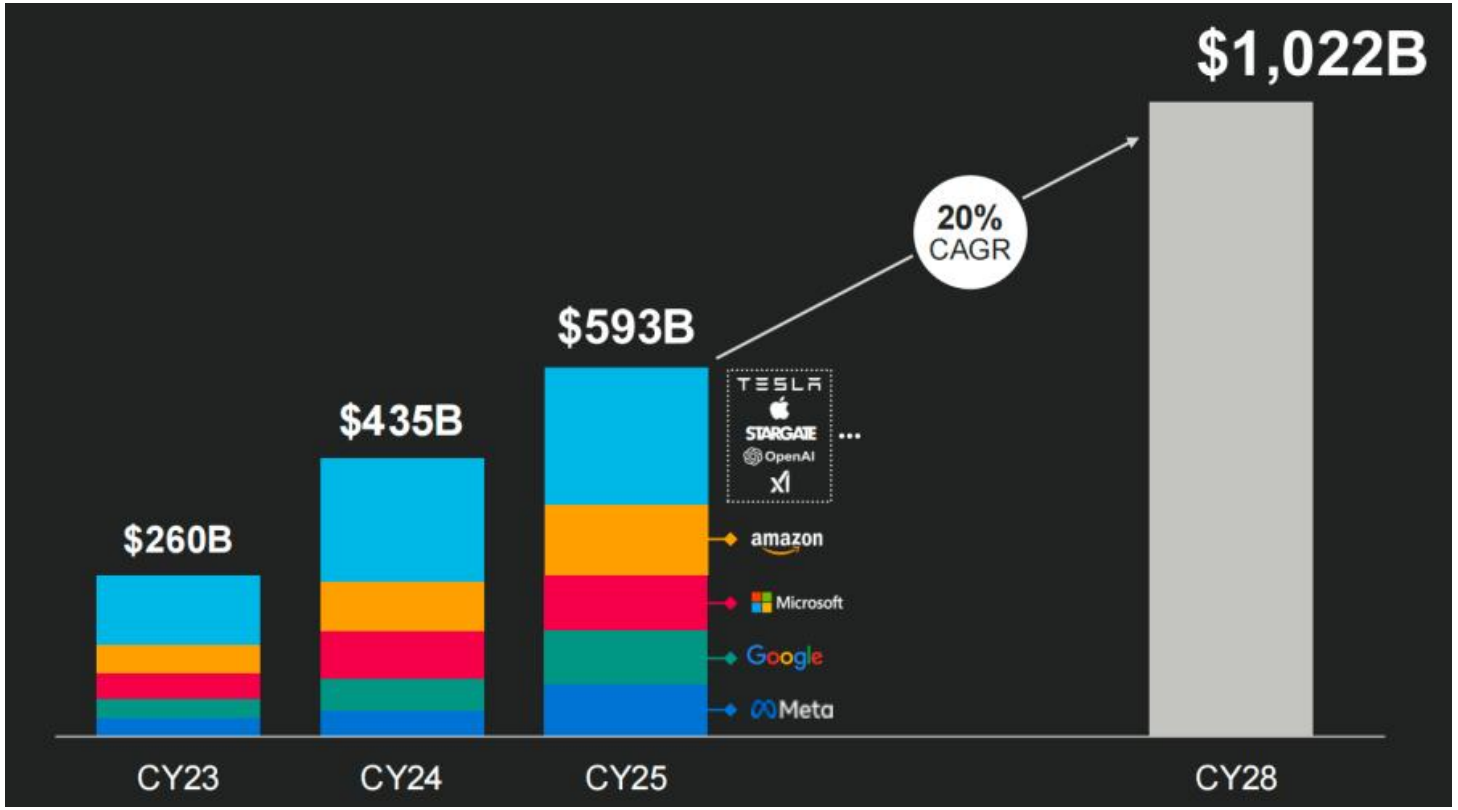


资料来源：IDC，国信证券经济研究所整理

大模型发展动力——全球算力建设资本开支持续加大

- 算力的扩充规模决定了大模型智能化的上限，海外大模型厂商持续加大AI资本开支来保障其产品的领先性。根据Marvell指引，2025-2028年全球AI算力资本开支仍将保持20%的年均增速成长。云服务商加快自身的AI基础设施投资。2025年谷歌、亚马逊资本开支总额领先，进入2026年预计谷歌（Alphabet）预计资本开支有望达1800-1900亿美元；Meta预计资本开支为1250-1450亿美元；亚马逊预计资本开支为2000亿美元。

图：北美主要大模型厂商的资本开支情况



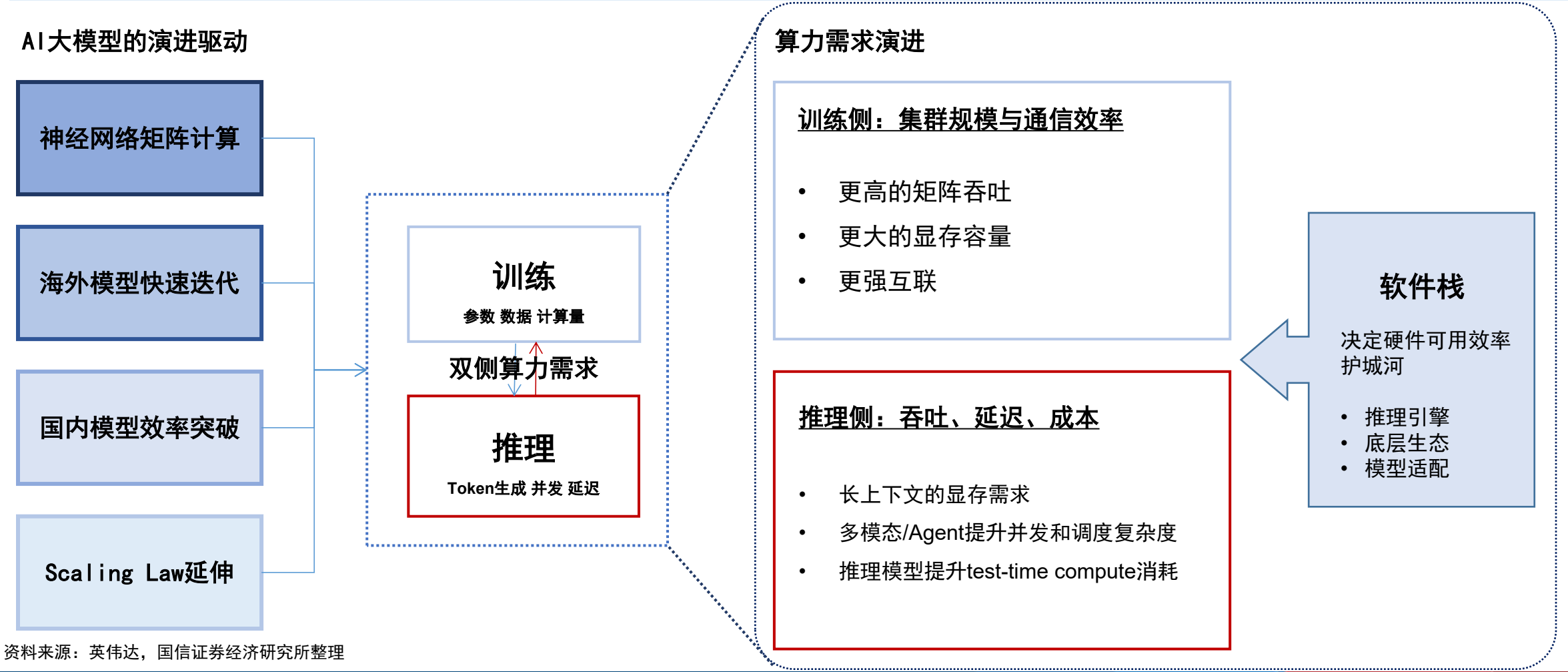
资料来源：Marvell，国信证券经济研究所整理

AI大模型驱动的产业链升级需求

训练侧的算力需求：更大规模、更高吞吐、更强互联将模型训练出来

推理侧算力需求：更低延迟、更高并发、更低成本将模型持续服务出去

图：AI大模型驱动算力需求

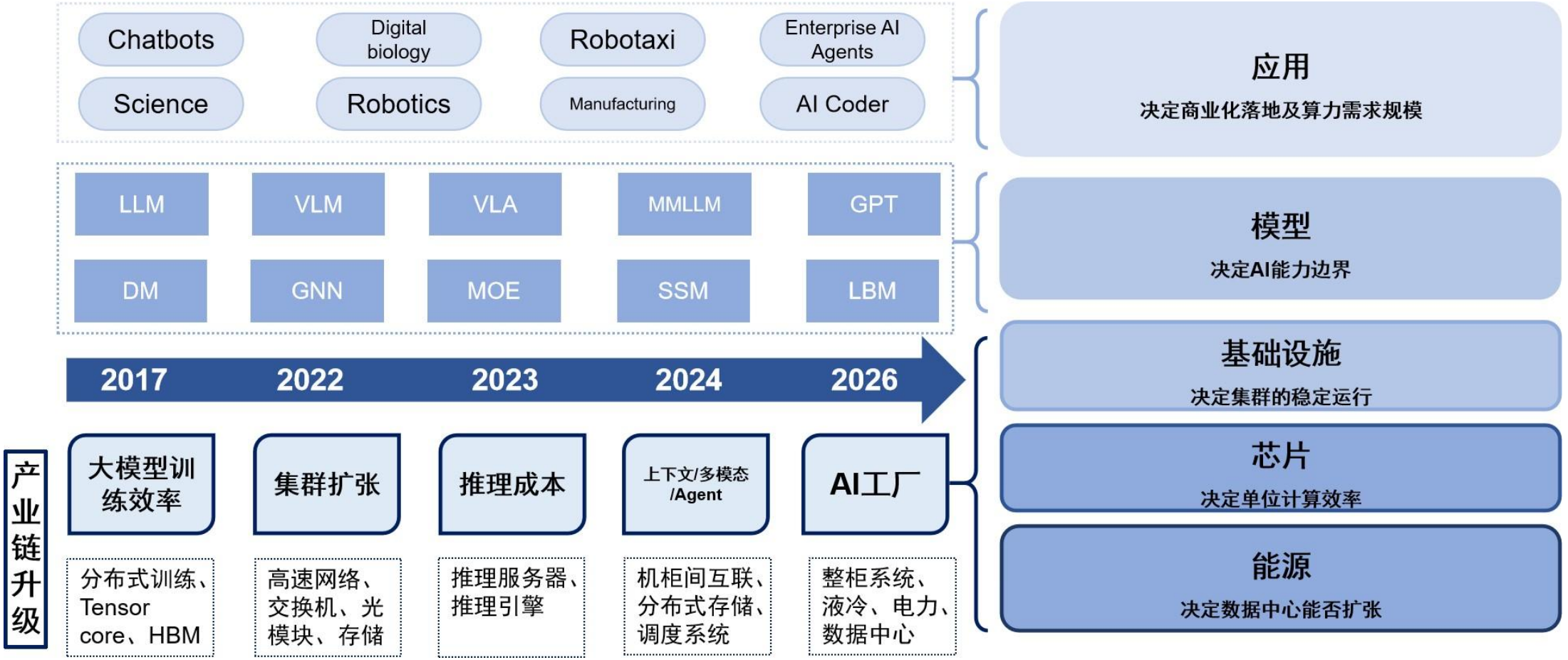


资料来源：英伟达，国信证券经济研究所整理

AI五层蛋糕

- AI不是单一技术或产业，而是一套从底层资源到上层应用逐层链接的基础设施体系，是一块五层“蛋糕”：能源→芯片→基础设施→模型→应用。

图：AI五层蛋糕



资料来源：英伟达官网、国信证券经济研究所整理

算力芯片：更大存储、更稠密算力、更大带宽

国信证券 GUOSEN SECURITIES

- 从芯片来看，以英伟达的芯片技术路径为例，更大的存储量、更强的稠密算力能力、更大的带宽速度是算力芯片一直以来的迭代方向。

图：算力芯片架构及性能指标

架构	A100	H100	H200	GH200	B100	B200	Full B200	GB200
	Ampere	Hopper			Blackwell			
显存大小	80GB	80GB	141GB	96/144GB	180/192GB	180/192GB	192GB	384GB
显存宽带	2TB/s	3.35TB/s	4.8TB/s	4/4.9TB/s	8TB/s	8TB/s	8TB/s	16TB/s
FP16稠密算力（FLOPS）	312T	1P	1P	1P	1.75P	2.25P	2.5P	5P
INT8稠密算力（OPS）	624T	2P	2P	2P	3.5P	4.5P	5P	10P
FP8稠密算力（FLOPS）	X	2P	2P	2P	3.5P	4.5P	5P	10P
FP6稠密算力（FLOPS）	X	X	X	X	3.5P	4.5P	5P	10P
FP4稠密算力（FLOPS）	X	X	X	X	7P	9P	10P	20P
NVLink宽带	600GB/s	900GB/s	900GB/s	900GB/s	1.8TB/s	1.8TB/s	1.8TB/s	3.6TB/s
功耗	400W	700W	700W	1000W	700W	1000W	1200W	2700W
备注	1个Die	1个Die	1个Die	1个Grace CPU 1个H200 CPU	2个Die	2个Die	2个Die	1个Grace CPU 2个Blackwell CPU

资料来源：vertiv《智算中心基础设施演进白皮书》、国信证券经济研究所整理

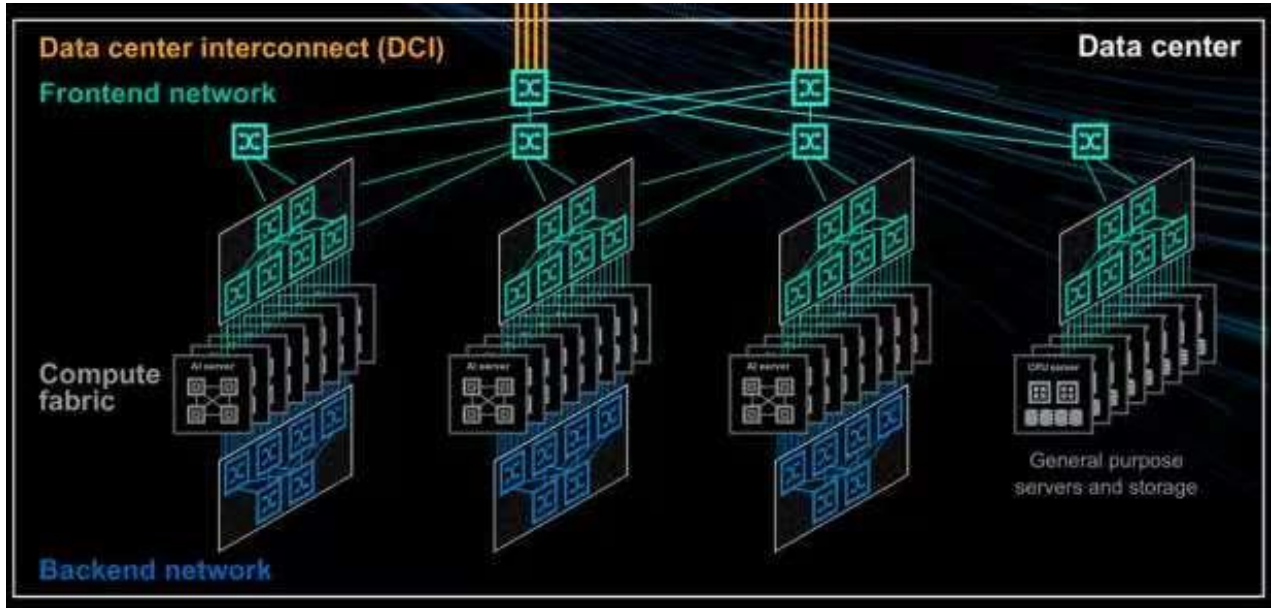
算力发展趋势——互联端

- 愈发“才思敏捷”的人工智能需要更强的算力支撑，而算力的升级并不局限于芯片制程升级，而是机柜级、集群级的整体化技术迭代。为了突破集群算力的瓶颈，互联技术沿着Scale-out（集群级升级）和Scale-up（机柜级升级）两个方向发展。

✓ **从机柜来看**，随着AI资本开支规模的持续上涨，人工智能软/硬公司愈发倾向于本地化部署算力，因此更一体化、更具性价比的算力部署单元更获得青睐，超节点需求应运而生。自B系列以来，GB200开辟了AI算力超节点时代，单机柜算力升级向超节点模式发展，且超节点带动液冷、铜缆等技术的升级。

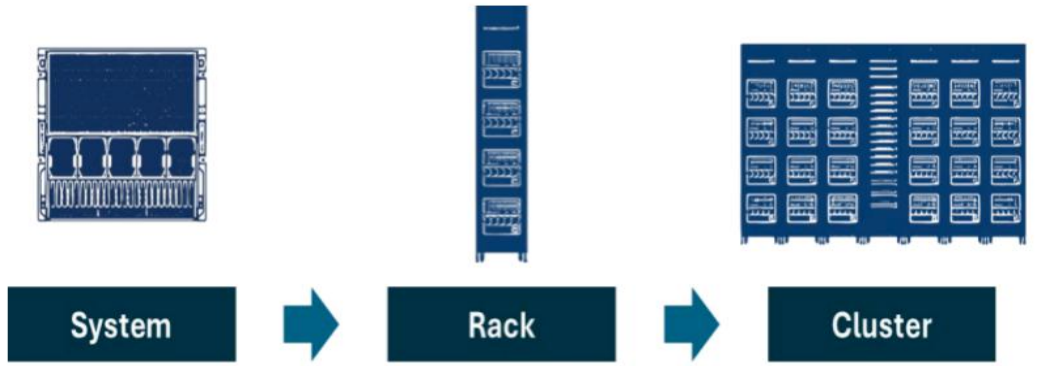
✓ **从集群来看**，随着大模型参数的持续扩张，算力集群规模越来越大。128节点的超级集群可由32个叶交换机和16个脊交换机构成，且交换机间需要较高的传输速度来满足快速通信需求。为了满足日益提升的传输速度需求，CPO技术和PCB材料的升级成为重要方向。

图：算力的Scale-Up和Scale-Out



资料来源：Marvell，国信证券经济研究所整理

图：算力从系统节点扩展到互联集群



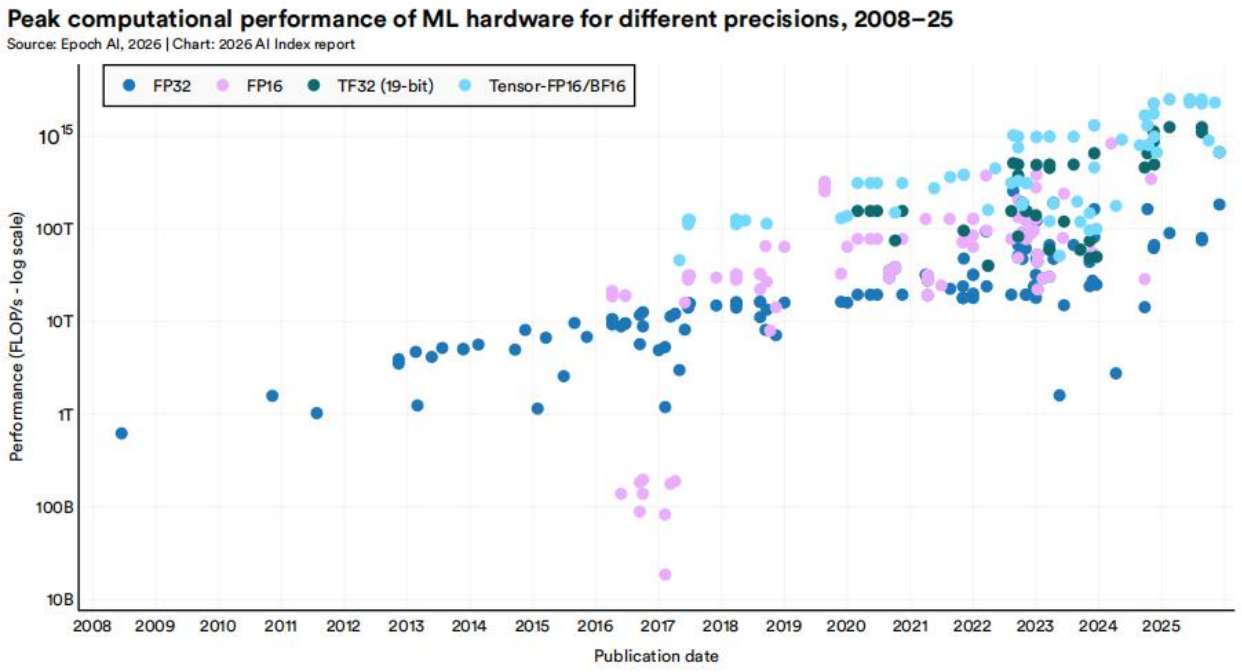
资料来源：SMCI，国信证券经济研究所整理

目录

AI大模型驱动的算力需求不断提升

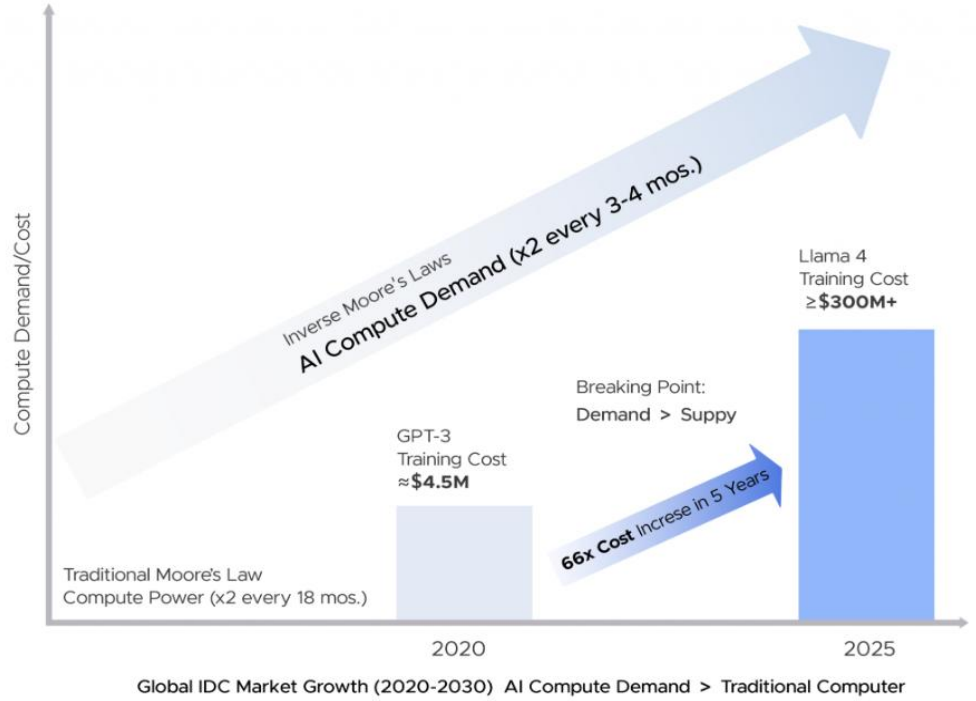
- 模型计算需求提升推动硬件峰值性能提升，系统效率成为关键。机器学习硬件的峰值计算性能随着产品迭代呈指数级增长。算力需求的增长速度已经超过传统摩尔定律的节奏。传统芯片依靠制程升级实现性能提升的边际效应正在减弱，而大模型训练、推理、多模态和Agent应用持续推高算力需求。我们认为，AI算力产业正从“单芯片性能提升”转向“系统级协同优化”：一方面，GPU/NPU、Chiplet、先进封装和高带宽存储成为硬件升级重点；另一方面，算子库、编译器、推理框架、并行训练和集群调度的重要性提升。中长期看，算力竞争将不再局限于芯片峰值性能，而是取决于芯片、软件生态、服务器、网络互联、液冷和大规模集群的综合效率。

图：机器学习硬件峰值计算性能



资料来源：Artificial Intelligence Index Report 2025，国信证券经济研究所整理

图：AI计算需求

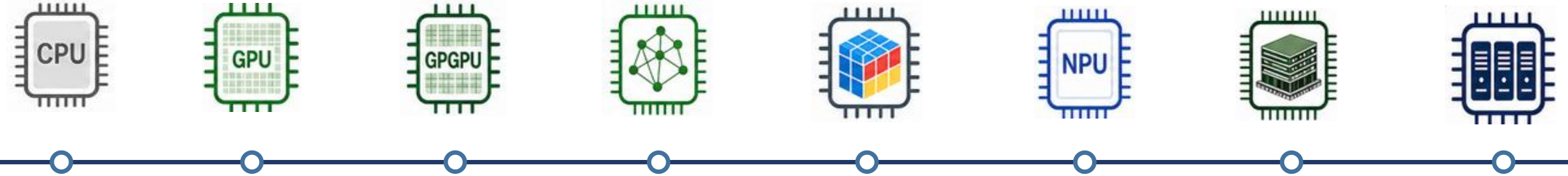


资料来源：艾邦半导体、中科算网，国信证券经济研究所整理

算力芯片的演进——通用计算→并行计算→矩阵加速→系统级算力

国信证券 GUOSEN SECURITIES

图：算力芯片演进



时间	1940-1970	1999-2005	2006-2012	2012	2016-2017	2017-2020	2020-2024	2024-2026
阶段	CPU通用计算	GPU并行计算萌芽	GPGPU平台化	深度学习GPU/专用张量加速		端侧NPU 云端ASIC	大模型与高密度算力集群	
主要厂商	IBM intel AMD	NVIDIA AMD	NVIDIA	NVIDIA	NVIDIA Google	aws HUAWEI	NVIDIA Google AMD HUAWEI	NVIDIA aws AMD Google HUAWEI
代表芯片	CPU：复杂控制与通用逻辑	GPU：图形渲染与大规模并行	GPGPU：通用并行加速	GPU训练：AI计算需求	TPU/Tensor Core GPU：矩阵乘法加速	NPU/Inferentia/Ascend	A100/H100/MI 300X/TPU：HBM+高速互联	Blackwell/MI系列/Trainium：机架级算力

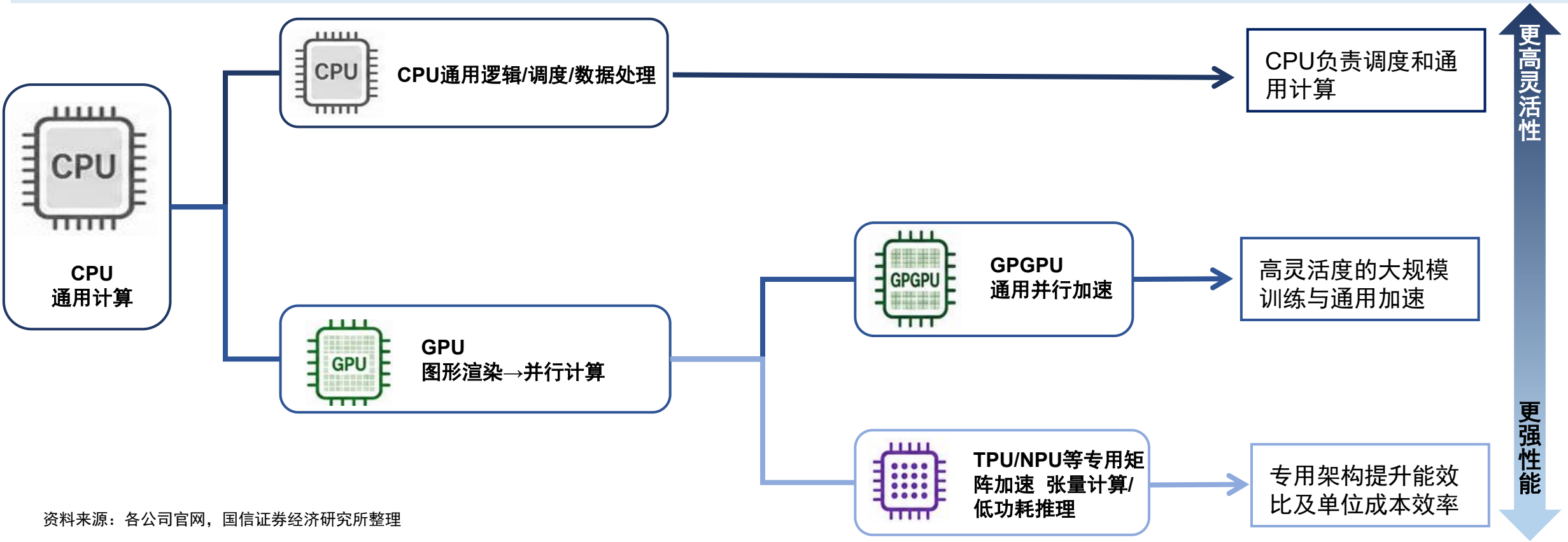
计算芯片演进方向：通用计算→并行计算→矩阵加速→系统级算力　　从单芯片→软件生态与系统级集群

资料来源：各公司官网，国信证券经济研究所整理

AI应用带动算力从通用算力到智能算力演进

- 通用算力以CPU为算力核心，智算算力采用芯片异构计算架构，结合CPU、GPU、NPU、TPU等多种芯片形成高并发分布式计算系统，应用于神经网络模型的训练及推理。大模型训练的效率和成本最优的诉求要求智算中心建立高度集中化的GPU集群。基于GPU分布式工作原理，在更小的物理空间内部署更多的GPU服务器，可以突破分布式计算因带宽和传输距离产生的运算瓶颈，提高集群算效，从而减少大模型的训练时间同时降低训练的成本。
- AI系统的本质是异构计算系统：AI需求推动下，计算芯片演进不是线性替代，而是围绕不同计算任务的分工深化。

图：算力芯片演进

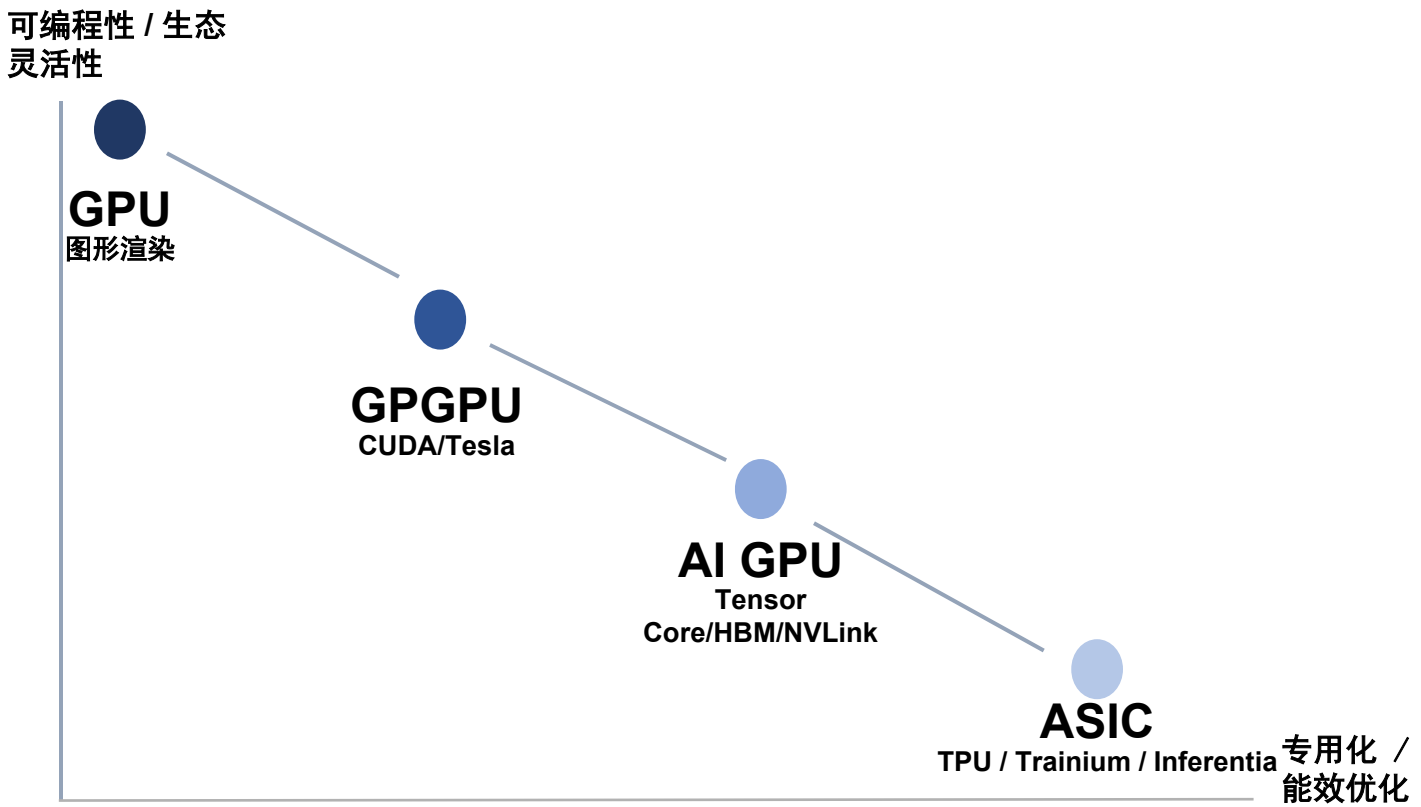


资料来源：各公司官网，国信证券经济研究所整理

AI应用带动算力从通用算力到智能算力演进

- AI需求推动下，计算芯片的演进并不是简单替代，而是不同架构围绕不同任务形成分工。CPU擅长通用计算、逻辑控制和系统调度，是计算体系的基础；GPU依靠大规模并行计算能力，适合神经网络训练和推理中的矩阵、向量运算，成为AI时代的重要加速器；TPU、NPU等ASIC芯片则针对张量计算、矩阵乘和特定AI算子做专门优化，在能效比、成本和特定场景性能上具备优势。
- AI系统更像是异构计算体系：CPU负责调度，GPU承担灵活的大规模加速，TPU/NPU在特定模型、推理和边缘场景中发挥效率优势。随着AI应用持续扩展，计算芯片将呈现长期共存、各司其职、百花齐放的发展格局。随着AI应用从训练扩展到推理、从云端扩展到边缘、从通用大模型扩展到行业模型，不同芯片路线会在性能、成本、功耗、生态和可编程性之间形成差异化竞争。最终格局更可能是CPU、GPU、TPU、NPU及各类AI ASIC长期共存，构成百花齐放的异构算力生态。

图：算力芯片演进



资料来源：各公司官网，国信证券经济研究所整理

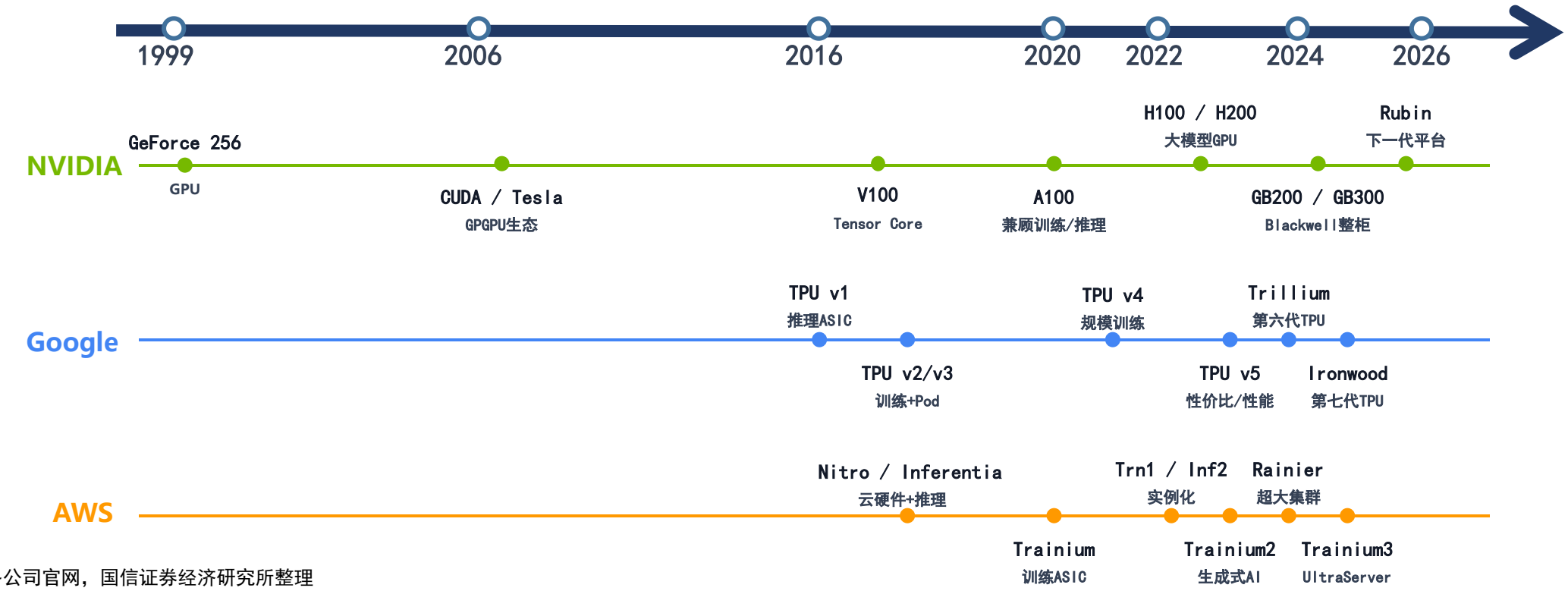
目录

算力芯片竞争已经从“单芯片跑分”转向“系统交付能力”

国信证券 GUOSEN SECURITIES

- 海外AI芯片竞争已从单颗芯片性能比拼演进为芯片、互联、软件、云服务、整机系统的综合交付能力竞争。英伟达依托GPU、CUDA、NVLink和整柜系统构建通用AI算力平台；谷歌以TPU为核心，服务自有模型和谷歌云客户；AWS通过Inferentia和Trainium将训练和推理成本拆解到云实例和自研ASIC体系中。GPU仍是前沿大模型训练和通用推理的重要底座，ASIC在云端降本和规模化部署中持续提升价值。

图：海外AI芯片历史发展



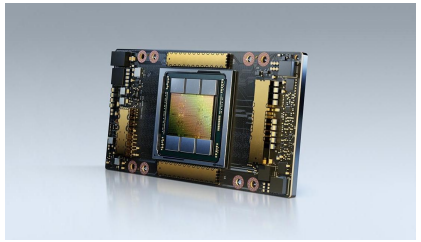
资料来源：各公司官网，国信证券经济研究所整理

英伟达：从GPU供给扩张，走向AI数据中心整体效率提升

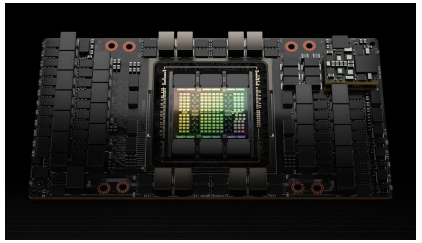
国信证券 GUOSEN SECURITIES

- 英伟达的AI芯片发展始于GPU通用计算。1999年GeForce推动图形处理器普及，2006年CUDA发布，使GPU可用于深度学习等并行计算。2012年AlexNet借助英伟达GPU取得突破，AI训练需求快速增长。此后，英伟达从游戏GPU转向数据中心AI芯片，推出Tesla、Volta架构V100，并加入Tensor Core，大幅提升矩阵运算效率。2020年前后，A100成为大模型训练核心硬件；随后H100进一步强化Transformer计算、互联和能效，支撑生成式AI爆发。近年英伟达继续推出更高性能的Blackwell系列，并通过NVLink、InfiniBand、CUDA生态和整机系统，把单颗芯片竞争扩展为“芯片+网络+软件+系统”的平台竞争。
- 趋势：性能从通用并行计算走向AI专用加速；应用从图形渲染转向数据中心和大模型；竞争重心从单芯片算力转向集群规模、能效、带宽和软件生态。

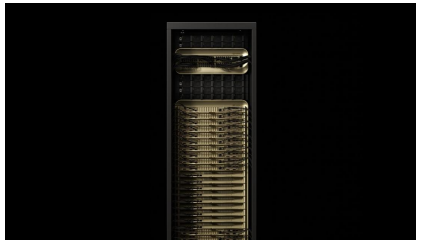
图：英伟达AI芯片历史发展



A100：数据中心GPU统一训练、推理、数据分析



H100：面向Transformer引入FP8与Transformer Engine



GB200：Blackwell从GPU升级为机柜级AI系统

GPU芯片 → 数据中心级系统交付能力

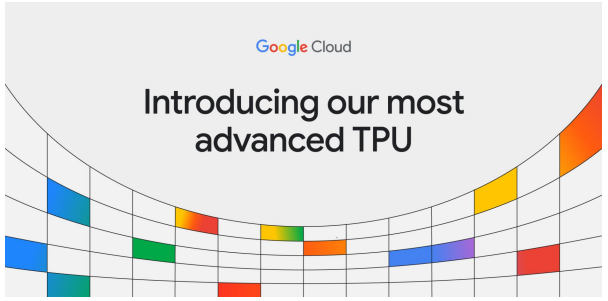
资料来源：公司官网，国信证券经济研究所整理

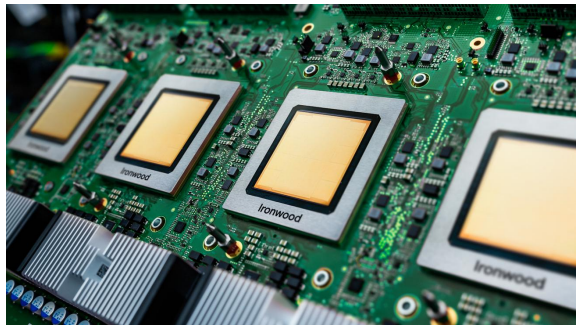
谷歌:从内部推理ASIC，演进为训练与推理兼顾的云端TPU平台

国信证券 GUOSEN SECURITIES

- 谷歌AI芯片发展以TPU为主线。2016年谷歌为搜索、翻译、推荐等内部AI任务推出第一代TPU，重点提升深度学习推理效率。随后TPU v2、v3开始支持大规模训练，逐步服务于Google Cloud客户。2021年前后TPU v4强化集群互联，适合更大模型训练；TPU v5e、v5p分别面向性价比和高性能训练及推理。2024年发布第六代Trillium，服务Gemini等生成式AI模型；2025年推出第七代Ironwood，定位为面向“推理时代”的 TPU，重点支持大规模低延迟推理、多模态生成等推理密集型工作负载。
- 趋势：从内部专用推理芯片走向云端通用AI基础设施；从单芯片性能转向TPU集群、软件栈和能效；应用重心从训练扩展到“训练+推理+多模态大模型”。

图：谷歌AI芯片历史发展





Trillium / TPU v6e：第六代TPU，面向下一代训练和推理

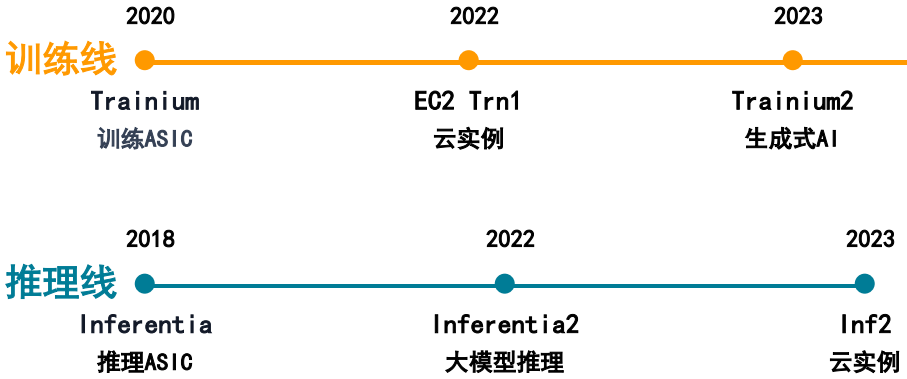
资料来源：公司官网，国信证券经济研究所整理

AWS:把云端AI成本拆成训练与推理两条ASIC产品线

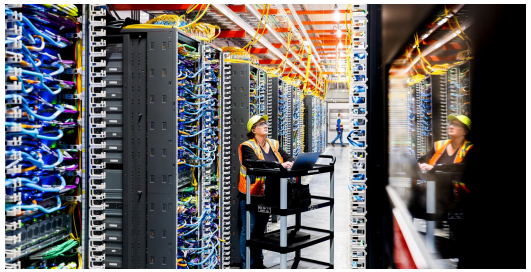
国信证券 GUOSEN SECURITIES

- 亚马逊AI芯片发展围绕AWS云服务展开。2018年，AWS推出Inferentia，主攻低成本、低延迟AI推理，服务云上模型部署。随后Inferentia2提升大模型推理能力。训练侧，AWS推出Trainium，目标是在云端提供比GPU更具成本优势的AI训练方案；Trainium2进一步面向生成式AI和大模型集群训练，并与Neuron软件栈、SageMaker等服务结合。2025年，AWS推出Trainium3和Trn3 UltraServers，强调3nm工艺、更大HBM、更高带宽，以及面向智能体、推理模型和视频生成的训练与推理成本优势。
- 趋势：从推理降本切入，扩展到训练芯片和整机集群；从单一芯片产品走向“芯片+云实例+软件SDK+AI服务”的云平台生态；降低大模型训练和推理的单位成本。

图：亚马逊芯片历史发展



Trainium2芯片：AWS自研训练芯片路线的第二代



Project Rainier：基于Trainium2的超大AI计算集群

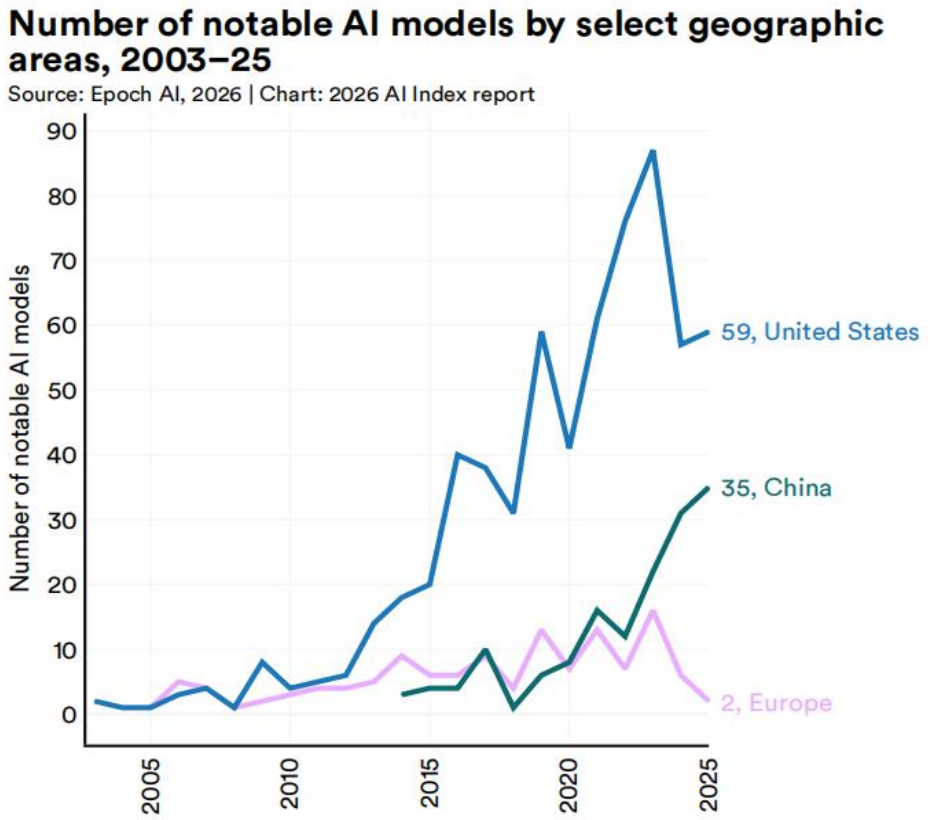
资料来源：公司官网，国信证券经济研究所整理

目录

国产大模型发展迅速

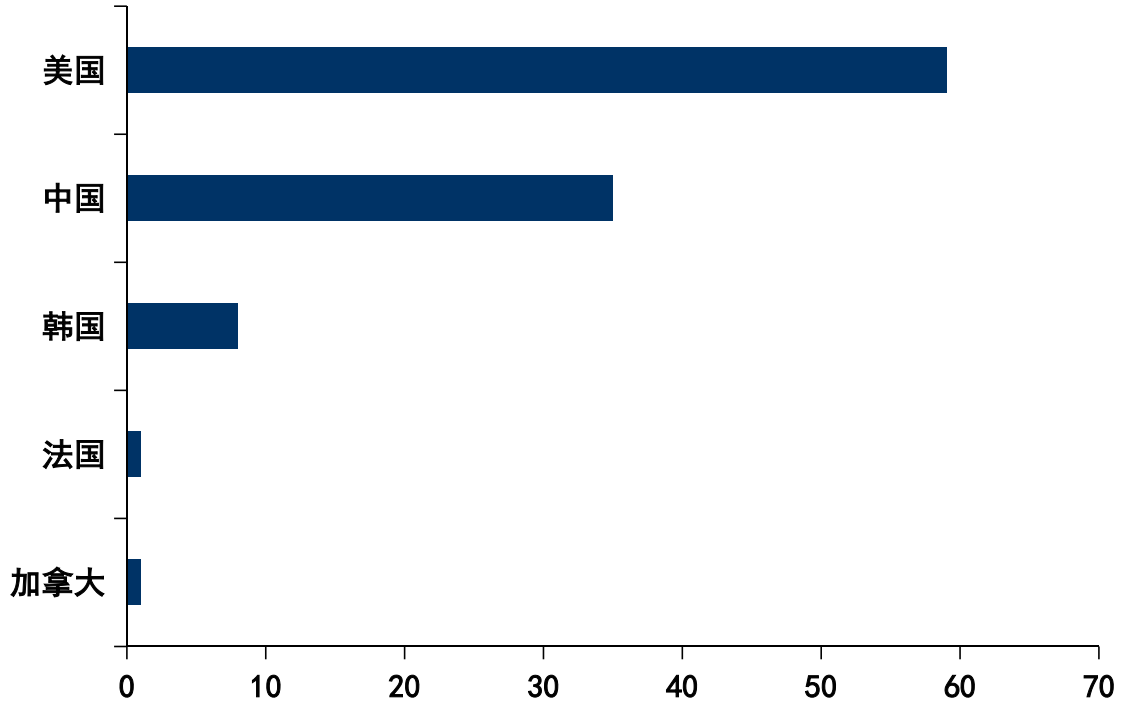
- 我国大模型发展迅速。在知名大模型数量中，2003-2024年美国一直保持领先态势，2023年开始中国实现了对欧洲的超越，中美成为全球大模型供给的两大核心力量，但美国在知名人工智能模型数量上仍领先。

图：2003-2025年各地知名人工智能模型数量



资料来源：Artificial Intelligence Index Report 2026，国信证券经济研究所整理

图：2025年各地知名人工智能模型数量

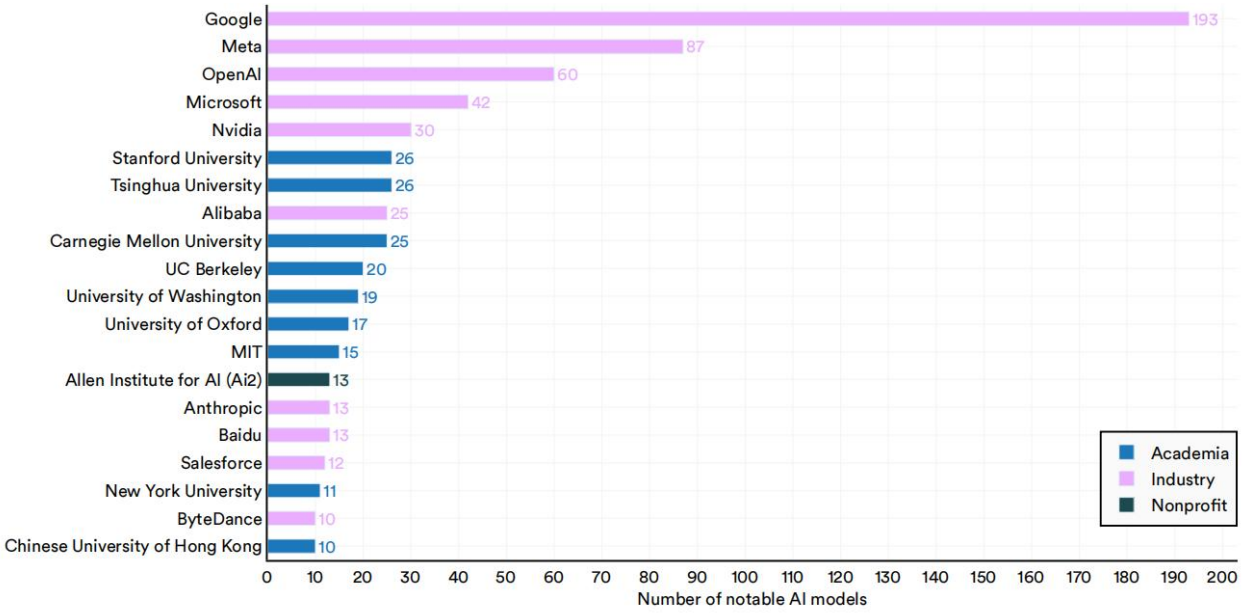


资料来源：Artificial Intelligence Index Report 2026，国信证券经济研究所整理

国产大模型加速追赶

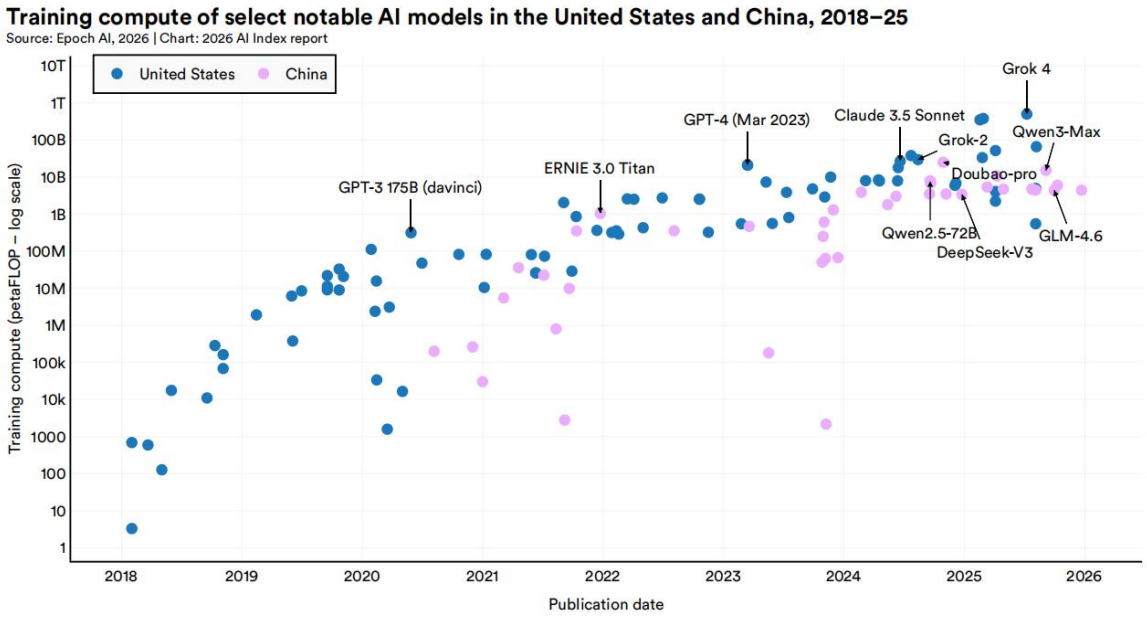
- 国产大模型加速追赶。从2014-2025年，全球前三大模型贡献方为谷歌（193个）、Meta（87个）和OpenAI（60个），中国区最多的是清华大学（26个）。2025年，全球前三大模型贡献方为OpenAI（20个）、谷歌（14个）和阿里巴巴（11个），字节跳动、DeepSeek和腾讯也上榜，国产大模型呈现加速追赶态势。
- 据Epoch AI，知名模型的算力需求已经上升多个数量级，美国模型比中国模型具有更高的计算密集度。

图：2025年知名大模型数量



资料来源：Artificial Intelligence Index Report 2026，国信证券经济研究所整理

图：美国大模型算力密集度高于中国

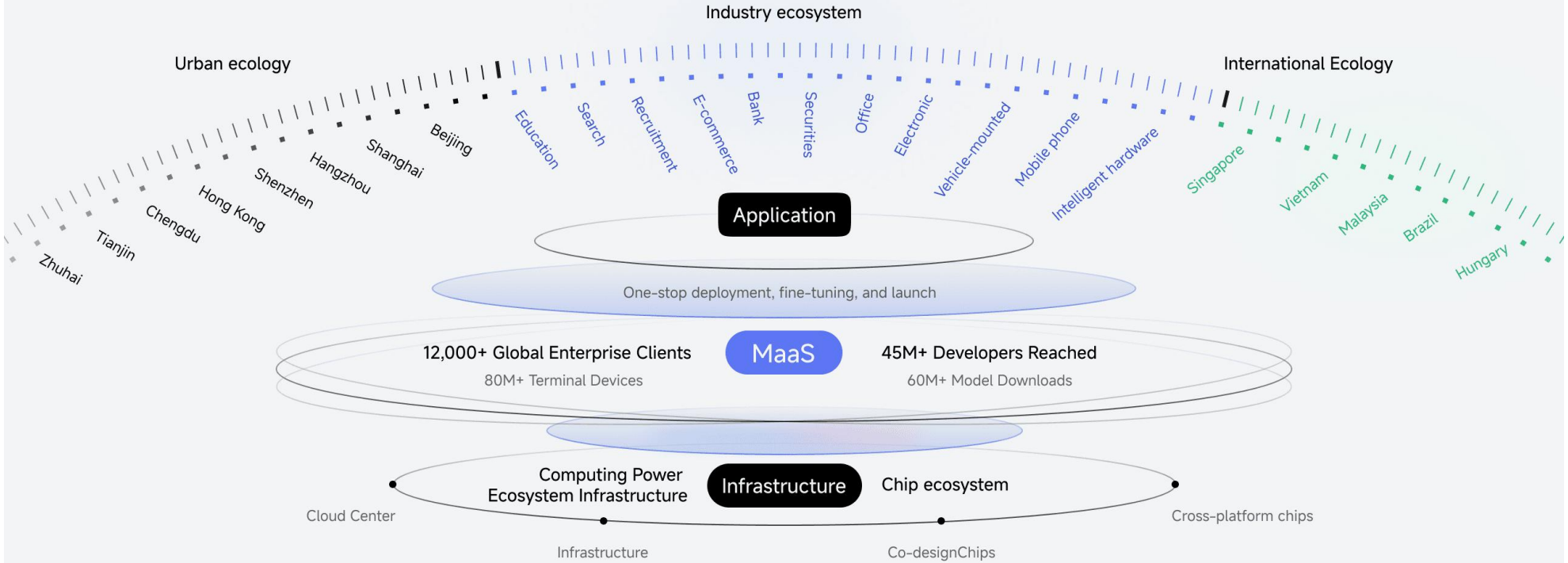


资料来源：Artificial Intelligence Index Report 2026，国信证券经济研究所整理

从国产大模型看算力硬件适配需求

- 国内大模型进入“Agent化、长上下文、多模态、低成本推理”阶段。智谱突出GLM系列在Coding、工具调用、长上下文任务处理和多模态Agent方向的能力，MiniMax-M3强调长上下文、原生多模态和自动化工作流，DeepSeek V3/R1则以MoE、MLA和高性价比见长，DeepSeek V4进一步将上下文扩展。我们认为，国产算力适配重点将从单卡峰值算力竞争，转向模型-芯片-框架-集群协同优化：围绕长上下文提升显存容量、带宽和KV Cache管理；围绕MoE强化专家并行、跨节点通信和负载均衡；完善算子生态、适配主流推理框架。中长期看，国产算力厂商若能深度绑定头部模型架构和应用场景，有望在私有化部署、行业Agent、低成本推理集群中形成突破。大模型体系连接应用、MaaS、基础设施、芯片生态和行业生态，国产大模型商业化不仅是模型能力竞争，也是“模型+平台+算力生态”的系统工程。

图：智谱业务生态图



资料来源：公司官网，国信证券经济研究所整理

从国产大模型看算力硬件适配需求

- 智谱做更完整的AI平台，覆盖办公、企业服务、代码、多模态和行业应用；MiniMax强化长文本、多模态和Agent能力，让模型能看长文档、看视频、调用工具、连续完成复杂任务；DeepSeek突出推理、代码和高性价比，让模型能力更强、使用成本更低。
- 国产算力适配不能仅关注“芯片算力有多高”。需要同时关注：稳定跑长文本任务、支撑多人同时使用、降低推理成本、兼容主流模型框架、支持图文视频等多模态应用的能力。

图：AI大模型及算力适配需求

公司	代表模型/能力	国产算力适配
智谱	GLM系列，覆盖文本、代码、智能体、多模态等能力。	不仅能跑大模型，还要支持企业部署、稳定服务、多用户同时使用，以及图文视频等多模态任务。
MiniMax	MiniMax-M3等模型，强调长上下文、多模态、工具调用和Agent能力。	重视大显存、长文本处理能力和推理效率，支持长时间、多步骤任务稳定运行。
DeepSeek	DeepSeek-V系列、DeepSeek-R系列等，强调推理、代码、长上下文和高性价比。	适配高效率模型架构，重点提升训练/推理速度、显存利用率和低成本部署能力。

资料来源：公司官网，国信证券经济研究所整理

AI加速芯片位于算力中心产业链上游

- AI加速芯片位于算力中心产业链上游，产业链中游是算力中心服务商，包含电信运营商、云服务商、第三方算力中心服务商，下游应用于互联网、AI、金融等领域。

- 华为是国内AI算力产业链布局最完整的厂商之一，已形成从昇腾AI芯片、Atlas加速卡/模块、AI服务器、AI集群，到CANN软件栈、MindSpore/MindIE推理训练工具链及华为云ModelArts平台的全栈布局。其优势在于软硬件一体化、端边云协同，构建国产算力底座。

- 根据华为官网，华为Atlas 800T A2训练服务器/800I A2推理服务器基于华为自研鲲鹏920处理器（CPU）和昇腾910AI处理器（NPU）。

图：算力中心产业链

图：华为AI算力布局

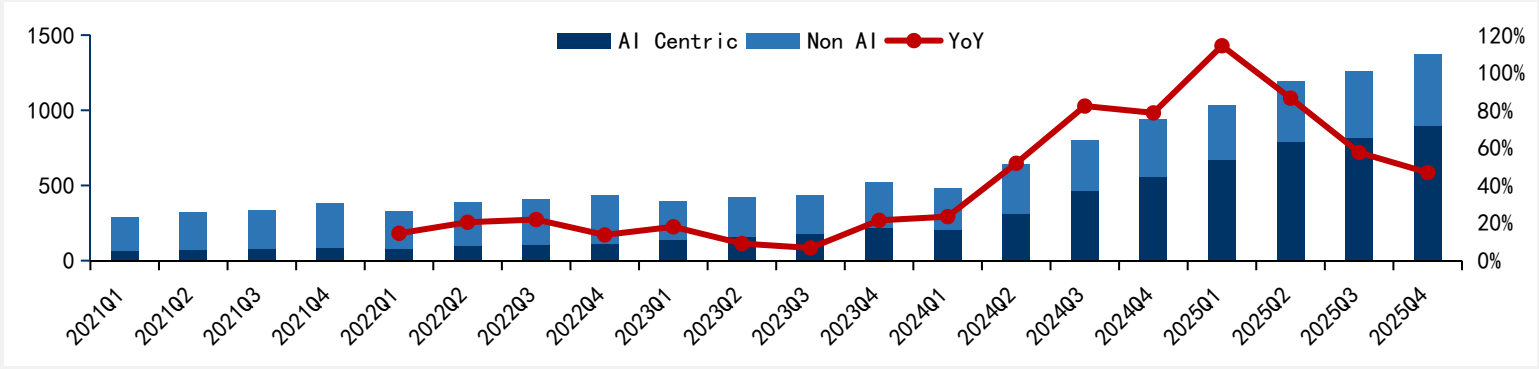
资料来源：灼识咨询、公司官网，国信证券经济研究所整理

中国与海外AI基础设施客户结构不同

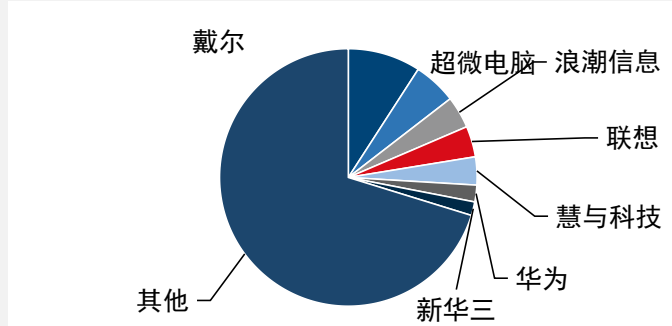
- 从AI基础设施客户看，中国和海外市场的需求结构不同：中国通信服务商占比高于全球，资本开支结构上从传统网络投资转向云、算力和AI基础设施；此外，数字服务提供商、非服务商客户相较海外占比较大。

图：AI基础设施市场规模、客户结构、市场份额、资本开支

AI基础设施市场规模（亿美元）

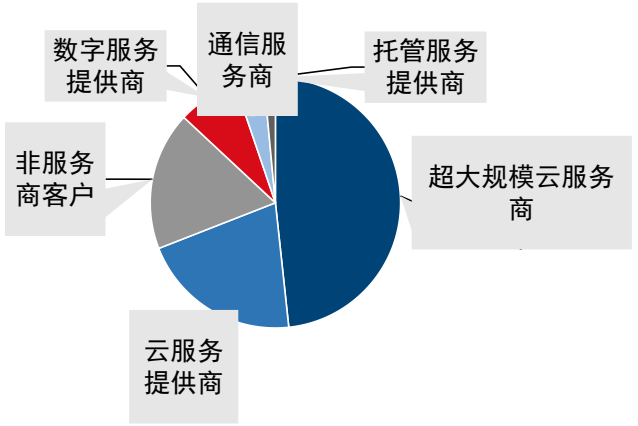


AI基础设施市场份额（2025）

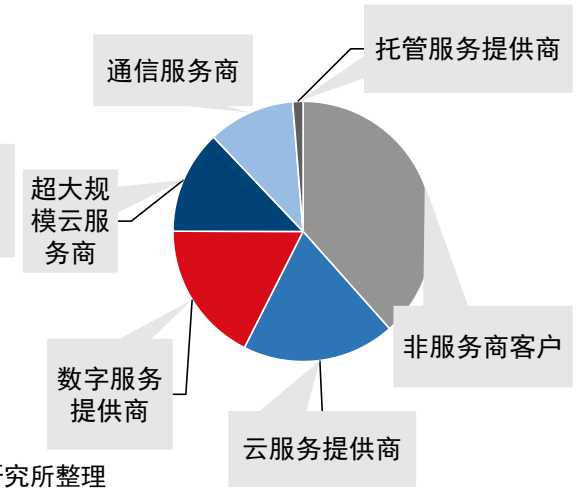


客户结构

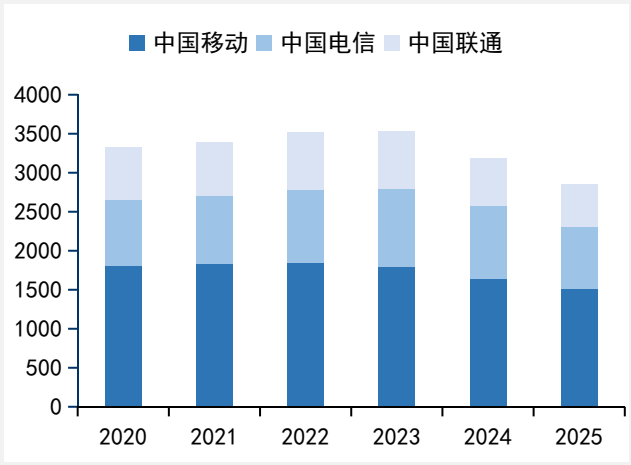
全球AI基础设施客户结构



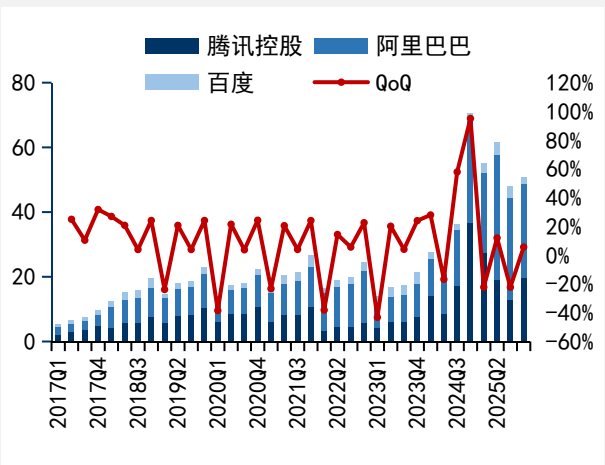
中国AI基础设施客户结构



运营商开支情况（亿美元）



互联网资本开支情况（十亿元）



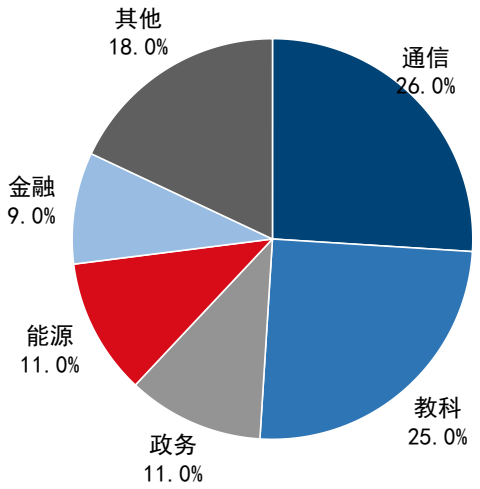
资料来源：IDC、公司官网，公司公告，国信证券经济研究所整理

信创行业注重AI算力基础设施安全可控

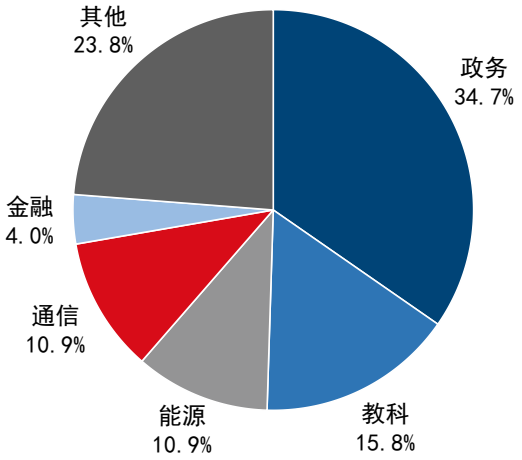
- AI大模型加速在各类信创新核心行业落地。根据亿欧智库，政务、金融、电信、能源、教育已经成为大模型落地的重点领域，业务中标金额和数量均有显著提升。需求侧看，大模型加速落地推动行业内信创应用软件产品能力升级，引入原生AI应用；从供给侧看，AI产品及服务供应商强调自研能力和AI产品与国产化软硬件适配能力，具备为传统行业提供全栈自研人工智能产品的业务能力。从中标项目看，通信、教科、政务位于前三位。人工智能发展推动国产化智能算力基础设施进化：以国产服务器为主的智算中心、智算集群、超级节点加速建设。
- 信创行业注重AI算力基础设施安全可控。2026年5月发布的《安全可靠测评结果公告（2026年第2号）》中将9款国产人工智能训练推理芯片纳入安全可靠等级Ⅰ级，首次在安全可靠认证框架下设立专门AI芯片品类，国产AI算力基础设施正式纳入国家信创安全认证体系。共有9款人工智能训练推理芯片通过安全可靠等级Ⅰ级认证，分别为华为海思昇腾310、昇腾910、平头哥真武M530、真武M890、壁仞科技壁砺™166、海光信息DCU-3G、天数智芯KCC-V100X、沐曦MXC600、摩尔线程PH100。

图：2024年大模型中标项目

2024年中国大模型中标项目数量行业分布



2024年中国大模型中标项目金额行业分布



资料来源：亿欧智库，国信证券经济研究所整理

图：安全可靠测评结果公告

序号	产品名称	送测单位	安全可靠等级
1	昇腾310	深圳市海思半导体有限公司	Ⅰ级
2	昇腾910	深圳市海思半导体有限公司	Ⅰ级
3	真武M530	平头哥（上海）半导体技术有限公司	Ⅰ级
4	真武M890	平头哥（上海）半导体技术有限公司	Ⅰ级
5	壁砺™ 166	上海壁仞科技股份有限公司	Ⅰ级
6	DCU-3G	海光信息技术股份有限公司	Ⅰ级
7	KCC-V100X芯片	上海天数智芯半导体股份有限公司	Ⅰ级
8	MXC600	沐曦集成电路（上海）股份有限公司	Ⅰ级
9	PH100	摩尔线程智能科技（北京）股份有限公司	Ⅰ级

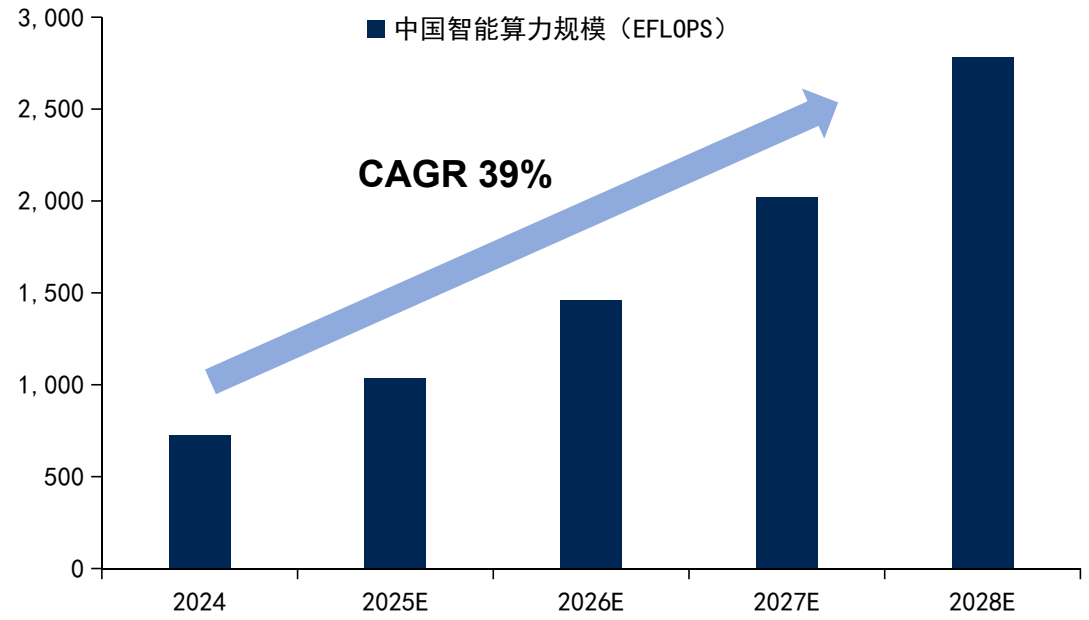
资料来源：中国信息安全测评中心，国信证券经济研究所整理

信创行业注重AI算力基础设施安全可控

国信证券 GUOSEN SECURITIES

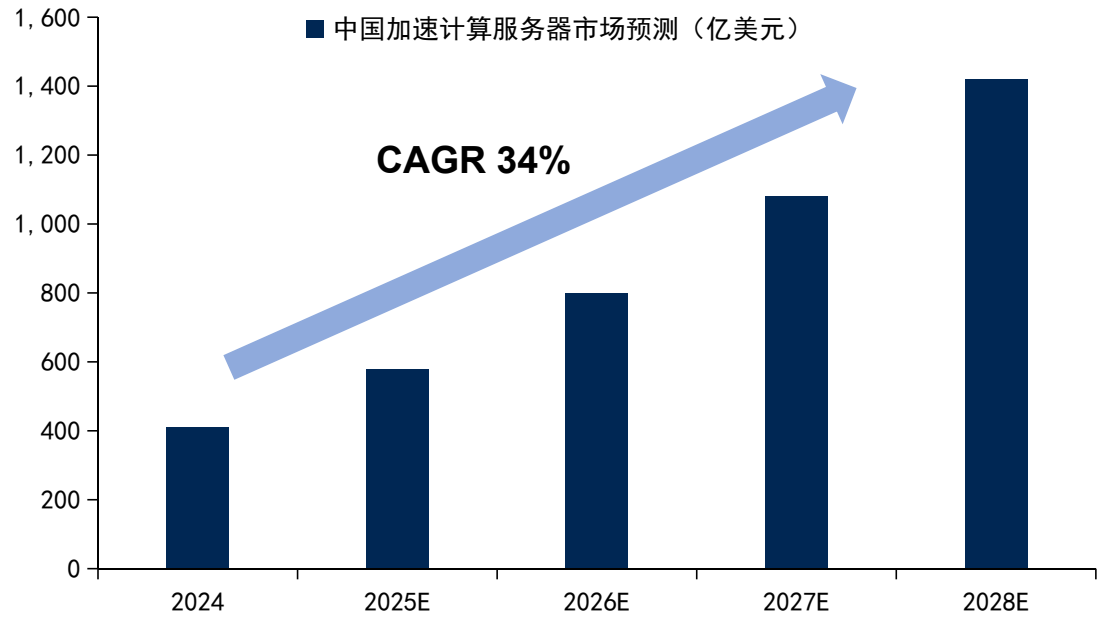
- 国内信创市场正从传统通用算力基础设施建设，逐步转向智能算力基础设施升级。过去信创的核心更多集中在CPU、服务器、操作系统、数据库等基础软硬件国产替代，目标是实现关键IT环节自主可控；而在大模型快速迭代、AI应用持续落地的背景下，各行业对训练、推理、数据处理和模型部署的需求显著提升，智能算力正成为信创产业升级的新重点。
- 加速计算服务器作为智能算力的重要载体，正在支撑本土大模型研发、行业模型训练以及AI原生应用运行。一方面，其带动国产AI芯片、服务器、存储、网络、散热等硬件环节升级；另一方面，也推动操作系统、数据库、中间件、AI框架、推理引擎等软件生态适配。未来信创建设将不再只是“可用替代”，而是向“高性能、可扩展、生态协同”的国产智能算力体系演进。

图：中国智能算力规模



资料来源：亿欧智库，国信证券经济研究所整理

图：中国加速计算服务器市场预测（亿美元）



资料来源：亿欧智库，国信证券经济研究所整理

国内本土芯片厂商份额持续扩大

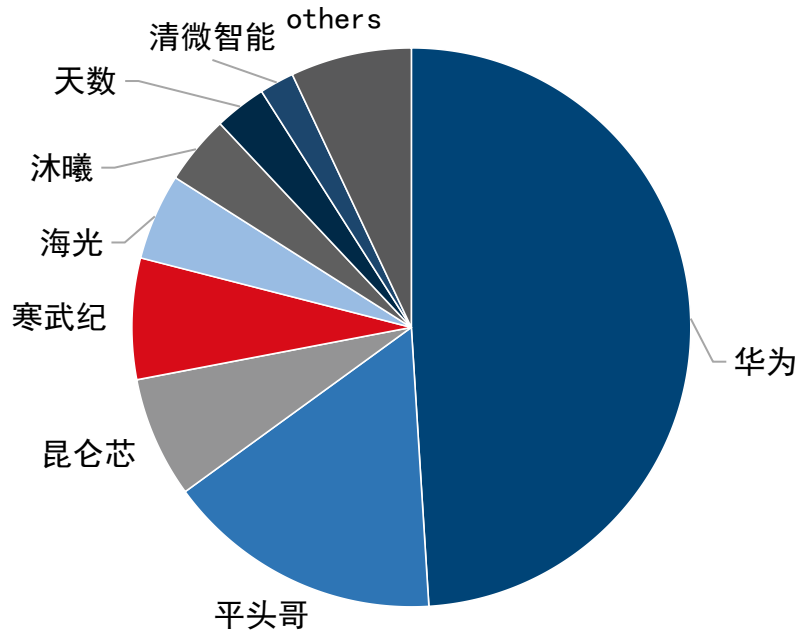
- 英伟达高端AI芯片销售受限，国内本土AI芯片厂商份额持续扩大。据IDC数据，2025年中国AI加速卡总出货量约400万张，其中英伟达占55.2%，华为占20.4%，平头哥占6.4%，AMD占3.5%，寒武纪占3.0%。国产品牌中，华为、平头哥、昆仑芯、寒武纪、海光信息位于前五位。国产算力的焦点不只是单卡峰值，而是“芯片 + HBM + 互联 + 服务器 + 编译器/算子库 + 推理引擎 + 模型适配”的全栈效率，产业生态完成从硬件主导向软硬件协同的转变。云端算力芯片侧重优化浮点运算单元、多芯片互联能力，针对大模型运算需求做深度定制；终端与边缘芯片则聚焦算力与功耗之间优化。

图：各厂商AI芯片产品布局

厂商	AI芯片/产品类型
华为昇腾	昇腾AI处理器；Atlas模块/板卡/服务器/集群；训练+推理
阿里平头哥	含光800推理NPU；真武M890/810E训推芯片；倚天710服务器CPU等
百度昆仑芯	昆仑芯1/2/3代；K/R系列加速卡；R480-X8加速器组
海光	深算DCU系列
寒武纪	思元MLU加速卡；边缘智能芯片；终端IP
沐曦	曦云C：训练/通用计算；曦思N：推理；曦彩G：图形渲染；曦索X：科学智能GPU；服务器；工作站；超节点
天数智芯	天垓：训练GPU；智铠：通用GPU推理产品
摩尔线程	全功能GPU/显卡：MTT S2000/S3000/S4000/S5000智算卡；模组/服务器/集群

资料来源：公司官网，国信证券经济研究所整理

图：2025年中国AI加速卡国产厂商出货量分布情况



资料来源：IDC，国信证券经济研究所整理

风险提示

一、国产替代进程不及预期。国内半导体企业相比海外半导体大厂起步较晚，在技术和人才等方面存在差距，在国产替代过程中产品研发和客户导入进程可能不及预期。

二、下游需求不及预期。

三、行业竞争加剧的风险。在政策和资本支持下，国内半导体企业数量较多，在部分细分市场可能出现竞争加剧的风险，从而影响企业盈利能力。

四、国际关系发生不利变化的风险。我国半导体产业链在部分环节需要依赖海外厂商，若未来国际关系发生不利变化，可能对半导体产业链运营产生重大影响。

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。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。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032