泽平宏观
ZEPING MACRO

AIDC 研究报告 2025

泽平宏观研究报告

行业研究　　专题报告　　2025/6/8

首席经济学家： 任泽平

研究员：喻楷文

导读：

本文为新基建、新能源研究系列。

随着 Deepseek、o 系列、Grok 等大模型不断升级，全球人工智能产业迎来了巨大变革。2025 年初，幻方发布的 Deepseek-R1 大模型，以极具竞争力的成本实现了相当的卓越性能，这极大地激发了国内 AI 产业的投资热情，**AIDC（人工智能数据中心）产业由此进入黄金发展期。**

AIDC 是传统 IDC 在 AI 算力驱动下的升级形态，核心是提供 AI 所需算力、数据和算法服务，堪称智能时代的“算力工厂”。与 IDC 相比，AIDC 在算力密度、散热技术上差异显著，以满足 AI 高算力、高稳定性需求。政策端，国家算力摸底推动资源向高效头部企业集中，加速中小厂商出清。

本轮 AIDC 的发展浪潮主要围绕两条核心主线展开：

主线一：训练端、推理端算力需求双重爆发。

一方面，Deepseek-R1 等大模型的涌现显著降低 AI 发展门槛，引发算力需求激增，形成“**效率提升-成本下降-需求扩张**”的杰文斯悖论循环。

另一方面，AI 技术正从 GenAI 向 Agentic AI、Physical AI 演进，Agent 智能体在编程、智能驾驶、A2A 等复杂场景的应用正快速普及。

展望未来，人形机器人的落地以及 AI for Science 的深入应用，将带来可观的增量算力需求。同时，多模态大模型驱动的现象级应用（如 AI 视频生成、对话）亦值得重点关注，其爆发式增长将显著拉动算力消耗。

主线二：全球云厂商掀起资本开支军备竞赛。

云厂商作为支撑 AI 应用的关键载体，资本开支直接决定了算力供给与技术落地的进程。从海外云厂商资本开支看，头部四家合计超 3000 亿美元，同比增长超 30%。从中国云厂商资本开支看，阿里、腾讯、百度等国内厂商的资本开支正在提速。

2025 年，在 Deepseek 推动下，国内 AI 叙事显著升温，国内云厂商资本开支有望迎来强劲增长拐点。阿里巴巴集团 CEO 吴泳铭表示，在云和 AI 的基础设施投入预计将超越过去十年的总和。**这一战略表态可以被认为是国内 AI 投资趋势的首个催化剂，未来 AIDC 的投资建设将迎来新一轮增长浪潮，成为推动 AI 产业发展的重要基石。**

目录

图表目录

1 AIDC：AI 发展的“猛禽发动机”

AIDC（人工智能数据中心，Artificial Intelligence Data Center）是传统 IDC（互联网数据中心，Internet Data Center）在 AI 算力需求驱动下的升级形态，其核心是基于最新人工智能理论，采用领先的人工智能计算架构，提供人工智能应用所需算力服务、数据服务和算法服务的新型算力基础设施。**简言之，IDC 是数字经济的“通用仓储”，满足广泛数字化需求，而 AIDC 是智能时代的“算力工厂”。**

图表：数据中心 IDC 走向 AIDC

资料来源：《AI DC 白皮书》华为，泽平宏观

1.1 AIDC 历史：从 IDC 到 AIDC

回顾过去几十年的发展历程，数据中心正走向智算数据中心。

技术萌芽期（1990 年代）：基础设施转型的起点。随着 TCP/IP 协议的全球普及和万维网技术的突破，全球信息化基础设施进入转型期。国内早期的分布式数据处理节点开始聚合，形成现代数据中心的雏形。这一阶段以技术探索为主导，基础设施部署呈现密度快速提升、覆盖范围扩张的特征，虽然规模较小，但为后续发展奠定了网络协议和基础架构的技术基础，标志着中国数据中心产业的萌芽。

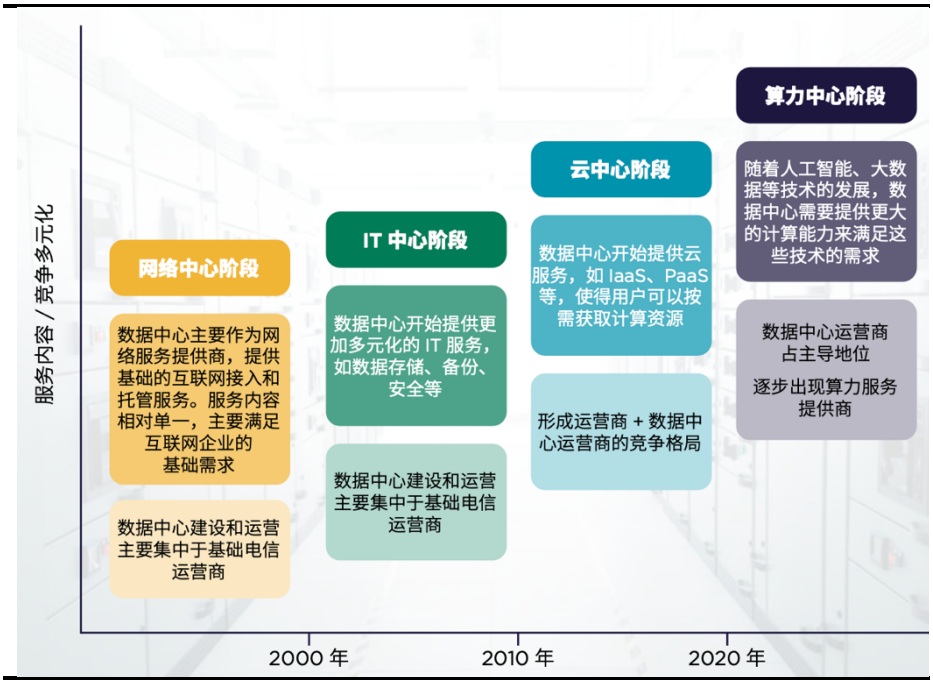
产业化培育期（2000-2010 年）：双轨模式的形成。进入 2000 年，中国信息化建设迎来黄金发展期，互联网应用从单一的门户网站向电商、社交等多元化场景拓展，推动基础设施服务标准升级。市场分化出两条清晰路径：企业级私有数据中心（EDC）满足大型企业个性化需求，第三方托管的互联网数据中心（IDC）开启商业化服务模式。此阶段以区域性分散部署的中小规模设施为主，初步构建起数字化经济的基础架构，形成"政企自建+第三方托管"的双轨发展格局，产业集中度较低但市场化进程加速。

云转型期（2010-2020 年）：集约化发展的变革。虚拟化技术的成熟引发 IT 资源供给模式革命，超大规模集群架构替代传统离散部署。行业竞争格局呈现三足鼎立：1）基础电信运营商依托网络资源优势布局基础设施；2）专业 IDC 服务商强化定制化服务能力；3）云服务巨头（如阿里云、腾讯云）通过技术创新引领行业方向。产业重心向 T3+以上高等级数据中心转移。

智能算力增长期（2020 年至今）：结构性升级的新周期。在 AI 技术革命与数据要素市场化的双重驱动下，数据中心产业发生结构性变革。需求端呈现两极分化：**超大规模数据中心**聚焦基础存储与通用计算，满足云计算、大数据等普惠需求；**异构算力中心**专注AI训练推理等专业场景，适配深度学习、大模型训练等高算力密度需求。具备全栈服务能力的第三方运营商凭借敏捷交付体系和技术中台优势快速扩张，行业集中度持续提升，标志着数据中心从"云化基础设施"向"智能算力枢纽"的战略升级。

四个阶段的演进本质上是技术驱动-需求升级-模式创新的螺旋上升过程。从早期技术导入形成产业雏形，到市场化驱动双轨发展，再到云化技术引发集约化变革，最终在AI和数据要素时代实现算力结构优化。每个阶段的核心矛盾不同，技术萌芽期解决"有没有"，产业化培育期解决"市场化"，云化转型期解决"效率提升"，**智能算力期解决"结构升级"。**

图表：AIDC 智算中心发展历程示意图



资料来源：《2024 数据中心行业投资与价值洞察》，泽平宏观

1.2 AIDC 变化：与 IDC 的差异哪？

AIDC 与 IDC 的本质差异源于 AI 算力需求对基础设施的重塑，其中有两大核心变化：

1）算力密度与功耗层面，IDC 以通用服务器为主，单机柜功率密度较低（4-8kW），而 AIDC 需部署高功率 GPU/TPU 服务器，单机柜功率达传统 IDC 的 5-10 倍（10-100kW 以上），硬件投入成本更高，但单位算力效率显著提升；

2）散热技术层面，IDC 多采用风冷技术，而 AIDC 因高功率密度需引入液冷方案（如冷板式或浸没式冷却），以降低 PUE 值（电能使用效率），同时满足长时间高负载运行的稳定性需求。根据英伟达，H200 在高负载任务下产生的热量较前代产品增加了约 30%，为确保其稳定运行和持续高性能输出，引入了液冷散热技术。实验对比显示，液冷散热效率较传统风冷提升了约 50%。

图表：IDC 通用数据中心与 AIDC 智算中心的差异

	IDC 通用数据中心	AIDC 智算中心
算力类型	提供通用算力（或称基础算力），以用于数据存储和虚拟化、通用（基础）计算、大数据分析等	提供智能算力，以用于人工智能的训练和推理计算，语言、图像和视频的智能处理
芯片	以 CPU 为中心	以 xPU 为中心
技术架构	采用冯·诺依曼的主从架构	采用更加先进的全互联对等架构
散热模式	一般采用传统的风冷散热。	主要采用液冷或风液混合的散热技术。
应用场景	电子商务平台、社交媒体、即时通讯、音视频与流媒体平台等	自动驾驶、科研计算、生成式 AI 智能语言模型、公共安全系统、智慧交通等
机柜规格	普遍在 2-10KW	12KW-24KW 或以上
市场玩家	万国数据、世纪互联、数据港、润泽科技、光环新网等	商汤科技、紫光集团、科大讯飞、浪潮集团、华为等

资料来源：《AI DC 白皮书》华为，《2024 数据中心行业投资与价值洞察》，泽平宏观

1.3 AIDC 用在哪：AIDC 主要用于 AI 模型的训推用

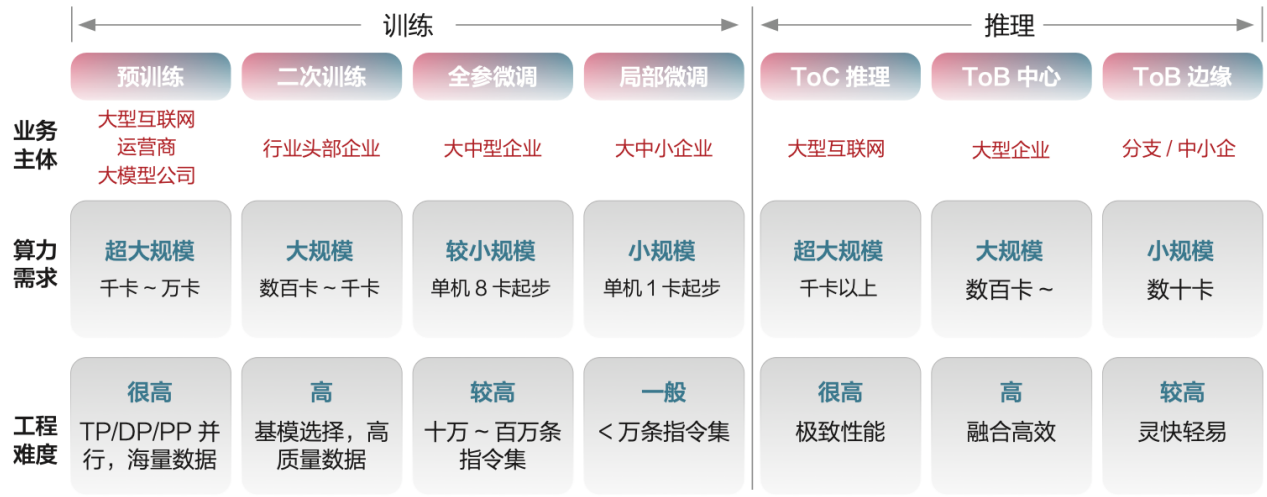
AI 模型的全面应用，是从训练到推理多环节紧密协作的过程。这个过程包括基础模型预训练、行业或企业模型的二次训练以及场景模型的微调，最终实现模型在实际环境中的部署与推理应用。**AIDC 最主要的是要围绕 AI 模型训练、推理和应用来规划设计和实施。**

基础模型预训练：大型互联网企业与专注大模型研发的公司，以构建通用基础模型为核心目标，其 AIDC 建设需打造具备十万甚至百万量级算力卡的超大规模集群平台。这类企业在训练过程中需处理万亿级 Token 数据，涵盖文本、图像、音视频等多模态信息，以实现模型对通用知识的深度学习。

行业模型二次训练：行业头部企业基于通用基础模型，叠加行业专属数据进行二次训练，以构建适配金融、医疗、制造等垂直领域的行业模型。此类训练虽数据规模降至数亿级 Token，但仍需数百至数千张 NPU/GPU 算力卡支撑，且需解决行业数据的合规处理、特征提取及模型参数优化问题。

模型微调与推理：多数企业将 AIDC 作为模型微调与推理的核心平台，结合自身业务场景数据对基础模型或行业模型进行针对性优化，使其满足客户服务、智能决策、自动化生产等具体需求。推理环节对 AIDC 的性能指标提出精细要求：面向个人用户的 ToC 服务需降低延迟以提升交互体验，面向企业客户的 ToB 服务强调高并发处理能力与稳定性，而企业内部应用则更注重算力使用效率与数据安全性。

图表：不同场景训练推理的算力需求及工程难度



资料来源：《AI DC 白皮书》华为，泽平宏观

1.4 AIDC 政策端：算力摸底减少市场低效供给

国家层面算力摸底，减少市场低效供给。根据财联社，4 月 16 日，地方发改委**《关于开展算力摸底有关工作的通知》**已下发，通知显示，摸底工作涉及已建、在建和拟建算力中心项目，摸底数据将作为国家算力资源统筹布局的重要依据。根据 IDC 圈统计，2025 年一季度，我国大陆共 165 个智算中心项目出现新动态，分布在 29 个省份，除 37 个未披露投资额的项目外，**128 个项目总规划投资总额超过 4300 亿元**。项目状态方面，58%的项目处于已审批筹建状态、33%处于在建或即将投产状态、仅 10%处于已投产/试运行状态。

根据《全国数据资源调查报告（2024 年）》，**2024 年，全国算力总规模达到 280EFLOPS（每秒百亿亿次浮点运算），智能算力规模达 90EFLOPS，占比提升至 32%，为海量数据计算提供智能底座。**其中，中央企业算力规模增长近 3 倍，智能算力占比为 40.22%；数据技术企业算力规模同比增长近 1 倍，智能算力占比为 43.63%。根据 IDC 预计，2025 年中国智能算力规模将达到 1,037.3EFLOPS，2028 年将达到 2,781.9EFLOPS，2023-2028 年中国智能算力规模和通用算力规模的五年年复合增长率分别达 46.2%和 18.8%。

回顾历史，本次算力摸底可以认为是此前政策延续。自 2020 年《关于加快构建全国一体化大数据中心协同创新体系的指导意见》发布以来，国家持续推动算力资源统筹，清退“小散老旧”数据中心，2022 年进一步明确 PUE≤1.3 的能效标准，淘汰高耗能低效产能。此次摸底将细化到“已建、在建、拟建”算力中心，动态掌握算力规模（如智算/通算占比）、利用率、能耗等指标，为国家层面统筹东西部资源匹配提供数据支撑。

我国数据中心投资主体主要包括互联网大厂、运营商、地方政府、第三方 IDC 厂商。算力摸底政策是国家层面推动算力基础设施高质量发展的关键一步，通过数据驱动的精准调控，将有效遏制盲目建设，引导资源向技术领先、运营高效的头部企业集中。

短期内可能加剧中小厂商出清，但长期看，龙头第三方 IDC 厂商和具备全栈能力的科技企业将主导市场，推动中国算力产业从“规模扩

张”转向“价值提升”。在此背景下，我们认为未来环一线城市、东数西算枢纽 AIDC 稀缺性凸显，具备大量优质资源储备的龙头企业将迎来巨大的 AIDC 盈利优势。

2 中国进入“AI+”叙事期，云厂商推动AIDC建设

2.1 Deepseek 加速 AI 平权，大模型爆发带动算力需求

2025 年初，Deepseek 横空出世，带动国内大语言模型爆发。DeepSeek 的引人注目之处在于它克服了计算能力的限制。DeepSeek 专注于算法效率，而非单纯的计算能力竞争。Deepseek 的出现有效降低了中国 AI 发展的门槛，中国突破美国的科技制裁限制，进而加速 AI 平权，使得算力需求在中国爆发。

2025 年 5 月，Deepseek 团队的新论文《Inference-Time Scaling for Generalist Reward Modeling》发布，引入了一种自我原则点评调优（SPCT）的方法，基于此方法推出 Deepseek GRM 模型，27B 的参数能跑出目前 R1 模型 671B 参数相当的性能。

从实验结果来看，Deepseek GRM 模型进一步压缩的硬件需求，采用 128 块 A100-80G GPU 训练，训练成本仅仅为 R1 的 1/6；推理阶段无需长链式推理的重复计算，降低了算力与显存的需求（GRM 模型全精度显存需求 108GB，R1 满血版模型显存需求 1300GB 以上），推理能耗为 R1 模型的 17%左右，大大降低了模型本地化部署的成本。Deepseek R2 有望在近期内发布，**此次 GRM 模型的发布或是其算法创新的雏形，Deepseek R2 的发布或将加速 AI 在各行各业普及。**

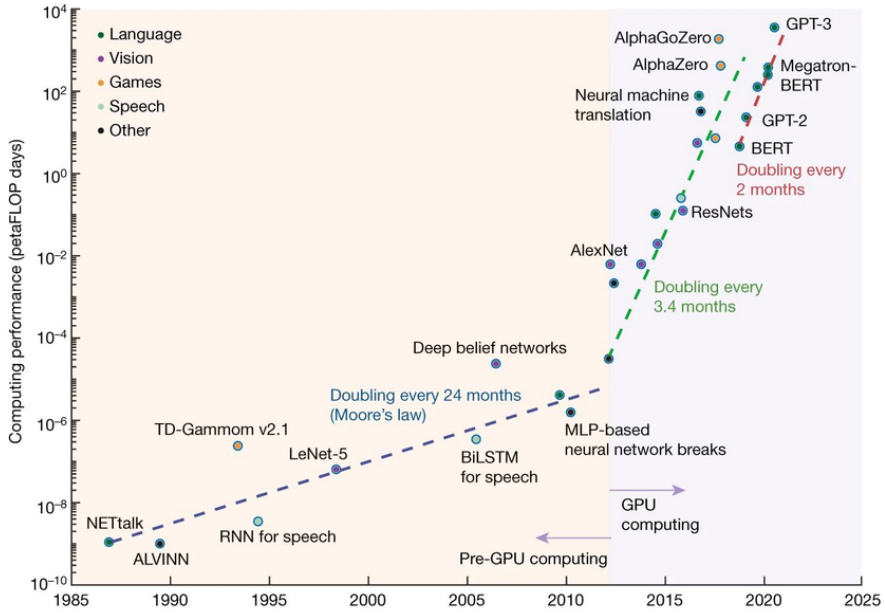
效率提升→成本下降→应用普及→需求激增→资源紧张，杰文斯悖论式的循环，在 AI 行业正在兑现。

杰文斯在《煤炭问题》一书中提出杰文斯悖论（Jevons paradox），其核心内涵在于，技术进步带来的资源利用效率提升未必会降低资源总消耗量，反而可能促使资源总需求呈现增长态势。其内在机制植根于需求端对成本变化的强敏感性。**当资源获取成本因效率提升而下降时，市场需求的扩张幅度往往会超越技术进步带来的节约效应，最终导致总消耗反增。**

以 19 世纪英国蒸汽机革命为例，通过改良燃烧效率使煤炭利用率大幅提升，**非但没有减少煤炭消耗，反而因工业生产规模的指数级扩张，**催生了冶铁、纺织、运输等多领域的用煤需求爆发，最终推动煤炭总消耗量迎来跨越式增长，**我们认为 AI 产业对于算力需求亦是如此。**

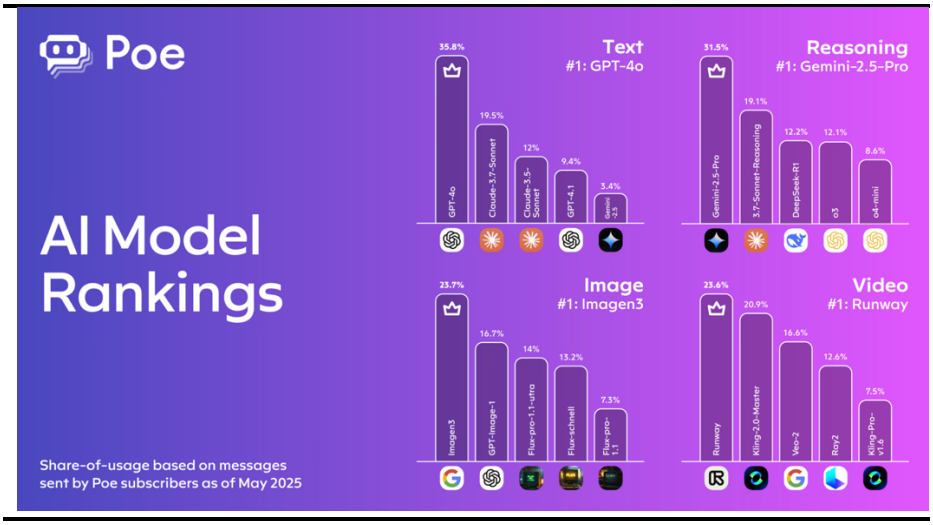
除此之外，我们认为训练端算力需求，后续值得期待。从近期的大模型迭代趋势来看，预训练（Pre-training）阶段的 Scaling Law 的放缓趋势。但我们认为预训练的算力需求仍有上升空间。**今年下半年全球大模型的迭代，会重回预训练上来，无论是模型参数的增加，数据集的增加，还是模型优化，如 GPT-5、Grok4、Claude 4 等。同时后训练（Post-training）的算力需求也值得关注。**

图表：不同阶段算力需求增速



资料来源：Sevilla et al.,2022;Amodei and Hernandez,2018，泽平宏观

图表：OpenAI 总份额持续提升

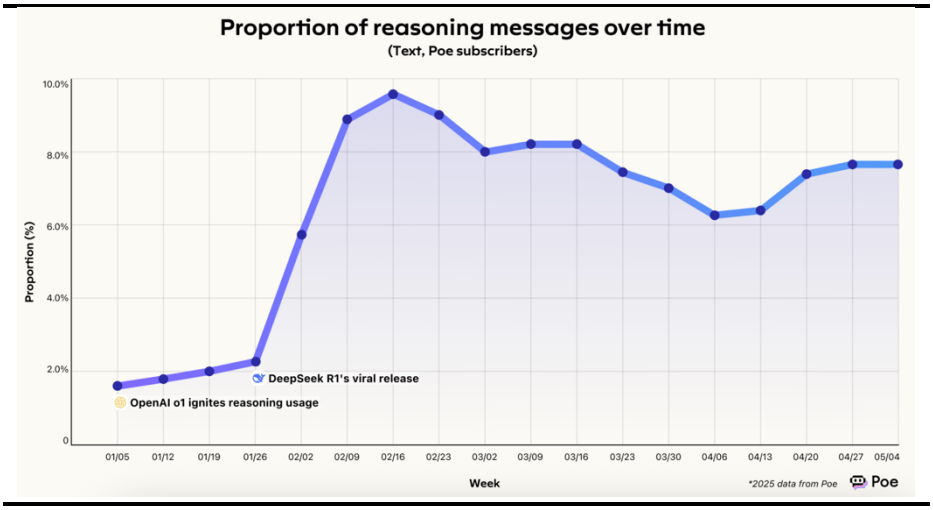


资料来源：POE：Spring 2025 AI Model Usage Trends，泽平宏观

图表：算力需求测算公式（FLOPs）

资料来源：OpenAI《Scaling Laws for Neural Language Models》，泽平宏观

图表：Deepseek 火爆以来，推理模型的使用率一直居高不下

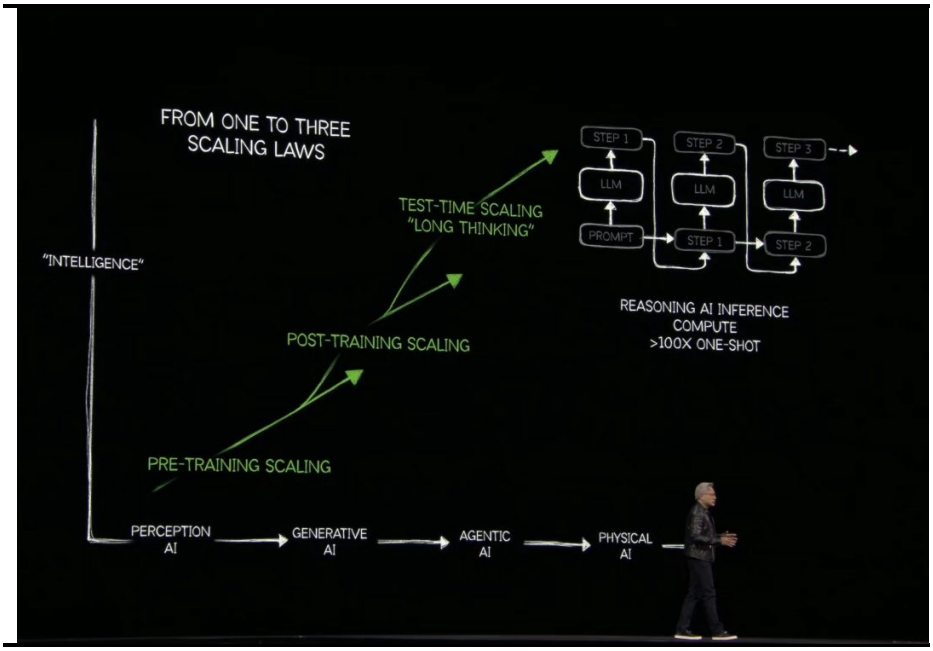


资料来源：POE：Spring 2025 AI Model Usage Trends，泽平宏观

2.2 新兴场景：Agent、智能驾驶元年，释放增量算力需求

AI 产业正从 AI Chatbot 转向 AI Agent。从 AlexNet 到 ChatGPT，是从检索的计算方式转变为生成的计算方式。而当 AI 从 ChatGPT 那种靠预测下一个 tokens、大概率出现幻觉的生成式 AI，**迈向 Deep Research、Manus 这样的 agentic AI 应用时，每一层计算都不同，所需要的 tokens 比想象的多 100 倍。**因为在 Agentic AI 应用中，上一个 token 是下一个 token 生成时输入的上下文、是感知、规划、行动的一步步推理。

图表：相较 Generative AI，Agentic AI 带来百倍级的计算量增长



资料来源：GTC 2025 英伟达发布会，泽平宏观

回顾 AI Agent 的产业发展，AI Agent 的发展轨迹恰似积木搭建的过程，需要将分散的技术模块精准拼接，方能构建完整的智能体形态。早期阶段，大模型智能性、多模态推理、代码生成能力、工具调用机制、Token 经济体系及算力支撑等核心模块长期处于孤立发展状态，单一技术因缺乏协同闭环而难以转化为实际行动能力，**导致 Agent 始终停留在“智能理论超前、落地执行滞后”的功能割裂阶段。**

当下产业环境已发生关键转变：以MCP为代表的工具调用协议完成标准化建构，为跨工具交互提供了统一技术语言；大模型代码生成能力突破产业级应用标准，能够支撑复杂业务逻辑的自动化实现；而 Token 调用成本的指数级下降，更从经济层面破除了大规模智能交互的落地壁垒。在技术标准统一化、开发框架工程化、核心能力实用化、应用成本亲民化的多重驱动下，**曾经散落的技术积木正按照产业级协同标准完成有机拼接，推动 AI Agent 从功能碎片迈向具备完整行动能力的智能体闭环，AI Agent 进入爆发元年。**

我们看来，Agent 发展也存在预期差，Agent 对于普罗大众来说可能感知不到，但它在特定场景、特定人群的使用量增长尤为迅速，比如 Coding Agent、Manus、智能招聘 Agent 等。Manus 使用客群大多集中在海外，因此实际增速可能比感知的要快一些。

进一步来看，大模型自身能力决定着，AI Agent 的智能高度和应用边界。随着它变得越来越智能、应用越来越广泛，更多的算力需求也随之而来。

图表：Coding Agent 打开大模型全新的想象空间

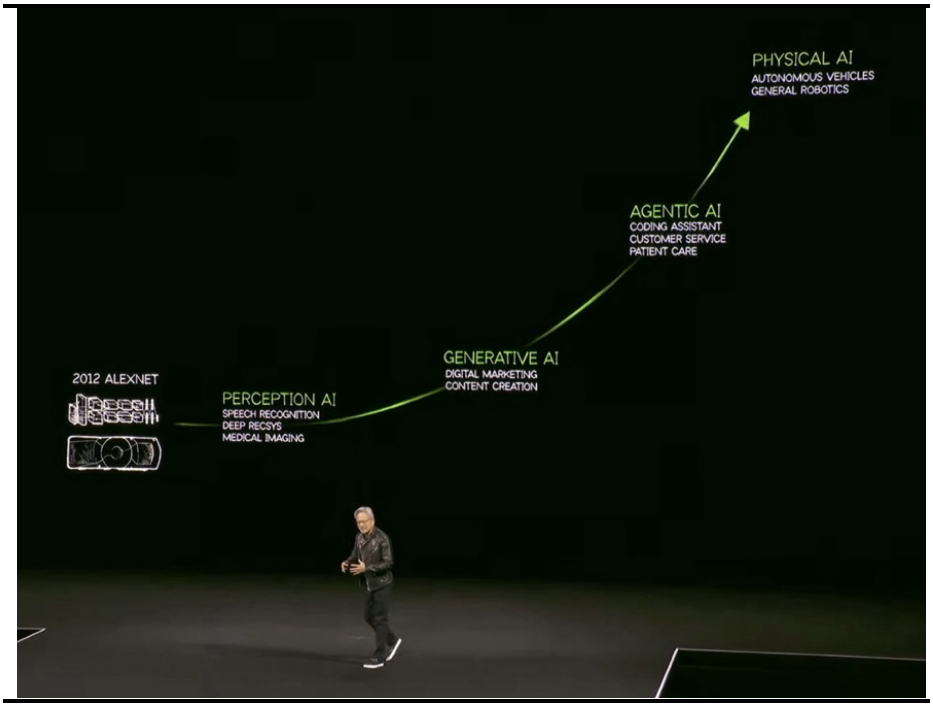
资料来源：GitHub，泽平宏观

图表：通用 AI Agent Manus 点燃中国 Agent 热情

资料来源：Manus 官网，泽平宏观

Physical AI 拥有更多的算力需求。AI 已经经历了三代技术范式的转移。最早是判别式 AI（语音识别、图像识别），接着是生成式 AI，然后就是当下我们身处的 Agentic AI，未来会是影响物理世界的 Physical AI，也就是智能驾驶、人形机器人。

图表：AI 正经历从 Agentic AI 到 Physical AI 的快速演进



资料来源：GTC 2025 英伟达发布会，泽平宏观

为了训练“全自动驾驶”神经网络，DOJO 2 性能预计指数级增长。特斯拉 DOJO 总计算能力或超 100exaflops，超 10 万英伟达 H100/H200 AI 芯片提供算力支持，旨在训练其“全自动驾驶”神经网络。特斯拉通过全球超 200 万辆车辆实时回传数据，构建了全球最大的自动驾驶数据库。依托 Dojo 超算中心，每天完成海量数据训练。Dojo 2 将于 2026 年发布，特斯拉希望将其计算机容量提高 10 倍。

国内新势力自建智算中心快速追赶应对算力挑战。理想汽车的理想智算中心算力已达到 8100 PFLOPS，华为车 BU 云智算中心的乾崑 ADS3.0 在算力方面已达到 7500 PFLOPS。2025 年小鹏云端的算力将会达到 1000 PFLOPS 以上。

对于后续的算力增量需求，人形机器人、AI for Science 的落地带来的算力需求值得期待。还需关注多模态大模型带来的现象级 AI 应用的算力需求，如 AI 视频对话等。

图表：国内车企与智驾供应商算力中心算力汇总（截止 2025 年初）

企业类型	企业	智算中心	形式	最新算力披露
车企	特斯拉	Dojo 智算中心	自建	100000PFLOPS
车企	小米	金米云智算中心	合作金米云	11450PFLOPS
车企	理想	理想智算中心	合作火山引擎	8100PFLOPS
车企	蔚来	“扶摇”智算中心	合作阿里云	2510PFLOPS
车企	吉利	吉利星睿智算中心	合作阿里云	1020PFLOPS
智驾供应商	商汤绝影	商汤绝影智算中心	自建	20000PFLOPS
智驾供应商	华为	车 BU 云智算中心	自建	7500PFLOPS
智驾供应商	Apollo	百度 Apollo 智算中心	共建	5500PFLOPS

资料来源：盖世汽车，汽车之心，小米公众号，36 氪，泽平宏观

2.3 全球云厂商资本开支军备竞赛，中国 AI 叙事加快 AIDC 建设

云厂商作为支撑 AI 应用的关键载体，资本开支（Capex）直接决定了算力供给与技术落地的进程。

2022 年底，OpenAI 正式推出 ChatGPT，成为生成式 AI 爆发的导火索，迅速点燃全球 AI 技术革命的热潮，由此为分水岭，进入 AI 大模型时代。自 2023 年起，北美科技巨头纷纷将 AI 列为战略核心，通过大幅增加资本开支，加速 AIDC 建设。微软、亚马逊、谷歌与 Meta 的资本投入呈现显著增长态势，在 AI 领域的布局力度空前。

相较之下，中国企业在 AI 发展初期面临双重挑战：一方面，高端 AI 算力芯片的进口受到限制；另一方面，国内头部企业在生成式大模型的技术研发上进度相对滞后。中国企业虽然在 AI 领域的资本开支增速较快，但投入的绝对值仍处于较低水平。

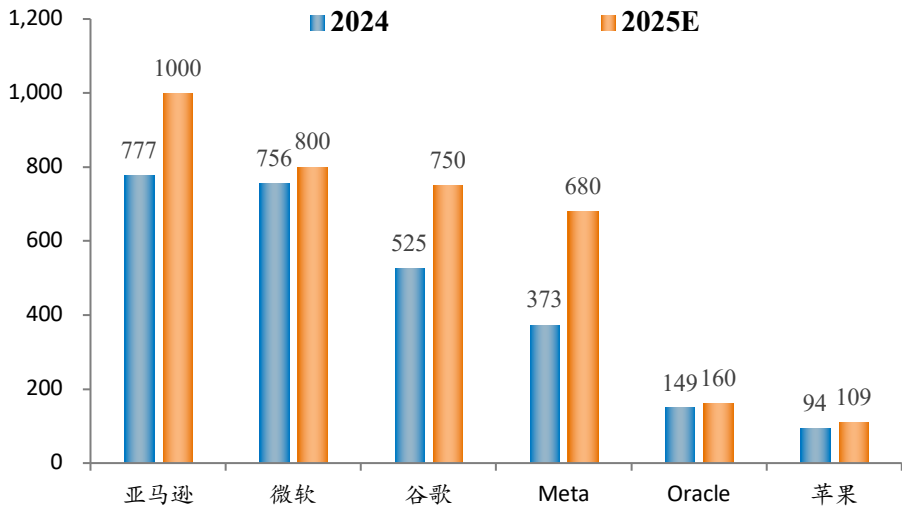
转折点出现在 2025 年 1 月，幻方发布的 Deepseek R1 大模型，以极具竞争力的成本实现了与 ChatGPT-o1 相当的卓越性能，由此国内芯片、模型、应用形成闭环，这极大地激发了国内 AI 产业的投资热情。**由此开始，全球科技开始分裂为两套制度，一个由美国 ChatGPT 主导，另一个则是中国 Deepseek 主导。**

我们有理由相信，在中美 AI 军备竞赛背景下，未来 AIDC 的投资建设将迎来新一轮增长浪潮，成为推动 AI 产业发展的重要基石。

从海外云厂商资本开支看，头部四家合计超 3000 亿美元，同比增长超 30%。Meta 预计将 2025 年的资本开支规划由之前的 600-650 亿美元提升到 640-720 亿美元，用于加大数据中心投资来支持 AI 发展以及要增加基础设施硬件的投资。微软预计 2025 财年投入 800 亿美元用于 AI 数据中心建设；亚马逊计划 2025 年投资超过 1000 亿美元，同比增速约 29%，

主要用于 AI 基础设施，以支持其云计算服务和 AI 服务的需求；谷歌预计 2025 年资本开支达 750 亿美元，同比增速约 43%，主要用于 AI 基础设施的建设。再如日本软银集团、OpenAI 和美国甲骨文公司三家企业联合在 2025 年宣布的星际之门计划，未来 4 年将投资 5000 亿美元。

图表：北美主要云厂商资本开支（亿美元）



资料来源：公司公告，公司财报电话会，泽平宏观

图表：北美四大 CSP2025 年资本支出指引乐观

公司	CAPEX 相关表述
亚马逊	亚马逊计划在 2025 年投资超过 1000 亿美元，同比增速约 29%，主要用于 AI 基础设施，以支持其云计算服务和 AI 服务的需求
微软	2025 年 2 月，据国际金融报报道，微软预计将在 2025 财年投入 800 亿美元用于 AI 数据中心建设，其中一半以上的支出将聚焦于美国市场。2026 财年，鉴于强劲需求信号，包括微软云客户合同积压订单，预计会继续投资，但增长率将低于 2025 财年，支出结构将开始转回与收入增长关联度更高的短期资产。
谷歌	随着加大人工智能领域投入，预计增加技术基础设施方面的资本支出投资，主要用于服务器，其次是数据中心和网络。预计 2025 年资本支出约为 750 亿美元，同比增速 43%，其中第一季度约为 160 亿至 180 亿美元。
Meta	预计 2025 年全年资本支出从 600-650 亿美元提升至 640-720 亿，主要是由于要加大数据中心投资来支持 AI 发展以及要增加基础设施硬件的投资。预计 2025 年资本支出增长由加大投资推动，以支持在生成式人工智能领域的工作以及核心业务。

资料来源：各公司业绩说明会，泽平宏观

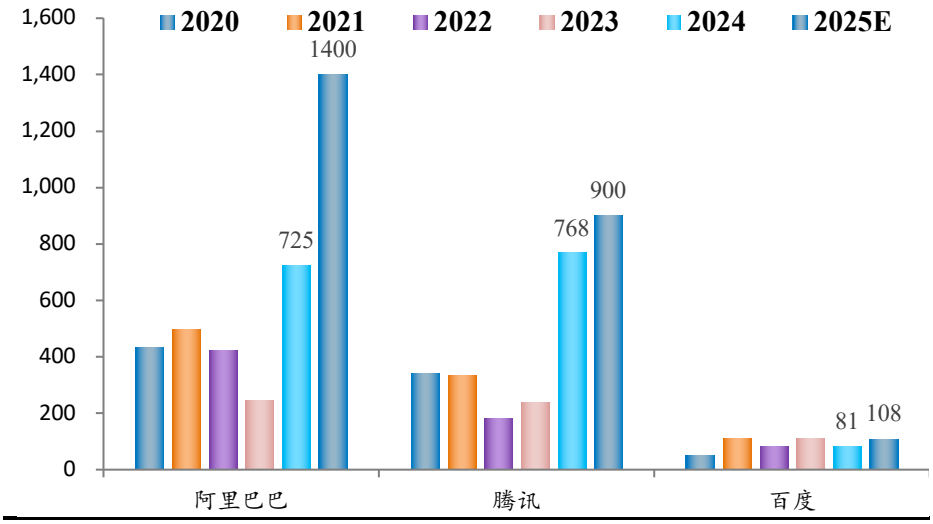
从中国云厂商资本开支看，与海外相比，阿里、腾讯、百度等国内厂商的资本开支高速增长阶段相对滞后于海外，存在错期。2025 年在 Deepseek 推动下开启 AI 叙事，国内云厂商资本开支有望大幅增长。

据新华社报道“阿里巴巴集团 CEO 吴泳铭 24 日宣布，**未来三年，阿里将投入超过 3800 亿元，用于建设云和 AI 硬件基础设施，总额超过去十年总和**”，意味着年均资本开支 1300 亿元。根据阿里巴巴财报，2024 年资本开支总额为 767 亿元，预计 2025 年的资本开支**增量**约 1300-767=533 亿元。

2025 年三大运营商资本开支计划合计达到 2898 亿元，但额度“不设限”。中国电信 2025 年预计算力投资同比增长 22%。中国电信同时表示，今年算力投资将根据需求灵活调整不设限。中国移动方面在算力领域投资 373 亿元，占资本开支的比例提升到 25%，对于推理资源将根据市场需求进行投资，不设上限。中国联通预计算力投资同比增长 28%，预算安排将根据智算和 6G 等需求，以及国内外发展趋势，及时调整投资规模。

字节跳动算力投入明显加大。2024 年的资本开支达为 800 亿人民币，2025 年有望达到 1600 亿元，大部分投向算力采购和 IDC 基础设施建设。根据 Omdia，2024 年字节和腾讯预计采购英伟达 H 系列算力卡合计约 46 万块，分别均采购了约 23 万块，仅次于微软采购量。

图表：中国主要云厂商资本开支（亿元）



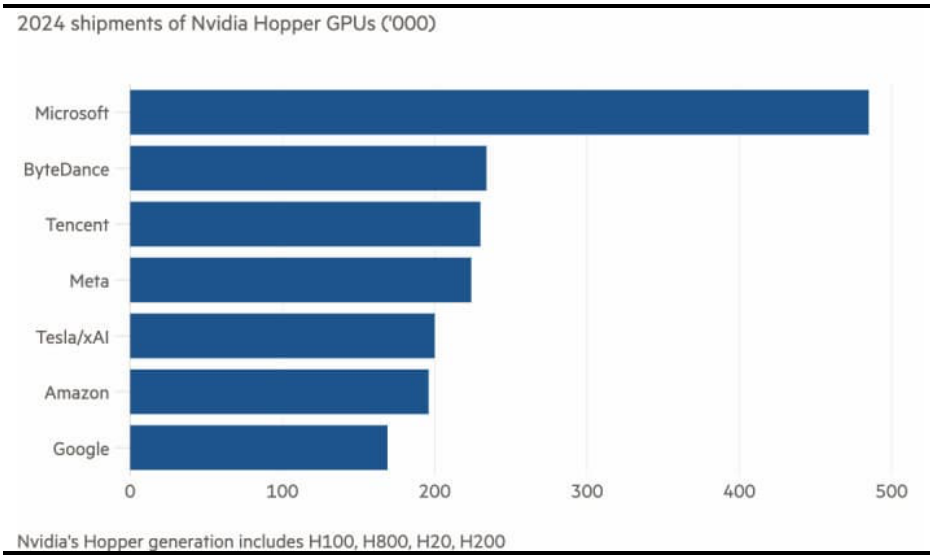
资料来源：公司公告，公司财报电话会，泽平宏观

图表：中国主要 CSP2025 年资本支出指引

公司	CAPEX 相关表述
字节跳动	计划在 2024 年资本开支达到 800 亿元，**2025 年增至 1600 亿元**，主要用于 AI 算力采购和数据中心建设。其中约 900 亿人民币将用于 AI 算力的采购，国内计划投入 400 亿人民币，国外（主要是东南亚地区）投入 500 亿人民币，以构建强大的 AI 计算能力基础；其余 700 亿人民币则分配给 IDC 基建以及网络设备如光模块、交换机的招标，国内 500 亿人民币，国外 200 亿人民币，旨在打造自主可控的大规模数据中心集群。
腾讯	在研发投入方面，腾讯 2024 年投入达 706.9 亿元，2018 年至今累计投入 3403 亿元。刘炽平表示“腾讯计划 2025 年进一步加大资本开支，预计会占 2025 年总收入的‘低两位数百分比’。”假设 2025 年的收入增速保持 8%，按照低两位数百分比（通常为 10%-20%）的中值 15%计算，2025 年的资本开支预计约 1070 亿元人民币。**2025 年第一季度资本开支 274.8 亿元，同比增长 91%，主要用于 GPU 及 AI 算力基础设施建设。**
阿里巴巴	2025 年第一季度公司资本开支 246.12 亿元，同比增长 121%，持续高增，主要用于 AI 算力基础设施及数据中心建设。据华尔街见闻报道，阿里巴巴集团 CEO 吴泳铭在 2 月 24 日宣布，**未来三年，阿里将投入超过 3800 亿元，用于建设云和 AI 硬件基础设施，总额超过去十年总和（约 3000 亿人民币）**。阿里巴巴官方表示将继续对客户和技术进行投入，尤其是在 AI 基建方面，以提升 AI 领域的云采用量，并维持市场领先地位。

资料来源：公司公告，Wind，泽平宏观

图表：2024 年字节采购英伟达 H 系列算力卡 20 万块以上

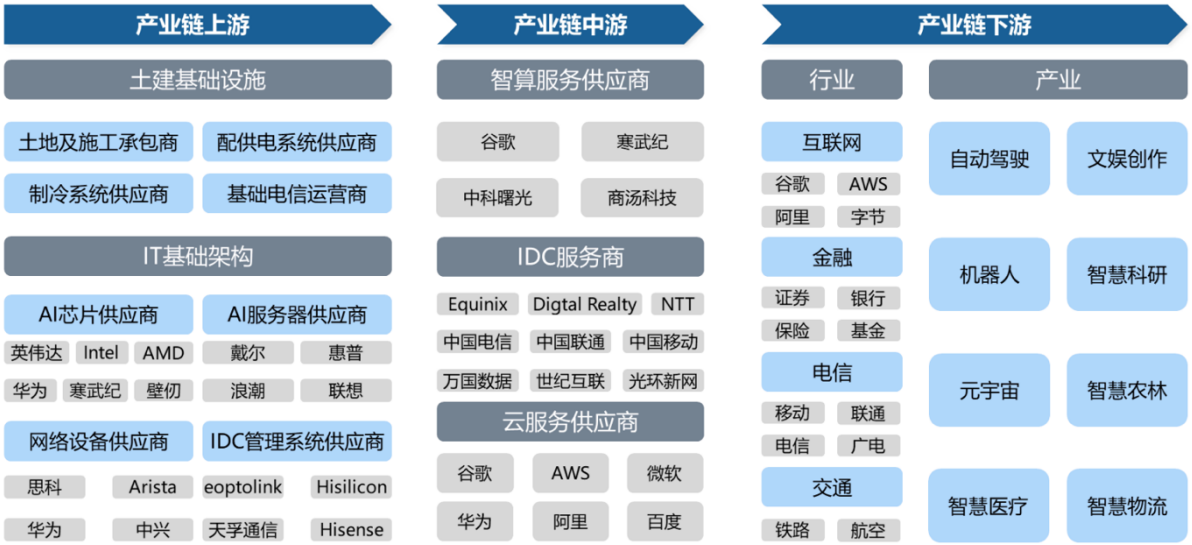


资料来源：Omdia，泽平宏观

3 AIDC 产业链：黄金发展新周期

智算中心产业链贯穿多个环节，从 AI 芯片、服务器等硬件的设计制造，到基础设施建设，再到智算服务供应，以及生成式 AI 大模型研发和行业应用。

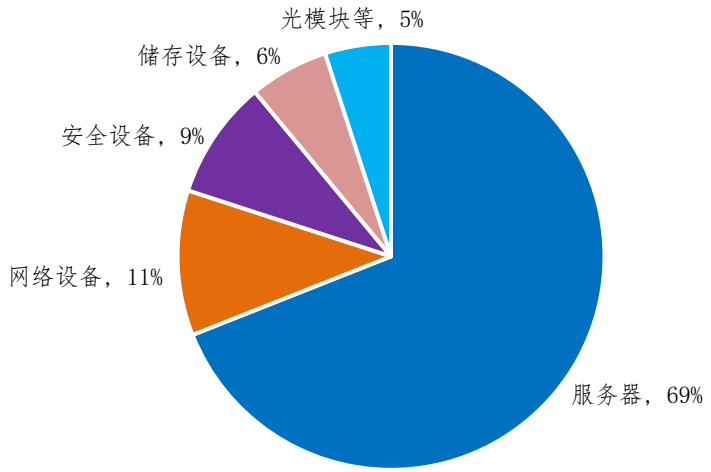
图表：AIDC 产业链上中下游示意图



资料来源：《中国智算中心产业发展白皮书》，泽平宏观

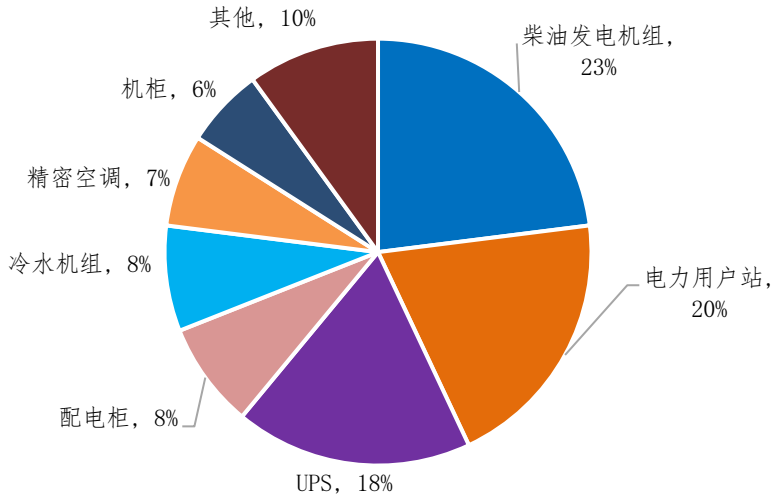
在产业链中，AIDC 资本开支大头仍在 IT 侧，服务器是算力承载的核心硬件。根据中商产业研究院，服务器约占 IT 侧成本的 70%；非 IT 侧主要为供电系统相关，其次为制冷系统相关，供电系统相关（柴油发电机组、电力用户站、UPS、配电柜）和制冷系统相关（冷水机组、精密空调、冷却塔）分别占 69%和 18%。

图表：数据中心 IT 设备成本占比



资料来源：中商产业研究院，泽平宏观

图表：数据中心非 IT 设备成本占比



资料来源：中商产业研究院，泽平宏观

3.1 AI 芯片：H20 封禁，国产加速替代

美国政府对英伟达“特供”中国市场的 AI 芯片 H20 实施出口管制，自 2025 年 4 月 14 日起无限期生效。此举导致英伟达预计 2026 财年计提 55 亿美元相关费用，并加剧国内市场对高性能算力芯片的供需矛盾。

国产替代“以量补质”实现性能赶超。2025 年 4 月，华为推出的 CloudMatrix 384 正式于芜湖数据中心上线，产品由 384 颗昇腾 901C 芯片组成，算力达 300PFLOPS。华为以超 5 倍于英伟达 GB200 NVL72 的芯片数量构建 AI 算力集群 CloudMatrix 384，实现了相当于该集群 1.7 倍的性能表现。尽管昇腾芯片单卡性能与 Blackwell 芯片仍有差距，但 CloudMatrix 集群性能已实现对英伟达 NVL72 的超越。

AI 大模型所仰赖的集群式算力特征，为国产 AI 算力产品孕育出全新发展契机。尽管现阶段中国大陆晶圆厂的先进制程工艺与国际一流厂商尚存差距，单颗 AI 芯片性能提升面临瓶颈，但依托高速网络设备实现多芯片算力堆叠，国产 AI 算力集群产品得以达成与海外竞品的性能对标。这种“以量补质”的集群化发展模式，正逐步成为国产算力产业破局突围的关键路径，为在制程工艺受限背景下实现技术追赶提供了有效解决方案。

对于算力芯片对 AIDC 需求压制，需要持续关注英伟达对于中国市场推出的新款定制降级芯片进展，以及国产芯片的放量速度。

图表：英伟达 GB200 和华为 Ascend 910C 性能对比

指标	Nvidia GB200	华为 Ascend 910C	GB200 vs 910C
BF16 稠密算力	2500 TFLOPS	780 TFLOPS	0.3x
HBM 内存容量	192 GB	128 GB	0.7x
HBM 内存带宽	8.0 TB/s	3.2 TB/s	0.4x
Scale up 带宽（单向）	7200 Gb/s	2800 Gb/s	0.4x
Scale out 带宽（单向）	400 Gb/s	400 Gb/s	1.0x

资料来源：SemiAnalysus，Nvidia，华为，泽平宏观

图表：英伟达 GB200 NVL72 和华为 Cloud Matrix 384 性能对比

指标	Nvidia GB200 NVL72	华为 Cloud Matrix CM384	CM384 vs NVL72
AI 芯片数量	72	384	5.3x
BF16 稠密算力	180 PFLOPS	300 PFLOPS	1.7x
HBM 内存容量	13.8 TB	49.2 TB	3.6x
HBM 内存带宽	576 TB/s	1,229 TB/s	2.1x
Scale up 带宽（单向）	518,400 Gb/s	1,075,200 Gb/s	2.1x
Scale out 带宽（单向）	28,800 Gb/s	153,600 Gb/s	5.3x
系统总功耗	145,000 W	599,821 W	4.1x
每 BF16 算力的功耗	0.81 W/TFLOP	2.0 W/TFLOP	2.5x

资料来源：SemiAnalysus，Nvidia，华为，泽平宏观

3.2 液冷：AI 芯片功耗提升，带动散热需求

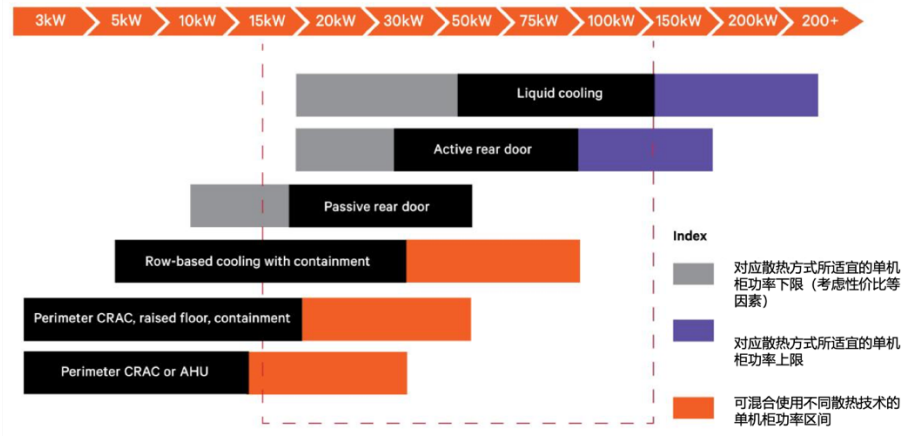
AI 芯片性能迭代带动功耗激增，液冷技术成为散热刚需。以英伟达 H100 为例，其热点功耗密度达 1.5W/cm²，远超风冷 0.3W/cm²的上限，若强行使用风冷需扩大机柜间距 50%，导致空间利用率下降 40%，显著推升土地和基建成本，液冷凭借热传导效率（液体比空气高 1000-3000 倍）成为必然选择。

根据英伟达估计，液冷数据中心的 PUE（能源使用效率）可以达到 1.15，远低于风冷数据中心的 1.6。工信部《新型数据中心发展三年行

动计划》要求 2025 年全国数据中心平均 PUE 降至 1.5 以下，单机柜功率超 30kW 必须强制采用液冷，并设定 PUE≤1.3 的准入红线，倒逼液冷技术普及。

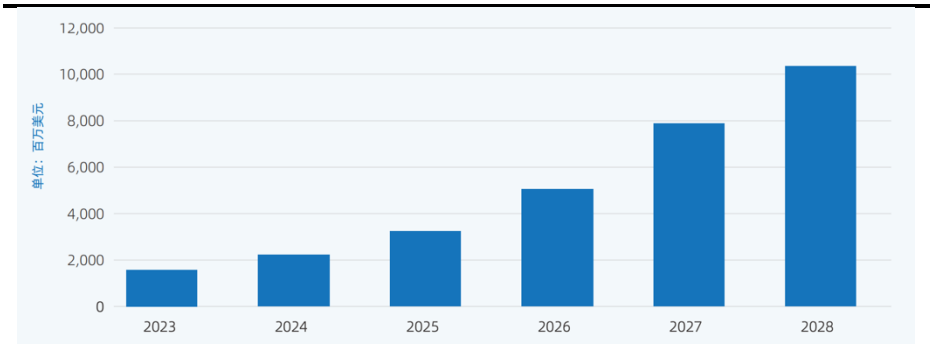
根据 IDC 预计，2028 年中国液冷服务器市场将达到 105 亿美元，2023-2028 年五年年复合增长率将达到 48.3%。中长期看，AI 算力密度提升与 ESG 要求深化将驱动液冷从“可选”迈向“必选”。

图表：单机柜功率密度与适宜的散热方式



资料来源：Vertiv 官网，泽平宏观

图表：2023-2028 年中国液冷服务器市场规模及预测



资料来源：IDC，泽平宏观

3.3 柴油发动机：AIDC 的最后防线

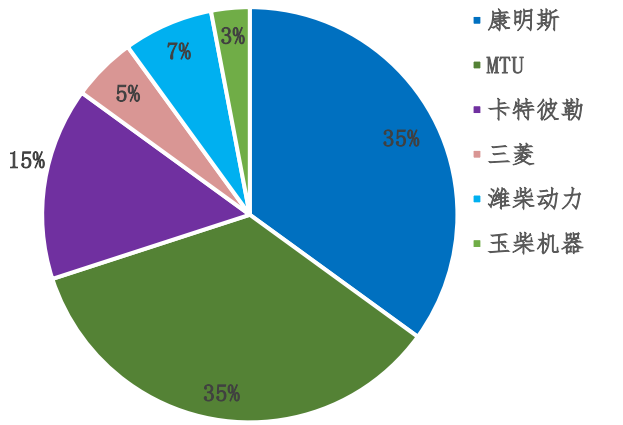
柴油发电机组作为数据中心电力冗余体系的核心设备，承担着应急备用电源“最后保障”的关键角色。数据中心供电系统通常采用“电网+UPS+柴发”三级保障架构：UPS 凭借毫秒级切换能力实现市电中断时的瞬时电力接续，但其储能容量有限（仅能维持数分钟至数十分钟）；柴发则在 UPS 续航耗尽后启动，以大功率、长时供能特性（可持续供电数小时至数天），成为数据中心应对长时间停电事故的唯一可靠备份电源，二者通过“瞬时响应+持续供能”的功能互补，构建起完整的电力保障链条。

柴发产业链具备技术壁垒高、扩产周期长的显著特征。扩产需要全产业链扩，由于柴发产业链长，要求标准高、工艺复杂，供应商导入严格，一个节点的扩产不及预期，都将拖累这条产业链的扩产进度。**而**

且大功率机组（1.6-2MW）对发动机、控制系统等核心部件的技术要求严苛，国内具备规模化交付能力的企业少。

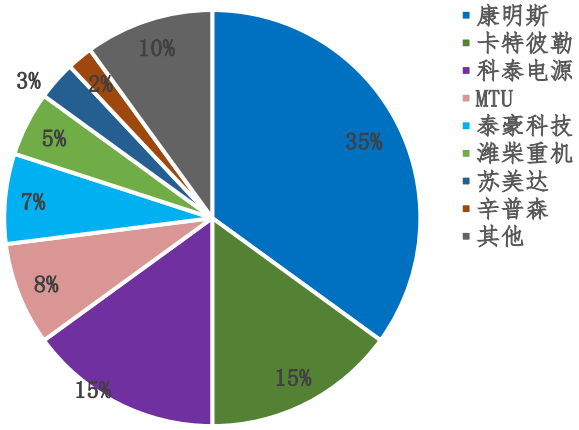
智算中心大型化有望推动柴油发动机需求曲线更加陡峭。AI 算力需求爆发推动数据中心建设加速，全球数据中心柴发冗余配置率从 80%提升至 120%-150%，当单柜功率密度从 20kW 跃升至 50-100kW，将带动大功率机组需求同比增长 150%。供需错配背景下，柴发设备价格已进入涨价周期，行业呈现量利齐升格局。

图表：2024 年中国 IDC 用柴油发动机竞争格局



资料来源：潍柴重机，泰豪科技，科泰电源，泽平宏观

图表：2024 年中国 IDC 用柴发机组竞争格局



资料来源：潍柴重机，泰豪科技，科泰电源，泽平宏观

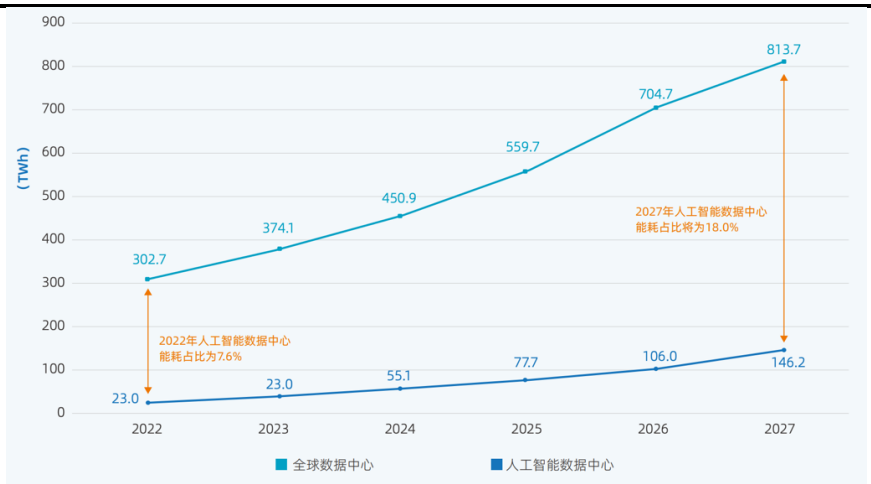
3.4 可控核聚变：算力的尽头是电力

AI 芯片国产替代“以量补质”的唯一缺点是，它需要 4.1 倍 GB200 NVL72 的功率，每 FLOP 的功率差 2.5 倍，每 TB/s 内存带宽的功率差 1.9 倍，每 TB HBM 内存容量的功率差 1.2 倍。**但这一模式的可行性植根于中国能源禀赋特性，相较于硅片制造受限的硬性瓶颈，电力供应体系具备更强弹性调节空间。**

人工智能大模型技术的研发和应用带来了更高的能耗需求。从 AI 对能源的需求来看，数据中心作为 AI 运行的核心载体，其电力消耗正经历迅猛增长。如美国的星际之门计划，其项目预计将有高达数千兆瓦的电力需求。

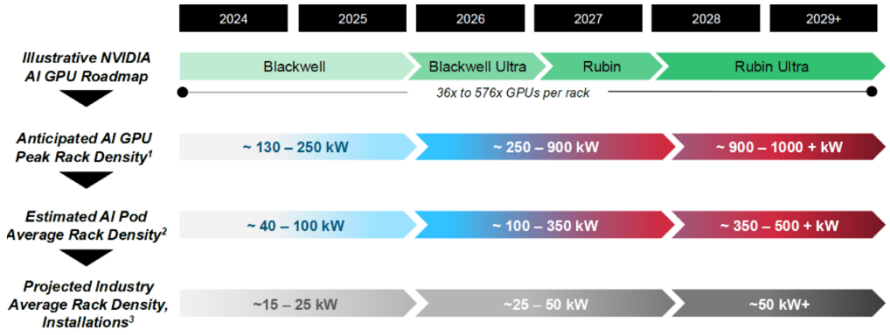
根据 IDC，2024 年中国人工智能数据中心 IT 能耗（含服务器、存储系统和网络）达到 55.1TWh，2025 年将增至 77.7TWh，是 2023 年能耗量的两倍，2027 年将增长至 146.2TWh，2022-2027 年五年年复合增长率为 44.8%，五年间实现六倍增长。据 Vertiv 预测，以能耗为单位，2023-2029 年全球新增智算中心总负载将达 100GW，每年新增约 13-20GW。在这一趋势下，保障电力供应尤为重要。

图表：2022-2027 年全球数据中心及人工智能数据中心能耗预测



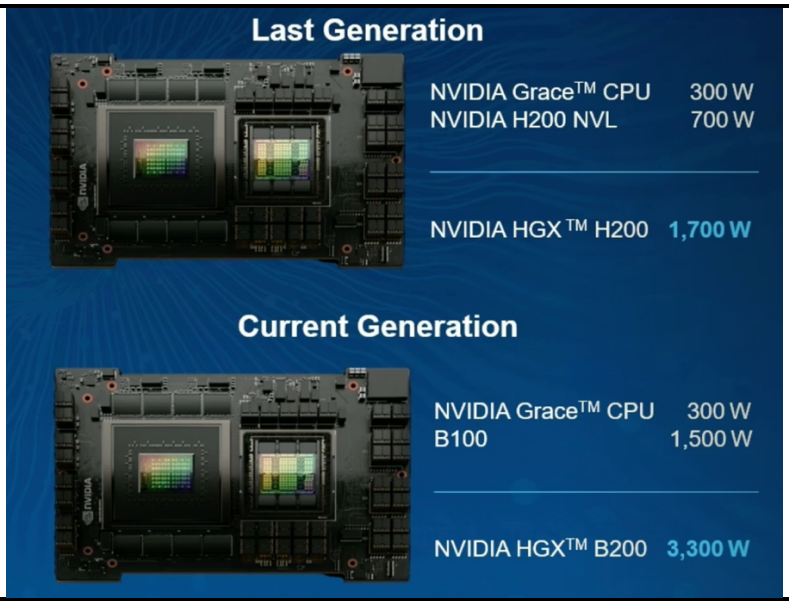
资料来源：IDC，泽平宏观

图表：未来单机柜能耗或将达到 MW 级



*资料来源：*Vertiv，*泽平宏观*

图表：英伟达 B200 能耗比 H200 翻一倍



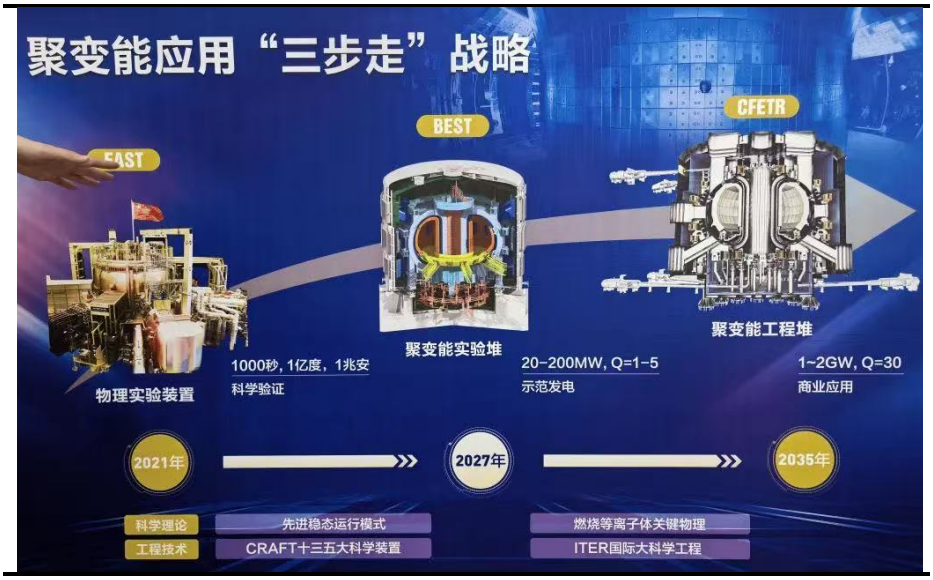
资料来源：《Delta’s Grid-to-Chip Power Solutions for Giga-Watt scale AI Data Centers》，泽平宏观

小型模块化反应堆（SMR）或是未来 AIDC 供电主力。在 2025 年中国春季核能论坛上，我国核能企业与 AIDC 应用方正式启动 SMR 为数据中心供电的示范项目前期技术论证与场景适配研究。与传统大型核反应堆相比，SMR 有望凭借其灵活部署、低碳高效的特性，成为我国核电领域对接新型算力基础设施的重要增量市场，为 AIDC 提供稳定、低成本的基荷电源。

中国在可控核聚变领域实现多项里程碑式突破，为 AIDC 能源需求奠定技术基础。5 月 1 日，根据合肥市官网，我国紧凑型聚变能实验装置（BEST）工程总装正式启动。5 月 4 日，根据央视，大科学装置聚变堆主机关键系统综合研究设施“夸父”（CRAFT）项目，建设已经进入关键阶段，预计将于 2025 年年底全面建成。

5 月以来核聚变催化不断，核能作为中国实现“双碳”目标、能源安全的重要战略方向，其可控核聚变及 SMR 是行业未来发展的重要增量，**未来，SMR 或将是除燃气轮机外的 AIDC 供电主力。**

图表：核聚变路线图



资料来源：中科院-合肥可控核聚变，泽平宏观