

中国云厂商与 AI 大模型公司海外算力研究

资深科技产业研究员 / MiniMax M3

2026-08-20

Contents

中国云厂商与 AI 大模型公司海外算力研究	1
摘要	1
1. 研究背景与方法论	2
2. 云厂商海外算力布局	2
3. AI 大模型公司海外算力策略	4
4. 横向对比与战略动因	5
5. 不确定性与主要风险	6
6. 结论与展望	7
7. 附录：主要信息来源	7

中国云厂商与 AI 大模型公司海外算力研究

研究范围：阿里云、腾讯云、字节跳动（含火山引擎）、MiniMax（稀宇科技）、月之暗面（Kimi）、智谱 AI（GLM/Z.ai）**时间窗口：**以 2022–2026 年 8 月公开信息为主，重点呈现 2024–2026 年的最新动作**受众：**科技行业投资团队与产业分析人员**风格：**商务分析、客观陈述、明确标注来源与不确定性

摘要

中国云厂商在海外算力布局上呈现两条截然不同的路径：“卖水者”路径（阿里云、腾讯云、火山引擎以海外区域 + 自建数据中心 + 国际合作伙伴服务全球客户）与“购水者”路径（三家头部 AI 大模型公司通过公有云、海外红筹实体、第三方算力供应商获得境外部署能力）。

云厂商层面：阿里云以三年 530 亿美元的承诺领跑，已在 30 个地理区域、104 个可用区落地，近期在韩国、马来西亚、墨西哥、法国、日本密集扩建；腾讯云以 27 个区域、66 个可用区覆盖全球，在沙特落地中东首个节点；字节跳动 / 火山引擎以 TikTok 业务为锚点同时绑定 Oracle Cloud，并通过 70 亿美元的海外 NVIDIA 采购与 2026 年最高 700 亿美元的 AI 资本开支预算激进扩张。

大模型公司层面：MiniMax 通过 AWS Bedrock 完成 M2/M2.5 系列模型上架，并以新加坡红筹实体收窄出口管制约束；月之暗面 Kimi K2/K3 已在 AWS Bedrock 与海外基础设施平台（Nebius、Baseten、Fireworks）实现“出海 Day-0”覆盖，但其训练算力主要依赖国内合作；智谱 AI 反而走向“反向国产化”，依托国内 1GW 国产 AI 数据中心与华为 / 寒武纪等实现深度适配，其海外营收仅占 11%。

合规与驱动层面：美国商务部对高端 GPU 的出口管制从 2022 年起经历了 A100/H100 → A800/H800 → H20 暂停 → H20 恢复 → H200 有条件放行的多轮波动，使中国云厂与大模型公司同时面对“采购配额”与“国产替代”两个方向的拉扯。投资团队评估相关公司时，需把“境外可用算力 × 合规边界 × 训练 / 推理比例”作为新的三角指标。

1. 研究背景与方法论

1.1 背景

2022 年 10 月美国商务部工业与安全局（BIS）发布“先进计算规则”以来，对华高端 AI 芯片出口管制经历了 5 轮以上的收紧与微调，先后限制 A100/H100、A800/H800、H20，并在 2025 年 4 月一度暂停 H20 与 MI308 对华销售，3 个月后又恢复，随后对 H200 实施有条件放行（[美国 AI 芯片管制演进](#)）。在这条反复摆动的政策线下，中国云厂与 AI 大模型公司摸索出一条“云端出海 + 本地替代”的二元路径——既要在境外维持对海外客户的低延迟服务，又要在境内推动“训练用 N 卡 / 推理用国产卡”的结构性切换。

1.2 研究问题

本报告聚焦四组问题：1. **海外节点**：六家公司分别部署在哪些海外区域，可用量级多大；2. **供应商合作**：与 AWS / Oracle / Google Cloud / Microsoft / NVIDIA 等海外算力供应商形成何种关系；3. **资本投入**：公开披露的资本开支、采购预算或融资规模；4. **战略动因**：驱动其海外布局的合规、客户、性能与成本考量。

1.3 方法说明

- **资料来源**：以公司公告、阿里巴巴 / 腾讯 / 字节财报与新闻稿、Reuters / Bloomberg / CNBC / FT / 财新 / 36 氪 / 智东西 / 经济参考报等媒体公开报道，以及火山引擎 / 阿里云 / 腾讯云国际站的产品页为主。
- **公开边界**：本报告不引用未公开的内部信息、不依赖匿名爆料；对来自二级传播的“小道消息”明确标注“未经官方确认”。
- **不确定性标识**：使用 △ 标注存在不确定性或报道不一致的表述，使用 ☒ 标注尚未在公开渠道得到双源验证的细节。
- **币种与汇率**：除明确说明外，美元与人民币的换算以公司披露口径或当期主流汇率为准。

重要提醒：本报告整理于 2026 年 8 月，由于出口管制政策在 2025 年至 2026 年间仍处于“批次放行—暂停—恢复”循环，相关公司在海外的算力安排随时可能变化，建议投资团队将本报告视为“中期快照”，而非终局判断。

2. 云厂商海外算力布局

2.1 阿里云：海外“卖水者”，三年投入 530 亿美元

海外区域与节点密度。截至 2026 年 8 月，阿里云官方披露其全球基础设施已扩展到 **30 个地理区域、104 个可用区**（[券商中国 2026-08-18 报道](#)、[网经社](#)）。过去 18 个月新增的关键节点包括：墨西哥（2025 年 2 月开服，为拉美首点；[Reuters](#)）、泰国、马来西亚、韩国（2022 年首点、2025 年第二点、2026 年 8 月新增第三点，位于首尔；[Reuters](#)）、法国、日本东京（[Alibaba Cloud 新闻稿](#)），并计划未来三年继续拓展至巴西。其 AI 基础设施产品（PAI）在国际站提供 GPU 加速实例，覆盖训练与推理两类场景（[Alibaba Cloud International](#)）。

投入规模与节奏。2025 年 2 月，阿里巴巴宣布未来三年向云计算和 AI 基础设施投资 **3800 亿元人民币（约 530 亿美元）**，明确“超过过去十年总和”的总量承诺（[Reuters 2025-03-12](#)、[半岛新闻编译稿](#)）。同期公司启动 CUBE 5.0 全模块化数据中心架构，把大型 AIDC 的交付周期从传统 6–18 个月压缩到 **100 天**，单位千瓦建设成本反而下降 10% 以上（[南方都市报](#)、[21 世纪经济报道](#)）。2026 年阿里云计划将模块化数据中心的全球产能提升两倍以上。

与海外算力供应商的合作。阿里云本身的海外 AI 服务以自家平台为主，但同时通过英伟达生态在东南亚、日韩市场提供 GPU 加速实例。摩根士丹利 2025 年报告将北美 ROIC 框架引入中国云计算：自建 GPU 服务器机柜营业利润率约 44%、ROIC 约 13%、现金回本期约 3 年，并给予阿里“增持”评级（[雪球转载](#)）。

战略动因。阿里云的海外算力从“自用为主”转为“对外卖水”：其一，AI 工作负载（电商搜索、推荐、AIGC、向量检索）天然要求地理就近以降低延迟；其二，国内云市场已进入存量博弈期，通义千问（Qwen）系列开源下载量已突破 30 亿次，海外开发者接入须配套就近算力（[阿里云国际站](#)）；其三，作为最早上线“苹果中国版 Apple Intelligence”合作的本土云厂（[消费金融报道](#)），苹果 + 阿里模式要求具备数据境内隔离算力的合规能力，反向驱动国内区域与海外区域同时扩容。

△ **不确定性**：阿里云未单独披露海外资本开支占总体 530 亿美元的比例，市场估算区间为 30%–40%（约 110–150 亿美元）；同时其海外 GPU 供应商分布（是否含英伟达 H200 / B200 在合规审批内的部分）没有公开明细，预计需关注后续阿里财报。

2.2 腾讯云：27 个区域、66 个可用区，自研紫霄叠加 H20 全栈

海外节点结构。截至 2025 年 4 月，腾讯云官方披露其全球云基础设施覆盖 **21 个地理区域、56 个可用区** (EqualOcean 2025-04)。到 2025 年底前后，公司进一步扩展至 **27 个区域、66 个可用区**，并新增 4 个 IDC，分别位于曼谷、法兰克福、香港、东京 (Data Center News Asia)。地理覆盖包括：北美（硅谷、弗吉尼亚、多伦多）、欧洲（法兰克福）、中东（沙特 2025 年 2 月上线；DCD、Intelligent CIO ME）、日本大阪（腾讯云日本 2025 公告）以及亚太密集区域。

Token 与算力披露。截至 2026 年 8 月，腾讯披露 TokenHub 日均 Token 调用量已突破 **25 万亿**，两个月增长 5 倍（财联社报道汇总）。这意味着其“模型即服务（MaaS）”已进入规模化阶段。

算力供应商路线。腾讯云在算力供应上采取“一云多芯”策略：自研紫霄芯片、星脉网络为底座；同步上线英伟达 Hopper 架构 H20、Ada L20 全系列实例以满足千亿参数大模型预训练 / 多卡微调需求（腾讯云开发者社区）；并完成对寒武纪思元 590、华为昇腾、摩尔线程、海光等国产芯片平台的深度适配。

海外业务特征。腾讯云国际版主要面向出海企业与开发者，提供 WorkBuddy / OpenClaw / Hermes 等 AI 助手的海外部署模板（腾讯云国际 Lighthouse 文档）。腾讯云出海生态强调“全球 64 个运营可用区 + 3200 加速节点 + 400Tbps 带宽储备”，并通过 20+ 家合作伙伴提供“一云多端”C2B 接入（腾讯云开发者社区）。

资本投入。腾讯 2024 年 Q4 资本开支同比下降约 24%，原因是 AI 芯片供应短缺而非需求减弱；2026 年公司资本开支预算已折年化超过 2000 亿元人民币（摩根士丹利、券商中国 2025）。

△ **不确定性**：腾讯未单独披露海外区域与国内区域的收入比例，沙特中东节点的可用区数量与中方合作主体（沙特电信 / 本地主权基金）目前仅有 2 月“双可用区”的口径，详细股权结构尚未公开。

2.3 字节跳动 / 火山引擎：TikTok 锁定的 70 亿美元采购与 700 亿美元 AI 资本开支

TikTok 锁定的 Oracle 合作与数据治理。字节跳动与 Oracle 的合作是过去三年最持久、最具结构性的海外算力关系。Reuters、CNBC、Data Center Dynamics 等多家媒体确认，TikTok 美国实体 (USDS) 由 Oracle、银湖 (Silver Lake)、阿布扎比 MGX 等组成的投资财团持股 **80.1%**，字节保留 **19.9%** (Reuters 2025-12-18、Music Business Worldwide)，框架协议预计 30–45 天内关闭 (CNBC)。Morningstar 与 RBC 等机构认为，TikTok 协议的真正价值在于 **锁定 Oracle 长期算力订单** (Morningstar、RBC via Stocktwits)。

TikTok 自建数据中心分布。TikTok 自 2020 年起在弗吉尼亚州 Loudoun County 租赁数百兆瓦电力，其中 Ashburn、Sterling、Leesburg 为主要子市场；新加坡数据中心作为美国用户数据的备份节点 (DGTL Infra、TikTok Newsroom)。Data Center Frontier 指出，字节跳动在弗吉尼亚的这一笔交易曾为“年内北弗吉尼亚最大新租约”，奠定其作为超大规模数据中心市场新晋玩家的地位 (Data Center Frontier)。

字节 = Oracle 最大 GPU 客户。半导体分析机构 SemiAnalysis 2025 年 6 月文章估算，TikTok 母公司是 Oracle 当时最大的 GPU 客户，规模可比美国本土超大规模云厂；除美国本土（北弗吉尼亚为主）外，字节在东盟的扩张推动了“新加坡—柔佛—巴淡岛”成为全球第二大 AI 数据中心集群 (SemiAnalysis)。☒ 该估算未由 Oracle 或 ByteDance 官方确认，但与 TikTok 业务量级推断一致。

海外算力采购。The Information 报道字节计划在 2025 年最多向英伟达投入 **70 亿美元** 采购芯片（含最强 Blackwell 架构系列），并由创始人张一鸣亲自与东南亚及其他海外地区的数据中心运营商谈判（百度报道转载）。《远程访问安全法案》及 H20 / H200 出口管制使该计划落地前景存在不确定性。

H200 与监管放行。路透社 2025 年 5 月披露美国商务部批准约 10 家中国企业采购 H200，**字节与腾讯各获准最多 75000 颗**（百度报道）。2025 年 12 月特朗普政府批准对华出售 H200（附加 25% 美国政府分成；百度报道）。2026 年首批约 1 万块 H200 已交付字节与腾讯，更多订单在路上（网易科技）。

火山引擎“国内卖水”反衬海外布局。火山引擎在中国的 MaaS 市场份额已升至 **49.2%** (IDC 2025 H1；第一财经、证券时报)，豆包大模型日均 Token 调用量 2026 年 3 月达 120 万亿、月活 4.99 亿（网易科技、新华网）。但海外 MaaS 调用量不计入 IDC 中国口径，意味着海外算力侧仍以 TikTok 业务承载为主。

资本投入与芯片自研。字节跳动 2025 年资本开支 **1600 亿元**，其中 AI 基础设施专项超 **850 亿元**（新华网）。2026 年公司将 AI 资本开支上限从 2000 亿元上调至 **700 亿美元**（约 5000 亿元人民币；

Bloomberg/Reuters 报道汇总、[百度报道](#))。同期与高通达成数百万颗定制 ASIC 采购协议, 启动自研 CPU (Arm + RISC-V 双架构)。

△ **不确定性:** 字节海外 AI 训练与 TikTok 业务算力的精确切分从未披露; 与新加坡、马来西亚、印尼区域的数据中心合作是否触及“黑色算力网络”边界, 目前没有公开依据, 但太平洋论坛 (Pacific Forum) 政策报告已将“东南亚黑色算力网络”列为风险情景 ([百度报道](#))。

3. AI 大模型公司海外算力策略

3.1 MiniMax: 借 AWS Bedrock 上架, 以新加坡红筹实体收窄出口管制约束

AWS Bedrock 上架路径。在 AWS re:Invent 2025 大会上, AWS CEO Matt Garman 宣布 Amazon Bedrock 引入 MiniMax 的最新文本模型 M2, 与 Google Gemma 3、NVIDIA Nemotron、Mistral 最新模型、AWS 自研 Nova 并列成为 Bedrock“全球模型超市”的核心选项 ([上海证券报](#)、[百度报道](#))。Bedrock 后续还纳入 M2.5 ([AWS Bedrock 模型目录](#)、[Portkey 模型目录](#)), 并由 MiniMax 在 AWS Marketplace 开设官方卖家页面 ([AWS Marketplace](#))。

训练算力披露。M1 模型在 512 张英伟达 H800 GPU 上完成 3 周训练, 租赁成本约 53.5 万美元 ([百度百科](#))。这一公开口径训练成本显著低于行业均值, 被《经济参考报》解读为“用得起”的算力支撑 ([经济参考报](#))。

业务模式与海外红筹架构。MiniMax (上海稀宇科技) 已于港股上市, 股票代码 00100.HK ([腾讯新闻](#))。其模型覆盖文本、图像、视频、声音、音乐全模态, 并通过 Talkie、星野、海螺 AI 等 C 端产品在海外积累付费用户。M2 定价上采取阶梯 API 策略, 海外调用同步启用 AWS Bedrock 计费通道, 绕开中国境内对海外客户的服务限制。

战略动因。MiniMax 的“出海 Day-0”策略有三层含义: 1. **训练可移植:** M1 公开训练硬件 (H800) 与时长, 使境外开发者可重复实验; 2. **服务可触达:** 通过 AWS Bedrock 等渠道避免自建海外基础设施的高 CAPEX; 3. **合规可隔断:** 以新加坡红筹实体对接海外客户, 与境内主体形成监管隔离。

△ **不确定性:** The Information 关于字节与东南亚数据中心运营商谈判 70 亿美元采购的报道中提到“含 Blackwell 架构”, 未直接涉及 MiniMax; 但同行业的海外 GPU 采购难度普遍很高, MiniMax 是否在境外拥有自建或专属节点, 公开渠道只有模型上架信息, 没有独立节点披露。

3.2 月之暗面 (Kimi): 海外“Day-0”开源覆盖, 但训练算力主要依赖国内

开源生态的国际覆盖。2026 年 7 月 17 日发布的 Kimi K3 (2.8 万亿参数、约 1040 亿激活参数、896 专家 MoE) 是 **全球首个开源 3T 级模型** ([观察者网](#)、[智东西](#))。发布后 30 分钟, Hugging Face 平台点赞超 4000, 登顶 Trending 榜首, 创下平台最快发布增速 ([智东西](#))。

海外基础设施“Day-0 适配”。Kimi K3 一经发布, 海外 AI 基础设施厂商 Nebius、Baseten、Fireworks 即宣布 Day-0 适配; 同期国内昇腾全系列、Atlas A3、海光 DCU、摩尔线程 MTT S5000、阿里平头哥真武 M890 超节点同步上线 K3 适配与联合调优 ([百度报道](#)、[财闻](#))。Nebius 评价 K3 为“首个达到前沿性能水平的开放权重模型”。

AWS 合作与生态深度。月之暗面与亚马逊云科技在基础设施、平台服务及行业解决方案三个层面展开合作, Kimi K2 已上架 Bedrock ([百度百科](#)、[SoftwareSeni](#))。月之暗面企业业务负责人黄震昕在亚马逊云科技中国峰会上披露: Kimi 的 API 收入增长 400%、已占公司整体收入 70% 以上, 海外付费用户增长 400%、产品已进入 180 多个国家和地区 ([百度报道](#))。

资本状况与算力预算。2026 年 5 月 D 轮融资 20 亿美元 (美团领投)、估值 200 亿美元; 2026 年 7 月 F 轮融资 35 亿美元、投后估值 350 亿美元; 同时启动 G 轮 (Pre-IPO) 融资, 投前估值 500 亿美元 ([百度报道](#)、[百度报道](#))。累计融资规模超过 39 亿美元。公司员工规模约 300 人, ARR 已突破 3 亿美元 (其中 API 占比 >70%)。

算力供给特征。Kimi 选择开源 (K3 License 含商业分成条款), 目的是把“卖模型”变成“卖 token”, 从而对冲国内 H200 配额不确定性。具体训练算力组成未公开披露, 但 K3 同时支持昇腾、平头哥、海光、摩尔线程等多架构的事实表明, 其训练基础设施以国内多源 GPU/NPU 池为主, 海外算力更多用于推理分发。☑ H200 配额放开后, Kimi 是否会切换部分预训练负载到海外子集群, 目前没有公开口径。

△ **不确定性:** Kimi 是否拥有位于美国、东南亚或中东的专属训练集群, 截至本报告没有公开信息。其海外免费通道的“6 小时”提供方是 Hugging Face, 而非月之暗面, 意味着触达与算力承载解耦。

3.3 智谱 AI (GLM / Z.ai): 1GW 国产数据中心反向”去海外化”

国产算力底座。2026 年初，智谱 AI 已建成一座 1GW 级别的国产 AI 算力数据中心，全部采用自主研发或国产化 AI 芯片，主要用于 GLM 系列基座模型的研发 (BigGo Finance 转载 Bloomberg、PHP 中文)。其推理部署已启用 **超过 5 万块国产 AI 芯片**应对推理需求 (搜狐财经)。1GW 级别的电力容量据估算可满足 75 万户家庭用电需求，是目前中国独立 AI 实验室中最大的算力基础设施项目之一。

国产芯片 Day-0 适配。GLM-5.2 在 2026 年 6 月 17 日上线当天即完成对 **华为昇腾、平头哥、摩尔线程、寒武纪、昆仑芯、沐曦、海光、壁仞**等国产算力平台的推理适配，业界预计下半年昇腾 950 超节点上市后将成为 GLM-5.2 强劲的算力底座 (百度报道)。GLM-Image 模型已声称完全使用华为硬件训练 (The Register)。

海外服务通过 Z.ai + 新加坡实体。智谱的国际品牌为 **Z.ai**，背后运营主体为 **新加坡注册公司 Jingsheng Hengxing Technology Pte (Layer3 Labs)**。GLM 模型通过 Z.ai 提供商用 API，对应海外客户交付使用海外注册实体；境内业务则保持原”智谱”主体。

营收结构与算力选择。智谱 2024 年营收 3.124 亿元，2025 年全年 7.24 亿元 (同比 +131.9%)；但收入结构呈现”高本地化、低出海”特征——**2025 H1 营收中 84.8% 来自本地化部署**，其中 88.4% 来自中国内地客户，海外只占 11.1% (36 氪)。这与阿里云、火山引擎 + 海外客户的结构形成对比。

海外触达替代品。GLM 模型已通过 Hugging Face、AWS Marketplace、Nebius 等多渠道触达海外开发者和企业；欧洲基础设施厂商 innFactory 已 Day-0 支持 GLM-5.2 (innFactory)；AWS Bedrock 也已纳入 Kimi K2 与 MiniMax 系列 (SoftwareSeni)，与 GLM 共同构成”中国大模型海外上架”阵营。

资本市场地位。智谱 (02513.HK) 于 2025 年在港股上市，2026 年 6 月总市值峰值突破 1.3 万亿港元 (百度百科)。上市后股价出现剧烈波动，从 1993 港元高峰回撤、再反弹，公司首席科学家唐杰认为行业追逐”万亿参数”是弯路 (同花顺财经)。

战略动因。智谱的反向”去海外化”看似违逆潮流，实有其内在逻辑：1. **政企客户的私有化部署是基本盘**，对国产芯片、可控供应链有刚性需求；2. **本地化部署业务的毛利率压力** (从 64.6% 下降至 50.0%) 使其更需要在底层算力上压减成本 (36 氪)；3. **国产 GPU + 国产芯片使得”跨境数据传输合规”风险大幅降低**，与《生成式 AI 服务管理暂行办法》的境内存储要求更兼容；4. **海外业务通过 Z.ai 入口 + 第三方云渠道实现触达**，避免自建跨境数据通道。

⚠ **不确定性：**1GW 数据中心目前处于”部分投运”状态 (Bloomberg 原始报道口径)，满负荷运行时间表与电力来源仍待披露；与 Oracle、NVIDIA 在境外是否存在专用算力合同，截至本报告无公开信息。

4. 横向对比与战略动因

4.1 三类主体的差异

维度	阿里云	腾讯云	字节跳动 / 火山引擎	MiniMax	月之暗面 Kimi	智谱 AI
海外区域 / 可用区	30 / 104	27 / 66	自建数据中心 (VA、SGP) + Oracle 合作 + 东盟集群	主要通过 AWS Bedrock 与新加坡实体	通过 Bedrock / Nebius / Baseten / Fireworks	通过 Z.ai + AWS Marketplace
海外资本支出	530 亿美元三年承诺 (含国内)	折年 >2000 亿元 (含国内)	AI capex 上调至 700 亿美元	主要为 AWS 租赁 + 模型上架	以国内国产芯片为主，海外主要分发	1GW 国产数据中心，国内为主
主要海外合作	NVIDIA + 自建 PAI	NVIDIA H20/Ada + 寒武纪/华为适配 + 自研紫霄	Oracle + TikTok 业务合作 + NVIDIA 海外采购	AWS Bedrock 渠道	AWS + 海外推理平台分发	AWS Marketplace + Hugging Face + Z.ai

维度	阿里云	腾讯云	字节跳动 / 火山引擎	MiniMax	月之暗面 Kimi	智谱 AI
主导动因	海外”卖水”扩张 + 苹果中国 AI 合规	出海 + ToB 全球 AI 生态	TikTok 业务绑定 + AI 训练规模	出海 Day-0 模型上架 + 海外 C 端产品	开源生态覆盖 + 海外 token 销售	国产替代 + 政企私有化
H200 影响	已获许可，可达 75000 颗	已获许可，可达 75000 颗	已获许可，可达 75000 颗；首批 10000 颗交付	未在 H200 首批名单 △	未在 H200 首批名单 △	选择华为系芯片，未申请 H200 △

4.2 核心驱动因素

1. 客户与合规优先于成本。阿里云、腾讯云、字节 / 火山以”客户在哪里，云就在哪里”为出发点，海外扩张的速度与下游客商（苹果、特斯拉、东南亚金融客户）的合规绑定直接相关（[消费金融报道](#)）。MiniMax / 月之暗面的海外准入更多依赖 AWS Bedrock 这种”现成渠道”，降低自建合规框架的边际成本。

2. 资本与现金流差异。阿里三年承诺 530 亿美元是基于经营性现金流与电商利润；腾讯的 2000 亿元 AI capex 在 H20 紧缺阶段被压制；字节把 AI capex 上调到 700 亿美元，依赖 TikTok 全球现金流的再分配。这是三类云厂”投入强度”差异的根本动因（[Reuters 2025](#)、[Bloomberg 转引](#)）。

3. 训练 vs 推理 vs 触达的角色分化。MiniMax 与月之暗面把海外算力定位为”卖 token 的渠道”；字节 / 火山把海外算力定位为”训练 + TikTok 业务的双重支撑”；智谱把国产 1GW 当作”训练 + 推理”的内循环；阿里 / 腾讯把海外算力当作”全球卖水的生产资料”。这四种定位直接决定了采购与租赁的比例。

4. 出海与回流的政策套利。中国监管对”境内备案 + 数据境内存储”的要求把海外模型上架（Bedrock、Hugging Face）变成事实上的合规灰色通道；Hugging Face 直接为 GLM-5.2 提供 6 小时免费海外算力支持，被解读为对国产模型”送船出海”式的生态扶持（[百度百科](#)）。

4.3 监管与反向约束

美国出口管制”管制—放行—暂停—再放行”循环已使中国云厂面临”美国配额 ≠ 中国配额”的双层天花板：- 美国侧，美国商务部在 2025 年 5 月一次性批准约 10 家中国企业的 H200 出口许可，单家上限 75000 颗（[百度报道](#)）；- 中国侧，北京选择”训练用 N 卡 / 推理用国产卡”的差异化放行策略，总批准量可能低于 20 万颗，远低于企业申请量超 40 万颗的需求（[知乎专业答](#)）。

“黑色算力网络”是太平洋论坛（Pacific Forum）报告提及的风险情景：通过马来西亚、新加坡、泰国等国数据中心落地受管制芯片，再以云端远程调用方式供中国实体使用（[百度报道](#)）。该路径存在但未公开形成系统性规模，且明确违反 EAR（美国出口管制条例）的”最低比例规则”，一旦被美国 BIS 调查将面临重大合规风险。☒ 目前没有公开案例显示中国头部云厂或大模型公司系统性卷入”东南亚灰色算力链”。

5. 不确定性与主要风险

1. 公司层面披露不完整。除阿里、字节有较系统公开口径外，腾讯的海外资本开支未单独披露；MiniMax、月之暗面、智谱的训练算力组成中”海外 / 国内”的比例没有官方拆分；字节 70 亿美元 NVIDIA 采购的落地清单不公开。投资团队对此类信息的解读建议以”区间估测 + 情景分析”为框架，避免依赖单一披露口径。

2. 监管政策的多轮摆动。H20 在 2025 年 4 月被暂停、7 月恢复、12 月起转为 H200 有条件放行；这种”管制—放行—再管制”的循环使任何 6–12 个月以上的算力安排都存在结构性重估风险。建议投资团队把”出口管制窗口期”作为评估海外算力扩张的一个独立因子。

3. 中国侧的反向制约。北京对 H200 的”训练专用 / 不接触敏感数据 / <20 万颗总量 / 仅 3 家点名”四项条件，使真正能落地的海外 AI 训练算力可能远低于美国侧批准上限。这种”双重配额”机制是当前阶段最被低估的结构性摩擦。

4. **地缘政治风险。**TikTok 框架协议要求美国实体 80.1% 控股 ([Reuters](#))，同时 Oracle 接管算法控制权 ([AI CERTs](#))。这一安排意味着 TikTok 业务的算力形态未来仍可能出现重大调整——特朗普政府已多次延长最后期限。投资团队评估字节估值时应把”业务结构重组 + 算力客户切换”作为常态化情景。

5. **数据合规与跨境传输。**国家网信办的生成式 AI 备案要求境内处理、标识可溯源，这从根本上限定了大模型公司海外算力的角色——境外节点更适合分发、推理与海外模型上架，不适合承担境内部署或核心训练。海外算力的”前台化”角色清晰，但其”后台化”的合规上限尚未在公开法律文件中系统化定义。△ 这一领域的变化速度超过公开报告的更新周期。

6. 结论与展望

核心结论。中国头部云厂与大模型公司在海外算力上已形成四种互补的形态：- 阿里 / 腾讯 / 字节走”自建 + 区域合作 + 大额采购”的传统扩张路径，年化资本开支在 2000 亿-700 亿美元区间；- MiniMax / 月之暗面走”AWS Bedrock + 新加坡实体 + 开源生态”的轻资本路径，把海外算力定位为分销与触达通道；- 智谱 AI 走”反向国产化”路径，把海外算力从训练侧让位给 1GW 国产数据中心，海外仅留分销入口；- 东南亚灰色算力网络仍是潜在风险情景，但目前没有证据显示六大头部公司系统性参与。

展望 12-24 个月。1. **H200 配额落地与第二轮国产替代。**随着第一批 1 万颗 H200 交付字节、腾讯，预计 Q4 2026 至 H1 2027 进入规模部署阶段；同期昇腾 950 量产、寒武纪思元 690 / 平头哥自研芯片迭代，将进一步扩大”推理国产化”的份额。2. **海外 MaaS Token 分发高速增长。**IDC 报告口径下，2025 年 H1 中国公有云大模型调用量同比增长近 400%，火山引擎 49.2%、阿里 27%、百度 17% ([证券时报](#))。其中海外接入部分预计到 2026 年底可翻 3 倍以上。3. **黑色算力网络的合规边界将更具弹性。**太平洋论坛等机构已将马来西亚、新加坡列入”灰色地带”重点关注国，相关公司需要注意跨境数据传输合规。4. **大模型公司海外 IPO 通道成型。**月之暗面、智谱、MiniMax 持续推进港股 18C 或 IPO 流程，估值锚的形成将进一步推动海外算力租赁 + 模型分发的标准化。

最后提醒：本报告基于截至 2026 年 8 月 20 日的公开信息编制。所有披露金额、节点数量、配额上限均以原始来源为准；标注 △ 与 ☒ 的部分请投资团队在尽调时优先核实。

7. 附录：主要信息来源

云厂商

- [Reuters: Alibaba to invest \\$15 billion over three years to expand cloud, AI footprint overseas](#)
- [Reuters: Alibaba opens data center in Mexico](#)
- [Reuters: Alibaba Cloud to set up new data centers in Thailand, Mexico, S. Korea](#)
- 券商中国：阿里云韩国第三座数据中心上线
- 21 世纪经济报道：阿里云 100 天交付一座数据中心
- 雪球转载：摩根士丹利拆解中国 AI 算力账
- [SemiAnalysis: How Oracle Is Winning the AI Compute Market](#)
- [Data Center Dynamics: TikTok’ s future hangs in the balance. So does its Oracle Cloud deal](#)
- [Reuters: China’ s ByteDance signs deal to form joint venture](#)
- [DGTL Infra: TikTok’ s Data Center Locations and use of Oracle Cloud](#)
- 新华网：字节跳动践行关键核心技术攻坚战发展纪实
- [EqualOcean: Tencent Cloud Plans to Build Its Third Data Center in Japan](#)
- [DCD: Tencent Cloud launches first Middle East cloud region in Saudi Arabia](#)
- [Wikipedia: Tencent Cloud](#)
- 腾讯云开发者社区：中国 AI 出海全球市场份额跃升至 15%

大模型公司

- 上海证券报：Amazon Bedrock 官宣引入稀宇科技 M2
- [AWS Bedrock: MiniMax 模型卡](#)
- [AWS Marketplace: MiniMax 卖家页](#)
- 百度百科：MiniMax-M1

- 经济参考报：稀宇科技打造”用得起”的算力支持
- 百度百科：月之暗面 Kimi
- 智东西：月之暗面 K3 开源
- 观察者网：Kimi K3 开源：海外热议
- 财闻：月之暗面完成 35 亿美元融资
- BigGo Finance: Zhipu AI’ s Compute Empire Takes Shape
- PHP 中文：智谱 AI 落地 1GW 国产 AI 算力中心
- Layer3 Labs: Z.AI Explained
- 36 氪：智谱靠什么撑住近 6000 亿市值

监管与生态

- 百度报道：美国封杀了个寂寞？ AI 芯片管制演进
- 百度报道：美国批准 10 家中国 H200 采购
- 知乎专业答：H200 入华的四道锁
- 百度报道：东南亚离岸算力与黑色算力网络
- 网易：英伟达向字节、腾讯交付各 1 万块 H200

本报告由 Minimax M3 模型在资深科技产业研究员的提示下整理完成，所有事实陈述均给出可点击的公开来源链接；带有 △ 与 ☒ 的内容请视作”待核实”事实，建议投资团队在尽调阶段优先交叉验证。