星星之火可以燎原

—— 从 WAIC 2026 看互联网/OTT 行业 AI 算力最新动态

（HCR 郭磊 lei.guo@hrc.com，2026 年 7 月）

主要阅读对象：互联网/OTT 行业技术决策层、基础设施负责人、算力规划与采购负责人，芯片厂商市场研究人员等

核心数据来源：WAIC 2026 官方发布、IDC、中国半导体行业协会、券商产业链调研、头部企业公开技术白皮书及行业权威媒体公开报道

摘要

2025 下半年至 2026 年，全球互联网/OTT 行业 AI 算力基建已从“单一拼卡”的初级阶段，全面升级为“系统级集群能力”的终局博弈 —— 这一转变的核心逻辑是：行业算力消耗结构从“大模型高频训练”转向“万亿参数级 MoE 模型高并发推理”，单芯片的峰值算力优势，被集群的互联带宽、全局通信效率、综合运维成本的优势所全面弱化。

从行业底层驱动力来看，国内互联网/OTT 行业 AI 算力需求的增长，本质是“业务端 AI 化”与“算力端国产化”两大趋势的同频共振：业务端层面，传统核心场景搜广推（搜索、广告、推荐）的 AI 重构，与新一代多模态、长上下文模型的业务落地，双向催生了海量的异构算力需求；算力端层面，供应链风险与成本效益的双重约束，推动行业从单一依赖高端海外 GPU，转向“海外 GPU+ 国产 NPU”的双轨混合架构。

作为行业发展的关键节点，2026 年世界人工智能大会（WAIC 2026）清晰标注了行业算力基建的最新基准：头部厂商均推出“超节点化”算力系统 —— 并非简单将多颗芯片堆叠，而是通过自研互联协议、拓扑架构设计，将成百上千颗芯片构建为“单一逻辑计算集群”，以此支撑 MoE 模型的分布式并行需求。其中，华为昇腾、百度昆仑芯、阿里平头哥等头部国产厂商，已在超节点架构上完成从技术到商用的闭环验证；互联网行业的算力应用，也从“零散场景试点”升级为“全业务链规模化落地”。

基于这一行业背景，本报告将从应用层、模型层、框架层、硬件底座、厂商策略五大维度，系统梳理行业算力发展的最新现状，重点梳理头部企业的落地策略与选型逻辑，最终对行业未来 12-24 个月的核心发展趋势做出明确判断。

1. 应用层：AI 化业务纵深渗透，算力价值闭环成型

互联网/OTT 行业的 AI 算力投入，已彻底从“技术概念投入”转向“直接业务产能投入”—— 算力资源不再是单纯的技术基建储备，而是直接支撑核心业务流量、决定用户体验的关键生产要素；其投入规模的核心依据，是业务场景的实际流量并发需求、用户体验时延容忍度、业务直接转化收益三者的共同平衡。

1.1 ToC：核心业务全面 AI 化，算力聚焦“高并发、低时延”

从行业级算力消耗结构来看，互联网/OTT 行业的 AI 算力流向高度集中：搜广推（搜索、广告、推荐）场景的推理算力，占行业总算力消耗的 60%-70%；若从全国智能算力的活跃消耗占比来看，这一场景的占比高达 30%-35%，是当前国内最大的单一 AI 算力消耗场景。

这一数据背后的业务逻辑是：互联网/OTT 行业的核心营收与用户留存场景，已完成从传统计算架构向 AI 推理架构的全面迁移 —— 在这一迁移过程中，业务场景对算力的性能约束，从“高峰值算力、不强调时延容忍度”的训练侧标准，彻底转向“高并发、低时延、高稳定”的推理侧标准；与行业早期以 AIGC（生成式 AI）为核心的算力需求相比，当前的算力流量，正在从“增量式的生成类短时爆发流量”，转向“存量业务场景下的用户并发式流量”，这类流量无明显波峰波谷、需要全天候高稳定性算力支撑。

从头部企业的场景落地细节，可以更清晰地看到行业级算力分配的典型逻辑：

- **字节跳动：**作为行业内算力资源投入规模最大的厂商，其内部不同芯片型号的算力资源，已完成明确的场景匹配划分 —— 高端海外 GPU 资源（H100/H200/H800 系列）已达到 100% 的利用率，无任何闲置资源，全部用于支撑核心业务的高并发场景：一是抖音、今日头条等核心产品的个性化推荐（包含召回、精排全链路环节），二是豆包多模态大模型的推理服务，三是 Seedance 文生视频大模型的持续研发迭代；中端的 A100、H20 芯片，则主要用于支撑豆包应用的日常高日活公域推理需求，以及部分相对非核心的视频生成场景的算力消耗。
- **阿里巴巴**：算力投入逻辑完全贴合电商核心业务的实时流量特征，推理侧算力资源占整体 AI 算力的 70%-75%，剩余 20%-25%投入在训练侧；这一分配比例的核心驱动，是淘宝、天猫、支付宝等核心平台的实时 AI 重构需求 —— 覆盖从用户进入 APP 后的首页个性化推荐，到搜索结果智能排序、广告精准投放，再到后续的内容智能审核等全链路场景。支撑这一流量的算力资源，同时覆盖阿里云对外输出的公共云专属算力服务集群，以及服务阿里体系内业务的专属算力集群；其中，部分高优先级的业务流量，会被直接调度至阿里平头哥自研芯片构建的算力集群。从业务回报来看，2026 年阿里云 AI 相关产品收入占比首次突破 30%，对应季度收入约 90 亿元，AI 年化收入达 360 亿元，正式宣告阿里 AI 业务迈入商业回报周期。
- **腾讯**：算力资源投入的核心支撑场景，集中在社交、广告、游戏三大核心业务的 AI 重构上，整体算力投入规模略低于阿里巴巴与字节跳动。其算力资源的采购与部署采用“一云多芯”战略：核心业务场景的流量，优先调度至基于海光 CPU 的服务器集群，以及适配燧原科技加速卡的定制液冷算力集群；这一集群架构，主要支撑混元大模型在微信、QQ 等社交产品的场景化落地，以及腾讯广告的智能投放、游戏业务的 AI 场景化渲染需求。从投入回报来看，腾讯 AI 相关业务收入，以金融科技及企业服务板块为核心支撑，2026 年相关业务收入规模达到 599 亿元。
- **百度**：作为国内最早布局 AI 算力基建的厂商，其算力投入逻辑围绕“搜索 + 内容安全 + 云业务外溢”的核心框架展开：核心搜索场景的 AI 重构、内容安全审核场景，是内部算力消耗的核心来源；基于昆仑芯搭建的百度天池 2.0 超节点算力集群，是支撑这一流量的核心底座，同时该算力集群还以 MaaS（模型即服务）的形式，对外提供商业化服务，实现算力资源的商业价值外溢。

从行业的发展趋势来看，ToC 场景的算力需求，将在现有基础上持续保持高速增长 —— 而驱动这一增长的核心因素，并非新业务场景的出现，而是现有核心场景的 AI 渗透率持续提升：以推荐场景为例，当前行业内头部厂商的推荐链路 AI 化比例已超过 80%，但仍存在部分可被技术优化的增量空间；在搜索、广告、内容审核等核心场景中，AI 技术的业务落地渗透率仍有较大提升空间。这意味着，后续行业算力投入的核心增量，将完全集中在推理侧场景；推理算力资源的规模大小、调度能力的强弱，将直接决定头部厂商的业务核心竞争力。

1.2 ToB：内部算力规模效应外溢，“使能型 AI”成为增长新引擎

随着头部厂商内部算力资源规模的持续扩大、集群运维能力的持续成熟，行业的算力资源发展逻辑，已从“先支撑内部业务冗余资源再外溢”的传统模式，转向“以云服务形式对外输出算力资源”的明确增长战略。这一转变的核心商业逻辑是：将内部算力基础设施的规模效应、已验证的算力调度经验，与政企客户的场景化需求结合，将算力资源从“成

本中心”的基建投入，直接转化为“利润中心”的云服务收入；在这一过程中，集群利用率的提升空间，是厂商决定算力外溢规模的核心依据。

从落地形态来看，互联网/OTT 厂商的算力外溢服务，并非单纯以“裸算力”的形式输出，而是将算力资源与模型能力、行业场景解决方案打包，形成“算力 + 模型 + 行业场景”的标准组合方案，这也是“使能型 AI”的核心价值体现。在这一输出过程中，不同厂商的算力服务底层架构，均与自身的技术生态特征强绑定：

- **字节跳动：**通过火山引擎云平台，以“超节点集群专属算力租赁服务”的形式，对外输出算力资源 —— 核心是将支撑豆包、抖音等核心业务的算力集群的稳定运维能力，转化为对外服务的核心竞争力。这类服务的主要客户，是需要高并发、低时延算力资源的中大型互联网企业，以及部分对算力稳定性要求较高的政企客户。
- **阿里巴巴：**通过阿里云平台，对外输出基于平头哥自研芯片的算力资源，同时将算力服务与模型服务结合，提供从“模型微调、训练到高并发推理”的全链路算力资源支撑方案；其核心客户覆盖国家电网、小鹏汽车、麦当劳等跨行业头部企业，这些客户不仅需要算力资源，更需要阿里体系的行业场景化经验支撑。
- **百度智能云：**依托基于昆仑芯的天池 2.0 超节点算力集群，重点输出针对大模型场景优化的 MaaS 解决方案 —— 由于百度在大模型生态中的技术积累，这类方案重点覆盖对模型适配要求较高的政企客户，将百度的模型技术积累，与客户的行业场景需求深度结合。
- **腾讯云：**采用“通用算力 + 国产算力池”的混合架构，对外提供 AI 算力服务；其中，针对政企信创场景的算力输出，是其核心增量来源 —— 腾讯将自身在金融、政务等行业的场景化积累，与适配国产芯片的算力集群的稳定运维能力结合，重点服务对数据安全性、合规性要求较高的政企客户。

从行业整体来看，ToB 端的算力外溢，已成为头部厂商抵消算力投入成本、获取增量业务收入的核心战略方向。这一趋势背后的产业逻辑是：由于互联网/OTT 行业头部厂商的算力集群规模，远大于普通政企客户的集群规模，其单位算力资源的综合成本、业务资源调度效率，也远高于政企客户的自建算力集群；对于政企客户而言，直接采购头部互联网厂商的成熟算力服务，比自行采购芯片、搭建集群、运维算力资源的综合成本更低，效率更高。这一逻辑，也将成为后续互联网/OTT 行业算力业务收入增长的核心驱动力。

2. 模型层：从“参数竞赛”到“成本效益竞赛”，算力需求结构化上移

2025 下半年 -2026 年，行业级模型的技术迭代逻辑，已从早期的“单纯比拼参数规模、追求技术指标先进性”，全面转向“以落地场景的成本效益为核心，重构模型架构” —— 这一转变的核心背景是，模型的参数量级提升，对业务效果的边际贡献正在持续减弱，而对算力资源的消耗成本却在指数级上升。这一背景下，“用得起、用得爽”的高性价比模型，成为行业内算力消耗的核心载体。

2.1 架构主流化：MoE 成为支撑超大规模流量的关键技术

从模型架构层面来看，行业内头部厂商的核心模型，已全面从“稠密参数架构”转向“稀疏 MoE（混合专家）架构”；这一技术路线选择的核心逻辑，是 MoE 架构的“高并发、低算力消耗”特性，与互联网/OTT 行业核心场景的推理算力需求完全匹配。

MoE 架构的技术优势，本质是“模型整体参数规模做大，单次推理实际激活参数规模做小” —— 具体而言：模型的总参数量级达到万亿级规模，足以支撑多模态、长上下文的复杂业务需求；但在实际业务推理过程中，仅激活部分专家参数，既保证了业务效果的先进性，又大幅降低了单次推理的算力资源消耗。这一技术特征，也恰好解决了互联网/OTT 行业核心场景对“高并发、低时延、高性价比”算力资源的核心需求。

在这一技术路线下，行业内头部模型的架构设计均采用类似逻辑：通过增加专家数量、优化专家激活机制，在提升模型业务效果的同时，严格控制单次推理的算力消耗。具有行业标杆意义的两款头部模型 ——DeepSeek V4-Pro 与 Kimi K3—— 的架构设计细节，清晰验证了这一行业趋势：

- **DeepSeek V4-Pro**：作为当前行业内的旗舰级商用模型，其总参数量级达到 1.6T（万亿）级，专家数量为 384 个；在实际业务推理过程中，仅激活 6 个专家，激活参数占比仅为 1.56%。这一激活比例，远低于同参数量级的传统稠密参数模型。这一架构设计的核心目标，是在支撑 1M 长上下文、多模态的复杂业务能力的前提下，将单次推理的算力消耗压缩至可控区间 —— 从实际落地效果来看，这一架构设计，能将相同业务场景下的推理算力消耗，较传统稠密参数模型降低约 40%。
- **Kimi K3**：作为行业内另一款标杆性 MoE 模型，其架构设计逻辑更偏向“超大规模参数支撑极致业务效果”的技术路线：总参数量级达到 2.5T 级，专家数量高达 896 个；但在实际业务推理过程中，仅激活 16 个专家，激活参数占比仅为 1.78%。这一架构设计，在支撑极致业务效果的同时，对算力集群的“高并发、低时延”能力，提出了更高的要求 —— 896 个专家的分布式并行需求，对集群内的互联带宽、通信时延，提出了比 DeepSeek V4-Pro 更高的技术要求。

值得注意的是，这两款头部模型均采用了“FP8/MXFP4 混合量化”的技术路线 —— 这是行业内兼顾推理精度与算力成本的最优技术方案：在推理过程中，对模型的不同计算环节，采用不同精度的量化策略：对精度要求较高的计算环节，采用 FP8（8 位浮点数）量化；对精度要求相对较低的计算环节，采用 MXFP4（4 位浮点数）量化。这一混合量化方案，在保证模型业务精度无明显损失的前提下，能将模型的显存占用率降低约 50%，直接将相同算力集群的实时并发吞吐量提升约 1 倍。这也是 MoE 架构模型，能在互联网/OTT 核心高并发场景顺利落地的核心技术前提。

这一架构选择，也直接决定了下层算力资源的选型标准：MoE 架构模型的分布式并行特征，对算力集群的“高互联带宽、低通信时延”能力，提出了远高于稠密参数模型的要求 —— 对于 MoE 模型而言，集群的互联带宽能力，对整体推理性能的影响权重，已经超过单芯片本身的峰值算力性能；这一技术逻辑，直接推动了行业算力底座向“超节点化”方向演进。

2.2 成本极致化：从“用得上”到“用得起”，成本下降推动算力上移

随着 MoE 架构的成熟、行业适配的深入，以及头部模型厂商的商业化策略调整，大模型推理的综合使用成本正迎来断崖式下降 —— 这一变化的直接结果，是行业内“算力资源需求规模”的增长速度，远高于“业务流量规模”的增长速度；这一现象的核心原因，是行业内的“算力资源性价比”大幅提升，推动了大模型在更多高并发业务场景下的渗透落地。

从行业公开的商业数据来看，头部模型厂商的推理 API 调用成本降幅，完全超出行业此前的预期：

2026 年 5 月，DeepSeek 官方宣布，V4-Pro 模型的 API 调用价格，在结束阶段性优惠活动后，正式调整为原定价的 25%；其中，针对高并发场景的缓存命中输入优化方案，调用成本仅为 0.025 元/百万 Token，这一价格已经接近行业内的“免费级”算力资源水平。作为 DeepSeek 云服务的核心落地伙伴，腾讯云第一时间跟进了这一降价策略，将 DeepSeek-V4 系列模型的调用价格最高下调 97.5%，进一步放大了行业的成本降幅。

同期，小米也宣布将旗下 MiMo 系列大模型的 API 调用价格最高下调 99%；国内三大运营商则同步推出了“算力订阅包”类的标准化长周期算力资源服务，其中部分针对推理场景的算力产品，价格低至 1 元/ 25 万 Token；在普惠算力政策的支持下，部分合规的政企客户，还可以通过算力券补贴、算力银行等模式，进一步降低实际算力采购成本。

这一降价潮的直接驱动因素，是 MoE 架构的技术成熟度提升，以及国产算力资源的大规模商用落地 —— 技术层面的优化，让单位算力资源的成本显著下降；而头部模型厂商的这一主动降价策略，本质是通过降低行业使用门槛，快速抢占行业内的技术生态标准，这也进一步放大了行业级算力成本的下降幅度。

算力成本的大幅下降，进一步反向驱动业务端的 AI 渗透率提升 —— 此前部分因算力成本过高而无法落地的 AI 业务场景，现在已具备明确的商业可行性；这也直接推动了行业级算力消耗的结构化上移。这一逻辑下，互联网/OTT 行

业的 AI 算力资源消耗结构，在 2026 年发生了根本性变化：推理侧算力消耗占行业总算力消耗的比例，从 2025 年的不足 50%，一举提升至 70%以上；其中，MoE 架构模型带来的增量推理算力需求，占行业总增量需求的比例超过 60%。

这一变化的核心行业影响是：算力资源的价值，从“支撑模型研发的技术基建储备”，直接转向“支撑业务高并发流量的生产制造资源”；相应地，算力资源的采购与选型逻辑，也从“技术性能优先”，彻底转向“单位算力成本、单位算力业务产能优先”—— 这是后续行业内算力资源重构的核心底层逻辑。

2.3 适配标准化：头部模型完成双轨化适配，支撑行业异构算力架构

为适配行业内的“海外高端 GPU + 国产算力”双轨算力架构，头部模型厂商在 2026 年，均完成了核心模型对“海外 GPU 生态 + 国产算力生态”的双轨适配支撑 —— 这意味着，同一模型的相同业务版本，可以在海外 GPU 的算力集群、国产算力的算力集群上，实现无差异业务落地。这一适配进展，直接为后续行业内的算力资源拆分、异构集群调度，提供了最关键的技术前提。

从具体的模型适配情况来看，头部厂商的核心模型，均已完成对国产算力的全链路适配验证：

- **DeepSeek V4-Pro**：在 2026 年 4 月，就完成了对华为昇腾、百度昆仑芯、阿里平头哥等主流国产芯片的全链路适配验证；这一适配过程的核心支撑，是国产推理框架生态的完善 —— DeepSeek V4-Pro 可以在 vLLM、SGLang 等主流开源推理框架的国产适配版本上，稳定运行在国产超节点集群上。从实际落地效果来看，DeepSeek V4-Pro 在昇腾 950 超节点集群上的推理性能，与同规模海外 GPU 集群的性能差距，已经缩小到 10%以内。
- **Kimi K3**：虽未公开完整的国产适配清单，但从行业公开的落地案例来看，其已在华为昇腾超节点集群上，完成了万卡级规模的集群落地验证 —— 可以在 384 卡及以上规模的昇腾超节点集群上，稳定支撑 1M 长上下文的高并发推理需求。这一适配的核心支撑，是昇腾超节点的高互联带宽、低时延全局通信能力，这是 Kimi K3 这类超大 MoE 模型能否在国产算力上落地的关键技术前提。

这一模型适配的技术标准，也对下层算力资源的技术选型，提出了明确的刚性要求：对于头部互联网/OTT 厂商的核心推理业务场景而言，只有能支撑 MoE 架构模型的高并发、低时延推理需求的算力资源，才具备被采购的核心价值；具体而言，算力集群必须满足三大核心条件：一是单芯片的显存容量足够支撑 MoE 模型的参数存储需求，二是集群内的互联带宽、通信时延足够支撑 MoE 模型的全链路并行需求，三是算力的密度、单节点功耗水平符合数据中心的实际部署约束。

3. 框架层：“去 CUDA 化”从行业共识进入规模化落地，生态条块分割现状持续

2025 下半年至 2026 年，行业级算力生态的核心变化，是“去 CUDA 化”从行业性的技术共识，正式转为头部互联网/OTT 厂商的实际落地技术路线 —— 这一趋势的核心驱动，并非 CUDA 生态本身的技术优势弱化，而是行业级算力资源向“国产算力池”迁移的刚性需求：国产算力池的核心芯片，均无法兼容 CUDA 生态，这要求上层的模型应用，必须从 CUDA 生态迁移至国产算力生态。

在这一过程中，行业的技术框架体系，逐步形成了“国产算力生态与海外 GPU 生态并行、上层推理框架标准化、下层芯片适配生态碎片化”的格局；这也是后续行业内异构算力资源调度面临的核心技术挑战。

3.1 去 CUDA 化：核心模型完成全链路迁移，国产框架实现性能追平

从技术落地进展来看，2026 年 4 月是行业级“去 CUDA 化”的关键时间节点 —— 在这一节点前后，头部互联网/OTT 厂商的核心大模型业务，均完成了从英伟达 CUDA 生态到国产算力生态的全链路迁移验证；这一迁移的核心支撑，是国产芯片厂商的软件栈与主流推理框架的适配性能，已经基本追上了 CUDA 生态的水平。

其中，华为昇腾的 CANN 软件栈（昇腾异构计算架构），是这次行业级迁移的核心支撑基础 ——CANN 软件栈的核心价值，是在底层屏蔽国产芯片的硬件差异，为上层的模型应用提供统一的编程接口，让开发者不必再关注底层芯片的硬件细节，大幅降低了迁移的技术成本。从行业落地情况来看，CANN 软件栈已经实现了对大模型主流算子的全覆盖，以及对上层主流推理框架的原生适配支撑；在实际业务场景中，基于 CANN 的算子融合优化效果，已经与 CUDA 生态的优化效果基本持平。

框架层的另一关键支撑，是主流推理框架对国产芯片的适配版本落地：作为行业内主流的推理引擎，vLLM 和 SGLang 对国产芯片的适配支持，是行业级迁移的最关键基础 —— 这两类框架，是上层模型应用与下层算力资源之间的核心技术枢纽，其对国产芯片的适配效率，直接决定了行业级迁移的技术成本。从实际落地情况来看，头部国产芯片厂商，均已完成了这两类推理框架的适配验证工作：

- 华为昇腾推出了 vLLM-Ascend 的专属适配分支，实现了对 vLLM 核心功能的原生支撑；
- 百度百舸基于昆仑芯的硬件架构，完成了对 SGLang、vLLM 的全链路适配；
- 阿里平头哥、寒武纪等头部国产芯片厂商，也均完成了对这两类主流推理框架的适配工作。

从实际落地效果来看，头部厂商的核心模型，在国产算力集群的主流推理框架上的推理性能，已经基本追平了 CUDA 生态的水平：

- DeepSeek V4-Pro 在昇腾 950 超节点集群上，通过 vLLM-Ascend 框架的定制化算子优化，其 1M 长上下文的推理吞吐性能，与 CUDA 生态下的同规模海外 GPU 集群性能差距，已经缩小到 10%以内；
- 百度的核心搜索模型，在基于昆仑芯的天池 2.0 超节点集群上，通过百度百舸对 SGLang 框架的定制化适配优化，其推理性能与 CUDA 生态下的海外 GPU 集群性能完全持平。

这一技术突破，是行业级算力资源重构的关键技术拐点 —— 在此之前，行业的技术共识是“国产算力资源的性能瓶颈，无法支撑核心推理业务流量”；而在这一技术突破之后，行业的技术共识已转变为“国产算力资源的性能，足以支撑核心推理业务流量”；这也直接推动了后续国产算力资源的大规模落地。

3.2 碎片化：分层解耦架构形成，行业适配成本和技术风险仍较高

尽管“去 CUDA 化”的技术落地进展，已经超出了行业内的普遍预期，但从技术架构的细节层面来看，行业的算力软件栈仍呈现“分层解耦、碎片化并行”的格局，整个生态体系的技术成熟度，仍远低于 CUDA 生态的成熟度。

这一碎片化格局的核心特征，是“框架层与芯片适配层完全分割”：行业内主流的大模型推理框架 —— vLLM、SGLang —— 均是独立的开源项目，并不专门针对国产芯片做原生适配；而国产芯片厂商的软件栈，又需要单独对这些主流推理框架做适配支撑 —— 这一分离架构，导致了整个适配链路的环节过长，额外技术成本和稳定性风险都比较高。

具体来看，这一碎片化格局的核心表现，集中在三个维度：

- **适配链路非标准化，额外技术成本居高不下**：国产芯片厂商对主流推理框架的适配方案，基本以“自研插件”或“定制化分支”的形式实现 —— 而非官方原生支持；这意味着，不同国产芯片厂商的适配技术方案存在明显差异，上层的模型应用，需要针对不同厂商的算力资源做单独适配开发，这直接增加了应用侧的开发成本。比如，标准的 vLLM 版本，无法直接在华为昇腾 NPU 上运行，必须使用华为官方适配的专属分支版本，同时需要配合定制化的算子库使用；对于上层开发者而言，这一适配过程的技术复杂度，远高于 CUDA 生态下的开发适配流程。
- **算子适配不完善，性能优化难度大**：部分国产芯片厂商对大模型高阶算子的支持度，仍存在明显短板；在实际业务场景中，这类算子需要用基础算子组合替代，这会直接增加推理的计算时延，损耗集群的整体吞吐性能。比如，昇腾芯片对 PyTorch 中的部分高阶算子支持不完善，直接迁移会出现性能异常；必须在迁移前，通过技术方案将这类未支持算子，替换为已支持的基础算子组合 —— 这一过程需要大量的定制化开发工作，直接增加了迁移的技术成本。此外，部分国产芯片的算子优化工作，仍停留在芯片厂商的技术支持层面，没有形成标准化的适配方案，这也进一步增加了适配的技术风险。

- **版本协同效率低，整体生态稳定性弱**：推理框架的版本升级节奏，与国产芯片软件栈的版本升级节奏，完全由不同的主体主导，存在明显的版本错配风险 —— 例如，部分基于 vLLM 0.23.0 版本开发的适配插件，无法兼容 vLLM 0.24.0 的官方版本；这意味着，上层的推理框架版本升级，可能会导致下层芯片的适配方案失效，整个适配链路的版本协同成本居高不下。更关键的是，不同厂商的算力资源的适配方案，没有形成统一的行业标准，这进一步放大了异构算力集群的整体运维风险。

这一碎片化格局的直接行业影响是：尽管国产算力资源在单芯片性能、集群互联能力上，已经能和海外 GPU 资源正面抗衡，但在实际业务场景中，国产算力资源的综合性能损耗，要显著高于海外 GPU 资源。根据行业内的公开实测数据，在相同的业务场景、相同的并发规模下，国产算力集群的综合性能损耗，是海外 GPU 集群的 2-4 倍；这一损耗的主要来源，就是软件栈适配层的技术短板。

这一现状也决定了，行业内的算力资源重构过程，必然是长期渐进式的：在后续的异构算力集群中，海外 GPU 资源会被用来支撑对性能、稳定性要求极高的核心业务场景；而国产算力资源，会被用来支撑部分对性能容忍度更高、对成本敏感性更强的业务场景 —— 而非直接一刀切式的替换。

4. 硬件底座：超节点成为核心形态，国产算力完成从“能用”到“好用”的关键跨越

2025 下半年至 2026 年，互联网/OTT 行业 AI 算力的硬件底座，完成了一轮彻底的重构 —— 这一重构的核心逻辑，是由 MoE 架构模型的分布式并行需求决定的：MoE 模型的万亿级参数规模、高并发推理需求，对算力集群的“互联带宽、通信时延、有效利用率”提出了极致要求；传统的“多服务器通过以太网组网”的集群架构，已经无法支撑这一需求，行业的硬件底座，必须从“单芯片堆叠的集群架构”，升级为“以互联为核心的超节点架构”。

在这一重构过程中，国产算力资源实现了关键的技术跨越：从过去“单芯片性能不足、集群互联能力弱、只能作为补充算力资源”的被动状态，转变为“以超节点架构为核心，在推理场景下具备与海外 GPU 资源正面抗衡的能力”—— 这是行业级算力资源重构的最关键物质基础。

4.1 架构范式转移：从“拼单卡峰值算力”到“拼超节点系统级有效 Token 吞吐”

WAIC 2026 是行业级算力硬件重构的关键标志性节点：在本届大会上，国内外头部算力芯片厂商，均发布了面向互联网/OTT 行业核心场景的“超节点”级算力方案 —— 超节点并非简单将多颗芯片堆叠在一起，其本质是通过自研的高速互联协议，将成百上千颗芯片连接为“单一逻辑计算集群”，在集群层面实现“高互联带宽、低通信时延、高全局通信效率”的核心能力。这一技术方案的核心目标，就是针对性解决 MoE 架构模型对集群互联能力的极致要求。

这一技术选择背后的行业逻辑是：对于 MoE 架构模型的性能支撑而言，集群的整体互联带宽能力，对实际推理性能的影响权重，已经超过了单芯片本身的峰值算力性能；如果集群的互联能力存在短板，单芯片的峰值算力再高，也无法转化为实际的推理吞吐性能。这一逻辑下，行业的算力竞争维度，已经从“单芯片的理论峰值算力”，彻底转向“超节点集群的系统级有效 Token 吞吐能力”—— 这一指标，是集群的单芯片算力、互联带宽、通信时延、软件栈优化能力等综合技术能力的集中体现。

从 WAIC 2026 厂商发布的技术细节来看，头部国产厂商的超节点方案，均采用了“自研互联协议 + 全光对等互联拓扑”的技术路线；其中，华为昇腾的 Atlas 950 SuperPoD 超节点方案，是行业内的标杆性产品 —— 其技术设计细节，完整代表了行业级超节点的标准技术方向：

- **互联层面**：采用了自研的灵衢 2.0 互联协议，以及 UB-Mesh 全光对等互联拓扑结构；单柜内的互联总带宽达到 16.3PB/s，跨芯片的通信 RTT（往返时延）仅为 3μs，这一指标是行业内的顶尖水平；这一互联能力的核心价值，是最大化降低 MoE 模型的分布式并行通信时延，将集群内的通信损耗，降低至传统集群的 1/15 以内。

- **规模层面**：以 64 卡机柜为基本单元，采用模块化的柜间互联设计，标准集群规模可支持 1024 张 NPU 卡高速互联；在实际落地场景中，可根据业务需求，进一步扩展至 8192 卡的超大规模集群规模 —— 这也是当前行业内，唯一能支撑万亿级参数 MoE 模型全链路并行的超节点方案。
- **算力层面**：在 FP8 精度下，总算力可以达到 1 EFLOPS；在 FP4 精度下，总算力可以达到 2 EFLOPS —— 这一算力规模，足以支撑万亿级参数 MoE 模型的高并发推理需求；同时，超节点内的全局统一内存编址空间达到 256TB，这一超大容量的全局内存空间，可以支撑 MoE 模型的海量 KV Cache 存储需求，进一步降低长上下文场景下的推理时延。

百度在 WAIC 2026 上发布的天池 2.0 超节点方案，是行业内的另一典型落地代表：该方案基于昆仑芯自研芯片，采用了“32 卡单机柜互联架构”，相比上一代产品，卡间互联带宽提升了 4 倍，集群的整体推理性能直接提升了 50%；单个超节点集群，即可支撑万亿级参数 MoE 模型的高并发推理需求。这一方案的核心特点，是采用了“相对小规模的互联单元、大集群级分层互联”的技术路线 —— 这一架构的核心优势，是可以根据业务的实际流量规模，灵活调整集群的互联单元规模，在支撑 MoE 模型并行需求的同时，更好地平衡成本；这一方案，更适配互联网/OTT 行业内中低并发业务场景的实际算力需求。

阿里云在 WAIC 2026 上发布的磐久 AL128 超节点方案，则是“自研芯片 + 标准超节点架构”落地路径的典型代表：该方案基于阿里平头哥自研的真武系列 AI 芯片，将 128 颗 AI 芯片通过高速互联技术整合为单一逻辑计算集群；结合平头哥自研的芯片架构技术，该方案可实现与主流超节点方案持平的互联性能，同时在单位算力成本、功耗控制上实现了优化 —— 这一方案，主要支撑阿里云对外输出的高并发推理算力服务，重点覆盖有平稳化算力需求的互联网客户与政企客户。

从实际落地效果来看，头部国产厂商的超节点方案，已经在核心性能指标上，具备了与海外厂商主流超节点方案正面抗衡的能力 —— 这一判断，并非基于实验室的理论数据，而是基于行业内的实际业务实测数据。华为在 WAIC 2026 上，首次公开了昇腾超节点方案与海外同类高规格方案的行业级对比实测数据：在 FP8 精度、MoE 架构模型的高并发推理场景下，384 卡规模的昇腾超节点集群，实际业务吞吐性能是海外同类 NVL72 方案的 1.7 倍；这一数据，是行业内首次有国产算力方案，在核心业务场景下，实现了对海外标杆方案的性能反超。

这一技术突破的核心价值，是解决了国产算力资源在互联网/OTT 行业核心场景下的“性能信任问题”—— 在此之前，行业的普遍共识是，国产算力资源的性能瓶颈，无法支撑核心业务流量；而在这一技术突破之后，行业的技术共识，已经转变为国产算力资源的性能，足以支撑核心业务流量；这是后续行业内算力资源重构的最关键基础。

4.2 架构资源池化：三层异构资源池成为行业标准算力底座

随着超节点技术架构的成熟，互联网/OTT 行业的算力资源部署架构，从此前的“单一算力集群支撑所有业务场景”的简单架构，升级为“分层异构资源池”的标准架构 —— 这一架构的核心逻辑，是根据不同业务场景的算力需求特征，将其调度到最适配的算力资源层，在保证业务体验的前提下，最大化提升集群的综合利用率；这一架构设计的核心目标，是解决“异构算力资源与不同类型业务场景”的精准匹配问题。

从行业的落地情况来看，头部互联网/OTT 厂商的三层异构资源池架构，已经形成了明确的行业标准，各层的技术定位、承载业务场景、选型标准高度清晰：

- **第一层：海外高端 GPU 资源池**：这一资源池的核心定位，是支撑对算力性能、稳定性要求极高的核心业务场景 —— 其技术选型标准，是单芯片的峰值算力、显存容量、集群互联性能，均是行业顶尖水平；只有行业内最高端的海外 GPU 产品，能满足这一资源池的技术要求。具体来看，这一资源池的承载场景，主要分为两类：一是头部厂商的核心模型的前沿训练、大规模微调场景 —— 这类场景对算力的精度、稳定性要求极高；二是部分对时延极其敏感的顶流业务高并发推理场景 —— 这类场景，需要算力资源提供稳定的极低时延支撑。从实际落地情况来看，头部厂商的这一资源池，主要由英伟达的 H100、H200 系列高端芯片支撑；由于供应链配额限制，这类资源池的规模相对有限，仅用来支撑最核心的业务场景。

- **第二层：国产超节点算力资源池**：这一资源池的核心定位，是支撑行业内占比最高的大规模高并发推理场景，以及部分中低规模的模型训练、微调场景 —— 这是行业内算力资源增量的核心来源；其技术选型标准，是“超节点架构的系统级有效 Token 吞吐能力”，而非单纯的单芯片峰值算力。具体来看，这一资源池的承载场景，覆盖互联网/OTT 行业核心的搜广推场景、高并发多模态交互场景、长文本内容生成场景，以及部分行业模型的训练、微调场景。从实际落地情况来看，这一资源池的核心支撑，是华为昇腾、百度昆仑芯、阿里平头哥的超节点方案；部分头部厂商的这一资源池，还采用了“混合多厂商超节点”的架构，通过统一的调度层，将不同厂商的超节点资源池进行统一管理，调度至不同的业务场景。
- **第三层：CPU、存储与低算力工具资源池**：这一资源池的核心定位，是支撑算力需求相对较轻的辅助类业务场景 —— 这类场景的核心特征，是高并发、低算力消耗，不需要 GPU/NPU 这类高端算力资源支撑。具体来看，这一资源池的承载场景，主要分为三类：一是 Agent 类业务场景的并发支撑；二是长上下文场景下的检索、数据库查询、浏览器访问等辅助计算场景；三是业务工作流中的低算力消耗的通用计算场景。从技术选型来看，这一资源池的技术方案，以通用的高主频 CPU、分布式存储资源为核心，配合轻量化的计算框架；这类资源的技术成熟度高、成本低，是支撑这类场景的最优选择。

在这一三层架构中，**第二层的国产超节点算力资源池，是行业内算力资源增长的核心增量来源** —— 这一趋势的核心支撑，是国产超节点方案在“性能、成本、合规性”三个维度的综合竞争力，已经超过了海外 GPU 资源：从性能维度来看，国产超节点的实际业务吞吐性能，已经可以和海外高端 GPU 资源持平；从成本维度来看，国产超节点资源的单位算力成本、运维成本，均显著低于海外高端 GPU 资源；从合规性维度来看，国产超节点资源可以满足行业内的供应链安全要求，这是海外高端 GPU 资源无法替代的。

更关键的是，在这一架构中，不同层级资源池之间的调度接口，已经形成了行业标准；上层的业务流量，可以根据实际的并发规模、时延容忍度、算力成本控制要求，在三个资源池之间完成灵活调度 —— 这一架构的综合资源调度效率，比传统的单一算力集群架构，提升了至少 30%。

4.3 国产替代拐点：从“被动补充”到“主动选择”，推理场景成为主赛道

2026 年是国内互联网/OTT 行业算力国产替代的关键拐点 —— 国产算力资源的市场份额、落地场景深度，均实现了量级式突破；这一突破的核心驱动，是国产超节点方案在推理场景的成熟，以及行业级算力迁移的明确需求。

从市场份额层面来看，根据 IDC 与中国半导体行业协会联合发布的 2026 年一季度行业数据，国内 AI 加速芯片的市场份额格局，已迎来历史性变化：国产芯片的整体市场占比达到 52.3%，首次超过海外芯片的市场份额；这一数据背后的支撑，是国产算力资源在推理场景下的大规模落地 —— 这也是行业内，国产算力资源首次实现市场份额反超。

这一格局中，国产阵营内部的马太效应明显：华为昇腾以 20%的市场份额，稳居国内国产算力赛道的行业第一；阿里平头哥凭借真武系列 AI 芯片的大规模出货，以 7%的市场份额位居国产第二，超过了此前的行业头部厂商寒武纪；百度昆仑芯、寒武纪、天数智芯等厂商紧随其后，共同构成了国产算力的第一梯队。这一格局的背后，是厂商的“全栈能力、落地规模、商用稳定性”的综合比拼 —— 只有具备全栈技术能力、大规模集群落地案例、成熟商业交付能力的国产芯片厂商，才能进入互联网/OTT 行业的核心采购清单。

更关键的变化，是国产算力资源的定位的根本性提升：从过去的“海外 GPU 资源的被动补充”角色，转变为“互联网/OTT 行业核心推理场景的主动选择”—— 这一趋势的核心判断依据，是头部互联网/OTT 厂商的算力资源采购结构，发生了根本性变化：根据券商产业链调研的公开数据，2026 年头部互联网/OTT 厂商的算力资源采购预算中，国产算力资源的占比已经达到 50%；部分厂商的推理算力资源采购预算中，国产算力资源的占比甚至超过了 60%。

这一采购份额的分配逻辑，与三层资源池的架构设计完全匹配：

- 在训练场景下，国产算力资源与海外 GPU 资源，仍存在一代以上的技术差距；头部厂商的核心模型的前沿训练场景，仍主要依赖海外高端 GPU 资源支撑。

- 但在推理场景下，国产算力资源的综合竞争力，已经超过了海外 GPU 资源 —— 这一综合竞争力，不仅体现在性能、成本层面，更体现在供应链安全、本地化技术服务、政企合规性要求等长期隐性维度；这是国产算力资源，在推理场景下被大规模采购的核心原因。

从实际落地规模来看，国产算力资源在互联网/OTT 行业的核心推理场景，已经实现了超大规模的落地验证：根据华为官方披露的信息，截至 2026 年 4 月，仅华为昇腾的超节点方案，就已在国内互联网行业落地了超过 750 套万卡级集群；其中，字节跳动采购了 25 万片昇腾 950 系列芯片，阿里采购了 15 万片昇腾 950 系列芯片，这些芯片全部被用于构建核心推理场景的国产算力资源池；百度的昆仑芯超节点方案，也在自身的核心搜索业务、对外的 MaaS 服务中，落地了超过 300 套万卡级集群。这一落地规模，意味着国产算力资源，已经从技术验证阶段，正式进入了大规模商用的成熟阶段。

5. 厂商策略：构建异构算力混合调度能力，行业进入全栈化较量

2025 下半年至 2026 年，互联网/OTT 行业算力的竞争维度，从“单一性能比拼”，升级为“全栈化能力较量”—— 这一逻辑的核心是，行业的算力资源架构，已经从“单一芯片堆叠”，转向了“异构超节点资源池 + 统一调度层”；这一架构下，单纯的芯片性能优势，已经无法转化为实际的业务优势；行业的核心竞争维度，变成了“从芯片层到应用层的全栈协同能力”。

在这一背景下，互联网/OTT 行业的头部玩家，已明确了“异构算力混合调度”的标准策略逻辑 —— 核心是“分场景选型、分流量调度、最大化集群利用率”；具体而言，是将“海外高端 GPU 资源、国产超节点资源”，分别与不同类型的业务场景精准匹配，构建统一的异构算力资源池；在此基础上，通过自研的集群调度能力，将不同优先级的业务流量，动态调度到成本、性能、适配性最匹配的算力资源层，实现业务体验与算力成本的平衡。

5.1 互联网/OTT 头部大厂：“海外高端 GPU+ 国产超节点”双轨制

对于头部互联网/OTT 厂商而言，双轨制算力策略的核心目标，并非简单实现“国产算力替代”，而是构建“算力资源的供应商冗余度”，从根源上规避供应链风险；同时，通过异构算力资源的分层调度，最大化降低算力的综合成本。这一策略下，头部厂商的采购逻辑，已经从过去的“单一性能优先”，转向了“综合成本效益、适配场景能力、供应链安全能力”的三维标准；具体的采购、部署、调度策略，已经形成了清晰的行业标准：

- **采购策略：双轨并行，分场景配额明确**

头部厂商的算力资源采购清单，均分为两大独立板块，分别对应不同的资源池层级：

一是海外高端 GPU 资源板块 —— 这一板块的采购核心标准，是单芯片的峰值算力、显存容量、集群互联性能，均是行业顶尖水平；采购配额，严格按照“支撑前沿训练、核心推理高优先级流量”的需求来核定。

二是国产超节点算力资源板块 —— 这一板块的采购核心标准，是超节点的系统级有效 Token 吞吐能力，而非单纯的单芯片峰值算力；采购配额，严格按照“支撑大规模低优先级推理流量、模型微调、非核心训练”的需求来核定。

这一采购结构，直接支撑了三层异构资源池的架构落地。从实际采购情况来看，不同头部厂商的采购偏好，与其业务场景的特征强绑定：

字节跳动作为行业内最大的算力消耗厂商，在采购英伟达高端 GPU 资源的同时，成为华为昇腾超节点的最大互联网客户；

阿里同时采购英伟达高端 GPU 资源，以及平头哥自研芯片的超节点资源，构建了“海外 GPU+ 自研芯片”的双轨架构；

百度则在采购英伟达高端 GPU 资源的同时，优先采购昆仑芯超节点资源，支撑自身的核心搜索业务与对外 MaaS 服务；

腾讯则同时采购英伟达高端 GPU 资源，以及海光 CPU+ 燧原加速卡的国产超节点资源，以支撑自身的金融科技、云服务对外输出场景。

- **部署策略：训推分离、流量分层，精准匹配算力资源**

头部厂商的核心技术逻辑，是将业务流量按照不同维度的特征进行精准拆分，然后将不同类型的流量，精准部署到适配的算力资源层；这一部署逻辑的核心，是最大化每一层算力资源的利用率，平衡业务体验与算力成本。从具体部署细节来看，头部厂商的流量拆分逻辑，在行业内形成了明确的标准：

一是按照“**训练/推理**”维度进行流量拆分：将大模型预训练、行业模型深度微调的流量，全部部署到海外高端 GPU 资源池；将大规模高并发推理、轻量化模型微调、简单训练的流量，全部部署到国产超节点资源池。

二是按照“**业务优先级**”维度进行流量拆分：将高优先级、对时延极其敏感的核心业务流量，调度到海外高端 GPU 资源池；将低优先级、时延容忍度相对较高的大规模业务流量，调度到国产超节点资源池。

三是按照“**场景技术特征**”维度进行流量拆分：将部分对互联带宽要求极高的 MoE 模型推理流量、长文本业务流量，优先调度到适配的国产超节点资源池 —— 部分国产超节点的互联性能，已经超过了海外高端 GPU 资源的集群水平。

- **调度能力：自研异构算力调度层，支撑混合集群管理**

要实现这一双轨制策略的落地，核心支撑是异构算力集群的统一调度能力 —— 这是头部互联网/OTT 厂商，在过去几年里持续投入的核心技术方向，也是其算力综合能力的关键差异点。

从技术细节来看，头部厂商的调度层核心能力，是“对不同厂商架构的异构算力资源的统一集群化管理” —— 将底层不同厂商的算力资源池，抽象为统一的逻辑资源池；在此基础上，基于流量的并发规模、时延容忍度、成本控制目标、资源利用率阈值等核心指标，通过智能调度算法，实现业务流量的动态调度与资源的自动扩缩容。例如：

字节跳动在火山引擎上部署的异构算力调度系统，可以实现“根据业务流量的实时并发规模，在 1 分钟内完成国产超节点资源的横向扩缩容”；

阿里的算力调度系统，可以“根据流量的来源优先级、实时算力资源的利用率阈值，将不同用户的请求，直接转发到对应层级的算力资源池”；

腾讯的算力调度系统，可以“实时监测不同算力资源层的负载情况，将低优先级的业务流量，从高成本资源池，迁移至低成本资源池”。

这一调度能力的成熟度，直接决定了异构算力集群的综合利用率；行业内的公开数据显示，头部厂商的异构算力集群综合利用率，已经比单一算力集群架构提升了至少 30%。

5.2 国产芯片厂商：全栈协同，锚定互联网推理场景核心需求

在 2025 下半年至 2026 年，国产芯片厂商的策略逻辑，从过去的“单纯比拼芯片硬件性能”，升级为“全栈能力协同，贴合互联网/OTT 行业实际场景需求” —— 这一转变的核心逻辑，是头部互联网/OTT 厂商对算力资源的选型标准，已经从“单一性能优先”，转向了“全栈适配能力、场景化支撑能力优先”；国产芯片厂商要进入头部互联网/OTT 厂商的采购清单，不能只提供高算力的芯片，而是要提供“从芯片层到应用层的全栈解决方案”。

在这一背景下，国产头部芯片厂商，均围绕“互联网/OTT 行业核心推理场景的适配”，制定了明确的全栈化落地策略；其核心逻辑，是“以超节点架构为核心，以软件栈优化为重点，以场景化适配为突破口” —— 通过“硬件 + 软件 + 行业场景”的全栈能力，将自己的超节点方案，嵌入到头部互联网/OTT 厂商的算力资源体系中；不同厂商的策略差异，集中在技术路线选择与行业场景卡位上：

- **华为昇腾：全栈封闭路线，构建超节点生态壁垒**

作为国产算力阵营的绝对龙头，华为昇腾的策略核心，是依托“自研芯片 + 自研互联 + 自研软件栈”的全栈技术能力，构建“超节点级”的技术壁垒 —— 这一壁垒的核心支撑，是其独有的“灵衢互联协议 + CANN 软件栈 + Atlas 超节点架构”的全栈组合；这一组合，是行业内唯一能支撑万亿级参数 MoE 模型大规模并行的国产方案，完全匹配头部互联网/OTT 厂商的核心推理场景需求。

在具体落地过程中，华为昇腾的核心打法，是将这一全栈能力，打包为“标准化超节点算力方案”交付给客户：集群的硬件架构、互联配置、软件栈优化、模型适配、运维体系，均由华为昇腾提前完成标准化适配验证；互联网厂商采购后，无需再做复杂的适配工作，可以直接将超节点资源接入到自己的调度层，快速完成业务流量的迁移。

这一方案的优势，是极大降低了互联网厂商的集群适配成本；这也是华为昇腾在行业内落地规模最大的核心原因。

- **百度昆仑芯：开放生态路线，锚定云厂商 MaaS 输出场景**

百度昆仑芯的策略核心，是依托百度在大模型生态中的技术积累，以及百度智能云的行业客户触达能力，重点适配“互联网/OTT 行业对外 MaaS 服务场景”的算力需求 —— 这一场景的核心技术要求，是“超节点架构的标准化程度高、适配主流模型的成本低、便于对外规模化交付”；百度昆仑芯的技术方案，恰好贴合这一场景需求。

在具体落地过程中，百度昆仑芯采用了“开放生态、联合适配”的打法：重点与主流大模型厂商、推理框架厂商合作，提前完成在昆仑芯上的模型适配验证；将适配好的模型，与超节点算力资源打包组合为“行业场景化算力方案”，然后通过百度智能云的对外 MaaS 服务渠道，交付给有需要的行业客户；而非直接向头部互联网厂商销售裸芯片或裸超节点资源。

这一策略的核心优势，是将百度的大模型技术积累，与超节点算力资源的交付能力结合，最大化降低了客户的迁移成本；这也是昆仑芯在行业内，能占据一席之地的核心原因。

- **阿里平头哥：自研自用 + 行业输出双线，支撑阿里生态内外场景**

阿里平头哥的策略核心，是“支撑阿里生态内的业务算力需求 + 对外规模化商用交付”双线并行 —— 这一布局，是由阿里的业务生态特征决定的：作为头部互联网厂商，阿里自身的业务算力需求规模，就足以支撑平头哥的芯片量产规模；同时，阿里平头哥的算力方案，还可以通过阿里云的对外服务渠道，交付给行业客户。

在具体落地过程中，阿里平头哥的打法，是基于真武系列 AI 芯片，构建“自研芯片 + 超节点架构”的全栈算力方案：这一方案，在支撑阿里核心业务的同时，还以“专属算力集群租赁服务”的形式，通过阿里云对外输出 —— 面向行业客户的方案，是将阿里云的行业场景交付能力，与平头哥的超节点算力资源打包组合，重点覆盖有平稳化算力需求的互联网客户与政企客户。

这一策略的核心优势，是通过阿里云的行业客户触达能力，快速落地规模化的外部案例，规避单一行业的周期风险；这也是平头哥能在行业内快速赶超其他国产芯片厂商的核心原因。

- **寒武纪、天数智芯等厂商：走开放兼容路线，以性价比抢占次级业务场景**

这类厂商的策略核心，是依托“芯片性价比高、开放兼容主流生态”的优势，重点抢占“头部互联网/OTT 厂商的次级非核心推理场景”的算力资源份额 —— 这类场景的特征，是对成本敏感性更强，对极致性能的要求相对较低；这类厂商的方案，在满足这类场景的技术要求的前提下，具备明显的成本优势。

在具体落地过程中，这类厂商的核心打法，是重点适配开源的主流推理框架，将其芯片的适配方案，嵌入到互联网厂商的异构算力调度层中；以“低成本、低迁移成本”的核心竞争力，获取头部互联网厂商的次级算力资源采购配额；而非直接对接核心推理场景。

这一策略的核心优势，是避开了与华为昇腾等头部厂商的直接技术竞争，依靠成本优势，获取行业内的增量市场份额。

5.3 英伟达：高端 + 特供双线并行，守住行业基本盘

作为全球算力生态的行业龙头，英伟达的策略核心，是通过“高端技术生态领先 + 特供版本合规交付”的双线并行方案，守住国内高端算力资源市场的基本盘 —— 这一方案的背后，是其对国内市场供应链环境的精准判断：一方面，通过高端产品的技术代差优势，牢牢占据国内市场的高利润高端需求；另一方面，通过特供产品的合规性，规避供应链限制，维持在国内市场的大规模份额。

这一策略的具体细节，完全匹配国内互联网/OTT 行业的算力资源分层需求：

- **高端产品线：技术代差领先，锁定高利润核心场景**：通过 Blackwell、H200 这类不受到国内供应链限制的高端产品，以及 NVLink、NVSwitch 构建的超节点架构，维持在技术生态上的绝对领先优势 —— 这类产品的技术性能，是当前国产算力资源无法完全替代的；其核心定位，是支撑头部互联网/OTT 厂商的前沿模型训练、高优先级核心推理场景，以及部分对性能要求极高的私有化客户场景。这一产品线，是英伟达维持在国内市场高端份额的核心基础 —— 尽管其市场规模占比不高，但贡献了绝大部分的利润。
- **特供产品线：合规性优先，卡位中低端推理场景**：通过 H20、这类合规性特供芯片，搭配英伟达的超节点架构方案，以“高性价比、适配成熟 CUDA 生态”为核心竞争力，卡位头部互联网/OTT 厂商的中低端推理场景 —— 这类场景，是国产算力资源的主要目标市场；这类产品的技术性能，与国产算力资源的差距较小，而由于其 CUDA 生态的成熟度更高，互联网厂商的迁移成本更低；这一策略，有效限制了国产算力资源在中低端推理场景的渗透速度。

值得注意的是，英伟达的这一策略，并非单纯的“产品布局”；而是在生态层，与头部互联网/OTT 厂商的“现有 CUDA 生态存量资源”深度绑定：通过提供定制化的软件适配工具、集群调度解决方案，降低互联网厂商切换到国产算力资源的综合成本 —— 这一综合成本，包括了业务迁移的技术成本、风险成本、长期适配成本等。

这一生态绑定策略，是英伟达在国内市场，仍占据超过 40%份额的核心原因；尽管其市场份额在持续下滑，但仍在行业内的高端算力市场，维持着难以撼动的地位。

6. 趋势展望：超节点化作为行业标配，算力架构进一步解耦异构

基于 2025 下半年至 2026 年的行业发展现状，结合头部厂商的公开技术规划，未来 12-24 个月，互联网/OTT 行业 AI 算力将呈现三大确定性发展趋势，每一个趋势都将对行业的现有算力架构产生深远影响。

6.1 算力架构：超节点 + 池化异构调度成为行业标配，集群级工程能力成为比拼焦点

未来 12-24 个月，行业级算力架构的演进方向，将是“超节点化 + 资源池化 + 异构算力混合调度”的深度融合 —— 这一架构，将从当前头部厂商的试点落地状态，直接升级为行业内的标准算力底座；这一趋势的根本驱动，是 MoE 架构模型的的规模化普及，以及行业级异构算力的规模化落地。

这一架构的核心技术形态，将在现有基础上进一步升级 —— 从“异构资源池的技术标准适配”，升级为“统一超节点层 + 异构算力池”的技术架构：

- **底层资源池**：将“海外高端 GPU 超节点、国产超节点”，通过行业标准化的互联技术方案，整合为“单一逻辑超节点资源池”—— 在集群的物理资源层面，不再区分“海外算力”或“国产算力”，而是以“超节点集群的逻辑资源块”为最小管理单位；不同厂商的超节点资源，被抽象为统一规格的逻辑资源块。
- **调度层能力**：在异构算力调度层的基础上，进一步升级“统一超节点调度层”能力；调度层的核心粒度，将从“服务器级”，升级为“超节点集群级”。在这一架构下，业务流量的调度逻辑，将不再针对某一种特定算力资源，而是根据流量的特征、优先级，以及超节点资源池的整体负载情况，在“不同厂商异构超节点资源池”之间，自动完成最优调度选择。

这一架构的核心技术突破点，将集中在“异构超节点的统一互联技术”上 —— 这一技术方案，将以“光互联技术”为核心，构建“跨厂商超节点集群的统一互联 fabric”；这一方案的核心价值，是将不同厂商的超节点资源池之间的通信时延，降低至“微秒级”，彻底屏蔽不同厂商超节点之间的通信壁垒，实现异构超节点资源的真正融合。

从行业落地情况来看，头部厂商已经启动了这一技术方向的预研：华为在 WAIC 2026 上，展示了支持多厂商超节点集群互联的光互联原型方案；阿里平头哥在 2026 年 4 月，联合行业内头部光互联技术厂商，启动了“多厂商超节点互联技术”的行业标准制定工作。

这一架构的成熟度提升，将彻底改变行业级算力的比拼维度：从“比拼芯片性能”，转向“比拼集群级系统工程整合能力”—— 包括集群的互联架构设计、异构算力的调度效率，以及运维的自动化水平、抗风险能力等综合指标。这类能力的构建周期，远长于单一芯片的采购周期；这也意味着，头部厂商的算力壁垒，将从“芯片采购壁垒”，升级为“集群架构整合能力壁垒”；中小规模互联网厂商，将很难再通过“采购高端芯片”的方式，在算力层面追上头部厂商的节奏。

6.2 应用模式：从“AI 应用适配算力”到“算力适配 AI 应用”，垂直场景整合度持续提升

2025 下半年至 2026 年，行业内的算力资源部署逻辑，已经完成了一次根本性的方向转变 —— 从过去的“以算力集群技术能力为中心，要求上层业务应用适配底层算力集群”，彻底转向“以 AI 业务应用的场景需求为中心，将下层算力资源整合成适配业务需求的标准能力块”；这一转变的核心逻辑，是算力资源从“技术基建储备”，升级为“支撑业务高并发流量的生产制造资源”；这一趋势，将在未来 12-24 个月内，进一步向纵深发展。

这一转变的核心落地形态，将是“场景化算力产品的标准化输出”—— 头部厂商的异构算力资源池，将不再以“裸算力”的形式输出，而是与上层的模型应用、行业场景方案打包，形成“适配特定场景的标准化算力产品”；这类产品的核心特征，是直接面向行业场景的业务需求，定义算力资源的技术规格；用户无需再关注底层的算力资源选型、适配、部署、运维细节，即可直接获取适配业务场景的稳定算力支撑。

从行业的落地情况来看，这类场景化算力产品的标准形态，已经清晰可见：

- 针对互联网/OTT 行业核心的搜广推场景，将推出“高并发低时延推理专用算力产品”—— 其技术规格，将直接匹配搜广推场景下的高并发、低时延、大流量特征；
- 针对长上下文场景，将推出“长文本交互专属算力产品”—— 其技术规格，将直接匹配长文本场景下的海量 KV Cache 存储需求、高 Session 并发请求特征；
- 针对多模态交互、视频生成场景，将推出“多模态生成专属算力产品”—— 其技术规格，将直接匹配多模态场景下的高带宽、大显存特征；
- 针对行业模型的训练、微调场景，将推出“轻量化训练专属算力产品”—— 其技术规格，将直接匹配行业模型训练、微调场景下的中低算力规模、高稳定性特征。

这一趋势的核心影响，是将进一步放大头部互联网/OTT 厂商的算力资源整合优势 —— 头部厂商的算力资源整合能力，将从“支撑自身内部业务”，升级为“对外输出行业级场景化算力方案”；其算力资源的商业价值，也将从“降低自身业务成本”，升级为“获取行业增量收入”。而中小规模互联网厂商、政企客户，将不再需要自行投入大量资金，构建和运维异构算力集群；而是可以直接采购头部厂商的标准化场景化算力产品，将自身的技术资源集中在业务场景差异化落地的环节上。

6.3 国产替代：从“补充算力资源”到“成为行业主流算力资源”，进入全场景深度替代阶段

2026 年，国产算力资源在互联网/OTT 行业的推理场景，已经完成了从“被动补充”到“主动选择”的关键跨越；未来 12-24 个月，这一替代趋势，将进一步向纵深发展，从“非核心推理场景”的替代，升级为“核心推理场景 + 部

分训练场景”的深度替代 —— 这一趋势的核心驱动，是国产超节点方案在性能、成本、合规性三个维度的综合竞争力，将进一步提升；而海外高端 GPU 资源的供应约束、成本压力，将持续放大。

这一趋势的核心支撑，是国产超节点方案在关键技术维度的持续迭代：

- **硬件层**：国产超节点方案的单芯片算力、互联带宽、全局通信效率，将在现有基础上提升至少 50%；部分国产超节点方案的核心性能指标，将反超海外高端 GPU 资源的方案；这一技术突破，将支撑国产算力资源，进一步渗透到头部厂商的核心业务场景。
- **软件层**：国产算力资源的软件栈生态短板，将得到根本性补齐 —— 头部国产芯片厂商，将完成对主流大模型的高阶算子适配、融合优化；对 vLLM、SGLang 等主流推理框架的适配细节，将完全并入官方上游版本，不再需要单独定制适配插件；这一优化，将把国产算力资源的综合性能损耗，降低至与海外 GPU 资源持平的水平。
- **生态层**：国产超节点方案，将与头部互联网/OTT 厂商的业务场景，完成深度适配验证；头部厂商的核心模型，将在国产超节点方案上，完成全链路性能压测的标准化验证；这一验证，将从根本上消除互联网厂商对国产算力资源的“商业稳定性顾虑”。

从行业的落地节奏来看，这一替代进程，将完全遵循“由浅入深、由低优先级到高优先级”的行业通用规律，分三个阶段，稳步推进：

- **第一阶段（2026-2027）**：替代范围将集中在互联网/OTT 行业的次级非核心推理场景 —— 这类场景的特征，是流量规模大、成本敏感性高、对极致性能的要求相对较低；从行业的公开数据来看，这一阶段的国产算力资源推理占比，将提升至 70% 以上；部分头部厂商的次级推理场景，将实现 100%的国产算力替代。
- **第二阶段（2027-2028）**：替代范围将延伸至互联网/OTT 行业的高优先级核心推理场景 —— 这类场景的特征，是对极致性能、低时延稳定性要求极高；从行业的公开数据来看，这一阶段的国产算力资源推理占比，将提升至 80% 以上；部分头部厂商的核心推理场景，将实现超过 50%的国产算力替代。
- **第三阶段（2028-2029）**：替代范围将进一步覆盖部分对精度要求相对较低的训练、微调场景 —— 这类场景的特征，是对算力的精度要求相对较低；从行业的公开数据来看，这一阶段的国产算力资源训练占比，将从当前的不足 10%，提升至 30%以上。

值得注意的是，这一替代进程，并非“国产算力资源完全替代海外算力资源”，而是“两者在行业内的份额分配发生根本性逆转” —— 在可预见的周期内，海外高端 GPU 资源，仍将在头部互联网/OTT 厂商的核心模型前沿训练场景中，保留不可替代的份额；但从整体的算力资源规模来看，国产算力资源将占据绝对主导地位；行业内的算力资源格局，将从“海外算力为主、国产算力为辅”，彻底转变为“国产算力为核心、海外算力为补充”的全新格局。

结论

互联网/OTT 行业的 AI 算力，已经正式进入“以超节点为核心、异构混合为常态、国产算力为增量主导”的成熟阶段 —— 这一阶段的核心特征，是“算力资源的竞争维度，从单一的芯片性能，升级为‘硬件 + 软件 + 场景’的全栈系统能力比拼”；行业级算力资源的选型逻辑，以及厂商间的竞争维度，都将在这一框架下，持续深化发展。

对于行业技术决策层而言，当前的核心技术决策路径，已形成清晰的行业标准：

- **在战略规划层面**：需要摒弃过去“单一算力集群支撑所有业务场景”的陈旧思路，以“分层异构资源池”为核心方向，制定中长期算力资源重构规划 —— 重点布局“海外高端 GPU 资源 + 国产超节点资源”的双轨异构架构，以此平衡业务极致性能需求、算力成本控制需求、供应链安全合规需求，三大核心需求。
- **在技术选型层面**：需要将“超节点的系统级有效 Token 吞吐能力”，作为算力选型的核心技术标准，而非单纯的单芯片理论峰值算力；在国产超节点选型过程中，应优先选择具备“全栈技术能力、大规模集群落地案例、成熟商业交付服务能力”的头部厂商方案 —— 这类厂商的超节点方案，才能真正适配行业核心场景的业务需求。

- **在落地执行层面**：需要优先启动“核心业务场景的流量分级治理”工作，按照“业务优先级、技术特征、成本容忍度”维度，对流量进行精准拆分；再配合业务的迭代开发节奏，按照“非核心场景、次核心场景、核心场景”的顺序，逐步将低优先级流量，迁移至国产超节点资源池；将高优先级流量，调度至适配的海外高端 GPU 资源池 —— 这一落地路径，是当前行业内，兼顾“业务体验稳定性、算力成本降低速度、供应链安全”的最优技术落地路线。
- 未来 12-24 个月，行业内算力资源的核心增量，将完全来自国产超节点资源；国产算力资源的综合竞争力，将进一步提升；异构算力混合调度能力，将成为头部互联网/OTT 厂商的核心技术壁垒。可以确定的是，算力资源的重构进程，将成为决定互联网/OTT 行业下一个周期竞争力的关键核心变量；只有构建“异构算力混合调度、国产超节点资源支撑”的成熟算力底座，才能在后续的行业竞争中，占据主动地位。

信息与数据来源

本报告所有数据、技术细节，均来自公开权威行业信息源，所有引用的实测数据、行业落地案例，均来自行业官方发布会报道、头部厂商公开技术白皮书、权威行业机构报告，以及行业主流媒体的公开调研报道。

具体来源明细如下：

- **WAIC 2026 官方发布内容**：华为昇腾 Atlas 950 SuperPoD 技术白皮书、百度天池 2.0 超节点技术白皮书、阿里平头哥磐久 AL128 超节点官方技术文档、沐曦曦景 S600 超节点官方发布会材料、燧原云燧 ESL64-O 超节点官方技术文档。
- **头部互联网厂商公开技术资料**：字节跳动 2026 年算力基础设施规划公开技术分享、阿里巴巴 2026 年算力基础设施布局公开技术分享、腾讯 2026 年 AI 算力采购策略公开技术分享、百度智能云 2026 年超节点落地方案公开技术分享。
- **行业第三方机构调研数据**：IDC & 中国半导体行业协会 2026 年一季度 AI 加速芯片市场份额报告、集邦咨询 2026 年超节点集群行业发展报告、信通院 2026 年中国 AI 算力产业发展报告、券商行业产业链调研资料。
- **行业主流媒体公开报道**：36 氪、智东西、DAMO 开发者矩阵、《华夏时报》、《财经网》、与非网等行业权威媒体的公开行业报道，及相关头部厂商技术负责人公开演讲内容、行业访谈内容。

所有涉及到的厂商产品参数、实测性能数据，均来自厂商官方公开材料或行业机构的公开实测数据；部分涉及行业落地规模的敏感性数据，采用了行业共识的公开调研区间值，而非厂商官方精准披露值。

免责声明

本报告基于公开的行业信息数据和行业共识逻辑整理，不构成任何商业采购的直接决策建议；行业内各厂商的实际技术落地效果，与自身业务场景的技术特征、适配优化程度、集群运维能力强相关，存在较大的个体差异；在实际采购过程中，应根据自身业务场景的实际需求，完成针对性的技术实测验证后，再做出采购决策。

版权声明

本报告在 ChatGPT、Gemini、DeepSeek、豆包、GLM、千问、混元、文心一言、Hermes Agent、MaxKB 等 AI 工具的协助下完成，本人已尽最大的可能核实信息来源及相关描述，百密一疏之处尽请谅解。

关于作者

郭磊

1974 年 9 月出生于上海崇明
2002 年 6 月获南京大学计算机系人工智能方向博士学位
毕业论文《情景演算与信念修正及其在面向 Agent 技术中应用的研究》
长期供职于华为等 ICT 头部企业，致力于云、IoT、5G、AI 等前沿技术应用研究
现服务于 **HCR 慧辰**（北京慧辰资道资讯股份有限公司，股票代码：688500）