FROST & SULLIVAN
沙利文

头豹
LeadLeo

中国推理算力市场追踪报告，2025年H1

2025年8月

头豹研究院
弗若斯特沙利文咨询（中国）

关键发现

01

■ **算力需求重心从训练转向推理，算力基础设施持续扩展与升级**

AI算力消耗已从集中式训练转向大规模推理，带来前所未有的增量需求。2025年被认为是算力爆发的元年，推理算力的需求将迎来井喷式增长。推理算力的需求将在未来几年内远超训练算力。

02

■ **2025年H1中国推理算力服务市场中，天翼云以【21.4%】的市场份额领先**

中国日均Tokens消耗量从2024年初的1000亿增长到截至今年6月底，日均Token消耗量突破30万亿，1年半时间增长了300多倍，这反映了中国人工智能应用规模快速增长。天翼云息壤一体化智算服务平台率先完成国产算力与DeepSeek-R1/V3系列大模型的深度适配优化，成为国内首家实现DeepSeek模型全栈国产化推理服务落地的运营商级云平台。

03

■ **未来推理算力长序列与超大模型推理优化成为关键，国产软硬件协同与生态成熟推动推理普及**

中国算力正朝着“训推一体”融合架构快速发展，以支撑大规模模型与多模态应用的高效低延迟推理。国产AI芯片与推理框架不断优化，结合模型压缩、量化、动态推理等技术，进一步提升能效比和部署灵活性。

研究框架

◆ 中国推理算力市场综述

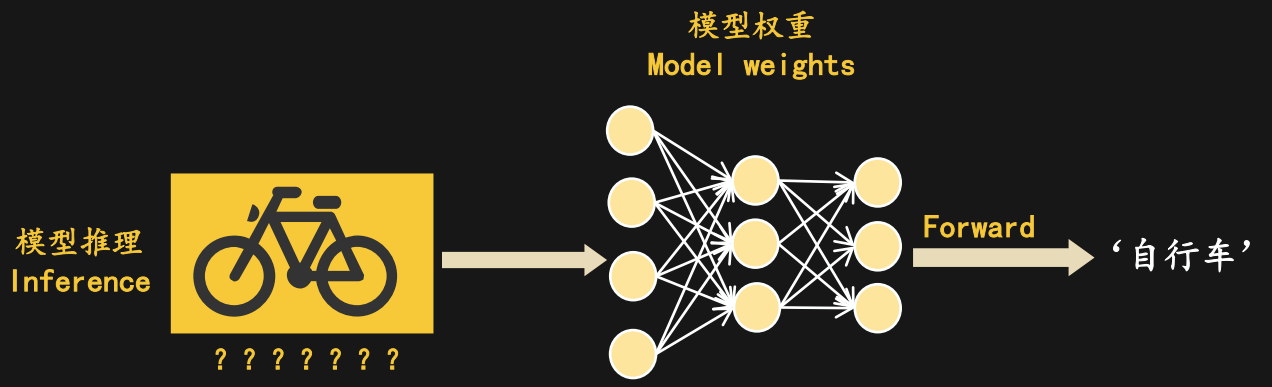
- 关键发现
- 中国推理算力定义及服务覆盖范围
- 算力需求重心从训练转向推理
- 中国推理算力市场规模分析
- 中国推理算力竞争格局分析
- 中国推理算力核心技术分析
- 中国推理算力相关政策分析
- 中国推理算力发展趋势分析
- 中国推理算力未来挑战分析

中国推理算力：定义与服务覆盖范围

关键发现

推理算力主要负责AI模型的推理任务，主要用于处理和执行已经训练好的模型进行实际应用。这包括执行推理任务、处理实时数据和提供预测结果。推理过程通常对计算资源需要快速响应，对实时性要求较高。

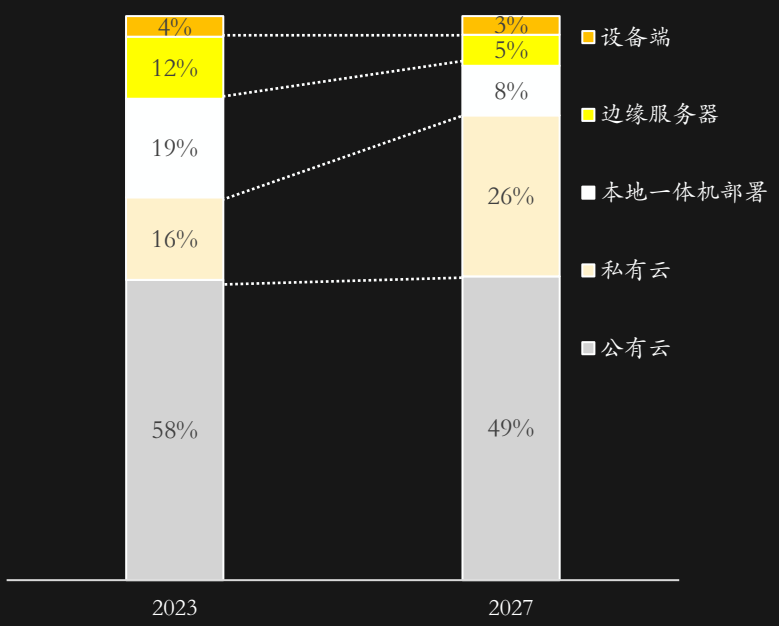
推理算力定义范围



- 推理是指利用训练好的大模型，使用新数据推理出各种结论。
- 推理芯片的目标是在已经训练好的模型上执行任务，推理芯片不需要进行复杂的学习过程，其设计重点是在保持高效计算的同时，尽可能减少功耗。
- 因此，推理芯片比较关注**低延时、低功耗**。可配置使用优化的推理硬件，高效能的服务器和网络设备如GPU、NPU或FPGA，这些硬件能够高效执行模型推理任务，以确保快速响应时间和稳定的服务。但不一定需要与训练时相同的硬件配置 。
- 推理型智算中心的硬件更注重处理速度和可靠性。

✓ 随着AI从训练为重走向推理为主，私有化环境及边缘的部署需求都在迎来爆发。

推理平台及应用部署偏好



来源：沙利文、头豹研究院

中国推理算力产业洞察——发展趋势

关键发现

中国智能算力正朝着“训推一体”融合架构快速发展，以支撑大规模模型与多模态应用的高效低延迟推理。

中国推理算力发展趋势分析

在当前国家高度重视人工智能发展的战略背景下，中国推理算力正迎来快速发展阶段。随着AI模型尤其是**大模型**和**多模态模型**的广泛应用，对高效、低延迟推理算力的需求持续攀升。从技术发展趋势来看，推理算力正呈现以下几个重要方向：

✓ 算力基础设施持续扩展与升级

国家政策和市场需求共同推动算力中心规模不断扩大，尤其是智能算力中心正在从“训练为主”向“训推一体”融合架构演进。这种架构不仅能支持大规模模型训练，还可高效完成模型推理任务，更好地适应多样化的业务场景需求。

✓ 长序列与超大模型推理优化成为关键

随着支持长序列（如32K甚至更长）的模型逐步进入商用，推理过程中对内存和计算资源的需求急剧上升。例如，处理超长文本或音视频输入时，KV Cache 等缓存机制面临巨大压力。多级缓存技术（如HBM + DRAM + 专业存储）通过“以存代算”策略显著减轻计算负担，提升推理效率，支持更长上下文理解和更复杂任务处理。

✓ 多机并行推理支撑超大模型与多模态应用

面对千亿级参数模型和百万级长度多模态输入带来的计算与内存挑战，多机并行推理成为必然选择。通过节点内NPU高速互联与节点间RoCE网络协同，实现计算资源的高效调度与通信优化，显著提升推理吞吐并降低延迟。

✓ 软硬件协同与生态成熟推动推理普及

国产AI芯片（如昇腾、寒武纪等）与推理框架（如MindSpore、PaddlePaddle）不断优化，结合模型压缩、量化、动态推理等技术，进一步提升能效比和部署灵活性。同时，开放算力生态建设和标准推进也加速了推理算力的普惠化应用。

➢ 国产算力正通过技术、生态与产业链的协同效应，为中国推理算力发展奠定坚实基础。

AI芯片实现多技术路线并行发展，训练与推理芯片性能快速提升。华为采取开放策略，公开芯片路线图并授权合作伙伴生产自有品牌服务器，吸引更多企业加入生态，共同提升国产解决方案竞争力。

以华为昇腾为代表的国产芯片迭代速度加快，通过“超级节点”集群架构，以多卡互联实现系统级算力突破，有效弥补单芯片性能差距。

从芯片制造（中芯国际、华虹半导体）、设备材料（中微公司、鼎龙股份），到整机、连接器、光模块、液冷等环节，已形成自主可控的算力基础设施体系，为推理算力发展提供全面保障。

来源：沙利文、头豹研究院

方法论

- 头豹研究院布局中国市场，深入研究19大行业，532个垂直行业的市场变化，已经积累了近100万行业研究样本，完成近10,000多个独立的研究咨询项目。
- 研究院依托中国活跃的经济环境，从纵深防御、快速响应、轻量化部署等领域着手，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。
- 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。
- 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。
- 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明

6