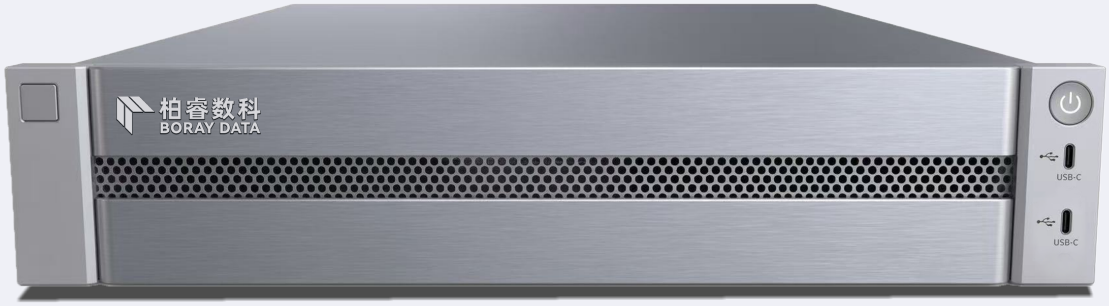

AI光存算一体机

新一代大模型推理基础设施

AI光存算一体机是一款面向生成式AI推理场景的软硬一体化产品。它以自研的分布式内存池化技术 Rapids FiberMem 为核心，通过“计算-内存-存储”分离架构，将集群内的物理内存转化为动态分配、全局共享的逻辑资源池。该产品旨在解决大模型推理中GPU资源利用率低、显存不足及“内存墙”等核心痛点，为企业提供高效、可扩展且低成本的AI推理基础设施解决方案。

产品亮点

01 分布式内存池化

- 通过 Rapids FiberMem 技术打破单机显存物理限制，实现全局 DRAM 资源的智能调度。

02 亚微秒级互联

- 自研 RDMA 高性能通信底座，跨节点访问延迟维持在 10-100 纳秒级。

03 软件定义内存（SDM）

- 将内存从物理服务器解耦，转化为可动态分配的逻辑资源。

04 多级异构池化

- 不仅能池化高速 DRAM，还能将 NVMe SSD 纳入统一虚拟地址空间，实现分层存储。

05 零拷贝数据交换

- 数据在池化内存中生成后可被其他节点直接读取，无需内核与用户空间间的反复搬运。

06 极高扩展性

- 独立部署模式下不受单机槽位限制，可提供 PB 级的统一内存访问空间。

010-64700868　400-088-1960
www.boraydata.cn　brmarketing@boraydata.com
北京市朝阳区酒仙桥北路乙10号院星地中心C座

特点优势

破解“内存墙”

- 消除算力增长与内存传输速度不匹配带来的瓶颈
- 提升算力集群整体吞吐量

高兼容性

- 全面适配 NVIDIA、AMD 及华为昇腾、寒武纪等国产 NPU
- 打破硬件生态锁死

动态热调度

- 支持秒级甚至毫秒级的内存配额调整
- 避免资源“静态死锁”造成的浪费

部署灵活

- 支持混合部署（充分利用既有硬件闲置内存）与独立部署（彻底存算分离，便于维护）两种模式

应用无感调用

- 确保分布式环境下数据一致性
- 上层应用无需修改代码逻辑即可调用扩展内存

降低成本 (TCO)

- 通过高效的资源利用和“内存-闪存”合一架构
- 显著降低大规模部署的硬件建设成本

应用场景

大模型推理加速

解决长文本推理和高并发请求下的显存溢出风险，提升响应时效（TTFT）

MoE 专家权重调度

解决长文本推理和高并发请求下的显存溢出风险，提升响应时效（TTFT）

模型加载与热切换

通过“虚拟参数池”实现万亿级模型权重的极速加载与多模型间的亚毫秒级切换

高性能数据缓存

作为计算单元间的极速数据中转站，减少数据搬运开销

异构算力资源池化

在混合部署多种品牌算力卡的复杂环境中提供统一的加速支持

行业应用

金融行业
支撑高并发的模型推理与实时风险评估

电信行业
用于大规模网络数据处理与智能化运营

工业制造
辅助复杂生产模型的实时推演与智能调度

医疗行业
支持医疗影像大模型及长文本病历分析的快速处理

应用能力

千亿级参数模型支撑	超长文本处理能力	高并发用户承载	PB 级内存扩展能力
有效支持如 Deepseek-671B、Qwen-397B 等超大规模模型的稳定运转	通过扩展 KV Cache 存储空间，支持高达 256K 甚至更长上下文的推理需求	在推荐配置下，单组集群可稳定支持 100 个以上并发用户的高强度访问	在独立部署模式下，具备支撑实时数据仓库或极高性能计算环境的扩展潜力

010-64700868　400-088-1960
www.boraydata.cn　brmarketing@boraydata.com
北京市朝阳区酒仙桥北路乙10号院星地中心C座