YC-WI-STOR

易存数智 · 易存科技

WS8000-V

KV Cache 加速存储

WS8000-V KV Cache Acceleration Storage System



180 GB/s
RAID6 后 OBD 读带宽

120 GB/s
RAID6 后 OBD 写带宽

24×NVMe
双端口 PCIe 5.0 SSD

典型应用场景

- **多轮对话服务** 保留会话前缀，降低连续交互中的重复预填充计算。
- **Agent 连续任务** 承载任务间可复用上下文，支持暂停会话后的 KV 回载。
- **长文档分析** 复用稳定文档前缀，服务合同审阅、代码分析与报告问答。
- **检索增强问答** 保留重复检索内容及提示前缀，减少公共上下文重复计算。
- **企业编码助手** 利用代码库、工具描述和长任务轨迹中的前缀复用机会。
- **多节点推理** 在模型、格式和授权匹配的条件下共享可回载 KV 。
- **公共提示复用** 复用稳定系统提示与任务模板，按上下文价值安排驻留。
- **企业私有 AI** 面向本地推理集群，结合权限边界组织共享缓存。

面向长上下文与多轮推理的企业级 KV 缓存存储系统

WS8000-V 是易存数智面向多节点大模型推理设计的 KV Cache 加速存储系统。CVALE 缓存引擎结合易存并行文件系统，为具有复用价值的上下文提供共享温冷缓存。系统采用双控制器与 24 块双端口 PCIe 5.0 SSD ，通过分层承载与上下文复用，减少重复预填充计算。

面向企业推理集群的共享缓存平台

- **双控全闪** 24 块双端口 PCIe 5.0 SSD ，元数据与对象服务分工承载共享 KV 。
- **缓存分层** GPU 显存、主机内存与共享闪存协同，按复用价值和回载成本组织数据。
- **上下文复用** 以模型权重、 KV 格式及租户权限匹配为前提，复用有效连续前缀。
- **RAID6 保护** 两组 10+2 RAID6 ，共 20 个数据盘与 4 个校验盘；每组最多容忍 2 块盘失效。

上下文缓存与并行文件协同

- **CVALE 缓存引擎** 以新近度、访问频率与老化管理上下文信用，采用上下文级置换策略。
- **易存并行文件系统** 全阵列 2 MDT 、 4 OST ；每控 1 个 MDS 与 2 个独立 OSS 服务实例。
- **KV 生命周期** 围绕身份、索引、提交、恢复及空间回收组织共享缓存。
- **应用效果评估** 关注 TTFT 、 TPOT 、有效输出吞吐与分层命中率；回载成本低于重算时获取收益。

管理与缓存观测

围绕推理服务观察 TTFT 、 TPOT 与满足 SLO 的有效吞吐，分别记录 GPU 、 DRAM 与共享闪存层的命中和回载字节。总体命中率与外部 SSD 贡献分开统计，并结合队列、 I/O 与资源使用定位瓶颈。

共享缓存与推理接入

通过推理连接器衔接 KV 读写与请求生命周期，由 CVALE 管理上下文身份、索引及缓存策略，易存并行文件系统承载共享数据。模型权重、 KV 格式与租户权限共同约束复用范围。

WS8000-V 系统拓扑



WS8000-V 技术规格

系统架构	双控制器共享架构；正常各负责一组 RAID 及所属元数据与对象目标
底层性能	RAID6 后 OBD： 读 180GB/s ，写 120GB/s （内部测试）
服务配置	全阵列 2 MDT 、 4 OST ；每控 1 MDS ＋ 2 个独立 OSS 服务实例
数据保护	2 组 10+2 RAID6 ； 20 个数据盘＋ 4 个校验盘，无额外热备盘
驱动器类型	24×NVMe 双端口 PCIe 5.0 SSD
软件架构	CVALE 独立缓存引擎＋易存并行文件系统；上下文价值置换、 KV 身份与索引、共享缓存分层、提交恢复与空间回收
容量示例	24×30.72TB： 原始 737.28TB ； RAID6 后 614.40TB （ 558.79TiB ）；扣最低 5% 元数据预算后约 530.85TiB 。

[1] OBD 为底层测试口径，不代表文件系统或 KV 端到端带宽。
[2] 容量为配置算例， 530.85TiB 尚未扣清格式化、日志及 GC 等空间开销。

获取更多资讯，请登录： http://www.wi-stor.com/ 资讯热线： 0510-88689859

关于易存数智

易存数智是一家专注于新一代高性能并行存储系统自主研发、生产与销售的高新技术企业。公司以国产化高性能存储为核心，形成“存储基础设施 + 数据治理应用”协同发展的产品路线，覆盖 AI 数据应用全流程。旗下拥有 WS 系列并行存储阵列、 AI-NAS 存算一体机、 AI 数据治理平台三大产品线。产品聚焦 AI 大模型训推、智算中心及企业私有 AI 等核心场景，为用户提供从存储底座到数据治理的全栈智算数据中心解决方案。

江苏省无锡市经济开发区金融八街 8 号联合金融大厦第 6 层 606 号 四川省成都市锦江区锦江大道 1696 号白鹭湾科技生态园科创一中心 3 栋 2 单元 403
浙江省宁波市鄞州区冬青路 528 号民和数字产业宁波高新区民和数字产业基地 B 幢 514

扫码关注易存数智

版权 ©2026 易存数智科技有限公司所有，保留一切权利。产品功能和规格如有变更，恕不另行通知。易存数智概不承担任何责任，并声明对于易存数智产品的销售与使用事宜（包括与产品适用性、商业品质或侵犯任何专利、版权或其他知识产权权利相关的责任或保证）不提供任何明示或暗示的担保。