WS-S5000-M8M
人工智能服务器
技术白皮书

4U · 8 GPU



AI Token 工厂算力与 KV 存储协同

易存数智 · 易存科技

V1.2 | 2026 年 11 月 11 日

目录

一 AI 存算一体化发展需求.....3
二 WS-S5000-M8M 产品简介.....4
2.1 产品形态与平台.....4
2.2 资源架构.....4
2.3 本地数据与管理资源.....5
三 WS-S5000-M8M 产品特点.....6
3.1 面向 Token 生产的 GPU 算力.....6
3.2 PCIe5.0 高速互联.....6
3.3 GPU 算力与 KV 存储协同.....6
3.4 企业级运行管理.....6
3.5 面向业务的资源组合.....6
3.6 可维护与可管理.....6
四 WS-S5000-M8M 关键技术与部署方案.....7
4.1 CPU 与 GPU 资源协同.....7
4.2 内存与本地数据分层.....7
4.3 网络与软件接入.....7
4.4 AI Token 工厂方案架构.....7
4.5 KV Cache 分层与上下文复用.....8
4.6 WS-8000-V 共享 KV 存储.....9
4.7 AI 服务运营与业务应用.....9
4.8 服务指标与运行管理.....9
五 WS-S5000-M8M 技术规格.....10

一 AI 存算一体化发展需求

企业 AI 从单点试用走向持续运行，计算资源的使用方式随之变化。模型服务既要处理输入文本和多模态内容，也要在多轮交互、知识检索和工具调用中保持上下文。 GPU 承担模型运算，CPU 负责请求处理、数据准备与系统服务，内存、本地盘和共享存储共同影响数据供给及服务响应。

单纯增加 GPU 数量不能覆盖所有应用需求。不同模型对显存容量、GPU 间通信、主机内存和网络的要求不同；多实例推理与多卡协同任务也需要不同的资源组织方式。平台建设应从模型规模、上下文长度、并发请求、数据集大小和管理方式出发，匹配计算与存储配置。

对企业私有 AI，业务数据的授权范围、模型版本、知识来源和结果记录同样需要纳入部署。知识问答、文档处理、编码助手、品质分析和设备运维可共用基础设施，同时保留各业务的模型、权限和资源边界。以代表性任务组织配置与评测，有利于形成可持续运行的 AI 服务。

AI Token 工厂把持续推理视为从请求接入到 Token 输出的服务过程。除 GPU 计算外，系统还需要组织模型服务、上下文缓存、共享 KV 存储和运营管理。长上下文、多轮对话及 Agent 连续任务中，具有复用价值的有效前缀可以通过缓存层组织驻留和回载；缓存策略应同时考虑复用机会、回载成本与资源占用。

WS-S5000-M8M 作为 AI Token 工厂算力节点，与 CVALE 缓存引擎和 WS-8000-V KV Cache 加速存储组合，形成 GPU 算力、上下文管理与共享数据承载的协同方案。服务评价围绕满足时延要求的有效 Token 输出，而非只比较 GPU 数量或存储峰值带宽。

二 WS-S5000-M8M 产品简介

WS-S5000-M8M 采用双路英特尔至强 6 6500/6700 平台，在 4U 空间内承载最多 8 张双宽 GPU，以 PCIe5.0 交换架构组织 GPU 与高速 I/O 资源。产品面向 AI Token 工厂，可与 CVALE 缓存引擎、WS-8000-V KV Cache 加速存储组合，构建算力与上下文数据协同的推理基础设施。

2.1 产品形态与平台

产品采用标准 4U 机架式形态，支持双路英特尔至强 6 6500/6700 系列处理器、32 个 DDR5 内存插槽及最多 8 张全高全长双宽 GPU。CPU 单颗最高 350W TDP，GPU 单卡最高 600W TDP；处理器、GPU、供电及散热按整机配置组合选择。



图 1 WS-S5000-M8M 4U 机架式平台

2.2 资源架构

PCIe5.0 交换架构组织 GPU 与扩展 I/O 资源，GPU 侧采用 x16 连接。可按业务所需选择 CPU 上行、GPU 分组及网络与存储分配，兼顾多卡推理、节点互联和 KV 数据访问。具体任务采用的 GPU 通信路径，由整机拓扑、GPU 能力和软件栈共同决定。



图 2 平台计算与 I/O 资源关系

2.3 本地数据与管理资源

12 个 3.5 英寸热插拔盘位兼容 2.5 英寸设备，可按背板配置 SAS、SATA 和 NVMe。系统盘、模型文件、工作数据、日志及临时文件可分别规划；M.2 适合系统与引导用途，NVMe 可承担本地高频数据访问。共享数据需求通过网络与外部存储组合。

AST2600 BMC 提供独立带外管理，支持 IPMI2.0、Redfish、SOL、KVM 和虚拟媒介。4 个 CRPS 电源模块支持热插拔及 2+2 或 3+1 冗余配置，风扇支持 N+1 冗余和热插拔，便于机房运行维护。

三 WS-S5000-M8M 产品特点

3.1 面向 Token 生产的 GPU 算力

双路 CPU 与 8 卡 GPU 组合，承载大模型预填充、解码和多实例推理，按模型服务需求组织 GPU 与 I/O 资源。

3.2 PCIe5.0 高速互联

以 PCIe5.0 交换架构连接 GPU 和扩展资源，结合网络与本地存储配置，匹配多卡推理和集群部署。

3.3 GPU 算力与 KV 存储协同

联动 CVALE 与 WS-8000-V，将 GPU 显存、主机内存和共享闪存组成分层 KV 数据架构，服务上下文复用。

3.4 企业级运行管理

4 个 CRPS 电源模块支持冗余供电；AST2600 提供带外管理，结合服务监控组织资源、任务和运行记录。

3.5 面向业务的资源组合

模型服务、数据处理和业务应用可以按任务分配 CPU、GPU、内存与本地盘资源。训练与微调关注数据供给、模型检查点和多 GPU 通信；推理关注显存驻留、上下文长度、批处理策略及请求排队；图形与分析任务则按对应软件的 GPU 和 I/O 需求配置。

3.6 可维护与可管理

电源、风扇和磁盘采用便于维护的模块化设计，配合远程控制、健康监控、事件记录和告警，组织从装机、配置到故障定位的运行管理。平台管理与业务管理分别配置网络和权限，降低业务负载对带外维护的影响。

方案交付覆盖数据接入、模型与工作流配置、平台部署、接口集成、业务评测和运维培训；数据授权、模型版本、评测结果和运行文档随项目归档。

四 WS-S5000-M8M 关键技术与部署方案

4.1 CPU 与 GPU 资源协同

PCIe5.0 互联为 GPU 和扩展资源提供高速数据路径。CPU 负责请求处理、预处理和系统服务，GPU 负责模型预填充、解码及相关计算。根据所选互联拓扑安排 GPU 卡组、网络资源和本地存储，可匹配单节点多卡推理或多节点协同。

GPU 互联配置应与张量并行、流水线并行、数据并行或多实例运行方式相匹配。推理框架的通信策略、GPU 的 P2P 支持、驱动以及系统 IOMMU 等设置共同影响数据路径。平台可按任务选择配置，不以交换架构本身替代具体软件兼容要求。

4.2 内存与本地数据分层

主机内存用于请求缓冲、数据预处理、模型加载及运行时数据；本地 NVMe 可用于模型文件、数据集、检查点和临时文件。将系统盘与高频业务数据分开，有利于组织系统维护、容量管理和任务隔离。

32 个 DDR5 RDIMM/3DS RDIMM 插槽提供内存扩展空间，最高速率 6400MT/s。实际速率随处理器、内存类型、每通道装配数量及 BIOS 设置确定。内存容量与带宽应与 GPU 显存和模型运行方式一起规划，避免只按插槽总数进行配置。

4.3 网络与软件接入

业务网卡通过标准 PCIe 或 OCP3.0 选配。模型文件传输、节点间通信、共享存储与管理流量可按需求分配网络；需要高速共享数据通路时，结合链路带宽、交换网络、MTU 及端到端软件设置进行配置。独立管理口用于带外运维。

操作系统可选择 Ubuntu、Red Hat Enterprise Linux 或 Windows Server 等平台，GPU 驱动、计算框架、容器运行环境和推理引擎按所选加速卡与业务应用适配。模型版本、容器镜像和配置文件应关联保存，便于复现部署和回滚。

4.4 AI Token 工厂方案架构

以 WS-S5000-M8M 承载大模型推理算力，结合 CVALE 缓存管理与 WS-8000-V KV Cache 加速存储，形成从模型服务、上下文复用到 Token 输出的存算协同架构。面向长上下文、多轮对话、Agent 连续任务与企业私有 AI，按业务规模组合计算节点、缓存层和存储节点。



图 3 AI Token 工厂算力与 KV 存储协同架构

业务与运营层：统一 API、任务队列、模型路由、租户配额、服务计量与审计，衔接知识问答、智能体和业务系统。

GPU 算力层：WS-S5000-M8M 提供双路 CPU 与 8 卡 GPU 资源，承载模型预填充、解码和多实例推理。

缓存管理层：CVALE 围绕上下文身份、有效前缀、缓存价值和生命周期组织 GPU 显存、主机内存与共享存储间的 KV 数据。

KV 存储层：WS-8000-V 以双控制器、24 块双端口 PCIe5.0 SSD 及并行文件系统，承载共享 KV 数据与模型相关文件。

4.5 KV Cache 分层与上下文复用

大模型推理包含预填充和解码两个阶段。KV Cache 承载模型对已处理上下文的中间状态；在模型权重、分词方式、KV 格式及授权范围匹配的条件下，有效连续前缀可以被后续请求复用。知识文本或检索结果相似，并不自动意味着 KV 能够直接复用，实际复用对象需要按连续前缀识别。

方案以 GPU 显存承担当前计算的热 KV，以主机内存组织节点内温缓存，以 WS-8000-V 共享闪存承载适合跨请求或跨节点回载的 KV 数据。CVALE 围绕上下文身份、索引、访问频率、驻留价值和生命周期管理缓存，推理连接器将 KV 读写与请求过程衔接。

请求到达后，连接器根据模型版本、租户和上下文身份查询可复用前缀。缓存命中且回载收益高于重新计算成本时，恢复对应 KV 并继续处理剩余输入；未命中或不适合回载时，执行正常预填充。新生成的 KV 按策略提交、索引和回收，不要求每个请求都写入共享层。

4.6 WS-8000-V 共享 KV 存储

WS-8000-V 采用双控制器与 24 块双端口 PCIe5.0 SSD，结合 CVALE 缓存引擎和易存并行文件系统，组织共享 KV 数据。存储层承载数据与元数据，缓存层负责上下文身份和价值策略，推理引擎负责模型执行及 KV 使用。计算与共享存储可按业务增长分别增加资源。

共享缓存的检索条件应包含模型与权重版本、KV 格式、精度、位置编码相关配置和租户授权；模型或缓存格式发生变化时，应识别新旧缓存边界。存储故障、缓存丢失或索引失效时，服务可按策略转入重新预填充，缓存数据的生命周期与原始业务数据分别管理。

4.7 AI 服务运营与业务应用

AI 服务运营层将数据、知识、模型、工具和资源关联到统一业务任务，配置模型路由、服务目录、身份权限、配额、计量和审计。文本 Token、图像视频、GPU 时长与存储容量分别组织统计，避免以单一指标混合不同服务的资源消耗。

企业知识问答、岗位助手、文档处理、品质分析、设备运维、AIGC 和 Agent 工作流可按场景接入。数据经过授权、解析、去重、切分和质量检查后形成有来源、有版本的知识资源；业务结果保留依据，反馈用于后续知识与模型配置更新。

4.8 服务指标与运行管理

围绕业务服务目标，记录 TTFT、TPOT、满足时延要求的有效输出吞吐、GPU 利用率及队列状态。缓存观测分别记录 GPU、DRAM 和共享存储的命中、回载字节、访问延迟和淘汰情况，区分总体缓存命中与共享存储层贡献。

评价方案时保持模型、精度、输入与输出长度、上下文分布和并发配置一致，分别观察冷缓存、可复用前缀和连续任务。单位有效 Token 成本同时纳入 GPU、CPU、内存、网络、存储及运行管理成本，使扩容决策与实际业务使用对应。

五 WS-S5000-M8M 技术规格

项目	规格参数
产品型号	WS-S5000-M8M
产品形态	标准 4U 机架式，最多 8 张全高全长双宽 GPU
处理器	双路英特尔至强 6 6500/6700 系列；单颗最高 350W TDP
GPU	最多 8 张全高全长双宽 GPU；单卡最高 600W TDP，按供电与散热配置选配
内存	32 个 DDR5 RDIMM/3DS RDIMM 插槽；速率最高 6400MT/s，随处理器与内存配置确定
GPU 互联	PCIe5.0 交换架构；GPU 侧 x16 连接，按工作负载选择 CPU、GPU 与 I/O 拓扑
PCIe 扩展	支持 PCIe 5.0 扩展与 OCP 3.0 网卡选配；可按节点互联和 KV 数据通路配置高速网络
本地存储	最多 12 个 3.5 英寸热插拔盘位（兼容 2.5 英寸）；按背板选配 NVMe；可选 2 个 M.2
存储控制器	可选 12Gb/s SAS HBA/RAID；支持 Intel VMD、VROC；可选 M.2 RAID 转接卡
网络	可选标准 PCIe 与 OCP3.0 业务网卡；1 个独立千兆 RJ45 管理口
I/O 端口	前置 2×USB 3.0、1×VGA；后置串口、2×USB 3.0、VGA、RJ45 管理口
电源	4 个 CRPS 热插拔电源模块；2000/2700/3200/3600W 可选，支持 2+2 或 3+1 冗余
风扇	N+1 冗余热插拔风扇；可选后置风扇模组
系统管理	AST2600 BMC；IPMI 2.0、Redfish、SOL、KVM、虚拟媒介；独立千兆管理口
安全功能	可选 TPM/TCM；机箱入侵检测；BIOS/BMC 双 Flash 冗余
操作系统	Ubuntu、Red Hat Enterprise Linux、Windows Server；按 GPU 与驱动选择版本
机箱尺寸	448（宽）×175（高）×822（深）mm；含挂耳深 842mm
工作环境	工作温度 5℃～35℃；相对湿度 10%～80%，无冷凝

盘位、NVMe、OCP 及标准 PCIe 扩展按所选背板和通道配置组合；GPU 型号、整机功耗与冗余模式按部署环境选择。