WS-S5000-M8
人工智能服务器
技术白皮书

4U · 8 GPU



企业 AI 训练与推理存算一体化平台

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-M8 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-M8 产品特点.....6
3.1 双路平台与 8 卡直连.....6
3.2 DDR5 内存与本地数据.....6
3.3 高速扩展与网络选配.....6
3.4 冗余供电与远程管理.....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-M8 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 模型服务与任务部署.....7
4.5 可靠运行与安全管理.....8
4.6 应用评测与交付.....8
五 WS-S5000-M8 技术规格.....9

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。

WS-S5000-M8 采用 4U 八卡计算平台，以 CPU 直连通道承载 GPU 资源，结合 DDR5 内存、本地存储和可选高速网络，适用于训练、推理、科学计算与数据分析。部署时将进程、GPU、内存及网卡的归属一并组织，使硬件配置与实际任务相匹配。

二 WS-S5000-M8 产品简介

WS-S5000-M8 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。通过 CPU 直连的 PCIe 5.0 x16 GPU 通道，组合计算、内存、本地存储与网络资源，适用于企业 AI 训练、推理、高性能计算及数据分析。

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-M8 4U 机架式平台

2.2 资源架构

8 张 GPU 分别以 PCIe5.0 x16 通道直连处理器。双路 CPU 各自管理对应 GPU 与内存资源域，CPU 之间通过 UPI 互联。针对独立 GPU 任务，可优先将服务进程和数据缓冲绑定到本地资源域；针对多 GPU 任务，应结合模型并行方式选择 GPU 分组。

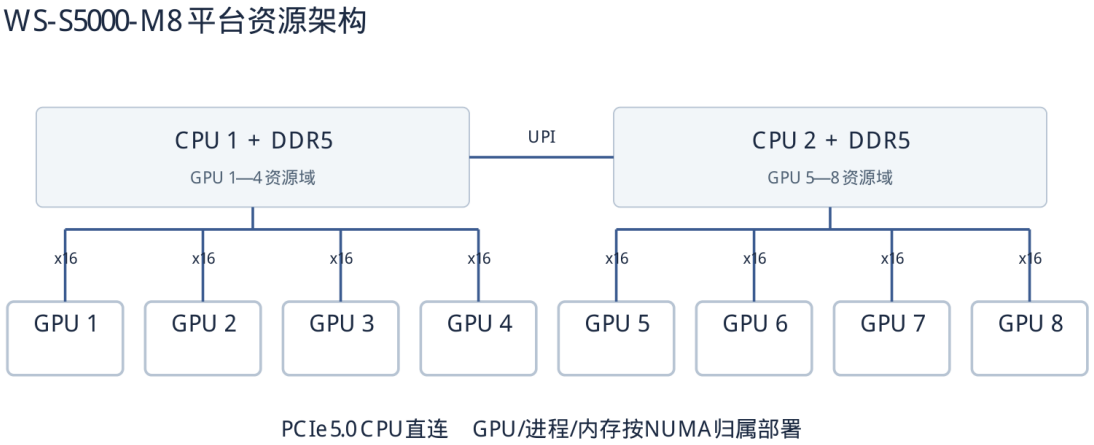


图 2 平台计算与 I/O 资源关系

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-M8 产品特点

3.1 双路平台与 8 卡直连

双路至强 6 与 8 卡 GPU 协同，按 CPU、GPU、内存和网卡归属组织任务，适配多实例推理、训练与异构计算。

3.2 DDR5 内存与本地数据

32 个 DDR5 内存插槽，最高 6400MT/s；本地盘位可分配系统、模型、数据集和临时文件空间。

3.3 高速扩展与网络选配

PCIe5.0 与 OCP3.0 支持按需配置网络及存储；结合工作负载选择带宽和数据路径。

3.4 冗余供电与远程管理

4 个 CRPS 电源模块支持冗余供电；AST2600 配合 IPMI2.0 和 Redfish，提供带外管理与健康监控。

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。

四 WS-S5000-M8 关键技术与部署方案

4.1 CPU 与 GPU 资源协同

CPU 直连结构将 GPU 分配到两个 NUMA 资源域。应用进程、内存分配及网络中断可结合 GPU 归属配置，减少不必要的跨资源域数据访问。多实例推理可以按单卡或卡组划分；模型需要跨卡协同时，应结合 GPU 能力、模型并行方式和通信库安排资源。

GPU 的可用显存、精度支持、驱动和计算框架决定模型承载能力。GPU 通道的代际与宽度属于硬件接口条件，不能单独推导模型吞吐；部署时应同时检查模型权重、 KV 占用、工作区和运行时资源。

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 模型服务与任务部署

单卡或多实例推理可按 GPU 划分服务，分别配置模型、显存使用和并发队列；多卡任务则按框架支持的并行方式组织。任务启动前核对 GPU 与 NUMA 归属、内存预留、本地数据目录、网络接口和监控指标，使应用配置与资源分配一致。

训练、微调和推理服务按环境与数据目录分离，模型权重、配置和检查点建立版本记录。面向企业知识应用时，数据授权、文档解析、知识索引与模型服务分别管理，避免把原始数据管理职责交由可重建的索引或缓存承担。

4.5 可靠运行与安全管理

供电按照所选 CPU、GPU、磁盘及网卡的整机功率需求选择 PSU 与冗余模式，校核机柜 PDU 和交流输入。散热配置结合 GPU 外形、功耗、风道和进风温度选择，电源或风扇告警通过 BMC 纳入机房运维。

业务网络、管理网络和存储访问分别设置权限。可选 TPM/TCM、机箱入侵检测与 BIOS/BMC 双 Flash 冗余；系统账户、远程控制和虚拟媒介操作按岗位授权。操作系统、GPU 驱动、BMC 及 BIOS 维护保存版本和回退记录。

4.6 应用评测与交付

训练任务关注作业完成时间、数据供给与检查点保存；推理关注上下文长度、并发、响应时延和有效输出；科学计算和图形任务采用实际业务数据评测。评测应记录 GPU 型号、模型与软件版本、精度、网络和存储配置，便于后续扩容或迁移。

交付内容包括整机配置、操作系统与驱动、模型或业务环境、网络与存储接入、健康监控、测试记录和运维文档。按代表性工作负载检查运行结果、异常恢复与权限边界，再扩展任务和资源范围。

五 WS-S5000-M8 技术规格

项目	规格参数
产品型号	WS-S5000-M8
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	8 张 GPU 以 PCIe5.0 x16 分别直连 CPU；按双路 NUMA 组织计算资源
PCIe 扩展	支持 PCIe 5.0 扩展与 OCP 3.0 网卡选配；GPU、本地 NVMe 和网卡通道按配置分配
本地存储	最多 12 个 3.5 英寸热插拔盘位（兼容 2.5 英寸）；最多 4 个 U.2 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。