WS-S5000-M8M

AI Token 工厂算力节点

4U 8-GPU AI Server

双路至强 6500/6700 平台

4U 高密度 8 卡 GPU

面向 AI Token 工厂

协同 WS-8000-V KV 存储



4U
机架式部署

8 × 600W
GPU 平台最高支持

PCIe 5.0
GPU 交换互联

典型应用场景

AI Token 工厂

承载模型服务与持续推理，组合算力和 KV 存储资源。

长上下文推理

面向文档分析、知识问答与有效前缀复用。

Agent 连续任务

服务工具调用、连续对话和任务上下文恢复。

企业私有 AI

按模型、租户与业务权限部署本地推理服务。

产品概述

WS-S5000-M8M 以 4U 8 卡 GPU 与 PCIe 5.0 交换架构承载推理算力，协同 CVALE 和 WS-8000-V，构建 AI Token 工厂。

面向 AI Token 生产

双路 CPU、8 卡 GPU 与 DDR5 内存协同，支持按模型服务组织预填充、解码与多实例推理资源。

GPU 算力与 KV 存储协同

结合 CVALE 缓存管理与 WS-8000-V 共享 KV 存储，面向长上下文、多轮对话及 Agent 连续任务。

高速扩展与企业级管理

PCIe 5.0 与 OCP 3.0 按需选配；4 个 CRPS 电源支持冗余，AST2600 提供远程管理与健康监控。

WS-S5000-M8M

4U · 8 GPU

技术规格

产品形态	标准 4U 机架式；最多 8 张双宽 GPU
处理器	双路英特尔至强 6 6500/6700 ； 单颗最高 350W TDP
GPU	全高全长双宽 GPU ； 单卡最高 600W TDP GPU 型号按供电与散热配置选配
内存	32 个 DDR5 RDIMM/3DS RDIMM 插槽；最高 6400MT/s 内存速率随处理器与装配配置确定
GPU 互联	PCIe 5.0 交换架构； GPU 侧 x16 连接
PCIe 扩展	支持 PCIe 5.0 扩展与 OCP 3.0 网卡选配 GPU 、网络与 NVMe 通道按配置分配
本地存储	最多 12 个 3.5 英寸盘位（兼容 2.5 英寸）； 按背板选配 SAS/SATA/NVMe ，可选 2 个 M.2
存储控制器	可选 12Gb/s SAS HBA/RAID ； 支持 Intel VMD/VROC
网络与 I/O	可选标准 PCIe / OCP 3.0 业务网卡；独立千兆管理口 前置 2×USB 3.0 、 VGA ； 后置串口、 2×USB 3.0 、 VGA
散热与供电	N+1 冗余热插拔风扇； 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
系统管理	AST2600 ； IPMI 2.0 / Redfish / SOL / KVM / 虚拟媒介
安全功能	可选 TPM/TCM ； 机箱入侵检测； BIOS/BMC 双 Flash
操作系统	Ubuntu / Red Hat Enterprise Linux / Windows Server
尺寸与环境	448 （宽） ×175 （高） ×822 （深） mm ； 含挂耳深 842mm 工作温度 5°C ～ 35°C ； 相对湿度 10% ～ 80% ，无冷凝

AI Token 工厂整体解决方案

WS-S5000-M8M 提供 GPU 算力， CVALE 组织上下文身份、索引与分层缓存， WS-8000-V 承载共享 KV 数据。通过推理连接器衔接 KV 读写与请求生命周期，面向多轮对话、 Agent 及长上下文场景组织 Token 生产；以有效输出吞吐、服务时延和单位 Token 成本衡量系统运行。

M8 GPU 算力
预填充与解码

→

CVALE 缓存管理
上下文身份与复用

→

WS-8000-V
共享 KV 存储

易存数智 · 易存科技（宁波 · 成都 · 北京 · 无锡）
www.wi-stor.com · marketing@wi-stor.com · 0510-88689859