

完全公开

鸿云智算大模型一体机
产品规格书
(MGP-8201s)

2025 年 1 月 15 日 首次发布

深圳市诺亚鸿云信息技术有限公司

● 版权声明

本文中出现的任何文字叙述、文档格式、插图、图片、方法、过程等内容，除另有特别注明，版权均归深圳市诺亚鸿云信息技术有限公司所有，受到有关产权及版权法保护。任何个人、机构未经深圳市诺亚鸿云信息技术有限公司的书面授权许可，不得以任何方式复制或引用本文的任何片段。

修订记录

版本	状态	修订理由和内容摘要	修订人	批准人	修订日期
v1.0	C	创建《鸿云智算大模型一体机 MGP-8201s产品规格书》	徐英		

状态：C-创建，A-增加，M-修改，D-删除

AGC 架构智算服务器

MGP-8201s

简介

MGP-8201s 是一款采用 AGC (AI computer system with the GPU at its Core) 架构设计的 8U20 卡革命性智算服务器。除适配 DeepSeek 全家桶外, 还实现了与 QwQ、LLama ChatGLM 等国内外主流开源 AI 大模型的适配。兼容多种形态 GPU 算力卡和多种类型处理器平台, 计算性能卓越、配置灵活且高效运维。适合加速



计算、科学计算、视频分析等应用场景。针对深度学习训练、推理以及科学计算等业务进行深度优化, 支持企业级和云部署模式。

亮点



架构领先

- 国内首创通过 GPU BOX 实现 GPU 算力卡热插拔架构, 提升运维连续性。
- 国内首创智慧能耗技术, 实现节能减排, 延长 GPU 算力卡寿命。
- 国内首创 GPU-RAID 架构, 实现单张 GPU 算力卡出现故障时, 自动将任务重定向到其他健康的 GPU 算力卡上, 确保业务不中断。



性能卓越

- 国内首创突破 PCIE 速率技术, GPU BOX 内置高速网络芯片, 实现 GPU 算力卡间的通信速率为 PCIE 速率与网络速率 (100Gb-400Gb) 的合集速率。
- 创新的卡间矩阵组网技术, 提升点 P2P 互联性能; 同时, I/O 缓存空间实现与 GPU 算力卡数据的直接读写, 且支持与外部存储高速数据交换。
- 支持国产 GPU 算力卡通用算子加速, 支持面向 KV 缓存 CPU 和 GPU 内存统一管理和全局共享。

地址: 深圳市龙岗区坂田街道岗头社区天安云谷产业园二期 4 栋 8 层 809 号

网址: www.nyhy-cloud.com



配置灵活

- ★ 最大支持 20 张 GPU 算力卡横向拓扑组网，实现 Host-to-Host 节点间 2Tb-8Tb 高速交换网络互联，提供充足 AI 算力。
- ★ 国内首创三段式 PCIE 结构，实现 GPU 与 CPU 解耦，支持多平台（如兆芯、海光、龙芯、飞腾等 CPU 处理器平台），不影响推理/训练性能。
- ★ GPU BOX 支持多种形态 GPU 算力卡，全高全长双宽/全高全长单宽/半高半长单宽等。



智能管理

- ★ 配置先进的电源管理芯片，与 GPUBOX 协同运作，实现对 GPUBOX 的精确电源控制，进而实现对 GPU 算力卡激活与休眠状态的有效管理。
- ★ 支持 GPU 分组管理功能，能够在同一台智算服务器的不同 GPU 算力卡上执行不同任务，实现硬件资源利用率的最大化。
- ★ 面板指示灯指引技术人员快速找到已经发生故障（或者正在发生故障）的组件，简化维护工作、加快解决问题速度，提高设备可用性。

规格

产品型号		MGP-8201s（训推一体）	MGP-8201s（推理）
形态		8U20 卡机器	8U20 卡机器
控制模组	处理器	支持兆芯、海光、龙芯、飞腾等 CPU 处理器平台	支持兆芯、海光、龙芯、飞腾等 CPU 处理器平台
	内存	最大支持 2TB 内存	最大支持 2TB 内存
	系统盘	配置 1 块 M.2 480GB NVME 硬盘	配置 1 块 M.2 480GB NVME 硬盘
	数据盘	配置 4 块 3.84TB SATA SSD 硬盘	配置 4 块 3.84TB SATA SSD 硬盘
算力模组	GPU	最大支持 20 张 GPU BOX，每张 GPU BOX 支持多种形态 GPU 算力卡，全高全长双宽/全高全长单宽/半高半长单宽等。	最大支持 20 张 GPU BOX，每张 GPU BOX 支持多种形态 GPU 算力卡，全高全长双宽/全高全长单宽/半高半长单宽等。
	算力卡	每张 GPU BOX 配置独立的 I/O 缓存空间，可选配 E1.S 规格的 1.92TB 或 3.84TB 存储容量。	/

地址：深圳市龙岗区坂田街道岗头社区天安云谷产业园二期 4 栋 8 层 809 号

网址：www.nyhy-cloud.com

产品型号	MGP-8201s (训推一体)	MGP-8201s (推理)
	每张 GPU BOX 配置 100Gb-400Gb 高速网络接口, 实现 GPU 算力卡与算力网络网卡链路绑定, 支持 RoCE v1 和 v2 协议。	/
网络模组	算力网络	最大支持 20 个 100Gbs-400Gb 光口, 支持 Roce v1 和 V2
	存储网络	可选配 2*10G/4*10G/2*25G/4*25G/1*100G 网络, 支持 Roce v1 和 V2
	业务网络	可选配 2*10G/4*10G/2*25G/4*25G/1*100G 网络, 支持 Roce v1 和 V2
电源模组	算力模组	配置 6 个 3200W 热插拔电源, 4+2 模式。
	控制模组	配置 2 个 1300W 热插拔电源, 1+1 模式。
其他	管理	支持 BMC 芯片集成专用管理 GE 网络, 提供全面的故障诊断、自动化运维、硬件安全加固等管理特性。
	GPU 运维	支持 GPU 算力卡热插拔技术。
	操作系统	根据配置不同处理平台, 支持 UOS、麒麟高级服务器操作系统、openEuler、Microsoft Windows Server、Ubuntu 等多种操作系统。
	尺寸	L(1000mm)*W(444mm)*H(355mm)
	工作温度	5°C~35°C (41°F~95°F), 符合 ASHRAE Class A1/A2

地址: 深圳市龙岗区坂田街道岗头社区天安云谷产业园二期 4 栋 8 层 809 号

网址: www.nyhy-cloud.com