KunLun

巅峰算力 全新互联

KunLun G8680 V3 超节点服务器

KunLun

简介

KunLun G8680 V3超节点服务器

KunLun G8680 V3超节点服务器面向互联网、金融等行业的大模型推理场景，提供高性能、高可靠、易部署的AI算力基础设施，满足通用风冷机房部署，以超节点架构高互联带宽为AI大模型集群推理提供超低通信时延支持，构筑AI风冷超节点产品竞争力 。

特性

旗舰性能，让训练、推理更高效

- 单机搭配8颗昇腾910C处理器提供超强算力，单卡128G 片上内存，对MOE稀疏专家模型等场景的架构适配性显著提升
- 实测性能相比上一代产品提升2.2倍，单机可部署DeepSeek 671B
- 支持单机/双机部署方案，可用于LLM大语言模型/多模态/搜索等场景

超节点架构，超大互联带宽

- 支持通过灵衢直连组成双机一体机
- 8颗昇腾910C处理器通过全新灵衢总线互联，双向互联带宽784GB/s
- 8*400GE RoCE v2高速接口

风冷机房直接部署，模块化设计高效运维

- 万瓦级商用LAAC风冷辅助液冷散热方案，风冷机房快速部署
- NPU 抽屉、灵衢总线板、计算抽屉、IO插框模块化设计，正交盲插架构实现快速安装和拆卸
- 计算-IO节点连接器三方向浮动设计，充分保证盲插对位精度，提升高速信号链路互联的健壮性和可靠性

技术规格

形态	10U AI服务器
NPU	8 * 昇腾910C处理器
CPU	4 * 鲲鹏920新型号处理器
显存	支持8 * 128GB
内存	64个DDR内存插槽，最高 5200 MT/s，单根内存条容量支持64 GB
本地存储	8 * 2.5 NVMe+2 * 2.5 SATA
网络	8 * 400GE QSFP 接口直出， RoCE v2 56* 400GE QSFP 接口直出，灵衢1.0
PCIe接口	最多支持5 个 PCIe 5.0 扩展插槽
电源	6个热插拔 3000W 电源模块，支持 5+1 冗余
供电	220VAC，336HVDC/240HVDC
散热方式	LAAC液冷辅助风冷
风扇	每个抽屉集成5个热拔插风扇模组，支持 4+1 冗余
功耗	最大输入功耗14.6kW
工作温度	5ºC～35ºC（41ºF～95ºF）
结构尺寸	442mm(高)* 447mm(宽)* 920mm(深)

河南昆仑技术有限公司

技术咨询电话：400-080-9000 **技术支持邮箱：**support@kunlunit.com

地址：河南省郑州市郑东新区龙子湖智慧岛中道东路时埂街北创智天地大厦10层

网址：www.kunlunit.com