

擎天系列 DeepSeek推理微调一体机

CQ7458-L / CQ7458-LT / 4U8卡PCIe



擎天系列 DeepSeek 推理微调一体机型号为 CQ7458-L 和 CQ7458-LT, 是超擎数智发布的新一代 DeepSeek (32B/70B/671B) 推理微调一体机。

擎天系列 DeepSeek 推理微调一体机基于 Intel 最新 Eagle Stream 平台, 搭载 NVIDIA L20 GPU, 4U8 卡 PCIe5.0 设计, 搭载 2 颗 Intel 第四 / 五代 CPU, 2 颗 CPU 连接 8 片 L20 GPU、1 片双口 25G 网卡, 具备业界领先的性能, 满足 DeepSeek 各参数模型推理与微调。

DeepSeek 各参数模型推荐配置								
模型大小	1.5B	7B	8B	14B	32B	70B	671B	671B
模型精度	FP16	FP16	FP16	FP16	FP16	FP16	FP8	FP16
推荐台数	1	1	1	1	1	1	4	4

注：推荐机型为 CQ7458-LT 或 CQ7458-L

产品特性



强大性能

采用 Intel 至强可扩展第四 / 五代处理器平台, DDR5 内存 + PCIe Gen5 技术支撑, 前后IO 维护兼容设计, 具备广泛的适用性, 满足多场景应用的高性能需求。



高可靠性

适配 NVIDIA 最新 GPU 方案和智能网卡, 专属团队全面支撑用户定制化开发需求, 确保系统的高稳定性和可靠性。

DeepSeek 性能测试

擎天系列 DeepSeek 推理微调一体机性能输出					
模型大小	模型精度	output_max token	用户并发数量	token 吞吐量 (Token/s)	台数
671B	FP8	512	512	2235	4
70B	FP16	512	100	990	1
32B	FP16	512	200	2790	1

产品规格

型号	CQ7458-L	CQ7458-LT
处理器	2 颗第四 / 五代英特尔至强可扩展处理器	
内存	16 条 DDR5 内存, 共 512G 内存, 每颗 CPU 最大支持 8 个内存通道(最大支持 32 条 DDR5 内存)	
存储	默认配置 2 块 480G SATA, 2 块 3.84T NVME 最大支持 12 块 3.5 寸硬盘 最大支持 24 块 2.5 寸硬盘, 最大支持 16 个 NVMe	
PCIe 扩展	后置: 最大支持 8 个双宽 GPU+4 个标准 PCIe 插槽 +1 个 OCP 插槽 前置: 最大支持 4 个标准 PCIe 插槽 +1 个 OCP 插槽	
网络	1 个 RJ45 千兆管理网口, 1 个双口 25G 网卡	
前置 I/O	2 个 USB3.0 端口 1 个 TypeC 端口 1 个 VGA 端口	
后置 I/O	1 个 RJ45 管理口	
尺寸	900mm*447mm*175mm	
电源	4 个 CRPS 电源, 支持 2+2 或 3+1 冗余	
环境参数	工作温度: 10°C ~ 35°C 非工作温度: -40°C ~ 70°C 非工作湿度: 20% ~ 90% (非冷凝)	
PCIe-Sw	采用 2 个 PCIe-Switch 分别连接 4 个 GPU	采用 2 颗 CPU 直通连接 8 个 GPU

