

元景系列 DeepSeek推理微调一体机 CQ7688-L / 8U8卡NVLink



元景系列 DeepSeek 推理微调一体机型号为 CQ7688-L，是超擎数智推出的智能算力旗舰产品。元景系列 DeepSeek 推理微调一体机搭载 NVIDIA H20(141G) GPU 模组 8U8 卡 NVLink，是基于全新一代 AI 超融合架构平台，面向超大规模数据中心的强劲性能，极致扩展的 AI 服务器，系统支持 4.0Tbps 网络带宽，满足万亿级参数超大模型并行训练需求。元景系列 H20 AI 服务器搭载 2 颗 Intel 第五代至强可扩展处理器，最高可达 4TB 系统内存，128TB NVMe 高速存储，支持高达 12 个 PCIe Gen5 x16/x8 扩展全高半长槽位，可灵活支持 OCP 3.0、CX7 多种智能网卡。

元景系列 DeepSeek 推理微调一体机已适配深度优化的 DeepSeek 大模型，可支持 1.5B、7B、8B、14B、32B、70B、671B 等规格的蒸馏模型以及原生大模型，依托自研 GPU 服务器的硬件垂直优化与 DeepSeek 模型的深度适配，大模型的推理性能可大幅提升。

产品特性



强劲性能

搭载 8 颗 NVIDIA 最新 Hopper 架构 GPU，2 颗 Intel 第五代至强可扩展处理器，大幅加速 GPT 大模型训练速度。



极致能效

机器采用科学的风道设计（前后风道），一些精密设备（如 GPU，网卡，模块，CPU 等等）位于前面板，机器后面板只有电源与风扇，此设计相较于其他厂家提高了这些精密设备的稳定性。



领先架构

节点内全 PCIe 5.0 高速链路，CPU 至 GPU 带宽提升 4 倍，GPU 节点之间通过 NVLink 高速互联扩展。



多元兼容

全模块化设计，一机多芯兼容，灵活配置支持本地和云端部署，领先支持多样化的 SuperPod 解决方案。

DeepSeek 性能测试

1 台元景系列(141G 显存)DeepSeek 推理微调一体机性能输出

模型大小	模型精度	用户并发数量	token 吞吐量 (Token/s)
671B	FP8	128	1052.92

产品规格

型号	CQ7688-L		
处理器	2 颗 Intel 第五代至强可扩展处理器 ,TDP350W		
GPU	H20(8*141G)Nvlink 模组		
内存	32 条 DDR5 DIMMs 4800MHz 内存,最大可扩展 4TB 内存		
存储	默认配置两张 960G SSD 硬盘和两张 7.68T NVME 硬盘, 最大支持 10 个 2.5 英寸硬盘其中 8 个 NVME,2 个 NVME/SSD 硬盘		
PCIe 扩展	支持 12 个 PCIe 扩展 (9 个总线带宽为 PCIe5.0 X16, 2 个总线带宽为 PCIe5.0 X8),每个插槽的槽位大小为全高半长。		
基础 IO	2x USB, Type-A 1x Status LED	1x Front VGA 1x UID BTN with LED	1x PWR BTN with LED 1x Console PORT, Type C
网络 1	25G Eth 双口网卡		
网络 2	1G MGMT		
尺寸	900.0mm * 448.0mm * 352.8mm (8U 机头)		
电源	2 * 2700W 12V CRPS 电源 (1+1 冗余),6 * 3200W 54V CRPS 电源 (3+3 冗余)		
工作环境参数	工作环境温度: 5 - 35°C 相对湿度: 10% - 85% 无冷凝 海拔: -50 - 3km (备注: 运行海拔高度,超过 900m 后,每上升 300 米,支持的环境温度降 1°C。)		

