CHAOQING 超擎数智

NVIDIA Elite Partner

元景系列 H20 AI服务器
CQ7688-L / 6U8卡PCIe

元景系列 H20 AI 服务器型号为 CQ7688-L，是超擎数智推出的智能算力旗舰新产品。

元景系列 H20 AI 服务器搭载 NVIDIA H20 GPU，6U8 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，最强算力密度 6U 空间内搭载 1 块 NVIDIA Hopper 架构 HGX-8GPU 模组，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 第四代至强可扩展处理器 (TDP 350W)，最高可达 4TB 系统内存，128TB NVMe 高速存储，支持高达 12 个 PCIe Gen5 x16 扩展槽位，可灵活支持 OCP 3.0、CX7 多种智能网卡，构建面向超大模型训练、元宇宙、自然语言理解、推荐、AIGC 等场景的最强 AI 算力平台。

产品特性

强劲性能

搭载 8 颗 NVIDIA 最新 Hopper 架构 GPU，2 颗 Intel 第四代至强可扩展处理器，集成 Transformer 引擎，大幅加速 GPT 大模型训练速度。

极致能效

散热性能极致优化，风道解耦设计提升 20% 系统能效比，12V 和 54V N+N 冗余电源分离供电设计，减少电源转换损耗，可选支持系统全液冷设计，液冷覆盖率高于 80%，PUE<1.15。

领先架构

节点内全 PCIe 5.0 高速链路，CPU 至 GPU 带宽提升 4 倍，节点间高速互联扩展，4.0Tbps 无阻塞带宽 IB/RoCE 组网，集群级优化架构设计，GPU: 计算 IB: 存储 IB=8:8:2。

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，支持大规模 GPT-3/LLaMA/ChatGLM 模型训练和推理，领先支持多样化的 SuperPod 解决方案，适用 AIGC, AI4Science 及元宇宙等丰富场景。

应用场景

可应用于超大模型训练和推理、元宇宙、自然语言理解、AIGC 等场景。

产品规格

型号	CQ7688-L
处理器	2 颗 Intel 第四代至强可扩展处理器 ,TDP350W
内存	32 条 DDR5DIMMs 内存，速率最高支持 4800MT/s，最大可扩展 4TB 内存
存储	最多支持 24 块 2.5 英寸 SSD 硬盘，其中最大支持 16 块 NVMe；2 块内置 M.2NVMe/SATA（可选）
PCIE 扩展	支持 10 个 PCIe5.0x16 插槽（其中 1 个 PCIe5.0x16 插槽可替换为 2 个 PCIe5.0x8 速率的 x16 槽位） 可选支持 Bluefield-3、CX7 以及多种类型智能网卡
网络	1 个 RJ45 管理口
前置 I/O	1 个 USB3.0 端口，1 个 USB2.0 端口，1 个 VGA 端口
后置 I/O	2 个 USB3.0 端口，1 个 MicroUSB 端口，1 个 VGA 端口
尺寸	860mm*447mm*263mm
电源	2 块 12V3200W 及 6 块 54V2700W 钛金级 CRPS 电源，支持 N+N 冗余
环境参数	工作温度：10℃ ~ 35℃ 非工作温度：-40℃ ~ 70℃ 非工作湿度：20% ~ 90%（非冷凝）