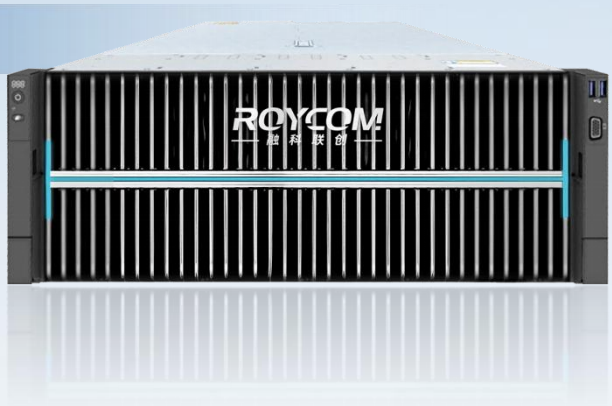
多元算力产品家族

ROYCOM 融科联创

RC4480-K20

高性能AI推理计算服务器

面向大于33B-175B量级大模型推理而设计的新一代硬件算力平台



- ☑ 专为大模型推理设计
- ☑ 支持33B-175B模型
- ☑ 提供0 . 6 PFLOPS/kw 的超高能效比
- ☑ 单机AI算力高达 2 . 2PFLOPS
- ☑ NPU 互联带宽达392GB/ s

产品概述

▶ 全新处理架构

采用国产自研64 bits高性能 鲲鹏 920 CPU 处理内部集成了DDR4、PCIe4.0、25GE、10GE、GE 等接口，提供完整的 SOC 功能。

▶ 强大的处理性能

采用昇腾8 * A I推理模组，提供256GB HBM内存，传输带宽高达800GB/s，专为AI大模型推理设计，提供了强大的AI推理能力，适用于33B-175B的大模型推理场景，主流模型全适配，同时适用内容生成、智能问答、搜索NLP、 内容审核、互联网推荐、智能客户等应用场景。

▶ 高速网络带宽

整机搭载了8 * 200GE RoCE v2高速接口；4 个 CPU与 8 个 NPU之间的双向互联总带宽为512GB/s；每一路 AI 处理器提供7 条HCCS互连链路，提供最大 392GB/s 带宽能力，保证了千卡规模集群部署。

▶ 高密度

采用4U机架式设计，独特的CPU/GPU独立散热结构设计，集高性能、高密度与高可靠性于一体。

ROYCOM 融科联创

产品规格	RC4480-K20
处理器	4 * Kunpeng 920 处理器
AI算力单元	8 *昇腾推理模组， 每一路AI处理器提供7条HCCS互连链路， 提供最大392GB/s带宽能力
AI算力	最大2.5PFLOPS FP16
内存	32个DDR4内存插槽， 最高3200MT/s， 单根内寸容量支持 16 / 32 / 64 GB
网络	8 * 200GE QSFP接口直出， RoCE协议
PCIe 扩展	最多支持 3 个 PCIE 4.0 扩展插槽
存储控制器	支持 RAID 0 / 1 / 10 / 5 / 50 / 6 / 60
硬盘方案	8 * 2.5 SATA + 2 * 2.5 NVMe 4 * 2.5 SATA + 6 * 2.5 NVMe
其他端口	前面板提供2个USB 2.0端口、1个DB15 VGA端口 后面板提供2个USB 3.0端口、1个DB15 VGA端口、1个RJ45串口、1个RJ45系统管理端口和4个RJ45板载网口
电源	4个热插拔2600W电源模块， 支持2+2冗余备份
散热	风冷， 支持8个热拔插风扇模组， 支持N+1冗余
管理功能	iBMC支持IPMI, SOL、 KVM over IP以及虚拟媒体,提供1个10/100/1000Mbps的RJ45管理网口。
机箱	机架式机箱 (4U)
机箱尺寸	175 mm（高）x 447 mm（宽）x 790 mm（深）
工作温度	5℃ ~ 35℃（ 41°F~95°F)
支持操作系统	CentOs 7.6 for ARM、EulerOs V2.0 SP10for ARM、Kylin V10 SP2 for ARM、Ubuntu 22.04等 （更多软件兼容性信息， 请联系融科联创公司技术中心）

更多产品信息，请查询公司官网或来电咨询
www.roycom.com.cn
400-018-8995