



ServMAX® G1088-K3推理超节点

产品描述

- 4 颗鲲鹏920新型号处理器
- 8个昇腾910C NPU模组
- 10U机架式服务器
- LAAC风液混合散热
- 6 x 3000W (5+1) 冗余电源，支持HVDC/AC供电
- 8 x 400GE RoCE v2 + 56 x 400GE灵衢直出接口

应用场景

大规模集群AI训练推理

多模态 AI 模型开发与部署

深度学习、自动驾驶

智慧医疗、科学计算

智能制造、金融AI分析等

产品特性

超节点架构

通过灵衢总线互联，组成16卡超节点，64卡超节点，96卡超节点，128卡超节点，192卡超节点，256卡超节点，384卡超节点等。超节点内任意两个NPU的互联带宽高达双向784GB/s，为大规模分布式训练或推理提供高效扩展，适用于LLM、多模态应用等高并发场景。

高效运维

采用LAAC风液混合散热方案，结合模块化正交设计、热插拔电源及风扇，有效提升系统可靠性与运维效率。

极致互联

单节点提供56个400GE灵衢直出接口和8个400GE RoCE接口，构建无阻塞通信网络，满足大规模分布式并行训练对通信带宽的极致要求。

处理器	- 四路鲲鹏920新型号处理器 - 单颗处理器支持≥64核
内存	- 64 x DDR5 DIMM - Up to 4TB DDR5 RDIMM
NPU	- 8 x 昇腾910C NPU模组，单卡128G片上内存
扩展槽	- Up to 5 x PCIe5.0 slots
存储	- 8 x 2.5" NVMe + 2 x 2.5" SATA热插拔硬盘
网络	- 8 x 400GE QSFP(RoCE v2) - 56 x 400GE QSFP(灵衢1.0)
I/O	- 2 x USB2.0 - 1 x VGA
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的管理网口汇聚到机柜管理板 - 灵衢总线板管理芯片（交换CPU）对交换芯片进行管理，提供1个10/100/1000Mbps的管理网口汇聚到机柜管理板
安全性	- 可选配TPM/TCM安全模块，支持数据加密与平台完整性校验 - 机箱开盖入侵检测，配备免工具拆卸机箱上盖板 - 支持BIOS/BMC双Flash冗余设计，单份固件异常可自动切换恢复
电源	- 6 x 3000W（5+1）冗余电源，支持HVDC/AC供电
尺寸	- 447mm x 442mm x 920mm (W x H x D)
机箱	- 10U
工作温度	- 5°C ~ 30°C（工作温度范围受不同配置影响，详情请咨询销售代表）
工作相对湿度	- 8% ~ 90%（无冷凝）
支持操作系统	- Ubuntu Linux、Red Hat Enterprise Linux、Microsoft Windows Server等主流操作系统
保修	- 可选现场保修服务

