

红芯聚创®昇腾AI一体机

G9680 PR



采用G9680 PR 一体机，您可以部署单一平台以解决各种企业工作负载：

- **高性能推荐系统**
例如电商商品推荐、短视频内容推荐、广告精准投放等场景。
- **大模型 Prefill 预填充**
例如长文本分析、知识库问答及超长上下文首次处理等场景。
- **在线大模型推理**
例如智能客服、企业知识助手、代码助手及实时问答等应用。
- **多模态推理**
例如图文理解、视觉问答、文档解析及视频内容分析等应用。
- **AIGC 内容生成**
例如文本创作、图像生成、数字人及营销素材生成等应用。
- **图像、视频智能处理**
例如图像识别、视频结构化分析、智能审核及画质增强等应用。

产品概述

红芯聚创®G9680 PR 昇腾AI一体机，基于第三代达芬奇架构，最大支持2颗鲲鹏950处理器和8个昇腾950PR NPU模组，单颗芯片提供84GB片上内存及1.4TB/s内存带宽，面向高性能推荐、大模型 Prefill及多模态推理等应用场景。通过NPU Full Mesh高速互联与灵衢Unified Bus技术，实现高带宽、低时延的算力协同。整机采用14U高密度风冷架构，配合最高96%的电源转换效率，为用户提供高性能、高可靠、易部署的智能算力平台。

产品特性

整机高密度算力

整机最高14.27 PFLOPS mxFP4算力
支持HiF8、mxFP8、mxFP4等格式
面向推荐与大模型推理高吞吐负载

大容量高速显存

整机配置 672GB 显存容量
单卡标配84GB，带宽1.4TB/s
配备最高128MB统一L2 Cache

高速NPU互联

单机8颗NPU Full Mesh互联
单机理论双向带宽约6.27TB/s
支持双机16颗NPU高速互联

产品规格

名称	规格参数
形态	14U机架式风冷AI服务器
NPU	8 × 昇腾950PR
CPU	2 × 鲲鹏950
*AI算力	14.27/12.49 PFLOPS@mxFP4
	6.43 PFLOPS@FP8
	3.89/3.40 PFLOPS@FP16
片上内存	8 × 84GB，总容量672GB；显存带宽1.4TB/s
系统内存	24个DDR内存插槽，最高6400MT/s；单根容量最高96GB
本地存储	最大8 × 2.5英寸NVMe + 2 × 2.5英寸SATA，支持配置RAID卡
NPU灵衢互联	单机：8 × 784GB/s（双向）；双机：2 × 8 × 1.68TB/s（双向）
网络	UBoE：每NPU 400Gbps（单向）；RoCE：每NPU 400Gbps（单向）
PCIe接口	最多支持5个PCIe 5.0扩展插槽
电源	6个热插拔3.0kW电源模块，支持5+1冗余，转换效率最高96%
供电	220VAC；336V HVDC/240V HVDC
散热方式	风冷，满足通用风冷机房部署要求
风扇	整机集成30个热插拔风扇模组，支持N+1冗余
功耗	最大输入功耗14.5kW
工作温度	5°C ~ 35°C（41°F ~ 95°F）
结构尺寸	618.8mm（高）×447mm（宽）×920mm（深）

* 理论算力值，实测可能存在小于1%误差

产品图片及规格仅供参考，具体配置以实际交付为准；性能表现受配置、软件和使用环境影响。