NEOW

AI 算力卡

XIN-10

产品概述

XIN智系列 GPU 算力卡，是完全的国产化算力卡，依托先进并行计算架构与细粒度指令集筑牢硬件根基，搭配专属软件平台的张量优化计算架构（TOPCA），让算力输出更 “精打细算” —— 既保障高性能算力支撑各领域需求，提供高性能、高效率和低成本的算力解决方案。

产品特性

- 国产化深度自主可控，贴合国内企业对数据安全与供应链稳定的核心需求
- 采用主流SIMT架构和细粒度指令集，兼容CUDA生态
- 支持AI推理、训练和科学计算等加速计算场景
- 支持Transformer、CNN、 RNN、 Stable Diffusion等模型架构
- 覆盖大语言模型、计算机视觉、多模态处理等应用领域
- 支持FP32、TF32、 BF16、 FP8、 INT8等计算精度
- 支持pyTorch、TVM、TGI等主流框架
- 拥有128GB超大显存超低功耗等特点，支持单卡部署千亿参数大模型
- 兼容Intel、AMD、海光、飞腾、鲲鹏、兆芯等CPU
- 兼容Linux、 Ubuntu、 Android、麒麟、统信等操作系

通用 GPU架构	兼容 CUDA生态	支持 主流大模型	超大显存 可达128GB	兼容 信创生态
细粒度 指令集	支持多种 计算精度	超低 推理功耗	超高 算力利用率	极致 性价比

技术规格

特征	规格
形态	全高、全长、单宽
计算Core	16
虚拟化实例	2
内存	32GB/64GB/128GB
带宽	384GB/s
AI算力	128TOPs@INT8，64TFLOPs@BF16，16TFLOPs@FP32
编解码能力	• 支持最大分辨率4K H.265 H.264 AVS2 VP9硬件解码 • 支持H.264/H.265硬件解码 60+路1080P 30FPS • 支持最大分辨率32K JPEG硬件解码 • 支持最大4K H.265 H.264硬件编码 20+路1080P 30FPS • 支持最大分辨率16K，JPEG硬件编码
PCIe接口	PCIe Gen 4.0 x16
功耗	• 典型功耗：70W • AI场景最大功耗：90W • 科学计算最大功耗：180W
重量	0.941 kg
散热	被动/主动散热
尺寸	281.33 mm x 99.33mm，单宽
工作环境温度	0 °C - 45 °C

模型支持- AI1.0

类别	模型
图像分类	Resnet50，Densetnet，Vgg16，Vgg19，MobileNetV1，MobileNetV2
图像检测	Yolov3，Yolov4，Yolov5，Yolov7，Yolov8
物体识别	PP-OCRv2-det，PP-OCRv3-det，PP-OCRv3-rec
其他	Tacotron2，Unet，Wave2lip，Bert base，Bge-m3，Bge-large-zh-noinstruct，bce-ranker-base，m3e-base

模型支持- AI2.0文生文

模型	版本
DeepSeek	DeepSeek-R1 满血版， DeepSeek-R1-Distill-Llama-70B，DeepSeek-R1-Distill-Qwen-32B ， DeepSeek-R1-Distill-Qwen-14B，DeepSeek-R1-Distill-Llama-8B DeepSeek-R1-Distill-Qwen-7B， DeepSeek-R1-Distill-Qwen-1.5B
智谱华章	ChatGLM3-6B ， GLM4-9B ， CodeGeeX2-6B，GLM130B，CodeGeeX4-9B
通义千问	Qwen3-32B，Owen2.5-14B-instruct， Owen2.5-32B-instruct，CodeOwen1.5-7B-chat， Qwen1.5-110B，Owen2.5-72B-instruct
Meta	Llama2-7B， Llama2-13B， Lama2-70B， Llama3-8B，Codellama-34B

模型支持- AI2.0 多模态/文生图

类别	模型
文生图	Stable-Diffusion-V1.4 Stable-Diffusion-V1.5 Stable-Diffusion-V2.1 Stable-Diffusion-XL Stable-Diffusion-XI Turbo Stable-Diffusion-V3.0(DiT) Flux(DiT)
文生视频	Stable Video Diffusion X
语音识别	Whisper-large-V3
多模态	Owen2-VI cogVLM2Lava-VL Videollama GOT-OCR2.0 GLM-4V-9B VideoClip Qwen2-VL-7B-instruct

总结

XIN10作为一款高性能、高能效的国产AI算力卡，在推理性能、显存带宽和国产生态适配方面展现出强劲竞争力。虽然在CUDA兼容性和软件生态广度上仍有提升空间，但其在性价比、供应链安全和特定模型精度上的优势，使其成为国产替代和绿色AI基础设施的理想选择。