Cambricon
寒 武 纪

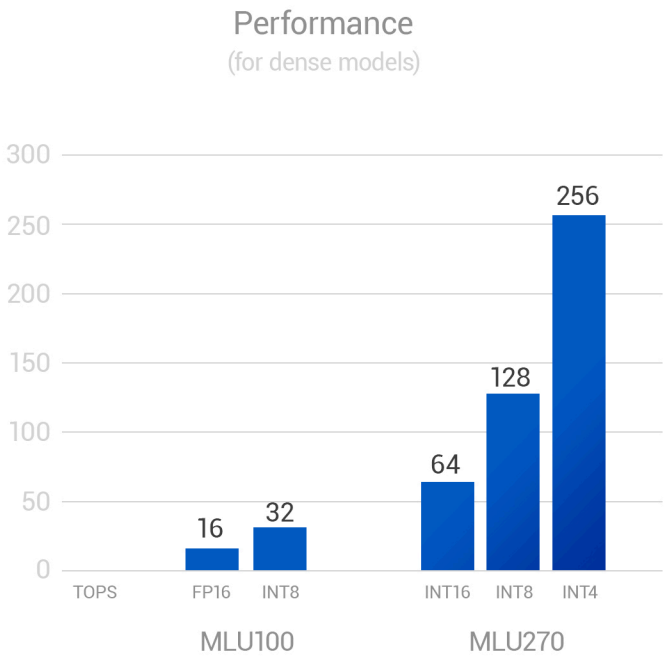
思元270
云端人工智能芯片

思元 270

MLU270-S4

面向高能效比
云端AI推理

思元270芯片基于寒武纪在处理器架构领域的一系列创新性技术，最新的特性被集成在半高半长的标准PCIe尺寸的加速卡中，可以轻松搭载于最先进的人工智能服务器，实现AI推理计算力横向扩展。MLU270-S4加速卡功耗仅为70W，比上一代加速芯片的计算能力提高了4倍，可广泛支持视觉、语音、自然语言处理以及传统机器学习等高度多样化的人工智能应用，帮助AI推理平台实现超高能效比。



MLU270-S4 产品规格

产品AI算力	芯片型号	MLU270 (MLUv02指令集)
	INT8 算力TOPS	128
	INT4 算力TOPS	256
	INT16 算力TOPS	64
计算精度支持	低精度、混合精度	支持INT16, INT8, INT4, FP32, FP16
内存规格	内存容量	16GB DDR4, ECC
	内存位宽	256-bit
	内存带宽	102GB/s
接口	PCIe接口	X16 PCIe Gen.3
功耗	最大热设计功耗	70w
	散热设计	被动散热
形态	半高半长，单宽	
尺寸	167.5mm x 68.9mm	
重量	310g	

为高能效比AI推理设计的
MLU270-S4数据中心级加速卡

全新寒武纪MLUv02架构

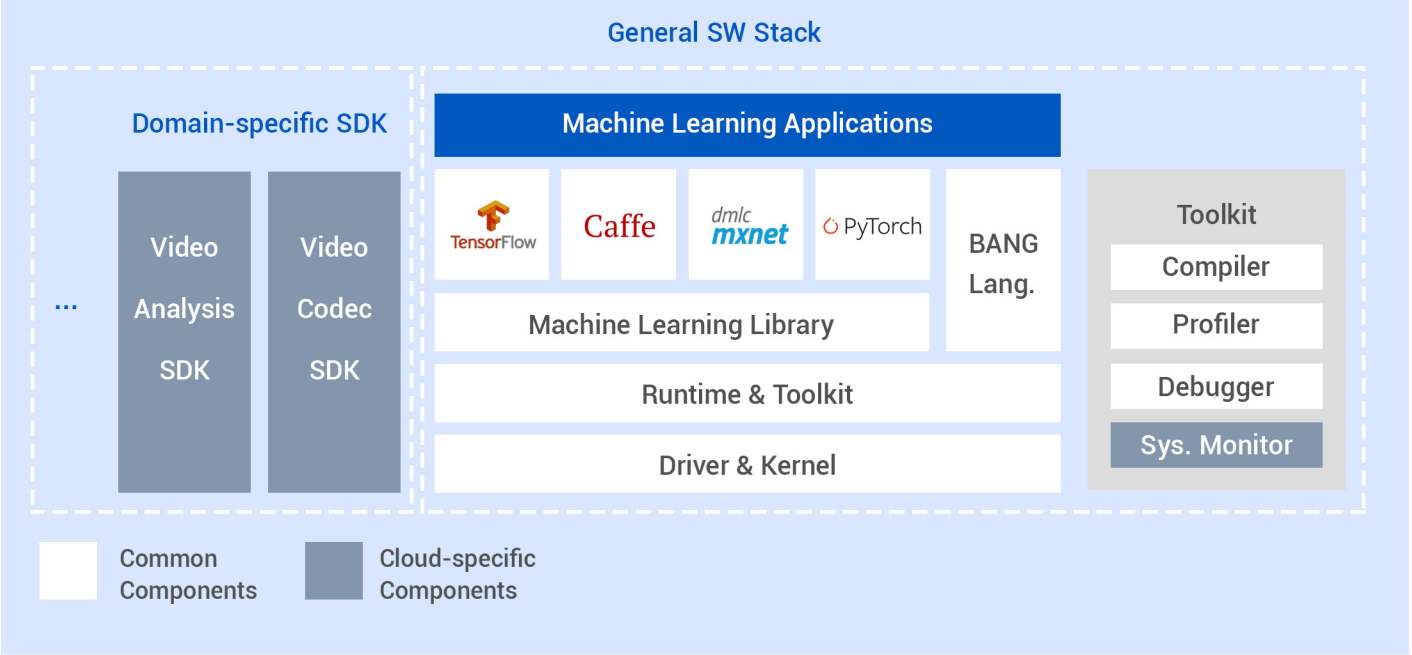
MLUv02架构不是简单的从上一代升级而来，新架构基于片上网络（NOC）构建，保证思元270芯片内多达16组IPU计算阵列的并行效率。基于硬件的片内数据压缩，提升缓存有效容量和带宽。新架构提供INT16, INT8, INT4, FP32, FP16的全面AI精度支持，满足多样化神经网络的计算力要求，通用、性能兼备。

推理性能再进一步

新架构在采用INT8精度进行AI推理计算时，非稀疏网络性能比第一代加速卡提升高达4倍，MLU270-S4可为系统提供40倍于CPU的超高能效比。其内置的全新硬件视频和图片编解码器，还可以在系统处理该类应用时，有效降低CPU前处理负载和PCIe带宽占用，帮助应用性能再进一步。

计算弹性和可编程

思元270芯片支持多类神经网络，寒武纪Neuware软件栈可以轻松部署推理环境。基于Bang语言的编程环境可对计算资源做直接定制，满足全多样化AI定制要求，专业而不专用。



访问 www.cambricon.com 了解更多产品信息　电话: 86-10-83030003　邮箱: business@cambricon.com
地址: 北京市海淀区知春路7号致真大厦D座11层