昇腾 Ascend AI 处理器

ZOMI

Ascend

About

- **昇腾全栈介绍：** 全栈全场景 vs AI 系统
- **Atlas 硬件介绍：** 边缘推理 – 云端推理 – 云端训练 – AI 集群
- **服务器爆炸视图：** 了解每个部件的细节内容

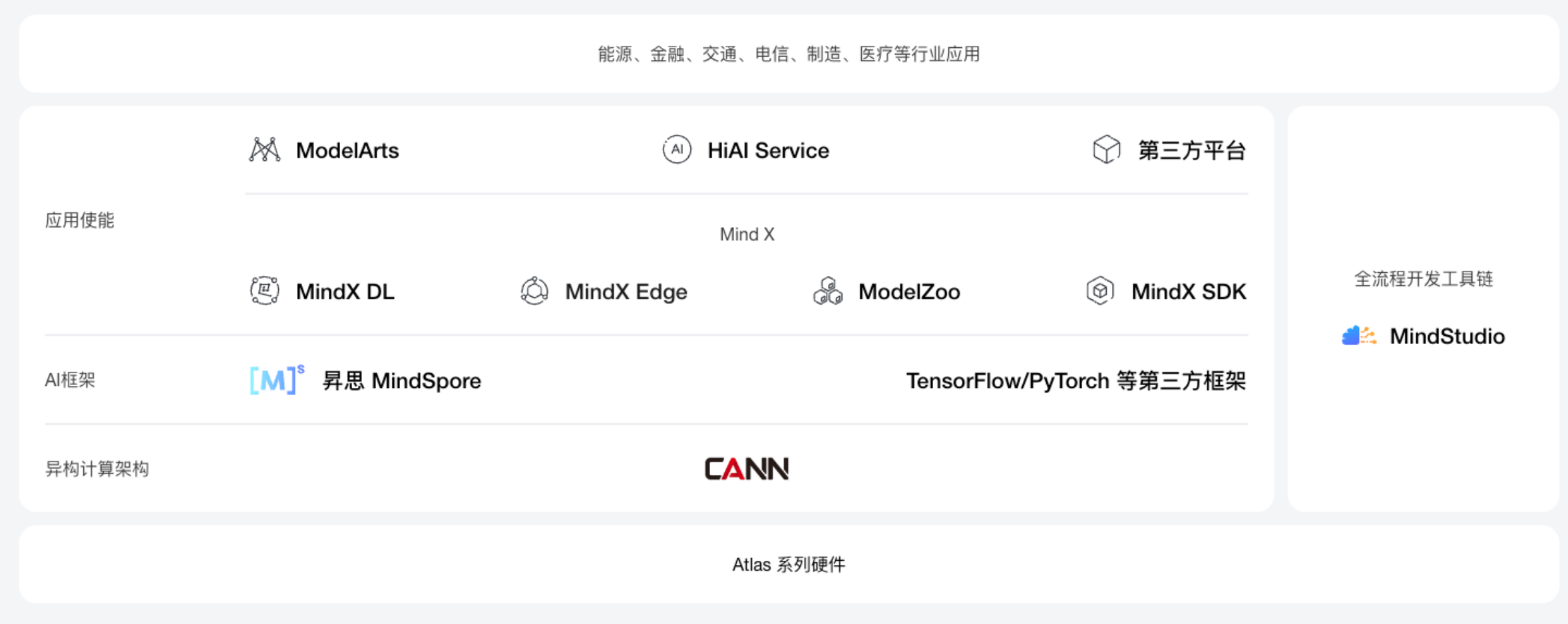
1. 昇腾全栈

什么是昇腾计算产业

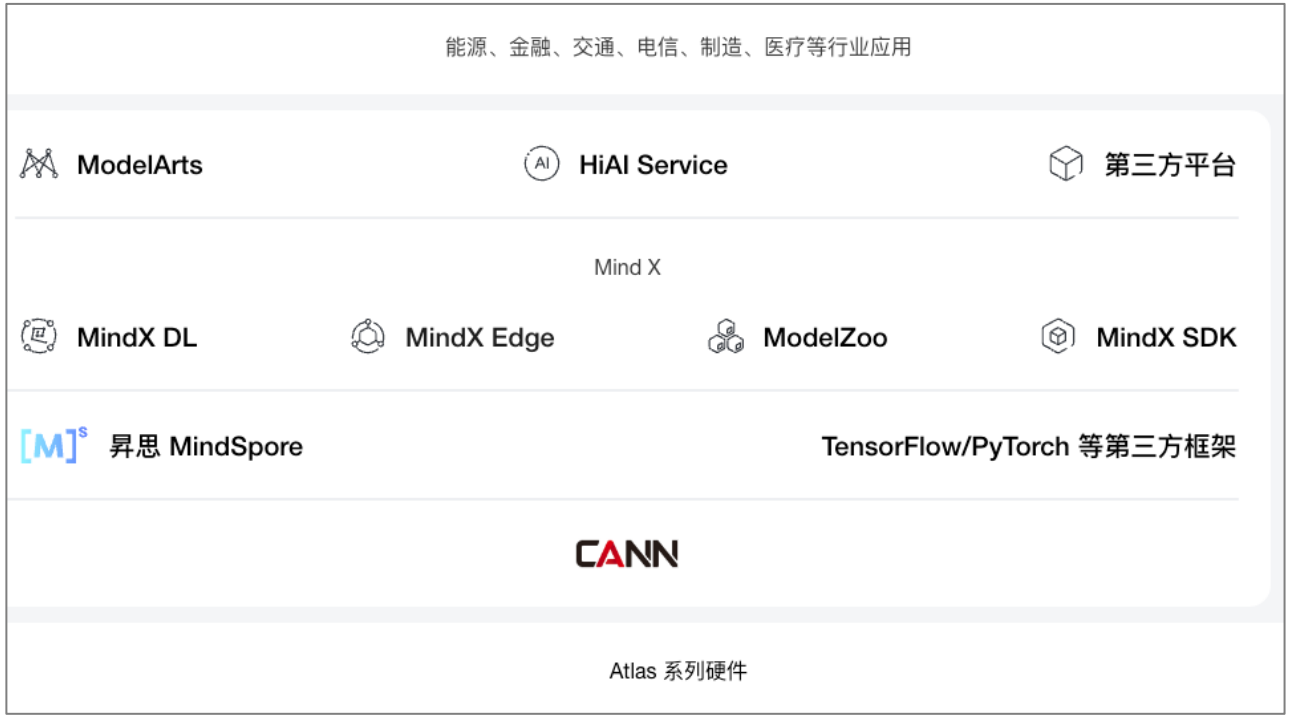
- 昇腾计算产业是基于昇腾系列（HUAWEI Ascend）处理器和基础软件构建的全栈 AI计算基础设施、行业应用及服务，包括昇腾系列处理器、系列硬件、CANN（Compute Architecture for Neural Networks，异构计算架构）、AI计算框架、应用使能、开发工具链、管理运维工具、行业应用及服务等全产业链。

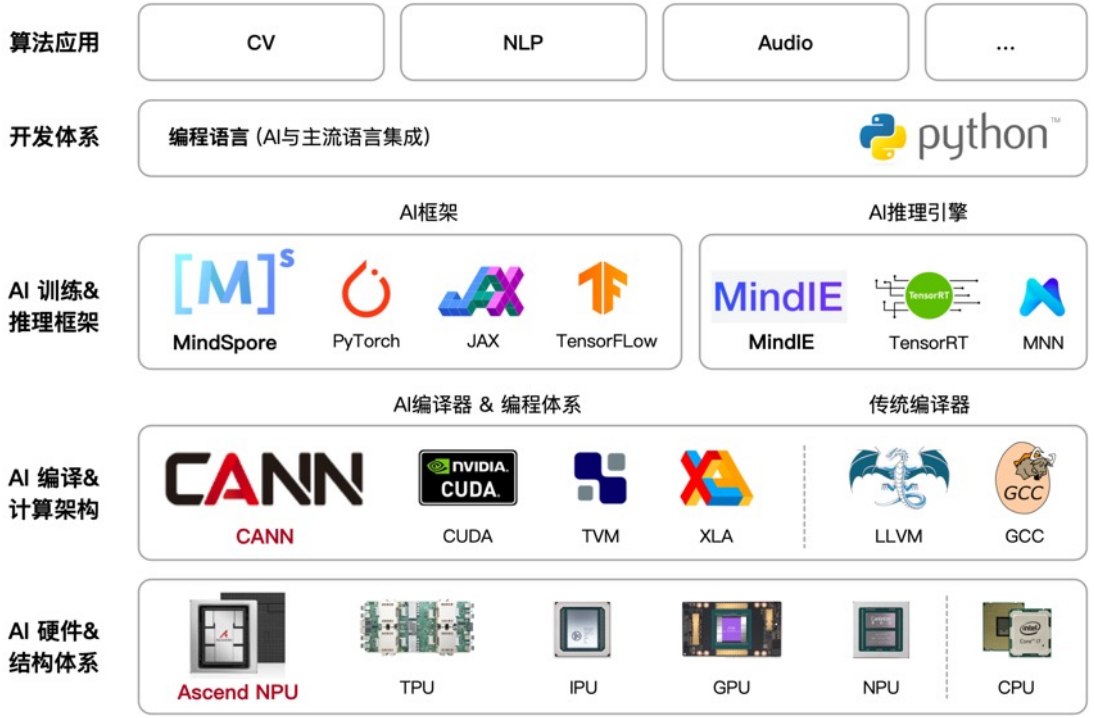
https://www.hiascend.com/

昇腾全栈 AI 软硬件平台，构筑智能世界的基石



昇腾全栈 AI 软硬件平台 vs AI 系统架构





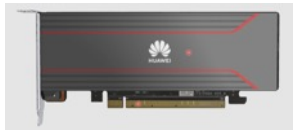
02 Atlas系列硬件

训练推理系列硬件产品

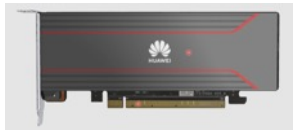
边缘端 **云端推理** **云端训练**

< 20 TFLOPS 1000 TFLOPS > 2000 TFLOPS

Atlas 200I A2 (20 TOPS)



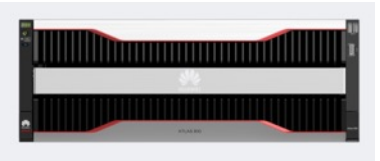
Atlas 300V Pro 视频解析卡



Atlas 300V Pro 视频解析卡

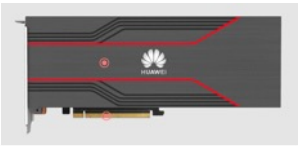
Atlas 800 推理服务器 (3000)

Atlas 500 Pro 智能边缘服务器



Atlas 800T A2训练服务器

Atlas 300I Pro 推理卡

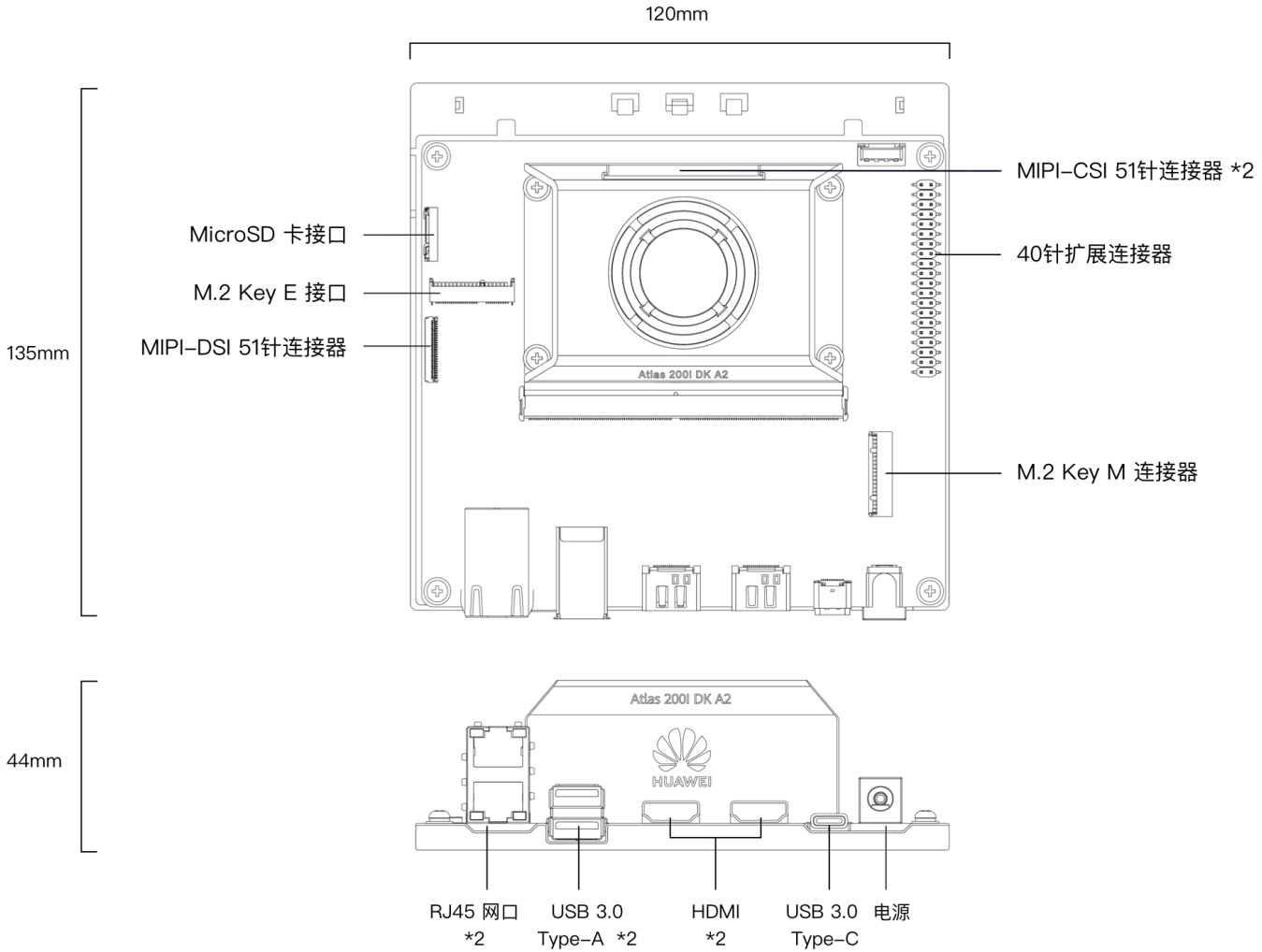


Atlas 300I Duo 推理卡

Atlas 800 推理服务器 (3010)

推理 Atlas 200I DK A2

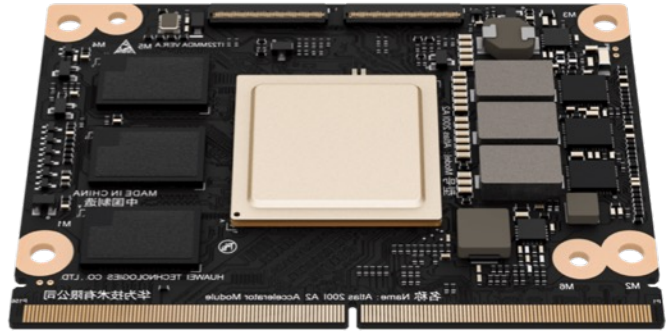
	Atlas 200I DK A2
AI算力	• 8 @ TOPS INT8 • 4 @ TFLOPS FP16
内存规格	• LPDDR4X @ 4GB • 支持ECC
CPU算力	• 4 core * 1.0 GHz
编解码能力	• 支持 H.264 / H.265 Decoder 硬件解码，20路 1080P 30FPS，2路 4K 75FPS • JPEG 解码 1080P 512FPS，编码 1080P 256FPS，分辨率最大 16384 x 16384
扩展接口	• MIPI-CSI 51针连接器 *2，可接摄像头模组 • MIPI-DSI 51针连接器 *1 • RJ45 网口 *2，支持自适应 100 / 1000M • HDMI 接口 *2 • USB 3.0 Type-A 接口 *2，兼容 USB 2.0 • USB 3.0 Type-C 接口 *1 • 40针 扩展接口 *1



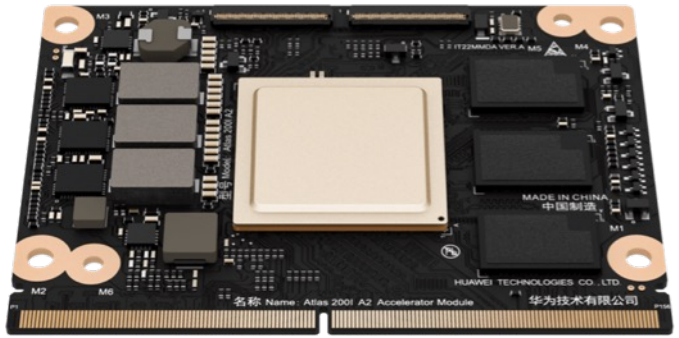
https://www.hiascend.com/hardware/developer-kit-a2

推理 Atlas 200I A2

	Atlas 200I A2（20 TOPS）	Atlas 200I A2（8 TOPS）
AI算力	• 20 @ TOPS INT8 • 10 @ TFLOPS FP16	• 8 @ TOPS INT8 • 4 @ TFLOPS FP16
内存规格	• LPDDR4X @ 12/8/4 GB • 总带宽 51.2/34.1/34.1 GB/s	• LPDDR4X @ 4GB • 总带宽 25.6GB/s
CPU算力	• 4core * 1.6GHz	• 4core * 1.0GHz
编解码	• 支持H.264 / H.265 硬件解码，40路 1080P 30FPS，4路 4K (3840 x 2160) 75FPS • 支持H.264 / H.265 硬件编码，20路 1080P 30FPS，3路 4K (3840 x 2160) 50FPS • JPEG解码能力1080P 512FPS，编码能力1080P 256FPS，最大分辨率：16384 x 16384	• 支持H.264 / H.265 硬件解码，20路 1080P 30FPS，2路 4K (3840 x 2160) 75FPS • 支持H.264 / H.265 硬件编码，12路 1080P 30FPS，2路 4K (3840 x 2160) 50FPS • JPEG解码能力1080P 512FPS，编码能力1080P 256FPS，最大分辨率：16384 x 16384
高速接口	• 高速解串器 SerDes：8 lane，支持 PCIe 3.0 * 4/SGMII * 4/USB 3.0 *4/SATA 3.0 * 4 等接口 • 数据接口：RGMII * 2（与SGMII共用4个MAC）	
低速接口	• UART * 5 / I2C * 4 / SPI * 2 / CAN * 4	
典型功耗	• 25W	• 21W
工作环境	• -20℃ ~ 80℃（-4℉ ~ 176℉）	
结构尺寸	• 采用MXM连接器：82mm（长）* 60mm（宽）* 7mm（高）	



Atlas 200I A2（20 TOPS）



Atlas 200I A2（8 TOPS）

推理 Atlas 300V

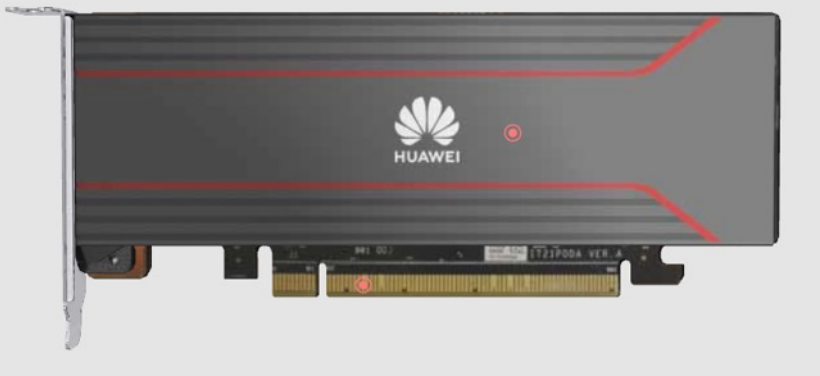
	Atlas 300V 视频解析卡	Atlas 300V Pro 视频解析卡
形态	• 单槽位半高半长PCIe卡	
AI算力	• 100 @ TOPS INT8 • 50 @ TFLOPS FP16	• 140 @ TOPS INT8 • 70 @ TFLOPS FP16
内存规格	• LPDDR4X 24GB，总带宽 204.8GB/s	• LPDDR4X 48GB，总带宽204.8GB/s
CPU算力	• 8 core * 1.9 GHz	
编解码	• 支持H.264 / H.265 硬件解码，100路 1080P 25FPS / 80路 1080P 30FPS / 10路 4K 60FPS • 支持H.264 / H.265 硬件编码，30路 1080P 25FPS / 24路 1080P 30FPS / 4路 4K 60FPS • JPEG解码能力4K 384FPS，编码能力4K 192FPS，最大分辨率：8192 x 8192	• 支持H.264硬件解码，128路 1080P 30FPS • 支持H.265硬件解码，128路 1080P 30FPS • 支持H.264硬件编码，24路 1080P 30FPS • 支持H.265硬件编码，24路 1080P 30FPS • JPEG解码能力4K 384FPS，编码能力4K 192FPS， 最大分辨率：8192 * 8192
PCIe接口	• PCIe x 16 Gen4.0	
最大功耗	• 72W	
工作温度	• 0℃ ~ 55℃（32℉ ~ 131℉）	
结构尺寸	• 169.5mm（长）x 18.45mm（宽）x 68.9mm（高）	

Atlas 300V Pro 视频解析卡

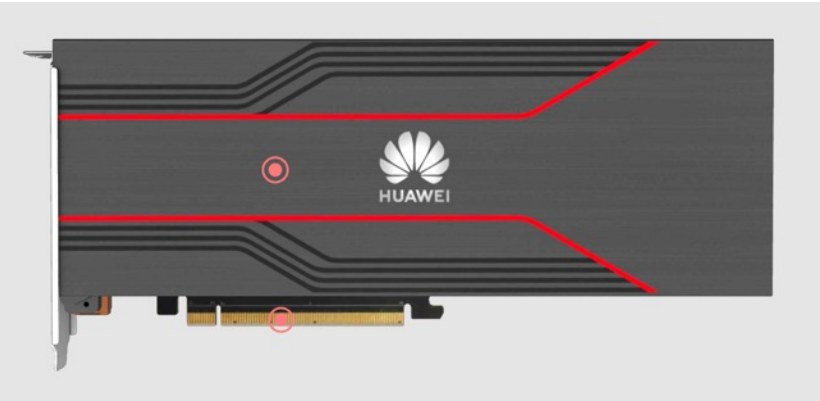
Atlas 300V Pro 视频解析卡

推理 Atlas 300I

	Atlas 300I Pro 推理卡	Atlas 300I Duo 推理卡
形态	• 单槽位半高半长PCIe卡	
AI算力	• 140 @ TOPS INT8 • 70 @ TFLOPS FP16	• 280 @ TOPS INT8 • 140 @ TFLOPS FP16
内存规格	• LPDDR4X 24 GB，总带宽204.8 GB/s	• LPDDR4X 48GB，总带宽 408GB/s
CPU算力	• 8 core * 1.9 GHz	
编解码	• H.264、H.265视频编解码 • JPEG图片编解码	• 支持H.264硬件解码，256路 1080P 30FPS（32路 3840 * 2160 60 FPS） • 支持H.265硬件解码，256路 1080P 30FPS（32路 3840 * 2160 60 FPS） • JPEG解码能力4K 768 FPS，编码能力4K 384 FPS
PCIe接口	• PCIe x 16 Gen4.0	
最大功耗	• 72W	
工作温度	• 0℃ ~ 55℃（32℉ ~ 131℉）	
结构尺寸	• 169.5mm（长）x 18.45mm（宽）x 68.9mm（高）	



Atlas 300I Pro 推理卡



Atlas 300I Duo 推理卡

Atlas 800 推理服务器

	Atlas 800 推理服务器（3000）	Atlas 800 推理服务器（3010）
CPU	• 鲲鹏920 * 2	• 1/2个Intel® Xeon® SP Skylake • Cascade Lake处理器，最高205W
CPU 内存	• 32个DDR4内存插槽，最高3200 MT/s	• 24个DDR4内存插槽，最高3200 MT/s
AI加速卡	• 最大支持8个Atlas 300I/V Pro	• 最大支持7个Atlas 300I/V Pro
AI算力	• 最大1120 TOPS INT8	• 最大980 TOPS INT8
RAID支持	• RAID 0/1/10/5/50/6/60等	• RAID 0/1/5/6/10/1E/50/60等
PCIe	• 最多支持9个PCIe4.0 接口，其中1个为RAID扣卡专用的PCIe扩展槽位，另外8个为标准的PCIe扩展槽位	• 10个PCIe Gen3.0接口（含1个RAID控制卡+1个灵活LOM）
工作温度	• 5℃ ~ 40℃ (41°F ~ 104°F)	• 5℃ ~ 45℃ (41°F ~ 113°F)
形态	• 2U AI服务器	• 2U AI服务器
结构尺寸	• 447mm * 790mm * 86.1mm	

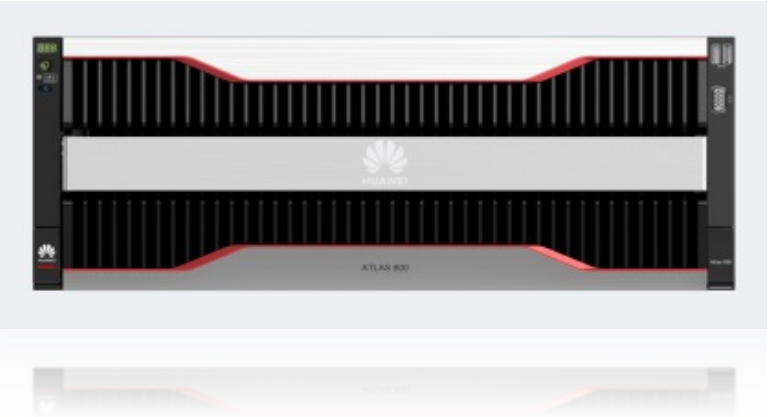
Atlas 800 推理服务器（3000）

Atlas 800 推理服务器（3010）

Atlas 800 训练服务器

	Atlas 500 Pro 智能边缘服务器	Atlas 800T A2训练服务器
CPU	• 1*鲲鹏920	• 4 * 鲲鹏920
CPU 内存	• 4个DDR4内存插槽，最高3200 MT/s	• 32个DDR4内存插槽，最高3200 MT/s
AI加速卡	• 最大支持3个Atlas 300I/V Pro 推理卡	
AI算力	• 最大420 TOPS INT8	
RAID支持	• RAID 1/5/6/10等	• 支持 RAID 0/1/10/5/50/6/60等
PCIe	• 最多4个PCIe 4.0 x8标准扩展槽位	• 最多支持3个PCIe 4.0扩展插槽
工作温度	• 长期：5℃ ~ 50℃ • 短期：0℃ ~ 55℃	• 5℃ ~ 35℃（41°F ~ 95°F）
形态	• 2U AI服务器	• 4U AI服务器
网络	• *10GE/25GE（光口）+2*GE（电口）	• 8 * 200GE QSFP接口直出，RoCE协议

Atlas 500 Pro 智能边缘服务器



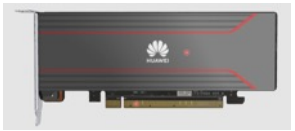
Atlas 800T A2训练服务器

训练推理系列硬件产品

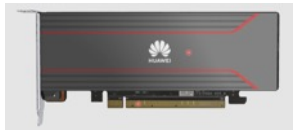
边缘端 **云端推理** **云端训练**

< 20 TFLOPS 1000 TFLOPS > 2000 TFLOPS

Atlas 200I A2 (20 TOPS)



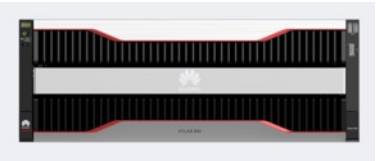
Atlas 300V Pro 视频解析卡



Atlas 300V Pro 视频解析卡

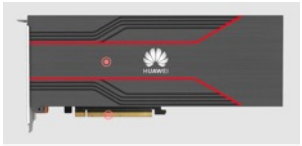
Atlas 800 推理服务器 (3000)

Atlas 500 Pro 智能边缘服务器



Atlas 800T A2训练服务器

Atlas 300I Pro 推理卡

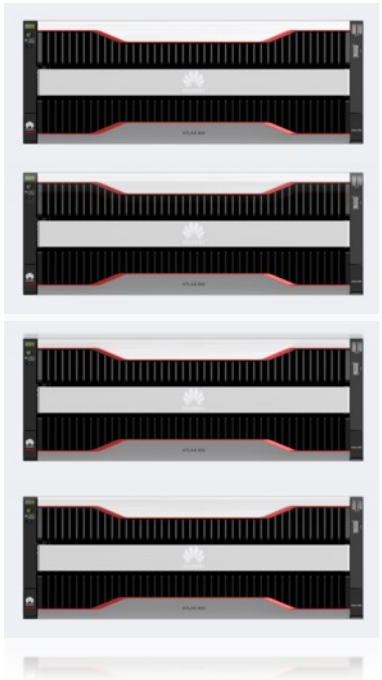


Atlas 300I Duo 推理卡

Atlas 800 推理服务器 (3010)

训练推理系列硬件产品

云端训练



训练推理系列硬件产品

云端训练

03 服务器硬件拆解介绍

Atlas 800 训练服务器

- https://support.huawei.com/enterprise/zh/ascend-computing/atlas-800-9000-a2-pid-254184887

04 小结与思考

思考

1. 那么多不同的训练和推理产品形态，对应的 AI 芯片到底有多少种？怎么组合？
2. 你还想了解昇腾哪些内容呢？

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem